Nonresponse Adjustment Using Logistic Regression: To Weight or Not To Weight?

Eric Grau[1], Frank Potter[1], Steve Williams[1], and Nuria Diaz-Tena[2]
[1]Mathematica Policy Research, Inc., Princeton, New Jersey 08543-2393
[2]TNS, Princeton, New Jersey 08540

Abstract

Unit nonresponse in sample surveys is accommodated by reallocating the weights of unit nonrespondents to respondents. One way of doing this is to develop logistic regression models to predict the probability of response. The inverses of the predicted probabilities from these models are then used to adjust the sampling weights. In rounds two and three of the Community Tracking Study (CTS) Household and Physician Surveys, nonresponse adjustments to the weights were carried out using weighted logistic regression models. In the fourth round of the survey, unweighted logistic regression models were used to adjust for nonresponse, with design variables, basic sampling weights, and higher order interactions included in the models, following a methodology introduced in papers by Vartivarian and Little (2003). In this paper, we compare nonresponse adjustments using the two methods.

Keywords: Nonresponse, weighting, propensity modeling, Community Tracking Study, physician surveys

1. Introduction

In sample survey methodology, weights are estimated as the inverse of the probability of selection. When there is unit nonresponse, these weights are commonly adjusted by a nonresponse weight (called an adjustment factor), which is the inverse of the probability of response. This probability is called a propensity score ϕ, and can be estimated using either weighting classes directly, or using logistic regression models (Little 1986). In the latter case, estimated propensity scores are often grouped into weighting classes, and the nonresponse weight recalculated, as was done in Smith et al (2001). This is done to reduce the variability that could potentially occur with individual propensity scores and to avoid complete reliance on the correct specification of the response propensity regression model (Little, 1986).

The logistic models used to compute the scores reflect the propensity to respond based on attributes of both respondents and nonrespondents. These propensity scores can be estimated with or without weights. In the case of weighting classes based directly on these covariates, the propensity score in weighting class c can be estimated by

$$\phi_c = r_c / n_c \qquad (1)$$

where r_c denotes the number of respondents in weighting class c, and $n_c$denotes the total number of sampled cases in weighting class c. An alternative estimate is given by

ϕ_c (2)
= (sum of weights for respondents in class c) /
(sum of weights for selected sample in class c)

Little and Vartivarian (2003) call (1) and (2) the unweighted and weighted response rates respectively. They argue that, although (2) is an unbiased estimate of the population response rate in weighting class c, this estimate does not ensure unbiased estimates of the variables of interest. In particular, they assert that the correct approach is to use (1) for weighting classes that condition on both covariate and design information. If weighting classes are created that are homogeneous with respect to the propensity to respond, then using (2) is unnecessary and inefficient. Moreover, they argue, if the weighting classes are not homogeneous with respect to the propensity to respond, then (2) will not yield unbiased estimates of the means of population outcomes.

If logistic regression models are used to estimate ϕ_c, then the unweighted estimate is equivalent to a predicted mean from an unweighted regression model. Moreover, the weighted estimate is equivalent to the weighted model predicted mean. In this paper, we evaluate whether bias or variance is greater with unweighted or weighted models in a Physician Survey. In particular, we address the question of whether propensity scores should be developed from weighted or unweighted models.

In this study the nonresponse weight is obtained directly from the propensity score obtained from the model. This is contrary to the recommendation of some (see, for example, Little, 1986). However, it does result in a “smoother” distribution of adjustment factors. (See Carlson and Williams, 2001.) In other words, with the weighting cell approach, there is a danger of having very different adjustment factors between weighting classes, and the same adjustment factors within weighting cells, even though differences between covariate values across weighting classes might be small. Indeed, the use of weighting classes requires a choice of arbitrary cutpoints, where adjustment factors on different sides of these arbitrary cutpoints might be large. Moreover, you are not limited by minimum cell sizes and ratios of responders to nonresponders, allowing for a greater pool of variables to be used in the nonresponse adjustment process. Finally, it avoids grouping respondents together in the same weighting class who are dissimilar in every other way, but have similar propensity scores. In order to avoid unnecessary increases in variance, each variable used in the models had at least twenty observations for each level. Variables resulting in very large or very small adjustment factors were removed from the models. A limited amount of trimming was performed after the adjustments were applied to avoid large variances associated with outlier weights resulting from large adjustments, alleviating the problem of large weights. Clusen et al (2005), Rizzo et al (1994), and Carlson and Williams (2001), found no major differences between a variety of weighting adjustment methods, including using raw propensity scores and using weighting classes based on propensity scores. Rizzo et al (1994) and Clusen et al (2005) conclude that the choice of variables is more important than the weighting methodology.

2. The Community Tracking Study Physician Survey, Round Four

The Community Tracking Study (CTS), which is funded by the Robert Wood Johnson Foundation, is designed to provide a sound information base for decision making by health leaders. It does so by collecting information on the United States health system, and how it is evolving, as well the effects of those changes on people. Begun in 1996, the CTS, is a longitudinal project that relies on periodic site visits and surveys of households, physicians, and employers. This survey consists of two samples, a site sample and a supplemental sample. The site sample is a national survey of 60 locations in the United States: 48 large Metropolitan Statistical Areas (MSAs), 3 small MSAs, and 9 non-MSAs. The supplemental sample includes all 48 contiguous states stratified in 10 different regions, as described in Potter et al. (2000). In Round Four of the Physician Survey, no supplemental sample was implemented, and was not considered for this study.

In the Physician Survey, there were three different subgroups of physicians for the site sample based on their Round Three interview status: (1) Round Three interviews (reinterviews), corresponding to physicians who completed the Round Three interview; (2) Round Three noninterviews, corresponding to physicians who were selected for the Round Three sample but who did not complete the interview for reasons such as ineligible, refusals or not located; and (3) new interviews in Round Four, corresponding to two groups of physicians, a) physicians in the Round Three sampling frame who were not selected for the Round Three sample, and b) physicians who were new to the frame in Round Four. Separate nonresponse adjustments were done for each of the three subgroups, where ϕ_c was estimated directly using logistic regression models, which are discussed in Section 3.

For all sampled physicians, demographic, personal, and practice characteristics are available from the American Medical Association (AMA), and the American Osteopathic Association (AOA) files that were used as the sample frame.

3. Logistic Regression Models to Estimate Propensity Scores, Round Four

There are two main causes of nonresponding: (1) when the physician could not be located, and (2) when the physician refused to complete the interview. For each cause of nonresponding, we first examined the pattern of nonresponse relative to the data available on sample members. We used different models in each of the three subgroups (reinterviews, noninterviews, and new interviews) to accommodate the different rates of nonlocation and noncooperation in each of the subgroups. ("Noncooperation" refers to nonresponse given that the sample member was located.) Logistic models were used to predict the probability of locating a physician (propensity score for location) for each subgroup. We then used other logistic models to predict the probability that a located physician would respond (propensity score for cooperation). The inverse of the location and cooperation propensities resulting from the application of those models was then used as the adjustment factor to the weights. There were too few reinterviews that could not be located for a separate location model for this group, so the unlocated reinterviews were included in the nonresponse model. The result was a total of five models, with a location and cooperation model for each of the noninterviews and new interviews, and a single nonresponse model for reinterviews.

In accordance with Little and Vartivarian's recommendation, unweighted models were used to estimate propensity scores in each subgroup, with design information included in the model in the form of covariates. In particular, this included sampling weights, binary variables identifying five sets of PSUs (five categories of the sixty sites mentioned in Section 2, categorized using CHAID[1]) and a stratification variable. The weights were partitioned into 11 categories, which were subsequently collapsed to 9 categories for the location model. The stratification variable was based on the physician respondent's survey status in the previous (third) round and whether the physician was a primary care physician (PCP) or a specialist. The physician's status in the previous round was defined by: (1) physician was a complete case in Round Three; (2) physician was selected for Round Three, but did not complete the interview (includes ineligibles); (3) physician was on the Round Three frame, but was not selected in Round Three; (4) physician was not on the Round Three frame. These four levels for PCPs and for specialists defined the eight sampling strata. Since models were fitted separately for reinterviews, noninterviews, and new interviews, clearly not all of the levels were evident for each 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 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 their significance. The final full model was developed using standard model-fitting procedures, including reviewing measures of goodness of fit and predictive power and eliminating nonsignificant predictors.

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.

4. Methodology of Study

For the purposes of the study described in this paper, comparisons were limited to new interviews among physicians who were new to the frame in Round 4. This was done to avoid complications due to the longitudinal nature of the study and to have a sample that would be most comparable to that which other users would encounter. This group, which included two sampling strata (new PCPs and new specialists), had some variability in sampling weights due to cross-site differences in the probability of selection. The frame, which in Round Four contained 559,967 eligible physicians, contained only 87,499 who were newly eligible in Round Four. Response rates among all respondents, new respondents in Round 4, and new respondents in Round 4 who were new to the frame are given in Table 1. The proportion of physicians who had 7 key attributes were calculated from the reduced sampling frame: (1) Did the physician attend medical school in the United States? ; (2) Was the physician an M.D. or an osteopath?; (3) Was the physician under the age of 45?; (4) Was the physician an primary care physician or specialist?; (5) Was the physician a gynecologist?; (6) Does the physician's practice engage in direct patient care; (7) Is the physician board-certified? Whereas the variables above were obtained from the frame alone, the variables used in the fitting of the models were obtained from both the questionnaire and the frame.

As stated earlier, separate models were fitted for reinterviews, noninterviews, and new interviews, with separate location and cooperation adjustments for two of these three groups. For the new interviews, weighted point estimates and confidence intervals for the seven variables listed above were calculated for the sampled cases from Round Four processing who were new to the Round Four frame, using both the location-adjusted weights and nonresponse-adjusted weights. The estimates using location-adjusted weights included values for all sample members who were located, regardless of whether they responded, since values were taken from the frame. The difference between

[1] Chi Squared Automatic Interaction Detector, discussed in Magidson (1993)

each point estimate and the value from the frame gives a sense of the bias in the estimate, and each confidence interval provides information about the variance.

For the purposes of the comparison, we refit the location and cooperation models among new interviews, using weighted models and unweighted models with sampling weights not included as covariates. In order to ensure that estimates were comparable, models were fitted with all new interviews, whereas point estimates and standard errors for the seven listed variables above were calculated only amongst new interviews who were new to the frame in Round Four. Estimates from the weighted models were obtained using SUDAAN software, to appropriately accommodate the sampling design. In summary, there were three sets of point estimates for the seven variables listed above: (1) weighted estimates where the weights were response-adjusted using weighted models; (2) weighted estimates where the weights were adjusted using unweighted models with sampling weights as covariates (taken from Round Four processing); and (3) weighted estimates where the weights were adjusted using unweighted models without sampling weights as covariates. Each set included point estimates and standard errors using location-adjusted and nonresponse-adjusted weights. The results are shown in Figures 1-7. In these graphs, "U1" refers to the unweighted model from Round Four processing, with all design information included; "U2" refers to the unweighted model with sampling weights not included as covariates; "W" refers to the weighted model; "loc" refers to the weighted estimates using location-adjusted weights; and "nr" refers to the weighted estimates using cooperation-adjusted weights. Interval widths, which allow us to graphically compare the variances, are given in Figures 8-14. Even though point and interval estimates using location-adjusted weights are on the same graphs as the point and interval estimates using nonresponse-adjusted weights, they should not be compared. Nonresponders who were located were included in the point and interval estimates using location-adjusted weights, resulting in a much larger sample size. It should not be surprising that the bias and variance would be less with the location-adjusted weights than with the nonresponse-adjusted weights.

5. Results

As is apparent from Figures 1-7, there does not appear to be a significant advantage of one method over the others in terms of bias. In Figure 2, the point estimates using location-adjusted weights significantly differ from the frame value, but the significant bias appears across the board. There is some evidence that the bias increases with the cooperation model, as is shown in Figures 3, 5, 6, and 7, though this increase doesn't appear to be significant. This apparent increase in bias appears to be greatest with the unweighted models, particularly in Figures 6 and 7, but again this may just be due to random variation.

With estimates weighted using location-adjusted weights, the interval widths consistently show slightly larger variance with weights adjusted using unweighted models without sampling weights as covariates, as shown in Figures 8-14. No discernable difference is apparent between the interval widths of estimates using weight adjustments from weighted models and estimates using weight adjustments from unweighted models with sampling weights used as covariates. The unweighted model without sampling weights as covariates lead to estimates with largest variances using cooperation-adjusted weights in Figures 8, 9, 11, 13, and 14, but the estimates using cooperation-adjusted weights with the largest variance in Figures 10 and 12 are associated with the unweighted models with sampling weights.

6. Discussion

The Little and Vartivarian paper suggests that, if weighted estimates of the propensity score are used, then bias may creep in to estimates of variables of interest if weighting classes are not homogeneous with respect to the propensity to respond. In our case, weighting classes consist of single observations, or profiles of observations with the same values for the covariates in the models. We would not expect to see an increase in bias due to the weighting adjustments, and our results bear this out.

Little and Vartivarian also suggest that using weighted estimates of the propensity scores is inefficient and unnecessary. There is no evidence, at least in this limited example, of increased variance with weighted models. In fact, there is some (weak) evidence to the contrary. This does not, however, refute the conclusions of Little and Vartivarian. There are other factors that may be at play.

As is apparent in Tables 2-4, the largest adjustment factors resulting from using individual predicted probabilities as propensity scores are very large, potentially having a major impact on the variance. Any differences in variance that are due to using a weighted model instead of an unweighted model might be overshadowed by the variance caused by these large adjustment factors. The largest adjustments for the weighted models are usually smaller for the weighted model than for the unweighted models, which may explain the slight advantage observed for the weighted models. This does tend to support the argument that propensity scores should be grouped in weighting classes, to avoid large variances due to large adjustment factors. Indeed, in a study using the same data, Diaz-Tena et al (2002) showed (before poststratification and trimming) that using directly modeled propensity scores resulted in slightly higher variances than using weighting cells based on propensity scores. Any advantage in levels of bias did not offset this higher variance. (They still recommended estimating propensity scores directly from logistic regression models, due to the increased level of effort required due to creating weighting cells from propensity scores, for a seemingly small decrease in the variance.) In the actual processing of the Physician Survey, however, trimming was employed to reduce the deleterious impact on the variance caused by outlier weights, which may include weights that became outliers because of large adjustment factors.

The models were fit using unweighted stepwise regression with weights as covariates. To ensure comparability, the same parameters that came out of this procedure were used for the weighted regression and the unweighted regression without weights as covariates. It is possible, though unlikely, that applying the same model-fitting procedures to all three scenarios may have given different results. Also, large adjustment factors may be due to overfitting of the logistic regression models. Note that in the model-fitting process, few model-fitting procedures were implemented after reviewing the result from the automated procedures of the stepwise regression software. There was little effort to actually find a parsimonious model, which may have resulted in models that fit closely to the sample data, but did not represent the population as well. This may have increased the value of the adjustment factors in some cases, which in turn increased the variance, regardless of how or whether weights were incorporated in the models.

7. Further Research

The next step for this research is to find ways to explore this question while dampening the effect of large adjustment factors on the variance. The Little and Vartivarian paper draws conclusions based on weighting classes. One approach would be to confirm (or not) the Little and Vartivarian result on the Physician Survey data using either using weighting classes determined directly from covariates, or using weighting classes defined from modeled propensity scores. In the former case, CHAID could be used to determine the covariates defining weighting classes, as was done by Tambau et al (1998). A second approach would be to make comparisons after trimming was implemented. Trimming alleviates the issue of increased variance due to large adjustment factors.

References

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.

Diaz-Tena, N., F. Potter, M. Sinclair, and S. Williams. “Logistic Propensity Models to Adjust for Nonresponse in Physician Surveys. ” Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]. Alexandria, VA: American Statistical Association, 2002

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.

Potter, F., M. Sinclair, and S. Williams. “Examining Attrition in the Physicians Component of the Community Tracking Study.” Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM]. Alexandria, VA: American Statistical Association, 2000.

Rizzo, L., G. Kalton, M. Brick, and R. Petroni. “Adjusting for Panel Nonresponse in the Survey of Income and Program Participation.” 1994 Proceedings of the American Statistical Association, Survey Research Methods Section. Alexandria, VA: American Statistical Assocation.

Smith, P.J., J.N.K. Rao, M.P. Battaglia, T.M. Ezzati-Rice, D. Daniels, and M. Khare. "Compensating for provider Nonresponse Using Response Propensity to Form Adjustment Cells: the National Immunization Survey." *Vital and Health Statistics*, vol. 2 2001, pp.133.

Tambau, J.-L., I. Schiopu-Kratina, J.Mayda, D. Stukel, and S. Nadon. (1998). "Treatment of Nonresponse in Cycle Two of the National Population Health Survey." *Survey Methodology*, vol 24 1998, pp. 147-156.

Table 1: Physician Survey Response Rates, Overall, Among New Interviews, and Among New Interviews for Physicians New to Frame

			Weighted	Weighted %	Weighted
	Total	Weighted	%	complete	%
Subgroup	sample	Total	Located	among located	complete
Total	15,063	559,967	90.8	57.7	52.4
New	4,675	147,872	87.0	52.6	45.7
New to frame	3,435	87,499	84.4	56.2	47.5

Figure 1. Percent of Physicians Who Attended Medical School in the United States: Frame Value, Point Estimates, and Confidence Intervals

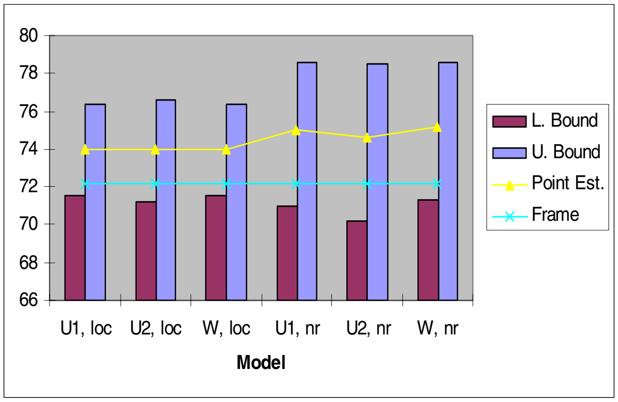


Figure 2. Percent of Physicians (MD's and Doctors of Osteopathy) Who Are MD's: Frame Value, Point Estimates, and Confidence Intervals

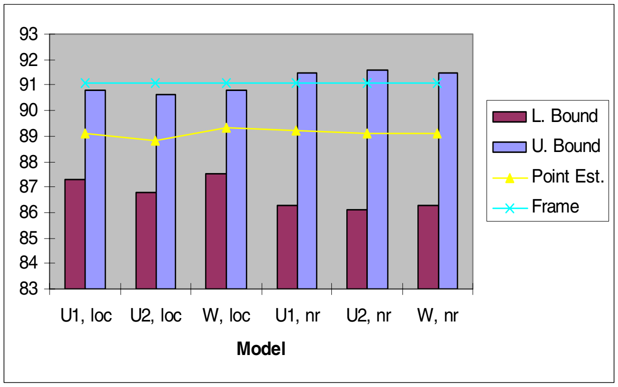


Figure 3. Percent of Physicians under Age 45: Frame Value, Point Estimates, and Confidence Intervals

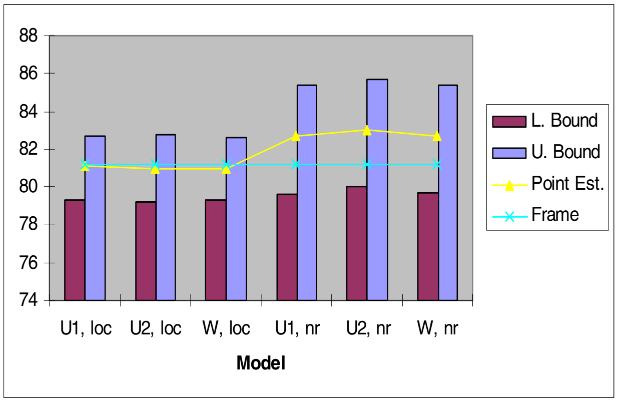


Figure 4. Percent Primary Care Physicians: Frame Value, Point Estimates, and Confidence Intervals

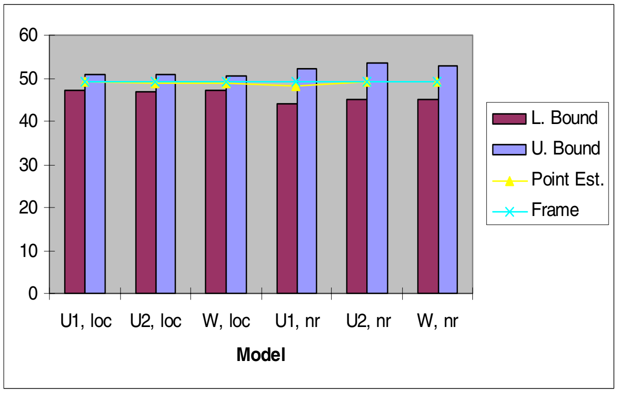


Figure 5. Percent of Physicians Who Are Gynecologists: Frame Value, Point Estimates, and Confidence Intervals

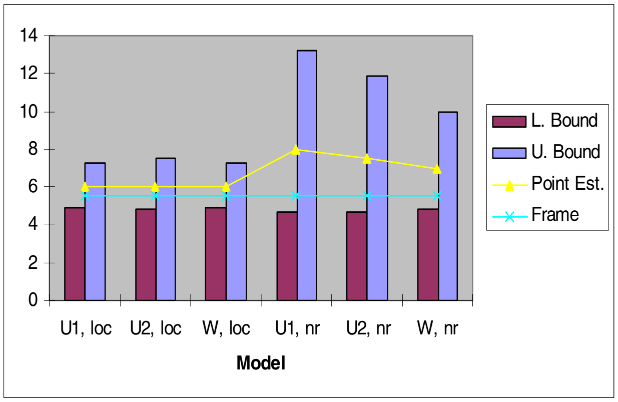


Figure 6. Percent of Physicians Whose Practices Engage in Direct Patient Care: Frame Value, Point Estimates, and Confidence Intervals

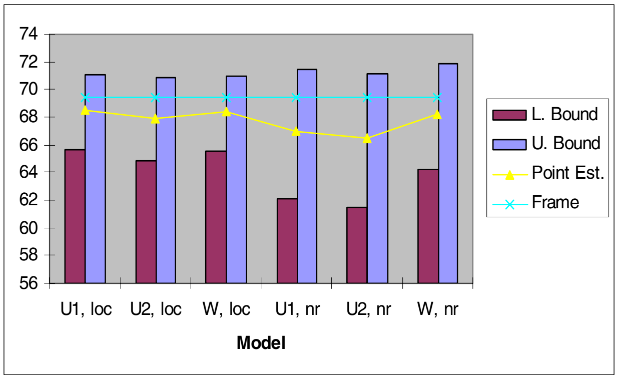


Figure 7. Percent of Physicians Who Are Board Certified: Frame Value, Point Estimates, and Confidence Intervals

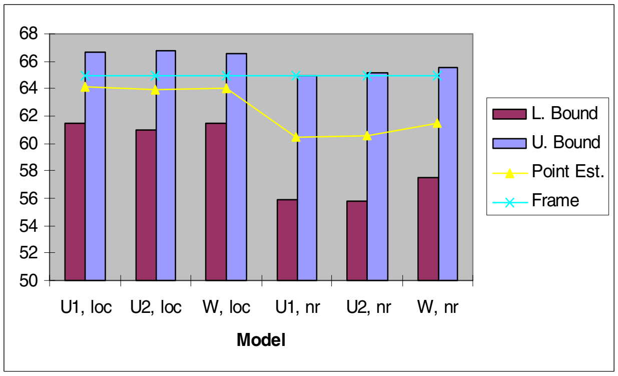


Figure 8. Percent of Physicians Who Attended Medical School in the United States: Interval Widths

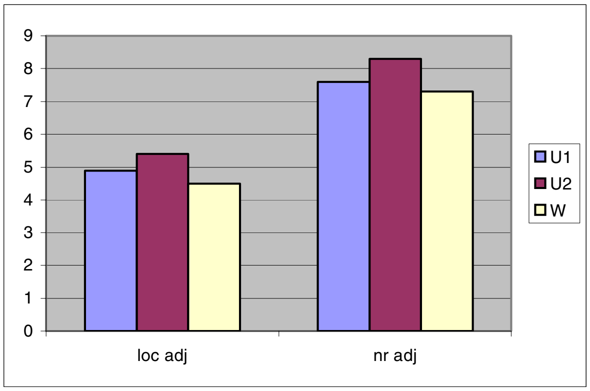


Figure 9. Percent of Physicians (MD's and Doctors of Osteopathy) Who Are MD's: Interval Widths

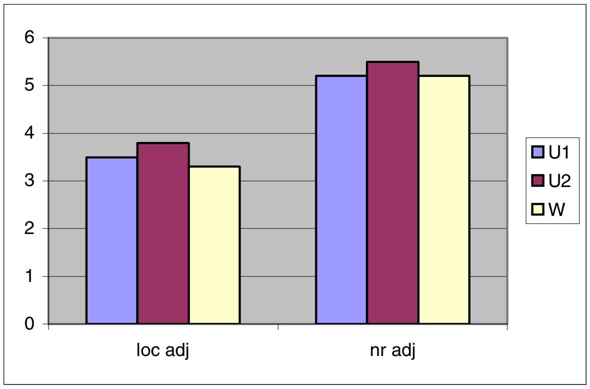
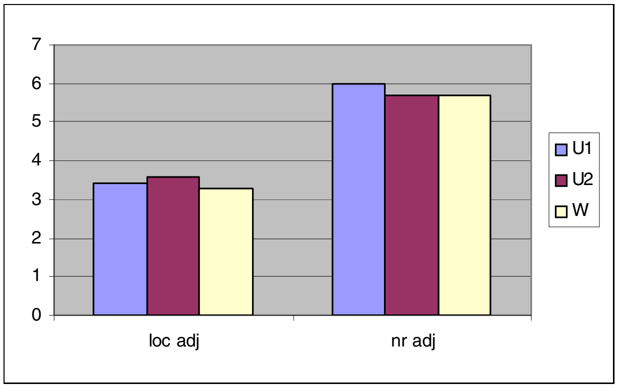


Figure 10. Percent of Physicians Under Age 45: Interval Widths

Figure 11. Percent Primary Care Physicians:



Interval Widths

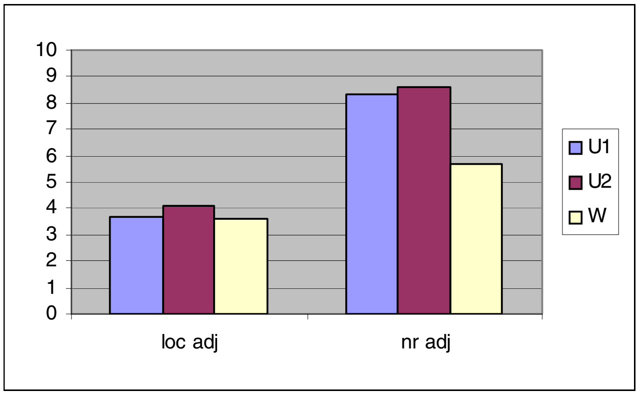


Figure 12. Percent of Physicians Who Are Gynecologists: Interval Widths

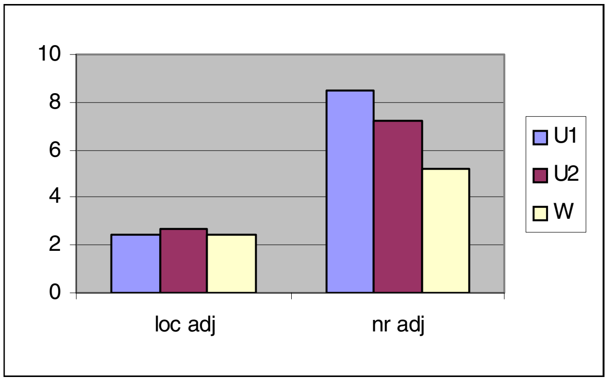


Figure 13. Percent of Physicians Whose Practices Engage in Direct Patient Care: Interval Widths

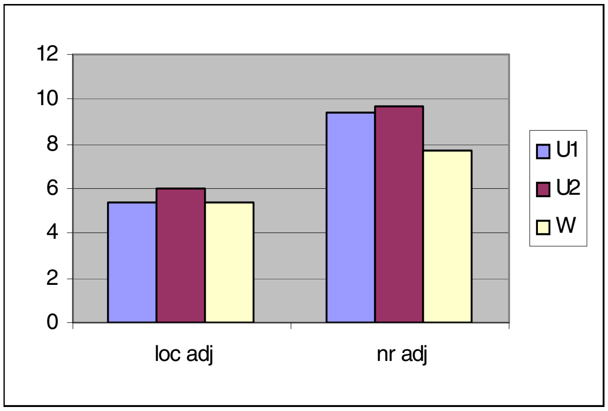


Figure 14. Percent of Physicians Who Are Board Certified: Interval Widths

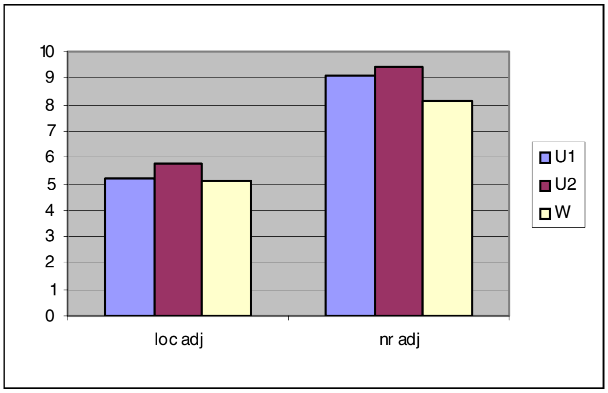


Table 2. Range of Adjustment Factors, Unweighted Model, Sampling Weights as Covariates

	Adjustment Factor		
Quantile	Location	Cooperation	Total
Maximum	19.84	33.96	61.59
95th %ile	3.90	13.68	18.32
Median	1.03	1.19	1.31
5th %ile	1.00	1.02	1.07
Minimum	1.00	1.01	1.03

Table 3. Range of Adjustment Factors, Unweighted Model, No Sampling Weights in Model

	Adjustment Factor		
Quantile	Location	Cooperation	Total
Maximum	18.27	31.02	57.02
95th %ile	3.74	12.75	16.81
Median	1.03	1.19	1.31
5th %ile	1.00	1.03	1.07
Minimum	1.00	1.01	1.03

Table 4. Range of Adjustment Factors, Weighted Model

	Adjustment Factor		
Quantile	Location	Cooperation	Total
Maximum	15.07	41.56	41.75
95th %ile	3.70	5.27	18.24
Median	1.03	1.20	1.32
5th %ile	1.00	1.02	1.07
Minimum	1.00	1.02	1.03