Evaluation of the Detectability and Inferential Impact of Nonresponse Bias in Establishment Surveys

Randall Powers, John Eltinge, and Moon Jung Cho, Bureau of Labor Statistics
Office of Survey Methods Research, PSB 1950, BLS, 2 Massachusetts Avenue NE, Washington, DC 20212
Powers.Randall@bls.gov

Abstract:

In construction of estimators from survey data, one often encounters important issues arising from nonresponse. For establishment surveys, methods to address these issues generally must account for important features of the sample design and weighting structure. For any given nonresponse adjustment procedure, an analyst makes implicit or explicit use of models for the nonresponse phenomenon and the outcome variables of primary interest. The performance of the adjustment procedure then depends on the extent to which the data deviate from the assumed models, the impact of these deviations on estimator bias, and the inferential power of diagnostics designed to detect these deviations. This paper presents a simulation study to evaluate trade-offs among the issues of model deviations, estimator performance and detectability for establishment surveys.

Keywords: Attrition; Incomplete data; Item nonresponse; Missing data; Sensitivity analysis; Simulation study.

1. Nonresponse Bias in Establishment Surveys

Survey nonresponse is an issue of concern in both household and establishment surveys. The cause of this concern is that nonresponse in surveys can lead to bias in point estimators as well as inflation in the variance of these estimators. In particular, under moderate conditions, when estimating a mean, bias is proportional to the population covariance of survey variable Y with the response probability p.

Nonresponse may have an impact on inference, such as the performance of confidence intervals. When bias is more than trivial (relative to standard error), it can lead to a reduction in true rates of coverage for confidence intervals. Another point of impact is inflation of confidence interval width due to a reduction in effective sample size. As the number of respondents decreases, the variance generally will increase.

There are special issues of concern for establishment surveys. When dealing with establishment surveys, we often encounter highly skewed distributions of unit sizes. Thus, we have a relatively small number of companies dominating the total dollar value of imports of a product. Stratification generally accounts for some, but not all, size effects. A customary approach to address this is that strata are often used as implicit weighting adjustment or imputation cells. Thus, within a certain size group, a sample of respondents will in effect be chosen to speak for nonrespondents.

2. Example: Import Price Index Series of Bureau of Labor Statistics International Price Program

In this study, nonresponse in the Import Price Index Series of the Bureau of Labor Statistics (BLS) International Price Program (IPP) was examined. Specifically, analysis was done on a stratified sample of import establishment x product classes, based on a frame that IPP obtains from Customs records. From this frame, population level estimators of price indexes are calculated. These estimators are a nonlinear function of weighted sums, based on the comparison of price reports from current and previous months, aggregated over months 1 to T. Several types of nonresponse occur and are examined by the IPP. The focus of this paper is nonresponse in the most recent month; we do not consider imputation based on data from subsequent months. For general background on the IPP, see Bureau of Labor Statistics (2003, Chapter 15).

3. Point Estimation for the International Price Program

The bias and variance properties of point estimators for the International Price Program depend on several factors, including the sample design; underlying population characteristics; response rates; the association between population characteristics and response rates; sample weights; and the formula for the IPP point estimators. The current section reviews the IPP point estimation formula for a "short term relative" (STR) quantity, while Section 4 specifies population characteristics and response rates used in the study.

Define θ^t_{sgpi} to be a short term relative (STR) of item i of probability weight group p in classification group (CG) g of stratum short s at time t. Let LTR^T_{sgpi} be a long term relative (LTR) of item i of probability weight group p in CG g of stratum short s at time T, computed as,

$$LTR^T_{sgpi} = \prod_{t=0}^{T} \theta^t_{sgpi}$$

Define θ^T_{sgp} to be a STR of probability weight group p in CG g of stratum short s at time T.

$$\theta^T_{sgp} = \frac{\sum_{i\in p} w_{sgpi} LTR^T_{sgpi}}{\sum_{i\in p} w_{sgpi} LTR^{T-1}_{sgpi}} = \frac{\sum_{i\in p} LTR^T_{sgpi}}{\sum_{i\in p} LTR^{T-1}_{sgpi}}$$

where LTR^T_{sgpi} is the long term price relative of item i of probability weight group p in CG g of stratum short s at time T, w_{sgpi} is the item weight of item i of probability weight group p in CG g of stratum short s at the base period 0, and the item weights are the same in the same weight group.

Similarly, let θ^T_{sg} be a STR of CG g of stratum short s at time T, computed as

$$\theta^T_{sg} = \frac{\sum_{p\in g} w_{sgp} LTR^T_{sgp}}{\sum_{p\in g} w_{sgp} LTR^{T-1}_{sgp}}$$

where w_{sgp} is a weight group weight , and LTR^T_{sgp} is the LTR of probability weight group p in CG g of stratum short s at time T.

The following describes the index formula for the indexes above the CG level. The IPP index computation is done in an aggregation tree structure. The formula is basically the same for all levels: each parent's index is computed from its children's indexes. For example, a stratum index is computed from the stratum's children's indexes. These children can be CGs, any number of strata, or CGs and strata. Indexes are computed in the same manner until they reach the desired aggregation level.

A stratum short is a two-digit stratum, and there may be several strata between the CG and the Stratum Short. Therefore, we need to take into account strata indices of middle steps between the CG and the Stratum Short. The number of middle steps from the CG to the Stratum Short varies depending on which stratum the specific CG belongs. Similarly, there may be several strata between the Stratum Short and the Overall, and the number of middle steps from the Stratum Short to the Overall level also varies.

Define $Child[h]$ to be the set of all strata or CGs directly below Stratum h in an aggregation tree. Let θ_h^T be a STR of stratum, h, at time T, computed as

$$\theta_h^T = \frac{\sum_c w_c LTR_c^T}{\sum_c w_c LTR_c^{T-1}} \quad (1)$$

where $c \in Child[h]$, w_c is the weight of c, and LTR_c^T is the LTR of c at time T.

This general formula (1) is used until the desired aggregation level index is obtained.

4. Design of the Simulation Study

One thousand replicates files were studied. Each replicate is an artificial set of approximately 16,000 sample units defined by the intersection of establishment x product classes based on true IPP import sample units. They include information on full-sample weights, unit size, and of particular important to this study, consistency rank. Consistency rank is a value that is assigned, and is based on the consistency with which a company imports a product in a specified product class.

In the IPP, one important quantity is called the "short term relative" or STR. The STR reflects the price change from one month to the next in a specified group of items. For a given month, each unit from a replicate is assigned an STR measure of price change. This simulated STR is generated by a Gaussian model.

Given the two aforementioned sources of information, we defined response probabilities and distortion factors for STR values for use in our simulation study. For this study, six cases of varying response probabilities and distortion factors were constructed.

All units were classified by consistency rank. The range of consistency rank values is one to seven, with seven being the most frequent importer. Upon closer examination, it was determined that units with a consistency rank of seven accounted for 90.86% of the sample, with the remaining 9.14% spread out over consistency ranks one through six, Thus, for the purpose of this study, consistency ranks one through six were aggregated into one group referred to as "less frequent importers," whereas units with consistency rank equal to seven were referred to as "frequent importers."

For this simulation, the consistency rank was used to determine two important terms. These terms were the response probability (p) and the distortion factor (d). If a unit was a frequent importer, it was assigned on response probability value, whereas if it was a less frequent importer, it was assigned a differing response probability.

The other term determined by consistency rank was the distortion factor. The distortion factor serves as a multiplier of the STR, thus distorting the initial value. If the d=1, there is no distortion factor present. If d>1, the distortion factor pushes up the STR value. If d<1, the distortion factor pushes down the STR value.

Six cases were derived, based on variation in the response probability and distortion factors. Table 1 at the end of this report summarizes these six cases. Case 1 was the full response case with no distortion. Thus p=1 and d=1 for all units. Case 2 through Case 6 each were assigned an average response rate p=.7 and an average distortion factor=1.0.

Each unit in Case 2 was assigned a response probability of p=.7 and a distortion factor of d=1.0. Thus for Case Two, there was no association between response probability and distortion factor. Each unit was assigned the same response probability and distortion factor, regardless of its consistency rank.

For Cases 3 through 6 less frequent importers were assigned a (below average) response probability p=.5, while frequent importers were assigned a response probability p=.72 (slightly above the p=.7 average). For Case 3, less frequent importers were assigned a distortion factor of slightly less than d=1 average (d=.995), whereas frequent importers were assigned a distortion factor slightly above the d=.1 average (1.00050). Thus, for Case 3, a mild positive association between p and d was induced. For Case 4, a smaller distortion factor of d=.99 was assigned to less frequent importers, thereby inducing stronger positive association between p and d than in Case 3.

Case 5 and Case 6 were essentially the reverse of Case 3 and Case 4, respectively. A

mild negative association between p and d was induced for Case 5, and a stronger negative association was induced for Case 6.

Several statistics were computed and used as sample performance evaluation criteria. For each establishment, a product is given an STR. This STR was aggregated to compute an overall full-population index estimate for each replicate.

First, we computed index estimates $\hat{\theta}_{cr}$ For each case c, and each replicate r = 1,…,1000, based on formula (1) from Section 3.

The mean STR of the 1000 replicates was computed, as was the variance.

$$\bar{\hat{\theta}}_c = (1000)^{-1} \sum_{r=1}^{1000} \hat{\theta}_{cr}$$

$$V_c^* = (1000-1)^{-1} \sum_{r=1}^{1000} (\hat{\theta}_{cr} - \bar{\hat{\theta}}_c)^2$$

Two summary statistics were used to evaluate other performance factors. Case 1 (full response) is assumed to produce unbiased point estimators. For the other cases (c), we approximated the bias by

$$b_c^* = \bar{\hat{\theta}}_c - \bar{\hat{\theta}}_1$$

Also, the relative contribution of squared bias to mean squared error was computed for each case:

$$RCB_c = (b_c^*)^2 / \{(b_c^*)^2 + V_c^*\}$$

Additionally, overall efficiency loss of case c relative to the full response case, where efficiency is evaluated through MSE is computed.

$$RE_c = \{(b_c^*)^2 + V_c^*\} / V_1^*$$

Finally, the idealized confidence interval coverage rates across the 1000 replicates were examined. We computed an idealized 95% confidence interval. The population level price index was computed from the point estimate from replicate r in case c, as well as the simulation based variance that was computed for the 1000 replicates as

$$\hat{\theta}_{cr} \pm 1.96 (V_c^*)^{1/2}$$

for each r=1,…,1000; and then computed the proportion of these intervals that contained

$$\bar{\hat{\theta}}_c \ .$$

5. Results for Month 1

Note: Due to an adjustment in the algorithm used the compute the final STR index, results reported here differ than those reported at the August 2006 JSM Meeting.

An examination of Table 2 at the end of this paper shows the following results for Month 1.

The table is ordered by the increase in value of the absolute value of the bias. By definition, the bias for Case 1 is 0. The bias for Case 2 is quite small. An increase in bias is seen in the positively associated cases (3 and 4), with more bias in the strong positive case. Bias is even larger in the negatively associated cases (5 and 6), with more bias in the strong negative case. It is interesting to note that bias appears more related to correlation than the strength of the correlation.

The relative efficiency loss (Rel Eff) increases as the bias increases. The value here is low through Case 4, but gets large in the final two cases. For Case 2, there is a pure variance inflation with little bias. This case has pure random nonresponse with p=.7. Thus, it is interesting to note the relative efficiency for Case 2 is 1.60, higher, but not significantly so, than the (1/.7=1.4) which would be expected. For Cases 3 through 6, the relative efficiency can't be predicted. This is due to the facts that a complicated nonlinear function is being dealt with, and a deviation from a pure random sample is being made.

The relative contribution of squared bias to mean squared error (RCB) calculations shows that bias is going up much faster than variance, and that bias begins to increasingly dominate as we go down the list of cases in the chart.

The coverage rate for the idealized 95% confidence interval (CI Coverage) is based on variance. As the bias increases, we see that the coverage rate decreases. The coverage rate

remains close to the nominal 95% level through the first four cases, begins dropping with Case 5, and declines markedly for the cases with the strongly negative association between the response probability and distortion factors (Case 6).

6. Results for Months 13 and 25

Month (k+12) draws from the same population as month k. Thus, it is of interest to additionally study the results for Month 13 and Month 25. The results for those months are shown in Tables 3 and 4 respectively, at the end of this paper.

The same general observations that were made for Month 1 can be made for Months 13 and 25 as well. The relative efficiency remains between 1 and 1.42 for Cases 1-4 in Month 13, and between 1 and 2.12 for Cases 1-4 in Month 25. Again, the relative efficiency does not increase greatly until Case 5. The relative contribution of bias to mean squared error is low until Case 6 in Month 13, and until Case 5 in Month 25. In case 6 for both Months 13 and 25, serious degradation in confidence interval performance is reached in Case 6. It should also be noted that the results displayed some heterogeneity across months, which we are not reporting here, was observed. Details of this heterogeneity will be considered further in later papers.

7. Conclusions

From this study, two conclusions can be drawn for a simulation-based evaluation of a price-index estimator under nonresponse. An association can be seen between response probabilities and sample unit-level STR's. Additionally, Cases 1 through 6 displayed increasing nonresponse bias, increasing loss of efficiency, increasing contribution of bias, and decreasing confidence interval coverage rates.

8. Future Work

This study lends itself to further study. One possibility is to link simulation conditions with empirical results from an IPP import survey that IPP has produced. Also, exploration of performance under alternative weighting adjustments is possible, especially in severe cases. Additionally, Te-Ching Chen, et. al. have considered alternative constructed populations of STR that come from the smoothed empirical distribution function for specific months and industry x product classes of true IPP data. Preliminary analysis indicates results qualitatively similar to those in the table above, but the data based on the empirical distribution function will warrant additional study.

Disclaimer and Acknowledgements:

The views expressed in this paper are those of the authors and do not necessarily reflect the policies of the Bureau of Labor Statistics. The authors thank Xun Wang and Te-Ching Chen for access to the IPP database and Bob Eddy and Steve Paben for many helpful comments on the International Price Program.

References:

Bobbitt, P., M.J. Cho, and R. Eddy, (2005). "Weighting Scheme Comparison in the International Price Program," *Proceedings on the Section on Survey Methods Research, American Statistical Association*, pp. 1006-1114.

Bobbitt, P., A. Cohen, R. Eddy, D. Slusher, "Lower Level Weights Proposal," BLS Internal Report, April 13, 2004.

Bureau of Labor Statistics (2003), *BLS Handbook of Methods*. Available at http://www.bls.gov/opub/hom/homch15_a.htm.

Groves, R.M. and M.P. Couper. (1998) *Nonresponse in Household Interview Surveys*,.New York:Wiley, pp 1-15.

Groves, R.M., D.A. Dillman, J.L. Eltinge, R.J.A. Little. (2002) *Survey Nonresponse*, New York:Wiley, pp 311-312, 352-354, 398-401, 431-443

Li, E., D. Boos, and M. Gumpertz. (2001). "Simulation Study in Statistics-Draft," NCSU Department of Statistics.

Wang, X., and J. Himelein, "A Top-Down Approach of Modeling IPP Short Term Relatives," BLS Internal Report, June 16, 2005.

Appendix: Tables

Table 1: Values of Response Probabilities p and Distortion Factors d for Consistency Ranks 1-6 and 7, Respectively, in Simulation Cases 1-6

Case	CR	p	d	Association of p, d
3	1-6	0.5	0.995	Positive
3	7	0.72	1.00050	
4	1-6	0.5	0.99	Stronger positive
4	7	0.72	1.001006	
5	1-6	0.5	1.050	Negative
5	7	0.72	0.995	
6	1-6	0.5	1.099	Stronger negative
6	7	0.72	0.99	

Table 2: Simulation Results on Bias, Relative Efficiency, Relative Contribution of Squared Bias to Mean Squared Error and Confidence Interval Coverage Rates for Month 1

Case	Bias	Rel Eff	RCB	CI Coverage
1	0.0	1.0	0.0	0.96
2	-0.00015	1.60	0.0039	0.94
3	0.00015	1.52	0.014	0.96
4	0.00041	1.68	0.10	0.93
5	-0.00273	8.97	0.81	0.44
6	-0.00531	29.60	0.93	0.05

Table 3: Simulation Results on Bias, Relative Efficiency, Relative Contribution of Squared Bias to Mean Squared Error and Confidence Interval Coverage Rates for Month 13

Case	Bias	Rel Eff	RCB	CI Coverage
1	0.0	1.0	0.0	0.95
2	-0.000159	1.42	0.00094	0.95
3	0.0000495	1.25	0.014	0.95
4	0.000298	1.31	0.03	0.95
5	0.0000462	3.02	0.00034	0.95
6	0.00733	35.82	0.71	0.71

Table 4: Simulation Results on Bias, Relative Efficiency, Relative Contribution of Squared Bias to Mean Squared Error and Confidence Interval Coverage Rates for Month 25

Case	Bias	Rel Eff	RCB	CI Coverage
1	0.0	1.0	0.0	0.95
2	0.0000325	2.12	0.000067	0.97
3	0.000124	1.50	0.0014	0.95
4	0.000395	1.48	0.01	0.96
5	0.00591	12.61	0.37	0.92
6	0.03984	248.86	0.86	0.32