Systematically missing data in individual participant data meta-analysis: a semiparametric inverse probability weighting approach

International Biometric Conference 2014

Jonathan Bartlett
www.missingdata.org.uk

London School of Hygiene and Tropical Medicine

Acknowledgements

I am grateful for input to this work from:

- Angela Wood (University of Cambridge, UK)
- Ian White & Shaun Seaman (MRC Biostatistics Unit, UK)
- Stijn Vansteelandt (Ghent University, Belgium)

Support for myself from an MRC fellowship (MR/K02180X/1).

I am also grateful to the MAGGIC Collaborative Group (funding from New Zealand Heart Foundation, University of Auckland and University of Glasgow), whose data I have used in an illustrative analysis.

Outline

The problem

Augmented IPW estimation

Simulations

Illustrative analysis of MAGGIC

Conclusions

Outline

The problem

Augmented IPW estimation

Simulations

Illustrative analysis of MAGGIC

Conclusions

Individual participant data meta-analysis

- Meta-analysis is traditionally performed using aggregate results from study publications.
- Increasingly, meta-analyses are performed using the individual participant data (IPD-MA) from contributing studies.
- IPD-MA are being used to (among other things):
  - estimate exposure effects, adjusted for a set of confounding variables
  - develop prognostic models
- Two stage and one stage analysis approaches are possible.
- Here we adopt a two stage approach.

Missing data in IPD-MA

- One difficult with IPD-MA is that of systematic missingness - some contributing studies do not measure one or more variables of interest.
- Analysing only complete studies is inefficient, and potentially biased.
- Multiple imputation (MI) is an obvious approach to take - we impute missing values (for all participants) in studies which did not record them.
- However, correctly specifying and fitting appropriate multi-level imputation models is difficult.

An augmented inverse probability weighting approach

- We therefore pursue an alternative approach based on augmented inverse probability weighting (AIPW).
- The augmentation function is similar to an imputation model, and enables information to be extracted from studies with systematic missingness.
- Importantly however, consistency will not rely on correct specification of the imputation type model.

Outline

The problem

Augmented IPW estimation

Simulations

Illustrative analysis of MAGGIC

Conclusions

Setup

- We assume the MA consists of n studies.
- In study i, there are N_i participants.
- For participant j in study i, we let Y_{ij} denote the outcome of interest, and X_{ij} and Z_{ij} (vectors of) covariates.
- Let $X_i = (X_{i1}^T, \ldots, X_{iN_i}^T)$ (and Z_i similarly) denote matrices of covariates for study i.
- For the moment suppose that there are no missing data.

Full data analysis

- A model for $Y_{ij}|X_{ij}, Z_{ij}$ is fitted to study i, giving estimates of $\hat{\mu}_i$ and corresponding variance $\hat{\sigma}_i^2$.
- Interest lies in $\mu = E(\mu_i)$ and $\tau^2 = Var(\mu_i)$.
- We adopt a method of moments estimation approach due to Paule and Mandel and recommended by DerSimonian [1].
- μ and τ^2 are estimated as the values solving

$$\sum_{i=1}^{n} m(\hat{\mu}_i, \hat{\sigma}_i^2, \mu, \tau^2) = 0$$

where

$$m(\hat{\mu}_i, \hat{\sigma}_i^2, \mu, \tau^2) = \begin{pmatrix} \frac{\hat{\mu}_i - \mu}{\sigma_i^2 + \tau^2} \\ \frac{(\hat{\mu}_i - \mu)^2}{\sigma_i^2 + \tau^2} - \frac{n-1}{n} \end{pmatrix}$$

Two stage MA with systematically missing covariates

- Now we suppose that X_i is entirely missing for some studies.
- R_i denotes whether study i recorded X_i ($R_i = 1$) or not ($R_i = 0$).
- We assume X_i is missing completely at random (MCAR).
- We can therefore model the distribution of R_i as $Bin(1, \pi)$.
- π can of course be trivially estimated by $\hat{\pi} = n^{-1} \sum_{i=1}^{n} R_i$.

Augmented inverse probability weighted estimators

- Using the same full data estimating function as before, augmented inverse probability weighted estimators [2] can be constructed as solving

$$\sum_{i=1}^{n} \frac{R_i}{\hat{\pi}} m(\hat{\mu}_i, \hat{\sigma}_i^2, \mu, \tau^2) - \left\{ \frac{R_i - \hat{\pi}}{\hat{\pi}} \right\} \phi(Y_i, Z_i, \mu, \tau^2) = 0$$

  where $\phi(Y_i, Z_i, \mu, \tau^2)$ is a function of the always observed variables in study i.
- Assuming MCAR is true, estimates are consistent irrespective of the choice of $\phi(Y_i, Z_i, \mu, \tau^2)$.

Augmented inverse probability weighted estimators

- In a more standard i.i.d. setting, the optimal choice of the augmentation function is given by

$$\phi^{\text{opt}}(Y_i, Z_i, \mu, \tau^2) = E\left[m(\hat{\mu}_i, \hat{\sigma}_i^2, \mu, \tau^2) | Y_i, Z_i\right]$$

- We adopt a pragmatic approach to approximating this:
  - impute X_i (in all studies) L times, based on a simple but easy to fit imputation model (e.g. using fixed study effects).
  - calculate $\hat{\mu}_i^{\text{imp}}$ and $\hat{\sigma}_i^{2^{\text{imp}}}$ based on the imputed X_i.
  - use

$$\hat{\phi}^{\text{opt}}(Y_i, Z_i, \mu, \tau^2) = \frac{1}{L}\sum_{l=1}^{L} m(\hat{\mu}_i^{(l)}, \hat{\sigma}_i^{2^{(l)}}, \mu, \tau^2)$$

- The sandwich variance estimator can be used, although we should be wary about relying on large n asymptotics.

Outline

Simulation study

- Simulations were conduced to assess the AIPW estimator.
- For each of 1,000 simulations, data were generated for $n = 15$ studies.
- Study size N_i was generated as $250 + 500\chi_3^2$ (rounded).
- Covariates X_i and Z_i (both scalar) were generated from a bivariate normal random-effects model.
- X_i was made MCAR with probability 0.5.

Time to event outcome

- A study-specific frailty random variable κ_i was generated from a gamma distribution with shape 2.5 and scale 0.4.
- An event time was generated for each participant, with hazard

$$h(t|X_{ij}, Z_{ij}, \kappa_i) = 0.1\kappa_i \exp(\eta_i X_{ij} + \mu_i Z_{ij})$$

with

$$\begin{pmatrix} \mu_i \\ \eta_i \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0.04 & 0 \\ 0 & 0.04 \end{pmatrix}\right)$$

- Study duration was generated from a $2 + Gamma(1, 1)$ distribution.
- Event times were censored at study duration.

Estimation methods

1. Complete studies analysis.
2. MI, 10 (proper) imputations, using linear regression imputation model with fixed study effects, including Z_{ij}, the event indicator and overall Nelson-Aalen cumulative hazard as covariates [3]. Studies missing X_i are imputed using the estimated constant, corresponding to the (arbitrary) first study which had X_i observed.
3. AIPW, assuming MCAR, and using 10 (improper) imputations to calculate $\hat{\phi}^{\text{opt}}(Y_i, Z_i, \mu, \tau^2)$.

Results for μ

	Mean (emp. SD)	Mean SE	CI coverage (%)
Complete studies	0.998 (0.081)	0.071	85.6
MI	0.939 (0.061)	0.055	75.8
AIPW	1.007 (0.069)	0.059	88.6

- MI is biased (due to imputation model mis-specification).
- AIPW is unbiased, more efficient than complete studies analysis, and has the best CI coverage.

Outline

MAGGIC study

- The Meta-analysis Global Group in Chronic Heart Failure (MAGGIC) is based on data from 39,372 patients from 30 studies with heart failure.
- The outcome is time to all cause mortality.
- We consider an illustrative Cox outcome model, with age, gender and BMI as covariates.
- Age and gender are fully observed.
- A previously developed risk score included BMI as a covariate, capped at 30 kg/m^2, i.e. a non-linear effect [4], and we do the same.

Missingness in BMI

- BMI was not recorded at all in 17 studies.
- In the remaining 13, it was mostly fully recorded, but in a few studies there was non-negligible 'sporadic' missingness.
- To remove the sporadic data problem:
  - we set BMI to missing in studies where it is recorded less than 80% of the time,
  - we delete records with BMI missing in studies where BMI is recorded $> 80\%$, to make it fully recorded.

Analysis approaches

We focus on the log hazard ratio for age (per 10 year increase).

We perform three analyses:

- complete studies analysis, using the 13 studies where BMI was recorded,
- multiple imputation (25 imputations), with a fixed study effect, and including the event indicator and Nelson-Aalen cumulative hazard estimate as covariates,
- augmented IPW (25 imputations)

Estimates of log hazard ratio for age (per 10 year increase)

	Estimate (SE)
Complete studies	0.333 (0.057)
MI	0.373 (0.031)
AIPW	0.367 (0.028)

- Similar estimates from MI and AIPW, and efficiency gain from both.
- Suggestion of more precision from AIPW compared to MI.

Outline

The problem

Augmented IPW estimation

Simulations

Illustrative analysis of MAGGIC

Conclusions

Conclusions

- AIPW estimator improves upon efficiency of complete studies analysis, but is robust to mis-specification of the imp. type model.
- Because it is based on two stage MA, it can be applied irrespective of the type of regression model being used.
- We are currently working on its extension to non-MCAR missingness mechanisms and to the setting wtih multiple variables subject to systematic missingness.
- To handle a combination of systematic and sporadic missingness, it may be possible to impute within study (for those with X measured), followed by application of the AIPW approach.

Meta-analysis Global Group in Chronic Heart Failure

CHARM ECHOS DIAMOND Kirk

Newton Andersson 1 & 2 Madsen

UK Heart Berry Tribouilloy

HFC Edmonton

Hillingdon Varela-Roman Euro HF

DIG Taffet

MUSIC Grigorian Guazzi IN-CHF

Rich 1 & 2

Hola Gotsman

Tsutsui

Executive Committee:

C Berry, R Doughty, C Granger, L Køber, B Massie, F McAlister, J McMurray, S Pocock, K Poppe, K Swedberg, J Somaratne, G Whalley

MAGGIC Steering Group:

A Ahmed, B Andersson, A Bayes-Genis, C Berry, M Cowie, R Cubbon, R Doughty, J Ezekowitz, J Gonzalez-Juanatey, M Gorini, I Gotsman, L Grigorian-Shamagian, M Guazzi, M Kearney, L Køber, M Komajda, A di Lenarda, M Lenzen, D Lucci, S Macín, B Madsen, A Maggioni, M Martínez-Sellés, F McAlister, F Oliva, K Poppe, M Rich, M Richards, M Senni, I Squire, G Taffet, L Tarantini, C Tribouilloy, R Troughton, H Tsutsui, G Whalley

Macin

AHFMS

NPC I

MAGGIC Coordinating Centre:

R Doughty, N Earle, GD Gamble, K Poppe, G Whalley, The University of Auckland, New Zealand

Battlescarred

Richards

MAGGIC Statistical Centres:

R Doughty, K Poppe, G Whalley, The University of Auckland, New Zealand
J Dobson, S Pocock, C Ariti, The London School of Hygiene and Tropical Medicine

MAGGIC

References I

[1] Rebecca DerSimonian and Raghu Kacker.
Random-effects model for meta-analysis of clinical trials: an update.
Contemporary clinical trials, 28(2):105–114, 2007.

[2] A A Tsiatis.
Semiparametric Theory and Missing Data.
Springer, New York, 2006.

[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] S. J. Pocock, C. A. Ariti, J. J. V. McMurray, L. Kber A. Maggioni, I. B. Squire, K. Swedberg, J. Dobson, K. K. Poppe, G. A. Whalley, and R. N. Doughty.

Predicting survival in heart failure: a risk score based on 39372 patients from 30 studies.

European Heart Journal, 34:1404–1413, 2013.