Efficient procedures for handling missing data in generalised linear models

Harvey Goldstein and Christopher Charlton

School of Education, University of Bristol

Background

- Seminal work by Rubin in 1980s resulted in the now standard approach via multiple imputation.
- Further work by various investigators introduced extensions and algorithmic simplifications
- This workshop will be using recent work at Bristol and London School of Hygiene extending this work to handle multilevel data, non-normal variables with missing values and interaction terms in the model of interest.
- For practicalities such as sensitivity analyses, distributional checking etc. see http://missingdata.org.uk/
- First, however, an overview of the problems, especially how ***not*** to handle missing data.

First of all: missingness is not just another category

- Suppose we have a true model. $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e$ (1)
- One recommendation is:

  If missing data in x_2 introduce $x_3 = 1\ if\ missing,\ \ else\ 0$ *and set* x_2 *to 0 if missing*.
- To give the extended model
- $y = \beta_0 + \beta_1^* x_1 + \beta_2 x_2 + \beta_3 x_3 + e$ (1a)
- So that where we have a missing value we get
- $y = (\beta_0 + \beta_3) + \beta_1^* x_1 + e$ (2)
- If x_2 missing completely at random (2) implies that, conditional on x_1 the (unknown) value of x_2 is actually unrelated to y, which is now just a function of x_1 (i.e. (2) is the marginal model for x_1), but this is only compatible for all values of x_2 in the true model (1) if $\beta_2 = 0$, which is of course not generally true, i.e. $\beta_1 \neq \beta_1^*$.
- Thus, where we are actually missing the value of x_2 using (1a) implies we are estimating β_1^* and whereas when not missing we are estimating β_1.
- Thus, fitting the combined model; (1) for non-missing and (1a) for missing, will produce a biased estimate lying somewhere between $\beta_1\ and\ \beta_1^*$.

Let’s illustrate other approaches using a simple regression of *y* on *x*

- $y_i = \alpha + \beta x_i + e_i \qquad y, x, bivariate\ normal$
- For example data might look like:

y	x
31.5	*
22.3	3.2
*	1.9
.	.
.	.

So now let’s generate a dataset – large enough so we can illustrate matters without having to go through tedious simulations.

The dataset

- $\binom{y}{x} \sim \mathrm{N}\left[\binom{0}{0}, \begin{pmatrix} 1 & \\ 0.5 & 1 \end{pmatrix}\right]$ (1)
- i.e. regression is:
- $y = 0.5x \qquad \sigma^2 = 0.75 \qquad (\hat{\beta} = cov(xy)/var(x))$
- Simulate 100,000 pairs of values and estimate regression:
- We get $\hat{\beta} = 0.502\ (0.00274) \quad \hat{\sigma}^2 = 0.751$
- Set about 20% of y's missing ***at random*** 20% x's missing – ***but not at random***. $\Pr(x\ missing) \propto |\mathrm{y}|)$
- We will now apply some popular (intuitive?) procedures to see what happens

What not to do - 1

- Calculate mean of observed and substitute for missing
- We get estimates (standard error in brackets);
- $\hat{\beta} = 0.455(0.00275)\quad \hat{\sigma}^2 = 0.632$
- This is biased.
- Also note standard error - as before but wrong since ~36% values are ‘imputed’ and not actually observed so estimate too small. Correct standard error in this case is elusive.

- *Original* $\hat{\beta} = 0.502\ (0.00274)\quad \hat{\sigma}^2 = 0.751$

What not to do - 2

- Use observed 'complete' records to predict y$|x$ & $x|y$ and use predicted values to plug in for missing. *Prediction imputation*
- We get estimates (standard error in brackets);
- $\hat{\beta} = 0.573\ (0.00264) \quad \hat{\sigma}^2 = 0.591$
- Parameters still biased as is standard error.
- Now use just the complete records
- $\hat{\beta} = 0.569\ (0.00357) \quad \hat{\sigma}^2 = 0.851$
- This is popular because simple but large bias still because complete records not a random sample
- *But,* if you have complete cases as a *random* sample we get;
- $\hat{\beta} = 0.500\ (0.00341) \quad \hat{\sigma}^2 = 0.748$
- So now unbiased but standard error larger than for full data since smaller sample.

- *original* $\hat{\beta} = 0.502\ (0.00274) \quad \hat{\sigma}^2 = 0.751$

How to do it better - 3

- For plug in and prediction imputation $\sigma_y^2, \sigma_x^2, \sigma_{xy}$ biased.
- So for regression impute let's add a random variable on to each imputed (predicted) value, drawn from the regression residual distributions, i.e. $f(y|x)$ & $f(x|y)$
- We now get estimates (standard error in brackets);
- $\hat{\beta} = 0.502\ (0.00279) \quad \hat{\sigma}^2 = 0.754$
- Bias now virtually gone but standard error still too small since takes no account of fact that imputed values are derived from data and not observed and hence similar to full data s.e.
- We call this random regression imputation

original $\hat{\beta} = 0.502\ (0.00274) \quad \hat{\sigma}^2 = 0.751$

How to do it better - 4

- Hot decking – as enjoyed by survey analysts.
- For each missing x (or y) find the set of y's (x's) that are 'similar' to the value of y, say y*, and for x , say x* associated with the missing x (y).
- Issue about how to define similar – in present case we shall take all those y's in the range $(\mathrm{y} * \mp 0.1)$ but we can do sensitivity analyses
- For that set of y's select one at random – or if you want to be sophisticated sample according to the distance from y*.
- This then becomes the imputed value.
- Results, when it works, similar to random regression imputation
- In practice the choice of range is crucial and for several variables we may not be able to find suitable pools of records from which to randomly select

How to do it better - 4 ctd.

- So: when missing not random the only procedure that gives unbiased parameter estimates, but incorrect standard errors is random imputation.
- When missingness is random, complete case analysis and random imputation are unbiased; the former is inefficient, the latter gives incorrect standard error.

How to do it properly

- Known as *multiple imputation* it basically does a random imputation but repeats it independently *n* times, where *n* is a suitably large number – traditionally 5, but more realistically up to 20. An MCMC chain is typically used.
- We therefore obtain *n* estimates of β, σ^2, and these are averaged – using Rubin's rules (below) - to obtain final values together with consistent standard error estimates. For *n=5* in our data we get
- $\hat{\beta} = 0.503\ (0.00290) \quad \hat{\sigma}^2 = 0.752$
- And we now have unbiased estimates with correct standard error (albeit of course larger) and is efficient.
- This is then the basis for a more general implementation (multilevel with mixed variable types) as in REALCOM and STATJR.
- Finally – a fully Bayesian procedure has been developed that is fast, very general and is also available in STATJR

Session 2: General approaches to MI

- Two approaches widely implemented.

1. Fully conditional or chained equation approach (STATA).
    - Take each variable in turn with missing data and regress on remainder (with suitable starting values and a burn-in)
    - Impute from conditional (residual) distribution
    - Repeat and select *n* completed datasets, suitably spaced
    - Fit to each model and combine (Rubin's rules)
    - Advantage is it is a series of univariate models and can handle discrete variables with missing value e.g. via logistic regression.
    - Disadvantage is it cannot deal with level 2 variables having missing values
2. Joint modelling approach (REALCOM, STATJR)
    - Basic assumption is set of variables with missing data are MVN or can be made so using a 'latent normal' transformation.
    - Form an MCMC chain, conditional on non-missing variables, where missing values are imputed from their posterior distribution at each iteration.

Joint modelling ctd.

- We have a multilevel MOI with a response (possibly >1) and covariates – possibly at several levels.
- We take all the variables at each level and make a multivariate response model with 'complete' variables either as responses or covariates.
- We finish up with a multilevel multivariate response model and at each higher level we allow the responses at that level to correlate with random effects derived from a lower level.
- At this point we can include 'auxiliaries' that are not in the MOI but might be associated with the propensity to be missing thus improving our ability to satisfy the missing at random (.MAR) – i.e. randomly missing given the other variables in the model. This assumption is needed.
- Within an MCMC chain we produce *n* 'complete data sets and fit the MOI to each one.
- Then we combine.

Combining MOIs Using Rubin's Rules

- Take the average of the point estimates

$$\hat{\beta} = \frac{1}{P}\sum_{n=1}^{N}\hat{\beta}^{(n)} \qquad \text{Within variance}(\hat{\beta}) = \frac{1}{N}\sum_{n=1}^{N}\text{var}(\hat{\beta}^{(n)})$$

- Take the average between-imputation variances

$$\text{Variance between}(\hat{\beta}) = \frac{1}{(N-1)N}\sum_{n=1}^{N}\left(\hat{\beta}^{(n)} - \bar{\hat{\beta}}\right)^2$$

$$\text{var}(\hat{\beta}) = \text{Within}(\hat{\beta}) + \left(1 + \frac{1}{N}\right)\text{Between}(\hat{\beta})$$

Handling a mixture of variable types

- All of this so far assumes normality. What if we also have categorical data?
- Most software assumes normality throughout.
- While STATA (chained equations) can handle mixed variable types it cannot handle higher level variables, nor interactions among predictors.
- Joint modelling can also handle higher level variables with missing values but not interactions.
- Handles discrete variables via threshold models, e.g. probit model for binary data. In MCMC this involves an extra step to sample underlying normal distributions.

- *Key reference:*
- Goldstein, H., Carpenter, J., Kenward, M. and Levin, K. (2009). Multilevel models with multivariate mixed response types. Statistical Modelling, 9,3, 173-197.
- Essentially works by assuming underlying (joint) normal distributions that

An efficient fully Bayesian joint modelling procedure

- We will deal with multilevel data but assume that missingness occurs only for variables defined at level 1 – although STATJR can deal with such variables defined at level 2 also.
- Our approach can deal properly with normal and categorical data with missing values and also with interaction terms for variables with missing values.
- It does not involve *multiple* imputation and inferences are made from a single MCMC parameter chain. – it is fully Bayesian. Recent work has extended the methodology to handle case weights such as typically occur in surveys.

See: Goldstein, H., Carpenter, J., and Kenward, M. (2018). Bayesian models for weighted data with missing values: a bootstrap approach. J. Royal Statistical Society, series C., DOI: 10.1111/rssc.12259

The one-pass Bayesian model - outline

- We create an MCMC chain where at each iteration we update both the missing values and the MOI.
- Consider the linear joint model:

$X_1 = X_2\alpha + \gamma_2$ - *Imputation component*

$Y = X\beta + e$ - *Model of Interest*

where $\mathrm{X}_1$are the variables that have missing values.

- For each missing value we propose a new value and then compare the joint likelihood for $X_1, Y$with value at previous iteration, in a metropolis step to decide whether to accept the new value.
- This joint updating of imputation model and MOI results in a single chain for the MOI from which we can make the usual inferences.
- We note that the MOI can include discrete responses, can be multivariate, contain interactions, multilevel terms etc, so is completely general.

STATJR (TREE)

- A menu based package.
- Running details for STATJR are available at

http://www.bristol.ac.uk/cmm/media/migrated/1-0-1/manual-tree-beginners.pdf

Download missing data files from
http://www.bristol.ac.uk/cmm/research/missing-data/

- Assume we have started STATJR and selected the tutorial dataset with two levels) in the datasets folder and just the regression template. Later we will use 2LevelMissOnePass template.

Running the model

- In the top menu bar we have several options for general settings, uploading a dataset etc.:
- Note that dataset should be in the STATJR datasets folder and in STATA format with DTA extension.
- After specifying the model we note that the model inputs have been summarised in a script under ‘current input string’. This is stored in **Templates/set inputs** so that it may be re-used or modified without going through the full model setup.
- It is usually a good idea to specify > 1 chain to verify convergence (e.g. 3)
- In outputs we can specify MCMC chain plots and e.g. residual estimates.

Imputation templates

- We will now use the *tutmiss* dataset with the one pass template – *2LevelMissingOnePass* (note that we could import the script if it has been already run, to set this up, from **templates/set inputs)**.
- We can watch progress through a script window.
- This template will not yet handle missing data for ordered or unordered categorical predictor variables, only binary ones.
- For multicategory data we need to do *multiple imputation* with the 2level (or Nlevel) impute template.
- But note that multiple imputation is slower and cannot handle interactions.

Practical session

- Users own data or tutmiss
- Fit both onepass and 2level impute
- Include interactions in onepass
- Plot chains
- Try fitting a model with missing value in a level 2 variable. You can use the data set tutmiss_lev2

References

- Realcom (2011) Developing multilevel models for REAListically COMplex social science data. University of Bristol. (Available from http://www.bristol.ac.uk/cmm/software/realcom.)
- Rubin, D. B. (1987) *Multiple Imputation for Non Response in Surveys*. Chichester: Wiley.
- Carpenter, J. R. and Kenward, M. G. (2013) *Multiple Imputation and Its Application*. Chichester: Wiley.
- Goldstein, H., Carpenter, J. R. and Browne, W. J. (2014), Fitting multilevel multivariate models with missing data in responses and covariates that may include interactions and non-linear terms. Journal of the Royal Statistical Society: Series A. 177(2), 553-564 doi: 10.1111/rssa.12022