Journal of the Nigerian Association of Mathematical Physics
Volume 10 (November 2006), **447 - 452**

Effective utilization of weighting adjustment for the estimates of means in survey non-response

**O. R. Oniyide and D. A. Agunbiade*
Department of Mathematical Sciences
Olabisi Onabanjo University, Ago-Iwoye, Nigeria.

Abstract

This paper provides a useful application for comparison on the use of Adjusted Estimates (Weighting Adjustment) as against Unadjusted Estimates for estimate of Mean in survey Non-response .The use of response propensity and the predicted mean of the outcome variable for cell creation are stressed .The results from our empirical study emphasize the efficacy of Weighting Adjustment over the Unadjusted estimates .We adopt the following criteria: Variance, Bias and Mean Square Error in reaching our conclusion.

Keywords: Weighting adjustment, potential stratifiers, adjustment cells, non-Response.

1.0 Introduction

It is not impossible to have individuals who failed to provide information to a questionnaire because of non contact or refusal to respond to the whole questionnaire (Unit Non response).

At times people provide incomplete information for which some items are missing (Item non response). The most common method of adjustment for unit non response is weighting. In weighting adjustment, respondents and non respondents are classified into adjustment cells based on covariate information known for all units in the sample and a non response weight is computed for cases in a cell proportional to the inverse of the response rate in that cell Little (2003). These weights often multiply the sample weight, and the overall weight is normalized to sum to the number of respondents in the sample. See Oh and Scheueren (1983), for detailed discussion on non response weighting. A Simple related approach to non response weighting is Post-stratification, Holt and Smith (1979), which applies when the distribution of the population over adjustment cells is available from eternal sources such as a census. The weight is then computed as the inverse of the ratio of the number of respondents in a cell to the population count in that cell.

Researchers view weighting as a means of reducing bias from unit non response, and this role is synonymous to the role of sampling weights, and is related to the design unbiasedness property of the Horvitz-Thompson (1952) estimator of the total for which units are weighted by the inverse of their selection probabilities. Non response weighting can be seen as a natural extension of this idea, where included units are weighted by the inverse of their inclusion probabilities, estimated as the product of the probability of selection and the probability of response given selection. In particular, nonresponse weight is simply the inverse of inclusion probability, where inclusion probability is the product of probability of selection and probability of response given selection. In actual fact the basic steps in nonresponse weighting adjustment includes the following:

1) Classifying Respondent and Non respondent into adjustment cells based on covariate information known for all units in the sample.
2) Compute the non response weight for cases in a cell proportional to the inverse of the response rate in the cell.

***Corresponding Author.**

3) Obtain the product of these weights and the sample weight and normalized the overall weight to sum to the number of respondents in the sample.

In particular, the basic aim of this paper is to examine the efficacy of weighting adjustment as a useful tool to handle non response problem over the unadjusted estimator.

We outlined this work as follows; Section 1 gives a brief overview of weighting adjustment method, Section 2 discusses the theoretical background of weighting adjustment. In section 3, we introduce our symbols as it is used in this study and the corresponding interpretation along side with the estimators, while Section 4 gives the results obtained from using these estimators for clear comparison.

2.0 Theoretical background

In work with population means, total and related parameters survey non response is of practical concern for many reasons such as

(1) Biases in point estimators
(2) Inflation of the variances of point estimators, and
(3) Biases in customary estimators of precision.

In order to illustrate these concerns, we consider a finite population U containing N units with items $Y_i, i = 1...N$ and define the associated population mean $\mu = N^{-1}\sum_{i=1}^{N} Y_i$ (2.1)

Let us in addition, define S to be set of indices of n sample units selected through a complex design D. This design may involve a combination of stratification, Clustering and Unequal probability of selection. For each unit i in the population, let π_i equal the probability that unit i is included in the sample S, and define the associated probability weight $w_i = 1/\pi_i$, then a standard point estimator of the population mean μ is

$$\hat{\mu}_f = \sum_{i\in s} w_i \Big/ \sum_{i\in s} w_i Y_i \qquad (2.2)$$

Under moderate regularity conditions, the full-sample point estimator $\hat{\mu}_f$ is evaluated with respect to the sample design for fixed characteristics, $Y_i, i = 1...N$.

Consider the presence of non-response by one or more elements; one cannot compute the full-sample estimator (2.1). A simple alternative is the unadjusted estimator

$$\hat{\mu}_{uA} = \sum_{i\in s} w_i \Big/ \sum_{i\in s} w_i Y_i \qquad (2.3)$$

Where now the sample S is partitioned into subsets R and M containing responding and missing units respectively. We may rewrite expression (2.3) as

$$\hat{\mu}_{uA} = \sum_{i\in s} w_i r_i \Big/ \sum_{i\in s} w_i r_i Y_i \qquad (2.4)$$

where r_i is the response indicator for element i define as follows $r_i = \begin{cases} 1, \text{ if element i responds} \\ 0, \text{ otherwise} \end{cases}$

The operating characteristics of $\hat{\mu}_{uA}$ depends on the non-response process, which can be formalized using the quasirandomization Oh and Scheuren, (1983).

One approach to reducing the bias of (2.4) is to apply a non-response weighting adjustment. Suppose that unbiased estimates $\hat{p}_i$ can be computed using available information from the survey or external sources, for full discussion on the estimation of p_i from auxiliary data see Little et. al (2002). One could compute the approximately unbiased estimator

$$\hat{\mu}_{uA} = \sum_{i\in R} wp_i \Big/ \sum_{i\in R} wp_i Y_i \qquad (2.5)$$

Where $wp_i = w_i / p_i$ is a weight that has been adjusted by the inverse of the estimated selection probability for unit *i*. Oh and Scheuren (1983) and Sarndal and Swenson (1987). The efficiency issue (2) then relates to the increase of the variance of the estimator $\hat{\mu}_p$ relative to the idealized estimator $\hat{\mu}_f$ and the inference issue (3) depends on the bias and stability properties of variance estimators applied to (2.5) and the coverage of confidence intervals and test constructed using (2.5) and these variance estimators.

3.0 Description of notation and estimators

Following the definitions and notation as used by Little (1986) with further modification, we present the following notation for population and sample quantities, R for respondents N – for non-respondents.

n_{cR}: The number of sample respondents in cell c, n_c the number of samples units in cell c, b_c is the response rate in cell c and can be written as $b_c = b_{cR}/n_c$. n is the total number of sampled units. b is the response rate in the entire sample and is given as $b = n_R/n$. P_{cR}, the proportion of respondents in cell c and is given as $P_{cR} = n_{cR}/n_R$.

P_c is the proportion of cell c is given as $P_c = n_c/n$,

$\bar{y}_{cR}$ is the respondent mean is cell and is given as $\bar{y}_{cR} = \frac{1}{n_{cR}}\sum_{i=1}^{n_{cR}} y_{ic}$

where y_{ic} is the characteristic value of the ith respondent unit in cell c $\bar{y}_c$, is the mean in cell c and $\bar{y}_c = \frac{1}{n_c}\sum_{i=1}^{nc} y_{ic}$

where,

$$y_{ic} = \begin{cases} y_i \text{ value if the ith unit responds} \\ 0, \text{ otherwise} \end{cases}$$

$\bar{y}_R$ is the respondent overall mean (unadjusted) and $\bar{y}$ is the overall mean. The symbol N represents population units. $\bar{Y}$ represents mean, b represent response rate and p represents cell proportion. The suffix c refers to adjustment cell and suffix R denotes restriction to respondents.

3.1 Estimators

Consider that the population N has be classified into adjustment cells based on covariate information known for all units in the population we have c adjustment cells. Thus the following estimators are obvious for unadjusted and adjusted estimator through weighting.

Estimator	Mean	Variance	Bias	M. S. E.
Unadjusted	$\bar{y}_R = \sum p_{cR}\bar{y}_{cR}$	$V(\bar{y}_R) = \sigma^2/n_R + \sum \pi_{CR}(\mu_{cR} - \bar{\mu}_R)^2/n_R$	$b(\bar{y}_R) = \mu_R - \mu$	$m.s.e.(\bar{y}_R) = b^2(\bar{y}_R) + V(\bar{y}_R)$
Adjusted (Weighting)	$\hat{\mu}_p = \sum p_{cR}\bar{y}_{CR}$	$V(\hat{\mu}_p) = (1+L)\sigma^2/n_R + \sum_{i=1}^{C} \pi_c(\mu_{cR} - \mu_R^2)^2$	$b(\hat{\mu}_p) = \sum \pi_c(\mu_{cR} - \mu_c)$	$m.s.e(\hat{\mu}_p)$ $b^2(\hat{\mu}_p) + V(\hat{\mu}_p)$

L is the variance of the non-response weight or equivalently the square of coefficient variation noted by Kish (1992).

Where σ^2 is the population variance and is given as

$\sigma^2 = \sum_{c=1}^{c}\sum_{c=1}^{nc}(Y_{ic} - Y_c)/N$

$\tilde{\mu} = \sum \pi_c \mu_{cR}$ is the mean of the responding unit using weighting adjustment.

$\bar{y}_{cR} = \mu_{cR}$ is the mean of the responding unit in cell c.

$\mu = \sum \pi_c \mu_c$ is the true population mean. π is the response rate analogous to b.

$\pi_{cR} = n_{cR}/n_c$ is the response rate in cell c.

4.0 Practical applications

In this section we present an analytical justification of the efficacy of weighting adjustment over the unadjusted estimates using a real life data collected from 500 sampled units. The purpose is to obtain the average monthly income of these units. These units were interviewed using Questionnaire Administration, 385 sampled individuals responds to a question on y (monthly income). A classification of respondent and non-respondent into three adjustment cells is done defined by the variable A: region (low, middle and high income earners). This pattern of cell classification is analogous to data categorization as done by Osungade (1989). Within each cell, the response rates were obtained and we estimate the respondent mean income. For the avoidance of sample weights, we assume that all individuals in the population have an equal chance of selection.

We also consider the choice of association with survey outcome intuitively our sampled population portray an appreciate degree of association with response propensity.

Table 4.1: Summarized data

	$C = 1$	$C = 2$	$C = 3$	Total
Total number in cell	230	150	120	500
Number of respondent to characteristic (y) in cell c.	202	115	68	385
Number of non response	28	35	52	115
Response rate	0.8783	0.7667	0.5667	
Mean income $\bar{y}_{cR}$ (₦'000)	15.8	35.4	68.2	
Total income	3,191.60	4,071.60	4,637.60	

As generally done in most data analysis, a simple estimate of the mean in the population is the respondent mean $\bar{y}_R = 30.89$. The outcome of the survey shows that non response rate is higher in high income region ($c = 3$) than in low income region $c = 3$. These trends follow the notion as claimed by Krosnick (1989). However we may be inclined to view $\bar{y}_R$ as an underestimates of the mean (Biased). To obtain an adjusted mean $\hat{\mu}_p = 34.32$, that plausibly reduces the bias from restriction to the respondent sample. It should be realised that the same adjusted mean $\hat{\mu}_p$ can be obtained by imputing the cell respondent for all non respondent in that cell. For example by imputing 35.4 for the entire non respondent in cell 2.

4.1 Variance of unadjusted estimates

Table 4.2: Estimates of variance for unadjusted estimates.

C	π_{cR}	$\mu_{cR} - \bar{\mu}_R$	$\pi_{cR}(\mu_{cR} - \bar{\mu}$	$\pi_{cR}(\mu_{cR} - \bar{\mu}_R)^2 / n_R$
C = 1	0.8783	15.09	199.996	0.5195
C = 2	0.7667	4.51	15.59	0.0405
C = 3	0.5667	37.3	788.44	2.0479

$$\sum \pi_{cR} \frac{(\mu_{cR} - \bar{\mu}_R)}{n_R} = 2.6079$$

Thus, $V(\bar{y}_R) = \sigma^2 / n_R + 2.6079$

4.1 Variance of the adjusted mean.

The estimator for variance of the adjusted estimates is given as $V(\hat{\mu}_p)=(1+L)\frac{\sigma^2}{n_R}+\sum\pi_c\frac{(\mu_{cR}-\tilde{\mu}_R)^2}{n}$

Table 4.3: Estimates of variance for adjusted estimate

c	π_c	$\mu_{cR}-\tilde{\mu}_R$	$\pi_c(\mu_{cR}-\tilde{\mu}_R)^2$	$\pi_c(\mu_{cR}-\bar{\mu}_R)^2/n_R$
c = 1	0.46	-18.456	156.69	0.3134
c = 2	0.30	1.144	0.3926	0.0008
c = 3	0.24	33.944	276.53	0.5531

$\sum\pi_c\frac{(\mu_{cR}-\tilde{\mu}_R)^2}{n}=0.8673$ and $V(\hat{\mu}_p)=\frac{(1+L)\sigma^2}{n_R}+\sum\pi_c\frac{(\mu_{cR}-\tilde{\mu}_R)}{n}=\frac{(1+L)\sigma^2}{n_R}+0.8673$

Table 4.4: Estimates of L (Squared Coefficient of Variation)

c	p_{cR}	w_c	w_c-1	$p_{cR}(w_c-1)$
c = 1	0.8783	0.8767	0.0152	0.0134
c = 2	0.7667	1.0043	1.849×10^{-5}	1.41×10^{-5}
c = 3	0.5667	1.3588	0.1288	0.00730

Since $L=p_{cR}(w_i-1)$ so we have $L=0.0864$

In particular the estimates of $V(\hat{\mu}_p)$ would be $V(\hat{\mu}_p)=1.864\frac{\sigma^2}{n_R}+0.8673$, $\sigma^2<<1$, Thus by comparing $V(\bar{y}_R)$ and $V(\hat{\mu}_p)$ we have $V(\hat{\mu}_p)<V(\bar{y}_R)$.

4.3 Estimates of the Bias

Similarly, for the Bias of unadjusted estimates we have, $b(\hat{\mu}_P)=\sum\pi_c(\mu_{cR}-\mu_c)$. This can be written in terms of $b(\hat{\mu}_P)$ as $b(\bar{y}_R)=b(\hat{\mu}_P)+\mu_R-\mu_{R_1}adj$ where $\mu_{R,Adj}$ the respondent mean adjusted and can be written as $\mu_{R.Adj}=\sum\pi_c\,\mu_{cR}$. Let μ_c the mean of y_p in cell c be represented as θ_c i.e. $\mu_c=\theta_c.\mu_{cR}$ is as defined in our notation. We have

Table 4.4

c	π_C	$\mu_{CR}\,\mu_C$	$\pi_C(\mu_{CR}-\mu_R)$
c =1	0.46	$15.8-\theta_1$	$0.46(15.8-\theta_1)$
c =2	0.30	$35.4-\theta_2$	$0.30(35.4-\theta_2)$
c =3	0.24	$68.2-\theta_3$	$0.24(68.2-\theta_3)$

$b(\hat{\mu}_P)=0.46(15.8-\theta_1)+0.30(35.4-\theta_2)+0.24(68.2-\theta_3)=7.268-0.46\theta_1+10.62-0.30\theta_2+16.368-0.24\theta_3$

$b(\hat{\mu}_P)=34.256-0.46\theta_1-0.30\theta_2-0.24\theta_3$

From the table below it can be seen $\mu_{R.Adj.}=34.256$. In addition the following table are obvious,

c	π_c	μ_{cR}	$\pi_c\,\mu_{cR}$
c =1	0.46	15.8	7.268
c =2	0.30	35.4	10.62
c =3	0.24	68.2	16.368

Following a simple comparison of Bias

$b(\bar{y}_R) - b(\hat{\mu}_P)$ is positive. Since $\mu_R - \mu_{R,Adj.} = \mu_R - 34.256$ and $\mu_R > \mu_{R,Adj.}$

Consequently, in the light of the above results of variances and bias and following an ordering properly of real number $\Re$, thus

$$m.s.e(\hat{\mu}_P) < m.s.e(\bar{y}_R)$$

5.0 Conclusion

The analysis indicates the reduction in both variance and bias of the weighted estimates as compared with the unadjusted estimates for the mean income. Thus, the argument that weighting increases variance is an oversimplification (if the stratifying variables are associated with the outcome variable). The analysis further reveals that the most important feature for inclusion in Weighting Adjustment is that they are predictive of the survey outcomes; prediction of the propensity to respond is a secondary (though useful) goal. In actual fact, the situation when Weighting Adjustment is most effective is when it reduces both variances and the bias.

References

[1] Holt, D and Smith, T.M.F (1979) Post stratification. Journal of the Royal Statistical Association, 81,945-61.

[2] Horvitz, D.G and Thompson, D.J (1952) A Generalization of Sampling without Replacement from Finite Universe, Journal of the American Statistical Association, 47,663-85.

[3] Kalton, G. and Kapryzyk, D. (1982): Imputing for Missing Survey Responses. Proc. Survey. Res. Method. Am. Statist Association 22-31.

[4] Little R. J. A. (1982): Models for Non-response in Sample Surveys. J. Am. Statist. Association 77, 327-350.

[5] Little R.J.A (1986): Survey Non-response Adjustment for Estimates of Means. International Statistical Review 54, 139-157.

[6] Little, R. J. A. and Rubin, D. B. (2002): Statistical Analysis with Missing Data 2nd Edition Wiley, New York.

[7] Oh, H. L. and Scheuren F. (1983): Weighting Adjustment for Unit Non-response. In Incomplete Data in Sample Surveys 2, Ed. W. G. Madow, I. Olkin and D. B. Rubin, pp. 143-184 New York Academic Press.

[8] Oniyide, O. R. and Adekanbi, D. B (2006); The efficacy of weighting adjustment for the estimates of population mean in survey nonresponse. Accepted for publication, (NAMP)

[9] Rubin D. B. (2002): Multiple Imputations for Non-response in Surveys, Wiley, New York.

[10] Sarndal, C.E and Swenson, B (1987). A general view of Estimation for two-phases of Selection with application to two-phase sampling and non-response, International Statistical review, 55, 279-94.

[11] Vartivarian, S. and Little, R. J. A. (2002): On the Formation of Weighting Adjustment Cells for unit non-response, proceeding of the survey Research methods section. American statistical Association 2002.