USE OF RATIO ESTIMATES TO COMPENSATE FOR NONRESPONSE BIAS IN CERTAIN ECONOMIC SURVEYS

Alan R. Tupek and W. Joel Richardson, Bureau of the Census

Introduction

In several economic surveys recently conducted by the Bureau of the Census response rates have been found to vary by the size of the establishment. Most of these surveys have extensive followup operations, but there is usually no subsampling of nonrespondents. Assumptions are made about the nonrespondents in order to account for the entire sampled population.

The sample design for many of the mail surveys are stratified designs with simple random sampling or probability proportional to size sampling within strata. Assumptions that are often made about nonrespondents in certain surveys can result in seriously biased estimates. Surveys which use ratio estimates in order to reduce sampling errors can also compensate for bias due to nonresponse.

The effect that ratio estimates have on nonresponse bias in several surveys and the assumptions necessary for this procedure to be useful are presented.

The Problem of Nonresponse

In many recent economic surveys conducted by the U.S. Bureau of the Census there has been a growing concern about how to handle the problem of nonresponse. Surveys of U.S. businesses have had steadily declining response rates in the mandatory surveys as well as voluntary ones. The Bureau of the Census is making attempts to increase response rates by improving relations with these businesses and by using better survey instruments and followup procedures. However, it appears that little can be done to increase response rates to the point where the statistical effects of nonresponse can be ignored.

Overall response rates for voluntary business surveys have been as low as 70% even with extensive followup operations. Many companies, especially large ones, have a policy not to respond to voluntary surveys. Followup operations are usually concentrated on large firms, since they have the greatest impact on economic data. These and other factors lead to varying response rates by employment size. Nonresponse rates by strata are presented in the appendix for three recent surveys conducted by the U.S. Bureau of the Census.

The percentage delinquent is presented in two ways for the Survey of Scientific and Technical Personnel. The percentage delinquent in the columns labeled "Estabs" is based on the number of establishments mailed in the stratum. The percentage delinquent in the columns labeled "Tot. Emp." is based on the total employment of establishments mailed in the stratum. In this survey the percentage of delinquent establishments increases as the total employment of the firm increases. It can be observed that in the "Large" and "Very Large" establishment columns that the "Tot. Emp." percentage is almost always larger than that of the "Estabs", which also demonstrates the directly proportional relationship of total employment and percentage delinquent in this survey.

The percentage delinquent is presented in two ways for the Survey of Domestic and International Transportation of U.S. Foreign Trade. The percentages are first given in terms of the number of shipments, then in terms of the value or weight of shipments. In this survey there is an inversely proportional relationship between the size of shipment and the percentage delinquent. In 9 of the 10 strata the percentage based on weight or value is less than or equal to the percentage based on the number of shipments. Thus, one could infer that the larger shipments had a better response rate.

The percentage delinquent for the New Jobs Tax Credit Survey is presented before and after a telephone followup of a subsample of 330 of the nonrespondents. In this survey no apparent relationship exists between the size of the firm and the percentage delinquent. Approximately 55% of the returns were received before a subsample of nonrespondents was taken. Even including the subsample of nonrespondents there was still a high percent of the sampled population not covered (29%). The timing of the survey was critical and hence the nonresponse problem was handled differently than if more time were available.

Followup Procedures

For most economic surveys that the Census Bureau conducts, an extensive followup is made of all nonrespondents. These operations include followup letters, remailing forms to all delinquent companies, telephone contacts, and even personal visits to large companies. However, this method of followup which usually focuses on larger companies creates differences in response rates between size groups.

Despite the followup procedures being geared heavily toward the large companies in the 1975 Survey of Scientific and Technical Personnel, the nonresponse rate was directly proportional to the size of the establishment.

Two probable explanations are: (1) many large companies will not participate in any voluntary survey and (2) the questionnaire was easy for a small company with zero or few technical personnel, but much harder to complete by large firms with a large number of technical personnel. This trend, although not typical in all recent economic surveys, is quite disturbing since large companies are more likely to use Census statistics than small businesses.

Subsampling of nonrespondents after extensive followup operations is usually infeasible. Large companies have already been telephoned or even visited and further contact may jeopardize their cooperation in future surveys. Small companies have also been followed up by mail and/or telephone. A subsampling of nonrespondents among small companies would probably be successful in terms of response, but the effect on the estimates would be minimal.

Assumptions About Nonrespondents' Characteristics

In a survey with many strata and SRS within strata it often is assumed that the nonrespondents within the strata have characteristics similar to the respondents. That is, treat the respondents within each stratum as an SRS of the stratum's population. To estimate a stratum total Y one could use $\frac{N}{n_r} \bar{y}_{n_r}$ where n_r is the number of respondents and $\bar{y}_{n_r}$ is the average of all respondents. With response rates of 90% or more the nonresponse bias will not be large in comparison to other nonsampling and sampling errors, unless the nonrespondents are considerably different from the respondents. With response rates for voluntary surveys as low as 70% the bias created by imputing for nonrespondents can be sizable. For example, in the New Jobs Tax Credit Survey the estimated percentage of all firms knowing about the tax credit changed from 32.1% to 34.4% after subsampling nonrespondents. This change represents approximately twice the standard error of the estimate.

When no followup of a subsample [3] of respondents can be made it is important to take advantage of any information about the nonrespondents that is available. Many techniques which take advantage of auxiliary data are possible, but also have limitations. The one examined here, ratio estimation, takes advantage of known characteristics of nonrespondents, but does not incur considerable cost by treating each nonrespondent as a special case.

Imputation of Nonrespondents in the Survey of Scientific and Technical Personnel

The sample design of the 1975 Survey of Scientific and Technical Personnel is stratified by employment and industry. The objective of the survey was to estimate the number of scientists and engineers in each industry group. Also, estimates were made of the types of scientists and engineers and their relation to research and development and energy-related activities. A ratio-type estimator was used to make all estimates. It was felt that within a stratum (industry x employment size class) the number of scientists and engineers would be highly correlated with the employment size. This survey had been previously conducted by the Bureau of Labor Statistics and from data available from their survey it was determined that a significant reduction in variance could be achieved by using the ratio estimate. An overall reduction in standard error was obtained, mostly because in those strata with a large percent of scientists and engineers the average correlation coefficient was 2/3 (correlation of total employment with total scientists and engineers). There were many strata where most establishments had no scientists and engineers. However, this resulted in little effect on the overall estimates, since in the larger size groups of every industry a correlation averaging .42 was obtained. The ratio estimate also provided a simplified way of handling the nonresponse problem. As noted previously, response varied by total employment within strata. Using an estimator which ratios responding establishments to the entire stratum population automatically adjusts for nonresponse.

That is, $\tilde{Y} = X \frac{\sum_i^{n_r} y_i}{\sum_i^{n_r} x_i}$, where x_i is the total employment (auxiliary variable) of the i^{th} establishment of the stratum, and n_r is the number of respondents in the stratum. In effect the following value is being imputed for all the nonrespondents:

$$\sum_i^{n-n_r} y_i = \sum_i^{n_r} y_i \frac{\sum_i^{n-n_r} x_i}{\sum_i^{n_r} x_i}$$

, where $n-n_r$ is the number of nonrespondents. This follows by observing that the estimator $\tilde{Y}$ is actually a substitute for another estimator, $\tilde{\tilde{Y}}$, which we could have used if all establishments responded. That is, $\tilde{Y}$ is a substitute for

$\tilde{\tilde{Y}} = X \frac{\sum_i^{n} y_i}{\sum_i^{n} x_i}$. Setting $\tilde{Y} = \tilde{\tilde{Y}}$ implies that

$$\frac{\sum_i^{n_r} y_i}{\sum_i^{n_r} x_i} = \frac{\sum_i^{n} y_i}{\sum_i^{n} x_i} = \frac{\sum_i^{n_r} y_i + \sum_i^{n-n_r} y_i}{\sum_i^{n_r} x_i + \sum_i^{n-n_r} x_i} ,$$

$$\text{OR} \quad \sum_i^{n-n_r} y_i = \frac{\sum_i^{n_r} y_i}{\sum_i^{n_r} x_i} \left[\sum_i^{n} x_i \right] - \sum_i^{n_r} y_i = \sum_i^{n_r} y_i \left[\frac{\sum_i^{n-n_r} x_i}{\sum_i^{n_r} x_i} \right]$$

That is, the nonrespondents are imputed by adjusting the respondents by the ratio of total employment among the nonrespondents to the total employment among the respondents. In strata where the response rate varies by employment, and when employment is highly correlated to the desired characteristic, this is an intuitively sound correction for nonresponse. Some mathematical results will be presented later in this paper to support this.

Another way to look at this situation is to note that for each missing y_i the value

$$x_i \frac{\sum_i^{n_r} y_i}{\sum_i^{n_r} x_i}$$

is imputed, where x_i is the total employment of the nonrespondent. That is, one could impute this value for each missing value into the records, and the same estimate would be obtained.

Conditions Which Affect the Reduction of Bias

In order to use the ratio estimate to reduce sampling error in large samples the following condition [1] must be true:

$$|\rho|>\left|\frac{S_x \bar{Y}}{2 \bar{X} S_y}\right|,$$ where X is the auxiliary variable,

and ρ is the correlation coefficient between X and Y

$$\rho = \frac{\left[\sum_i^N (X_i-\bar{X})\ (Y_i-\bar{Y})\right]}{\left[\sum_i^N (X_i-\bar{X})^2 \sum_i^N (Y_i-\bar{Y})^2\right]^{\frac{1}{2}}}$$

$$S_x = \left(\frac{\sum_i^N (X_i-\bar{X})^2}{N-1}\right)^{\frac{1}{2}} \qquad S_y = \left(\frac{\sum_i^N (Y_i-\bar{Y})^2}{N-1}\right)^{\frac{1}{2}}$$

Basically, the ratio estimate has a smaller variance if the correlation is high, and only in this situation will any claims be made about bias reduction.

To compare the nonresponse bias of the ratio estimate to the nonresponse bias of the usual estimate some additional assumptions must be made. First, assume that the relation between X and Y is linear, but does not necessarily pass through the origin. That is, the relationship between X and Y can be written as:

$$Y_i = b_o + b_1 X_i + e_i \qquad i = 1,\ldots N$$

where b_o is the Y intercept, b_1 is the slope of the line and e_i is an error term. N is the total number of establishments in the population.

If we assume the best linear relationship between X and Y is the one which minimizes $\sum_i^N e_i^2$ then the following conditions (normal equations [2]) must be met:

$$b_o = \bar{Y} - b_1 \bar{X}$$

$$b_1 = \frac{\sum_i^N (X_i-\bar{X})\ (Y_i-\bar{Y})}{\sum_i^N (X_i-\bar{X})^2}$$

Note that $b_1 = \frac{S_y}{S_x}\rho$, hence if $\bar{X}$ and $\bar{Y}$ are positive as is true of most practical situations then

$$|\rho| > \left|\frac{S_x \bar{Y}}{2 \bar{X} S_y}\right| \Rightarrow b_1 > \frac{\bar{Y}}{2 \bar{X}} > 0 \text{ if } \rho > 0$$

That is, the overall regression in the population has a positive slope.

Let us further assume that the population represented by the respondents also follows the above model, but with different parameters. That is,

$$Y_i = b_o' + b_1' X_i + e_i' \qquad i = 1,\ldots N_R$$

N_R is the total number of establishments in the population represented by the respondents, b_1' is the slope of the line, and e_i' is an error term.

Given such a model it is reasonable to assume that if the following 3 conditions hold:

(1) $b_1 \bar{X}_R \geq b_o$

(2) $\bar{X}^2 \geq \bar{X}_R^{\,2}$

(3) $b_o + b_1 \bar{X}_R \geq b_o' + b_1' \bar{X}_R$

then the bias of the ratio estimate will be less than the bias of the usual estimate.

NOTE: If the inequality is simultaneously reversed in both conditions (2) and (3) then we still have the same results.

To see that this is true we will first observe the definition of the bias of the ratio and usual estimates.

The usual estimator is $\bar{y}_{n_r} = \sum_{i=1}^{n_r} \frac{y_i}{n_r}$

with $E(\bar{y}_{n_r}) = \bar{Y}_R$, BIAS (USUAL) $= \bar{Y}_R - \bar{Y}$.

The ratio estimator is $\overset{\approx}{Y} = \bar{X}\ (\bar{y}_{n_r})/(\bar{x}_{n_r})$

with $E(\tilde{\bar{Y}}) = \bar{X}\left(\frac{\bar{Y}_R}{\bar{X}_R} - \frac{\operatorname{cov}(\hat{R},\bar{x}_{n_r})}{\bar{X}_R}\right) \doteq \bar{X}\frac{\bar{Y}_R}{\bar{X}_R}$

hence its nonresponse bias is

$$\text{BIAS (RATIO)} = \frac{\bar{X}\,\bar{Y}_R}{\bar{X}_R} - \bar{Y}$$

Now observe that conditions (2) and (3) imply

$$\left(b_o' + b_1'\,\bar{X}_R\right)\left(\bar{X}^2 - \bar{X}_R{}^2\right) \le \left(b_o + b_1\bar{X}_R\right)\left(\bar{X}^2 - \bar{X}_R{}^2\right) \qquad \{A\}$$

Condition (1) implies $b_o\left(\bar{X}^2 - 2\bar{X}_R\bar{X} + 2\bar{X}_R{}^2 - \bar{X}_R{}^2\right)$ $- b_1\bar{X}_R\left(-\bar{X}^2 + 2\bar{X}^2 - 2\bar{X}\,\bar{X}_R + \bar{X}_R{}^2\right) \le 0$ which implies

$$b_o\left(\bar{X}^2 - \bar{X}_R{}^2\right) - 2b_o\left(\bar{X}_R\,\bar{X} - \bar{X}_R{}^2\right) + b_1\bar{X}_R\left(\bar{X}^2 - \bar{X}_R{}^2\right) - 2b_1\bar{X}\left(\bar{X}_R\,\bar{X} - \bar{X}_R{}^2\right) \le 0 \qquad \{B\}$$

$$\{A\} \text{ and } \{B\} \Rightarrow \bar{Y}_R\left(\bar{X}^2 - \bar{X}_R{}^2\right) - 2\bar{Y}\left(\bar{X}_R\bar{X} - \bar{X}_R{}^2\right) \le 0$$

As stated previously we will assume all X's and Y's are ≥ 0. Hence, multiplying by

$\frac{\bar{Y}_R}{\bar{X}_R{}^2}$ and adding and subtracting $\bar{Y}^2$ imply

$$\left(\frac{\bar{X}\,\bar{Y}_R}{\bar{X}_R} - \bar{Y}\right)^2 - \left(\bar{Y}_R - \bar{Y}\right)^2 \le 0$$

$$\Rightarrow \text{BIAS}^2\text{ (RATIO)} \le \text{BIAS}^2\text{ (USUAL)}$$

What is the meaning of the 3 conditions and do they have any practical use? Condition (1) is not very limiting. Few distributions with high correlation would not have this relationship. In fact it can be shown that given the normal equations and the conditions on ρ, $b_1\,\bar{X} > b_o$. Note however that the condition is based on $\bar{X}_R$ rather than $\bar{X}$.

Conditions (2) and (3) should be examined together. They state that when the respondents are small in terms of the auxiliary variable then the bias of the ratio estimate is less when the regression estimate of Y at $\bar{X}_R$ is smaller for the population represented by respondents than for the entire population.

Some special situations involving these conditions are worth looking at. When $b_1 = b_1'$, that is, the slopes of the two lines are identical, then condition (3) becomes $b_o \geq b_o'$. Thus when the respondents are small in terms of X and the regression line for the respondents is parallel to and "below" the regression line for the entire population the bias of the ratio is less. Similarly if we assume $b_o = b_o'$, that is, the y intercept of the two lines is the same, then condition (3) becomes $b_1 \geq b_1'$. Thus, where the respondents are small in terms of X and the regression line for the respondents has the same y intercept, but is always below the regression line for the entire population, the bias of the ratio is less.

It should also be noted that if the respondents are larger in terms of the auxiliary variable and the regression line for the respondents is above that of the entire population, then again the bias of the ratio is less. This follows from the reversal of the inequalities in conditions (2) and (3).

The PPS Estimator

A probability proportional to size estimator(PPS) within strata can also be used to impute for nonresponse. This was what we used to adjust for nonresponse in the 1976 Survey of Domestic and International Transportation of U.S. Foreign Trade.

The estimator employed was $\bar{\bar{Y}} = A \sum_i^{n_r} \frac{y_i}{P_i}$ where the adjustment $A = \frac{\sum_i^{n} \frac{1}{P_i}}{\sum_i^{n_r} \frac{1}{P_i}}$ takes into consideration response rates which vary by size (P_i) within each stratum. Hence, if a disproportionately small number of exporters of large shipments respond then the nonresponse adjustment will be greater than if the response rate were distributed evenly across all size groups.

Areas for Future Study

The ratio estimator was determined to be useful in adjusting for nonresponse bias in the surveys discussed in this paper. How well the ratio estimator will work for other surveys can be seen by determining if the conditions given in this paper are satisfied. Its application to other surveys may necessitate the development of new conditions from different assumptions.

The usefulness of the PPS estimator in reducing bias could be studied beyond the intuitive level presented here. Development of conditions under specified assumptions for the PPS estimator and other estimators such as the regression estimator are areas for future research.

References

[1] Cochran, W. G. Sampling Techniques, John Wiley and Sons, Inc., 1963.

[2] Draper, N. R. and Smith H. Applied Regression Analysis, John Wiley and Sons, Inc., 1967.

[3] Hansen, M. H. and Hurwitz, W. H. "The Problem of Non-Response in Sample Surveys," Journal of the American Statistical Association, December 1946, Vol. 41, pp. 517-529.

[4] Hendricks, W. A. "Adjustments for Bias Caused by Non-Response in Mailed Surveys," Agricultural Economics Research, April 1949, Vol. 1, No. 2, pp. 52-56.

TABULAR APPENDIX

Survey of Scientific and Technical Personnel (1975)

Percentage Delinquent

Industry	Size of Establishment 1/							
	Small		Medium		Large		Very Large	
	Estabs	Total Emp.	Estabs	Total Emp.	Estabs	Total Emp.	Estabs	Total Emp.
Food and Kindred Products	19	20	19	19	30	31	36	38
Textiles and Apparel	23	22	21	23	29	30	42	46
Lumber Products & Furniture	29	24	17	17	22	24	33	38
Paper Products	19	23	25	24	34	35	38	47
Chemicals	19	17	31	31	--	--	46	54
Petroleum Refining	21	18	22	23	43	40	40	41
Rubber, Plastic Products & Leather Products	28	35	28	29	33	34	57	61
Stone, Clay and Glass	15	14	18	19	34	35	41	42
Primary Metals	17	13	34	38	51	53	64	61
Fabricated Metal Products	21	18	16	17	32	34	57	60
Machinery	15	12	20	21	30	31	44	47
Electrical Machinery	9	7	29	30	34	34	44	51
Motor Vehicles	24	22	--	--	23	29	49	72
Aircraft and Parts	--	--	--	--	--	--	25	59
Other Transportation Equip.	27	31	24	23	44	48	54	59
Instruments	24	20	20	23	35	41	42	49
Other Nondurable Goods	21	18	18	18	24	24	30	32
Petroleum Extractions	17	19	28	30	--	--	47	50
Other Minerals	26	22	--	--	31	33	36	39
Construction	21	18	18	19	--	--	37	46
Transportation	21	31	--	--	--	--	25	36
Telephone & Telegraph	78	47	--	--	--	--	78	71
Radio and Television Broadcasting	21	21	--	--	--	--	20	29
Electric, Gas and Sanitary Services	21	32	--	--	--	--	29	33
Research & Development Labs	24	28	20	22	--	--	22	12
Commercial Testing Labs	27	29	30	32	--	--	41	55
Miscellaneous Business Serv.	34	35	22	24	24	24	31	19
Engineering & Architectural Services	17	17	18	18	--	--	22	32
Medical & Dental Labs	16	17	15	14	23	24	--	--
Other Nonmanufactures	32	29	26	22	30	32	30	42

1/ Size is based on total employment, and varies by industry.

Survey of Domestic and International Transportation of U.S. Foreign Trade (1976)

Percentage Delinquent

Stratum	Based on Number of Shipments Mailed	Based on Total Weight or Value of Shipments Mailed
Exports - All	23	--
Vessel - General Cargo	24	22
Vessel - Bulk	20	20
Air	26	21
Transshipments - General Cargo	20	21
Transshipments - Bulk	22	22
Imports - All	25	--
Vessel - General Cargo	24	24
Vessel - Bulk	16	12
Air	30	27
Transshipments - General Cargo	33	23
Transshipments - Bulk	42	23

New Jobs Tax Credit Survey (1977)

Percentage Delinquent Based on the Number of Establishments Mailed

Stratum	Before Subsampling Nonrespondents	After Subsampling Nonrespondents
Single Units		
0 - 9	45	31
10 - 49	41	23
50 - 249	41	20
250 - 499	43	12
500 and over	50	10
Multiunits		
0 - 9	42	17
10 - 49	40	1
50 - 249	35	0
250 - 499	40	10
500 and over	44	28