The Use of Propensity Scores to Adjust for Nonignorable Nonresponse Bias

Leigh Ann Harrod, Virginia Lesser
Department of Statistics, Oregon State University

Abstract

Propensity score methodology has been applied in the analysis of survey data to account for differences in covariate distributions between respondents and nonrespondents. When a subsample of nonrespondents is obtained, response propensity scores may be used to account for a nonignorable missing-data mechanism. For a complex survey of elk hunters and a binary outcome, we examine nonresponse bias adjustment techniques using response propensity scores that are conditional on the outcome propensity.

Keywords: Missing data, nonresponse, propensity scores, not missing-at-random

1. Introduction

When unit nonresponse occurs in a survey, the associated estimates may be biased if the responding units are appreciably different from the nonresponding units (Lessler and Kalsbeek, 1992). For a finite population survey, let Y be the outcome variable, R be the indicator of response, and X be a set of fully-observed covariates associated with Y.

The mechanism that causes the nonresponse may be completely unrelated to the outcome, any related covariates, or the survey design. In this case, the data are missing completely at random (MCAR) and the obtained sample may be considered random and representative. In this case, Y and R are independent. If the response mechanism is related to the outcome variable but this relationship may be modeled on related covariates, the data are missing at random (MAR). Here, Y is independent of R conditional on X. Several analysis procedures are available for MAR missingness, such as poststratification adjustment, weighting class adjustment, and raking (Oh and Scheuren, 1983). When the mechanism that causes the missingness is related to the outcome of interest, the mechanism is referred to as not missing at random (Little and Rubin, 2002). If data are not missing at random (NMAR), a subsample of the nonrespondents must be obtained or a model of nonrespondent outcomes must be assumed.

1.1 Propensity score methodology

Propensity score methodology was developed by Rubin and Rosenbaum (1983) so that units with similar covariates could be matched between treatment groups and an unbiased estimate of the average treatment effect may be obtained. Rubin and Rosenbaum (1983) define a balancing score as a function *b*(X) such that the covariates are independent of the response mechanism conditional on the balancing score, or Y is independent of $R|b(X)$. Rubin and Rosenbaum (1983) state that the finest balancing score is the complete set of covariates X, and the coarsest balancing score is the propensity score, defined as in Little (1986) as the probability of response given the covariates:

$$p(X) = pr\left(R = 1 \middle| X\right).$$

The response propensity is generally unknown and must be estimated, usually by logistic regression of the response indicator on the covariates X.

Little (1986) found that adjustments for MAR data could be made within adjustment cells based on the response propensity. Coarsening the response propensity into five classes creates a response propensity score that may be used as adjustment classes. Within each propensity score class, all units have similar probabilities of response. Little (1986) determined that response propensity classification controls bias but not variance. He also examined the effectiveness of creating adjustment classes based on the predicted mean. Predicted mean stratification was found to control both bias and variance but weighting techniques may yield biased estimates.

1.2 Joint classification

Vartivarian and Little (2002) examined the case in which data are MAR but a wealth of covariate information exists. Here, using all of the covariates to create adjustment classes is problematic because some cells contain no respondents or estimated response probabilities are highly variable. Vartivarian and Little (2002) found that creating two sets of adjustment cells based on the predicted mean and the response propensity had a favorable property that Vartivarian and Little (2002) called "double robustness." If at least one model is correctly specified, then a reduction in bias occurs. If the predicted mean model is correctly specified, then there are additional increases in efficiency.

In this paper we consider the scenario in which data are NMAR and a subsample of nonrespondents is available. Our objective is to extend the technique of Vartivarian and Little (2002) to the case of NMAR data when the outcome of interest is a binary response.

2. Conditional response propensity score methodology

When data are NMAR, the response mechanism depends on the outcome of interest and this relationship cannot be fully described by the information obtained in the sample. When a subsample of nonrespondents is attainable, this additional information is useful in determining if the outcome of interest differs considerably between the respondents and nonrespondents.

Assume that the outcome of interest is a binary random variable. Then a predicted mean model could be obtained using a logistic regression of the outcome indicator on related covariates. The predicted mean is here analogous to estimating the success propensity conditional on covariates. Success propensity scores may then be obtained and used as adjustment classes. In the proposed method, the success propensity scores are used as a covariate in a model to estimate response propensity. The conditional response propensity is calculated as conditional on success propensity because data are NMAR and the response mechanism is not independent of success. Both sets of propensity scores (i.e. conditional response propensity scores and success propensity scores) may be used as adjustment variables. This approach differs from the Vartivarian and Little (2002) methods in that they estimate the response and success propensities independently for MAR data, while we assume NMAR data and account for the dependence of the response mechanism on success.

3. Case study: New Mexico elk hunter questionnaire

3.1 Background

New Mexico Department of Game and Fish (NMDGF) oversees annual elk hunts. When a person purchases an elk license in New Mexico, s/he receives a questionnaire with her/his license. The survey instrument includes questions on the licensee's hunt activity, success, and effort, and questionnaire return is not mandatory. NMDGF managers have felt that successful hunters are more likely to return their surveys, causing overestimation of elk harvest from the returned surveys. For this reason, a nonrespondent subsample was conducted to determine if bias exists in the elk harvest estimates and to adjust for any significant bias that might be found.

For the 2001-02 hunt year, 38,211 people bought licenses to hunt elk in the state of New Mexico and 29.5% of these licensees returned completed surveys. A random subsample of 3,078 of the 26,953 nonrespondents representing roughly a third of all elk hunts was obtained using a two-stage cluster sample with individual hunts as primary sampling units and licensees within each hunt as secondary sampling units. All hunts were stratified by *weapon type* (rifle, bow, muzzle-loader, or impaired), *landowner type* (public or private), and *hunt size* (see Table 1).

Table 1: Hunt size categories

Hunt size	*Hunt size category*	*Hunt size code*	*Sampling rate*
< 30 licensees	Small	1	1.00
31 to 170 licensees	Medium	2	0.30
>170 licensees	Large	3	0.15

Three mailing waves were used to contact the subsampled licensees. Licensees who had still not responded were telephoned up to five times. A total of 188 licensees had out-of-date contact information and were removed from the subsample. The overall response rate for the nonrespondent subsample after accounting for undeliverables was 82.2%.

This survey was repeated in the 2003-04 hunt year using the same survey design. Of a total nonrespondent subpopulation of 30,690 licensees, 2,747 licensees were subsampled and an 81% response rate was observed after accounting for undeliverables.

3.2 Adjusted estimates of harvest

For each subsample, some nonresponse still remained after three mailing waves and up to five telephone attempts. In each year, less than five licensees refused to participate in the telephone survey. Chi-square tests of association revealed that the age of the licensee was related to his/her response to the subsample survey ($\chi^2 = 1066.87$, df = 4, p<0.0001). The remaining nonresponse in the subsample was assumed to be unrelated to the outcome of interest and a poststratification adjustment on age class was used to account for the missing outcomes.

Because the population of licensees is censused, the subpopulations of respondents and nonrespondents are completely known. The results from the nonrespondent subsample were used to estimate the harvest for the subpopulation of nonrespondents only, and these estimates were combined with the reported harvest from the questionnaire respondents to obtain statewide estimates of elk harvest. The adjusted estimate of harvest is then a model-assisted estimator because it is calculated from the stratified two-stage cluster design but also employs a poststratification adjustment to account for the nonresponse in the subsample. Table 2 provides the estimates of harvest from the original questionnaire and the model-assisted estimates based on the nonrespondent subsample. It is important to note that the model-assisted estimator is likely still subject to nonresponse error; however, the estimates are considered an improvement on the estimates calculated from the original questionnaire survey because they incorporate the information from the nonrespondent subsample.

Table 2: Elk harvest estimates

Data source	*2001*	*2003*
Original questionnaire	13520 (SE = 134)	14323 (SE = 159)
Model-assisted estimator	9915 (SE = 183)	10304 (SE = 223)

3.3 Assessment of relative difference

The relative difference between the original questionnaire estimate and the model-assisted estimate was estimated for the 2001 harvest as 26.7% (SE = 1.4%) and for the 2003 harvest as 28.2% (SE = 1.4%). Therefore, game managers appear to have been quite justified in their concerns that the elk harvest questionnaire was overestimating annual elk harvest.

Nonrespondent subsamples are expensive and not feasible for every year. NMDGF game managers would like to develop a method of bias adjustment that will allow correction of past and future estimates and estimates of domains such as bag type (for example, harvests of mature bulls versus cow harvests). Given the consistency of the relative difference for the two years examined (z=0.76, p-value = 0.2243), creating an adjustment model to use retroactively is reasonable if hunts remain fairly consistent over the years for which adjustment will occur.

4. Application of the conditional response propensity methodology to the NMDGF elk harvest data

The conditional response propensity method is applied to the NMDGF elk harvest data. The objective is to develop a model of success that is robust enough to use in years for which a nonrespondent subsample is not available. Success propensities scores may be obtained and used in a model of response for a particular year using the response indicator observed during that hunt year. This allows factors that affect response to change over time but assumes that the factors that affect success do not change from year to year.

4.1 Model selection

The 2001 data were used to develop the success propensity model, and the 2003 data was used to test the performance of the 2001 model. The indicator of response to the original questionnaire was included as a covariate in the success propensity model because *(i)* these data are available for the entire population of licensees and *(ii)* response and success are related for NMAR data. Models for success and response conditional on success were selected using the Bayes Information Criterion (BIC).

For the 2001 data, the best model to predict success as determined by BIC included the indicator of response to the initial questionnaire, weapon type, landowner type, bag type, and month. The success propensity scores estimated from the covariates in the model just described was found to be the only significant covariate in modeling 2001 questionnaire response. Therefore, for the 2001 data, there is only one set of distinct adjustment classes because the success propensity scores and the conditional response propensity scores are equivalent. For the 2003 data, the model for initial response to the 2003 questionnaire depended on the success propensity score and weapon type. Here, joint classification on the success propensity score and the conditional response propensity score was possible.

4.2 Results

To compare the effect of using the conditional response propensity versus independent modeling of the success and response, we present the estimates from the methods used by Vartivarian and Little (2002) for MAR data as well as the estimates from the conditional response propensity approach for NMAR data. Vartivarian and Little (2002) use observed response probabilities within each adjustment class as estimates of response rate; this is analogous to a poststratification adjustment. We use the poststratification adjustment for the adjustment classes based on propensity scores; however, the associated standard errors will be underestimated because the variance components for modeling and coarsening each set of scores was not incorporated into the variance form.

When success and response models are developed independently, the model estimates are actually larger than the estimates from the original questionnaire. This is likely due to the overestimation of success rates for certain subgroups that reported high success rates in the original questionnaire. Because the missingness is NMAR, the development of independent models for success and response propensities does not account for the association between success and response. The conditional response propensity method is effective in reducing the relative difference in elk harvest estimates.

Comparing Table 3 results to those in Table 2, one may observe that the 2003 estimate from the conditional response propensity method is still large compared to the model-assisted estimator found from the nonrespondent subsample. This discrepancy may be due to a different underlying success mechanism, and the success model calculated from the 2001 subsample data was not robust enough to accurately model success in 2003. More modeling will be needed to develop a robust model that may be used in years for which a subsample is not available. However, the 2001 success model produces 2003 harvest estimates more similar to the 2003 model-assisted estimates. This improvement is conservative, which is acceptable to NMDGF game managers who will base future license numbers on these estimates.

Table 3: Adjusted elk harvest estimates

Modeling of success and response	*2001*	*2003*
Independent models	14537 (SE = 147)	15123 (SE = 173)
Conditional response model	9836 (SE = 140)	11854 (SE = 258)

5. Conclusions

When data are NMAR, additional information is needed to obtain unbiased estimates. The method proposed here represents a variation of a MAR adjustment technique to NMAR data. Propensity score methodology may be extended to include NMAR data if the association between the outcome of interest and the response mechanism is accounted for and accurately modeled. Application of this method to New Mexico Department of Game and Fish elk harvest data indicates that the method is effective in reducing bias for NMAR data. Further development will include incorporating variance components for the estimation and coarsening of propensity scores.

Current work includes developing a modified Horvitz-Thompson estimator that incorporates the conditional response propensity score and the use of Bayesian models when subsamples are not available.

Acknowledgements

This research was supported by the EPA STAR Cooperative Agreement, CR82-9096-01, through the DAMARS program at Oregon State University. The authors would like to acknowledge New Mexico Department of Game and Fish for their cooperation and dedication to obtaining unbiased estimates of elk harvest. We would also like to thank Dr. N. Scott Urquhart for his insights on the sampling design.

References

Lessler, JT and WD Kalsbeek. *Nonsampling Error in Surveys*, John Wiley and Sons, Inc: New York, 1992.

Little. RJA. Survey nonresponse adjustments for estimates of means. *International Statistical Review*, 1986; 54(2):139-157.

Little, RJA and DB Rubin. *Statistical Analysis with Missing Data*, 2nd Edition, John Wiley and Sons, Inc: New York, 2002.

Oh, HL and FJ Scheuren. Weighting adjustment for unit nonresponse. In *Incomplete Data in Sample Surveys*; WG Madow, I Olkin, and DB Rubin (eds), New York: Academic Press. 1983: 143-184.

Rubin, DB and PR Rosenbaum. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 1983; 70(1):41-55.

Vartivarian, S and Little, RJA. On the formation of weighting adjustment cells for unit nonresponse. American Statistical Association, 2002, Proceedings of the Survey Research Methods Section, 3553-3558.