Multilevel multiple imputation allowing for survey weights

James R. Carpenter
London School of Hygiene & Tropical Medicine
james.carpenter@lshtm.ac.uk

www.missingdata.org.uk

JRC is funded by ESRC fellowship RES-063-27-0257

July 1, 2010

Acknowledgements

Jonathan Bartlett (LSHTM)

Vernon Gayle (Stirling)

Harvey Goldstein (Bristol)

Outline

- Factors affecting GCSE score in the Youth Cohort Study
- Brief review of multilevel multiple imputation
- Including weights in the imputation model
- Application to Youth Cohort Study analysis
- Discussion

Youth Cohort Study

– The Department for Education and Skills (DFES) conducts the Youth Cohort Study (YCS) on a sample of young people (aged 16-19) in the year after they are eligible to leave compulsory schooling.

– Data are collected about their activity status, i.e. whether they are in a full-time job, full or part-time education, on a training scheme, unemployed or doing something else. Also collected is information about their qualifications (gained and studying for), family background and other socio-economic and demographic data.

– For further details see, for example, http://www.statistics.gov.uk/STATBASE/Source.asp?vlnk=668

Analysis

Our aim is to use data from 1990s cohorts of the Youth Cohort Study of England and Wales (YCS) to model relationships between educational attainment (Year 11 GCSE results) and key social stratification measures (e.g. gender, ethnicity and social class).

This work follows on from Connolly (2006) [1], where factors affecting GCSE attainment are explored using logistic regression models for three cohorts of YCS data separately.

For this talk we focus on

- handling missing data, and
- using the weights provided with the data appropriately.

Substantive model

We use a regression model to explain variability in year 11 GCSE points score by

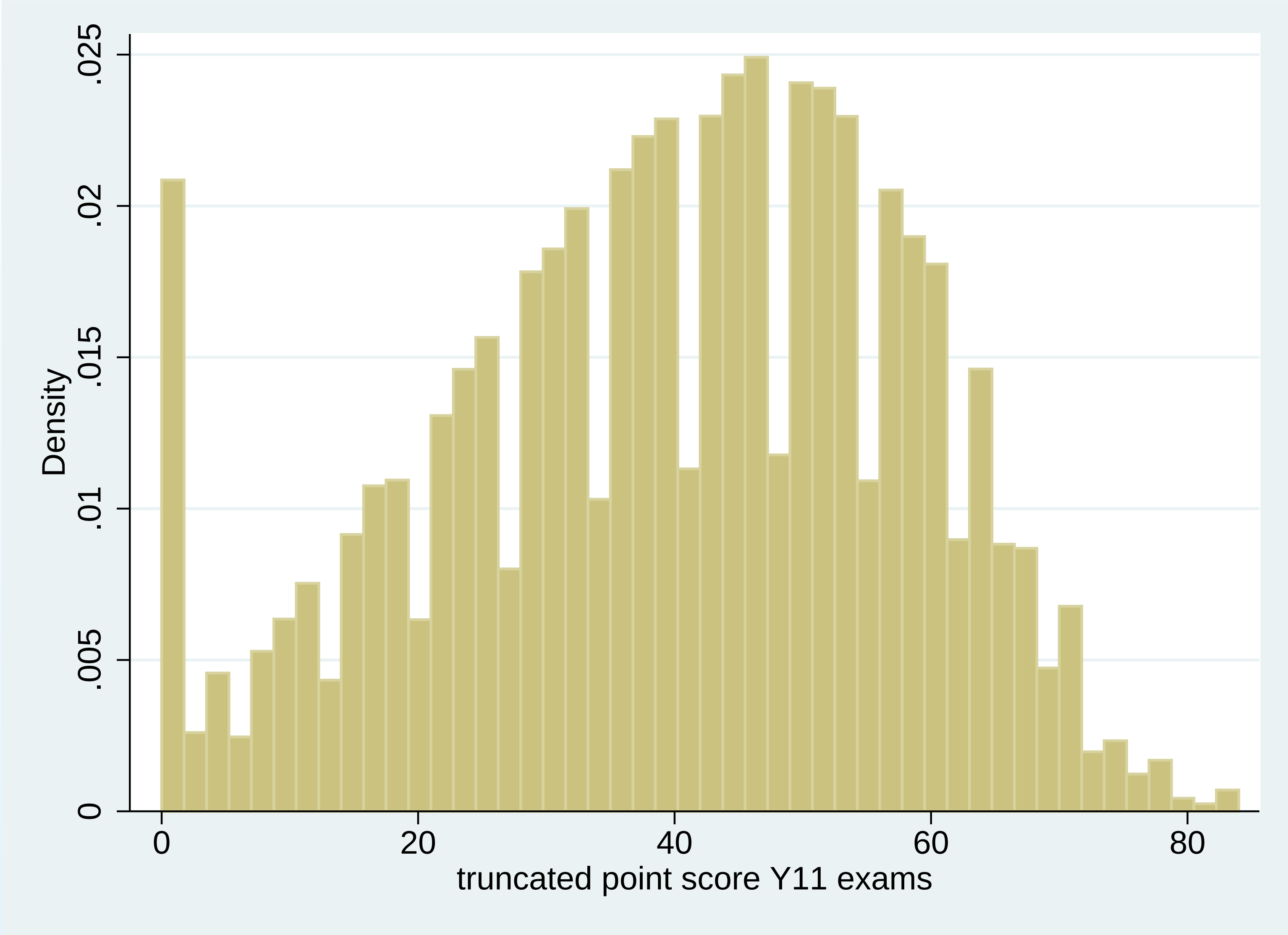
- cohort: 1990, 1993, 1995, 1997, 1999
- Gender
- Parental occupation (managerial, intermediate, working)
- Ethnicity: Bangladeshi, Black, Indian, other Asian, Other, Pakistani, White

The GCSE points score is calculated by weighting each GCSE grade by A/A*=7 through to grade G=1.

This is truncated to 12 GCSEs at A/A*.

The mean is 39.71; range: 0–84.

Distribution of GCSE score



Missing data patterns

Unfortunately, there is a non-trivial proportion of missing data.

The key patterns are:

GCSE score	Parental occupation	Ethnicity	N
+	+	+	66965
+	.	+	7523
.	+	+	760
+	.	.	651

Key predictors of missing parental occupation (adjusted) include GCSE score, gender, cohort and ethnicity (ROC=0.73).

As the reason for missing data includes the response, a complete case analysis is likely to be biased.

In addition, each wave of the cohort is supplied with weights. These range between 0.2 and 3.8, and are standardised to have mean 1.

Complete case analysis

Variable	Unweighted CC	Weighted CC
Cohort90	reference	
Cohort93	5.27 (0.20)	5.01 (0.23)
Cohort95	9.35 (0.21)	8.20 (0.23)
Cohort97	8.08 (0.21)	7.36 (0.23)
Cohort99	12.69 (0.22)	11.18 (0.24)
Boys	−3.42 (0.13)	−4.42 (0.15)
White	reference	
Black	−5.62 (0.57)	−5.43 (0.63)
Indian	3.60 (0.44)	4.13 (0.50)
Pakistani	−41.89 (0.58)	−1.95 (0.65)
Bangladeshi	0.42 (1.04)	0.46 (1.32)
Other Asian	5.36 (0.68)	6.22 (0.86)
Other	−0.36 (0.70)	−0.17 (0.84)
Managerial	reference	
Intermediate	−7.46 (0.15)	−8.06 (0.18)
Working	−13.85 (0.17)	−14.33 (0.19)

55145 out of 64045 cases used.

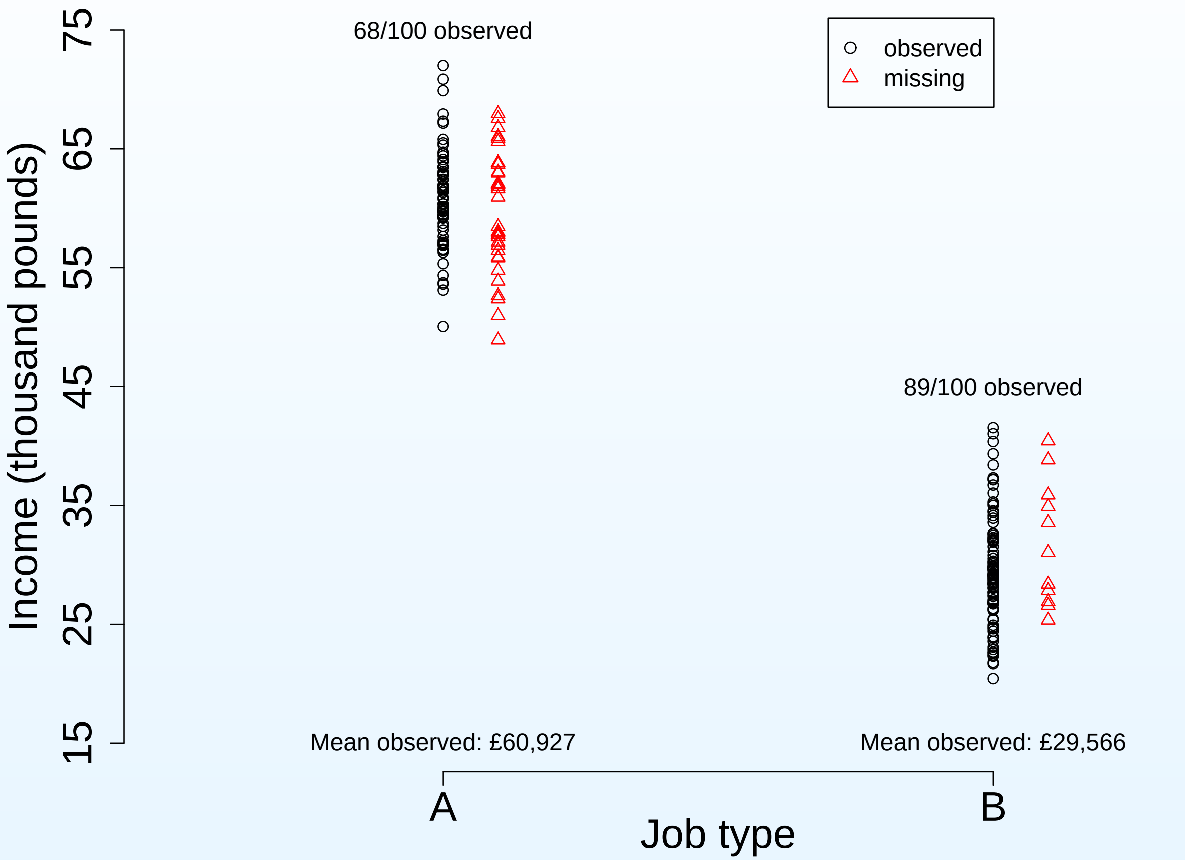
Key assumption for multiple imputation: MAR

Multiple imputation (MI) usually rests on the assumption that data are *missing at random* (MAR).

This means two things:

- two individuals with the same (similar) observed data, x, have the same (similar) conditional distribution of other variables, Y, *given* x, *whether* Y *is observed or not*, and
- while the probability of observing Y depends on Y, once we take x into account this is no longer the case.

Example: true mean income £45,000



Observed income: $\pounds 43,149.$

MAR estimate: $$\frac{100 \times 60,927 + 100 \times 29,566}{200} = \pounds 45,246$$

Multiple Imputation: intuition

Consider two variables X, Y with some Y values MAR given X.

Under the assumption that data are MAR, using only units with both observed we can get valid estimates of the regression of Y on X.

However, inference based on observed values of Y alone (eg sample mean, variance) is typically biased.

This suggests the following idea

1. Fit the regression of Y on X
2. Use this to impute the missing Y
3. With this completed data set, calculate our statistic of interest (eg sample mean, variance, regression of X on Y).

As we can only ever know the *distribution* of missing data (given observed), steps 2,3 have to be repeated, and the results ‘averaged’ in some way—Rubin’s rules are appropriate for this.

Joint modelling approach

To implement MI, we need to choose and fit the *imputation model.*

This is a a multivariate response model, where partially observed variables are responses, and fully observed variables are covariates.

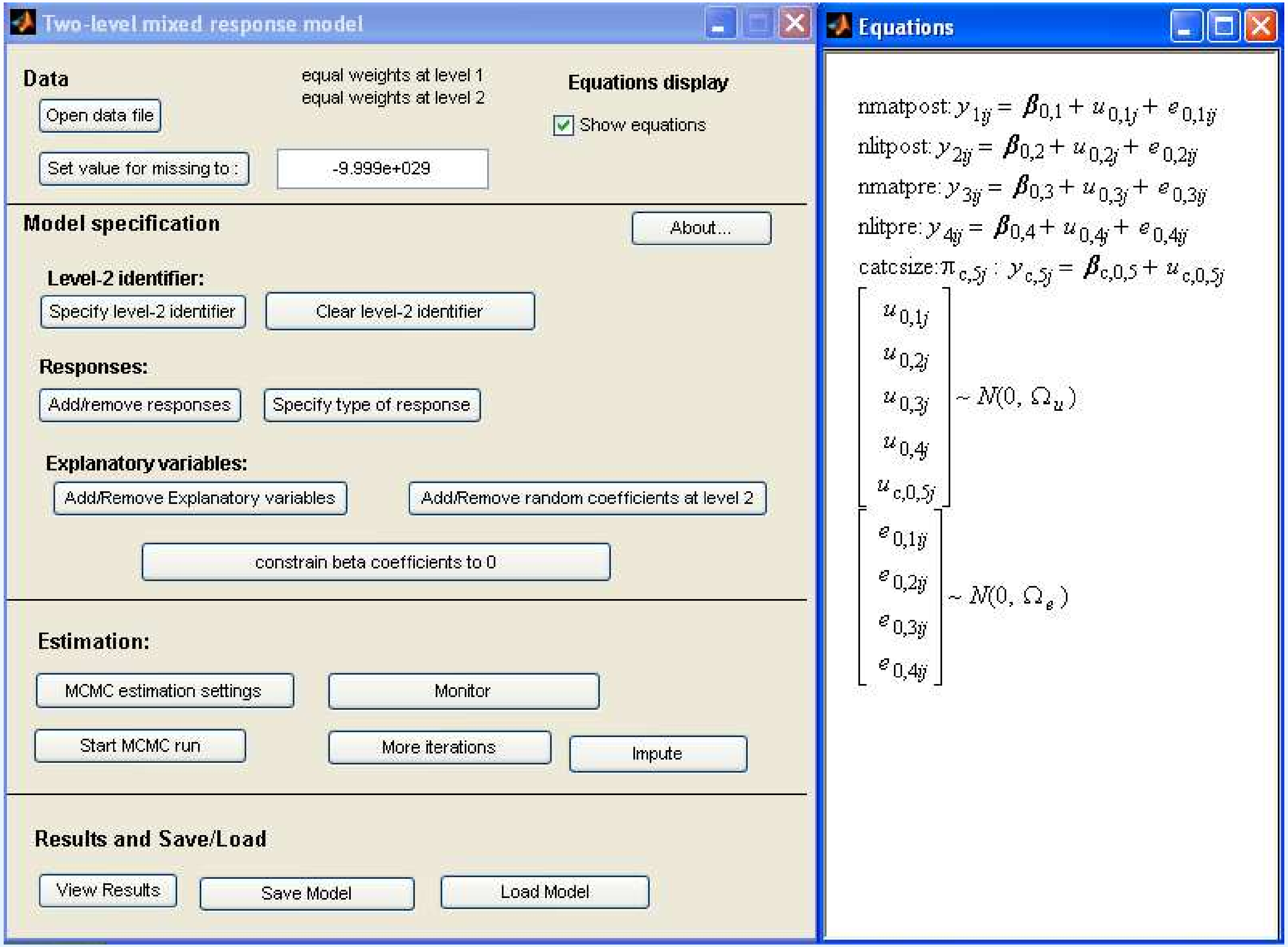
The responses will generally be a mix of continuous, binary, ordinal and unordered categorical variables.

The imputation model is usually multilevel, with partially observed responses at both levels.

The model is fitted using Markov Chain Monte Carlo methods, and this naturally allows imputed data to be generated *taking full account of the uncertainty*.

Freely available software for doing this, called `REALCOM`,z can be downloaded from www.cmm.bristol.ac.uk; macros for use with `MLwiN` can be downloaded from www.missingdata.org.uk

Snapshot of REALCOM



Preliminary step

Consider a 2-level setting, and let j index level 2 units and i index level 1 units.

Suppose we have n_j level 1 units in each level 2 unit, m level 2 units, and $N = \sum_j n_j$ level 1 units in total.

Let w_j be the weight attached to level 2 unit j, and $w_{i|j}$ the weight attached to level 1 unit i within level two unit j.

We first scale the weights so that the level 1 weights within each level two unit have mean 1, i.e. $\sum_i w_{i|j} = n_j$, and likewise for the higher levels, $\sum_j w_j = m$.

We then define the composite weight as

$$w_{ij} = N w_{i|j} w_j \Bigg/ \sum_j n_j w_j \, .$$

Incorporating the weights

Let Z_u and Z_e respectively denote the sets of explanatory variables for the level 2 and level 1 random coefficients.

Define W_u as the $m \times m$ matrix with diagonal $\{w_j^{-0.5}\}$ and zero elsewhere.

Likewise define W_e as the $N \times N$ matrix with diagonal $\{w_{ij}^{-0.5}\}$ and zero elsewhere.

We simply replace Z_u, Z_e in the estimation process by $Z_u^\star = W_u Z_u$ and $Z_e^\star = W_e Z_e$.

In the single level case this is equivalent to the usual procedure for weighted regression. Pfeffermann *et al* (1998) [2] carry out simulations and show that this procedure has good coverage properties, even though it is not equivalent to the full weighted likelihood procedure.

We also note that for the case of equal level two weights, this procedure does give weighted maximum likelihood estimates.

Relationship to weighting in Stata

For those who use Stata, we make the explicit link with the various weighting options.

The weights we use above

(a) are not ’frequency weights’ which indicate replicated units;

(b) are not strictly inverse variance weights, though for a single level analysis this and (c) below are the same, and

(c) are effectively ’inverse probability of observation’ weights - ie weighting for unequal selection.

Analyses

We did not include any auxiliary variables in these analyses, though this is usually a good idea.
We carried out the following:

- unweighted and weighted complete case analysis;
- unweighted multiple imputation and unweighted analysis of imputed data;
- weighted multiple imputation and unweighted analysis of imputed data;
- unweighted multiple imputation and weighted analysis of imputed data, and
- weighted imputation and weighted analysis of imputed data.

Results

We focus on the results for ethnic group:

Analysis	Race (reference: white)				
	Black	Indian	Pakistani	Bangladeshi	Other Asian
CC-U	−5.6 (0.6)	3.6 (0.4)	−1.9 (0.6)	0.4 (1.0)	5.4 (0.7)
CC-W	−5.4 (0.6)	4.1 (0.5)	−2.0 (0.7)	0.5 (1.3)	6.2 (0.9)
MI-U, M-U	−7.0 (0.5)	2.7 (0.4)	−4.6 (0.5)	−5.2 (0.7)	4.3 (0.6)
MI-U, M-W	−6.7 (0.5)	2.9 (0.4)	−4.6 (0.5)	−5.1 (0.7)	4.7 (0.7)
MI-W, M-U	−6.8 (0.5)	2.8 (0.5)	−4.7 (0.5)	−5.2 (0.7)	4.3 (0.6)
MI-W, M-W	−6.7 (0.5)	3.0 (0.4)	−4.7 (0.4)	−5.1 (0.7)	4.8 (0.7)

Key: -W: weighted; -U: unweighted
MI: multiple imputation; M: model of interest
CC: complete case

Conclusions

Computational

- We have generalised the REALCOM software to allow the inclusion of weights to adjust for unequal selection.
- When analysing data in MLwiN using weights, these are automatically picked up by REALCOM for MI.

Practical

- MI makes efficient use of partially observed data, and corrects bias when the missingness mechanism includes the response.
- A key requirement with multiple imputation is that the model of interest and the imputation model are consistent, or equivalently congenial.
- Thus, if weights are intended for the analysis, they should be used for the imputation.
- Allowing for the weights, and using MI, this preliminary analysis of the YCS cohorts from the 1990's suggests that average GCSE score among Bangladeshi, Black and Pakistani children is markedly below that of whites, with Indian and other Asian ethnic groups having the best average score.

References

[1] P Connolly. The effects of social class and ethnicity on gender differences in gcse attainment: a secondary analysis of the youth cohort study of england and wales 1997-2001. *British Educational Research Journal*, 32:3–21, 2006.

[2] D Pfeffermann, C J Skinner, D Holmes, H Goldstein, and J Rasbash. Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society, Series B*, 60:23–40, 1998.