Missing covariates in competing risks analysis

Jonathan Bartlett

London School of Hygiene and Tropical Medicine
www.missingdata.org.uk
www.thestatsgeek.com

Centre for Biostatistics
University of Manchester
7th October 2015

Acknowledgements

I am grateful for support from the UK Medical Research Council (MR/K02180X/1).

This is joint work with Jeremy Taylor, University of Michigan.

Outline

Missing covariates in competing risks analysis

Validity of complete case analysis

Multiple imputation

Simulations

NHANES III analysis

Conclusions

Outline

Missing covariates in competing risks analysis

Validity of complete case analysis

Multiple imputation

Simulations

NHANES III analysis

Conclusions

Competing risks analysis

- A set of independent individuals is followed up over time.
- For each, we follow them until the first of a set of events occurs.
- Examples include time to death, with cause of death defining the type of failure, or time to cancer recurrence, with death as a competing risk.
- We record the time of first event Y and the type of event $D \in \{0, 1, .., K\}$, where $D = 0$ corresponds to censoring.

Modelling cause specific hazards

- Typically we have baseline covariates, and want to model how the hazards for the competing risks depend on these covariates.
- Model each competing hazard, treating failures from other failure types as censoring events.
- A popular approach is to fit a Cox proportional hazard model for each cause specific hazard function. i.e. for cause k

  $$h_k(t|X,Z) = h_{0k}(t)\exp(g_k(X,Z,\beta_k))$$

  where $g_k(X,Z,\beta_k)$ gives the linear predictor and $h_{0k}(t)$ is an arbitrary baseline hazard function.
- The parameters β_k are log hazard ratios of interest.

Ignoring competing risks

- When covariates are fully observed, to fit the model for cause 1 (say), we can fit a Cox model where we treat failures from other causes as censorings.
- This means that if we are only interested in modelling failure from one cause, there is no need to model the hazards for the other causes.

Illustrative example

- The third National Health and Nutrition Examination Survey (NHANES III) was conducted in the US between 1988 and 1994.
- Survey of health and nutrition status of adults and children, obtained from physical exam and interview.
- The overall study involved around 40,000 individuals.
- Mortality at end of 2011 has been ascertained by linkage to the US National Death Index.

Illustrative example

- Here I focus on a subset of individuals aged between 60 and 70 at the time of the original survey.
- I ignore the complex survey design here - all results are intended to be purely illustrative.
- Data are available on 2,583 individuals.
- I have categorised death into cardiovascular disease (CVD), cancer, and other causes:

Cause of death	Number (%)
CVD	358 (13.9%)
Cancer	379 (14.7%)
Other	755 (29.2%)

Missingness in covariates

- Aim: model hazard for death due to CVD, with baseline risk factors.
- Inevitably, for a variety of reasons, there is non-trivial missingness in many:

Variable	Mean (SD) / no. (%)	No. missing (%)
Sex, female	1,302 (50.4)	0
Age (years)	64.4 (2.9)	0
Current smoker	597 (38.9)	1,048 (40.6)
Diabetes	427 (16.6)	3 (0.1)
Alcohol consumer	992 (55.0)	778 (30.1)
SBP (mm Hg)	137.8 (19.4)	297 (11.5)
Total chol. (mg/dl)	225.6 (45.2)	355 (13.7)
CRP > 0.21 mg/dl	946 (42.7)	368 (14.2)
Fibrinogen (mg/dl)	330.8 (96.0)	387 (15.0)

Missingness in covariates

- We can perform complete case analysis, dropping those with missing covariate values.
- Here a complete case analysis uses data from only 1,106 individuals, 42.8% of the total sample.
- It is clearly inefficient.
- It could be biased too, if data are not missing completely at random.
- An alternative we will consider later is to use multiple imputation.

Outline

Setup

- We assume there exists a failure time T and failure type indicator $D^* \in \{1, .., K\}$.
- Typically some individuals are censored.
- We let C denote the potential censoring time for each individual.
- We then observe $Y = \min(T, C)$ and $D = 1(T < C) \times D^*$, i.e. we only observe time to first of censoring or failure.
- So $D \in \{0, 1, .., K\}$, with $D = 0$ indicating censoring.

Validity of complete case analysis

- We assume there are some covariates X which are partially observed, while the covariate(s) Z are fully observed.
- Let $R = 1$ denote that all covariates are observed, $R = 0$ that some are missing.
- We want to fit a Cox model for hazard of failure due to cause k, i.e.:

$$h_k(t|X, Z) = h_{0k}(t) \exp(g_k(X, Z, \beta_k))$$

- If values are missing completely at random (MCAR), i.e. $R \perp\!\!\!\perp (T, D^*, C, X, Z)$, then complete case analysis (CCA) is valid.

Validity of complete case analysis

- CCA is also valid under weaker conditions.
- Provided $R \perp\!\!\!\perp (T, D^*) | (C, X, Z)$, CCA is valid.
- This means that missingness in X can depend on time to censoring C, fully observed covariates Z, and even X itself.
- Thus, CCA can be valid even under certain missing not at random mechanisms [1].

Plausibility of covariate dependent missingness

- An assumption that missingness in baseline covariates is unrelated to future time of failure T, conditional on covariates X and Z, may sometimes be plausible.
- Indeed, missingness can only be independently associated with the future time of failure T if there exists other variables V which affect hazard of failure and also missingness in X.

Assessing missingness assumptions

- Unfortunately the CCA validity assumption $R \perp\!\!\!\perp (T, D^*) | (C, X, Z)$ cannot be verified from the observed data.
- We can however check whether the data are consistent with a stronger assumption, that $R \perp\!\!\!\perp (T, D^*, X) | (C, Z)$ and that $X \perp\!\!\!\perp C | Z$.
- To check, first fit a Cox model where censoring corresponds to failure, with X and Z as covariates, in those with $R = 1$, and check that X is not an important predictor.
- Second, fit a Cox model for failure of any type, with R and Z as covariates, in all individuals, and check R is not an important predictor.

Assessing missingness assumptions - NHANES data

- In the NHANES data, we fitted a Cox model for death from any cause, with R and the fully observed variables sex, age, diabetes (dropping the three observations with diabetes missing) as covariates.
- Unfortunately this showed that R (i.e. missingness) was an independent predictor of hazard of death.
- The data are thus not consistent with the stronger assumption that $R \perp\!\!\!\perp (T, D^*, X) | (C, Z)$.
- Note however, that this does not necessarily mean the CCA is invalid.
- Our findings may have arisen because, for example, missingness in some covariates depends on their own values (i.e. MNAR).

Outline

Imputation of a single covariate

- We now consider multiple imputation of missing covariate values.
- We first assume there are missing values in only one covariate X.
- We assume the missing values in X are missing at random.
- Here this means $R \perp\!\!\!\perp X | (Y, D, Z)$, where R denotes whether X is recorded $(R = 1)$ or not $(R = 0)$.
- We assume we have specified a Cox model for each competing risk, as described earlier.

Multiple imputation of X

- To impute the missing values in X, we must specify a model for $f(X|Y, D, Z)$.
- The question is, how should we specify this model, in light of how we will be analysing the data?
- If X were continuous, we might try a linear regression imputation model, with Y, D (as a factor variable) and Z as covariates.
- The problem with such a model is that it is *incompatible* with our outcome or substantive model for $f(Y, D|X, Z)$ (the Cox models).

Compatibility between imputation and substantive models

- An imputation model $f(X|Y, D, Z)$ is said to be compatible with the substantive model $f(Y, D|X, Z)$ if (loosely speaking) there exists a joint model $f(Y, D, X|Z)$ which has these models as its conditionals.
- Assuming we believe in our substantive model being (at least approximately) correctly specified, unless our imputation model for X, or a model nested within it, is compatible with the substantive model, our imputation model is misspecified [2].
- Essentially, incompatibility means the two models (imputation and substantive) conflict – they can't both be right!
- Our previously posited imputation model for X, it turns out, is not compatible with the Cox models for the competing risks.
- Using it would therefore expect to result in biased estimates and invalid inferences.

Imputation of covariates in survival analysis

- In the simpler survival analysis setting, White and Royston showed that an approximately compatible imputation model for X, when the Cox outcome model contains main effects of X and Z, is one which includes D (the event indicator) and $H_0(t) = \int_0^t h_0(u)du$ as covariates [3].
- Recently, Resche-Rigon *et al* have extended these results to the competing risks setting, showing that one should include D (as a factor variable) and $H_{0k}(Y)$ $(k = 1, .., K)$ as covariates [4].
- The unknown baseline hazard function can be approximated by the marginal Nelson-Aalen estimates of the cause specific hazard functions.
- A drawback of their results is that they are only approximate, and do not obviously generalize when the Cox models contain interactions or non-linear covariate effects.

Imputing compatibly

- To derive an imputation model for X which is compatible with the outcome model, we can express the conditional distribution $f(X|Y, D, Z)$ as:

$$\begin{aligned} f(X|Y,D,Z) &= \frac{f(X,Y,D|Z)}{f(Y,D|Z)} \\ &\propto f(Y,D|X,Z)f(X|Z) \end{aligned}$$

- The first component, $f(Y, D|X, Z)$, is determined by the assumed models for the cause specific hazard functions.
- The imputation distribution specification is thus completed by specifying a model $f(X|Z, \phi)$.
- This can be chosen according to the type of variable, e.g. linear regression for continuous X, logistic regression for binary X, etc.

Imputing compatibly with the substantive model

- MI is derived from a Bayesian perspective, with draws taken from the posterior of the missing data given the observed data and priors for model parameters.
- Typically the priors are chosen as 'standard' noninformative ones.
- Here we can assume independent standard priors for the parameters in the Cox models and for parameter ϕ in the model $f(X|Z,\phi)$.
- To sample from the posterior, we use a Gibbs sampling approach, where we iterate between:
  1. imputing X from the previously described distribution, conditional on current values of model parameters
  2. sampling new parameters from their posteriors given priors, observed data, and current imputed values of X
- We run multiple independent chains, taking last set of imputed values in each to create each imputed dataset.

Sampling from the imputation distribution

- In the case of binary/categorical X, it is easy to work out the required probabilities $P(X = x|Y, D, Z)$.
- More generally, the imputation distribution, which is compatible with the substantive (Cox) models, does not belong to a standard parametric family.
- We use rejection sampling to draw from the distribution, with $f(X|Z, \phi)$ as the proposal distribution (details omitted).

Advantages of substantive model compatible imputation

- Imputing the partially observed covariate compatibly with the substantive model is desirable since incompatibility implies the imp. model is misspecified.
- If the Cox models include interactions or non-linear effects involving partially observed covariates, it is very difficult, if not impossible, to specify direct imputation models $f(X|Y, D, Z)$ which are compatible with the substantive Cox models.
- Our approach can automatically handle such situations.

Missingness in multiple covariates

- So far we have assumed we have missing values in only one covariate, X.
- Of course in practice often multiple covariates have missing values, so that X is vector valued.
- In principle we could specify a multivariate model $f(X|Z, \phi)$, and extend the Gibbs sampling approach developed earlier.
- However, specifying such multivariate models directly becomes tricky when some components of X are continuous and some are discrete.

Chained equations / fully conditional specification MI

- More generally, the chained equations / fully conditional specification approach to MI has become popular for imputing when there are variables of different types.
- This involves specifying a separate conditional imputation model for each partially observed variable.
- i.e. for each partially observed variable X_j, $j = 1, .., p$, we specify a model for $f(X_j|Y, D, X_{-j}, Z)$, where $X_{-j} = (X_1, .., X_{j-1}, X_{j+1}, .., X_p)$.
- The problem, as in the case of one missing variable, is how to ensure each of these models is compatible with the substantive model.

Substantive model compatible fully conditional specification imputation

- Recently we proposed a modification of this, called substantive model compatible fully conditional specification imputation (SMC-FCS), which combines the flexibility of FCS MI with the concept of ensuring compatibility between imputation and substantive models [2].
- We specify a separate model $f(X_j|X_{-j}, Z, \phi)$ for $j = 1, .., p$ where there are p partially observed covariates.
- This approach readily incorporates our earlier results for the case of competing risks outcomes.
- There is however a potential concern, since the models $f(X_j|X_{-j}, Z, \phi)$ may be mutually incompatible. Whether or not such incompatibility causes a problem in practice requires further research.

Substantive model compatible fully conditional specification imputation

- The SMC-FCS approach is implemented in both Stata (from SSC) [5] and R (from CRAN).
- See www.missingdata.org.uk for instructions on installing the latest development version.
- As well as competing risks outcomes, linear regression, logistic regression, and Cox models for time to event data are supported.
- Covariates can be imputed using normal, logistic, ordinal logistic, multinomial logistic, Poisson, and negative binomial models.

Outline

Simulation 1 - setup

- Samples of size $n = 1000$.
- $X_1 \sim \text{Bernoulli}(0.5)$.
- $X_2|X_1 \sim \text{Bernoulli}(0.25 + 0.5X_1)$.
- $X_3|X_1, X_2 \sim N(-1 + X_1 + X_2, 1)$
- Probability of X_3 being missing $0.25 + 0.5X_1$ (so 50% missing)

Simulation 1 - setup

- Two competing events. First with hazard

$$h_1(t|X_1, X_2, X_3) = 0.002 \exp(\beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3)$$

and second with

$$h_1(t|X_1, X_2, X_3) = 0.002 \exp(\beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3)$$

- Random censoring, with hazard 0.002.

Methods

- Full data (results not shown here)
- Complete case analysis (results not shown here)
- Direct imputation, assuming $f(X_3|T, D, X_1, X_2)$ is normal, with covariates X_1, X_2, D (factor variable) and Nelson-Aalen estimates of $H_{01}(T)$ and $H_{02}(T)$.
- Substantive model compatible MI, assuming the Cox models for cause specific hazards, and that $f(X_3|X_1, X_2)$ is normal linear regression.

5 imputations for both imputation methods

Results based on 1,000 simulations

	Direct MI			SMC MI		
	Mean	SD	CI	Mean	SD	CI
$\beta_{11} = 1$	0.92	0.12	0.93	1.04	0.14	0.94
$\beta_{12} = 1$	1.03	0.12	0.96	1.01	0.14	0.95
$\beta_{13} = 1$	0.66	0.06	0.06	0.99	0.09	0.94
$\beta_{21} = 0.5$	0.44	0.21	0.94	0.52	0.21	0.94
$\beta_{22} = -1$	-1.03	0.25	0.95	-1.00	0.25	0.94
$\beta_{23} = 0.75$	0.62	0.11	0.83	0.76	0.13	0.95

Simulation conclusions

- The directly specified imputation approach gives slightly biased estimates for fully observed covariate effects, but badly biased for effect of partially observed covariate.
- The imp. model it uses is only approximately compatible with the Cox substantive models.
- Particularly when covariate effects are large, the approximation breaks down, leading to bias.
- In contrast, the substantive model compatible MI gives unbiased estimates, and CIs have correct coverage.

Simulation 2 - setup

Same as before, except

- binary covariate X_2 also made missing (MCAR 25%).
- hazard functions include interaction between X_2 and X_3:

$$h_k(t|X_1, X_2, X_3) = 0.002 \exp(\beta_{k1}X_1 + \beta_{k2}X_2 + \beta_{k3}X_3 + \beta_{k4}X_2X_3)$$

for $k = 1, 2$

Methods

- Chained equations / FCS MI, using logistic imp. model for X_2 and normal model for X_3, adjusting for event indicator and Nelson-Aalen cumulative hazards as before.
- Substantive model compatible FCS, using logistic imp. model for X_2 and normal model for X_3, accounting for interaction in cause specific hazard functions.

Results based on 1,000 simulations

	FCS MI			SMC-FCS MI		
	Mean	SD	CI	Mean	SD	CI
$\beta_{11} = 1$	0.94	0.13	0.94	1.03	0.14	0.94
$\beta_{12} = 1$	1.08	0.15	0.93	0.99	0.15	0.96
$\beta_{13} = 1$	0.64	0.10	0.21	1.02	0.14	0.95
$\beta_{14} = -1$	-0.56	0.08	0.08	-1.03	0.17	0.94
$\beta_{21} = 0.5$	0.51	0.18	0.96	0.55	0.20	0.94
$\beta_{22} = -1$	-0.07	0.20	0.05	-0.93	0.31	0.94
$\beta_{23} = 0.75$	0.72	0.10	0.97	0.74	0.13	0.96
$\beta_{24} = 1$	0.14	0.09	0.00	0.96	0.21	0.96

Simulation conclusions

- Standard FCS fails to allow for interactions in the Cox models, leading to substantial bias for some parameters.
- SMC-FCS is essentially unbiased, with confidence interval coverage attaining nominal 95% level.

Outline

NHANES III - illustrative analysis

- Returning to the NHANES III data, we would like to fit a Cox model for hazard of death due to CVD, with the risk factors listed earlier as covariates.
- We use the study time scale, with adjustment for age at baseline.
- We will analyse using the following approaches:
  - Complete case analysis (CCA)
  - Imputing using FCS (chained equations), with failure indicator and Nelson-Aalen estimates of the three cumulative hazards as predictors
  - SMC-FCS

NHANES III - selected results

Estimate (SE) of log hazard ratios

	Complete case	FCS	SMC-FCS
Male	0.51 (0.18)	0.69 (0.12)	0.69 (0.12)
Age	0.086 (0.027)	0.09 (0.019)	0.092 (0.019)
Current smoker	0.59 (0.15)	0.63 (0.13)	0.63 (0.13)
Diabetic	0.26 (0.2)	0.74 (0.13)	0.75 (0.13)
Alcohol consumer	0.38 (0.16)	0.37 (0.14)	0.35 (0.14)
SBP (per 10mmHg)	0.96 (0.38)	1.38 (0.28)	1.36 (0.29)
Cholesterol (mg/ml)	0.34 (0.16)	0.31 (0.12)	0.31 (0.12)
CRP (>0.21mg/dl)	0.45 (0.17)	0.45 (0.12)	0.45 (0.12)
Fibrinogen (mg/dl	0.19 (0.08)	0.13 (0.06)	0.13 (0.06)

NHANES III - illustrative analysis conclusions

- Substantial gains in precision through imputing missing covariates.
- Some material changes between estimates from CCA and MI approaches.
- FCS and SMC-FCS give similar estimates (since no interactions/non-linear covariate effects).
- Unclear which missingness assumption (CCA or MAR) is more reasonable, but arguably missingness in smoking/alcohol could be MNAR.
- In this case, one might argue that the CCA is more plausibly valid.

Outline

Conclusions

- Missing covariates are a common issue in competing risks analysis.
- Complete case analysis is valid provided missingness does not depend on time to failure and failure type.
- To a certain extent this assumption can be investigated using the observed data.

Conclusions

- Multiple imputation, under the MAR assumption, provides an alternative approach.
- We gain efficiency by imputing missing values, compared to CCA.
- In certain cases the MAR assumption is arguably more questionable however.
- The SMC-FCS approach ensures missing covariates are imputed from models which are compatible with the competing risks models we specify.
- Software is available in Stata and R - see www.missingdata.org.uk

References I

[1] J W Bartlett, J R Carpenter, K Tilling, and S Vansteelandt.
Improving upon the efficiency of complete case analysis when covariates are MNAR.
Biostatistics, 15:719–730, 2014.

[2] J W Bartlett, S R Seaman, I R White, and J R Carpenter.
Multiple imputation of covariates by fully conditional specification: Accommodating the substantive model.
Statistical Methods in Medical Research, 24:462–487, 2014.

[3] I. R. White and P. Royston.
Imputing missing covariate values for the Cox model.
Statistics in Medicine, 28:1982–1998, 2009.

References II

[4] M Resche-Rigon, I White, and S Chevret.

Imputing missing covariate values in presence of competing risk.

presentation at the International Society for Clinical Biostatistics Conference, 2012.

[5] J W Bartlett and T P Morris.

Multiple imputation of covariates by substantive model-compatible fully conditional specification.

The Stata Journal, 15(2):437–456, 2015.