Longitudinal and Life Course Studies 2009 Volume 1 Issue 1 Pp 63-72

Handling attrition and non-response in longitudinal data

Harvey Goldstein
University of Bristol

Correspondence.
Professor H. Goldstein
Graduate School of Education
University of Bristol
Bristol BS8 1JA
UK
Email: h.goldstein@bristol.ac.uk

Abstract

Procedures for handling attrition and missing data values in longitudinal studies are discussed. A multiple imputation (MI) strategy is developed that can be applied to complex multilevel data. It is both general and statistically efficient and estimation software is available. An example of its use is given.

Keywords

Multilevel, attrition, longitudinal, multiple imputation, weighting, latent normal model.

Acknowledgements

I am most grateful to Jon Rasbash and referees for helpful comments.

Introduction

Attrition in longitudinal studies is often seen as a serious problem for two reasons. First, the loss of individuals over time will often result in a sample size, after a few occasions or 'sweeps', very much smaller than the initial sample size. Thus, for example Hawkes and Plewis (2006) report that just 71% of the target sample at sweep 6 (42 years) of the National Child Development Study provided information, compared with 99% at the start of the study. They also point out that 11% of the sample at sweep 6 had missed one or more earlier sweeps. For those analyses that utilise data at more than one occasion, if only those individuals with data at all such occasions are used in the analysis this will result in a loss of efficiency. Note that we use the term 'attrition' to mean any pattern of loss of individual records over time, including those cases where individuals may return to a study after missing measurement occasions.

Secondly, loss may not occur at random so that the remaining sample may be biased with respect to the variables being analysed. In longitudinal studies, at any given occasion the characteristics of subsequent losses will be known and these can be compared with those who are followed up. If biases are detected then suitable weights can be introduced to compensate for this, and this is the traditional approach to dealing with attrition.

The present paper sets out a general model-based approach to dealing with attrition in longitudinal studies. It does this by embedding the problem within a general approach to handling missing data and the procedure will, in principle, handle both the loss of individual records over time and the loss of individual data items. In the next section, we summarise briefly the weighting approach. For further details of weighting, including the use of auxiliary variables see, for example, Schouten and De Nooig (2005). Following this we then describe our general model for attrition and give an example.

Weighting procedures

The methodology for computing weights specifically in order to compensate for 'informative' attrition that leads to biases, involves procedures for estimating the probability of a sample member responding, for each sample member at each occasion, as a function of sample member characteristics. Hawkes and Plewis (2006) provide a useful description of a model-based approach to this. The resulting (inverse) probabilities are then used in standard ways in subsequent analyses. One of the problems with such procedures is that response may be partial. Thus, for example, all individuals may respond to a set of educational variables but not to health variables at a particular occasion. In this case, we would not use weights to adjust for attrition when analysing the educational data but we would wish to do so if health variables were also being analysed. This will complicate matters generally, for example because successive analyses may be based upon different numbers of individual cases. A further problem is that for a particular individual the weights will generally change from occasion to occasion, depending on the possibly changing response patterns across the sample, which may alter the joint distribution of respondent characteristics, and this will create difficulties when conducting analyses across several occasions.

Another set of issues when using weights is that a traditional weighting approach will generally lose data information. For example, suppose we wish to regress a time 2 variable on a set of time 1 variables with attrition occurring. Weights can be obtained in various ways but the weighted analysis will then be carried out using only the cases with measurements at both occasions. This ignores the information, from one occasion even though lost at another, that is not incorporated in weights but may be available, for example, from auxiliary variables or further covariates in a model. A similar problem arises in the case mentioned above where there is a differential response to, say, health and educational variables. As we will show, our proposed procedure avoids this problem.

In the next section we describe how attrition can be formally considered within a missing data framework.

Missing data and attrition

We shall start by making the simplifying assumption that data are missing at random (MAR) conditional on a set of 'conditioning variables' that have been collected for both respondents and non-respondents. The conditioning variables might be variables collected at the start of the survey or they may be 'auxiliary variables' that are of no analytical interest but may be available and associated with the propensity to be missing. In addition to their use for attrition purposes, such auxiliary variables may also be used to deal with non-response at the first stage of data collection. As we shall explain below, these auxiliary variables will not generally belong to the statistical model being fitted, the so called 'Model of Interest' (MOI) and are introduced alongside the MOI and linked to it. These might be interviewer characteristics, neighbourhood characteristics etc. The MAR assumption is crucial and underlies our modelling procedure (see for example Little and Rubin, 1987).

An individual record that is missing is a special case of missing data where all the variables at that occasion are missing. The existence of known conditioning variables for such records allows us to use multiple imputation (MI) where we condition on these variables. Thus, formulating attrition in this way will allow us to use a common consistent approach to handling item missing data and attrition within a single model. It will also be efficient since it allows us to use all the available data on individuals and not simply those data on variables that are present across all the occasions being used in an analysis, as in the weighting approach described above.

In the next section we look at a full model-based procedure for handling the attrition problem and this is followed by a discussion of multiple imputation methods.

A full model-based procedure for informative attrition and item missing data

Where we cannot assume MAR and attrition is informative it may be possible to adopt a fully model-based procedure. In this case we need to write down an explicit model for the probability of attrition and link this to the model of interest (MOI) that we wish to fit. Consider a simple situation where the MOI is given by (1) and the response probability model is given by (2).

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \quad (1)$$

$$\Pr(response\ observed) = \alpha_0 + \alpha_1 z_i + e_{2i} \quad (2)$$

where the ***Z*** is an auxiliary variable (or more generally a set of variables) that predicts the probability of a response (***Y***) being observed and is uncorrelated with the random effects e_{1i}, e_{2i} in (1) and (2). The two models are linked by assuming

$$E(e_{1i}e_{2i}) = \sigma_{e12} \neq 0$$

In practice we would use a nonlinear link function for the response probability such as the *probit*, and that particular function would allow us to formulate the situation conveniently in terms of a bivariate normal model in the case where e_{1i} is normally distributed.

The existence of $\sigma_{e12} \neq 0$ is a statement that the propensity for a record to be missing is related to the response in the MOI or that the attrition is informative. In such a case a joint analysis of (1) and (2) will yield unbiased estimates. This is conveniently carried out in a Bayesian framework where the missing values are treated as parameters to be estimated and samples from the appropriate posterior distributions are chosen at each iteration of a Markov Chain Monte Carlo (MCMC) algorithm (see for example Gilks et al., 1996).

While such a full model based procedure is often attractive, it does not deal with the case where individual items in a record additionally may be missing. In order to extend this model to the case where there is item missing data, we can use MCMC, assuming MAR for the missing items, and treating all the missing values as additional parameters, but this becomes computationally very time-consuming. Further references are those of Nathan (1983), Nathan and Holt (1980) and Pfeffermann (2001). The multiple imputation procedure described below will handle attrition where whole records are missing as well as missing items.

Random multiple imputation for missing data

Imputation is a procedure for handling missing data that works by constructing a complete data set, replacing every missing value by an 'imputed' value that is generated by a specified algorithm. The completed data set can then be analysed in the usual manner. Several imputation methods have been proposed but here we shall consider only the now standard method of Random multiple imputation first introduced by Rubin (1987) and an application to survey data can be found in Rubin (2004). A useful introduction to imputation can be found at www.missingdata.org.uk which also provides macros for fitting certain kinds of multiple imputation for multilevel structures using the multilevel modelling package MLwiN (Rasbash et al, 2008).

If we have a survey where some individuals do not respond, but we do have some auxiliary information about them, for example characteristics of where they live, reason for non-response, characteristics of interviewer, or in a longitudinal study earlier variable values, then formally we can view this as a missing data problem. The full set of variables is considered to be the survey items plus the auxiliary data and in the case of attrition all the survey items at that occasion are missing. In addition we can have missing survey items for respondents, perhaps by design as in rotation sampling, and even missing data for some auxiliary variables. In imputation we condition on all the observed variables, including the auxiliary variables, when creating our imputed values, and we describe how this is actually done below. If we have an efficient procedure for handling general patterns of missing data then this will lead to a single comprehensive analysis model that simultaneously will handle what is conventionally described as attrition in longitudinal data (complete missingness at certain occasions) and conventional missing data situations where just some variable values are missing.

When carrying out imputation, while it is important that all relevant auxiliary variables are conditioned upon, these do not all have to be included in the final model of interest (MOI). It is necessary, however, that any variable used in the MOI is used in the imputation stage. Thus, for example, if a multilevel structure is part of the MOI, then a relevant multilevel imputation procedure should be used. This general requirement may create problems if imputed data sets are being created for secondary data analysis so that care needs to be taken.

A simple example of multiple imputation

We consider dividing into sets the variables of interest in our data. Set A are those variables that constitute the response and predictors in the MOI, which may for instance be a linear or generalised linear model. If the model is multilevel the predictors will include those variables defining the random effects. Set B is a set of further conditioning variables that do not feature in the model of interest but which are correlated with those in set A. Set C is the union of sets A and B. Note that we include here as conditioning or auxiliary variables any variable not in the model of interest, whether collected within the survey or outside it; the sole requirement is that for some set B variables, information is available on both respondents and non-respondents. Even this assumption can be relaxed, however, if some auxiliary variables have missing values, by including these variables as responses rather than predictors at the MI stage (see below).

The procedure has two stages. In the first stage we set up a (possibly multilevel) model where the set A variables are treated as a multivariate response vector, each response regressed on just an intercept.[1]

To illustrate, consider a simple regression model where we have a model with a single, normal, response, *y* and a single, normal predictor, *x*. The model of interest is

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad (3)$$

and we may have missing data in both *X* and *Y*. We now set up an 'imputation' model that has all the variables as responses with just an intercept predictor, i.e.

$$\begin{aligned} y_i &= \alpha_1 + e_{1i} \\ x_i &= \alpha_2 + e_{2i} \\ \begin{pmatrix} e_1 \\ e_2 \end{pmatrix} &\sim N\begin{pmatrix} 0 & \sigma_1^2 \\ 0, & \sigma_{12} \quad \sigma_2^2 \end{pmatrix} \end{aligned} \qquad (4)$$

and (4) is just a bivariate normal model where each response is modelled by its mean.

By fitting model (4) we can incorporate missing as well as observed responses using the procedure described below. If we fit this model we will obtain the intercept estimates, the residual variances and the covariance. We will also have estimates for the residuals $\hat{e}_{1i}$, $\hat{e}_{2i}$, obtained by subtraction.

Suppose now that an *x* value is missing so that we can no longer estimate the corresponding residual by subtraction. Nevertheless we do have an estimate of the *distribution* of *x*, namely $N(\hat{\alpha}_2, \hat{\sigma}_2^2)$, that is a normal distribution with estimated mean $\hat{\alpha}_2$ and estimated variance $\hat{\sigma}_2^2$.

We can therefore generate a value at random from this distribution and this becomes our imputed value. In practice we fit (4) using Markov Chain Monte Carlo Methods. After each of a set of suitable chain intervals, for example at iteration 1000, 2000….., we randomly sample a complete set of imputed values for each missing value. The intervals between these sampled sets should be long enough to guarantee (approximate) independence for the sampled values. This will be done n times, providing a single set of imputed plus observed values after each of n intervals, yielding multiple (n) 'complete' datasets. The value of n required will depend on the application, but for a multilevel model may need to

[1] We may choose at this stage to place one or more fully observed set B variables as predictors for this set of responses. This is the standard procedure for traditional weighting methods. There may be computational advantages where there are a large number of variables. In multilevel modelling we will also have further random coefficients for some of these predictors. Note, however, that where we have missing values in the set B variables, either for those cases that are missing or those that are observed, they must be treated as responses.

be as high as 20 or more. The original model of interest (3) is fitted this number of times yielding n sets of parameters. These are then averaged to provide the final estimates. The details of this procedure are given at www.missingdata.org.uk.

We note that in fact the parameters for model (3) can, for this simple case, be derived from the parameters fitted to model (4), the latter being sufficient. For more complex models including multilevel ones, however, there will typically be no straightforward way to do this, as in the following case.

Suppose that we now have an auxiliary variable, *Z*, that is associated with the propensity to be missing. We can extend (4) as follows

$$
\begin{aligned}
y_i &= \alpha_1 + \gamma_1 z_i + e_{1i} \\
x_i &= \alpha_2 + \gamma_2 z_i + e_{2i} \\
\begin{pmatrix} e_1 \\ e_2 \end{pmatrix} &\sim N\begin{pmatrix} 0 & \sigma_1^2 \\ 0 & \sigma_{12} \quad \sigma_2^2 \end{pmatrix}
\end{aligned}
\qquad (5)
$$

where the missing values now depend on *Z* so that, if the relationships are as assumed in (5), conditional on *Z* we can assume we have MAR for the imputed values. If any of the values of *Z* were missing then we can incorporate *Z* as a response variable in (5) and this would have a similar effect. This procedure is readily extended to more complex structures, including multilevel ones such as that described later in an example.

We have assumed above that the data are multivariate normal and most treatments of multiple imputation make such an assumption. However, many predictor (and response) variables are binary, ordered or nominal. Treating binary or ordered variables as normal can lead to biases, especially where there are very few cases in one or more categories. Treating a *p*-category nominal variable using a set of *p-1* (0,1) indicator categories and assuming multivariate normality for these, likewise can lead to biases. This is where a 'latent Normal' variable approach may be used. To illustrate this, consider the case where *Y* is normal but *X* is a binary variable, for example whether or not a student passes an examination.

Suppose we have a normally distributed underlying variable, with a variance fixed at 1 to ensure identifiability, and mean μ

$$z_i \sim N(\mu, 1),$$

Where we observe an exam pass, that is a positive (=1) response for our binary variable *x* if *z* is positive, that is

$$z_i = \mu + e_i > 0 \quad \text{or}$$

$$e_i > -\mu$$

So that we have

$$\text{Prob}(x = 1) = \text{Prob}(e_i > -\mu) = \int_{-\mu}^{\infty} \phi(t)dt = \int_{-\infty}^{\mu} \phi(t)dt \qquad (6)$$

where $\phi(t)$ is the standard normal density function. Equation (6) is a probit characterisation and given an observed binary response (0 or 1) and an estimate for μ we can randomly sample a value from the underlying normal distribution *Z* at each cycle of our MCMC algorithm. This can be done in such a way that *Y, Z* have a bivariate normal distribution and we can then apply the missing data procedure we have already described. This provides us with imputed values for the *Z* distribution, and we can then invert the procedure we used for sampling *Z* to obtain a randomly imputed value for *X*.

Similar procedures can be used for ordered or unordered categorical data and also for non-normal continuous data and details are given in Goldstein et al (2009).

Software for carrying out the computations has been developed under the auspices of an ESRC project REALCOM, and details together with software can be found at http://www.cmm.bristol.ac.uk/research/Realcom/index.shtml. In release 2.1 of MLwiN Rasbash et al., (2008) it is possible to utilise the REALCOM extensions straightforwardly from within MLwiN itself by specifying the appropriate features of the MOI.

Note that so far we have made no distinction between type of non-response (refusal, non contact etc.), and that is equivalent to assuming that the relationship (as expressed in the parameters of our imputation/prediction model) is the same for different types. If this is felt to be unreasonable then we can allow for this in the imputation model. This can be done by including auxiliary variables associated with different types of non-response.

Care needs to be taken to ensure that the conditioning for the missing values is adequate. The advantage of the MI approach is that auxiliary variables not in the model of interest can be conditioned upon at the imputation stage. An alternative to MI, known as double robustness estimation (see Carpenter et al., 2006) uses weights based upon estimated missingness probabilities and may be useful in cases where the pattern of missingness is relatively simple (e.g. confined to a single variable), but difficult to implement in the general case. It has the theoretical advantage that, if either (i) a correct model for the probability of being missing is specified, or (ii) a correct model for just the conditional mean of the missing given observed data is specified, it then provides consistent and nearly efficient estimates. Thus only one of the two criteria needs to be satisfied. In practice, however, if a correct model is available for the probability of being missing (i.e. the variables responsible for missingness can be measured) then the relevant variables can also be incorporated into the MI process.

An example

We use an educational data set of measurements on 4059 students in 65 schools in London who have test scores and other measurements made at two occasions, at the end of primary schooling and prior to starting at secondary school at the age of 11 years (year 6), and at age 16 (year 11). In particular, at age 11 we have a reading test score – the London Reading Test (LRT) and at 16 an average examination score (EXAMSCORE) derived from grades obtained in the General Certificate of Secondary Examination (GCSE). Full details of the dataset can be found in Goldstein et al (1993).

To illustrate the imputation procedure we fit a simple 2-level model where EXAMSCORE (Y) is related to LRT ($\boldsymbol{X}_1$) and gender ($\boldsymbol{X}_2$) as follows

$$y_{ij} = \beta_0 + \beta_1 x_{1ij} + \beta_2 x_{2ij} + u_j + e_{ij}, \quad u_j \sim N(0,\sigma_u^2), \quad e_{ij} \sim N(0,\sigma_e^2) \qquad (7)$$

We first fit the model to the full data set and we will then drop data values and examine different procedures for handling the resulting dataset that includes missing values.

Table 1 gives the maximum likelihood parameter estimates for the model fitted to the full dataset.

Table 1. Exam score related to gender and reading test score.
Two level variance components model (7). Maximum likelihood estimates.

Parameter	**Estimate** (standard error)
Intercept	-0.095 (0.043)
Reading test	0.560 (0.012)
Gender (girl-boy)	0.171 (0.033)
Level 2 variance	0.088 (0.017)
Level 1 variance	0.562 (0.013)

We see from these results that girls do significantly better than boys after adjusting for initial reading test score. In this sense we can infer that girls make more progress between ages 11 and 16 than boys.

We also have a measure of verbal reasoning ability at age 11 for these students that is associated with progress. There are 3 categories that were originally defined to comprise the lowest 25% of ability scores, the next 50% and the highest 25%. The middle category in our sample contains 58% of the students and for these we assume that a random 50% drop out and for these we set the EXAMSCORE to be missing. The original values are retained for use in the imputation model as explained below.

In addition we randomly set a third of the LRT values to be missing. Altogether 53% of the student records have at least one missing value. The resulting estimates using 'listwise deletion' of all such records gives the results in
Table 2.

Table 2. Exam score related to gender and reading test score with listwise deletion of pupils with any missing data. Two level variance components model (7). Maximum likelihood estimates.

Parameter	**Estimate** (standard error)
Intercept	-0.041 (0.052)
Reading test	0.576 (0.017)
Gender (girl-boy)	0.125 (0.047)
Level 2 variance	0.103 (0.022)
Level 1 variance	0.571 (0.019)

We see that, the girl – boy difference has decreased, the LRT coefficient has increased and both variances have increased, with increases in all the standard errors resulting from the smaller sample size.

We now carry out an imputation analysis where, in the imputation model, the exam score and LRT are responses, since these contain missing data, and we condition both on gender and the (known) verbal reasoning group with the results shown in Table 3. The verbal reasoning group variable is here treated as an auxiliary variable known to be related to the propensity to be missing and also to the LRT score and EXAMSCORE, and is available both for those with full data and those with missing data. Thus, while all the variables in the MOI are used in the imputation model, the additional use of the verbal reasoning group is to correct for the bias we have (artificially) introduced which depends on the verbal reasoning group value. We note that our analysis is for illustration purposes only. If we wished to demonstrate how MI in general recaptures the original parameter estimates we would need to carry out a full set of simulations using multiply generated datasets, rather than just one illustrative analysis as here.

For the MCMC estimation we have used a 500 burn in with 10000 iterations, sampling every 500 to give 20 completed data sets. We see that the final parameter estimate for gender in Table 3 (below) is rather closer to the original value than in Table 2, and the reading test coefficient and the variance estimates in Tables 1, 2 & 3 vary, reflecting the sampling variability associated with a single randomly simulated set of missing values. It is, however, the reduction in the standard errors that is most noticeable, showing that we have gained in precision.

Table 3. Exam score related to gender and reading test score. Two level variance components model (7). Multiple imputation estimates with 20 completed datasets.

Parameter	**Estimate (standard error)**
Intercept	-0.077 (0.049)
Reading test	0.544 (0.017)
Gender (girl-boy)	0.164 (0.038)
Level 2 variance	0.112 (0.022)
Level 1 variance	0.572 (0.016)

Discussion

We have described a model-based procedure for handling quite general patterns of missingness and attrition in longitudinal data. We have used recent developments that combine existing multiple imputation techniques with procedures for transforming data to an underlying multivariate normal distribution. By considering attrition as a special case of missingness, and by assuming that auxiliary variables are available, we can set up an imputation procedure that will deal simultaneously with both attrition and item missingness. Since this procedure utilises all the available data it can be expected to provide maximum efficiency, and we have illustrated the efficiency gain with a simple example.

In longitudinal data we almost always have auxiliary data that can be conditioned upon and which is collected at the first measurement occasion. In addition, other data such as interviewer characteristics may also be available. This suggests that particular attention should be given to collecting auxiliary data that may be potentially associated with the propensity to be lost to a study, even if it is not intended to use such data in the substantive analyses. In terms of study resource allocation it could even be more efficient to devote resources to the collection of such data at the expense of attempts to secure repeated cooperation, at least where such attempts have low chances of success.

In the case of attrition, since the imputation procedure is Bayesian, we can also envisage the incorporation of prior information about missing data. For each individual, for missing items, we may have a prior distribution for the unknown values and this can also be incorporated into the imputation procedure. Such prior information could come from data that have been linked, for example from administrative records. Goldstein et al (2009) discuss this under the heading of 'partially observed data'. This can be used also with those individuals suffering attrition where such information might also come from linked data sets or be available from sources that have not been incorporated already into the auxiliary variables, such as interviewer observations. In the case of attrition a possible, less direct, alternative is to formulate a prior distribution for the imputation model parameters for each non-respondent and then combine this with the data using the current respondents' parameter values (see Rubin, 2004).

One of the problems with the techniques we have explored is that they tend to be computationally time-consuming. Nevertheless, since, in principle, we need carry out the imputation step just once for all the variables that we will be using in all our analyses, this will be less of a problem since we will simply reuse the same set of completed data sets.

The analyses in this paper were carried out, as described earlier, using multiple imputation commands newly available in MLwiN V2.10 (Rasbash et al., 2008) linked to REALCOM (Goldstein et al., 2008).

References

Carpenter, J. R., M. G. Kenward, et al. (2006). "A comparison of multiple imputation and doubly robust estimation for analyses with missing data." *Journal of the Royal Statistical Society, Series A* **169**: 571-584.

W. R. Gilks, S. Richardson and D. J. Spiegelhalter (1996). *Markov chain monte carlo in practice*. London, Chapman and Hall.

Goldstein, H., J. Rasbash, et al. (1993). "A multilevel analysis of school examination results." *Oxford Review of Education* **19**: 425-433.

Goldstein, H., J. Carpenter, et al. (2009). "Multilevel Models with multivariate mixed response types." *Statistical Modelling*. (To appear).

Goldstein, H., Steele., F., Rasbash, J., and Charlton, C. (2008). REALCOM: methodology for realistically complex multilevel modelling. Centre for Multilevel Modelling, University of Bristol. http://www.cmm.bristol.ac.uk

Hawkes, D. and I. Plewis (2006). "Modelling non-response in the National Child Development Study." *Journal of the Royal Statistical Society*, *A.* **169**: 479-492.

Heckman, J. (1976). "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables," *Annals of Economic and Social Measurement,* (December 1976).

R. J. A. Little and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. New York, Wiley.

Nathan, G. (1983). A simulation comparison of estimators for a regression coefficient under differential non- response. *Communications in Statistics: Theory and Methods*, 11, 645-659.

Nathan, G. and D. Holt (1980). The effect of survey design on regression analysis. *Journal of the Royal Statistical Society* B, vol. 42, pp. 377-386.

Pfeffermann, D. (2001). *Multilevel modelling under informative probability sampling*. 53rd Session of International Statistical Institute.

Rasbash, J., Browne, W., Cameron, B., and Charlton, C. (2008). MLwiN 2.10. Software for multilevel modelling. Centre for Multilevel Modelling, Bristol.

Rubin, D. B. (1987). Multiple imputation for non-response in surveys. Chichester, Wiley.

Rubin, D. B. (2004). "The design of a general and flexible system for handling non-response in sample surveys." *American Statistician* 58: 298-302.

Schouten, B. and De Nooij, G. (2005). *Non-response adjustment using classification trees.* Voorburg, Statistics Netherlands.