COUNTERING ESTIMATION BIAS DUE TO RESPONSE ERRORS: A SIMULATION EXAMPLE

Naihua Duan, Kent H. Marquis, and M. Susan Marquis
The Rand Corporation

INTRODUCTION

Social policy research often uses interview survey data to estimate relationships between variables. However, response errors in the survey observations can lead to bias in the estimated relationships. Different strategies have been proposed to minimize these estimation biases, and, in this paper, we evaluate how well they overcome the biasing effects of common kinds of survey response errors on multiple regression coefficients. Our approach is to use computer simulation to illustrate how the various strategies behave. To give realism to the example, we pose a specific health policy question and simulate survey responses that contain realistic error structures.

The paper is organized as follows: We describe the policy example to be considered, the alternative correction strategies, the simulated data, and our criteria for evaluating the strategies. Then, we will show selected results from the simulations and discuss the strengths and weaknesses of the correction strategies.1/

POLICY EXAMPLE

We assume that health policy planners want to know how use of dental care services varies by income. To this end, the policy analyst will estimate the coefficients of the regression of the log of annual family dental expenditures regressed on family income, family size, education of the head, and race of the head:2/

$$\ln(D) = \beta_0 + \beta_1 I + \beta_2 F + \beta_3 E + \beta_4 R + w$$

where: $\ln(D)$ = natural log family dental expenditures
I = income,
F = family size,
E = education of head,
R = race of head,
β_i = the coefficients to be estimated, and
w = equation error (residual).

We assume that a household survey obtains observations on the variables. The income observations contain survey response error. We address two questions:

- What are the effects of the survey response errors on the estimated ordinary least squares regression coefficients?
- Can correction strategies improve the estimates and the analysts' inferences?

STRATEGIES: DESIGN FEATURES

Each correction strategy we consider combines a special data collection feature with a special analysis method.3/ The three design features we investigate are reinterviews, internal consistency, and a constructed income design.4/

The reinterview design collects a second measure of income in a follow-up interview taken on a subsample of respondents.

The internal consistency design obtains responses to two similar questions about income in the same interview.

The constructed income design collects measures on a set of predictors of income. The predictors are used to construct an "error free" income measure.

STRATEGIES: ANALYSIS METHODS

We employ three analysis methods as part of the empirical correction strategies. Each analysis is appropriate for one or more data collection design.

One method is to correct the variance-covariance martrix of the raw measures using estimates of the response error properties (Fuller and Hidiroglou, 1978). We employ this strategy with the reinterview design and the internal consistency design. The repeated income measurements provide the estimate of income response error variance used for the correction.

We use instrumental variables analysis (see Johnston, 1963) with the constructed income strategy. We regress measured income on the predictor variables collected and then form predicted or constructed income which is used as the instrument.

The empirical Bayes method can be used to correct the individual income measurements directly given an estimate of the response error variance. We use this technique with the reinterview design strategy as a comparison with the more usual method of matrix correction.

SIMULATION

We have generated synthetic data on the computer to illustrate the effects of response error on ordinary least squares regression coefficients and the ability of the strategies to overcome the effects. To produce the synthetic data, we assign specific numerical values to population parameters and then draw samples from the hypothetical population.

First, true values for family dental expenditures, family income, and the other family characteristics are generated for a sample of 1500 families.5/

The next step is to introduce response error into the income measures. The other variables contain no measurement error.

The third step is to generate the second income measure for the reinterview and internal consistency designs. The repeated income observation, in the reinterview and internal consistency design, has the same error properties as the initial measure. However, we allow response errors to be correlated across the two trials. Reinterview observations are generated fcr a 10 percent subsample.

Forty-six replicates of the synthetic sample are generated. On each sample, we carry out an uncorrected regression analysis and each correction strategy and then observe the distribution of the coefficient estimates for each method.

The assumptions about the income response errors we used to generate our data are based on a review of the measurement literature.6/ We used two response error models. For model 1, income response errors are random. For response model 2, income response errors are systematically related to education, age, race, and sex and hence indirectly correlate with income. In both models, the variance of the income measurement error is 27 percent of total measured variance. In both models, original and reinterview response errors are correlated .25; response errors for the internal consistency items are correlated .63.

EVALUATION STATISTICS

For each strategy we ask: How accurate are the parameter estimates? How reliable are the estimates? How valid and efficient are the analysts' inferences? The four indicators used to answer these questions are: the percent bias, the standard error of the estimate, the coverage probability, and the root mean square error.

The estimate of coefficient bias is the difference between the true population parameter and the mean of the coefficient estimate over the 46 trials. The percent bias is the bias relative to the true population parameter.

The precision of a coefficient is described by its standard error which we directly estimate using the sample-to-sample variability in the coefficients.

The coverage probability we define to be the probability that an analyst's constructed 95 perdent confidence interval for the parameter estimates will include the true population value. We use the coverage probability to assess the validity of inferences.

We use the root mean square error (RMSE)--which combines bias and sampling variability--to compare the efficiency of inferences using alternative estimation strategies. We show the RMSE relative to the RMSE based on data containing no measurement error.

RESULTS: INCOME COEFFICIENT

Model 1: Response Error Uncorrelated with True Value

Model 1 assumes that response errors in income are random. This error structure leads to a significant attenuation in the ordinary least squares (OLS) estimate of the income coefficient (Table 1). The expectation of the estimate is 32 percent less than the true population value. The analyst's 95 percent confidence interval will include the true value only about 10 percent of the time.

All of the strategies do reduce the coefficient bias. The internal consistency approach is least effective because of the high correlation of response errors (.63) across measures. We deliberately introduce the correlation to investigate the kind of measures we expect field workers to obtain. However, the correlation does result in an underestimate of the income error variance and hence to an incomplete adjustment. Similarly, the matrix correction analysis of reinterview data does not achieve a full correction. However, because the interview and reinterview data contain only a moderate correlation, the remaining bias in the corrected coefficient is only -8 percent.

The empirical Bayes strategy using reinterview data is not as effective as the matrix correction analysis. The constructed income technique yields an unbiased estimate of the income coefficient when the income response errors are random.

Most of the strategies reduce bias without a loss in precision. The instrumental variable analysis using predicted income, however, results in a standard error which is almost double the standard error for the other strategies. The standard error of the coefficient estimated by the instrumental variables technique will decrease as the correlation between the measure and the instrument increases. The set of predictors is only moderately correlated with true income; the predictors explain 36 percent of the "population" variance in true income. Because the predictors are imperfect proxies for true income, the standard error is larger than the other strategies.

The reinterview matrix correction analysis, and the constructed income design, yield coverage probabilities that exceed 90 percent. For these designs, the analysts' inferences are likely to be valid. However, the analyst is also concerned about the efficiency of his inferences. The constructed income analysis is less preferred between these two strategies because it produces very high standard errors. The relative RMSE is 2.3, which exceeds the total error for all correction strategies.

Model 2: Response Error Correlated with True Value

Model 2 assumes response error depends on some of the demographic characteristics, although it is not directly a function of income. The amount of dependence is that observed in real survey data. It can be shown that the OLS estimate of the income coefficient will still be attenuated when estimated from data conforming to this response model.7/ The attenuation depends on the ratio of random error variance to true income variance. Although there is systematic error variance in the second model, random error still dominates. As a result, the bias in the ordinary least squares estimate of the income coefficient is about the same for model 2 as for model 1 (see Table 2).

The effect of the strategies on the income coefficient bias under model 2 is the same as for model 1 with one exception. The constructed income strategy yields an income coefficient with a significant positive bias. The assumption for the instrumental variables analysis is that the predictors used to construct income are uncorrelated with the income error. This assumption is violated for our model 2, since the variables included in the predictors of income are determinants of response error.

EDUCATION COEFFICIENT

Response error will not only bias the regression coefficient on the variable measured with error; it can also bias coefficients of other variables correlated with the said variable. We use the education coefficient as an example.

Model 1:

In our random response error income model, the estimated OLS coefficient on education is biased because of random response errors in the income observations; the uncorrected analysis results in an estimate which is 12 percent too high (Table 3).

Table 1

EFFECT OF CORRECTION STRATEGIES ON ESTIMATED INCOME COEFFICIENT

MODEL 1: Income Response Error is Random
No Response Error in Other Variables

Correction Strategy	Percent Bias	Standard Error	Coverage Probability	Relative RMSE
Uncorrected OLS	-32*	.0049	.10	2.8
Internal Consistency	-23*	.0052	.42	2.1
Reinterview-Matrix Correction	-08*	.0075	.91	1.5
Reinterview-Bayes	-14*	.0070	.82	1.7
Constructed Income	+03	.0135	.95	2.3

*Coefficient estimate is biased p less than .05.

Table 2

EFFECT OF CORRECTION STRATEGIES ON ESTIMATED INCOME COEFFICIENT

MODEL 2: Income Response Error Contains Systematic Bias
No Response Error in Other Variables

Correction Strategy	Percent Bias	Standard Error	Coverage Probability	Relative RMSE
Uncorrected OLS	-31*	.0051	.15	2.7
Internal Consistency	-21*	.0053	.49	2.0
Reinterview-Matrix Corection	-05*	.0075	.94	1.3
Reinterview-Bayes	-10*	.0071	.89	1.5
Constructed Income	+21*	.0257	.93	4.7

*Coefficient estimate is biased p less than .05.

Table 3

EFFECT OF CORRECTION STRATEGIES ON ESTIMATED EDUCATION COEFFICIENT

MODEL 1: Income Response Error is Random
No Response Error in Education Responses

Correction Strategy	Percent Bias	Standard Error	Coverage Probability	Relative RMSE
Uncorrected OLS	12*	.0130	.85	1.3
Internal Consistency	08*	.0133	.91	1.2
Reinterview-Matrix Corection	02	.0138	.95	1.0
Reinterview-Bayes	13*	.0129	.83	1.3
Constructed Income	-04	.0184	.94	1.4

*Coefficient estimate is biased p less than .05.

The bias reflects a "transmission" of the income random error caused by the correlation of education and income.

To the extent that the strategies correct for the random response error in income, they also improve the estimated education coefficient. Our findings about the effectiveness of the alternative strategies in producing valid, efficient inferences about the size of the relationship between education and dental expenditures follows the conclusions for the income coefficient. The reinterview matrix correction analysis yields valid conclusions about the size of the education coefficient without a sacrifice in precision. The relative RMSE is about 1.0 for this design. The constructed income strategy yields unbiased estimates of the education coefficient, but the estimate is imprecise. The relative RMSE of the education coefficient is 1.4 for the constructed income design. The other strategies are not as effective in producing valid conclusions, as seen from their lower coverage probabilities.

Model 2:

Using measurement model 2, the OLS estimate of the education coefficient is biased both by the "transmission" effect and also directly because response errors in income are a function of education.

The uncorrected education coefficient is biased upward by 32 percent; the probability that a 95 percent confidence interval will include the true value is only 23 percent (Table 4).

Because most of the bias in the education coefficient is due to the effects of systematic response bias, our strategies, which are designed only to correct for random errors, result in only small improvements in the education coefficient. The empirical Bayes strategy does not significantly reduce the bias. The other random error correcting strategies do result in significant reductions in bias; however, the analysts' inferences about the size of the education coefficient are likely to remain valid.

Table 4

EFFECT OF CORRECTION STRATEGIES ON ESTIMATED EDUCATION COEFFICIENT

MODEL 2: Income Response Error Contains Systematic Bias
No Response Error in Education Responses

Correction Strategy	Percent Bias	Standard Error	Coverage Probability	Relative RMSE
Uncorrected OLS	+32*	.0117	.23	2.4
Internal Consistency	+30*	.0117	.30	2.3
Reinterview-Matrix Corection	+28*	.0125	.41	2.2
Reinterview-Bayes	+33*	.0122	.25	2.2
Constructed Income	+23*	.0155	.70	2.0

*Coefficient estimate is biased p less than .05.

CONCLUSIONS

We have examined how response errors affect ordinary least squares estimates of regression coefficients and whether alternative design and estimation strategies can yield valid and efficient inferences about the coefficients.

We found that if a variable contains random response errors, its OLS coefficient is directly biased. These same errors also bias the coefficients of any other predictors correlated with the variable.

Alternative design and estimation strategies can reduce the coefficient bias. Among the two repeated measurement strategies applied to data containing realistic random errors, the reinterview approach was more likely to yield valid conclusions than the internal consistency approach. The analysis method also mattered; the matrix correction strategy produced better estimates than the empirical Bayes method.

The instrumental variables strategy produced valid inferences when income errors were random, but the coefficients were imprecise.

Allowing even small amounts of systematic error in the income responses, we found that these strategies were not effective because they are designed to correct only for random errors. If systematic bias is expected, there are other strategies, such as record checks, that should be considered.

FOOTNOTES

1/ This paper summarizes results reported in Marquis, et al., 1981. Greater detail about the methods, assumptions, and findings can be found in that report.

2/ The researcher will use transformed expenditures rather than raw dollars so that the distribution of the residual, w, is approximately normal. A constant of $1 is added to raw expenditures to avoid trying to take the log of 0 for families with no expenditures.

3/ There are also design-only solutions which seek to find a measurement procedure that eliminates response error. We focus on empirical approaches which accept the inevitability of error and attempt to counteract it through a combination of special data collection and analysis features.

4/ Other special data collection features that might be used to correct for the effects of response errors include record checks, randomized response, and multiplicity designs. Marquis, et al. (1981) consider these additional designs. However, in the interest of time, we have restricted our attention here to the three designs for which we have the best estimates of the relevant input parameters.

5/ The generated values and their relationships are like those observed in the 1971 health survey conducted by the Center for Health Administration Studies described in Andersen, Kravitz, and Anderson. Family income and dental expenditures are inflated to 1979 dollars.

6/ Details are in Marquis, et al., 1981.

7/ Assuming the determinants of error are among the explanatory variables in the regression.

REFERENCES

Andersen, Ronald, Joanna Kravits, and Odin W. Anderson (Eds.), Equity in Health Services: Empirical Analyses in Social Policy, Ballinger, Cambridge, Mass., 1975.

Fuller, Wayne A. and Michael A. Hidiroglou, "Regression Estimates After Correcting for Attenuation," Journal of the American Statistical Association, 73, 99-104, 1978.

James, W. and C. Stein, "Estimation with Quadratic Loss," Proceedings of the Fourth Berkeley Symposium, Vol. 1, University of California Press, Berkeley, 361-379, 1961.

Johnston, J., Econometric Methods, McGraw Hill, New York, 1963.

Marquis, K. H., N. Duan, M. S. Marquis, and J. M. Polich with J. E. Meshkoff, D. S. Schwarzbach, and C. M. Stasz, Response Errors in Sensitive Topic Surveys, The Rand Corporation, R-2710/2-HHS, 1981.