An Application of Propensity Modeling: Comparing Unweighted and Weighted Logistic Regression Models for Nonresponse Adjustments

Frank Potter,[1] Eric Grau,[1] Stephen Williams,[1] Nuria Diaz-Tena,[2] and Barbara Lepidus Carlson[1]
[1]Mathematica Policy Research, Inc., Princeton, New Jersey
[2]TNS-Global, Princeton, New Jersey

Abstract

Using logistic regression models to predict the probability that a unit will respond is one method for adjusting for survey nonresponse. The inverse of the propensity score can be the weight adjustment factor. This method can make use of more predictive variables than in the weighting class method. Having used this method for two previous rounds of a large physician survey, this paper describes the results from the most recent round, round four. The logistic regression models used to estimate the propensity score were unweighted in round four, and the independent variables were expanded in round four to include design variables, basic sampling weights, and higher-order interactions. The predictive power of the propensity models was substantially improved over previous rounds, but also presented some interesting issues. The more effective models produced more extreme adjustment factors. In this paper, we evaluate the impact of using weights as covariates, rather than for weighting the models, on the adjustment factors and various measures of predictive power and goodness of fit.
Keywords: Nonresponse, weighting, propensity modeling, Community Tracking Study, physician surveys

1. Introduction

Unit response in surveys is almost inevitable. Not all sampled units will be willing to participate, and some may not even be given the chance to agree to participate because they cannot be located and/or contacted. To compensate for nonresponse, we use weighting adjustments so that those that did respond to the survey better represent the entire sample. When combined with the sampling weights that account for the probability of selection into the sample, the nonresponse adjustments can make the respondents represent the sampling frame.

There are various types of nonresponse adjustments that are commonly used. One can form weighting cells consisting of sample members thought to be similar in terms of response propensity, and then weight up the respondents to represent themselves and the nonrespondents in the cell. Or one can model response propensity using a logistic regression model, and use the inverse of the propensity score as an adjustment factor. A third method combines the two methods above, using the propensity score resulting from the logistic regression model to form weighting cells, and then weighting up the respondents within each cell.

The first method, the weighting cell adjustment, places restrictions on the variables that can be used to form cells. First, the variables must be available for both respondents and nonrespondents. Second, the cells must be large enough so that there are a sufficient number of respondents within the cell (say 20 or more), making the adjustment factor stable. The first restriction on variables applies to all methods of nonresponse adjustment, but the second restriction is a particular limitation of the weighting cell approach, where weighting cells are formed based on design variables and/or descriptive variables available for all sample members. When cells must be collapsed to ensure sufficient numbers of respondents in the cell, the variance of the weight-adjusted estimate may be reduced, but the bias may increase.

The second method, the inverse propensity score adjustment, allows for more variables to be used to predict nonresponse. One must find the best-fitting model to predict nonresponse using the most appropriate set of available variables. This results in a "smooth" distribution of adjustment factors (see Carlson and Williams, 2001), with no requirement to select arbitrary cutpoints. However, the propensity score can have extreme values, creating adjustment factors that can in turn create highly variable weights, and therefore highly variable weight-adjusted estimates. This can be remedied by either trimming the adjustment factor, or trimming the nonresponse-adjusted weight. But these remedies increase the possibility of bias. Moreover, this method puts considerable faith in the correct specification of the model, which Little (1986) had warned was an additional disadvantage.

The third method, using propensity scores to form weighting cells, falls in between the two methods in terms of limitations and variance-bias tradeoffs.

More descriptive variables can be used to predict nonresponse than can be used in the traditional weighting cell approach. And extreme propensity scores are winsorized by forming cells based on quantiles of the score, in effect implicitly trimming the factor. The adjustment factor within cell is generally the inverse of the response rate within cell or the mean inverse propensity score for the cell. The disadvantage of this method is that arbitrary cutpoints based on quantiles are used to categorize the propensity scores. It is possible to end up with very different adjustment factors between weighting classes, and the same adjustment factors within weighting classes, even though differences between covariate values across weighting class boundaries might be small. Moreover, this method groups respondents together in the same weighting classes who may be dissimilar in every other way, but have similar propensity scores.

The question of which method to use has been reviewed by others. Clusen et al. (2005), Rizzo et al. (1994), and Carlson and Williams (2001), found no major differences between a variety of weighting methods, including the three given above. Rizzo et al. (1994) and Clusen et al. (2005) conclude that the choice of variables is more important than the weighting methodology.

In all three methods, there is a choice of whether to use sampling weights—the inverse of the probability of selection—as part of the process. For the weighting cell approach, one generally uses the inverse of the response rate within cell as the adjustment factor, but should this be a weighted or unweighted response rate? For the inverse propensity score approach, should the model be weighted by the sampling weight, or should the model be unweighted? If unweighted, should the sampling weight be included as a covariate in the model? For the propensity score-based weighting cell approach, all of the above questions apply.

In Little and Vartivarian (2003), they conclude based on a simulation that cell-based approaches should use unweighted inverse response rates for the weighting adjustment, and that the cells should be formed in such a way that they not only predict response propensity, but also predict key survey variables. The cells should make use of design variables (which would in effect account for differential probabilities of selection) and variables that are available for all samples members, regardless of response status. They infer from these findings that weighting a logistic regression model to predict nonresponse does not offer any advantages over using an unweighted model.

2. Methods

2.1 Analysis

We decided to evaluate whether weighting the response propensity model made sense for a physician survey we have been involved with for a number of years, the Community Tracking Study (CTS) Physician Survey. In prior rounds of the CTS physician survey, we used the inverse propensity score as the adjustment factor for nonresponse, and the model used to derive the propensity score was weighted by the sampling weight. Based partly on the recommendations in the 2003 Little and Vartivarian paper, we decided to try a new approach for round four. We would use an unweighted model, but include the categorized sampling weight as a covariate in the model. We also would include design and operational variables as covariates, which was not done in prior rounds. When we did this, we saw larger adjustment factors (larger inverse propensity scores) than had been seen in prior rounds. We wondered whether these larger adjustment factors were due to our new unweighted model approach, so we decided to try to disaggregate the impact of (a) using an unweighted vs. weighted model, and (b) including the sampling weight as a covariate within the unweighted models.

There are two stages of nonresponse adjustments for the physician survey: (1) adjusting for those physicians who could not be located, and (2) adjustment for nonresponse among located physicians. We used the CHAID procedure[1] to find covariates that predicted the two types of nonresponse, including design and operational variables. Interaction terms that were found to be significant by the CHAID procedure were included in the pool of covariates considered for the final model.

In the round four processing of the Physician Survey, we used an unweighted forward stepwise logistic regression procedure from SAS to select variables, where the original pool of variables included the design variables (sampling weights, stratification variables, and PSU identifiers) in both the location and cooperation model. This procedure indicates the significance of main effects, second and third order interactions when they are introduced into the model. We obtained a full logistic regression model using the more significant main effects, second and third

[1] Chi Squared Automatic Interaction Detector, discussed in Biggs et al. (1991) and Magidson (1993).

order interactions. Any combination of main effects and second order interactions involved in the third order interactions was included in the full model, regardless of its significance. For the final weighted logistic regression models, we used SUDAAN, which computes the correct sampling variances for the estimates of the models and takes into account the sampling design of the survey.

In addition to the design variables, the variables included in the pool of covariates considered for the regression models included: age, gender, nature of practice (solo, partnership, group, hospital, etc.), number of calls required to locate (or attempt to locate) the physician, geographic location (Census region or division), specialty, time between the release of the sampled case and the date the case was completed (or the end of the processing for that round); and binary indicators of whether (1) the physician was an MD or osteopath; (2) a phone number could be found for the physician; (3) the physician was board-certified; (4) the physician attended medical school in the United States; and (5) the physician participated in an experiment investigating pre-paying the physicians taking part in the survey. Besides these variables, second and third order interactions were included if significant in the model.

For each stage, we ran three models using the covariates selected in the process described above for the unweighted model with design variables. This resulted in a total of six models to be compared. For each model, we looked at the fit of the model (using the generalized R-square, the concordance rate, and the significance of the Hosmer-Lemeshow statistic), the sizes of the adjustments (maximum adjustment factor and number of large adjustment factors), and the sizes and variability of the nonresponse-adjusted weights (maximum weight, weight at the 99th percentile, and the design effect due to unequal weighting). Because we found no bias in estimates across the three types of models (Grau et al. 2006), we focused only on the variance component of the mean square error in this paper.

2.2 Survey Data

The CTS is designed to provide a sound information base for decision making by health care leaders. It does so by collecting information on how the health system is evolving in 60 nationally representative communities across the United States and on the effects of those changes on people. The CTS, which has been under way since 1996, is a longitudinal project that relies on periodic site visits and surveys of households, physicians, and employers. The CTS addresses two broad questions that are important to public and private health decision makers:

1. How is the health system changing? How are hospitals, health plans, physicians, safety net providers, and other provider groups restructuring, and what key forces are driving organizational change?

2. How do these changes affect people? How are insurance coverage, access to care, use of services, health care costs, and perceived quality of health care changing over time? The CTS includes independent surveys of households, physicians, and employers in all 60 sites, thereby enabling researchers to explore relationships among purchasers, providers, and consumers of health care at the site level. The CTS is sponsored by the Center for Studying Health System Change (www.hschange.org), a nonpartisan policy research organization committed to providing objective research on the nation's changing health care system. The CTS is funded by the Robert Wood Johnson Foundation.[2]

The physician survey is designed to document changes that allopathic (MD) and osteopathic (DO) physicians are experiencing in the health care system and to learn how these changes are affecting physicians, their practices, and the way they deliver medical care to their patients. The goal is to provide information to public and private leaders that will enable them to make better policy decisions. Some of the analytic areas include:

-Impact of managed care participation on physician behavior, perceptions of quality of care provided and physician satisfaction.

-Effects of physician practice arrangements, ownership and risk-bearing on the practice of medicine.

-Relationships between the distribution of practice revenue and physician practice style and satisfaction.

[2] Under HSC's direction, Mathematica Policy Research (MPR) designed the household and physician survey samples and weights. MPR conducted the household survey and Gallup collected the physician surveys. Final data processing and file production were carried out by Social and Scientific Systems.

-Effects of socio-demographic or market factors on physicians' practice revenues or income.

-Impact of federal, state and local policies affecting physician practice (including Medicare and Medicaid policy) on physician behaviors and perceptions of their impact on quality of care.

The survey is a nationally representative telephone survey of non-federal, patient care physicians in the 48 contiguous United States and the District of Columbia. Eligible physicians included MDs and DOs who had completed their medical training, provide 20 or more hours of direct patient care per week, and practice in an office or hospital. There were two key domains: primary care physicians (family physicians, general practitioners, internists, and pediatricians) and specialist (for specialties designated as eligible).

Each round of the Physician Survey contains observations from more than 12,000 physicians who spend at least 20 hours a week in direct patient care. The sampling frame is comprised of the American Medical Association Masterfile (which contains MDs and the majority of DOs in the U.S.) and a list of DOs from the American Osteopathic Association. The CTS had a multistage sample design, where the primary sampling units were 60 sites. In the first three rounds of CTS, approximately 90 percent of the interviews were collected from physicians practicing in the 60 CTS sites; the remaining 10 percent were with physicians selected from a nationally representative supplement designed to improve national estimates. In round four, all physicians included in the survey were those practicing within the 60 sites. Round four had 5,828 completed interviews. Unlike prior rounds, round four had approximately proportional allocation across the 60 sites (to better achieve national estimates), whereas prior rounds had disproportionate sampling across sites (to optimize site-specific estimates).

The majority of interviews in all but the first round were conducted with prior round respondents, to allow for panel estimates; however, a refresher sample was added each round for the purposes of cross-sectional estimates. For this methodological paper, we only included those physicians who were new to the round four sample because they appeared for the first time on the round four frame (although the model itself also included physicians who were in the round four sample for the first time because they were not selected in round three). This sample component had variability in the sampling weights, mostly due to cross-site differences in probabilities of selection. Limiting the sample to the new physicians avoided the additional sampling and weighting complexities associated with the "overlap" sample, and allowed us to focus on the more straightforward sampling weights. More importantly, this sample component is most similar to the type of sample that others would encounter in their own surveys, thereby making our findings more generalizable.

3. Results

Table 1 shows the results for the first stage of nonresponse adjustments, that accounts for physicians that could not be located. The three models (unweighted model with categorized sampling weight as covariate, unweighted model without weight as covariate, and weighted model) had fairly comparable fits. All had R-square values around 0.3, concordance rates of about 92 percent, and all had Hosmer-Lemeshow statistics that were not statistically significant. The weighted model had the lowest maximum adjustment factor (15, compared to 18 and 19 for the unweighted models), and had the lowest number of factors that were larger than 10 (five such cases, compared to 10 or 11 for the unweighted models). The weight that combined the sampling weight and the adjustment for locatability was also lowest for the weighted model. The weighted model had a lower maximum weight (732, compared to over 900 for the unweighted models), a lower value at the 99th percentile (134, compared to about 145 for the unweighted), and a lower design effect due to unequal weighting (1.8, compared to about 2 for the unweighted models). Note that the design effect due to the sampling weights alone was 1.276.

Table 2 shows comparable statistics for the weighting adjustment due to nonresponse among located physicians. The three models had fairly comparable fits. All had R-square values around 0.4, concordance rates of about 88 percent, and all had Hosmer-Lemeshow statistics that *were* statistically significant. Unlike the locatability adjustment, the weighted model had the *highest* maximum adjustment factor (42, compared to 31 and 34 for the unweighted models), and had the *highest* number of factors that were larger than 10 (44 such cases, compared to 36 or 39 for the unweighted models), and had nine factors that were greater than 20. The weight that combined the sampling weight, the adjustment for locatability, and the adjustment for response was also lowest for the weighted model, by two of the three measures. The weighted model had a lower maximum weight (1,159, compared to over 2,400 for the unweighted models), a *higher* value at the 99th percentile (611, compared to about 550 for

the unweighted), and a lower design effect due to unequal weighting (3.2, compared to about 4.5 for the unweighted models). This second adjustment factor, for nonresponse among located physicians, had less of a consistent pattern when compared to the factor to adjust for the inability to locate physicians.

4. Discussion

While we were looking for evidence that the unweighted model caused the large adjustment factors, our findings do not necessarily support that assertion. The unweighted model seemed to result in a slightly larger nonresponse adjustment, but not enough to explain these findings. Instead, the inclusion of design and operational variables, also new to round four, may have "over-fitted" the model. All three models had good fits (high R-square and concordance values), but also had large adjustment factors, which increase the variance and the need for trimming (and the resulting potential bias). Care needs to be taken in variable selection and modeling.

References

Biggs, D., B. de Ville, and E. Suen. "A Method of Choosing Multiway Partitions for Classification and Decision Trees," *Journal of Applied Statistics*, vol. 18, 1991, pp. 49-62.

Carlson, B.L., and S. Williams. "A Comparison of Two Methods to Adjust Weights for Nonresponse: Propensity Modeling and Weighting Class Adjustments." Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]. Alexandria, VA: American Statistical Association, 2001.

Clusen, N. A., H. Xu, and M. Hartzell. "Adjusting for Nonresponse in the Healthcare Survey of DoD Beneficiaries." Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]. Alexandria, VA: American Statistical Association, 2005.

Grau, E., F. Potter, S. Williams, N. Diaz-Tena. "Nonresponse Adjustment Using Logistic Regression: To Weight Or Not To Weight?" Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]. Alexandria, VA: American Statistical Association, 2006.

Little, R.J.A. "Survey Nonresponse Adjustments for Estimates of Means." *International Statistical Review*, vol. 54, 1986, pp. 139-157.

Little, R.J.A., and S. Vartivarian. "On Weighting the Rates in Non-Response Weights." *Statistics in Medicine*, vol. 22, 2003, pp. 1589-1599.

Magidson, J. (1993). "SPSS for Windows CHAID Release 6.0." Belmont MA: Statistical Innovations Inc.

Table 1. Nonresponse Adjustments for Locatability Among Sampled Physicians

	Unweighted Model Using Sampling Weight as a Covariate	Unweighted Model Not Using Sampling Weight as a Covariate	Weighted Model Not Using Sampling Weight as a Covariate
Model Fit			
R-square	.323	.320	.304
Concordance rate	92.3	92.0	91.9
Hosmer-Lem. p-value	.42	.75	.67
Adjustment Factors			
Maximum	19.8	18.3	15.1
Count >10 / >5	10 / 25	11 / 24	5 / 23
Adjusted Weights			
Maximum	912	981	732
99^{th} percentile	141	146	134
Deff due to weighting	1.97	2.05	1.77

N.B. For reference, the design effect due to unequal weighting for the unadjusted sampling weight is equal to 1.276.

Table 2. Nonresponse Adjustments for Response Among Located Physicians

	Unweighted Model Using Sampling Weight as a Covariate	Unweighted Model Not Using Sampling Weight as a Covariate	Weighted Model Not Using Sampling Weight as a Covariate
Model Fit			
R-square	.413	.411	.382
Concordance rate	88.5	88.4	88.2
Hosmer-Lem. p-value	<.01	<.01	<.01
Adjustment Factors			
Maximum	33.9	31.0	41.6
Count >20 / >10	4 / 36	3 / 39	9 / 44
Adjusted Weights			
Maximum	2,684	2,497	1,159
99^{th} percentile	588	537	611
Deff due to weighting	4.76	4.45	3.21

N.B. For reference, the design effect due to unequal weighting for the unadjusted sampling weight is equal to 1.276.