
Correction for classical covariate measurement error and extensions to life-course studies

Jonathan William Bartlett



A thesis submitted to the University of London for the degree of
Doctor of Philosophy

London School of Hygiene and Tropical Medicine, 2010

Declaration

I, Jonathan Bartlett, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Measurement error in the covariates of regression models is a common problem in epidemiology, generally causing bias in estimates of covariate effects. In the first part of the thesis we examine the effects of, and methods to allow for, classical covariate measurement error in regression models for continuous, binary, and censored time-to-event outcomes. We describe the most commonly used estimation methods, including regression calibration (RC), maximum likelihood (ML), the conditional score method, multiple imputation (MI), and moment reconstruction.

We demonstrate how, for continuous and binary outcomes, MLEs for particular parametric specifications can be obtained by fitting a simple linear mixed model to the error-prone measurements. We also illustrate how MLEs for certain parametric specifications can be obtained by the Monte-Carlo Expectation Maximization (MCEM) algorithm. This includes a novel proposal for multiply imputing the covariate measured with error for Cox proportional hazards outcome models using rejection sampling. Simulations are used to compare the performance of the methods.

In the second part of the thesis we consider the extension of these methods to life-course studies. We show that our proposal for ML estimation in the case of classical covariate measurement error and the MCEM algorithm can both be extended to this more general setting. In applications we typically do not know which aspects of a longitudinal trajectory influence the outcome of interest, leading us to fit a number of different models. We show that naive application of RC gives biased parameter estimates when important aspects of the longitudinal trajectory are omitted from the outcome model. We show how multiple imputations, which are a by-product of the MCEM algorithm, may be used to obtain consistent estimates in such situations.

In the final part of the thesis we use data from the Framingham Heart Study to illustrate the application of RC and our proposed estimation approaches.

Acknowledgements

I am tremendously grateful to my two supervisors, Chris Frost and Bianca De Stavola, for their advice, guidance, and huge investment of time in me. I am also grateful to the members of my Advisory Committee, Ian White, Angela Wood, and Dave Leon, for their helpful advice and suggestions.

Thank you to Daniela, my partner, and my parents, David and Evelyn, for their love and support.

I would like to thank the Medical Research Council for providing my PhD studentship.

Lastly, I also would like acknowledge the contribution made by the participants and investigators of the Framingham Heart Study. The Framingham Heart Study - Cohort (FHS-C) is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with the FHS-C Study Investigators. This manuscript was prepared using a limited access dataset obtained from the NHLBI and does not necessarily reflect the opinions or views of the FHS-C or the NHLBI.

Acronyms

CS Conditional score

CHD Coronary heart disease

CVD Cardiovascular disease

EM Expectation Maximization

MAR Missing at random

MCAR Missing completely at random

MCEM Monte-Carlo Expectation Maximization

MI Multiple imputation

ML Maximum likelihood

MNAR Missing not at random

MOM Method of moments

RC Regression calibration

RRC Risk-set regression calibration

SD Standard deviation

Contents

Declaration	1
Abstract	2
Acknowledgements	3
Acronyms	4
1 Introduction	13
1.1 Classical covariate measurement error in regression models	13
1.2 Extensions to life-course studies	15
I Classical covariate measurement error	19
2 Background	20
2.1 The outcome model	20
2.2 The measurement model	21
2.3 Non-differential measurement error	25
2.4 Validation data	25
2.5 Replication data	26
3 Continuous outcomes	31
3.1 Linear regression	32
3.2 The effects of classical covariate measurement error	33
3.3 Method of moments correction	36
3.4 Regression calibration	39
3.5 Maximum likelihood	47
3.6 Maximum likelihood estimation using linear mixed models	53
3.7 Multiple imputation	61
3.8 Moment reconstruction	67
3.9 Simulations	71
3.10 Conclusions	76

4	Binary outcomes	80
4.1	Logistic regression	81
4.2	The effects of classical covariate measurement error	81
4.3	Regression calibration	83
4.4	Maximum likelihood	85
4.5	Ascent-based Monte-Carlo Expectation Maximization	92
4.6	Maximum likelihood estimation using linear mixed models	99
4.7	Multiple imputation	103
4.8	Moment reconstruction	104
4.9	Conditional score method	104
4.10	Simulations	106
4.11	Conclusions	113
5	Survival outcomes	116
5.1	Cox’s proportional hazards model	117
5.2	The effects of classical covariate measurement error	119
5.3	Regression calibration	122
5.4	Maximum likelihood	125
5.5	Ascent-based Monte-Carlo Expectation Maximization	129
5.6	Multiple imputation	134
5.7	Conditional and corrected score methods	136
5.8	Simulations	137
5.9	Conclusions	141
II	Extensions to life-course studies	144
6	Background	145
6.1	Linear mixed models for longitudinal error-prone measurements	146
6.2	Outcome model specification	149
6.3	Missing longitudinal error-prone measurements	151
7	Continuous outcomes	152
7.1	A naive two-stage approach	153
7.2	Regression calibration	154
7.3	Maximum likelihood	155
7.4	Maximum likelihood estimation using linear mixed models	157
7.5	Multiple imputation	158
7.6	Alternative outcome model covariate specifications	159
7.7	Simulations	164
7.8	Conclusions	167

8	Binary outcomes	172
8.1	Regression calibration	172
8.2	Maximum likelihood	173
8.3	Maximum likelihood using standard linear mixed models	175
8.4	Multiple imputation	175
8.5	Conditional score method	176
8.6	Alternative outcome model covariate specifications	178
8.7	Simulations	184
8.8	Conclusions	190
9	Survival outcomes	193
9.1	Overview	193
9.2	Modelling assumptions and notation	197
9.3	Regression calibration	199
9.4	Maximum likelihood	204
9.5	Ascent-based Monte-Carlo Expectation Maximization	208
9.6	Multiple imputation	211
9.7	Conditional score method	212
9.8	Simulations	214
9.9	Modelling assumptions revisited	221
9.10	Conclusions	224
III Systolic blood pressure and cardiovascular disease - analyses of data from the Framingham Heart Study		228
10	The Framingham Heart Study	230
10.1	Study recruitment procedure	231
10.2	Characteristics at the initial examination	231
10.3	All-cause mortality	232
10.4	Death due to cardiovascular disease	233
11	Systolic blood pressure measurements in men	236
11.1	Systolic blood pressure measurements in the Framingham study	236
11.2	Data reduction	240
11.3	Classical measurement error assumptions	242
12	Risk of death due to cardiovascular disease between age 70 and 80 and its relationship with systolic blood pressure levels in earlier life	244
12.1	Longitudinal model	245
12.2	Estimation methods	250

12.3 Results	254
12.4 Conclusions	257
13 The effects of current and past systolic blood pressure on the hazard of cardiovascular disease	262
13.1 Longitudinal model	263
13.2 Estimation methods	265
13.3 Results	268
13.4 Conclusions	271
IV Conclusions	273
14 Conclusions	274
14.1 Classical covariate measurement error	274
14.2 Extensions to life-course studies	277
14.3 Analyses of data from the Framingham Heart Study	282
14.4 Future research	285
Appendix: R code	287

List of Figures

9.1	Diagrammatic representation of three situations in which longitudinal and time-to-event data are observed	195
10.1	Framingham Heart Study: Kaplan-Meier curve for all-cause mortality survival using data from 5,079 subjects	233
10.2	Framingham Heart Study: estimated hazard function for all-cause mortality using data from 5,079 subjects	234
10.3	Framingham Heart Study: estimated cumulative incidence of death due to CVD, using data from 5,078 subjects	235
10.4	Framingham Heart Study: estimated cause-specific hazard function for death due to CVD, using data from 5,078 subjects	235
11.1	Framingham Heart Study: within-subject SD versus mean for SBP measurements from visits 1 and 2, based on data from 1,099 men . .	242
11.2	Framingham Heart Study: histogram of difference in SBP measurements from visits 1 and 2, based on data from 1,099 men	243
12.1	Framingham Heart Study: Histogram of SBP measurements at age 60, based on data from 779 men	247
12.2	Framingham Heart Study: Estimated population mean evolution of SBP, the SBP measurements for a randomly chosen man, and the best linear unbiased prediction of his SBP trajectory	249
12.3	Framingham Heart Study: Estimated population mean evolution of SBP, the SBP measurements for a second randomly chosen man, and the best linear unbiased prediction of his SBP trajectory	250

List of Tables

3.1	Linear regression simulation results with normally distributed covariate: bias and variability	74
3.2	Linear regression simulation results with normally distributed covariate: confidence interval coverage	75
3.3	Linear regression simulation results with log-normally distributed covariate: bias and variability	76
3.4	Linear regression simulation results with log-normally distributed covariate: confidence interval coverage	77
4.1	Logistic regression simulation results with normally distributed covariate: bias and variability	109
4.2	Logistic regression simulation results with normally distributed covariate: confidence interval coverage	111
4.3	Logistic regression simulation results with covariate conditionally normal given outcome: bias and variability	112
4.4	Logistic regression simulation results with covariate conditionally normal given outcome: confidence interval coverage	112
5.1	Cox regression simulation results with normally distributed covariate	140
7.1	Linear regression simulation results with longitudinal error-prone measurements – adjusted effect estimates	167
7.2	Linear regression simulation results with longitudinal error-prone measurements – unadjusted effect estimates assuming conditional independence	168
7.3	Linear regression simulation results with longitudinal error-prone measurements – unadjusted effect estimates not assuming conditional independence	169
8.1	Logistic regression simulation results with longitudinal error-prone measurements – adjusted effect estimates	188
8.2	Logistic regression simulation results with longitudinal error-prone measurements – unadjusted effect estimates assuming conditional independence	189

8.3	Logistic regression simulation results with longitudinal error-prone measurements – unadjusted effect estimates not assuming conditional independence	190
9.1	Cox regression simulation results with longitudinal error-prone measurements – estimates of adjusted effects	219
9.2	Cox regression simulation results with longitudinal error-prone measurements – estimates of unadjusted effects	220
10.1	Framingham Heart Study: Descriptive statistics for 5,079 Framingham subjects at entry to study	232
11.1	Framingham Heart Study: Number of systolic blood pressure measurements at follow-up visits 1-13 for men in the Framingham study .	238
11.2	Framingham Heart Study: Number of systolic blood pressure measurements at follow-up visits 14-26 for men in the Framingham study	239
11.3	Framingham Heart Study: Missing follow-up visits by age for 2,294 men in the Framingham study	240
11.4	Framingham Heart Study: Mean and SDs of systolic blood pressure (mmHg) measurements from visit 2 of the Framingham study, based on data from 1,895 men who had all three measurements available . .	241
11.5	Framingham Heart Study: Estimated correlations between systolic blood pressure measurements from visit 2 of the Framingham study, based on data from 1,895 men who had all three measurements available	241
12.1	Framingham Heart Study: Values of covariates used to parametrize the linear mixed model for longitudinal SBP measurements at 5 yearly intervals from age 40 to 70	246
12.2	Framingham Heart Study: Estimates of fixed effect parameters and random-effects parameters from linear mixed model for SBP measurements	247
12.3	Framingham Heart Study: Estimated correlations between random-effects in linear mixed model for SBP measurements	248
12.4	Framingham Heart Study: Estimates of mutually adjusted odds ratios for the effects of a 10mmHg increase in SBP at ages 40, 50, 60 and 70 on odds of death due to CVD	256
12.5	Framingham Heart Study: Estimates of unadjusted odds ratios for the effects of a 10mmHg increase in SBP at ages 40, 50, 60 and 70 on odds of death due to CVD	258
12.6	Framingham Heart Study: Estimates of mutually adjusted odds ratios for for the effects of a 10mmHg increase in SBP at ages 40 and 70 on odds of death due to CVD	259

13.1 Framingham Heart Study: Estimates of fixed effect parameters and random-effects parameters from linear mixed model for SBP measurements	265
13.2 Framingham Heart Study: Estimated correlations between random-effects in linear mixed model for SBP measurements	265
13.3 Framingham Heart Study: Estimates of mutually adjusted hazards ratios for the effects of SBP at ages t , $t - 10$, and $t - 20$ on hazard of CVD at age t	269
13.4 Framingham Heart Study: Estimates of unadjusted hazards ratios for the effects of SBP at ages t , $t - 10$, and $t - 20$ on hazard of CVD at age t	270
13.5 Framingham Heart Study: Estimates of mutually adjusted hazards ratios for the effects of SBP at age t and $t - 20$ on hazard of CVD at age t	271

Chapter 1

Introduction

1.1 Classical covariate measurement error in regression models

Regression models are the workhorses of statistics. Typically, a regression model specifies how the conditional distribution of an outcome of interest depends on one or more variables, which are referred to as covariates. This dependence is usually characterized by a small number of parameters. Given a suitable sample of data on the outcome and covariates, these parameters can be estimated, the method of maximum likelihood being the most popular approach.

Such regression models by default assume that the values of the covariates are known exactly, that is they are perfect observations of the variable we wish to include in the model as a covariate. Often the assumption that covariates are measured without error is valid or reasonable - gender and age at study entry would be examples in most settings. However, sometimes the value we observe can be considered an error-prone measurement of the underlying covariate of interest. The difference between the value we observe and the ‘true’ value of the covariate is termed measurement error. In certain situations it may be impossible to ever observe the value of the underlying covariate, and we only have information about it through error-prone measurements. The reasons observed values differ from the underlying ‘true’ value obviously depend on the covariate in question. For example, a single blood pressure measurement will differ from the person’s mean value (defined over some period) because of both instrument error and because of genuine temporal biological variation around the mean blood pressure for that period.

Ignoring such covariate measurement error generally results in biased estimates of the parameters of the regression model. In the simplest case of a continuous outcome variable and a single continuous covariate measured with classical non-differential error, the effect of measurement error is to attenuate estimates of the slope towards the null value, so that the association between the outcome and covariate appears less strong. Such attenuation was termed ‘regression dilution’ by MacMahon *et*

al [1], since the magnitude of associations are diluted towards the null. However, once the regression model has two or more covariates, it is generally not the case that parameter estimates are always attenuated [2], and so the term ‘regression dilution’ has the potential to mislead.

Errors in the measurement of covariates are common in epidemiological studies. Examples include dietary epidemiology (e.g. dietary records as error-prone measures of true intake of particular nutrients [3]), cardiovascular epidemiology (e.g. single measurements of biological risk factors such as blood pressure as measurements of an ‘underlying’ level [1]), and environmental epidemiology (e.g. radiation exposure as a result of atomic bombs, estimated on the basis of a person’s location and shielding [4]). While it is often recognised that measurement error in the exposure of interest causes bias, it is perhaps less appreciated that measurement error in confounders also in general results in biased estimates of exposure effects [5]. If a confounder is measured with error, the resulting estimates for exposure effects are only partially adjusted for the confounder.

In response, during the 20th century much research was conducted into both the effects of covariate measurement error and methods to allow for these effects. The initial focus was given to covariate measurement error in linear regression models for continuous outcomes [6, 7]. As use of generalized linear and survival models has become widespread in the the last thirty years, research efforts have correspondingly addressed the effects of covariate measurement errors in such models [8].

Recognition of the effects of covariate measurement error in epidemiology led to the application of some of these methods, with the aim of obtaining unbiased estimates of the effects of exposures of interest. Examples include correction for measurement error in dietary exposures in relation to breast cancer risk [3] and coronary heart disease risk [9]. Other studies have investigated correction for measurement error and temporal variability in measurements of blood pressure [1], cholesterol [10], and recently homocysteine [11] and fibrinogen [12].

Arguably, covariate measurement error pervades epidemiology to a far greater extent than is implied by the relatively limited application of methods to deal with covariate measurement errors. This is likely due to a number of factors. A large number of different approaches to dealing with covariate measurement error have been proposed in the statistical literature. Although some are simple to use, such as the so called method of moments correction method [13], many are complex to implement, which likely acts as a barrier to their widespread application in epidemiology. Linked to this is the relative scarcity of the implementation of such methods in standard statistical software packages. As there are now so many different methods for dealing with covariate measurement error, it may be difficult for researchers to decide which is the most appropriate to use. Another reason for the limited use

of measurement error correction methods may be lack of appropriate data which are needed to make a correction.

In Part I of this thesis we review the statistical approaches which have been proposed for dealing with classical measurement error in continuous covariates. In Chapter 2 we describe the general setting for the following chapters, which includes defining classical and Berkson errors, the distinction between measurement error and misclassification in categorical covariates, the concept of non-differential errors, and the types of data which can be used to estimate the magnitude of measurement errors. We focus on the situation in which the exact value of the underlying covariate(s) cannot be observed, but replicate error-prone measurements are available. In Chapters 3, 4, and 5 we deal with continuous, binary, and censored survival time outcomes respectively. For each outcome type, we review the known results for the effects of classical covariate measurement error on parameter estimates. We then describe some of the estimation methods which have been proposed, and show how they relate to each other. We demonstrate how, for continuous and binary outcomes, maximum likelihood (ML) estimates can be obtained, under certain parametric assumptions, by fitting a standard linear mixed model to the error-prone measurements of covariates. We also show how the Monte-Carlo Expectation Maximization (MCEM) algorithm can be used to obtain ML estimates. This includes a novel proposal for how rejection sampling can be used to multiply impute the covariate measured with error when the outcome model is Cox's proportional hazards model. For each outcome type, we compare the performance of the various estimation methods through simulation, and give details of their availability in statistical software packages, with the aim of aiding researchers in both choosing an appropriate method and giving guidance regarding their implementation.

1.2 Extensions to life-course studies

1.2.1 Motivation

In recent years there has been interest in estimating the associations between the levels of putative risk factors during life with subsequent development of disease. In 1990, MacMahon *et al* [1] published a paper examining the associations of diastolic blood pressure with stroke and coronary heart disease (CHD) from nine large prospective studies. This was the first study in which an attempt was made to estimate associations between what MacMahon *et al* termed the 'usual' level of a variable (diastolic blood pressure (DBP)) and an outcome (incidence of stroke or CHD). MacMahon *et al* described how most epidemiological studies relate a single baseline measurement of a risk factor to subsequent incidence of disease, but that

“because such measurements are subject to substantial random fluctuations (due partly to the measurement process and partly to some real but temporary deviations from an individual’s usual DBP), uncorrected use of just the baseline DBP can result in systematic and substantial underestimation of the strength of the real association of disease with usual DBP (the “regression dilution”).”

One of the included studies (the Framingham Heart Study [14]) included both baseline and repeat measurements of DBP made every two years post-baseline. The repeat measurements were used to estimate a correction factor which was used to correct the attenuated associations obtained from using a single baseline DBP measurement, which was claimed to result in estimates of the association between ‘usual’ DBP and incidence of stroke or CHD. Such an approach implicitly assumes that each individual has an underlying constant blood pressure throughout their lifetime and that it is only this level that is related to risk of stroke or CHD. Measurements of their blood pressure differ from the underlying level because of measurement error and because of “temporary deviations from an individual’s usual DBP”. The correction for measurement error resulted in relative risk estimates that were increased by around 60% in the study by MacMahon *et al*. The authors thus concluded that diastolic blood pressure makes a larger contribution to risk of stroke and CHD than had previously been thought from studies which did not allow for the effects of measurement error and within-subject variability over time.

If individuals do have a single ‘usual’ level of blood pressure over their adult lifetime, the correction factor for measurement error resulting from repeat measurements made at varying times since a baseline measurement would be the same, apart from sampling variability, irrespective of the time interval between the two measurements. However, the results of MacMahon *et al* using repeat DBP measurements made at two and four years post-baseline from the Framingham Study indicated that a larger correction for measurement error would be needed if repeat measurements four years post-baseline were used instead of two years post-baseline, although this was not discussed in their paper. If the correction factors vary with the time interval between baseline and repeat measurements, this means that individuals do not have a constant underlying DBP level, casting doubt on the magnitude of the correction made by McMahon and colleagues.

In 1999 Clarke *et al* [15] reported further results for measurement error correction in cohort studies, using data from the Framingham and Whitehall studies. They examined how the magnitude of the correction for measurement error in risk factors varied as the interval between baseline and repeat measurements increased. Their results confirmed that the magnitude of the correction in parameter estimates increased as the interval between baseline and repeat measurements increased, for systolic and diastolic blood pressure, and blood cholesterol. Since the amount by

which uncorrected estimates are corrected by depended on the time interval between the baseline and repeat measurements, Clarke and colleagues proposed how an appropriate correction factor should be obtained:

“To assess the relevance of the usual levels of a risk factor during some particular exposure period (e.g. the second decade of follow-up) to disease risk, correction factors may need to be based on remeasurements made after an interval approximately equivalent to the midpoint of the relevant period (e.g. after about 15 years of follow-up).”

Clarke *et al* thus implicitly acknowledged that a risk factor such as DBP may undergo long term changes throughout an individual’s lifetime, with a different usual level in each decade of exposure:

“uncorrected associations of disease risk with baseline measurements underestimate the strength of the real associations with usual levels of these risk factors during the first decade of exposure by about one-third, the second decade by about one-half, and the third decade by about two-thirds.”

Recently Frost and White have shown that such corrections for measurement error and within-subject variability do not give unbiased estimates of a meaningful parameter, unless current risk of the outcome of interest depends only on the risk factor’s current level, and earlier levels of the risk factor have no independent effect [16].

1.2.2 A life-course approach

It is increasingly recognised that when interest lies in estimating the associations between risk factors which vary over time with subsequent disease, statistical methods should explicitly account for such variation. The term ‘life-course epidemiology’ has been used to refer to the study of the long term effects of physical and social exposures at different points in life on subsequent disease [17]. A life-course approach involves consideration of the inter-relationships between different exposure variables, in addition to the trajectories over time of individual exposures. Exposures many years previous to disease incidence may be of interest, including those occurring during “gestation, childhood, young adulthood and later adult life” [17] or even inter-generational exposures. From a practical perspective, answering such questions first requires that longitudinal studies collect sufficient information regarding the longitudinal trajectories of exposures of interest, by making repeated measurements on subjects over time.

The statistical modelling of such data presents a number of methodological challenges, as discussed by De Stavola *et al* [18]. First, appropriate models must be specified for the longitudinal measurements which allow for the evolution of the exposures and confounders of interest to differ between subjects. Such models must

allow both for the correlations between levels of a particular exposure over time, as well as for correlations between levels of different variables. Often the times at which longitudinal measurements are made differs between subjects, leading to unbalanced data. Similarly some subjects may miss scheduled measurements at particular time points, for a variety of reasons, leading to missing data.

In response to such challenges, over the last 15 years there has been a large amount of research effort into statistical modelling and estimation approaches for such studies. By viewing the aspects of longitudinal trajectories thought to influence the outcome of interest as the outcome model's true covariates, methods originally conceived for dealing with classical measurement error have been extended to the life-course/longitudinal setting. This has resulted in an array of different estimation approaches. However, analogous to the simpler case of classical measurement error, these statistical methods have had less penetration into applied work than might be expected. As for classical covariate measurement error, this is likely due to the perceived complexity and lack of availability in statistical software packages for such methods.

In Part II of the thesis we consider the extension of the methods described for dealing with classical covariate measurement error to such life-course settings. In Chapter 6 we set up a framework for our subsequent developments in which the longitudinal error-prone measurements are assumed to follow a linear mixed model. In Chapters 7, 8, and 9 we consider estimation in the case of continuous, binary, and censored survival time outcomes. We show that our earlier proposals for ML estimation can be extended to the life-course setting. We again use simulations to compare the performance of some of the estimation methods, and discuss their relative merits in terms of bias, efficiency, and ease of use.

In applications we typically do not know which aspects of a longitudinal trajectory influence the outcome of interest, leading us to fit a number of different models for how the outcome depends on the longitudinal measurements. We discuss the implications of this, and show why naive application of one estimation method (regression calibration) to fit a number of different models may result in biased estimates, and we propose approaches to overcome this.

In Part III we use data from the Framingham Heart Study to illustrate the application of our proposed estimation approaches, and compare the results with those obtained using a simpler approach. We consider models which relate the risk of cardiovascular disease (CVD) to both current and earlier levels of systolic blood pressure (SBP).

We summarize our contributions and conclusions in Chapter 14, and outline areas for future research.

Part I

Classical covariate measurement error

Chapter 2

Background

In this chapter we describe the general setting for our investigation of classical covariate measurement error. We first outline the basic building blocks for a model which incorporates covariate measurement error. We describe the classical measurement error model, and contrast it with Berkson error. We introduce the concept of non-differential measurement error, an assumption on which many methods which allow for covariate measurement error rely. Lastly we consider the typical data sources from which information on measurement errors can be obtained and discuss estimation of measurement error model parameters.

2.1 The outcome model

We denote by Y_i the outcome of interest for subject i . Our primary interest lies in estimating the parameters of a regression model relating Y_i to one or more covariates. We call this the *outcome model*. We divide the outcome model covariates into those which are observed with measurement error, $\mathbf{X}_i = (X_{i1}, X_{i2}, \dots, X_{ip})^T$, and those observed without error, denoted $\mathbf{Z}_i = (Z_{i1}, Z_{i2}, \dots, Z_{iq})^T$. We sometimes refer to \mathbf{X}_i as the true, or underlying covariate, in contrast to an error-prone measurement of it.

In this thesis, we focus on the case where the components of \mathbf{X}_i are continuous. In Section 2.2.4 we explain some of the differences between measurement error and misclassification, and the impact this has on methods of correction. The outcome model defines how \mathbf{X}_i and \mathbf{Z}_i are assumed to influence the distribution of the outcome Y_i . Often this involves specifying how the conditional distribution of Y_i depends on \mathbf{X}_i and \mathbf{Z}_i , but this is not necessarily the case. For example in a linear regression outcome model we may specify only the conditional mean function, without specifying the distribution of the residual errors.

We assume the effects of \mathbf{X}_i and \mathbf{Z}_i on Y_i are parametrized by β_X and β_Z respectively. If Y_i , \mathbf{X}_i and \mathbf{Z}_i were observed, we assume that estimation of β_X and β_Z could be carried out by some method, such as maximum likelihood.

2.1.1 Covariates fixed or random

Covariates are usually treated as fixed in repeated sampling in the context of regression models. In some cases, this is appropriate because the covariates represent design features which would be fixed in repeated sampling. However, in many studies, such as in observational epidemiology, subjects are ideally randomly sampled from the population of interest. The observed values of covariates are then random samples from the covariate distribution in the population. For example, in an epidemiological study of the relationship between blood pressure and risk of coronary heart disease, we might measure blood pressure once at the beginning of the study in each subject. If we were to repeat the study by sampling a new set of study subjects, we would obtain a new set of covariate values X_i , which are samples from the distribution of true blood pressures in the population from which we have sampled.

In such cases the covariates might be more appropriately treated as random for the purposes of estimation and inference. However, standard practice is to treat covariates as fixed, even in situations where the covariates are random. The frequentist justification for this practice is as follows. If the covariates are random variables, then their distribution in the population is characterized by a set of parameters. If these parameters are variationally independent from the outcome model parameters, we lose no information by basing estimation and inference on the conditional distribution $f(Y_i|\mathbf{X}_i)$ [19].

Covariate measurement error means that \mathbf{X}_i is usually unobserved. In principle, we can proceed by treating the \mathbf{X}_i as unknown, but fixed constants, or as unobserved values which are random draws from the covariate distribution in the population. Methods to allow for covariate measurement error which are based on treating \mathbf{X}_i as fixed unknown parameters in most cases result in inconsistent estimators (see page 151 of [8]). This is perhaps unsurprising, given the fact that the usual properties of maximum likelihood estimators do not hold when the number of parameters is of the same order of magnitude as the sample size [20]. For this reason, we treat \mathbf{X}_i throughout the thesis as random.

2.2 The measurement model

The covariates \mathbf{X}_i are not, in general, observed on subjects. Instead, each subject has one or more error-prone measurements of \mathbf{X}_i . The *measurement model* specifies how the observed error-prone measurements are related to the unobserved \mathbf{X}_i . Carroll *et al* (page 26 of [8]) distinguish between two broad alternative approaches to this specification. In the first, we specify how the distribution of \mathbf{W}_i depends on \mathbf{X}_i . This approach includes the classical measurement error model (Section 2.2.1). The alternative is the so called ‘regression calibration model’ approach, whereby we

specify the conditional distribution of \mathbf{X}_i given \mathbf{W}_i and if present, \mathbf{Z}_i . This approach includes the so called Berkson error model (Section 2.2.2).

2.2.1 Classical measurement error

Single covariate

We first define classical measurement error in the case of an error-prone measurement of a scalar covariate X_i . For a single error-prone measurement W_i of X_i , the measurement error in W_i is classical if

$$W_i = X_i + U_i, \tag{2.1}$$

such that $\mathbb{E}(U_i|X_i) = 0$ (\mathbb{E} denotes expectation), so that the measurements are unbiased for X_i . This means that $\mathbb{E}(U_i) = \mathbb{E}(\mathbb{E}(U_i|X_i)) = 0$ and that $\text{Cov}(X_i, U_i) = 0$, since:

$$\begin{aligned} \text{Cov}(X_i, U_i) &= \mathbb{E}((X_i - \mu_X)(U_i - 0)) \\ &= \mathbb{E}(X_i U_i) - \mu_X \mathbb{E}(U_i) \\ &= \mathbb{E}(\mathbb{E}(X_i U_i | X_i)) \\ &= \mathbb{E}(X_i \mathbb{E}(U_i | X_i)) \\ &= 0 \end{aligned}$$

where μ_X denotes the mean of X_i . The fact that X_i and U_i are uncorrelated means that the variance of the error-prone measurements W_i is given by

$$\text{Var}(W_i) = \sigma_X^2 + \sigma_U^2 \tag{2.2}$$

where σ_X^2 and σ_U^2 are the variances of X_i and U_i respectively. The reliability coefficient of W_i is defined as the ratio of the variance of X_i to the variance of W_i :

$$\lambda = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} \tag{2.3}$$

In particular applications we may wish to make the stronger assumption that X_i and U_i are independent. An example of non-independence between X_i and U_i is where the variance $\text{Var}(U_i|X_i)$ depends on X_i .

Extensions of the classical error model defined here have been used in a number of applications. For example, Freedman *et al* considered a measurement error model whereby error-prone measurements $W_i = \gamma_0 + \gamma_X X_i + U_i$ [21]. By setting $\gamma_0 = 0$ and $\gamma_X = 1$, we see that the standard classical error model is a special case of this more general error model.

Multiple covariates

In the case of multivariate \mathbf{X}_i , a single (multivariate) error-prone measurement \mathbf{W}_i of \mathbf{X}_i is subject to classical error if:

$$\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i \quad (2.4)$$

where $\mathbb{E}(\mathbf{U}_i|\mathbf{X}_i) = \mathbf{0}$. Again this implies $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$ and $\text{Cov}(\mathbf{X}_i, \mathbf{U}_i) = \mathbf{0}$. We denote the variance covariance matrix of \mathbf{U}_i by Σ_U . This covariance matrix need not be diagonal. Non-zero off diagonal elements then correspond to correlation between the measurement errors of a single (multivariate) measurement of \mathbf{X}_i . As an example, suppose \mathbf{X}_i corresponds to a subject's true systolic and diastolic blood pressure. Then it may be the case that in a single error-prone blood pressure measurement, if the error in the systolic measurement is positive it is more likely that the error in the diastolic measurement is also positive. Hereafter we denote the mean vector of \mathbf{X}_i by $\boldsymbol{\mu}_X$ and the variance covariance matrix by Σ_X .

2.2.2 Berkson measurement error

The defining feature of classical measurement error is that the error-prone measurements are equal to the true value X_i plus a measurement error, which, if not independent of X_i , is at least uncorrelated with X_i . An alternative measurement error model is the Berkson model [22], which assumes that:

$$X_i = W_i + U_i \quad (2.5)$$

where $\mathbb{E}(U_i|W_i) = 0$. This means that W_i and U_i are uncorrelated. This is in contrast to classical measurement error, in which the error-prone measurement is correlated with the measurement error. One consequence of Berkson error is that the true values X_i have a larger variance than the error-prone measurements. A further consequence of Berkson error is that $\mathbb{E}(X_i|W_i) = W_i$, whereas for classical error $\mathbb{E}(X_i|W_i) \neq W_i$. This means that for a linear regression outcome model, Berkson error does not causes bias in parameter estimates.

The Berkson model is a special case of a more general regression calibration model proposed by Rosner *et al* [2]:

$$\mathbf{X}_i = \alpha_0 + \alpha_Z \mathbf{Z}_i + \alpha_W \mathbf{W}_i + \mathbf{U}_i,$$

where $\mathbb{E}(\mathbf{U}_i|\mathbf{Z}_i, \mathbf{W}_i) = \mathbf{0}$ (Rosner *et al* [2] made the additional assumption that $\mathbf{U}_i \sim N(\mathbf{0}, \Sigma_U)$ where Σ_U is an unstructured variance covariance matrix).

2.2.3 Classical or Berkson?

The type of measurement model which we assume has major implications for the consequences of the covariate measurement error in our outcome model of interest. It is therefore important in any given application to decide if the covariate measurement errors are classical or Berkson in nature. One of the features which often distinguishes between classical or Berkson error is whether replicate measurements can, if only in principle, be made. Measurement errors are usually classical if the error-prone measurements are unique to an individual, and if, at least in principle, replicate measurements of the underlying value can be made on the same individual. An example of classical measurement error is blood pressure measurements made on individuals, where X_i may represent average blood pressure over a particular time period, and W_{ij} (with j indexing repeated error-prone measurements) are replicate measurements of blood pressure.

Berkson errors are typically found when the same value W_i is assigned to groups of subjects, with their individual level X_i varying as a function of W_i . Examples include radiation exposure studies where for each individual, X_i is estimated based on characteristics such as location and number of hours of exposure. In this case, subjects in the same location are assigned the same value of W_i , about which their true values vary. A further example is designed experiments, in which the same dosage W_i of a treatment is applied to a group of subjects or units. Here, X_i represents the actual dosage received by the i th subject, which is a function of the applied dosage and other factors. For further discussion of the distinction between classical and Berkson error, we refer to the monograph by Carroll *et al* [8].

2.2.4 Misclassification

We now briefly describe the problem of covariates subject to misclassification, and contrast this with the situation of classical measurement error. Recall that for classical measurement error, an assumption of independence (or a weaker assumption of zero correlation) is made between the true covariate and the measurement error. For discrete covariates which are subject to misclassification this assumption cannot hold. Suppose that a binary covariate X_i is subject to misclassification, giving a measurement $W_i = X_i + U_i$. If $X_i = 1$, U_i must be equal to 0 or -1. If $X_i = 0$ then U_i must be 0 or 1. Thus U_i is negatively correlated with X_i , violating one of the assumptions of the classical error model. This correlation must be accounted for when allowing for the effects of misclassification, as shown by White *et al* [23]. In particular, applying methods developed for allowing for measurement error in continuous X_i will result in biased estimates when X_i is in fact a discrete variable subject to misclassification.

2.3 Non-differential measurement error

The consequences of, and methods to deal with, covariate measurement error usually depend on whether there is a relationship between the outcome Y_i and the measurement errors in error-prone measurements. The measurement error in \mathbf{W}_i as a measurement of \mathbf{X}_i is said to be non-differential if Y_i is conditionally independent of \mathbf{W}_i , given \mathbf{X}_i and \mathbf{Z}_i . Using $f(\cdot|\cdot)$ to denote a probability density function, where the arguments make clear the density we are referring to, the non-differential error-assumption means:

$$f(Y_i|\mathbf{W}_i, \mathbf{X}_i, \mathbf{Z}_i) = f(Y_i|\mathbf{X}_i, \mathbf{Z}_i) \quad (2.6)$$

Non-differential error means that all the information in \mathbf{W}_i about Y_i is contained in \mathbf{X}_i . An equivalent expression of the assumption of equation (2.6) is that:

$$f(\mathbf{W}_i|Y_i, \mathbf{X}_i, \mathbf{Z}_i) = f(\mathbf{W}_i|\mathbf{X}_i, \mathbf{Z}_i) \quad (2.7)$$

It is perhaps easiest to think of examples of differential error when Y_i is a binary variable from a case-control study, representing whether subjects are a case $Y_i = 1$ or a control $Y_i = 0$. Typically in case-control studies measurements of exposure are made retrospectively, for example by asking patients questions about previous exposure to a variable of interest. One example of differential error in such a context would be where the classical measurement errors of the exposure of interest have a larger variance in cases than in controls. A further example, often known as ‘recall bias’, would be where measurement errors have a positive bias in cases but not in controls.

2.4 Validation data

With a few exceptions, such as Berkson measurement error in linear regression, covariate measurement error causes bias in the parameter estimates of the outcome model of interest, and so a vast array of methods have been developed for allowing for the effects of covariate measurement error in regression models. All methods are based on an assumed measurement model, which depends on certain parameters. For example, the basic classical measurement error model of equation (2.1) depends on parameters σ_X^2 and σ_U^2 . To allow for covariate measurement error, these parameters must either be known, or be estimable from data. In most cases parameters are not known (although exceptions do exist, see 1.6.9 of [8]), and so we must estimate these. We can estimate these parameters using either internal data, i.e. our primary dataset contains information which permits their estimation, or externally, from other studies which have already been conducted. In order to use estimates from

external studies, we must be confident that the same measurement model applies to both the external study and our study, and that, at least approximately, the measurement model parameters are the same.

In this thesis we focus on classical error models, and so we now describe the types of data which can be used to assess the validity of our chosen measurement model and, given an assumed model, estimate the model parameters. Validation data contain observations of both the true \mathbf{X}_i and at least one error-prone measurement \mathbf{W}_i . Validation data permit an empirical assessment of any modelling assumptions we may make about the measurement errors and true covariates \mathbf{X}_i , because we jointly observe \mathbf{X}_i and \mathbf{W}_i . Thus we can assess:

- the distribution of \mathbf{X}_i
- the distribution of measurement errors \mathbf{U}_i , including:
 - whether the errors are unbiased
 - whether errors within a single multivariate measurement \mathbf{U}_i are correlated with each other
- whether the distribution of errors \mathbf{U}_i depends on \mathbf{X}_i . This includes the possibility of multiplicative measurement error and subject-specific bias.

Furthermore, with internal validation data, we can assess the non-differential error assumption by examining the joint distribution of Y_i and the measurement errors \mathbf{U}_i .

The specification of an appropriate measurement model can be made in light of these assessments. The fact that many assumptions can be empirically assessed with validation data means that, if we wish, we can use estimation methods which make fewer assumptions than are typically necessary when it is not possible to observe \mathbf{X}_i . The approach used to estimate the measurement model parameters depends on the particular model which is assumed. For example, in the previously described model used Rosner *et al* [2], the parameters can be estimated by multivariate linear regression of \mathbf{X}_i on \mathbf{Z}_i and \mathbf{W}_i , in those subjects included in the validation study.

2.5 Replication data

With replication data we have available at least two error-prone measurements on a group of subjects, but, in contrast to validation data, we are unable to observe the true covariate(s) \mathbf{X}_i . Internal replication data refers to the situation in which some of the subjects which form our main data under study have replicate measurements of their covariate(s). In contrast, an external replication study is when subjects who do not make up our main dataset have replicate measurements. Often this means that for subjects in an external replication study we do not observe the outcome Y_i .

In Part I of the thesis, we focus on the situation of classical covariate measurement error when internal replication data are available. We now extend the classical measurement error model to the case of internal replication data.

2.5.1 The classical error model for replication data

Scalar X_i

In the case of scalar X_i , we let n_i denote the number of error-prone measurements that are observed for subject i . To extend the classical measurement error model to this situation, we denote the j th measurement of X_i by W_{ij} . We then assume that:

$$W_{ij} = X_i + U_{ij} \quad (2.8)$$

such that $\mathbb{E}(U_{ij}|X_i) = 0$. Again, this implies that $\mathbb{E}(U_{ij}) = 0$ and that $\text{Cov}(X_i, U_{ij}) = 0$. We assume that errors within subjects are uncorrelated: $\text{Cov}(U_{ij}, U_{ij'}) = 0$ for $j \neq j'$. We also assume that measurement errors within subjects are identically distributed, so that in particular the variance of measurement errors does not depend on the j index. We then write σ_U^2 for the (common) variance of the measurement errors U_{ij} . A further assumption which we will sometimes make is that the errors U_{ij} are independent of X_i , rather than merely being uncorrelated with X_i .

Multivariate \mathbf{X}_i

In the case of multivariate $\mathbf{X}_i = (X_{i1}, \dots, X_{ip})^T$, we let n_{ij} denote the number of error-prone measurements available for subject i of covariate X_{ij} . In this case we let $n_i = \sum_{j=1}^p n_{ij}$ denote the total number of error-prone measurements available from subject i . In general, we allow for the possibility that different components of \mathbf{X}_i may have more error-prone measurements than others. We assume that the k th error-prone measurement of X_{ij} , denoted W_{ijk} , is given by:

$$W_{ijk} = X_{ij} + U_{ijk}$$

where U_{ijk} is measurement error with variance $\sigma_{U_j}^2$ that depends on the component of \mathbf{X}_i being measured. We now write $\mathbf{W}_{ij} = (W_{ij1}, \dots, W_{ijn_{ij}})^T$ for the vector of measurements of covariate X_{ij} , and $\mathbf{W}_i = (\mathbf{W}_{i1}^T, \dots, \mathbf{W}_{ip}^T)^T$ for the vector of all error-prone measurements for subject i . We can then express the classical error model as:

$$\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i \quad (2.9)$$

where \mathbf{D}_i is a known design matrix, given by:

$$\mathbf{D}_i = \mathbf{1}_{n_{i1} \times 1} \oplus \mathbf{1}_{n_{i2} \times 1} \oplus \dots \oplus \mathbf{1}_{n_{ip} \times 1},$$

and

$$\mathbf{U}_i = (\mathbf{U}_{i1}^T, \dots, \mathbf{U}_{ip}^T)^T,$$

where

$$\mathbf{U}_{ij} = (\mathbf{U}_{ij1}, \dots, \mathbf{U}_{ijn_{ij}})^T.$$

for $j = 1, \dots, p$. The design matrix \mathbf{D}_i is simply a matrix of zeros and ones which indicate which component of \mathbf{X}_i each measurement in \mathbf{W}_i corresponds to. Thus a row of \mathbf{D}_i corresponding to a measurement W_{ijk} consists of zeros, except for the j th column, which equals one.

The measurement error is classical if $\mathbb{E}(\mathbf{U}_i | \mathbf{X}_i) = \mathbf{0}$, so that $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$ and $\text{Cov}(\mathbf{X}_i, \mathbf{U}_i) = \mathbf{0}$. This means that:

$$\text{Var}(\mathbf{W}_i) = \mathbf{D}_i \boldsymbol{\Sigma}_X \mathbf{D}_i^T + \text{Var}(\mathbf{U}_i).$$

The structure of $\text{Var}(\mathbf{U}_i)$ will depend on the assumptions we wish to make. For example, we may assume that all the measurement errors U_{ijk} are uncorrelated with each other, so that:

$$\text{Var}(\mathbf{U}_i) = \sigma_{U_1}^2 \mathbf{I}_{n_{i1}} \oplus \sigma_{U_2}^2 \mathbf{I}_{n_{i2}} \oplus \dots \oplus \sigma_{U_p}^2 \mathbf{I}_{n_{ip}}. \quad (2.10)$$

Alternatively, we may wish to relax some of the zero correlation assumptions. For example, we might assume that measurement errors U_{ijk} and $U_{ij'k'}$ of different components of \mathbf{X}_i are correlated when $k = k'$.

2.5.2 Checking model assumptions

With replication data we cannot directly estimate the distribution of \mathbf{X}_i or of the measurement errors, but only the distribution of the error-prone measurements themselves. With replication data we are forced to make some assumptions which can either be checked indirectly or perhaps not at all. For example, with replication data we cannot assess whether error-prone measurements are biased. A number of the assumptions of the classical error model can however be checked with replication data. For example, plotting the within-subject SD of error-prone measurements against their mean can be used to assess an assumption of constant variance for the measurement errors. Carroll *et al* suggest plotting histograms of within-subject dif-

ferences in error-prone measurements (e.g. $\mathbf{W}_{i1} - \mathbf{W}_{i2}$) to indirectly assess whether the measurement errors are normally distributed [8].

2.5.3 Estimation

For univariate X_i and the model of equation (2.8), the classical measurement error model is a standard one-way variance components model. If n_i is the same for all subjects, it is a balanced one-way model, whereas in general it is unbalanced. The parameters (μ_X , σ_X^2 , and σ_U^2) can be estimated using the standard analysis of variance (ANOVA) estimators [24]. The ANOVA estimators of the variance components are unbiased and consistent without requiring distributional assumptions for X_i or U_{ij} . Alternatively, maximum likelihood (ML), or restricted maximum likelihood (REML) may be used for estimation, under the assumption of normality for X_i and U_{ij} . Although derived under these normality assumptions, Westfall has shown that, under suitable assumptions regarding the finiteness of moments, these ML estimators are consistent regardless of the distribution of X_i and U_{ij} [25]. In the case of balanced data (i.e. $n_i = n_\bullet$ for some $n_\bullet > 1$ for all i), the REML likelihood score equations are identical to the ANOVA estimating equations, so that they give identical estimates (unless the ANOVA estimator gives a negative variance estimate for σ_X^2) [24].

In the case of multivariate \mathbf{X}_i , the model of equation (2.9) falls within the linear mixed model framework. As such, under normality assumptions for \mathbf{X}_i and \mathbf{U}_i , its parameters can be estimated using linear mixed model commands, such as the SAS PROC MIXED command. Alternatively, analogous to the univariate X_i case, moment based estimators can be used (see Section 4.4.2 of [8]).

2.5.4 Number of replicates

A common study design is for all subjects to have one error-prone measurement W_{i1} , and for a randomly selected subset of the subjects to have a second error-prone measurement W_{i2} . Often this subset represents a relatively small (i.e. 10%) proportion of the total sample. In this design, the number of error-prone measurements for a particular subject i can be viewed as a random variable N_i , taking values in a finite discrete set (e.g. 1,2). In the Part I of the thesis, we treat the number of error-prone measurements available for each subject as fixed, and condition on the observed values n_i (this is analogous to treating the design matrices as fixed in linear mixed models, when often the number of observations available from a subject is the result of some random process). This is justified in particular if N_i is independent of \mathbf{X}_i , \mathbf{Z}_i , Y_i , and \mathbf{W}_i . By viewing those subjects with only one error-prone measurement as having a hypothetical second measurement ‘missing’, this assumption is the same

as the missing completely at random (MCAR) assumption from the missing data literature.

2.5.5 Asymptotics

For the internal replication setup that we focus on in the first part of this thesis, the information regarding the model parameters is not only a function of the number of subjects n . For example, if as $n \rightarrow \infty$, the number of subjects for which $n_i > 1$ is bounded above by some fixed value, then the precision of estimators of the measurement error variance will not increase with increasing n . When we say that methods are consistent, we must therefore specify how the number of replicate measurements n_i (or n_{ij} in the multivariate \mathbf{X}_i case) increases with increasing n . In order to ensure consistency of estimators, we assume that the number of subjects with $n_i > 1$ tends to infinity as $n \rightarrow \infty$. This ensures that estimates of the measurement error parameters, for which information comes from subjects with more than one error-prone measurement, become ever more precise as the number of subjects increases. In the case of multivariate \mathbf{X}_i , this requirement must be satisfied for each component of \mathbf{X}_i .

Chapter 3

Continuous outcomes

In this chapter we review the consequences of, and methods to allow for, classical covariate measurement error in linear regression outcome models. The study of errors in the independent variable of linear regression has a long history, dating back at least to the late 19th century [26]. We begin by reviewing the linear regression model and the standard ordinary least squares estimators for the regression coefficients (Section 3.1). We then examine the consequences of ignoring classical covariate measurement error, which leads naturally to the method of moments correction method (Section 3.3). In Section 3.4 we review regression calibration (RC), a conceptually appealing approach to dealing with measurement error which involves predicting the true covariate using error-prone measurements. We then discuss a structural ML approach to dealing with covariate measurement error for the linear regression model in Section 3.5. In Section 3.6, we show how standard linear mixed models, available in modern statistical packages, can be used to find ML estimates for a parametric model which assumes joint normality. The material of this section forms part of a paper which has been recently published in the journal *Statistics in Medicine* [27]. Covariate measurement error can be viewed as a type of missing data problem, and recently multiple imputation (MI) has been proposed as a method to deal with covariate measurement error. In Section 3.7 we discuss the application MI to deal with covariate measurement error. We show how the linear mixed model described for finding ML estimates can be used to multiply impute the unobserved covariates when replication data are available. These imputations can then be used to estimate the outcome model parameters. Using results from the missing data literature, we explain the relationship between the ML and MI estimators. We then examine moment reconstruction (MR), a new approach recently proposed for measurement error correction. We use simulations to compare these methods under a variety of scenarios (Section 3.8). We conclude the chapter by comparing the estimation methods in terms of efficiency and availability of software, and mentioning some further methods for dealing with covariate measurement error.

3.1 Linear regression

Linear regression models involve modelling the expected value of the outcome Y_i given the covariates \mathbf{X}_i and \mathbf{Z}_i . When \mathbf{X}_i is a scalar X_i , and there are no \mathbf{Z}_i , the simplest linear regression model is:

$$\mathbb{E}(Y_i|X_i) = \beta_0 + \beta_X X_i \quad (3.1)$$

This means that the expectation of Y_i varies as a linear function of X_i , with the mean function depending on two parameters, an intercept β_0 and a slope β_X . The model can also be expressed as:

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i \quad (3.2)$$

where ϵ_i is a residual error term with $\mathbb{E}(\epsilon_i|X_i) = 0$. This means that $\mathbb{E}(\epsilon_i) = 0$ and also that $\text{Cov}(X_i, \epsilon_i) = 0$. The ordinary least squares estimators (OLS) of β_0 and β_X are those values which minimize the residual sum of squares, and are given by:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_X \bar{X} \quad (3.3)$$

$$\hat{\beta}_X = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad (3.4)$$

where \bar{Y} and \bar{X} denote the sample means of the Y_i and X_i respectively. The OLS estimators are unbiased, and assuming that X_i has finite mean and variance, are consistent and asymptotically normal [28].

The parameters β_0 and β_X can be expressed in terms of the means, variances and covariance of Y_i and X_i . First, the covariance of Y_i and X_i is equal to:

$$\begin{aligned} \text{Cov}(Y_i, X_i) &= \text{Cov}(\beta_0 + \beta_X X_i + \epsilon_i, X_i) \\ &= \beta_X \sigma_X^2 + \text{Cov}(\epsilon_i, X_i) \\ &= \beta_X \sigma_X^2 \end{aligned} \quad (3.5)$$

where σ_X^2 denotes the variance of X_i in the population. This means that:

$$\beta_X = \frac{\sigma_{YX}}{\sigma_X^2} \quad (3.6)$$

where σ_{YX} denotes the covariance between Y_i and X_i . Similarly, it is simple to show that

$$\beta_0 = \mu_Y - \beta_X \mu_X \quad (3.7)$$

where μ_Y and μ_X denote the means of Y_i and X_i in the population.

3.1.1 Model mis-specification

In practice we posit a particular regression model on the basis of external subject matter knowledge, but we can rarely be sure that the specification is correct. What happens if we mis-specify the conditional mean function in the linear regression model? For example, suppose the expectation of Y_i is a linear function of X_i^2 , but that we fit a model which assumes the expectation is a linear function of X_i using the OLS estimators given in equations (3.3) and (3.4). In this case, $\hat{\beta}_X$ is an unbiased and consistent estimator of $\frac{\sigma_{YX}}{\sigma_X^2}$ and $\hat{\beta}_0$ is an unbiased and consistent estimator of $\mu_Y - \frac{\sigma_{YX}}{\sigma_X^2}\mu_X$. This means that even if the expectation of Y_i does not vary linearly with X_i , the OLS estimator $\hat{\beta}_X$ gives a consistent estimate of $\frac{\sigma_{YX}}{\sigma_X^2}$, which is the mean increase in the expectation of Y_i for a 1-unit increase in X_i , the mean being taken across the distribution of X_i .

Regardless of the joint distribution of Y_i and X_i , we can always write:

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i \quad (3.8)$$

where the residual error ϵ_i is uncorrelated with X_i and has mean zero, the simple proof of which is given in Appendix A4 of the monograph by Carroll *et al* [8]. This expression is known as the ‘best linear prediction’ of Y_i using X_i , where ‘best’ is used in the sense of minimizing mean squared error of prediction.

3.2 The effects of classical covariate measurement error

We now examine the effects of classical covariate measurement error in linear regression. These results have historically been derived assuming that the posited linear regression model is correctly specified, including normality assumptions for X_i and ϵ_i [7]. In fact, such strong assumptions about the validity of the outcome model of interest are not always necessary in order to quantify the effects of covariate measurement error. Since in practice we can rarely be sure that our proposed outcome model is correctly specified, it is preferable, where possible, to quantify the effects of covariate error without assuming that the outcome model is correctly specified [29].

We therefore assume that in the absence of covariate measurement error, we would fit a linear regression model to Y_i given X_i , that is:

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i \quad (3.9)$$

and that we would use the OLS estimators of β_0 and β_X , as given in equations (3.3) and (3.4), to estimate β_0 and β_X .

We assume that a single error-prone measurement W_i is available for each subject and that this follows the classical measurement error model (see Section 2.2.1). We thus assume that $W_i = X_i + U_i$ where U_i is measurement error such that $\mathbb{E}(U_i|X_i) = 0$, so that $\mathbb{E}(U_i) = 0$ and $\text{Cov}(X_i, U_i) = 0$. To quantify the effects of classical covariate measurement error in a linear regression outcome model, the non-differential error assumption need not be as strong as requiring conditional independence of Y_i and W_i given X_i (as shown below). Instead we proceed under the weaker assumption that the measurements errors U_i are uncorrelated with the outcome regression errors ϵ_i , i.e. $\text{Cov}(U_i, \epsilon_i) = 0$.

Suppose we fit the linear regression model given in equation (3.9), using the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_X$, but using W_i in place of X_i as the covariate. Recall that in general, with X_i as covariate, $\hat{\beta}_X$ is an unbiased estimate of

$$\frac{\sigma_{YX}}{\sigma_X^2}. \quad (3.10)$$

It therefore follows that using the OLS estimator for the linear regression slope with W_i as covariate is an unbiased estimator of

$$\beta_W = \frac{\sigma_{YW}}{\sigma_W^2} \quad (3.11)$$

where σ_{YW} and σ_W^2 denote the covariance of Y_i with W_i and the variance of W_i respectively. Under the stated assumptions for W_i , β_W is equal to:

$$\begin{aligned} \beta_W &= \frac{\text{Cov}(\beta_0 + \beta_X X_i + \epsilon_i, X_i + U_i)}{\text{Var}(X_i + U_i)} \\ &= \frac{\beta_X \sigma_X^2}{\sigma_X^2 + \sigma_U^2} \\ &= \beta_X \lambda \end{aligned} \quad (3.12)$$

where λ denotes the reliability of the error-prone measurements (see equation (2.3)). Since variances are non-negative, $0 < \lambda \leq 1$. We thus have the well known result that the effect of classical, non-differential covariate measurement error in linear regression models with a single covariate is to bias estimates towards the null. The larger the measurement error variance σ_U^2 relative to σ_X^2 , the greater the bias.

3.2.1 Multiple covariates

We now consider the effects of classical covariate measurement error in the more general case, in which multiple covariates are measured with error, and the outcome model also contains error-free covariates. As outlined in Chapter 2, we denote the covariates subject to classical measurement error by \mathbf{X}_i and those measured without error by \mathbf{Z}_i . As before, we assume that in the absence of covariate measurement

error we would fit a linear regression model:

$$Y_i = \beta_0 + \beta_X^T \mathbf{X}_i + \beta_Z^T \mathbf{Z}_i + \epsilon_i \quad (3.13)$$

using the OLS estimators. Analogous to the univariate X_i case, the OLS estimators for the regression coefficients of \mathbf{X}_i and \mathbf{Z}_i are consistent estimators of β_X and β_Z , where:

$$\begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} = \begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{XY} \\ \Sigma_{ZY} \end{pmatrix} \quad (3.14)$$

where Σ_X and Σ_Z denote the covariance matrices of \mathbf{X}_i and \mathbf{Z}_i , Σ_{ZX} denotes the $q \times p$ covariance matrix of \mathbf{Z}_i with \mathbf{X}_i , and Σ_{XY} and Σ_{ZY} denote the $p \times 1$ and $q \times 1$ covariance vectors of \mathbf{X}_i and \mathbf{Z}_i with Y_i .

We assume that \mathbf{X}_i is measured with classical error by $\mathbf{W}_i = \mathbf{X}_i + \mathbf{U}_i$ where $\mathbb{E}(\mathbf{U}_i | \mathbf{X}_i) = \mathbf{0}$, where $\text{Var}(\mathbf{U}_i) = \Sigma_U$. Furthermore, we assume that $\text{Cov}(\mathbf{U}_i, \epsilon_i) = \mathbf{0}$ and $\text{Cov}(\mathbf{U}_i, \mathbf{Z}_i) = \mathbf{0}$. We now denote by β_W and β_Z^* the regression coefficients which are estimated consistently by fitting the linear regression model using \mathbf{W}_i and \mathbf{Z}_i as covariates. Using equation (3.14), these are equal to:

$$\begin{aligned} \begin{pmatrix} \beta_W \\ \beta_Z^* \end{pmatrix} &= \begin{pmatrix} \Sigma_W & \Sigma_{WZ} \\ \Sigma_{ZW} & \Sigma_Z \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{WY} \\ \Sigma_{ZY} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_X + \Sigma_U & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{XY} \\ \Sigma_{ZY} \end{pmatrix} \\ &= \begin{pmatrix} \Sigma_X + \Sigma_U & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix} \begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} \end{aligned} \quad (3.15)$$

where the assumption that \mathbf{U}_i is uncorrelated with ϵ_i implies that $\Sigma_{WY} = \Sigma_{XY}$ and that it is uncorrelated with \mathbf{Z}_i implies $\Sigma_{ZW} = \Sigma_{ZX}$.

If there are no error-free covariates \mathbf{Z}_i , we have that:

$$\beta_W = \left(\Sigma_X + \Sigma_U \right)^{-1} \Sigma_X \beta_X \quad (3.16)$$

where $(\Sigma_X + \Sigma_U)^{-1} \Sigma_X$ can be seen as a multivariate generalisation of the reliability ratio λ .

To illustrate some of the consequences of classical measurement error in this setting, we consider an example in which there is a single covariate X_i measured with error and a single error-free covariate Z_i . We assume that

$$\begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$$

and

$$\begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} = \begin{pmatrix} 1 \\ 0.5 \end{pmatrix}.$$

Now with $\Sigma_U = (1)$, we have from equation (3.15) that:

$$\begin{pmatrix} \beta_W \\ \beta_Z^* \end{pmatrix} = \begin{pmatrix} 0.43 \\ 0.79 \end{pmatrix}$$

This example illustrates a number of important consequences of classical covariate measurement error. First, although Z_i is measured without error, the naive estimate of its adjusted effect on Y_i is biased due to the measurement error in W_i . Recall that β_Z represents the effect of Z_i adjusted for X_i . However, since we are adjusting for W_i , a noisy version of X_i , the effect of Z_i is only partially adjusted. Because X_i and Z_i are positively correlated in our example, this lack of adjustment causes bias away from the null for the effect of Z_i . If X_i and Z_i were uncorrelated, the estimate of β_Z would be unaffected by measurement error in W_i . Second, we see that the effect estimate of X_i is biased downwards by a factor greater than the univariate reliability ratio for the measurements W_i , which here is equal to 0.5. Thus the effect of measurement error for the variable measured with error is greater in the presence of a correlated error-free covariate Z_i . This is because it is the conditional variance of X_i given Z_i which is relevant, which is less than the unconditional variance.

These results have serious implications for epidemiological studies, where typically multivariable regression models are used to estimate the associations of exposures of interest with an outcome, having adjusted for variables thought to be confounders [5]. One is that if confounding variables are measured with error, effect estimates for other variables will only be partially adjusted for confounding. The second is that the impact of measurement error on regression coefficients for exposures of interest is greater when the outcome model contains other error-free covariates which are correlated with the exposures of interest, which is often the case.

3.3 Method of moments correction

The OLS slope estimate from the regression of Y_i on W_i gives an unbiased and consistent estimate of $\lambda\beta_X$, where β_X is the slope of the regression of Y_i on X_i , the parameter of interest, and λ is the reliability ratio of the error-prone measurements W_i . Equating the expected value of the naive estimator $\hat{\beta}_W$ with its point estimate immediately suggests that to remove bias we can divide the estimate $\hat{\beta}_W$ found by ignoring covariate measurement error (regressing Y_i on W_i) by λ . This approach to adjusting for classical covariate measurement error is known as method of moments

correction (MOM). For a linear regression outcome model it gives a consistent estimate of β_X . This corrected estimate of β_X can also be used to give a corrected estimate of β_0 , the intercept term of the linear regression model. Usually the reliability ratio λ will not be known exactly. However, we may estimate it using internal validation or replication data, or from an external study, as discussed in Chapter 2. Provided λ is estimated consistently, the corrected estimates of β_X and β_0 are also consistent.

If the value of the reliability ratio λ is known, we then have that:

$$\text{Var}\left(\frac{\hat{\beta}_W}{\lambda}\right) = \frac{\text{Var}(\hat{\beta}_W)}{\lambda^2} \quad (3.17)$$

Since $\lambda < 1$, the MOM corrected estimator has greater variance than the biased naive estimator – in order to remove bias we must accept an increase in sampling variability, although the value we are estimating is of course also larger in magnitude.

3.3.1 Inference

If λ is known, confidence intervals for β_X can be constructed by dividing the ‘naive’ confidence interval limits for $\hat{\beta}_W$ by λ . When λ is estimated, inference for the corrected estimate $\hat{\beta}_X$ should incorporate the uncertainty in $\hat{\lambda}$. The delta method can be used to estimate the variance of the corrected estimate, taking into account uncertainty in $\hat{\lambda}$ and $\hat{\beta}_W$ [13]. Alternatively, Frost and Thompson described how Fieller’s theorem can be used to find confidence intervals for the corrected estimate [30]. If the estimates of λ and β_W cannot be considered independent, because they are estimated using the same dataset, this should be allowed for in inference for the corrected estimate, although simulation results from Frost and Thompson suggested ignoring this dependence may be reasonable in certain settings [30].

An alternative approach, which can allow for such dependency, is to use non-parametric bootstrapping [31]. This involves repeatedly sampling subjects (i.e. $(Y_i, \mathbf{W}_i, \mathbf{Z}_i)$) with replacement from the original dataset to create a number of ‘bootstrap’ datasets. The process of estimating the measurement error parameters (i.e. λ), fitting the naive regression model, and correcting the biased estimates, is repeated for each bootstrap dataset. The variation in the estimates of the parameters of interest across the bootstraps can then be used to estimate standard errors which can be used to form normal-based confidence intervals. Alternatively, with a sufficiently large number of bootstraps, confidence intervals can be constructed based on the percentiles of the bootstrap distribution. These have superior coverage properties and allow for non-normality in the sampling distribution of estimators. This may be particularly useful for measurement error corrected estimates, because they usually have a skewed distribution, which in small samples can be quite severe [7].

We suppose for a moment that a subset of subjects in our dataset have independent replicates W_{i1} and W_{i2} , which we use to estimate the reliability ratio λ . Simple sampling with replacement means the number of observations with two replicates will vary between bootstrap samples. In theory, the bootstrap sampling process should respect the original design of the study. Thus if a fixed number of subjects were to be included in the replication study, we should bootstrap separately from subjects who were in the replication study and those who were not, thus creating bootstrap samples with the same number of subjects with and without replicates. Carroll *et al* reported however that in large samples, whether one uses such stratified re-sampling or not typically makes little difference [8].

3.3.2 Efficiency

In the case when internal replication data are available, simple MOM correction is generally inefficient when n_i differs between subjects. Suppose, for example that a randomly selected subset of subjects have $n_i = 2$ error-prone measurements of X_i and the remainder have $n_i = 1$ error-prone measurement of X_i . MOM correction then consists of dividing the naive estimate of β_X , obtained by regressing Y_i on W_{i1} using all subjects, by an estimate of λ . This is inefficient because the second error-prone measurements which are available are only used in the estimation of λ , but are ignored in the naive estimation of β_X .

It is however relatively simple to construct an improved MOM estimator with greater efficiency [7, 32]. First, we estimate σ_X^2 using the error-prone measurements \mathbf{W}_i , either by ANOVA methods or likelihood methods. Then recalling that $\beta_X = \sigma_{YX}/\sigma_X^2$, we see that we only need an estimate of σ_{YX} in order to construct an estimate of β_X . This covariance can be estimated by the sample covariance of Y_i and \bar{W}_i , since:

$$\begin{aligned} \text{Cov}(Y_i, \bar{W}_i) &= \text{Cov}\left(Y_i, X_i + \frac{\sum_{j=1}^{n_i} U_{ij}}{n_i}\right) \\ &= \sigma_{YX}. \end{aligned}$$

The sample covariance of \bar{W}_i with Y_i can thus be divided by an estimate of σ_X^2 , giving a consistent estimator of β_X . This estimator is more efficient than the simple MOM estimator because it uses all of a subject's error-prone measurements to estimate the two parameters which β_X depends on. In contrast, the simple MOM estimator only uses a single error-prone measurement to estimate σ_{YX} and $\sigma_X^2 + \sigma_U^2$.

This improved MOM estimator is however still somewhat inefficient because each subject is given equal weight in the estimation of σ_{YX} . This is not optimal because subjects with more error-prone measurements provide more information about the covariance than those with fewer measurements, because the variance of \bar{W}_i is a decreasing function of n_i . To further improve efficiency, Fuller proposed a weighted

estimator of β_X which uses the previously described improved MOM estimator as a preliminary estimator of β_X [7].

In Section 3.4, we introduce regression calibration, which, unlike simple MOM correction, also uses all of a subject's available error-prone measurements in the estimation process, and is simple to implement.

3.3.3 Multiple covariates

As for the univariate X_i case, with either multivariate \mathbf{X}_i or if there are error-free covariates \mathbf{Z}_i , the naive estimates of β_X and β_Z can be corrected by re-arranging equation (3.15), giving:

$$\begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} = \begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_X + \Sigma_U & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix} \begin{pmatrix} \beta_W \\ \beta_Z^* \end{pmatrix}.$$

Of course, as for the univariate X_i case, inferences for corrected estimates should allow for the estimation of not just β_W and β_Z^* , but of the parameters involved in the matrices above. Rosner and colleagues showed how the multivariate delta method can be used for this, both in the case of validation data [2], and replication data [33].

3.4 Regression calibration

Regression calibration (RC) is a widely applicable intuitive approach for adjusting for the effects of measurement error. RC involves fitting the outcome model of interest but in place of the unobserved true covariate X_i we use a prediction of its value based on either one error-prone measurement W_i or multiple measurements W_{ij} . There are at least two approaches to predicting X_i using W_{ij} . Under certain modelling assumptions the two approaches give identical estimates of β_X (and β_Z).

3.4.1 Prediction using conditional expectation

Within the more general framework of a generalised linear outcome model, Armstrong proposed what is now called RC [34]. In the simplest case of a single covariate X_i measured with error by W_i , RC involves fitting the outcome regression model of interest, substituting $\mathbb{E}(X_i|W_i)$ for the unobserved X_i . To show why RC works when the outcome model of interest is linear regression, we assume that the conditional expectation of Y_i given X_i is linear in X_i , so that:

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i \tag{3.18}$$

where $\mathbb{E}(\epsilon_i|X_i) = 0$. Now suppose that we do not observe X_i , but instead an error-prone measurement W_i . The validity of RC relies on the non-differential error assumption, i.e. that $f(Y_i|X_i, W_i) = f(Y_i|X_i)$. If Y_i follows the linear regression model given in equation (3.18), the non-differential error assumption means that W_i is independent of the residual error ϵ_i . Then taking expectations of equation (3.18) conditional on W_i , it follows that:

$$\begin{aligned}\mathbb{E}(Y_i|W_i) &= \beta_0 + \beta_X \mathbb{E}(X_i|W_i) + \mathbb{E}(\epsilon_i|W_i) \\ &= \beta_0 + \beta_X \mathbb{E}(X_i|W_i)\end{aligned}$$

where the independence of W_i and ϵ_i (non-differential error) implies $\mathbb{E}(\epsilon_i|W_i) = 0$.

Multiple error-prone measurements

One of the advantages of RC compared to MOM correction is its flexibility in handling various types of measurement data and different measurement models. In particular, RC allows a subject's value of X_i to be predicted by multiple error-prone measurements. Suppose subject i has n_i error-prone measurements of X_i , denoted by $\mathbf{W}_i = (W_{i1}, \dots, W_{in_i})^T$. As before, we can then show that:

$$\mathbb{E}(Y_i|\mathbf{W}_i) = \beta_0 + \beta_X \mathbb{E}(X_i|\mathbf{W}_i)$$

assuming that the errors in \mathbf{W}_i are non-differential. With more error-prone measurements, a subject's value of X_i can be predicted more accurately. As n_i increases, $\text{Var}(\mathbb{E}(X_i|\mathbf{W}_i))$ increases, with an upper bound of $\text{Var}(X_i)$. The residual error variance of the regression of Y_i on $\mathbb{E}(X_i|\mathbf{W}_i)$ therefore decreases as n_i increases. It follows that the precision of the OLS slope estimator increases as n_i increases, since this is a decreasing function of the residual error variance and an increasing function of the variance of the explanatory variable. Furthermore, the OLS estimator is consistent even if the residual variance differs between subjects, and so the fact that n_i may differ between subjects does not affect the consistency of the resulting estimates.

Specifying the calibration function

To use RC, we must specify the form of, and calculate $\mathbb{E}(X_i|\mathbf{W}_i)$. The conditional distribution of X_i given \mathbf{W}_i can be expressed in terms of the marginal density $f(X_i)$ and the conditional density $f(\mathbf{W}_i|X_i)$. This means that the conditional distribution $f(X_i|\mathbf{W}_i)$ depends not only on the distribution of the measurement errors, but also on the distribution of the covariate X_i in the population of interest. It is often assumed that the conditional expectation is a linear function of the mean of available measurements, which we denote \overline{W}_i . This is for example the case when

X_i is normal and the measurement errors U_{ij} are normally distributed, independent of X_i . However, there is no reason why the conditional expectation $\mathbb{E}(X_i|\mathbf{W}_i)$ need necessarily be a linear function of \overline{W}_i . For example if the variance of the errors U_{ij} increases with X_i , $\mathbb{E}(X_i|\mathbf{W}_i)$ is non-linear in \overline{W}_i .

3.4.2 Best linear prediction

With a single error-prone measurement W_i , a popular approach in the literature is to assume that the conditional expectation $\mathbb{E}(X_i|W_i)$ is a linear function of W_i . Fitting this calibration function is equivalent to using the ‘best linear prediction’ of X_i given W_i . We will show that using the best linear prediction of X_i given W_i as covariate results in consistent estimates of β_0 and β_X when the outcome model is linear regression, irrespective of whether the best linear prediction is equal to $\mathbb{E}(X_i|W_i)$. The expressions we derive are also of interest because they coincide with those obtained when one assumes that X_i and U_i are independent and both normally distributed. Using the best linear prediction of X_i given W_i in a RC type approach was proposed in the more general setting of generalized linear models by Gleser [35] and Carroll and Stefanski [36].

In contrast to RC using $\mathbb{E}(X_i|W_i)$, for which the justification requires the correct specification of $\mathbb{E}(Y_i|X_i)$, here we merely assume that in the absence of covariate measurement error we would fit the linear regression model:

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i \quad (3.19)$$

where β_0 and β_X , defined in equations (3.6) and (3.7), are estimated by the usual OLS estimators. We recall from Section 3.1.1 that the residual error ϵ_i is uncorrelated with X_i and has expectation zero. Assume that $W_i = X_i + U_i$ is an error-prone measurement of X_i where $\mathbb{E}(U_i|X_i) = 0$ and that U_i is uncorrelated with ϵ_i .

Analogous to the best linear prediction of Y_i given X_i (see Section 3.1.1), the best linear prediction of X_i using W_i is given by:

$$\begin{aligned} X_i^{blp} &= \mathbb{E}(X_i) + \frac{\text{Cov}(X_i, W_i)}{\text{Var}(W_i)}(W_i - \mathbb{E}(W_i)) \\ &= \mu_X + \frac{\text{Cov}(X_i, X_i + U_i)}{\sigma_X^2 + \sigma_U^2}(W_i - \mu_X) \\ &= \mu_X + \lambda(W_i - \mu_X). \end{aligned} \quad (3.20)$$

If we regress Y_i on X_i^{blp} , the slope parameter that we consistently estimate is given by

$$\begin{aligned} \frac{\text{Cov}(Y_i, \mu_X + \lambda(W_i - \mu_X))}{\text{Var}(\mu_X + \lambda(W_i - \mu_X))} &= \frac{\text{Cov}(Y_i, \lambda W_i)}{\lambda^2(\sigma_X^2 + \sigma_U^2)} \\ &= \frac{\text{Cov}(Y_i, X_i)}{\lambda(\sigma_X^2 + \sigma_U^2)} \\ &= \frac{\text{Cov}(Y_i, X_i)}{\sigma_X^2} = \beta_X. \end{aligned}$$

Thus consistent estimates of β_X are obtained if we regress Y_i on X_i^{blp} , assuming we can estimate the parameters involved in X_i^{blp} consistently. A similar exercise shows that the OLS estimate of the intercept term is a consistent estimate of β_0 .

From equation (3.20), we see that for subjects with values of W_i above the mean, W_i is shrunk in towards the mean to give X_i^{blp} , whereas those with below average W_i values are pushed up towards the mean. This can be thought of as an example of regression to the mean – for subjects with large values of W_i , the fact they have an above average value means in expectation their measurement has positive measurement error ($U_i > 0$), and so the predicted value of their value of X_i is smaller than their measurement W_i . The phenomenon is also known as shrinkage in the linear mixed model literature in the prediction of realisations of unobserved random effects [37].

Multiple error-prone measurements

The best linear prediction of X_i is easily extended to the case where multiple error-prone measurements are available. Thus suppose that subject i has n_i error-prone measurements of X_i , which are subject to classical error. As described in Section 2.5, we assume the j th measurement $W_{ij} = X_i + U_{ij}$ where $\mathbb{E}(U_{ij}|X_i) = 0$ and $\text{Cov}(U_{ij}, U_{ik}) = 0$ for $j \neq k$. We then consider the best linear prediction of X_i given the mean \bar{W}_i . Since:

$$\begin{aligned} \text{Var}(\bar{W}_i) &= \text{Var}\left(X_i + \frac{\sum_{j=1}^{n_i} U_{ij}}{n_i}\right) \\ &= \sigma_X^2 + \frac{\sigma_U^2}{n_i} \end{aligned}$$

and:

$$\text{Cov}(X_i, \bar{W}_i) = \text{Cov}\left(X_i, X_i + \sum_{j=1}^{n_i} U_{ij}\right) = \sigma_X^2,$$

the best linear prediction of X_i given \overline{W}_i is equal to:

$$X_i^{blp} = \mu_X + \frac{\sigma_X^2}{\sigma_X^2 + \frac{\sigma_U^2}{n_i}}(\overline{W}_i - \mu_X). \quad (3.21)$$

When $n_i = 1$ this reduces to (3.20). As before, subjects with large values of \overline{W}_i are shrunk downwards towards the mean, while those with low values are increased upwards. Also, as n_i increases, the predicted value converges to \overline{W}_i , since with more measurements there is greater information about a subject's value of X_i . As with prediction using a single W_i , the OLS slope estimator for the regression of Y_i on X_i^{blp} is a consistent estimator of β_X .

3.4.3 Best linear prediction and conditional expectation

As stated earlier, in certain cases best linear prediction coincides with conditional expectation. One case in which this occurs is when X_i and U_{ij} are assumed to be independent of each other and normally distributed. Since RC using best linear prediction gives consistent estimates of β_X , this means that if one uses RC and assumes X_i and U_{ij} are independent and normally distributed, RC gives consistent estimates of β_X even if these assumptions do not hold, providing the parameters involved in $\mathbb{E}(X_i|\mathbf{W}_i)$ are estimated consistently. As discussed earlier in Section 2.5, the ANOVA estimators of the variance components σ_X^2 and σ_U^2 and of μ_X are unbiased regardless of the distributions of X_i and U_{ij} and without requiring independence between X_i and U_{ij} . Westfall has shown that the ML and REML estimators of these parameters for the one-way model are also consistent without requiring these assumptions [25]. Thus, at least for a linear regression outcome model, assuming that X_i and U_{ij} are normally distributed and independent and using RC gives consistent estimates even if these assumptions do not hold.

When $\mathbb{E}(X_i|\mathbf{W}_i)$ differs from the best linear prediction, intuitively we would expect use of best linear prediction to result in less precise estimates of β_X compared to using $\mathbb{E}(X_i|\mathbf{W}_i)$. This follows from the fact that:

$$\text{Var}(\mathbb{E}(X_i|\mathbf{W}_i)) \geq \text{Var}(X_i^{blp}).$$

3.4.4 Implementation

Usually the parameters required to predict X_i are estimated, using internal validation or replication data, or from an external study, as discussed in Chapter 2, and substituted in place of the unknown true values. RC then involves regressing Y_i on $\hat{\mathbb{E}}(X_i|\mathbf{W}_i)$ or \hat{X}_i^{blp} . So long as the parameters needed to calculate $\mathbb{E}(X_i|\mathbf{W}_i)$ or X_i^{blp} are consistently estimated, the resulting RC estimates will also be consistent.

3.4.5 Inference

Whichever method is used to predict the true covariate X_i , standard errors which are reported from the fitted regression of Y_i on the predicted X_i values will be too small if the parameters needed to predict X_i have been estimated. Furthermore, if multiple measurements \mathbf{W}_i are used to predict X_i , the conditional variance of Y_i given the prediction of X_i will vary between subjects, invalidating the basis for the usual estimates of precision for the OLS estimators. The easiest approach to obtaining valid standard errors and confidence intervals for parameter estimates for RC is probably to bootstrap the whole procedure. Again, we refer to section A9 of the monograph by Carroll *et al* [8]. An alternative is to view the RC estimator as the solution to unbiased estimating equations. Asymptotic standard errors can then be obtained using the sandwich estimator of variance for estimating equations (see Appendix A6.6 of [8]), although this approach is more complicated.

3.4.6 Efficiency

Compared to simple MOM correction (i.e. correcting an estimate of β_W found using a single error-prone W_i), RC gives more efficient estimates of β_X in the case of internal replication data. This is because all of a subject's error-prone measurements are used to predict X_i . However, when n_i varies between subjects, even if the error ϵ_i in the linear regression of Y_i on X_i has constant variance, the regression of Y_i on $\mathbb{E}(X_i|\mathbf{W}_i)$ will have errors with differing variances because the variance of $\mathbb{E}(X_i|\mathbf{W}_i)$ depends on n_i . Using unweighted least squares to regress Y_i on $\mathbb{E}(X_i|\mathbf{W}_i)$ (or $\hat{\mathbb{E}}(X_i|\mathbf{W}_i)$) will thus be inefficient to some extent.

We now briefly consider how the residual variance of the regression of Y_i on $\mathbb{E}(X_i|\mathbf{W}_i)$ varies depending on n_i . We assume that X_i is normally distributed, and is measured with independent errors U_{ij} which are also normally distributed with variance σ_U^2 . It then follows that the regression of Y_i on $\mathbb{E}(X_i|\mathbf{W}_i)$ has residual variance:

$$\begin{aligned} \text{Var}(Y_i|\mathbf{W}_i) &= \text{Var}(Y_i) - \beta_X^2 \text{Var}(X_i|\mathbf{W}_i) \\ &= \beta_X^2 \sigma_X^2 + \sigma_\epsilon^2 - \beta_X^2 \frac{\sigma_X^4}{\sigma_X^2 + \sigma_U^2/n_i} \\ &= \sigma_\epsilon^2 + \beta_X^2 \sigma_X^2 \left(1 - \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2/n_i} \right). \end{aligned} \quad (3.22)$$

This shows that the weights for subjects will differ most when the measurement error variance σ_U^2 is large in comparison to the variance σ_X^2 and the number of measurements n_i is large for some subjects but small for others. Conversely, when the residual variance σ_ϵ^2 is large in comparison to β_X and σ_X^2 , the effects of different values of n_i on subjects' weights will be less important. This suggests that if X_i

only explains a small part of the variation in Y_i (as is typical in chronic disease epidemiology), using RC may not result in a large loss in efficiency. Our simulation results (see Section 3.9) support this.

We note that more complicated RC estimators can be constructed which should have greater efficiency than standard RC. This would involve first using RC to form a preliminary estimate of β_X , followed by weighted least squares, using equation (3.22) to estimate the required weights. Weighted least squares regression could then be applied iteratively, updating the weights using the updated estimate of β_X until the estimates converge to a fixed point.

3.4.7 Multiple covariates

Prediction using conditional expectation

We now suppose that \mathbf{X}_i is a p -column vector of covariates which are measured with error and \mathbf{Z}_i is a q -column vector of covariates that are observed without error, and that:

$$\mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{Z}_i) = \beta_0 + \beta_X^T \mathbf{X}_i + \beta_Z^T \mathbf{Z}_i.$$

Suppose that \mathbf{W}_i is a vector of error-prone measurements of \mathbf{X}_i . Analogous to the single X_i case, it follows that:

$$\mathbb{E}(Y_i|\mathbf{W}_i, \mathbf{Z}_i) = \beta_0 + \beta_X^T \mathbb{E}(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i) + \beta_Z^T \mathbf{Z}_i$$

assuming that \mathbf{W}_i is independent of Y_i , conditional on \mathbf{X}_i and \mathbf{Z}_i . We see that in the presence of error-free covariates \mathbf{Z}_i , we must substitute the conditional expectation of \mathbf{X}_i given both \mathbf{W}_i and \mathbf{Z}_i . In general, the error-free covariates \mathbf{Z}_i will not be independent of the unobserved \mathbf{X}_i , and so \mathbf{Z}_i provides information about \mathbf{X}_i in addition to that provided by the error-prone measurements \mathbf{W}_i . This means that we must specify how the expectation of \mathbf{X}_i varies as a function of \mathbf{Z}_i .

We now consider a particular model specification which we will use repeatedly throughout the thesis. We assume that \mathbf{X}_i is multivariate normal given \mathbf{Z}_i , with:

$$\mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) = \mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z \mathbf{Z}_i \tag{3.23}$$

$$\text{Var}(\mathbf{X}_i|\mathbf{Z}_i) = \mathbf{\Sigma}_{X|Z}. \tag{3.24}$$

We then assume that a vector of error-prone measurements \mathbf{W}_i is available for each subject which follows the classical error model $\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i$ as described in Section 2.5.1. We assume that $\mathbb{E}(\mathbf{U}_i|\mathbf{X}_i) = \mathbf{0}$, that $\mathbb{E}(\mathbf{U}_i|\mathbf{Z}_i) = \mathbf{0}$, and that $\mathbf{U}_i|\mathbf{Z}_i \sim N(\mathbf{0}, \text{Var}(\mathbf{U}_i))$, where:

$$\text{Var}(\mathbf{U}_i) = \sigma_{U_1}^2 \mathbf{I}_{n_{i1} \times n_{i1}} \oplus \sigma_{U_2}^2 \mathbf{I}_{n_{i2} \times n_{i2}} \oplus \dots \oplus \sigma_{U_p}^2 \mathbf{I}_{n_{ip} \times n_{ip}}. \tag{3.25}$$

It then follows that:

$$\mathbb{E}(\mathbf{W}_i|\mathbf{Z}_i) = \mathbf{D}_i(\boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_Z\mathbf{Z}_i), \quad (3.26)$$

and that:

$$\text{Var}(\mathbf{W}_i|\mathbf{Z}_i) = \mathbf{D}_i\boldsymbol{\Sigma}_{X|Z}\mathbf{D}_i^T + \text{Var}(\mathbf{U}_i). \quad (3.27)$$

We now show that the conditional mean function $\mathbb{E}(\mathbf{W}_i|\mathbf{Z}_i)$ can be expressed as a product of a known design matrix and a vector of unknown parameters. First, if we let $\boldsymbol{\Gamma}_{Z_1}, \dots, \boldsymbol{\Gamma}_{Z_q}$ denote the columns of $\boldsymbol{\Gamma}_Z$, we can write:

$$\boldsymbol{\Gamma}_Z\mathbf{Z}_i = Z_{i1}\boldsymbol{\Gamma}_{Z_1} + \dots + Z_{iq}\boldsymbol{\Gamma}_{Z_q}$$

and so

$$\mathbb{E}(\mathbf{W}_i|\mathbf{Z}_i) = \mathbf{D}_i\boldsymbol{\Gamma}_0 + Z_{i1}\mathbf{D}_i\boldsymbol{\Gamma}_{Z_1} + \dots + Z_{iq}\mathbf{D}_i\boldsymbol{\Gamma}_{Z_q} \quad (3.28)$$

Therefore, letting $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_0^T, \boldsymbol{\Gamma}_{Z_1}^T, \dots, \boldsymbol{\Gamma}_{Z_q}^T)^T$, we can express the fixed effects of the model for \mathbf{W}_i given \mathbf{Z}_i as the product of a known design matrix and vector of parameters as:

$$\mathbb{E}(\mathbf{W}_i|\mathbf{Z}_i) = \mathcal{D}_i\boldsymbol{\Gamma} \quad (3.29)$$

where:

$$\mathcal{D}_i = \begin{pmatrix} \mathbf{D}_i & Z_{i1}\mathbf{D}_i & \dots & Z_{iq}\mathbf{D}_i \end{pmatrix}. \quad (3.30)$$

We have therefore shown that \mathbf{W}_i can be expressed as a linear mixed model given \mathbf{Z}_i , with fixed effects design matrix \mathcal{D}_i , random effects design matrix \mathbf{D}_i , random effects variance covariance matrix $\boldsymbol{\Sigma}_{X|Z}$, and residual variance covariance matrix $\text{Var}(\mathbf{U}_i)$. This linear mixed model for \mathbf{W}_i given \mathbf{Z}_i can be fitted using ML or REML using SAS's PROC MIXED command.

Given the estimates of the model's parameters, the conditional expectation of the random effects can be calculated for each subject. These follow from standard results for multivariate normal distributions, see for example Section 7.2 of [37]. The predicted random effects can then be added to $\mathcal{D}_i\hat{\boldsymbol{\Gamma}}$ to give $\hat{\mathbb{E}}(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$. The regression of Y_i on $\hat{\mathbb{E}}(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ and \mathbf{Z}_i can then be fitted, yielding consistent estimates of $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$.

Prediction using best linear prediction

As before, instead of using predicting \mathbf{X}_i by $\mathbb{E}(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$, we may use its best linear prediction. Expressions for the best linear prediction in the case of replication data are given in Section 4.4 of [8], in which moment based estimators of the required parameters are also given.

3.5 Maximum likelihood

The method of maximum likelihood (ML) is arguably the most popular approach to finding estimates of parameters of statistical models. This is due to the fact that, subject to regularity conditions, ML estimators have a number of desirable properties - they are consistent and have the smallest asymptotic variance amongst consistent estimators [28]. In this section we consider a parametric likelihood approach to allow for covariate measurement error in a linear regression outcome model. The use of ML to accommodate covariate measurement error has been advocated repeatedly over the last 25 years. Early examples are papers by Schafer and colleagues, who suggested using the expectation maximization algorithm to obtain ML estimates of a model which allows for covariate measurement error [32, 38, 39]. Despite these, and other proposals, the ML method has been used much less frequently in epidemiology than simpler methods such as RC. This is presumably due to a number of factors, including lack of availability in standard statistical software packages, concerns about the robustness of ML procedures to their parametric assumptions, and a belief that simpler methods such as RC often perform as well as ML.

In this section we describe a particular parametric model, based on an assumption of joint normality, for a linear regression outcome model in which the covariate X_i is unobserved, but is measured with classical error by internal replicate error-prone measurements (Section 3.5.1). We then define the observed data likelihood function, on which ML estimates and inferences are based (Section 3.5.2). Next we describe the Newton Raphson method (Section 3.5.3) and the Expectation Maximization (EM) algorithm (Section 3.5.4), both of which can be used to find the maximum likelihood estimates of the model parameters. In Section 3.6, we show how the maximum likelihood estimates for the model we define in Section 3.5.1 can be found by fitting a simple linear mixed model, using standard statistical packages. This means that researchers can find ML estimates for this model using standard statistical software, without having to write custom Newton Raphson or EM programs.

3.5.1 Model specification

We first consider the simple case in which the outcome Y_i is assumed to be related via a linear regression model to a single covariate X_i . As before, X_i is unobserved, but for subject i , a vector of n_i error-prone measurements of X_i is available, which we denote \mathbf{W}_i . Under the assumption that the measurement errors in \mathbf{W}_i are non-differential, the density function of the complete data, i.e. including X_i , can be decomposed in terms of conditional densities as:

$$\begin{aligned} f(Y_i, X_i, \mathbf{W}_i) &= f(Y_i|X_i, \mathbf{W}_i)f(X_i, \mathbf{W}_i) \\ &= f(Y_i|X_i)f(\mathbf{W}_i|X_i)f(X_i). \end{aligned} \quad (3.31)$$

Thus under the non-differential error assumption, a ‘joint model’ can be defined by specifying three so called sub-models: the outcome model, $f(Y_i|X_i)$, the measurement model, $f(\mathbf{W}_i|X_i)$, and the covariate model, $f(X_i)$ [40].

Outcome model

For the outcome model $f(Y_i|X_i)$ we assume a linear regression model:

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i \quad (3.32)$$

where $\epsilon \sim N(0, \sigma_\epsilon^2)$ is independent of X_i . Note that in contrast to MOM or RC, for a parametric likelihood approach we must specify the conditional distribution of Y_i given X_i , and not merely the conditional mean function.

Covariate model

We assume that the covariate X_i is normally distributed $N(\mu_X, \sigma_X^2)$.

Measurement model

For the measurement model, $f(\mathbf{W}_i|X_i)$, we assume that the error-prone measurements $\mathbf{W}_i = (W_{i1}, \dots, W_{in_i})^T$ are subject to unbiased normally distributed classical error:

$$W_{ij} = X_i + U_{ij}. \quad (3.33)$$

We thus assume that $U_{ij} \sim N(0, \sigma_U^2)$, and that measurement errors U_{ij} are independent of X_i and other measurement errors $U_{ij'}$. In addition, we assume the errors are non-differential, so that $f(Y_i|X_i, \mathbf{W}_i) = f(Y_i|X_i)$.

3.5.2 The likelihood function and the maximum likelihood estimate

We now consider the likelihood function when only Y_i and \mathbf{W}_i are observed, i.e. X_i is unobserved. The observed data likelihood, which is sometimes called the marginal likelihood function, corresponding to the observed data is then given by:

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^n L(\boldsymbol{\theta}|Y_i, \mathbf{W}_i) \quad (3.34)$$

$$= \prod_{i=1}^n f(Y_i, \mathbf{W}_i|\boldsymbol{\theta}) \quad (3.35)$$

where $\boldsymbol{\theta} = (\beta_0, \beta_X, \sigma_\epsilon^2, \mu_X, \sigma_X^2, \sigma_U^2)^T$ denotes the vector of parameters of the joint model. We condition on the observed values of n_i throughout (see Section 2.5.4). The observed data likelihood therefore depends on the density $f(Y_i, \mathbf{W}_i)$, which can be found by marginalizing the joint distribution $f(Y_i, X_i, \mathbf{W}_i)$ over X_i . As a product of normal densities (equation (3.31)), the joint distribution $f(Y_i, X_i, \mathbf{W}_i)$ is multivariate normal with mean vector

$$\mathbb{E} \begin{pmatrix} Y_i \\ X_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_X \mu_X \\ \mu_X \\ \mu_X \mathbf{1}_{n_i \times 1} \end{pmatrix} \quad (3.36)$$

and variance covariance matrix

$$\text{Var} \begin{pmatrix} Y_i \\ X_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \beta_X^2 \sigma_X^2 + \sigma_\epsilon^2 & \beta_X \sigma_X^2 & \beta_X \sigma_X^2 \mathbf{1}_{1 \times n_i} \\ \beta_X \sigma_X^2 & \sigma_X^2 & \sigma_X^2 \mathbf{1}_{1 \times n_i} \\ \beta_X \sigma_X^2 \mathbf{1}_{n_i \times 1} & \sigma_X^2 \mathbf{1}_{n_i \times 1} & \sigma_X^2 \mathbf{1}_{n_i \times n_i} + \sigma_U^2 \mathbf{I}_{n_i \times n_i} \end{pmatrix} \quad (3.37)$$

where $\mathbf{I}_{n_i \times n_i}$ denotes a $n_i \times n_i$ identity matrix and $\mathbf{1}_{n_i \times 1}$ denotes a $n_i \times 1$ vector of 1s. It then follows from standard results for the multivariate normal distribution (e.g. Appendix S3 of [24]), that the marginal density $f(Y_i, \mathbf{W}_i)$ is also multivariate normal, with mean vector and variance covariance matrix given by the appropriate sub-vector and matrices of equations (3.36) and (3.37):

$$\mathbb{E} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_X \mu_X \\ \mu_X \mathbf{1}_{n_i \times 1} \end{pmatrix}, \quad (3.38)$$

$$\text{Var} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \beta_X^2 \sigma_X^2 + \sigma_\epsilon^2 & \beta_X \sigma_X^2 \mathbf{1}_{1 \times n_i} \\ \beta_X \sigma_X^2 \mathbf{1}_{n_i \times 1} & \sigma_X^2 \mathbf{1}_{n_i \times n_i} + \sigma_U^2 \mathbf{I}_{n_i \times n_i} \end{pmatrix} \quad (3.39)$$

The maximum likelihood estimate (MLE) of the model parameters $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}$ which maximizes the observed data likelihood function, or equivalently the observed data log likelihood. Under certain regularity conditions, the MLE is the value

which solves the likelihood score equations:

$$S_n(\boldsymbol{\theta}) = \sum_{i=1}^n S(\boldsymbol{\theta}|Y_i, \mathbf{W}_i) = 0 \quad (3.40)$$

where

$$S(\boldsymbol{\theta}|Y_i, \mathbf{W}_i) = \frac{\partial l(\boldsymbol{\theta}|Y_i, \mathbf{W}_i)}{\partial \boldsymbol{\theta}}$$

where

$$l(\boldsymbol{\theta}|Y_i, \mathbf{W}_i) = \log(L(\boldsymbol{\theta}|Y_i, \mathbf{W}_i))$$

denotes the contribution to the log likelihood from the observed data on subject i .

If the assumed parametric model is correctly specified, the MLE is a consistent estimator of $\boldsymbol{\theta}_0$, the true value of $\boldsymbol{\theta}$, under the conditions described in Section 2.5.5. Furthermore, the MLE is asymptotically normally distributed with variance covariance matrix equal to the inverse of the expected information matrix, denoted $I_n(\boldsymbol{\theta}_0)$ which is equal to:

$$I_n(\boldsymbol{\theta}) = - \sum_{i=1}^n \mathbb{E} \left(\frac{\partial^2 l(\boldsymbol{\theta}|Y_i, \mathbf{W}_i)}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}} \right). \quad (3.41)$$

The expected information matrix can be used to estimate the variance covariance matrix of the MLE $\hat{\boldsymbol{\theta}}$ by using $I_n(\hat{\boldsymbol{\theta}})^{-1}$. Alternatively, we may base inferences on the observed information matrix $I_n(\hat{\boldsymbol{\theta}})$, where:

$$I_n(\boldsymbol{\theta}) = \sum_{i=1}^n - \frac{\partial^2 l(\boldsymbol{\theta}|Y_i, \mathbf{W}_i)}{\partial \boldsymbol{\theta}^T \partial \boldsymbol{\theta}}. \quad (3.42)$$

In general, the observed information matrix is preferred to the expected information matrix [20]. Based on the asymptotic normality of MLEs, confidence intervals can be calculated and significance tests can be performed.

3.5.3 Maximum likelihood estimation via Newton Raphson

The most common approach to finding the root of likelihood score equations (equation (3.40)) in general is the Newton Raphson method [41]. Starting with an initial estimate of the model parameters, $\boldsymbol{\theta}^{(0)}$, successive estimates of $\boldsymbol{\theta}$ are given by:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + I_n(\boldsymbol{\theta}^{(t)})^{-1} S_n(\boldsymbol{\theta}^{(t)})$$

An alternative to using the observed information matrix $I_n(\boldsymbol{\theta}^{(t)})$ is to use the expected information matrix $I_n(\boldsymbol{\theta}^{(t)})$, which is known as the method of Fisher scoring.

In order to use Newton Raphson to find the MLE, we must have available the score and information matrices, which are found as the first and second derivatives of the log likelihood function with respect to the model parameters $\boldsymbol{\theta}$. Although the log likelihood function for the model we have defined in Section 3.5.1 can be differentiated twice with respect to the model parameters, the expressions are relatively complicated because of the non-standard mean and variance covariance structure corresponding to $f(Y_i, \mathbf{W}_i)$.

We note that there are many other methods for solving non-linear optimization problems, some of which only require evaluation of the function to be maximized. Others may require only the first derivatives of the log likelihood with respect to the parameters. Another approach is to use a gradient based maximization method, using numerical differentiation to estimate the required derivatives.

3.5.4 Maximum likelihood estimation via Expectation Maximization

The Expectation Maximization algorithm (EM) is a procedure which can be used to find MLEs for problems where directly maximizing the likelihood function is difficult but where maximizing the complete data likelihood is easier [42]. It is particularly suited to situations in which data are missing, or where unobserved latent variables can be treated as missing data. Schafer was one of the first to propose using the EM algorithm to find ML estimates for a parametric model which included covariate measurement error [38]. In a later paper, Schafer and Purdy described the implementation of the EM algorithm for the parametric model we described in Section 3.5.1. This involves treating the unobserved X_i as missing data. In contrast to the Newton Raphson method, the EM algorithm does not require the calculation of the observed data score function or the information matrix. This however comes at a cost. The EM algorithm does not automatically provide a measure of precision for the MLE, and its rate of convergence is generally much slower than methods such as Newton Raphson.

As with Newton Raphson, EM is an iterative algorithm that requires an initial estimate of the model parameters, which we denote $\boldsymbol{\theta}^{(0)}$. At iteration $t + 1$, given the current parameter estimate $\boldsymbol{\theta}^{(t)}$, the updated parameter estimate, $\boldsymbol{\theta}^{(t+1)}$ is found as follows. In the E step of EM, we find the expected value of the complete data log likelihood, conditional on the observed data and the current estimate of the parameter vector $\boldsymbol{\theta}^{(t)}$. This is denoted by $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$, and is given by:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^n \mathbb{E}(l(\boldsymbol{\theta}|Y_i, X_i, \mathbf{W}_i)|Y_i, \mathbf{W}_i, \boldsymbol{\theta}^{(t)}),$$

where $l(\boldsymbol{\theta}|Y_i, X_i, \mathbf{W}_i)$ denotes the contribution to the log likelihood function from observing Y_i, X_i, \mathbf{W}_i . The M step then consists of maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t)$ in $\boldsymbol{\theta}$, giving an updated parameter estimate $\boldsymbol{\theta}^{(t+1)}$. The process then repeats, with another application of the E step followed by the M-step. It can be shown that at each step, the observed data log likelihood of the updated parameter value is not less than that for the preceding parameter value [43]. Convergence can be judged by changes in parameter estimates between iterations, or alternatively by the increase in the observed data likelihood, if this is easily evaluated.

We briefly describe the EM algorithm for the parametric model we described previously in Section 3.5.1. The complete data log likelihood for this model is given by:

$$\begin{aligned}
& -\frac{n}{2}\log(\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2} \sum_{i=1}^n (Y_i - \beta_0 - \beta_X X_i)^2 \\
& \quad - \frac{\log(\sigma_U^2)}{2} \sum_{i=1}^n n_i - \frac{1}{2\sigma_U^2} \sum_{i=1}^n \sum_{j=1}^{n_i} (W_{ij} - X_i)^2 \\
& \quad - \frac{n}{2}\log(\sigma_X^2) - \frac{1}{2\sigma_X^2} \sum_{i=1}^n (X_i - \mu_X)^2 \quad (3.43)
\end{aligned}$$

In the E step of EM we find the expected value of this expression with respect to the distribution of (for each subject i) X_i given the observed data Y_i, \mathbf{W}_i and the current estimate of the parameter, $\boldsymbol{\theta}^{(t)}$. Since the complete data log likelihood is linear in X_i and X_i^2 , in the E-step we must find $\mathbb{E}(X_i|Y_i, \mathbf{W}_i, \boldsymbol{\theta}^{(t)})$ and $\mathbb{E}(X_i^2|Y_i, \mathbf{W}_i, \boldsymbol{\theta}^{(t)})$. These expectations follow from standard results for multivariate normal distributions, and are given by Schafer and Purdy [32].

In the M-step, we maximize this expected complete data log likelihood function. Because the three parts of the complete data log likelihood corresponding to the outcome model, the measurement error model, and the true covariate model, have no parameters in common, we can maximize the three parts independently. We illustrate the M-step for the last part of the complete data model, i.e. that for the true covariate. Expanding this part of the complete data log likelihood we have

$$-\frac{n}{2}\log(\sigma_X^2) - \frac{1}{2\sigma_X^2} \sum_{i=1}^n (X_i^2 + \mu_X^2 - 2\mu_X X_i)$$

In the E-step we take expectations conditional on $Y_i, \mathbf{W}_i, \boldsymbol{\theta}^{(t)}$, giving (suppressing the conditioning argument):

$$-\frac{n}{2}\log(\sigma_X^2) - \frac{1}{2\sigma_X^2} \sum_{i=1}^n (\mathbb{E}(X_i^2) + \mu_X^2 - 2\mu_X \mathbb{E}(X_i))$$

In the M-step we maximize this expected complete data log likelihood. To find the updated estimate of μ_X we differentiate with respect to μ_X as usual and set the resulting expression to zero, resulting in an updated estimate:

$$\frac{\sum \mathbb{E}(X_i)}{n}$$

This is just the sample mean, with the unobserved X_i replaced by their conditional expectations given the observed data and current parameter estimate. For the variance σ_X^2 , differentiating and setting to zero gives an updated estimate:

$$\frac{\sum \mathbb{E}(X_i^2)}{n} - \left(\frac{\sum \mathbb{E}(X_i)}{n} \right)^2$$

If the X_i were observed, recall that the usual MLE of σ_X^2 is given by:

$$\frac{\sum X_i^2}{n} - \left(\frac{\sum X_i}{n} \right)^2$$

and so an iteration of EM does *not* involve simply calculating the usual complete data MLE, replacing the unobserved X_i by $\mathbb{E}(X_i)$. Instead, here we see that the updated parameter estimate is given by the usual MLE, but with X_i replaced by $\mathbb{E}(X_i)$, and X_i^2 replaced by $\mathbb{E}(X_i^2)$. This result applies whenever the complete data density (or in this case a component of it which shares no parameters with the remainder) is from the regular exponential family [43]. In these cases, the complete data log likelihood is a linear function of the complete data sufficient statistics. The E-step thus involves calculating the conditional expectation of the complete data sufficient statistics. The M-step then consists of calculating the usual complete data MLE, with the sufficient statistics replaced by their conditional expectations, as calculated in the E-step. In the case of the parametric model we have described, closed form expressions are available for the MLEs based on complete data.

3.6 Maximum likelihood estimation using linear mixed models

In general, statistical packages do not include specific commands to fit models which allow for covariate measurement error using ML. Notable exceptions include the `cme` wrapper for GLLAMM in Stata [44], the latent variable modelling software `Mplus` [45], and `PROC NLMIXED` in SAS, which can be adapted to fit joint models [46]. If the observed data likelihood function is available in closed form, as it is for the model described in Section 3.5.1, then it is also possible to use general purpose maximization routines, such as Stata's `ml` command, to find MLEs. This however still usually requires the user to supply the expressions for calculating the

first and possibly second derivatives of the likelihood function, although numerical differentiation routines could be used. Alternatively, the EM algorithm can be used, which requires some programming effort.

In this section we shall show that the MLE for the parametric model we described in Section 3.5.1 can instead be found by fitting a standard linear mixed model [27]. Standard linear mixed models can be fitted using most modern statistical packages, and our approach does not require the user to write any custom programs.

First we recall that under our assumed model, the observed data for subject i , which consists of their outcome Y_i and error-prone measurements \mathbf{W}_i are jointly normal, with mean and variance covariance matrix:

$$\begin{aligned}\mathbb{E} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} &= \begin{pmatrix} \beta_0 + \beta_X \mu_X \\ \mu_X \mathbf{1}_{n_i \times 1} \end{pmatrix}, \\ \text{Var} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} &= \begin{pmatrix} \beta_X^2 \sigma_X^2 + \sigma_\epsilon^2 & \beta_X \sigma_X^2 \mathbf{1}_{1 \times n_i} \\ \beta_X \sigma_X^2 \mathbf{1}_{n_i \times 1} & \sigma_X^2 \mathbf{1}_{n_i \times n_i} + \sigma_U^2 \mathbf{I}_{n_i \times n_i} \end{pmatrix}\end{aligned}$$

Our strategy is to factorise the contribution to the log likelihood function from subject i as:

$$\begin{aligned}l(\boldsymbol{\theta}|Y_i, \mathbf{W}_i) &= \log(f(Y_i, \mathbf{W}_i|\boldsymbol{\theta})) \\ &= \log(f(Y_i|\boldsymbol{\theta})) + \log(f(\mathbf{W}_i|Y_i, \boldsymbol{\theta}))\end{aligned}$$

where $\boldsymbol{\theta} = (\beta_0, \beta_X, \sigma_\epsilon^2, \mu_X, \sigma_X^2, \sigma_U^2)^T$ is the vector of model parameters. Marginally, Y_i is normally distributed $Y_i \sim N(\mu_Y, \sigma_Y^2)$ where:

$$\mu_Y = \beta_0 + \beta_X \mu_X \quad (3.44)$$

$$\sigma_Y^2 = \beta_X^2 \sigma_X^2 + \sigma_\epsilon^2 \quad (3.45)$$

Furthermore, since X_i and Y_i are jointly normal, X_i given Y_i is also normal. We can therefore write:

$$X_i = \gamma_0 + \gamma_Y Y_i + b_i \quad (3.46)$$

where $b_i \sim N(0, \sigma_{X|Y}^2)$ is an independent normally distributed residual and:

$$\gamma_0 = \mu_X - \frac{\beta_X \sigma_X^2 (\beta_0 + \beta_X \mu_X)}{\beta_X^2 \sigma_X^2 + \sigma_\epsilon^2} \quad (3.47)$$

$$\gamma_Y = \frac{\beta_X \sigma_X^2}{\beta_X^2 \sigma_X^2 + \sigma_\epsilon^2} \quad (3.48)$$

$$\sigma_{X|Y}^2 = \sigma_X^2 - \frac{\beta_X^2 \sigma_X^4}{\beta_X^2 \sigma_X^2 + \sigma_\epsilon^2} \quad (3.49)$$

follow from standard results for multivariate normal distributions (e.g. Appendix S3 of [24]). Then since $W_{ij} = X_i + U_{ij}$, substituting for X_i using equation (3.46) we have

$$W_{ij} = \gamma_0 + \gamma_Y Y_i + b_i + U_{ij}$$

where $U_{ij} \sim N(0, \sigma_U^2)$ is independent of b_i . This shows that W_{ij} given Y_i follows a standard random-intercepts model, with random-intercepts variance $\sigma_{X|Y}^2$, within-subject variance σ_U^2 , and a fixed effect of Y_i .

This new parameter vector $\boldsymbol{\phi} = (\mu_Y, \sigma_Y^2, \gamma_0, \gamma_Y, \sigma_{X|Y}^2, \sigma_U^2)$ is a one-to-one function of the original model parameter vector $\boldsymbol{\theta} = (\beta_0, \beta_X, \sigma_\epsilon^2, \mu_X, \sigma_X^2, \sigma_U^2)$. Furthermore, in terms of the new parametrization $\boldsymbol{\phi}$, $f(Y_i)$ and $f(\mathbf{W}_i|Y_i)$ share no parameters. Since the two subsets of parameters are variationally independent, it follows that the ML estimate of $\boldsymbol{\phi}$ can be obtained by maximizing the two likelihood components separately [47].

For the first part of the log likelihood, the MLEs of μ_Y and σ_Y^2 are simply the sample mean and the biased sample variance (i.e. with n in the denominator rather than $n-1$). For the second part of the log likelihood, standard software can be used to fit the random intercept model for \mathbf{W}_i given Y_i , using ML, giving estimates $\hat{\gamma}_0$, $\hat{\gamma}_Y$, $\hat{\sigma}_{X|Y}^2$, $\hat{\sigma}_U^2$.

Using equation (3.46), we can express σ_X^2 as $\gamma_Y^2 \sigma_Y^2 + \sigma_{X|Y}^2$. Then, since γ_Y can be expressed as $\text{Cov}(Y_i, X_i)/\sigma_Y^2$, it follows that β_X can be expressed in terms of $\boldsymbol{\phi}$ by:

$$\beta_X = \frac{\text{Cov}(Y_i, X_i)}{\sigma_X^2} = \frac{\gamma_Y \sigma_Y^2}{\gamma_Y^2 \sigma_Y^2 + \sigma_{X|Y}^2}. \quad (3.50)$$

The ML estimate of β_X can thus be calculated as:

$$\hat{\beta}_X = \frac{\hat{\gamma}_Y \hat{\sigma}_Y^2}{\hat{\sigma}_{X|Y}^2 + \hat{\gamma}_Y^2 \hat{\sigma}_Y^2} \quad (3.51)$$

3.6.1 Inference

We now show how a Wald-type confidence interval for $\hat{\beta}_X$, the MLE of β_X , can be calculated using output from the fitted linear mixed model and a standard formula for the variance of the biased sample variance. First we note that $\hat{\sigma}_Y^2$ is asymptotically uncorrelated with both $\hat{\gamma}_Y$ and $\hat{\sigma}_{X|Y}^2$. This follows from the fact that the likelihood function factors into two components such that σ_Y^2 is only involved in one part and γ_Y and $\sigma_{X|Y}^2$ are only involved in the other part. Furthermore, a standard result for linear mixed models is that the estimators of fixed effects parameters are asymptotically uncorrelated with the estimators of the variance component parameters [24]. Thus $\hat{\gamma}_Y$ and $\hat{\sigma}_{X|Y}^2$ are asymptotically uncorrelated, and so for large sample

sizes:

$$(\hat{\sigma}_Y^2, \hat{\gamma}_Y, \hat{\sigma}_{X|Y}^2)^T \sim N \left((\sigma_Y^2, \gamma_Y, \sigma_{X|Y}^2)^T, \begin{pmatrix} \text{Var}(\hat{\sigma}_Y^2) & 0 & 0 \\ 0 & \text{Var}(\hat{\gamma}_Y) & 0 \\ 0 & 0 & \text{Var}(\hat{\sigma}_{X|Y}^2) \end{pmatrix} \right).$$

Then by the multivariate delta method [28], it follows that in large samples:

$$\hat{\beta}_X \sim N(\beta_X, \mathbf{J} \text{Var}(\hat{\sigma}_Y^2, \hat{\gamma}_Y, \hat{\sigma}_{X|Y}^2) \mathbf{J}^T)$$

where \mathbf{J} denotes the Jacobian matrix of the transformation $(\sigma_Y^2, \gamma_Y, \sigma_{X|Y}^2) \mapsto \beta_X$:

$$\mathbf{J} = \begin{pmatrix} \frac{\partial \beta_X}{\partial \sigma_Y^2} & \frac{\partial \beta_X}{\partial \gamma_Y} & \frac{\partial \beta_X}{\partial \sigma_{X|Y}^2} \end{pmatrix}$$

Partial differentiation of equation 3.50 then gives, after some simplification:

$$\text{Var}(\hat{\beta}_X) = \frac{\gamma_Y^2 \sigma_{X|Y}^4 \text{Var}(\hat{\sigma}_Y^2) + \sigma_Y^4 (\sigma_{X|Y}^2 - \sigma_Y^2 \gamma_Y^2)^2 \text{Var}(\hat{\gamma}_Y) + \gamma_Y^2 \sigma_Y^4 \text{Var}(\hat{\sigma}_{X|Y}^2)}{(\sigma_{X|Y}^2 + \gamma_Y^2 \sigma_Y^2)^2}$$

This variance can be estimated consistently by replacing each parameter by its respective estimate. Using the observed information matrix one can estimate the variance of $\hat{\sigma}_Y^2$ by $\frac{2\hat{\sigma}_Y^4}{n}$. A standard error for $\hat{\gamma}_Y$ is given by the linear mixed model output of software packages, which is usually based on inverting the block of the observed information matrix corresponding to the fixed effects. Packages such as Stata and SAS also give estimated standard errors for parameters corresponding to variance components (e.g. $\sigma_{X|Y}^2$ and σ_U^2), which are based either on inverting the observed information matrix for all model parameters, or on inverting the sub-matrix of the observed information corresponding to the variance components. An approximate 95% confidence interval for $\hat{\beta}_X$ can be thus found as $\hat{\beta}_X \pm 1.96 \sqrt{\widehat{\text{Var}}(\hat{\beta}_X)}$.

3.6.2 Robustness to modelling assumptions

One reason that researchers may be wary of using techniques such as parametric ML to deal with covariate measurement error is that they may make stronger parametric modelling assumptions compared to simpler methods such as MOM and RC. It is therefore important to consider the robustness of an approach which makes strong parametric assumptions to violations of these assumptions. The assumed parameter model includes an assumption of normality for X_i , U_{ij} and ϵ_i . If X_i and ϵ_i are not both normally distributed, the linear mixed model for \mathbf{W}_i given Y_i will in general be misspecified in a number of ways. First, the random effects b_i are not normally distributed. Second, the variance of these random effects may vary as a function of Y_i . Lastly, the conditional mean function $E(X_i|Y_i)$ may be a more complicated function of Y_i , and so the fixed effects structure of the mixed model may be misspec-

ified. In this section we show that the unbiasedness of the likelihood score equations for the linear mixed model does not rely on these normality assumptions. Similarly, the ML estimator of variance for σ_Y^2 is clearly consistent regardless of the marginal distribution of Y_i . It thus follows that $\hat{\beta}_{ML}$ remains consistent even if some or all of the normality assumptions are violated.

Irrespective of whether X_i and ϵ_i are both normally distributed, we can always express X_i in terms of its best linear prediction given Y_i as:

$$X_i = \gamma_0 + \gamma_Y Y_i + b_i \quad (3.52)$$

where $\mathbb{E}(b_i) = 0$, $\text{Cov}(Y_i, b_i) = 0$, and γ_0 and γ_Y are as given in equations (3.47) and (3.48). The random term b_i is no longer necessarily normally distributed, and its variance may vary as a function of Y_i . The expression in equation (3.49) for $\sigma_{X|Y}^2$ then gives the mean conditional variance of the b_i . Thus Y_i and b_i are not necessarily independent, although they are uncorrelated. We can therefore write:

$$\begin{aligned} \mathbf{W}_i &= X_i \mathbf{1}_{n_i \times 1} + \mathbf{U}_i \\ &= \gamma_0 \mathbf{1}_{n_i \times 1} + \gamma_Y Y_i \mathbf{1}_{n_i \times 1} + b_i \mathbf{1}_{n_i \times 1} + \mathbf{U}_i. \end{aligned} \quad (3.53)$$

We now let $\boldsymbol{\gamma} = (\gamma_0, \gamma_Y)^T$ and let \mathcal{X}_i denote the $n_i \times 2$ design matrix for the fixed effects in the above model. This is simply a matrix whose first column has entries equal to one and second column with entries Y_i . In this notation:

$$\mathbf{W}_i = \mathcal{X}_i \boldsymbol{\gamma} + b_i \mathbf{1}_{n_i \times 1} + \mathbf{U}_i. \quad (3.54)$$

From standard linear mixed model theory (e.g. Chapter 6 of [24]), the likelihood score corresponding to the fixed effects of the linear mixed model for \mathbf{W}_i given Y_i can be expressed as:

$$\boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathbf{W}_i - \boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathcal{X}_i \boldsymbol{\gamma} \quad (3.55)$$

where $\mathbf{V}_i = \sigma_{X|Y}^2 \mathbf{1}_{n_i \times n_i} + \sigma_U^2 \mathbf{I}_{n_i}$ denotes the variance covariance matrix of \mathbf{W}_i given Y_i under the assumed model. We now substitute for \mathbf{W}_i using equation (3.54), which gives

$$\begin{aligned} \boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} (\mathcal{X}_i \boldsymbol{\gamma} + b_i \mathbf{1}_{n_i \times 1} + \mathbf{U}_i) - \boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathcal{X}_i \boldsymbol{\gamma} \\ = \boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathbf{1}_{n_i \times 1} b_i + \boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathbf{U}_i. \end{aligned} \quad (3.56)$$

Now since $\mathbb{E}(b_i) = 0$, the expectation of the first term (conditional on n_i) gives:

$$\mathbb{E}(\boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathbf{1}_{n_i \times 1} b_i) = \text{Cov}(\boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathbf{1}_{1 \times n_i}, b_i) = \mathbf{0}$$

because the only random component of χ_i is Y_i , and $\text{Cov}(Y_i, b_i) = 0$. Because of the assumed independence between U_{ij} and both X_i and ϵ_i , $\mathbb{E}(\boldsymbol{\chi}_i^T \mathbf{V}_i^{-1} \mathbf{U}_i) = 0$.

The likelihood score corresponding to $\sigma_{X|Y}^2$ is given by:

$$-0.5\text{tr}(\mathbf{V}_i^{-1} \mathbf{1}_{n_i \times n_i}) + 0.5(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma})^T \mathbf{V}_i^{-1} \mathbf{1}_{n_i \times n_i} \mathbf{V}_i^{-1} (\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}) \quad (3.57)$$

where $\text{tr}()$ denotes the trace of a square matrix. By a standard result for the expectation of quadratic forms (e.g. Appendix S5 of [24]), the expectation of the second part of the likelihood score is equal to:

$$0.5(\mathbb{E}(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma})^T \mathbf{V}_i^{-1} \mathbf{1}_{n_i \times n_i} \mathbf{V}_i^{-1} \mathbb{E}(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}) + \text{tr}(\mathbf{V}_i^{-1} \mathbf{1}_{n_i \times n_i} \mathbf{V}_i^{-1} \text{Var}(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}))). \quad (3.58)$$

Then since $\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma} = \mathbf{1}_{n_i \times 1} b_i + \mathbf{U}_i$, it follows that $\mathbb{E}(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}) = \mathbf{0}$. Furthermore, the assumption of independence between the errors U_{ij} and both X_i and ϵ_i imply that b_i is uncorrelated with U_{ij} . This in turn means that $\text{Var}(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}) = \mathbf{V}_i$, and so the expectation of the second part of the score is equal to:

$$0.5\text{tr}(\mathbf{V}_i^{-1} \mathbf{1}_{n_i \times n_i} \mathbf{V}_i^{-1} \mathbf{V}_i) = 0.5\text{tr}(\mathbf{V}_i^{-1} \mathbf{1}_{n_i \times n_i}). \quad (3.59)$$

The expectation of the likelihood score corresponding to $\sigma_{X|Y}^2$ thus has expectation zero.

Lastly we consider the likelihood score component corresponding to σ_U^2 , which is given by:

$$-0.5\text{tr}(\mathbf{V}_i^{-1}) + 0.5(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma})^T \mathbf{V}_i^{-1} \mathbf{V}_i^{-1} (\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}). \quad (3.60)$$

Taking expectations, and as above using the fact that $\mathbb{E}(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}) = \mathbf{0}$, that $\text{Var}(\mathbf{W}_i - \mathcal{X}_i \boldsymbol{\gamma}) = \mathbf{V}_i$, and the rule for the expectation of quadratic forms, we have:

$$-0.5\text{tr}(\mathbf{V}_i^{-1}) + 0.5\text{tr}(\mathbf{V}_i^{-1} \mathbf{V}_i^{-1} \mathbf{V}_i) = 0. \quad (3.61)$$

Thus the unbiasedness of the likelihood score equations for the linear mixed model for \mathbf{W}_i given Y_i does not depend on any of the normality assumptions, and so the MLE of β_X (under the previously stated normality assumptions) is consistent even if some or all of the normality assumptions are violated.

3.6.3 Multiple covariates

We now extend our approach to the case of a p -dimensional \mathbf{X}_i and a q -dimensional error-free covariate \mathbf{Z}_i . We first define the joint model.

Outcome model

We assume that Y_i follows a linear regression model given \mathbf{X}_i and \mathbf{Z}_i :

$$Y_i = \beta_0 + \boldsymbol{\beta}_X^T \mathbf{X}_i + \boldsymbol{\beta}_Z^T \mathbf{Z}_i + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$ is independent of \mathbf{X}_i and \mathbf{Z}_i .

Covariate model

As in Section 3.4.7, we assume that \mathbf{X}_i is multivariate normal given \mathbf{Z}_i with:

$$\mathbb{E}(\mathbf{X}_i | \mathbf{Z}_i) = \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_Z \mathbf{Z}_i \quad (3.62)$$

$$\text{Var}(\mathbf{X}_i | \mathbf{Z}_i) = \boldsymbol{\Sigma}_{X|Z} \quad (3.63)$$

where $\boldsymbol{\Gamma}_Z$ is a $p \times q$ matrix of regression coefficients. It follows that $Y_i \sim N(\delta_0 + \boldsymbol{\delta}_Z^T \mathbf{Z}_i, \sigma_{Y|Z}^2)$ where:

$$\delta_0 = \beta_0 + \boldsymbol{\beta}_X^T \boldsymbol{\Gamma}_0 \quad (3.64)$$

$$\boldsymbol{\delta}_Z^T = \boldsymbol{\beta}_X^T \boldsymbol{\Gamma}_Z + \boldsymbol{\beta}_Z^T \quad (3.65)$$

$$\sigma_{Y|Z}^2 = \boldsymbol{\beta}_X^T \boldsymbol{\Sigma}_{X|Z} \boldsymbol{\beta}_X + \sigma_\epsilon^2. \quad (3.66)$$

Since Y_i and \mathbf{X}_i are jointly normal given \mathbf{Z}_i , and since $\text{Cov}(\mathbf{X}_i, Y_i | \mathbf{Z}_i) = \boldsymbol{\Sigma}_{X|Z} \boldsymbol{\beta}_X$, it follows from standard results for multivariate normal distributions that \mathbf{X}_i is normally distributed given both Y_i and \mathbf{Z}_i , with mean:

$$\mathbb{E}(\mathbf{X}_i | Y_i, \mathbf{Z}_i) = \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_Y Y_i + \boldsymbol{\gamma}_Z \mathbf{Z}_i \quad (3.67)$$

and conditional variance covariance matrix:

$$\text{Var}(\mathbf{X}_i | Y_i, \mathbf{Z}_i) = \boldsymbol{\Sigma}_{X|Y,Z} \quad (3.68)$$

where:

$$\boldsymbol{\gamma}_0 = \boldsymbol{\Gamma}_0 - \boldsymbol{\Sigma}_{X|Z} \boldsymbol{\beta}_X \sigma_{Y|Z}^{-2} \delta_0 \quad (3.69)$$

$$\boldsymbol{\gamma}_Y = \boldsymbol{\Sigma}_{X|Z} \boldsymbol{\beta}_X \sigma_{Y|Z}^{-2} \quad (3.70)$$

$$\boldsymbol{\gamma}_Z = \boldsymbol{\Gamma}_Z - \boldsymbol{\Sigma}_{X|Z} \boldsymbol{\beta}_X \sigma_{Y|Z}^{-2} \boldsymbol{\delta}_Z^T \quad (3.71)$$

$$\boldsymbol{\Sigma}_{X|Z,Y} = \boldsymbol{\Sigma}_{X|Z} - \boldsymbol{\Sigma}_{X|Z} \boldsymbol{\beta}_X \sigma_{Y|Z}^{-2} \boldsymbol{\beta}_X^T \boldsymbol{\Sigma}_{X|Z} \quad (3.72)$$

Measurement model

We assume the classical measurement error model as described in Section 2.5.1, so that $\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i$, where \mathbf{D}_i is a known design matrix, of dimension

$n_i = \sum_{j=1}^p n_{ij}$, consisting of 0s and 1s denoting which component of \mathbf{X}_i each error-prone measurement corresponds to. As in Section 3.4.7, we assume that $\mathbf{U}_i \sim N(\mathbf{0}, \text{Var}(\mathbf{U}_i))$ where:

$$\text{Var}(\mathbf{U}_i) = \sigma_{U_1}^2 \mathbf{I}_{n_{i1}} \oplus \sigma_{U_2}^2 \mathbf{I}_{n_{i2}} \oplus \dots \oplus \sigma_{U_p}^2 \mathbf{I}_{n_{ip}}. \quad (3.73)$$

The linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i

It then follows that:

$$\mathbb{E}(\mathbf{W}_i | Y_i, \mathbf{Z}_i) = \mathbf{D}_i(\gamma_0 + \gamma_Y Y_i + \gamma_Z \mathbf{Z}_i) \quad (3.74)$$

As in Section 3.4.7, we can express the mean function $\mathbb{E}(\mathbf{W}_i | Y_i, \mathbf{Z}_i)$ as a product of a known design matrix and an unknown parameter vector. To do this, we first express the $p \times q$ matrix of regression coefficients γ_Z in terms of its column vectors:

$$\gamma_Z = \begin{pmatrix} \gamma_{Z_1} & \dots & \gamma_{Z_q} \end{pmatrix} \quad (3.75)$$

where each column is a $p \times 1$ vector of regression coefficients. Then we can write:

$$\mathbb{E}(\mathbf{W}_i | Y_i, \mathbf{Z}_i) = \mathbf{D}_i \gamma_0 + Y_i \mathbf{D}_i \gamma_Y + Z_{i1} \mathbf{D}_i \gamma_{Z_1} + \dots + Z_{iq} \mathbf{D}_i \gamma_{Z_q} \quad (3.76)$$

$$= \begin{pmatrix} \mathbf{D}_i & Y_i \mathbf{D}_i & Z_{i1} \mathbf{D}_i & \dots & Z_{iq} \mathbf{D}_i \end{pmatrix} \begin{pmatrix} \gamma_0 \\ \gamma_Y \\ \gamma_{Z_1} \\ \vdots \\ \gamma_{Z_q} \end{pmatrix} \quad (3.77)$$

Therefore if we let:

$$\mathcal{D}_i = \begin{pmatrix} \mathbf{D}_i & Y_i \mathbf{D}_i & Z_{i1} \mathbf{D}_i & \dots & Z_{iq} \mathbf{D}_i \end{pmatrix} \quad (3.78)$$

and:

$$\boldsymbol{\gamma} = \left(\gamma_0^T \quad \gamma_Y^T \quad \gamma_{Z_1}^T \quad \dots \quad \gamma_{Z_q}^T \right)^T, \quad (3.79)$$

we have that:

$$\mathbb{E}(\mathbf{W}_i | Y_i, \mathbf{Z}_i) = \mathcal{D}_i \boldsymbol{\gamma} \quad (3.80)$$

for the $n_i \times (p \times (q + 2))$ design matrix \mathcal{D}_i and $(p \times (q + 2)) \times 1$ vector of unknown parameters $\boldsymbol{\gamma}$. Lastly, the variance covariance matrix is given by:

$$\text{Var}(\mathbf{W}_i | Y_i, \mathbf{Z}_i) = \mathcal{D}_i \boldsymbol{\Sigma}_{X|Z,Y} \mathcal{D}_i^T + \text{Var}(\mathbf{U}_i).$$

This means that \mathbf{W}_i follows a linear mixed model given Y_i and \mathbf{Z}_i , with fixed effects design matrix \mathcal{D}_i , random effects design matrix \mathbf{D}_i , random effects with variance covariance matrix $\boldsymbol{\Sigma}_{X|Z,Y}$, and residual errors with variance covariance matrix $\text{Var}(\mathbf{U}_i)$.

The likelihood component corresponding to $f(\mathbf{W}_i|Y_i, \mathbf{Z}_i)$ can thus be maximized by fitting this linear mixed model. The likelihood component corresponding to Y_i given \mathbf{Z}_i can be maximized by fitting the least squares regression of Y_i on \mathbf{Z}_i , giving the ML estimates of δ_0 , $\boldsymbol{\delta}_Z$ and $\sigma_{Y|Z}^2$. Analogous to the univariate X_i case, one can show that

$$\boldsymbol{\beta}_X = (\boldsymbol{\Sigma}_{X|Z,Y} + \sigma_{Y|Z}^2 \boldsymbol{\gamma}_Y \boldsymbol{\gamma}_Y^T)^{-1} \boldsymbol{\gamma}_Y \sigma_{Y|Z}^2 \quad (3.81)$$

and that

$$\boldsymbol{\beta}_Z = \boldsymbol{\delta}_Z - (\boldsymbol{\gamma}_Z^T + \boldsymbol{\delta}_Z \boldsymbol{\gamma}_Y^T) \boldsymbol{\beta}_X. \quad (3.82)$$

The MLEs of $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ can thus be calculated by inserting the MLEs of the relevant parameters using these formulae.

Although in principle Wald type confidence intervals can be found for the components of $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$, in practice it may be easier to use non-parametric bootstrapping.

3.7 Multiple imputation

One way of viewing the covariate measurement error problem is as a missing data problem, whereby the true covariate X_i is missing for all subjects (unless an internal validation dataset is available). Missing data in the conventional sense and covariate measurement error are examples of a more general process, known as ‘data coarsening’, where the data we observe (in the case of internal replication data, Y_i and \mathbf{W}_i) are a coarsened version of the underlying data (Y_i, X_i) [48]. Missing data, rounding, and covariate measurement error are all types of data coarsening. One of the most popular approaches to dealing with missing data has been the method of multiple imputation (MI). We first briefly introduce MI in the context in which it was originally conceived, that of missing data. We then discuss the application of MI to deal with covariate measurement error.

3.7.1 Multiple imputation for missing data

MI was first proposed by Rubin in the context of large population surveys, in which there are often missing values for variables for some subjects [49]. To describe MI we adopt the notation that \mathbf{Y}_{obs} and \mathbf{Y}_{mis} denote the observed and missing data

vectors for a given dataset (\mathbf{Y}_{obs} and \mathbf{Y}_{mis} are not used here to indicate that the variables are necessarily ‘outcomes’). The validity of approaches which ignore the missing data mechanism rely on the so called missing at random (MAR) assumption. This assumption states that the probability of missingness only depends on observed data \mathbf{Y}_{obs} , and not on the unobserved \mathbf{Y}_{mis} .

Bayesian multiple imputation

MI was originally conceived by Rubin within the Bayesian framework. For parametric MI, this involves specification of a parametric imputation model for the missing data given the observed, $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \boldsymbol{\xi})$, parametrized by $\boldsymbol{\xi}$. In Bayesian MI, we specify an uninformative prior distribution for $\boldsymbol{\xi}$, denoted $f(\boldsymbol{\xi})$. A number $M > 1$ imputations of the missing data \mathbf{Y}_{mis} are then generated by taking random draws from the posterior distribution $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, f(\boldsymbol{\xi}))$. One way of doing this is to first draw a new parameter value $\boldsymbol{\xi}^{(m)}$ from the posterior distribution $f(\boldsymbol{\xi}|\mathbf{Y}_{obs}, f(\boldsymbol{\xi}))$, and then to generate the imputed values from $f(\mathbf{Y}_{mis}|\mathbf{Y}_{obs}, \boldsymbol{\xi}^{(m)})$, treating the model parameter as if it is known to be equal to $\boldsymbol{\xi}^{(m)}$. Such an approach creates ‘proper’ imputations, meaning that the uncertainty about $\boldsymbol{\xi}$ is reflected in variability between imputed datasets [43].

Usually the most difficult aspect of MI is creating the M imputations. For certain missing data patterns and choices of imputation model, the posterior distributions are standard distributions, thus allowing one to easily generate imputations [43]. In general however, especially when a dataset has a complicated pattern of missingness, the posterior distributions needed to generate imputations cannot be expressed in closed form. This means that one cannot directly sample from the posterior distributions of interest. To overcome this problem, Markov Chain Monte Carlo (MCMC) methods such as Gibbs sampling have successfully have been used. These are a collection of methods which can be used to draw from a ‘target distribution’ from which it is difficult or impossible to draw directly [50]. MCMC methods involve repeatedly drawing from distributions which are easy to sample from, creating a chain of draws. Under certain conditions, the distribution of the draws from the chain converges to the ‘target distribution’. Judging whether the chain has converged to the target distribution is however, non-trivial, and many tools have been developed which can be used to check for non-convergence [50].

Having created the imputations, we now suppose that we are interested in estimating some scalar parameter ψ . The extension of these results to vector valued $\boldsymbol{\psi}$ is given for example by Schafer [43]. Given the M completed datasets, we apply the estimation procedure which we would have used had there been no missing data to each of the M completed datasets, giving M estimates $\psi^{(m)}, m = 1, \dots, M$. These M parameter estimates are then averaged to give an overall point estimate. Rubin

proposed the MI estimator of ψ as:

$$\hat{\psi} = \frac{\sum_{m=1}^M \hat{\psi}^{(m)}}{M}$$

One of the reasons that MI has proved so popular with researchers is that Rubin proposed a variance estimator for $\hat{\psi}$ that is particularly simple to compute, and from which significance tests and confidence/credible intervals can be calculated. Rubin's variance formula only requires that an estimate of the full data standard error of parameter estimates can be obtained from each imputed dataset, which are usually available from statistical packages when a standard analysis method is used to estimate ψ . Denoting an estimate of the variance of $\hat{\psi}^{(m)}$ based on the m th imputed dataset, referred to as a within-imputation estimate of variance, by $\widehat{\text{Var}}(\hat{\psi}^{(m)})$, we define the average within-imputation variance as:

$$W = \frac{1}{M} \sum_{m=1}^M \widehat{\text{Var}}(\hat{\psi}^{(m)})$$

Similarly, the between-imputation variance is defined as the sample variance of the estimates $\hat{\psi}^{(m)}$, $m = 1, \dots, M$:

$$B = \frac{1}{M-1} \sum_{m=1}^M (\hat{\psi}^{(m)} - \hat{\psi})^2$$

Rubin's variance estimator for $\hat{\psi}$ is then given by:

$$W + \frac{1}{1 + M^{-1}} B$$

The variance of the MI estimator is thus greater than the average of the within-imputation variances, unless $B = 0$, showing that it is essential to allow for between-imputation variability in point estimates. If there is little variability in point estimates of ψ between imputations, $B \approx 0$ and the variance of the MI estimator can be estimated by the average of the within-imputation variances.

An attractive feature of MI is that in terms of efficiency of parameter estimation, a relatively small number M of imputations may be sufficient if the information in the missing data \mathbf{Y}_{mis} is small relative to that in the observed data \mathbf{Y}_{obs} [49]. However, researchers may be uncomfortable with the fact that if they were to re-run the analysis slightly different point estimates for ψ would be obtained. For this reason, using a larger number of imputations may be preferable [51].

Frequentist multiple imputation

An alternative to Bayesian multiple imputation is so called 'frequentist multiple imputation' [52]. Rather than creating the imputations using different values of the

imputation model parameter ξ , in frequentist MI we create M imputed datasets by using the same estimated value of ξ . To do this, we must estimate ξ using the observed data, one example being the observed data MLE.

There are a number of possible advantages to frequentist MI over Bayesian MI. First, one does not need to consider the issue of prior distributions for the imputation model parameters. Second, for a fixed number of imputations M , frequentist MI results in more efficient estimates of ψ than Bayesian MI [53]. This occurs because in Bayesian MI extra random variability is introduced through the use of different values of the imputation model parameter ξ when creating the imputations. This difference in efficiency goes to zero however as $M \rightarrow \infty$. The major disadvantage of frequentist MI however is that applying Rubin’s rules does not result in valid inferences, because the imputations are all generated using a single value of the imputation model parameter ξ .

3.7.2 Multiple imputation for measurement error

Multiple imputation has recently been proposed as an approach for correcting for measurement error and covariate misclassification, whereby the unobserved covariate X_i is treated as missing data. Messer and Natarajan [46] and Cole and Greenland [54] both considered the case in which internal validation data are available. In this case, because X_i is observed on a subset of subjects, standard software for MI can be used to impute the missing X_i values.

In the case of internal replication data, X_i is missing for every subject, and so standard software for MI cannot be used. We now show how frequentist MI can be implemented in the case of internal replication data, for a particular parametric model, and compare it with ML. We then discuss the recent results of Freedman *et al* [21], who investigated implementations of MI, moment reconstruction (MR) (see Section 3.8), and RC.

Multiple imputation with internal replication data

In order to use MI, we must impute X_i from its conditional distribution given all other observed data, i.e. $f(X_i|\mathbf{W}_i, Y_i)$. We assume the parametric model defined in Section 3.5.1. We showed in Section 3.6 that under this model, \mathbf{W}_i given Y_i follows a random-intercepts model:

$$W_{ij} = \gamma_0 + \gamma_Y Y_i + b_i + U_{ij}$$

where $b_i \sim N(0, \sigma_{X|Y}^2)$ is a random effect representing the residual from the regression of X_i on Y_i . Under this model, the conditional distribution $f(X_i|\mathbf{W}_i, Y_i)$ is

normal, with mean:

$$\gamma_0 + \gamma_Y Y_i + \mathbb{E}(b_i | \mathbf{W}_i, Y_i)$$

and variance $\text{Var}(b_i | \mathbf{W}_i, Y_i)$. The conditional mean of the random effect, $\mathbb{E}(b_i | \mathbf{W}_i, Y_i)$, is the BLUP of the random effect, which from standard results for linear mixed models (e.g. Section 7.2 of [37]) is given by:

$$\mathbb{E}(b_i | \mathbf{W}_i, Y_i) = \sigma_{X|Y}^2 \mathbf{1}_{1 \times n_i} \mathbf{V}_i^{-1} (\mathbf{W}_i - \gamma_0 \mathbf{1}_{n_i \times 1} - \gamma_Y Y_i \mathbf{1}_{n_i \times 1})$$

where $\mathbf{V}_i = \sigma_{X|Y}^2 \mathbf{1}_{n_i \times n_i} + \sigma_U^2 I_{n_i \times n_i}$ denotes the variance covariance matrix of \mathbf{W}_i given Y_i . By a result from matrix algebra (e.g. Appendix M1 of [24]):

$$\mathbf{V}_i^{-1} = \frac{1}{\sigma_U^2} \left(I_{n_i \times n_i} - \frac{\sigma_{X|Y}^2}{n_i \sigma_{X|Y}^2 + \sigma_U^2} \mathbf{1}_{n_i \times n_i} \right).$$

After some algebra, it then follows that:

$$\mathbb{E}(b_i | \mathbf{W}_i, Y_i) = \frac{\sigma_{X|Y}^2}{\sigma_{X|Y}^2 + \frac{\sigma_U^2}{n_i}} (\bar{W}_i - \gamma_0 - \gamma_Y Y_i) \quad (3.83)$$

where \bar{W}_i denotes the mean of subject i 's n_i measurements. We note that the factor by which \bar{W}_i is multiplied by is equal to the reliability ratio of the error-prone measurements but with σ_X^2 replaced by the conditional variance of X_i given Y_i . The conditional variance $\text{Var}(b_i | \mathbf{W}_i, Y_i)$ is then given by:

$$\begin{aligned} \text{Var}(b_i | \mathbf{W}_i, Y_i) &= \sigma_{X|Y}^2 - \left(\frac{\sigma_{X|Y}^2}{\sigma_{X|Y}^2 + \frac{\sigma_U^2}{n_i}} \right)^2 \text{Var}(\bar{W}_i | Y_i) \\ &= \sigma_{X|Y}^2 - \left(\frac{\sigma_{X|Y}^2}{\sigma_{X|Y}^2 + \frac{\sigma_U^2}{n_i}} \right)^2 \left(\sigma_{X|Y}^2 + \frac{\sigma_U^2}{n_i} \right) \\ &= \sigma_{X|Y}^2 \left(1 - \frac{\sigma_{X|Y}^2}{\sigma_{X|Y}^2 + \frac{\sigma_U^2}{n_i}} \right). \end{aligned} \quad (3.84)$$

Thus, having fitted the linear mixed model for \mathbf{W}_i given Y_i using ML, we can create M multiple imputations of X_i by drawing from a normal distribution with mean and variance as given in equations (3.83) and (3.84). We can then regress Y_i on the M imputed datasets, yielding M estimates of β_X . The arithmetic mean of these estimates can then be used as an estimate of β_X .

Relationship to maximum likelihood

Wang and Robins investigated the frequentist properties of both Bayesian MI, and what we (following Tsiatis [52]) have called frequentist MI (termed improper MI by Wang and Robins) [53]. In particular, Wang and Robins showed that, asymptotically (as the number of subjects increases), the frequentist MI estimator (imputing using the observed data MLE, as we have proposed) is inefficient when a finite number M of imputations is used. This is due to the Monte-Carlo error caused by using a finite number of imputations. As the number of imputations $M \rightarrow \infty$, Wang and Robins state that using frequentist MI corresponds to a single step of the EM algorithm [53]. Since our proposal involves imputing using the observed data MLE, the results of Wang and Robins imply the MI estimator will converge to the MLE as $M \rightarrow \infty$. In simulations (described later in Section 3.9), we found that the point estimate obtained from frequentist MI with $M = 25$ imputations was almost identical to the MLE. Of course, for this model, the observed data MLEs can be easily found by fitting the previously described mixed model to \mathbf{W}_i given Y_i . Since these are asymptotically optimal in terms of efficiency, there is little to gain in using MI.

Inference

The MI we have described for measurement error thus far is frequentist MI. Robins and Wang have described a sandwich estimator of variance for use with frequentist MI, but this requires additional analytical work [55]. Alternatively, one could use non-parametric bootstrapping [31]. This would involve applying MI with a fixed number M imputations to each bootstrap sample, which may therefore be prohibitively computationally expensive. A further alternative is to use proper MI and then use Rubin's rules. One simple approximate approach to creating proper imputations is to sample new values of the imputation model parameters from a normal distribution, with mean equal to the MLE of the imputation model parameters and variance covariance equal to the estimated variance covariance of the estimates [56]. Rubin's rules can then be applied to the resulting MI estimator to perform significance tests and find confidence intervals. Another approach to creating approximate draws from the posterior of the imputation model parameters is, for each imputation m , to use the the MLE of the imputation model parameters found from a bootstrapped sample of the original data.

Multiple imputation for other model setups

Freedman *et al* recently compared the performance of RC, a version of moment reconstruction (see Section 3.8), and a version of MI to deal with covariate measurement error in linear (and logistic) regression [21]. Freedman *et al* considered a

situation in which each subject has a measurement which adheres to a ‘non-classical’ error model:

$$W_i = \gamma_0 + \gamma_X X_i + \delta_i$$

where δ_i is mean zero error, independent of X_i and Y_i . Note that γ_0 and γ_X are not necessarily equal to zero and one respectively, as they are under the classical error model. In their setup, a random subset of subjects in an internal calibration study had two error-prone measurements M_{i1} and M_{i2} which follow a classical error model, with error variance different to that of W_i . Under multivariate normality assumptions, Freedman *et al* showed how an MI approach could be implemented for their particular assumed setting. Their approach involved fitting a multivariate normal model for the two unbiased measurements M_{i1} and M_{i2} given the biased error-prone measurement W_i and Y_i in internal calibration study subjects. Based on the parameter estimates from this, multiple imputations of X_i were generated, both for subjects in the main study and those in the internal calibration study.

Crucially, their implementation of MI did not utilize the assumption of non-differential error. In simulations Freedman *et al* found that their MI estimator could be substantially less efficient than a weighted RC estimator. They validated their simulation results by deriving approximate expressions for the asymptotic variances of the various estimators. They concluded that the inefficiency of their MI and MR estimators was likely due to the fact their implementation of these did not utilize the non-differential error assumption. Although not shown in the paper, Freedman *et al* reported that they also considered versions of MI and MR which did utilize this assumption. Although more complicated to implement, Freedman *et al* reported that these versions appeared to have comparable efficiency to the weighted RC estimator.

We note that for the model considered by Freedman *et al*, the joint distribution is multivariate normal and therefore the likelihood function is available in closed form. We do not believe our approach using linear mixed models (Section 3.6) can be applied to this specific model, because the coefficient of X_i in the model for W_i is not constrained to be one. However, since the likelihood function is equal to a multivariate normal density, we believe numerical maximization routines, such as Stata’s `ml` command or SAS PROC NLMIXED, could be used to find the MLE.

3.8 Moment reconstruction

Recently Freedman *et al* proposed a novel method for dealing with measurement error called moment reconstruction (MR) [57]. It combines features of regression calibration, through non-stochastic imputation of a new variable to substitute for

X_i , and multiple imputation, by using the outcome Y_i to make the imputation. It can be applied when the outcome model is either linear or logistic regression.

For simplicity, we first describe MR in the case in which a single scalar covariate X_i is measured with classical error by a single error-prone measurement $W_i = X_i + U_i$. Furthermore, we assume that $\mathbb{E}(W_i|Y_i) = \mathbb{E}(X_i|Y_i)$. Since:

$$\begin{aligned}\mathbb{E}(W_i|Y_i) &= \mathbb{E}(X_i + U_i|Y_i) \\ &= \mathbb{E}(X_i|Y_i) + \mathbb{E}(U_i|Y_i),\end{aligned}$$

this assumption is satisfied if the measurement errors U_i are non-differential with respect to Y_i , since this means $\mathbb{E}(U_i|Y_i) = 0$.

The aim of MR is to construct a random variable X_i^{mr} such that the first two moments of X_i^{mr} are the same as those of X_i given Y_i . Freedman *et al* showed that this can be done by defining:

$$X_i^{mr} = \mathbb{E}(X_i|Y_i)(1 - G) + W_iG \quad (3.85)$$

where G is given by:

$$G = \sqrt{\frac{\text{Var}(X_i|Y_i)}{\text{Var}(W_i|Y_i)}}.$$

Since $\mathbb{E}(W_i|Y_i) = \mathbb{E}(X_i|Y_i)$, it follows that:

$$\begin{aligned}\mathbb{E}(X_i^{mr}|Y_i) &= \mathbb{E}(X_i|Y_i)(1 - G) + \mathbb{E}(W_i|Y_i)G \\ &= \mathbb{E}(X_i|Y_i).\end{aligned}$$

It therefore also follows that the unconditional first moment $\mathbb{E}(X_i^{mr}) = \mathbb{E}(X_i)$. Then it follows that:

$$\begin{aligned}\text{Cov}(X_i^{mr}, Y_i) &= \mathbb{E}(X_i^{mr}Y_i) - \mathbb{E}(X_i^{mr})\mathbb{E}(Y_i) \\ &= \mathbb{E}(\mathbb{E}(X_i^{mr}Y_i|Y_i)) - \mathbb{E}(X_i)\mathbb{E}(Y_i) \\ &= \mathbb{E}(Y_i\mathbb{E}(X_i^{mr}|Y_i)) - \mathbb{E}(X_i)\mathbb{E}(Y_i) \\ &= \mathbb{E}(Y_i\mathbb{E}(X_i|Y_i)) - \mathbb{E}(X_i)\mathbb{E}(Y_i) \\ &= \text{Cov}(X_i, Y_i).\end{aligned}$$

The conditional variance of X_i^{mr} given Y_i is also the same as that of X_i , since:

$$\begin{aligned}\text{Var}(X_i^{mr}|Y_i) &= \text{Var}(\mathbb{E}(X_i|Y_i)|Y_i)(1 - G)^2 + \text{Var}(W_i|Y_i)G^2 \\ &= 0 + \frac{\text{Var}(X_i|Y_i)}{\text{Var}(X_i|Y_i) + \sigma_U^2}(\text{Var}(X_i|Y_i) + \sigma_U^2) \\ &= \text{Var}(X_i|Y_i).\end{aligned}$$

Then it also follows that:

$$\begin{aligned}
\text{Var}(X_i^{mr}) &= \mathbb{E}(\text{Var}(X_i^{mr}|Y_i)) + \text{Var}(\mathbb{E}(X_i^{mr}|Y_i)) \\
&= \mathbb{E}(\text{Var}(X_i|Y_i)) + \text{Var}(\mathbb{E}(X_i|Y_i)) \\
&= \text{Var}(X_i).
\end{aligned}$$

To use MR we must specify the conditional mean function $\mathbb{E}(X_i|Y_i)$ and the conditional variance function $\text{Var}(X_i|Y_i)$, and either know the values or have estimates of the parameters involved in these. Assuming these functions are correctly specified, and that we have consistent estimates of the parameters they involve, asymptotically X_i^{mr} will have the same mean, variance and covariance with Y_i as the original X_i values. It follows that parameters which are only functions of these first and second moments will be consistently estimated if X_i^{mr} is used in place of the unobserved X_i .

The specification of $\mathbb{E}(X_i|Y_i)$ and $\text{Var}(X_i|Y_i)$ depends on the assumed model for X_i and for Y_i given X_i . Freedman *et al* considered a linear regression outcome model, in which X_i , ϵ_i and U_i were assumed mutually independent. Freedman *et al* gave expressions for estimators of $\mathbb{E}(X_i|Y_i)$ and $\text{Var}(X_i|Y_i)$, stating that they made no normality assumptions. We believe however that the stated estimators are equal to the best linear prediction of X_i given Y_i and the variance of this given Y_i , but as we have previously discussed, these are not in general equal to $\mathbb{E}(X_i|Y_i)$ and $\text{Var}(X_i|Y_i)$. Freedman *et al* showed that the resulting MR estimate of β_X is equal to the MOM estimate (which is consistent), demonstrating that for a linear regression outcome model, one can assume that $\mathbb{E}(X_i|Y_i)$ is linear in Y_i and need not worry about the validity of this assumption. Furthermore, we believe the assumptions of independence of X_i , ϵ_i and U_i can be relaxed to assumptions of pair-wise zero correlation.

Freedman *et al* described X_i^{mr} as a variance preserving empirical Bayes estimate of X_i , shrinking W_i back towards its expectation conditional on Y_i . In contrast to RC, MR thus utilizes Y_i in predicting the unobserved X_i . Also, whereas RC creates predicted values of X_i which have variance less than σ_X^2 , MR creates predicted values which (asymptotically) have the same variance as X_i .

3.8.1 Moment reconstruction with internal replication data

We now consider the implementation of MR for the model as described in Section 3.5.1. In this setup, we recall that each subject has n_i error-prone measurements of X_i , with n_i potentially varying between subjects. Under this model specification, the mean of the measurements, \overline{W}_i is sufficient for X_i , and so MR can be adapted

by defining:

$$X_i^{mr}(\bar{W}_i, Y_i) = \mathbb{E}(X_i|Y_i)(1 - G_i) + \bar{W}_i G_i \quad (3.86)$$

where

$$G_i = \sqrt{\frac{\text{Var}(X_i|Y_i)}{\text{Var}(\bar{W}_i|Y_i)}}.$$

We recall from Section 3.6 that

$$\begin{aligned} \mathbb{E}(\mathbf{W}_i|Y_i) &= \gamma_0 \mathbf{1}_{n_i \times 1} + \gamma_Y Y_i \mathbf{1}_{n_i \times 1} \\ \text{Var}(\mathbf{W}_i|Y_i) &= \sigma_{X|Y}^2 \mathbf{1}_{n_i \times n_i} + \sigma_U^2 \mathbf{I}_{n_i}, \end{aligned}$$

so that

$$\begin{aligned} \mathbb{E}(X_i|Y_i) &= \gamma_0 + \gamma_Y Y_i \\ \text{Var}(X_i|Y_i) &= \sigma_{X|Y}^2 \\ \text{Var}(\bar{W}_i|Y_i) &= \sigma_{X|Y}^2 + \frac{\sigma_U^2}{n_i}. \end{aligned}$$

As we showed in Section 3.6, a linear mixed model for \mathbf{W}_i can be fitted, with a fixed effect of Y_i and a single random effect $b_i \sim N(0, \sigma_{X|Y}^2)$ representing the deviation in X_i from its mean given Y_i . The ML (or REML) estimates of the parameters can then be used to calculate X_i^{mr} , and the linear regression model for Y_i fitted with these values as the covariate.

3.8.2 Efficiency

Having fitted the linear mixed model for \mathbf{W}_i given Y_i , the ML estimate of β_X (for the previously described parametric model) can be calculated as described in Section 3.6. Thus, at least in large samples, there can be no efficiency gain from proceeding to calculate X_i^{mr} and using MR to estimate β_X . When $n_i = n_\bullet$ for all subjects, as shown by Freedman *et al*, the MR and MOM estimates of β_X coincide. When n_i differs between subjects, this equivalence no longer holds. Through calculation of both the ML and MR estimates of β_X on a simulated dataset, we have found that the MR estimate of β_X is not identical to the ML estimate. However, in our simulations (see Section 3.9), we found that MR, implemented as described, had efficiency which was effectively identical to ML.

3.8.3 Moment reconstruction for other setups

As previously discussed, Freedman *et al* have recently compared the performance of MR to RC and MI (see Section 3.7) [21]. Freedman *et al* considered a situation in which each subject has a measurement which adheres to a ‘non-classical’ error model:

$$W_i = \gamma_0 + \gamma_X X_i + \delta_i$$

where δ_i is mean zero error, independent of X_i and Y_i . Note that γ_0 and γ_X are not necessarily equal to zero and one respectively, as they are under the classical error model. In their setup, a random subset of subjects in an internal calibration study had two error-prone measurements M_{i1} and M_{i2} which follow a classical error model, with error variance different to that of W_i .

Their implementation of MR involved estimating the parameters needed to calculate the modified expression for X_i^{mr} , using data both from the calibration sub-study and the main study, and also allowed for the possibility that the measurement error in W_i was related to Y_i conditional on X_i , i.e. differential measurement error. Their MR estimate of β_X was then based on regressing Y_i on X_i^{mr} only in main study subjects. This means that the data from subjects in the calibration sub-study is only used to estimate the parameters needed to calculate X_i^{mr} , but are not used to directly estimate β_X . Freedman *et al* found that a weighted RC estimator (see Section 3.4.6) was usually more efficient than their implementation of MR. However it would seem that this is a somewhat unfair comparison, since the weighted RC estimator makes use of both calibration and main study subjects in the second stage estimation of β_X , whereas their implementation of MR only uses main study subjects in this second stage. Furthermore, the weighted RC estimator utilizes the non-differential error assumption for the measurements W_i , whereas their implementation of MR does not. Indeed, Freedman *et al* commented in their discussion that they considered versions of MR for this setting which utilize the non-differential error assumption, and reported that in simulations not shown that it was as efficient as the weighted RC estimator.

3.9 Simulations

In this section we report the results of simulations to compare the performance of RC, ML, MI, and MR for a linear regression outcome model. We simulated data in the simplest setting of a single covariate which is measured with error and no error-free covariates. As is typical in many epidemiological studies, we assumed that all subjects had at least one error-prone measurement, and that a 10% subset of subjects had a second error-prone measurement. Our aim was to compare the methods’

relative performance, in terms of bias and efficiency, and to see how these are affected by the strength of the association between X_i and Y_i and by the reliability of the error-prone measurements. Although in typical epidemiological studies there will usually be error-free covariates in the outcome model of interest and possibly multiple covariates measured with error, we believe our simulation results can be used to inform about the relative advantages and disadvantages of the estimation methods in the more general setup (of multiple covariates, one or more of which are measured with error).

3.9.1 Simulation setup

We simulated data for $n = 5,000$ independent subjects with a single covariate $X_i \sim N(0, 1)$. We chose this sample size to ensure that the probability of obtaining estimates of σ_X^2 equal to zero was sufficiently close to zero. The outcome Y_i was simulated according to the linear regression model:

$$Y_i = \beta_0 + \beta_X X_i + \epsilon_i$$

where $\beta_0 = 0, \beta_X = 1$, and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. We varied the value of σ_ϵ^2 to consider either weak, moderate or strong associations between Y_i and X_i , corresponding to a correlation between Y_i and X_i of 0.2, 0.5 and 0.8. We simulated measurements of X_i which were subject to independent, classical normally distributed error. Specifically, each subject had a first error-prone measurement W_{i1} where:

$$W_{i1} = X_i + U_{i1}$$

and $U_{i1} \sim N(0, \sigma_U^2)$. In addition, a random subset of 500 subjects had a second error-prone measurement W_{i2} , which was also subject to independent, classical normally distributed error, also with variance σ_U^2 . We varied the value of σ_U^2 to give reliability ratios of 2/3, 1/2 and 1/3.

The results for each scenario are based on 10,000 simulations.

3.9.2 Estimation methods

We estimated β_X using RC, ML, MI, and MR. For each method we report the mean and standard deviation of its estimates of β_X over the 10,000 simulations for each scenario.

Regression calibration

To implement RC, we fitted a one-way random-intercepts model to the error-prone measurements available from each subject using maximum likelihood. We used the `lmer` command from the `lme4` package to fit this model. Based on the estimates of

σ_X^2 and σ_U^2 , we calculated $\hat{\mathbb{E}}(X_i|\mathbf{W}_i)$ (see equation (3.21)) for each subject. We then estimated β_X using by the OLS slope estimate from the regression of Y_i on $\hat{\mathbb{E}}(X_i|\mathbf{W}_i)$. For the confidence intervals for RC, we calculated 95% Wald-type confidence intervals using the naive standard error obtained from regressing Y_i on $\hat{\mathbb{E}}(X_i|\mathbf{W}_i)$. These confidence intervals thus ignore the uncertainty in the parameter estimates needed to calculate $\hat{\mathbb{E}}(X_i|\mathbf{W}_i)$, and are thus not expected to give correct coverage. Despite this, researchers may be tempted to use such intervals given their ease of calculation, and so it is of interest to examine their performance.

Maximum likelihood

For the ML estimate of β_X , we fitted the linear mixed model to the error-prone measurements as described in Section 3.6 using the `lmer` command. Wald-type confidence intervals were calculated also as described in Section 3.6, using the reported standard error for $\hat{\gamma}_Y$ from `lmer`. To find a standard error for $\hat{\sigma}_{X|Y}^2$, we calculated the sub-matrix of the observed information matrix corresponding to the variance components. This was then inverted to find a standard error for $\hat{\sigma}_{X|Y}^2$. This ignores variability in the fixed effects parameters, which is asymptotically valid due to the fact that estimators of fixed effects and variance components in linear mixed models are asymptotically uncorrelated. This matches the approach used by Stata's `xtmixed` command.

Frequentist multiple imputation

Using the estimates of the parameters of the mixed model for \mathbf{W}_i given Y_i , we created $M = 25$ imputations of X_i for each subject, as described in Section 3.7.2. We then regressed Y_i on these imputed X_i for each of the 25 imputations, and estimated β_X by the arithmetic mean of these 25 estimates.

Moment reconstruction

Again, using the estimates of the parameters of the mixed model for \mathbf{W}_i given Y_i , we calculated X_i^{mr} for each subject, as described in Section 3.8. We then estimated β_X by the OLS slope estimated from the regression of Y_i on X_i^{mr} .

All simulations were performed using R. For reasons of brevity, we include only the R code for the simulations of Chapter 4, which are given in Listing 14.1.

3.9.3 Simulation results

Table 3.1 shows the results of the simulations. All the methods considered are consistent for a linear regression outcome model, and our simulations showed little bias for the sample size and scenarios considered. The variability of RC was very similar to ML - ML only being appreciably more efficient for scenario 9, when the

Table 3.1: Linear regression simulation results with normally distributed covariate. Mean (SD) of estimates of β_X (true value 1) from regression calibration (RC), maximum likelihood (ML), frequentist multiple imputation using 25 imputations (MI), and moment reconstruction (MR). λ denotes the reliability ratio of the error-prone measurements.

Scenario	$\text{corr}(Y_i, X_i)$	λ	RC	ML	MI	MR
1	0.2	2/3	1.002 (0.090)	1.002 (0.090)	1.002 (0.090)	1.002 (0.090)
2	0.2	1/2	1.007 (0.114)	1.007 (0.114)	1.007 (0.115)	1.007 (0.114)
3	0.2	1/3	1.015 (0.168)	1.015 (0.168)	1.015 (0.169)	1.015 (0.168)
4	0.5	2/3	1.001 (0.043)	1.001 (0.043)	1.001 (0.043)	1.001 (0.043)
5	0.5	1/2	1.003 (0.070)	1.003 (0.069)	1.004 (0.069)	1.003 (0.069)
6	0.5	1/3	1.015 (0.124)	1.014 (0.121)	1.014 (0.121)	1.014 (0.121)
7	0.8	2/3	1.001 (0.035)	1.001 (0.033)	1.002 (0.033)	1.001 (0.033)
8	0.8	1/2	1.004 (0.063)	1.003 (0.057)	1.003 (0.057)	1.003 (0.057)
9	0.8	1/3	1.011 (0.122)	1.010 (0.110)	1.010 (0.110)	1.010 (0.110)

reliability ratio was 1/3 and the association between X_i and Y_i was very strong. Indeed for most scenarios, the RC and ML estimates for a given dataset were very similar. For example, for scenario 1, the standard deviation of the difference between them was 0.002. The bias and variability of the MR estimates were effectively the same as ML. The MI estimates had identical bias to ML, and were marginally more variable than ML. We believe this small difference would reduce if the number of imputations M were increased further.

Table 3.2 shows the empirical coverage rates of the naive 95% confidence intervals obtained using RC and the two-sided and one-sided coverage rates for the ML Wald interval. The coverage of the naive 95% confidence interval for RC was reasonable when the correlation between X_i and Y_i was weak. Our results show however that ignoring the uncertainty in the parameters needed to calculate $\hat{\mathbb{E}}(X_i|\mathbf{W}_i)$ affects the coverage of the confidence intervals by a larger amount as the reliability ratio decreases and also as the correlation between X_i and Y_i increases. For a moderate or strong correlation, the coverage properties for the particular data setup we have used is poor, with coverage of only 34.9% for scenario 9. In contrast, the coverage of the ML Wald intervals was good for all scenarios, with the empirical coverage slightly less when the reliability ratio was lower. However, as λ decreased and the correlation between Y_i and X_i increased, the coverage rates of the one-sided intervals using the lower limit of the 95% interval became closer to 100% while the coverage of the one-sided interval using the upper limit became closer to 95%, rather than their 97.5% nominal coverage rates. This was a symptom of the fact that the sampling distribution of the ML estimator of β_X was positively skewed. The result was that the two-sided Wald intervals had close to the correct coverage level.

Table 3.2: Linear regression simulation results with normally distributed covariate. Empirical coverage rates of 95% confidence intervals for β_X (coverage of lower and upper one-sided 97.5% intervals): naive Wald intervals found using regression calibration (RC) and Wald intervals for the maximum likelihood (ML) estimate. λ denotes the reliability ratio of the error-prone measurements.

Scenario	corr(Y_i, X_i)	λ	RC		ML
			Naive	Wald	Wald
1	0.2	2/3	93.5	(96.6, 96.9)	95.2 (97.7, 97.4)
2	0.2	1/2	90.8	(94.9, 95.9)	95.1 (98.0, 97.1)
3	0.2	1/3	85.0	(92.2, 92.8)	95.4 (99.4, 96.0)
4	0.5	2/3	84.7	(92.5, 92.2)	95.1 (98.3, 96.7)
5	0.5	1/2	70.8	(85.4, 85.4)	95.1 (99.1, 96.0)
6	0.5	1/3	55.7	(77.7, 78.0)	94.9 (99.9, 95.0)
7	0.8	2/3	64.3	(82.2, 82.2)	95.0 (98.4, 96.6)
8	0.8	1/2	47.3	(74.0, 73.3)	95.2 (99.2, 96.0)
9	0.8	1/3	34.9	(68.5, 66.5)	94.3 (99.7, 94.7)

3.9.4 Robustness to modelling assumptions

As previously discussed in Sections 3.4.3 and 3.6.2, we believe the implementations of RC and ML which we have used, which are predicated on normality of X_i , should give consistent estimates of β_X even when this assumption is violated. We performed a further set of simulations to explore this claim. We used the same setup as previously described, except that X_i was generated with a log-normal distribution. To emphasize that the ML estimator assumes marginal normality for X_i , and is therefore no longer ML for the true data-generating model, we refer to it in this sub-section as ‘normal-model ML’.

Table 3.3 shows the results of the simulations, where again 10,000 simulations were performed for each scenario. Compared to the simulations under which X_i was normally distributed, the bias and sampling variability of both RC and normal-model ML were larger. As before, there was little evidence of bias in RC or normal-model ML. The variability of both RC and normal-model ML were similar to when X_i was normally distributed for a reliability ratio of 2/3 or 1/2. For a reliability ratio of 1/3, both RC and normal-model ML were more variable than they were for normal X_i . The efficiency advantage of normal-model ML over RC was greater compared to that when X_i was normally distributed. As before, MI and MR had bias and variability which were almost identical to that of normal-model ML.

Table 3.4 shows the empirical coverage rates of the RC and normal-model ML confidence intervals. Under non-normality we would not necessarily expect the coverage of the normal-model ML Wald intervals to equal the nominal level. However, in our simulations the coverage of these confidence intervals was quite acceptable, being close to 95% for most of the scenarios. The coverage rates of the one-sided

Table 3.3: Linear regression simulation results with log-normally distributed covariate. Mean (SD) of estimates of β_X from regression calibration (RC), normal-model maximum likelihood (ML), frequentist multiple imputation using 25 imputations (MI), and moment reconstruction (MR). λ denotes the reliability ratio of the error-prone measurements.

Scenario	$\text{corr}(Y_i, X_i)$	λ	RC	ML	MI	MR
1	0.2	2/3	1.000 (0.091)	1.003 (0.091)	1.004 (0.092)	1.003 (0.091)
2	0.2	1/2	1.003 (0.119)	1.010 (0.120)	1.011 (0.121)	1.010 (0.120)
3	0.2	1/3	1.012 (0.184)	1.024 (0.185)	1.024 (0.186)	1.023 (0.185)
4	0.5	2/3	0.998 (0.046)	1.002 (0.045)	1.002 (0.045)	1.002 (0.045)
5	0.5	1/2	0.999 (0.078)	1.006 (0.075)	1.006 (0.075)	1.006 (0.075)
6	0.5	1/3	1.013 (0.145)	1.019 (0.134)	1.019 (0.134)	1.019 (0.134)
7	0.8	2/3	0.999 (0.038)	1.002 (0.034)	1.002 (0.034)	1.002 (0.034)
8	0.8	1/2	1.000 (0.072)	1.003 (0.059)	1.004 (0.059)	1.003 (0.059)
9	0.8	1/3	1.010 (0.145)	1.011 (0.112)	1.011 (0.112)	1.011 (0.112)

intervals deviated from their nominal 95.7% levels in the same way as for normal X_i .

3.10 Conclusions

In this chapter we have examined the consequences of classical covariate measurement error in linear regression models and reviewed some of the methods which can be used to correct for these effects. For a linear regression model with a single covariate, we saw that classical measurement error causes attenuation in the observed association between outcome and the error-prone measurement of the covariate. In the more usual setting in which there are multiple covariates, biases can be either towards or away from the null, depending on the correlations between the underlying covariates. We conclude the chapter by comparing the estimation methods we have considered.

3.10.1 Statistical efficiency

MOM correction for classical covariate measurement error is attractive because of its simplicity of implementation. However, if the measurement model parameters are estimated using internal replication data, the simple MOM estimator of β_X is inefficient because only a single error-prone measurement is used to obtain the naive estimate of β_X . RC improves on this by using all of a subject's available error-prone measurements to predict X_i . In general however, RC is inefficient compared to the ML (under normality assumptions) estimator, although our simulations results suggest that the efficiency advantage of ML over RC may be small in typical situations. In RC, unweighted least squares is used to estimate β_X using data in which the

Table 3.4: Linear regression simulation results with log-normally distributed covariate. Empirical coverage rates of 95% confidence intervals for β_X (coverage of lower and upper one-sided 97.5% intervals): naive Wald intervals found using regression calibration (RC) and Wald intervals for the normal-model maximum likelihood (ML) estimate. λ denotes the reliability ratio of the error-prone measurements.

Scenario	$\text{corr}(Y_i, X_i)$	λ	RC	Normal-model ML
1	0.2	2/3	93.1 (96.8, 96.4)	94.6 (97.5, 97.1)
2	0.2	1/2	89.5 (94.7, 94.8)	94.6 (97.8, 96.9)
3	0.2	1/3	82.5 (91.7, 90.8)	94.2 (99.4, 94.5)
4	0.5	2/3	82.5 (92.0, 90.5)	94.4 (98.0, 96.4)
5	0.5	1/2	65.3 (84.6, 80.7)	93.6 (98.6, 95.0)
6	0.5	1/3	50.0 (76.4, 73.6)	93.8 (99.8, 94.0)
7	0.8	2/3	60.3 (82.3, 77.9)	94.3 (98.0, 96.2)
8	0.8	1/2	42.9 (74.5, 68.4)	94.7 (98.9, 95.8)
9	0.8	1/3	30.8 (68.6, 62.2)	94.2 (99.7, 94.5)

residual variance differs between subjects who have different values of n_i , which we believe is the cause of its (often small) inefficiency. We have shown that under a particular parametric model based on joint normality, the ML estimate of β_X can be obtained by first fitting a linear mixed model for \mathbf{W}_i given Y_i . This same linear mixed model can be used to estimate the parameters needed to multiply impute X_i , and also to calculate the predictions of X_i defined by the method of MR. Our simulations results suggest that the efficiency of both MI (with a suitably large number of imputations) and MR are very close to that of ML.

3.10.2 Assumptions

Consistent estimates of β_X can be obtained using MOM or RC (using best linear prediction) without making any distributional assumptions. In contrast, methods such as ML and MI are predicated on specific distributional assumptions. However, we have shown that for a joint model predicated on normality assumptions, the ML estimator of β_X remained consistent even if these normality assumptions are violated. Thus, even though the ML estimator is derived on the basis of stronger assumptions than are necessary for MOM or RC, these assumptions are not needed to ensure consistency of estimates. This robustness to modelling assumptions occurs because the parameters on which β_X depend are estimated consistently regardless of the distributions of the various variables [58].

3.10.3 A situation when the methods are equivalent

Although in general MOM, RC and ML give different estimates of β_X , if $n_i = n_\bullet$ for all i , so that each subject has n_\bullet error-prone measurements available, they give

identical point estimates under certain parametric assumptions. To see that MOM and RC give the same estimate, recall that the MOM estimate is found by dividing the naive estimate of β_X , obtained by regressing Y_i on \bar{W}_i , by:

$$\frac{\sigma_X^2}{\sigma_X^2 + \frac{\sigma_U^2}{n_{\bullet}}}, \quad (3.87)$$

or rather this expression with the parameters replaced by their respective estimates. The RC estimate under an assumption of normality for X_i and U_{ij} , or equivalently RC using the best linear prediction, is given by regressing Y_i on X_i^{blp} , where X_i^{blp} is as given in equation (3.21). In the definition of X_i^{blp} , \bar{W}_i is multiplied by the same factor as given above in equation (3.87). Since multiplying the independent variable of a linear regression model by a particular factor causes the estimated slope parameter to be multiplied by the reciprocal of the factor, we see that the MOM and RC estimates of β_X are identical if the same estimates of σ_X^2 and σ_U^2 are used. Under the parametric model described in Section 3.5.1 which assumes joint normality, Wang *et al* have shown that the MLE of β_X is identical to the MOM and RC estimates of β_X , if in MOM and RC, σ_X^2 and σ_U^2 are estimated using ML methods (Proposition 1 of [59]). In fact, we believe the estimates are only identical when the solutions to the estimating equations result in non-negative variance estimates. However, the probability of this occurring tends to zero as $n \rightarrow \infty$ and the number of subjects with $n_i \geq 2$ increases.

3.10.4 Software

MOM correction and RC can be easily implemented using standard statistical packages, although for Stata a user-written command `rcal` can be used to perform RC for generalized linear outcome models in a number of settings, including that in which internal replication data are available [60]. In Stata the `cme` wrapper for the GLLAMM package uses ML to allow for covariate measurement error in generalized linear outcome models. For the model considered in Section 3.5.1, in which the observed data log likelihood is available in closed form, commands for maximizing user defined likelihoods could also be used, such as Stata's `ml` command or the `NLMIXED` command in SAS. Such models can also be fitted using the latent variable software `Mplus` [45]. Our approach to ML estimation described in Section 3.6 only requires that one can fit a random-intercepts mixed model, which is available in most, if not all, modern statistical packages. Implementations of MI in statistical packages all require (as far as we are know) that the variable subject to missingness is observed on at least some subjects. The case of covariate measurement error with internal replicate error-prone measurements cannot therefore be dealt with using standard MI software. However, as described in Section 3.7, for the parametric model we considered, having found the ML estimates for the joint model, it is relatively easy

to create imputations of X_i . MR is also easy to implement for the setting we have considered, where as for MI, the parameters needed to calculate the values of X_i^{mr} can be estimated by fitting the linear mixed model for \mathbf{W}_i given Y_i , as described in Section 3.6.

3.10.5 Other methods

We note that this chapter is not an exhaustive account of all estimation methods which allow for classical covariate measurement error in linear regression outcome models. Other methods include semi-parametric methods such as the conditional score method (see Section 4.9) and simulation extrapolation (SIMEX) (see Chapter 5 of [8]). These methods are likely to be more useful for outcome models other than linear regression, for which obtaining consistent estimates of the parameters of interest is typically more difficult.

Chapter 4

Binary outcomes

In this chapter we consider the effects of covariate measurement error for binary outcome models. The generalised linear model family provides an extremely flexible framework for regression models, including models for outcomes which are binary, or Bernoulli distributed. Although other link functions are available, the canonical logit link is by far the most popular, giving the logistic regression model. Because of the popularity of logistic regression for binary outcomes, the majority of research into covariate measurement error for binary outcomes has focused on logistic regression, and it is this outcome model which we too focus on.

When considering binary outcomes, we assume that subjects are sampled randomly, and specifically not dependent on their outcome status Y_i (i.e. case-control studies). In the absence of covariate measurement error, case-control studies can be analyzed as if the data were obtained prospectively. However, if the covariates are measured with error, one cannot necessarily ignore the retrospective sampling when adjusting for covariate measurement error [61].

In Section 4.1 we begin by briefly reviewing the logistic regression model. We then review the literature examining the effects of covariate measurement error on logistic regression parameter estimates, which lead to MOM correction (Section 4.2). We then examine RC for logistic regression, for which estimates are only approximately consistent, in contrast to the case of linear regression (Section 4.3). Next, we describe the ML approach, which under the standard parametric assumptions is complicated by the intractability of the observed data likelihood function (Section 4.4). We then explore one simulation based approach to finding the MLEs for this parametric model by using a recently proposed version of Monte-Carlo EM (Section 4.5). Then we show how our approach to ML estimation based on fitting a linear mixed model can be adapted to the case of logistic regression, under a particular parametric model (Section 4.6), which forms part of a paper recently published in *Statistics in Medicine* [27]. This mixed model can also be used to multiply impute the unobserved covariates (Section 4.7) and to implement MR (Section 4.8). In Section 4.9 we describe the semi-parametric conditional score method, which gives

consistent estimates without requiring specification of the distribution of X_i . In Section 4.10 we report results of simulations, and in Section 4.11 discuss the advantages and disadvantages of the methods described.

4.1 Logistic regression

A binary outcome Y_i follows a logistic regression given a covariate X_i if:

$$P(Y_i = 1|X_i) = \mathbb{E}(Y_i|X_i) = \frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)} \quad (4.1)$$

In the absence of covariate measurement error, and conditioning on the observed values of X_i , the MLEs of β_0 and β_X are found by maximizing the log likelihood function:

$$\sum_{i=1}^n Y_i(\beta_0 + \beta_X X_i) - \log(1 + \exp(\beta_0 + \beta_X X_i)). \quad (4.2)$$

Under suitable regularity conditions, the MLE of β_0 and β_X are those values which solve the likelihood score equations:

$$\sum_{i=1}^n \psi_{ML}(Y_i, X_i, \beta_0, \beta_X) = 0 \quad (4.3)$$

where

$$\psi_{ML}(Y_i, X_i, \beta_0, \beta_X) = \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(Y_i - \frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)} \right). \quad (4.4)$$

Since the score equations are non-linear in β_0 and β_X , there is no closed form solution, and so an iterative method is required, such as Newton-Raphson.

4.2 The effects of classical covariate measurement error

For a linear regression outcome model, we saw earlier that the bias induced by classical covariate measurement error could be quantified exactly. Because the parameters in a linear regression only depend on means, variances, and covariances, the bias induced by classical covariate measurement error depends only on the variance of the measurement errors relative to the variance of the true covariate. Unfortunately the same is not true more generally for non-linear outcome regression models, such as logistic regression.

4.2.1 Induced model for Y_i given W_i

We first show why there is no exact expression for the bias induced by classical covariate measurement error in logistic regression. We assume X_i is measured by classical error by $W_i = X_i + U_i$ where U_i is independent of X_i and independent of Y_i conditional on X_i (i.e. non-differential). The conditional distribution of Y_i given W_i can then be expressed as:

$$\begin{aligned}
 P(Y_i = 1|W_i) &= \mathbb{E}(Y_i|W_i) \\
 &= \mathbb{E}(\mathbb{E}(Y_i|X_i, W_i)|W_i) \\
 &= \mathbb{E}(\mathbb{E}(Y_i|X_i)|W_i) \\
 &= \mathbb{E}\left(\frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)}|W_i\right)
 \end{aligned} \tag{4.5}$$

where we use the non-differential error assumption to go from the second to the third line. The induced model relating Y_i to the error-prone measurement W_i thus depends on the conditional distribution of the true covariate X_i given W_i , or equivalently, on the marginal distribution $f(X_i)$ and the distribution of the measurement error, $f(W_i|X_i)$. The distribution of Y_i given W_i does not in general follow a logistic regression. Also, in general the expectation in equation (4.5) cannot be evaluated in closed form, and so no exact expression for the bias caused by classical covariate measurement error can be found. With additional assumptions however, exact or approximate expressions for bias may be found.

4.2.2 Normal discriminant model

One case in which Y_i given W_i does follow a logistic regression model, and where quantification of the bias caused by covariate measurement error is simple, is the normal discriminant model where X_i is measured with independent normal error [62, 63]. Under the normal discriminant model, $X_i|Y_i \sim N(\gamma_0 + \gamma_Y Y_i, \sigma_{X|Y}^2)$. It can be easily shown that this model implies a logistic regression model for Y_i given X_i :

$$P(Y_i = 1|X_i) = \mathbb{E}(Y_i|X_i) = \frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)}$$

where:

$$\beta_0 = \log\left(\frac{\pi_1}{1 - \pi_1}\right) - \frac{1}{2\sigma_{X|Y}^2}(\gamma_Y^2 + \gamma_Y \gamma_0) \tag{4.6}$$

$$\beta_X = \frac{\gamma_Y}{\sigma_{X|Y}^2} \tag{4.7}$$

for $\pi_1 = P(Y_i = 1)$. If X_i is measured with normally distributed error with mean zero and variance σ_U^2 , then $W_i|Y_i \sim N(\gamma_0 + \gamma_Y Y_i, \sigma_{X|Y}^2 + \sigma_U^2)$. This thus implies that

Y_i given W_i also follows a logistic regression with log odds ratio β_W given by:

$$\beta_W = \frac{\gamma_Y}{\sigma_{X|Y}^2 + \sigma_U^2} = \frac{\gamma_Y}{\sigma_{X|Y}^2} \frac{\sigma_{X|Y}^2}{\sigma_{X|Y}^2 + \sigma_U^2}. \quad (4.8)$$

Therefore for the normal discriminant model, independent normally distributed measurement error causes attenuation of parameter estimates, by a factor $\frac{\sigma_{X|Y}^2}{\sigma_{X|Y}^2 + \sigma_U^2}$. We note that this factor is not the same as the reliability ratio of the error-prone measurements, because the expression involves the conditional variance $\sigma_{X|Y}^2$ rather than the marginal variance σ_X^2 . However, the two are similar if the association between Y_i and X_i is weak or if the outcome is rare or very common. In this case, the naive estimates of β_X are then biased by approximately the same factor as in linear regression.

4.2.3 Other conditions under which bias can be approximated

In Section 4.3 we review the work of Rosner *et al* [2, 13, 33], who showed that RC can be justified for logistic regression if the outcome Y_i is rare and X_i given W_i is normal, with conditional mean function linear in W_i . Subsequently, Kuha showed that RC also gives approximately consistent estimates of β_X providing $\beta_X^2 \text{Var}(X_i|W_i)$ is small [64].

When $\mathbb{E}(X_i|W_i)$ is linear in W_i , the RC estimate of β_X is identical to a MOM estimate obtained by dividing the naive estimate of β_X (found by using W_i as covariate) by $\lambda = \sigma_X^2 / (\sigma_X^2 + \sigma_U^2)$. Therefore if $\mathbb{E}(X_i|W_i)$ is linear in W_i , if either the outcome is rare and X_i is normally distributed given W_i , or if $\beta_X^2 \text{Var}(X_i|W_i)$ is small, the naive estimator of β_X is biased approximately by a factor of λ , the reliability ratio of the error-prone measurements W_i .

4.3 Regression calibration

The method which is now referred to as RC was first proposed in the context of a generalized linear outcome model by Armstrong [34]. Its use for logistic regression outcome models was made popular by a series of papers by Rosner *et al* in the early 1990s [2, 13, 33]. We review the conditions under which RC can be expected to give approximately consistent estimates of β_X for logistic regression. For simplicity, we consider RC using a single error-prone measurement W_i , subject to classical measurement error, although the extension to using multiple error-prone measurements \mathbf{W}_i applies as for continuous outcomes.

4.3.1 $\text{Var}(X_i|W_i)$ small

Armstrong's original proposal for using RC in generalized linear models was based on a simple first order Taylor series expansion, i.e. using the delta method, to approximate the expectation in equation (4.5):

$$\begin{aligned} P(Y_i = 1|W_i) &= \mathbb{E} \left(\frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)} | W_i \right) \\ &\approx \frac{\exp(\beta_0 + \beta_X \mathbb{E}(X_i|W_i))}{1 + \exp(\beta_0 + \beta_X \mathbb{E}(X_i|W_i))} \end{aligned}$$

which is valid providing $\text{Var}(X_i|W_i)$ is small. This conditional variance is small when the measurement error variance σ_U^2 is small.

4.3.2 Rare outcome and X_i given W_i normal

Rosner *et al* proposed using RC (or equivalently, under their assumptions, MOM correction) in logistic regression [2, 13, 33]. Under the assumptions that the outcome is rare and that X_i is normal given W_i (with conditional mean linear in W_i), Rosner *et al* showed that Y_i approximately follows a logistic model given W_i , with log odds ratio $\lambda\beta_X$ for W_i . Under the assumption that the outcome is rare:

$$\begin{aligned} P(Y_i = 1|X_i) &= \frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)} \\ &\approx \exp(\beta_0 + \beta_X X_i), \end{aligned}$$

using the first-order Taylor series expansion of $x/(1+x)$ around 0, which gives $x/(1+x) \approx x$. Then under the assumption that $X_i|W_i \sim N(\mathbb{E}(X_i|W_i), \text{Var}(X_i|W_i))$, where $\mathbb{E}(X_i|W_i) = \mu_X + \lambda(W_i - \mu_X)$ and $\text{Var}(X_i|W_i) = \sigma_X^2(1 - \lambda)$, it follows that $\exp(\beta_0 + \beta_X X_i)$ follows a log-normal distribution conditional on W_i , with parameters $\beta_0 + \beta_X \mathbb{E}(X_i|W_i)$ and $\beta_X^2 \text{Var}(X_i|W_i)$. Then using the fact that the mean of a log-normal distribution is equal to $\exp(\mu + \sigma^2/2)$ where μ and σ^2 are the mean and variance of the underlying normal distribution, it follows, using equation (4.5), that:

$$\begin{aligned} \mathbb{E}(Y_i = 1|W_i) &= \mathbb{E}(\mathbb{E}(Y_i|X_i)|W_i) \\ &\approx \mathbb{E}(\exp(\beta_0 + \beta_X X_i)|W_i) \\ &\approx \exp(\beta_0 + \beta_X \mathbb{E}(X_i|W_i) + \beta_X^2 \text{Var}(X_i|W_i)/2) \\ &\approx \frac{\exp(\beta_0 + \beta_X \mathbb{E}(X_i|W_i) + \beta_X^2 \text{Var}(X_i|W_i)/2)}{1 + (\exp(\beta_0 + \beta_X \mathbb{E}(X_i|W_i) + \beta_X^2 \text{Var}(X_i|W_i)/2))}. \end{aligned}$$

Since $\text{Var}(X_i|W_i)$ does not depend on W_i , this means Y_i follows a logistic regression given W_i , with intercept $\beta_0 + \beta_X^2 \text{Var}(X_i|W_i)/2$ and log odds ratio corresponding to W_i of $\beta_X \mathbb{E}(X_i|W_i)$. Thus approximately consistent estimates of β_X can be obtained by fitting the logistic regression model with $\mathbb{E}(X_i|W_i)$, under the stated assumptions.

Under the assumption that $\mathbb{E}(X_i|W_i) = \mu_X + \lambda(W_i - \mu_X)$, the same estimate of β_X is obtained by dividing the naive estimate of β_X by λ , justifying the conclusion that the naive estimate is biased approximately by λ under the given assumptions.

4.3.3 $\beta_X^2 \text{Var}(X_i|W_i)$ small

In 1994, Kuha showed that RC could be justified for logistic regression under weaker assumptions than the ones used by Rosner *et al* [64]. Recall from equation (4.5) that:

$$\mathbb{E}(Y_i = 1|W_i) = \mathbb{E}\left(\frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)}|W_i\right).$$

By replacing $\exp(\beta_0 + \beta_X X_i)/(1 + \exp(\beta_0 + \beta_X X_i))$ by its second order Taylor series expansion about $\mathbb{E}(X_i|W_i)$, Kuha showed that RC gives approximately consistent estimates of β_X providing $\beta_X^2 \text{Var}(X_i|W_i)$ is small. This holds if either β_X is small, $\text{Var}(X_i|W_i)$ is small, or both are small.

When subjects have multiple error-prone measurements, the small conditional variance condition $\text{Var}(X_i|W_i)$ becomes $\text{Var}(X_i|\mathbf{W}_i)$. This conditional variance is a decreasing function of the number of error-prone measurements n_i , and so we would expect the bias in RC to decrease when subjects have more replicate error-prone measurements of X_i .

4.4 Maximum likelihood

In this section we consider a maximum likelihood approach to deal with classical covariate measurement error in a logistic regression outcome model. We first define the parametric model which is most often considered, which is based on marginal normality for the covariate X_i (Section 4.4.1). The observed data or marginal likelihood function cannot be expressed in closed form for this model, precluding direct application of maximization methods such as Newton-Raphson. We consider adaptations of these methods to deal with the intractability of the likelihood function in Sections 4.4.3 and 4.4.4 respectively. We end the section by briefly describing a probit outcome model for Y_i , for which the likelihood function can be evaluated and the ML estimates can be more easily found. In Section 4.5 we show how the Monte-Carlo EM algorithm can be used to find ML estimates for the joint model which assumes marginal normality for X_i . In Section 4.6 we consider an alternative parametric model for which a standard linear mixed model can be fitted to find the ML estimates.

4.4.1 Model specification and the observed data likelihood function

As for continuous outcomes, under the non-differential error assumption we can specify the joint distribution of the complete data (i.e. including X_i) by separately defining the outcome model, the covariate model, and the measurement model. As described at the beginning of the chapter, we assume that Y_i follows a logistic regression model given X_i , with intercept β_0 and log odds ratio corresponding to X_i equal to β_X . We assume the same covariate and measurement model as in Section 3.5.1, which for convenience we briefly recall here. We assume $X_i \sim N(\mu_X, \sigma_X^2)$, and that subject i has n_i error-prone measurements of X_i , subject to normally distributed error with variance σ_U^2 . The measurement errors are assumed independent of X_i , and to be independent of Y_i conditional on X_i . As before, we denote the vector of parameters for this joint model by $\boldsymbol{\theta} = (\beta_0, \beta_X, \mu_X, \sigma_X^2, \sigma_U^2)^T$. We note that in contrast to a linear regression outcome model, there is no variance parameter in logistic regression because the variance function of a Bernoulli random variable is a fixed function of the mean.

As for continuous outcomes, given data from n independent subjects, the observed data likelihood function is equal to the probability density function of the observed data, considered as a function of the parameter vector $\boldsymbol{\theta}$:

$$\begin{aligned}
 L_n(\boldsymbol{\theta}) &= \prod_{i=1}^n f(Y_i, \mathbf{W}_i | \boldsymbol{\theta}) \\
 &= \prod_{i=1}^n \int f(Y_i | X_i) f(\mathbf{W}_i | X_i) f(X_i) dX_i \\
 &= \prod_{i=1}^n \int \frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)} \\
 &\quad \times n_i (2\pi\sigma_U^2)^{-1/2} \exp\left(-\sum_{j=1}^{n_i} (W_{ij} - X_i)^2 / 2\sigma_U^2\right) \\
 &\quad \times (2\pi\sigma_X^2)^{-1/2} \exp(-(X_i - \mu_X)^2 / 2\sigma_X^2) dX_i \tag{4.9}
 \end{aligned}$$

Recall that when $f(Y_i | X_i)$ is the conditional normal linear regression model, Y_i and \mathbf{W}_i are jointly normally distributed (see Section 3.5.1). Unfortunately, when Y_i given X_i follows a logistic regression model, the integral in equation (4.9) cannot be expressed in closed form, and the joint distribution of (Y_i, \mathbf{W}_i) is not a standard distribution. This means that the observed data likelihood function for this model cannot be expressed in a tractable form. This makes finding the MLE considerably more difficult.

The observed data likelihood of equation (4.9) is similar to that for generalized linear mixed models [41]. Generalized linear mixed models extend the generalized linear model by allowing the linear predictor to contain both fixed effects and random

effects. The observed data likelihood function for such models similarly involves integrating over the distribution of the unobserved random effects, and apart from the case of continuous outcomes, this is usually an intractable integral. Because of the similarities of the likelihood functions in the two cases, many of the methods which can be used to maximize the likelihood functions of generalized linear mixed models can also be applied to joint models which include covariate measurement error.

4.4.2 Approximating integrals

We briefly review the main approaches to approximating intractable integrals, and then examine how these can be used within the Newton-Raphson and EM algorithms. Methods to approximate integrals can broadly be divided into deterministic and those based on pseudo-random simulation, or Monte-Carlo methods [41].

Deterministic approximations

The method of Gaussian quadrature involves approximating integrals by a weighted sum of the integrand evaluated at a number of locations (quadrature points) in the domain of integration. The approximation can be made more accurate by increasing the number of quadrature points. The location of the quadrature points and weights depend on the integrand and the domain of integration. For integrals such as the one in equation (4.9), where the domain of integration is the entire real line, and the integrand is the product of a function and a normal density, the locations of the quadrature points are solutions to the Hermite polynomial functions [41, 65].

The location of the quadrature points for Gauss-Hermite quadrature depend on the normal density involved in the integrand, but not on the function with which it is multiplied. Depending on this function, the value of the integrand may be small for many of the quadrature locations, and so in effect, many of the quadrature points are wasted. To overcome this, adaptive Gaussian quadrature shifts the location of the quadrature points so that they are centred at the centre of the integrand and scaled depending on the support of the integrand. For a given number of quadrature points, this results in a more accurate approximation of the integral compared to standard Gaussian quadrature. However, to use adaptive quadrature we must calculate how much to shift and scale the quadrature points, which entails additional computation.

Monte-Carlo approximations

The integral involved in equation (4.9) can be viewed as the expectation of $f(Y_i|X_i)f(\mathbf{W}_i|X_i)$ with respect to the marginal distribution of X_i . In classical Monte-Carlo integration [50], a number M random draws $X_i^{(m)}$, $m = 1, \dots, M$ are generated from the density $f(X_i)$ and we approximate the expectation by the sample mean of

$f(Y_i|X_i)f(\mathbf{W}_i|X_i)$ evaluated over the M randomly drawn values of X_i :

$$\mathbb{E}_{f(X)}(f(Y_i|X_i)f(\mathbf{W}_i|X_i)) \approx \frac{1}{M} \sum_{m=1}^M f(Y_i|X_i^{(m)})f(\mathbf{W}_i|X_i^{(m)})$$

The approximation can be made arbitrarily accurate by increasing M , and we can assess the error in the approximation using the sample variance of the summand about the mean.

Sometimes it is not possible to directly sample from the required density $f(X_i)$, in which case a variety of alternative approaches may be possible. The method of rejection sampling can sometimes be used to generate samples from the desired distribution by sampling from an alternative distribution, for which it is easy to generate samples, and only accepting particular draws on the basis of some criterion [50]. We describe rejection sampling in greater detail in Section 4.5.2. The method of importance sampling also involves sampling from a distribution which differs from the target distribution. In importance sampling, the integrand is multiplied by the ratio of the target density to the candidate density (evaluated at each sample), so that the approximation is a valid estimate of the integral of interest [50].

‘The curse of dimensionality’

Deterministic quadrature methods can be usefully employed to approximate both one dimensional and multi-dimensional integrals of low dimension. For multi-dimensional integrals, the number of points at which the integrand has to be evaluated grows exponentially with the dimension of the integral. Recall that in the case of covariate measurement error, the density of the observed data involves integrating over the distribution of the unobserved covariate X_i . With multiple covariates measured with error, we must integrate over the distribution of the multivariate random variable \mathbf{X}_i . Thus for even moderate dimensions of \mathbf{X}_i , the time required to find the MLE may be prohibitively large because of the number of points at which the integrand must be evaluated.

In contrast, Monte-Carlo methods may still be feasible for approximating higher dimensional integrals. This is because the error in a Monte-Carlo estimate does not depend on the dimension of the integral, in contrast to deterministic methods such as Gaussian quadrature. Of course to use Monte-Carlo integration, we must be able to sample from the required multivariate density, or resort to methods such as rejection sampling or importance sampling. However, as the dimension of \mathbf{X}_i increases, use of these latter methods to generate random draws may also become problematic, as we discuss further in Section 4.5.5.

4.4.3 Newton-Raphson

As described in Section 3.5.3, the Newton-Raphson method requires that we can evaluate the likelihood score and information matrices for given values of the model parameters $\boldsymbol{\theta}$. Since we have no closed form expression for the observed data likelihood function for the specified model, we cannot derive closed form expressions for the first and second derivatives of the log observed data likelihood, and so approximate methods must be used.

The NLMIXED procedure in SAS can be used to maximize likelihood functions such as the one in equation (4.9) [46]. The GLLAMM package for Stata, developed by Rabe-Hesketh, Skrondal and Pickles [66], can also be used to maximize the likelihood function in equation (4.9). A wrapper command called `cme`, which calls the GLLAMM package, is available for fitting generalized linear outcome models in which covariates are measured with error [44]. Both NLMIXED and GLLAMM use adaptive Gaussian quadrature (see Section 4.4.2) to approximate the observed data likelihood function. A Newton-Raphson type algorithm is then used, with the first and second derivatives of the log likelihood approximated by the numerical derivatives of the approximated log likelihood [66].

4.4.4 Expectation Maximization

In one of the earliest proposals to use ML to allow for covariate measurement error, Schafer proposed using the EM algorithm as a convenient approach to finding the MLE [38]. Recall that the E-step consists of finding the expectation of the complete data log likelihood, conditional on observed data and the current estimate of the model parameters. As for the continuous outcome model we considered in Section 3.5, the complete data log likelihood can be decomposed into the sum of three parts, corresponding to the outcome model, the covariate measurement error model, and the model for the true covariate. Since we assume here that the true covariate X_i is normally distributed, and measured with error by classical normal error, the second and third components of the complete data log likelihood are the same as for when the outcome model is linear regression.

We recall from Section 3.5.4 that in the E-step we required $\mathbb{E}(X_i|Y_i, \mathbf{W}_i)$ and $\mathbb{E}(X_i^2|Y_i, \mathbf{W}_i)$. For the model specification we are assuming, in which Y_i given X_i follows a logistic regression, the conditional distribution $f(X_i|Y_i, \mathbf{W}_i)$ is no longer normal, and there is no closed form expression for $\mathbb{E}(X_i|Y_i, \mathbf{W}_i)$ and $\mathbb{E}(X_i^2|Y_i, \mathbf{W}_i)$. Furthermore, from equation (4.2), the component of the complete data log likelihood corresponding to $f(Y_i|X_i)$ is given by:

$$l(\beta_0, \beta_X) = \sum_{i=1}^n \{Y_i(\beta_0 + \beta_X X_i) - \log(1 + \exp(\beta_0 + \beta_X X_i))\}.$$

Taking expectations conditional on the observed data (Y_i and \mathbf{W}_i) we have:

$$\sum_{i=1}^n \{Y_i(\beta_0 + \beta_X \mathbb{E}(X_i|Y_i, \mathbf{W}_i)) - \mathbb{E}(\log(1 + \exp(\beta_0 + \beta_X X_i))|Y_i, \mathbf{W}_i)\}.$$

Thus the E-step also requires computation of $\mathbb{E}(\log(1 + \exp(\beta_0 + \beta_X X_i))|Y_i, \mathbf{W}_i)$.

Schafer proposed approximating this by $\log(1 + \exp(\beta_0 + \beta_X \mathbb{E}(X_i|Y_i, \mathbf{W}_i)))$, and approximating $\mathbb{E}(X_i|Y_i, \mathbf{W}_i)$ and $\text{Var}(X_i|Y_i, \mathbf{W}_i)$ by the mean and variance of a normal distribution which approximates the conditional distribution $f(X_i|Y_i, \mathbf{W}_i)$. The resulting estimator of β_X is not equal to the MLE due to these approximations, although Schafer reported promising performance in simulations [38]. The required expectations can instead be approximated using quadrature methods. Higdon and Schafer describe in detail the implementation of the EM algorithm for GLMs subject to covariate measurement error, using Gauss-Hermite quadrature to approximate the required expectations [67].

Monte-Carlo Expectation Maximization

In 1990 Wei and Tanner proposed using Monte-Carlo approximation to deal with an intractable E-step in the EM algorithm, leading to an algorithm which they termed Monte-Carlo Expectation Maximization (MCEM) [68]. At iteration $(t + 1)$ of the MCEM algorithm, the expected complete data log likelihood is approximated by its Monte-Carlo average, the average being taken over draws from $f(X_i|Y_i, \mathbf{W}_i, \boldsymbol{\theta}^{(t)})$. The M-step then involves maximizing this approximation with respect to the model parameters $\boldsymbol{\theta}$. This means maximizing the combined complete data log likelihood corresponding to $n \times M$ pseudo-observations.

We describe MCEM in greater detail in Section 4.5, where we also describe a recent proposal for how the number of imputations M should be increased as the algorithm moves towards convergence.

4.4.5 Robustness to modelling assumptions

For a linear regression outcome model, we saw in Section 3.6.2 that the ML estimator based on a model of joint normality was consistent even if some of the normality assumptions are violated. As discussed by Huang *et al* [58], such robustness occurs when the parameter of primary interest only depends on moments of the true model and when these are consistently estimated even if the distributional assumptions are violated.

A small number of papers have investigated the robustness of the model described in Section 4.4.1 to non-normality of X_i . Thoresen and Laake reported that if X_i has a chi-squared distribution with one degree of freedom, the ML estimator based on normality for X_i has little bias [69]. In contrast, Schafer found that the

ML estimator based on normality was biased when the distribution of X_i was a mixture of two normals, although he did not describe how the the MLE was found [70]. Schafer’s results also suggest that RC based on normal X_i had substantial bias, which given the size of the measurement error used in the simulations is quite surprising, given RC has generally been found to have relatively small biases for logistic regression. When X_i had a skewed distribution, Rabe-Hesketh *et al* also found that the ML estimator based on normality for X_i was biased. Using a re-measurement technique similar to SIMEX, where the observed data are repeatedly contaminated with additional measurement error, Huang *et al* found that the ML estimator assuming normality for X_i is biased when the true distribution of X_i is a mixture of normals [58].

4.4.6 Relaxing assumptions

Since it is difficult to empirically check distributional assumptions for X_i when only internal replication data are available, and the ML approach is not robust to non-normality of X_i , a number of approaches have been proposed to relax the distributional assumptions for X_i . Perhaps the earliest were semiparametric approaches proposed by Stefanski and Carroll *et al* [71]. This includes the conditional score method, which we describe later in Section 4.9.

Carroll *et al* proposed using parametric likelihood, but with the distribution of X_i assumed to belong to a flexible class of densities [72]. Specifically, Carroll *et al* used a mixture of normals to model the distribution of X_i , and used MCMC methods to estimate the model parameters in a Bayesian approach. Such models can be fitted using software such as WinBUGS, within the Bayesian framework.

An alternative approach which makes no assumptions about the distribution of X_i is to use so called ‘non-parametric maximum likelihood’ (NPML) to estimate the distribution of X_i non-parametrically [70, 73]. The non-parametric MLE of the distribution is discrete, with non-zero probabilities at a finite number of locations. The locations, their probabilities, and the number of them, can be estimated jointly, as described by Rabe-Hesketh *et al* [73]. Schafer suggested an EM approach for estimation [70], whereas Rabe-Hesketh *et al* proposed using the Newton-Raphson method, which is implemented as an option in the Stata command GLLAMM [73]. With a normal covariate X_i , Rabe-Hesketh *et al* ’s simulations suggest the NPML estimator is no less efficient than the correctly specified parametric ML estimator [73]. For simulations with skewed X_i , the NPML estimates showed little bias, while the ML estimator incorrectly predicated on normality for X_i was biased towards the null. In addition to robustness of estimation of the outcome model parameters, a further potential benefit of this approach is that one is able to examine the estimated distribution of X_i . In contrast, predictions of X_i under a parametric model will

not reflect the underlying distribution unless the assumed distribution is correct or $\text{Var}(X_i|\mathbf{W}_i)$ is small.

4.4.7 Probit outcome model

The logistic regression model is the most commonly used model for relating a binary outcome to covariates of interest. Schafer showed that the ML estimates using the EM algorithm can be obtained in which no intractable integrals are involved if instead a probit model is assumed for Y_i given X_i [39]. Instead of using the logit link function, the probit model uses the normal cumulative distribution function as link. Since the logistic and cumulative normal functions are very similar, the resulting estimates can then be transformed into odds ratios (see Section 4.8.2 of [8]).

4.5 Ascent-based Monte-Carlo Expectation Maximization

For scalar X_i , the MLE of the model which assumes normality for X_i can be found using Gaussian quadrature methods. This is implemented in Stata's GLLMM command and the NLMIXED command in SAS. However, when \mathbf{X}_i is multivariate, as discussed in Section 4.4.2, quadrature methods may become impractical because the number of evaluations of the integrand necessary to accurately approximate the required integrals becomes very large. Instead, we can resort to Monte-Carlo approximations, which may still be computationally feasible for larger dimensions.

In this section we describe the implementation of ascent-based Monte-Carlo EM, as recently proposed by Caffo *et al* [74], to find MLEs for the parametric model previously described. We first describe the Monte-Carlo Expectation Maximization (MCEM) method in more detail. The MCEM method involves approximating the intractable E-step of the EM algorithm using Monte-Carlo simulations of the missing, or unobserved values. We then describe the ascent-based version of MCEM, which recovers one of the key properties of the standard EM algorithm – that at each iteration the observed data likelihood value is no lower than the value at the last iteration.

4.5.1 Monte-Carlo Expectation Maximization

In the E-step of EM for the model defined in Section 4.4.1, given the estimate of the model parameters at the $t - 1$ th iteration, $\boldsymbol{\theta}^{(t-1)}$, we need to calculate:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)}) = \sum_{i=1}^n \mathbb{E} \left(\log(f(Y_i|X_i)) + \log(f(\mathbf{W}_i|X_i)) + \log(f(X_i)) \right). \quad (4.10)$$

For each $i = 1, \dots, n$, for now assume we can randomly draw M values $X_i^{(1)}, \dots, X_i^{(M)}$ from $f(X_i|Y_i, \mathbf{W}_i, \boldsymbol{\theta}^{(t-1)})$. Then we can approximate $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t-1)})$ by:

$$\tilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) = \frac{1}{M} \sum_{i=1}^n \sum_{m=1}^M \log(f(Y_i|X_i^{(m)})) + \log(f(\mathbf{W}_i|X_i^{(m)})) + \log(f(X_i^{(m)})) \quad (4.11)$$

In the M-step, we maximize $\tilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$ in $\boldsymbol{\theta}$ to obtain updated estimates of the model parameters. One way of viewing this is that the unobserved X_i are imputed from their conditional distribution given the observed data and the current parameter estimates. Since the complete data log likelihood is the sum of components corresponding to the outcome model (logistic regression), the measurement model (classical, independent normally distributed error) and the covariate model (normal distribution), and these components share no parameters, the M-step of MCEM involves maximizing each of these sub-model log likelihoods. Maximizing the Monte-Carlo estimate of each component is equivalent to maximizing the log likelihood with a single ‘pseudo clustered dataset’, whereby each subject has M observations, but for the purposes of maximization these are all treated as independent.

An alternative to maximizing the complete data log likelihood corresponding to this pseudo clustered dataset is to separately maximize the complete data log likelihoods corresponding to each of the M imputed datasets and average the resulting M sets of parameter estimates. This algorithm thus involves multiply imputing the unobserved X_i M times, maximizing the M completed datasets and averaging the resulting M parameter estimates. This is the same as frequentist MI (see Section 3.7.1), with M imputations. Wang and Robins showed that this is asymptotically (as $n \rightarrow \infty$) equivalent to maximizing the function in equation (4.11) [53]. As pointed out by Nielsen, maximizing the function in equation (4.11) may sometimes encounter difficulties due to the likelihood being multi-modal [75]. With large M , it will also usually be computationally cheaper to separately maximize the likelihoods of the M completed datasets and average the resulting estimates, rather than maximize the combined likelihood function of equation (4.11). We therefore use this approach in our implementation of MCEM.

For the parametric model of Section 4.4.1, this approach involves fitting M logistic regressions to the M imputed datasets $(Y_i, X_i^{(m)}, i = 1, \dots, n)$, maximizing the M normal likelihoods corresponding to $(\mathbf{W}_i, X_i^{(m)}, i = 1, \dots, n)$, and maximizing the M normal likelihoods corresponding to $(X_i^{(m)}, i = 1, \dots, n)$. The M sets of estimates for each of the parameters in $\boldsymbol{\theta}$ are then averaged to obtain the updated parameter estimates. This process can then be iterated in the same way as standard EM.

4.5.2 Generating imputations using rejection sampling

To use MCEM for the model defined in Section 4.4.1 we need to be able to draw from $f(X_i|Y_i, \mathbf{W}_i)$. The method of rejection sampling can be used to do this. Tsiatis recently described the implementation of rejection sampling for this model (in the case of internal validation data) [52]. We first describe the method of rejection sampling in general, and then show how it can be used to sample from $f(X_i|Y_i, \mathbf{W}_i)$ for the model of Section 4.4.1.

Suppose that we can sample values from a density $g(X)$ (the ‘candidate density’), but that we actually want to sample values from a density $f(X)$ (the ‘target density’), from which it is either impossible or difficult to sample from directly. Furthermore, we assume that we can find a constant c such that

$$\frac{f(X)}{g(X)} \leq c \quad \forall X.$$

Then we can create samples from $f(X)$ in the following way:

1. Draw a random value x from $g(X)$, and a random draw u from the uniform $U(0, 1)$ distribution.
2. If $u \leq \frac{f(x)}{cg(x)}$ then accept x as a sample from $f(X)$. Otherwise repeat the process until the inequality is satisfied.

It can be shown that the accepted samples are distributed according to the target distribution $f(X)$ [50].

Recall that for the model defined in Section 4.4.1, we assume that X_i is marginally normally distributed and that $W_{ij} = X_i + U_{ij}$ are error-prone measurements subject to unbiased independent normal errors with variance σ_U^2 . The covariate X_i and error-prone measurements \mathbf{W}_i are thus jointly normal, and so the conditional distribution $f(X_i|\mathbf{W}_i)$ is also normal, with mean and variance coinciding with that of the best linear prediction of X_i given \mathbf{W}_i , as given in Section 3.4.2. It is thus simple to sample from $f(X_i|\mathbf{W}_i)$, which we use as our candidate density.

To use rejection sampling we must bound:

$$\begin{aligned} \frac{f(X_i|Y_i, \mathbf{W}_i)}{f(X_i|\mathbf{W}_i)} &= \frac{f(Y_i, X_i, \mathbf{W}_i)}{f(\mathbf{W}_i, Y_i)f(X_i|\mathbf{W}_i)} \\ &= \frac{f(Y_i|X_i)f(\mathbf{W}_i, X_i)}{f(Y_i|\mathbf{W}_i)f(\mathbf{W}_i)f(X_i|\mathbf{W}_i)} \\ &= \frac{f(Y_i|X_i)}{f(Y_i|\mathbf{W}_i)}. \end{aligned} \tag{4.12}$$

Since Y_i is binary, $f(Y_i|X_i)$ is less than or equal to one, and so we have that:

$$\frac{f(X_i|Y_i, \mathbf{W}_i)}{f(X_i|\mathbf{W}_i)} = \frac{f(Y_i|X_i)}{f(Y_i|\mathbf{W}_i)} \leq \frac{1}{f(Y_i|\mathbf{W}_i)} = c. \tag{4.13}$$

Thus if we sample x_i from $f(X_i|\mathbf{W}_i)$, we accept x_i if a random sample u from the uniform distribution on $(0, 1)$ satisfies:

$$u \leq \frac{f(x_i|Y_i, \mathbf{W}_i)}{cf(x_i|\mathbf{W}_i)} = f(Y_i|\mathbf{W}_i) \frac{f(Y_i|x_i)}{f(Y_i|\mathbf{W}_i)} = f(Y_i|x_i). \quad (4.14)$$

The latter conditional probability is simply that given by the logistic regression model specification:

$$f(Y_i|x_i) = \frac{(\exp(\beta_0 + \beta_X x_i))^{Y_i}}{1 + \exp(\beta_0 + \beta_X x_i)}$$

4.5.3 Ascent-based Monte-Carlo Expectation Maximization

Because $\tilde{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$ is only an estimate of $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$, $\tilde{\boldsymbol{\theta}}^{(t)}$ is only an estimate of $\boldsymbol{\theta}^{(t)}$, the next parameter estimate which would be obtained if one could perform the E-step exactly, without Monte-Carlo approximation. If the number of imputations M used at each iteration is held fixed, the resulting iterates of the parameter $\boldsymbol{\theta}$ do not converge because of persistent Monte-Carlo error. Instead, the iterations converge to a stationary Markov chain [75]. In Wei and Tanner's original proposal for MCEM, they suggested that in the early iterations of EM, when the increase in log likelihood is largest, small values of M are suitable. As the EM algorithm converges and each iteration brings a smaller increase in the log likelihood, the number of imputations M must be increased. However, they did not give a specific prescription for how the number of imputations should be increased.

Over the last 10 years a number of proposals have been made for how to increase the number of imputations M as the MCEM algorithm converges. Booth and Hobert showed that given the current parameter estimate $\boldsymbol{\theta}^{(t-1)}$, the parameter estimate that will be obtained by Monte-Carlo EM by imputing the missing data is approximately a random draw from a normal distribution, with mean equal to the next parameter estimate which would be obtained as $M \rightarrow \infty$ ($\boldsymbol{\theta}^{(t)}$) and a variance which can be estimated [76]. They proposed constructing a confidence interval for $\boldsymbol{\theta}^{(t)}$. If the previous parameter value $\boldsymbol{\theta}^{(t-1)}$ lies in this confidence interval, Booth and Hobert suggested that 'the EM step was swamped by Monte Carlo error', and that the number of imputations M needs to be increased.

More recently, Caffo *et al* have proposed an alternative implementation of MCEM, which they call 'ascent-based MCEM', which we now describe [74]. In the standard EM algorithm, we maximize the Q function (equation (4.10)), which is the expected complete data log likelihood. In fact, so long as:

$$Q(\tilde{\boldsymbol{\theta}}^{(t)}|\boldsymbol{\theta}^{(t-1)}) \geq Q(\boldsymbol{\theta}^{(t-1)}|\boldsymbol{\theta}^{(t-1)}) \quad (4.15)$$

it follows by an application of Jensen's inequality that:

$$L_n(\tilde{\boldsymbol{\theta}}^{(t)}) \geq L_n(\boldsymbol{\theta}^{(t-1)}) \quad (4.16)$$

i.e. that the observed data likelihood value at $\tilde{\boldsymbol{\theta}}^{(t)}$ is at least as great as that at $\boldsymbol{\theta}^{(t-1)}$. This is what Caffo *et al* refer to as the ascent property of EM. As noted earlier, when the E-step is approximated using Monte-Carlo integration, there is no guarantee that the new parameter estimate, $\tilde{\boldsymbol{\theta}}^{(t)}$ has a higher observed data likelihood value than the previous value. Caffo *et al* thus proposed a version of MCEM in which the ascent property holds with high probability, by checking that the Q function is increased at each iteration.

The idea of ascent-based MCEM is that at each iteration, an asymptotic confidence interval is calculated for:

$$Q(\tilde{\boldsymbol{\theta}}^{(t)} | \tilde{\boldsymbol{\theta}}^{(t-1)}) - Q(\tilde{\boldsymbol{\theta}}^{(t-1)} | \tilde{\boldsymbol{\theta}}^{(t-1)}) \quad (4.17)$$

If the lower limit of the confidence interval is positive, then the new parameter estimate $\tilde{\boldsymbol{\theta}}^{(t)}$ is accepted. If not, additional imputations are generated to increase M to obtain another new parameter estimate. This process is then repeated until the lower limit of the confidence interval is positive.

Suppose that for each subject we generate M_t imputations $X_i^{(m)}$, $m = 1, \dots, M_t$ from the conditional distribution $f(\mathbf{X}_i | Y_i, \mathbf{W}_i, \tilde{\boldsymbol{\theta}}^{(t-1)})$. Using this Monte-Carlo sample to approximate the E-step, we find an updated parameter estimate $\tilde{\boldsymbol{\theta}}^{(t)}$ as previously described. We can estimate the increase in the Q function:

$$\Delta Q(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) = Q(\tilde{\boldsymbol{\theta}}^{(t)} | \tilde{\boldsymbol{\theta}}^{(t-1)}) - Q(\tilde{\boldsymbol{\theta}}^{(t-1)} | \tilde{\boldsymbol{\theta}}^{(t-1)}) \quad (4.18)$$

by the empirical mean:

$$\Delta \tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) = \tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)} | \tilde{\boldsymbol{\theta}}^{(t-1)}) - \tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t-1)} | \tilde{\boldsymbol{\theta}}^{(t-1)}) \quad (4.19)$$

$$= \frac{\sum_{m=1}^{M_t} l_n(\tilde{\boldsymbol{\theta}}^{(t)} | Y_i, X_i^{(m)}, \mathbf{W}_i, i = 1, \dots, n)}{M_t} \quad (4.20)$$

$$- \frac{\sum_{m=1}^{M_t} l_n(\tilde{\boldsymbol{\theta}}^{(t-1)} | Y_i, X_i^{(m)}, \mathbf{W}_i, i = 1, \dots, n)}{M_t} \quad (4.21)$$

which converges to $\Delta Q(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)})$ as $M_t \rightarrow \infty$. Caffo *et al* showed that:

$$\sqrt{M_t}(\Delta \tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) - \Delta Q(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)})) \quad (4.22)$$

is asymptotically normal with mean zero and variance σ_Δ^2 which can be estimated by the sample variance of the M_t terms:

$$\sum_{i=1}^n \left(\log(f(Y_i, X_i^{(m)}, \mathbf{W}_i | \tilde{\boldsymbol{\theta}}^{(t)})) - \log(f(Y_i, X_i^{(m)}, \mathbf{W}_i | \tilde{\boldsymbol{\theta}}^{(t-1)})) \right) \quad (4.23)$$

Given $\hat{\sigma}_\Delta^2$, we can calculate an asymptotic standard error (ASE) for

$$\Delta\tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) - \Delta Q(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) \quad (4.24)$$

by $ASE = \hat{\sigma}_\Delta^2 / \sqrt{M_t}$. Denoting by z_α the value such that $P(Z > z_\alpha) = \alpha$, where $Z \sim N(0, 1)$, it follows that

$$\Delta\tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) - z_\alpha ASE < \Delta Q(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) \quad (4.25)$$

with probability $1 - \alpha$ as $M_t \rightarrow \infty$. Thus if $\Delta\tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) - z_\alpha ASE$ is greater than zero, we conclude that there is a high probability that $\tilde{\boldsymbol{\theta}}^{(t)}$ increases the likelihood, and so it is accepted as the next parameter estimate. If not, Caffo *et al* suggested that the Monte-Carlo sample of size M_t is appended, by adding a further M_t/k (rounded, if necessary, to the nearest integer) imputations, for some chosen positive integer k .

Choosing the number of imputations

In our simulations and subsequent analyses, we arbitrarily chose to use 10 imputations at the start of the first iteration of MCEM. At the start of subsequent iterations, we must choose the initial number M_{t+1} of imputations to use to approximate the expectation in the E-step. Recall that

$$\Delta\tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t+1)}, \tilde{\boldsymbol{\theta}}^{(t)}) \sim N \left(\Delta Q(\boldsymbol{\theta}^{(t+1)}, \tilde{\boldsymbol{\theta}}^{(t)}), \frac{\sigma_\Delta^2}{M_{t+1}} \right) \quad (4.26)$$

Caffo *et al* therefore proposed using a standard sample size calculation to choose M_{t+1} as:

$$M_{t+1, start} = \max \left(M_{t, start}, \frac{\hat{\sigma}_\Delta^2 (z_\alpha + z_\beta)^2}{(\Delta\tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}))^2} \right) \quad (4.27)$$

where we use the estimate $\hat{\sigma}_\Delta^2$ from the last iteration and $\Delta\tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)})$ as an estimate of $\Delta Q(\boldsymbol{\theta}^{(t+1)}, \tilde{\boldsymbol{\theta}}^{(t)})$. By taking the maximum of $M_{t, start}$ and the estimated sample size, we force the starting number of imputations between iterations to increase. In our implementation of the method, we only use the number of imputations given by equation (4.27) if it is greater than the total number used in the previous iteration, i.e. we never decrease the number of imputations.

Determining convergence

Since we cannot evaluate the observed data likelihood function directly, we cannot use the change in observed data likelihood to judge whether the MCEM algorithm has converged. Prior to the proposal of ascent-based MCEM, Booth and Hobert suggested terminating MCEM when the relative change in parameter estimates between consecutive iterations is small [76]. One problem with this approach is that there may be a small change in parameter estimates between two iterations in MCEM by chance, since the new parameter estimate is a stochastic estimate of the next parameter value which would found with exact EM.

Caffo *et al* therefore proposed using the estimated increase in the Q -function to decide when to declare convergence of MCEM. If we let z_γ denote the value such that $P(Z > z_\gamma) = \gamma$ for a standard normal deviate Z , we can calculate the upper limit of a confidence interval for $\Delta Q(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)})$ as:

$$\Delta\tilde{Q}(\tilde{\boldsymbol{\theta}}^{(t)}, \tilde{\boldsymbol{\theta}}^{(t-1)}) + z_\gamma ASE. \quad (4.28)$$

Caffo *et al* proposed declaring convergence of MCEM when this confidence upper limit is less than a specified positive value. This means stopping when the increase in Q afforded by additional iterations of MCEM is small.

4.5.4 Standard errors

As we have previously noted, the EM algorithm does not provide estimates of precision automatically. A number of proposals have therefore been made as to how the observed information matrix might be calculated using quantities that are available from the EM algorithm. Based on an expression given by Oakes [77], Robert and Casella [50] showed that:

$$I_n(\boldsymbol{\theta}|Y_i, \mathbf{W}_i, i = 1, \dots, n) = \mathbb{E}(I_n(\boldsymbol{\theta}|Y_i, X_i, \mathbf{W}_i, i = 1, \dots, n)) - \text{Var}(S_n(\boldsymbol{\theta}|Y_i, X_i, \mathbf{W}_i, i = 1, \dots, n)) \quad (4.29)$$

where the expectation and variance are with respect to the conditional distribution $f(X_i|Y_i, \mathbf{W}_i)$. For each imputation generated in the final iteration of MCEM, we can calculate the complete data information matrix and score vector, which are standard expressions for logistic regression and normal models. The required expectation and variance covariance can then be approximated by their respective Monte-Carlo estimates over the imputations, and thus the observed data information matrix $I_n(\boldsymbol{\theta}|Y_i, \mathbf{W}_i, i = 1, \dots, n)$ can be estimated. This can then be inverted in the usual way to find asymptotic variances.

4.5.5 MCEM compared to EM using quadrature

We briefly consider some of the potential advantages and disadvantages of MCEM compared to EM using quadrature methods to approximate the E-step. Looking ahead to situations in which \mathbf{X}_i is a p -dimensional multivariate vector, the E-step involves a p -dimensional integral. The Gaussian quadrature technique requires evaluating the integrand at a number of points in the domain of integration. When the integral is multi-dimensional, the number of evaluations required in quadrature grows as a power of the dimension p . Thus for p greater than four or five, quadrature with a reasonable number of points requires evaluating the integrand an extremely large number of times. This means that while quadrature may be feasible when p is small (e.g. one to three say), it rapidly becomes infeasible for a higher-dimensional vector \mathbf{X}_i . As the dimension of \mathbf{X}_i increases, use of rejection sampling to generate the required imputations may also become problematic however, because the upper bound c in general increases exponentially with the dimension of the integral (see Chapter 29 of [78]).

An advantage of MCEM is that having imputed the unobserved X_i in the E-step, standard complete data commands can be used to maximize the complete data log likelihoods for each imputation. This is especially advantageous when a component of the complete data log likelihood can only be maximized using an iterative method such as Newton-Raphson. For example, for the model we have defined, a logistic regression model can be fitted to each imputed dataset, using the imputed values of X_i . In contrast, when using quadrature methods, one must typically write a custom-routine for maximizing the component. Although there are no statistical software commands for MCEM, only a moderate amount of programming in a package such as R or Stata is necessary.

One potential issue for MCEM is that as the algorithm converges to the MLE, the number of imputations required to approximate the E-step becomes very large. Even when X_i is a scalar, this means that we must generate and (temporarily) store a matrix of size $n \times M$, where M denotes the number of imputations. Thus it is usually not possible to determine convergence as precisely as would normally be possible using deterministic methods. However, this is mitigated by the ever increasing computational power and memory capacity available to researchers.

4.6 Maximum likelihood estimation using linear mixed models

In parametric approaches to covariate measurement error, the most common model assumed for the covariate X_i is a normal distribution. As we have seen in Section 4.4, this leads to an intractable observed data likelihood function, which makes

finding the MLE much more difficult. In this section we describe an alternative model specification for which the MLE can be obtained by fitting a simple random-intercepts model.

The approach we have taken thus far to specifying a joint model for Y_i , X_i and \mathbf{W}_i has, under the non-differential error assumption, involved specifying $f(Y_i|X_i)$, $f(\mathbf{W}_i|X_i)$, and $f(X_i)$. Here, we specify the joint distribution by instead specifying $f(X_i|Y_i)$ and $f(Y_i)$. Since Y_i is binary, it is marginally Bernoulli with $P(Y_i = 1) = \pi_1$. For $f(X_i|Y_i)$ we assume X_i is conditionally normal given Y_i , with $\text{Var}(X_i|Y_i = 0) = \text{Var}(X_i|Y_i = 1) = \sigma_{X|Y}^2$. Then we can write:

$$X_i = \gamma_0 + \gamma_Y Y_i + b_i$$

where $b_i \sim N(0, \sigma_{X|Y}^2)$ is a normally distributed residual that is independent of Y_i . The assumption of conditional normality for X_i given Y_i with constant variance implies that Y_i given X_i follows a logistic regression with intercept β_0 and coefficient for X_i β_X , where (see Section 4.2.2):

$$\beta_0 = \log\left(\frac{\pi_1}{1 - \pi_1}\right) - \frac{1}{2\sigma_{X|Y}^2}(\gamma_Y^2 + \gamma_Y \gamma_0) \quad (4.30)$$

$$\beta_X = \frac{\gamma_Y}{\sigma_{X|Y}^2}. \quad (4.31)$$

As before, we assume subject i has $j = 1, \dots, n_i$ error-prone measurements:

$$W_{ij} = X_i + U_{ij}$$

where $U_{ij} \sim N(0, \sigma_U^2)$. As usual, we assume the measurement errors are independent of each other, of X_i and of Y_i . These assumptions imply that:

$$W_{ij} = \gamma_0 + \gamma_Y Y_i + b_i + U_{ij}$$

Thus, as for a linear regression outcome model (see Section 3.6), W_{ij} follows a standard random-intercepts model, with a fixed effect of Y_i , random-intercept variance $\sigma_{X|Y}^2$, and within-subject variance σ_U^2 . This model can thus be fitted using linear mixed model commands in statistical packages, using either maximum likelihood or restricted maximum likelihood. By the invariance property of maximum likelihood, the maximum likelihood estimates of β_0 and β_X can be obtained by replacing the parameters in equations (4.30) and (4.31) by their respective estimates.

We can find an asymptotic Wald confidence interval for $\hat{\beta}_X$ using the reported standard errors for $\hat{\gamma}_Y$ and $\hat{\sigma}_{X|Y}^2$. Using the multivariate delta method [28], after

partial differentiation of equation (4.31) we have:

$$\text{Var}(\hat{\beta}_X) = \frac{\text{Var}(\hat{\gamma}_Y)}{\sigma_{X|Y}^4} + \frac{\gamma_Y^2 \text{Var}(\hat{\sigma}_{X|Y}^2)}{\sigma_{X|Y}^8}$$

Substituting estimates of γ_Y and $\sigma_{X|Y}^2$ and their standard errors gives an estimate of $\text{Var}(\hat{\beta}_X)$, from which a Wald type confidence interval can be calculated. This can then be transformed to the odds-ratio scale by exponentiating.

An alternative confidence interval for $\hat{\beta}_X$ can be found using Fieller's theorem [79], since asymptotically it is the ratio of two normally distributed random variables. The estimators of fixed effects and variance components in linear mixed models are asymptotically uncorrelated [24], and so $\hat{\gamma}_Y$ and $\hat{\sigma}_{X|Y}^2$ are asymptotically uncorrelated. Using Fieller's theorem, a 95% confidence interval for $\hat{\beta}_X$ can be found as:

$$\left(\frac{f_1 - \sqrt{f_1^2 - f_0 f_2}}{f_2}, \frac{f_1 + \sqrt{f_1^2 - f_0 f_2}}{f_2} \right)$$

where

$$\begin{aligned} f_0 &= \hat{\gamma}_Y^2 - 1.96^2 \text{Var}(\hat{\gamma}_Y) \\ f_1 &= \hat{\gamma}_Y \hat{\sigma}_{X|Y}^2 \\ f_2 &= \hat{\sigma}_{X|Y}^4 - 1.96^2 \text{Var}(\hat{\sigma}_{X|Y}^2) \end{aligned}$$

4.6.1 Multiple covariates

The approach can be extended to a more general setup in which there are multiple covariates measured with error, \mathbf{X}_i , and error-free covariates \mathbf{Z}_i , under the assumption that these are jointly normal given Y_i . Thus suppose that:

$$\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Z}_i \end{pmatrix} \sim N \left(\begin{pmatrix} \gamma_{X0} + \gamma_{XY} Y_i \\ \gamma_{Z0} + \gamma_{ZY} Y_i \end{pmatrix}, \begin{pmatrix} \Sigma_{X|Y} & \Sigma_{XZ|Y} \\ \Sigma_{ZX|Y} & \Sigma_{Z|Y} \end{pmatrix} \right)$$

This implies that Y_i follows a logistic regression given \mathbf{X}_i and \mathbf{Z}_i , with log odds ratios:

$$\begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} = \begin{pmatrix} \Sigma_{X|Y} & \Sigma_{XZ|Y} \\ \Sigma_{ZX|Y} & \Sigma_{Z|Y} \end{pmatrix}^{-1} \begin{pmatrix} \gamma_{XY} \\ \gamma_{ZY} \end{pmatrix} \quad (4.32)$$

It also follows that \mathbf{X}_i is normal given \mathbf{Z}_i and Y_i , with:

$$\begin{aligned} \mathbb{E}(\mathbf{X}_i | \mathbf{Z}_i, Y_i) &= \gamma_0 + \gamma_Y Y_i + \gamma_Z \mathbf{Z}_i \\ \text{Var}(\mathbf{X}_i | \mathbf{Z}_i, Y_i) &= \Sigma_{X|Z,Y} \end{aligned}$$

where

$$\begin{aligned}
\boldsymbol{\gamma}_0 &= \boldsymbol{\gamma}_{X0} - \boldsymbol{\Sigma}_{XZ|Y} \boldsymbol{\Sigma}_{Z|Y}^{-1} \boldsymbol{\gamma}_{Z0} \\
\boldsymbol{\gamma}_Y &= \boldsymbol{\gamma}_{XY} - \boldsymbol{\Sigma}_{XZ|Y} \boldsymbol{\Sigma}_Z^{-1} \boldsymbol{\gamma}_{ZY} \\
\boldsymbol{\gamma}_Z &= \boldsymbol{\Sigma}_{XZ|Y} \boldsymbol{\Sigma}_{Z|Y}^{-1} \\
\boldsymbol{\Sigma}_{X|Z,Y} &= \boldsymbol{\Sigma}_X - \boldsymbol{\Sigma}_{XZ|Y} \boldsymbol{\Sigma}_{Z|Y}^{-1} \boldsymbol{\Sigma}_{ZX|Y}
\end{aligned}$$

Now we assume error-prone measurements of \mathbf{X}_i are available as described in Section 3.6.3. In this case we have:

$$\begin{aligned}
\mathbb{E}(\mathbf{W}_i | \mathbf{Z}_i, Y_i) &= \mathbf{D}_i \boldsymbol{\gamma}_0 + \mathbf{D}_i \boldsymbol{\gamma}_Y Y_i + \mathbf{D}_i \boldsymbol{\gamma}_Z \mathbf{Z}_i \\
\text{Var}(\mathbf{W}_i | \mathbf{Z}_i, Y_i) &= \mathbf{D}_i \boldsymbol{\Sigma}_{X|Z,Y} \mathbf{D}_i^T + \text{Var}(\mathbf{U}_i)
\end{aligned}$$

As in Section 3.6.3, if we now define:

$$\mathcal{D}_i = \begin{pmatrix} \mathbf{D}_i & Y_i \mathbf{D}_i & Z_{i1} \mathbf{D}_i & \dots & Z_{iq} \mathbf{D}_i \end{pmatrix} \quad (4.33)$$

and:

$$\boldsymbol{\gamma} = \left(\boldsymbol{\gamma}_0^T \quad \boldsymbol{\gamma}_Y^T \quad \boldsymbol{\gamma}_{Z_1}^T \quad \dots \quad \boldsymbol{\gamma}_{Z_q}^T \right)^T$$

then:

$$\mathbb{E}(\mathbf{W}_i | \mathbf{Z}_i, Y_i) = \mathcal{D}_i \boldsymbol{\gamma}.$$

Hence this again is a linear mixed model, with fixed effects design matrix \mathcal{D}_i , random effects design matrix \mathbf{D}_i , and diagonal residual covariance matrix $\text{Var}(\mathbf{U}_i)$ as given in equation (3.73).

Multivariate regression of \mathbf{Z}_i on Y_i can be used to find the ML estimates of $\boldsymbol{\gamma}_{Z0}$, $\boldsymbol{\gamma}_{ZY}$ and $\boldsymbol{\Sigma}_{Z|Y}$. Fitting the above linear mixed model to \mathbf{W}_i given \mathbf{Z}_i and Y_i then gives the ML estimates of $\boldsymbol{\gamma}_0$, $\boldsymbol{\gamma}_Y$, $\boldsymbol{\gamma}_Z$ and $\boldsymbol{\Sigma}_{X|Z,Y}$. It is then straightforward to show that:

$$\begin{aligned}
\boldsymbol{\gamma}_{XY} &= \boldsymbol{\gamma}_Y + \boldsymbol{\gamma}_Z \boldsymbol{\gamma}_{ZY} \\
\boldsymbol{\Sigma}_{XZ|Y} &= \boldsymbol{\gamma}_Z \boldsymbol{\Sigma}_{Z|Y} \\
\boldsymbol{\Sigma}_{X|Y} &= \boldsymbol{\Sigma}_{X|Z,Y} + \boldsymbol{\Sigma}_{XZ|Y} \boldsymbol{\Sigma}_{Z|Y}^{-1} \boldsymbol{\Sigma}_{ZX|Y}
\end{aligned}$$

These formulae can be used to calculate the MLEs of these parameters, and then these can be inserted into equation (4.32) to calculate the MLEs of $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$.

As for a linear regression outcome model, although Wald type intervals can in principle be calculated for $\hat{\boldsymbol{\beta}}_X$ and $\hat{\boldsymbol{\beta}}_Z$, it may be programmatically simpler to instead use non-parametric bootstrapping for inference.

4.7 Multiple imputation

In the case of scalar X_i and no error-free covariates \mathbf{Z}_i , under the parametric model defined in Section 4.6, which assumes conditional normality for $X_i|Y_i$, the MLE for β_X can be found by fitting a linear mixed model for \mathbf{W}_i given Y_i . Given the estimated parameters of this mixed model, we can multiply impute the unobserved X_i from its conditional distribution given the observed data Y_i and \mathbf{W}_i , in exactly the same way as described for a linear regression outcome model in Section 3.7.

Given M imputations we can fit M logistic regression models for Y_i given the imputed X_i values, and average the resulting estimates of β_X across imputations. If the assumed model is valid, this gives consistent estimates of β_X . Even as $M \rightarrow \infty$, we believe the resulting estimator is inefficient, because for the assumed normal discriminant model, estimating β_X by fitting the logistic regression model does not utilize the assumption of conditional normality for $X_i|Y_i$ [80]. A more efficient estimate would be obtained by fitting the normal discriminant model to the imputed data, i.e. estimating the conditional variance of $X_i|Y_i$ and the difference in the conditional means of X_i between $Y_i = 1$ and $Y_i = 0$.

4.7.1 Multiple covariates

A major drawback of the ML approach described in Section 4.6, based on fitting a linear mixed model for \mathbf{W}_i given Y_i , is that when error-free covariates \mathbf{Z}_i are present, the method can only be used under the restrictive assumption that \mathbf{X}_i and \mathbf{Z}_i are jointly normal given Y_i . Often \mathbf{Z}_i may contain discrete covariates, for which an assumption of normality is clearly untenable. To address this limitation, in this section we describe how the previously described linear mixed model for \mathbf{W}_i given \mathbf{Z}_i and Y_i can be used to multiply impute \mathbf{X}_i , under certain parametric assumptions, and the resulting imputations used to consistently estimate the logistic regression model parameters.

Thus, as in Section 4.6.1, we assume that \mathbf{X}_i is normal given \mathbf{Z}_i and Y_i , with:

$$\begin{aligned}\mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i, Y_i) &= \boldsymbol{\gamma}_0 + \boldsymbol{\gamma}_Y Y_i + \boldsymbol{\gamma}_Z \mathbf{Z}_i \\ \text{Var}(\mathbf{X}_i|\mathbf{Z}_i, Y_i) &= \boldsymbol{\Sigma}_{X|Z, Y}.\end{aligned}$$

This implies that \mathbf{X}_i given \mathbf{Z}_i (marginalized over Y_i) is a mixture of two multivariate normals with mixing proportions equal to $P(Y_i = 1)$ and $1 - P(Y_i = 1)$. We then assume that Y_i follows a logistic regression given \mathbf{X}_i and \mathbf{Z}_i , with corresponding vectors of log odds ratios $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$. Note that, as in the case of a continuous outcome, we do not need to specify the marginal distribution of the error-free covariates \mathbf{Z}_i . We then assume that \mathbf{X}_i is measured with normally distributed error by \mathbf{W}_i , as

described in Section 3.6.3, so that \mathbf{W}_i follows the same linear mixed model given Y_i and \mathbf{Z}_i as described in Section 4.6.1.

We can fit the linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i using ML or REML, using standard linear mixed model commands. Analogous to the scalar X_i case, described for continuous Y_i in Section 3.7, we can multiply impute \mathbf{X}_i from its conditional distribution (which is normal) given Y_i , \mathbf{Z}_i , and \mathbf{W}_i . Given M imputations $\mathbf{X}_i^{(m)}$, $m = 1, \dots, M$, we can then fit the logistic regression of Y_i to \mathbf{X}_i and \mathbf{Z}_i . If the parametric assumptions hold, the resulting estimates of β_X and β_Z are consistent.

4.8 Moment reconstruction

Moment reconstruction (MR) (see Section 3.8) can also be used with a logistic regression outcome model [57]. MR for logistic regression is based on the normal discriminant model for X_i given Y_i (see Section 4.2.2). Thus we assume that $X_i|Y_i \sim N(\gamma_0 + \gamma_Y Y_i, \sigma_{X|Y}^2)$. Since a normal distribution is defined entirely by its mean and variance, it follows that X_i^{mr} (defined in Section 3.8 given Y_i) has the same conditional distribution as X_i given Y_i . It also follows that the joint distribution of X_i^{mr} and Y_i is the same as the joint distribution of X_i and Y_i , and so parameters that are consistently estimated using data on (Y_i, X_i) can be consistently estimated using (Y_i, X_i^{mr}) data. In particular, we can fit the logistic regression of Y_i on X_i^{mr} to estimate β_0 and β_X .

If each subject has n_i error-prone measurements of X_i , which are subject to independent normally distributed error with variance σ_U^2 , we have shown in Section 4.6 that the MLE of β_X can be obtained by fitting a linear mixed model for \mathbf{W}_i given Y_i . Fitting this model also gives the MLEs of the parameters required to calculate X_i^{mr} , as we have previously described.

Although the MR estimates of β_X are consistent, under the assumed modelling assumptions, we believe they must be inefficient, compared to the MLE, for the same reason as we believe fitting the logistic regression for Y_i given multiply imputed values of X_i is inefficient. This is because by fitting a logistic regression model for Y_i given X_i^{mr} , we ignore the additional assumption that X_i is conditionally normal given Y_i . By not using this additional assumption we believe efficiency must be lost in the same way as using logistic regression is inefficient compared to normal discriminant analysis when the normal discriminant model holds [80].

4.9 Conditional score method

In 1987 Stefanski and Carroll proposed a number of methods for allowing for covariate measurement error in generalized linear outcome models which make no

assumptions about the distribution of the true covariate X_i [71]. We describe one of their proposed methods, the so called conditional score (CS) method.

4.9.1 Estimating equations

Before describing the CS estimator, we recall that in the absence of measurement error, the logistic regression likelihood score function is given by

$$\psi_{ML}(Y_i, X_i, \beta_0, \beta_X) = \begin{pmatrix} 1 \\ X_i \end{pmatrix} \left(Y_i - \frac{\exp(\beta_0 + \beta_X X_i)}{1 + \exp(\beta_0 + \beta_X X_i)} \right)$$

We assume that $W_i = X_i + U_i$ is an error-prone measurement error with $U_i \sim N(0, \sigma_U^2)$ which is independent of X_i and Y_i . For now we assume that σ_U^2 is known. Now we define

$$\tilde{X}_i = W_i + Y_i \sigma_U^2 \beta_X. \quad (4.34)$$

The motivation for this definition is that one can show that if X_i is viewed as a parameter, \tilde{X}_i is a complete and sufficient statistic for X_i . Stefanski and Carroll showed that:

$$\begin{aligned} \mathbb{E}(Y_i | X_i, \tilde{X}_i) &= \frac{\exp(\beta_0 + \beta_X \tilde{X}_i - \beta_X^2 \sigma_U^2 / 2)}{1 + \exp(\beta_0 + \beta_X \tilde{X}_i - \beta_X^2 \sigma_U^2 / 2)} \\ &= \mathbb{E}(Y_i | \tilde{X}_i). \end{aligned} \quad (4.35)$$

We now define the CS estimating function by:

$$\psi_{CS}(Y_i, W_i, \beta_0, \beta_X) = \begin{pmatrix} 1 \\ \tilde{X}_i \end{pmatrix} (Y_i - \mathbb{E}(Y_i | \tilde{X}_i)). \quad (4.36)$$

Then it follows that:

$$\mathbb{E}(\psi_{CS}(Y_i, W_i, \beta_0, \beta_X) | \tilde{X}_i) = 0.$$

It then also follows that the CS function has marginal (unconditional) mean of zero. This means that we can form unbiased estimating equations:

$$\sum_{i=1}^n \psi_{CS}(Y_i, W_i, \beta_0, \beta_X) = 0.$$

Stefanski and Carroll showed that the CS estimating equations have a solution which is consistent for β_0 and β_X . Note that when $\sigma_U^2 = 0$, the CS estimating equations reduce to the usual likelihood score equations for logistic regression.

4.9.2 Implementation

The CS estimating equations are non-linear and must therefore be solved by iterative methods, such as the Newton-Raphson method. Depending on the method used to solve the equations, the derivatives of the estimating functions with respect to β_0 and β_X may be required. At least two problems may occur when implementing the method. First, methods such as Newton-Raphson may diverge if the initial parameter estimate is not sufficiently close to the root of the equations. Carroll *et al* reported (Section 7.5.1) that using the naive estimator of β_0 and β_X as the initial estimates in Newton-Raphson often works well, but that sometimes a bias-corrected initial estimate, such as that from RC, may be necessary. Second, we may find a solution to the estimating equations, but it may not be the root which leads to consistent estimates.

In our description of the CS method we have assumed that the measurement error variance is known and that each subject has a single error-prone measurement W_i . If replication data are available, we can first obtain an estimate $\hat{\sigma}_U^2$. The CS estimating equation is then modified by replacing W_i by \bar{W}_i , and σ_U^2 by σ_U^2/n_i .

4.9.3 Inference

For inference, the standard sandwich estimator of variance for M-estimators can be used, as described in Section 7.5.1 of Carroll *et al* [8]. This is a function of the CS estimating function and its derivatives, evaluated at the CS estimate. If σ_U^2 has been estimated however, this should be accounted for. Carroll *et al* [8] show how this can be done by stacking the CS estimating equations with those used to estimate σ_U^2 when internal replication data are used (Section 7.5.2).

4.10 Simulations

4.10.1 Simulations with X_i marginally normal

We first conducted simulations in which data were generated according to the model defined in Section 4.4.1. We simulated $X_i \sim N(0, 1)$, and Y_i given X_i as a logistic regression with $\beta_X = 0.1$ or $\beta_X = 1$, representing a weak and moderately strong association. We varied β_0 so that $P(Y_i = 1)$ was either 0.1 or 0.5, representing a relatively rare outcome and a common outcome. As for the simulations in Section 3.9, we simulated data for $n = 5,000$ subjects, 500 of which had two measurements of X_i subject to normally distributed error, while the remaining 4,500 had a single error-prone measurement of X_i . As before, we varied σ_U^2 to give values of the reliability ratio of 2/3, 1/2, and 1/3. We performed 10,000 simulations for each scenario.

We estimated β_X using RC, ML via ascent-based MCEM, mis-specified ML under the assumption of $X_i|Y_i$ normal, and using the CS method. The R code used for the simulations is given in Listing 14.1.

Regression calibration

RC was implemented in exactly the same way as for the linear regression simulations (see Section 3.9).

Maximum likelihood

We found the MLE of β_X for the data-generating model using ascent-based MCEM, as described in Section 4.5. For each simulated dataset we used the estimated parameter values from RC at the beginning of MCEM, and used 10 imputations in the first iteration. To use ascent-based MCEM we must specify values of the control parameters α , β and γ as described in Section 4.5. We used $\alpha = 0.25$, $\beta = 0.25$ and $\gamma = 0.05$, as these were reported by Caffo *et al* [74] to work well in examples in which they applied ascent-based MCEM. We declared convergence either when the number of imputations reached 1,000 or if the upper confidence interval limit for the increase in the Q function was less than 0.01. We used the final set of imputations to estimate the standard error of the estimate of β_X , as described in Section 4.5.4.

Maximum likelihood assuming $X_i|Y_i$ normal

We also found the MLE of β_X under the assumption that $X_i|Y_i$ is normal, by fitting the linear mixed model for \mathbf{W}_i given Y_i , as described in Section 4.6. We calculated 95% Wald confidence intervals for β_X , and also intervals using Fieller's theorem, as described in Section 4.6.

Conditional score

Initially, we used the Newton-Raphson algorithm to solve the CS estimating equations, using the estimates of β_0 , β_X and σ_U^2 obtained in RC as starting values. However, for scenarios 6 and 12, in which the effect size is moderate and the measurement error large, the Newton-Raphson algorithm often diverged, indicating that the RC estimates of β_0 and β_X were not sufficiently close to the root of the equations. We therefore used the R package 'nleqslv' to solve the CS estimating equations, using its default configuration. This procedure implements a so called global search strategy which aims to solve the equations even when the initial estimate leads (by simple Newton-Raphson) to divergence. For the purposes of reporting the simulation results, in those simulations for which a root of the estimating equations was not found, we substituted the RC estimate of β_X in place of the CS estimate.

Simulation results

Table 4.1 shows the empirical mean and standard deviation of the various estimators from the simulations. RC had little bias when $\beta_X = 0.1$. For $\beta_X = 1$, RC was biased towards the null by up to 10% of the true value, with the bias larger when the outcome was more common and when the reliability ratio of the error-prone measurements was lower. Ascent-based MCEM had little bias for all of the scenarios. The ML estimator based on an assumption that $X_i|Y_i$ is normal also had only small biases, despite being misspecified. It was less variable than the ascent-based MCEM estimator, except for scenarios 11 and 12. Except for scenarios 6 and 12, the CS estimator had little bias, and had variability which was within 15% of that of the correctly specified MLE, found by MCEM. For scenarios 6 and 12, no solution to the CS estimating equations was found in 8 and 33 (of the 10,000) simulations respectively. The bias in the CS estimates in scenarios 6 and 12 is presumably due to the Newton-Raphson algorithm converging to non-consistent roots of the CS estimating equations.

We now consider the reasons for the lack of bias in the ML estimator based on an assumption that $X_i|Y_i$. If $X_i \sim N(\mu_X, \sigma_X^2)$, then unless $\beta = 0$, the conditional distribution of X_i given Y_i is not normal, and so the ML estimator which assumes $X_i|Y_i$ normal is not ML for the true data generating process. However, as noted by Freedman *et al* [21, 57], if either the outcome Y_i is rare or the effect of X_i on Y_i is small, the distribution of X_i is approximately both marginally and conditionally normal given Y_i . An additional consequence of marginal normality for X_i is that $\text{Var}(X_i|Y_i)$ depends on Y_i . If we fit the linear mixed model for \mathbf{W}_i given Y_i , but assume that $\text{Var}(X_i|Y_i)$ is independent of Y_i , the ML estimator of $\sigma_{X|Y}^2$ is a consistent estimator of the mean of the conditional variances $\text{Var}(X_i|Y_i)$ over the distribution of Y_i . This follows by the arguments given in Section 3.6.2. The bias in the ML estimator which assumes conditional normality of X_i when X_i is marginally normal thus reduces to the bias of the method of normal discriminant analysis when the assumptions of conditional normality and constant variance are violated. As noted by Freedman *et al* [57], normal discriminant analysis is known to be relatively robust to the assumption of conditional normality [81]. Freedman *et al* found that MR, which is also based on an assumption of normality for $X_i|Y_i$, had little bias in simulations even when X_i was marginally normally distributed, and suggested that this is likely due to the robustness of normal discriminant analysis. Since MR and the ML estimator based on fitting a mixed model for \mathbf{W}_i given Y_i both assume the normal discriminant model for X_i given Y_i , the two methods should consistently estimate the same value under model misspecification. We have also performed further simulations with larger values of β_X , in which the ML estimator predicated on normality of $X_i|Y_i$ had larger bias.

Table 4.1: Logistic regression simulation results with normally distributed covariate. Mean (SD) of estimates of β_X from regression calibration (RC), maximum likelihood (ML) using ascent-based Monte-Carlo EM, ML assuming $X_i|Y_i$ is normal using linear mixed models, and the conditional score method (CS). λ denotes the ratio of variance of X_i to variance of error-prone measurements.

Scenario	$P(Y_i = 1)$	β_X	λ	RC	ML via ascent-based MCEM	ML assuming $X_i Y_i$ normal	CS	
1	0.1	0.1	2/3	0.100 (0.058)	0.100 (0.058)	0.100 (0.058)	0.100 (0.058)	
2			1/2	0.100 (0.066)	0.102 (0.067)	0.101 (0.066)	0.102 (0.067)	
3			1/3	0.100 (0.082)	0.100 (0.083)	0.100 (0.082)	0.101 (0.085)	
4	0.5	1	2/3	0.968 (0.069)	1.006 (0.079)	0.985 (0.073)	1.009 (0.081)	
5			1/2	0.954 (0.091)	1.011 (0.109)	0.988 (0.100)	1.025 (0.117)	
6			1/3	0.948 (0.138)	1.025 (0.172)	0.999 (0.159)	1.143 (0.497)	
7	0.5	0.1	2/3	0.100 (0.035)	0.100 (0.035)	0.100 (0.035)	0.100 (0.035)	
8			1/2	0.100 (0.040)	0.100 (0.040)	0.100 (0.040)	0.100 (0.040)	
9			1/3	0.101 (0.051)	0.102 (0.051)	0.102 (0.051)	0.102 (0.052)	
10			1	2/3	0.938 (0.050)	1.001 (0.061)	1.000 (0.060)	1.006 (0.063)
11				1/2	0.911 (0.071)	0.997 (0.089)	1.003 (0.092)	1.020 (0.100)
12				1/3	0.893 (0.118)	0.997 (0.150)	1.018 (0.163)	1.710 (2.186)

Table 4.2 shows the empirical coverage rates of the various confidence intervals. The Wald-based naive confidence intervals from RC, which ignored the imprecision in the parameters needed to calculate $\mathbb{E}(X_i|\mathbf{W}_i)$, performed well for $\beta_X = 0.1$, but had poor coverage for $\beta_X = 1$. Especially when $\beta_X = 1$, the estimated variance for the MLE found using ascent-based MCEM of β_X using the method described in 4.5.4 was sometimes negative. When calculating the coverage of the Wald intervals for ascent-based MCEM, we therefore calculated the coverage proportion using those simulations in which the variance estimate was positive. We believe the negative variance estimates occurred because we only used a maximum of 1,000 imputations, and that with more imputations the probability of this occurring would reduce. For the MLE which assumes $X_i|Y_i$ normal, both the Wald confidence intervals and those based on Fieller’s theorem had coverage close to the nominal 95%. Inspection of the one-sided coverage rates showed however that the one-sided Wald intervals had the wrong coverage level when $\beta_X = 1$, whereas the Fieller intervals had approximately the correct one-sided coverage rates.

4.10.2 Simulations with $X_i|Y_i$ normal

We also performed simulations in which $X_i|Y_i$ was normally distributed. The simulation setup was otherwise the same as the previous simulations, except that we did not find the MLE using ascent-based MCEM which assumes marginal normality for X_i . For this data-generating model, the estimate of β_X obtained by fitting the mixed model for \mathbf{W}_i given Y_i (see Section 4.6) is the correctly specified MLE. We again performed 10,000 simulations per scenario. Tables 4.3 and 4.4 show the results of the simulations.

The performance of RC was very similar to when X_i was marginally normal, despite the fact that with $X_i|Y_i$ normally distributed, X_i was no longer marginally normal, and thus the parametric assumptions made by our implementation of RC (of marginal normality for X_i) were violated in these simulations. However, as previously discussed, when either $P(Y_i = 1)$ or β_X are small, the marginal distribution of X_i will not be far from normal (when $X_i|Y_i$ is normal). As expected, the ML estimator based on fitting a linear mixed model for \mathbf{W}_i given Y_i had little bias. While the RC estimator was biased towards the null, it was less variable than the MLE. The CS method performed similarly to that when X_i was simulated as marginally normal, with little bias for all scenarios, except 6 and 12.

As for the simulations in which X_i was marginally normal, the naive confidence intervals obtained using RC only had coverage rates close to the 95% nominal level when $\beta_X = 0.1$. The coverage of the ML Wald and Fieller confidence intervals was close to the nominal 95% level, although as before, the coverage rates of the one-sided Wald intervals deviated from the 97.5% level when $\beta_X = 1$, while the one-sided intervals based on Fieller’s theorem had the correct coverage.

Table 4.2: Logistic regression simulation results with normally distributed covariate. Empirical coverage rates of 95% confidence intervals for β_X (coverage of lower and upper one-sided 97.5% intervals): naive Wald intervals found using regression calibration (RC), Wald intervals from maximum likelihood (ML) using ascent-based Monte-Carlo EM, and Wald and Fieller intervals for the MLE assuming $X_i|Y_i$ is normal using linear mixed models. λ denotes the ratio of variance of X_i to variance of error-prone measurements.

Scenario	$P(Y_i = 1)$	β_X	λ	RC	ML via ascent-based MCEM	ML assuming $X_i Y_i$ normal		
				Naive Wald	Wald	Wald	Fieller	
1	0.1	0.1	2/3	94.8 (97.4, 97.4)	94.9 (97.4, 97.5)	94.8 (97.4, 97.4)	94.8 (97.4, 97.4)	
2			1/2	94.9 (97.4, 97.5)	94.7 (97.4, 97.3)	95.3 (97.8, 97.5)	95.0 (97.4, 97.5)	
3			1/3	94.8 (97.5, 97.3)	94.5 (97.5, 97.0)	95.5 (98.1, 97.4)	95.0 (97.5, 97.5)	
4	0.5	1	2/3	89.2 (98.8, 90.4)	93.9 (97.0, 96.9)	94.5 (99.0, 95.5)	94.8 (98.6, 96.2)	
5			1/2	81.1 (98.0, 83.2)	90.4 (96.0, 94.4)	93.9 (99.3, 94.6)	94.6 (98.6, 96.0)	
6			1/3	71.8 (95.3, 76.5)	81.9 (91.9, 90.0)	93.9 (99.9, 94.0)	94.9 (98.3, 96.7)	
7	0.5	0.1	2/3	95.0 (97.6, 97.4)	95.0 (97.7, 97.3)	95.1 (97.7, 97.4)	95.0 (97.6, 97.4)	
8			1/2	94.9 (98.0, 97.0)	94.8 (98.0, 96.9)	95.3 (98.3, 97.0)	95.2 (98.0, 97.2)	
9			1/3	94.3 (97.0, 97.2)	93.5 (96.7, 96.8)	95.5 (98.3, 97.2)	94.8 (97.2, 97.6)	
10			1	2/3	63.7 (99.8, 64.0)	93.0 (97.5, 95.6)	95.0 (98.7, 96.3)	95.0 (97.9, 97.2)
11				1/2	46.9 (99.3, 47.7)	90.8 (97.7, 93.2)	95.0 (99.5, 95.6)	95.0 (98.0, 97.0)
12				1/3	40.9 (96.1, 44.8)	80.3 (94.1, 86.2)	94.4 (100, 94.4)	95.0 (97.9, 97.1)

Table 4.3: Logistic regression simulation results with covariate conditionally normal given outcome. Mean (SD) of estimates from regression calibration (RC), ML assuming $X_i|Y_i$ is normal using linear mixed models, and the conditional score (CS) method. λ denotes the ratio of variance of X_i to variance of error-prone measurements.

Scenario	$P(Y_i = 1)$	β_X	λ	RC	ML ($X_i Y_i$ normal)	CS	
1	0.1	0.1	2/3	0.100 (0.058)	0.100 (0.058)	0.100 (0.058)	
2			1/2	0.100 (0.066)	0.101 (0.067)	0.101 (0.067)	
3			1/3	0.101 (0.082)	0.102 (0.083)	0.103 (0.085)	
4	0.5	0.1	2/3	0.977 (0.069)	1.004 (0.074)	1.008 (0.078)	
5			1/2	0.966 (0.091)	1.008 (0.103)	1.025 (0.114)	
6			1/3	0.962 (0.139)	1.019 (0.162)	1.126 (0.424)	
7	0.5	0.1	2/3	0.100 (0.034)	0.100 (0.035)	0.100 (0.035)	
8			1/2	0.099 (0.040)	0.100 (0.040)	0.100 (0.040)	
9			1/3	0.101 (0.050)	0.101 (0.050)	0.102 (0.051)	
10			1	2/3	0.939 (0.049)	1.002 (0.060)	1.005 (0.062)
11				1/2	0.913 (0.071)	1.007 (0.093)	1.021 (0.100)
12				1/3	0.892 (0.118)	1.020 (0.164)	1.176 (0.813)

Table 4.4: Logistic regression simulation results with covariate conditionally normal given outcome. Empirical coverage rates of 95% confidence intervals for β_X (coverage of lower and upper one-sided 97.5% intervals): naive Wald intervals found using regression calibration (RC) and Wald and Fieller intervals for the MLE assuming $X_i|Y_i$ is normal using linear mixed models. λ denotes the ratio of variance of X_i to variance of error-prone measurements.

Scenario	$P(Y_i = 1)$	β_X	λ	RC		ML assuming $X_i Y_i$ normal				
				Naive	Wald	Wald	Fieller			
1	0.1	0.1	2/3	95.0	(97.7, 97.3)	95.2	(97.8, 97.3)	95.1	(97.8, 97.3)	
2			1/2	94.7	(97.4, 97.3)	95.0	(97.7, 97.3)	94.8	(97.4, 97.4)	
3			1/3	94.9	(97.3, 97.6)	95.8	(98.2, 97.6)	95.0	(97.4, 97.7)	
4	0.5	0.1	2/3	90.4	(98.4, 92.1)	95.2	(98.0, 97.2)	95.2	(97.5, 97.7)	
5			1/2	83.6	(97.3, 86.3)	95.4	(98.9, 96.5)	95.0	(97.5, 97.5)	
6			1/3	73.0	(93.7, 79.2)	95.3	(99.9, 95.4)	95.2	(97.8, 97.4)	
7	0.5	0.1	2/3	95.0	(97.5, 97.5)	95.2	(97.7, 97.5)	95.1	(97.5, 97.6)	
8			1/2	95.0	(97.5, 97.5)	95.5	(98.0, 97.5)	95.1	(97.6, 97.6)	
9			1/3	94.4	(97.4, 97.1)	95.4	(98.3, 97.1)	95.1	(97.6, 97.5)	
10			1	2/3	63.3	(99.7, 63.7)	95.5	(98.6, 97.0)	95.3	(97.8, 97.6)
11				1/2	48.3	(99.2, 49.1)	94.9	(99.2, 95.7)	94.7	(97.5, 97.2)
12				1/3	40.0	(96.2, 43.8)	94.4	(100, 94.4)	94.9	(97.5, 97.4)

4.11 Conclusions

4.11.1 Effects of classical covariate measurement error

Unlike for linear regression, the bias caused by classical covariate measurement error in logistic regression cannot in general be quantified exactly (or at least not without making further distributional assumptions). However, under certain assumptions and conditions which may often hold in practice, the bias is approximately the same as for linear regression.

4.11.2 Regression calibration

Whereas RC gives consistent estimates of β_X for a linear regression outcome model, it only gives approximately consistent estimates for logistic regression. However, RC is widely established as method for logistic regression models because its bias is thought to be small in settings which are typical of many epidemiological studies. Our simulations show that in the simple case of a single covariate measured with error, the bias of RC only becomes appreciable in situations in which the the covariate effects are moderate in magnitude. With stronger covariate effects we would expect RC to show larger biases, and so in studies in which large effects of the covariates measured with error are expected, use of other correction methods may be advisable.

4.11.3 Maximum likelihood

Unlike the case of continuous normally distributed outcomes, the observed data likelihood function for the most commonly assumed model for binary outcomes is intractable, due to the presence of an integral over the unobserved X_i . In cases where the dimension of \mathbf{X}_i is small to moderate, quadrature methods, as implemented in commands such as NLMIXED in SAS, GLLAMM in Stata, or Mplus, can be used to maximize the resulting the likelihood. These procedures can however be slow, due to the large number of calculations required to approximate the integrals involved.

4.11.4 Ascent-based Monte-Carlo Expectation Maximization

We have shown in Section 4.5 how MCEM can be implemented to find the MLEs for a particular parametric model. Furthermore, we have applied a recent proposal for controlling the MCEM procedure. This is not available in statistical packages, and requires a moderate amount of programming effort. The approach is therefore unlikely to appeal to applied researchers unless it is implemented into software packages, for example as an R package. A further limitation is that it is also relatively

slow, and may be prohibitive for large datasets because of the requirement to store a large number of imputations for the X_i , an issue further compounded in the case of multivariate \mathbf{X}_i . However, we have implemented it in an arguably inefficient way using R. The multiple imputation of X_i involves many for loops, for which R is notoriously slow. The algorithm could thus undoubtedly be made more computationally efficient by optimizing the program code, perhaps by writing the imputation part in C++.

4.11.5 A Bayesian approach

One possible approach to dealing with the computational difficulties of finding parameter estimates for the models such as that which assumes marginal normality for X_i , is to take a Bayesian approach, as demonstrated by Kuha [82]. The main benefit of such an approach is the availability of software such as WinBUGS, which given the model and prior specification, performs the required sampling for the user, rather than any intrinsic computational benefit. However, this requires specification of priors for all model parameters, and involves the non-trivial step of determining when Markov chains have converged to their stationary distributions.

4.11.6 Maximum likelihood using linear mixed models

We have shown that our novel approach of fitting a linear mixed model for \mathbf{W}_i given Y_i can be extended to the case of logistic regression under an assumption that $X_i|Y_i$ is normal, i.e. the normal discriminant model. This approach is easily implemented using existing commands for fitting linear mixed models, and gives the MLE for the joint model which assumes conditional normality of X_i given Y_i . In simulations we have shown that with weak or moderate covariate effects this estimator has little bias even when X_i is marginally normal. In situations in which RC had non-negligible bias (moderate effect of X_i and moderately large measurement error), our approach had little bias, but like RC, is still easy to implement. However, we would not expect our approach to have less bias than RC in all possible scenarios. In particular, in situations where X_i is marginally normal, the measurement error variance is small (and so RC is expected to have little bias), but β_X is large, we would expect the mis-specified ML approach based on an assumption of normality for $X_i|Y_i$ to have greater bias. It is not therefore possible to claim that the ML approach based on normality for $X_i|Y_i$ is less biased than RC in all situations.

4.11.7 Multiple imputation

The main limitation of the approach based on fitting a linear mixed model for \mathbf{W}_i given Y_i is its extension to a general setup in which there may be error-free covariates, some of which may be discrete, or in any case not jointly normal with \mathbf{X}_i given Y_i .

In Section 4.7 we have described how having fitted a linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i , we can multiply impute \mathbf{X}_i , fit logistic regression models using the imputations, and average the resulting parameter estimates. The benefit of this approach is that we do not need to specify the marginal distribution of \mathbf{Z}_i .

4.11.8 Moment reconstruction

Moment reconstruction gives consistent estimates for logistic regression under an assumption of conditional normality for X_i given Y_i . We have shown how, for the setting of internal replication data, MR can be implemented by first fitting the same linear mixed model for \mathbf{W}_i given Y_i as used to obtain the MLEs for the conditional normal model. As discussed in Section 4.8, having fitting this linear mixed model, the MLE for β_X can be obtained directly from the fitted mixed model estimates, and so perhaps at least for the internal replication data setting, there is little to gain by using MR.

4.11.9 Conditional score

Semi-parametric methods such as the CS method offer the potential for consistent estimates without making distributional assumptions for the unobserved X_i . The price for this is a loss of efficiency relative to the correctly specified MLE. However, our simulation results show that when X_i is marginally or conditionally normal given Y_i , the efficiency of the CS method is often very similar, or at least not substantially larger than the efficiency of the MLE, at least when measurement error is small to moderate. The main barrier to greater use of the CS method is its lack of availability in statistical packages. Based on our simulation results, applying a simple Newton-Raphson procedure, using the RC estimates of β_0 and β_X as initial values, a consistent estimate of β_X is obtained. The exception to this was when β_X was moderate and the measurement error was large. In these cases the Newton-Raphson procedure often diverges from a root. To address this we used a more sophisticated non-linear equation solving algorithm, which backtracks when such divergence occurs. In our simulations this algorithm failed to find a root to the CS estimating equations in a very small proportion of simulations. However, even excluding these simulations, the CS estimates had large upward bias. We believe this is due to the algorithm finding a root of the CS estimating equations which is not consistent for β_X .

Chapter 5

Survival outcomes

In this chapter we examine the effects of classical covariate measurement error in regression models for censored survival outcomes. The most popular approach to modelling censored survival data is Cox’s proportional hazards model, in which no parametric form need be specified for the baseline hazard function. Because of its popularity, the majority of research into the effects of covariate measurement error in regression models for survival data has focused on Cox’s model. In Section 5.1 we introduce Cox’s proportional hazards model and review the partial likelihood estimator. In Section 5.2 we review the literature concerning the effects of classical covariate measurement error on parameter estimates in the Cox model. We then examine the conditions under which RC can be used in Cox regression (Section 5.3). The use of ML to allow for covariate measurement error in survival regression models has received much interest, particularly in the context of longitudinal error-prone measurements (see Chapter 9). In Section 5.4 we therefore describe the ML approach for the Cox model. In Section 5.5 we show how Monte-Carlo EM can be used to find the MLEs. This involves a novel proposal for using rejection sampling to sample from the relevant conditional distribution. Next, we show how recent results regarding the use of MI for missing data when the outcome follows a Cox model can be combined with our approach based on fitting a linear mixed model for \mathbf{W}_i conditional on the outcome (Section 5.6). We then briefly describe some of the semi-parametric (with respect to $f(X_i)$) approaches which have been proposed, which make no assumptions regarding the distribution of X_i (Section 5.7). In Section 5.8 we report the results of simulations, and conclude in Section 5.9 with some discussion of the various methods.

5.1 Cox's proportional hazards model

5.1.1 The hazard and survival functions

We let T_i denote a non-negative random variable which represents the survival time, or time to an event of interest for subject i . When considering the distribution of survival times, rather than considering density or distribution functions, it is often more convenient to instead consider the hazard function or rate:

$$h_i(t) = \lim_{dt \downarrow 0} \frac{P(t \leq T_i < t + dt | T_i \geq t)}{dt} = \lim_{dt \downarrow 0} \frac{P(t \leq T_i < t + dt)}{P(T_i \geq t)dt} \quad (5.1)$$

which can be interpreted as the instantaneous risk that subject i experiences the event at time t , conditional on being event free up to time t . The cumulative hazard function is then defined by:

$$H_i(t) = \int_0^t h_i(s) ds. \quad (5.2)$$

The survival function $S_i(t) = P(T_i > t)$, equals the probability that T_i exceeds t , and is therefore equal to one minus the distribution function of T_i . For convenience we recall the relationships between the density function $f_i(t)$, hazard function $h_i(t)$, and survival function $S_i(t)$:

$$h_i(t) = \frac{f_i(t)}{S_i(t)} \quad (5.3)$$

$$S_i(t) = \exp(-H_i(t)). \quad (5.4)$$

5.1.2 Censoring

Perhaps the key feature of survival data is the presence of censoring, which means that for some reason we are unable to observe a subject's survival time, but we know only that it exceeds a certain value. We observe $V_i = \min(T_i, C_i)$, where C_i denotes a potential censoring time for subject i , and we let $Y_i = 1(T_i \leq C_i)$ be an indicator variable showing whether the subject was censored ($Y_i = 0$) or was observed to have had the event ($Y_i = 1$). We assume throughout that censoring is 'independent' [83]. This means that the hazard rate observable in the population which is subject to censoring is equal to the hazard rate which would be observed in the absence of censoring:

$$h(t|X_i, C_i \geq t) = h(t|X_i). \quad (5.5)$$

In type I censoring, the censoring time is a non-random value $C_i = c_i$ (which may differ between subjects). Type II censoring occurs when subjects are followed up (and hence survival times observed) until a given number of events have been ob-

served. Both of these censoring types are independent. Lastly, in random censoring, C_i is a random variable. So long as T_i and C_i are independent, conditional on X_i , random censoring is also independent.

5.1.3 The Cox model

Usually interest focuses on relating the distribution of T_i to covariates. Due to the one-to-one correspondence between the hazard function and the density function of T_i , we can model the hazard function of subject i , conditional on covariate X_i , which we denote $h(t|X_i)$. A popular class of models are so called proportional hazards models, in which a subject's hazard function is equal to the product of a baseline hazard function $h_0(t)$ and some function of X_i . The most popular specification assumes that:

$$h(t|X_i) = h_0(t) \exp(\beta_X X_i). \quad (5.6)$$

In a landmark paper in 1972, Cox proposed a method of estimation for the parameter β_X which requires no assumptions regarding the baseline hazard function $h_0(t)$ [84]. Cox argued that if one was a priori ignorant regarding the form of the baseline hazard function $h_0(t)$, then $h_0(t)$ could conceivably be zero between survival times, and that therefore most of the information regarding β_X is contained in the ranking of the survival times, rather than the times themselves. Based on this argument, Cox proposed a likelihood function for β_X based on the ranking of the survival times, and thus which makes no assumptions regarding the form of the baseline hazard function $h_0(t)$. Cox later termed this function a 'partial likelihood function', to stress the fact that it is not a standard likelihood function [85]. For simplicity of exposition, we assume in the following that there are no tied event times, although the methods can be adapted to deal with ties. Assuming there are no tied events, the partial likelihood function corresponding to data from a sample of n independent subjects is given by:

$$\prod_{i=1}^n \left(\frac{\exp(\beta_X X_i)}{\sum_{j \in R_i} \exp(\beta_X X_j)} \right)^{Y_i} \quad (5.7)$$

where $R_i = \{j : V_j \geq V_i\}$ denotes the *risk set* at time V_i , that is all subjects who have not experienced the event and have not been censored just before time V_i . The log partial likelihood function is then given by:

$$\sum_{i=1}^n Y_i \left(\beta_X X_i - \log \left(\sum_{j \in R_i} \exp(\beta_X X_j) \right) \right), \quad (5.8)$$

and, differentiating with respect to β_X and setting the resulting expression to zero, the estimate of β_X is found as the value $\hat{\beta}_X$ which satisfies the partial log likelihood score equation:

$$\sum_{i=1}^n Y_i \left(X_i - \frac{\sum_{j \in R_i} X_j \exp(\hat{\beta}_X X_j)}{\sum_{j \in R_i} \exp(\hat{\beta}_X X_j)} \right) = 0 \quad (5.9)$$

Cox proposed that the partial likelihood be treated as if it were a standard likelihood function. In particular he suggested the estimator solving equation (5.9) is consistent and asymptotically normal, with variance which can be estimated in the usual way, i.e. by the inverse of minus the second derivative of the log partial likelihood function, evaluated at the estimated value of β_X . Subsequently these results were proved rigorously using the theory of counting processes.

5.2 The effects of classical covariate measurement error

We begin, as for linear and logistic regression, by considering the induced model for the outcome given an error-prone measurement W_i . Thus suppose that a single error-prone measurement W_i is available which is subject to classical measurement error. Prentice [86] stated that the hazard function conditional on W_i can in general be expressed as the expectation of $h(t|W_i, X_i)$ conditional on W_i and $T_i \geq t$, a result we now derive. Using the relationship between the hazard, probability density, and probability distribution functions:

$$\begin{aligned} \mathbb{E}(h(t|W_i, X_i)|T_i \geq t, W_i) &= \int h(t|W_i, X_i) f(X_i|T_i \geq t, W_i) dX_i \\ &= \int \frac{f(t|W_i, X_i)}{S(t|W_i, X_i)} f(X_i|T_i \geq t, W_i) dX_i \\ &= \int \frac{f(t|W_i, X_i)}{f(T_i \geq t|W_i, X_i)} \frac{f(X_i, T_i \geq t, W_i)}{f(T_i \geq t, W_i)} dX_i \\ &= \int \frac{f(t|W_i, X_i)}{f(T_i \geq t|W_i, X_i)} \frac{f(T_i \geq t|W_i, X_i) f(W_i, X_i)}{f(T_i \geq t, W_i)} dX_i \\ &= \int \frac{f(t|W_i, X_i) f(W_i, X_i)}{f(T_i \geq t, W_i)} dX_i \\ &= \int \frac{f(t|W_i, X_i) f(W_i, X_i)}{f(T_i \geq t|W_i) f(W_i)} dX_i \\ &= \frac{1}{S(t|W_i)} \int \frac{f(t, W_i, X_i)}{f(W_i)} dX_i \\ &= \frac{1}{S(t|W_i)} \int f(t, X_i|W_i) dX_i \\ &= \frac{f(t|W_i)}{S(t|W_i)} = h(t|W_i). \end{aligned} \quad (5.10)$$

In terms of the hazard function, the non-differential error assumptions means that

$$h(t|W_i, X_i) = h(t|X_i) \quad (5.11)$$

so that

$$h(t|W_i) = \mathbb{E}(h(t|X_i)|T_i \geq t, W_i). \quad (5.12)$$

Under the proportional hazards model, $h(t|X_i) = h_0(t) \exp(\beta_X X_i)$, so that

$$h(t|W_i) = h_0(t) \mathbb{E}(\exp(\beta_X X_i)|T_i \geq t, W_i). \quad (5.13)$$

Naively this may be seen as a proportional hazards model, with the baseline hazard function, $h_0(t)$ as in the original outcome model. However, the hazard ratio is the average of the original hazard ratio, averaged across the conditional distribution of the true covariate X_i , given that $T_i \geq t$ and the observed error-prone measurement W_i . The conditioning on $T_i \geq t$ means that in general the expectation in equation 5.13 depends on the baseline hazard function, $h_0(t)$ and β_X [86]. This is because $f(X_i|W_i, T_i \geq t)$ in general does not equal $f(X_i|W_i)$. If X_i is related to survival, the event $T_i \geq t$ provides information about X_i . For example, if large values of X_i are associated with increased survival, X_i is more likely to be large given $T_i \geq t$ and W_i . Of course we do not know β_X and $h_0(t)$, so the required expectation cannot usually be calculated directly. This means that, as for logistic regression, it is not possible to quantify the effects of classical covariate measurement error in Cox regression models exactly. However, as for logistic regression, we can make progress under particular additional assumptions.

When the survival times are subject to censoring, the observable hazard is $h(t|W_i, V_i \geq t)$, which by the result in equation (5.10) is equal to:

$$\begin{aligned} h(t|W_i, V_i \geq t) &= \mathbb{E}(h(t|W_i, X_i, V_i \geq t)|V_i \geq t, W_i) \\ &= \mathbb{E}(h(t|X_i)|T_i \geq t, C_i \geq t, W_i), \end{aligned}$$

provided that $h(t|W_i, X_i, V_i \geq t) = h(t|X_i)$. This means that, conditional on X_i , W_i and $V_i \geq t$ are jointly uninformative regarding the hazard at time t .

5.2.1 Method of moments type correction

In Section 5.3 we describe the conditions under which RC is justified for Cox proportional hazards models. In particular, RC gives approximately consistent estimates of β_X if $\beta_X \approx 0$, if $\text{Var}(X_i|W_i)$ is small (small measurement error), or if most subjects are censored, i.e. $P(Y_i = 0) \approx 1$. If one assumes that X_i and U_i are normally distributed and uncorrelated, then as we have seen previously,

$\mathbb{E}(X_i|W_i) = \mu_X + \lambda(W_i - \mu_X)$ where λ denotes the reliability ratio of the measurements W_i . Then because of the equivalence between RC and the method of moments type correction, dividing the naive estimate of β_X , $\hat{\beta}_W$, by an estimate of the reliability ratio λ gives an approximately consistent estimate of β_X under one or more of the conditions under which RC is justified.

In 1993 Hughes considered the validity of this correction method in detail [87]. Hughes first considered the situation in which no censoring occurs, i.e. all subjects are observed until failure. If W_i is subject to classical independent normal error, Hughes showed that the relationship between β_W and β_X does not depend on $h_0(t)$, the baseline hazard function, but the relationship does depend on the marginal distribution of the true covariate, $f(X_i)$. When X_i is normally distributed, Hughes investigated how β_W depends on β_X and λ , the reliability ratio of W_i as measurements of X_i . Hughes showed that for small β_X (after standardizing X_i), $\beta_W \approx \lambda\beta_X$, i.e. the bias caused by normally distributed classical error is the same (on the log hazard ratio scale) as for linear regression. However, for increasingly large values of β_X , the attenuation increases further, i.e. $|\beta_W| < |\lambda\beta_X|$. This means that unless β_X is small, dividing the naive estimate by λ results in estimates which are still attenuated towards the null.

These results were proved under the assumption of no censoring. Hughes also considered the relationship between β_W and β_X under two censoring mechanisms. Hughes first considered type I censoring, in which all subjects are censored at a fixed time. In this case, Hughes showed that parameter estimates are biased towards the null by an amount greater than the reliability ratio λ , but that the amount of bias depends on the proportion of subjects who are censored. As the proportion censored increases, the attenuation factor converges to λ . Hughes also considered a particular type of random censoring, in which subjects are entered into a study uniformly over time, and followed up until recruitment is complete. In this case, the relationship between β_W and β_X depends on the baseline hazard function, $h_0(t)$, although Hughes reported that the influence of the baseline hazard function is generally small. Furthermore, Hughes' results suggested that the attenuation factor was similar for both types of censoring considered.

5.2.2 An alternative expression for bias

Later, Kong derived an alternative approximate expression for the bias of the partial likelihood estimator which ignores classical covariate measurement error [88]. In contrast to Hughes, Kong did not make assumptions about the measurement error distribution or the type of censoring. The expression for bias can be easily calculated using quantities which are produced by statistical software when the naive Cox model is fitted, leading to an alternative, simple approximate approach to reducing the bias of the naive estimator $\hat{\beta}_W$.

We write τ for the maximum possible follow-up time for a subject, i.e. we preclude the possibility of arbitrarily long follow-up. The approximate expressions for bias are derived under the assumption that $n \rightarrow \infty$ and $\sigma_U^2 \rightarrow 0$. Despite this latter assumption, Kong claimed that the expression for bias often works reasonably well when the measurement error variance is ‘moderate’. Kong showed that β_W , the value which is consistently estimated by the naive estimator $\hat{\beta}_W$, is approximately equal to:

$$\beta_W = \beta_X \left(1 - \sigma_U^2 \Sigma \int_0^\tau \mathbb{E}(\exp(\beta_X X_i) | T_i \geq s) h_0(s) ds \right) \quad (5.14)$$

where Σ denotes the normalized limiting variance of the estimator $\hat{\beta}_X$ which would be obtained by maximizing the partial likelihood with X_i as covariate. Based on this approximation, Kong proposed that a bias-corrected estimate of β_X be obtained from the naive estimate $\hat{\beta}_W$ using:

$$\hat{\beta}_X = \hat{\beta}_W \left(1 + \hat{\sigma}_U^2 \widehat{\text{Var}}(\hat{\beta}_W) \bar{N}(\tau) \right) \quad (5.15)$$

where $\widehat{\text{Var}}(\hat{\beta}_W)$ denotes the usual estimate of the variance of $\hat{\beta}_W$ obtained from inverting the partial likelihood information matrix and $\bar{N}(\tau)$ is equal to the proportion of subjects who were observed to have the event of interest during follow-up. It is of interest to note that the expression and correction for bias does not depend on, or require an estimate of, the value of σ_X^2 .

Kong performed simulations under type II censoring, whereby all subjects are censored after a particular number of events have been observed. Kong assumed that the measurement error variance σ_U^2 is known. The results showed that for small σ_U^2 , the bias correction was successful in removing most of the bias from the naive estimator. For larger values of σ_U^2 however, the corrected estimator only removed some of the bias of the naive estimator. Kong’s correction for bias is appealing because it only requires quantities which are automatically produced by fitting the naive Cox regression model using W_i as covariate, in addition to an estimate of σ_U^2 .

5.3 Regression calibration

5.3.1 Simple regression calibration

Prentice’s investigation into the effects of covariate measurement error in Cox regression resulted in the first proposal of what we now refer to as RC [86]. Prentice considered the possible situations in which the induced hazard function (equation (5.13)) might be expected not to depend on $h_0(t)$ or β_X . Prentice suggested that when either $\beta_X \approx 0$, the measurement error is small in magnitude, or if the probability of being censored is close to one, then the event that $T_i \geq t$ provides little

information about X_i after conditioning on W_i , so that:

$$\begin{aligned} h(t|W_i) &= h_0(t)\mathbb{E}(\exp(\beta_X X_i)|T_i \geq t, W_i) \\ &\approx h_0(t)\mathbb{E}(\exp(\beta_X X_i)|W_i). \end{aligned} \tag{5.16}$$

In such cases, the expectation only involves the conditional distribution of X_i given W_i . If, for example, we assume that $f(X_i|W_i)$ is normal, it follows from the moment generating function of the normal distribution that

$$h(t|W_i) \approx h_0^*(t) \exp(\beta_X \mathbb{E}(X_i|W_i))$$

where a different baseline hazard function, $h_0^*(t)$, absorbs a factor of $\exp(0.5\beta_X^2 \text{Var}(X_i|W_i))$, assuming the latter conditional variance is constant. Given estimates of the relevant parameters, one can thus calculate $\hat{\mathbb{E}}(X_i|W_i)$ for each subject and maximize the partial likelihood function with this as covariate, in order to estimate β_X .

Wang *et al* considered the asymptotic properties of this simple regression calibration estimator [89]. They showed the regression calibration estimator is asymptotically normal, but that in general it is not a consistent estimator of β_X . Wang *et al* performed simulations in which internal validation data are available to estimate the regression calibration function $\mathbb{E}(X_i|W_i)$. They performed simulations using a uniformly distributed X_i , and both normal or uniformly distributed measurement errors U_i , but assumed that $\mathbb{E}(X_i|W_i) = \alpha_0 + \alpha_1 W_i$, which they estimated using OLS. For subjects in the validation sample, they used the observed X_i as covariate, and for those subjects not in the validation study, $\hat{\mathbb{E}}(X_i|W_i)$. Their simulation results suggested that RC performed well, giving estimates which had little bias. The bias in RC estimates in this setting obviously depends on how large the validation sample is. Wang *et al* also proposed a sandwich estimator of the variance of RC, which is robust to misspecification of the calibration function $\mathbb{E}(X_i|W_i)$, and which allows for the estimation of the parameters involved in $\mathbb{E}(X_i|W_i)$.

Subsequently, Wang considered the case in which internal replication data are available, and where the conditional expectation $\mathbb{E}(X_i|W_i)$ is approximated by the best linear approximation [90]. Wang proved that the simple RC estimator in this case is also asymptotically normal, and proposed a sandwich estimator for its variance which takes into account estimation of the parameters required to calculate the best linear prediction of X_i given W_i . Wang's simulations confirmed that the simple RC estimator may be substantially biased when the percentage of subjects who are censored is low.

5.3.2 Risk set calibration

Clayton proposed a modification to RC which does not require the assumption that the proportion of subjects whose survival times are censored is high [40]. To incorporate the conditioning information that $T_i \geq t$ in equation (5.13), Clayton proposed modelling the distribution of X_i in each risk set as normal, with a separate mean for each risk set but common variance σ_X^2 . This approach thus allows for the fact that if the hazard function depends on X_i , the distribution of X_i in survivors changes as time moves forwards, as subjects with large values of $\beta_X X_i$ tend to fail earlier than subjects with small values of $\beta_X X_i$.

Clayton considered the situation where a subset of subjects have $n_i = 2$ error-prone measurements of X_i , from which the usual moment based estimators of σ_X^2 and σ_U^2 can be used. Clayton then proposed that the mean of X_i in each risk set be estimated by the sample mean of \bar{W}_i for those subjects in the corresponding risk set. The conditional expectation $\mathbb{E}(X_i | \mathbf{W}_i, T_i \geq t)$ can then be calculated for each subject at risk at each risk set time t , using the risk-set specific mean of \bar{W}_i to take into account the information that $T_i \geq t$. The partial likelihood function can then be maximized, replacing the unobserved value of X_i for each subject who is at risk at each event time t by $\mathbb{E}(X_i | \mathbf{W}_i, T_i \geq t)$. Standard software can be used to maximize the partial likelihood function by noting that $\mathbb{E}(X_i | \mathbf{W}_i, T_i \geq t)$ can be treated as a time-dependent covariate, which changes value at each distinct event time.

More recently Xie *et al* considered a more flexible version of risk set RC by allowing for the variance of X_i to vary as time proceeds [91]. For the risk set at time t , Xie *et al* proposed using the best linear prediction of X_i based on \mathbf{W}_i , and using an estimate of the mean and variance of X_i given $T_i \geq t$. As we have discussed previously, the best linear prediction matches the conditional expectation of X_i given \mathbf{W}_i if X_i and \mathbf{W}_i are jointly normal. However, as Xie *et al* noted, even if they are jointly normal at $t = 0$, if X_i is related to survival, their conditional joint distribution will not remain normal as time moves forward. This means that even under marginal normality assumptions for the joint distribution of X_i and \mathbf{W}_i , the best linear prediction is only an approximation to the true conditional expectation given survival to time t .

The proposed method of Xie *et al* is relatively simple to implement. The measurement error variance σ_U^2 can be estimated in the usual way, using data from all n subjects. Xie *et al* proposed that the mean of X_i at each risk set be estimated by the unweighted mean of subjects' values of \bar{W}_i and suggested a simple moment estimator for the variance of X_i in each risk set, based on the variance of \bar{W}_i in the risk set and the estimate of σ_U^2 .

Xie *et al* proved that the resulting estimator is asymptotically normal and derived a sandwich estimator of variance. They also derived analytical expressions for its asymptotic bias. They used these results to illustrate the asymptotic bias of their

risk set RC estimator and compared it with the asymptotic bias of the simple RC estimator when X_i and U_{ij} are normally distributed, $n_i = 4$ for all subjects, the baseline hazard function is constant, and there is no censoring. Their results showed that when β_X is small, both the risk set RC estimator and simple RC show little bias. For larger β_X and larger values of σ_U^2 , the bias of both methods increases, although the risk set RC estimator is uniformly less biased than the simple RC estimator, as we would expect.

They also carried out simulations to examine the finite-sample performance of the two RC estimators in the presence of fixed time censoring. These results also showed that for small measurement error variances or small β_X , simple RC performs as well as risk set RC, and only for large σ_U^2 and β_X does risk set RC offer substantially less bias. In this latter situation, risk set RC was also more variable than simple RC. Xie *et al* also performed simulations in which X_i and/or U_{ij} were uniformly distributed. Their results suggested that both simple RC and risk set RC still perform reasonably well.

Xie *et al* also performed simulations in which censoring depended on X_i . In these simulations, simple RC performed badly, whereas risk set RC performed well. This is to be expected, because simple RC ignores the conditioning information that $V_i \geq t$ (which implies that $C_i \geq t$), which when censoring depends on X_i , is informative. Xie *et al* concluded that risk set RC is likely to perform adequately in situations which are typically found.

5.4 Maximum likelihood

5.4.1 Justifications for Cox's partial likelihood

Before describing the ML approach for dealing with covariate measurement error in the Cox model, it is first useful to consider an alternative justification for Cox's partial likelihood function for when the covariate X_i is observed, i.e. without covariate measurement error. To do this, we follow the developments outlined in Section 25.10 of [28] and Exercise 4.7 of [83].

Cox's partial likelihood function is usually heuristically justified by considering the probability of the observed ranking of failure times, given subjects' covariates, ignoring the actual observed event times. However, Cox's partial likelihood function can also be derived as a profile likelihood by considering the likelihood of the full data [92]. Using the relations between the probability density function, hazard function, and survival function, a subject who is observed to fail at time V_i contributes a factor $f(V_i|X_i) = h(V_i|X_i)S(V_i|X_i)$ to the likelihood function, while a subject who is censored at time V_i contributes a factor $S(V_i|X_i)$, where $S(V_i|X_i)$ denotes the probability of survival to time V_i given X_i . The likelihood function for data from n

independent subjects is then given by:

$$\prod_{i=1}^n (h_0(V_i) \exp(\beta_X X_i))^{Y_i} \exp(-H_0(V_i) \exp(\beta_X X_i)). \quad (5.17)$$

As we have previously described, the model is semi-parametric, due to the presence of the infinite-dimensional parameter $h_0(t)$. Just as in the case of parametric models, the principle of maximum likelihood can often be used to find consistent and efficient estimators in semi-parametric models (see Section 25.10 of [28]). However, for semi and non-parametric models, the function which we maximize has to be carefully defined. The likelihood function of equation (5.17) can be made arbitrarily large by letting $h_0(t)$ be zero except from close to the observed event times, where we let it become larger and larger.

To overcome this, we instead consider the modified likelihood function:

$$\prod_{i=1}^n (\Delta H_0(V_i) \exp(\beta_X X_i))^{Y_i} \exp(-H_0(V_i) \exp(\beta_X X_i)) \quad (5.18)$$

where $H_0(t)$ is a step-function, with jumps at the observed event times, which at time t we denote $\Delta H_0(t)$. Thus if $Y_i = 0$ (subject i is censored), $\Delta H_0(V_i) = 0$, whereas if $Y_i = 1$, the increment $\Delta H_0(V_i)$ is treated as a parameter to be estimated. We now show how Cox's (log) partial likelihood can be derived from this modified likelihood function. First, we take logarithms of equation (5.18), giving:

$$\sum_{i=1}^n Y_i \log(\Delta H_0(V_i)) + Y_i \beta_X X_i - \exp(\beta_X X_i) \sum_{j: V_j \leq V_i} \Delta H_0(V_j). \quad (5.19)$$

Denoting the observed failure times by $P_1 < P_2 < \dots$, we can re-express this as:

$$\sum_{P_j} \log(\Delta H_0(P_j)) + \sum_{i=1}^n Y_i \beta_X X_i - \sum_{P_j} S^{(0)}(\beta_X, P_j) \Delta H_0(P_j), \quad (5.20)$$

where $S^{(0)}(\beta_X, P_j) = \sum_{i=1}^n 1(V_i \geq P_j) \exp(\beta_X X_i)$.

Holding β_X fixed, partial differentiation of equation (5.20) with respect to $\Delta H_0(P_j)$ shows that the likelihood is maximized by choosing:

$$\Delta H_0(P_j) = \frac{1}{S^{(0)}(\beta_X, P_j)}. \quad (5.21)$$

This matches an estimator for the cumulative baseline hazard function proposed by Breslow [93]. Finally, we can find a profile log likelihood for β_X by inserting this

into equation (5.20), giving:

$$-\sum_{P_j} \log(S^{(0)}(\beta_X, P_j)) + \sum_{i=1}^n Y_i \beta_X X_i - \sum_{P_j} 1. \quad (5.22)$$

Ignoring $\sum_{P_j} 1$, which does not involve any model parameters, this is identical to Cox's log partial likelihood function, as given in equation (5.8). We have thus shown that Cox's partial likelihood can be derived as a profile likelihood function.

In parametric models where the number of parameters increases at the same rate as the sample size, maximum likelihood estimators are typically asymptotically biased (Chapter 10 of [20]). A legitimate question is therefore why Cox's estimator for β_X is consistent, despite the model containing as many parameters as unique event times. According to Johansen [92], the Cox model 'enjoys a property similar to S-ancillarity, which can also be used to argue that one should use Cox's partial likelihood to estimate β_X even though it is estimated at the same time as the infinitely many parameters $H_0(t)$ '.

5.4.2 Maximum likelihood with covariate measurement error

Although Prentice first hinted at taking a full likelihood approach to deal with covariate measurement error in a Cox proportional hazards model [86], only 15 years later did the proposal receive serious attention [94, 95]. As with continuous and binary outcomes, we first specify a joint model for the full data, including X_i . As usual, we assume that X_i is measured with classical error by n_i error-prone measurements $W_{ij} = X_i + U_{ij}$ where U_{ij} is independent of X_i , T_i , C_i , and other measurement errors U_{ik} . With a parametric density assumed for X_i , which we denote $f(X_i)$, the likelihood function can be formed as the product of the outcome model density, the measurement error model $f(\mathbf{W}_i|X_i)$ and the marginal density $f(X_i)$. The likelihood function corresponding to the observed data is given by:

$$\prod_{i=1}^n \int (\Delta H_0(V_i) \exp(\beta_X X_i))^{Y_i} \exp(-H_0(V_i) \exp(\beta_X X_i)) f(\mathbf{W}_i|X_i) f(X_i) dX_i. \quad (5.23)$$

Estimation

As is the case for a logistic regression outcome model (Section 4.4), the likelihood function in equation (5.23) cannot be expressed in closed form due to the intractable integral over X_i . Thus, as for binary outcomes, either deterministic or Monte-Carlo methods can be used to approximate the required integrals, and methods such as Newton-Raphson, or EM, can be used to maximize the likelihood. Wulfsohn and Tsiatis, in the more general setting of longitudinal error-prone measurements, pro-

posed using the EM algorithm to obtain the MLEs [94]. In the E-step, expectations of functions of X_i are required, conditional on the observed data, Y_i, V_i, \mathbf{W}_i , which can be approximated using Gaussian quadrature. In the M-step, the parameter values are updated. For μ_X , σ_X^2 , and σ_U^2 , the expressions for their updated estimates are the same as in the cases of continuous and binary outcomes. Wulfsohn and Tsiatis proposed using one-step of Newton-Raphson to update the estimate of β_X . The increment of the cumulative baseline hazard function at time t is then updated using:

$$\hat{\Delta}(H_0(t)) = \sum_{i:V_i=t} \frac{Y_i}{\sum_{j \in R_i} \mathbb{E}(\exp(\hat{\beta}_X X_j))}. \quad (5.24)$$

In simulations, Hu *et al* found that this ML approach gave estimates with little bias, in contrast to simple RC, which showed some bias [95].

Inference and asymptotic properties

Hu *et al* proposed using profile likelihood to estimate the asymptotic variance of $\hat{\beta}_X$, although recently Hsieh *et al* have claimed that this approach may underestimate the variance, and they therefore recommend using bootstrapping [96]. Within the more general framework which allows for longitudinal error-prone measurements (see Section 9.4), Zeng and Cai have recently shown that the MLE of the joint model parameters is consistent and asymptotically normal [97]. They also appear to have proved the validity of the profile likelihood approach proposed by Hu *et al* [95] (and earlier, in the longitudinal setting, Wulfsohn and Tsiatis [94]), apparently in disagreement with the claims of Hsieh *et al* [96].

Software

The NLMIXED command in SAS can be used to maximize the likelihood function. As far as we are aware, until recently, this was the only way the model could be fitted using a standard statistical package. For the more general model in which error-prone measurements are made longitudinally, Guo and Carlin described taking a Bayesian approach and using the WinBUGS software to fit the model, and also discussed using NLMIXED to fit the model [98]. Recently, the JM package for R has been released for fitting joint models of longitudinal error-prone measurements and a censored survival time. Since the classical error model is a special case of this more general framework, it should be possible to use this package to fit joint models in which covariates are measured by simple classical error.

5.4.3 Relaxing assumptions

Hu *et al* considered approaches to relax the normality assumption for X_i [95]. They considered both a fully non-parametric approach and a semi-parametric approach with respect to the distribution of X_i . In the non-parametric model, no assumptions are made regarding the density of X_i . In the semi-parametric approach, the density of X_i is assumed to belong to a large class of smooth densities, referred to as the semi-non-parametric (SNP) class in the econometrics literature. The flexibility of the class of densities is controlled by a tuning parameter, which is usually fixed rather than estimated. Hu *et al* described how a Fortran command NLMIX, developed previously by Davidian and Gallant, could be used to find MLEs for this semi-parametric model. We believe the SAS NLMIXED procedure could also be used to fit this model – the only modification required is to modify the normal density of X_i to the SNP density function given by Hu *et al*.

Estimation for the non-parametric model is more complicated, and Hu *et al* proposed using an EM algorithm, which necessitates additional programming. In simulations, Hu *et al* found that when X_i was normally distributed, the non-parametric approach was unbiased but more variable than the MLE based on assuming X_i is normal, as we would expect. Paradoxically, the estimate based on a semi-parametric model for $f(X_i)$ was less variable than the estimate based on assuming X_i is normal. In simulations in which X_i was simulated as the mixture of two normals, the misspecified MLE based on assuming normality for X_i had some bias, but much less than RC. The semi-parametric estimate had little bias, and was only slightly more variable than the misspecified MLE. The estimate based on making no assumptions for $f(X_i)$ had a small bias, and was also more variable than the other likelihood estimates.

5.5 Ascent-based Monte-Carlo Expectation Maximization

As for binary outcomes (see Section 4.4), quadrature methods are feasible for finding the MLEs of joint models in which X_i is scalar, or has low dimension. However, for higher dimensional \mathbf{X}_i quadrature methods become infeasible because the number of required evaluations of the integrand grows exponentially. As previously discussed, Monte-Carlo methods may be a feasible alternative in such settings. The use of Monte-Carlo methods to find the MLEs of joint models with survival outcomes and covariates measured with error is not new. For the more general setting in which error-prone measurements are made longitudinally over time (see Chapter 9), Henderson *et al* proposed using Monte-Carlo methods to approximate the integrals involved in the E-step of the EM algorithm [99]. Earlier, Wulfsohn and Tsiatis [94],

again in the setting of longitudinal error-prone measurements, had shown that the expectations required in the E-step could be expressed in terms of expectations with respect to the conditional density $f(X_i|\mathbf{W}_i)$, which, with normality assumptions for $f(X_i)$ and $f(\mathbf{W}_i|X_i)$, is also normal. Henderson *et al* therefore proposed sampling randomly from $f(X_i|\mathbf{W}_i)$ and to use these draws to approximate the expected complete data log likelihood. Henderson *et al* proposed using a fixed number of samples of $f(X_i|\mathbf{W}_i)$ throughout the EM procedure, stating that ‘the value of M can be tuned to balance numerical accuracy with computational speed’. As we have previously discussed, using a fixed number of Monte-Carlo samples/imputations is far from ideal, as initially a smaller number of draws is usually sufficient, when the increases in log-likelihood are largest, whereas as convergence is approached, a greater number are required to ensure parameter estimates are found which increase the likelihood function.

In this section, we show how ascent-based MCEM can be implemented for the joint model defined in Section 5.4. We first show, in Section 5.5.1, how standard rejection sampling can be used to sample from $f(X_i|Y_i, V_i, \mathbf{W}_i)$. As in the case of a binary outcome, this then means that the M-step consists of (with the exception of the baseline hazard function) maximizing the complete data likelihood separately for each imputation and averaging parameter estimates across imputations. This means that built-in commands for maximizing the components of the complete data log likelihood - specifically that for the Cox partial likelihood, can be used to update the estimate of β_X . However, it is not clear whether averaging parameter estimates across imputations is appropriate for parameters which represent the increments in the cumulative baseline hazard function. In Section 5.5.2 we describe the results of our investigations into this question, and describe an alternative approach that can be used to update the estimate of the baseline hazard function. We conclude the section by describing how the ascent-based scheme for controlling the number of imputations used in each iteration is used for the joint model under consideration.

5.5.1 Generating imputations using rejection sampling

To use MCEM we need to sample from $f(X_i|Y_i, V_i, \mathbf{W}_i)$. For the related problem of missing covariates in a Cox proportional hazards model, Herring and Ibrahim showed how adaptive rejection sampling [100] can be used to sample from the required conditional distribution [101]. We now show that, when the outcome model is a proportional hazards model, standard rejection sampling can in fact be used to sample from $f(X_i|Y_i, V_i, \mathbf{W}_i)$, assuming we can sample from $f(X_i|\mathbf{W}_i)$. As far as we are aware, this has not been previously proposed, either in the literature for covariate measurement error, or in the literature concerning missing covariates.

Analogous to the case of a binary outcome, under the non-differential error assumption, the conditional density $f(X_i|Y_i, V_i, \mathbf{W}_i)$ can be expressed as:

$$\begin{aligned}
f(X_i|V_i, Y_i, \mathbf{W}_i) &= \frac{f(X_i, V_i, Y_i, \mathbf{W}_i)}{f(V_i, Y_i, \mathbf{W}_i)} \\
&= \frac{f(V_i, Y_i|X_i, \mathbf{W}_i)f(X_i|\mathbf{W}_i)f(\mathbf{W}_i)}{f(V_i, Y_i|\mathbf{W}_i)f(\mathbf{W}_i)} \\
&= \frac{f(V_i, Y_i|X_i)f(X_i|\mathbf{W}_i)}{f(V_i, Y_i|\mathbf{W}_i)}. \tag{5.25}
\end{aligned}$$

This conditional distribution depends on $f(V_i, Y_i|X_i)$, which in turn depends on the baseline hazard function $h_0(t)$. With no restriction on the baseline hazard function $h_0(t)$, $f(V_i, Y_i|X_i)$ cannot belong to a particular parametric family of distributions. However, we now show that rejection sampling can be used to sample from the required density. This can be used either when a parametric form is assumed for the baseline hazard function, or when a non-parametric estimate of the hazard function is used. As in Section 4.5, we use $f(X_i|\mathbf{W}_i)$ as our candidate distribution for rejection sampling, since it is normal, with mean and variance which are readily calculated. We now consider separately the cases where $Y_i = 0$ (censored) and $Y_i = 1$ (event observed).

Censored subjects

We first consider the case when $Y_i = 0$, so that the i th subject has been censored at time V_i . Then using equation (5.25) we have that:

$$f(X_i|V_i, Y_i = 0, \mathbf{W}_i) = \frac{S(V_i|X_i)f(X_i|\mathbf{W}_i)}{S(V_i|\mathbf{W}_i)}$$

To use rejection sampling we must bound:

$$\begin{aligned}
\frac{f(X_i|V_i, Y_i = 0, \mathbf{W}_i)}{f(X_i|\mathbf{W}_i)} &= \frac{S(V_i|X_i)}{S(V_i|\mathbf{W}_i)} \\
&\leq \frac{1}{S(V_i|\mathbf{W}_i)}
\end{aligned}$$

since $S(V_i|X_i) \leq 1$. Rejection sampling then proceeds by simulating a value $x_i \sim f(X_i|\mathbf{W}_i)$ and $u \sim U(0, 1)$. Then we accept x_i if

$$\begin{aligned}
u &\leq S(V_i|\mathbf{W}_i) \frac{f(x_i|V_i, Y_i = 0, \mathbf{W}_i)}{f(x_i|\mathbf{W}_i)} \\
&= S(V_i|\mathbf{W}_i) \frac{S(V_i|x_i)}{S(V_i|\mathbf{W}_i)} \\
&= S(V_i|x_i).
\end{aligned}$$

Thus we accept a sample x_i if $u \leq S(V_i|x_i) = \exp(-H_0(V_i) \exp(\beta_X x_i))$.

Uncensored subjects

Now we consider the case when $Y_i = 1$, so that the event was observed to occur at time V_i for subject i . From equation (5.25) we have that:

$$f(X_i|V_i, Y_i = 1, \mathbf{W}_i) = \frac{f(V_i, Y_i = 1|X_i)f(X_i|\mathbf{W}_i)}{f(V_i, Y_i = 1|\mathbf{W}_i)}.$$

Then using the relationship between the hazard function and probability density function:

$$\begin{aligned} f(V_i, Y_i = 1|X_i) &= h(V_i|X_i) \exp(-H_0(V_i) \exp(\beta_X X_i)) \\ &= h_0(V_i) \exp(\beta_X X_i - H_0(V_i) \exp(\beta_X X_i)) \end{aligned} \quad (5.26)$$

Therefore we must bound:

$$\frac{f(X_i|V_i, Y_i = 1, \mathbf{W}_i)}{f(X_i|\mathbf{W}_i)} = \frac{h_0(V_i) \exp(\beta_X X_i - H_0(V_i) \exp(\beta_X X_i))}{f(V_i, Y_i = 1|\mathbf{W}_i)}.$$

Differentiating the numerator with respect to X_i we find that it takes its maximum when $\exp(\beta_X X_i) = 1/H_0(V_i)$, so that:

$$\frac{f(X_i|V_i, Y_i = 1, \mathbf{W}_i)}{f(X_i|\mathbf{W}_i)} \leq \frac{h_0(V_i) \exp(-1)}{H_0(V_i) f(V_i, Y_i = 1|\mathbf{W}_i)}.$$

Rejection sampling can thus be performed by simulating a value $x_i \sim f(X_i|\mathbf{W}_i)$ and $u \sim U(0, 1)$ and accepting x_i if

$$u \leq \frac{f(V_i, Y_i = 1|x_i)}{f(V_i, Y_i = 1|\mathbf{W}_i) \frac{h_0(V_i) \exp(-1)}{H_0(V_i) f(V_i, Y_i = 1|\mathbf{W}_i)}} \quad (5.27)$$

$$= \frac{H_0(V_i)}{h_0(V_i)} \exp(1) f(V_i, Y_i = 1|x_i), \quad (5.28)$$

with $f(V_i, Y_i = 1|x_i)$ as given in equation (5.26).

When the baseline hazard function is left unspecified, as previously described in Section 5.4, the likelihood function is maximized by allowing the cumulative baseline hazard function to be a step-function, with increases at each observed failure time V_i equal to $\Delta H_0(V_i)$, with the size of the increments to be estimated. To implement rejection sampling using the previously described expressions, we substitute $\Delta H_0(V_i)$ for $h_0(V_i)$ and $\sum_{j:V_j \leq V_i} \Delta H_0(V_j)$ for $H_0(V_i)$.

5.5.2 The M-step

In our implementation of the MCEM algorithm described in Section 4.5 we updated each of the model parameters at each iteration of EM by finding the MLE of the parameter using each imputation, and taking the mean of these values. If the

complete data MLE is linear in the data (as it is for the mean for example), this approach gives an identical estimate to that found by maximizing the likelihood corresponding to the ‘clustered’ dataset using all M imputations. For those MLEs (such as for the normal distribution variance) which are not exactly linear in the data, they are asymptotically linear [52], so that in large samples the two approaches are equivalent.

In the joint model we have described we treat the increments in the cumulative baseline hazard function as parameters. There is thus an increment to be estimated for each unique failure time. In implementing the MCEM algorithm, maximizing the complete data likelihoods separately for each imputation results in M (assuming M imputations are used at a particular iteration) estimates of β_X and M estimates of each increment in the cumulative baseline hazard function. We may then update the estimates of these increments by taking their mean across the M imputations. We have investigated this approach using simulations, but found that updating the estimate of the cumulative baseline hazard function in this way often did not result in an increase of the expected complete data log likelihood. We believe this may be because for the parameters corresponding to the increments in the cumulative baseline hazard function at later event times the number of subjects in the risk set is small when there is little censoring. In this case, there is likely to be a larger discrepancy between taking the mean of the estimates across imputations and that found using the expression derived by Wulfsohn and Tsiatis [94] (equation (5.24)). That is, the asymptotic justification for the equivalence between maximizing the likelihood corresponding to the ‘clustered dataset’ formed by combining the data from M imputations and maximizing the likelihoods separately for each imputation and averaging is not reasonable.

For the purposes of updating the estimate of the baseline hazard function, we therefore used equation (5.24), as given by Wulfsohn and Tsiatis, and replaced the expectation involved by its Monte-Carlo approximation. That is, assuming we have generated M imputations of X_i for $i = 1, \dots, n$, we updated the increment in the cumulative baseline hazard function at time t by:

$$\hat{\Delta}(H_0(t)) = \sum_{V_i=t} \frac{Y_i}{\sum_{j \in R_i} \left(\frac{\sum_{m=1}^M (\exp(\hat{\beta}_X X_j^{(m)}))}{M} \right)}, \quad (5.29)$$

where $\hat{\beta}_X = (1/M) \sum_{m=1}^M \hat{\beta}_X^{(m)}$ denotes the most recently updated estimate of β_X , based on the M imputations.

5.5.3 Ascent-based MCEM

To control how many imputations are required at each iteration of MCEM, we can again apply the ascent-based proposal of Caffo *et al* [74], as previously described in

Section 4.5.3. The only modification is to the complete data log likelihood component corresponding to the outcome model. Whereas for the binary model we use the log likelihood corresponding to a logistic regression model for Y_i given X_i , we now use the (log) of the Cox model modified-likelihood, as given in equation (5.18).

5.6 Multiple imputation

5.6.1 Multiple imputation for missing data

For the related problem of missing data, MI has been used to impute missing covariate values when the outcome model of interest is Cox's proportional hazards model. In a paper concerning the imputation of missing blood pressure values in a survival analysis, van Burren *et al* proposed using $\log(V_i)$ and Y_i as covariates in the imputation model for the missing values [102]. Including the logarithm of the survival/censoring time and censoring indicator in the imputation model for the missing covariate(s) has subsequently become the accepted approach when imputing missing covariates when the outcome model is Cox regression [103].

As discussed in Section 5.5.1, the conditional distribution of X_i given V_i , Y_i and \mathbf{W}_i does not belong to a particular parametric class of distributions. This follows from the fact that this conditional distribution depends on the baseline hazard function $h_0(t)$, about which we make no parametric assumptions in Cox's proportional hazards model.

Recently White and Royston have investigated the problem of how best to incorporate a censored survival time, which is assumed to follow a Cox model given covariates, in an imputation model [103]. White and Royston showed that, using a Taylor series approximation, if $X_i \sim N(\mu_X, \sigma_X^2)$, then $X_i|V_i, Y_i$ is approximately normally distributed, with mean a linear function of Y_i and $H_0(V_i)$, the cumulative baseline hazard function at time V_i . The approximation is valid providing $\text{Var}(\beta_X X_i)$ is small and/or $H_0(V_i) \exp(\beta_X \bar{X}_i)$ (approximately the marginal cumulative hazard at time V_i) is small.

By using Cox's proportional hazards model we do not assume a particular parametric form for $h_0(t)$, and hence also do not assume a particular form for $H_0(t)$. If we are willing to assume constant hazards for example, we can enter V_i in the linear predictor of the imputation model for X_i . Alternatively, White and Royston propose using a marginal (i.e. not conditional on covariates) estimator of the cumulative hazard function $H(t)$, such as the Nelson-Aalen estimator, as covariate in the imputation model for X_i . So long as the covariate effects on survival are small, the cumulative baseline hazard function and marginal cumulative hazard function will be similar.

5.6.2 Multiple imputation for measurement error

Recently Cole and Greenland proposed using MI to allow for misclassification in a binary covariate in a Cox model [54]. Cole and Greenland assumed the presence of a validation dataset in which the true covariate is measured, so that established MI commands in statistical software packages could be used. They assumed an imputation model for X_i with the censoring indicator and logarithm of the survival/censoring time as covariates, following the earlier proposals in the missing data literature. In simulations, Cole and Greenland reported that MI gave almost unbiased estimates in a number of scenarios, although the recent results of White and Royston suggest that this method will in general give biased estimates.

We now consider how the results of White and Royston can be adapted to use MI to deal with classical measurement error when the outcome model is Cox regression and we have internal replication data. As in Sections 3.7 and 4.7, we can fit a linear mixed model to the error-prone measurements \mathbf{W}_i . To implement White and Royston's proposal, we include Y_i and the Nelson-Aalen estimator of $H(V_i)$ as fixed effects. We explore the performance of this method using simulations, which are reported in Section 5.8.

5.6.3 A spline-based imputation approach

Li and Ryan recently proposed an imputation based approach, whereby to impute the unobserved X_i a linear spline model is assumed for the baseline hazard function [104]. Their motivation is that approaches which make no assumptions about the form of the baseline hazard function may have low efficiency, whereas specifying simple parametric forms for the baseline hazard function will result in bias when this specification is incorrect. For the purposes of imputing the unobserved X_i , Li and Ryan assumed a piece-wise linear spline model for the logarithm of the baseline hazard function. The estimating equations which are solved are equal to the expectation of those that would be used if X_i were observed, taken with respect to $f(X_i|V_i, Y_i, \mathbf{W}_i)$. Since these expectations are not available in closed form, Li and Ryan described numerical approaches to approximating them.

Li and Ryan's proposed method appears to be very similar to the Monte-Carlo EM algorithm, and rather than making no assumptions regarding $h_0(t)$, they assume a linear spline model. Li and Ryan also claim that the method is expected to be relatively robust even when the linear spline model for the baseline hazard is invalid because the complete data partial likelihood score equation for β_X does not depend on it. Li and Ryan claimed that their algorithm is easy to implement, but arguably the method is at least as complicated to implement as Monte-Carlo EM for the model described in Section 5.4, if not more so because of the need to choose the number and position of knots for the linear spline model.

5.7 Conditional and corrected score methods

5.7.1 Conditional score method

The conditional score (CS) approach, original devised in the context of a generalised linear outcome model (see Section 4.9), was adapted for use with Cox's proportional hazard model by Tsiatis and Davidian [105]. For Cox regression, the CS method involves the sufficient statistic $\tilde{X}_i(t) = W_i + \sigma_U^2 \beta_X dN_i(t)$ where $dN_i(t)$ is the counting process increment for subject i at time t which equals one if subject i experiences the event at time t and is zero otherwise. Song *et al* [106] showed that the CS estimating equation for Cox regression can be expressed as:

$$\sum_{i=1}^n Y_i \left(W_i + \sigma_U^2 \beta_X - \frac{\sum_{j \in R_i} (W_j + \sigma_U^2 \beta_X dN_j(V_i)) \exp(\beta_X (W_j + \sigma_U^2 \beta_X dN_j(V_i)))}{\sum_{j \in R_i} \exp(\beta_X (W_j + \sigma_U^2 \beta_X dN_j(V_i)))} \right) = 0 \quad (5.30)$$

When $\sigma_U^2 = 0$ the CS estimating equation reduces to the usual partial likelihood score equation (equation (5.9)). Tsiatis and Davidian gave a heuristic proof that there exists a consistent solution to the CS estimating equation, which then by the theory of estimating equations is asymptotically normal. As for logistic regression, with multiple error-prone measurements \mathbf{W}_i , we can substitute \bar{W}_i in place of W_i and σ_U^2/n_i in place of σ_U^2 in equation (5.30).

5.7.2 Corrected score method

The corrected score method was introduced by Nakamura as a general approach to dealing with bias caused by covariate measurement error [107, 108]. The idea of Nakamura's corrected score method is to adjust the naive score function based on using W_i as covariate so that the estimating equations have expectation zero at the true value of β_X . In the absence of covariate measurement error the outcome model parameter β_X would be estimated as the value which solves the likelihood score equation. When X_i is measured with classical error by W_i , as we have seen, solving the score equation with W_i in place of X_i results in asymptotically biased estimates of β_X . This means that the score equations have non-zero mean at β_X . Nakamura proposed a modified estimating equation which can be shown to have expectation zero at the true value β_X . Although Nakamura claimed that his corrected score method was only asymptotically unbiased if $\sigma_U^2 \rightarrow 0$ as $n \rightarrow \infty$, Kong later showed that it is in fact consistent for fixed σ_U^2 [109]. The corrected score estimating equation

for β_X is given by:

$$\sum_{i=1}^n Y_i \left(W_i + \sigma_U^2 \beta_X - \frac{\sum_{j \in R_i} W_j \exp(\beta_X W_j)}{\sum_{j \in R_i} \exp(\beta_X W_j)} \right) = 0. \quad (5.31)$$

5.7.3 Conditional versus corrected score

Neither the conditional nor the corrected score methods make any assumption about the distribution of X_i . However, they both assume normality for the measurement errors U_i (or, in the case of multiple error-prone measurements, \bar{W}_i as a measurement of X_i). The two methods are asymptotically equivalent [105], but in small samples and relatively large measurement errors, Song *et al* found that the CS method was considerably less variable [106]. Unfortunately, it is not possible to solve either estimating equation by fitting a standard Cox model, and so they must be solved by methods such as Newton-Raphson. Furthermore, as for the CS method for logistic regression, the estimating equations may have multiple solutions.

5.8 Simulations

In this section we report the results of simulations to compare the methods we have described in this chapter.

5.8.1 Simulation setup

The covariate X_i was simulated as $X_i \sim N(0,1)$. The survival time of each subject, T_i , was then generated according to an exponential distribution (i.e. constant baseline hazard), with constant hazard equal to $\exp(\beta_X X_i)$, where $\beta_X = 0.1$ or 0.9 , representing weak and moderately strong covariate effects. The survival times were then censored according to a type II censoring scheme, by censoring either the largest 10% or the largest 90% of survival times. As before, we simulated data for $n = 5,000$ subjects, 500 of which had two measurements of X_i subject to normally distributed error, while the remaining 4,500 had a single error-prone measurement of X_i . We varied σ_U^2 to give values of the reliability ratio of $2/3$, $1/2$, and $1/3$. We again performed 10,000 simulations per scenario.

5.8.2 Estimation methods

Regression calibration

We first estimated β_X using simple RC, i.e. not risk-set calibration, in the same way as for linear and logistic regression outcome models.

Risk-set regression calibration

We used Xie *et al*'s proposed implementation of risk-set regression calibration. Specifically, we estimated μ_X separately in each risk-set by the mean of \bar{W}_i . We estimated σ_X^2 in each risk-set using the moment estimator proposed by Xie *et al*. For σ_U^2 we used the estimate obtained from fitting the random-intercepts model to the error-prone measurements \mathbf{W}_i of all subjects. As described in Section 5.3.2, risk-set RC can be implemented using standard Cox regression commands, by splitting the data at each event time, calculating the predicted value of X_i for each subject in each risk-set, and using this as a time-dependent covariate. Unfortunately, we were unable to do this because the expanded dataset was too large for the available memory on our system, illustrating the additional computational burden of the method (in terms of systems resources). We therefore found the risk-set RC estimate of β_X by using the `nleqslv` function, passing to it functions that manually calculate the likelihood and score functions (without expanding, or splitting the dataset at each event time).

Maximum likelihood via ascent-based Monte-Carlo Expectation Maximization

We used ML to estimate β_X using ascent-based MCEM, as described in Section 5.5. As for logistic regression, the control parameters for ascent-based MCEM were set to $\alpha = 0.25$, $\beta = 0.25$ and $\gamma = 0.05$, and convergence was declared either when the number of imputations reached 1,000 or if the upper confidence interval limit for the increase in the Q function was less than 0.01. Since no valid analytical method is available for estimation of standard errors, we do not report standard errors.

Multiple imputation

We also used frequentist MI to estimate β_X , as described in Section 5.6.2. We report the results from two MI estimators. The first is based on including V_i and Y_i as fixed effects in the linear mixed model for \mathbf{W}_i . As the survival times were generated according to an exponential distribution, $H_0(t) = kt$ for some constant k , this MI estimator avoids use of the Nelson-Aalen estimator as an approximation to the baseline cumulative hazard function. This first MI estimator thus allows us to examine how much bias is induced by the Taylor series approximation used to derive the results of White and Royston [103], as opposed to any bias induced by using the Nelson-Aalen estimator as an approximate estimator of the baseline cumulative hazard function. This second MI estimator is that which would be used in practice, when we wish to make no assumption regarding the form of the baseline hazard function. For both MI estimators we used 25 imputations.

Conditional score

As for logistic regression, we fixed σ_U^2 at its estimated value from fitting the random-intercepts model to \mathbf{W}_i . We then solved the CS estimating equation (as expressed by Song *et al* [106]), by using the `nleqslv` function in R, using the RC estimate of β_X as the initial value. As for the logistic regression simulations, for simulations where the estimating equation could not be solved, we substituted the RC estimate of β_X for the purposes of the table of results (Table 5.1). We have not included the corrected score method in our simulations, since it is asymptotically equivalent to the CS method and Song *et al* have demonstrated that its performance in finite samples is inferior to the CS method [106].

5.8.3 Simulation results

Table 5.1 shows the results of the simulations. RC had little bias when $\beta_X = 0.1$. For $\beta_X = 1$ and 90% of survival times censored, RC showed a small downward bias, with the amount of bias increasing with increasing measurement error. Conversely, when only 10% of survival times were censored, estimates from RC were downwardly biased by a substantial amount. The results for RC are similar to those for logistic regression, which perhaps is not surprising given the close connection between logistic and Cox regression [110]. The risk-set RC estimator performed well, with less bias but increased variability compared to standard RC. For scenarios 10-12, although the risk-set RC estimator had some bias, the biases were considerably smaller than those for standard RC.

In preliminary simulations we found that for the scenarios in which the reliability ratio was 1/3, the ascent-based MCEM estimates of β_X had some bias. Our hypothesis is that the observed data likelihood function is not concave, and that in these cases the initial estimates (from RC) were farther from the MLE, leading the EM algorithm to converge to a local maximum. To overcome these difficulties we found that a small modification to the MCEM algorithm could be used to give estimates which had little bias across all scenarios. We first used MCEM in which all model parameters, except β_X and the cumulative baseline hazard function increments, were held fixed at the values estimated by fitting the random-intercepts mixed model to \mathbf{W}_i . This algorithm is ML for the model in which it is assumed that μ_X , σ_X^2 and σ_U^2 are known. Since we substitute consistent estimates of these parameters in place of the known values this algorithm should itself give consistent estimates of β_X , although it may not be optimal in terms of efficiency. Such a two-stage likelihood approach was proposed in the context of a logistic regression outcome model by Carroll *et al* [111]. When this pseudo-ML MCEM algorithm was deemed to have converged, we then re-started the full MCEM algorithm, whereby

Table 5.1: Cox regression simulation results. Mean (SD) of estimates of β_X from regression calibration (RC), risk-set regression calibration (RRC), maximum likelihood via ascent-based Monte-Carlo Expectation Maximization (ML-MCEM), multiple imputation (MI) using censoring indicator and time to event or censoring, MI using censoring indicator and estimated cumulative hazard function, and the conditional score (CS) method. λ denotes the ratio of variance of X_i to variance of error-prone measurements

Scenario	$P(Y_i = 1)$	β_X	λ	RC	RRC	ML-MCEM	MI using Y_i and V_i	MI using Y_i and $\hat{H}(V_i)$	CS
1	0.1	0.1	2/3	0.100 (0.055)	0.100 (0.055)	0.100 (0.055)	0.100 (0.055)	0.100 (0.055)	0.100 (0.055)
2	0.1	0.1	1/2	0.101 (0.063)	0.101 (0.063)	0.102 (0.063)	0.102 (0.063)	0.101 (0.063)	0.101 (0.063)
3	0.1	0.1	1/3	0.100 (0.078)	0.100 (0.078)	0.100 (0.078)	0.100 (0.079)	0.100 (0.079)	0.100 (0.079)
4	0.1	1	2/3	0.973 (0.063)	0.985 (0.066)	1.005 (0.071)	0.970 (0.061)	0.969 (0.061)	1.004 (0.077)
5	0.1	1	1/2	0.963 (0.086)	0.983 (0.091)	1.009 (0.100)	0.965 (0.084)	0.965 (0.084)	1.010 (0.133)
6	0.1	1	1/3	0.961 (0.135)	0.989 (0.147)	1.022 (0.162)	0.969 (0.133)	0.969 (0.133)	1.109 (0.949)
7	0.9	0.1	2/3	0.100 (0.019)	0.100 (0.019)	0.100 (0.019)	0.100 (0.019)	0.100 (0.019)	0.100 (0.019)
8	0.9	0.1	1/2	0.100 (0.022)	0.101 (0.022)	0.101 (0.022)	0.100 (0.022)	0.100 (0.022)	0.101 (0.022)
9	0.9	0.1	1/3	0.101 (0.028)	0.102 (0.029)	0.102 (0.029)	0.101 (0.029)	0.101 (0.029)	0.102 (0.029)
10	0.9	1	2/3	0.842 (0.033)	0.951 (0.049)	1.003 (0.051)	0.828 (0.033)	0.861 (0.038)	1.001 (0.067)
11	0.9	1	1/2	0.787 (0.052)	0.943 (0.086)	1.005 (0.085)	0.794 (0.058)	0.840 (0.067)	1.008 (0.144)
12	0.9	1	1/3	0.747 (0.092)	0.951 (0.166)	1.012 (0.156)	0.782 (0.112)	0.838 (0.128)	1.093 (0.773)

all model parameters are updated at each iteration. As shown in Table 5.1, the resulting algorithm had little bias across all scenarios.

The two MI methods had bias and SD which were practically identical for scenarios 1-9. Interestingly, for scenarios 10-12, MI using the Nelson-Aalen estimator $\hat{H}(V_i)$ in the imputation model had less bias than MI using V_i in the imputation model. White and Royston's results suggest that $H_0(t)$ should be used in the imputation model [103]. Since the survival times were generated according to an exponential distribution, using V_i in the imputation model meant that the cumulative baseline hazard function did not need to be approximated, and so we would expect this approach to give less bias. One possible explanation for the greater bias found for MI using V_i in the imputation model is that because the survival times were exponentially distributed, when there was only 10% censoring, a small number of subjects' values of V_i have large influence over the estimated coefficient for the association between V_i and X_i .

A solution to the CS estimating equation was found for all simulations in all scenarios, with the exception of 6 simulations in scenario 6 and 2 simulations in scenario 12. The CS method showed little bias, except for scenarios 6 and 12, in which $\beta_X = 1$ and $\lambda = 1/3$ and for which the method was biased upwards. We presume this bias, analogous to the logistic regression simulations, is due to solutions of the CS estimating equation being found which are not the consistent solution. It is of interest to note that for scenarios 1-3 and 6-9, in which $\beta_X = 0.1$, the CS estimator had the same efficiency as ML. For scenarios 4-5 and 10-11, in which $\beta_X = 1$, the CS estimator showed little bias, but was less efficient than ML.

5.9 Conclusions

5.9.1 Effects of classical covariate measurement error

If the proportion of censored subjects is high and the covariate effects are small to moderate in magnitude, the effect of classical error on estimates from Cox's proportional hazards model is to bias effect estimates towards the null (in the case of a single covariate) by approximately the same amount as in linear regression outcome models, i.e. by the reliability ratio of the error-prone measurements. However, as we described in Section 5.2, when these conditions do not hold, the bias can often deviate substantially from that implied by the reliability ratio.

5.9.2 Regression calibration

Our simulation results confirm that RC provides approximately unbiased estimates when β_X is small. For larger effects, RC has non-trivial biases, which is larger still when a smaller proportion of subjects are censored. Risk-set RC, whereby predic-

tions of X_i are updated at each risk-set, substantially reduces the bias of standard RC. Risk-set RC can also be implemented using standard commands for fitting Cox proportional hazards models, although in large datasets one may encounter difficulties with regards to the memory resources required when ‘splitting’ or expanding the dataset at each event time.

5.9.3 Maximum likelihood

Given correct parametric specifications, ML can be used to give consistent estimates of the Cox model parameters. The NLMIXED command in SAS can in principle be used to fit such models, requiring the user to write a function to evaluate the complete data likelihood function. Recently however, the JM package for R has been released, which fits joint models for longitudinal and survival data (see Chapter 9) via ML, using quadrature methods to approximate the required integrals. The classical measurement error model is a special case of the longitudinal models which these packages accommodate, and so it should be possible to use these to find ML estimates for Cox models which are subject to classical covariate error. We did attempt to use the JM package to find ML estimates for the simulations reported here. Unfortunately the command gave an error message reporting that the available computer system memory was insufficient.

5.9.4 Ascent-based Monte-Carlo Expectation Maximization

We have shown in Section 5.5 that MCEM can be used to obtain ML estimates, in which rejection sampling is used to multiply impute X_i at each iteration of EM. As discussed in the conclusions to Chapter 4, the implementation of ascent-based MCEM requires a moderate amount of programming. Our implementation in R is quite slow, although as previously noted, it could be made more efficient by writing the function which creates multiple imputations using rejection sampling using a language such as C++, which the R program could then call. For univariate X_i use of quadrature methods to find ML estimates may well be computationally faster. However, for multivariate \mathbf{X}_i , as previously discussed, quadrature methods rapidly become infeasible, whereas Monte-Carlo approximation may still be feasible.

5.9.5 Multiple imputation

In Section 5.6 we have applied recent results regarding imputing missing covariates in Cox models to deal with covariate measurement error. Our proposed implementation is easy to use, and involves adaption of our linear mixed model approach (see Sections 3.6 and 4.6) to deal with the fact that the survival time distribution is not specified parametrically and that survival times may be censored. Somewhat disappointingly, in simulations the two implementations of MI that we considered performed very

similarly to RC. In particular, the estimates using MI had moderately large biases when there was little censoring and the covariate effects were moderately strong. However, given that RC and this particular implementation of MI are justified for the Cox model under the same assumptions (small covariate effects or low risk of events), that both methods performed badly in these scenarios is perhaps unsurprising.

5.9.6 Conditional score method

Our simulation results suggest the CS method will perform well in a wide range of scenarios. Only when the effect of X_i was moderate and the measurement error was large did the CS method show some bias, which we presume to be due to solutions of the CS estimating equation being found which are not the consistent root. Despite making no assumption about the distribution of X_i , the CS estimator also had comparable efficiency to ML, except when the effect of X_i and the measurement error were large. The computational burden of the CS method is far less than for ML, and so it should be considered more often as an alternative to simpler methods such as RC, especially when few subjects are censored and the effect sizes are moderate. Its implementation is relatively straightforward, and can make use of built-in commands for solving non-linear equations, such as the `nleqslv` package in R.

5.9.7 Alternatives to Cox's proportional hazards model

As discussed at the beginning of the chapter, the vast majority of research into the effects of classical measurement error has focused on Cox's proportional hazards model, but some research has been conducted for other types of survival time regression models. Classical covariate measurement error is easily accommodated in Aalen's non-parametric additive model, where the effects of classical error are the same as for linear regression [112]. Gimenez *et al* investigated the effects of classical error in Weibull regression models (a parametric proportional hazards model) [113]. They showed that with X_i and U_i both normally distributed and with no censoring, the naive estimator $\hat{\beta}_W$ is biased by the reliability ratio λ , as for linear regression. Recently He *et al* have investigated the consequences of covariate error in general accelerated failure time models, and explored the use of SIMEX to allow for the resulting bias [114].

Part II

Extensions to life-course studies

Chapter 6

Background

In Chapters 2 to 5 we have examined the consequences of classical covariate measurement error in models where the outcome is continuous, binary or is a censored survival outcome. In many applications the simple classical measurement error model may not be appropriate however. As discussed in Chapter 1, one example which has received a large amount of interest in the last 20 years is when error-prone measurements are made longitudinally over time, and interest lies in the associations between trajectories of one or more underlying longitudinal processes and an outcome of interest. This is sometimes rather unspecifically referred to as ‘joint modelling’. Underlying these approaches is an assumption that for each subject, there exists an unobserved (latent) vector of random effects. These random effects are assumed to influence both the longitudinal measurements, and the outcome of interest, and are defined such that conditional on them, the longitudinal measurements and outcome are independent.

In this chapter we review the linear mixed model family and describe how linear mixed models can be used to model error-prone measurements of one or more longitudinal processes. The random effects (or possibly a function of them) of the linear mixed model for the longitudinal error-prone measurements are then treated as covariates in a regression model for the outcome of interest. In Section 6.2 we discuss the specification of the outcome model. In particular, we note that in applications, while we may have some confidence in our specification of a model for the longitudinal error-prone measurements, we may be uncertain as to which aspects of the longitudinal process influence the outcome. In consequence, analyses will often be somewhat exploratory, and we may wish to fit a number of different models relating the outcome to different aspects of the longitudinal process. We conclude in Section 6.3 by discussing assumptions concerning missingness of longitudinal error-prone measurements.

6.1 Linear mixed models for longitudinal error-prone measurements

Linear mixed models represent an extremely flexible family of models which allow data with complex dependency structures to be modelled. An extensive account is given by Verbeke and Molenberghs [37]. One main application of linear mixed models is to longitudinal settings in which multiple observations are made on subjects over time. A linear mixed model is specified by a sum of functions of ‘fixed’ and ‘random’ effects. For longitudinal data, the fixed effects structure models the mean evolution of the longitudinal process in the population. We then suppose that there exist one or more, unobserved (latent) subject-specific random effects, which determine how a subject’s underlying longitudinal trajectory deviates from the population evolution. Within the structural equation literature, such models are sometimes referred to as latent-variable models for longitudinal data [115].

For continuity with the preceding chapters, we adopt notation similar to that used for classical measurement error. This helps make clear that, as we show shortly, the latter can be viewed as a special case of this more general model framework. As before, we denote by $\mathbf{W}_i = (W_{i1}, \dots, W_{in_i})^T$ the vector of n_i error-prone measurements for subject i . We then assume that:

$$\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i \tag{6.1}$$

where \mathbf{D}_i is a known $n_i \times p$ design matrix, \mathbf{X}_i is a $p \times 1$ vector of random effects for the i th subject with mean $\boldsymbol{\mu}_X$ and variance covariance matrix $\boldsymbol{\Sigma}_X$, and $\mathbf{U}_i = (U_{i1}, \dots, U_{in_i})^T$ is a $n_i \times 1$ vector of within-subject ‘errors’. We assume the ‘errors’ \mathbf{U}_i are independent of \mathbf{X}_i and that $\mathbb{E}(\mathbf{U}_i) = \mathbf{0}$.

Linear mixed models are usually parametrized such that the random effects have mean zero, so that they represent subject deviations from the mean evolution. We do not follow this convention, in order to maintain consistency with our notation in the classical measurement error setting. It is of course simple to re-express equation (6.1) in terms of a mean evolution and a mean zero random-effects vector by:

$$\mathbf{W}_i = \mathbf{D}_i \boldsymbol{\mu}_X + \mathbf{D}_i (\mathbf{X}_i - \boldsymbol{\mu}_X) + \mathbf{U}_i, \tag{6.2}$$

where $\mathbf{X}_i - \boldsymbol{\mu}_X$ is a mean zero random effects vector with variance covariance matrix $\boldsymbol{\Sigma}_X$, and which is equal to the deviation of \mathbf{X}_i from the population mean.

Depending on the estimation approach, we may make distributional assumptions for \mathbf{X}_i and \mathbf{U}_i . A common assumption is that the random effects are multivariate normal, $\mathbf{X}_i \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$, with $\boldsymbol{\Sigma}_X$ an unstructured variance covariance matrix. Alternatively, structured covariance matrices can be used for \mathbf{X}_i . A common assumption is that the errors are $\mathbf{U}_i \sim N(0, \text{diag}(\sigma_U^2))$, indicating that errors U_{ij}, U_{ik}

are independent for $j \neq k$. With such normality assumptions, the parameters of the linear mixed model can be estimated by ML or REML. Most modern statistical packages provide facilities for fitting linear mixed models, such as the SAS command PROC MIXED, Stata's xtmixed command, and the lmer command in R.

6.1.1 Examples

Classical measurement error – single covariate

If X_i is a scalar, and \mathbf{D}_i is a column vector of 1's, with length n_i , we see that the linear mixed model of equation (6.1) is identical to the classical measurement error model for a single covariate.

Classical measurement error – multiple covariates

The linear mixed model of equation (6.1) has exactly the same form as the classical measurement error model in the case of multiple covariates \mathbf{X}_i , as described in Section 2.5.1. For classical measurement error with multiple covariates, the design matrix \mathbf{D}_i consists of columns of ones and zeros, which denote which component of \mathbf{X}_i each error-prone measurement corresponds to.

Random-intercepts and slopes

The random-intercepts and slopes model assumes there exists a bivariate random effects vector $\mathbf{X}_i = (X_{i1}, X_{i2})^T$, in which X_{i1} and X_{i2} represent subject i 's intercept and rate of increase in the longitudinal process over time. Assume that measurement W_{ij} is made at time t_{ij} . The random-intercepts and slopes model is then specified by defining the j th row of the matrix \mathbf{D}_i to be equal to $(1, t_{ij})$. The model specification is completed by assuming $\text{Var}(\mathbf{U}_i) = \text{diag}(\sigma_U^2)$. This gives the so called random-intercepts and slopes model.

Polynomial and spline models

The random-intercepts and slopes model assumes that the underlying longitudinal process evolves linearly in time. While this may be a reasonable model for certain settings, we may wish to use models which allow subjects' trajectories to vary in a more flexible way over time. Higher-order time effects can be included to model the longitudinal process as a polynomial function of time. Furthermore, so called spline models, such as a piece-wise linear function of time, can be fitted within the linear mixed model framework [116].

6.1.2 Multiple longitudinal processes

The model of equation (6.1) also encompasses models for error-prone measurements of more than one longitudinal process. For example, we may measure both systolic and diastolic blood pressure over time in subjects. Following Li *et al* [117], suppose that we observe error-prone measurements of g longitudinal processes, and that for the j th process the corresponding error-prone measurements, $\mathbf{W}_i^{(j)}$ follow a linear mixed model:

$$\mathbf{W}_i^{(j)} = \mathbf{D}_i^{(j)}\mathbf{X}_i^{(j)} + \mathbf{U}_i^{(j)}$$

where $\mathbf{D}_i^{(j)}$ is a known design matrix, $\mathbf{X}_i^{(j)}$ is a vector of random-effects, and $\mathbf{U}_i^{(j)}$ is a vector of mean zero within-subject errors, assumed independent of $\mathbf{X}_i^{(j)}$. Now define $\mathbf{W}_i = (\mathbf{W}_i^{(1)T}, \dots, \mathbf{W}_i^{(g)T})^T$, and similarly define \mathbf{X}_i and \mathbf{U}_i for the $\mathbf{X}_i^{(j)}$ s and $\mathbf{U}_i^{(j)}$ s. If we let $\mathbf{D}_i = \mathbf{D}_i^{(1)} \oplus, \dots, \oplus \mathbf{D}_i^{(g)}$, then \mathbf{W}_i follows a linear mixed model:

$$\mathbf{W}_i = \mathbf{D}_i\mathbf{X}_i + \mathbf{U}_i,$$

as in equation (6.1). This setting is more complicated than when a single longitudinal process is measured, as it usually implies a particular structure amongst the $\mathbf{X}_i^{(j)}$ s and $\mathbf{U}_i^{(j)}$ s. We would usually expect there to be correlation between different longitudinal processes, which is dictated by the assumed structure for $\text{Var}(\mathbf{X}_i)$. Since each $\mathbf{X}_i^{(j)}$ may be a vector, it may often be necessary to impose structure on $\text{Var}(\mathbf{X}_i)$ to avoid modelling a prohibitively large number of covariance parameters. The measurement errors of different longitudinal processes would usually be expected to have different variances. Depending on the application, there may also be correlations between the errors of measurements of different longitudinal processes which are made at the same time. Returning to the blood pressure example, it is quite likely that if a systolic blood pressure measurement taken on a subject on a particular day is above their current underlying or true level, we might expect the diastolic measurement made on the same day to also be above their true level.

6.1.3 Definition of ‘errors’ and model extensions

Whilst we have used the term ‘errors’ for \mathbf{U}_i , we emphasize that in the context of longitudinal error-prone measurements its elements may often be the sum of both technical measurement error and within subject temporal variability in measurements which are not explained by the subject-level random effects. A key assumption we make is the elements of \mathbf{U}_i are independent. We must therefore define the subject-level random-effects \mathbf{X}_i in a way in which we hope captures the underlying trajectory of the process under measurement, such that the elements of the ‘er-

ror' vector \mathbf{U}_i are mutually independent. One approach to this is to specify a rich subject-level random-effects structure, for example using spline models.

An alternative is to incorporate a serial correlation term. One way to do this is to decompose the residual errors \mathbf{U}_i into $\mathbf{U}_{(1)i} + \mathbf{U}_{(2)i}$ where $\mathbf{U}_{(1)i}$ represents pure measurement errors which are independent of each other and $\mathbf{U}_{(2)i}$ represents serial correlation. Serial correlation is achieved by assuming a variance covariance structure for $\mathbf{U}_{(2)i}$ whereby the covariance between two elements is a function of their times of measurement. A variety of functional specifications are possible (see [37]), but they are usually a decreasing function of the time difference between measurements. Incorporating serial correlation terms and specifying a richer subject random-effects structure are two alternatives (although they can be combined) to allowing for (residual) correlation between measurements within subjects.

Although the model defined in equation (6.1) is flexible, there are numerous ways in which it may be extended. The model is a so called two-level model, meaning that measurements (level 1) are clustered within subjects (level 2). Multi-level mixed models allow specification of additional sets of random effects which partition variability in the data into that which can be attributed to each level of clustering. For example, suppose in an observational study that subjects are followed-up repeatedly over time, and that at each follow-up visit a variable of interest is measured repeatedly with error. A three-level model could then be specified which includes random subject effects, random visit effects, and random residual errors. Such a model decomposes variability in the error-prone measurements into that which can be ascribed to differences between subjects in their overall longitudinal trajectories, that which can be ascribed to variability between follow-up visits around this trajectory (assumed mutually independent), and the remainder which is within visit variability. In such a model the role of the random visit effects is to allow for the fact that a subject's underlying value on a given day will usually deviate somewhat from the value implied by their subject-level random-effects.

6.2 Outcome model specification

The specification of a joint model is completed by defining the outcome model. This specification obviously depends on the type of outcome – we consider continuous, binary, and survival time outcomes in Chapters 7, 8, and 9 respectively. Over the last 20 years there has been a great deal of research effort into developing estimation methods for such models, particularly for survival or time-to-event outcomes.

For continuous and binary outcomes, the methodological literature regarding estimation methods almost exclusively considers the situation in which the distribution of the outcome Y_i is assumed to potentially depend on all of the elements of \mathbf{X}_i (and error-free \mathbf{Z}_i , if present) (e.g. [59, 118]). For censored survival time outcomes,

the longitudinal measurements are often observed at the same time as subjects are at-risk for the event of interest. This leads to models in which hazard is modelled as a function of time-dependent covariates, which themselves are functions of the underlying longitudinal process(es). Methodological papers consider estimation of model parameters given a particular specification for how the hazard at time t depends on the values of the longitudinal process up to and including time t . A common specification for this is that the hazard at time t is assumed to depend on the longitudinal process only through its current value at time t [94].

In contrast, in applications, we rarely know a priori which aspects of a longitudinal process(es) affect the outcome of interest, and we may therefore be interested in fitting a number of different models. As a motivating example, Boshuizen *et al* recently used data from the Seven Countries Study to investigate how mortality risk depends on both current, and levels 25 years earlier of systolic blood pressure (SBP) and total cholesterol [119]. In the Seven Countries Study, 6,518 men aged 40-59 were recruited between 1958 and 1964, and were followed up for 35 years for mortality. Blood pressure and cholesterol was measured at recruitment into the study, at 5 and 10 years after baseline, and then at numerous later times which varied by country. Boshuizen *et al* first fitted linear mixed models to the longitudinal measurements of SBP and cholesterol. They then fitted various Cox proportional hazards models for the hazard of death due to coronary heart disease (and also models for the hazard of death due to stroke), using an RC type approach to correct for measurement error and intra-subject variation. Boshuizen *et al* first fitted a Cox model to examine the effect of average SBP/cholesterol on hazard. Since subjects only had periodic measurements of SBP and cholesterol, and these were subject to measurement error and within-subject variation, at time t , Boshuizen *et al* used the BLUP of the mean level of SBP/cholesterol between time zero and time t as a time-dependent covariate. They then fitted models, again using RC, with either current SBP/cholesterol, or SBP/cholesterol level 25 years earlier as time-dependent covariates. Lastly, they fitted models in which both current and levels 25 years earlier were included simultaneously as covariates, to examine whether both current and past levels of SBP or cholesterol have independent effects on hazard.

The investigation by Boshuizen *et al* is typical of many epidemiological studies, whereby we may want to estimate the parameters of a number of different regression models for the outcome of interest. The components of the random-effects \mathbf{X}_i may be highly correlated, such that estimation of their independent effects, through fitting the ‘full model’ for Y_i given \mathbf{X}_i (and \mathbf{Z}_i), leads to imprecise estimates. Failure to find evidence of an independent effect of one component of \mathbf{X}_i is then likely due to a lack of power, as opposed to the lack of an independent effect. In such a situation, we may wish to proceed to fit a model for Y_i in which the corresponding component of \mathbf{X}_i is not included as a covariate. However, we may be wary of making

an assumption that this component has no independent effect on Y_i , since we have limited power to detect such an effect. In the following chapters we therefore explore how the various methods can be used for estimation of a model for Y_i in which some function $\mathbf{X}_i^* = \mathbf{A}\mathbf{X}_i$ of \mathbf{X}_i (such as a subset of its components) enters as covariate. In particular, we consider estimation of the parameters of this model when it is assumed Y_i only depends on \mathbf{X}_i via \mathbf{X}_i^* , but also when we do not wish to make this assumption.

6.3 Missing longitudinal error-prone measurements

In most longitudinal cohort studies, subjects are scheduled to have measurements of the longitudinal process(es) at particular times. For example, in the Framingham Heart Study (see Part III), subjects were scheduled to return for follow-up visits every two years after a baseline examination. In practice, some subjects will miss one or more scheduled visits, for a possibly large variety of different reasons, leading to missing error-prone measurements. In the estimation methods we consider in the following three chapters, analogous to conditioning on the number n_i of error-prone measurements in the classical covariate error setting, we do not specify a model for the missingness mechanism of longitudinal measurements, by treating the design matrix \mathbf{D}_i for subject i 's longitudinal measurements as fixed. In reality, the mechanism determining which error-prone measurements are missing for a particular subject is a random process, so that the matrices \mathbf{D}_i are stochastic. Assumptions regarding this mechanism which are sufficient for consistent estimation differ between the different estimation methods. We thus give details regarding sufficient conditions about this missingness mechanism as we describe each estimation method.

Chapter 7

Continuous outcomes

In this chapter we consider estimation of the parameters of joint models for longitudinal error-prone measurements and continuous outcomes Y_i . We assume that longitudinal error-prone measurements \mathbf{W}_i are available, and that a linear mixed model of the form described in Section 6.1 has been specified. We first consider how the methods previously described for allowing for classical covariate measurement error can be extended to this setting when Y_i is assumed to follow a linear regression model given \mathbf{X}_i and \mathbf{Z}_i :

$$Y_i = \beta_0 + \boldsymbol{\beta}_X^T \mathbf{X}_i + \boldsymbol{\beta}_Z^T \mathbf{Z}_i + \epsilon_i, \quad (7.1)$$

where $\mathbb{E}(\epsilon_i | \mathbf{X}_i, \mathbf{Z}_i) = 0$. In Section 7.1 we note why a naive two-stage approach to estimation results in biased estimates. In Section 7.2 we describe how RC is easily extended to the longitudinal setting. In Section 7.3 we consider ML estimation for a particular parametric model, and give the implementation details of the EM algorithm for this model. In Section 7.4 we note that our approach based on fitting a linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i , described in Section 3.6 for the case of classical measurement error, can also be used in this more general setting in which the measurement model is a linear mixed model. In Section 7.5 we describe how MI can be implemented after first fitting this mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i .

As discussed in Section 6.2, in applications we may often be interested in estimating the parameters of a number of different regression models for Y_i , in which different functions of \mathbf{X}_i enter as covariates, rather than \mathbf{X}_i itself. In Section 7.6 we introduce a framework in which rather than \mathbf{X}_i , a lower dimension linear function \mathbf{X}_i^* of \mathbf{X}_i is used as covariate in the regression model for Y_i . We then consider how each of the estimation methods can be used when interest lies in the estimating the parameters of such an alternative outcome model.

7.1 A naive two-stage approach

The first method we consider is an analogue of the ‘naive method’ described for classical error in Section 3.2. For covariates measured with classical error a naive analysis can be performed by treating an error-prone measurement as if it were equal to the unobserved covariate of interest. With error-prone measurements which are made longitudinally there is no direct analogue of such a naive analysis, as each subject has multiple error-prone measurements which may be a complicated function of the covariates \mathbf{X}_i . However, one so called ‘naive approach’, considered by Wang *et al* [59], is to substitute in place of the unobserved \mathbf{X}_i the least squares estimate of \mathbf{X}_i based on the available error-prone measurements \mathbf{W}_i . Following Wang *et al* [59], we briefly outline why this approach gives biased estimates by showing it is biased in a particular special case.

Wang *et al* assumed that \mathbf{U}_i is multivariate normal with covariance matrix parametrized up to a finite number of parameters, and that \mathbf{U}_i is independent of \mathbf{X}_i . They considered a two-stage approach where the unobserved \mathbf{X}_i are treated as unknown parameters to be estimated. Using the observed \mathbf{W}_i and design matrix \mathbf{D}_i , we can use standard least squares regression to estimate \mathbf{X}_i as:

$$\mathbf{X}_i^{(ls)} = (\mathbf{D}_i^T \mathbf{D}_i)^{-1} \mathbf{D}_i^T \mathbf{W}_i. \quad (7.2)$$

In order to calculate $\mathbf{X}_i^{(ls)}$, the design matrix \mathbf{D}_i must be of full rank, so that $\mathbf{D}_i^T \mathbf{D}_i$ is invertible. This condition effectively requires that there is some information with which to estimate all elements of \mathbf{X}_i . In the second stage, we fit the outcome model using $\mathbf{X}_i^{(ls)}$ as covariate in place of the unobserved \mathbf{X}_i .

To illustrate the method, suppose that we have chosen to model the longitudinal error-prone measurements \mathbf{W}_i by the random-intercepts and slopes model, as described in Section 6.1.1. In this case, the two elements of $\mathbf{X}_i^{(ls)}$ are equal to the OLS estimates of the intercept and slope of the regression of \mathbf{W}_i with design matrix \mathbf{D}_i . The requirement that \mathbf{D}_i be of full rank then corresponds to subject i having at least two error-prone measurements, and that these measurements were not made at the same time. In the second stage, we then fit the regression of Y_i with the predicted intercepts and slopes as covariates.

In general the resulting estimates of the parameters in the outcome model are inconsistent however. A simple way to see why this approach yields biased estimates is to consider the case of univariate X_i when $\mathbf{D}_i = (1, \dots, 1)^T$ and $\mathbf{U}_i \sim N(0, \text{diag}(\sigma_U^2))$, i.e. classical normally distributed measurement error. In this case, the least squares estimate of X_i is simply the mean of the n_i error-prone measurements, \bar{W}_i . Thus the naive two-stage approach in this case consists of fitting the outcome model using the mean of the error-prone measurements as covariate, a procedure which we have seen earlier produces biased estimates of the outcome model parameters.

7.2 Regression calibration

RC can be applied in the case of a linear mixed measurement model in exactly the same way as for multiple covariates measured with classical error (see Section 3.4.7). Taking expectations of equation (7.1) conditional on \mathbf{W}_i and \mathbf{Z}_i gives:

$$\mathbb{E}(Y_i|\mathbf{W}_i, \mathbf{Z}_i) = \beta_0 + \beta_X^T \mathbb{E}(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i) + \beta_Z^T \mathbf{Z}_i \quad (7.3)$$

Thus, as for classical measurement error, we must at least specify the conditional mean function $\mathbb{E}(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ in order to use RC.

For example, suppose that we model the longitudinal measurements \mathbf{W}_i using the random-intercepts and slopes model (see Section 6.1.1). We assume bivariate normality for the random-intercepts and slopes, with an unstructured covariance matrix, and assume independent, normally distributed measurement errors with constant variance σ_U^2 . Furthermore, assume that there are no error-free covariates \mathbf{Z}_i . To implement RC, we first fit the random-intercepts and slopes model to the error-prone measurements using either ML or REML. We can then calculate $\hat{\mathbb{E}}(\mathbf{X}_i|\mathbf{W}_i)$ by adding the estimated mean intercept and slope, $\boldsymbol{\mu}_X$, to the BLUPs of $\mathbf{X}_i - \boldsymbol{\mu}_X$, which follow from standard results for linear mixed models (for example, see Chapter 7 of [37]). We can then fit the OLS regression of Y_i on $\hat{\mathbb{E}}(\mathbf{X}_i|\mathbf{W}_i)$ to estimate β_X .

As for classical measurement error, RC is expected to be inefficient, relative to ML, since in the second stage all subjects are given equal weight, irrespective of the amount of error in the prediction of their value of \mathbf{X}_i . This may be a particular issue in the longitudinal setting, since in some applications the amount of information regarding \mathbf{X}_i may differ greatly between subjects because of differences in the number and timing of their error-prone measurements \mathbf{W}_i .

7.2.1 Missingness assumptions

In the first stage of RC we fit the linear mixed model (using either ML or REML) for the observed longitudinal measurements \mathbf{W}_i , conditioning on \mathbf{Z}_i if present. This ignores the missing data mechanism, but gives valid parameter estimates under the missing at random (MAR) assumption [37]. In our assumed model, this means that missingness of longitudinal measurements depends at most on the observed longitudinal measurements \mathbf{W}_i and the error-free covariates \mathbf{Z}_i . If missingness depends on Y_i , the unobserved (possibly hypothetical) longitudinal measurements, or the unobserved random-effects \mathbf{X}_i , the estimates in general will be biased. Thus, providing missingness depends at most on the observed longitudinal measurements and the error-free covariates, we expect RC to give consistent estimates of the outcome model parameters.

7.3 Maximum likelihood

As for classical measurement error, ML can be used for parameter estimation, given a correctly specified parametric model. We first define a joint parametric model, which is identical, except in the specification of the measurement model, to that used for the case of multiple covariates measured with classical measurement error (Section 3.5). We give the details of the observed data likelihood function, and then discuss approaches to finding the MLEs. In the following section (Section 7.4) we show that our novel proposal for ML estimation in the case of classical measurement error (Section 3.6) can also be used to find ML estimates with a linear mixed measurement error model.

7.3.1 Model specification

As in Section 3.6.3 we assume that \mathbf{X}_i is multivariate normal given \mathbf{Z}_i with:

$$\begin{aligned}\mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) &= \mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z\mathbf{Z}_i \\ \text{Var}(\mathbf{X}_i|\mathbf{Z}_i) &= \mathbf{\Sigma}_{X|Z}\end{aligned}$$

where $\mathbf{\Gamma}_Z$ is a $p \times q$ matrix of regression coefficients. We assume that Y_i follows the linear regression given in equation (7.1), with independent error $\epsilon \sim N(0, \sigma_\epsilon^2)$. Lastly, we assume that:

$$\mathbf{W}_i|\mathbf{X}_i \sim N(\mathbf{D}_i\mathbf{X}_i, \sigma_U^2\mathbf{I}_{n_i \times n_i}). \quad (7.4)$$

7.3.2 Likelihood function

The observed data likelihood function is given by the product of $f(Y_i, \mathbf{W}_i|\mathbf{Z}_i)$ over the n subjects. The joint density $f(Y_i, \mathbf{W}_i|\mathbf{Z}_i)$ is multivariate normal. Taking expectations of equation (7.1) and \mathbf{W}_i conditional on \mathbf{Z}_i , it follows that the mean function of $(Y_i, \mathbf{W}_i^T)^T$ is given by:

$$\mathbb{E} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \beta_0 + \boldsymbol{\beta}_X^T(\mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z\mathbf{Z}_i) + \boldsymbol{\beta}_Z^T\mathbf{Z}_i \\ \mathbf{D}_i(\mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z\mathbf{Z}_i) \end{pmatrix}$$

and similarly that the variance covariance matrix is given by:

$$\text{Var} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_X^T\mathbf{\Sigma}_{X|Z}\boldsymbol{\beta}_X + \sigma_\epsilon^2 & \boldsymbol{\beta}_X^T\mathbf{\Sigma}_{X|Z}\mathbf{D}_i^T \\ \mathbf{D}_i\mathbf{\Sigma}_{X|Z}\boldsymbol{\beta}_X & \mathbf{D}_i\mathbf{\Sigma}_{X|Z}\mathbf{D}_i^T + \sigma_U^2\mathbf{I}_{n_i \times n_i} \end{pmatrix}. \quad (7.5)$$

7.3.3 Estimation via the EM algorithm

As for classical error, the observed data likelihood function for this model is tractable, and can be maximized by gradient based methods or the EM algorithm. We now give the details of the EM algorithm for the previously defined model.

The outcome model and model for $f(\mathbf{X}_i|\mathbf{Z}_i)$ are the same as that assumed by Schafer and Purdy [32], who give expressions for the updated estimates of $\beta_0, \boldsymbol{\beta}_X, \boldsymbol{\beta}_Z, \sigma_\epsilon^2, \boldsymbol{\Gamma}_0, \boldsymbol{\Gamma}_Z, \boldsymbol{\Sigma}_{X|Z}$ (they assume the first element of \mathbf{Z}_i corresponds to an intercept, and so there is no β_0 in their formulation).

The complete data log likelihood component corresponding to $f(\mathbf{W}_i|\mathbf{X}_i)$ (as specified in equation (7.4)) is given by:

$$-0.5 \sum_{i=1}^n \log |\sigma_U^2 \mathbf{I}_{n_i \times n_i}| + (\mathbf{W}_i - \mathbf{D}_i \mathbf{X}_i)^T (\sigma_U^2 \mathbf{I}_{n_i \times n_i})^{-1} (\mathbf{W}_i - \mathbf{D}_i \mathbf{X}_i).$$

Using the fact that the determinant of a multiple of the identity matrix is equal to the scalar raised to the dimension of the matrix and that the inverse of the identity matrix is the identity matrix, this expression equals

$$-0.5 \sum_{i=1}^n n_i \log(\sigma_U^2) + (\mathbf{W}_i - \mathbf{D}_i \mathbf{X}_i)^T (\mathbf{W}_i - \mathbf{D}_i \mathbf{X}_i) / \sigma_U^2,$$

and expanding the matrix term gives:

$$-0.5 \sum_{i=1}^n n_i \log(\sigma_U^2) + (\mathbf{W}_i^T \mathbf{W}_i - 2\mathbf{W}_i^T \mathbf{D}_i \mathbf{X}_i + \mathbf{X}_i^T \mathbf{D}_i^T \mathbf{D}_i \mathbf{X}_i) / \sigma_U^2.$$

We now take expectations conditional on the observed data (which we suppress in the notation), which using a standard result for the expectation of quadratic forms (e.g. Appendix S3 of [24]), gives:

$$\begin{aligned} -0.5 \sum_{i=1}^n n_i \log(\sigma_U^2) + (\mathbf{W}_i^T \mathbf{W}_i - 2\mathbf{W}_i^T \mathbf{D}_i \mathbb{E}(\mathbf{X}_i) \\ + \text{tr}(\mathbf{D}_i^T \mathbf{D}_i \text{Var}(\mathbf{X}_i)) + \mathbb{E}(\mathbf{X}_i)^T \mathbf{D}_i^T \mathbf{D}_i \mathbb{E}(\mathbf{X}_i)) / \sigma_U^2. \end{aligned}$$

Differentiating with respect to σ_U^2 and solving the resulting equation for zero then results in the updated estimate of σ_U^2 given by:

$$\hat{\sigma}_U^2 = \frac{\sum_{i=1}^n (\mathbf{W}_i^T \mathbf{W}_i - 2\mathbf{W}_i^T \mathbf{D}_i \mathbb{E}(\mathbf{X}_i) + \text{tr}(\mathbf{D}_i^T \mathbf{D}_i \text{Var}(\mathbf{X}_i)) + \mathbb{E}(\mathbf{X}_i)^T \mathbf{D}_i^T \mathbf{D}_i \mathbb{E}(\mathbf{X}_i))}{\sum_{i=1}^n n_i}$$

where $\text{tr}()$ denotes the trace of a matrix.

The M-step of EM for this model thus requires evaluation of $\mathbb{E}(\mathbf{X}_i|Y_i, \mathbf{Z}_i, \mathbf{W}_i)$ and $\text{Var}(\mathbf{X}_i|Y_i, \mathbf{Z}_i, \mathbf{W}_i)$, where the unknown model parameters are replaced by their

current estimates. These follow from standard results for multivariate normal distributions:

$$\begin{aligned} \mathbb{E}(\mathbf{X}_i|Y_i, \mathbf{Z}_i, \mathbf{W}_i) &= \mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) + \text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)\text{Var}((Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)^{-1} \\ &\quad \times \left(\begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} - \begin{pmatrix} \beta_0 + \boldsymbol{\beta}_X^T \mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) + \boldsymbol{\beta}_Z^T \mathbf{Z}_i \\ \mathbf{D}_i \mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) \end{pmatrix} \right) \end{aligned} \quad (7.6)$$

and

$$\begin{aligned} \text{Var}(\mathbf{X}_i|Y_i, \mathbf{Z}_i, \mathbf{W}_i) &= \boldsymbol{\Sigma}_{X|Z} \\ &\quad - \text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)\text{Var}((Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)^{-1}\text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)^T. \end{aligned} \quad (7.7)$$

The covariance matrix:

$$\begin{aligned} \text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i) &= \begin{pmatrix} \text{Cov}(\mathbf{X}_i, Y_i|\mathbf{Z}_i) & \text{Cov}(\mathbf{X}_i, \mathbf{W}_i|\mathbf{Z}_i) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{X|Z}\boldsymbol{\beta}_X & \boldsymbol{\Sigma}_{X|Z}\mathbf{D}_i^T \end{pmatrix} \end{aligned}$$

and $\text{Var}((Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)$ was given in equation (7.5).

7.3.4 Missingness assumptions

The likelihood approach we have described is based on the likelihood of the observed data $(Y_i, \mathbf{W}_i, \mathbf{Z}_i)$. Provided that missingness of longitudinal error-prone measurements depends at most on the observed data, the likelihood approach we have described gives consistent parameter estimates. We note that in contrast to RC, the likelihood approach which ignores the missing data mechanism (for longitudinal measurements) is still valid if missingness depends on the outcome Y_i . If however missingness (conditional on observed data) depends on the unobserved (possibly hypothetical) longitudinal measurements, or the random-effects \mathbf{X}_i , biased results are expected in general.

7.4 Maximum likelihood estimation using linear mixed models

In Section 3.6.3 we showed how ML estimates (for a particular parametric model specification) could be obtained in the case of multiple covariates measured by classical error by fitting a linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i . The derivation of Section 3.6.3 did not however depend on the particular structure of the design matrix \mathbf{D}_i or $\text{Var}(\mathbf{U}_i)$. It thus follows that the same approach can be used to obtain

ML estimates when \mathbf{W}_i follows a general linear mixed model as described in Section 6.1.

For completeness, we recall from Section 3.6.3 that \mathbf{W}_i follows a linear mixed model given Y_i and \mathbf{Z}_i with:

$$\mathbb{E}(\mathbf{W}_i|Y_i, \mathbf{Z}_i) = \mathcal{D}_i\boldsymbol{\gamma}$$

where:

$$\mathcal{D}_i = \begin{pmatrix} \mathbf{D}_i & Y_i\mathbf{D}_i & Z_{i1}\mathbf{D}_i & \dots & Z_{iq}\mathbf{D}_i \end{pmatrix}$$

is the fixed effects design matrix, with corresponding vector of parameters $\boldsymbol{\gamma}$. The conditional variance covariance matrix is given by:

$$\text{Var}(\mathbf{W}_i|Y_i, \mathbf{Z}_i) = \mathbf{D}_i\boldsymbol{\Sigma}_{X|Z,Y}\mathbf{D}_i^T + \text{Var}(\mathbf{U}_i),$$

which implies a p -dimensional random effects vector, with design matrix \mathbf{D}_i , unstructured covariance matrix $\boldsymbol{\Sigma}_{X|Z,Y}$, and residual covariance matrix $\text{Var}(\mathbf{U}_i)$. Calculation of the MLEs of $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ then proceeds as described in Section 3.6.3. As described in that sub-section, although Wald-type intervals can in principle be found, non-parametric bootstrapping may be a more convenient approach for inference.

7.5 Multiple imputation

Our implementation of MI, described in Section 3.7, can be easily extended to the case of a linear mixed measurement model. As for classical error, the linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i , described in Section 7.4 can first be fitted using ML or REML. Frequentist MI then consists of using the estimated parameter values to draw from the conditional distribution $f(\mathbf{X}_i|Y_i, \mathbf{Z}_i, \mathbf{W}_i)$. For the same reasons discussed in Section 3.7.2, as the number of imputations M is increased, the resulting estimates of $\beta_0, \boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ are expected to (asymptotically, as $n \rightarrow \infty$) have the same efficiency as the MLEs.

7.5.1 Missingness assumptions

As the implementation of MI as described is based on imputing from the model with parameters estimated by ML, estimates from MI are consistent under the same assumptions required for validity of ML (see Section 7.3.4).

7.6 Alternative outcome model covariate specifications

In this section we consider the estimation of the parameters of regression models for Y_i in which, rather than \mathbf{X}_i , a vector \mathbf{X}_i^* , which is some function of \mathbf{X}_i is used as covariate. In Section 7.6.1 we develop a framework for investigating this. We restrict attention to \mathbf{X}_i^* which are a linear function of \mathbf{X}_i . We then derive expressions for the parameters of the regression model which uses \mathbf{X}_i^* and \mathbf{Z}_i as covariates in terms of the parameters of the ‘full’ regression model which has \mathbf{X}_i and \mathbf{Z}_i as covariates. In the remaining sections, we consider how the previously described estimation methods (RC, ML, MI) can be adapted to estimate the parameters of this alternative outcome model.

7.6.1 Alternative outcome model covariate specifications

As previously stated, we restrict attention to \mathbf{X}_i^* which are linear functions of \mathbf{X}_i :

$$\mathbf{X}_i^* = \mathbf{A}\mathbf{X}_i \quad (7.8)$$

where \mathbf{A} is a full-rank matrix with p (the length of \mathbf{X}_i) columns and number of rows less than than p . The matrix \mathbf{A} thus defines a linear transformation of \mathbf{X}_i , i.e. the elements of \mathbf{X}_i^* consist of linear functions of the elements of \mathbf{X}_i .

To illustrate the flexibility of this framework, we give some simple examples. Suppose $\mathbf{X}_i = (X_{i1}, X_{i2})^T$ is a bivariate random effects vector. Then by defining:

$$\mathbf{A} = (0.5, 0.5),$$

we have $X_i^* = 0.5X_{i1} + 0.5X_{i2}$. This definition therefore corresponds to including the mean of X_{i1} and X_{i2} as covariate in the alternative regression model for Y_i . Alternatively, defining:

$$\mathbf{A} = (1, 0),$$

yields $X_i^* = X_{i1}$, corresponding to omitting X_{i2} from the regression model for Y_i . For example, in the random-intercepts and slopes model, such a definition of \mathbf{A} corresponds to the outcome model for Y_i only containing the random-intercepts as covariate, but not the slopes.

We now consider the regression model for Y_i in which we use \mathbf{X}_i^* in place of \mathbf{X}_i , plus \mathbf{Z}_i , as covariates. Regardless of the specification of \mathbf{X}_i^* , as in Section 3.1.1 we can express Y_i in terms of \mathbf{X}_i^* and \mathbf{Z}_i as:

$$Y_i = \beta_{0^*} + \boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^* + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i + \epsilon_i^*, \quad (7.9)$$

where ϵ_i^* has mean zero and has zero correlation with \mathbf{X}_i^* and \mathbf{Z}_i . This expression holds irrespective of whether Y_i is conditionally independent of \mathbf{X}_i given \mathbf{X}_i^* and \mathbf{Z}_i .

We now show how β_{X^*} and β_{Z^*} can be expressed in terms of β_X, β_Z , the covariance matrix of $(\mathbf{X}_i^T, \mathbf{Z}_i^T)^T$ and its covariance with Y_i . Recall from equation (3.14) in Section 3.2.1 that β_X, β_Z are equal to:

$$\begin{pmatrix} \beta_X \\ \beta_Z \end{pmatrix} = \begin{pmatrix} \Sigma_X & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_Z \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{XY} \\ \Sigma_{ZY} \end{pmatrix}. \quad (7.10)$$

The variance covariance matrix of $(\mathbf{X}_i^T, \mathbf{Z}_i^T)^T$ can be inverted block-wise (see for example Section 3.4 of [120]), giving:

$$\beta_X = (\Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX})^{-1}(\Sigma_{XY} - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZY}) \quad (7.11)$$

and:

$$\begin{aligned} \beta_Z &= \Sigma_Z^{-1}\Sigma_{ZX}(\Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX})^{-1}\Sigma_{XY} \\ &\quad + (\Sigma_Z^{-1} + \Sigma_Z^{-1}\Sigma_{ZX}(\Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX})^{-1}\Sigma_{XZ}\Sigma_Z^{-1})\Sigma_{ZY} \\ &= \Sigma_Z^{-1}\Sigma_{ZX}(\Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX})^{-1}(\Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZY} - \Sigma_{ZY}) + \Sigma_Z^{-1}\Sigma_{ZY} \\ &= \Sigma_Z^{-1}(\Sigma_{ZY} - \Sigma_{ZX}\beta_X). \end{aligned} \quad (7.12)$$

We can substitute \mathbf{X}_i^* for \mathbf{X}_i in equation (7.11), and using the definition $\mathbf{X}_i^* = \mathbf{A}\mathbf{X}_i$, it follows that:

$$\begin{aligned} \beta_{X^*} &= (\Sigma_{X^*} - \Sigma_{X^*Z}\Sigma_Z^{-1}\Sigma_{ZX^*})^{-1}(\Sigma_{X^*Y} - \Sigma_{X^*Z}\Sigma_Z^{-1}\Sigma_{ZY}) \\ &= (\mathbf{A}\Sigma_X\mathbf{A}^T - \mathbf{A}\Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX}\mathbf{A}^T)^{-1}(\mathbf{A}\Sigma_{XY} - \mathbf{A}\Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZY}) \\ &= (\mathbf{A}(\Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX})\mathbf{A}^T)^{-1}\mathbf{A}(\Sigma_{XY} - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZY}) \\ &= (\mathbf{A}(\Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX})\mathbf{A}^T)^{-1}\mathbf{A}(\Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX})\beta_X. \end{aligned} \quad (7.13)$$

In the special case in which $\mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) = \Gamma_0 + \Gamma_Z\mathbf{Z}_i$ with constant covariance matrix $\Sigma_{X|Z} = \Sigma_X - \Sigma_{XZ}\Sigma_Z^{-1}\Sigma_{ZX}$, this gives:

$$\beta_{X^*} = (\mathbf{A}\Sigma_{X|Z}\mathbf{A}^T)^{-1}\mathbf{A}\Sigma_{X|Z}\beta_X. \quad (7.14)$$

Lastly, using equation (7.12) we can express β_{Z^*} as:

$$\begin{aligned} \beta_{Z^*} &= \Sigma_Z^{-1}(\Sigma_{ZY} - \Sigma_{ZX}\mathbf{A}^T\beta_{X^*}) \\ &= \beta_Z + \Sigma_Z^{-1}\Sigma_{ZX}(\beta_X - \mathbf{A}^T\beta_{X^*}). \end{aligned} \quad (7.15)$$

Again in the special case in which $\mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) = \Gamma_0 + \Gamma_Z\mathbf{Z}_i$, $\Sigma_Z^{-1}\Sigma_{ZX} = \Gamma_Z^T$, giving:

$$\beta_{Z^*} = \beta_Z + \Gamma_Z^T(\beta_X - \mathbf{A}^T\beta_{X^*}). \quad (7.16)$$

Thus β_{X^*} and β_{Z^*} can be expressed in terms of parameters of the regression model for Y_i which has \mathbf{X}_i and \mathbf{Z}_i as covariates, β_X and β_Z , the variance covariance matrix $\Sigma_{X|Z}$, and the matrix of regression coefficients Γ_Z .

In the special case in which there are no \mathbf{Z}_i , \mathbf{X}_i is two-dimensional, and we omit X_{i2} by specifying $\mathbf{A} = (1, 0)$ (which we use in our simulations), this gives:

$$\beta_{X^*} = \beta_{X_1} + \frac{\beta_{X_2} \text{Cov}(X_{i1}, X_{i2})}{\text{Var}(X_{i1})}. \quad (7.17)$$

7.6.2 Regression calibration

Assuming conditional independence

One of the appeals of RC is that it separates the modelling of the longitudinal error-prone measurements, without reference to the outcome Y_i , and the fitting of the outcome model for Y_i , into two stages. Having calculated $\mathbb{E}(\mathbf{X}_i | \mathbf{W}_i, \mathbf{Z}_i)$, we can predict \mathbf{X}_i^* by:

$$\mathbb{E}(\mathbf{X}_i^* | \mathbf{W}_i, \mathbf{Z}_i) = \mathbf{A} \mathbb{E}(\mathbf{X}_i | \mathbf{W}_i, \mathbf{Z}_i).$$

An obvious approach to estimating β_{X^*} and β_{Z^*} is then to regress Y_i on these predicted values. Taking expectations of equation (7.9) conditional on \mathbf{W}_i and \mathbf{Z}_i :

$$\mathbb{E}(Y_i | \mathbf{W}_i, \mathbf{Z}_i) = \beta_{0^*} + \beta_{X^*}^T \mathbb{E}(\mathbf{X}_i^* | \mathbf{W}_i, \mathbf{Z}_i) + \beta_{Z^*}^T \mathbf{Z}_i + \mathbb{E}(\epsilon_i^* | \mathbf{W}_i, \mathbf{Z}_i).$$

Therefore, providing $\mathbb{E}(\epsilon_i^* | \mathbf{W}_i, \mathbf{Z}_i) = 0$, we can thus substitute $\hat{\mathbb{E}}(\mathbf{X}_i^* | \mathbf{W}_i, \mathbf{Z}_i)$ in place of the unobserved \mathbf{X}_i^* to obtain consistent estimates of β_{0^*} , β_{X^*} and β_{Z^*} .

The validity of RC here relies on the assumption that $\mathbb{E}(\epsilon_i^* | \mathbf{W}_i, \mathbf{Z}_i) = 0$. Assuming that the measurement errors \mathbf{U}_i are independent of \mathbf{X}_i , \mathbf{Z}_i and ϵ_i , this follows if \mathbf{X}_i is independent of Y_i conditional on \mathbf{X}_i^* and \mathbf{Z}_i . This means that \mathbf{X}_i is uninformative regarding Y_i once \mathbf{X}_i^* and \mathbf{Z}_i are known. In particular it means that:

$$\mathbb{E}(Y_i | \mathbf{X}_i, \mathbf{Z}_i) = \mathbb{E}(Y_i | \mathbf{X}_i^*, \mathbf{Z}_i).$$

Thus if we are willing to assume that \mathbf{X}_i^* contains all the information in \mathbf{X}_i regarding Y_i , RC gives consistent estimates of the regression coefficients β_{0^*} , β_{X^*} and β_{Z^*} . If this conditional independence assumption does not hold, using RC is expected in general to give asymptotically biased estimates of β_{0^*} , β_{X^*} and β_{Z^*} . We therefore refer to these estimates in our simulations (see Section 7.7) by the term ‘naive RC’.

Not assuming conditional independence

Although RC is only expected to give consistent estimates of β_{X^*} and β_{Z^*} under the previously stated conditional independence assumption, it is still possible to obtain

consistent estimates by using the expressions given in Section 7.6 and estimates which are obtained in using RC to fit the regression model for Y_i which has \mathbf{X}_i and \mathbf{Z}_i as covariates. It is particularly simple if we assume $\mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) = \mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z\mathbf{Z}_i$ and $\text{Var}(\mathbf{X}_i|\mathbf{Z}_i) = \mathbf{\Sigma}_{X|Z}$. For in this case, $\boldsymbol{\beta}_{X^*}$ and $\boldsymbol{\beta}_{Z^*}$ can be estimated using equations (7.14) and (7.16). Equation (7.14) involves only the transformation matrix \mathbf{A} , the parameter $\boldsymbol{\beta}_X$, which can be estimated by using standard RC, and $\mathbf{\Sigma}_{X|Z}$, which is estimated in the first stage of RC. Similarly, $\boldsymbol{\beta}_{Z^*}$ can be estimated by substituting in the previously found estimates for $\boldsymbol{\beta}_Z$, $\boldsymbol{\beta}_{X^*}$ and $\mathbf{\Gamma}_Z$ (which is estimated in the first stage of RC) into equation (7.16). We refer to this approach in our simulation results as ‘corrected RC’.

7.6.3 Maximum likelihood

Assuming conditional independence

To incorporate an assumption of independence between Y_i and \mathbf{X}_i , conditional on \mathbf{X}_i^* and \mathbf{Z}_i , the mean and covariance matrices corresponding to the observed data are modified as follows. Taking expectations of equation (7.9) we have:

$$\begin{aligned}\mathbb{E}(Y_i|\mathbf{Z}_i) &= \beta_{0^*} + \boldsymbol{\beta}_{X^*}^T \mathbb{E}(\mathbf{X}_i^*|\mathbf{Z}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i \\ &= \beta_{0^*} + \boldsymbol{\beta}_{X^*}^T \mathbf{A} \mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i \\ &= \beta_{0^*} + \boldsymbol{\beta}_{X^*}^T \mathbf{A} \mathbf{\Gamma}_0 + (\boldsymbol{\beta}_{X^*}^T \mathbf{A} \mathbf{\Gamma}_Z + \boldsymbol{\beta}_{Z^*}^T) \mathbf{Z}_i\end{aligned}$$

and so the mean of $(Y_i, \mathbf{W}_i^T)^T$, conditional on \mathbf{Z}_i , is given by:

$$\mathbb{E} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \beta_{0^*} + \boldsymbol{\beta}_{X^*}^T \mathbf{A} \mathbf{\Gamma}_0 + (\boldsymbol{\beta}_{X^*}^T \mathbf{A} \mathbf{\Gamma}_Z + \boldsymbol{\beta}_{Z^*}^T) \mathbf{Z}_i \\ \mathbf{D}_i (\mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z \mathbf{Z}_i) \end{pmatrix}.$$

Similarly, the variance covariance matrix is given by:

$$\text{Var} \begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_{X^*}^T \mathbf{A} \mathbf{\Sigma}_{X|Z} \mathbf{A}^T \boldsymbol{\beta}_{X^*} + \sigma_{\epsilon^*}^2 & \boldsymbol{\beta}_{X^*}^T \mathbf{A} \mathbf{\Sigma}_{X|Z} \mathbf{D}_i^T \\ \mathbf{D}_i \mathbf{\Sigma}_{X|Z} \mathbf{A}^T \boldsymbol{\beta}_{X^*} & \mathbf{D}_i \mathbf{\Sigma}_{X|Z} \mathbf{D}_i^T + \sigma_U^2 \mathbf{I}_{n_i \times n_i} \end{pmatrix}. \quad (7.18)$$

If the conditional independence assumption is incorrect, in general we will obtain biased estimates of $\boldsymbol{\beta}_{X^*}$ and $\boldsymbol{\beta}_{Z^*}$.

The EM algorithm can be easily modified to incorporate the conditional independence assumption, and therefore to obtain the MLEs of this modified model. The M-step corresponding to $f(Y_i|\mathbf{X}_i^*, \mathbf{Z}_i)$ is as before, but with \mathbf{X}_i^* in place of \mathbf{X}_i . The conditional independence assumption also affects the expressions for $\mathbb{E}(\mathbf{X}_i|Y_i, \mathbf{W}_i)$

and $\text{Var}(\mathbf{X}_i|Y_i, \mathbf{W}_i)$. These are then given by:

$$\begin{aligned} \mathbb{E}(\mathbf{X}_i|Y_i, \mathbf{Z}_i, \mathbf{W}_i) &= \mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) + \text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)\text{Var}((Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)^{-1} \\ &\quad \times \left(\begin{pmatrix} Y_i \\ \mathbf{W}_i \end{pmatrix} - \begin{pmatrix} \beta_{0^*} + \boldsymbol{\beta}_{X^*}^T \mathbb{E}(\mathbf{X}_i^*|\mathbf{Z}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i \\ \mathbf{D}_i \mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) \end{pmatrix} \right) \end{aligned} \quad (7.19)$$

and

$$\begin{aligned} \text{Var}(\mathbf{X}_i|Y_i, \mathbf{Z}_i, \mathbf{W}_i) &= \boldsymbol{\Sigma}_{X|Z} \\ &\quad - \text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)\text{Var}((Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)^{-1}\text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)^T. \end{aligned} \quad (7.20)$$

where:

$$\begin{aligned} \text{Cov}(\mathbf{X}_i, (Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i) &= \begin{pmatrix} \text{Cov}(\mathbf{X}_i, Y_i|\mathbf{Z}_i) & \text{Cov}(\mathbf{X}_i, \mathbf{W}_i|\mathbf{Z}_i) \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{\Sigma}_{X|Z} \mathbf{A}^T \boldsymbol{\beta}_X & \boldsymbol{\Sigma}_{X|Z} \mathbf{D}_i^T \end{pmatrix} \end{aligned}$$

and $\text{Var}((Y_i, \mathbf{W}_i^T)^T|\mathbf{Z}_i)$ was given in equation (7.18).

Not assuming conditional independence

If we do not wish to make the conditional independence assumption, we can first find the MLEs of the model parameters as described in Section 7.3. This does not involve the specification of the transformation matrix \mathbf{A} . Second, by the invariance principle of ML, the MLEs of $\boldsymbol{\beta}_{X^*}$ and $\boldsymbol{\beta}_{Z^*}$ (for the model which does not assume conditional independence between Y_i and \mathbf{X}_i given \mathbf{X}_i^* and \mathbf{Z}_i) can be found by substituting in the MLEs of the required parameters into equations (7.14) and (7.16).

This approach yields consistent estimates of $\boldsymbol{\beta}_{X^*}$ and $\boldsymbol{\beta}_{Z^*}$ regardless of whether Y_i is conditionally independent of \mathbf{X}_i given \mathbf{X}_i^* and \mathbf{Z}_i . However, if the conditional independence assumption is valid, the resulting estimates are expected to be inefficient to some extent, relative to their MLEs for the model in which the conditional independence assumption is used.

7.6.4 Maximum likelihood using standard linear mixed models

Assuming conditional independence

We do not believe our approach to ML estimation based on fitting a linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i can be used to find the MLEs for the joint model in which it is assumed that $\mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{Z}_i) = \mathbb{E}(Y_i|\mathbf{X}_i^*, \mathbf{Z}_i)$. This assumption implies constraints between the fixed effects and random effects parameters in the mixed

model for \mathbf{W}_i given Y_i and \mathbf{Z}_i , and such constraints cannot be specified in linear mixed model commands in statistical packages.

Not assuming conditional independence

To find the MLEs of β_{X^*} and β_{Z^*} for the model which does not assume conditional independence, we can first find the MLEs of the joint model, as described in Section 7.4, and then substitute the relevant estimates into equations (7.14) and (7.16).

7.6.5 Multiple imputation

Not assuming conditional independence

As described in Section 7.5, multiple imputations of \mathbf{X}_i can be generated, by first finding the MLEs of the joint model in which Y_i is assumed to follow a linear regression given \mathbf{X}_i and \mathbf{Z}_i . After having created M multiple imputations of \mathbf{X}_i , denoted $\mathbf{X}_i^{(m)}$, $m = 1, \dots, M$, multiple imputations of \mathbf{X}_i^* can then be easily generated as:

$$\mathbf{X}_i^{*(m)} = \mathbf{A}\mathbf{X}_i^{(m)}$$

for $m = 1, \dots, M$. Because of the previously described asymptotic equivalence between ML and MI, we expect the resulting estimates to have the same efficiency as ML (as $M \rightarrow \infty$) for the model which does not assume conditional independence between Y_i and \mathbf{X}_i , given \mathbf{X}_i^* and \mathbf{Z}_i . MI may be particularly useful because having created the imputations of \mathbf{X}_i , imputations of \mathbf{X}_i^* can be generated for different specifications of the transformation \mathbf{A} using the same imputations of \mathbf{X}_i .

7.7 Simulations

In this section we report the results of simulations investigating the performance of RC and ML. We omit MI from the simulations because of the previously described asymptotic equivalence between MI and ML. We simulated the longitudinal data using the random-intercepts and slopes model we have previously discussed, and simulated a continuous outcome Y_i according to a linear regression model, with the random-intercepts and slopes as covariates.

7.7.1 Simulation setup

For $n = 1,000$ subjects we simulated the random-intercepts and slopes $\mathbf{X}_i = (X_{i1}, X_{i2})^T$ from a multivariate normal distribution with mean $(1, 1)^T$ and covariance

matrix:

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

We then generated Y_i according to a linear regression model:

$$Y_i = \beta_{X_1}X_{i1} + \beta_{X_2}X_{i2} + \epsilon_i$$

where $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. We performed simulations with $\beta_{X_1} = 1, \beta_{X_2} = 1$ and with $\beta_{X_1} = 1, \beta_{X_2} = 0$, corresponding to independent and equal effects of the intercepts and slopes, and no effect of slopes given intercepts respectively. For both, we performed simulations with either $\sigma_\epsilon^2 = 25$ or $\sigma_\epsilon^2 = 1$, to investigate the performance of the methods when the covariates are either weakly or strongly predictive of Y_i .

For each subject, their number of error-prone measurements, n_i , was generated as a random draw from the discrete uniform distribution, taking values between 1 and 20 (inclusive). For the j th measurement of subject i , we simulated the ‘time’ t_{ij} that the measurement took place from the continuous uniform distribution, taking values between 0 and 1. We then generated \mathbf{W}_i according to:

$$W_{ij} = X_{i1} + t_{ij}X_{i2} + U_{ij}$$

where the measurement errors U_{ij} were simulated independently from $N(0, 1)$. By having some subjects with only one or a few error-prone measurements and some with many more, we ensure that for some subjects \mathbf{X}_i can be predicted well while for others there is much greater uncertainty. We performed 10,000 simulations for each scenario.

7.7.2 Estimation methods

Regression calibration

To estimate the adjusted effects of X_{i1} and X_{i2} on Y_i using RC, we fitted a random intercepts and slopes model to the error-prone measurements \mathbf{W}_i using ML via the R command `lmer`. We then found the BLUPs of X_{i1} and X_{i2} using the `ranef()` command. We regressed Y_i on these BLUPs to estimate β_{X_1} and β_{X_2} .

To estimate the unadjusted effects of X_{i1} and X_{i2} on Y_i , we first regressed Y_i on the BLUPs of X_{i1} and then regressed Y_i on the BLUPs of X_{i2} . As described in Section 7.6.2, this approach gives consistent estimates if Y_i is conditionally independent of \mathbf{X}_i given \mathbf{X}_i^* . Here, this means that Y_i must be independent of the omitted component of \mathbf{X}_i conditional on the included component. We refer to these estimates by the term ‘naive RC’.

We also estimated the unadjusted effects of X_{i1} and X_{i2} on Y_i without making the previously described conditional independence assumptions. To do this, we substituted the RC estimates of the adjusted effects of X_{i1} and X_{i2} and the estimate of Σ_X from fitting the mixed model for \mathbf{W}_i into equation (7.14). We refer to these estimates by the term ‘corrected RC’.

Maximum likelihood

To estimate the adjusted effects of X_{i1} and X_{i2} on Y_i , we fitted the linear mixed model \mathbf{W}_i , conditional on Y_i , and calculated the MLEs of β_{X_1} and β_{X_2} as described in Section 7.4.

To estimate the unadjusted effects of X_{i1} and X_{i2} on Y_i , assuming that Y_i is independent of X_{i2} (respectively X_{i1}) conditional on X_{i1} (respectively X_{i2}), we used the EM algorithm, as described in Sections 7.3.3 and 7.6.3. We used the estimated values from the model fitted to the longitudinal measurements \mathbf{W}_i , i.e. the first stage of RC, as initial values for the mean vector and covariance matrix of \mathbf{X}_i and for σ_U^2 . For the linear regression outcome model parameters we used the corresponding values estimated which had been previously found by RC. At each iteration we calculated the improvement in the observed data log likelihood, and declared convergence when the change was less than 0.01.

To estimate the unadjusted effects of X_{i1} and X_{i2} on Y_i without making the previously described conditional independence assumptions, we substituted the ML estimates of the adjusted effects of X_{i1} and X_{i2} and the estimate of Σ_X from fitting the mixed model for \mathbf{W}_i into equation (7.14).

7.7.3 Simulation results

Adjusted effect estimates

Table 7.1 shows the mean (SD) of estimates of the adjusted effects of X_{i1} and X_{i2} on Y_i , using either RC or ML. Both methods showed little bias across all four scenarios, as expected. Of interest, the efficiency of RC was no worse than ML for $\sigma_\epsilon^2 = 25$ and only slightly worse than ML for $\sigma_\epsilon^2 = 1$. Based on our findings in Section 3.4.6, we expected there to be little difference in efficiency between RC and ML when the covariates explain only a small part of the variability in the outcome. However, for $\sigma_\epsilon^2 = 1$ we expected ML to have a larger efficiency advantage compared to RC since in our simulations there was large variability in the number n_i of error-prone measurements available for each subject. The lack of a difference may be due to the relatively small measurement error variance $\sigma_U^2 = 1$, although in these simulations there is much greater variability between subjects in the number of error-prone measurements available. Indeed, inspection of values of $\hat{\text{Var}}(\mathbf{X}_i|\mathbf{W}_i)$ for a single

Table 7.1: Adjusted effect estimates for linear regression simulations. Mean (SD) of estimates found using regression calibration (RC) and maximum likelihood (ML).

Scenario	β_{X_1}	β_{X_2}	σ_ϵ^2	$\hat{\beta}_{X_1}$		$\hat{\beta}_{X_2}$	
				RC	ML	RC	ML
1	1	1	25	0.995 (0.332)	0.995 (0.333)	1.004 (0.384)	1.007 (0.385)
2	1	1	1	0.989 (0.083)	0.990 (0.082)	1.009 (0.094)	1.012 (0.092)
3	1	0	25	1.003 (0.331)	1.003 (0.331)	-0.010 (0.382)	-0.009 (0.382)
4	1	0	1	1.009 (0.084)	1.009 (0.083)	-0.009 (0.097)	-0.008 (0.095)

simulated dataset showed large variation in this quantity, suggesting RC may be expected to be inefficient.

Unadjusted effect estimates

Table 7.2 shows the results for the unadjusted effect estimates of X_{i1} and X_{i2} using naive RC and ML (assuming conditional independence). As expected, in scenarios 1 and 2, in which X_{i1} and X_{i2} both have independent effects on Y_i , both methods gave biased estimates of the unadjusted effects. In scenarios 3 and 4, the naive RC and ML estimates of the unadjusted effect of X_{i2} had little bias. This is because for these scenarios X_{i2} has no independent effect on Y_i . As before, the efficiency of the naive RC estimates were very similar to that of ML, even for $\sigma_\epsilon^2 = 1$. The estimates of the unadjusted effect of X_{i2} are biased because the corresponding conditional assumption no longer holds – X_{i1} has an independent effect on Y_i , conditional on X_{i2} .

Table 7.3 shows the results for the unadjusted effect estimates of X_{i1} and X_{i2} found using corrected RC and ML, which do not assume conditional independence. As expected, the estimates for both methods showed little bias for all parameters and across all four scenarios. The RC based estimates had very similar efficiency to ML. For scenarios 3 and 4, the estimates of the unadjusted effect of X_{i1} on Y_i had greater variability than those in Table 7.2. This increased variability is the price paid for relaxing the conditional independence assumption in the case in which such an assumption is appropriate. However, the efficiency loss is not drastic, with the standard deviation of estimates 11% (scenario 3) and 14% (scenario 4) greater when the conditional independence assumption was not made.

7.8 Conclusions

In this chapter we have shown that RC and ML can be easily extended to the more general case of a linear mixed measurement error model.

Table 7.2: Unadjusted effect estimates assuming conditional independence for linear regression simulations. Mean (SD) of estimates found using naive regression calibration (naive RC) and maximum likelihood (ML).

Scenario	β_{X_1}	β_{X_2}	σ_ϵ^2	Unadjusted effect of X_{i1}			Unadjusted effect of X_{i2}		
				True	Naive RC	ML	True	Naive RC	ML
1	1	1	25	1.5	1.730 (0.180)	1.735 (0.181)	1.5	1.962 (0.227)	1.989 (0.229)
2	1	1	1	1.5	1.729 (0.061)	1.815 (0.064)	1.5	1.965 (0.115)	1.861 (0.095)
3	1	0	25	1	0.994 (0.174)	0.995 (0.174)	0.5	0.959 (0.213)	0.970 (0.217)
4	1	0	1	1	1.001 (0.041)	1.002 (0.041)	0.5	0.963 (0.081)	1.049 (0.074)

Table 7.3: Unadjusted effect estimates not assuming conditional independence for linear regression simulations. Mean (SD) of estimates found using corrected regression calibration (RC) and maximum likelihood (ML).

Scenario	β_{X_1}	β_{X_2}	σ_ϵ^2	Unadjusted effect of X_{i1}			Unadjusted effect of X_{i2}		
				True	Corrected RC	ML	True	Corrected RC	ML
1	1	1	25	1.5	1.503 (0.198)	1.505 (0.199)	1.5	1.506 (0.270)	1.510 (0.271)
2	1	1	1	1.5	1.503 (0.063)	1.504 (0.063)	1.5	1.513 (0.118)	1.516 (0.116)
3	1	0	25	1	0.995 (0.194)	0.996 (0.194)	0.5	0.501 (0.254)	0.502 (0.254)
4	1	0	1	1	1.002 (0.047)	1.003 (0.046)	0.5	0.500 (0.073)	0.502 (0.072)

7.8.1 Software

RC can be easily implemented using most modern statistical packages which include commands for fitting linear mixed models. Alternatively, we have shown how our estimation approach based on fitting a linear mixed model for \mathbf{W}_i given Y_i (and \mathbf{Z}_i , if present), can be extended to this setting to give ML estimates.

7.8.2 Statistical efficiency

Our simulation results indicate that, at least in some scenarios, RC estimates of the linear regression coefficients may have efficiency which is very similar to ML. This was the case even when the covariates were strongly predictive of the outcome and despite the fact that the number of error-prone measurements differed widely between subjects in our simulations. We may have observed a greater difference in efficiency had we simulated data with a larger value of σ_U^2 .

7.8.3 Assumptions

For classical measurement error we saw in Section 3.4 that RC gives consistent estimates even if the normality assumptions for \mathbf{X}_i are violated. In simulations with a random-intercepts and slopes model, Li *et al* reported that RC, implemented assuming normality of \mathbf{X}_i , showed little bias under non-normal distributions for \mathbf{X}_i [118]. Given the findings of Li *et al*, and our earlier robustness results in the case of classical measurement error, we believe that RC for the longitudinal setting, predicated on normality for \mathbf{X}_i , may also be robust to deviations from the normality assumption.

7.8.4 Alternative outcome model covariate specifications

Often we do not know which aspects of a covariate's longitudinal trajectory affect the outcome of interest Y_i , and so interest often lies on the exploration of different specifications for the outcome conditional on the longitudinal trajectory. RC is a particularly appealing approach in this case, since having calculated the BLUPs of \mathbf{X}_i , different models for Y_i can be fitted using a linear transformation \mathbf{X}_i^* of \mathbf{X}_i as covariate. The simplest example of this is when \mathbf{X}_i^* contains a subset of the components of \mathbf{X}_i . As we have described in Section 7.6.2, in this case, the validity of the resulting RC estimates relies on \mathbf{X}_i being conditionally independent of the outcome given \mathbf{X}_i^* (and \mathbf{Z}_i , if present). Our simulations results agree with this, showing that in general bias may occur if this conditional independence assumption does not hold. In particular, the values which are unbiasedly estimated differ from those values which would be unbiasedly estimated if we were able to regress Y_i on the true values of \mathbf{X}_i^* .

As our simulation results show, the same applies to ML estimation when one incorrectly assumes that Y_i only depends on \mathbf{X}_i through the value of \mathbf{X}_i^* , i.e. in general bias occurs if the assumption does not hold. Here however, in contrast to RC, the assumption is arguably more transparent, since it directly affects the specification of the assumed model and therefore also the implementation of the estimation algorithm. We have shown however that it is relatively easy to find estimates of these parameters without making such a conditional independence assumption, using either the RC or ML estimates of the ‘full’ model, which assumes Y_i may depend jointly on all elements of \mathbf{X}_i .

7.8.5 Inference

As for classical error, non-parametric bootstrapping can be used to find standard errors and confidence intervals for estimates from all of the methods described. As in the classical error case, when using non-parametric bootstrapping, we should sample subjects (with replacement). For ML, asymptotic standard errors and Wald intervals can be found using the observed information matrix, as in the case of classical error. Alternatively, all of the estimators we have described can also be expressed as solutions to unbiased estimating equations, and so the sandwich estimator of variance can in principle be used.

Chapter 8

Binary outcomes

We now consider the extension of the methods described previously for dealing with classical measurement error for binary outcomes to the longitudinal setting. As in Chapter 7, we first consider estimation when the regression model for Y_i is assumed to depend jointly on all of the elements of \mathbf{X}_i , given \mathbf{Z}_i . As for classical error, we therefore assume that Y_i follows a logistic regression given \mathbf{X}_i and \mathbf{Z}_i :

$$\text{logit}(P(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i)) = \beta_0 + \beta_X \mathbf{X}_i + \beta_Z \mathbf{Z}_i. \quad (8.1)$$

In Section 8.1 we discuss the use of RC. In Section 8.2 we consider ML estimation. This includes describing how ascent-based MCEM can be implemented for joint models with longitudinal error-prone measurements. In Section 8.3 we note that our previously described proposal for ML estimation by fitting a mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i also applies to the longitudinal setting. However, this approach is only appropriate when \mathbf{X}_i and \mathbf{Z}_i are multivariate normal given Y_i . In Section 8.4, as in Section 4.7, we describe how the fitted mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i can nevertheless be used to multiply impute \mathbf{X}_i . This approach permits consistent estimation under the weaker assumption of conditional normality of \mathbf{X}_i given \mathbf{Z}_i and Y_i . In Section 8.5 we describe the extension of the CS method to the longitudinal setting, and note a requirement of the method which may limit its applicability in certain settings.

As for continuous outcomes, in Section 8.6 we consider estimation when the logistic regression model for Y_i has a reduced subset, or a reduced subset of linear combinations of the original \mathbf{X}_i , $\mathbf{X}_i^* = \mathbf{A}\mathbf{X}_i$, and \mathbf{Z}_i as covariates.

8.1 Regression calibration

The implementation of RC proceeds as for a linear regression outcome model (see Section 7.2). As for classical measurement error in logistic regression, RC gives only approximately consistent estimates of β_X and β_Z . The bias of RC estimates is

expected to be small under the same conditions as for classical error with a logistic regression outcome model, namely that the effects of \mathbf{X}_i are small or $\text{Var}(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ is small. Wang *et al* reported simulation results for the performance of RC for a logistic regression outcome model and linear mixed measurement model [59]. They simulated data with bivariate \mathbf{X}_i representing subject-specific intercepts and slopes, which had correlation -0.1. Each subject had four error-prone measurements which followed the random-intercepts and slopes model, with measurement times normally distributed around ‘visit’ times 0, 1, 2 and 3, and normally distributed measurement error. Wang *et al*’s simulations results showed that RC had relatively small bias, but that the bias increased with larger effects of X_{i1} and X_{i2} , and to a lesser extent, when the measurement error variance was larger. Li *et al* reported that RC was ‘unacceptably biased’ in simulations with \mathbf{X}_i normally distributed [118]. However, they performed simulations in which the effects of the intercepts and slopes on the outcome Y_i had standardized odds ratios of $\exp(3) \approx 20$ and $\exp(2.5) \approx 12$ respectively, which would ordinarily be considered to be quite large effects.

8.2 Maximum likelihood

In this section we outline some of the issues regarding use of ML for joint models which include a logistic regression outcome model and linear mixed measurement error model.

8.2.1 Model specification

The most common parametric assumption is that of conditional normality for \mathbf{X}_i given \mathbf{Z}_i . Given the assumed logistic regression model for Y_i given \mathbf{X}_i and \mathbf{Z}_i , the only remaining specification is that for the measurement errors \mathbf{U}_i in the linear mixed model $\mathbf{W}_i = \mathbf{D}_i\mathbf{X}_i + \mathbf{U}_i$. The simplest assumption is that the errors U_{ij} of W_{ij} are normally distributed with constant variance σ_U^2 , and that they are independent of the errors U_{ik} of other measurements.

8.2.2 Estimation using Monte-Carlo Expectation Maximization

As for classical measurement error (see Section 4.4), the observed data likelihood function for the model which assumes normality for \mathbf{X}_i given \mathbf{Z}_i involves an intractable integral with respect to $f(\mathbf{X}_i|\mathbf{Z}_i)$. Thus the same computational difficulties are found as for logistic regression subject to classical measurement error case, and either deterministic quadrature methods or Monte-Carlo techniques can be used to find the MLEs. Monte-Carlo EM can be used to find the MLEs as described in Section 4.5. This involves multiply imputing \mathbf{X}_i from its conditional distribution given

$\mathbf{W}_i, \mathbf{Z}_i$ and Y_i using rejection sampling, followed by maximizing (for each imputation) the complete data likelihoods corresponding to the logistic regression outcome model, the multivariate normal model for \mathbf{X}_i given \mathbf{Z}_i , and the measurement model for \mathbf{W}_i given \mathbf{X}_i .

8.2.3 Semiparametric likelihood methods

As previously discussed, MLEs based on a fully parametric model are expected to be biased when the assumed model is incorrect. Semi-parametric methods such as the CS method (see Section 8.5) can be used, which provide consistent estimates without making any assumptions about the distribution of \mathbf{X}_i , but this comes at the expense of loss of efficiency. To address these concerns Li *et al* recently proposed a likelihood approach which makes weaker assumptions regarding the distribution of \mathbf{X}_i [121]. The approach of Li *et al* was to assume that the random effects distribution belongs to the SNP (seminonparametric) class of smooth densities (see Section 5.4.3) which includes ‘normal, multimodal, skewed, and heavy or thin tailed but not ‘unusual’ densities’. Li *et al* suggested using the EM algorithm for estimation, but, like the joint model which assumes marginal normality for \mathbf{X}_i , the E-step involves intractable integrals. Estimation for this model is therefore as computationally demanding as when marginal normality for \mathbf{X}_i is assumed, but with the additional complications created by assuming the density for \mathbf{X}_i belongs to the semiparametric SNP class. To simplify the estimation process, Li *et al* also proposed a two-stage pseudo-likelihood approach, in which the linear mixed model for \mathbf{W}_i , which assumes the density of \mathbf{X}_i belongs to the SNP class, is first estimated without reference to the outcome Y_i . The likelihood function is then maximized, treating the parameters involved in the mixed model for \mathbf{W}_i as known at their estimated values, using the probit approximation to the logistic function. This simplified procedure is still relatively complex, and is not trivial to implement.

Li *et al* performed a number of simulations with different distributions for a bivariate \mathbf{X}_i . They simulated \mathbf{X}_i according to a bivariate normal, a bimodal mixture of normals, a bivariate skew-normal distribution, and a bivariate t-distribution. They found that wrongly assuming bivariate normality for \mathbf{X}_i caused substantial bias when \mathbf{X}_i was a bimodal mixture of normals, but under the other distributions the bias was usually small. Their approach based on using the SNP class of densities for \mathbf{X}_i showed good performance, with little bias under all scenarios, and increased efficiency compared to the CS method. Their pseudo-likelihood two-stage approach also performed well compared to the full likelihood approach.

8.2.4 Sensitivity to assumptions

The simulation results of Li *et al* [121] showed that the MLEs for the logistic regression parameters for a joint model in which the \mathbf{X}_i are assumed to be normal are sometimes robust to violations of this normality assumptions. Recently, Huang *et al* showed that when the longitudinal information about \mathbf{X}_i is sufficiently rich, misspecifying the distribution of \mathbf{X}_i does not cause bias in parameter estimates [122]. They showed that as the information about \mathbf{X}_i increases, the observed data likelihood (and hence the MLE) does not depend on the underlying density of the random effects. Huang *et al* give a rigorous definition of the meaning of the longitudinal information increasing, but as with typical asymptotic arguments, whether the information is sufficient in any given situation may be difficult to assess. To address this, Huang *et al* proposed a re-measurement technique, similar to SIMEX, which can be used to investigate whether the analysis of a particular dataset is sensitive to the parametric assumptions for \mathbf{X}_i [122].

8.3 Maximum likelihood using standard linear mixed models

Our approach to ML estimation based on fitting a linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i when \mathbf{W}_i is subject to classical error (see Section 4.6.1) can also be used when the measurement model is a general linear mixed model, assuming that \mathbf{X}_i and \mathbf{Z}_i are multivariate normal given Y_i . The derivation in Section 4.6.1 did not rely on the particular specification of the design matrix \mathbf{D}_i and $\text{Var}(\mathbf{U}_i)$ which was appropriate in the case of multiple covariates measured with classical error, and so the results apply to the more general case in which \mathbf{W}_i follows a linear mixed model as described in Section 6.1.

8.4 Multiple imputation

As for classical measurement error, a major limitation of our ML approach based on fitting a linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i is the requirement that \mathbf{X}_i and \mathbf{Z}_i are jointly normal given Y_i . However, as in Section 4.7, having fitted the linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i as described in Section 8.3, we can create multiple imputations of \mathbf{X}_i . These multiple imputations can be used to fit the logistic regression model for Y_i using the imputed values of \mathbf{X}_i and the values of \mathbf{Z}_i , and the resulting parameter estimates averaged over imputations, as previously described. This approach will give consistent estimates of the logistic regression parameters providing the imputation model, which assumes that \mathbf{X}_i is normally distributed given Y_i and \mathbf{Z}_i , is correct.

8.5 Conditional score method

The CS method (and another method, called the sufficiency score method), originally conceived within the simple classical measurement error setting (see Section 4.9), has recently been extended to the case of a linear mixed measurement error model by Li *et al* [118]. Li *et al* assumed a linear mixed model $\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i$ for the longitudinal error-prone measurements. They assumed that $\mathbf{U}_i \sim N(\mathbf{0}, \sigma_U^2 \mathbf{I}_{n_i \times n_i})$, but made no assumptions regarding the distribution of the \mathbf{X}_i . Analogous to the classical measurement error setting, Li *et al* showed that $\tilde{\mathbf{X}}_i = \mathbf{D}_i^T \mathbf{W}_i + Y_i \sigma_U^2 \boldsymbol{\beta}_X$ is a complete sufficient statistic for \mathbf{X}_i when the parameters σ_U^2 and $\boldsymbol{\beta}_X$ are known. Using this property, Li *et al* derived the CS estimator as the solution to a set of unbiased estimating equations. In simulations with a random-intercepts and slopes model, Li *et al* found that the CS method gave consistent estimates of the logistic model parameters under a number of different distributions for \mathbf{X}_i .

In the case of classical measurement error, recall that the error variance σ_U^2 can be first estimated (using replication data) by ANOVA methods, regardless of the distribution of \mathbf{X}_i . The CS estimating equations in the case of classical error, as described in Section 4.9, then treat the estimated value of σ_U^2 as the true value. Li *et al* proposed augmenting the CS estimating equations with an additional estimating equation for σ_U^2 which does not require normality for \mathbf{X}_i [118]. However, in a more recent paper, Li *et al* gave simulation evidence suggesting that σ_U^2 can be estimated consistently by fitting the linear mixed model for \mathbf{W}_i which assumes normality for \mathbf{X}_i , even when this normality assumption does not hold [117]. This is in agreement with the earlier robustness results for mixed models of Verbeke and Lesaffre [123]. Thus an alternative approach is to first fit the linear mixed model to \mathbf{W}_i , assuming normality of \mathbf{X}_i , which yields an estimate of σ_U^2 . The estimating equations corresponding to the logistic regression model can then be solved, treating the estimated value of σ_U^2 as if it is known. Li *et al* (Web Appendix B of [117]) reported that in simulations the efficiency difference between estimates obtained in this way and those obtained by simultaneously estimating the logistic model parameters and σ_U^2 was small.

8.5.1 Multiple longitudinal processes and correlated errors

The CS estimating equations originally proposed by Li *et al* [118] only allowed a single error variance σ_U^2 , meaning that the approach could not be applied in the case of multiple longitudinal processes or when the measurement errors are correlated. In a more recent paper, the same authors have considered the extension of the CS method to the case of multiple longitudinal processes measured with error [117]. Here, Li *et al* assumed a separate mixed model $\mathbf{W}_i^{(j)} = \mathbf{D}_i^{(j)} \mathbf{X}_i^{(j)} + \mathbf{U}_i^{(j)}$ for each longitudinal process. They assumed that the errors were normally distributed, and have variance

covariance $\text{Var}(\mathbf{U}_i^{(j)})$ which may include (within longitudinal process) correlations. They assumed however that the errors of different longitudinal processes are uncorrelated. This means that the linear mixed model for each longitudinal process can be fitted (assuming normality of $\mathbf{X}_i^{(j)}$), and consistent estimates obtained of the parameters involved in $\text{Var}(\mathbf{U}_i^{(j)})$. A consistent estimate of the variance covariance matrix $\text{Var}(\mathbf{U}_i)$ can then be obtained as a block diagonal matrix, formed by the corresponding estimates of $\text{Var}(\mathbf{U}_i^{(j)})$. Li *et al* then generalized the CS estimating equations to this more general measurement model, which rather than depending on σ_U^2 , depend on $\text{Var}(\mathbf{U}_i)$ [117]. This was termed the generalized conditional score (GCS) method by Li *et al* [117].

Li *et al* [117] also noted that multiplying the mixed model equation for the combined longitudinal processes ($\mathbf{W}_i = \mathbf{D}_i\mathbf{X}_i + \mathbf{U}_i$) on the left by $\text{Var}(\mathbf{U}_i)^{-1/2}$ yields:

$$\begin{aligned}\tilde{\mathbf{W}}_i = \text{Var}(\mathbf{U}_i)^{-1/2}\mathbf{W}_i &= \text{Var}(\mathbf{U}_i)^{-1/2}\mathbf{D}_i\mathbf{X}_i + \text{Var}(\mathbf{U}_i)^{-1/2}\mathbf{U}_i \\ &= \tilde{\mathbf{D}}_i\mathbf{X}_i + \tilde{\mathbf{U}}_i,\end{aligned}$$

in which $\tilde{\mathbf{U}}_i \sim N(0, \mathbf{I})$ (where \mathbf{I} denotes the identity matrix, with number of rows equal to the total number of error-prone measurements). Thus the CS estimating equations for longitudinal data in the case of i.i.d. measurement errors can be applied, using $\tilde{\mathbf{W}}_i$ (or rather a consistent estimate of this, found by replacing $\text{Var}(\mathbf{U}_i)$ by its estimate) as the vector of error-prone measurements, and setting $\sigma_U^2 = 1$.

Li *et al* investigated, via simulation, the impact of correlation between measurement errors when only a single longitudinal process is observed, when it is assumed the errors are independent [117]. They found, perhaps unsurprisingly, that estimates found under the assumption of independence were biased, whereas using the GCS method described above showed much less bias.

Although the GCS method relaxes an assumption of independence between errors of the same longitudinal process, it still makes an assumption of independence between errors of the different longitudinal processes, when more than one is being modelled. However, Li *et al* noted that this can also in principle be relaxed, providing that $\text{Var}(\mathbf{U}_i)$ can be consistently estimated. Li *et al* conjectured that incorrectly assuming independence between errors of different longitudinal processes might at worst cause inefficiency, but not bias, and they reported that additional simulations appeared to support this [117].

8.5.2 Estimation and inference

The CS estimating equations, as in the case of classical measurement error, must be solved using an iterative method, such as Newton-Raphson. As in the case of classical measurement error, because the CS estimator is defined as the solution of

unbiased estimating equations, the resulting estimator is asymptotically normal and has variance which can be estimated by the usual sandwich estimator. When σ_U^2 or $\text{Var}(\mathbf{U}_i)$ is estimated separately in the first stage, and then treated as known, the estimating equations used to obtain these estimates (e.g. the linear mixed model likelihood score equations when this is used to estimate σ_U^2) can be stacked with the CS estimating equations for the logistic model parameters and treated as a single set of estimating equations.

8.5.3 Identification conditions for the design matrix \mathbf{D}_i

An important requirement of the conditional score method is that the design matrix \mathbf{D}_i in the linear mixed model for the longitudinal measurements must be of full-rank for each subject. This means that for each subject there is some information about all the components of \mathbf{X}_i . Depending on the application and assumed model, it is possible that this assumption may not be met for all subjects. For example, for the random-intercepts and slopes model, this requirement means that each subject must have at least two error-prone longitudinal measurements (and they must be made at different times), so that there is some information regarding the intercept and slope for each subject. Alternatively, if instead the underlying longitudinal process were modelled by piece-wise linear splines, the rank requirement for \mathbf{D}_i would mean that each subject must have at least one measurement between each pair of knots.

8.5.4 Missingness assumptions

Li *et al* [117,118] did not give details regarding the validity of the CS approach when longitudinal measurements are subject to missingness. Of course they are consistent if measurements are MCAR, but presumably they are also consistent under weaker assumptions. This is an area for future research.

8.6 Alternative outcome model covariate specifications

Analogous to Section 7.6, we now consider estimation when, instead of \mathbf{X}_i , interest lies in the parameters of a logistic regression model for Y_i in which the covariates are a function of \mathbf{X}_i (plus \mathbf{Z}_i , if present). We saw in Section 7.6.1 that for continuous outcomes Y_i it is easy to express the parameters of the linear regression of Y_i on the transformed $\mathbf{X}_i^* = \mathbf{A}\mathbf{X}_i$ and \mathbf{Z}_i in terms of the parameters of regression model given \mathbf{X}_i and \mathbf{Z}_i and their covariance matrices. Unfortunately, this property does not hold in general for non-linear models, including logistic regression. For example, if Y_i follows a logistic regression given X_{i1} and X_{i2} , unless X_{i2} has no independent effect on Y_i , Y_i does not follow a logistic regression given only X_{i1} . However, by using

the same methods as were used to find the model for Y_i given W_i (see Section 4.2), under certain conditions Y_i approximately follows a logistic regression given \mathbf{X}_i^* and \mathbf{Z}_i . We describe these approximate expressions in Section 8.6.1.

One parametric model in which Y_i does follow a logistic regression (exactly) given \mathbf{X}_i^* and \mathbf{Z}_i is when \mathbf{X}_i and \mathbf{Z}_i are jointly normal given Y_i , i.e. the normal discriminant model. In Section 8.6.2 we give expressions for the parameters of the logistic regression of Y_i on \mathbf{X}_i^* and \mathbf{Z}_i in this case.

Following this, we discuss how each of the previously described estimation methods can be used to estimate the parameters of the logistic regression model which has \mathbf{X}_i^* and \mathbf{Z}_i as covariates. For each method, as for linear regression, we treat separately the cases in which Y_i is assumed to be conditionally independent of \mathbf{X}_i given \mathbf{X}_i^* and \mathbf{Z}_i , and when it is not.

8.6.1 Approximate results

To find approximate expressions, we consider the model induced for Y_i given \mathbf{X}_i^* and \mathbf{Z}_i , assuming that Y_i follows a logistic regression given \mathbf{X}_i and \mathbf{Z}_i . Since \mathbf{X}_i^* is a function only of \mathbf{X}_i , Y_i is independent of \mathbf{X}_i^* , conditional on \mathbf{X}_i and \mathbf{Z}_i . Thus, analogous to Section 4.2, we can write:

$$\mathbb{E}(Y_i|\mathbf{X}_i^*, \mathbf{Z}_i) = \mathbb{E}\left(\frac{\exp(\beta_0 + \boldsymbol{\beta}_X^T \mathbf{X}_i + \boldsymbol{\beta}_Z^T \mathbf{Z}_i)}{1 + \exp(\beta_0 + \boldsymbol{\beta}_X^T \mathbf{X}_i + \boldsymbol{\beta}_Z^T \mathbf{Z}_i)} \mid \mathbf{X}_i^*, \mathbf{Z}_i\right)$$

Then using the results of Kuha for the justification of RC for logistic regression [64], and viewing \mathbf{X}_i^* as an imperfect ‘measurement’ of \mathbf{X}_i , we can approximate this by:

$$\mathbb{E}(Y_i|\mathbf{X}_i^*, \mathbf{Z}_i) = \frac{\exp(\beta_0 + \boldsymbol{\beta}_X^T \mathbb{E}(\mathbf{X}_i|\mathbf{X}_i^*, \mathbf{Z}_i) + \boldsymbol{\beta}_Z^T \mathbf{Z}_i)}{1 + \exp(\beta_0 + \boldsymbol{\beta}_X^T \mathbb{E}(\mathbf{X}_i|\mathbf{X}_i^*, \mathbf{Z}_i) + \boldsymbol{\beta}_Z^T \mathbf{Z}_i)}, \quad (8.2)$$

when either $\boldsymbol{\beta}_X^T \text{Var}(\mathbf{X}_i|\mathbf{X}_i^*, \mathbf{Z}_i)\boldsymbol{\beta}_X$ is small or when $P(Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i)$ is small and \mathbf{X}_i given \mathbf{X}_i^* and \mathbf{Z}_i is normal.

Given a specification for $\mathbb{E}(\mathbf{X}_i|\mathbf{X}_i^*, \mathbf{Z}_i)$, we can collect the coefficients \mathbf{X}_i^* and \mathbf{Z}_i in equation (8.2) to find expressions for the parameters $\boldsymbol{\beta}_{X^*}$ and $\boldsymbol{\beta}_{Z^*}$, which we define as those values that would be consistently estimated by fitting a logistic regression of Y_i on \mathbf{X}_i^* and \mathbf{Z}_i .

Suppose for example that we assume \mathbf{X}_i is normally distributed given \mathbf{Z}_i , with $\mathbb{E}(\mathbf{X}_i|\mathbf{Z}_i) = \boldsymbol{\Gamma}_0 + \boldsymbol{\Gamma}_Z \mathbf{Z}_i$ and $\text{Var}(\mathbf{X}_i|\mathbf{Z}_i) = \boldsymbol{\Sigma}_{X|Z}$. It then follows that \mathbf{X}_i^* and \mathbf{X}_i are jointly normal given \mathbf{Z}_i , with $\text{Var}(\mathbf{X}_i^*) = \mathbf{A}\boldsymbol{\Sigma}_{X|Z}\mathbf{A}^T$ and $\text{Cov}(\mathbf{X}_i^*, \mathbf{X}_i) = \mathbf{A}\boldsymbol{\Sigma}_{X|Z}$.

Then it follows that:

$$\begin{aligned}
\mathbb{E}(\mathbf{X}_i|\mathbf{X}_i^*, \mathbf{Z}_i) &= \mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z \mathbf{Z}_i + \mathbf{\Sigma}_{X|Z} \mathbf{A}^T (\mathbf{A} \mathbf{\Sigma}_{X|Z} \mathbf{A}^T)^{-1} (\mathbf{X}_i^* - \mathbf{A}(\mathbf{\Gamma}_0 + \mathbf{\Gamma}_Z \mathbf{Z}_i)) \\
&= \mathbf{\Gamma}_0 - \mathbf{\Sigma}_{X|Z} \mathbf{A}^T (\mathbf{A} \mathbf{\Sigma}_{X|Z} \mathbf{A}^T)^{-1} \mathbf{A} \mathbf{\Gamma}_0 \\
&\quad + (\mathbf{\Gamma}_Z - \mathbf{\Sigma}_{X|Z} \mathbf{A}^T (\mathbf{A} \mathbf{\Sigma}_{X|Z} \mathbf{A}^T)^{-1} \mathbf{A} \mathbf{\Gamma}_Z) \mathbf{Z}_i \\
&\quad + \mathbf{\Sigma}_{X|Z} \mathbf{A}^T (\mathbf{A} \mathbf{\Sigma}_{X|Z} \mathbf{A}^T)^{-1} \mathbf{X}_i^*.
\end{aligned}$$

Collecting together the coefficients of \mathbf{X}_i^* and \mathbf{Z}_i it follows that:

$$\boldsymbol{\beta}_{X^*} = (\mathbf{A} \mathbf{\Sigma}_{X|Z} \mathbf{A}^T)^{-1} \mathbf{A} \mathbf{\Sigma}_{X|Z} \boldsymbol{\beta}_X \quad (8.3)$$

and that:

$$\boldsymbol{\beta}_{Z^*} = \boldsymbol{\beta}_Z + \mathbf{\Gamma}_Z^T (\boldsymbol{\beta}_X - \mathbf{A}^T \boldsymbol{\beta}_{X^*}). \quad (8.4)$$

We note that these expressions for $\boldsymbol{\beta}_{X^*}$ and $\boldsymbol{\beta}_{Z^*}$ are identical to those found in Section 7.6 for linear regression. However, whereas for linear regression the expressions are exact, the expressions are only approximate for logistic regression, under the previously described conditions.

As in Section 7.6.1, in the special case in which there are no \mathbf{Z}_i , \mathbf{X}_i is two-dimensional, and we omit X_{i2} by specifying $\mathbf{A} = (1, 0)$, this gives:

$$\beta_{X^*} = \beta_{X_1} + \frac{\beta_{X_2} \text{Cov}(X_{i1}, X_{i2})}{\text{Var}(X_{i1})}. \quad (8.5)$$

8.6.2 Normal discriminant model

One case where the induced model for Y_i given \mathbf{X}_i^* and \mathbf{Z}_i does follow a logistic regression (exactly, rather than approximately) is when \mathbf{X}_i and \mathbf{Z}_i are multivariate normal given Y_i , i.e. the normal discriminant model. Thus, as in Section 4.6.1 we assume that \mathbf{X}_i and \mathbf{Z}_i are multivariate normal given Y_i :

$$\begin{pmatrix} \mathbf{X}_i \\ \mathbf{Z}_i \end{pmatrix} \sim N \left(\begin{pmatrix} \gamma_{X0} + \gamma_{XY} Y_i \\ \gamma_{Z0} + \gamma_{ZY} Y_i \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{X|Y} & \mathbf{\Sigma}_{XZ|Y} \\ \mathbf{\Sigma}_{ZX|Y} & \mathbf{\Sigma}_{Z|Y} \end{pmatrix} \right)$$

This implies that Y_i follows a logistic regression given \mathbf{X}_i and \mathbf{Z}_i , with log odds ratios:

$$\begin{pmatrix} \boldsymbol{\beta}_X \\ \boldsymbol{\beta}_Z \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{X|Y} & \mathbf{\Sigma}_{XZ|Y} \\ \mathbf{\Sigma}_{ZX|Y} & \mathbf{\Sigma}_{Z|Y} \end{pmatrix}^{-1} \begin{pmatrix} \gamma_{XY} \\ \gamma_{ZY} \end{pmatrix} \quad (8.6)$$

Then since \mathbf{X}_i^* is jointly normally distributed with \mathbf{Z}_i , conditional on Y_i , Y_i follows a logistic regression model given \mathbf{X}_i^* and \mathbf{Z}_i . This is true irrespective of whether $\mathbb{E}(Y_i|\mathbf{X}_i, \mathbf{Z}_i) = \mathbb{E}(Y_i|\mathbf{X}_i^*, \mathbf{Z}_i)$. That is, even if \mathbf{X}_i is informative about Y_i , conditional

on \mathbf{X}_i^* and \mathbf{Z}_i , Y_i still follows a logistic regression given \mathbf{X}_i^* and \mathbf{Z}_i . The reduced covariate \mathbf{X}_i^* has conditional mean function:

$$\begin{aligned}\mathbb{E}(\mathbf{X}_i^*|Y_i) &= \mathbb{E}(\mathbf{A}\mathbf{X}_i|Y_i) \\ &= \mathbf{A}\mathbb{E}(\mathbf{X}_i|Y_i) \\ &= \mathbf{A}(\boldsymbol{\gamma}_{X0} + \boldsymbol{\gamma}_{XY}Y_i),\end{aligned}$$

and conditional covariance matrix:

$$\begin{aligned}\text{Var}(\mathbf{X}_i^*|Y_i) &= \text{Var}(\mathbf{A}\mathbf{X}_i|Y_i) \\ &= \mathbf{A}\text{Var}(\mathbf{X}_i|Y_i)\mathbf{A}^T \\ &= \mathbf{A}\boldsymbol{\Sigma}_{X|Y}\mathbf{A}^T.\end{aligned}$$

Its covariance with \mathbf{Z}_i given Y_i is equal to:

$$\begin{aligned}\text{Cov}(\mathbf{X}_i^*, \mathbf{Z}_i|Y_i) &= \text{Cov}(\mathbf{A}\mathbf{X}_i, \mathbf{Z}_i|Y_i) \\ &= \mathbf{A}\text{Cov}(\mathbf{X}_i, \mathbf{Z}_i|Y_i) \\ &= \mathbf{A}\boldsymbol{\Sigma}_{XZ|Y}.\end{aligned}$$

Thus Y_i follows a logistic regression given \mathbf{X}_i^* and \mathbf{Z}_i , with corresponding log odds ratio vectors $\boldsymbol{\beta}_{X^*}$ and $\boldsymbol{\beta}_Z$ given by:

$$\begin{pmatrix} \boldsymbol{\beta}_{X^*} \\ \boldsymbol{\beta}_{Z^*} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{X^*|Y} & \boldsymbol{\Sigma}_{X^*Z|Y} \\ \boldsymbol{\Sigma}_{ZX^*|Y} & \boldsymbol{\Sigma}_{Z|Y} \end{pmatrix}^{-1} \begin{pmatrix} \boldsymbol{\gamma}_{X^*Y} \\ \boldsymbol{\gamma}_{ZY} \end{pmatrix} \quad (8.7)$$

where:

$$\begin{aligned}\boldsymbol{\gamma}_{X^*Y} &= \mathbb{E}(\mathbf{X}_i^*|Y_i = 1) - \mathbb{E}(\mathbf{X}_i^*|Y_i = 0) \\ &= \mathbf{A}\boldsymbol{\gamma}_{XY}.\end{aligned}$$

Then using the derived mean and covariance functions for \mathbf{X}_i^* we have that:

$$\begin{pmatrix} \boldsymbol{\beta}_{X^*} \\ \boldsymbol{\beta}_{Z^*} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\boldsymbol{\Sigma}_{X|Y}\mathbf{A}^T & \mathbf{A}\boldsymbol{\Sigma}_{XZ|Y} \\ \boldsymbol{\Sigma}_{ZX|Y}\mathbf{A}^T & \boldsymbol{\Sigma}_{Z|Y} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{A}\boldsymbol{\gamma}_{XY} \\ \boldsymbol{\gamma}_{ZY} \end{pmatrix}. \quad (8.8)$$

In the special case in which there are no \mathbf{Z}_i , \mathbf{X}_i is two-dimensional, and we omit X_{i2} by specifying $\mathbf{A} = (1, 0)$, this gives:

$$\beta_{X^*} = \beta_{X_1} + \frac{\beta_{X_2} \text{Cov}(X_{i1}, X_{i2}|Y_i)}{\text{Var}(X_{i1}|Y_i)}. \quad (8.9)$$

We note that this expression is identical to that in equation (8.5), except that the covariance and variance here are conditional on Y_i . Whenever $P(Y_i = 1)$ is small, or

the effects of X_{i1} and X_{i2} are small, as previously discussed, the difference between the marginal and conditional covariances will be small, and so in these cases the two expressions are equivalent.

8.6.3 Regression calibration

Assuming conditional independence

Since the implementation of RC is the same for linear and logistic regression, we can follow the same procedure as described for continuous outcomes in Section 7.6.2. This involves fitting the logistic regression with $\hat{\mathbb{E}}(\mathbf{X}_i^* | \mathbf{W}_i, \mathbf{Z}_i)$ and \mathbf{Z}_i as covariates. In addition to the fact that RC is only approximately consistent for logistic regression, as for linear regression, the resulting estimates are expected to be biased estimates of β_{X^*} and β_{Z^*} if Y_i is not conditionally independent of \mathbf{X}_i , given \mathbf{X}_i^* and \mathbf{Z}_i . We again refer to this approach by the term ‘naive RC’.

Not assuming conditional independence

If we do not wish to make this conditional independence assumption, as for linear regression, we can use equations (8.3) and (8.4) to estimate β_{X^*} and β_{Z^*} , substituting the RC estimates of β_X and β_Z and the estimates of $\Sigma_{X|Z}$ and Γ_Z which are obtained from fitting the linear mixed model to \mathbf{W}_i . We again refer to this by the term ‘corrected RC’.

8.6.4 Maximum likelihood

Assuming conditional independence

Finding the MLEs of the joint model in which conditional independence is assumed between Y_i and \mathbf{X}_i (conditional on \mathbf{X}_i^* and \mathbf{Z}_i) requires two minor modifications of the MCEM algorithm. First, when multiply imputing \mathbf{X}_i using rejection sampling, the acceptance probability is equal to $P(Y_i = 1 | \mathbf{X}_i^*, \mathbf{Z}_i)$, rather than $P(Y_i = 1 | \mathbf{X}_i, \mathbf{Z}_i)$. Second, in the M-step, rather than fitting logistic regressions for Y_i with \mathbf{X}_i and \mathbf{Z}_i as covariates, we use the imputations of \mathbf{X}_i^* (found by pre-multiplying the imputations of \mathbf{X}_i by \mathbf{A}) and \mathbf{Z}_i as covariates.

As described in Section 8.3, our novel approach to ML estimation based on fitting a mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i can be used when \mathbf{X}_i and \mathbf{Z}_i are jointly normal given Y_i . However, as for linear regression, we do not believe this approach can be used to incorporate an assumption of independence between Y_i and \mathbf{X}_i conditional on \mathbf{X}_i^* and \mathbf{Z}_i . To see why, we consider the implications of the assumption for a particularly simple case. Thus suppose that $\mathbf{X}_i = (X_{i1}, X_{i2})^T$, that there is no \mathbf{Z}_i , and that the transformation $\mathbf{A} = (1, 0)$ so that \mathbf{X}_i^* is equal to X_{i1} . Then the conditional independence assumption means that $\beta_{X_2} = 0$. Then using

equation (8.6), after some simple algebra, this implies that:

$$\gamma_{X_2Y} = \frac{\text{Cov}(X_{i1}, X_{i2})\gamma_{X_1Y}}{\text{Var}(X_{i1})}.$$

Thus in this case the conditional independence assumption implies a certain constraint between the fixed effects and random effects parameters of the mixed model for \mathbf{W}_i given Y_i . Such constraints cannot be specified in the standard linear mixed model commands of statistical software packages, and thus we do not believe that our approach can be used to find the MLEs for the model which makes the assumption of conditional independence between Y_i and \mathbf{X}_i given \mathbf{X}_i^* and \mathbf{Z}_i .

Not assuming conditional independence

In general, if we do not assume conditional independence between Y_i and \mathbf{X}_i (conditional on \mathbf{X}_i^* and \mathbf{Z}_i), then as previously described, Y_i does not follow a logistic regression given \mathbf{X}_i^* and \mathbf{Z}_i . In this case, β_{X^*} and β_{Z^*} are not parameters of a correctly specified parametric model. However, if as before, we define β_{X^*} and β_{Z^*} as the vectors which would be estimated consistently by fitting a logistic regression model to Y_i with \mathbf{X}_i^* and \mathbf{Z}_i as covariates, we can obtain approximately consistent estimates by using equations (8.3) and (8.4), substituting the ML estimates of the required parameters. Alternatively, we can create multiple imputations of \mathbf{X}_i after finding the full model's MLEs via MCEM (see Section 8.6.5 below).

As previously described, if \mathbf{X}_i and \mathbf{Z}_i are multivariate normal given Y_i , it follows that $\mathbf{X}^* = \mathbf{A}\mathbf{X}_i$ and \mathbf{Z}_i are also multivariate normal given Y_i , with corresponding logistic regression log odds ratios β_{X^*} and β_{Z^*} as given by equation (8.8). The MLEs of β_{X^*} and β_{Z^*} can thus be calculated by inserting the MLEs (for the joint model which assumes conditional normality of \mathbf{X}_i and \mathbf{Z}_i , which can be obtained using our approach of fitting a mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i) of the required parameters into this equation.

8.6.5 Multiple imputation

Not assuming conditional independence

As discussed previously, a by-product of fitting the full model which assumes marginal normality for \mathbf{X}_i via MCEM is a large set of multiple imputations, created assuming that Y_i may depend jointly on all of the components of \mathbf{X}_i . Multiple imputations of \mathbf{X}_i^* can then be generated by pre-multiplying imputations of \mathbf{X}_i by the transformation matrix \mathbf{A} . Logistic regression models can then be fitted to Y_i , with the imputations of \mathbf{X}_i^* and \mathbf{Z}_i as covariates, and the resulting estimates of β_{X^*} and β_{Z^*} averaged over the imputations. Assuming the parametric assumptions hold, this

will give consistent estimates of the values that would be consistently estimated if we were able to observe \mathbf{X}_i^* directly, without error.

As described in Section 8.4, if we are willing to assume \mathbf{X}_i is normally distributed given Y_i and \mathbf{Z}_i , \mathbf{X}_i can be multiply imputed after fitting the linear mixed model for \mathbf{W}_i given Y_i and \mathbf{Z}_i described in Section 8.3. Multiple imputations of \mathbf{X}_i^* can then be generated as previously described. Providing the assumption of conditional normality for \mathbf{X}_i given Y_i and \mathbf{Z}_i holds, we believe the resulting estimates are consistent for the vectors β_{X^*} and β_{Z^*} which would be consistently estimated if we were able to fit the logistic regression for Y_i with \mathbf{X}_i^* and \mathbf{Z}_i as covariates. In Section 8.7 we report the results of simulations to investigate this.

8.7 Simulations

In this section we report the results of simulations to examine the performance of the previously described methods.

8.7.1 Simulation setup

Apart from specification of the outcome model and the minimum number of error-prone measurements, we used the same simulation setup as for linear regression, as described in Section 7.7, and which we recall here for completeness. We simulated the random-intercepts and slopes $\mathbf{X}_i = (X_{i1}, X_{i2})^T$ for $n = 1,000$ subjects from a multivariate normal distribution with mean $(0, 0)^T$ and covariance matrix:

$$\begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

For each subject, their number of error-prone measurements, n_i , was generated as a random draw from the discrete uniform distribution, taking values between 2 and 20 (inclusive). The minimum of two error-prone measurements was chosen so that the identifiability requirement for the CS method would be satisfied for each each subject. For the j th measurement of subject i , we simulated the ‘time’ t_{ij} that the measurement took place from the continuous uniform distribution, taking values between 0 and 1. We then generated W_{ij} according to:

$$W_{ij} = X_{i1} + t_{ij}X_{i2} + U_{ij}$$

where the measurement errors U_{ij} were simulated independently from a normal distribution with mean zero and variance one.

Lastly, we generated a binary outcome Y_i according to the logistic regression model:

$$P(Y_i = 1|\mathbf{X}_i) = \frac{\exp(\beta_{X_1}X_{i1} + \beta_{X_2}X_{i2})}{1 + \exp(\beta_{X_1}X_{i1} + \beta_{X_2}X_{i2})}.$$

Since X_{i1} and X_{i2} had mean zero, this meant that $P(Y_i = 1) = 0.5$. We simulated data under four scenarios: scenario 1: $\beta_{X_1} = 0.1$, $\beta_{X_2} = 0.1$, scenario 2: $\beta_{X_1} = 1$, $\beta_{X_2} = 1$, scenario 3: $\beta_{X_1} = 0.1$, $\beta_{X_2} = 0$, scenario 4: $\beta_{X_1} = 1$, $\beta_{X_2} = 0$. We performed 10,000 simulations for each scenario.

8.7.2 Estimation methods

Ideal

Since closed form expressions do not exist for some of the unadjusted parameters which we consider, we report the mean (SD) estimates of the ‘ideal’ method when considering the unadjusted effects of X_{i1} or X_{i2} . By this, we mean that we fitted the logistic regression model using either X_{i1} or X_{i2} as covariate. In reality it is of course not possible to use this estimator, as \mathbf{X}_i is unobserved, or latent. However, we use the mean of the ideal estimator over simulated datasets to give an estimate of the values which would be estimated consistently if it were possible to observe \mathbf{X}_i directly.

Regression calibration (RC)

To estimate the adjusted effects of X_{i1} and X_{i2} on Y_i using RC, we fitted a random intercepts and slopes model to the error-prone measurements \mathbf{W}_i using ML via the R command `lmer`. We then found the BLUPs of X_{i1} and X_{i2} using the `ranef()` command and then fitted the logistic regression model for Y_i with these BLUPs as covariates to estimate β_{X_1} and β_{X_2} .

To estimate the unadjusted effects of X_{i1} and X_{i2} on Y_i (implicitly under the assumption of conditional independence), we first regressed Y_i on the BLUPs of X_{i1} and then regressed Y_i on the BLUPs of X_{i2} . As described in Section 8.6.3, this approach gives approximately consistent estimates if Y_i is conditionally independent of \mathbf{X}_i given \mathbf{X}_i^* . Here, this means that Y_i must be independent of the omitted component of \mathbf{X}_i conditional on the included component. We refer to these estimates by the term ‘naive RC’.

We also estimated the unadjusted effects of X_{i1} and X_{i2} on Y_i without making the previously described conditional independence assumptions. To do this, we substituted the RC estimates of β_{X_1} and β_{X_2} , the estimate of $\text{Cov}(X_{i1}, X_{i2})$, and of $\text{Var}(X_{i1})$ (or $\text{Var}(X_{i2})$) from fitting the mixed model for \mathbf{W}_i , into equation (8.5). We refer to these estimates by the term ‘corrected RC’.

Maximum likelihood assuming marginal normality for \mathbf{X}_i (ML1)

To estimate the adjusted effects of X_{i1} and X_{i2} on Y_i , we used ascent-based MCEM, as described in Section 8.2.2. We used the same convergence criteria and control parameters for ascent-based MCEM as in the simulations for classical measurement error, as described in Section 4.10.

To estimate the effect of X_{i1} (or X_{i2}) under an assumption of independence for Y_i and X_{i2} (respectively X_{i1}) given X_{i1} (respectively X_{i2}), we modified the MCEM algorithm as described in Section 8.6.4.

Maximum likelihood assuming marginal normality + multiple imputation (ML1+MI)

Following our proposal in Section 8.6.5, we multiply imputed X_{i1} and X_{i2} , from the model fitted using MCEM (which assumes Y_i depends on both X_{i1} and X_{i2}). We then fitted logistic regression models for Y_i , with either the imputations of X_{i1} or X_{i2} as covariate.

Maximum likelihood assuming normality for \mathbf{X}_i given Y_i (ML2)

We found the MLEs of β_{X_1} and β_{X_2} for the model which assumes conditional normality for \mathbf{X}_i given Y_i , by fitting the linear mixed model to \mathbf{W}_i , conditional on Y_i as described in Section 8.3.

We also used these estimates to calculate the MLEs (under the conditional normal model) of the unadjusted effects of X_{i1} and X_{i2} . To do this, we substituted the ML estimates of the adjusted effects of X_{i1} and X_{i2} and the estimate of $\text{Cov}(X_{i1}, X_{i2}|Y_i)$ from fitting the mixed model for \mathbf{W}_i into equation (8.9).

Conditional score

We used the CS method to estimate the adjusted effects of X_{i1} and X_{i2} on Y_i . We used the estimate of σ_U^2 obtained in the first stage of RC from fitting the linear mixed model to \mathbf{W}_i , and treated this as fixed in the CS estimating equations. We used the RC estimates of β_0 and β_{X_1} and β_{X_2} as starting values, and used the `nleqslv` function in R to solve the estimating equations. For the derivatives, we used those described by in the Web Appendix B to Li *et al* [117]. These expressions contain an error (verified by personal communication with Li). In the expressions for the derivatives, wherever the matrix $(\mathbf{D}_i \boldsymbol{\Sigma}_i \mathbf{D}_i)^{-1}$ appears it should be replaced by $(\mathbf{D}_i \boldsymbol{\Sigma}_i^{-1} \mathbf{D}_i)^{-1}$. As before, if no solution to the estimating equations was found, we used the RC estimates for the purposes of reporting mean and SD.

8.7.3 Simulation results

Adjusted effect estimates

Table 8.1 shows the mean (SD) of estimates of the adjusted effects of X_{i1} and X_{i2} on Y_i , using RC, ML assuming marginal normality of \mathbf{X}_i (ML1), ML assuming conditional normality of \mathbf{X}_i given Y_i (ML2) and CS.

For scenarios 1 and 3, there was little or no suggestion of bias in the RC estimates of β_{X_1} and β_{X_2} . For scenarios 2 and 4, in which the effects were larger, RC showed downward bias in the estimates of β_{X_1} .

The ML1 estimates, based on marginally normality for \mathbf{X}_i , showed some upward bias in the estimates of the adjusted effect of X_{i1} in scenario 4 and in the adjusted effect of X_{i2} in scenario 2, but otherwise had negligible bias. The ML1 estimates were more variable than the ML2 estimates, despite ML1 being the correctly specified MLE for the data-generating process. We suspect this may be due to Monte-Carlo error inherent in the MCEM algorithm.

The MLEs based on assuming normality for \mathbf{X}_i given Y_i (labeled as ML2), which is misspecified for the data-generating model, also showed little suggestion of bias for scenarios 1 and 3. Furthermore, the estimates of β_{X_1} in scenarios 2 and 4 were less biased than the RC estimates. This is in agreement with our simulation results of Section 4.10.

For scenarios 1 to 4 respectively, no solution was found for the CS estimating equations in 35.3%, 26.8%, 34.2% and 21% of the 10,000 simulations. The mean (SD) values shown in Table 8.1 for CS thus are based on estimates which approximately 25% of the time are the RC estimates. To ensure that our implementation of the CS estimating equations was correct, we ran simulations under the setup used by Li *et al* [118]. Under the setup of Li *et al* [118] a solution was found in virtually all simulations, and our results were similar to those of Li *et al* [118]. In the simulation setup of Li *et al* [118] most subjects have $n_i = 5$ longitudinal measurements, whereas in our setup subjects have between 2 and 20 measurements. We found that if we fixed n_i to be the same for all subjects a solution to the CS estimating equations was found in all simulations. This suggests that finding a solution to the CS estimating equations may be problematic in cases where the number of error-prone measurements differs widely between subjects.

Unadjusted effect estimates assuming conditional independence

Table 8.2 shows the results for the unadjusted effect estimates of X_{i1} and X_{i2} using the true \mathbf{X}_i values (ideal), and naive RC and ML1 (which assumes marginal normality). The naive RC and ML1 estimators here assume (respectively, implicitly and explicitly) that Y_i is independent of the omitted covariate given the included covariate. Comparing the naive RC and ML1 estimates with the ideal estimates, as

Table 8.1: Adjusted effect estimates for logistic regression simulations. Mean (SD) of estimates found using regression calibration (RC), maximum likelihood assuming marginal normality for \mathbf{X}_i using MCEM (ML1), maximum likelihood assuming normality for \mathbf{X}_i given Y_i (ML2), and the conditional score method (CS).

Scenario	β_{X_1}	β_{X_2}	$\hat{\beta}_{X_1}$			
			RC	ML1	ML2	CS
1	0.1	0.1	0.097 (0.130)	0.097 (0.133)	0.098 (0.131)	0.098 (0.480)
2	1	1	0.946 (0.171)	0.999 (0.171)	0.973 (0.164)	0.996 (0.643)
3	0.1	0	0.098 (0.132)	0.099 (0.136)	0.098 (0.133)	0.097 (0.345)
4	1	0	0.985 (0.156)	1.031 (0.184)	1.019 (0.178)	1.025 (0.215)

Scenario	β_{X_1}	β_{X_2}	$\hat{\beta}_{X_2}$			
			RC	ML1	ML2	CS
1	0.1	0.1	0.104 (0.149)	0.106 (0.154)	0.105 (0.151)	0.108 (0.281)
2	1	1	0.973 (0.191)	1.036 (0.224)	0.996 (0.209)	1.015 (0.682)
3	0.1	0	0.003 (0.151)	0.002 (0.156)	0.002 (0.152)	0.000 (0.551)
4	1	0	-0.011 (0.172)	-0.020 (0.183)	-0.013 (0.179)	-0.020 (0.256)

expected, in scenarios 1 and 2, in which X_{i1} and X_{i2} both have independent effects on Y_i , both methods give biased estimates of the unadjusted effects. In scenarios 3 and 4, the naive RC and ML1 estimates of the unadjusted effect of X_{i2} have little bias. This is because for these scenarios X_{i2} has no independent effect on Y_i . The estimates of the unadjusted effect of X_{i2} are biased because the corresponding conditional assumption no longer holds – X_{i1} has an independent effect on Y_i , conditional on X_{i2} .

Unadjusted effect estimates without assuming conditional independence

Table 8.3 shows the results for the unadjusted effect estimates of X_{i1} and X_{i2} , but which are found without assuming that Y_i is necessarily independent of the omitted covariate, given the included covariate. For scenario 1, the corrected RC estimates of the unadjusted effects, obtained by substituting the RC estimates of the adjusted effects into equation (8.5), showed little bias relative to the ideal estimates. In contrast, in scenario 2, the corrected RC estimates were biased upwards relative to the ideal estimates, due to the fact that the effects of X_{i1} and X_{i2} were larger, thus making the assumption used to justify equation (8.5) less tenable. If the RC estimates of β_{X_1} and β_{X_2} were unbiased, we would expect the estimator based on equation (8.5) to have expectation 1.5. However, the downward bias in the RC estimates of β_{X_1} and β_{X_2} act to reduce the bias in the estimate of the unadjusted effect of X_{i1} . In scenario 3, the corrected RC estimate of the effect of X_{i1} had little bias, but the estimate of the unadjusted effect of X_{i2} was biased upwards slightly. In scenario 4, the estimated effect of X_{i1} was biased downwards slightly. This can be explained by consideration of equation (8.5). The RC estimator of β_{X_2} is expected

Table 8.2: Unadjusted effect estimates for logistic regression simulations assuming conditional independence. Mean (SD) of estimates found using X_{i1} and X_{i2} (ideal), naive regression calibration (naive RC), and maximum likelihood assuming marginal normality of \mathbf{X}_i (ML1).

		Unadjusted effect of X_{i1}				
Scenario	β_{X_1}	β_{X_2}	Ideal	Naive RC	ML1	
1	0.1	0.1	0.150 (0.064)	0.173 (0.070)	0.174 (0.070)	
2	1	1	1.322 (0.097)	1.597 (0.124)	1.760 (0.153)	
3	0.1	0	0.099 (0.063)	0.099 (0.069)	0.100 (0.069)	
4	1	0	1.005 (0.084)	0.974 (0.089)	1.008 (0.098)	

		Unadjusted effect of X_{i2}				
Scenario	β_{X_1}	β_{X_2}	Ideal	Naive RC	ML1	
1	0.1	0.1	0.151 (0.064)	0.197 (0.081)	0.199 (0.083)	
2	1	1	1.320 (0.097)	1.800 (0.166)	2.422 (0.286)	
3	0.1	0	0.050 (0.063)	0.096 (0.079)	0.097 (0.080)	
4	1	0	0.434 (0.069)	0.887 (0.115)	1.050 (0.145)	

to have mean zero here, and so the expected value of the corrected RC estimate of the unadjusted effect of X_{i1} reduces to the expected value of the RC estimate of β_{X_1} , which from Table 8.2, is biased downwards. For the unadjusted effect of X_{i2} , from equation (8.5) it follows that the corrected RC approach would give unbiased estimates of 0.5, providing the RC estimates of β_{X_1} and β_{X_2} were unbiased. Their downward bias then causes the corrected RC estimates of the effect of X_{i2} to be less than 0.5.

The estimates based on MI from the fitted model using ML1 showed little bias compared to the ideal estimator, except for scenario 2, in which estimates were biased upwards slightly.

The ML2 estimates, based on an assumption of normality for \mathbf{X}_i given Y_i , showed little bias for both effects across all four scenarios.

Lastly, we compare the efficiency of the estimates of the unadjusted effect of X_{i1} , between when an assumption of conditional independence is made between Y_i and X_{i2} , given X_{i1} (Table 8.2), and when this assumption is relaxed (Table 8.3). For the RC estimates, we see that the standard deviation of the estimates in scenario 3 increases from 0.069 to 0.076, and in scenario 4 increases from 0.090 to 0.098. The loss in efficiency is the price to be paid for relaxing the conditional independence assumption, although at least under the simulation setup considered, this price is not particularly large.

Table 8.3: Unadjusted effect estimates for logistic regression simulations not assuming conditional independence. Mean (SD) estimates found using X_{i1} or X_{i2} (ideal), corrected regression calibration (corrected RC), maximum likelihood assuming marginal normality for \mathbf{X}_i + multiple imputation (ML1+MI), and maximum likelihood assuming normality for \mathbf{X}_i given Y_i (ML2).

Scenario	β_{X_1}	β_{X_2}	Unadjusted effect of X_{i1}			
			Ideal	Corrected RC	ML1+MI	ML2
1	0.1	0.1	0.150 (0.064)	0.150 (0.077)	0.150 (0.077)	0.150 (0.077)
2	1	1	1.322 (0.097)	1.437 (0.126)	1.333 (0.137)	1.313 (0.130)
3	0.1	0	0.099 (0.063)	0.099 (0.077)	0.099 (0.078)	0.099 (0.077)
4	1	0	1.005 (0.084)	0.977 (0.097)	1.013 (0.113)	1.006 (0.110)
Scenario	β_{X_1}	β_{X_2}	Unadjusted effect of X_{i2}			
			Ideal	Corrected RC	ML1+MI	ML2
1	0.1	0.1	0.151 (0.064)	0.153 (0.099)	0.155 (0.101)	0.154 (0.100)
2	1	1	1.320 (0.097)	1.450 (0.168)	1.359 (0.213)	1.331 (0.200)
3	0.1	0	0.050 (0.063)	0.052 (0.099)	0.052 (0.100)	0.052 (0.099)
4	1	0	0.434 (0.069)	0.491 (0.120)	0.440 (0.114)	0.440 (0.113)

8.8 Conclusions

As we have seen, the extension of estimation methods to the case of longitudinal error-prone measurements is also relatively simple for binary outcomes. We conclude with some comments on the various methods when interest lies in estimating the parameters of the logistic regression model for Y_i given \mathbf{X}_i and \mathbf{Z}_i , and then we discuss the issue of fitting models with different functions of the longitudinal process(es) as covariate.

8.8.1 Regression calibration

In our simulations RC performed well in estimating β_X , with biases of the same order of magnitude as in our simulations for classical measurement error. While some authors (e.g. Li *et al* [118]) report that RC has large biases for logistic regression, even when the normality assumptions for \mathbf{X}_i hold, this only occurs when the effects of \mathbf{X}_i are quite large. In the simulations by Li *et al*, the standardized log odds ratio for X_{i1} was 3, corresponding to an odds ratio of around 20, which may not be typical of most epidemiological studies.

8.8.2 Maximum likelihood

We have shown that ascent-based MCEM can be used to find MLEs for joint models with longitudinal error-prone measurements and binary outcomes, in which marginal normality is assumed for \mathbf{X}_i . However, our implementation of the method, in R, is

very slow to run. As discussed in Section 4.11, this is due to the large amount of looping required to multiply impute \mathbf{X}_i . With further programming effort however, this could be overcome.

8.8.3 Maximum likelihood using standard linear mixed models and multiple imputation

Our simulation results suggest our novel approach to ML estimation, based on an assumption of conditional normality for \mathbf{X}_i given Y_i , gives estimates with smaller biases than RC, even when \mathbf{X}_i is marginally normal. However, as in Section 4.11, with larger effects of \mathbf{X}_i , we would expect biases to increase. In contrast to the model which assumes marginal normality for \mathbf{X}_i , our approach involves fitting a linear mixed model, for which standard mixed model commands in packages such as R, Stata, and SAS, can be used. It is thus computationally efficient, and may be less biased than RC. In the presence of error-free covariates \mathbf{Z}_i , this fitted model can be used to multiply impute \mathbf{X}_i , from which β_X and β_Z can be estimated.

8.8.4 Conditional score method

The conditional score method offers the potential for computationally efficient estimation, while relaxing distributional assumptions for \mathbf{X}_i . Our simulations suggest however that in certain settings it may be difficult to find solutions to the CS estimating equations. As previously discussed in Section 4.11, this is a numerical problem which could potentially be overcome by using more sophisticated root-finding algorithms. Our investigations suggest however that successful implementation of the method is not quite as easy as some methodological papers imply. In particular, our results suggest that solving the estimating equations becomes problematic when the number of longitudinal error-prone measurements differs substantially between subjects.

8.8.5 Alternative outcome model covariate specifications

Our simulation results confirm that naive application of RC gives biased estimates of the parameters of a logistic regression model for Y_i given $\mathbf{X}_i^* = \mathbf{A}\mathbf{X}_i$ and \mathbf{Z}_i if Y_i and \mathbf{X}_i are not conditionally independent given \mathbf{X}_i^* and \mathbf{Z}_i . Specifically, estimates of the unadjusted effects, obtained by omitting the predicted values of some components of \mathbf{X}_i from the logistic regression model for Y_i , are biased estimates of those values which would be obtained by fitting the model with \mathbf{X}_i^* as covariate. Our simulation results suggest that the expressions given in Section 8.6.1 can be used, together with the RC estimates of β_X and β_Z , to give approximately consistent estimates of β_{X^*} and β_{Z^*} . However, these expressions may not be valid when the effects of \mathbf{X}_i are large.

These expressions can of course also be used with estimates of β_X and β_Z obtained by ML. However, in this case, we can use the fitted model to multiply impute the \mathbf{X}_i from its conditional distribution, and estimate β_{X^*} and β_{Z^*} using the imputed values of \mathbf{X}_i^* . For the model which assumes marginal normality \mathbf{X}_i , imputations of \mathbf{X}_i are a by-product of the MCEM algorithm. Providing the parametric assumptions are correct, this provides consistent estimates of β_{X^*} and β_{Z^*} , without requiring the small effect size conditions made in justifying equation (8.3).

8.8.6 Inference

Approaches to estimate standard errors and confidence intervals for the resulting estimates are the same as described for continuous outcomes, as discussed in Section 7.8.

Chapter 9

Survival outcomes

Joint modelling of longitudinal and survival (or more generally time-to-event) outcomes has been the subject of a great deal of research over the last 15 years. We first give a brief overview of this field, before summarizing the contents of the chapter. Excellent overviews of the field have been given by Tsiatis and Davidian [124] and more recently, by Diggle *et al* [125].

9.1 Overview

The large research effort into models for longitudinal and survival outcomes is motivated by the fact that many studies generate both longitudinal measurement data and data regarding the time to an event of interest. A particularly popular framework involves specification of a vector of unobserved, subject-specific random-effects, which are responsible for inducing associations between the longitudinal process and the time to the event of interest. Usually, the occurrence of the event of interest precludes any further longitudinal measurements being made, and we will assume this to be the case. Extensions to multiple longitudinal processes are possible, but for simplicity we restrict attention to the case of a single longitudinal process.

Joint modelling of longitudinal and survival (or time to event) outcomes may be used to answer at least two different questions [99]. The first is to describe how the longitudinal process influences the time to the event of interest. This coincides with our developments of the previous two chapters, in which certain aspects of the longitudinal process enter as covariates in the outcome model. Here the outcome model describes the dependency of time to the event of interest on those aspects of the longitudinal process thought to be important. Some of the earliest research into methods for such models was motivated by studies in AIDS, in which interest lay in the relationship between a patient's CD4 count profile over time and their survival [126, 127].

An alternative focus for joint models is when the longitudinal process is itself of primary interest, but the occurrence of the event, such as losing a subject to

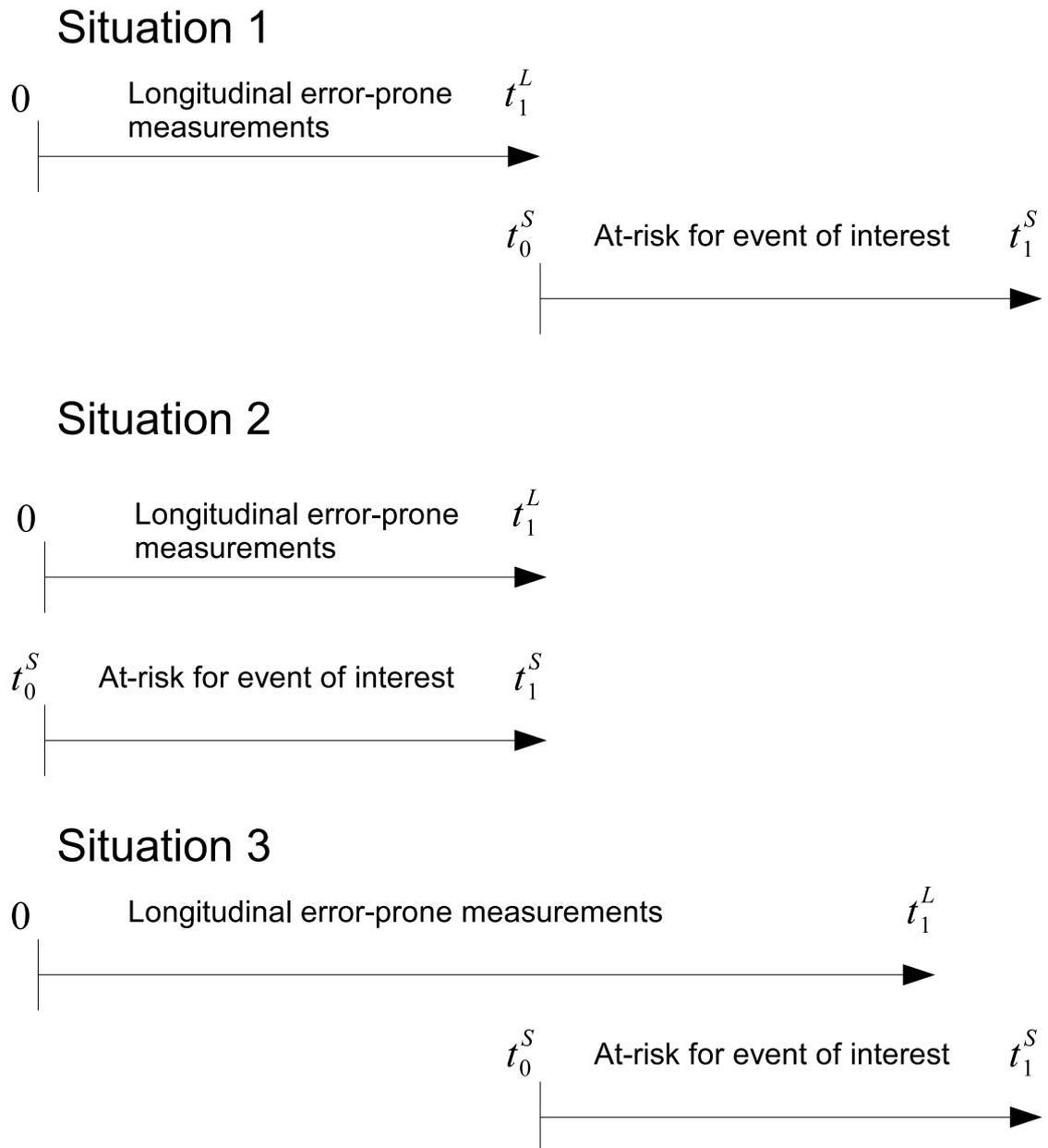
follow-up, or ‘dropping-out’, means that no further longitudinal measurements are made. In this situation, the occurrence of the event causes all subsequent planned longitudinal measurements to be missing. In this case, analysis of the observed longitudinal measurements, for example by fitting a linear mixed model by likelihood methods, and which ignores the drop-out (missing data) process, may result in biased estimates. For example, Henderson *et al* used data from a clinical trial of schizophrenia patients to investigate the trajectory of scores from a measure of psychiatric disorder [99]. As the trial progressed, some patients dropped out from the trial, following which no further scores were observed. Analysis of the longitudinal measurements by likelihood methods, which ignore the drop-out process, are consistent under the MAR assumption. If the data are missing not at random (MNAR), so that conditional on observed longitudinal measurements, the future trajectories differ systematically between subjects who drop-out and those who do not, an analysis which ignores the drop-out process, such as based on a linear mixed model, in general results in biased estimates. One way in which this occurs is if drop-out depends on the unobserved random-effects. The joint models we consider are usually referred to as ‘shared parameter models’ within the missing data literature, the shared parameters being the unobserved random-effects which are assumed to induce the dependency between the longitudinal process and the time-to-event outcome [128].

The particular specification of a joint model obviously depends on the study design and inferential objectives, but Figure 9.1 depicts three common situations, in which the longitudinal and survival, or time-to-event outcome, are observed in different windows of time. We assume that n subjects are recruited into a study at time 0, with longitudinal measurements made periodically from time 0 onwards until either the subject experiences the event of interest, is censored, or until time t_1^L . Subjects are considered at risk for the event of interest from time $t_0^S \geq 0$ until time t_1^S . The three situations depicted in Figure 9.1 correspond to:

1. $t_1^L \leq t_0^S$: longitudinal error-prone measurements are made in a time period which precedes the ‘at-risk’ period, and so there is no overlap between these time periods
2. $t_0^S = 0$ and $t_1^L = t_1^S$: the time period in which longitudinal measurements take place is identical to the ‘at-risk’ period
3. $t_0^S > 0$ and $t_1^L > t_0^S$: longitudinal measurements are made at times preceding the at-risk period, but are also made during the at-risk period

Situation 1 is analogous to the models considered in Chapters 7 and 8 for continuous and binary outcomes. Since the longitudinal process occurs prior to when subjects become at risk of experiencing the event of interest, the random-effects

Figure 9.1: Diagrammatic representation of three situations in which longitudinal and time-to-event data are observed



which characterize the longitudinal process may enter as time-independent covariates in a model for the survival time / time to event. This situation has been considered less frequently, probably because if the hazard function for the event of interest depends on the longitudinal process values preceding the at-risk period, it is also likely to depend on more recent values of the longitudinal process as time proceeds, i.e. situation 2 or 3.

The majority of research into methods for jointly modelling longitudinal and survival or time-to-event data has focused on what we call situation 2 [94, 99, 105]. Here the longitudinal process and time-to-event outcome are observed over the same time period. The relevant aspects of the longitudinal process would then usually enter the model for the time to the event of interest as time-dependent covariates. An example of such a model is that used by Tsiatis *et al* , who fitted models using data from HIV patients in which the hazard of death at time t was assumed to depend on the patient's historical CD4 levels only through its current value at time t [127].

In situation 3, observations of the longitudinal process can occur both prior to the at-risk period and also concurrently with the at-risk period. Of course if subjects are observed for the longitudinal process, they will usually also be observed for the event of interest at the same time. However, when we are interested in relating hazard at time t to the value of the longitudinal process at some time prior to t , we must necessarily consider the at-risk period to start at a time t_0^S , and thus our analysis is conditional on being event free at time t_0^S . An example of this are the analyses by Boshuizen *et al* [119], previously described in Section 6.2. Using data from the Seven Countries Study, Boshuizen *et al* investigated how, given survival to age 65, the hazard for mortality due to coronary heart disease (and in a second analysis, death due to stroke) at ages $t > 65$ was related to systolic blood pressure (SBP) at time t and to SBP at time $t - 25$, i.e. 25 years earlier.

In Section 9.2 we define the modelling assumptions and notation for the chapter. In Section 9.3 we review the application of an RC type approach to joint models for longitudinal and survival outcomes. The investigation of Boshuizen *et al* [119], as previously discussed in Chapters 6, 7, and 8, is typical of many epidemiological studies, in that we usually do not know a priori how the hazard for the event of interest at a particular time t depends on the history of the longitudinal process, e.g. underlying SBP. In Section 9.3 we therefore also discuss the application of RC when interest focuses on estimation of a number of different specifications for the outcome model covariates, and describe the implicit assumptions which are required for (approximately) consistent parameter estimation. In Section 9.4 we review the ML approach to estimation for such joint models. In Section 9.5 we show how the MCEM algorithm, described in the case of time-independent covariates measured with classical error in Section 5.5, can be extended to the current setting,

in which in general the model for survival (or time to the event of interest) has time-dependent covariates which are a function of the longitudinal process. This involves showing how multiple imputations of the subject-specific random-effects can be generated using rejection sampling. In Section 9.6 we show how we can use such multiple imputations to find estimates when a number of different outcome model specifications are of interest, such as in the analyses performed by Boshuizen *et al* , under weaker assumptions than implicitly required for RC. We review the extension of the CS method to the setting of longitudinal and survival outcomes in Section 9.7. In Section 9.8 we report the results of simulations, and in Section 9.9 consider the implications of using such models when the occurrence of the event of interest means the longitudinal process is undefined at times following the event's occurrence. We summarize the chapter in Section 9.10.

9.2 Modelling assumptions and notation

In this section we describe a particular modelling framework for a single longitudinal process measured with error, and the time to survival or an event of interest. In Section 9.10 we discuss extensions to multiple longitudinal processes.

9.2.1 Longitudinal process

In this chapter we shall use a slightly different notation to that previously used, in order to make dependency on time explicit. Thus we now assume that the (single) longitudinal process has true value $M_i(t)$ at time t . The trajectory of $M_i(t)$ over time is assumed to be given by:

$$M_i(t) = \mathbf{g}(t)^T \mathbf{X}_i \tag{9.1}$$

where $\mathbf{g}(t)$ is a known vector-valued function of time, and \mathbf{X}_i is a vector of subject-specific random effects. We let $M_i^H(t) = \{M_i(u) : u \leq t\}$ (H stands for history) denote the set of values of $M_i(t)$ at times up to and including t .

We give two examples for illustrative purposes. First, if \mathbf{X}_i is a two-dimensional random-effects vector, by defining:

$$\mathbf{g}(t) = (1, t)^T, \tag{9.2}$$

we assume that the longitudinal process is a linear function of time, with intercept X_{i1} and slope X_{i2} . As a second example, suppose \mathbf{X}_i is a four-dimensional random-effects vector, representing the underlying average systolic blood pressure level in the first four consecutive decades of a person's life. In this case, we would define

$\mathbf{g}(t)$, with t in years, by:

$$\mathbf{g}(t) = (1(t < 10), 1(10 \geq t < 20), 1(20 \geq t < 30), 1(30 \geq t < 40))^T, \quad (9.3)$$

where $1(\cdot)$ denotes the indicator function which takes value one if its argument is true and zero otherwise.

Rather than observing $M_i(t)$ directly, we measure it periodically, and with error. We assume measurements can be made at any time from time 0 until either the subject experiences the event of interest, is censored, or time t_1^L . As before we assume subject i has error-prone measurements W_{ij} , $j = 1, \dots, n_i$. Measurement W_{ij} is made at time t_{ij} , and is assumed to be an unbiased measurement of $M_i(t_{ij})$:

$$\begin{aligned} W_{ij} &= M_i(t_{ij}) + U_{ij} \\ &= \mathbf{g}(t_{ij})^T \mathbf{X}_i + U_{ij}. \end{aligned} \quad (9.4)$$

We assume that the measurement errors U_{ij} are $N(0, \sigma_U^2)$ and that they are independent of each other. We let $\mathbf{W}_i^H(t) = \{W_{ij} : t_{ij} \leq t\}$ denote the vector of error-prone measurements made at times up to and including time t , and denote by $\mathbf{t}_i^H(t) = \{t_{ij} : t_{ij} \leq t\}$ the corresponding vector of measurement times.

Together equations (9.1) and (9.4) imply that $\mathbf{W}_i = (W_{i1}, \dots, W_{in_i})^T$ follow a linear mixed model. While we could express the model for \mathbf{W}_i in the familiar form $\mathbf{W}_i = \mathbf{D}_i \mathbf{X}_i + \mathbf{U}_i$, this formulation does not make explicit how the mean function of the longitudinal process depends on time, which is important for subsequent developments.

9.2.2 Survival process

As in Chapter 5, for each subject i , we denote by T_i and C_i the time to the event of interest and time to censoring respectively. We observe $V_i = \min(T_i, C_i)$, and a failure indicator $Y_i = 1(T_i \leq C_i)$. In order to accommodate each of the three scenarios previously described, we assume that each subject is at risk from time t_0^S . We are then interested in modelling the hazard for times $t > t_0^S$.

Except in situation 1, it does not make sense for the hazard at time t to depend jointly on the random-effects vector \mathbf{X}_i . This is because (especially for the models we consider later) some of the random-effects may only influence the longitudinal process at later times, and it does not make sense to allow the hazard at time t to depend on future values of the longitudinal process. This can be seen most clearly in the second example given earlier, whereby \mathbf{X}_i is a four-dimensional random-effects vector representing the mean systolic blood pressure levels in the first four decades of a subject's life. If we consider the subject to be at-risk from $t = 0$, it would not make sense to model the hazard in the first decade as depending on mean blood

pressure levels in future decades. To incorporate this assumption, we define a (in general vector valued) time-dependent covariate $\mathbf{X}_i^*(t)$ which is a function of the available longitudinal history at time t , $M_i^H(t)$:

$$\mathbf{X}_i^*(t) = \mathbf{G}(M_i^H(t)), \quad (9.5)$$

where $\mathbf{G}(\cdot)$ is some function of the history at time t . Since the longitudinal process is a deterministic function of the random-effects \mathbf{X}_i , as given in equation (9.1), we may also view \mathbf{X}_i^* as a direct function of time and \mathbf{X}_i :

$$\mathbf{X}_i^*(t) = \mathbf{A}(t, \mathbf{X}_i). \quad (9.6)$$

The notation $\mathbf{X}_i^*(t) = \mathbf{G}(M_i^H(t))$ makes explicit however that the hazard at time t cannot depend on future values of the longitudinal process. Continuing the random-intercepts and slopes example, we might define $X_i^*(t) = M_i(t) = X_{i1} + X_{i2}t$ to be the value of the longitudinal process at time t . In the piece-wise constant model for systolic blood pressure, if we considered subjects to be at risk from $t = 10$ years onwards, we might assume $X_i^*(t) = M_i(t - 10)$, so that hazard depended on the mean systolic blood pressure in the preceding decade.

Given the specification of $\mathbf{X}_i^*(t)$, we assume a Cox proportional hazards model:

$$h(t|M_i^H(t), \mathbf{Z}_i) = h_0(t) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(t) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i), \quad (9.7)$$

where $h_0(t)$ denotes an arbitrary baseline hazard function and \mathbf{Z}_i denotes, as usual, a vector of fully observed error-free covariates. This equation makes explicit an assumption that the hazard at time t only depends on the history of the longitudinal process, $M_i^H(t)$ via the value of $\mathbf{X}_i^*(t)$. This means that $\mathbf{X}_i^*(t)$ captures those aspects of the longitudinal process up to time t which are related to the hazard of the event of interest.

9.2.3 Assumptions

Additional assumptions which are sufficient to give valid parameter estimates, using the observed data, differ somewhat between the various estimation methods. We therefore defer statement of sufficient assumptions to our description of each of the estimation approaches.

9.3 Regression calibration

The idea of RC was extended to the setting of joint models for longitudinal measurements and time to event outcomes by Tsiatis *et al* [127]. Tsiatis *et al* considered what we have described as situation 2 (see Figure 9.1), but the methods

apply in all three situations. In the case of time-independent covariates measured with classical error (Section 5.13), we derived the induced hazard function, given an error-prone measurement of the unobserved covariate. In the current setting, with time-dependent covariates which are a function of a longitudinal process, the ‘observable hazard’, with $\mathbf{X}_i(t)$ measured with error, and in the presence of censoring, is equal to $h(t|\mathbf{W}_i^H(t), \mathbf{t}_i^H(t), \mathbf{Z}_i, V_i \geq t)$ [127]. Then, analogous to the induced function derived in Section 5.13), we can write:

$$\begin{aligned} & h(t|\mathbf{W}_i^H(t), \mathbf{t}_i^H(t), \mathbf{Z}_i, V_i \geq t) \\ &= \mathbb{E}(h(t|\mathbf{W}_i^H(t), \mathbf{t}_i^H(t), M_i^H(t), \mathbf{Z}_i, V_i \geq t)|\mathbf{W}_i^H(t), \mathbf{t}_i^H(t), \mathbf{Z}_i, V_i \geq t). \end{aligned} \quad (9.8)$$

For RC to give valid estimates, we must make the following conditional independence assumption:

$$h(t|\mathbf{W}_i^H(t), \mathbf{t}_i^H(t), M_i^H(t), \mathbf{Z}_i, V_i \geq t) = h(t|M_i^H(t), \mathbf{Z}_i), \quad (9.9)$$

which means that, conditional on the longitudinal process history $M_i^H(t)$ and \mathbf{Z}_i , the hazard at time t does not depend on the observed longitudinal measurements made up to time t , the number or timing of the measurements, and the fact that subject i has not yet been censored. In this case the observable hazard is equal to:

$$\begin{aligned} & \mathbb{E}(h(t|\mathbf{X}_i^*(t), \mathbf{Z}_i)|\mathbf{W}_i^H(t), \mathbf{Z}_i, V_i \geq t) \\ &= h_0(t)\mathbb{E}(\exp(\boldsymbol{\beta}_X^T \mathbf{X}_i^*(t) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i)|\mathbf{W}_i^H(t), \mathbf{Z}_i, V_i \geq t). \end{aligned} \quad (9.10)$$

To enable use of standard routines for fitting Cox proportional hazards models, as in Section 5.3, we can approximate this expectation using the delta method, thus giving:

$$h_0(t) \exp(\boldsymbol{\beta}_X^T \mathbb{E}(\mathbf{X}_i^*(t)|\mathbf{W}_i^H(t), \mathbf{Z}_i, V_i \geq t) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i). \quad (9.11)$$

9.3.1 Risk-set regression calibration

In order to calculate the required expectation, we must make distributional assumptions for the random-effects \mathbf{X}_i . Typically these are assumed to be multivariate normal. To approximate the required expectation, Tsiatis *et al* [127] proposed that, at each event time t , the linear mixed model specified by equations (9.1) and (9.4) is fitted to the longitudinal measurements $\mathbf{W}_i^H(t)$ (i.e. all measurements up to time t), using only data from those subjects who were at risk at time t (thus incorporating the conditioning information that $T_i \geq t$). The BLUPs of the random-effects can then be calculated, from which the required expectation of the covariates $\mathbf{X}_i^*(t)$ can be approximated. This can be viewed as the extension of risk-set RC, described for

time-independent covariates measured with classical error in Section 5.3.2, to the setting of time-dependent covariates which are functions of a longitudinal process.

The risk-set RC approach proposed by Tsiatis *et al* [127] requires that a separate linear mixed model be fitted at each event time, and so is relatively computationally intensive. Its implementation in any given situation may also encounter difficulties, depending on the specification of the linear mixed model for the longitudinal measurements and the availability of measurements. For example, suppose we are in ‘situation 2’, where error-prone measurements of an underlying longitudinal process are made concurrently with the at-risk period. Suppose further that we assume a random-intercepts and slopes model for the longitudinal process, and that we specify that the hazard at time t depends on the longitudinal process via its current value, $X_{i1} + X_{i2}t$. Assume that each subject has an error-prone measurement available at $t = 0$. Now suppose that a subject dies (the event of interest) before any subjects have had a second error-prone measurement. With only a single error-prone measurement available from each subject, it is not possible to fit a random-effects model. In this situation one approach may be to ignore the earliest deaths from the analysis, and only analyse deaths which occur after a time at which the linear mixed model can be fitted. This illustrates that the implementation of risk-set RC will need to take into account the particular modelling assumptions and availability of longitudinal measurements in any given application.

An advantage of risk-set RC over more complex estimation methods is that it can be implemented using standard software for fitting linear mixed models and Cox proportional hazards models with time-dependent covariates. As Tsiatis *et al* noted [127], the approach is only expected to give approximately consistent estimates of β_{X^*} and β_{Z^*} . As for time-independent covariates measured with classical error, risk-set RC involves at least two approximations. First, we approximate the expression in equation (9.10) by equation (9.11). Second, even if the \mathbf{X}_i are normally distributed in the population, once we condition on being event free to a particular time t , the distribution of \mathbf{X}_i in this so far event-free sub-group will no longer be normally distributed.

9.3.2 Simplifications

As previously noted, the risk-set RC approach requires that a separate linear mixed model be fitted at each unique event time. A variety of ad-hoc modifications have been proposed in the literature to reduce the computational demands of this approach.

Dafni and Tsiatis performed simulations in which longitudinal measurements were made in subjects at 8 week intervals [129]. They assumed a random-intercepts and slopes model for the longitudinal process, and assumed that the hazard at time t depended on the current value of the longitudinal process at time t . Rather than

fitting a separate mixed model at each event time, Dafni and Tsiatis refitted the mixed model once for each 8 week interval, using data from those subjects at risk at the start of the interval, and using only their measurements preceding that time. Their simulation results suggested that while there was some bias in the resulting estimates, the bias was much smaller compared to that of a naive method, in which at time t a subject's last measurement of the longitudinal process was used as covariate. Aside from other considerations, the latter naive approach makes no allowance for the measurement error, and so is expected to be biased.

Tsiatis and Davidian considered an approach in which the mixed model is fitted four times, using the available longitudinal data at times corresponding to the 25%, 50%, 75% and 100% percentiles of the ordered failure times [105]. At each event time, the covariate $X_i^*(t)$ was predicted using the mixed model fit corresponding to which quartile the event occurred in. Tsiatis and Davidian reported that they found through simulations that this simplified approach gave estimates which differed negligibly compared to the 'full regression calibration' approach.

Boshuizen *et al* simplified the approach further, by fitting the linear mixed model once, to all subjects' available longitudinal measurements [119]. They reported that the resulting estimates were very similar to those obtained by refitting the mixed model at each risk set. We refer to this simplified version by the term 'RC'.

Recently Ye *et al* have compared the performance of risk-set RC with the simplified version used by Boshuizen *et al*, which involves fitting the linear mixed model once, with simulations (although they used a so called semiparametric stochastic mixed model, rather than a standard linear mixed model, for the longitudinal measurements) [130]. Mirroring the investigations of Xie *et al* [91] in the case of time-independent covariates measured with classical error, Ye *et al* found that the risk-set RC estimator had smaller bias but larger variability than the non-risk-set RC estimator.

If the failure rate is low, as in the case of time-independent covariates measured with classical error, we might expect that ignoring the conditioning information $V_i \geq t$ is reasonable. However, note that in these simplified versions of RC, at any time t , future longitudinal measurements are used to predict $\mathbf{X}_i^*(t)$, contrary to equation (9.11). It is not clear how use of this use of information from the future may affect the resulting estimates. We investigate this simplified version of RC, as well as 'risk-set' RC, using simulations, which are described in Section 9.8.

It is important to note that, as described in Section 9.1, fitting the linear mixed model to the observed longitudinal measurements does not, under our modelling assumptions, give valid inferences. We have assumed that no longitudinal measurements take place after a subject experiences the event of interest. Since the occurrence of the event of interest is assumed to be driven by the unobserved random-effects \mathbf{X}_i , these 'missing' longitudinal measurements are not missing at random,

which is the weakest condition under which a likelihood based analysis of the longitudinal measurements, which ignores the mechanism giving rise to the missing data, gives valid inference.

9.3.3 Different specifications of $\mathbf{X}_i^*(t)$

As previously discussed in Chapters 6, 7, and 8, in a typical epidemiological study we usually do not know a priori how the hazard at time t depends on the longitudinal process history $M_i^H(t)$, i.e. we are uncertain regarding the specification of $\mathbf{X}_i^*(t) = \mathbf{G}(M_i^H(t))$. We may often be interested in estimation for a number of different specifications for $\mathbf{X}_i^*(t) = \mathbf{G}(M_i^H(t))$.

As previously described in Section 6.2, Boshuizen *et al* recently used data from the Seven Countries Study to investigate how mortality risk depends on both current, and levels 25 years earlier, of systolic blood pressure (SBP) and total cholesterol [119]. They first fitted a linear mixed model to the longitudinal measurements of SBP and cholesterol. They then fitted various Cox proportional hazards models for the hazard of death due to coronary heart disease (and also models for the hazard of death due to stroke), using the simplified RC approach previously described. Boshuizen *et al* first examined the effect of average SBP/cholesterol on hazard. Since subjects only had periodic measurements of SBP and cholesterol, and these were subject to measurement error and within-subject variation, at time t , Boshuizen *et al* used the BLUP of the mean level of SBP/cholesterol between time zero and time t as a time-dependent covariate. They then fitted models, again using RC, with either current SBP/cholesterol, or SBP/cholesterol level 25 years earlier as time-dependent covariates. Lastly, they fitted models in which current and levels 25 years earlier were included simultaneously as covariates, to examine whether current and past levels of SBP or cholesterol have independent effects on hazard.

For each specification of $\mathbf{X}_i^*(t) = \mathbf{G}(M_i^H(t))$, the validity of the resulting RC estimates of β_{X^*} and β_{Z^*} relies on the assumption of equation (9.7), which says that the hazard at time t depends on $M_i^H(t)$ only through the value of $\mathbf{X}_i^*(t) = \mathbf{G}(M_i^H(t))$. If this is not the case, RC is not expected to give consistent estimates of a meaningful parameter. In the context of the analyses performed by Boshuizen *et al*, this means for example that when fitting the Cox model with current SBP as a time-dependent covariate, RC only gives (approximately) consistent estimates of the effect of current SBP if the past SBP levels have no independent influence on hazard. If they do, RC does not give valid estimates of the unadjusted effect of current SBP on hazard, which is the value we would consistently estimate if \mathbf{X}_i were observed directly. We have highlighted this issue in a letter regarding the paper by Boshuizen *et al* [131]. Boshuizen *et al* replied to our letter [132], in which they explained that for example when fitting the Cox model with only predicted current SBP as covariate, they were implicitly assuming that past levels of SBP had no independent

effect on hazard. Boshuizen *et al* also presented results for the effect of SBP 25 years earlier on current hazard, without adjustment for current SBP. Analogous to above, the resulting RC estimates for this analysis are then only valid if current SBP has no independent effect on hazard. These two sets of analyses thus only provide estimates of meaningful parameters under two sets of conflicting assumptions.

The magnitude of the bias in such RC estimates will depend both on the effect of the omitted covariate and the strength of the association between the omitted (true) covariate and the included covariate. We investigate how large the biases may be through simulations, which are described in Section 9.8. In Section 9.6 we show how multiple imputation might be used to overcome such difficulties. Specifically, in the example examined by Boshuizen *et al*, we would propose imputing \mathbf{X}_i from a fitted model in which the hazard is assumed to potentially depend jointly on current and past SBP levels. The imputations can then be used to fit models in which either current or past SBP level is entered as a time-dependent covariate. In the case when the omitted covariate (current or past SBP) has an independent effect on hazard, this approach is expected to give consistent estimates of the crude, unadjusted effect of the included covariate, in contrast to RC, provided that hazard depends on the longitudinal history only through its current value and its value 25 years earlier.

9.4 Maximum likelihood

In this section we give an overview of the ML approach for joint models of longitudinal and survival outcome models. In Section 9.5 we show how MCEM can be used to obtain ML estimates.

9.4.1 Model specification and likelihood function

Wulfsohn and Tsiatis were the first to propose using maximum likelihood to estimate the parameters of a joint model in which the outcome model is Cox's proportional hazards model, the covariates of this model are time-dependent, and these are measured periodically and with error [94]. Wulfsohn and Tsiatis considered what we have described as 'situation 2', in which the longitudinal measurements are observed concurrently with the event of interest, although the method also applies for all three of the situation types we described earlier. They assumed a random-intercepts and slopes model for the longitudinal process, and that the hazard at time t depended on the longitudinal process only via its current value $M_i(t) = X_{i1} + X_{i2}t$.

Later Henderson *et al* considered estimation within a more general model specification [99]. Whereas Wulfsohn and Tsiatis considered $X_i^*(t) = X_{i1} + X_{i2}t$, Henderson *et al* considered models in which the hazard at time t depends jointly on a subject-specific intercept X_{i1} , slope X_{i2} , and current value $X_{i1} + X_{i2}t$. Henderson *et al* also allowed for a frailty (subject-specific random-effect) term in the Cox proportional

hazards model and considered models in which the longitudinal process is also a function of a stochastic process, in addition to subject-specific random-effects and independent measurement error.

We now review the model and likelihood function specifications for the modelling assumptions described in Section 9.2. In addition to the assumptions previously stated, we must specify a parametric model for the random-effects \mathbf{X}_i , conditional on \mathbf{Z}_i . Due to computational convenience, an assumption of multivariate normality has typically been made for \mathbf{X}_i . As before, the observed data likelihood function is then equal to the integral of the complete data (i.e. when the \mathbf{X}_i is observed) likelihood function:

$$\prod_{i=1}^n \int f(V_i, Y_i | \mathbf{X}_i, \mathbf{Z}_i) f(\mathbf{W}_i | \mathbf{X}_i) f(\mathbf{X}_i | \mathbf{Z}_i) d\mathbf{X}_i. \quad (9.12)$$

The component corresponding to the Cox proportional hazards outcome model is given by:

$$\begin{aligned} & (h_0(V_i) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(V_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i))^{Y_i} \\ & \times \exp \left(- \int_{t_0^S}^{V_i} h_0(s) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(s) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) ds \right). \end{aligned} \quad (9.13)$$

Under our assumed model, the longitudinal measurements \mathbf{W}_i are multivariate normal conditional on \mathbf{X}_i , with density:

$$f(\mathbf{W}_i | \mathbf{X}_i) = \prod_{j=1}^{n_i} \frac{1}{\sqrt{2\pi\sigma_U^2}} \exp \left(- \frac{(W_{ij} - \mathbf{g}(t_{ij})^T \mathbf{X}_i)^2}{2\sigma_U^2} \right). \quad (9.14)$$

Lastly, we must specify $f(\mathbf{X}_i | \mathbf{Z}_i)$. As in previous chapters, one possible specification assumes that \mathbf{X}_i is multivariate normal with mean a linear function of \mathbf{Z}_i and constant covariance matrix $\boldsymbol{\Sigma}_{X|Z}$.

As in the case of time-independent covariates measured with classical measurement error (Section 5.4), if no assumptions are made regarding the form of the baseline hazard function, MLEs can be found by assuming that the cumulative baseline hazard function is a step function. At each failure time t , it increases by an amount $\Delta H_0(t)$, which is treated as a parameter to be estimated. In this case, the component of the modified observed data likelihood function corresponding to the Cox proportional hazards model (equation (9.13)) is equal to:

$$\begin{aligned} & (\Delta H_0(V_i) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(V_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i))^{Y_i} \\ & \times \exp \left(- \sum_{j: V_j \leq V_i} \Delta H_0(V_j) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(V_j) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) \right). \end{aligned} \quad (9.15)$$

9.4.2 Estimation

The integral involved in the observed data likelihood function given by equation (9.12) cannot be expressed in closed form. Wulfsohn and Tsiatis proposed that the EM algorithm be used to maximize the observed data likelihood, using Gaussian quadrature to approximate the intractable integrals [94]. In the M-step (as previously described in Section 5.4.2), Wulfsohn and Tsiatis showed that closed form expressions exist for the updated estimate of σ_U^2 and the mean and covariance parameters for \mathbf{X}_i . Wulfsohn and Tsiatis proposed using one-step of Newton-Raphson to update the estimates of β_{X^*} (in their particular model formulation \mathbf{X}_i^* was scalar, and there were no \mathbf{Z}_i). The increments $\Delta H_0(t)$ of the cumulative baseline hazard function are then given by Breslow's estimator, but with the denominator terms replaced by their estimated expectation (see equation (5.24) of Section 5.4.2).

Henderson *et al* proposed using Monte-Carlo integration to approximate the required integrals, and hence to use the MCEM algorithm [99]. Wulfsohn and Tsiatis had shown earlier that the integrals involved could be expressed as expectations with respect to the conditional distribution of \mathbf{X}_i given the longitudinal measurements \mathbf{W}_i . With normality assumptions for \mathbf{X}_i and the measurement errors U_{ij} , this conditional distribution is normal, and thus easy to generate draws from. Henderson *et al* thus proposed generating draws from this conditional distribution and using these to approximate the required integrals. In the M-step, the parameters are then updated in the same way as proposed by Wulfsohn and Tsiatis [94].

9.4.3 Inference and asymptotic properties

Zeng and Cai have recently shown [97] that the MLE of the joint model parameters is consistent and asymptotically normal, and that the profile likelihood approach, proposed by Wulfsohn and Tsiatis [94], can be used to give asymptotically valid inferences for the finite dimensional parameters of the model. This appears however to conflict with more recent findings by Hsieh *et al* [96], who claim that the approach underestimates the variance of estimates.

9.4.4 Assumptions and likelihood justification

The likelihood function given in equation (9.12) was loosely justified in the paper by Wulfsohn and Tsiatis [94] as being valid provided that the timing of error-prone measurements is uninformative. However, except in what we described as situation 1, the longitudinal error-prone measurements play a role analogous to 'internal' time-dependent covariates [133]. That is, their timing is predictive of survival, in the sense that if a subject has an error-prone measurement at time t , we know they did not fail prior to time t . More recently, Tsiatis and Davidian have given a more formal justification for such a likelihood function [124]. In the same way as the

likelihood function in the case of time-dependent covariates for standard survival models is derived, Tsiatis and Davidian showed that a likelihood function such as that given in equation (9.12) is valid under the assumptions [124]:

- $h(t|V_i \geq t, \mathbf{X}_i, \mathbf{Z}_i, W_i^H(t), \mathbf{t}_i^H(t)) = h(t|T_i \geq t, \mathbf{X}_i, \mathbf{Z}_i)$
- measurement errors U_{ij} are independent of all other variables, conditional on \mathbf{X}_i and \mathbf{Z}_i
- at each time t , the hazard of censoring may depend on the past observed longitudinal history and \mathbf{Z}_i , but not on \mathbf{X}_i
- at each time t , the probability of a longitudinal error-prone measurement taking place may depend on the past observed longitudinal history and \mathbf{Z}_i , but not on \mathbf{X}_i

9.4.5 Software

A significant issue in relation to the use of ML for longitudinal and survival models is the lack of availability to fit such models in statistical software packages. As far as we are aware, the SAS PROC NLMIXED command is the only way ML estimates can be obtained using standard statistical software. To address the need for estimation routines for such models, Rizopoulos has recently released the JM package for R, which allows a number of joint survival/longitudinal models to be fitted, including the one used by Wulfsohn and Tsiatis [94]. Unfortunately, the JM package only permits estimation for models in which $X_i^*(t)$ is scalar and equal to the current value $M_i(t)$ of the longitudinal process.

9.4.6 Flexible parametric approach and sensitivity to parametric assumptions

The ML approach described thus far is based on making parametric assumptions, typically of multivariate normality, for the random-effects vector \mathbf{X}_i . To relax this assumption Song *et al* proposed a likelihood approach in which the distribution of \mathbf{X}_i is assumed to belong to the semi-nonparametric class of distributions [134]. They proposed using the EM algorithm for estimation, which like the model which assumes normality for \mathbf{X}_i , involves intractable integrals at the E-step. They described the use of Gaussian quadrature to approximate the required integrals. Unexpectedly, Song *et al* found that ML estimates for the model which assumes normality for \mathbf{X}_i appeared to remain unbiased even when the normality assumption was violated, which they said required further investigation. Additional simulations by Tsiatis and Davidian further supported this finding of robustness to parametric assumptions for \mathbf{X}_i [124].

This phenomenon has been investigated further from various theoretical perspectives by a number of authors [96, 122, 135]. As discussed earlier in Section 8.2.4, the results of these papers suggest that the ML approach may be robust to violations of the parametric assumptions for the random-effects \mathbf{X}_i when the information available about the random-effects for each subject is large. This can occur either because measurement errors are small, or because subjects have many longitudinal measurements. Huang *et al* have recently proposed an approach similar to SIMEX, which can be used to investigate the sensitivity of the ML estimator to the distributional assumptions for \mathbf{X}_i in a particular dataset [122].

9.4.7 Parametric models for the hazard function

As previously discussed use of quadrature methods is feasible when the dimension of \mathbf{X}_i is small. However, there is probably considerable interest in fitting models in which the dimension of \mathbf{X}_i is larger. In particular, this is the case if we attempt to model the trajectory of the longitudinal process using splines or higher-order functions of time (see Section 9.9). To address the computational difficulties of finding the MLE in such joint models, Rizopoulos has recently proposed using a version of the Laplace approximation technique to approximate the integrals [136]. Rather than make no assumptions about the baseline hazard function, Rizopoulos *et al* proposed a model in which the baseline hazard function is specified parametrically using B-splines [136]. Since in typical applications the baseline hazard function can be expected to be smooth, use of a fully parametric model can be expected to have greater efficiency compared to leaving its form unspecified. Furthermore, inference for parameter estimates follows from standard asymptotic theory, in contrast to the situation when the baseline hazard function is estimated non-parametrically.

9.5 Ascent-based Monte-Carlo Expectation Maximization

In this section we show how MCEM, described in the case of a time-independent covariate measured with classical measurement error in Section 5.5, can be used to find MLEs for the parametric model described in Sections 9.2 and 9.4.1. We show in Section 9.5.1 that, as in the case of time-independent covariates measured with classical measurement error, rejection sampling can be used to draw from the conditional distribution of \mathbf{X}_i given the observed data. In Section 9.5.2 we describe the necessary modifications to the M-step of EM. The rules of ascent-based MCEM, as previously described in Section 4.5.3, can again be applied without modification.

9.5.1 Generating imputations using rejection sampling

We first restate equation (5.25), but in more generality here with multivariate \mathbf{X}_i and potentially error-free \mathbf{Z}_i :

$$f(\mathbf{X}_i|V_i, Y_i, \mathbf{W}_i, \mathbf{Z}_i) = \frac{f(V_i, Y_i|\mathbf{X}_i, \mathbf{Z}_i)f(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)}{f(V_i, Y_i|\mathbf{W}_i, \mathbf{Z}_i)}. \quad (9.16)$$

As in Section 5.5.1, we consider the cases of $Y_i = 0$ and $Y_i = 1$ separately. We use $f(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ as our candidate density, as this is multivariate normal under the previously stated assumptions (of conditional normality for \mathbf{X}_i given \mathbf{Z}_i and normality for W_{ij} given \mathbf{X}_i).

For censored subjects, the derivation of Section 5.5.1 applies without modification. Thus, we can draw $\mathbf{x}_i \sim f(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ and $u \sim U(0, 1)$, and accept \mathbf{x}_i if:

$$u \leq S(V_i|\mathbf{x}_i, \mathbf{Z}_i) = \exp \left(- \sum_{j:V_j \leq V_i} \Delta H_0(V_j) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{x}_i^*(V_j) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) \right), \quad (9.17)$$

where $\mathbf{x}_i^*(V_j) = \mathbf{A}(V_j, \mathbf{x}_i)$.

We now consider the case in which subject i experiences the event of interest, so that $Y_i = 1$. To use rejection sampling with $f(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)$ as the candidate density, we must bound:

$$\frac{f(\mathbf{X}_i|V_i, Y_i = 1, \mathbf{W}_i, \mathbf{Z}_i)}{f(\mathbf{X}_i|\mathbf{W}_i, \mathbf{Z}_i)}, \quad (9.18)$$

which by equation (9.16) is equal to:

$$\frac{f(V_i, Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i)}{f(V_i, Y_i = 1|\mathbf{W}_i, \mathbf{Z}_i)}. \quad (9.19)$$

The density $f(V_i, Y_i|\mathbf{X}_i, \mathbf{Z}_i)$ for $Y_i = 1$ is, according to equation (9.15):

$$\begin{aligned} f(V_i, Y_i = 1|\mathbf{X}_i, \mathbf{Z}_i) &= \Delta H_0(V_i) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(V_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) \\ &\times \exp \left(- \sum_{j:V_j \leq V_i} \Delta H_0(V_j) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(V_j) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) \right) \end{aligned} \quad (9.20)$$

We now substitute $\mathbf{X}_i^*(t) = \mathbf{A}(t, \mathbf{X}_i)$ to make clear the dependence on \mathbf{X}_i :

$$\begin{aligned} \Delta H_0(V_i) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{A}(V_i, \mathbf{X}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) \\ \times \exp \left(- \sum_{j:V_j \leq V_i} \Delta H_0(V_j) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{A}(V_j, \mathbf{X}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) \right) \end{aligned} \quad (9.21)$$

We do not believe it is possible to tightly bound this expression without taking into account the particular specification of the function $\mathbf{A}(t, \mathbf{X}_i)$. However, since the increments $\Delta H_0(V_j)$ are always positive, the expression is always less than or equal to:

$$\begin{aligned} & \Delta H_0(V_i) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{A}(V_i, \mathbf{X}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) \\ & \times \exp(-\Delta H_0(V_i) \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{A}(V_i, \mathbf{X}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i)). \end{aligned} \quad (9.22)$$

Analogous to the derivation in Section 5.5.1, by differentiating with respect to \mathbf{X}_i it follows that this expression takes its maximum when:

$$\exp(\boldsymbol{\beta}_{X^*}^T \mathbf{A}(V_i, \mathbf{X}_i) + \boldsymbol{\beta}_{Z^*}^T \mathbf{Z}_i) = 1/\Delta H_0(V_i), \quad (9.23)$$

so that the expression in equation (9.22) can be bounded above by $\exp(-1)$. Thus equation (9.19) can be bounded above by:

$$\frac{\exp(-1)}{f(V_i, Y_i = 1 | \mathbf{W}_i, \mathbf{Z}_i)}. \quad (9.24)$$

Rejection sampling can then be used by drawing $\mathbf{x}_i \sim f(\mathbf{X}_i | \mathbf{W}_i, \mathbf{Z}_i)$ and $u \sim U(0, 1)$, and accepting \mathbf{x}_i if:

$$\begin{aligned} u & \leq \frac{f(V_i, Y_i = 1 | \mathbf{x}_i, \mathbf{Z}_i)}{f(V_i, Y_i = 1 | \mathbf{W}_i, \mathbf{Z}_i) \frac{\exp(-1)}{f(V_i, Y_i = 1 | \mathbf{W}_i, \mathbf{Z}_i)}} \\ & = \exp(1) f(V_i, Y_i = 1 | \mathbf{x}_i, \mathbf{Z}_i), \end{aligned} \quad (9.25)$$

where $f(V_i, Y_i = 1 | \mathbf{x}_i, \mathbf{Z}_i)$ is as given in equation (9.20). While this procedure allows us to generate imputations from $f(\mathbf{X}_i | V_i, Y_i = 1, \mathbf{W}_i, \mathbf{Z}_i)$, the looseness of the bound used means that a large number of draws will be required before the (u, \mathbf{x}_i) pair satisfies the inequality in equation (9.25). We recall that in the case of a time-independent covariate (measured with classical error), the inequality found in Section 5.5.1 was $u \leq \frac{H_0(V_i)}{h_0(V_i)} \exp(1) f(V_i, Y_i = 1 | x_i)$, which is much more likely to be satisfied, due to the presence of the term $\frac{H_0(V_i)}{h_0(V_i)}$.

9.5.2 M-step

Having multiply imputed \mathbf{X}_i , the model parameter estimates must be updated. As before, we let $\mathbf{X}_i^{(m)}$ denote the m th imputation of \mathbf{X}_i . The updated estimate of the measurement error variance σ_J^2 for the m th imputation is then given by:

$$\frac{\sum_{i=1}^n \sum_{j=1}^{n_i} (W_{ij} - \mathbf{g}(t_{ij})^T \mathbf{X}_i^{(m)})^2}{\sum_{i=1}^n n_i}. \quad (9.26)$$

The mean and covariance parameters of the normal distribution corresponding to $f(\mathbf{X}_i|\mathbf{Z}_i)$ similarly follow from standard results for MLEs in the multivariate normal model, depending of course on the specification for how the mean varies with \mathbf{Z}_i .

The Cox model parameters β_X and β_Z can be estimated using standard routines to fit a Cox proportional hazards model with time-updated covariate $\mathbf{X}_i^* = \mathbf{A}(t, \mathbf{X}_i^{(m)})$ (in the m th imputation) and time-independent covariate \mathbf{Z}_i . Lastly, the increments of the cumulative baseline hazard function are, analogous to the expression in Section 5.5.2, updated by:

$$\hat{\Delta}(H_0(t)) = \sum_{V_i=t} \frac{Y_i}{\sum_{j \in R_i} \mathbb{E}(\exp(\beta_{X^*}^T \mathbf{X}_j^*(t) + \beta_{Z^*}^T \mathbf{Z}_j))}, \quad (9.27)$$

where the expectation is replaced by its Monte-Carlo estimate based on the multiple imputations of \mathbf{X}_i .

9.6 Multiple imputation

Analogous to our proposals of Sections 7.6.5 and 8.6.5, we propose that MI can be used to estimate the Cox proportional hazard model parameters β_{X^*} and β_{Z^*} corresponding to a number of different specifications for $\mathbf{X}_i^*(t) = \mathbf{G}(M_i^H(t)) = \mathbf{A}(t, \mathbf{X}_i)$. First, we define $\mathbf{X}_i^*(t)$ such that all of our alternative specifications of interest for $\mathbf{X}_i^*(t)$ are special cases. For example, in the analyses of Boshuizen *et al*, it appeared that the authors were interested in at least the three following specifications for $\mathbf{X}_i^*(t)$:

1. $X_i^*(t) = M_i(t)$
2. $X_i^*(t) = M_i(t - 25)$
3. $\mathbf{X}_i^*(t) = (M_i(t), M_i(t - 25))^T$

In this case, the first and second models are nested within the third more general model specification. To obtain valid parameter estimates for β_{X^*} and β_{Z^*} under all three of these specifications, we propose that the parameters of the joint model be first estimated under the general specification (for example by ML), within which the other specifications are nested (i.e. $\mathbf{X}_i^*(t) = (M_i(t), M_i(t - 25))^T$).

Having estimated the parameters of the joint model for the general specification of $\mathbf{X}_i^*(t)$, we can then multiply impute \mathbf{X}_i , using the estimates of the model parameters. This can be done using rejection sampling, as described in Section 9.5.1. For the m th imputation, we can then fit the Cox model for each of the specifications of $\mathbf{X}_i^*(t)$, where we use the imputation $\mathbf{X}_i^{(m)}$ to generate $\mathbf{X}_i^*(t) = \mathbf{A}(t, \mathbf{X}_i^{(m)})$. The parameter estimates can then be averaged over the imputations. Providing that the model used to generate the imputations is not misspecified, we believe this approach

will yield valid parameter estimates for a given specification of $\mathbf{X}_i^*(t)$, even if the conditional independence assumption:

$$h(t|\mathbf{X}_i^*(t), M_i^H(t), \mathbf{Z}_i) = h(t|\mathbf{X}_i^*(t), \mathbf{Z}_i) \quad (9.28)$$

does not hold. In the models considered by Boshuizen *et al*, we may for example fit the Cox model corresponding to $X_i^*(t) = M_i(t)$, and the parameter estimates should be consistent for the unadjusted effect of current SBP, even if SBP levels 25 years earlier have an independent effect on hazard at time t .

One of the advantages of using the MCEM procedure described in Section 9.5 to find ML estimates is that a large set of imputations of \mathbf{X}_i is available after the procedure has converged. Providing the specification of $\mathbf{X}_i^*(t)$ made encompasses the various alternative specifications of $\mathbf{X}_i^*(t)$ as special cases, we can use these multiple imputations to estimate β_{X^*} and β_{Z^*} under the various alternatives for the definition of $\mathbf{X}_i^*(t)$. We explore this proposal in simulations, which we report in Section 9.8.

Since our proposed MCEM procedure is very computationally intensive, an alternative approach to creating imputations would be to first estimate the model parameters using a method such as RC, and then to create multiple imputations (using rejection sampling as described) using these parameter estimates.

9.7 Conditional score method

The conditional score method, originally devised by Carroll *et al* for generalised linear outcome models (see Section 4.9), was first adapted to the case of a Cox proportional hazards model in the setting of longitudinal error-prone measurements, by Tsiatis and Davidian [105]. Tsiatis and Davidian described the implementation of the method when $\mathbf{X}_i = (X_{i1}, X_{i2})^T$ represent random-intercepts and slopes for a longitudinal process, and when the hazard at time t is assumed to depend on the longitudinal process history via its current value $X_i^*(t) = X_{i1} + X_{i2}t$.

In the implementation of the CS method for binary outcomes, we recall the requirement that the design matrix \mathbf{D}_i in the linear mixed model $\mathbf{W}_i = \mathbf{D}_i\mathbf{X}_i + \mathbf{U}_i$ be of full-rank. In the case of a time-to-event outcome with time-dependent covariates, the analogous conditions mean that a subject may be able to contribute to estimation at some risk-sets, but not others. For the random-intercepts and slopes model considered by Tsiatis and Davidian [105], for an event to contribute to estimation, the subject who died must have had at least two longitudinal error-prone measurements preceding their event time, so that their intercept and slope are identifiable. Furthermore, at each event time, only subjects who at the event time in question have at least two longitudinal error-prone measurements contribute to the risk-set.

Tsiatis and Davidian proposed an estimator for the measurement error variance σ_U^2 , and gave a heuristic argument of the consistency and asymptotic normality of the CS estimator. In simulations, Tsiatis and Davidian found that the CS estimator gave unbiased estimates under a wide range of different distributions for \mathbf{X}_i [105]. One of the prices for obtaining consistency without requiring distributional assumptions for \mathbf{X}_i is inefficiency, relative to the correctly specified ML estimator. In simulation results by Tsiatis and Davidian [124], the standard deviation of the CS estimator was 15% larger than the MLE.

9.7.1 Multiple longitudinal processes and alternative model specifications

Subsequently Song *et al* generalized the CS method to a much more general model setting [137]. Song *et al* developed the method when there may be multiple longitudinal processes, each of which follow a model as given in equation (9.1) (i.e. not restricting attention to the random-intercepts and slopes model), and in which measurements of the different longitudinal processes made at the same time may be correlated. In this more general setting, a subject i can only contribute to the risk-set at time t if $\mathbf{X}_i^*(t)$ is identifiable based on their preceding measurements. Song *et al* stated that at a given time, a subject must have (in the case of a single longitudinal process) at least as many error-prone measurements as the dimension of \mathbf{X}_i , although this does not seem to be a strong enough sufficient condition to allow identifiability. In our simulations (see Section 9.8), we consider a piece-wise constant model for $M_i(t)$, with a five-dimensional random-effects vector \mathbf{X}_i representing the longitudinal process value in each of five equally size periods of time. We then assume the hazard at time t depends on the value of the longitudinal process in the preceding two intervals over which the process is constant. In this case, even if a subject had at least five longitudinal measurements, if they had no measurements from one of the two preceding intervals, the value of $\mathbf{X}_i^*(t)$ would not be identifiable. We therefore believe subjects could only contribute to the CS estimating equation in a particular period if they had at least one error-prone measurements in the two preceding periods. Such a requirement may be overly restrictive in particular applications.

9.7.2 Assumptions

Tsiatis and Davidian later detailed assumptions under which the CS estimator gives consistent estimates [124]. In summary:

- $h(t|V_i \geq t, \mathbf{X}_i, \mathbf{Z}_i, W_i^H(t), \mathbf{t}_i^H(t)) = h(t|T_i \geq t, \mathbf{X}_i, \mathbf{Z}_i)$
- Measurement errors U_{ij} are independent of all other variables

- At each time t , the hazard of censoring, after conditioning on $\mathbf{X}_i, \mathbf{Z}_i, \mathbf{t}_i^H(t)$, does not depend on the past measurement errors
- At each time t , the probability of a longitudinal error-prone measurement taking place, after conditioning on $\mathbf{X}_i, \mathbf{Z}_i, \mathbf{t}_i^H(t)$, does not depend on the past measurement errors

As noted by Tsiatis and Davidian, the third and fourth assumptions differ compared to the third and fourth assumptions which are sufficient to justify the likelihood approach [124]. Whereas for the likelihood approach, censoring and timing of measurements may depend on past longitudinal measurements but not \mathbf{X}_i , for the CS method, censoring and timing of measurements may depend on \mathbf{X}_i , but conditional on \mathbf{X}_i , not on the past longitudinal measurements (which conditional on \mathbf{X}_i is equivalent to being independent of the measurement errors). As Tsiatis and Davidian pointed out, if censoring and the timing of measurements is thought not to depend on \mathbf{X}_i or the past measurement history, then the assumptions required for either the likelihood or CS approach are satisfied [124].

9.8 Simulations

In this section we give the results of simulations to investigate the performance of RC methods and ML via ascent-based MCEM. Rather than simulate under the random-intercepts and slopes model, as we have done previously for continuous and binary outcomes, we adopt the piece-wise constant model we have previously given as an example. Our motivation for this is discussed further in Section 9.9.

9.8.1 Simulation setup

Longitudinal process

For each of $n = 1,000$ subjects, we simulated a five dimensional random-effects vector \mathbf{X}_i , with mean $(0, 0, 0, 0, 0)^T$ and variance covariance matrix:

$$\Sigma_X = \begin{pmatrix} 1 & 0.5 & 0.4 & 0.3 & 0.2 \\ 0.5 & 1 & 0.5 & 0.4 & 0.3 \\ 0.4 & 0.5 & 1 & 0.5 & 0.4 \\ 0.3 & 0.4 & 0.5 & 1 & 0.5 \\ 0.2 & 0.3 & 0.4 & 0.5 & 1 \end{pmatrix} \quad (9.29)$$

We assumed that the longitudinal process is observed from time 0 until potentially $t_1^L = 50$ ‘years’ and we divided time into 5 intervals of 10 years. We assumed a

piece-wise constant model for the longitudinal process, so that for a given time t ,

$$M_i(t) = X_{i(\lfloor t/10 \rfloor + 1)}, \quad (9.30)$$

where $\lfloor x \rfloor$ denotes x rounded down to the next smallest integer. We then generated error-prone measurements W_{ij} according to:

$$W_{ij} = M_i(t_{ij}) + U_{ij}, \quad (9.31)$$

with $U_{ij} \sim N(0, 1)$. All subjects had an error-prone measurement at $t = 0$, as would typically occur at the beginning of a study. We then simulated from a discrete uniform distribution to determine how many additional error-prone measurements each subject had, ranging from 0 to 19 (to give an overall potential maximum of 20 error-prone measurements). The times of these additional measurement, t_{ij} , were then generated according to the continuous uniform distribution on $(0, 50)$. Any measurements which took place after the time when a subject experienced the event of interest (see below) were then removed.

Survival process

We assumed the risk-period started at $t_0^S = 20$ years, and that the hazard at time t depended jointly on the value of $M_i(\cdot)$ in the preceding two decades. For a given time t , let j denote the integer such that $10j < t < 10(j+1)$, i.e. $j = \lfloor t/10 \rfloor$. Then:

$$\begin{aligned} \mathbf{X}_i^*(t) &= (M_i(t-10), M_i(t-20))^T \\ &= (X_{ij}, X_{i(j-1)})^T, \end{aligned} \quad (9.32)$$

We assumed that the hazard for subject i at time t was equal to:

$$h_0 \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(t)) \quad (9.33)$$

where $h_0 = 0.1$ denotes an assumed constant baseline hazard. The survival function is then given by:

$$S(t|\mathbf{X}_i) = \exp\left(-\int_{20}^t h_0 \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(s)) ds\right). \quad (9.34)$$

To evaluate the integral we decompose it based on the partition of time into intervals of length 10, over which \mathbf{X}_i^* , and hence the integrand, is constant:

$$\int_{20}^t h_0 \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(s)) ds = \sum_{j=2}^{\lfloor t/10 \rfloor - 1} \int_{10j}^{10(j+1)} h_0 \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(s)) ds + \int_{10\lfloor t/10 \rfloor}^t h_0 \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(u)) du \quad (9.35)$$

Since the integrand is constant in the domains of each of these integrals:

$$\int_{10j}^{10(j+1)} h_0 \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(s)) ds = 10h_0 \exp(\beta_{X_1^*} X_{ij} + \beta_{X_2^*} X_{i(j-1)}), \quad (9.36)$$

and

$$\int_{10\lfloor t/10 \rfloor}^t h_0 \exp(\boldsymbol{\beta}_{X^*}^T \mathbf{X}_i^*(s)) ds = (t - 10 \lfloor t/10 \rfloor) h_0 \exp(\beta_{X_1^*} X_{i(10\lfloor t/10 \rfloor)} + \beta_{X_2^*} X_{i(10\lfloor t/10 \rfloor - 1)}) \quad (9.37)$$

To simulate the event time for each subject, we simulated a draw $u \sim U(0, 1)$ and then found the value t (using the above expressions) such that $S(t|\mathbf{X}_i) = u$. If $u < S(50|\mathbf{X}_i)$, then the subject was censored (and the event indicator $Y_i = 0$) at $t = 50$.

We simulated data under four scenarios; scenario 1: $\beta_{X_1^*} = 0.1$, $\beta_{X_2^*} = 0.1$, scenario 2: $\beta_{X_1^*} = 1$, $\beta_{X_2^*} = 1$, scenario 3: $\beta_{X_1^*} = 0.1$, $\beta_{X_2^*} = 0$, scenario 4: $\beta_{X_1^*} = 1$, $\beta_{X_2^*} = 1$.

9.8.2 Estimation methods

Analogous to the simulations in Chapters 7 and 8, we considered estimation of:

- $\beta_{X_1^*}$ and $\beta_{X_2^*}$, the mutually adjusted effects of $M_i(t - 10)$ and $M_i(t - 20)$
- the unadjusted effect of $M_i(t - 10)$
- the unadjusted effect of $M_i(t - 20)$

We now describe the estimation methods used. Due to the computational time required for the ML (via MCEM) approach, we performed 100 simulations per scenario.

Ideal

As for the simulations with a binary outcome, closed form expressions are not available for some of the unadjusted effect parameters. For the unadjusted effects, we

therefore present the mean (SD) of the ‘ideal’ estimator, which consists of fitting a Cox proportional hazards model using the true value of the longitudinal process in the preceding decade (or the value two decades earlier) as time-dependent covariate. This estimator is thus of course not possible in practice, but it allows us to estimate the unadjusted parameter which would be consistently estimated if \mathbf{X}_i were observed without error.

Regression calibration

To implement RC, we fitted a single linear mixed model to all the longitudinal measurements from all subjects, using the lmer command in R. Based on this fit, the BLUPs of their five dimensional random-effects vector \mathbf{X}_i were calculated, and added to the estimated mean $\hat{\boldsymbol{\mu}}_X$. These were then used to predict the time-dependent covariate \mathbf{X}_i^* , as given by equation (9.32). That is, for events occurring between 20 and 30 years, the first covariate was set equal to each subject’s predicted value of X_{i2} , and the second covariate was set equal to their predicted value of X_{i1} . For events between 30 and 40 years, the predictions of, respectively, X_{i3} and X_{i2} were used as covariates. To estimate $\beta_{X_1^*}$ and $\beta_{X_2^*}$, we fitted the Cox proportional hazards model with the predictions of the longitudinal process in the preceding two decades as time-dependent covariates. To estimate the unadjusted effects of either $M_i(t-10)$ or $M_i(t-20)$, we re-fitted the Cox model with the corresponding predicted values as covariate.

Risk-set regression calibration

To implement risk-set RC, in closer spirit to the induced hazard given in equation (9.10), at each event time, we refitted the linear mixed model, using data only from subjects who were at risk at that event time. Furthermore, at each event time, we fitted the linear mixed model using only those measurements made in the decades preceding that in which the event occurred. For example, for events occurring between 20 and 30 years, we fitted linear mixed models using only measurements occurring between 0 and 20 years (and by implication with only a two-dimensional random-effects vector). The linear mixed model fit was then used to predict $\mathbf{X}_i^*(t)$ at the particular event time. We then fitted the same Cox proportional hazards models as described for RC.

If we had attempted to include longitudinal measurements made in the ‘current’ decade, estimation problems would occur at event times early in the decade. For example, suppose an event occurred at 20.1 years. Then it is likely that very few subjects would have any error-prone measurements between times 20 and 20.1, meaning there would be very little information with which to estimate the variance and covariances of X_{i3} .

Maximum likelihood via ascent-based Monte-Carlo Expectation Maximization (ML)

We used ascent-based MCEM to estimate the mutually adjusted effects of $M_i(t-10)$ and $M_i(t-20)$ on the hazard at time t , as described in Section 9.5. We used the estimates of $\boldsymbol{\mu}_X$, $\boldsymbol{\Sigma}_X$ and σ_U^2 found in the first stage of RC as initial estimates. For $\boldsymbol{\beta}_{X^*}$ and the baseline hazard function, we used the RC estimates. We used the same values of the control parameters for ascent-based MCEM as described previously.

When running the simulations we found that the MCEM algorithm did not converge, even after many iterations. In our previous simulations, the number of imputations increased as the estimated increase in likelihood at each iteration become smaller. However, for the current simulations, with each iteration the estimate of the increase in likelihood did not decrease, so that the number of imputations did not increase from the initial 10. We suspected that the increase in the likelihood (strictly the expected complete data log likelihood) at each iteration was being over estimated. To investigate this, at each iteration we generated a second, independent set of imputations, from which we estimated the increase in the likelihood. This confirmed that the estimate based on the imputations used to update the model parameters was over-optimistic. This occurs because the same set of imputations are used to find updated values of the model parameters and to assess the increase in likelihood. This optimism converges in probability to zero as the number of imputations increases [74]. By using an independent set of imputations to estimate the increase in the likelihood function, we found that the algorithm converged reliably. We are unsure as to why this occurred in these simulations but not in our earlier simulations. However, the issue also occurred in our data analyses, both when considering a binary outcome (Chapter 12) and a time to event outcome (Chapter 13), in which the dimension of \mathbf{X}_i was moderate. This suggests the issue is related to the higher dimensionality of \mathbf{X}_i , although we are unsure as to exactly why this occurs.

We did not find the ML estimates of the unadjusted effects for the joint model which assumes hazard at t depends only on $M_i(t-10)$ or $M_i(t-20)$ (as we did for the analogous simulations in Chapters 7 and 8), because the MCEM algorithm here was so much slower in terms of computational time.

Maximum likelihood via ascent-based Monte-Carlo Expectation Maximization plus multiple imputation (ML+MI)

To estimate the unadjusted effects of $M_i(t-10)$ and $M_i(t-20)$ on the hazard at time t , we fitted time-dependent Cox models using the multiple imputations of \mathbf{X}_i which were available after termination of the MCEM algorithm. These estimates thus do not rely on the omitted covariate having no independent effect on hazard.

Table 9.1: Estimates of adjusted effects for Cox regression simulations. Mean (SD) of estimates found using regression calibration (RC), risk-set regression calibration (RRC), and maximum likelihood via ascent-based MCEM (ML).

Scenario	$\beta_{X_1^*}$	$\beta_{X_2^*}$	$\hat{\beta}_{X_1^*}$		
			RC	RRC	ML
1	0.1	0.1	0.099 (0.058)	0.103 (0.061)	0.100 (0.059)
2	1	1	0.795 (0.071)	0.864 (0.079)	1.008 (0.089)
3	0.1	0	0.106 (0.054)	0.110 (0.056)	0.107 (0.055)
4	1	0	0.869 (0.058)	0.919 (0.069)	0.994 (0.076)

Scenario	$\beta_{X_1^*}$	$\beta_{X_2^*}$	$\hat{\beta}_{X_2^*}$		
			RC	RRC	ML
1	0.1	0.1	0.098 (0.055)	0.097 (0.056)	0.099 (0.056)
2	1	1	0.731 (0.073)	0.854 (0.080)	0.992 (0.096)
3	0.1	0	-0.004 (0.050)	-0.005 (0.051)	-0.004 (0.051)
4	1	0	-0.011 (0.055)	0.012 (0.061)	0.003 (0.063)

9.8.3 Simulation results

Estimates of adjusted effects

Table 9.1 shows the mean (SD) of estimates of the adjusted effects of $M_i(t - 10)$ and $M_i(t - 20)$ using RC, RRC, and ML via ascent-based MCEM. RC showed little bias in scenarios 1 and 3, in which the covariate effects were small. In contrast, in scenario 2 the estimates were biased towards the null by approximately 20% (effect of $M_i(t - 10)$) and 27% (effect of $M_i(t - 20)$). In scenario 4, the RC estimate of the effect of $M_i(t - 10)$ was still biased towards the null, but to a smaller extent (approximately 13%).

The risk-set RC estimates showed less bias but greater variability than the RC estimates, consistent with our findings in the case of a time-independent covariate measured with classical error (see Chapter 5). However, in scenarios 2 and 4, the RRC estimates still showed substantial bias.

In contrast, the ML estimates, obtained using ascent-based MCEM, showed little bias across all four scenarios. For those parameters where RC/RRC had little bias (and therefore a comparison of SD is reasonable), the variability of the ML estimates was similar to that of RC.

Estimates of unadjusted effects

Table 9.2 shows the estimates of the unadjusted effects of $M_i(t - 10)$ or $M_i(t - 20)$, using RC, RRC and ML+MI. For scenario 1, the estimates from RC were biased upwards, due to the fact that the conditional independence assumption was invalid in this scenario. For scenario 2, the mean of the RC estimates was slightly higher than

Table 9.2: Estimates of unadjusted effects for Cox regression simulations. Mean (SD) of estimates found using true \mathbf{X}_i (ideal), regression calibration (RC), risk-set regression calibration (RRC), and maximum likelihood via ascent-based MCEM plus multiple imputation (ML+MI).

Scenario	$\beta_{X_1^*}$	β_{X_2}	Unadjusted effect of $M_i(t - 10)$			
			Ideal	RC	RRC	ML+MI
1	0.1	0.1	0.152 (0.030)	0.167 (0.038)	0.170 (0.040)	0.148 (0.042)
2	1	1	1.161 (0.049)	1.176 (0.070)	1.205 (0.090)	1.164 (0.082)
3	0.1	0	0.103 (0.031)	0.103 (0.040)	0.106 (0.042)	0.104 (0.042)
4	1	0	1.004 (0.035)	0.860 (0.051)	0.925 (0.061)	0.993 (0.069)

Scenario	$\beta_{X_2^*}$	β_{X_2}	Unadjusted effect of $M_i(t - 20)$			
			Ideal	RC	RRC	ML+MI
1	0.1	0.1	0.148 (0.030)	0.158 (0.036)	0.158 (0.037)	0.148 (0.038)
2	1	1	1.176 (0.050)	1.110 (0.068)	1.178 (0.084)	1.163 (0.081)
3	0.1	0	0.051 (0.028)	0.061 (0.037)	0.062 (0.038)	0.049 (0.038)
4	1	0	0.393 (0.035)	0.471 (0.051)	0.483 (0.056)	0.395 (0.049)

that of the ideal estimator. We believe this reflects a fortuitous combination of bias towards the null inherent in RC for the Cox model, and bias away from the null (for this particular simulation setup) due to the conditional independence assumption being invalid. Indeed, for the effect of $M_i(t - 20)$, RC was biased downwards in scenario 2. For scenario 3, the RC estimates of the effect of $M_i(t - 10)$ showed little bias, since the conditional assumption was valid here, and the effect size was small. In contrast, the estimates of the effect of $M_i(t - 20)$ were biased upwards, due to the conditional independence assumption being false. In scenario 4, the estimates of the effect of $M_i(t - 10)$ were biased downwards (usual bias of RC for the Cox model), whereas the estimates of the effect of $M_i(t - 20)$ were biased upwards (due to invalidity of conditional independence assumption outweighing usual bias of RC for the Cox model).

In the absence of violations of the conditional independence assumption, the RRC estimator showed less bias than RC for estimates of the adjusted effects of $M_i(t - 10)$ and $M_i(t - 20)$. Consequently, when the conditional independence was violated (scenarios 1 and 2, and in scenarios 3 and 4, the estimates of the effect of $M_i(t - 20)$), the RRC estimator had greater bias than RC. When the conditional independence assumption was not violated (estimates of the effect of $M_i(t - 10)$ in scenarios 3 and 4), RRC was less biased than RC.

The ML+MI estimates had a similar mean to the ideal estimator for all effects. In particular, the ML+MI estimator, which uses the RC estimates as initial values, corrected for the upward bias in the RC estimates.

9.9 Modelling assumptions revisited

Throughout our developments we have assumed that for all subjects there exists an unobserved (latent) random-effects vector \mathbf{X}_i , which determines the trajectory of the longitudinal process, according to equation (9.1). The majority of papers concerning joint models of this kind have considered the simple setting in which \mathbf{X}_i is two-dimensional, representing random-intercepts and slopes [94, 105, 137]. As discussed by Tsiatis and Davidian [124], one interpretation of this assumption is that it implies that the longitudinal trajectory is an inherent characteristic of a subject.

In contrast, in many epidemiological applications, we may envisage that the longitudinal process evolves stochastically over time. To address this, Henderson *et al* proposed a model in which in addition to the random-effects \mathbf{X}_i , the longitudinal process path also depends on a continuous time stochastic process [99]. Such a model allows the longitudinal process trajectory to evolve stochastically in time, over and above the ‘long term’ trend dictated by the subject-specific random-effects \mathbf{X}_i . However, although this may be attractive from a modelling perspective, being closer to our understanding of how data arise, the addition of the stochastic process to the model substantially increases the computational complexity involved in finding MLEs [99, 125].

Both Tsiatis and Davidian [124], and more recently Diggle *et al* [125], have suggested that it may be possible to specify models which are closer in spirit to that proposed by Henderson *et al* [99] (in which a stochastic process allows the longitudinal process to evolve over time) within the type of modelling framework we have considered (i.e. with only random-effects \mathbf{X}_i determining the longitudinal trajectory). Diggle *et al* specifically suggested that spline models, expressed in terms of a random-effects model, might be used as a compromise between models which include a stochastic process (which are computationally problematic) and the types of simple random-effects specifications (e.g. random-intercepts and slopes) which had previously dominated the literature regarding joint modelling of longitudinal and survival data [125]. In fact, in one of the earliest papers proposing joint models for longitudinal and survival data, DeGruttola *et al* suggested that for longitudinal data measured at a small set of common time points, a saturated mean and covariance model could be used for the longitudinal measurements, thus making no assumptions such as linearity over time for the trajectory [126].

The setup used in our simulations of Section 9.8 takes up these suggestions, by assuming that the longitudinal process is assumed to follow a piece-wise constant trajectory, with the elements of \mathbf{X}_i representing the constant levels in the time periods 0 to 10, 10 to 20, 20 to 30, 30 to 40, 40 to 50. In certain situations, although the occurrence of the time-to-event outcome precludes further longitudinal error-prone measurements, it may not preclude the existence of the longitudinal process after the event. The most obvious example is when the event of interest is drop-out

from a clinical trial or observational study. In this case, assuming subjects are still alive, the longitudinal process of interest still exists after the event occurs, even if it is not observable. In this case, a model of the kind used in our simulations may be appropriate.

In some situations however, the occurrence of the event of interest may terminate the existence of the longitudinal process. In Chapter 13 we consider models for the relationship between SBP levels and the hazard of cardiovascular disease (either non-fatal or death due to cardiovascular disease). After death, a person's blood pressure does not exist, and is therefore not a well defined quantity. Applying the model used in our simulations to this setting, for subjects who 'die' between $t = 20$ and $t = 30$, what do the random-effects X_{i4} and X_{i5} represent? From one perspective, our modelling assumptions for the 'complete data', which include \mathbf{X}_i , imply that the SBP trajectory for a particular subject is defined at all times (e.g. from $t = 0$ to $t = 50$), irrespective of if, and when, the event of interest occurs. For a particular subject who dies at $t = 25$, we may then view the random-effects X_{i4} and X_{i5} as representing a subject's SBP levels in the periods 30 to 40 and 40 to 50, in the hypothetical situation in which death does not preclude further observation (and indeed existence) of SBP. In settings such as this, a legitimate question is whether use of models which assume the existence of such quantities is appropriate.

We now give a brief argument explaining why, at least in terms of parameters of the survival model, using estimation methods which assume the existence of the vector \mathbf{X}_i when the event of interest terminates the existence of the longitudinal process, may still be reasonable. To illustrate our argument, we continue with the setup used in the simulations of Section 9.8. The data in our simulations were generated in the following way:

1. Generate $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5})^T \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$
2. Generate survival time T_i according to model with hazard at time t given by $0.1 \exp(\beta_{X_1^*} M_i(t - 10) + \beta_{X_2^*} M_i(t - 20))$ (and zero hazard for $t < 20$), as previously described in Section 9.8
3. Set $V_i = T_i$, $Y_i = 1$, unless $T_i > 50$, in which case set $V_i = 50$, $Y_i = 0$
4. Generate times longitudinal error-prone measurements as previously described in Section 9.8

This underlying data-generating model is consistent with the modelling framework we have used in this chapter. In particular, the vector \mathbf{X}_i is well-defined for all subjects. The occurrence of the event of interest merely prevents any further observation of the longitudinal process. The observed data consist of (V_i, Y_i, \mathbf{W}_i) (plus the times of the measurements \mathbf{W}_i). Now consider the following, alternative, data-generating model:

1. Generate $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3})^T \sim N(\boldsymbol{\mu}_{X_{1-3}}, \boldsymbol{\Sigma}_{X_{1-3}})$, i.e. from a multivariate normal distribution with mean and covariance matrices given by the appropriate sub-vector/matrix of $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$
2. Generate survival up to time $t = 30$ according to an exponential distribution with hazard $0.1 \exp(\beta_{X_1^*} X_{i2} + \beta_{X_2^*} X_{i1})$
3. For those subjects who survived to $t = 30$, generate $X_{i4}|X_{i1}, X_{i2}, X_{i3}$ from a normal distribution, with conditional mean and variance implied by $\boldsymbol{\mu}_{X_{1-4}}$ and $\boldsymbol{\Sigma}_{X_{1-4}}$ from standard results for the multivariate normal
4. For those subjects who survived to $t = 30$, generate survival from $t = 30$ to time $t = 40$ according to an exponential distribution with hazard $0.1 \exp(\beta_{X_1^*} X_{i3} + \beta_{X_2^*} X_{i2})$
5. For those subjects who survived to $t = 40$, generate $X_{i5}|X_{i1}, X_{i2}, X_{i3}, X_{i4}$ from a normal distribution, with mean and variance implied by $\boldsymbol{\mu}_X$ and $\boldsymbol{\Sigma}_X$
6. For those subjects who survived to $t = 40$, generate survival from $t = 40$ to time $t = 50$ according to an exponential distribution with hazard $0.1 \exp(\beta_{X_1^*} X_{i4} + \beta_{X_2^*} X_{i3})$
7. Set $V_i = T_i$, $Y_i = 1$, unless $T_i > 50$, in which case set $V_i = 50$, $Y_i = 0$
8. Generate times longitudinal error-prone measurements as previously described in Section 9.8

The observed data again consist of (V_i, Y_i, \mathbf{W}_i) (plus the times of the measurements \mathbf{W}_i).

These two different data-generating models give rise to observed data with the same joint distribution. From this it follows that whatever estimation procedure is applied to the observed data, the same values are consistently estimated irrespective of which underlying data-generating model is correct. In particular, this suggests that we can obtain valid estimates of the parameters of the survival model, which is the same in the two data-generating models, by using estimation methods which are based on the first data-generating model, even if the second data-generating model more closely matches our belief of how the data arise.

The implications for the parameters of the longitudinal model are somewhat different however. In the first data-generating model, the random-effects vector $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4}, X_{i5})^T$ is defined for all subjects, with corresponding population mean $\boldsymbol{\mu}_X$ and covariance matrix $\boldsymbol{\Sigma}_X$. In the second data-generating model, X_{i4} and X_{i5} are only defined for those subjects who survive to $t = 30$ and $t = 40$ respectively, and their means in particular (in these subsets of the original population) differ from the corresponding elements of $\boldsymbol{\mu}_X$. For example, if SBP is associated with increased

hazard, those who survive to $t = 30$ will tend to be subjects with lower values of X_{i1} and X_{i2} (which affect survival from $t = 20$ to $t = 30$).

Consider for example, the parameter corresponding to $\mathbb{E}(X_{i4})$. Estimates of this parameter from either RC or the ML approach (which assume the first data-generating model) are estimates of the corresponding elements of the marginal mean vector $\boldsymbol{\mu}_X$. In particular, they are not estimates of the mean of X_{i4} in survivors to $t = 30$. For those subjects who die prior to $t = 30$, both RC and ML implicitly impute what the value of X_{i4} would have been on the basis of X_{i1} , X_{i2} and X_{i3} (which themselves are implicitly imputed based on error-prone measurements, and in addition, for ML, using the fact the subject had died). This imputation is made using the relationship between X_{i4} and the preceding levels as estimated using data from those who have longitudinal error-prone measurements between $t = 30$ and $t = 40$. In situations in which the event of interest may terminate the existence of the longitudinal process, we therefore do not believe the estimates of parameters relating to the longitudinal process at later times correspond to meaningful population parameters. The estimate of $\mathbb{E}(X_{i4})$ for example can be viewed as a weighted mean of the X_{i4} in the fourth decade in those who survive to the start of the decade and the value of X_{i4} that would have occurred in those who died prior to $t = 30$, assuming the relationship between X_{i4} and (X_{i1}, X_{i2}, X_{i3}) is the same as it was in the sub-group who survived to $t = 30$.

This issue has been recently addressed in a paper by Kurland *et al* [138], who considered the slightly different objective of the analysis of longitudinal data which is subject to missingness due to death. They similarly concluded that methods which marginalize over the missing data patterns are usually inappropriate when data are missing due to death, since, as we have described, such methods implicitly impute the missing data after death, based on the relationship estimated in those who do not die. Further research is needed to investigate whether use of models such as those considered in this chapter are appropriate when the focus is instead on the relationship between the longitudinal process and the occurrence of the event of interest.

9.10 Conclusions

The extension of estimation methods from the classical measurement error to the longitudinal setting is arguably most complicated (except for what we described in Section 9.1 as situation 1, in which covariates are time-independent) in the case of time-to-event outcomes, compared to continuous or binary outcomes. This is due to the fact that in the longitudinal setting the event process is often observed concurrently with the longitudinal process, leading to models for the event of interest in which the longitudinal process enters as one or more time-dependent covariates.

9.10.1 Regression calibration

Mirroring the results of our simulations for time-independent covariates measured with classical error, our simulation results suggest that an RC approach (based on a single linear mixed model fit) may carry substantial biases when the time-dependent covariate effects are moderate or large. For a time-independent covariate measured with classical error, our simulation results (Section 5.8) suggested that recalibrating in each risk-set removed much of the bias of the RC estimator. In our simulations using a piece-wise constant model for the longitudinal process, while the bias of the risk-set RC estimator was less than that of RC, considerable bias remained when the effects of the time-dependent covariates were moderate. Despite the approximate nature of RC, its continued popularity is due to the fact that it can be easily implemented using routines in standard statistical packages.

Our simulation results confirmed that if RC is used to estimate the parameters of a Cox proportional hazards model in which one or more influential time-dependent covariates are omitted, bias results in general. In our simulations such biases acted in the opposite direction of the inherent bias in the RC estimator, resulting in less overall bias. This will obviously not always be the case. For example, had the effect of the omitted time-dependent covariate on hazard been in the opposite direction to the effect of the included covariate, the bias would have been in the same direction as that inherent in RC estimates for Cox models (i.e. towards the null). We note that, whereas for continuous and binary outcomes, we were able to construct a ‘corrected RC’ estimator, such a simple correction is not possible in the case of a Cox proportional hazards outcome model with time-dependent covariates. This is because the time-dependent covariates do not usually have a constant variance covariance matrix as time moves forward, as they do in the case of time-fixed covariates.

9.10.2 Maximum likelihood

Maximum likelihood methods have received a large amount of attention in the context of joint models of time-to-event outcomes and longitudinal processes measured periodically with error. However, a barrier to wider use has been the lack of implementation in standard statistical packages. Developments such as the JM package in R should help encourage more widespread usage, although this is limited to models with a single time-dependent covariate.

9.10.3 Maximum likelihood using ascent-based Monte-Carlo Expectation Maximization and multiple imputation

We have shown how MCEM can be implemented for joint models with a Cox proportional hazards outcome model by using rejection sampling to generate samples from the conditional distribution of the random-effects \mathbf{X}_i given the observed data.

However, as discussed in Section 9.5, the bound used in rejection sampling is a very loose one. Consequently, depending on the number of subjects in the dataset (which affects the magnitude of the jumps in the non-parametric estimate of the cumulative baseline hazard function), it may take thousands of proposed draws before one is found which satisfies the stated inequality. This means that rejection sampling is even more computationally intensive than in the case of time-independent covariates measured with classical error. As discussed in Chapter 8, the algorithm could certainly be implemented more efficiently from a computational perspective by programming the rejection sampler in a language such as C++. Since R is an ‘interpreted’ package, loops are performed extremely slowly in comparison to compiled code. Nevertheless, our simulation results suggest that use of rejection sampling, as part of ascent-based MCEM, can be used to find consistent estimates of the parameters of the Cox proportional hazards model. Furthermore, unlike quadrature methods, the method may be expected to still be feasible for models in which the dimension of the random-effects vector \mathbf{X}_i is larger.

A benefit of using the proposed rejection sampling method and estimating the model parameters using ascent-based MCEM is that upon convergence, a large number of multiple imputations of \mathbf{X}_i is available as a by-product of the algorithm. As we have shown in our simulations, these can be used to estimate the parameters of alternative outcome models with different specifications for the time-dependent covariates \mathbf{X}_i^* , provided that the model is nested within the model used to generate the imputations. Of course, the validity of such estimates still rests on the assumption that the model used to create the imputations is correctly specified.

9.10.4 Conditional score method

Unfortunately we were unable to implement the CS method in our survival simulations due to time constraints. The CS method provides consistent estimates without requiring any distributional assumptions for the random-effects \mathbf{X}_i . The price for not needing to make a distributional assumption for \mathbf{X}_i is inefficiency compared to the correctly specified MLE. In the case of binary outcomes, a subject can contribute to estimation in the CS method if and only if the design matrix \mathbf{D}_i of their longitudinal error-prone measurements is of full-rank. In the case of a time-to-event outcome with time-dependent covariates, the issue is slightly more subtle, as previously described. A subject’s event can contribute only if the value of the time-dependent covariates are identifiable for the subject. At each event time, the risk-set consists of those subjects who were both at risk and for whom the time-dependent covariates are identifiable. Depending on model specification and the availability of longitudinal error-prone measurements, these requirements may mean that few events can contribute to the analysis, and that in the events that can, few subjects contribute to the risk-set. However, in such cases, Tsiatis and Davidian commented

(in the context of the random-intercepts and slopes model) that attempting to estimate the model parameters ‘may be a fruitless enterprise regardless of estimation method’ [105]. As we have previously discussed, we believe further efforts are needed to make such methods more widely used, including coding the method into packages or commands that users can use off the shelf, and perhaps making more accessible explanations of the methodology.

9.10.5 Modelling assumptions

As discussed in Section 9.9, much of the statistical literature concerning joint models of longitudinal and survival data have assumed simple random-intercepts and slopes models for the longitudinal process. Others have proposed models which, in addition to subject-specific random-effects, allow the longitudinal process trajectory to be affected by a stochastic process. Estimation for such models is an order of magnitude more difficult however. The use of spline type models, which can be expressed as random-effects models, has recently been advocated as a potential compromise. For settings in which the longitudinal process is no longer well defined after the occurrence of the event of interest (e.g. death, for many longitudinal processes), it is not entirely clear that use of methods which assume a model for the longitudinal process at all times (i.e. even after the occurrence of the event of interest) is appropriate. In such situations, we have argued that although the methods we have considered in this chapter may not give estimates of meaningful parameters in the longitudinal model, they may be appropriate when interest lies in the relationship between the longitudinal process and the occurrence of the event of interest. Further research regarding these issues is needed.

Part III

Systolic blood pressure and cardiovascular disease - analyses of data from the Framingham Heart Study

In the following chapters we report the results of two analyses of data from the Framingham Heart Study, with the aim of illustrating the application of some of the estimation methods we have previously described to a real dataset. In Chapter 10 we describe the Framingham Heart Study and summarize all-cause mortality and mortality due to cardiovascular disease of the subjects in Framingham. In Chapter 11 we describe the protocol for measurement of blood pressure in the study, and summarize the available blood pressure measurements in the men recruited into the Framingham Study. In Chapters 12 and 13 we use the data from the men in the Framingham Study to illustrate some of the estimation methods described in Chapters 8 and 9 respectively. In Chapter 12 we report the results of analyses investigating the associations between risk of death due to cardiovascular disease between age 70 and 80 (a binary outcome) and underlying systolic blood pressure at earlier ages. In Chapter 13 we report analyses examining how the hazard of experiencing a cardiovascular event (a time-to-event outcome) between ages 60 and 80 depends on subjects' current and past underlying systolic blood pressure. These analyses are intended to be illustrative of some of the methods we have previously described, and are not in any way intended to be a definitive epidemiological analysis of the Framingham Study data.

Chapter 10

The Framingham Heart Study

In this chapter we describe the background to the Framingham Heart Study, describe the baseline characteristics of the cohort, and summarize their all-cause and cardiovascular mortality. The Framingham Heart study was one of the earliest large scale epidemiological studies to be initiated, and was set up to investigate the epidemiology of cardiovascular disease (CVD) [14]. The study was conducted in the town of Framingham in the state of Massachusetts, in the USA. The aim of the study was to investigate risk factors for the development of CVD. In total, 5,209 subjects were recruited to the study, between 1948 and 1951. Subjects were then invited to return to a hospital clinic every two years following their initial baseline examination. At each follow-up visit, a medical history, physical examination, blood studies and other laboratory tests were performed. Symptoms of illness that had occurred since the previous follow-up visit were recorded, including any interim hospitalization or medical visits. Inevitably, at each follow-up visit some subjects did not attend, causing missing data in the variables measured at each visit. Furthermore, the protocol specifying which variables were to be measured varied as the study progressed, as new variables of interest were added and variables no longer thought to be of interest were removed.

The limited access Framingham dataset

Our analyses are based on the ‘limited access’ Framingham Heart Study data. The limited access dataset is provided by the US National Heart Lung and Blood Institute (NHLBI) to researchers who wish to use the Framingham Heart Study data for epidemiological or methodological research. Of the 5,209 subjects in the Framingham Study, 130 subjects did not give consent for the data to be shared. The limited-access dataset provided by the NHLBI thus includes data on 5,079 subjects. Furthermore, the dataset does not contain variables which may potentially enable study participants to be identified. This means for example that dates of when subjects were recruited to the study or when their follow-up visits took place are

not available. Instead, variables are provided which specify how many days elapsed between a subject's baseline examination and his/her subsequent follow-up visits.

10.1 Study recruitment procedure

We briefly describe the inception and recruitment procedure used in the Framingham Study, based on the book by Dawber [14]. The town of Framingham was chosen because it was considered at the time to be a relatively self-contained community. In 1948, Framingham had an estimated population of 28,000. Based on the expected rates of CVD and the expected strengths of association between risk factors, it was decided that between 5,000 and 6,000 subjects would be required to estimate associations between risk factors and CVD with reasonable precision. Based on the town census list, the study investigators drew up a list of each family and arranged them alphabetically. From this, two thirds of the families were selected, who together formed a group of 6,507 subjects who were invited to participate in the study. The study investigators anticipated that almost all selected subjects would respond to the invitation to come for the initial examination. A community organisation was established, which involved giving an initial examination to subjects who volunteered to be involved (as opposed to being selected). These volunteers were in turn asked to approach some of the selected individuals to try and encourage them to participate in the study. As a result, 4,469 (68.7%) of the subjects who were selected to participate did attend the initial examination. In addition, the study included 740 subjects who volunteered to be involved in the study, giving a total sample of 5,209 subjects. The fact that 31.3% of those invited to participate chose not to potentially limits the representativeness of the sample. Based on interviewing a sample of the nonparticipants, Dawber concluded that the main reason that people did not want to participate was for fear of medical conditions being discovered in the examinations.

10.2 Characteristics at the initial examination

Table 10.1 shows descriptive statistics for a number of variables recorded at entry to the Framingham Study. Bearing in mind that study recruitment took place between 1948 and 1951, many biochemical variables which would be measured in a modern epidemiological study were not measured. More women than men were recruited into the study. The youngest subject was just under 29 years of age at recruitment, while the oldest was just under 63 years of age. The age at entry was reasonably uniformly distributed between the ages of 30 and 60, and was well balanced between men and women. As expected, the men were taller and heavier than the women on average. The mean systolic blood pressure (SBP) was similar

Table 10.1: Descriptive statistics for 5,079 Framingham subjects at entry to study

Variable (units)	Number missing	Mean (SD)	
		Males (n=2,294)	Females (n=2,785)
Age (years)	0	44.7 (8.6)	44.6 (8.6)
Height (cm)	6	171.6 (6.9)	159.0 (6.1)
Weight (kg)*	6	76.2 (11.5)	64.5 (11.9)
Systolic blood pressure (mm Hg)**	0	137.1 (20.7)	137.2 (26.2)
Diastolic blood pressure (mm Hg)**	0	86.3 (12.5)	84.8 (13.4)
Cholesterol (mg / dL)***	612	227.2 (42.5)	229.7 (46.8)

* - weight was recorded in 5lb categories

** - blood pressure designated as first physician measurement at baseline examination (see Chapter 11)

*** - due to large number of missing cholesterol values at baseline examination, mean (SD) is based on cholesterol values from the first follow-up visit (visit 2)

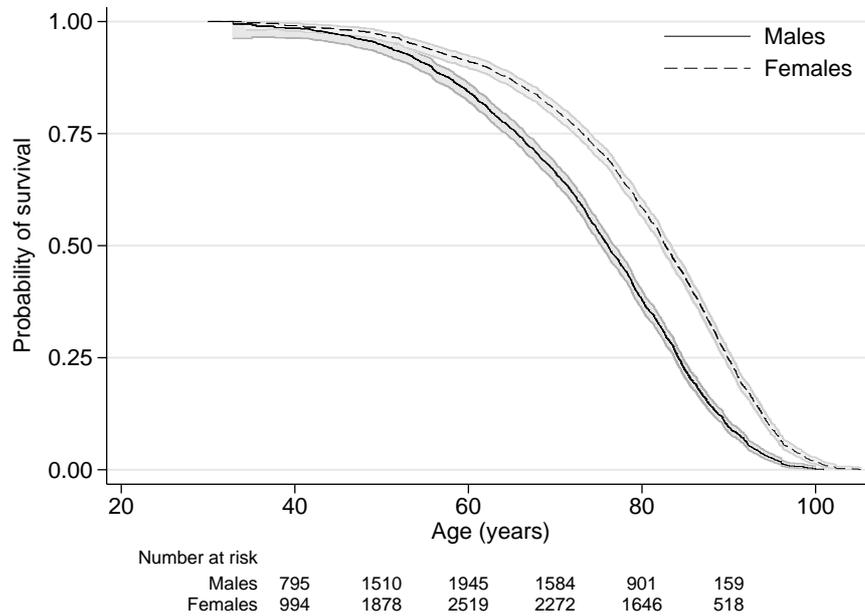
in men and women, but there was greater variability in the women. The mean diastolic blood pressure (DBP) was slightly higher in men than women, but again the variability was greater in women. The mean SBP and DBP are in the range that today is classified as high-normal [139]. Cholesterol was measured at the initial examination, but only in 3,092 subjects. Table 10.1 therefore shows the mean and standard deviation cholesterol based on measurements at subjects' first follow-up visit (visit 2). The mean cholesterol level is also slightly above the level which would currently be considered ideal [139].

10.3 All-cause mortality

The dataset we used has survival information up to 31st December 2003. Four subjects were lost to follow-up immediately following their initial examination at entry to the study. Of the 5,079 subjects, 4,483 were known to have died by 31st December 2003. The limited-access dataset does not contain the date of recruitment to the study. However, of the 596 subjects who have been censored, 528 (88.6%) have been censored at least 50 years after recruitment, suggesting that relatively few subjects have been censored due to loss to follow-up.

Since age has a strong influence on mortality risk we examine mortality on the age time scale. In Figures 10.1 and 10.2 we show the Kaplan-Meier curve (conditional on being alive at age 30) and smoothed estimate of the hazard function. The Kaplan-Meier plot shows that the probability of survival to age 100 was close to zero for both men and women, but that women survived on average longer than men. For example, for men the estimated probability of survival to age 70 was 0.663 (95% CI 0.641 to 0.685), while for women the estimated probability was 0.803 (95% CI 0.786

Figure 10.1: Kaplan-Meier curve for all-cause mortality survival using data from 5,079 subjects in the Framingham Study, conditional on survival to age 30, with 95% pointwise confidence bands



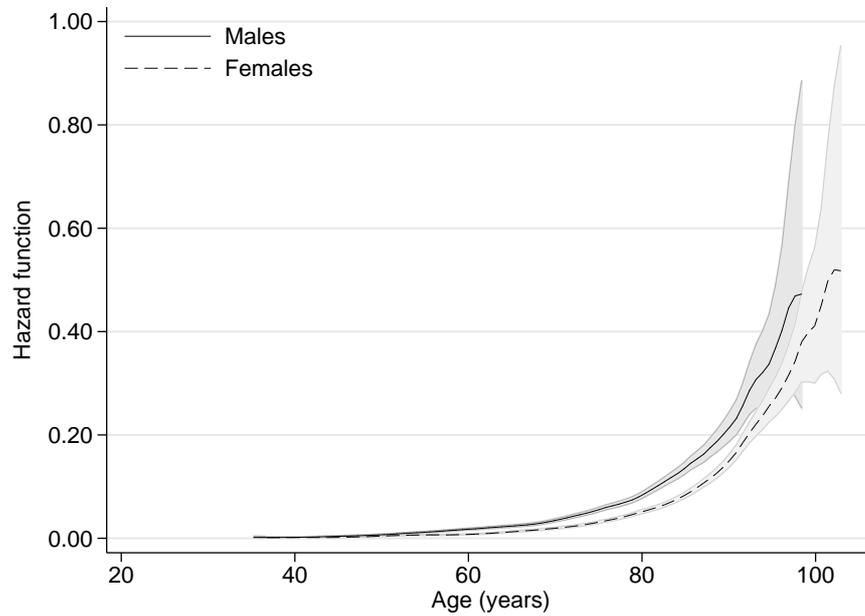
to 0.819). Figure 10.2 shows, as expected, that the hazard of death increased with age, and also that the rate accelerated with age. The plots were produced using Stata's `sts graph` command, with the default Epanechnikov kernel used to create a smoothed estimate of the hazard function.

10.4 Death due to cardiovascular disease

For the purposes of examining cause-specific mortality, we used follow-up data up to 18,600 days after baseline, because (in the dataset we used) cause of death has been ascertained by the study's event review committee for all but one of the subjects who have died by that follow-up time. Censoring all subjects at this time meant that we discarded the known time and cause of death from 129 subjects, and the known time of death from 46 subjects, leaving 4,308 deaths.

We categorised deaths into those due to cardiovascular disease (CVD), those due to cancer, and those due to any other cause (including deaths of unknown cause). Of the 4,308 deaths, 1,687 (39.2%) were due to CVD, 1,016 (23.6%) were due to cancer, and 1,605 (37.3%) were due to other causes. Figure 10.3 shows the estimated cumulative incidence function for death due to CVD, separately for men and women. The cumulative incidence function gives the absolute probability of death due to a cause of interest by a particular time, properly allowing for the risk of death due to other causes [133]. The figure shows that given survival to age 30, men were more likely to die from CVD by any future age compared to women. Figure 10.4 shows

Figure 10.2: Estimated hazard function for all-cause mortality using data from 5,079 subjects in the Framingham Study, using the Epanechnikov kernel, with 95% pointwise confidence bands



the estimated cause-specific hazard function for death due to CVD. This shows that the cause-specific hazard for death due to CVD was higher for men compared to women, with the hazard rate accelerating with age in both groups. The plots were produced using Stata's `stcompet` and `sts` graph commands.

Figure 10.3: Estimated cumulative incidence of death due to CVD, conditional on survival to age 30, using data from 5,078 subjects in the Framingham Study, with 95% pointwise confidence bands

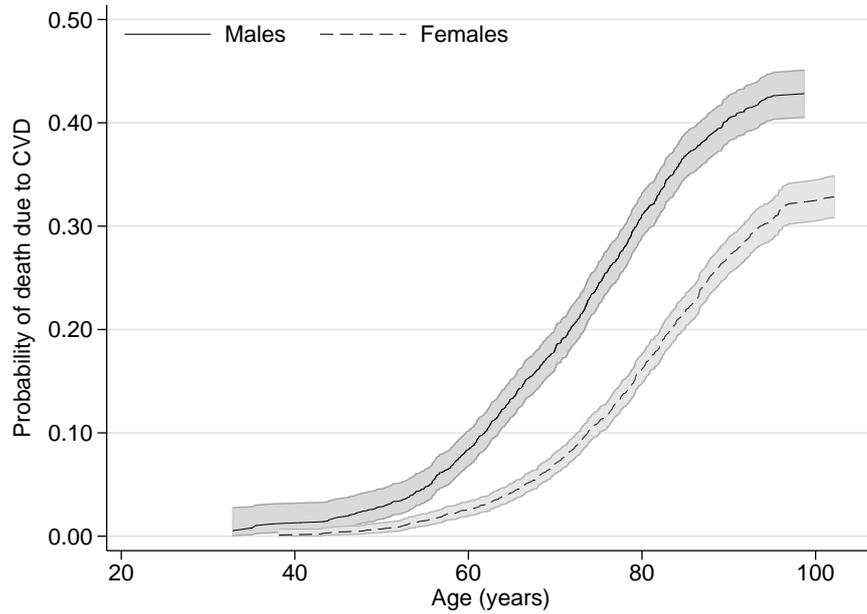
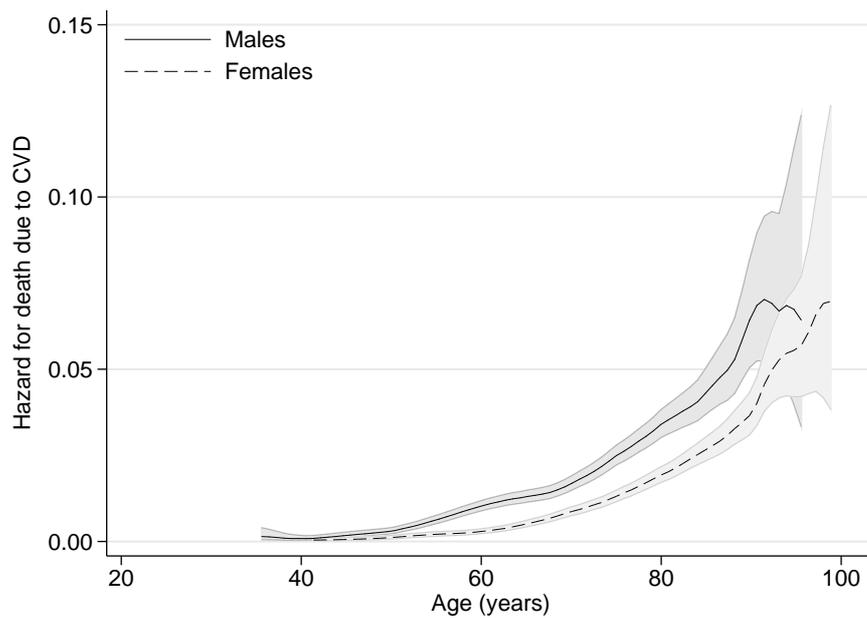


Figure 10.4: Estimated cause-specific hazard function for death due to CVD using data from 5,078 subjects in the Framingham Study, using the Epanechnikov kernel, with 95% pointwise confidence bands



Chapter 11

Systolic blood pressure measurements in men

In this chapter we first describe how blood pressure (BP) was measured in the Framingham study, the schedule for follow-up visits at which blood pressure was due to be measured and how closely this schedule was adhered to (Section 11.1). In Section 11.2 we investigate the variability of systolic blood pressure (SBP) measurements within a single follow-up visit, and consider how the multiple measurements within a follow-up visit can be summarized, with the aim of reducing the modelling complexity and computational demands of our subsequent analyses. Upon concluding that using the mean of the three measurements of SBP that were made (at most visits) is reasonable, in Section 11.3 we examine the extent to which the resulting SBP measurements satisfy the classical measurement error assumptions, when viewed as error-prone measurements of a subject's current underlying SBP.

In this and the following chapters we use only data from the men in the Framingham study. We do this for a number of reasons. First, it has been found that there are differences in the trajectories of SBP between men and women [14], and as shown in Chapter 10, men and women differ substantially in terms of both all-cause and CVD mortality. By using data only from the men in the study, we need not concern ourselves with modelling such differences between men and women. Second, using data only from the men reduces the computational demands of our analyses, which as previously noted, are intended to be illustrative of some of the estimation methods previously described and proposed.

11.1 Systolic blood pressure measurements in the Framingham study

Systolic (SBP) and diastolic (DBP) blood pressure (BP) were measured in the Framingham study using a mercury manometer, in the left arm, with the subject seated and the arm at heart level. Blood pressure was measured when subjects were re-

cruited into the study and subsequently at each two yearly follow-up visit. At most visits subjects had an initial BP measurement performed by a nurse, followed by two measurements by physicians (sometimes the same physician for both measurements). However, as one might expect given the duration of the Framingham study, there is some heterogeneity in the number and type of BP measurements which are available between the visits.

At their baseline examination, subjects who were recruited prior to April 1950 only had a BP measurement performed by a nurse and a single BP measurement performed by a physician. Subjects recruited after April 1950 had three measurements: one nurse measurement, one measurement was performed by the first examining physician and another measurement was performed by a second physician. Unfortunately, for those subjects who were recruited in the earlier period, their nurse BP measurement was recorded in the dataset as their first physician measurement (according to the documentation accompanying the dataset). It is not possible, using the limited access dataset we have, to identify which subjects this applies to. At visit 4, no nurse BP measurements were made, and for 966 of the men their second physician measurement was made by the same physician as their first physician measurement (personal communication with NHLBI). At the time of visit 11 the National Institute for Health funding for the Framingham study ended, meaning that many of the subjects did not have a follow-up visit at this time (personal communication with NHLBI). Subsequently funding was obtained and the study resumed as normal, hence the limited number of subjects with a SBP measurement at visit 11. At visit 24, there was also no nurse measurement, but instead each physician took two BP measurements. Lastly, in the documentation for visit 26, it is stated that for 'off-site' visits, the physician BP measurements were performed by a technician. Off-site visits are where subjects had their visit measurements performed at their home or care facility, rather than in hospital.

The numbers of SBP measurements available from the men in the Framingham study by follow-up visit are given in Tables 11.1 and 11.2. In total, there are 82,935 SBP measurements from the 2,294 men in Framingham. The tables show that, with the exception of visits 1, 4 and 11, of those subjects who have any SBP measurements at a given visit, the vast majority have three SBP measurements. We also see that the proportion of men who were still alive at the time of a scheduled visit but who did not have any SBP measurements steadily increased throughout the study. For visits 1-13, the proportion of at risk men with no SBP measurements at each visit increased from around 10% to around 17%. By visit 26 the proportion increased to around 35%.

Table 11.3 shows the number of missed follow-up visits by age. For the purposes of this table a missed follow-up visit for a particular man means no SBP measurement was recorded for the man within one year of the scheduled time of the follow-up

Table 11.1: Number of systolic blood pressure measurements at follow-up visits 1-13 for men in the Framingham study

Time since study entry (years)	Visit number	Number at risk†	Number of men with given measurement			Number (%) of men with given number of measurements			
			Nurse	1st physician	2nd physician	0	1	2	3
0	1	2,294	1,468	2,294 *	2,119	0 (0)	62 (2.7)	877 (38.2)	1,355 (59.1)
2	2	2,271	1,998	2,091	1,985	180 (7.9)	3 (0.1)	193 (8.5)	1,895 (83.4)
4	3	2,230	969	1,921	1,796	309 (13.9)	86 (3.9)	905 (40.6)	930 (41.7)
6	4	2,197	**	1,971	1,750 ***	226 (10.3)	221 (10.1)	1,750 (79.7)	0 (0)
8	5	2,141	1,878	1,904	1,704	237 (11.1)	5 (0.2)	216 (10.1)	1,683 (78.6)
10	6	2,085	1,801	1,821	1,751	264 (12.7)	1 (0.0)	88 (4.2)	1,732 (83.1)
12	7	2,025	1,763	1,771	1,741	253 (12.5)	1 (0.0)	39 (1.9)	1,732 (85.5)
14	8	1,957	1,696	1,698	1,662	255 (13.0)	2 (0.1)	46 (2.4)	1,654 (84.5)
16	9	1,890	1,605	1,608	1,579	281 (14.9)	1 (0.1)	33 (1.7)	1,575 (83.3)
18	10	1,805	1,494	1,502	1,494	303 (16.8)	1 (0.1)	14 (0.8)	1,487 (82.4)
20	11	1,721	639	1,210	642	510 (29.6)	563 (32.7)	16 (0.9)	632 (36.7)
22	12	1,627	1,336	1,339	1,336	288 (17.7)	1 (0.1)	4 (0.2)	1,334 (82.0)
24	13	1,520	1,272	1,273	1,266	247 (16.3)	0 (0)	8 (0.5)	1,265 (83.2)

† - number of men who were still alive at the visit's scheduled time and had not been lost to follow-up

* - visit 1 - the first subjects to be recruited to Framingham had only a nurse and single physician measurement. For these subjects, their nurse measurement was classified as the 1st physician measurement, but we are not able to identify which subjects this applies to.

** - visit 4 - no nurse measurements were recorded at visit 4

*** - visit 4 - for 966 men their second physician measurement was made by the same physician as their first measurement

Table 11.2: Number of systolic blood pressure measurements at follow-up visits 14-26 for men in the Framingham study

Time since study entry (years)	Visit number	Number at risk†	Number of men with given measurement			Number (%) of men with given number of measurements			
			Nurse	1st physician	2nd physician	0	1	2	3
26	14	1,411	1,142	1,142	1,135	268 (19.0)	0 (0)	10 (0.7)	1,133 (80.3)
28	15	1,298	1,008	1,010	1,005	287 (22.1)	0 (0)	10 (0.8)	1,001 (77.1)
30	16	1,188	897	898	890	290 (24.4)	1 (0.1)	7 (0.6)	890 (74.9)
32	17	1,078	797	797	788	277 (25.7)	2 (0.2)	17 (1.6)	782 (72.5)
34	18	961	659	684	683	277 (28.8)	0 (0)	26 (2.7)	658 (68.4)
36	19	841	547	559	559	282 (33.5)	0 (0)	12 (1.4)	547 (65.0)
38	20	734	470	499	499	235 (32.0)	0 (0)	29 (4.0)	470 (64.0)
40	21	630	421	430	429	200 (31.7)	0 (0)	10 (1.6)	420 (66.7)
42	22	533	362	367	367	166 (31.1)	0 (0)	5 (0.9)	362 (67.9)
44	23	452	263	305	305	147 (32.5)	0 (0)	42 (9.3)	263 (58.2)
46	24	368	*	224, 218	193, 190	132 (35.9)	0 (0)	55 (14.9)	9 (2.4) **
48	25	309	177	215	215	94 (30.4)	0 (0)	38 (12.3)	177 (57.3)
50	26	243	114 ***	164 ***	164	79 (32.5)	0 (0)	50 (20.6)	114 (46.9)

† - number of men who were still alive at the visit's scheduled time and had not been lost to follow-up

* - visit 24 - there were no nurse measurements, but each of the physician measurements was repeated, giving up to four measurements for each man

** - visit 24 - 172 (46.7%) men had four measurements

*** - visit 26 - for offsite visits, 'physician' blood pressures were taken by technicians

Table 11.3: Missing follow-up visits by age for 2,294 men in the Framingham study

Age (years)	Number (% of men at risk) of men who missed given number of visits			
	0	1	2	3
30-35	355 (94.4)	20 (5.3)	1 (0.3)	0 (0)
35-40	719 (89.4)	54 (6.7)	20 (2.5)	11 (1.4)
40-45	1,045 (86.3)	101 (8.3)	40 (3.3)	25 (2.1)
45-50	1,330 (86.0)	102 (6.6)	90 (5.8)	24 (1.6)
50-55	1,573 (83.6)	143 (7.6)	100 (5.3)	65 (3.5)
55-60	1,724 (82.8)	175 (8.4)	127 (6.1)	55 (2.6)
60-65	1,564 (77.7)	229 (11.4)	126 (6.3)	95 (4.7)
65-70	1,324 (72.9)	262 (14.4)	170 (9.4)	60 (3.3)
70-75	1,076 (67.9)	261 (16.5)	140 (8.8)	107 (6.8)
75-80	777 (61.5)	260 (20.6)	176 (13.9)	51 (4.0)
80-85	556 (61.7)	195 (21.6)	107 (11.9)	43 (4.8)
85-90	326 (67.5)	106 (22.0)	40 (8.3)	11 (2.3)
90-95	116 (73.0)	30 (18.9)	9 (5.7)	4 (2.5)
95-100	26 (83.9)	5 (16.1)	0 (0)	0 (0)

visit (i.e. from one year preceding the scheduled time to one year following the scheduled time), provided they were alive and not lost to follow-up one year after the scheduled visit time. This shows that the proportion of men who missed at least one visit increased steadily with age, from around 5% at ages 30-35 to around 40% by age 75-80. The number of men missing two or three follow-up visits is relatively small compared to the number missing a single visit. This suggests that when men missed a visit, they tended to return for the subsequent visit, so that few men are missing two or three consecutive visits. We note that the table also indicates that relatively few men have SBP measurements in the earliest age band of 30-35 years. This is of course a reflection of the fact that the majority of men were recruited into the study at ages greater than 35.

11.2 Data reduction

In this section we consider whether the multiple SBP measurements made within a follow-up visit can be reasonably summarized by a smaller dimensional function, e.g. by their mean. Doing so will reduce the complexity of models for SBP, since we will not need to model within-visit variability. It will also reduce the effective number of error-prone measurements, and thus decrease the time required to fit models.

To investigate within-visit variability we used the SBP measurements from visit 2, since more SBP measurements are available compared to visit 1, and because whether each measurement was taken by a nurse or physician was recorded (in contrast to visit 1, as previously described). Table 11.4 shows the mean and SDs of

Table 11.4: Mean and SDs (95% CIs) of systolic blood pressure (mmHg) measurements from visit 2 of the Framingham study, based on data from 1,895 men who had all three measurements available

Measurement	Estimated mean SBP	SD
Nurse	134.72 (133.76, 135.67)	21.26 (20.23, 22.45)
Physician 1	133.17 (132.23, 134.11)	20.82 (19.78, 21.85)
Physician 2	130.79 (129.85, 131.72)	20.76 (19.66, 22.06)

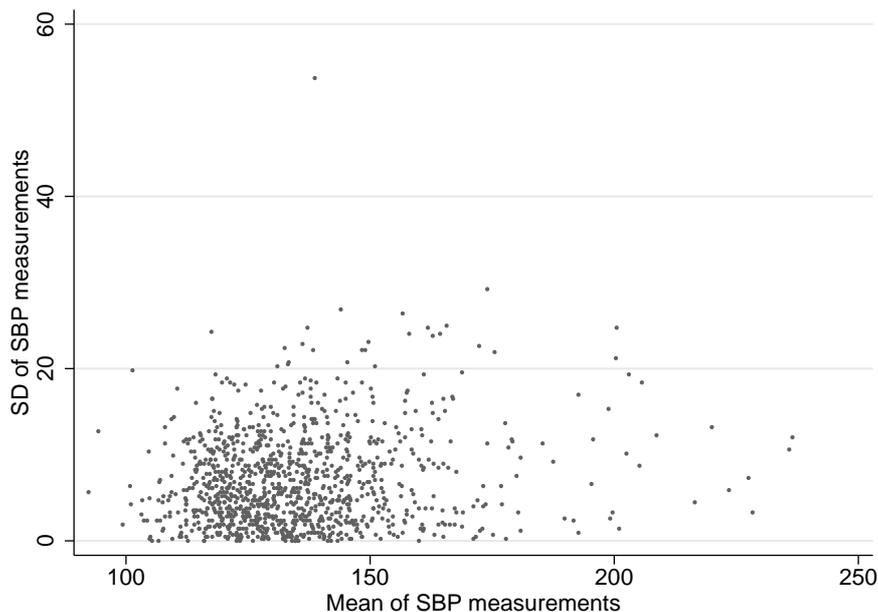
Table 11.5: Estimated correlations (95% CIs) between systolic blood pressure measurements from visit 2 of the Framingham study, based on data from 1,895 men who had all three measurements available

	Nurse	Physician 1	Physician 2
Nurse	1		
Physician 1	0.844 (0.825, 0.860)	1	
Physician 2	0.821 (0.801, 0.842)	0.867 (0.852, 0.881)	1

the SBP measurements from visit 2, using data only from the 1,895 men who had a nurse and two physician measurements. Although each of the pair-wise differences in means is statistically significant (each $p < 0.001$), the differences are small in magnitude. Similarly, the SDs of the three measurement types differ only slightly. Table 11.5 shows the estimated correlations between the three measurements at visit 2, which indicates that measurements within subjects are highly correlated, as we would expect. The estimated correlation between the two physician measurements is slightly higher than that between the nurse measurement and either of the physician measurements, but the difference is relatively small.

On the assumption that within-visit variability in SBP is neither of interest nor related to the incidence of CVD, we decided to use the mean of a subject’s SBP measurements in our analyses. However, as shown in Tables 11.1 and 11.2, the number of SBP measurements available varies between subject and between the different follow-up visits. The large correlations between SBP measurements within visit 2 suggest that there is little variation between SBP measurements at visit 2, relative to between-subject variability. However, the within-visit variations are not necessarily small relative to within-subject changes in underlying SBP over time. We therefore decided to use the mean of a fixed number of SBP measurements, rather than the mean of a subject’s available measurements. This ensured that SBP ‘observations’ had constant variance. At each follow-up visit, of those men who had any SBP measurements, a large majority had three SBP measurements. We therefore chose to use the mean of the nurse and two physician SBP measurements in our subsequent analyses. This means that we discarded a subject’s SBP measurements at a visit if they had only one or two SBP measurements. With the exception of visit 4, at which no subjects had three SBP measurements, this means discarding a relatively

Figure 11.1: Within-subject SD versus mean for SBP measurements from visits 1 and 2 in the Framingham study, based on data from 1,099 men

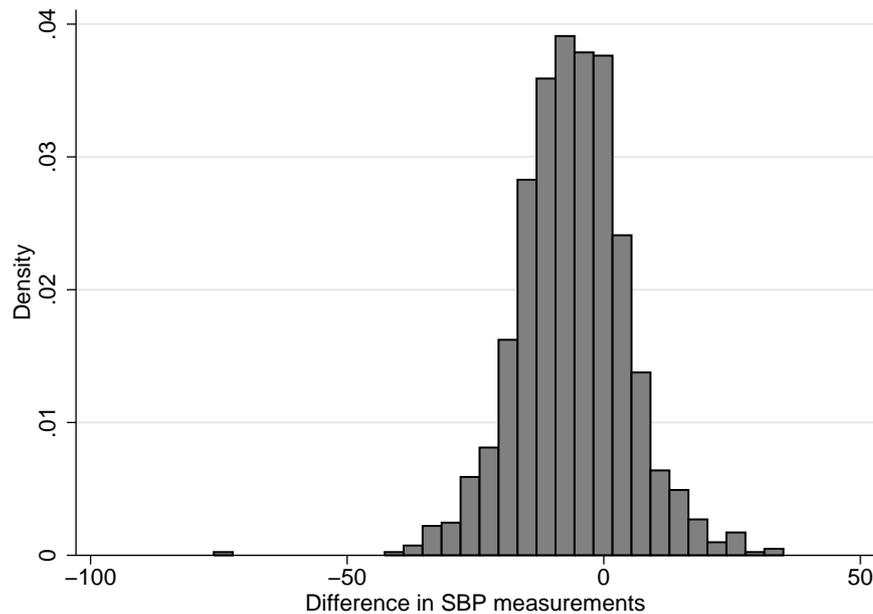


small proportion of SBP measurements. In the following, we use the term ‘SBP measurement’ to refer to a mean of a nurse and two physician SBP measurements.

11.3 Classical measurement error assumptions

In previous analyses of the Framingham data log transformations have been used to improve the normality of the SBP measurements, e.g. Carroll *et al* [111]. We considered whether the SBP measurements (i.e. the mean of the nurse and two physician measurements) required transformation in order to better satisfy the assumptions of the classical error model, when considered as error-prone measurements of a subject’s underlying SBP at the follow-up visit at which a measurement was made. Since a subject’s SBP undergoes changes throughout life, we do not have true replicate measurements for which to assess this. However, particularly at younger ages, it may be reasonable to assume subjects’ underlying SBP did not change by a large amount between two consecutive follow-up visits which are two years apart, and therefore that measurements of SBP from two consecutive visits can be considered independent error-prone replicates of subjects’ underlying SBP. There were 1,099 men who had a SBP measurement at both visit 1 and visit 2 (i.e. they had a nurse and two physician measurements at both visits). The mean SBP at visit 2 for these men was 131.97, in contrast to a mean of 138.11 at baseline. This decrease has been previously documented [14], and may have been due to subjects being less nervous at their second visit, resulting in less elevated SBP measurements. Assuming the apparent decrease is due to a reason such as this, we can still use the measurements to

Figure 11.2: Histogram of difference in SBP measurements from visits 1 and 2 in the Framingham study, based on data from 1,099 men



examine whether the variability in subject-level differences between SBP at the two visits (about the mean decrease of $138.11 - 131.97 = 6.14$) varied with the mean of the two SBP measurements. Figure 11.1 therefore shows a plot of the within-subject SD of a subject's two SBP measurement versus their mean, as suggested by Carroll *et al* [8]. The figure shows that subjects with a larger mean SBP tended to have greater variability between their visit 1 and visit 2 measurements ($p < 0.001$ from linear regression of SD on mean), suggesting that the error variability increased with underlying SBP. However, since the increase in variability with mean is not particularly large, we felt that transformation of the SBP measurements was not warranted, given the consequent loss in interpretability in parameter estimates if a transformation is used.

Figure 11.2 shows the histogram of the differences in the SBP measurements from visits 1 and 2 for the 1,099 men who have measurements at both visits. This suggests that their difference, which is equal to the difference in their two errors, is approximately normally distributed. Although this does not necessarily mean the errors themselves are normally distributed, as discussed by Carroll *et al* [8], that the plot does not show evidence of gross non-normality gives some reassurance that an assumption of normality for the 'errors' may be reasonable.

Chapter 12

Risk of death due to cardiovascular disease between age 70 and 80 and its relationship with systolic blood pressure levels in earlier life

In this chapter we report the results of analyses examining the associations between SBP levels between the ages of 40 and 70 and the odds of death due to cardiovascular disease (CVD) between age 70 and 80. As in the previous chapter, we restricted our analyses to men. Below we give our motivation for our choice of outcome and longitudinal measurements, and describe the data available. In Section 12.1 we describe a linear mixed model for the SBP measurements, and report the estimated model parameters. In Section 12.2 we describe the methods we used to estimate the effects of SBP levels on odds of death due to CVD. We report the results of our analyses in Section 12.3, and give some concluding remarks in Section 12.4.

We decided to consider odds of death due to CVD because CVD was the leading cause of death in the Framingham Study, and is known (partly on the basis of analyses from Framingham) to be strongly related to BP levels. We chose to consider odds of death due to CVD after age 70 because the majority of CVD mortality occurs after this age. We considered odds of death in the 10 years following age 70 because this ensured a sufficient number of deaths due to CVD, and because the mortality status at age 80 is known for all but one man. We restricted our analyses to those men who were alive at age 70 and had had no previous non-fatal CVD events recorded, since the associations between SBP after a non-fatal CVD event with subsequent CVD mortality may differ from those for men who have not had non-fatal CVD events.

In preliminary analyses of the longitudinal measurements made between ages 30 and 70, we found that there was little information with which to estimate parameters corresponding to the underlying SBP values at age 30, due to the fact that very few men had SBP measurements available between the ages of 30 and 35. We therefore decided to use only SBP measurements made after age 40 in our analyses.

There were 1,087 men who survived to age 70, had had no previous CVD events prior to age 70, and had at least one SBP measurement (meaning a nurse and two physician measurements) between age 40 and 70. The vital status of one man at age 80 was not known (i.e. he was censored), and so we did not include this man in the analyses reported in this chapter. From the remaining 1,086 men, there were 9,046 SBP measurements available between age 40 and 70. As before, we denote the error-prone SBP measurements by \mathbf{W}_i (which are the mean of a nurse and two physician SBP measurements, as described in Chapter 11) for subject i . Between age 70 and 80, 141 (13.0%) of the men died due to CVD, which we denote by Y_i (1 if died due to CVD, 0 if alive at age 80 or died due to another cause), for subject i .

12.1 Longitudinal model

12.1.1 Model specification

To model the longitudinal SBP measurements between ages 40 and 70, we assumed a piece-wise linear mean trajectory with age, with knots at age 50 and 60 (see solid line in Figure 12.2). We parametrized this mean structure by defining four covariates:

$$\begin{aligned} age_{40}(t) &= 1 - \min(\text{abs}(t - 40)/10, 1), \\ age_{50}(t) &= 1 - \min(\text{abs}(t - 50)/10, 1), \\ age_{60}(t) &= 1 - \min(\text{abs}(t - 60)/10, 1), \\ age_{70}(t) &= 1 - \min(\text{abs}(t - 70)/10, 1), \end{aligned}$$

and by having no intercept term. Table 12.1 shows the values of these four covariates at 5 yearly intervals from age 40 to 70. This definition thus meant that the corresponding fixed effect parameters represent the mean SBP at ages 40, 50, 60 and 70.

To allow for between-subject variability in SBP trajectories, we allowed each of these four covariates to be random at the subject level. We assumed this four-dimensional random-effects vector to be multivariate normal, with an unstructured variance covariance matrix. These random-effects thus represented the difference between a subject's underlying SBP at ages 40, 50, 60, and 70 and the mean SBP at these ages. To be consistent with our earlier notation, we let \mathbf{X}_i denote the vector of random-effects representing subject i 's true SBP at ages 40, 50, 60 and 70. The elements of \mathbf{X}_i are thus equal to the combination of the population mean SBP at

Table 12.1: Values of covariates used to parametrize the linear mixed model for longitudinal SBP measurements at 5 yearly intervals from age 40 to 70

t	$age_{40}(t)$	$age_{50}(t)$	$age_{60}(t)$	$age_{70}(t)$
40	1	0	0	0
45	0.5	0.5	0	0
50	0	1	0	0
55	0	0.5	0.5	0
60	0	0	1	0
65	0	0	0.5	0.5
70	0	0	0	1

these ages and the mean-zero random-effects representing the deviation of subject i 's true SBP at these ages from the population mean. We assumed that the residual errors were independent of each other, and of the subject-specific random-effects, and that they were normally distributed. Denoting an SBP measurement for subject i at age t_{ij} by W_{ij} , our model thus assumed that:

$$W_{ij} = age_{40}(t_{ij})X_{i1} + age_{50}(t_{ij})X_{i2} + age_{60}(t_{ij})X_{i3} + age_{70}(t_{ij})X_{i4} + U_{ij}$$

where $U_{ij} \sim N(0, \sigma_U^2)$ denotes an independent residual error.

12.1.2 Random-effects assumptions

As a first check on the normality assumption for the random effects, we examined the distribution of the observed SBP measurements available at age 60. Of the 1,086 men included in the analyses of this chapter, 779 had an SBP measurement within a year of age 60, whose histogram is shown in Figure 12.1. The distribution is positively skewed, with a minority of men having very high SBP, indicating that a logarithmic transformation may improve normality. However, as the distribution is not grossly non-normal, our analyses are intended to be illustrative, and because of the previously described robustness results (see Section 8.2.4), we proceeded with an assumption of normality for the underlying SBP levels.

12.1.3 Model estimates

Tables 12.2 and 12.3 show the estimates (and 95% confidence intervals) of the linear mixed model parameters, estimated using ML with Stata's `xtmixed` command. The estimated mean SBP at ages 40, 50, 60 and 70 were, respectively, 126.40mmHg (95% CI 125.10 to 127.69), 129.70mmHg (128.66, 130.74), 136.10mmHg (134.92, 137.28) and 142.59mmHg (141.31, 143.88). SBP was thus estimated to increase on average by 3.30mmHg (1.91, 4.69) between ages 40 and 50, by 6.40mmHg (5.33, 7.47) between 50 and 60, and by 6.50mmHg (5.25, 7.74) between 60 and 70.

Figure 12.1: Histogram of SBP measurements at age 60 in the Framingham Study, based on data from 779 men

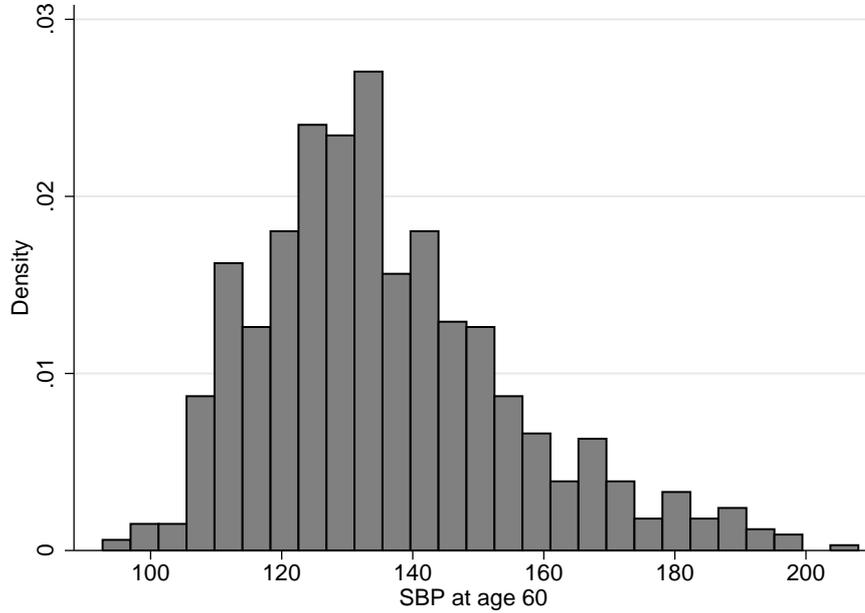


Table 12.2: Estimates of fixed effect parameters and random-effects parameters from linear mixed model for SBP measurements (95% CI). Units are mmHg

Fixed effects parameter	Estimate (95% CI)
SBP at 40	126.40 (125.10, 127.69)
SBP at 50	129.70 (128.66, 130.74)
SBP at 60	136.10 (134.92, 137.28)
SBP at 70	142.59 (141.31, 143.88)
Random effects parameter	Estimate (95% CI)
SD(SBP at 40)	12.09 (10.84, 13.48)
SD(SBP at 50)	14.24 (13.39, 15.15)
SD(SBP at 60)	18.30 (17.42, 19.23)
SD(SBP at 70)	18.75 (17.74, 19.81)
SD(Residual)	9.01 (8.84, 9.17)

Table 12.3: Estimated correlations between random-effects in linear mixed model for SBP measurements (95% CI)

	SBP at 40	SBP at 50	SBP at 60	SBP at 70
SBP at 40	1			
SBP at 50	0.84 (0.72, 0.91)	1		
SBP at 60	0.71 (0.61, 0.78)	0.79 (0.74, 0.83)	1	
SBP at 70	0.55 (0.43, 0.65)	0.50 (0.43, 0.57)	0.67 (0.62, 0.72)	1

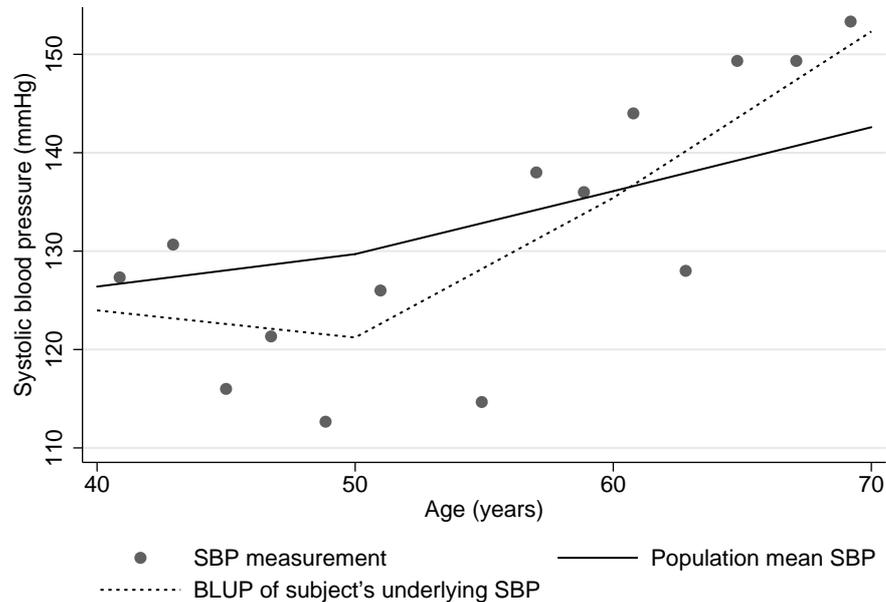
The estimated SD of the random-effects representing the difference between a subject's SBP and the population mean SBP at age 40 was 12.09mmHg (10.84, 13.48). The corresponding estimates for ages 50, 60 and 70 were 14.24mmHg (13.39, 15.15), 18.30mmHg (17.42, 19.23) and 18.75mmHg (17.74, 19.81), indicating that between-subject heterogeneity increased with age.

The estimated residual SD, which can be viewed as representing a combination of deviations from the linear trajectory implied by the fixed and random-effects, plus within-visit variability, was 9.01mmHg (8.84, 9.17). To investigate how much of this is due to within-visit variability, we fitted a simple one-way random-intercepts model to the individual (as opposed to mean of three) SBP measurements at visit 2. This gave an estimated within-visit SD of 8.49mmHg. This implies an SD for the mean of three SBP measurements (about a visit-specific true value) of 4.90mmHg. This therefore suggests that of the estimated residual variance (9.01^2), around 30% is due to within-visit variation, and 70% is due to variation around the linear trajectory implied by the fixed and random-effects.

Table 12.3 shows the estimated correlations between the subject specific random-effects. As we would expect, SBP is highly correlated within subjects over time, and the correlation between SBP at two times decreases as the interval between the times increases.

Figures 12.2 and 12.3 show the SBP measurements for two randomly chosen men, and in addition, the trajectory implied by the combination of the estimated population mean evolution of SBP with the BLUPs of the two men's random-effects. For the first man (Figure 12.2), who had 14 SBP measurements available between ages 40.9 years and 69.2 years, the predicted trajectory of his underlying SBP is determined predominantly from his SBP measurements. In contrast, the second man (Figure 12.3) has only three SBP measurements, which were made at ages 60.6, 66.4, and 68.4 years. The predicted trajectory for SBP for this man in the earlier decades thus borrows heavily from information from other subjects, in the form of the estimated population trajectory and the estimated correlations between the random-effects representing SBP at the different ages. The plot shows how the predicted trajectory of underlying SBP between age 60 and 70 for this man is closer

Figure 12.2: Estimated population mean evolution of SBP, the SBP measurements for a randomly chosen man, and the best linear unbiased prediction of his SBP trajectory

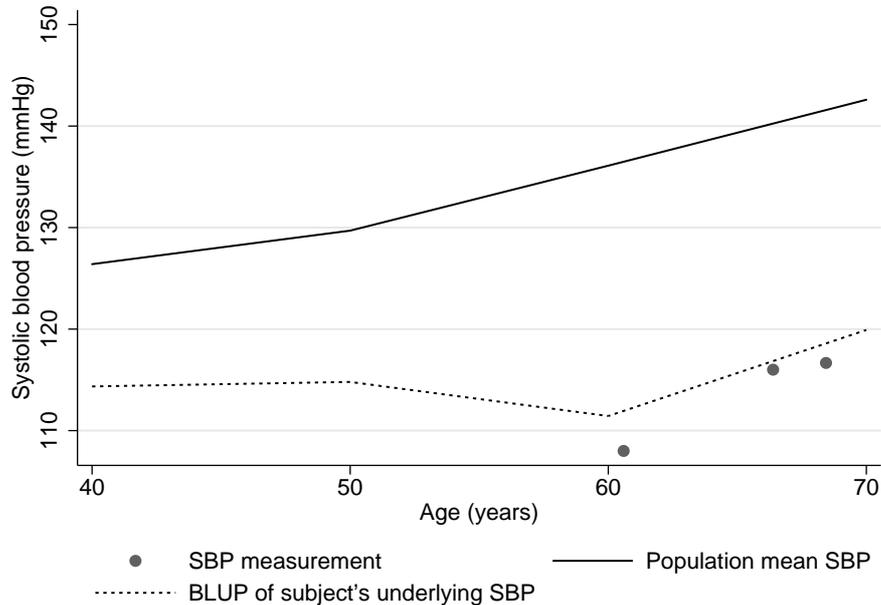


to the population mean than the observed SBP measurements – the phenomenon commonly referred to as ‘shrinkage’.

12.1.4 Alternative model specifications

The linear mixed model we have specified is obviously a simplification of the truth (as most models are), since there is no particular reason why underlying SBP should increase at constant rates over time, with changes in slope only at the start of decades of age. Frost and White considered a model for the SBP measurements from Framingham which assumes a piece-wise constant trajectory for each subject, with instantaneous jumps every five years (although they considered models on the follow-up time scale, rather than age) [16]. The piece-wise linear model we have used can be considered a more flexible specification, since it implies a continuous trajectory for underlying SBP, although our model only allows changes in the slope of the underlying SBP trajectories every 10 years. In principle more complicated spline models could be fitted in which the implied SBP trajectories are differentiable (at least once) in the entire age range. However, given the frequency with which SBP measurements were made (every two years), and the fact that some subjects do not have SBP measurements in the earlier decades (due to being recruited at an older age), the Framingham SBP data probably do not contain sufficient information to enable fitting of such models with any degree of precision.

Figure 12.3: Estimated population mean evolution of SBP, the SBP measurements for a second randomly chosen man, and the best linear unbiased prediction of his SBP trajectory



12.2 Estimation methods

In this section we describe the SBP effects we aimed to estimate, before describing the different estimation methods that we used. We modelled the longitudinal SBP measurements using the linear mixed model previously described. We modelled the odds of death due to CVD between age 70 and 80 using a series of logistic regressions, with covariates consisting of various subsets of underlying SBP at ages 40, 50, 60 and 70. We attempted to estimate the following sets of effects:

- the mutually adjusted effects of SBP at ages 40, 50, 60, 70 on the odds of death due to CVD between age 70 and 80
- the unadjusted effects of SBP at ages 40, 50, 60, 70 on the odds of death due to CVD between age 70 and 80
- the mutually adjusted effects of SBP at ages 40 and 70 on the odds of death due to CVD between age 70 and 80

The first set corresponds to a logistic regression model for Y_i with \mathbf{X}_i as a four-dimensional covariate. The second set corresponds to four separate logistic regression models for Y_i , with X_{i1} , X_{i2} , X_{i3} , and X_{i4} as the single covariate. The third set corresponds to a logistic regression with X_{i1} and X_{i4} as covariates.

Given the estimated high correlations between SBP at the different ages, we expected the estimation of the independent effects of SBP at the four ages to be problematic. In contrast, we expected to be able to estimate the unadjusted effects

of SBP at these ages with reasonable precision. We chose to estimate the mutually adjusted effects of SBP at ages 40 and 70, as a compromise between the fully adjusted and unadjusted effects: since SBP at ages 40 and 70 is less correlated than SBP levels separated by 10 years, and because there are only two covariates, we expected that estimation of the mutually adjusted effects of SBP at 40 and 70 may be feasible.

We estimated these sets of parameters using the following methods. We did not use the CS method because the identifiability requirement for the random-effects \mathbf{X}_i would severely limit the number of subjects who could contribute to the analysis.

12.2.1 Regression calibration

Mutually adjusted effects of SBP at 40, 50, 60 and 70

To implement RC we first fitted the linear mixed model to the SBP measurements \mathbf{W}_i as previously described in Section 12.1, using the `lmer` command from the R package `lme4`. We found the BLUPs of each subject's SBP at ages 40, 50, 60 and 70 years of age, using the fitted linear mixed model. We then fitted a logistic regression model for the binary outcome (death due to CVD between age 70 and 80), using the BLUPs of SBP at ages 40, 50, 60 and 70 as covariates.

Unadjusted effects of SBP at 40, 50, 60 and 70

We estimated the unadjusted effects of SBP at each age (40, 50, 60 and 70) by fitting four separate logistic regression models for the binary outcome, with the BLUPs of SBP at each age (found using the same linear mixed model fit as used to estimate the adjusted effects, as previously described) as the only covariate. As we have discussed in Section 8.6, the RC estimates of the unadjusted effect of SBP at one age on odds of death due to CVD are only approximately consistent if SBP at the other three ages have no independent effect on the odds of death. We therefore refer to these estimates as 'naive RC'.

Mutually adjusted effects of SBP at 40 and 70

We estimated the mutually adjusted effects of SBP at ages 40 and 70 by fitting a logistic regression model with the BLUPs of SBP at ages 40 and 70 as covariates. Again, we refer to these estimates as 'naive RC', as their approximate consistency relies on SBP at 50 and 60 having no independent effects on the outcome.

12.2.2 Corrected regression calibration

Unadjusted effects of SBP at 40, 50, 60 and 70

To estimate the unadjusted effects of SBP at the four ages without making the previously described assumption (of no effects of SBP at other ages), we used equation

(8.3) of Section 8.6.1. This uses the RC estimate of the adjusted effects of SBP at 40, 50, 60, and 70, and the estimate of the variance covariance matrix of \mathbf{X}_i , which is available from the linear mixed model fitted in the first stage of RC. The transformation matrix \mathbf{A} involved in equation (8.3) corresponding to the four unadjusted effects were given by:

$$\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

We refer to these estimates by the term ‘corrected RC’.

Mutually adjusted effects of SBP at 40 and 70

The same approach was used as for the unadjusted effects of SBP at 40, 50, 60 and 70, except with the transformation matrix \mathbf{A} given by:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

12.2.3 Maximum likelihood using ascent-based Monte-Carlo Expectation Maximization (ML1) + multiple imputations (ML1+MI)

Mutually adjusted effects of SBP at 40, 50, 60 and 70

We found ML estimates for the joint model which assumes the linear mixed model described in Section 12.1, with \mathbf{X}_i marginally normal, and in which Y_i may depend jointly on SBP levels at 40, 50, 60 and 70. We used the ascent-based MCEM algorithm, as previously described in Sections 4.5 and 8.2.2. We used the same values for the control parameters of the ascent-based method as in our simulations.

As in the simulations described in Section 9.8, the MCEM algorithm failed to increase the number of imputations from the initial 10, even after many iterations. As previously discussed in Section 9.8, we believe this was due to the fact that the same set of imputations are used to update the model parameters and to assess how much the updated parameters increase the expected complete data log likelihood. As in Section 9.8, we modified the algorithm by, at each iteration, generating a second set of imputations from which we estimated the increase in the expected complete data log likelihood. As in the simulations of Section 9.8, this modification

succeeded in alleviating the problem. As previously noted, that this issue occurred for the simulations in Section 9.8 and for the analyses here suggests that it is related to the dimensionality of the random-effects vector \mathbf{X}_i .

Unadjusted effects of SBP at 40, 50, 60 and 70

As in the simulations of Section 8.7, a by-product of the MCEM algorithm used to estimate the adjusted effects of SBP at the four ages on odds of death due to CVD is a large set of multiple imputations of \mathbf{X}_i . We used these imputations to fit four logistic regression models to estimate the unadjusted effect of SBP at each of the four ages. Because the imputations are generated from the fitted joint model which assumes the odds of death due to CVD may depend jointly on SBP at all four ages, the resulting estimates are consistent for the unadjusted effect of SBP at a given age even if SBP at one or more of the other ages has an independent effect.

Mutually adjusted effects of SBP at 40 and 70

As for the estimation of the unadjusted effects of SBP at 40, 50, 60 and 70, we used the multiple imputations of \mathbf{X}_i to fit the logistic regression model for Y_i , with imputed SBP at ages 40 and 70 as the two covariates.

12.2.4 Maximum likelihood assuming conditional normality for \mathbf{X}_i given Y_i (ML2)

Mutually adjusted effects of SBP at 40, 50, 60 and 70

We found ML estimates for the joint model which assumes multivariate normality for \mathbf{X}_i given Y_i by using our novel approach which consists of fitting a linear mixed model to the longitudinal measurements \mathbf{W}_i and conditioning on Y_i , as previously described in Sections 4.6 and 8.3. This involved modifying the fixed effects specification of the linear mixed model used in the first stage of RC. Specifically, we added four additional fixed effects, equal to the four original covariates multiplied by the binary outcome Y_i . We then estimated the log odds ratios corresponding to SBP at 40, 50, 60 and 70 using equation (4.32) of Section 4.6.

Unadjusted effects of SBP at 40, 50, 60 and 70

We estimated the unadjusted effects of SBP at each of the four ages by substituting the ML2 estimates of the adjusted effects of SBP at 40, 50, 60 and 70, and the estimated variance conditional covariance matrix $\Sigma_{\mathbf{X}|Y}$ found from the fitted linear mixed model, into equation (8.8) of Section 8.6.2. This involved using the same four transformation matrices as for the corrected RC estimates. As previously described

in Section 8.6.2, the resulting estimates are valid for the effect of SBP at one age even if SBP at another age has an independent effect.

Mutually adjusted effects of SBP at 40 and 70

As for the estimation of the unadjusted effects of SBP at 40, 50, 60 and 70, except using the transformation matrix \mathbf{A} given by:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

12.2.5 Inference

We found bootstrap 95% percentile confidence intervals by performing 2,000 non-parametric bootstrap re-samples of the data. Following the sampling scheme which gave rise to the data, we re-sampled subjects, rather than individual SBP measurements. For some bootstrap samples, some of the methods failed, due to estimated correlations between SBP at ages 40 and 50 being close to 1. For each estimate, we report the number of bootstrap samples for which estimation failed, and report 95% bootstrap percentile confidence intervals based on the centiles of the distribution from the bootstrap samples for which estimation did not fail.

12.3 Results

12.3.1 Mutually adjusted effects of SBP at 40, 50, 60 and 70

Table 12.4 shows the estimates of the mutually adjusted effects of SBP at ages 40, 50, 60 and 70, on the odds of death due to CVD between age 70 and 80, expressed as odds ratios for a 10mmHg increase in SBP. The point estimates from the three estimation methods are very similar. However, the confidence intervals are extremely wide. This is not unexpected, as SBP levels at the four ages are highly correlated within subjects (see Table 12.3). Given these high correlations, very large samples would be required to estimate the independent effects of SBP at each age, adjusted for SBP at the other three points, with any reasonable degree of precision.

The linear mixed model used in the first stage of RC converged successfully for all bootstrap samples. In 14 bootstrap samples, the estimated correlation between SBP at age 40 and 50 was either equal to 1, or was very close to 1. For these bootstrap samples, R's glm function (used to fit the logistic regression model) dropped either SBP at 50 (for one sample) or SBP at 70 (for 13 samples) from the model, due to the covariance matrix of the logistic model covariates being almost singular. The MCEM (ML1) algorithm similarly failed (for 56 bootstrap samples) when one of

the estimated correlations between SBP levels was close to 1. This occurred if the initial estimates of the log odds ratios from RC contained a missing value (due to the previously described co-linearity in the BLUPs). It also occurred for other bootstrap samples in which the RC estimates of the adjusted log odds ratios were finite but extremely large. This resulted in numerical problems when evaluating the rejection probabilities when multiply imputing \mathbf{X}_i . The ML2 method, based on fitting a linear mixed model for \mathbf{W}_i given Y_i , gave estimates for all bootstrap samples, although for some bootstrap samples estimates were sometimes zero or infinite.

12.3.2 Unadjusted effects of SBP at 40, 50, 60 and 70

Table 12.5 shows the estimates of the unadjusted odds ratios for the effects of SBP at ages 40, 50, 60 and 70 on odds of death due to CVD. With the exception of the ML2 estimate of the effect of SBP at age 40, and ignoring for a moment the fact that estimation failed for some bootstrap samples, all estimates were statistically significant at the 5% level, with higher SBP at each age associated with greater odds of death due to CVD. The estimated odds ratios were largest for the effect of SBP at age 40, while SBP at age 60 had the smallest estimated effect. Assuming that SBP at each age has some positive independent effect on the odds of death due to CVD between age 70 and 80, we expect the naive RC estimates to be biased upwards. The estimates support this, with the naive RC estimates larger than the corrected RC estimates. The biggest difference is for the effect of SBP at 40, for which the naive RC estimate (on the log odds ratio scale) is 13% larger than the corrected RC estimate. This may be because only a minority of men have SBP measurements earlier in life, and so the implicit conditional independence assumption is only violated in this minority of men. In contrast, when predicting SBP at age 40, for a majority of men the prediction is based on SBP measurements later in life (because they were only recruited at an older age), and so the conditional independence assumption is violated to a greater extent.

Compared to the corrected RC estimates, the ML1+MI and ML2 estimates are slightly larger in magnitude. This is consistent with our previous findings (and those in the literature) that RC usually gives estimates which are biased towards the null. Comparing the confidence intervals for naive RC and ML2 (for which both estimation methods did not fail for any bootstrap samples, and which therefore permit a fair comparison), the intervals for ML2 are wider than those for RC. This is consistent with the fact that the ML2 estimates do not rely on an assumption that SBP at the other ages has no independent effect on odds of death due to CVD, which is implicitly made by naive RC. We note that the confidence intervals for the effects of SBP at age 50, 60 and 70 are only slightly wider for ML2 compared to RC, whereas for SBP at age 40 the ML2 interval is substantially wider. We conjecture

Table 12.4: Estimates of mutually adjusted odds ratios (OR) (95% bootstrap percentile CI) for the effects of a 10mmHg increase in SBP at ages 40, 50, 60 and 70 on odds of death due to CVD. Estimates found using regression calibration (RC), maximum likelihood assuming marginal normality for random-effects via ascent-based MCEM (ML1), and maximum likelihood assuming conditional normality for random-effects given outcome (ML2). # failures denotes the number of bootstrap samples (out of 2,000) for which estimation failed. Bootstrap percentile CIs are based on estimates from bootstrap samples in which estimation did not fail.

SBP at age	RC		ML1		ML2	
	OR (95% CI)	# failures	OR (95% CI)	# failures	OR (95% CI)	# failures
40	1.16 (0.14, > 100)	0	1.16 (0.17, 26.73)	56	1.25 (0.03, > 100)	0
50	1.09 (0.001, 6.12)	1	1.11 (0.12, 5.36)	56	1.09 (< 0.001, 18.15)	0
60	0.92 (0.51, 4.62)	0	0.91 (0.52, 1.74)	56	0.89 (0.36, > 100)	0
70	1.17 (0.29, 1.80)	13	1.18 (0.67, 1.71)	56	1.20 (< 0.001, 2.48)	0

that this may be due to the same reason which caused the corrected RC estimates to differ most from the naive RC estimates for SBP at 40, i.e. because SBP at age 40 has to be predicted for many men on the basis of SBP measurements at later ages.

Estimates of the unadjusted effects were found for all bootstrap samples for RC and ML2. For the 14 bootstrap samples for which the RC estimates of the adjusted effects of SBP at the four ages were not available (see Table 12.4), the corrected RC estimates could not be calculated, since this uses the RC estimates of the adjusted effects. Similarly, the ML1+MI estimates are based on the imputations available at the convergence of the MCEM algorithm. Since this failed for 56 bootstrap samples as previously described, the ML1+MI estimates were unavailable for 56 bootstrap samples.

12.3.3 Mutually adjusted effects of SBP at 40 and 70

Table 12.6 shows the estimates of the mutually adjusted effects of SBP at ages 40 and 70. The point estimates from all four methods are similar, suggesting SBP at both 40 and 70 was positively associated with odds of death due to CVD. The largest difference in point estimates between the methods was the ML2 estimate of the independent (of SBP at 70) effect of SBP at 40, which was 25% larger than the estimates from the other methods. Although all of the 95% confidence intervals include 1, their lower limits are only slightly below 1, suggesting weak evidence that SBP at the two ages have independent effects on the odds of death due to CVD. As for the estimates of the unadjusted effects of SBP at the four ages, the confidence intervals for the RC estimates are narrower than those for the ML2 estimates. Also mirroring the results of the unadjusted analyses, the difference in confidence interval width between ML2 and RC is largest for the effect of SBP at 40.

12.4 Conclusions

Our results show that mean SBP levels increased with age in the men recruited into the Framingham Study. As one would expect, a man's underlying SBP levels at ages 40, 50, 60 and 70 are highly correlated. Between-subject variability in SBP increased with age, with the estimated between-subject SD in underlying SBP at age 70 around 55% greater than the estimated SD of underlying SBP at age 40. By comparing the estimated within-subject residual SD with an estimate of the within-visit SD of individual SBP measurements from visit 2, we found that around 70% of the within-subject variance is attributable to variations around a subject's linear trajectory. Our estimates are valid provided these variations are independent and unrelated to the odds of death due to CVD. This finding does however suggest that more complex models might be appropriate for these data.

Table 12.5: Estimates of unadjusted odds ratios (OR) (95% CI) for the effects of a 10mmHg increase in SBP at ages 40, 50, 60 and 70 on odds of death due to CVD. Estimates found using naive regression calibration (naive RC), corrected RC, maximum likelihood assuming marginal normality for random-effects via ascent-based MCEM followed by multiple imputation (ML1+MI), and maximum likelihood assuming conditional normality for random-effects given outcome (ML2). # failures denotes the number of bootstrap samples (out of 2,000) for which estimation failed. Bootstrap percentile CIs are based on estimates from bootstrap samples in which estimation did not fail.

SBP at age	Naive RC		Corrected RC		ML1+MI		ML2	
	OR (95% CI)	# failures						
40	1.39 (1.17, 1.64)	0	1.33 (1.03, 1.76)	14	1.35 (1.03, 1.77)	56	1.41 (0.98, 2.20)	0
50	1.27 (1.12, 1.44)	0	1.25 (1.08, 1.43)	14	1.26 (1.09, 1.45)	56	1.28 (1.08, 1.55)	0
60	1.19 (1.08, 1.31)	0	1.17 (1.05, 1.29)	14	1.17 (1.05, 1.30)	56	1.18 (1.05, 1.34)	0
70	1.22 (1.09, 1.35)	0	1.21 (1.08, 1.35)	14	1.22 (1.08, 1.36)	56	1.24 (1.09, 1.43)	0

Table 12.6: Estimates of mutually adjusted odds ratios (OR) (95% CI) for the effects of a 10mmHg increase in SBP at ages 40 and 70 on odds of death due to CVD. Estimates found using naive regression calibration (naive RC), corrected RC, maximum likelihood assuming marginal normality for random-effects via ascent-based MCEM followed by multiple imputation (ML1+MI), and maximum likelihood assuming conditional normality for random-effects given outcome (ML2). # failures denotes the number of bootstrap samples (out of 2,000) for which estimation failed. Bootstrap percentile CIs are based on estimates from bootstrap samples in which estimation did not fail.

SBP at age	Naive RC		Corrected RC		ML1+MI		ML2	
	OR (95% CI)	# failures						
40	1.20 (0.94, 1.52)	0	1.19 (0.83, 1.79)	14	1.20 (0.83, 1.77)	56	1.25 (0.76, 2.38)	0
70	1.13 (0.97, 1.31)	0	1.14 (0.94, 1.35)	14	1.14 (0.95, 1.37)	56	1.15 (0.86, 1.47)	0

The high correlations between SBP at different ages meant that our estimates of the independent effects of SBP at one age, adjusted for SBP at other ages, on the odds of death due to CVD between age 70 and 80, were very imprecise. We do not believe therefore that any conclusions can be drawn from our estimates of the mutually adjusted effects of SBP at each of the four ages on the outcome.

In contrast, we were able to estimate the unadjusted associations between SBP at ages 40, 50, 60 and 70 with odds of death due to CVD, much more precisely. Our results suggest that SBP at each of these ages, without adjustment for SBP at other ages, is positively associated with odds of death due to CVD between 70 and 80, with estimated odds ratios for a 10mmHg increase in SBP ranging from around 1.17 for the unadjusted effect of SBP at age 60, to 1.35 for the effect of SBP at age 40. These estimates are of a similar magnitude to the hazard ratio estimates reported by the Prospective Studies Collaboration [140], which aimed to estimate associations between underlying BP levels at the start of a decade of age with CVD events in the following decade. Although their analysis made an allowance for measurement error, they used correction methods similar to those used by Clarke *et al* [15], and as discussed in Section 1.2.1, the validity of these adjustment rests on number of questionable assumptions.

Lastly, we considered the mutually adjusted effects of SBP at ages 40 and 70 on odds of death due to CVD. Although the confidence intervals were relatively wide, our results suggest that SBP at both ages may have an independent effect on odds of death due to CVD between 70 and 80. The point estimates suggest that a 10mmHg increase in SBP at age 40 may have a larger influence, conditional on SBP at 70, than SBP at age 70 has (conditional on SBP at 40). However, it is important to note that there was less variability between-subjects in SBP at age 40 compared to at age 70.

When estimating the associations between SBP levels at one age and odds of death due to CVD, without adjustment for SBP at all other time points, the naive RC estimates were larger than the corrected RC estimates. This is in agreement with our simulation results. The naive RC estimates implicitly rely on the assumption that SBP at the other ages have no association with odds of death due to CVD, conditional on the SBP level included as a covariate in the logistic regression model. If we assume that SBP at each age is independently and positively associated with odds of death due to CVD, the univariate estimates from naive RC will be biased upwards, since SBP levels at different ages are positively correlated. The confidence intervals for the ML2 estimates, which do not make such a conditional independence assumption, were wider than those for the naive RC estimates, which we believe is a reflection of the cost in terms of efficiency of relaxing this assumption.

On the basis of the results here, in broad agreement with our simulation results, we conclude that RC gives estimates which are similar to those obtained by more

complicated estimation methods, such as ML, when all of the random-effects \mathbf{X}_i are included as covariates in the logistic regression outcome model. When fitting outcome models with subsets of the random-effects \mathbf{X}_i as covariates, although the naive RC estimates may be biased, one may argue on the basis of these results that these biases are often small enough to ignore. We had anticipated that correcting the naive RC estimates would have a larger effect. As with our simulations in Chapter 8 however, the biases in the naive RC estimates may have been beneficial in acting in the opposite direction to the bias towards the null inherent in RC estimates for logistic regression outcome models. This cannot be expected to occur in all situations – if for example the covariates are negatively correlated with each other, but each is positively associated with the outcome, the bias in the naive RC estimates due to the violation of the conditional independence will be negative.

The corrected RC estimates were easy to calculate from the RC estimates of the ‘full model’, in which SBP at 40, 50, 60 and 70 were included as covariates, and the estimated variance covariance matrix of \mathbf{X}_i . This approach does not rely on an assumption that SBP levels which are not included as covariates in the logistic regression outcome model have no independent effect on odds. Since estimation of the independent effects of SBP at the four ages is so difficult, due to their high correlations, we cannot determine with any degree of confidence which SBP levels have no independent effect. It therefore seems prudent, when estimating other associations which are not adjusted for SBP at all other ages, to not make any assumptions regarding lack of independent effects. The corrected RC approach is therefore, in our view, an appealing approach.

Ascent-based MCEM gave similar estimates to the RC approach. Whereas the latter can be easily implemented using standard statistical software, the former required a custom-written program in R, and took longer to run. Given that the two methods gave very similar estimates, there seems little benefit in using ascent-based MCEM for estimation for the analyses we have considered in this chapter.

Our proposed method to find ML estimates for the model which assumes conditional normality for \mathbf{X}_i given Y_i also gave similar estimates to the RC method. Like RC, our approach can be implemented by fitting a modified version of the linear mixed model fitted in the first stage of RC. While for the data considered here this approach gave similar estimates to the corrected RC method, on the basis of our earlier simulations (see Section 8.7), we believe our approach may be useful when the effect sizes are larger, where RC gives estimates with non-negligible biases.

Chapter 13

The effects of current and past systolic blood pressure on the hazard of cardiovascular disease

In this chapter we report the results of analyses to investigate how the hazard rate for the occurrence of cardiovascular disease depends on current and past levels of underlying SBP. We begin by explaining our choice of outcome and exposure, and describing the data available. In Section 13.1 we describe a linear mixed model for the longitudinal measurements, and report the parameter estimates based on the available longitudinal SBP measurements. In Section 13.2 we describe the various effects which we aimed to estimate and the two estimation methods used. We report the results of analyses in Section 13.3, and give concluding remarks in Section 13.4.

As for the analyses described in the previous chapter, we used data only from the men in Framingham. In order to increase the number of events (compared to the binary outcome used in Chapter 12), we considered time from age 60 to a man's first CVD event (non-fatal or fatal) as our outcome. Since we wish to exclude men who had previously had non-fatal CVD events from our analysis, by choosing age 60 as the time origin (as opposed to 70), we increased the number of subjects which could contribute to the analysis. We thus included only those men who at age 60 had had no previous non-fatal CVD events. To simplify the analysis we did not use data from 67 men who were recruited after the age of 60 (although they could have been treated as 'delayed entry' observations from the age at which they were recruited). Men who died from causes other than CVD, and who had had no previous CVD events, were censored at the time of their death. There were 1,587 men who satisfied the above criteria, and were hence at risk at age 60. We censored all those who were still at risk at age 80, to avoid difficulties in modelling SBP at ages greater than 80, for which there were far fewer measurements.

In order to relate hazard of experiencing a CVD event to current and past levels of SBP, we considered models for SBP measurements made from age 40 up to age 80.

The 1,587 men included in the analysis collectively had 17,176 SBP measurements between age 40 and 80. Since SBP following a non-fatal CVD event may change as a result of the event itself or because of treatment given in response to the event, we excluded 1,811 SBP measurements which took place after such an event from our analyses, leaving 15,365 SBP measurements from 1,582 men. Of these 1,582 men, 783 (49.5%) experienced either a non-fatal CVD event or died due to CVD before age 80. By considering hazard of CVD from age 60 onwards, we thus only considered the influence of SBP in the preceding 20 years, in contrast to our analyses in Chapter 12, in which we considered SBP levels in the 30 years preceding the ‘risk period’.

13.1 Longitudinal model

13.1.1 Model specification

We adopted the same linear mixed model for the longitudinal SBP measurements as used in Chapter 12, extended to allow for measurements up to age 80. We thus assumed a piece-wise linear mean structure with age, with knots at ages 50, 60 and 70. We also assumed the existence of a five-dimensional subject-specific random-effects vector $\mathbf{X}_i \sim N(\boldsymbol{\mu}_X, \boldsymbol{\Sigma}_X)$ such that an SBP measurement W_{ij} made at age t_{ij} :

$$W_{ij} = age_{40}(t_{ij})X_{i1} + age_{50}(t_{ij})X_{i2} + age_{60}(t_{ij})X_{i3} \\ + age_{70}(t_{ij})X_{i4} + age_{80}(t_{ij})X_{i5} + U_{ij} \quad (13.1)$$

where the covariates $age_{40}(t)$, $age_{50}(t)$, $age_{60}(t)$, $age_{70}(t)$, $age_{80}(t)$ are defined as in Section 12.1 and $U_{ij} \sim N(0, \sigma_U^2)$ denotes an independent residual error.

13.1.2 Model estimates

Tables 13.1 and 13.2 show the estimated parameters obtained by fitting the linear mixed model to the observed SBP measurements using Stata’s `xtmixed` command (via ML).

Before considering the interpretation of these estimates, we first note that, as discussed in the introduction to Chapter 9, for the current analysis we can view the occurrence of a CVD event as censoring the observation of the longitudinal SBP process, i.e. subjects ‘drop-out’ from longitudinal observation when they experience CVD. The linear mixed model fitted to the observed SBP measurements ignores the time-to-CVD event outcome, and inferences are based on the likelihood function of the longitudinal measurements. Such analyses are valid providing that ‘drop-out’, or missingness, depends at most on observed data, i.e. data are missing at

random (MAR). Under our modelling assumptions, time-to-CVD is driven by the unobserved random-effects \mathbf{X}_i . The longitudinal SBP measurements are thus subject to missingness which is not at random (MNAR), implying that estimates based on the likelihood function of the SBP measurements is biased. However, one may argue that if the error-prone SBP measurements are a good proxy for the underlying \mathbf{X}_i , then after conditioning on observed SBP measurements, \mathbf{X}_i is less informative about missingness, and so the MAR assumption is violated to a lesser extent.

We now consider interpretation of the linear mixed model parameters, assuming that the missing not at random issue is small enough to be ignored. The estimated mean SBP at ages 40, 50 and 60 were 127.55mmHg (95% CI 126.44, 128.65), 131.86mmHg (130.91, 132.81), and 138.76mmHg (137.71, 139.81) respectively. Since all subjects were alive at age 60, we may (with caveats regarding the MAR assumption) interpret these as the estimated mean SBP levels in the population of men who survived to age 60 and who had had no previous CVD.

Following our discussion in Section 9.9, we do not believe the estimates corresponding to parameters concerning SBP at ages 70 and 80 (means, variances, and correlation) correspond to meaningful population parameters. For those men who experience CVD between ages 60 and 80, their SBP levels at times following their event are implicitly imputed using their available SBP measurements when we fit a linear mixed model to the observed SBP measurements. They are imputed on the basis of the relationship between SBP levels at 70 and 80 and SBP levels earlier in life, the estimation of which uses data from those subjects who have SBP measurements in both the earlier decades and the later decades (and hence did not experience CVD). For those who experienced CVD, their later SBP levels are thus imputed after the time of their CVD event, assuming that the relationship is between SBP early and later levels of SBP is the same as in those who do not experience CVD. For those men who experience a non-fatal CVD event but who survive to age 80, their SBP to age 80 existed, and if we are willing to assume that SBP levels following a non-fatal CVD has the same relationship with earlier SBP levels as those men who survived to age 80 without experiencing CVD, the estimates for mean SBP at 70 and 80 may legitimately be interpreted as the mean SBP levels in the population.

However, many of the men died between 60 and 80, due to both CVD and other causes. For these men, their SBP ceased to exist at the time of death. By fitting the linear mixed model to the observed SBP measurements, we are implicitly imputing their SBP levels at age 70 and 80 using their earlier SBP measurements, based on the relationship estimated in those men who survived to age 80 and thus had SBP measurements across the entire age range. In effect, for those men who died, the model imputes what their SBP would have been, assuming the relationship between earlier and later SBP levels is the same as in survivors. This makes little sense,

Table 13.1: Estimates of fixed effect parameters and random-effects parameters from linear mixed model for SBP measurements (95% CI). Units are mmHg

Fixed effects parameter	Estimate (95% CI)
SBP at 40	127.55 (126.44, 128.65)
SBP at 50	131.86 (130.91, 132.81)
SBP at 60	138.76 (137.71, 139.81)
SBP at 70	144.44 (143.32, 145.57)
SBP at 80	148.06 (146.41, 149.71)
Random effects parameter	Estimate (95% CI)
SD(SBP at 40)	12.88 (11.77, 14.10)
SD(SBP at 50)	16.05 (15.26, 16.88)
SD(SBP at 60)	19.64 (18.85, 20.47)
SD(SBP at 70)	18.81 (17.88, 19.79)
SD(SBP at 80)	20.12 (18.56, 21.81)
SD(Residual)	9.79 (9.66, 9.93)

Table 13.2: Estimated correlations between random-effects in linear mixed model for SBP measurements (95% CI)

	SBP at 40	SBP at 50	SBP at 60	SBP at 70
SBP at 40	1			
SBP at 50	0.86 (0.76, 0.91)	1		
SBP at 60	0.73 (0.66, 0.79)	0.80 (0.76, 0.84)	1	
SBP at 70	0.59 (0.49, 0.68)	0.52 (0.46, 0.58)	0.71 (0.66, 0.75)	1
SBP at 80	0.57 (0.40, 0.69)	0.49 (0.39, 0.59)	0.53 (0.44, 0.60)	0.65 (0.57, 0.72)

and so we urge considerable caution in interpreting the estimates of the parameters corresponding to SBP at 70 and 80.

13.2 Estimation methods

We assumed a Cox proportional hazards model for the time to a CVD event. We considered estimation of three different sets of effects:

- the mutually adjusted effects of SBP at ages t , $t - 10$ and $t - 20$ on the hazard of CVD at age t
- the unadjusted effects of SBP at ages t , $t - 10$ and $t - 20$ on the hazard of CVD at age t
- the mutually adjusted effects of SBP at ages t and $t - 20$ on the hazard of CVD at age t

As in Chapter 9, let $M_i(t)$ denote the underlying SBP value at age t . The first set corresponds to a Cox proportional hazards model with a three-dimensional

time-dependent covariate $\mathbf{X}_i^*(t) = (M_i(t), M_i(t - 10), M_i(t - 20))^T$. The second set corresponds to three separate Cox proportional hazards models, with scalar time-dependent covariate equal either to $X_i^*(t) = M_i(t)$, $X_i^*(t) = M_i(t - 10)$, or $X_i^*(t) = M_i(t - 20)$. The third set corresponds to a Cox proportional hazards model with the bivariate time-dependent covariate $\mathbf{X}_i^*(t) = (M_i(t), M_i(t - 20))^T$.

We used RC and ML (or ML+MI) to estimate the hazard ratios corresponding to these three sets of effects.

13.2.1 Regression calibration

Mutually adjusted effects of SBP at ages $t - 20$, $t - 10$ and t

To implement RC we fitted the linear mixed model to all of the available SBP measurements (i.e. not risk-set RC). From this fitted mixed model we found the BLUPs of each subject's SBP at ages 40, 50, 60, 70 and 80. We used these BLUPs to calculate the value of $\mathbf{X}_i^*(t) = (M_i(t), M_i(t - 10), M_i(t - 20))^T$ at each event time. We then fitted a time-updated Cox model using R's `coxph` command, using the predicted value of $\mathbf{X}_i^*(t)$ as a (3 dimensional) time-dependent covariate.

Unadjusted effects of SBP at ages $t - 20$, $t - 10$ and t on hazard of CVD at age t

To estimate the unadjusted effects of SBP at ages t , $t - 10$ and $t - 20$ on the hazard of CVD at age t we fitted three separate time-dependent Cox models, with the predicted values of $M_i(t)$, $M_i(t - 10)$ and $M_i(t - 20)$ as the single (time-dependent) covariate.

Mutually adjusted effects of SBP at ages $t - 20$ and t on hazard of CVD at age t

To estimate the mutually adjusted effects of SBP at age t and $t - 20$ on hazard of CVD at age t we fitted a time-dependent Cox model with the predicted values of $\mathbf{X}_i^*(t) = (M_i(t), M_i(t - 20))^T$ as a bivariate time-dependent covariate.

13.2.2 Maximum likelihood using ascent-based Monte-Carlo Expectation Maximization (ML) + multiple imputation (ML+MI)

Mutually adjusted effects of SBP at ages $t - 20$, $t - 10$ and t on hazard of CVD at age t

We found ML estimates for the joint model which assumes the hazard at age t may depend jointly on $M_i(t)$, $M_i(t - 10)$ and $M_i(t - 20)$, using ascent-based MCEM,

as described in Section 9.5. We used the same values for the control parameters of the ascent-based method as in our simulations. As for the analyses described in Chapter 12, we generated a second set of imputations at each iteration, which was used to estimate the increase in the expected complete data log likelihood. Due to the computational demands of the algorithm for this model we declared convergence either when the number of imputations reached 100 or, as before, if the upper confidence interval limit for the increase in the Q function was less than 0.01.

Unadjusted effects of SBP at ages $t - 20$, $t - 10$ and t

To estimate the unadjusted effects of SBP at ages t , $t - 10$, and $t - 20$ on hazard of CVD at age t , we used the multiple imputations of \mathbf{X}_i which were available as a by-product of the MCEM algorithm. We fitted three separate Cox models (for each imputation) with time-dependent covariate equal to either $M_i(t)$, $M_i(t - 10)$ or $M_i(t - 20)$, with these values calculated using the imputed value of \mathbf{X}_i . We then averaged (on the log hazard ratio scale) the parameter estimates across the imputations to give an overall point estimate.

Mutually adjusted effects of SBP at ages $t - 20$ and t

As for the unadjusted effects, to estimate the mutually adjusted effects of SBP at ages t and $t - 20$ on hazard of CVD at age t we fitted a Cox model (for each imputation) using $M_i(t)$ and $M_i(t - 20)$ as time-dependent covariates, the latter being calculated using the imputed value of \mathbf{X}_i . The estimates were then averaged across the multiple imputations (on the log hazard scale) to give an overall point estimate.

13.2.3 Inference

Due to the additional computational burden of estimating the parameters of the joint model (relative to the case of a binary outcome), we used non-parametric bootstrapping (200 bootstrap samples) to estimate the standard error of the RC, ML, and ML+MI estimates. We used these estimated standard errors to form Wald-type 95% confidence intervals for the estimated log hazard ratios, and back-transformed these to give 95% confidence intervals for the estimated hazard ratios. Inspection of histograms of the 200 bootstrap estimates confirmed that an assumption of normality for the RC and ML estimators was reasonable.

13.3 Results

13.3.1 Mutually adjusted effects of SBP at ages $t - 20$, $t - 10$ and t

Table 13.3 shows the estimates of the mutually adjusted effects of SBP at ages t , $t - 10$, and $t - 20$ on hazard of CVD at age t . We begin by noting that estimation of these adjusted effects was much more precise than the corresponding estimates in the case of a binary outcome (Chapter 12), as indicated by the confidence intervals in Table 13.3. We comment further on the likely reasons for this in our Conclusions (Section 13.4).

The point estimates from both RC and ML for the effects of SBP at ages t , $t - 10$ and $t - 20$ on hazard of CVD at age t are all greater than one, suggesting higher SBP levels were independently associated with greater hazard of CVD. Both the RC and ML estimates of the adjusted (for SBP 10 and 20 years earlier) effect of current SBP are statistically significant at the 5% levels, with a 10mmHg increase in current SBP estimated to increase hazard of CVD by 13% (95% CI 5.3%, 21.3%) using RC and by 16% (6.2%, 26.7%) using ML. The estimated effects of SBP both 10 and 20 years earlier were smaller (for both RC and ML), although their 95% confidence intervals indicate the data are consistent with them having no independent effects.

The point estimates from RC suggest that SBP 10 years earlier has an independent effect on current hazard that is approximately 80% greater than the independent effect of SBP 20 years earlier. In contrast, the ML estimate of the adjusted effect of SBP 10 years earlier is similar (in fact slightly smaller) to that of the effect of SBP 20 years earlier. It is unclear why this difference in the relative importance of SBP levels between RC and ML occurred.

It is important to note that the validity of the both the RC and ML estimates rely (in addition to assumptions regarding the model for SBP) on the assumption that the hazard depends on the longitudinal SBP history only via its current value, its values 10 years earlier, and its value 20 years earlier. If, between ages 70 and 80, the hazard depends on SBP levels between 40 and 50, we would expect bias to occur in general.

13.3.2 Unadjusted effects of SBP at ages $t - 20$, $t - 10$ and t

Table 13.4 shows the estimates of the unadjusted effects of SBP at age t , $t - 10$, and $t - 20$ on hazard of CVD at age t . All of the estimates are highly statistically significant on the basis of the lower limits of the 95% bootstrap confidence intervals, with an increase in either current, 10 years earlier, or 20 years earlier SBP level by 10mmHg associated with around a 25% increase in current hazard.

Table 13.3: Estimates of mutually adjusted hazards ratios (95% bootstrap normal CI) for the effects of a 10mmHg increase in SBP at ages t , $t-10$, and $t-20$ on hazard of CVD at age t . Estimates found using regression calibration (RC) and maximum likelihood assuming marginal normality for random-effects via ascent-based MCEM (ML).

SBP at age	RC	ML
t	1.130 (1.053, 1.213)	1.160 (1.062, 1.267)
$t-10$	1.087 (0.951, 1.242)	1.055 (0.909, 1.226)
$t-20$	1.048 (0.917, 1.197)	1.057 (0.923, 1.210)

The RC point estimates are all larger than the ML+MI estimates, with the difference being largest for the unadjusted effect of SBP 20 years earlier. Assuming that SBP at ages t , $t-10$ and $t-20$ are all independently positively associated with hazard at age t , for the reasons discussed in Section 9.3 and our simulation evidence (Section 9.8), we expect the RC estimates of the unadjusted effects to be positively biased, although this may be partly negated by the inherent bias towards the null in RC estimates for Cox models. The point estimates are thus consistent with our earlier findings regarding the bias of RC when outcome models are fitted which exclude important covariates (which are a function of the longitudinal process) when fitting the Cox outcome model, although the differences are not particularly large.

We conjecture that the difference between RC and ML+MI may be largest for the effect of SBP 20 years earlier partly for the same reason as the difference between RC and corrected RC/ML estimates for the effect of SBP on odds of death due to CVD between 70 and 80 were largest for the unadjusted effect of SBP at age 40 (Chapter 12). For events occurring between the ages of 60 and 70, SBP levels 20 years earlier correspond to SBP between ages 40 and 50. For those men who were recruited after the age of 50, and who necessarily have no SBP measurements prior to age 50, their SBP levels at ages between 40 and 50 must be predicted solely on the basis of SBP measurements made at later ages. If SBP at these later ages has an independent effect on hazard at current age, we expect RC to give biased estimates, as discussed in Section 9.3.

Conversely, for the unadjusted effect of current SBP, for events occurring between the ages of 60 and 70, aside from intermittent missingness of follow-up visits, men should have SBP measurements available with which to predict their current underlying SBP. Conditional on such measurements, earlier (and indeed later) SBP measurements provide less information, and therefore the bias in the RC estimates of the unadjusted effect of current SBP should be smaller. A second contributory reason for the largest difference occurring for the effect of SBP 20 years earlier may be if current SBP has a larger independent influence on current hazard than SBP 20 years earlier. In this case, the conditional independence assumption on which

Table 13.4: Estimates of unadjusted hazards ratios (95% bootstrap normal CI) for the effects of a 10mmHg increase in SBP at ages t , $t - 10$, and $t - 20$ on hazard of CVD at age t . Estimates found using regression calibration (RC) and maximum likelihood assuming marginal normality for random-effects via ascent-based MCEM plus multiple imputation (ML+MI)

SBP at age	RC	ML+MI
t	1.248 (1.203, 1.294)	1.246 (1.199, 1.295)
$t - 10$	1.257 (1.210, 1.306)	1.248 (1.200, 1.298)
$t - 20$	1.281 (1.223, 1.343)	1.253 (1.193, 1.317)

RC estimates of unadjusted effects is based is violated to a greater extent, and so we might expect the consequent bias to be greater.

Lastly, we note that the confidence intervals for the ML+MI estimates of the unadjusted effects are of a similar width to those for RC. This is despite the fact that in contrast to RC, the ML+MI estimates are valid for the unadjusted effect of SBP at one age even if SBP at the two ages have an independent effect on current hazard.

13.3.3 Mutually adjusted effects of SBP at ages $t - 20$ and t

Table 13.5 shows the estimates of the mutually adjusted effects of SBP at age t and $t - 20$ on hazard of CVD at age t . Using either RC or ML, the estimates and confidence intervals suggest evidence that current SBP (relatively strong evidence) and SBP 20 years earlier (weak evidence) are independently and positively associated with current hazard of CVD. The point estimates from both RC and ML suggest that current SBP has a larger independent effect than SBP 20 years earlier on current hazard of CVD. The ML+MI estimates suggest a greater relative importance of current SBP compared to SBP 20 years earlier than the RC estimates, although the point estimates do not differ by a large amount.

Analogous to the unadjusted effects, we expect the RC estimates here to be positively biased if SBP 10 years earlier has an independent effect on current hazard of CVD. That the RC estimates place more relative importance on SBP levels 20 years earlier than the ML+MI estimates may again be due to the fact that SBP at ages between 40 and 50 had to be predicted on the basis of SBP measurements made at older ages for those men who were recruited to the study after age 50. Thus as for the unadjusted effect estimates, the RC estimates may place more importance on SBP 20 years earlier due to the violation of the implicit conditional independence assumption on which it relies.

Table 13.5: Estimates of mutually adjusted hazards ratios (95% bootstrap normal CI) for the effects of a 10mmHg increase in SBP at age t and $t - 20$ on hazard of CVD at age t . Estimates found using regression calibration (RC) and maximum likelihood assuming marginal normality for random-effects via ascent-based MCEM plus multiple-imputation (ML+MI)

SBP at	RC	ML+MI
t	1.170 (1.101, 1.242)	1.186 (1.112, 1.265)
$t - 20$	1.113 (1.030, 1.202)	1.096 (1.010, 1.188)

13.4 Conclusions

The estimates of the adjusted effects were much more precise than those of the adjusted odds ratios of Chapter 12. This is likely to due to a number of factors. First, the analysis in this chapter is based on a larger number of subjects. Second, by considering the occurrence of events over a longer period (20 compared to 10 years), and also by considering non-fatal CVD events, a greater proportion of the subjects in the analyses of this chapter experienced the event of interest. Third, the binary outcome used in Chapter 12 can be viewed as a coarsened version of the outcome ‘time to death due to CVD’, and such coarsening would in general be expected to result in a loss of information. Fourth, in the analyses of this chapter we considered the mutually adjusted effects of three correlated covariates, rather than four.

Our estimates suggest that, conditional on earlier SBP levels, current SBP is positively associated with current hazard of CVD. The estimated effect of a 10mmHg increase in current SBP, holding earlier SBP levels constant, was to increase current hazard of CVD by 13.0% (95% CI 5.3%, 21.3%) using RC, and by 16.0% (6.2%, 26.7%) using ML. Although the point estimate for the effect of SBP 20 years earlier on current hazard, conditional on current and SBP 10 years earlier, suggested a positive effect, the 95% confidence interval included the null value of 1, indicating the data are consistent with SBP level 20 years earlier having no independent effect on current hazard. The same was true for SBP 10 years earlier.

There was strong evidence that unconditionally (i.e. not adjusted for SBP at other times), current SBP, SBP 10 years earlier, and SBP 20 years earlier are each positively associated with current hazard of CVD, with unadjusted hazard ratio estimates of around 1.25 for a 10mmHg increase in SBP. There was also evidence that, current SBP and SBP 20 years earlier have independent effects on the current hazard of CVD.

Broadly speaking, the RC and ML estimates were quite similar. However, when estimating the unadjusted effects of SBP, the RC estimates were larger in magnitude than the ML+MI estimates. As previously discussed in Section 9.3 and shown in our simulations in Section 9.8, we expect RC to give biased estimates of effects when important time-dependent covariates which are a function of the longitudinal process

are omitted from the Cox outcome model. That the RC estimates of the unadjusted effects were larger than the corresponding ML+MI estimates is consistent with this, assuming that current SBP, SBP 10 years earlier, and SBP 20 years earlier have positive independent effects on current hazard of CVD.

However, as in our simulations of Section 9.8, these biases may have acted in the opposite direction to the bias towards the null inherent in RC estimates of effect in Cox models. The net effect was that the differences between the RC and ML+MI estimates were not particularly large. However, on the basis of our earlier simulation results (Section 9.8) we would expect greater discrepancies in situations in which the longitudinal process has a greater effect on hazard. Furthermore, in other situations the biases in RC estimates caused by omitting an important time-dependent covariate may act in the same direction as RC's inherent bias towards the null, causing larger differences between RC and estimates based on ML+MI. It is also important to note that the ML approach used assumes that hazard at time t only depends on past SBP via its current level, its level 10 years earlier, and its level 20 years earlier. In particular, it assumes that at ages between 70 and 80, SBP levels between 40 and 50 have no independent influence on current hazard of CVD.

Given these findings, at least for the Framingham data we have considered, we would not advocate use of the complicated ML or ML+MI method over RC, since the two methods gave estimates which differed by only relatively small amounts. The fact that RC is so easy to implement, relative to methods such as ML, and that it has often been found to give similar estimates to more complicated methods [94, 130], explains its continued attraction for estimating the parameters of joint models for longitudinal and time-to-event outcomes.

As discussed in Section 9.10, our implementation of ML, via ascent-based MCEM, is extremely slow, due to the inefficiency of the rejection sampling used to multiply impute \mathbf{X}_i (see Section 9.5.1). While a more computationally efficient implementation could no doubt be produced, our proposal is inherently computationally demanding due to the looseness of the bound used in the rejection sampler. We would not therefore advocate its use for finding ML estimates for such joint models, unless a more efficient approach to generating the required imputations could be found.

The analyses of this chapter are based on a model in which we assume that for each subject there exists a five-dimensional random-effects vector \mathbf{X}_i , of which the last two elements determine SBP at ages 70 and 80. Since some subjects died between the ages of 60 and 80, it is not clear how these random-effects can be unambiguously defined for all subjects. As discussed earlier in Section 9.9, we believe further research is needed regarding this issue.

Part IV

Conclusions

Chapter 14

Conclusions

In this thesis we have investigated methods to allow for classical covariate measurement error, and considered extensions to life-course studies. In Part I we considered the effects of, and methods to allow for, classical covariate measurement error in regression models for continuous, binary, and censored survival time outcomes. We summarize our contributions to this area and give some broad conclusions in Section 14.1. In Part II we considered the extension of the methods considered for classical covariate measurement to life-course studies. In such studies, interest lies in the relationships between aspects of the trajectories of longitudinal processes of interest, which are measured periodically and with error, with subsequent outcomes. We summarize our contributions from this part in Section 14.2. We illustrated some of the analysis methods we have described and proposed in two analyses of data from the Framingham Heart Study, which we summarize in Section 14.3. Lastly, in Section 14.4 we outline areas of interest for future research.

14.1 Classical covariate measurement error

In Part I of the thesis we described and compared some of the many estimation methods which have been proposed for dealing with classical covariate measurement error in continuous covariates. We have considered arguably the three most common types of outcome: continuous, binary, and censored survival times, focusing on the setting in which internal replication data are available.

14.1.1 Continuous outcomes

In Chapter 3 we reviewed the effects of classical covariate measurement error in linear regression models, the simple method of moments (MOM) correction approach, and introduced the regression calibration (RC) method. We compared these two approaches, highlighting the efficiency advantage of RC compared to MOM.

We then described the parametric maximum likelihood (ML) approach. The ML approach has been used less than RC in applied work, do to the increased

computational complexity and scarcity of software for finding ML estimates. We have shown how maximum likelihood (ML) estimates of the model (which assumes normality for the true covariate X_i measured with error, conditional on the outcome Y_i) parameters can be obtained by fitting a standard random-intercepts model to the error-prone measurements \mathbf{W}_i , with the outcome Y_i entering the model as a fixed effect. This work has recently been published in the journal *Statistics in Medicine* [27].

We reviewed the method of multiple imputation to deal with missing data, and discussed its application to deal with covariate measurement error. We showed how the fitted random-intercepts model involved in our novel approach to find ML estimates can be used to multiply impute the unobserved true covariate X_i . We noted however that having found the MLE, which is a simple function of the estimates of the fitted random-intercepts model, no efficiency can be gained by multiply imputing X_i , since ML estimators are asymptotically optimal, assuming the model is correctly specified. This was confirmed in our simulation results. We similarly showed how the recently proposed moment reconstruction (MR) method could be implemented in the context of internal replication data by using the estimated parameter values from the random-intercepts model used in our novel approach to find ML estimates, and confirmed in simulations that the resulting estimates had the same efficiency as ML.

For continuous outcomes, RC is an efficient and intuitive approach to dealing with classical covariate measurement error. It can be easily implemented in two stages using commands in modern statistical packages. Based on our simulation results, and in accordance with the findings of other studies, in many situations RC has efficiency which is very similar to that of more complicated (from an implementation and software perspective) methods such as ML. We would therefore recommend its use for dealing with classical covariate measurement error with continuous outcome models.

14.1.2 Binary outcomes

In Chapter 4 we considered binary outcomes, focusing on the popular logistic regression model. We showed how the Monte-Carlo Expectation Maximization (MCEM) algorithm can be used to find ML estimates for models in which a binary outcome is assumed to follow a logistic regression model. The implementation we considered involved, at each iteration of the MCEM algorithm, multiply imputing the true covariate X_i from its conditional distribution given observed data. By doing this, standard commands for fitting logistic regression models can be used in the M-step of EM. We also showed ‘ascent-based’ MCEM, can be implemented to control the number of imputations used in the EM algorithm and for deciding when the algorithm has converged.

We extended our novel approach for obtaining ML estimates by fitting a random-intercepts model to the case of binary outcomes, which was included in our Statistics in Medicine paper [27]. This approach provides ML estimates in the setting of internal replication data under the normal discriminant model that X_i is conditionally normal given the binary outcome Y_i . A limitation of this approach is that in the presence of error-free covariates \mathbf{Z}_i , the method can only be used under the restrictive assumption that X_i and \mathbf{Z}_i are multivariate normal given Y_i . To address this, we proposed using the fitted random-intercepts model to multiply impute X_i , and to estimate the logistic regression parameters by fitting logistic regression models using the imputed X_i . As the number of imputations is increased, the resulting estimates have the same asymptotic efficiency as the MLEs for the joint model which assumes conditional normality for X_i given Y_i and \mathbf{Z}_i .

The MR method for binary outcomes is based on the same parametric assumption of conditional normality for X_i given Y_i as our novel approach to ML estimation. Analogous to the case of continuous outcomes, in the case of internal replication data, there may be no benefit in using MR for estimation, given that the MLEs are available having fitted the random-intercepts model previously described.

We reviewed the conditional score (CS) method, which makes no distributional assumptions about the unobserved covariate X_i . Although developed over 20 years ago, this method has hardly been used as far as we aware in applied work, presumably due to its perceived complexity in implementation and a lack of available software commands. We were able to implement the method with relatively little programming in R, and to use commands to solve non-linear equations to find the CS estimates. In simulations, except when measurement errors were very large, the method was fast and had little bias.

RC gives approximately consistent estimates of the model parameters when the outcome model is logistic regression for a binary outcome. However, consistent with findings in the literature, we found in simulations that when the effect of the covariate measured with error is large estimates using RC are biased by a non-negligible amount. This bias is greater when the outcome is not rare. However, even in these situations, the biases were arguably still moderate, validating the view that for logistic regression RC is an attractive approach to estimation. However, our novel approach to ML estimation, based on fitting a random-intercepts model to the error-prone measurements, may provide an alternative approach in situations in which RC is expected to suffer from non-negligible biases.

14.1.3 Survival outcomes

In Chapter 5 we considered censored survival outcomes, focusing on Cox's proportional hazards model. We reviewed the application of RC to Cox's proportional hazards model, and the justification for the risk-set modified version of RC.

We showed how the MCEM algorithm can be used to find ML estimates for such models when covariates are measured with classical error. This included a novel proposal for how rejection sampling can be used to draw imputations from the conditional distribution of X_i given error-prone measurements \mathbf{W}_i and the censored survival outcome.

We applied recent results by White and Royston [103] concerning the use of multiple imputation to deal with missing covariates in Cox's proportional hazards model to the case of classical covariate measurement error. We showed how the random-intercepts model used in our novel approach to finding ML estimates in the case of continuous or binary outcomes could be adapted to include the censored survival outcome as fixed effects. However, in simulations we found that this approach carried considerable bias, and gave estimates which were at least as biased as those obtained using RC.

Our simulations confirmed previous findings that RC provides approximately consistent estimates for Cox proportional hazards outcome models, providing the covariate effect is not too large. As the effect size increases, and the proportion of subjects who are censored decreases, RC becomes increasingly biased. In contrast, MCEM, using our novel proposal to use rejection sampling to multiply impute X_i , resulted in estimates with little bias across all scenarios investigated. We also demonstrated through simulations that the conditional score (CS) method, which has been apparently little used in applied work, performed well in the simulation set up used, providing estimates with little bias, except when the covariate effects were moderate and measurement errors were large. Particularly for Cox proportional hazards outcome models, where RC may carry appreciable bias, methods such as ML (for example by MCEM) or CS should arguably be considered more often as an alternative to RC.

14.2 Extensions to life-course studies

In Part II of the thesis we considered the extension of these estimation methods to the more general setting where error-prone measurements follow a linear mixed model, in which the random-effects \mathbf{X}_i play the role of the covariates measured with error. This provides a framework for models used in life-course epidemiology, in which interest lies in relating characteristics of the trajectories of longitudinal processes to an outcome of interest. We showed how the estimation methods described for dealing with classical covariate measurement error can be extended to this more general setting.

14.2.1 Continuous outcomes

For continuous outcomes we showed how our novel approach to ML estimation in the case of classical measurement error, based on fitting a random-intercepts model for \mathbf{W}_i with Y_i as a fixed effect, can be easily extended to this more general setting. Our approach involves fitting the linear mixed model which is assumed for the longitudinal error-prone measurements, with a modification to the fixed effects structure to include the outcome Y_i .

Using simulations, mirroring our results from the classical covariate measurement error, we found that RC had efficiency which was very similar to ML. This was the case even in the situations in which we expected the efficiency difference between the two would be greatest: when the covariates \mathbf{X}_i were strongly predictive of the outcome Y_i , and when the number of error-prone measurements differed widely between subjects. On the basis of our simulations, and as for the case of classical covariate measurement error, RC is a highly efficient and easy to use estimation approach, and we would recommend its use in the the case of longitudinal error-prone measurements and continuous outcomes.

14.2.2 Binary outcomes

For binary outcomes we noted that the MCEM approach to ML estimation, described in detail in the case of classical covariate measurement error, is easily extended to the life-course setting in which error-prone measurements follow a general linear mixed model. Most implementations of ML for such models involve using quadrature techniques to approximate the intractable integrals which are involved in the EM algorithm. Quadrature methods rapidly become computationally infeasible as the dimension of the random-effects \mathbf{X}_i increases. The MCEM approach based on multiple imputation may therefore offer an attractive alternative approach to estimation when the dimension of \mathbf{X}_i is moderately large.

We showed how our novel approach to ML estimation in the case of classical measurement error, can also be adapted when the outcome is binary, under the normal discriminant model. However, this approach is limited because in the presence of error-free covariates \mathbf{Z}_i , it can only be used to find ML estimates under the restrictive assumption that \mathbf{X}_i and \mathbf{Z}_i are jointly normal given Y_i . To overcome this, we therefore proposed, analogously to the case of classical covariate measurement error, that the linear mixed model which includes Y_i and \mathbf{Z}_i as fixed effects can be used to multiply impute \mathbf{X}_i . Logistic regression models can then be fitted using these imputations of \mathbf{X}_i and the observed error-free covariates \mathbf{Z}_i .

In reviewing recent proposals extending the CS method to the longitudinal setting, we noted that a condition of the method for the available longitudinal measurements may limit its use in certain applications. Specifically, a subject can only

contribute to the analysis if all of the elements of \mathbf{X}_i are identifiable from their longitudinal measurements.

Using simulations we found that RC performed well, with biases of a similar magnitude to those in the case of classical measurement error. ML estimates obtained from both MCEM (assuming marginal normality for \mathbf{X}_i) or based on our novel proposal based on fitting a linear mixed model for the error-prone measurements \mathbf{W}_i conditional on the outcome Y_i (assuming conditional normality for \mathbf{X}_i given Y_i), had little bias. Mirroring our results from classical covariate measurement error, RC had efficiency similar to these more complicated techniques.

14.2.3 Survival outcomes

Estimation for models in which a longitudinal process and survival, or time-to-event outcome are observed has received a vast amount of attention in the methodological literature over the last 15 years. We aimed to concisely summarize the findings from this literature, including the extension of the RC and ML methods to this setting. For RC, we highlighted the different versions which have been employed, including RC based on a single linear mixed model fit to all available longitudinal measurements and the risk-set version. While the latter has been found to be less biased, it is more variable, and is far more computationally intensive.

We were able to generalize our rejection sampling approach from the simpler setting of time-independent covariates measured with classical error to the life-course setting in which the covariates are time-dependent and are measured longitudinally. However, the resulting algorithm is extremely slow and inefficient, due to the looseness of the bound derived for the rejection sampler. Although in a small simulation study the algorithm gave estimates with little bias, we would not recommend using our implementation of MCEM for finding ML estimates of such joint models due to its high computational cost.

In simulations, with small effects of \mathbf{X}_i non-risk set RC (i.e. based on a single linear mixed model fit for the longitudinal measurements) performed well, giving estimates with little bias. However, as with the case of time-independent covariates measured with classical error, we found that with larger effect sizes RC gave estimates with relatively large biases. With moderate effect sizes therefore, use of more complex methods such as ML may be required. A major limiting factor for the latter however is the lack of available software for fitting such models, although as we have noted, this is gradually being addressed.

In much of the literature regarding joint modelling of longitudinal and survival data, simple random-intercepts and slopes model are assumed for the longitudinal trajectory. We have explored recent suggestions to use random-effects models to allow more complicated evolutions in subject-specific trajectories over time. However, when modelling data in which subject die, following which the longitudinal process

is usually undefined, the use of such models may not be inappropriate. This is because some of the random-effects \mathbf{X}_i may be defined such that they only affect the value of the longitudinal process at later time points, by which time some subjects may have died. It is then not clear how to unambiguously define the meaning such random-effects for all subjects. We have suggested however, that this issue may not necessarily prevent valid inferences when interest lies in estimating the associations between the longitudinal process and the hazard of the event of interest, as opposed to marginal inference for the longitudinal process.

14.2.4 Alternative outcome model specifications

In applications we rarely, if ever, know which characteristics of a longitudinal process' trajectory are important for the outcome of interest. Furthermore, the components of the random-effects vector \mathbf{X}_i may be highly correlated, so that estimation of their independent effects on Y_i is difficult. If, upon fitting the 'full model' for Y_i given \mathbf{X}_i , we find that there is no evidence of an independent effect of one of the components of \mathbf{X}_i , this may well be as a result of very low power, as opposed to there being no independent effect. Thus while we may then wish to proceed to fit models for Y_i in which this component is no longer a covariate, we may be wary of making an assumption that it has no independent effect on Y_i . We may therefore be interested in fitting a number of different models, which have different specifications for how the longitudinal process influences the outcome, and where we may not necessarily wish to make the assumption that the omitted components of \mathbf{X}_i have no independent effect on the outcome.

RC apparently offers an appealing approach in this situation. Having specified a suitable linear mixed model for longitudinal measurements, we can predict the unobserved random-effects \mathbf{X}_i by their best linear unbiased predictions (BLUPs). We can then fit a variety of alternative outcome models for the outcome of interest, using different subsets, or functions of \mathbf{X}_i , as covariates. This approach has recently been used in an analysis of a longitudinal study, to investigate the relationship between SBP levels throughout adult life and subsequent mortality, by Boshuizen *et al* [119]. We have shown why, from analytical considerations, and empirically verified through simulations, such an approach may produce biased estimates of the parameters of interest. Such biases occur whenever, conditional on the subset or function of \mathbf{X}_i included in the outcome model, the longitudinal measurements \mathbf{W}_i are predictive of the outcome. The simplest example of this is when an element of \mathbf{X}_i is omitted from the outcome model when it is independently (of the included components of \mathbf{X}_i) predictive of the outcome. Our letter raising this issue in relation to the analysis by Boshuizen *et al* was recently published in the American Journal of Epidemiology [131].

For linear and logistic outcome models, we have shown how the parameters of alternative outcome models, with subsets or functions of \mathbf{X}_i as covariates can be expressed (sometimes approximately) in terms of the parameters of the outcome model in which \mathbf{X}_i is entered as a covariate (the ‘full’ model). These expressions can be used to find estimates of the parameters of alternative outcome models, by using estimates from the ‘full’ model and the estimated variance covariance matrix of \mathbf{X}_i . In simulations we confirmed that these expressions can be used to provide either unbiased (linear regression) or, under appropriate conditions, approximately consistent (logistic regression), estimates of the parameters of such alternative outcome models. We termed this approach ‘corrected RC’.

An alternative solution is to first create multiple imputations of \mathbf{X}_i , under the ‘full’ model which assumes the outcome may depend jointly on all elements of \mathbf{X}_i . Such imputations can be created using estimates of the model’s parameters. In the case of continuous outcomes, imputation is simple under an assumption of multivariate normality. For binary and survival outcomes, the rejection sampling scheme we have used as part of the MCEM algorithm can be used. Indeed, a by-product of the MCEM algorithm we used to find ML estimates is a large set of multiple imputations of \mathbf{X}_i , imputed under the ‘full’ model.

Multiple imputations of \mathbf{X}_i can be used to directly estimate the parameters of alternative outcome models. In the simplest case in which interest lies in an outcome model with one or more elements of \mathbf{X}_i not included as covariates, we can simply fit the outcome model using the imputations of only those elements of \mathbf{X}_i we wish to include. More generally, if we are interested in the effect of a function of \mathbf{X}_i on the outcome, we can apply this function to the imputations of \mathbf{X}_i and fit the outcome model with the resulting values as covariates. This approach may be particularly appealing in the context of Cox proportional hazards models with time-dependent covariates, for which we do not believe simple expressions relating the parameters of the full model to those of alternative outcome models are available. As we have discussed previously, our proposal for using the MCEM algorithm to find MLEs when the outcome model is Cox’s proportional hazards regression with time-dependent covariates, and as a by-product, multiple imputations of \mathbf{X}_i , is very computationally intensive. A more viable alternative may therefore be to estimate the full model parameters using an approach such as RC, and then to create multiple imputations, using the RC estimates of the full model parameters in the rejection sampling procedure, as opposed to the MLEs.

14.3 Analyses of data from the Framingham Heart Study

In Part III we reported the results of two illustrative analyses, using data from men in the Framingham Heart Study. In Chapter 12 we reported analyses considering the odds of death due to CVD between ages 70 and 80 in those men who were alive and had had no previous CVD at age 70. We proposed a piece-wise linear model for the trajectories of men's SBP between ages 40 and 70, with knots at ages 50 and 60. Our analyses illustrated the application of our proposed corrected RC approach when some components of the random-effects vector \mathbf{X}_i are omitted from the logistic regression outcome model. We also illustrated our proposals to combine ML via MCEM followed by MI to estimate the parameters of such models, and also estimates based on our proposal involving fitting a mixed model to the longitudinal measurements conditional on the binary outcome. With the caveat that the analyses are illustrative and ignore many important epidemiological aspects, our results suggested that SBP levels between the ages of 40 and 70 are associated with increase odds of death due to CVD between ages 70 and 80. Estimates of the independent effects of SBP at the various ages were very imprecise, due to the high correlations of SBP within subjects over time.

In Chapter 13 we considered models for relating the hazard of CVD events (non-fatal and fatal) between the ages of 60 and 80 to both current and earlier levels of SBP. We illustrated the application of non-risk-set RC, and also our proposed approach to ML estimation based on the MCEM algorithm. However, given the very high computational burden of the latter approach, we would not advocate its use. Our results suggested that systolic blood pressure (SBP) is positively associated with the incidence of cardiovascular disease (CVD), and that, conditional on current SBP, SBP levels 20 years earlier may have an independent effect on current hazard of CVD.

Our analyses raised a number of important practical issues in studying the relationship between longitudinal processes and outcomes of interest. The trajectories of longitudinal processes during life are often relatively smooth, so that levels of the process at different times are highly correlated. On the basis of our illustrative analyses of data from the Framingham Study, SBP levels at different ages during adult life were highly correlated within men. Such high correlations mean that estimation of the independent contributions of the longitudinal process levels at different times to the outcome of interest is difficult. One approach may be to model the longitudinal process in a simpler way, using a random-effects vector of smaller dimension. Such models may however be overly simplistic for longitudinal processes measured over long periods of life, such as SBP. Thus unless subject matter knowledge indicates particular features of trajectories which determine the outcome, this naturally

leads to richly specified longitudinal models and exploratory analyses of how the outcome depends on various aspects of the longitudinal trajectory. We believe that some of the methodology (i.e. that for alternative outcome model specifications) we have proposed may be useful in this context.

Another pervasive issue, which is particularly relevant in the context of longitudinal processes measured periodically with error, is that of missing data. In our analyses of the Framingham data many men had ‘missing’ SBP measurements from the earlier decades of adult life due to the fact they were recruited into the study at an older age. Other measurements were missing because men missed one or more follow-up visits for some reason.

The RC approach can be used in such situations where the longitudinal data are highly unbalanced, provided the probability of missing measurements depends at most on the observed longitudinal measurements \mathbf{W}_i and the error-free covariates \mathbf{Z}_i . The ML approach also accommodates such missing data. For continuous and binary outcomes ML is valid under the less restrictive assumption that missingness depends at most on the observed longitudinal measurements \mathbf{W}_i , the error-free covariates \mathbf{Z}_i , and the outcome Y_i . For censored survival time outcomes, the ML approach is valid providing the probability of a subject missing a longitudinal measurement depends at most on the past longitudinal measurements and any error-free covariates \mathbf{Z}_i .

The CS approach, which makes no distributional assumptions regarding the random-effects \mathbf{X}_i is a potentially a useful alternative estimation approach. However, in the longitudinal setting it requires (for binary outcomes) that each subject has sufficient longitudinal measurements to identify \mathbf{X}_i . This may limit its applicability for certain applications, for as in our analysis of odds of death due to CVD using the Framingham data, a large number of subjects may not satisfy this requirement. For survival outcomes, subjects can only contribute to a risk-set if their preceding longitudinal measurements identify the time-dependent covariate \mathbf{X}_i^* . Unfortunately, due to time limitations, we were unable to investigate the use of this method using the Framingham data.

14.3.1 Limitations

Our analyses of the Framingham data were intended to be illustrative of some of the estimation methods previously described, and there were a number of limitations to them.

In our analyses we chose to reduce the data by using the mean of the three SBP measurements which took place at most follow-up visits in our analyses, rather than modelling the individual measurements themselves. A benefit of this was that the computational burden for estimation was reduced, partly due to the fitting of simpler models, and partly because the number of error-prone measurements in the analyses was reduced. It also meant that the R programs we had written for our

earlier simulation studies could be adapted with minimal modifications. There are however a number of disadvantages to this choice. First, we lost information from discarding SBP measurements at visits at which a subject had only one or two SBP measurements made. Second, it meant that we were unable to decompose within-subject variability in SBP measurements. Instead of modelling the mean of three SBP measurements at a visit, we could have specified three level longitudinal models for the individual SBP measurements. Such models would allow us to decompose within-subject variability into within-visit variability, which might be thought of as ‘pure’ measurement error, and deviations of the visit specific ‘true’ values around a subject’s underlying trajectory (as defined by subject level random effects \mathbf{X}_i). The magnitude of the variance corresponding to these visit effects would give an indication as to the adequacy of a particular specification for the underlying trajectories of SBP, as dictated by the random subject effects \mathbf{X}_i . We believe that each of the estimation methods discussed in Chapters 8 and 9 could be used with such a three-level model for SBP measurements, including our approach to ML estimation based on fitting a mixed model for the error-prone measurements conditional on the outcome Y_i . Although the MCEM algorithm could also be applied with such a model, this would involve generating and storing multiple imputations of random visit effects for each visit at which a subject had SBP measured. Given the large number of follow-up visits in Framingham, this would substantially increase the computational burden of using the MCEM algorithm.

We found evidence that within-subject variability in SBP measurements increased with the underlying mean being measured, but since the relationship was moderately small in magnitude, we ignored this in our analyses. Ideally this should be accommodated either by modelling the variance function, or by applying some kind of transformation. We also found evidence that the distribution of subjects’ underlying SBP levels was somewhat skewed. Our analyses assumed normality for the distribution of underlying SBP, since it has been shown that parametric methods are robust to non-normality of the random-effects \mathbf{X}_i when information on them from longitudinal measurements is sufficiently rich. However, we did not investigate whether in the Framingham study there was sufficient information in subject’s longitudinal measurements for this robustness property to hold. Again applying a transformation to the SBP measurements may have improved the normality assumption. Alternatively, we could have used the SIMEX methods as proposed by Huang *et al* [122] to assess whether the violation of the normality assumption adversely affected our estimators.

A further limitation of our analyses is that we did not include any covariates in our models. A more thorough analysis would include other covariates which are predictive of CVD, such as weight, cholesterol, and smoking status. Weight and cholesterol both change throughout the life-course, and are measured with varying

degrees of measurement error, and so would need to be included as secondary longitudinal processes. Similarly, smoking status often changes during life. However, assuming self-reported smoking status can be assumed to be error-free, this might be included in outcome models as a directly observed, time-varying covariate. The inclusion of such additional time-varying covariates introduces the complexity of time-varying confounding, a subject which we have not discussed, but for which there is a growing statistical literature [141].

14.4 Future research

14.4.1 Software and implementation

We believe our novel proposal for finding ML estimates by fitting a linear mixed model to the error-prone measurements \mathbf{W}_i , conditional on the outcome Y_i , offers an appealing approach to finding ML estimates. Although it can be implemented using standard linear mixed model commands, the method could be programmed as a single command (similar to the `gllamm` command in Stata) to enable researchers to use it routinely.

The ascent-based MCEM approach we have used required a moderate amount of programming effort in R, and certainly more than many researchers would be willing, or perhaps able, to do. Packaging the algorithm into a single command, for a statistical package such as Stata or R, would enable greater use, and would permit a more computationally efficient implementation. However, we believe that additional experience is needed to ensure its performance (statistically and computationally) is satisfactory, before programming the algorithm into a command/package.

The availability (or rather lack) of software is probably the main factor why other methods, such as the CS approach, are not used more routinely. Thus it would be desirable if such methods were wrapped into single commands. This would require relatively little programming effort, since it would consist only of writing functions to evaluate the estimating equations, their first derivatives, and then passing these to a non-linear equation solver which is usually available in statistical packages.

14.4.2 Methodology

In our investigations into classical covariate measurement error we focused (exclusively in our simulations) predominantly on the case in which a single covariate X_i is measured with error. Although we indicated how the various methods could be extended to the case of multivariate \mathbf{X}_i , it would be of interest to investigate the performance of the various methods in this case. Our simulations in Chapters 7, 8, and 9 addressed this to a certain extent, but they assumed at most a correlation of

0.5 between elements of \mathbf{X}_i . As our analyses of the Framingham data showed, in applications the correlations between covariates may often be much larger.

Further research is needed regarding the use of models for longitudinal and survival data in which random-effects are defined which only affect the value of the longitudinal process at later times, by which point some subjects may have died. In such settings, it is not clear how such random-effects can be defined without ambiguity or resorting to definitions in terms of a hypothetical situation where death does not terminate the existence of the longitudinal process in question.

In recent years statistical methods have been developed to analyse data in which there are time-varying exposures and confounders [141]. In these methods it is usually assumed that exposures and confounders are measured at discrete time points and without any measurement error. These methods, and the principles they are founded on, undoubtedly have implications for the analysis of data in which the exposures and confounders are measured periodically, and with measurement error.

14.4.3 Framingham Heart Study

Our analyses of the Framingham Study data have suggested that risk of CVD may depend jointly on both recent levels of SBP and levels of SBP earlier in life. However we have ignored many important factors which would need to be considered in a more thorough epidemiological investigation of the data. We would therefore plan to perform a more detailed analysis of the Framingham Study data, to investigate more fully the contributions of blood pressure levels throughout the life-course to CVD risk later in life.

Appendix: R code for simulations with logistic regression subject to classical covariate measurement error

Listing 14.1: R code for logistic regression simulations with X_i marginally normal, subject to classical measurement error, corresponding to results in Tables 4.1 and 4.2

```
args <- commandArgs(TRUE)
#arguments to R program
# 1 = fileNamePrefix
simulationSet <- Sys.getenv("SGE_TASK_ID")
fileNamePrefix <- args[1]
library(lme4)
library(nleqslv)
runSimulations <- function() {
  createImputations <- function(numberOfImputationsNeeded) {
    newImputations <- array(0, dim=c(n,
      numberOfImputationsNeeded))
    for (i in 1:n) {
      numberNeeded <- numberOfImputationsNeeded
      numberAlreadyFound <- 0
      while (numberNeeded>0) {
        #generate new draw from X given W
        newDrawXGivenW <- rnorm(numberNeeded, mean=
          xBLUP[i], sd=VarXGivenW[i]^0.5)
        newUniform <- runif(numberNeeded)
        rejectionProb <- (1-y[i])*(1/(1+exp(beta0+
          betaX*newDrawXGivenW)))+y[i]*exp(beta0+
          betaX*newDrawXGivenW)/(1+exp(beta0+betaX
          *newDrawXGivenW))
        acceptReject <- 1*(newUniform<=
          rejectionProb)
        if (sum(acceptReject)>0) {
          successfulImputations <-
            newDrawXGivenW[acceptReject==1]
```

```

        newImputations[i,(
            numberAlreadyFound+1):(
                numberAlreadyFound+sum(
                    acceptReject))] <-
                newDrawXGivenW[acceptReject==1]
        numberAlreadyFound <-
            numberAlreadyFound + sum(
                acceptReject)
        numberNeeded <- numberNeeded-sum(
            acceptReject)
    }
}
}
newImputations
}
#conditional score equations
equation1 <- expression( y - exp(CSbeta0+CSbetaX*(wMean+y*(sigma_u_
sq/n_i)*CSbetaX)-0.5*CSbetaX*2*(sigma_u_sq/n_i))/(1+exp(CSbeta0+
CSbetaX*(wMean+y*(sigma_u_sq/n_i)*CSbetaX)-0.5*CSbetaX*2*(sigma_
u_sq/n_i))))
equation2 <- expression( (y - exp(CSbeta0+CSbetaX*(wMean+y*(sigma_u_
sq/n_i)*CSbetaX)-0.5*CSbetaX*2*(sigma_u_sq/n_i))/(1+exp(CSbeta0
+CSbetaX*(wMean+y*(sigma_u_sq/n_i)*CSbetaX)-0.5*CSbetaX*2*(sigma
_u_sq/n_i))) )*(wMean+y*(sigma_u_sq/n_i)*CSbetaX))
D11 <- D(equation1, "CSbeta0")
D12 <- D(equation1, "CSbetaX")
D21 <- D(equation2, "CSbeta0")
D22 <- D(equation2, "CSbetaX")
CSestimatingEquations <- function(parameter) {
    CSbeta0 <- parameter[1]
    CSbetaX <- parameter[2]
    array(rowMeans(rbind(eval(equation1), eval(equation2))),
        dim=c(2,1))
}
CSgradient <- function(parameter) {
    CSbeta0 <- parameter[1]
    CSbetaX <- parameter[2]
    array(c(mean(eval(D11)), mean(eval(D21)), mean(eval(D12)),
        mean(eval(D22))), dim=c(2,2))
}
fileName <- paste(fileNamePrefix, "_", simulationNumber, "_",
    simulationSet, ".dat", sep="")
mu_x_true <- 0
sigma_x_sq_true <- 1
sigma_u_sq_true <- sigma_x_sq_true*(1/reliability-1)
simulations <- 100
n_i <- c(rep(2,n1), rep(1,(n-n1)))
rcalEstimates <- array(0, c(simulations))
rcalSE <- array(0, c(simulations))

```

```

conditionalScoreEstimates <- array(0, c(simulations))
conditionalScoreConverged <- array(0, c(simulations))
mlEstimates <- array(0, c(simulations))
mcmEstimates <- array(0, c(simulations))
pseudoMcmEstimates <- array(0, c(simulations))
GammaY <- array(0, c(simulations))
sigmabArray <- array(0, c(simulations))
betaSE <- array(0, c(simulations))
varPhi13Array <- array(0, c(simulations))
idealEstimates <- array(0, c(simulations))
fiellerCILower <- array(0, c(simulations))
fiellerCIUpper <- array(0, c(simulations))
mcmSE <- array(0, c(simulations))
#MCEM configuration
maxNumEMIterations <- 100
maxImputations <- 1000
imputations <- array(0, dim=c(n,maxImputations))
beta0Estimates <- array(0, dim=maxImputations)
betaXEstimates <- array(0, dim=maxImputations)
sigma_u_sqEstimates <- array(0, dim=maxImputations)
sigma_x_sqEstimates <- array(0, dim=maxImputations)
alphaLevel <- 0.25
betaLevel <- 0.25
gammaLevel <- 0.05
zAlpha <- qnorm(1-alphaLevel)
zBeta <- qnorm(1-betaLevel)
zGamma <- qnorm(1-gammaLevel)
QIncreaseConvergenceCriterion <- 0.01
for (simul_n in 1:simulations) {
  newSeed <- (as.numeric(simulationSet)-1)*(simulations*
    scenarios)+(simulationNumber-1)*simulations + simul_n
  set.seed(newSeed)
  #print(simul_n)
  #simulate data
  x <- rnorm(n, mean=mu_x_true, sd=sigma_x_sq_true^0.5)
  w1 <- x+rnorm(n, mean=0, sd=sigma_u_sq_true^0.5)
  w2 <- x+rnorm(n, mean=0, sd=sigma_u_sq_true^0.5)
  xb <- beta0True+betaXTrue*x
  prob <- exp(xb)/(1+exp(xb))
  y <- 1*(runif(n)<prob)
  analysisMod <- glm(y ~ x, family="binomial")
  idealEstimates[simul_n] <- analysisMod$coef[2]
  longData <- array(c(1:n, y, w1), dim=c(n,3))
  longData <- rbind(longData, array(c(1:n1, y[1:n1], w2[1:n1
    ]), dim=c(n1,3)))
  longData <- data.frame(longData)
  colnames(longData) <- c("id", "y", "w")
  wMean <- w1
  wMean[1:n1] <- (w1[1:n1]+w2[1:n1])/2

```

```

#regression calibration
rcalMod <- lmer(w~1+(1|id), data=longData, REML=FALSE)
vc <- VarCorr(rcalMod)
mu_x <- fixef(rcalMod)
sigma_u_sq <- attr(vc, "sc")^2
sigma_x_sq <- vc[[1]][1]
xBLUP <- (sigma_x_sq/(sigma_x_sq+sigma_u_sq/n_i))*wMean
VarXGivenW <- sigma_x_sq*(1-sigma_x_sq/(sigma_x_sq+sigma_u_sq/n_i))
analysisMod <- glm(y ~ xBLUP, family="binomial")
rcalEstimates[simul_n] <- analysisMod$coef[2]
rcalSE[simul_n] <- vcov(analysisMod)[2,2]^0.5
#conditional score
fit <- nleqslv(c(analysisMod$coef[1], analysisMod$coef[2]),
  CSestimatingEquations, CSgradient)
conditionalScoreEstimates[simul_n] <- fit$x[2]
conditionalScoreConverged[simul_n] <- fit$termcd
#ML under assumption of marginal normality for X, using
  ascent-based MCEM
beta0 <- analysisMod$coef[1]
betaX <- analysisMod$coef[2]
EMIteration <- 0
EMConverged <- FALSE
while (EMConverged==FALSE & EMIteration<maxNumEMIterations)
  {
    EMIteration <- EMIteration + 1
    if (EMIteration==1) {
      additionalNumImputations <- 10
    } else {
      additionalNumImputations <- max(
        totalNumImputations, round( (var(
          QIncrease)*(zAlpha+zBeta)^2) / ((mean(
            QIncrease))^2) ))
      additionalNumImputations <- min(
        maxImputations, additionalNumImputations
      )
    }
    print(paste("Iteration number:", EMIteration, sep=
      ""))
    print(paste("Starting number of imputations:",
      additionalNumImputations, sep=""))
    iterationCompleted <- FALSE
    numImputations <- 0
    xBLUP <- mu_x+(sigma_x_sq/(sigma_x_sq+sigma_u_sq/n_i))*
      (wMean-mu_x)
    VarXGivenW <- sigma_x_sq*(1-sigma_x_sq/(sigma_x_sq+
      sigma_u_sq/n_i))
    while (iterationCompleted==FALSE) {
      #generate additional imputations required

```

```

totalNumImputations <- numImputations +
  additionalNumImputations
additionalImputationsIndexVector <- (
  numImputations+1):totalNumImputations
imputations[,
  additionalImputationsIndexVector] <-
  createImputations(
    additionalNumImputations)
#maximize complete data likelihood in
  additional imputations
#first maximize f(y|x)
for (i in additionalImputationsIndexVector)
{
  logisticModel <- glm(y ~
    imputations[,i], family=binomial
  )
  beta0Estimates[i] <- logisticModel$
    coef[1]
  betaXEstimates[i] <- logisticModel$
    coef[2]
}
beta0New <- mean(beta0Estimates[1:
  totalNumImputations])
betaXNew <- mean(betaXEstimates[1:
  totalNumImputations])
#maximize f(w|x)
for (i in additionalImputationsIndexVector)
{
  u <- w1-imputations[,i]
  u <- c(u, w2[1:n1]-imputations[1:n1
    ,i])
  sigma_u_sqEstimates[i] <- sum(u^2)/
    (n+n1)
}
sigma_u_sqNew <- mean(sigma_u_sqEstimates
  [1:totalNumImputations])
#maximize f(x)
mu_xNew <- mean(imputations[,1:
  totalNumImputations])
for (i in additionalImputationsIndexVector)
{
  sigma_x_sqEstimates[i] <- (var(
    imputations[,i])*(n-1))/n
}
sigma_x_sqNew <- mean(sigma_x_sqEstimates
  [1:totalNumImputations])
#estimate increase in expected complete
  data log likelihood

```

```

#first estimate Q at new parameter
  estimates
logLikXNew <- colSums(-0.5*log(sigma_x_
  sqNew)-((imputations[,1:
  totalNumImputations]-mu_xNew)^2)/(2*
  sigma_x_sqNew))
logLikUNew <- -(n+n1)*0.5*log(sigma_u_sqNew
  )-(1/(2*sigma_u_sqNew))*sigma_u_
  sqEstimates[1:totalNumImputations]*(n+n1
  )
logLikYXNew <- y %%% (beta0New+betaXNew*
  imputations[,1:totalNumImputations])-
  colSums(log(1+exp(beta0New+betaXNew*
  imputations[,1:totalNumImputations])))
logLikNew <- logLikXNew+logLikUNew+as.
  vector(logLikYXNew)
#estimate Q function at last iteration's
  estimates
logLikX <- colSums(-0.5*log(sigma_x_sq)-((
  imputations[,1:totalNumImputations]-mu_x
  )^2)/(2*sigma_x_sq))
logLikU <- -(n+n1)*0.5*log(sigma_u_sq)-(1/
  (2*sigma_u_sq))*sigma_u_sqEstimates[1:
  totalNumImputations]*(n+n1)
logLikYX <- y %%% (beta0+betaX*imputations
  [,1:totalNumImputations])-colSums(log(1+
  exp(beta0+betaX*imputations[,1:
  totalNumImputations])))
logLik <- logLikX+logLikU+as.vector(
  logLikYX)
#estimate increase in Q function
QIncrease <- logLikNew-logLik
print(paste("QIncrease: ", mean(QIncrease),
  sep=""))
#estimate asymptotic standard error
ase <- sd(QIncrease)/(totalNumImputations
  ^0.5)
#calculate CI lower bound for increase
QIncreaseLowerBound <- mean(QIncrease) -
  zAlpha*ase
print(paste("QIncrease CI lower bound: ",
  QIncreaseLowerBound, sep=""))
if ((QIncreaseLowerBound>0) | (
  totalNumImputations==maxImputations)) {
  iterationCompleted <- TRUE
} else {
  #increase number of imputations
  numImputations <-
    totalNumImputations

```

```

        additionalNumImputations <- round(
            numImputations/3)
        if (numImputations+
            additionalNumImputations >
            maxImputations) {
            additionalNumImputations <-
                maxImputations -
                numImputations
        }
        print(paste("Increasing number of
            imputations to:", numImputations
                +additionalNumImputations, sep="
            "))
    }

}

#save updated parameter estimates
mu_x <- mu_xNew
sigma_x_sq <- sigma_x_sqNew
sigma_u_sq <- sigma_u_sqNew
beta0 <- beta0New
betaX <- betaXNew
#decide whether MCEM has converged
QIncreaseUpperBound <- mean(QIncrease)+zGamma*ase
if ((QIncreaseUpperBound <
    QIncreaseConvergenceCriterion) | (
    totalNumImputations==maxImputations)) {
    EMConverged <- TRUE
}

}

mcemEstimates[simul_n] <- betaX
#estimate standard errors
outcomeInformationSum <- array(0, dim=c(2, 2))
scoreArray <- array(0, dim=c(totalNumImputations, 5))
for (imputation in 1:totalNumImputations) {
    #outcome model
    X <- cbind(array(1, dim=c(n,1)), imputations[,
        imputation])
    mu_i <- exp(X %*% c(beta0,betaX))/(1 + exp(X %*% c(
        beta0,betaX)))
    v_i <- mu_i*(1-mu_i)
    outcomeScore <- t(array(y, dim=c(n,1))-mu_i) %*% X
    outcomeInformationSum <- outcomeInformationSum +
        array(c(sum(v_i), sum(imputations[,imputation]*v
            _i), sum(imputations[,imputation]*v_i), sum(
            imputations[,imputation]^2*v_i)), dim=c(2,2))
    #true covariate model
    trueCovariateScore <- c(sum(imputations[,imputation
        ]-mu_x)/sigma_x_sq, -n/(2*sigma_x_sq)+sum((

```

```

        imputations[,imputation]-mu_x)^2)/(2*sigma_x_sq
        ^2))
    #measurement model
    measurementScore <- -(n+n1)/(2*sigma_u_sq) + sigma_
        u_sqEstimates[imputation]*(n+n1)/(2*sigma_u_sq
        ^2)
    scoreArray[imputation,] <- c(outcomeScore,
        trueCovariateScore, measurementScore)
}
trueCovariateInformation <- array(c(n/sigma_x_sq, 0, 0, n/
    (2*sigma_x_sq^2)), dim=c(2,2))
measurementInformation <- (n+n1)/(2*sigma_u_sq^2)
observedInformation <- bdiag((1/totalNumImputations)*
    outcomeInformationSum, trueCovariateInformation,
    measurementInformation) - var(scoreArray)
varcov <- solve(observedInformation)
mcmSE[simul_n] <- varcov[2,2]^0.5
# ML (under assumption of X/Y normal)
mlMod <- lmer(w ~ y +(1|id), data=longData, REML=FALSE)
#extract parameter estimates
Gamma <- array(fixef(mlMod), dim=c(2,1))
GammaY[simul_n] <- Gamma[2,1]
vc <- VarCorr(mlMod)
sigma_u_sq <- attr(vc, "sc")^2
sigma_b_sq <- as.numeric(vc[[1]])
sigmabArray[simul_n] <- sigma_b_sq
#calculate ML estimate (under assumption of X/Y normal) of
    beta
mlEstimates[simul_n] <- Gamma[2,1]/sigma_b_sq
phi12 <- Gamma[2,1]
phi13 <- sigma_b_sq
varPhi12 <- vcov(mlMod)[2,2]
#first calculate elements for those subjects with 1 w
k <- 1
V <- array(1, dim=c(k,k))*sigma_b_sq+diag(1,k)*sigma_u_sq
v_inv <- solve(V)
res <- w1[(n1+1):n]-Gamma[1]-Gamma[2]*y[(n1+1):n]
resSumSq <- t(res) %*% res
lsigma_b_sqsigma_b_sq <- (n-n1)*0.5*sum(diag(v_inv %*%
    array(1, dim=c(k,k)) %*% v_inv %*% array(1, dim=c(k,k)))
    ) - sum(diag( v_inv %*% array(1, dim=c(k,k)) %*% v_inv %
    %* array(1, dim=c(k,k)) %*% v_inv %*% resSumSq ))
lsigma_u_sqsigma_u_sq <- (n-n1)*0.5*sum(diag(v_inv %*% v_
    inv)) - sum(diag( v_inv %*% v_inv %*% v_inv %*% resSumSq
    ))
lsigma_b_sqsigma_u_sq <- (n-n1)*0.5*sum(diag(v_inv %*%
    array(1, dim=c(k,k)) %*% v_inv)) - sum(diag( v_inv %*%
    array(1, dim=c(k,k)) %*% v_inv %*% v_inv %*% resSumSq ))
#now for subjects with 2 ws

```

```

k <- 2
V <- array(1, dim=c(k,k))*sigma_b_sq+diag(1,k)*sigma_u_sq
v_inv <- solve(V)
res <- cbind(w1[1:n1]-Gamma[1]-Gamma[2]*y[1:n1], w2[1:n1]-
  Gamma[1]-Gamma[2]*y[1:n1])
resSumSq <- t(res) %*% res
lsigma_b_sqsigma_b_sq <- lsigma_b_sqsigma_b_sq + n1*0.5*sum
  (diag(v_inv %*% array(1, dim=c(k,k)) %*% v_inv %*% array
  (1, dim=c(k,k)))) - sum(diag( v_inv %*% array(1, dim=c(k
  ,k)) %*% v_inv %*% array(1, dim=c(k,k)) %*% v_inv %*%
  resSumSq ))
lsigma_u_sqsigma_u_sq <- lsigma_u_sqsigma_u_sq + n1*0.5*sum
  (diag(v_inv %*% v_inv)) - sum(diag( v_inv %*% v_inv %*%
  v_inv %*% resSumSq ))
lsigma_b_sqsigma_u_sq <- lsigma_b_sqsigma_u_sq + n1*0.5*sum
  (diag(v_inv %*% array(1, dim=c(k,k)) %*% v_inv)) - sum(
  diag( v_inv %*% array(1, dim=c(k,k)) %*% v_inv %*% v_inv
  %*% resSumSq ))
varcov <- array(c(lsigma_b_sqsigma_b_sq, lsigma_b_sqsigma_u
  _sq,
  lsigma_b_sqsigma_u_sq, lsigma_u_sqsigma_u
  sq), dim=c(2,2))
varPhi13 <- solve(-varcov)[1,1]
varPhi13Array[simul_n] <- varPhi13
term1 <- 1/sigma_b_sq
term2 <- -Gamma[2,1]/(sigma_b_sq^2)
betaSE[simul_n] = (varPhi12*term1^2 + varPhi13*term2^2 )
  ^0.5
f0=phi12^2-1.96^2*varPhi12
f1=phi12*phi13
f2=phi13^2-1.96^2*varPhi13
fiellerCILower[simul_n] =(f1-(f1^2-f0*f2)^0.5)/f2
fiellerCIUpper[simul_n]=(f1+(f1^2-f0*f2)^0.5)/f2
}
write.table(cbind(idealEstimates, rcalEstimates, rcalSE,
  mlEstimates, betaSE, fiellerCILower, fiellerCIUpper,
  mcemEstimates, mcemSE, conditionalScoreEstimates,
  conditionalScoreConverged),
  file=fileName)
}
scenarios <- 12
#parameters of simulation
n <- 5000
n1 <- 500
#beta0True=-2.2 needed for betaXTrue=0.1 to get 10% prevalence
#beta0True=-2.57 needed for betaXTrue=1
simulationNumber <- 1
beta0True <- -2.2
betaXTrue <- 0.1

```

```

reliability <- 2/3
runSimulations()
simulationNumber <- 2
beta0True <- -2.2
betaXTrue <- 0.1
reliability <- 1/2
runSimulations()
simulationNumber <- 3
beta0True <- -2.2
betaXTrue <- 0.1
reliability <- 1/3
runSimulations()
simulationNumber <- 4
beta0True <- -2.57
betaXTrue <- 1
reliability <- 2/3
runSimulations()
simulationNumber <- 5
beta0True <- -2.57
betaXTrue <- 1
reliability <- 1/2
runSimulations()
simulationNumber <- 6
beta0True <- -2.57
betaXTrue <- 1
reliability <- 1/3
runSimulations()
simulationNumber <- 7
beta0True <- 0
betaXTrue <- 0.1
reliability <- 2/3
runSimulations()
simulationNumber <- 8
beta0True <- 0
betaXTrue <- 0.1
reliability <- 1/2
runSimulations()
simulationNumber <- 9
beta0True <- 0
betaXTrue <- 0.1
reliability <- 1/3
runSimulations()
simulationNumber <- 10
beta0True <- 0
betaXTrue <- 1
reliability <- 2/3
runSimulations()
simulationNumber <- 11
beta0True <- 0

```

```
betaXTrue <- 1
reliability <- 1/2
runSimulations()
simulationNumber <- 12
beta0True <- 0
betaXTrue <- 1
reliability <- 1/3
runSimulations()
```

Bibliography

- [1] S. MacMahon, R. Peto, J. Cutler, R. Collins, P. Sorlie, J. Neaton, R. Abbott, J. Godwin, A. Dyer, and J. Stamler. Blood pressure, stroke, and coronary heart disease. Part 1, Prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet*, 335:765–774, 1990.
- [2] B. Rosner, D. Spiegelman, and W. C. Willett. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology*, 132:734–743, 1990.
- [3] W. C. Willett, D. J. Hunter, M. J. Stampfer, G. Colditz, J. E. Manson, D. Spiegelman, B. Rosner, C. H. Hennekens, and F. E. Speizer. Dietary fat and fiber in relation to risk of breast cancer. an 8-year follow-up. *Journal of the American Medical Association*, 268:2037–2044, 1992.
- [4] D. A. Pierce, D. O. Stram, M. Vaeth, and D. W. Schafer. The errors-in-variables problem: Considerations provided by radiation dose-response analyses of the a-bomb survivor data. *Journal of the American Statistical Association*, 87:351–359, 1992.
- [5] A. Phillips and G. Davey-Smith. How independent are "independent" effects? relative risk estimation when correlated exposures are measured imprecisely. *Journal of Clinical Epidemiology*, 44:1223–1231, 1991.
- [6] A. Wald. The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, 11(3):284–300, sep 1940.
- [7] W. A. Fuller. *Measurement error models*. John Wiley & Sons Inc, 1987.
- [8] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu. *Measurement Error in Nonlinear Models*. Chapman & Hall/CRC, 2nd edition, 2006.
- [9] E. B. Rimm, A. Ascherio, E. Giovannucci, D. Spiegelman, M. J. Stampfer, and W. C. Willett. Vegetable, fruit and cereal fiber intake and risk of coronary heart disease among men. *JAMA*, 275:447 – 451, 1996.

- [10] C. Iribarren, D. Sharp, C. M. Burchfiel, P. Sun, and J. H. Dwyer. Association of serum total cholesterol with coronary disease and all-cause mortality: multivariate correction for bias due to measurement error. *American Journal of Epidemiology*, 143:463–471, 1996.
- [11] Homocysteine Studies Collaboration. Homocysteine and risk of ischemic heart disease and stroke. *JAMA*, 288:2015–2022, 2002.
- [12] The Fibrinogen Studies Collaboration. Regression dilution methods for meta-analysis: assessing long-term variability in plasma fibrinogen among 27 247 adults in 15 prospective studies. *International Journal of Epidemiology*, 35:1570–1578, 2006.
- [13] B. Rosner, WC. Willett, and D. Spiegelman. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine*, 8:1051–1069, 1989.
- [14] T. R. Dawber. *The Framingham Study. The epidemiology of atherosclerotic disease*. Harvard University Press, 1980.
- [15] R. Clarke, M. Shipley, S. Lewington, L. Youngman, R. Collins, M. Marmot, and R. Peto. Underestimation of risk associations due to regression dilution in long-term follow-up of prospective studies. *American Journal of Epidemiology*, 150:341–353, 1999.
- [16] C. Frost and I. R. White. The effect of measurement error in risk factors that change over time in cohort studies: do simple methods overcorrect for ‘regression dilution’? *International Journal of Epidemiology*, 34:1359–1368, 2005.
- [17] Y. Ben-Shlomo and D. Kuh. A life course approach to chronic disease epidemiology: conceptual models, empirical challenges and interdisciplinary perspectives. *International Journal of Epidemiology*, 31:285–293, 2002.
- [18] B. L. De Stavola, D. Nitsch, I. dos Santos Silva, V. McCormack, R. Hardy, V. Mann, T. J. Cole, S. Morton, and D. A. Leon. Statistical issues in life course epidemiology. *American Journal of Epidemiology*, 163:84–96, 2006.
- [19] D. R. Cox. *Principles of Statistical Inference*. Cambridge University Press, 2006.
- [20] Y. Pawitan. *In All Likelihood*. Oxford University Press, Oxford, 2001.
- [21] L. S. Freedman, D. Midthune, R. J. Carroll, and V. Kipnis. A comparison of regression calibration, moment reconstruction and imputation for adjusting for covariate measurement error in regression. *Statistics in Medicine*, 2008.

- [22] J. Berkson. Are there two regressions? *Journal of the American Statistical Association*, 45:164–180, 1950.
- [23] I. White, C. Frost, and S. Tokunaga. Correcting for measurement error in binary and continuous variables using replicates. *Statistics in Medicine*, 20:3441–3457, 2001.
- [24] S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. John Wiley & Sons Inc, 1992.
- [25] P. H. Westfall. A comparison of variance component estimates for arbitrary underlying distributions. *Journal of the American Statistical Association*, 82:866–874, 1987.
- [26] R. J. Adcock. Note on the method of least squares. *The Analyst*, 4:183–184, 1877.
- [27] J. W. Bartlett, B. L. de Stavola, and C. Frost. Linear mixed models for replication data to efficiently allow for covariate measurement error. *Statistics in Medicine*, 28:3158–3178, 2009.
- [28] A. W. van der Vaart. *Asymptotic Statistics*. Cambridge University Press, 2000.
- [29] P. Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. Chapman & Hall / CRC, 2003.
- [30] C. Frost and S. G. Thompson. Correcting for regression dilution bias: comparison of methods for a single predictor variable. *Journal of the Royal Statistical Society A*, 163:173–189, 2000.
- [31] B. Efron and R. J. Tibshirani. *An introduction to the bootstrap*. Chapman & Hall / CRC, 1993.
- [32] D. W. Schafer and K. G. Purdy. Likelihood analysis for errors-in-variables regression with replicate measurements. *Biometrika*, 83:813–824, 1996.
- [33] B. Rosner, D. Spiegelman, and W. C. Willett. Correction of logistic regression relative risk estimates and confidence intervals for random within-person measurement error. *American Journal of Epidemiology*, 136:1400–1413, 1992.
- [34] B. Armstrong. Measurement error in the generalised linear model. *Communications in Statistics - Simulation and Computation*, 14:529–544, 1985.
- [35] L. J. Gleser. Improvements of the naive approach to estimation in nonlinear errors-in-variables regression models. In P. J. Brown and W. A. Fuller, editors,

- Statistical analysis of measurement error models and applications.* American Mathematics Society, Providence, 1990.
- [36] R. J. Carroll and L. A. Stefanski. Approximate quasi-likelihood estimation in models with surrogate predictors. *Journal of the American Statistical Association*, 85:652–663, 1990.
- [37] G. Verbeke and G. Molenberghs. *Linear Mixed Models for Longitudinal Data*. Springer-Verlag, 2000.
- [38] D. W. Schafer. Covariate measurement error in generalized linear models. *Biometrika*, 74:385–391, 1987.
- [39] D. W. Schafer. Likelihood analysis for probit regression with measurement errors. *Biometrika*, 80:899–904, 1993.
- [40] D. Clayton. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In J. H. Dwyer, M. Feinleib, P. Lipsert, and H. Hoffmeister, editors, *Statistical models for longitudinal studies of health*, pages 301–331. Oxford University Press, New York, 1991.
- [41] A. Skrondal and S. Rabe-Hesketh. *Generalized Latent Variable Modeling*. Chapman & Hall / CRC, 2004.
- [42] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B*, 39:1–38, 1977.
- [43] J. L. Schafer. *Analysis of Incomplete Multivariate Data*. Chapman & Hall / CRC, 1997.
- [44] S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Maximum likelihood estimation of generalized linear models with covariate measurement error. *The Stata Journal*, 3:386–411, 2003.
- [45] L. K. Muthén and B. O. Muthén. *Mplus user’s guide*. Muthén & Muthén, 1998.
- [46] K. Messer and L. Natarajan. Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment. *Statistics in Medicine*, 27:6332–6350, 2008.
- [47] J. K. Lindsey. *Parametric Statistical Inference*. Oxford University Press, 1996.
- [48] D. F. Heitjan and D. B. Rubin. Ignorability and coarse data. *The Annals of Statistics*, 19:2244–2253, 1991.

- [49] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, 1987.
- [50] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer-Verlag, 2nd edition, 2004.
- [51] M. G. Kenward and J. Carpenter. Multiple imputation: current perspectives. *Statistical Methods in Medical Research*, 16:199–218, 2007.
- [52] A. A. Tsiatis. *Semiparametric Theory and Missing Data*. Springer-Verlag, 2006.
- [53] N. Wang and J. M. Robins. Large-sample theory for parametric multiple imputation procedures. *Biometrika*, 85:935–948, 1998.
- [54] S. R. Cole, H. Chu, and S. Greenland. Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, 35:1074–1081, 2006.
- [55] J. M. Robins and N. Wang. Inference for imputation estimators. *Biometrika*, 87:113–124, 2000.
- [56] R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons Inc, 2nd edition, 2002.
- [57] L. S. Freedman, V. Fainberg, V. Kipnis, D. Midthune, and R. J. Carroll. A new method for dealing with measurement error in explanatory variables of regression models. *Biometrics*, 60:172–181, 2004.
- [58] X. Huang, L. A. Stefanski, and M. Davidian. Latent-model robustness in structural measurement error models. *Biometrika*, 93:53–64, 2006.
- [59] C. Y. Wang, N. Wang, and S. Wang. Regression analysis when covariates are regression parameters of a random effects model for observed longitudinal measurements. *Biometrics*, 56:487–495, 2000.
- [60] J. W. Hardin, H. Schmiediche, and R. J. Carroll. The regression calibration method for fitting generalized linear models with additive measurement error. *Stata Journal*, 3:361–372, 2003.
- [61] R. J. Carroll, S. Wang, and C. Y. Wang. Prospective analysis of logistic case-control studies. *Journal of the American Statistical Association*, 90:157–169, 1995.
- [62] J. E. Michalek and R. C. Tripathi. The effect of errors in diagnosis and measurement on the estimation of the probability of an event. *Journal of the American Statistical Association*, 75:713–721, 1980.

- [63] B. G. Armstrong, A. S. Whittemore, and G. R. Howe. Analysis of case-control data with covariate measurement error: application to diet and colon cancer. *Statistics in Medicine*, 8:1151–1163, 1989.
- [64] J. Kuha. Corrections for exposure measurement error in logistic regression models with an application to nutritional data. *Statistics in Medicine*, 13:1135–1148, 1994.
- [65] G. Molenberghs and G. Verbeke. *Models for Discrete Longitudinal Data*. Springer-Verlag, 2005.
- [66] S. Rabe-Hesketh, A. Skrondal, and A. Pickles. Reliable estimation of generalized linear mixed models using adaptive quadrature. *The Stata Journal*, 2:1–21, 2002.
- [67] R. Higdon and D. W. Schafer. Maximum likelihood computations for regression with measurement error. *Computational Statistics & Data Analysis*, 35:283–299, 2001.
- [68] G. C. G. Wei and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man’s Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85:699–704, 1990.
- [69] M. Thoresen and P. Laake. A simulation study of measurement error correction methods in logistic regression. *Biometrics*, 56:868–872, 2000.
- [70] D. W. Schafer. Semiparametric maximum likelihood for measurement error model regression. *Biometrics*, 57:53–61, 2001.
- [71] L. A. Stefanski and R. J. Carroll. Conditional scores and optimal scores for generalized linear measurement- error models. *Biometrika*, 74:703–716, 1987.
- [72] R. J. Carroll, K. Roeder, and L. Wasserman. Flexible parametric measurement error models. *Biometrics*, 55:44–54, 1999.
- [73] S. Rabe-Hesketh, A. Pickles, and A. Skrondal. Correcting for covariate measurement error in logistic regression using nonparametric maximum likelihood estimation. *Statistical Modelling*, 3:215–232, 2003.
- [74] B. S. Caffo, W. Jank, and G. L. Jones. Ascent-based Monte Carlo expectation-maximization. *Journal of the Royal Statistical Society B*, 67:235–251, 2005.
- [75] S. F. Nielsen. The stochastic EM algorithm: estimation and asymptotic results. *Bernoulli*, 6:457–489, 2000.

- [76] J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society B*, 61:265–285, 1999.
- [77] D. Oakes. Direct calculation of the information matrix via the EM algorithm. *Journal of the Royal Statistical Society B*, 61:479–482, 1999.
- [78] D. J. C. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.
- [79] J. Buonaccorsi. Fiellers theorem. In *Encyclopedia of Biostatistics*. John Wiley & Sons, 1998.
- [80] B. Efron. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association*, 70:892–898, 1975.
- [81] S. J. Press and S. Wilson. Choosing between logistic regression and discriminant analysis. *Journal of the American Statistical Association*, 73:699–705, 1978.
- [82] J. Kuha. Estimation by data augmentation in regression models with continuous and discrete covariates measured with error. *Statistics in Medicine*, 16:189–201, 1997.
- [83] O. O. Aalen, Ø. Borgan, and H. K. Gjessing. *Survival and Event History Analysis*. Springer-Verlag, 2008.
- [84] D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*, 34:187–220, 1972.
- [85] D. R. Cox. Partial likelihood. *Biometrika*, 62:269–276, 1975.
- [86] R. L. Prentice. Covariate measurement errors and parameter estimation in a failure time regression model. *Biometrika*, 69:331–342, 1982.
- [87] M. D. Hughes. Regression dilution in the proportional hazards model. *Biometrics*, 49:1056–1066, 1993.
- [88] F. H. Kong. Adjusting regression attenuation in the Cox proportional hazards model. *Journal of Statistical Planning and Inference*, 79:31–44, 1999.
- [89] C. Y. Wang, L. Hsu, Z. D. Feng, and R. L. Prentice. Regression calibration in failure time regression. *Biometrics*, 53:131–145, 1997.
- [90] C. Y. Wang. Robust sandwich covariance estimation for regression calibration estimator in cox regression with measurement error. *Statistics & Probability Letters*, 45:371–378, 1999.

- [91] S. X. Xie, C. Y. Wang, and R. L. Prentice. A risk set calibration method for failure time regression by using a covariate reliability sample. *Journal of the Royal Statistical Society B*, 63:855–870, 2001.
- [92] S. Johansen. An extension of Cox’s regression model. *International Statistical Review*, 51:165–174, 1983.
- [93] N. E. Breslow. Contribution to the discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B*, 34:187, 1972.
- [94] M. S. Wulfsohn and A. A. Tsiatis. A joint model for survival and longitudinal data measured with error. *Biometrics*, 53:330–339, 1997.
- [95] P. Hu, A. A. Tsiatis, and M. Davidian. Estimating the parameters in the Cox model when covariate variables are measured with error. *Biometrics*, 54:1407–1419, 1998.
- [96] F. Hsieh, T. Yi-Kuan, and J-L. Wang. Joint modeling of survival and longitudinal data: Likelihood approach revisited. *Biometrics*, 62:1037–1043, 2006.
- [97] D. Zeng and J. Cai. Asymptotic results for maximum likelihood estimators in joint analysis of repeated measurements and survival time. *Annals of Statistics*, 33:2132–2163, 2005.
- [98] X. Guo and B. P. Carlin. Separate and joint modeling of longitudinal and event time data using standard computer packages. *The American Statistician*, 58:16–24, 2004.
- [99] R. Henderson, P. Diggle, and A. Dobson. Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1:465–480, 2000.
- [100] W. R. Gilks and P. Wild. Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 41:337–348, 1992.
- [101] A. H. Herring and J. G. Ibrahim. Likelihood-based methods for missing covariates in the cox proportional hazards model. *Journal of the American Statistical Association*, 96:292–302, 2001.
- [102] S. van Buuren, H.C. Boshuizen, and D. L. Knook. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18:681–694, 1999.
- [103] I. R. White and P. Royston. Imputing missing covariate values for the cox model. *Statistics in Medicine*, 28:1982–1998, 2009.

- [104] Y. Li and L. Ryan. Inference on survival data with covariate measurement error - an imputation-based approach. *Scandinavian Journal of Statistics*, 33:169–190, 2006.
- [105] A. A. Tsiatis and M. Davidian. A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error. *Biometrika*, 88:447–458, 2001.
- [106] X. Song and Y. Huang. On corrected score approach for proportional hazards model with covariate measurement error. *Biometrics*, 61:702–714, 2005.
- [107] T. Nakamura. Corrected score function for errors-in-variables models: methodology and application to generalized linear models. *Biometrika*, 77:127–137, 1990.
- [108] T. Nakamura. Proportional hazards model with covariates subject to measurement error. *Biometrics*, 48:829–838, 1992.
- [109] F. H. Kong and M. Gu. Consistent estimation in cox proportional hazards model with covariate measurement errors. *Statistica Sinica*, 9:953–969, 1999.
- [110] M. S. Green and M. J. Symons. A comparison of the logistic risk function and the proportional hazards model in prospective epidemiologic studies. *Journal of Chronic Diseases*, 36:715–724, 1983.
- [111] R. J. Carroll, C. H. Spiegelman, K. K. G. Lan, K. T. Bailey, and R. D. Abbott. On errors-in-variables for binary regression models. *Biometrika*, 71:19–25, 1984.
- [112] O. O. Aalen. A linear regression model for the analysis of life times. *Statistics in Medicine*, 8:907–925, 1989.
- [113] P. Gimenez, H. Bolfarine, and E. A. Colosimo. Estimation in Weibull regression model with measurement error. *Communications in Statistics - Theory and Methods*, 28:495–510, 1999.
- [114] W. He, G. Y. Yi, and J. Xiong. Accelerated failure time models with covariates subject to measurement error. *Statistics in Medicine*, 26:4817–4832, 2007.
- [115] B. Muthén. Latent variable modelling of longitudinal and multilevel data. *Sociological Methodology*, 27:453–480, 1997.
- [116] D. Ruppert, M. P. Wand, and R. J. Carroll. *Semiparametric regression*. Cambridge University Press, 2003.
- [117] E. Li, N. Wang, and N. Wang. Joint models for a primary endpoint and multiple longitudinal covariate processes. *Biometrics*, 63:1068–1078, 2007.

- [118] E. Li, D. Zhang, and M. Davidian. Conditional estimation for generalized linear models when covariates are subject-specific parameters in a mixed model for longitudinal measurements. *Biometrics*, 60:1–7, 2004.
- [119] H. C. Boshuizen, M. Lanti, A. Menotti, J. Moschandreas, H. Tolonen, A. Nissinen, S. Nedeljkovic, A. Kafatos, and D. Kromhout. Effects of past and recent blood pressure and cholesterol level on coronary heart disease and stroke mortality, accounting for measurement error. *American Journal of Epidemiology*, 165:398–409, 2007.
- [120] J. E. Gentle. *Matrix Algebra*. Springer-Verlag, 2007.
- [121] E. Li, D. Zhang, and M. Davidian. Likelihood and pseudo-likelihood methods for semiparametric joint models for a primary endpoint and longitudinal data. *Computational Statistics & Data Analysis*, 51:5776–5790, 2007.
- [122] X. Huang, L. A. Stefanski, and M. Davidian. Latent-model robustness in joint models for a primary endpoint and a longitudinal process. *Biometrics*, 65:719–727, 2009.
- [123] G. Verbeke and E. Lesaffre. The effect of misspecifying the random-effects distribution in linear mixed models for longitudinal data. *Computational Statistics & Data Analysis*, 23:541–556, 1997.
- [124] A. A. Tsiatis and M. Davidian. Joint modelling of longitudinal and time-to-event data: an overview. *Statistica Sinica*, 14:809–834, 2004.
- [125] P. Diggle, R. Henderson, and P. Philipson. Random-effects models for joint analysis of repeated-measurement and time-to-event outcomes. In *Longitudinal Data Analysis*. Chapman & Hall / CRC, 2008.
- [126] V. DeGruttola and X. M. Tu. Modelling progression of CD4-lymphocyte count and its relationship to survival time. *Biometrics*, 50:1003–1014, 1994.
- [127] A. Tsiatis, V. DeGruttola, and M. S. Wulfsohn. Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS. *Journal of the American Statistical Association*, 90:27–37, 1995.
- [128] P. S. Albert and D. A. Follmann. Shared-parameter models. In *Longitudinal Data Analysis*. Chapman & Hall / CRC, 2008.
- [129] U. G. Dafni and A. A. Tsiatis. Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, 54:1445–1462, 1998.

- [130] W. Ye, X. Lin, and J. M. G. Taylor. Semiparametric modeling of longitudinal measurements and time-to-event data - a two-stage regression calibration approach. *Biometrics*, 64:1238–1246, 2008.
- [131] J. Bartlett, B. De Stavola, I. White, and C. Frost. RE: "Effects of past and recent blood pressure and cholesterol level on coronary heart disease and stroke mortality, accounting for measurement error". *American Journal of Epidemiology*, 167:502–503, 2008.
- [132] H. C. Boshuizen, M. Lanti, A. Menotti, J. Moschandreas, H. Tolonen, A. Nissinen, S. Nedeljkovic, A. Kafatos, and D. Kromhout. The authors reply. *American Journal of Epidemiology*, 167:503–504, 2008.
- [133] J. D. Kalbfleisch and R. L. Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, 2002.
- [134] X. Song, M. Davidian, and A. A. Tsiatis. A semiparametric likelihood approach to joint modeling of longitudinal and time-to-event data. *Biometrics*, 58:742–753, 2002.
- [135] D. Rizopoulos, G. Verbeke, and G. Molenberghs. Shared parameter models under random effects misspecification. *Biometrika*, 95:63–74, 2008.
- [136] D. Rizopoulos, G. Verbeke, and E. Lesaffre. Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society B*, 71:637–654, 2009.
- [137] X. Song, M. Davidian, and A. A. Tsiatis. An estimator for the proportional hazards model with multiple longitudinal covariates measured with error. *Biostatistics*, 3:511–528, 2002.
- [138] B. F. Kurland, L. L. Johnson, B. L. Egleston, and P. H. Diehr. Longitudinal data with follow-up truncated by death: match the analysis method to research aims. *Statistical Science*, 24:211–222, 2009.
- [139] British Cardiac Society, British Hypertension Society, Diabetes UK, HEART UK, Primary Care Cardiovascular Society, and Stroke Association. JBS 2: Joint British Societies' guidelines on prevention of cardiovascular disease in clinical practice. *Heart*, 91:v1–v52, 2005.
- [140] Prospective Studies Collaboration. Age-specific relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet*, 360:1903–1913, 2002.

- [141] J. M. Robins and M. A. Hernán. Estimation of the causal effects of time-varying exposures. In *Longitudinal Data Analysis*. Chapman & Hall / CRC, 2008.