TRACKING, WEIGHTING, AND SAMPLE SELECTION MODELING TO CORRECT FOR ATTRITION

Kimberly A. McGuigan, Phyllis L. Ellickson, Ronald D. Hays and Robert M. Bell, RAND
Kimberly A. McGuigan, RAND, 1700 Main Street, Santa Monica CA 90407-2138

Key Words: unit nonresponse, attrition, nonresponse bias, tracking, sample selection

1. INTRODUCTION

When taking multiple measures of the same subjects over time, some subjects may be lost to follow-up. The effects of attrition can range from modest to considerable, depending on the length of time between data collections, the number of subsequent data collections, and characteristics of the subjects under study. If attrition is systematically related to outcomes of interest and if nonresponse adjustments are not made, bias may result. The internal validity of a study may be threatened, as observed differences may be due to differential nonresponse. Attrition may also effect the generalizability of a study's results: if the composition of the respondents varies from the original sample because of differential rates of loss for certain kinds of subjects, then the respondents may not reflect the target population of interest.

In school-based substance use applications, attrition is usually positively correlated with substance use: those who are lost to follow-up are more likely to use target substances. Pirie et al. (1988) found that data which did not account for nonresponse showed considerably lower prevalence levels of daily smoking than data which included subjects who might otherwise have been lost. In school-based studies, the percentage of subjects lost to attrition may be large, depending on the location of data collection (e.g. urban centers may have higher rates of student transfers, absentees, or dropouts) and time from baseline to follow-up. In Pirie et al. (1988), only 78% of seventh graders surveyed at baseline were found in their original school districts five years later, and only 68% of the baseline sample provided responses to the follow-up survey.

For this study we compare alternative approaches to correct for attrition using data from a school-based substance use prevention program, Project ALERT (Ellickson and Bell, 1990; Ellickson, Bell, and McGuigan, 1993). We assess the performance of three different methods that can reduce bias associated with nonresponse: tracking, weighting, and sample selection modeling. The first of these, tracking, is implemented as part of the data collection phase. Weighting and sample selection modeling, in contrast, are statistical modeling approaches that can be performed *ex post.* Modeling approaches could save substantial costs, but they require assumptions that may not be correct or testable. Then, this study addresses the questions: In school-based substance use research, are the costs of tracking avoidable? And, for this and similar applications, are the assumptions underlying modeling methods acceptable?

2. BACKGROUND

Pirie et al. (1989) provide an excellent overview of methods used to implement tracking, and the success rates of tracking in school-based prevention research. Methods include telephone searches, high school records, postal forwarding and record updates, personal contacts, and public record searches. Tracking requires considerable time and effort. While tracking may substantially reduce the percentage of subjects lost to follow-up, frequently multiple methods and multiple attempts at contact are required to achieve good coverage rates. For example, in the Waterloo Smoking Prevention Project, more than 80% of the subjects who were tracked required three or more sources in order to be located two and one-half years after the previous data collection (Pirie, et al., 1989).

2.1 Overview of Alternative Methods to Correct for Attrition

2.1.1 Tracking

Where possible and affordable, successful tracking is the ideal solution to nonresponse. For example, if simple random sampling was employed at baseline and all cases are located through tracking, then the sample becomes self-weighting. Tracking, if completely successful, results in unbiased estimates. These estimates are also the most efficient as they are based on the largest attainable sample size and have no additional variation introduced.

However the cost required for tracking can be considerably higher than the cost of the main follow-up data collection. In an illustrative example, Graham and Donaldson (1993) cite a ratio of 5 times the cost of the normal data collection; this ratio is consistent with estimates found for Project ALERT. While this ratio will differ across applications, tracking is nevertheless likely to be comparatively more resource intensive and thus more costly on a per-case basis. Further, it is then likely that tracking will result in less than complete coverage of all subjects lost to follow-up, and so the researcher may still be forced to rely upon additional approaches to adjust for missing observations.

2.1.2 Nonresponse Weights

The first model-based estimation method to be considered is the use of nonresponse weights. We use inverse propensity score weighting, although other approaches, such as weighting cell adjustments can be taken (e.g. Little and Rubin, 1987; Kalton, 1983). We use a logistic model to estimate a subject's propensity to be present at the follow-up wave as a function of baseline variables measured for all subjects. The inverse of this predicted probability is used to calculate the sampling weights. In this way, nonresponse weights resemble sampling weights.

For a dichotomous outcome y that indicates response at follow-up, and $\{x_i\}$, a vector of baseline measures, the estimated propensity $p(y=1 \mid x_1 \ldots x_i)$, is given by :

$$p = \frac{e^{\Sigma\beta_i x_i}}{1 + e^{\Sigma\beta_i x_i}}$$

where $\Sigma\beta_i x_i$ is obtained from the estimated coefficients and covariates in the logistic regression model. The weight associated with each observation is 1/p, is inversely proportional to the propensity of a subject with a given set of characteristics, as defined by the x_i's, to be present at follow-up. Subjects with characteristics correlated with low likelihood of attrition--those with an estimated p near one--receive smaller weights, and subjects who are similar to those who were likely to drop out--those with a small p--receive larger weights. Intuitively, the weights corresponding to those who should have been present for follow-up are redistributed to follow-up respondents in a way which reflects the characteristics of the original sample. To the degree this redistribution is successful, it minimizes the bias associated with attrition.

This method assumes that, conditional on covariates known for both the respondents and nonrespondents, nonresponse is ignorable. Once we have adjusted for nonresponse based on the covariates used to generate nonresponse weights, the respondents are then assumed to be a random sample of respondents plus nonrespondents (Little and Rubin, 1987). Nonresponse weights may reduce bias. However, use of these weights adds an additional source of variation. Thus, weighting will increase the standard error of estimates.

2.1.3 Heckman Sample Selection Model

The second model-based estimation method, the Heckman sample selection model (Heckman, 1979), tests nonignorable nonresponse: there may be unmeasured factors that determine the level of an outcome, given that the outcome is not observed. The sample selection method is often used as an approach to correct estimates in the presence of self-selection, including attrition (e.g. Leigh, Ward, and Fries, 1993; Stolzenberg and Relles, 1990). This analysis has two steps. First, a probit model estimates the probability of being present for follow-up. The results of this probit model are used to calculate an Inverse Mill's Ratio, a hazard term inversely related to the probability of being observed. This hazard term is then used as a continuous covariate in a subsequent Ordinary Least Squares (OLS) regression model to predict the outcome of interest.

Typically one is concerned about "exclusion restrictions" between the independent variables in the probit model and in the OLS regression. That is to say, in the extreme case where the same predictors are used in both steps, one will find that the hazard term may be highly correlated with the predictors in the second step. This is because the hazard is a function of those same predictors from the probit model. To the degree that some of the predictors are shared by the two models, this introduces multicollinearity into the OLS model. Because our focus will be point estimation without covariates, our preferred model would not require covariates other than the hazard term in the OLS regression. However, this would assume that there is no information in the probit covariates that would also predict our target outcomes in the OLS regression. These model specification issues are discussed in greater detail below.

Though there may appear to be some similarities across the weighting and sample selection methods, there are a number of distinctions between the two. First, the distributional forms of residual terms differ between the logistic and probit models. This difference is not particularly noteworthy, especially if the predicted probability does not fall close to one of the extreme tails (e.g. $p < .10$, $p > .90$). The second distinction is more interesting: this is in the way the predicted probability of response is integrated into the second estimation step. The weighting approach uses the predicted probability to redistribute weights across the nonmissing observations. The use of weights will increase the variance of estimates. In contrast, the sample selection method maintains the original weights associated with each observation, but uses the hazard function as a linear predictor of the outcome of interest. Lastly, as noted, weighting assumes nonresponse is ignorable, conditional on the covariates. Sample selection models do not make this assumption; rather, they are recommended as a test for nonignorable nonresponse (e.g. Rubin, 1987).

The derivation of the sample selection method shows that this correction will result in unbiased estimates under the assumption of independent, identically distributed normal errors [i.i.d. $N \sim (0, \sigma^2)$] (Heckman, 1979). Also, one requires exclusion restrictions, noted above: ideally, the variables used to predict attrition are not a subset of the variables used to estimate the outcome variable in the

second step. The sample selection method is not particularly robust to deviations from these assumptions (Manning, Duan, and Rogers, 1987). Stolzenberg and Relles (1990) found that even when normality assumptions are met, the Heckman method often resulted in worse--more biased--estimates under cases of small samples and large percentage attrition.

In summarizing the differences between the two statistical approaches, the weighting method should result in estimates that are less precise than estimates obtained by the sample selection method. Yet, the weighting method may prove to be more robust, and may result in less biased estimates than the sample selection method. For this reason we will address both the bias and efficiency of estimates resulting from each of these two methodological approaches, and will compare these with the estimates from the tracking approach.

3. METHODS

Project ALERT included a total of 6527 students at baseline (grade 7). At a three-year follow-up wave (grade 10), 62.6% of the original baseline sample was located through the main, school-based data collection, 22.0% of the baseline sample was located through tracking, and 15.4% was missing.

To compare the effects of attrition on sample estimates, we chose measures of lifetime use of three substances as outcome variables: cigarettes, alcohol, and marijuana. Each of these was measured using a twelve-point scale, ranging from zero (no use) to eleven (frequent).

Predictors used for weighting and sample selection models include subject demographics (gender, ethnicity, parents' education), geographical location, school achievement, and substance use-related questions (exposure to substance use, attitudes towards substance use).

Weights were developed using forward stepwise logistic regression, regressing presence/absence at follow-up on baseline information. The model results in approximately 12% of deviance (deviance = -2* loglikelihood) accounted for by these variables (Table 1). The fitted model shows a statistically significant improvement over the null; the difference in deviances between the fitted model and the null model containing no covariates is statistically significant (-2*loglikelihood = 1021.104, 30 d.f., $p<.0001$). The resulting weights have a mean of 1.60, with a range from 1.05 to 15.87.

Standard errors for weighted estimates are calculated using Huber's (1967) robust estimation method in the software package Stata. Sample selection models were also performed using Stata.

4. RESULTS

To validate the estimates obtained by each method, we reconstructed mean levels of cigarette, alcohol, and marijuana use variables at baseline, simulating the effects of attrition under each condition (Table 2). At the baseline data collection we have information for all study subjects ("Actual"). This is our "gold standard" since we know with certainty the actual responses for all subjects. Against this standard, we compare results obtained using the three proposed methods to adjust for nonresponse.

The second column of Table 2 reports unweighted estimates for school-based respondents ("Respondents") This column simulates the effect of these subjects being missing at baseline with no other adjustment. These results show that unweighted estimates of substance use prevalence for initial respondents understates actual levels of use by 7% for alcohol, 23% for cigarettes, and 42% for marijuana. ([actual-estimate]/actual).

The third column incorporates information from tracked subjects along with the school respondents. We see underreporting of 3% for alcohol, 10% for cigarettes, and 24% for marijuana--roughly half the level of underreporting seen in the unadjusted column. While these numbers are based on a twelve-point ordinal scale rather than on an interval metric, the magnitude of bias is similar when measures are prevalence rates rather than levels (data not shown).

Weighting data from school-based respondents to correct for nonresponse bias, shown in column four, improves the estimates substantially. Differences between estimated levels and actual rates are only (+) 0.3% for alcohol, (+) 0.2% for cigarettes, and (+) 2% for marijuana. These numbers show that nonresponse weights nearly eliminated the underreporting bias. As noted above, the standard errors for the weighted estimates are larger than those observed for the unweighted estimates for the respondents--i.e. the same sample sizes are used, and the increase in variance is due to weighting.

As Table 2 shows, our initial sample selection results yielded poor estimates. Note the results for cigarettes and marijuana give implausible negative estimates, since our scales were bounded at zero. These results reflect the sensitivity of the sample selection model to the underlying assumption that, conditional on the hazard term estimated in the first equation, there is no other information which predicts substance use levels. These results indicate that the hazard term alone is insufficient to adjust for nonresponse bias, and that additional information is required to improve the model. The constraint here is exclusion restrictions--using the same covariates in the probit and OLS equations induces multicollinearity. While complete collinearity is not a likely risk (e.g.

Olsen, 1980), the coefficients for the selection term and for the other predictors become unstable as multicollinearity increases. Using a simple approach to stabilize the estimates obtained, we included one additional continuous covariate to the OLS regression model. This brought the estimates closer to the actual values, especially for cigarette use, although this method was still considerably less accurate than the weighting method. This example underscores the importance of model specification when using the sample selection method: while our goal was simple point estimation, the selection of covariates into each step of the two-step model must be carefully considered.

Table 3 reports estimates for the same target substances, using data from the three-year follow-up. We do not have an objective criterion against which to assess accuracy of estimates, because "true" substance use levels unaffected by attrition are unknown. We observe the same rank ordering of estimates for alcohol and marijuana as was observed in the validation analyses: school-based respondents only show the lowest levels, followed by respondents plus tracked subjects, then by weighted means of respondents.

5. DISCUSSION

The magnitudes of differences between weighting and tracking are smaller in the three year follow-up results than observed in the validation results. This may be attributable to one of two causes: the correction provided by weighting may not be as strong in the three-year follow-up, or differences between respondents and nonrespondents may not be as large for these substances in the tenth grade as was observed in the seventh grade. This second explanation is plausible given the decrease in differences between the school-based respondents estimates and the respondents plus tracked estimates. At baseline, these groups differ by 4.5%, 17% and 30% for cigarettes, alcohol, and marijuana, respectively. At follow-up these differences have decreased to 1%, 11% and 12%.

School-based respondents use substances less frequently than nonrespondents, consistent with Pirie (1988). Consequently, estimates that do not take attrition into account seriously underestimate overall levels of substance use. Although use of data from tracked subjects decreases this bias, substantial bias remained when evaluated on baseline outcomes.

One concern at the outset was the relative performance of weighting versus sample selection modeling with respect to bias and precision. The precision lost due to weighting is not appreciable here, and the sample selection model, as implemented, is not strictly appropriate as a sensitivity analysis for nonignorable nonresponse. Our application of the two-step sample selection model would be appropriate if the variables that predict nonresponse did not also predict of level of use. More traditional uses of the sample selection model are to predict wages: the first equation predicts entry into the workforce while the second predicts wage amounts (e.g. Maddala, 1985). In such applications it is plausible that these factors have different determinants; here, we know *a priori* that the nonresponse and substance use processes are related.

These analyses provide insight into the relative performance of comparatively inexpensive corrections for subject nonresponse. As indicated at the outset of this study, nonrespondents can be difficult to track, and the resources required for this effort can be considerable. Our validation analyses show excellent agreement between weighted estimates under simulated attrition and the true values observed at baseline, and good adjustments at the three-year follow-up. While tracking has good face validity, it apparently would not have been an adequate substitute for statistical adjustment to correct for attrition. This may not be true in all applications. However, in large studies, the costs associated with a full-scale tracking effort may not be justified by the amount of information gained by such an investment. An alternative compromise might be to track a relatively small sample of nonrespondents, as suggested by Graham and Donaldson (1993). Then characteristics of these subjects may be used for modeling adjustments. Also, tracking would still be important if information is especially sparse for specific subgroups in the sample. Lastly, these results support the importance of evaluating the effects of assumptions underlying any approach to nonresponse adjustment.

6. REFERENCES

Ellickson, P.L. and Bell, R.M. (1990). Drug prevention in junior high: A multi-site, longitudinal test.
Science, **247**.

Ellickson, P.L., Bell, R.M., and McGuigan, K.M. (1993). Drug prevention over 6 years: Long-term results from a multi-site test. *American Journal of Public Health.*

Graham, J. W. and Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, **78**(1): 119-128.

Heckman, J. J.(1979). Sample selection bias as a specification error. *Econometrica*, **47**(1): 153-161.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under non-standard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**: 221-233.

Kalton, G. (1983). *Compensating for Missing Survey Data.* Ann Arbor, MI: Institute for Social Research.

Lee, J. P., Ward, M.M., and Fries, J.F. (1993). Reducing attrition bias with an instrumental variable in a regression model: Results from a panel of rheumatoid arthritis patients. *Statistics in Medicine,* **12**: 1005-1018.

Little, R. J. and Rubin, D. B. (1987). *Statistical Analysis With Missing Data.* New York: John Wiley and Sons, Inc.

Maddala, G.S. (1985). A survey of the literature on selectivity bias as it pertains to health care markets. In R.M. Scheffler and L.F. Rossiter (eds.) *Advances in Health Economics and Health Services Research.* Greenwich, CT: JAI Press, Inc.

Manning, W.G., Duan, N., and Rogers, W.H. (1987). Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics,* **35**: 59-82.

Olsen, R. J. (1980). A least-squares correction for selectivity bias. *Econometrica,* **48**(7): 1815-1820.

Pirie, P.L., Murray, D.M., and Luepker, R.V. (1988). Smoking prevalence in a cohort of adolescents, including absentees, dropouts, and transfers. *American Journal of Public Health,* **78**(2):176-178.

Pirie, P.L., Thomson, S.J., Mann, S.L., Peterson, A.V., Murray, D.M., Flay, B.R., and Best, J.A. (1989). Tracking and attrition in longitudinal school-based smoking prevention research. *Preventive Medicine,* **18**: 249-256.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys* New York: John Wiley and Sons, Inc.

Stolzenberg, R. M., and Relles, D.A. (1990). Theory testing in a world of constrained research design: The significance of Heckman's censored sampling bias correction for nonexperimental research. *Sociological Methods and Research,* **18**(4): 395-415.

Table 1: Assessing weighting logistic regression model fit

Criteria for Assessing Model Fit

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	8630.218	7669.115	
SC	8637.002	7879.409	
-2 LOG L	8628.218	7607.115	1021.104 with 30 DF (p=0.0001)
Score			977.078 with 30 DF (p=0.0001)

Table 2: Validating estimates against "gold standard" (actual baseline levels of use)

	Sample and Estimation Method					
	Actual (N=6527)	Respondents (N=4087)	Respondents + Tracked (N=5365)	Weighting (N=4087)	Sample Selection #1 (N=6527)	Sample Selection #2 (N=6527)
Alcohol	2.676 (.0349)	2.489 (.0425)	2.602 (.0375)	2.684 (.0537)	1.341 (.0397)	2.487 (.0130)
Cigarettes	1.960 (.0373)	1.501 (.0402)	1.760 (.0382)	1.964 (.0656)	-.9385 (.0823)	1.887 (.0219)
Marijuana	.783 (.0263)	.457 (.0231)	.593 (.0247)	.798 (.0581)	-1.192 (--)	0.808 (.0238)

Table 3: Comparing estimates at three-year follow-up wave

	Sample and Estimation Method					
	Actual (N=6527)	Respondents (N=4087)	Respondents + Tracked (N=5365)	Weighting (N=4087)	Sample Selection #1 (N=6527)	Sample Selection #2 (N=6527)
Alcohol	n/a	5.146 (.0541)	5.194 (.0463)	5.299 (.0604)	4.313 (.2584)	5.269 (.0107)
Cigarettes	n/a	2.979 (.0592)	3.299 (.0539)	3.259 (.0722)	1.160 (.0526)	2.554 (.0119)
Marijuana	n/a	1.950 (.0498)	2.184 (.0450)	2.270 (.0645)	-0.292 (.0426)	1.431 (.0163)