Handling attrition and non-response in longitudinal data with an application to a study of Australian youth

Joy Cumming   Australian Catholic University
Joy.cumming@acu.edu.au
Harvey Goldstein   Australian Catholic University and University of Bristol, UK

Abstract

A standard concern with long term longitudinal studies is that of attrition over time. Together with initial non-response this typically leads to biased model estimates unless a suitable form of adjustment is carried out. The standard approach to this has been to compute weights based upon the propensity to respond and to drop out and then carry out weighted analyses to compensate for response bias. In the present paper we argue that this approach is statistically inefficient, because it drops incomplete data records, is inflexible, and in practice gives rise to undue complexity involving a proliferation of weighting systems for different analyses. Instead we set out an alternative approach that makes use of recently developed imputation procedures for handling missing data and show how this can be used to improve the quality of the statistical analysis. An example analysis is given using the Longitudinal Study of Australian Youth (LSAY).

Keywords

Longitudinal data, life course, further education, attrition, missing data, imputation, weighting

Introduction

Attrition in longitudinal studies is typically viewed 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. For those analyses that utilise data at more than one occasion, the use of only those individuals with data at all such occasions in the analysis will result in a loss of efficiency. 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, as well as initial non-response at wave 1. Further, this approach provides efficient estimates based upon a one-pass Markov Chain Monte Carlo (MCMC) algorithm that avoids the creation of multiply imputed data sets.

Data

These procedures are applied to a long term study the "Longitudinal Study of Australian Youth' (LSAY) with up to 12 waves of data collection. This is a study that is tracking the pathways of young Australians as they move from school to further study, work and other destinations. Data are collected on variables related to education, training, work, financial matters, health, social activities and attitudes. LSAY started in 1995 by sampling year nine students, average age 14.5 years, in Australian secondary schools and following them up every year on a further 11 occasions (LSAY, 2013). In the present paper we analyse the second cohort commencing in 1998 using data up to wave 6 in 2003. The principal aim of the paper is to explore ways in which the richness of the dataset can be utilised efficiently, in the light of extensive non-random sample attrition over time. We concentrate on a single outcome, whether or not the respondent is in part time or full time study and relate this to a number of characteristics. A series of reports by Sheldon Rothman presents basic tabulations for outcomes up to wave 6 (Rothman, 2005).

Australia has six states and two territories with very unequal population sizes. The sample design used in LSAY is one where there was oversampling from some smaller states and territories leading to unequal student selection probabilities (table 1). Thus a set of sample design weights was derived so that inferences could be made to the actual Australian population. For every child in a school the weights are equal so that we can treat this as a single level weight in any weighted analyses. In addition the study provides weights that attempt to control for non-random attrition. A detailed discussion of how to derive such weights can be found in Plewis (2007).

Table 1 shows the characteristics of the sample, including the distribution by state. The initial sample consists of 296 schools.

Table 1: Characteristics of 1998 LSAY cohort

		Unweighted sample	All Australian schools
Average age of respondents 30 /6/1998 (yrs)		14.5	
Total no of respondents		14117	
State/territory (%)	New South Wales	24.0	32.8
	Victoria	20.9	23.5
	Queensland	22.0	20.1
	South Australia	8.8	7.8
	Western Australia	12.0	10.8
	Tasmania	5.1	2.7
	Northern Territory	3.3	0.9
	Australian Capital Territory	4.0	1.9
Sex (%)	Male	51.2	
	Female	48.2	
	Unknown	0.6	
Indigenous status (%)	Indigenous	3.1	
	Non-Indigenous	96.9	
Geographic region (%)	Metropolitan	60.0	
	Regional	22.4	
	Rural and remote	17.5	
Country of birth (%)	Australia	85.0	
	Other	15.0	
School sector (%)	Government	63.0	
	Catholic	22.1	
	Independent	14.9	

Source: LSAY - Longitudinal Surveys of Australian Youth, Y98 cohort to 2009, released April 2010, updated January 2011.

Since the corresponding sample weights apply at the state level, we can effectively eliminate the need for weights by explicitly fitting state (as a series of dummy variables) in our models. Only if we wish to provide country level estimates will we then need to marginalise over the state distributions, but we do not consider this at the present stage. In fact, the use of weights does not change any of our inferences very much. Ignoring weights and not fitting effects for state will provide inferences for a (hypothetical) population where the state population sizes are proportional to the chosen state sample sizes.

In all our subsequent models we have also studied whether there are interactions among the explanatory (predictor) variables and found little evidence for these, so that they will be omitted.

Table 2. Percentages of original sample of 14117 year 9 students in 1998 remaining at waves 2-6

Year (wave)	Per cent remaining
1999 (2)	65.8
2000 (3)	67.6
2001 (4)	62.2
2002 (5)	55.0
2006 (6)	48.9

As table 2 shows the attrition is considerable, especially in the first year of the study with just 49% remaining by wave 6, and it is clearly necessary to make adjustments for this. We note that attrition is not completely monotonic with some students returning to the study. This will not affect the estimation model.

The LSAY user guide (LSAY, 2013) also discusses deriving weights to compensate for non-random attrition over time in the cohort. Weighting, however, is not entirely satisfactory (Goldstein, 2009) since it will in general require a different set of weights to be computed for every different combination of waves entering any given (longitudinal) model and this is not practically very feasible. What LSAY itself provides is essentially only weights that correct the data at each wave separately. Thus these will allow adjustment for cross sectional analyses at each wave but not for longitudinal analyses. In any case these weights are only computed based on overall achievement (average of student outcomes on tests of mathematics and reading comprehension) and gender at wave 1. Plewis (2007) proposes an extension of this by computing a set of weights specific for each of a number of longitudinal analyses involving different combinations of occasions or waves and different sets of variables. Such approaches are intended to deal with the problem of biases that may arise from differential attrition. In the present paper we are additionally concerned with efficiency, and in particular the loss of efficiency that is implicit in such weighting procedures, as we explain below.

Modelling the probability of participation in further and higher education

As noted, our principal aim in this article is methodological, namely to explore an alternative approach to weighting to improve efficiency. A secondary aim is to explore factors that predict the probability of participating in further or higher education after school. To address the methodological issue, in the present analysis we are only considering two time points, so that we can use weights based upon the probability of remaining in the sample at wave 6 as described below. Inspection of the data shows that other variables are associated with attrition, as shown in table 3.

Table 3. Variables predicting the probability of remaining in the sample at wave 6. Two level model. Reference categories in brackets. Probit link function. 9027 out of 14117 cases used.

Parameter	Estimate	Standard error
Intercept	-0.617	0.050
Female (male)	0.111	0.030
Catholic (Government) school	0.179	0.050
Independent (Government) school	-0.019	0.059
Maths score year 9	0.023	0.004
Reading score year 9	0.030	0.004
Non-Australia country of birth of mother (Australia)	-0.116	0.031
SES ANU3 score father	0.0022	0.0007
Level 2 variance	0.060	0.009

We see that additionally to the test scores and gender, country of birth of mother, type of school and father's socioeconomic status (SES) are predictors of attrition. We shall be using these variables in subsequent analyses either fitting into the model or using in the imputation model to correct for attrition bias. Adding some other terms in this model such as home language (English/other) does not add to the prediction so we will use the above variables as standard.

One way to utilise the results in table 3 is to use the inverse predicted probabilities of inclusion as weights when modelling wave 6 outcomes in order to correct for any attrition bias. A problem with such an analysis is that it is based only upon the students who remain at wave 6 and hence does not increase efficiency, especially as the predictors in table 3 are also those used in our model of interest given in table 4.

The methodological approach we use as an alternative to weighting is a recently proposed extension of multiple imputation (Goldstein, Carpenter, & Browne, 2014). In longitudinal data Goldstein (2009) discusses procedures for handling attrition and item missing data in longitudinal studies and contrasts weighting with multiple imputation. Seaman and White (2014) discuss the use of inverse probability weighting to adjust for biases that may arise when values are missing not at random, where the weights are derived from a model that predicts the probability of an individual having no missing values in the model of interest. Both these latter two papers point out that the weighted analysis uses only individuals with complete cases. For this reason, apart from issues of bias, imputation is generally recognised to be more efficient and the statistically most satisfactory method for handling attrition and in fact any type of 'missing' data (Moodie, Delany, Lefevre, & Platt, 2008). Carpenter and Kenward (2012) also discuss this and in particular consider the question of bias reduction and increasing efficiency using a combination of multiple imputation and inverse probability weighting, doubly robust estimation, although in the present case there is little to be gained since the prediction of the weights utilises the same variables as are in the model of interest.

In the present paper we demonstrate the increase in efficiency from using imputation. A particular advantage of imputation is that, in addition to attrition, nonresponses to individual questions can also be incorporated straightforwardly (see Goldstein, 2009 for a discussion). In our final analyses we will compare imputation with a complete cases only model, with and without weighting. We shall be utilising a

recent development in imputation that obviates the need for multiply imputed datasets and allows quite general models with missing data, as described below. Software for this has been developed by the Centre for Multilevel Modelling in Bristol and specifically the STATJR software package (STATJR, 2015), Alternative imputation methods, notably 'chained equation' procedures, are based upon multiple imputed datasets and are unable properly to handle functions of variables having missing data, such as power and interaction terms (Goldstein et al., 2014).

The imputation algorithm

The method described by Goldstein et al. (2014) extends the standard joint model for multiple imputation procedure (Carpenter & Kenward, 2013) by obviating the need to produce a set of multiply imputed datasets and also allows interaction and polynomial terms in the model of interest. It is a one-pass method using a single MCMC algorithm and is fully Bayesian with a faster implementation in software. The following is a simple summary of the procedure, avoiding undue technicality.

For simplicity consider a single level model of interest where the response and explanatory variables have a joint normal distribution. If there were no missing data then this can be readily fitted using standard multiple regression, either via maximum likelihood or in a Bayesian model, for example with default diffuse prior distributions. Consider the simple regression, joint model

$$Y = \beta_0 + \beta_1 X + e_{Y|X}$$
$$X = \alpha_0 + e_X$$
$$\tag{1}$$

The first line of (1) is the usual regression model and the second line specifies a model for the explanatory variable X. The residual terms express the usual conditional distribution $Y|X$ and the distribution for X itself. Note that, unlike in standard linear regression where it is not needed, we have explicitly introduced a distribution for X, since this will be required when we have missing data in this predictor. Where there are missing values in the response the record is omitted, although where there are several responses in a multivariate model we would impute missing responses where other responses are present. It is assumed that $e_X, e_{Y|X}$ are independent.

We initiate a MCMC algorithm, with suitable starting values, that at each iteration uses a 'metropolis' step, to propose, for each missing value in X (or Y in the multivariate case), a value, which is used in (1) to update both lines if accepted using the implied likelihood values. Suitable starting values can be derived from the complete cases. For each missing value, in turn, the Metropolis step will accept a proposed value based upon a comparison of the joint likelihood for X and Y based respectively on the current and proposed value. Where there are several X variables, the second line of (1) will express a conditional relationship among the X variables for each one in turn.

Model (1), with several predictors, is readily generalised to several explanatory variables and to further hierarchical levels or cross classifications. If there are interactions or polynomial terms in the model of interest these are only present in the first line of (1) and so are easily incorporated. Where we have non-normally distributed variables we adopt a 'latent normal' transformation that introduces additional steps in the algorithm to sample from underlying normal distributions: for a binary variable this is the usual probit model, and is used in our models. Note that the requirement for compatibility (congeniality) of the imputation model and the model of interest in standard multiple imputation is automatically satisfied using this algorithm. One of the assumptions made in imputation models is that, conditionally on the variables included in the imputation component, any propensity to missingness is effectively random, the so called Missing at Random (MAR) assumption. We may also include auxiliary variables, not required in the model of interest, in the imputation component of the joint model where these are needed to ensure MAR. Full details are given by Goldstein et al. (2014), including choice of prior distributions.

Our subsequent analyses, therefore, include all the variables from table 3 associated with attrition.

Predicting education and training participation at wave 6

Our model of interest is a comparison between those in tertiary education currently studying part time or full time, and those not.

The model of interest is
$$g(\pi_{ij}) = X_{ij}\beta + u_j$$
$$\pi_{ij} = \int_{-\infty}^{X_{ij}\beta} \phi(t)dt$$
$$u_j \sim N(0, \sigma_u^2) \tag{2}$$

where π_{ij} is the probability that the ith student in the jth (1998) school is in tertiary education, and $\phi(t)$ is the standard normal distribution with g the probit link function, relating this probability to the set of covariates listed in table 4. The subscripts i, j index students and schools respectively. The probit function rather than the logit is convenient for modelling with missing data and allows interpretation of parameters on an underlying standard normal (N(0,1)) scale.

The imputation model can be written, excluding subscripts, as

$$X^* \sim MVN(\alpha Z, \Omega_X)$$

$$(3)$$

This is a two level multivariate normal model where the responses are the predictors, apart from the intercept and school type, Z, where there are no missing values, in table 4. For the maths and reading scores, these are already normalised. For the remaining binary variables the X^* are obtained

using a step in the algorithm that randomly samples from the underlying 'latent' normal distribution corresponding to the binary variable. The joint model is thus the combination of (2) and (3).

We have explored a number of analyses with different predictors and table 4 presents a final fitted model that contains a set of predictor variables that jointly predict propensity to remain studying. Gender is included for completeness. We looked at student's country of birth, father's country of birth and also student location in urban/semi urban/rural region and these had small effects and have been omitted. The continuous ANU3 score is used for SES status as recommended by Marks (1999). In addition to variables shown in table 3 on probability of remaining in the sample at wave 6, two additional variables, Home language not English (English), and mother's SES, are in the model of interest.

Table 4. Full or part time study in 2003 by student, parental, environmental and school characteristics in year 9, 1998. MCMC estimates with burnin=5,000, iterations=25,000. Listwise deletion with 3407 out of 14117 cases used. Two level model. Reference categories in brackets. Single level estimates in brackets.

Parameter	Estimate	Standard error
Intercept	-0.938 (-0.952)	0.098 (0.088)
Female (male)	0.052 (0.053)	0.047 (0.046)
Catholic (Government) school	0.182 (0.174)	0.062 (0.052)
Independent (Government) school	0.236 (0.222)	0.080 (0.070)
Maths score year 9	0.049 (0.048)	0.006 (0.006)
Reading score year 9	0.024 (0.023)	0.006 (0.006)
Non-Australia country of birth of mother (Australia)	0.185 (0.181)	0.059 (0.058)
Home language not English (English)	0.412 (0.408)	0.119 (0.115)
SES ANU3 score father	0.0049 (0.0041)	0.0011 (0.0011)
SES ANU3 score mother	0.0025 (0.0027)	0.0014 (0.0013)
Level 2 variance	0.034	0.016

The results in the above model are not unexpected. Country of birth of mother outside Australia is associated with a greater propensity to remain studying and this could be explored further in terms of actual country of origin.

Being in Catholic or independent schools likewise is positively associated with increased propensity to remain studying, as is having a main language in the home that is not English, and having high maternal and paternal occupational status.

We see that the school level variance is 0.034 with an equivalent standard deviation of 0.18. This is on a standard normal scale and is relatively small. Omitting the school level we have also fitted a single level model whose estimates are given in brackets and it can be seen that the inferences are almost identical. Thus the actual school attended in 1998 appears to have little effect on propensity to be studying five years later. The overall percentage studying is 61% and a variance component analysis, just fitting an intercept term in the model gives us an estimate for the between school variance (standard deviation) of 0.121 (0.35) so we see that most of the school effect is accounted for by the predictor variables.

The main methodological problem with this analysis is that the effective sample size is only 3407 out of a possible 14117 (24%) of all cases. In particular 33% of father's and 45% of mother's SES are missing (table 5).

Table 5. Percentage of missing values for selected variables

Studying full or part time at wave 6	51.1
Sex	0.6
School type	0
Maths score year 9	2.3
Reading score year 9	2.5
Country of birth of mother	4.7
Mother's SES ANU3 score	44.7
Father's SES ANU3 score	33.5

Cases with any missing data are omitted from the analysis. Thus, even if we had appropriate weights that corrected for any biases we would still have a very reduced efficiency due to the fact that only 24% of the sample cases can be used.

We demonstrate this in table 6 by repeating the analysis in table 4 with one that uses the inverse probability weights derived from table 3 where we predict the probability of remaining in the sample until wave 6. Bayesian models do not allow us to incorporate weights so that the estimates are second order quasi-likelihood estimates (Goldstein, 2011) which do approximate very closely, in the unweighted case, to the Bayesian MCMC estimates.

Table 6 Full or part time study in 2003 by student, parental, environmental and school characteristics in year 9, 1998. Second order PQL (Quasi-likelihood) estimates. Listwise deletion with 3,407 out of 14,117 cases used. Two level model. Reference categories in brackets. Standard errors in brackets

Parameter	Unweighted estimates	Inverse probability weighted estimates
Intercept	-0.942(0.093)	-0.927(0.095)
Female (male)	0.054(0.047)	0.040(0.047)
Catholic (Government) school	0.184(0.057)	0.209(0.057)
Independent (Government) school	0.232(0.090)	0.230(0.090)
Maths score Year 9	0.050(0.006)	0.050(0.006)
Reading score Year 9	0.024(0.006)	0.024(0.006)
Non-Australia country of birth of mother (Australia)	0.184(0.060)	0.177(0.061)
Home language not English (English)	0.412(0.120)	0.409(0.121)
SES ANU3 score father	0.0050(0.0011)	0.0048(0.0011)
SES ANU3 score mother	0.0026(0.0013)	0.0025(0.0013)
Level 2 variance	0.034(0.014)	0.036(0.014)

We see that the weighted and unweighted point estimates are very similar, with virtually identical standard errors.

In the next model, in table 7, we have used our imputation procedure to both adjust for bias and utilise all the information in a statistically efficient manner.

Incorporating missing data using imputation

Out of the total sample, 51% are not present at wave 6 and hence missing the response in the model of interest, 'in full or part time study', which is available for all those remaining in the study. In the present case, even though we fit only the sample of 6901 who remain at wave 6, we do actually obtain results where the parameter estimates and standard errors are essentially the same as fitting the full data, that is, imputing responses for those not present at wave 6. This is to be expected since where the response, in a model with covariates, is missing, there is no further information available from such records.

Table 7. Full or part time study in 2003 by student, parental, environmental and school characteristics in year 9, 1998. Missing data model. MCMC estimates with burnin=500, iterations=2,500. Two level model. Reference categories in brackets. Single level estimates in brackets. Sample size=6,901.

Parameter	Estimate	Standard error
Intercept	-1.040 (-0.951)	0.065 (0.057)
Female (male)	0.058 (0.047)	0.034 (0.032)
Catholic (Government) school	0.210 (0.159)	0.047 (0.040)
Independent (Government) school	0.255 (0.214)	0.057 (0.049)
Maths score year 9	0.049 (0.051)	0.004 (0.004)
Reading score year 9	0.028 (0.026)	0.004 (0.004)
Non-Australia country of birth of mother (Australia)	0.179 (0.186)	0.040 (0.040)
Home language not English (English)	0.457 (0.497)	0.089 (0.072)
SES ANU3 score father	0.0046 (0.0045)	0.0009 (0.0009)
SES ANU3 score mother	0.0025 (0.0018)	0.0010 (0.0006)
Level 2 variance	0.036	0.009

While there are some small differences in the parameters estimates, notably for school type, the reading score, home language and father's SES, the main difference lies in the considerable reduction in standard errors. These reductions are of an order up to about 50%, reflecting the efficiency gain from the imputation-based modelling. Fitting the imputation model is not onerous, the two level model took approximately 80 minutes to fit, on a 2.4 Ghz PC running windows 7.

Discussion

The Longitudinal Study of Australian Youth represents a major investment in data collection over a 12 year period to track the fortunes of students from school year 9 annually into early adulthood. In the present paper we have explored a very limited set of variables looking at antecedents of whether in the years after leaving school, the students are still in full or part time education. We find that test scores, parental country of birth, home language and SES all affect the propensity to be studying.

The particular focus of the paper, however, lies in its methodology. By wave 6 half the students who were sampled originally in wave 1 had left the study, and the propensity to leave was not random. We have shown how the use of complete cases, whether using inverse probability weights or not, results in estimates that are considerably less efficient than a fully imputation based approach. In the present case bias appears not to be an issue since neither the weighted analysis nor the imputation analysis lead to very different estimates from the unweighted complete cases analysis.

More generally, our analysis illustrates the usefulness of a missing data approach to both missing data item values and dropout. When there are a relatively small number of complete cases it is not efficient to base an analysis solely on such cases. Utilising weights in a complete cases analysis may provide a means of correcting potential biases, but where such weights, as in the case of attrition,

are derived from observed data at wave 1 (or later) we show that greater simplicity and greater statistical efficiency can be obtained by an imputation based approach that incorporates all the observed data in a single model.

The assumption of missingness at random is an important one. To deal with non-random attrition in a weighting approach we would normally seek to satisfy this by incorporating 'auxiliary' variables when modelling the propensity. In an imputation approach we would incorporate such auxiliary variables in the imputation component of the model. In our analysis all the variables that are significant predictors of dropout by wave six are also in the model of interest.

In some longitudinal studies the data analyst may be confronted with data that contain design or non-response weights of unknown provenance. In such cases, unlike the present, these weights will generally need to be incorporated into both the imputation component and that for the model of interest. Within a fully Bayesian framework it is not clear how to do this, and this problem is currently the focus of further research. Carpenter and Kenward (2012) provide some guidance.

Acknowledgements

References

Carpenter, J., & Kenward, M. (2012). *Multiple imputation and its application*. Chichester: John Wiley & Sons.

Carpenter, J., Goldstein, H., & Kenward, M. (2011). REALCOM-IMPUTE software for multilevel multiple imputation with mixed response types. *J. Statistical Software*, *45*(5)*,* 1–14. http://dx.doi.org/10.18637/jss.v045.i05

Goldstein (2009). Handling attrition and non-response in longitudinal data. *International Journal of Longitudinal and Life Course Studies, 1,* 63–72.

Goldstein, H. (2011). Multilevel statistical models (Fourth ed.). Chichester, England: 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* (Statistics in Society). http://dx.doi.org/10.1111/rssa.12022

Goldstein, H., Carpenter, J., Kenward, M., & Levin, K. (2009). Multilevel models with multivariate mixed response types. *Statistical Modelling*, *9*(3), 173–197. http://dx.doi.org/10.1177/1471082X0800900301

LSAY. (2013). *Longitudinal Surveys of Australian Youth.* 1998 cohort user guide. Accessed 12th February 2015 from http://www.lsay.edu.au/publications/2199.html

Marks, G. (1999). *The measurement of socioeconomic status and social class in the LSAY project [Technical Report #14].* Melbourne: ACER.

Moodie, E.F.M., Delany, J.A.C., Lefevre, G., & Platt, R.E. (2008). Missing confounding data in marginal structural models: a comparison of inverse probability weighting and multiple imputation. *International Journal of Biostatistics*, *4*, 1–23. http://dx.doi.org/10.2202/1557-4679.1106

Plewis, I. (2007). Non-response in a birth cohort study: the case of the Millennium Cohort Study. *International Journal of Social Research Methodology*, *10*, 325–334. http://dx.doi.org/10.1080/13645570701676955

Rothman, S. (2005). Accessed 12th February 2015 from http://www.lsay.edu.au/publications/1844.html

Seaman, S., & White, I. (2014). Inverse probability weighting with missing predictors of treatment assignment or missingness. *Communications in Statistics—Theory and Methods*, *43*, 3499–3515. http://dx.doi.org/10.1080/03610926.2012.700371

STATJR. (2015). Centre for Multilevel Modelling, University of Bristol. Accessed 12th February 2015 from http://www.bristol.ac.uk/cmm/software/statjr/