Adjusting for Nonignorable Missing Data with Nonignorable Sampling Design

Moh Yin Chang
University of Nebraska-Lincoln

Abstract

Current practice in survey arena handles attrition in longitudinal surveys assuming that the data is missing at random (MAR), despite some evidences that show the data are not missing at random (NMAR). When the missing data is treated as MAR when they are in fact NMAR it results in biased inference. This paper reviews some methods for handling nonignorable missing data, examines the ignorability in missing health outcomes in the Health and Retirement Study (HRS) using propensity scores balance tests, and demonstrates bias of treating the missing data as MAR when they are in fact NMAR.

Keywords: attrition bias, nonignorable missing data, not missing at random, pattern-mixture model, propensity scores

1. Purpose of the Study

Analysis of longitudinal data is often complicated by attrition bias in which the attriters demonstrate different responses from the nonattriters. This phenomenon causes data to be not missing at random (NMAR) (Little and Rubin, 2002). Ignoring the missing data mechanism in this case can result in biased inferences, rendering the missingness mechanism nonignorable. Evidences of attrition bias have been found in health (e.g. Reynolds et al., 2005) and economic studies (e.g. Zabel, 1994). Statistical methods dealing with nonignorable missing data have been studied extensively in the area of biostatistics. Nonetheless, nonignorable missing data models are rarely used in survey statistics because of lack of study on the joint treatment of nonignorable missing data and complex sampling designs. This paper aims to identify potential solutions for handling nonignorable missing data and complex sampling simultaneously.

2. Review of Nonignorable Missing Data Models

Statistical analysis when data is NMAR requires joint modeling of dependent variable and response (or missing data) processes. This is problematic because researchers have to make unverifiable assumption about the data distribution of the nonrespondents. Nevertheless, there has been an explosive growth in statistical methods dealing with NI missing data in the last decade. These methods can be broadly classified into selection models versus pattern-mixture models (Little, 1993).

1.1 Selection Model

Selection models specify the full-data distribution and thus require users to assume a data distribution for nonrespondents. For a longitudinal study with m waves, let $Y=(Y_{i1},\ldots,Y_{ik})$ be a $(1\times k)$ complete-data vector of outcomes for subject i for m waves of measurements. Y may be completely observed so that $1\le k_i\le m_i$. Partition $Y=(Y_{obs},Y_{mis})$, where Y_{obs} is the observed part of Y, Y_{mis} is the missing part of Y. X represents a set of covariates that are fully observed. R is a missing-data indicator, $R=1$ if a subject responds, and $R=0$ for nonresponse. With selection models, the joint distribution of Y_i, R_i are factorized as: $f(Y,R\mid X)=f(Y\mid X)f(R\mid X,Y)$
where $f(Y\mid X)=f(Y_{obs},Y_{mis})$ is the complete-data model; $f(R\mid Y,X)$ is the response process.

Examples of selection models include Heckman's (1976) and Diggle and Kenward's (1994).

Making assumption on the unobserved data is necessary for a selection model to be identified. However, this criterion is not always sufficient for the model to be identified. Moreover, selection models often run into convergence problems due to too many parameters to be estimated.

1.2 Pattern-Mixture Model

Pattern-mixture models stratify the responses by missing data patterns, that is, the response process is a mixture model of varying missing data patterns. The models are factorized as: $f(Y,R\mid X)=f(Y\mid X,R)f(R\mid X)$, where $f(Y\mid X,R)$ represents the data distribution of the outcome variable conditional for each patterns of missing data. $f(R\mid X)$ is the marginal distribution of each missing data patterns predicted by a set of covariates. A popular way of controlling for attrition bias is to stratify the data by wave in which respondents drop out (e.g. Little

1993, Hogan and Laird, 1997). Little (1994) and Little and Wang (1995) showed that pattern-mixture models can be simpler to fit than selection models when a parametric distribution is assumed for the response/missing data model. Nevertheless, our substantive interest is almost always in parameters averaged over patterns. This situation resembles the problem of obtaining averaged estimates from a stratified random sample. Researchers may obtain marginal estimates over all patterns using weighted average method, and calculate the variances for marginal estimates by Taylor linearization or replication methods.

3. Model Estimation

The nonignorable missing data models may be estimated using a likelihood-based methods, generalized estimating equations (GEE), and two-step methods for Heckman's selection model.

With respect to likelihood-based estimation, missing data is estimated by maximizing the log likelihood functions. Theoretical solution to the complex likelihood function is difficult to find, especially in the context of missing data. To solve this issue, EM algorithm is proposed to compute maximum likelihood estimates (Dempster, Laird, and Rubin, 1977).

When analysing survey data, a weighted likelihood function is maximized, called pseudo maximum likelihood (PML) estimation. The idea behind the pseudo-likelihood estimator is to treat the weighted data as a census. The asymptotic covariance matrix of these estimates are estimated using Huber-White "sandwich" estimator, which has be shown to be robust of any sampling designs (Skinner, 1989). Specifically, the PML estimates

$$\log(L) = \sum_i w_i \log(L_i)$$

and the covariance matrix is estimated by

$$\left(\partial^2(\log(L))/\partial\theta\partial\theta'\right)^{-1}\left[\sum_i w_i^2\left(\partial(\log(L_i)/\partial\theta\right)\left(\partial(\log(L_i))/\partial\theta\right)'\right]\left(\partial^2(\log(L))/\partial\theta\partial\theta'\right)^{-1}$$

The likelihood-based estimation do not fill in values for the missing data but attempt to identify the sufficient statistics from the likelihood function and then maximize the likelihood through estimating values for the sufficient statistics given the observed data. EM algorithm performs these steps in a iterative manner to gradually increase the likelihoods at each iteration to eliminate having to find a theoretical solution to the likelihood function.

When estimating a pattern-mixture model, different likelihood functions are maximized separately within patterns.

4. Methods

4.1 Data

This study use data from the Health and Retirement Study (HRS). The HRS is a longitudinal survey collected biennially since 1992 to collect data on Americans aged 50 and older. A total of 9825 individuals responded the baseline survey in 1992. This survey has very modest attrition rates of 10%, 9%, 7%, 7%, 6% and 4% respectively in 1994, 1996, 1998, 2000, 2002, and 2004. Nevertheless, the cumulative attrition rate is still high, about 40%. Hereafter refers each year of study as wave 1 to wave 7.

Health outcomes are the main interest of this paper. Two health outcomes are studied: self-rated health and physical limitation. Self-rated health is reported on a five-point scale: excellent, very good, good, fair, and poor. Physical health is measured by a scale of twelve "yes-no" questions on limitation in terms on performing physical tasks. The tasks vary in terms of strength and parts of body used. These tasks are walking a block, walking several blocks, stooping, climbing a flight of stair, climbing several flights of stairs, jogging a mile, getting up from a chair after sitting for two hours, getting up after sitting for a long time, extending arms above shoulder, picking a dime from a table, pulling or pushing an object, and lift ten-pound weights. Selected baseline information is used as covariates. The covariates are respondent's age, gender, race, level of education (in grades), lifetime smoking, current smoking status, BMI, and marital/cohabitation status.

4.2 Statistical Analysis

A series of propensity scores balance tests are performed to examine the ignorability of the missing data. The propensity scores are a scalar summary of the covariates that "balance" the subjects between the treated and untreated groups, assuming that the treatment is independent of the outcome of interest conditional on the covariates (Rosenbaum and Rubin, 1983).

In the context of attrition, the state of drop-out is analogous to a binary treatment. The attrition propensity scores are estimated based on the selected covariates mentioned above as well as the baseline measures of the outcome variables.

If the propensity score balance tests suggest the missing data to be potentially NMAR, the pattern-mixture model stratified by time of attrition is performed using Mplus software. Mplus take into accounts the sampling design and sampling weights using PML estimation with EM algorithm.

5. Results

Table 1 indicates the weighted inferences for covariates and baseline outcomes in 1992.

Table 1. Descriptive statistics (weighted) of baseline covariates from HRS

Variable	Baseline proportion/mean	SE
male	0.478	0.005
white	0.818	0.009
lifetime smoker (= yes)	0.641	0.006
current smoking (= yes)	0.270	0.006
married/cohabitating (= yes)	0.754	0.006
age	55.61	0.04
education	12.37	0.08
BMI	26.96	0.07
self-rated health	2.51	0.02
physical limitation	3.59	0.06

Note. Coding of self-rated health status: 1 = excellent; 5 = poor.
Coding of physical limitation: 0 = no limitation;
12 = maximum limitation.

Table 2 indicates the raw mean outcome measures comparing the attriters and nonattriters before they drop out from the study. There exists a consistent trend in which the attriters have reported worse health prior to withdrawal from the study relative to the nonattriters.

Propensity scores are created to balance the covariates between the attriters and nonattriters. The number of attriters is cumulative at each wave. Table 3 indicates the balance tests of the baseline self-rated health and physical limitation values between attriters and nonattriters by wave. Nonattriters are the reference group.

Table 2. Weighted mean outcome measures between attriters and nonattriters prior to attrition

Year	Self-rated health		Physical limitation	
	Attriters	Nonattriters	Attriters	Nonattriters
1992	2.937	2.558	4.065	3.542
1994	2.985	2.626	3.095	2.727
1996	3.009	2.613	3.527	3.029
1998	3.255	2.826	4.053	3.130
2000	3.180	2.728	4.103	3.244
2002	3.314	2.783	4.202	3.500

With respect to self-rated health, the attriters in 1994 on average reported 2.538 worse health status at baseline than the nonattriters, controlling for their attrition propensities in 1994. Nevertheless, the difference in baseline self-rated health between the attriters and the nonattriters is decreasing overtime and eventually becomes insignificant.

On the other hand, the attriters in 1994 reported an average of 3.003 more physical limitations at baseline than the attriters, controlling for their attrition propensities. The difference is not significant in 1994, but becomes more stable overtime.

The balance tests suggest the missing values in both outcome measures to be NMAR. Pattern-mixture model is performed on the self-rated health to demonstrate the potential bias from treating NMAR missing values as MAR. Table 4 compares the unadjusted longitudinal estimates (i.e., mean initial status, linear and quadratic growth rates) under MAR assumption and those of pattern-mixture model. Three missing data patterns are identified: early attriters who withdraw at wave 2, 3, 4 and 5; late attriters who withdraw at wave 5 and 7; and nonattriters.
Under MAR assumption, the mean baseline estimate of self-rated health is 2.506. The linear and quadratic growth rates at each wave are 0.087 and -0.003 respectively. Under NMAR assumption, early attriters have worse health than late attriters, and late attriters have worse health than nonattriters. The marginal estimate of the mean baseline health status is 2.616, which is 0.11 point higher than the estimate under MAR. The corresponding linear and quadratic growth rates are 0.104 and 0.001 respectively, compared to 0.087 and -0.003under MAR assumption.

6. Discussion

Baseline difference in outcomes between the attriters and nonattriters controlling for attrition propensity suggests that the missing outcomes values to be potentially NMAR. Early attriters appear to account for the major part of the baseline difference since the difference between the attriters and nonattriters are diminishing in both health outcomes over time. In addition, although the baseline difference in health outcomes are decreasing over time but the difference is consistent, which leads to biased inference if omitted. The degree of bias in longitudinal estimates as demonstrated in the pattern-mixture model on the self-rated health measure, are 0.11, 0.03, and 0.008 points respectively comparing the unadjusted estimates and the marginal estimates from the pattern-mixture model.

Note that this study does not attempt to identify the best nonignorable missing data model for the problem. The pattern-mixture model used here may not be the best model to reduce attrition bias. It is expected that a more thoughtful model building may further reduce the attrition bias in self-rated health. Besides, a selection or shared-parameter model may produce better adjusted estimates than a patter-mixture model in this case. These are references for future research.

7. References

Dempster, A.P., Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*(1), 1-38.

Diggle, P., & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied Statistics, 43*(1), 49-93.

Hogan, J. W., & Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in Medicine, 16*(1-3), 239-257.

Hogan, J.W., Roy, J., & Korkontzelou, C. (2004). Tutorial in biostatistics: Handling drop-out in longitudinal studies. *Statistics in Medicine*, *23*, 1455-1497.

Little, R. J. A. (1993). Pattern-mixture nodels for multivariate incomplete data. *Journal of the American Statistical Association, 88*(421), 125-134.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. (2nd ed.) New York: Wiley-Interscience.

Reynolds, J., Frank, K., & Heyman, K. (2005). The problem of attrition in survey research on health. *Paper presented at the Conference of American Sociological Association*, Philadelphia, PA.

Rosenbaum, P. R., & Rubin, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of Royal Statistical Society. Series B (Methodological), 45*(2), 212-218.

Skinner, C. J. (1989). Domain mean, regression and multivariate analysis. In C. J. Skinner, D. Holt & T. M. F. Smith (Eds.), *Analysis of complex surveys*. Essex: John Wiley & Sons.

Zabel, J. E. (1994). *An analysis of attrition in the PSID and SIPP with an application to a model of labor market behavior.* Paper presented at the BLS Conference on the Consequences of Attrition in Longitudinal Data, Washington, D.C.

Table 3. Propensity score balance test on ignorability of attrition

	Self-rated health		Physical limitation	
Year	Attriters coef. (SE)	p-value	Attriters coef. (SE)	p-value
1994	-2.538 (0.471)	0.000	3.003 (1.966)	0.132
1996	-1.384 (0.300)	0.000	1.800 (1.076)	0.099
1998	-0.620 (0.177)	0.001	2.370 (0.754)	0.003
2000	-0.279 (0.129)	0.035	1.572 (0.588)	0.010
2002	-0.179 (0.095)	0.064	1.422 (0.472)	0.004
2004	-0.080 (0.088)	0.363	1.318 (0.414)	0.002

Note. Reference group= nonattriters.

Coding of self-rated health status: 1 = excellent; 5 = poor.

Coding of physical limitation: 0 = no limitation; 12 = maximum limitation

Coefficients are controlling for the propensity scores estimated on age, gender, race, education, smoking BMI, marital/cohabitation status, and baseline measures of overall health status and physical limitation

Table 4. Unadjusted and pattern-mixture model estimates for self-rated health

	Unadjusted		Attrite W2-W5		Attrite W6-W7		Nonattriters		Marginal	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
Baseline mean	2.506	0.022	2.927	0.057	2.721	0.072	2.501	0.052	2.616	0.052
Growth rate : linear	0.087	0.007	0.058	0.043	0.113	0.03	0.117	0.02	0.104	0.019
quadratic	-0.003	0.001	0.039	0.014	-0.001	0.005	-0.011	0.003	0.001	0.005

Note. W = wave. Self-rated health status: 1 = excellent; 5 = poor.

Baseline mean = mean overall health status at wave 1. Growth rate refers to growth of overall health status by wave.