Issue 65: Autumn 2014

social research Update

Multiple Imputation for handling missing data in social research

Ian Brunton-Smith, James Carpenter, Mike Kenward, and Roger Tarling

Ian Brunton-Smith and Roger Tarling are quantitative social researchers and members of the Criminology and Criminal Justice research group in the Department of Sociology, University of Surrey.

James Carpenter and Mike Kenward are statisticians at the London School of Hygiene and Tropical Medicine. Together they run the website http://www.missingdata.org.uk, and support medical and social researchers with analysis of incomplete datasets. The site includes details of all aspects of missing data, an extensive bibliography, slides of talks and notices of meetings.
James Carpenter is supported by ESRC grant RES-063-27-0257.

<I.R.Brunton-Smith@surrey.ac.uk>

- **Missing data are ubiquitous in quantitative social research and can lead to incorrect inferences**
- **Statistically principled approaches have been developed to address the problem of missing data**
- **Of the available approaches, we favour multiple imputation (MI) for its flexibility, accessibility and ease of use**
- **MI is described and a worked example, using the statistical software Stata, is presented**

Missing data frequently occurs in quantitative social research. For example, in a survey of individuals, some of those selected for interview will not agree to participate (unit non-response) and others who do agree to be interviewed will not always answer all the questions (item non-response).

At its most benign, missing data reduces the achieved sample size, and consequently the precision of estimates. However, missing data can also result in biased inferences about outcomes and relationships of interest. Broadly, if the underlying, unseen, responses from those individuals in the survey frame who have one or more missing responses differ systematically from those individuals in the survey frame whose responses are all observed, then any analysis restricted to the subset of individuals whose responses are all observed runs the risk of producing biased inferences for the target population.

Thus every researcher needs to take seriously the potential consequences of missing data. This paper describes the use of *Multiple Imputation* (MI) to correct estimates for missing data, under a general assumption about the cause, or reason for missing data. This is generally termed the *missingness mechanism*. MI has robust theoretical properties while being flexible, generalisable and readily available in a range of statistical software.

Missingness Mechanisms

The missingness mechanism is central to the potential impact of missing data on the precision and accuracy of results from an analysis. If data are Missing Completely at Random (MCAR), such that the chance that data is missing on an occasion does not depend on any variables (whether fully or partially observed) an analysis carried out using the

subset of complete records is valid and unbiased, if potentially inefficient because it will be based on a smaller sample than intended.

If data are not MCAR they may yet be Missing at Random (MAR). Broadly speaking, this means any systematic pattern of missingness can be 'explained' by observed data. For example, people on high incomes may be less willing to provide details of how much they earn in a survey (and hence missingness on income depends on a respondent's income). If we also have observed data on each respondent's occupation (distinguishing between high paid and low paid jobs), and within each job type we have a random sample of all income levels in that occupation, the data are MAR – controlling for job type. In other words, job type fully explains the association between income and the chance of observing income. Crucially, if data can be assumed MAR, valid and unbiased analyses can still be conducted if the additional variables (i.e. job type in this example) are included and appropriate statistical methods are adopted.

Methods for imputing missing data

There has been much statistical research on the problem of missing data, and a range of statistically principled approaches have been proposed. The aim is to maximise the efficiency of the analysis and/or to correct for bias due to missingness by using some form of adjustment. This uses all the observed values of variables in the statistical model addressing the analyst's substantive question (the *substantive model*), together with additional variables, not in the substantive model, if available. These are termed auxiliary variables. Auxiliary variables should help predict missing values (this improves precision). If they also predict the probability of a variable being missing, they will correct bias. For example, occupation type above does both of these, and thus increases the plausibility of the MAR assumption. These statistically principled approaches move beyond traditional approaches like mean imputation (which simply replaces all missing values with the mean of the variable) and which fail to take into account the variability that would naturally occur amongst the values had they not been missing.

The three broad classes of method are *Inverse Probability Weighting* (IPW), *Full Information Maximum Likelihood* (FIML) and *Multiple Imputation* (MI). Despite recent advances, IPW adjustments are potentially very inefficient, because they only weight the complete records; information from individuals with partial data is discarded. Appropriately specified FIML methods produce similar results to MI when data are missing in the dependent variable in a regression, but are much more technically challenging when data are missing in independent variables. We therefore focus on MI.

MI proceeds as follows. First, the imputation process takes all the variables in the substantive model, together with any auxiliary variables, and defines an imputation model (often implicitly). This is fitted, and then used to impute the missing values, thus resulting in a 'completed' dataset. Importantly, this imputation process takes full account of the uncertainty in the imputation model parameters.

This process is then continued to create a number (say 40) of 'completed' datasets (each with different, yet plausible, imputed missing values). Together these represent the distribution of the missing data, given the observed data. The substantive model is fitted to each completed data set (using the standard complete data method), and the relevant parameter estimates and measures of precision obtained. In the second step (using 'Rubin's rules'), the sets of estimates and variances are combined to give one overall estimate and a valid measure of precision. Most importantly the precision estimates take account of the process of imputation, so there is no sense of data being 'created' or 'made up', in contrast to ad-hoc single imputation approaches.

An important advantage of the MI approach is the opportunity to have auxiliary variables in the imputation model that need not be in the substantive model. Further, the same imputation model, and hence imputed data sets, may be appropriate for several different substantive analyses. MI also has the conceptual advantage that the same substantive model, and fitting method, which would have been used had there been no missing values, is retained. Hence there is a sense in which the MI analysis is 'closer' conceptually to the originally conceived substantive analysis.

Example of Multiple Imputation

The Surveying Prisoner Crime Reduction (SPCR) survey is a four wave longitudinal panel survey of prisoners across England and Wales, split between interviews within prison and interviews in the community on release from prison. To illustrate multiple imputation, we only consider here the first two waves (the prison stage of the study), restricted to a subset of the sample comprising 2,841 prisoners serving sentences of over six months. At wave 1, interviews were conducted with prisoners on reception into prison between November 2005 and November 2006, with the second interview conducted in the two weeks prior to release from prison. Although all should have been interviewed at wave 2, a significant number, 1,053 (37%) were not.

Our substantive research question is to identify factors that are associated with reoffending one year after release from prison. Our response is thus a binary variable and we

have complete data for it from the Police National Computer. In this illustrative example we wish to examine the effects of: ethnicity, age, gender, whether the offender was a regular crack user prior to going to prison, and an indicator of previous offending as measured by whether the prisoner had received a prior prison sentence. These five explanatory variables were measured at wave 1 and there are no missing data. In addition, we examine whether involvement in education and training or behaviour management programmes as part of a prison sentence results in a different propensity to re-offend within a year of release. Both are binary variables and were measured at wave 2, where a significant proportion of the data were missing.

In the Table we present the results from fitting three logistic models. The first is restricted to the five wave 1 variables and is fitted to the full sample of 2,841 prisoners. The results indicate that the odds of reoffending fall for 'non-whites', for older prisoners and females while the odds of re-offending are significantly higher for those who had served prior prison sentences and those identified as having a drug dependency prior to incarceration.

Model II extends the analysis to incorporate the two measures observed at wave 2. Based on complete case analysis (the default when no adjustment is made for missing data), the sample size for the full analysis has dropped to 1,788 (losing information from the 1,053 respondents that were not interviewed at wave 2). Reflecting this reduced sample size, the standard errors from the model have been inflated a little, reducing the precision of results. A number of parameter estimates and odds ratios have also been altered by a small amount in the revised model. Notable is that non-white is no longer significant in Model II. Of the two new wave 2 variables included, the model reveals statistically significant, lower odds of reoffending amongst those enrolled on an educational training course during their prison sentence, but no significant differences in the likelihood of reoffending for those enrolled on a behaviour management course.

Model III refits the second model but after multiple imputation. Before discussing the model and comparing it with models I and II we describe how MI was conducted.

The first step was to examine the nature of missing data and to identify the factors associated with it. An initial analysis found that item non-response was not a problem; that is, if a prisoner was interviewed, he or she answered all questions. The major problem was thus unit non-response, conducting an interview with the prisoner in the first place. Furthermore, it was found that unit non-response had two distinct elements. First, at wave 2 it proved not possible to contact 837 of the eligible sample and second, 216 of those contacted refused to be interviewed

Analysis was then undertaken to isolate those variables that were independently and significantly associated with missingness. In consultation with the survey company, Ispos-MORI, one problem identified in making contact with prisoners at wave 2 was that many were released early and before the interview was scheduled to take place. This issue was spotted early on and at later stages of the project, wave 2 interviews were timed to take place earlier in the sentence to ensure that the prisoner had not been released. Thus non-contact was higher for prisoners taking part in the first 12 months of the survey. This problem was compounded for

	Model I: Complete Cases - wave 1 only				**Model II: Complete cases - wave 1 and 2**				**Model III: Imputed analysis**			
	Beta	S.E	Odds Ratio	Sig	Beta	S.E	Odds Ratio	Sig	Beta	S.E	Odds Ratio	Sig
Cons	-1.73	0.09		0.00	-1.60	0.12			-1.64	0.10		0.00
Non-White	-0.31	0.13	0.73	0.02	-0.20	0.16	0.82	0.21	-0.29	0.13	0.75	0.02
Age	-0.06	0.01	0.94	0.00	-0.07	0.01	0.94	0.00	-0.06	0.01	0.94	0.00
Female	-0.87	0.16	0.42	0.00	-1.04	0.22	0.35	0.00	-0.90	0.16	0.41	0.00
Prior prison sentence	1.44	0.10	4.22	0.00	1.34	0.13	3.80	0.00	1.43	0.10	4.17	0.00
Daily crack use	0.92	0.14	2.51	0.00	0.62	0.17	1.87	0.00	0.93	0.14	2.52	0.00
Educational training (w2)					-0.26	0.12	0.77	0.03	-0.28	0.12	0.76	0.02
Behaviour management course (w2)					0.12	0.13	1.13	0.37	0.07	0.14	1.07	0.62
Sample size	2841				1788				2841			

short term prisoners (those serving a sentence of less than 18 months). Prisoners imprisoned for a theft offence or for a short-term following conviction for a drug offence were also less likely to be contacted. Two variables improved the likelihood of contact: whether the prisoner had served a previous prison sentence or had received a prior sentence for burglary. Young prisoners were more likely to agree to the interview. In addition, achieving contact at seven (of the 117) prisons and compliance at five proved to be more difficult. Thus two binary variables were created for those two groups of prisons (PrisonNoncontact and PrisonRefusal).

Thus nine auxiliary variables were identified as related to missingness. Note that as well as being auxiliary variables, both age and having received a previous prison sentence are also variables of interest in our substantive model. This is perfectly permissible in MI analysis. Further, the above discussion is consistent with missing data being MAR (taking into account the information in the auxiliary variables).

The imputation model must therefore include all variables in the substantive model including the fully observed outcome (Model II above), but also incorporate the list of auxiliary variables linked to the process of missing data and predictive of the missing values. MI was carried out in Stata using the *Imputation using Chained Equations* (ICE) algorithm implemented by Patrick Royston (see below for details). This is more flexible than the built in Stata commands for MI, although in practice for most applications the results will be the same. The relevant command (with self-explanatory variable names) to run the imputation model is:

```
mi ice Reoffend NonWhite Age
Male PriorPrison Dailycrack
EducationW2 BehaviourManW2
EarlyRelease Sentlt18 Theftoff
ShortSentdrug PriorSentBurglary
PrisonNoncontact PrisonRefusal,
add(40)
```

Although age and previous prison sentence are both auxiliary and variables of substantive interest, they only need to be entered once in the command. (`mi ice` automatically recognises which variables are complete and which have missing data.) `add(40)` requests 40 random imputations. Having performed the imputations, we continue by estimating the 40 separate models from the imputed datasets and combining them into one overall model with the Stata command:

```
mi estimate: logit Reoffend
NonWhite Age Male PriorPrison
Dailycrack EducationW2
BehaviourManW2
```

The output is shown as Model III in the Table and includes details of the combined estimates from the imputed datasets. The first thing to note is that the sample size has increased to the original 2,841, confirming the inclusion of all the respondents that were not re-interviewed at wave 2. As a result of the increased sample size, the standard errors in the imputed model return almost to the same magnitude as Model I. The odds ratios for the wave1 variables are more similar to the Model I values than to Model II. Notably, the variable Non-White returns from non-significance in Model II to statistical significance as in Model I.

Thus the imputed analysis now more closely resembles Model I and the advantage of MI is clearly seen in the recovery of the information on the wave 1 variables. Precision in the wave 1 variables is primarily being recovered because we bring into the analysis the information from all those with wave 1 data but with wave 2 data missing. Furthermore, coefficients from the wave 2 variables are similar in Models II and III, suggesting that given the other variables in the model the chance of these being missing is not strongly associated with whether the prisoner reoffended within one year of release.

Bibliography

An introductory text which explains missing mechanisms and methods for imputation is:

Enders, C. K. (2010) *Applied Missing Data Analysis*. London: The Guilford Press.

A more technical text which includes methods for complex datasets (multilevel and longitudinal) is:

Carpenter J. R. and Kenward M. G. (2013) *Multiple Imputation and its Application*. Chichester: Wiley..

ICE is fully described on Patrick Royston's website (http://www.homepages.ucl.ac.uk/~ucakjpr/). To download the relevant packages type "net from http://www.homepages.ucl.ac.uk/~ucakjpr/stata" in Stata.

Further details of SCPR and additional examples of MI can be found in:

Brunton-Smith, I., Carpenter, J., Kenward, M., and Tarling, R. (2014). *Surveying Prisoner Crime Reduction (SPCR) Missing Data Recovery.* Ministry of Justice Research Series X/12.

social research UPDATE is distributed without charge on request to social researchers in the United Kingdom by the Department of Sociology at the University of Surrey as part of its commitment to supporting social research training and development.

Contributions to *social research UPDATE* that review current issues in social research and methodology in about 2,500 words are welcome. All UPDATE articles are peer-reviewed.

social research UPDATE (ISSN: 1360-7898) is published by the Department of Sociology, University of Surrey
Guildford GU2 7XH, United Kingdom. tel: +44 (0)1483689450
Edited by Nigel Gilbert (n.gilbert@surrey.ac.uk)