THE EFFECT OF WEIGHT TRIMMING ON NONLINEAR SURVEY ESTIMATES

Frank J. Potter, Research Triangle Institute
Research Triangle Institute, P. O Box 12194, Research Triangle Park, NC 27709

KEY WORDS: sampling weights

I. INTRODUCTION

In survey sampling practice, unequal sampling weights (the inverse of the selection probabilities) can be both beneficial and deleterious. Extreme variation in the sampling weights can result in excessively large sampling variances when the data and the selection probabilities are not positively correlated.

In some survey situations, the survey statistician may impose a trimming strategy for excessively large weights. Because of the weight trimming, the survey statistician will usually expect an increased potential for a bias in the estimate and a decrease in the sampling variance. The ultimate goal of weight trimming is to reduce the sampling variance more than enough to compensate for the possible increase in bias and, thereby, reduce the mean square error (MSE).

In this research, I investigated the effect of the weight trimming on simple linear regression coefficients using a population that can be fully enumerated. The specific empirical goal is to evaluate the effect of the weight trimming on estimates of linear regression coefficients in terms of; (a) the consistency and variability of trimming levels, (b) the change in the sampling variance and in the MSE, and (c) the coverage probability of confidence intervals.

II. WEIGHT TRIMMING PROCEDURES

A. Overview

Five weight trimming procedures are discussed in this paper: The procedures are:

(1) the estimated MSE procedure using regression coefficients
(2) the estimated MSE procedure using mean estimates
(3) Taylor Series (TS) procedure using mean estimates
(4) the "NAEP procedure" (using only the sampling weights)
(5) the Weight Distribution procedure (using only the sampling weights).

The "NAEP procedure" has been reported in conjunction with the National Assessment of Educational Progress (NAEP) and, for brevity is referred to here as the NAEP procedure (Benrud et al. 1978). An alternative version of the NAEP procedure, which uses data, is described by Johnson et al. (1987).

B. Procedures

Assume a sampling frame of N units and

Y_k = the observed value for the kth unit.
π_k = the selection probability for the kth unit when a sample of size n is selected and assume π_k is less than 1 for all k;
w_k = the untrimmed sampling weight for the kth unit; that is, $w_k = 1 / \pi_k$.
w_{kt} = the sampling weight for the kth unit when a weight trimming strategy is used.

1. Estimated MSE Trimming Using Regression Coefficients

In this procedure, an estimate of the MSE for selected data items is evaluated at various trimming levels to empirically determine the trimming level (Cox & McGrath 1981; Potter 1988). The assumption underlying this procedure is that, for a set of weights and data, a point exists at which the reduction in the sampling variance resulting from the trimming is offset by the increase in the square of the bias introduced into the estimate. In this procedure, the $MSE(\hat{\beta}_t)$ is estimated by

$$\hat{MSE}(\hat{\beta}_t) = \hat{Var}(\hat{\beta}_t) + (\hat{\beta}_t - \hat{\beta})^2$$

where

$\hat{\beta}$ = the estimate of the regression coefficient using the untrimmed weights;
$\hat{\beta}_t$ = the estimate of the regression coefficient using trimmed weights; and

$\hat{Var}(\hat{\beta}_t)$ = the estimated variance of $\hat{\beta}_t$ using the standard TS variance approximation.

The procedure is implemented by repeatedly computing the estimate of the MSE for regression coefficients at differing levels of weight truncation and the 'optimal' level of truncation is the point that minimizes estimated MSE.

In the empirical study, 20 candidate trimming levels were used and 20 sets of weights were computed for each sample. The following procedure was implemented:

1. For each regression, the estimated MSE is computed for each set of weights and assigned a rank (1 to smallest value and 20 to the largest value).
2. The estimated squared relative bias is computed for each regression coefficient using the equation

$$\textit{Rel Bias}\ (\hat{\beta}_t)^2 = [(\hat{\beta}_t - \hat{\beta}) / \hat{\beta}]^2 .$$

3. An average rank is computed for the estimated MSEs and relative bias across the regression coefficients for each set of weights.
4. These two average ranks are then averaged and the lowest combined average rank is the trimming level.

2. Estimated MSE Trimming Using Mean Estimates

In this procedure, I used basically the same procedure described above except for using estimates of the mean. The MSE estimate for selected data items is evaluated at various trimming levels to determine empirically the trimming level. In this procedure, the $MSE(\hat{\bar{Y}}_t)$ was estimated essentially by

$$\hat{MSE}(\hat{\bar{Y}}_t) = \hat{Var}(\hat{\bar{Y}}_t) + (\hat{\bar{Y}}_t - \hat{\bar{Y}})^2$$

where

$(\hat{\bar{Y}}_t)$ = the estimate of the mean using the untrimmed (trimmed) weights; and

$\hat{Var}(\hat{\bar{Y}}_t)$ = the estimated variance of $\hat{\bar{Y}}_t$.

As stated previously, the estimated MSE is used to identify a trimming level among a set of candidate trimming levels that, jointly for multiple data items, has the smallest estimated MSE. The following procedure was implemented:

1. The estimated MSE is computed for each data item for each set of weights and assigned a rank.
2. An average rank is computed for the estimated MSEs across the data items for each set of weights and the lowest average rank is defined as the trimming level.

3. The TS Procedure Using Mean Estimate

The TS procedure uses the estimated MSE and the estimated relative bias computed for each data item at multiple candidate trimming levels. The estimated MSE is computed using a derived form for TS linearized variate. This procedure is described in detail in Potter (1990).

Assume a sample of size n is selected with unequal probability and with replacement. The following derivations are conditional on a fixed weight trimming value of w_o for all possible samples of size n. All weights below this value are adjusted by a factor A_s so that the original weight sum (W_s) for the sample is preserved. The usual estimator can be written as a function of weighted totals. That is,

$$\hat{\bar{Y}}_t = \Sigma_n\, w_{kt}\, Y_k \,/\, \Sigma_n\, w_{kt}$$

where Σ_n denotes the sum over the sample n.

The linearized variate for variance estimation is:

$$\hat{z}_k = (1/\hat{N}) \{ w_{kt} (Y_k - \hat{\bar{Y}}_{NT}) - w_k [\Sigma_n\, \tau_k\, w_o (Y_k - \hat{\bar{Y}}_{NT}) / \hat{N}]\}.$$

where

$\tau_k = 1$ if weight is trimmed; 0 otherwise; and

$$\hat{\bar{Y}}_{NT} = \Sigma_n\, (1 - \tau_k)\, w_k\, Y_k \,/\, \Sigma_n\, (1 - \tau_k)\, w_k$$

For the TS trimming procedure, the estimated MSE and the relative bias are computed for each data item ℓ (ℓ=1,...,m):

1. an estimated MSE measure:

$$MSE_1 = Var(\hat{\bar{Y}}_{t\ell}) + \hat{Bias}(\hat{\bar{Y}}_{t\ell})^2,$$

where

$$\hat{Bias}(\hat{\bar{Y}}_{tl}) = (\hat{\bar{Y}}_l - \hat{\bar{Y}}) .$$

2. the relative bias:

$$\text{RelBias}_l = \hat{Bias}(\hat{\bar{Y}}_{tl})/\hat{\bar{Y}}_l.$$

For the empirical study, the same procedure described for the estimated MSE of the regression coefficients were used. Because multiple data items were used, the identified trimming level may not be the smallest estimated MSE and bias for all or any of the data items.

4. The NAEP Procedure

In the NAEP procedure, the relative contribution is limited to a specific value by comparing the squared of each weight to a multiple of the sum of the square weights. That is,

$$w_k^2 \le c\ \Sigma_n\ w_k^2\ /\ n \text{ or}$$
$$w_k \le K_n . \qquad (1)$$

where $K_n = (c\ \Sigma_n\ w_k^2\ /\ n)^{1/2}$.

The value for c is arbitrary and can be chosen empirically by looking at the distribution of the square root of the values of

$$n\ w_k^2\ /\ \Sigma_n\ w_k^2.$$

In the NAEP algorithm, each weight in excess of K_n is given this value and the other weights are adjusted to reproduce the original weight sum. The sum of square adjusted weights is computed and each weight is again compared using equation (1). For the empirical study, the NAEP procedure was allowed to go through 10 iterations, using c = 10. Smaller or larger values of c will generate different trimming levels.

5. Weight Distribution

This trimming procedure is based on an assumed distribution for the sampling weights, and no survey data are used. If the selection probabilities are assumed to follow a Beta distribution, the sampling weight distribution can be shown to be of a form that is essentially an inverse of a beta variate. The percentiles for the cumulative distribution function ($F_w(w)$) of the weight distribution can be computed using the complete Beta distribution. The values for the cumulative distribution function of the weight distribution F_w (w) is

$$F_w(w_o) = \int_0^{1/nw_o} (1 - u)^{\beta-1}\ u^{\alpha-1}\ du/B(\alpha,\beta).$$

Estimates for alpha and beta can be computed as

$$\hat{\alpha} = [\bar{w}\ (n\bar{w} - 1)/\ ns_w^2] + 2 \qquad (2)$$

$$\hat{\beta} = (n\bar{w} - 1)\ [\bar{w}\ (n\bar{w} - 1)/\ ns_w^2 + 1] \qquad (3)$$

where

$$\bar{w} = \Sigma_n\ w_i/n$$

$$s_w^2 = \Sigma_n\ (w_i - \bar{w})^2/n.$$

The weight distribution trimming procedure compares the distribution of the weights relative to the theoretical distribution.

For the empirical study, the probability of occurrence criterion was set at 0.01; that is, a weight with a value in excess of w_{op} where $1 - F(w_{op}) = 0.01$ was trimmed to w_{op}. For the first of 10 iterations, the original weights were used to estimate α and β using equations (2) and (3), respectively. For the second to the tenth iteration, α and β was estimated using the weights from the prior iteration.

III. EMPIRICAL INVESTIGATION

A. Overview

The goals of the empirical study are to investigate and evaluate the effect of weight trimming procedures on simple linear regression coefficients from a population that can be fully enumerated.

B. Empirical Study Design

For the empirical study, county-level data on medical resources and demographic characteristics of the county population were obtained from the Area Resource File (ARF) data base developed by the Health Resources and Services Administration of the U.S. Department of Health and Human Services. A total of 2,989 of the 3,080 county units in the ARF data base was used. Excluded county units either had a very large (greater than 200,000 households) or a very small (less than 250 households) count of households, or were likely to have a zero value for one or more of the data items (county units in Hawaii and Alaska). Two hundred (200) samples of 100 units each were

selected using the probability minimal replacement sampling procedure developed by Chromy (1979) for the PPS selection using the number of households in 1980 as the size measure.

Seven data items were used for the simple linear regressions:

1. Average temperature in July
2. Average temperature in January
3. Birth rate among teenagers
4. Percentage white age 10-17
5. Percentage of households with access to a telephone
6. Per capita income
7. Median family income.

Six regression equations (R1 - R6) were computed with the following dependent (DEP) and independent (IND) variable (using the number from the list above to denote the variable and r is the correlation):

R1: DEP: 1, IND: 2 (r = 0.733)
R2: DEP: 3, IND: 4 (r = -0.458)
R3: DEP: 5, IND: 4 (r = 0.537)
R4: DEP: 5, IND: 6 (r = 0.583)
R5: DEP: 3, IND: 6 (r = -0.315)
R6: DEP: 7, IND: 6 (r = 0.702)

For the two weight trimming procedures using the mean estimator as the trimming estimator, four data items were used to identify the trimming level. The four variables were chosen because of the estimated correlation between the data item and the sampling weight across the 200 samples: (1) median family income (negative), (2) birth rate among teenagers (positive), (3) percentage of 5 to 17 year old population that are white (zero), and the average temperature in July (zero).

The TS procedure and the estimated MSE procedures evaluate statistics for predetermined candidate trimming levels. For the empirical study, 20 candidate trimming levels were computed for each sample. The candidate trimming levels were computed as follows:

a. Trimming level 1 is the next to largest weight;

b. Trimming level 2 is the sum of the next to largest and the third largest weight, divided by 2;

c. Trimming level 3 is the sum of the next to largest, the third largest, and the fourth largest weight, divided by 3;

d. Trimming level 4 to 20 were computed as similar averages of the largest weights (excluding the largest weight).

For each trimming candidate level, a set of trimmed adjusted weights were computed. For the other procedures (as described in Section II), the trimming levels were generated within the procedure.

The regression coefficients were computed using the untrimmed weights and each set of trimmed weights. The sampling variance of each regression coefficient was also computed and the MSE, where the bias term was computed as the difference between the trimmed weight estimator and the untrimmed weight estimator.

C. Summary of Results

The findings of the empirical study show that, of the five procedures, the estimated MSE procedure using the regression coefficients imposed less weight trimming and, as was expected, performed the best in terms of the effect on the smallest MSE. The TS and the estimated MSE procedures using the mean estimator tended to perform similarly. Also, the NAEP procedure and the weight distribution procedure operated almost identically to each other.

In comparing the TS and the estimated MSE procedures, the estimated MSE procedure using the mean estimator resulted in the larger average reduction (8.0 percent reduction over the 6 regressions) in the variance over the 200 replicated samples than estimated MSE procedure using the regression coefficients (5.3 percent reduction) and the TS procedure (5.9 percent reduction) for the six regressions (Table 1.). As seen in Table 1, the range of reduction in the sampling variance was the greatest for the estimated MSE procedure using the mean estimator (1.5 percent to 11.3 percent) than the estimated MSE procedure using the regression coefficients or the TS procedure (approximately 2 percent to 8 percent reduction). In terms of the MSE, the estimated MSE procedure using the regression coefficients had an average MSE change (relative to the sampling variance using untrimmed weights) of 10.2. The procedures using the mean estimator exhibited an average change in MSE of 26.1 and 19.3 percent for the estimated MSE procedure and the TS procedure, respectively. Therefore, although the average

variance reduction was larger for the estimated MSE procedure using the mean estimator, the average change in the MSE was substantially smaller for the estimated MSE procedure using the regression coefficients than for the estimated MSE using the mean estimator or the TS procedure.

Both the NAEP procedure and the weight distribution procedure produced almost identical trimming levels over the 200 samples. The NAEP procedure and the weight distribution procedure resulted in an average variance reduction of 3.2 and 2.9 percent (Table 1) and average MSE change of 16.3 and 15.7 percent, respectively. Therefore, in comparison to the estimated MSE procedure using the regression coefficients, the NAEP procedure and the weight distribution procedure resulted in more weight trimming. The added weight trimming of these procedures resulted is less variance reduction and larger MSEs than the estimated MSE procedure using the regression coefficients.

The effect of weight trimming on both the estimate and its variance can be evaluated by the coverage probabilities of confidence intervals. The 95 percent and the 99 percent confidence intervals were computed for each of the 200 samples using the untrimmed weights and each of the alternative trimmed weights. Table 2 shows the proportion of the confidence intervals from the 200 samples that contained the population value of the regression coefficient (using data from 2,989 counties). All confidence intervals performed at below the nominal level. The intervals based on the untrimmed weights generally covered the population value less frequently than the intervals based on any of the trimmed weights. The weights trimmed using the estimated MSE of the regression coefficient estimate generally resulted in confidence intervals with higher coverage proportions than the intervals based on the untrimmed weights. However, the intervals based on the other weight trimming procedures tended to perform better (higher coverage proportions) than the intervals computed based on weights trimmed using the estimated MSE with regression coefficients. The weights trimmed using the estimated MSE with the mean estimate generally resulted in confidence intervals with the highest coverage proportions of all the weight trimming procedures.

IV. CONCLUSIONS

In terms of the five weight trimming procedure as evaluated in the empirical study, the two estimated MSE procedures and the TS procedure utilize the data and an estimate of the MSE, and these procedures are preferable to the other two procedures. Among these three procedures, the results show the importance of using the estimator of interest. In previous research (Potter 1990), the TS procedure showed some improvement over the estimated MSE procedure for estimators of a mean. Both fared poorly when compared to the estimated MSE procedure using the regression coefficients for estimates of the MSE. However, these two procedures resulted in weights which achieved better coverage probabilities for interval estimates.

Similarly, the weight trimming from the weight distribution procedure and the NAEP procedure fared poorly both in sampling variance reduction and the change in the MSE relative to the estimated MSE procedure using the regression coefficient estimator. The coverage probabilities computed using weights generated by these procedures were as good or better than the coverage probabilities computed from the weights trimmed using these estimated MSE procedure with regression coefficients.

The primary conclusion based on the empirical study results is that weight trimming can have both advantages and disadvantages. The advantages (for example, the reduction in the sampling variances) occurred on average for all procedures, but all procedures also resulted in an increase in the estimated sampling variance for at least some of the 200 replicated samples. The use of the specific estimator of interest (in this case the regression coefficient) was shown to have an important effect. Global trimming of sampling weights based on one estimator can produce potentially misleading results when another estimator is used in analyses. Therefore, the survey analyst needs to be cautious when trimming sampling weights because, unless weight trimming is conducted carefully and evaluated for various estimators and data items, larger sampling variance or substantial bias can result for some survey estimates.

REFERENCES

Benrud, C.H., et al. (1978). "Final Report on National Assessment of Educational Progress:

Sampling and Weighting Activities for Assessment Year 08." Prepared for National Assessment of Educational Progress, Research Triangle Park, N.C.: Research Triangle Institute.

Chromy, J. R. (1979). "Sequential Sample Selection Methods," in Proceedings of the *American Statistical Association, Section on Survey Research Methods*, pp. 401-406.

Cox, B.G. and McGrath, D.S. (1981). "An Examination of the Effect of Sample Weight Truncation on the Mean Square Error of Survey Estimates." Presented at Biometrics Society ENAR Meeting, Richmond, VA.

Johnson, E.G. et al. (1987). "Weighting Procedures in Implementing the New Design," in The NAEP 1983-1984 Technical Report, (ed.) A. E. Beaton, Princeton, N.J.: National Assessment of Educational Progress, pp. 493-504.

Potter, F. (1988). "Survey of Procedures to Control Extreme Sampling Weights," in Proceedings of the *American Statistical Association, Section on Survey Research Methods*, pp. 453-458.

Potter, F. (1990). "A Study of Procedures to Identify and Trim Extreme Sampling Weights," in Proceedings of the *American Statistical Association, Section on Survey Research Methods*, pp. 225-230.

Table 1. Average Percentage Change in Variance and Mean Square Error Based on Estimates from 200 Replicated Samples

	Procedure				
	Reg	Mean		Weights Only	
Var	Est. MSE	Est. MSE	Taylor Series	NAEP	Weight Dist'n
	% Change in Estimated Sampling Variances				
1 2*	-6.2	-10.1	-6.9	-5.4	-4.9
3 4	-3.6	-4.1	-5.3	-0.8	-0.6
5 4	-2.9	-1.5	-2.2	-0.2	0.1
3 6	-5.1	-11.1	-7.4	-3.7	-3.3
5 6	-5.8	-10.0	-5.8	-3.6	-3.1
7 6	-8.3	-11.3	-7.9	-5.8	-5.4
Avg	-5.3	-8.0	-5.9	-3.2	-2.9
	% Change in Estimated Mean Square Error				
1 2*	6.1	13.5	12.1	9.0	8.8
3 4	13.1	37.1	16.0	20.3	19.4
5 4	8.7	29.3	16.6	14.7	14.3
3 6	10.1	19.0	18.7	15.6	15.0
5 6	7.4	22.3	20.0	14.8	14.3
7 6	15.6	35.4	32.6	23.4	22.3
Avg	10.2	26.1	19.3	16.3	15.7

% Change = 100 * (Trimmed - Untrimmed) / Untrimmed.

*Variables in Regression (dependent and independent)
1 Temperature in July
2 Temperature in January
3 Birth rate among teenagers
4 Percentage white ages 10-17
5 Percentage of households with telephones
6 Per capita income
7 Median family income

Table 2. Coverage Probabilities for CI Error Based on Variance Estimates from 200 Replicated Samples

		Trimming Procedure				
		Reg	Mean		Weights Only	
Var	Orig. Wts.	Est. MSE	Est. MSE	Taylor Series	NAEP	Weight Dist'n
		Coverage of 95% Confidence Interval				
1 2*	84.5	86.0	88.0	87.0	87.5	88.0
3 4	80.0	82.5	84.0	83.0	83.5	82.5
5 4	86.0	86.5	90.0	87.5	89.5	89.0
3 6	85.5	87.5	88.5	88.5	88.5	88.5
5 6	79.5	83.5	86.5	86.0	85.0	85.0
7 6	64.0	66.0	62.5	65.0	67.5	67.5
		Coverage of 99% Confidence Interval				
1 2*	92.0	94.0	95.0	94.5	94.5	94.5
3 4	91.0	92.5	94.0	92.0	92.5	93.0
5 4	92.5	95.0	96.0	95.5	94.5	94.5
3 6	92.5	95.5	96.5	97.0	96.0	96.0
5 6	90.5	94.5	97.0	96.5	95.5	95.5
7 6	77.5	79.5	79.5	82.0	80.5	80.5

95% CI = $\theta_s \pm 1.96 * \sqrt{Var(\theta_s)}$

99% CI = $\theta_s \pm 2.58 * \sqrt{Var(\theta_s)}$

* Variables in Regression (dependent and independent)
1 Temperature in July
2 Temperature in January
3 Birth rate among teenagers
4 Percentage white ages 10-17
5 Percentage of households with telephones
6 Per capita income
7 Median family income