Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2005 : Methodological Challenges for Future Information needs

2005

Statistics Canada
Canada

Statistique
Canada

Canada

Proceedings of Statistics Canada Symposium 2005
Methodological Challenges for Future Information Needs

DEALING WITH MISSING SURVEY DATA IN LONGITUDINAL ANALYSIS

Robert M. Baskin[1]

ABSTRACT

This paper reviews techniques for dealing with missing data from complex surveys when conducting longitudinal analysis. In this context, longitudinal data will refer to repeated observations on the same unit of measurement. In addition to incurring the same types of missingness as cross sectional data, longitudinal observations also suffer from drop out missingness. For the purpose of analyzing longitudinal data, random effects models are most often used to account for the longitudinal nature of the data. However, there are difficulties in incorporating the complex design with typical multi-level models that are used in this type of longitudinal analysis, especially in the presence of dropout missingness. In this situation, drop out missingness is often handled by serially modelling the 'attrition samples' with or without the use of longitudinal weights. An alternative is to fill in the missing data by imputation. In the presence of imputation, standard variance estimators will underestimate the variance. Thus, methods for accounting for imputation variance such as multiple imputation of dropout missing data should be considered.

KEY WORDS: attrition sample; complex survey; missing at random; multi-level models; multiple imputation

1. INTRODUCTION

This paper reviews techniques for dealing with missing data from complex surveys when conducting longitudinal analysis. This is a very broad topic and in order to limit the scope of discussion, there will be no discussion of time series or of modelling techniques. Only selected techniques for dealing with missing data in a longitudinal setting from a complex survey will be reviewed. However, some specific longitudinal models are used as illustrative examples. This paper is approached from the point of view that the analyst knows what kind of analysis is appropriate but may need help with the issue of missing data.

Since the subject is missing survey data, it is assumed that the reader has some knowledge of estimation methodology in surveys from a standard reference such as Cochran (1977). This methodology is often referred to as finite population methodology but here it will typically be referred to as *design based* methodology. In contrast, what is sometimes referred to as *mainstream statistics* or infinite population methodology will be referenced as *model based* or model dependent methodology. Also, the term *likelihood inference* will be employed to denote inference based on a likelihood function. Finally, a combination of the two methodologies, as presented in Sarndal, Swensen, and Wretman (1992), will be called *model assisted*, which is distinct from the *superpopulation approach* referenced, for example, in Ghosh and Meeden (1997).

The pitfalls of ignoring the sample design in modeling survey data are well documented, but the reader is referred to Skinner, Holt and Smith (1987) or Brogan (1998) for a review of issues such as model robustness and unbiased estimation of variance from a complex survey. For an introduction to the use of sample weights in modeling survey data the reader is referred to Korn and Graubard (1999) and for a higher level discussion of weighting and survey data the reader is referred to Pfeffermann (1993). The pitfalls of ignoring the missingness, especially of complete case analysis, can be found in Korn and Graubard (1999) or in Little and Rubin (2002). Little and Rubin (2002) also discuss the concepts of missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR), which will be occasionally mentioned in this paper.

[1] Agency for Healthcare Research and Quality, 540 Gaither Road, Rockville, MD 20850. The views expressed in this paper are those of the author and no official endorsement by the Department of Health and Human Services or the Agency for Healthcare Research and Quality is intended or should be inferred.

The discussion will proceed along the following lines. First, purely design-based methods for dealing with missing data will be discussed. Second, model dependent methods that incorporate some aspect of the design-based ideas will be presented. Next, Bayesian methods will be presented. After that the methods involving imputation will be discussed. Finally, an example of modelling missing data will be given. Much of the material can be found in Skinner, Holt, and Smith (chapters 12-14, 1989) or in Chambers and Skinner (2003).

2. DESIGN BASED METHODS

A more detailed discussion of the material in this section can be found in Skinner, Holt, and Smith (1989, chapter 14). The only approach for purely design-based methodology, other than to fill in missing data with imputation, is to use reweighting. In cross-sectional analysis the preferred approaches are to reweight for unit missingness and to use imputation for item missingness. However, in longitudinal analysis a unit may be missing in one time period and not in another. The situation referred to as *dropout missingness*, in which a unit is present for early rounds of collection and later drops out, is often the greatest concern. It is often believed that the dropouts are different in some systematic way from the full-case units available for analysis. However, other patterns of missingness are possible. A new unit can enter the sample after early rounds of collection. This situation, sometimes referred to as a *statistical birth* or *drop-in missingness,* is usually not considered a major source of bias. Often, other patterns of missingness are assumed to be missing completely at random.

In order to deal with longitudinal analysis in a purely design based manner, the set of longitudinal observations for each individual unit is treated as a vector of observations on the unit and a covariance matrix is estimated using weighting methodology. If each unit has no missing observations this method can be used directly for linear models or in conjunction with linearization to estimate many non-linear models. In the presence of dropout missingness, if wave weights are available for a survey, the following method can be used to model the attrition samples over time. For each time period, say *t*, the units, which have not dropped out up to time t, are modeled using the wave weights to estimate the covariance matrix of the observations for the units. This produces a set of models for each time period *t*, which attempts to account for the dropout missingness.

While this method retains many advantages of design-based methodology, there are drawbacks to the method. One difficulty is that only one pattern of missingness, typically dropout missingness, is accounted for in any analysis. Another problem that often prevents the use of this method is that not all surveys have wave weights for every round of collection. Perhaps the biggest issue here is that any comparison of the models is on an ad hoc basis.

3. RANDOM EFFECTS MODELS

Random effects models have become the de facto standard for performing longitudinal analysis in mainstream statistics. Verbeke and Molenberghs (2000), for example, describe the use of random effects models for longitudinal analysis with missing data. Random effects models have the property that valid inference still results if the data are missing at random whereas other models may require the assumption of missing completely at random to ensure valid inference. But the approach of Verbeke and Molenberghs (2000), as is typical with mainstream statistical texts, does not address the issues of modeling data from a complex survey.

Building longitudinal random effects models that account for a complex survey design is a difficult topic. It might be argued from a technical point of view that a design-based approach cannot produce a random effects model. However, it is possible to produce a *design consistent* random effects model that is likelihood based. The term design consistent used here refers to a model that produces estimates that may not be design unbiased but are consistent in the sense of asymptotically design unbiased. This is the approach for cross sectional data proposed in Pfeffermann and Lavange (1998) and modified in Pfeffermann, et al, (1998). Skinner and Holmes (2003) extend the previous works to include random effects models for longitudinal data. Presumably, because the random effects models in Skinner and Holmes are likelihood based they would retain the property that the models are valid for data that is missing at random as opposed to missing completely at random. This sequence of papers takes an evolutionary path starting from Generalized Least Squares (GLS) to probability weighted GLS to Iterative

Generalized Least Squares (IGLS) to probability weighted Iterative Generalized Least Squares (PWIGLS). References for each step can be found in Skinner and Holmes (2003).

A sketch of the ideas behind PWIGLS is as follows. If we want to build a random effects model on our data, then the model will have regression parameters, denoted by $\boldsymbol{\beta}$, and variance components denoted by $\boldsymbol{\theta}$. If we start by ignoring any design weights and build a model using IGLS then we make some initial guess at $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_0$. Using the value of $\boldsymbol{\theta}_0$, we produce an estimate of $\boldsymbol{\beta}$, say $\boldsymbol{\beta}_1$ using generalized least squares. Now, treating $\boldsymbol{\beta}_1$ as a constant, we go back and produce a new estimate of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}_1$. This process is iterated, updating each set of parameters consecutively until convergence is achieved.

At this point, the method looks promising because it is a likelihood based approach, however it doesn't account for the sample weights. The next step is to incorporate the sample weights. Pfeffermann, et al, (1998) modified the IGLS method so that sample weights are included in the following way. First the weights are normalized in a clever way as described in Pfeffermann, et al, (1998). Then, the normalized weights are included in the IGLS random effects model as a covariate but <u>not</u> as design weights (p-weights). This produces point estimates of the $\boldsymbol{\beta}$ that are simultaneously model unbiased and design consistent. Unfortunately, if the design is not equally weighted (self-weighting) this modification produces estimates of variance that are neither model consistent nor design consistent, so one further modification is required. At this point, in the second step of the IGLS approach, a design consistent estimator of variance needs to be employed. This has the unfortunate effect that standard software for IGLS has to be stopped at each iteration and a new variance estimate inserted into the process or that the iteration has to done by hand. Nevertheless, this method does provide a way of fitting random effects models for complex survey data that is likelihood based but still design consistent.

There is one further modification in Skinner and Holmes (2003) that accounts for the longitudinal aspect of data. The adjustment is made to account for serially correlated data in the longitudinal model. Altogether this produces models with the desired theoretical properties but actually fitting the model is quite tedious.

One final technical aspect of the method that needs clarification is how to determine the original estimate of variance. Pfeffermann, et al, (1998) suggest starting with a design consistent estimator of the variance components $\boldsymbol{\theta}_0$ conditional on a design consistent estimate of $\boldsymbol{\beta}$ from a standard design based model.

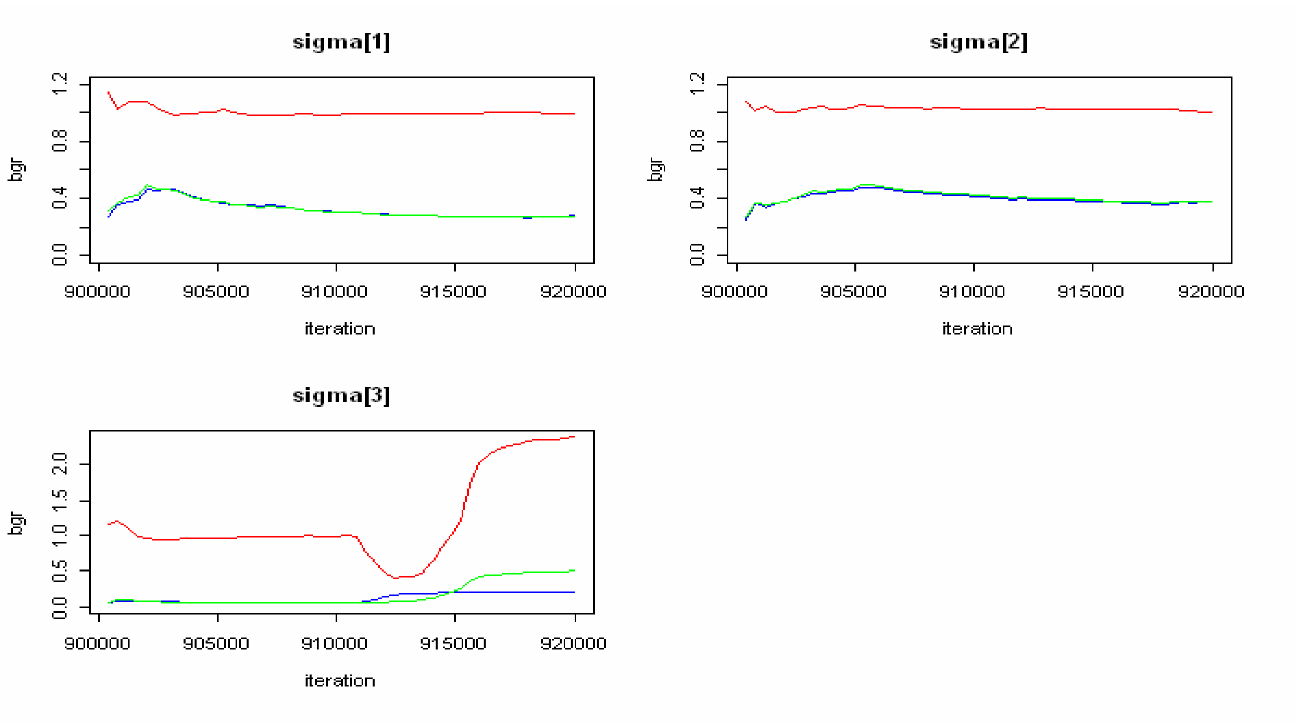
4. BAYESIAN MODELS

Bayesian methods are increasing in popularity and have nice theoretical properties. The increase in popularity of Bayesian methods is largely due to the use of Monte Carlo methods such as Gibbs sampling. These computer intensive methods for Bayesian models are available in several software packages and in particular the package WinBUGS. The open source version, OpenBUGS, is becoming very popular. The R package, R Development Core Team (2005), allows direct access to OpenBUGS with many diagnostic tools easily accessible.

As mentioned, Bayesian methods have nice theoretical properties. An introduction to Bayesian methods in a finite population setting can be found in Ghosh and Meeden (1997). The methods in Ghosh and Meeden (1997) correspond to superpopulation models in the sense that both finite population sampling and a random model for the observed values is accounted for. These models for the observed values have similar properties to random effects models and thus the data are assumed to be missing at random if the missingness is not modeled directly in the model. There is one distinction between the assumptions on Bayesian models and frequentist models. For valid inference from a random effects model, the analyst must assume that the data are missing at random and that any model parameters are distinct from any non-response mechanism parameters. However, for valid inference from a Bayesian model, in addition to the missing at random assumption, the analyst must assume that model parameters and non-response mechanism parameters have *independent priors*. This is a stronger assumption than having distinct parameters and if the analyst doesn't believe the independent prior assumption then valid inference from a Bayesian model would require modelling the data assuming it is not missing at random. This also raises the issue of prior distributions for the parameters. A discussion of prior distributions and any controversy surrounding such distributions is far beyond the scope of this paper. For a discussion of prior distributions from a Bayesian perspective the reader is referred to Bernardo and Smith (2000).

Assuming the analyst wants to undertake a Bayesian analysis, software now exists that makes the process of building a Bayesian model much easier than in the past. However, besides the usual problems of model checking, the analyst needs to be aware that the Monte Carlo approach to model building in packages such as BRugs, the R interface to OpenBUGS, requires checking of convergence. BRugs has built in diagnostic tools for checking the convergence.

As a specific example, consider the following longitudinal analysis using Self Reported Health Status (SRHS) from the 2002 Medical Expenditure Panel Survey (MEPS). MEPS is a complex national probability sample survey sponsored by the Agency for Healthcare Research and Quality. MEPS is designed to provide nationally representative estimates of health care use, expenditures, sources of payment, and insurance coverage for the U.S. civilian noninstitutionalized population. MEPS consists of a family of three interrelated surveys with the Household Component (HC) as the core survey. MEPS-HC is an overlapping panel survey and eligible respondents remain in the survey for five consecutive rounds of data collection. The fact that MEPS-HC is collected in ongoing panels provides the advantage of performing longitudinal analysis on data but also introduces the challenges of addressing missing data issues in a longitudinal dataset. SRHS is collected in five consecutive rounds, over two years, and has both dropouts and drop-ins as well as other patterns of missingness. A simple mean model with and without a time covariate was fit to this data using non-informative priors. Using BRugs, the model was allowed to run for 900,000 warm up iterations before the diagnostic tools were turned on. For a simple mean model this would seem to be an extreme number of iterations but this large number was used to stress the point. With diagnostics in place the model was run for another 20,000 iterations. At the end of the 20,000 iterations, the diagnostics were examined for convergence. As is typical, the diagnostic tools indicated convergence for all of the mean parameters in the model. However, the diagnostic tools for the variance parameters indicated a problem with convergence of the variance parameters. The Brooks-Gelman-Rubin statistic (BGR) is displayed for the variance components in Figure One. It has three diagnostic lines. The top line should converge to 1.0 and the bottom two lines should converge together to 0.4. As can be seen from the diagnostic graph below, the third variance component initially appears to be converging but after 10,000 iterations, the top line suddenly moves away from the ideal point of 1.0 and diverges wildly. If the analyst had only monitored the graph for 10,000 iterations the analyst might be fooled into believing that convergence had been achieved. Thus, it is necessary to be extremely cautious when diagnosing convergence of Monte Carlo methods for Bayesian models. In particular, the variance components are notorious for having problems. For recent work on variance parameters in Bayesian models see, for example, Gelman (2005).



5. IMPUTATION METHODS

Imputation is a statistical technique for filling in missing data and is commonly used with item non-response in survey data. Imputation has many advantages but also has disadvantages. Many different imputation procedures have been developed but there are two general classes or types of imputation procedures that have different properties. Deterministic imputation such as mean imputation and some forms of nearest neighbor imputation have no random component to the imputation process. Deterministic imputation may be response distribution unbiased but will attenuate covariance structure in the data. This may not be a problem for some simplistic forms of analysis but in longitudinal analysis relationships among the variables and covariance structure is typically important to preserve.

A second class of imputation procedures involves a random component in the imputation process such as hot deck and some forms of nearest neighbor imputation. This type of imputation procedure can be response distribution unbiased and can retain covariance under certain assumptions. However there is a major drawback to this type of imputation that statisticians have been struggling to overcome. For forms of imputation that involve a random component, naive variance estimation methods that treat the imputed data as if it were observed will underestimate the variance.

In order to overcome this problem of underestimating the total variance, two classes of methods have been proposed. In one approach, adjustments within replicates, either jackknife replicates or BRR replicates, have been proposed by Rao and Shao (1992) and Shao (1996) for means and totals. It is not clear if these methods can easily accommodate longitudinal analysis in a complex survey. Second, methods of imputing data more than one time have been proposed. The best known of these is multiple imputation proposed by Rubin (1987) but Fay (1996) has proposed fractionally weighted imputation and Shao and Sitter (1996) have proposed using imputation in conjunction with the bootstrap. For the bootstrap imputation, the imputation must be repeated in each bootstrap replicate. This is a computationally costly form of imputation and may not be a satisfactory solution, especially in the context of longitudinal analysis. It is not clear if Fay's fractionally weighted imputation can be extended to the situation of longitudinal analysis. While the controversy over multiple imputation has extended for almost two decades, the article by Kim, Brick, Fuller, and Kalton (2006) demonstrates that the variance, when using multiple imputation for survey data, is biased. This is unfortunate since, other than a biased variance estimator, multiple imputation has many other desirable properties for longitudinal analysis in the presence of missing data, especially relative ease of use.

Note that there is a unique problem in using imputation for longitudinal analysis. Most imputation procedures are developed for item missingness and base a prediction of the missing item on other current values available for the sampled unit with the missing item. In longitudinal analysis, as previously mentioned dropout missingness is a major concern. For dropouts there are no other current items available for prediction of the missing value so the standard techniques may not be adaptable to the situation of imputing items for dropout missingness. Some research has been done on imputation for dropout missingness with techniques such as Lost Observation Carried Forward but currently the evaluation of these techniques is not hopeful.

In spite of the fact that imputation is a commonly used technique that provides many desirable features, in terms of longitudinal analysis of survey data with missing observations, it has many problems that the analyst should carefully consider before embarking on an analysis with imputed data.

6. MODELLING THE MISSING DATA

Missing data can be modeled under given assumptions as part of the modeling process. If the data are not missing at random this is the only way to properly deal with the missingness. Examples of these techniques can be found in Little and Rubin (2002) for non-survey data. Examples of modeling the missingness in survey data are not common but a few examples do exist. Beckett, et. al. (1993) use an interesting idea of modelling the missingness using Markov models. Individuals observed over time are modelled as being in state 0, 1, or missing using logistic regression. The logistic models used are design based. This work provides an interesting example for categorical data, but for continuous and ordinal type data the techniques do not apply.

7. CONCLUSION

This paper provides a short review of some techniques which have been used to deal with missingness in longitudinal analysis of data from a complex survey. Since the topic is too broad to cover in one paper, only highlights of techniques that the author has investigated have been covered. It is clear that dealing with missing data is a very difficult subject in general. In conjunction with longitudinal data as well as complex survey data, the approaches to overcoming problems of missing data present a unique set of challenges.

REFERENCES

Beckett, Laurel A., Brock, Dwight B., Scherr, Paul A. and de Leon, Carlos Mendes (1993), "Markov models for longitudinal data from complex samples", *ASA Proceedings of the Section on Survey Research Methods*, 921-925.

Bernardo, J.M. and Smith, A.F.M. (2000), *Bayesian Theory*, New York: Wiley.

Binder, D.A. (1983), "On the variance of asymptotically normal estimators from complex surveys." *International Statistical Review*, 51, 279-92.

Brogan, Donna (1998), "Pitfalls of using standard statistical software packages for sample survey data", http://www.rti.org/sudaan/pdf_files/brogan.pdf

Chambers, R. and Skinner, C.J. (2003), *Analysis of Survey Data*, New York: Wiley.

Cochran, W.G. (1977), *Sampling Techniques*, Wiley & Sons, New York

Fay, Robert E. (1996), "Alternative paradigms for the analysis of imputed survey data", *Journal of the American Statistical Association*, 91, 490-498.

Gelman, A. (2005) "Prior distributions for variance parameters in hierarchical models", *Bayesian Analysis*, 1, 1-19.

Ghosh, Malay and Meeden, Glen (1997), *Bayesian Methods for Finite Population Sampling*, New York: Chapman Hall.

Kim, J.K, Brick, J.M., Fuller, W.A. and Kalton, G. (2006), "On the bias of the multiple imputation variance estimator in survey sampling", *Journal of the Royal Statistical Society, Series B*, to appear.

Korn, E.L. and Graubard, B.I. (1999), *Analysis of Health Surveys*, New York: Wiley.

Little, R.J.A. and Rubin, D.B. (2002), *Statistical Analysis With Missing Data* (2nd Ed.), New York: Wiley.

Pfeffermann, D. (1993), "The role of sampling weights when modeling survey data." *International Statistical Review*, 61, 317-337.

Pfeffermann, D. and Lavange, L. (1989), "Regression Models for Stratified Multi-Stage Cluster Samples", in *Analysis of Complex Surveys* (Skinner, C. J., Holt, D., and Smith T.M.F. eds.), chapter 12, New York: Wiley.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rasbash, J. (1998), "Weighting for Unequal Selection Probabilities in Multi-Level Models", *Journal of the Royal Statistical Society, Series B*, 60, 23-40.

R Development Core Team (2005), R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rao, J. N. K. and Shao, J. (1992), “Jackknife variance estimation with survey data under hot deck imputation”, *Biometrika*, 79, 811-822.

Sarndal, C.-E., Swensen, B. and Wretman, J. (1992), *Model Assisted Survey Sampling*, New York: Springer-Verlag.

Shao, Jun (1996), “Resampling methods in sample surveys”, *Statistics*, 27, 203-237.

Shao, Jun and Sitter, Randy R. (1996), “Bootstrap for imputed survey data”, *Journal of the American Statistical Association*, 91, 1278-1288.

Skinner, C.J. and Holmes, D.J. (2003), “Random Effects Models for Longitudinal Survey Data”, in *Analysis of Survey Data* (Chambers, R. and Skinner, C.J. eds.), chapter 14, New York: Wiley.

Skinner, C. J., Holt, D. and Smith, T.M.F (1989), *Analysis of Complex Surveys*, New York: Wiley.

Verbeke, G. and Molenberghs, G. (2000), “Linear mixed models for longitudinal data”, Springer-Verlag Inc (Berlin; New York)