Journal of Physical Sciences,
Vol. 1, No. 1, 1 – 6, April 2010

A Class of Product-Type Estimator when there is Unit Non-Response in the Study Variable

Fabian C. Okafor
Dept of Statistics University of Nigeria, Nsukka, Nigeria
fc_okafor@yahoo.com

Abstract

We have proposed a class of product-type estimator for the estimation of the population mean when there is unit non-response. We are able to show, under a given model, that our estimator is more efficient than the sample mean given by Hansen and Hurwitz (1946). We also established theoretically the condition under which our estimator is more efficient than other particular estimators derivable from the given class of estimators. It has also been shown that it is better to adjust the auxiliary variable for non-response even when there is complete response in the auxiliary variable.

Key word: *Product-type, Estimator, Variable, Non-response, Bias, Mean square error.*

1. Introduction

Consider a finite population $U=\{U_1,U_2,...U_N\}$ of N units. Let (x_i,y_i) $i=1,2,...,n$ be observations on a sample of n units selected from U by simple random sampling without replacement (srs). Ratio estimation is normally used to improve the estimate of the population mean of the study variate , y, when the auxiliary variable, x, is positively correlated with y. If the auxiliary variable is negatively correlated with the study variable the product estimator $\bar{y}_p=\overline{yx}/\bar{X}$ is used. $\bar{X}$ is the population mean of x. $\bar{x}$ and $\bar{y}$ are the sample means of x and y respectively. This estimator assumes total non-response in the variables. In practical situation not all the sample units agree to take part in a survey; this gives rise to unit or total non-response.

Rao (1986) and Okafor and Lee (2000) have discussed the treatment of unit non-response in ratio estimation of the mean. Discussions on product estimators so far have been under the condition of total response of all the sample units. Sisodia and Dwivedi (1981) proposed a class of ratio-cum-product estimator where there is complete response of the form $T=(1-a)\bar{y}+a\bar{y}\left(\bar{X}/\bar{x}\right)^{\alpha}$. Shah and Shah (1978) and Singh (1967) also proposed their own type of estimators. In this study a class of product-type estimator of the population mean is proposed for a situation when there is unit or total non-response in the study variable but complete response in the auxiliary variable.

2. The Proposed Estimator

Suppose x and y are negatively correlated and that the population mean of x is known. We propose a class of product-type estimator of the population mean $\bar{Y}$ of y when some sample units fail to supply information on the study variate but all supply information on the auxiliary variable. Let n_1 units out of the n units in the sample supply information on y and n_2 fail to respond. A subsample of size $m=n_2/k$ (k is a given constant less than one) is selected from the n_2 units and revisited. With better interviewing technique we assume that all the m units supplied the required information

The class of product-type estimator of the population mean is then given by

$$T^*=\bar{y}^*\left\{\frac{\bar{x}^*}{\bar{x}^*-\phi\left(\bar{x}^*-\bar{X}\right)}\right\}^{\alpha},\quad \alpha>0,\phi>0 \qquad (2.1)$$

$$\bar{z}^*=\frac{n_1\bar{z}_1+n_2\bar{z}_m}{n},\quad z=x,y$$

is the Hansen and Hurwitz (1946) estimator of the population mean adjusted for non-response..

$\bar{z}_1$ is sample mean based on n_1 respondents at the first attempt.

$\bar{z}_m$ is the sample mean based on the m subsample of non-respondents who respond at the second attempt. α and ϕ are suitably chosen constants.

Bias and Mean Square Error of

(2.1) can be rewritten in terms of delta notation as $T^* = \bar{Y}\left\{\frac{(1+\delta\bar{y}^*)(1+\delta\bar{x}^*)}{1+\delta\bar{x}^* - \phi\delta\bar{x}^*}\right\}^{\alpha}$

Now expanding the above equation to an infinite series, the first order approximation to the bias of T^*, $E\left(T^* - \bar{Y}\right)$, after simplification is

$$B(T^*) = \frac{g\alpha\phi C_x S_y}{n}\left[\frac{\gamma}{2}\{(\alpha-1)-(1-\phi)(1+\alpha)\}+\rho\right] + \frac{w}{n}C_{2x}S_{2y}\left(\gamma_2+\rho_2\right) \quad (2.2)$$

$C_z = S_z/\bar{Z}$ is the population coefficient of variation. $C_{2z} = S_{2z}/\bar{Z}$; $z = x, y$

S_{2z} is the population standard deviation in the non-respondent stratum.

ρ and ρ_2 are the population correlation coefficients for the entire population and the non-respondent population respectively.

$\gamma = C_x/C_y$; $\gamma_2 = C_{2x}/C_{2y}$; $g = 1 - n/N$ and $w = W_2(k-1)$. W_2 is the proportion of non-respondents in the population of N units.

And the first order approximation of the mean square error (mse) of T^* is given by

$$E\left(T^* - \bar{Y}\right)^2 = V(T^*) = \frac{g}{n}S_\gamma^2 + \frac{w}{n}S_{2\gamma}^2 \quad (2.3)$$

$$S_\gamma^2 = S_y^2\left\{1+(\alpha\phi\gamma)^2 + 2\alpha\phi\gamma\rho\right\}; \quad S_{2\gamma}^2 = S_{2y}^2\left\{1+(\alpha\phi\gamma_2)^2 + 2\alpha\phi\gamma_2\rho_2\right\}$$

By differentiating (2.3) with respect to ϕ, equating to zero and solving the optimum value of ϕ for a given α that minimizes the mse of T^* is

$$\phi_o = -\frac{gS_y^2\rho\gamma + wS_{2y}^2\rho_2\gamma_2}{\alpha\left\{gS_y^2\gamma + wS_{2y}^2\gamma_2\right\}} = -\frac{\beta^*}{\alpha} \quad (2.4)$$

For a given value of ϕ, the optimum $\alpha_o = -\beta^*/\phi$

Since there is a complete availability of information on the auxiliary variable and if the sample mean of x is not adjusted for non-response, the estimator T^* boils down to

$$T = \bar{y}^*\left\{\frac{\bar{x}}{\bar{x} - \phi\left(\bar{x} - \bar{X}\right)}\right\}^{\alpha}, \quad \alpha > 0, \phi > 0 \quad (2.5)$$

with bias equal to

$$B(T) = \frac{g\alpha\phi C_x S_y}{2}\left[\frac{\gamma}{2}\{\alpha - 1 - (1-\phi)(1+\alpha)\} + \rho\right] \quad (2.6)$$

and approximate mse is

$$V(T) = \frac{g}{n}S_\gamma^2 + \frac{w}{n}S_{2y}^2 \quad (2.7)$$

Here the optimum ϕ is $-\rho/\alpha\gamma$ for a given α.

3. Optimum Values of *n* and *k* for a Given Cost Function

Consider the simple cost function given by Cochran (1977, p. 371).

$$C = c_o n + c_1 n_1 + c_2 n_2/k$$

c_o = the cost of first attempt on each of the n sample units.

c_1 = the cost of processing the result from the first attempt.

c_2 = cost of collecting and processing result from the non-respondents stratum.

The expected cost of the survey will be

$$C_e = c_o n + c_1 n W_1 + c_2 n W_2 / k \quad (3.1)$$

W_1 is the population proportion of the respondents and, $W_2 = 1 - W_1$ is the population proportion of the non-respondents.

Using this expected cost function and the mse of T^* in (2.3), and by Cauchy-Schwarz inequality, the optimum value of k is

$$k_o = \sqrt{\frac{c_2(1+W_2\delta)}{\delta(c_o + c_1 W_1)}}; \quad \delta = \frac{S_{2\gamma}^2}{S_\gamma^2} \quad (3.2)$$

And for fixed cost, C, the optimum sample size n is $\quad n_o = C(c_o + c_1 W_1 + c_2 W_2 / k_o)^{-1} \quad (3.3)$

Whereas for a specified mse, V_o of T^*, the optimum n will be

$$N\{S_\gamma^2 + W_2(k_o - 1)S_{2\gamma}^2\}(NV_o + S_\gamma^2)^{-1} \quad (3.4)$$

Optimum values of the sample size, n and k for given values of C, c_o, c_1, c_2, δ and W_2 are presented in Table1 below.

Table 1: Optimum Values k_o and n_o

C	c_o	c_1	C_2	W_2	δ	k_o	n_o
500	1	1.5	3	0.3	0.5	1.835	197
500	1	1.5	3	0.6	0.5	2.208	207
500	1	1.5	4.5	0.3	0.5	2.247	189
500	1	1.5	4.5	0.6	0.5	2.704	192
500	1	1.5	3	0.3	1	1.379	185
500	1	1.5	3	0.6	1	1.732	189
500	1	1.5	4.5	0.3	1	1.689	175
500	1	1.5	4.5	0.6	1	2.121	174
500	1	1.5	3	0.3	1.5	1.189	178
500	1	1.5	3	0.6	1.5	1.541	181
500	1	1.5	4.5	0.3	1.5	1.457	168
500	1	1.5	4.5	0.6	1.5	1.887	165

4. Particular Estimators from T^*

We shall now give some of the product estimators of the population mean derivable from the composite estimator T^* in (2.1). $t_1^* = \bar{y}^*$ for $\phi = 0$ or $\alpha = 0$. This is the Hansen & Hurwitz (1946) estimator of the mean corrected for unit non-response. The bias of this estimator is zero and variance is.

$$V(t_1^*) = (gS_y^2 + wS_{2y}^2)/n \qquad (4.1)$$

This estimator, t_1^* is used when there is no auxiliary variable.

$t_2^* = \bar{x}^* \bar{y}^* / \bar{X}$ for $\alpha = 1, \phi = 1$ is the usual product estimator adjusted for unit non-response. The approximate bias and mse of t_2^* are obtained by setting $\alpha = \phi = 1$ in (2.2) and (2.3) respectively. The bias of t_2^* is then

$$B(t_2^*) = \frac{gS_y C_x \rho + wC_{2x} S_{2y}(\gamma_2 + \rho_2)}{n} \qquad (4.2)$$

with approximate mse as

$$V(t_2^*) = \frac{gS_y^2(1+\gamma^2+2\gamma\rho) + wS_{2y}^2(1+\gamma_2^2+2\gamma_2\rho_2)}{n} \qquad (4.3)$$

The third estimator is

$$t_3^* = \frac{\bar{x}^* \bar{y}^*}{\bar{x}^* - \phi(\bar{x}^* - \bar{X})} \quad \text{for } \alpha = 1$$

The approximate bias and mse are as follows:

$$B(t_3^*) = \frac{gC_x S_y[\phi\rho - \phi(1-\phi)\gamma] + wC_{2x} S_{2y}(\gamma_2 + \rho_2)}{n} \qquad (4.4)$$

when $\alpha = 1$ is substituted in (2.2) and the variance is

$$V(t_3^*) = \frac{gS_y^2(1+\phi^2\gamma^2+2\phi\gamma\rho) + wS_{2y}^2(1+\phi^2\gamma_2^2+2\phi\gamma_2\rho_2)}{n} \qquad (4.5)$$

when $\alpha = 1$ is substituted in (2.3).

The last estimator is

$$t_4^* = \bar{y}^* \left(\frac{\bar{x}^*}{\bar{X}} \right)^{\alpha} \quad \text{for } \phi = 1$$

Its approximate bias is derived from (2.2) by setting $\phi = 1$ and this gives

$$B(t_4^*) = \frac{gC_x S_y[\alpha\rho - \alpha(\alpha-1)\gamma/2] + wC_{2x} S_{2y}(\gamma_2 + \rho_2)}{n} \qquad (4.6)$$

While the variance from (2.3) is

$$V(t_4^*) = \frac{gS_y^2(1+\alpha^2\gamma^2+2\alpha\gamma\rho) + wS_{2y}^2(1+\alpha^2\gamma_2^2+2\alpha\gamma_2\rho_2)}{n} \qquad (4.7)$$

5. Comparison of the Proposed Estimators.

In the comparison of the estimators we shall first adopt the procedure of Sisodia & Dwivedi (1981) and later we shall use simulated data to carry out an empirical comparison.

T^* is more efficient than t_1^* if $0 < \phi < -2\beta^*/\alpha$ for a given value of α. Since we are concerned with the product estimator, that is at its best when β^* is negative, the above condition is better presented as $0 < \phi < 2\phi_o$ or $|\phi - \phi_o| < |\phi_o|$
From (2.3) and (4.3) the variance of T^* will always be less than the variance of t_2^* provided that $\alpha^{-1} < \phi < 2\phi_o$ or $|\phi - \phi_o| < |\phi_o - \alpha^{-1}|$.
Using (2.3) and (4.5), T^* will be more precise than t_3^* if the following condition holds:

$$0 < \phi < 2\phi_o \frac{\alpha}{1+\alpha} \text{ or } |\phi - \phi_o| < \left|\phi_o \frac{\alpha - 1}{\alpha + 1}\right|.$$

The condition under which T^* is preferred to t_4^* is derived from (2.3) and (4.7). And this is $|\phi - \phi_o| < |\phi_o - 1|$.
Finally, from (2.3) and (2.7), T^* will be more precise than T if $\phi < -\frac{2\rho_2}{\alpha\gamma_2}$.

Empirical Study

To study the relative performance of the proposed estimators, we adopt the model used by Rao and Sitter (1995). This model is of the form

$$y_{hi} = \beta_h x_{hi} + x_{hi}^{1/2}\varepsilon_{hi} \qquad (5.1)$$

where $\varepsilon_{hi} \sim N(0, \sigma_h^2)$ and $x_{hi} \sim gamma(\lambda_h, \theta_h)$ and ε_i is independent of x_i
h=1 is the stratum of respondents while h=2 is the stratum of non-respondents. The means of x and y are $\mu_{xh} = \lambda_h\theta_h$, and $\mu_{yh} = \beta_h\mu_{xh}$ respectively; and the variances of x and y, and the covariance of x and y are $\sigma_{xh}^2 = \lambda_h\theta_h^2$, $\sigma_{yh}^2 = \beta_h^2\sigma_{xh}^2 + \mu_{xh}\sigma_h^2$ and $Cov(x_h, y_h) = \beta_h\sigma_{xh}^2 = \rho_h\sigma_{xh}\sigma_{yh}$.

Coefficient of variation is $C_{xh} = \lambda_h^{-1/2}$. The population mean and variance of z and covariance of x and y based on the model (5.1) are:

$$\mu_z = \sum_{h=1}^{2} W_h\mu_{zh},\ \sigma_z^2 = \sum_{h=1}^{2} W_h^2\sigma_{zh}^2 \quad \text{and}$$

$$\sigma_{xy} = Cov(x, y) = \sum_{h=1}^{2} W_h^2\beta_h\sigma_{xh}^2\ ;\ z = x, y$$

W_h is the population proportion of units in stratum h.
Let us consider a particular model in which $\varepsilon_{1i} \sim N(0,1)$, $\varepsilon_{2i} \sim N(0,2)$; $x_{1i} \sim$ gamma(7,3) and $x_{2i} \sim$ gamma(3,2) with initial values $\beta_1 = -0.7$ and $\beta_2 = -1.8$, $W_2 = 0.3$, $W_1 = 0.7$. Using these values we have $\rho_2 = -0.87$ $\rho = -0.761$, $\sigma_y^2 = 29.996$, $\sigma_{2y}^2 = 50.88$ $C_x = 0.343$, $C_{2x} = 0.577$, $C_y = 0.405$, $C_{2y} = 0.66$, $\sigma_{xy} = -23.553$, $\gamma_2 = 0.874$ $\gamma = 0.847$, $\Delta = \sigma_{2y}^2/\sigma_y^2 = 1.696$. The performances of the estimator T^* and its particular estimators, for optimum values of ϕ, $n = 197$ and $k = 1.835$ (n and k are from table 1), based on the given particular model, are presented in tables 2 and 3 below.
The estimators t_1^* and t_2^* are the sample mean and the ordinary product estimator respectively; these estimators since they do not involve α and ϕ give the same values for their respective biases and mse's for different values α and ϕ. For the estimator t_3^*, α is always 1 and so the optimum ϕ remains the same so also the bias and the mse.

Table 2 Performance of the Estimators for ρ_2 = -0.87 and ρ = – 0.761

Estimator	ρ	ρ_2	α	ϕ_o	Variance	Bias	MSE	Rel bias%
T^*	-0.761	-0.87	2	0.465	0.08013	-0.0084	0.07174	2.966
T	-0.761	-0.87	2	0.449	0.12877	-0.00829	0.12048	2.31
t_1^*	-0.761	-0.87	2	0	0.21695	0	0.216953	0
t_2^*	-0.761	-0.87	1	1	0.08094	-0.00675	0.07419	2.372
t_3^*	-0.761	-0.87	1	0.929	0.08013	-0.00677	0.07337	2.39
t_4^*	-0.761	-0.87	2	1	0.26222	-0.00598	0.25624	1.169
T^*	-0.761	-0.87	1	0.929	0.08013	-0.00677	0.07337	2.39
T	-0.761	-0.87	1	0.898	0.12877	-0.00677	0.12200	1.887
t_4^*	-0.761	-0.87	1	1	0.08094	-0.00675	0.07419	2.372
T^*	-0.761	-0.87	0.5	1.857	0.08013	-0.00352	0.07661	1.243
T	-0.761	-0.87	0.5	1.797	0.12877	-0.00373	0.12505	1.039
t_4^*	-0.761	-0.87	0.5	1	0.10928	-0.00431	0.10498	1.303
T^*	-0.761	-0.87	0.25	3.716	0.08013	0.00299	0.08312	1.0561
T	-0.761	-0.87	0.25	3.594	0.12877	0.00235	0.13113	0.656
t_4^*	-0.761	-0.87	0.25	1	0.15320	-0.00238	0.15082	0.608

Table 3. Performance of the Estimators for ρ_2 = - 0.95 and ρ = – 0.92

Estimator	ρ	ρ_2	α	ϕ_o	Variance	Bias	MSE	Rel bias%
T^*	-0.92	-0.95	2	0.543	0.02970	-0.01081	0.01889	6.271
T	-0.92	-0.95	2	0.543	0.08808	-0.01041	0.07767	3.507
t_1^*	-0.92	-0.95	2	0	0.21695	0	0.21695	0
t_2^*	-0.92	-0.95	1	1	0.03088	-0.00858	0.02230	4.885
t_3^*	-0.92	-0.95	1	1.086	0.02970	-0.00858	0.02111	4.982
t_4^*	-0.92	-0.95	2	1	0.16211	-0.00923	0.15287	2.293
T^*	-0.92	-0.95	1	1.086	0.02970	-0.00858	0.02111	4.982
T	-0.92	-0.95	1	1.086	0.08808	-0.00818	0.07989	2.758
t_4^*	-0.92	-0.95	1	1	0.03088	-0.00858	0.02230	4.885
T^*	-0.92	-0.95	0.5	2.173	0.02970	-0.00414	0.02556	2.401
T	-0.92	-0.95	0.5	2.172	0.08808	-0.00374	0.08434	1.26
t_4^*	-0.92	-0.95	0.5	1	0.08425	-0.00543	0.07882	1.872
T^*	-0.92	-0.95	0.25	4.346	0.02970	0.00476	0.03445	2.7614
T	-0.92	-0.95	0.25	4.345	0.08808	0.00515	0.09323	1.7355
t_4^*	-0.92	-0.95	0.25	1	0.14069	-0.00315	0.13754	0.84

From tables 2 and 3, we observe that the proposed estimator, T^* has smaller variance than the Hansen and Hurwitz estimator, t_1^*. The proposed estimator has the same variance as t_3^* (for α =1) but different biases. It is more efficient than the estimator T, the estimator obtained when only the variable of interest is adjusted for non-response. This shows that even if there is total response in the auxiliary variable it is still advantageous to adjust it for non-response. We notice that no matter the value of α, T^* has the same variance for the optimum value of the weight, ϕ_0 but different bias and hence different mean square error (mse). It also has the same variance as t_3^* but different bias. The precision of the proposed estimator increases for increasing value of the correlation coefficient. The bias decreases also as the correlation coefficient increases except when α = 0.25.

The mean square errors of t_2^* and t_3^* are less than that of the proposed estimator when α = 0.25, 0.5; the reverse is the case when α = 1, 2. However, the mean square error of T^* is less than the rest of the other particular estimators no matter the value of α.

The ratio of the absolute bias and the standard error of T^*, the relative bias, for a given α increases for increasing value of the correlation; it also increases for increasing value of α.

In general the T^* has higher relative bias than its particular estimators, except when α = 0.25, 0.5. For practical purposes we recommend the use of t_3^*.

References

Cochran, W.G. (1977): Sampling Technique, 3rd *Ed. John Wiley and Sons New York.*

Hansen, M.H. and Hurwitz, W.N. (1946): The Problem of Non-response in Sample Survey. *Journal of American Statistical Association*, 41, pp 517 – 529

Okafor, F.C. and Lee, H. (2000) Double Sampling for Ratio/Regression Estimation with Subsampling for Non-response. *Survey Methodology, Canada*, vol. 26, No 2, 183 – 188.

Rao, P. S.R.S. (1986): Ratio Estimation with Subsampling the Non-respondents. *Survey Methodology*, 12, 2, pp 217 – 230.

Rao, J.N.K and Sitter, R.R. (1995): Variance Estimation Under Two-phase Sampling with Application to Imputation for Missing Data. *Biometrika*, 82, 2, pp 453 – 460.

Shah, S.M. and Shah, D.N. (1978):. Ratio- cum-Product Estimators for Estimating Ratio (Product) of Two Population Parameters *Sankhya: The Indian Journal of Statistics*, 40, C, 2, pp 156 - 166.

Singh, M. P. (1967): Ratio cum Product Method of Estimation. *Metrika*, 12, pp 34 – 42.

Sisodia, B.V.S. and Dwivedi, V.K. ((1981): A Class of Ratio – Product - Type Estimator. *Biometrical Journal*, 23, 2. pp 133 – 139.