A Method for Analyzing Categorical Data with Panel Attrition

Tim F. Liao

University of Illinois/University of Essex

First draft paper prepared for the 2006 Methodology of Longitudinal Surveys conference at the University of Essex.

Contact information: Tim F. Liao, Department of Sociology, University of Illinois, 326 Lincoln Hall, Urbana, IL 61801, USA; email: tfliao@uiuc.edu.
**Introduction**

Panel studies have given social scientists an edge for data analysis over cross-sectional designs because panel data allow the researcher to analyze social, economic, and behavioral dynamics over time. However, the strengths of panel data at the same time are weakened by the problem of attrition. Given that panel attrition is a rule rather than an exception in longitudinal studies, researchers have long been aware of the issue and have attempted methodological approaches to dealing with the problem (e.g., Hausman and Wise, 1979).

Broadly speaking, panel attrition can be viewed as a type of missing data, which may or may not be generated at random, and as such the nature of missing mechanisms as defined by Little and Rubin (1987) may affect our analyses in undesirable ways. Specifically, if attrition is not a random phenomenon, then there are two kinds of undesirable consequences: it affects the representativeness of the original sample, and it may create bias in the analysis of association between variables.

This paper focuses on the second situation when association between categorical variables in a two-wave panel is the object in a research project. For analyzing such association, log-linear models are typically used. However, when there exist missing data, be they due to panel attrition or some other mechanisms, estimation can be biased if only complete data are analyzed. Here we follow the tradition of analyzing partially classified contingency tables (Baker and Laird, 1988; Fay, 1986; Fuchs, 1982; Liao, 2004; Molenberghs et al., 1999; Vermunt, 1997).

The difficulty for analyzing partially missing data in general, and panel data with attrition in particular, is that the data analyst does not know exactly what mechanism is
responsible for generating the missing data. To solve this conundrum, we conduct a Monte Carlo simulation in this paper. Partially missing data representing a varying proportion of attrition in a two-wave panel are generated according to three different assumptions of missingness for three different sample sizes. The data are then analyzed by models assuming three different missing mechanisms, with the object of identifying which model may fare better under which sets of conditions.

**Missing Mechanisms of Panel Attrition**

Let $A_t$ and $B_t$ denote two categorical variables observed at time $t$. For a two-wave panel with two dichotomous variables of interest, panel attrition means that some respondents are observed on both occasions while some others are not observed at time point 2. Table 1 illustrates the situation where completely observed individuals are contained in a completely cross-classified contingency table, and cases with attrition, a partial table.

---Table 1 about here---

The $A_1$ and $B_1$ variables are observed on the first wave, and the $A_2$ and $B_2$ variables, the second wave. Individuals that do not experience attrition appear twice in the table, once in the upper panel and once in the lower panel, in one of the $C_{111}$ to $C_{122}$ cells and again in one of the $C_{211}$ to $C_{222}$ cells. Individuals that experience attrition, however, are found only once in the upper panel of the $C_{111}$ to $C_{122}$ cells. Therefore, the contingency table for a two-wave panel with two categorical variables consists of two (sub)tables, one being the completely cross-classified table of both the upper and lower panels and the other being the upper panel only for the cases with attrition. It is obvious
that the common practice of using only complete data ignores useful information, and is
prone to bias, depending on the missing mechanism generating the attrition.

Let $R$ represent response patterns (response vs. nonresponse or 1 vs. 2) in the
table. When $R=1$, a case is observed; when $R=2$, a case is unobservable or missing.
Missing mechanisms are defined by the relation between $R$ and Let $A_t$ and $B_t$. Three
common mechanisms are Missing Completely At Random (MCAR), Missing At Random
(MAR), and Missing Not At Random (MNAR, or Not Missing At Random, NMAR).
These well-known patterns of missing mechanisms are summarized in Rubin and Little

Panel attrition may be considered as a special case of missing data. Whereas
missing data may occur to any individual on any (i.e., some but not necessarily all)
variables, when panel attrition is present, the individual is unobservable for all the
variables in the data. Thus, we may define how $R$ relates to $A_t$ and $B_t$ in terms of whether
attrition is due to current (though unobserved) variables or variables of an earlier wave
(Hausman and Wise, 1979).

For a two-wave data set with two categorical variables, we define the three
common missing mechanisms as follows:
MCAR:

\[
R \perp A_1, A_2, B_1, B_2
\]
\[
P(R = 1 | A_1, A_2, B_1, B_2) = P(R = 1)
\]

In other words, panel attrition is independent of or unrelated to any of the $A_t$ and $B_t$
variables.
MAR:
\[ R \perp A_2, B_2 \mid A_1, B_1 \]
\[ P(R = 1 \mid A_1, A_2, B_1, B_2) = P(R = 1 \mid A_1, B_1) \]

In this scenario, the conditional probability of attrition, \( R \), depends only on \( A_1 \) and \( B_1 \) but not on \( A_2 \) and \( B_2 \).

HW/MNAR:
\[ R \perp A_1, B_1 \mid A_2, B_2 \]
\[ P(R = 1 \mid A_1, A_2, B_1, B_2) = P(R = 1 \mid A_2, B_2) \]

Here the conditional probability of attrition, \( R \), depends only on \( A_1 \) and \( B_1 \) but not on \( A_2 \) and \( B_2 \) (once \( A_1 \) and \( B_1 \) are controlled). This is the data situation originally described and analyzed by Hausman and Wise (1979), and the model under study by Hirano et al. (2001) who also attributed the model to Hausman and Wise (1979).

For the data in Table 1, we specify a log-linear model for analyzing the association between \( A_t \) and \( B_t \) and whether such association would be consistent over time. We present below the log-linear models for the three situations with missing data:

**MCAR:**
\[
\log m_{u,a,b} = u + u_{a,b}^{A_{t}} + u_{a,b}^{A_{t}B_{t}} + u_{r}^{R}
\]  
(1)

where \( u \) is the grand mean (when variables are effect coded) or the reference category (when variables are dummy coded), \( u_{a,b}^{A_{t}} \) is the two-way interaction between \( A_1 \) and \( B_1 \), and \( u_{a,b}^{A_{t}B_{t}} \) are the four-way interaction term between \( A_1, B_1, A_2, \) and \( B_2 \). For the sake of brevity, lower-order terms are not presented in the formula even though they are included in the estimation. Note that the MCAR model is different from the one relying on complete data only, where the \( u_{a,b}^{A_{t}} \) and \( u_{r}^{R} \) terms are excluded.

**MAR:**
\[ \log m_{a_i,b_j} = u + u_{rb_i} + u_{ra_i} + u_{rb_j} + u_{ra_j} \]  

(2)

where \( u_{ra_i} + u_{rb_i} \) terms captures the association between attrition pattern (R) and \( A_1 \) and/or \( B_1 \).

\[ \log m_{a_i,b_j} = u + u_{rb_i} + u_{ra_i} + u_{rb_j} + u_{ra_j} + u_{rb_2} \]  

(3)

where \( u_{ra_1} + u_{rb_1} \) terms captures the association between attrition pattern (R) and \( A_2 \) and/or \( B_2 \). Once again, equations (2) and (3) do not include any lower-order terms even though such terms are included in the estimation.

**A Monte Carlo Simulation of Two-Wave Panel Attrition with Two Binary Variables**

In the simulation we explicitly use three missing mechanisms—MCAR, MAR, and HW/MNAR—for data generation, and we estimate models assuming each of the same three types of missingness. In addition, we simulate the data at the three sample sizes of 200, 500, and 1,000. The size of 200 may be consider the smallest possible for meaningfully analyzing the relation between two binary variables over two time points with attrition being present while the sample size of 1,000 should provide some resemblance to large samples.

To assess how estimation is affected, we allow for five attrition rates in the panel data generation: 0.1, 0.2, 0.3, 0.4, and 0.5. Such different attrition rates may roughly represent the gradual increase in the loss to the original sample over the subsequent panels (i.e., panels 2 to 6).

Thus, data are generated according to the scenarios above, and models assuming MCAR, MAR, and HW/MNAR are estimated by analyzing the simulated data of both
completely classified and partially classified contingency tables (and subtables). For analyzing these simulated data, log-linear models as described by Vermunt (1997) and Liao (2004) are estimated via the EM algorithm (Dempster, Laird, and Rubin, 1977). Simply put, the three log-linear models of the forms of (1), (2), and (3) are estimated in the simulation study, with two statistics reported: the likelihood ratio (LR) statistic for assessing the overall model fitting, and log-odds ratios formed by $A_{11}B_{11}, A_{22}B_{22}, A_{11}B_{22}$ and $A_{22}B_{11}$:

$$\log \left( \frac{P[A_1(1)B_2(1)|A_1(1)B_1(1)] P[A_1(2)B_2(2)|A_1(2)B_1(2)]}{P[A_2(1)B_2(2)|A_2(1)B_1(1)] P[A_2(1)B_2(1)|A_2(2)B_1(2)]} \right)$$

for understanding possible bias in estimating the association between the two categorical variables (hereafter, log=natural logarithm, as is the case in log-linear models). Note that this log-odds ratio is one of many possible choices. It is however, the cleanest choice when association is of concern. A positive value indicates a positive association between the two variables over the two time points. By assuming stability over time (i.e., respondent who chose category 1 in variable $A$ at time 1 would more likely keep that choice) and using a joint probability of 0.7 in the diagonal cells and 0.1 in the off-diagonal cells (indicating mobility in choosing the categories), we set the log-odds ratio at $\log(49)$ or 3.89. The parameter will allow us to assess bias and consistency in estimation.

First, data are generated by assuming MCAR, that is, attrition is unrelated to any of the variables. The simulated results are presented in Figures 1-6. In each figure, the first five boxplots are for the data situations of attrition rates of 10% to 50%, respectively, all with a sample size of 200; the next five boxplots are again for the five different attrition rates, but with a sample size of 500; the final five boxplots in each figure are of sample size 1,000. In Figures 1 and 2, the MCAR model is used to analyzing the data; in
Figures 3 and 4, the MAR model is used; in Figures 5 and 6, the HW/MNAR model is used. All odd numbered figures report LR statistics while even numbered figures report estimated log-odds ratios.

---Figures 1 to 6 about here---

Overall, it is clear that the larger the sample size, the better the performance of the model, judged either by LR statistics or by estimated log-odds ratios. It is also obvious that the greater the attrition rate becomes, the greater uncertainty there is with model estimation, regardless of the sample size. Estimation becomes least stable when the sample size is small and when attrition is high. When the sample size is small and especially when attrition is high, log-odds ratio estimates can be biased, with a mean value above 4 and a large amount of uncertainty. When the data are generated by MCAR, the difference in performance between the three models is not very large, though MCAR and MAR models are slightly more preferable to MNAR models. How would the models behave when data are MAR?

---Figures 7 to 12 about here---

Most of the earlier observations can also be applied to the data scenario when missing is at random (MAR). That is, the general remarks about sample size and attrition rate are still valid when the data are MAR, and both MCAR and MAR models are still slightly preferable to MNAR models. However, there is a new phenomenon common to all three types of models: when the attrition rate is low, overall modeling fitting can actually be poorer than when attrition rates are high, somewhat counter-intuitively, and performance is even poorer when the sample size is larger. This is probably due to a larger available sample all contributing to the calculation of LR statistics though why it
happens to the scenario when missing is MAR is unclear. (The bias in log-odds ratio estimation is not affected.) Among the three models, MAR models are least affected by this adverse effect when the data are MAR, followed by MNAR models while MCAR models behave the poorest. Would the same effect be found when data are MNAR? 

---Figures 11 to 18 about here---

Not really. Overall model fitting is affected only slightly for MAR and MNAR models when attrition rates are low and the increase in LR statistic for larger samples is also rather moderate. However, the same cannot be said for MCAR models, when the increase in LR statistic is much greater. When the concern is log-odds ratio estimation, both MCAR and MAR models can do fairly well when sample sizes are 500 or greater. MNAR models may be slightly more preferable though the difference is rather small.

**Concluding Remarks**

Overall, a higher attrition rate leads to a poorer model fitting, and to less efficiency and greater bias in estimating log-odds ratio. A larger sample size in general gives a more efficient estimator and generates less bias. However, a larger sample size combined with a lower attrition rate when data are MAR can produce a poorer model fitting though log-odds ratio estimation is not affected.

MCAR and MAR models are indistinguishable when it comes to log-odds ratio estimation although they perform differently as far as overall model fitting is concerned. When data are MCAR, MCAR and MAR models will do equally well, and HW/MNAR models will do well for samples size $\geq 500$. When data are MAR, MAR models will be more preferable (in modeling fitting, not necessarily in unbiased
estimation) than MNAR models and MCAR models, in that order. When data are HW/MNAR, MAR models may actually slightly outperform HW/MNAR models for model fitting; when sample size=200, MCAR and MAR models can do better than HW/MNAR ones in log-odds ratio estimation.

Which model to use when the missing mechanism is unknown? It often makes no difference, but when there is a difference, it appears that the MAR model may come slight ahead when all things considered. The most important point is that partially observed cases are included in partially classified tables in the analysis. When they are excluded, bias in estimation is bound to occur. The next draft of the paper will include as well estimation based on complete cases only.

References


Table 1: Missing Patterns in a Two-Variable, Two-Wave Setting

<table>
<thead>
<tr>
<th></th>
<th>( B_1 = 1 )</th>
<th>( B_1 = 2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( A_1 = 1 )</td>
<td>( C_{111} )</td>
<td>( C_{112} )</td>
</tr>
<tr>
<td>( A_1 = 2 )</td>
<td>( C_{121} )</td>
<td>( C_{122} )</td>
</tr>
<tr>
<td></td>
<td>( B_2 = 1 )</td>
<td>( B_2 = 2 )</td>
</tr>
<tr>
<td>( A_2 = 1 )</td>
<td>( C_{211} )</td>
<td>( C_{212} )</td>
</tr>
<tr>
<td>( A_2 = 2 )</td>
<td>( C_{221} )</td>
<td>( C_{222} )</td>
</tr>
</tbody>
</table>
Fig. 1: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MCAR/MCAR
Fig. 2: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MCAR/MCAR

Note: Outliers at infinity in boxplots 1-5 excluded.
Fig. 3: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MAR/MCAR
Fig. 4: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MAR/MCAR

Note: Outliers at infinity in boxplots 1-5 excluded.
Fig. 5: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MNAR/MCAR
Fig. 6: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MNAR/MCAR

Note: Outliers at infinity in boxplots 1-5 excluded.
Fig. 7: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MAR/MAR
Fig. 8: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MAR/MAR

Note: Outliers at infinity in boxplots 1-5 and 10 excluded.
Fig. 9: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MCAR/MAR
Fig. 10: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MCAR/MAR

*Note:* Outliers at infinity in boxplots 1-5 and 10 excluded.
Fig. 11: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MNAR/MAR
Fig. 12: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MNAR/MAR

Note: Outliers at infinity in boxplots 1-5 and 10 are excluded.
Fig. 13: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MNAR/MNAR
Fig. 14: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MNAR/MNAF

Note: Outliers at infinity in boxplots 1-5, 10, and 13 are excluded.
Fig. 15: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MCAR/MNAR
Fig. 16: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MCAR/MNAR

Note: Outliers at infinity in boxplots 1-5 and 10 are excluded.
Fig. 17: LR Statistics by Attrition Rate 10%-50% & sample size 200-1,000: MAR/MNAR
Fig. 18: Log-Odds Ratios by Attrition Rate 10%-50% & sample size 200-1,000: MAR/MNAR

Note: Outliers at infinity in boxplots 1-5 and 10 are excluded.