Global Journal of Science Frontier Research: F
Mathematics and Decision Sciences
Volume 15 Issue 7 Version 1.0 Year 2015
Type : Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
Online ISSN: 2249-4626 & Print ISSN: 0975-5896

A Robust Regression Type Estimator for Estimating Population Mean under Non-Normality in the Presence of Non-Response

By Sanjay Kumar
Central University of Rajasthan, India

Abstract- In sampling theory, regression type estimators are extensively used to estimate the population mean when the correlation between study and auxiliary variables is high. In this study, we incorporate robust modified maximum likelihood estimators (MMLEs) into regression type estimator in the presence of non-response and their properties have been obtained theoretically. For the support of the theoretical outcomes, simulations under several super-population models have been made. We study the robustness properties of these modified estimators. We show that utilization of MMLEs in estimating finite populations mean leads to robust estimates, which is very advantageous when we have non-normality or other common data anomalies such as outliers.

Keywords: *regression type estimator, modified maximum likelihood, robust linear regression, super-population, simulation study, non-response.*

GJSFR-F Classification : *MSC 2010: 93D21*

Strictly as per the compliance and regulations of :

A Robust Regression Type Estimator for Estimating Population Mean under Non-Normality in the Presence of Non-Response

Sanjay Kumar

Abstract- In sampling theory, regression type estimators are extensively used to estimate the population mean when the correlation between study and auxiliary variables is high. In this study, we incorporate robust modified maximum likelihood estimators (MMLEs) into regression type estimator in the presence of non-response and their properties have been obtained theoretically. For the support of the theoretical outcomes, simulations under several super-population models have been made. We study the robustness properties of these modified estimators. We show that utilization of MMLEs in estimating finite populations mean leads to robust estimates, which is very advantageous when we have non-normality or other common data anomalies such as outliers.

Keywords: *regression type estimator, modified maximum likelihood, robust linear regression, super population, simulation study, non-response.*

I. Introduction

The use of auxiliary information in sample survey have been considered mainly in the field of agricultural, biological, medical and social sciences at the stage of planning, designing, selection of units and devising the estimation procedure. In sampling theory, the ratio method of estimation uses the auxiliary information which is correlated with the study variable to improve the precision which results in improved estimators when the regression of y on x is linear and passes through origin. When the regression of y on x is linear, it is not necessary that the line should always passes through origin. Under such conditions, it is more appropriate to use the regression type estimators and the correlation between study and auxiliary variables is high.

Sometimes, it may not be possible to collect complete information for all the units selected in the sample due to non-response. Estimation of the population mean in sample in the presence of non-response has been considered by Hansen and Hurwitz (1946), Rao (1986, 1987) and several other authors.

Let $\overline{Y}$ and $\overline{X}$ be the population mean of the main study variable y and the auxiliary variable x for the population $U: (U_1, U_2, \ldots U_N)$. The population U is supposed to be composed of N_1 responding and N_2 non-responding units. From the population of size N, a sample of size n is selected by using SRSWOR method of sampling and it was observed that n_1 units respond and n_2 units don't respond. Further, by making extra

Author: *Department of Statistics, Central University of Rajasthan Bandarsindri, Kishangarh-305817, Ajmer, Rajasthan, India.*
e-mail: sanjay.kumar@curaj.ac.in

effort, a sub-sample of size $r = \frac{n_2}{K}(K > 1)$ is drawn from n_2 non-responding units by using SRSWOR method of sampling. Hence, we have n_1 units from respondent group and r units from non-respondent group of the population in the sample for which the value of the y character is obtained. Hansen and Hurwitz (1946) proposed the unibiased estimator for $\overline{Y}$, which is given as follows:

$$\overline{y}^* = \frac{n_1}{n}\overline{y}_1 + \frac{n_2}{n}\overline{y}_2', \tag{1.1}$$

where $\overline{y}_1$ and $\overline{y}_2'$ are the sample means based on n_1 and r units respectively.
The estimator $\overline{y}^*$ is unbiased and the $V(\overline{y}^*)$ is given by

$$V(\overline{y}^*) = \frac{f}{n}S_y^2 + \frac{W_2(K-1)}{n}S_{y(2)}^2 \tag{1.2}$$

where $f = \frac{N-n}{N}; W_i = \frac{N_i}{N}(i = 1,2)$, S_y^2 and $S_{y(2)}^2$ are the population mean squares of the character y for the whole population and for the non-responding part of the population.

The regression type estimator in the presence of non-response (Rao1990)when the population mean $\bar{X}$ is known, is given by

$$t_{lr} = \bar{y}^* + \hat{\theta}_L(\bar{X} - \bar{x}) , \tag{1.3}$$

where, $\hat{\theta}_L = \frac{\hat{S}_{yx}}{s_x^2}$is the regression coefficient obtained by least square estimation.$\hat{S}_{yx}$ and s_x^2(sample mean square) denote the unibiased estimates of $S_{yx}$and S_x^2 based on $n_1 + r$ observations and n observations respectively.

The bias and mean square error (MSE) of the traditional regression estimator is given by

$$B(t_{lr}) = -Cov(\bar{x}, \theta_L) \tag{1.4}$$

$$MSE(t_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 + \theta_L^2 S_x^2 - 2\theta_L S_{yx}\right) + \frac{W_2(K-1)}{n}S_{y(2)}^2, \tag{1.5}$$

where, $\theta_L = \frac{S_{yx}}{S_x^2}$, $S_x^2$is the population variance of the auxiliary variable, $S_{yx}$is the population covariance between the study variable and the auxiliary variable.

We know that the regression estimator is useful in estimating the finite population mean when the information on the auxiliary variable is available, however this is known to be quiet sensitive to outliersas studied by Farrell and Barrera(2006) and Gwet and Rivest (1992).

In sample survey studies, non-normal distributions are very common in practice as found in Cochran (1977), Jenkinset. al.(1977), Chambers (1986)and Farrell and Barrera(2007).

In this paper, we study robust modified maximum likelihood estimator (MMLE) into regression type estimator (Rao 1990)in the presence of non-response and provide their properties theoretically.

We specially focus on the situation where the error term is not normally distributed. We obtain the mean square error of the proposed regression estimator theoretically and found the conditions under which the proposed regression type estimator in the presence of non-response has less mean square error than the

corresponding regression type estimator. We support the theoretical result with simulations under several super population models and study the robustness property of the modified regression estimator. We show that utilization of MMLE for estimating finite populations mean results to robust estimate, which is very fruitful when we have non-normality or other common data anomalies such as outliers.

II. Non-Normal Errors and Proposed Regression Estimator.

Notes

For the linear regression model, $y_i = \theta x_i + e_i\ ;\ i = 1,2,\dots,n$, let the distribution of the error term follows the long tailed symmetric family.

$$f\ (e) = LTS(p,\sigma) = \frac{\Gamma p}{\sigma\sqrt{K}\,\Gamma\left(\frac{1}{2}\right)\Gamma\left(p-\frac{1}{2}\right)}\left\{1+\frac{1}{K}\left(\frac{e}{\sigma}\right)^2\right\}^{-p}\ ;-\infty < e < \infty, \tag{2.1}$$

where, $K = 2p-3, p \geq 2$ is the shape parameter ($p$is known) with $E(e_i) = 0$and $V(e_i) = \sigma^2$.

Here it can be obtained that the kurtosis of (2.1) is $\frac{\mu_4}{\mu_2^2} = 3K/(K-2)$.

The coefficients of kurtosis of the LTS family that we consider in this family are ∞, 6, 4.5, 4.0 for p= 2.5, 3.5, 4.5, 5.5 respectively.

We realize that when $p = \infty$, (2.1) reduces to a normal distribution. The likelihood equations obtained from the likelihood function of (2.1) are expressions in terms of the intractable functions.

$$g\ (z_i) = z_i/\left\{1+\frac{1}{K}(z_i^2)\right\},$$

where, $z_i = \frac{e_i}{\sigma}\ \ (i = 1,2,\dots,n)$and do not have explicit solutions.

The robust MMLE which is known to be asymptotically equivalent to the MLE are obtained in following three steps:

1. The likelihood equations are expressed in terms of the ordered variate $z_{(i)} = \frac{e_{(i)}}{\sigma}$.
2. The function $g(z_i)$ are replaced by their linear approximations and
3. The resulting equations are solved for the parameters.

The solutions which are explicit functions of the concomitant observations$(y_{[i]}, x_{[i]}), i = 1,2,\dots,n$ are

$$\hat{\theta}_T = K + L\hat{\sigma}_T,\ \ \text{and}\hat{\sigma}_T = G + \frac{\sqrt{G^2+4nC}}{2\sqrt{n(n-2)}}, \tag{2.2}$$

where, $K = \sum_{i=1}^{n}\beta_i y_{[i]}x_{[i]}\ /\ \sum_{i=1}^{n}\beta_i x_{[i]}^2$

$$L = \frac{\sum_{i=1}^{n}\alpha_i x_{[i]}}{\sum_{i=1}^{n}\beta_i x_{[i]}^2}, G = (2p/K)\sum_{i=1}^{n}\alpha_i\ (y_{[i]} - Kx_{[i]}), \tag{2.3}$$

$$C = (2p/K)\sum_{i=1}^{n}\beta_i(y_{[i]} - Kx_{[i]})^2$$

$$\alpha_i = \left(\frac{2}{K}\right)\frac{t_{[i]}^3}{\{1+(1/K)t_{[i]}^2\}^2}\text{and}\beta_i = \frac{1-(1/K)t_{[i]}^2}{\{1+(1/K)t_{[i]}^2\}^2}, \tag{2.4}$$

where, the approximate $t_{(i)}$ values are obtained from the equation

$$\int_{-\infty}^{t_{(i)}} h(z)dz = \frac{i}{n+1}\ ; 1 \le i \le n,$$

where $h(z)$ is the distribution of $z = e/\sigma$

In the same linear model, $y_i = \theta x_i + e_i\ ;\ i = 1,2,\dots,n$, now we suppose that the error term has one of the distributions in the skewed family namely, generalised logistic distribution which is given by

$$f(e) = \frac{r}{\sigma}\frac{\exp(-e/\sigma)}{\{1+\exp(-e/\sigma)\}^{r+1}}; -\infty < e < \infty, \qquad (2.5)$$

where, r is the shape parameter with $E(e_i) = \sigma\{\Psi(r) - \Psi(1)\}andV(e_i) = \sigma^2\{\Psi'(r) + \Psi'(1)\}$.

Here $\Psi(x) = \Gamma'(x)/\Gamma(x)$ is the psi function and $\Psi'(x)$ is its derivative.

For,$r < 1, r = 1,$ and$r > 1$,(2.5) represents negatively skewed, symmetric and positively skewed distribution respectively.

The coefficient of skewness and kurtosis of the generalised logistic distribution which we consider in this study are computed from the moment generating function $M_e(t) = \frac{r\,\Gamma(r+t\sigma)\Gamma(1-t\sigma)}{\Gamma(r+1)}$ and $r$is given below:

r-values	0.5	1.5	2.0	4.0	5.0
Skewness	- 0.855	0.380	0.577	0.868	0.924
Kurtosis	5.400	4.188	4.332	4.758	4.870

The likelihood equations obtained from (2.5) can be expressed in terms of the ordered variates $z_{(i)}, (i = 1,2,\dots,n)$, and in whole the intractable function $g(z_{(i)}) = \frac{1}{1+E\{z_{(i)}\}}$. These functions are linearised as we have done in the LTS family case. The solutions of the MMLE equations are the MMLEs which are given as follows:

$$\hat{\theta}_T = K - M\hat{\sigma}_T \text{ and } \hat{\sigma}_T = \frac{-D+\sqrt{D^2+4nE}}{2\sqrt{n(n-1)}}\ , \qquad (2.6)$$

where, K can be calculated from the formula (2.3) by replacing $\alpha_i, \beta_i and t_i$with

$$\alpha_i = \frac{1+e^{t(i)}+t_{(i)}e^{t(i)}}{(1+e^{t(i)})^2},\ \beta_i = \frac{e^{t(i)}}{(1+e^{t(i)})^2}\ , \qquad (2.7)$$

and $t_{(i)} = -\log(q_i^{-\frac{1}{r}} - 1), q_i = \frac{1}{n+1}$ respectively.

In equation (2.6) for calculating $\hat{\theta}_T$ and $\hat{\sigma}_T$, M, D and Evalues are calculated from the following equations

$$M = \frac{\sum_{i=1}^{n}\Delta_i x_{[i]}}{\sum_{i=1}^{n} p_i x_{[i]}^2},\ D = (r+1)\sum_{i=1}^{n}\Delta_i(y_{[i]} - Kx_{[i]}) \qquad (2.8)$$

and

$$E = (r+1)\sum_{i=1}^{n}\beta_i\left(y_{[i]} - Kx_{[i]}\right)^2, \qquad (2.9)$$

where $\Delta_i = \alpha_i - (r+1)^{-1}\ for\ 1 \le i \le n$

Islam et. al. (2001) showed that the MMLEs given in (2.6) are more efficient and robust then their corresponding least square estimators (LSEs) when the error term is from the skewed family (2.5).

In this study we calculate the MMLE $\hat{\theta}_T$ from (2.2) if the error term is from $LTS(p,\sigma)$or from (2.6) if the error is from $GL(r,\sigma)$and modify the traditional regression estimator in the presence of non-response as given in (1.1) to achieve efficient estimator under non-normality, which is given by

Notes

$$T_{lr} = \bar{y}^* + \hat{\theta}_T(\bar{X} - \bar{x}) \ , \tag{2.10}$$

The bias and mean square error of the proposed estimator (2.10) are given by

$$B(T_{lr}) = -Cov(\bar{x}, \theta_T) \tag{2.11}$$

$$MSE(T_{lr}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left(S_y^2 + {\theta_T}^2 S_x^2 - 2\theta_T S_{xy}\right) + \frac{W_2(K-1)}{n} S_{y(2)}^2 \tag{2.12}$$

In order to compare the MSE of the proposed estimator in (2.10) with the MSE of the regression type estimator $t_{lr} = \bar{y}^* + \hat{\theta}_L(\bar{X} - \bar{x})$, we get the following conditions under which the proposed estimator is more efficient than the regression type estimatort_{lr}.

$$MSE(T_{lr}) \leq MSE(t_{lr}) \text{if} \theta_T \leq \theta_L \text{ or } \theta_T \geq \theta_L \tag{2.13}$$

NOTE:- In general the shape parameter p in (2.1) and r in (2.5) may not be known. Using the least square estimator and constructing q-q plots (with the observed values as in Hamilton(1992), one can easily determine the closest distribution for the error term. Since the families (2.1) and (2.5) include a very large variety of location scale distribution, one can easily determine an approximate distribution for the error by using one of the two families given in study.

III. Simulation Study

In this study for the simulation, we have used R-programming software. In the super population models generated, we use the model

$$y_i = \theta x_i + e_i, i = 1,2, \dots, N, \tag{3.1}$$

where, we generate e_i and x_i independently and calculate y_i for $i = 1,2, \dots, N$.

Let the errors $e_1, e_2, \dots , e_N$ be the random observations from a super population either from (2.1) or (2.5). Let U_N denotes the corresponding finite population consists of Npairs (x_1, y_1), (x_2, y_2), ... , (x_N, y_N). To calculate the MSE of the proposed estimator in (2.10), we calculate T_{lr} for all possible samples $\binom{N}{n}$ simple random samples of size $n$from U_N. Since $\binom{N}{n}$is extremely large, so we conduct all Monte-Carlo studies as follows.

We take $N = 500$ in each simulation and generate U_{500} pairs from an assumed super population. From the generated finite population U_{500}, we have selected a sample of size $(n = 14, 19, 26, 40, 70)$ by simple random sampling without replacement. From each selected sample, the last 43% (3, 4, 6, 8, 11 respectively) of units have been considered as non-responding units. Now, we choose at random S= 15000 samples for all the possible $\binom{500}{n}$ samples of size n $(n = 14, 19, 26, 40, 70)$, which gives 15000 values

of T_{lr}. To compare the efficiency of the proposed estimator under different models for a given n, we calculate the values of mean square errors as follows:

$MSE(T_{lr}) = \frac{1}{S}\sum_{j=1}^{S}(T_{lrj} - \bar{Y})^2$, $MSE(t_{lr}) = \frac{1}{S}\sum_{j=1}^{S}(t_{lrj} - \bar{Y})^2$ and

$$MSE(\bar{y}^*) = \frac{1}{S}\sum_{j=1}^{S}(\bar{y}^* - \bar{Y})^2,$$

where, $\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$

For setting the population correlation ρ_{yx} is sufficiently high, which choose the value of parameter θ in the model $= \theta x + e$, such that the correlation coefficient between study variable (y) and auxiliary variable (x) is ρ_{yx} to determine the value of θ that satisfied this condition, we follow a similar way given by Rao and Beegle (1967) and write the population correlation between the study variable (y) and the auxiliary variable (x) .For example if $X \sim U(0,1)$, the value of θ for which the population correlation between y and x becomes $\theta^2 = \frac{12\sigma^2\rho_{yx}^2}{1-\rho_{yx}^2}$ for the LTS family and $\theta^2 = \frac{12\sigma^2(\varphi'(r)+\varphi(1))\rho_{yx}^2}{1-\rho_{yx}^2}$ for the skewed family. Similarly, if $x$is generated from $Exp(1)$, the value of θ for which the population correlation becomes $\theta^2 = \frac{\sigma^2\rho_{yx}^2}{1-\rho_{yx}^2}$ for the symmetric family and $\theta^2 = \frac{\sigma^2(\varphi'(r)+\varphi(1))\rho_{yx}^2}{1-\rho_{yx}^2}$ for the skewed family.

Here we take $\sigma^2 = 1$, in all situations without loss of generality and calculate the require parameter θ for which $\rho_{yx} = 0.75$.

IV. Comparison of Efficiencies of the Proposed Estimator

We consider four different super-population models given below to see how much efficiency we gain with the proposed modified estimator, when the condition (2.13) is satisfied under
non-normality:

I. $x \sim U(0,1)$ and $e \sim LTS(p,1)$ and independent of x.
II. $x \sim exp(1)$ and $e \sim LTS(p,1)$ and independent of x.
III. $x \sim U(0,1)$ and $e \sim GL(r,1)$ and independent of x.
IV. $x \sim exp(0.5)$ and $e \sim GL(r,1)$ and independent of x.

For the models (1) to (4), the values of θwhich makes the population correlation $\rho_{yx} = 0.75$ are given in table 1.

Table 1 : Parameter values of θused in models (1)–(4) that give $\rho_{yx} = 0.75$

Population	p		
	2.5	4.5	5.5
Model (1)	3.928	3.928	3.928
Model (2)	1.133	1.133	1.133
Population	r		
	0.5	1.5	2.0
Model (3)	10.076	6.309	5.944
Model (4)	2.057	1.288	1.213

Here, we note that for the LTS family (2.1), the value of θ does not depend on the shape parameter$p$.

To verify that the super-population are generated appropriately, we provide a scatter graph and error distribution of model to $p = 4.5$ for model (2) in the figure 1 and in the figure 2. Similarly for the model (3) with $r = 0.5$ in the figure 3 and in the figure 4, a scatter graph and error distribution is provided.

Notes

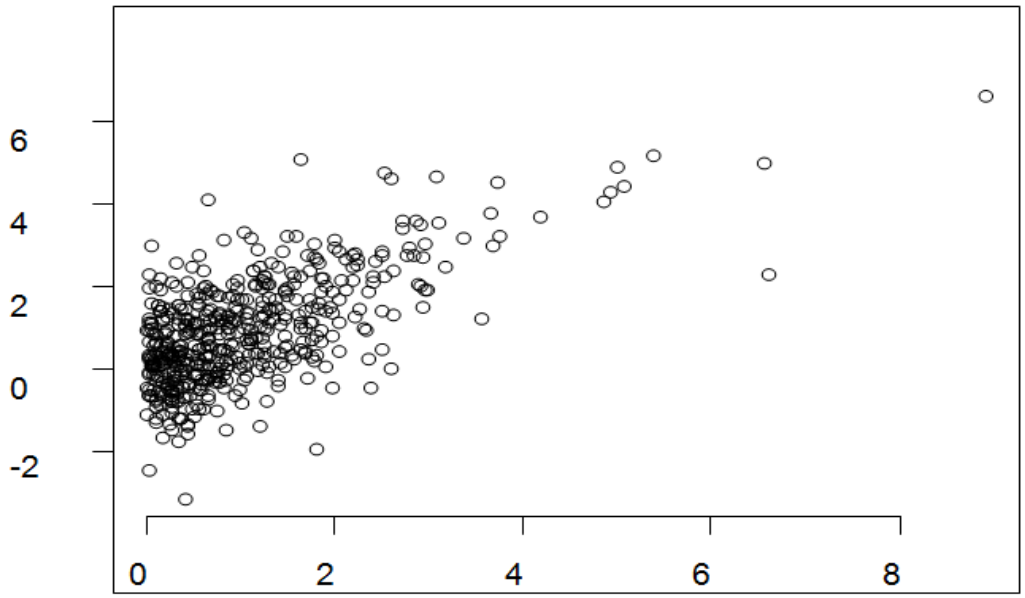

Figure 1 : A scatter graph of the study variable and auxiliary variable obtained from model (2) for $p = 4.5$.

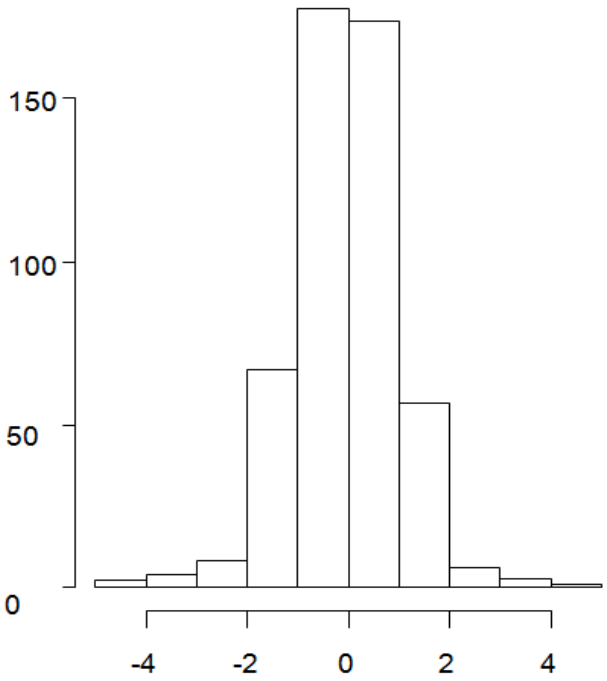

Figure 2 : Generated error distribution obtained from model (2) for $p = 4.5$

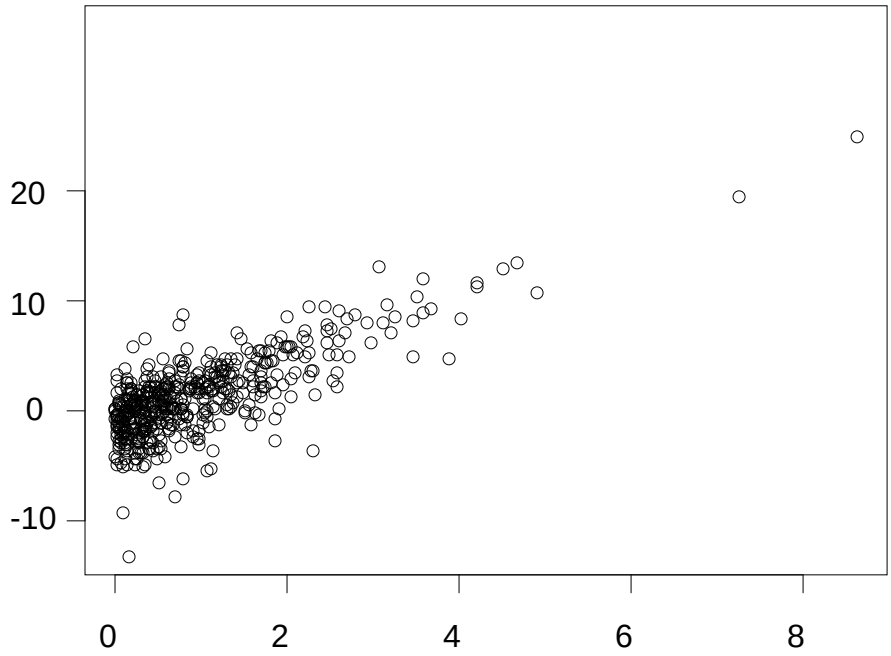

Figure 3 : A scatter graph of the study variable and auxiliary variable obtained from model (3) for $r = 0.5$.

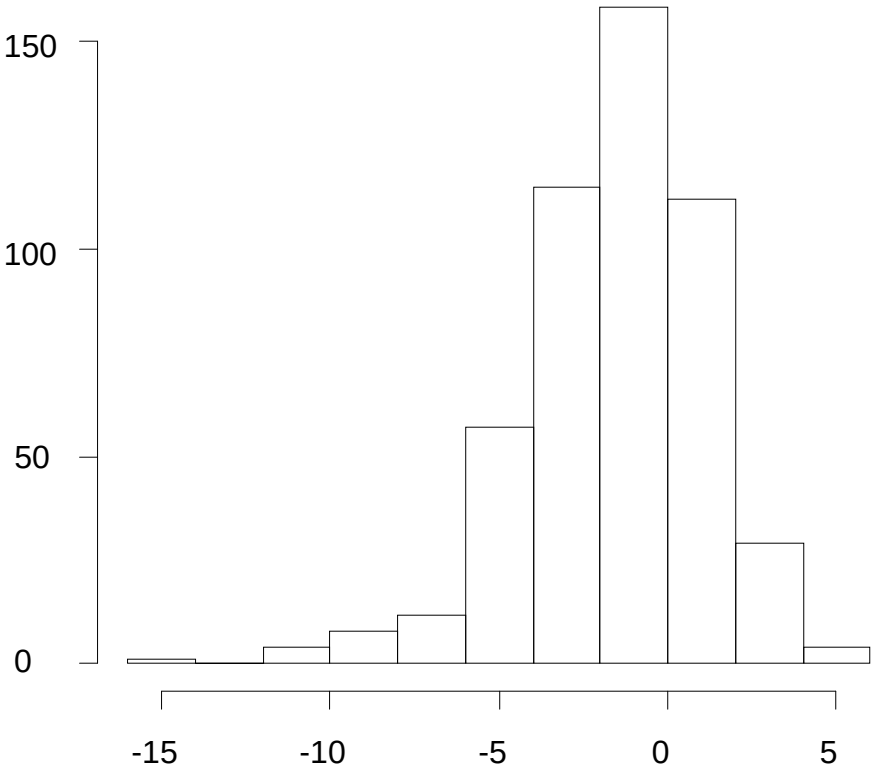

Figure 4 : Generated error distribution obtained from model (3) for $r = 0.5$.

Relative efficiencies are calculated as $RE = \frac{MSE(\bar{y}^*)}{MSE(.)} * 100$,

where, MSE(.)and relative efficiency (RE) are given in the table 2 for the model (1) and (2) and in the table 3 for the model (3) to and (4).

Notes

From the table 2, we see that the proposed estimator T_{lr} is more efficient than the regression estimator t_{lr} in the presence of non-response because the theoretical condition is satisfied. We also observe that when sample size increases, mean square error decreases.

From the table 3, we also observe that the proposed estimator T_{lr} in the presence of non-response is more efficient than the regression estimator t_{lr} because the theoretical condition is satisfied. It is also clear that when sample size increases, mean square error decreases.

Notes

Table 2 : Mean square error and efficiencies of the estimators under super-populations (1-2).

		$\boldsymbol{Model(1): x\sim U(0,1) and\ e\sim LTS(p,1)}$				
	Est.	$\boldsymbol{n}$				
		11	**15**	**21**	**31**	**51**
$p=2.5$	$\bar{y}^*$	0.1999 (100.00)*	0.1550 (100.00)	0.1339 (100.00)	0.0898 (100.00)	0.0518 (100.00)
	t_{lr}	0.1220 (163.85)	0.1041 (148.90)	0.0881 (151.99)	0.0626 (143.45)	0.0346 (149.71)
	T_{lr}	0.1177 (169.84)	0.1019 (152.11)	0.0862 (155.34)	0.0620 (144.84)	0.0345 (150.15)
$p=4.5$	$\bar{y}^*$	0.2212 (100.00)	0.1704 (100.00)	0.1208 (100.00)	0.0781 (100.00)	0.0541 (100.00)
	t_{lr}	0.1398 (158.79)	0.1081 (157.63)	0.0723 (167.08)	151.95 (151.95)	0.0382 (141.62)
	T_{lr}	0.1343 (164.71)	0.1058 (161.06)	0.0714 (169319)	0.0510 (153.14)	0.0381 (142.00)
$p=5.5$	$\bar{y}^*$	0.2598 (100.00)	0.1690 (100.00)	0.1146 (100.00)	0.0823 (100.00)	0.0533 (100.00)
	t_{lr}	0.1776 (146.28)	0.1077 (156.92)	0.0677 (169.28)	0.0549 (149.91)	0.0374 (142.51)
	T_{lr}	0.1723 (150.78)	0.1054 (160.34)	0.0667 (171.81)	0.0544 (151.29)	0.0372 (143.28)
		$Model(2): x\sim exp(1) and\ e\sim LTS(p,1)$				
	Est.	n				
		11	15	21	31	51
$p=2.5$	$\bar{y}^*$	0.2211 (100.00)*	0.1811 (100.00)	0.1219 (100.00)	0.0639 (100.00)	0.0496 (100.00)
	t_{lr}	0.1630 (135.64)	0.1192 (151.93)	0.0719 (169.54)	0.0424 (150.71)	0.0329 (150.76)
	T_{lr}	0.1583 (139.67)	0.1184 (152.96)	0.0718 (169.78)	0.0420 (152.14)	0.0327 (151.68)
$p=4.5$	$\bar{y}^*$	0.1856 (100.00)	0.1719 (100.00)	0.1336 (100.00)	0.0823 (100.00)	0.0584 (100.00)
	t_{lr}	0.1346 (137.89)	0.1141 (150.66)	0.0803 (166.38)	0.0499 (164.93)	0.0430 (135.81)
	T_{lr}	0.1320 (140.61)	0.1130 (152.12)	0.0797 (167.63)	0.0498 (165.26)	0.0429 (136.13)
$p=5.5$	$\bar{y}^*$	0.2419 (100.00)	0.1747 (100.00)	0.1125 (100.00)	0.0808 (100.00)	0.0479 (100.00)
	t_{lr}	0.1697 (142.55)	0.1166 (149.83)	0.0720 (156.25)	0.0501 (161.28)	0.0351 (136.47)
	T_{lr}	0.1664 (145.37)	0.1161 (150.47)	0.0716 (157.12)	0.0500 (161.60)	0.0349 (137.25)

*(*Efficiencies are in the parenthesis)*

Table 3 : Mean square error and efficiencies of the estimators under super-populations (3-4).

		$\boldsymbol{Model(3): x\sim U(0,1) and\ e\sim GL(r,1)}$				
$r=0.5$	**Est.**	$\boldsymbol{n}$				
		11	**15**	**21**	**31**	**51**
	$\bar{y}^*$	1.5083 (100.00)*	1.0495 (100.00)	0.6897 (100.00)	0.5187 (100.00)	0.3088 (100.00)

	t_{lr}	0.9555 (157.85)	0.6606 (158.87)	0.4183 (164.88)	0.3257 (159.26)	0.2246 (137.49)
	T_{lr}	0.9457 (159.49)	0.6603 (158.94)	0.4161 (165.75)	0.3252 (159.50)	0.2242 (137.73)
$r = 1.5$	$\bar{y}^*$	0.5654 (100.00)	0.3904 (100.00)	0.2970 (100.00)	0.2071 (100.00)	0.1505 (100.00)
	t_{lr}	0.3881 (145.68)	0.2668 (146.33)	0.1655 (179.46)	0.1308 (158.33)	0.0988 (152.33)
	T_{lr}	0.3758 (150.45)	0.2630 (148.44)	0.1605 (185.05)	0.1293 (160.17)	0.0983 (153.10)
$r = 2.0$	$\bar{y}^*$	0.5506 (100.00)	0.4200 (100.00)	0.2795 (100.00)	0.1734 (100.00)	0.1228 (100.00)
	t_{lr}	0.3337 (165.00)	0.2544 (165.09)	0.1694 (164.99)	0.1032 (168.02)	0.0835 (147.07)
	T_{lr}	0.3155 (174.52)	0.2458 (170.87)	0.1652 (169.19)	0.1028 (168.68)	0.0830 (147.95)
	Model(4): $x \sim exp(0.5)$ *and* $e \sim GL(r,1)$					
	Est.	n				
		11	15	21	31	51
$r = 0.5$	$\bar{y}^*$	2.2570 (100.00)	1.6369 (100.00)	1.0539 (100.00)	0.6654 (100.00)	0.4526 (100.00)
	t_{lr}	1.1878 (190.02)	0.8950 (182.89)	0.6039 (174.52)	0.4048 (164.38)	0.3119 (145.11)
	T_{lr}	1.1738 (192.28)	0.8803 (185.95)	0.5971 (176.50)	0.3982 (167.10)	0.3117 (145.20)
$r = 1.5$	$\bar{y}^*$	0.9732 (100.00)	0.7423 (100.00)	0.4968 (100.00)	0.3323 (100.00)	0.2227 (100.00)
	t_{lr}	0.5078 (191.65)	0.4077 (182.07)	0.2485 (199.92)	0.1788 (185.85)	0.1271 (175.22)
	T_{lr}	0.4963 (196.09)	0.4038 (183.83)	0.2463 (201.71)	0.1770 (187.74)	0.1270 (175.35)
$r = 2.0$	$\bar{y}^*$	0.7822 (100.00)	0.6629 (100.00)	0.3742 (100.00)	0.2942 (100.00)	0.2207 (100.00)
	t_{lr}	0.4033 (193.95)	0.3245 (204.28)	0.1720 (217.59)	0.1590 (185.03)	0.1280 (172.42)
	T_{lr}	0.3899 (200.62)	0.3240 (204.60)	0.1719 (217.69)	0.1573 (187.03)	0.1272 (173.51)

*(*Efficiencies are in the parenthesis)*

V. Robustness of the Proposed Estimator

The outliers in sample data are normally a in centered problem for survey statistician [1]. In practice, the shape parameters p in $LTS(p,\sigma)$, and r in $GL(r,\sigma)$ might be mis-specified. Therefore, it is very important for estimators to have efficiencies of robustness estimates such as an estimator is full efficient or nearly so for an assumed model and maintains high efficiencies for plausible to the assumed model.

Here, we take $N = 500$ and $\sigma^2 = 1$ without loss of generality and we study the robustness property of proposed estimator under different outlier models as follows.

We assume $x \sim U(0,1)$ and the error term $e \sim LTS(p = 3.5, \sigma^2 = 1)$. We determine our super-population model as follow:

(5). True model: $LTS(p = 3.5, \sigma^2 = 1)$

(6). Dixon's outliers model: $N - N_o$ observations from $LTS(3.5,1)$and N_o (we don't know which) form $LTS(3.5,2.0)$

(7). Mis-specified model: $LTS(4.0,1)$

Here, we realize that the model (5), the assumed super population model is given for the purpose of comparison and the models (6) and (7) are taken as its plausible alternatives. Here we have assumed the super population model $LTS(3.5,1)$for estimating θ_T. The coefficients (α_i, β_i) from (2.4) are calculated with $p = 3.5$ and are used in models (5) and (6). $N_o$in model (6) is calculated from the formula$(|0.5 + 0.1 *$

Ref

1. Chambers, R. L. (1986), Outlier robust finite population estimation. Journal of the American Statistical Association, 81, 1063–1069.

$N| = 50)$ for $N = 500$.We standardised the generated $e_i's, (i = 1,2, \dots, N)$ in all the models to have the same variance as that of $LTS(3.5, 1)$ i.e. it should be equal to 1. The simulated values of MSE and the relative efficiency are given in table 4. Here the estimators $\tilde{\theta}_L, \hat{\theta}_T$ are both location invariant estimators so that both of them are the estimators of θ under all the models described above. Here theoretical condition (2.13) is satisfied for model (5).

Notes

From the table 4, we see that in the presence of non-response, the proposed estimator T_{lr} is more efficient than the regression estimator t_{lr} because the theoretical condition is satisfied. We also observe that when sample size increases, mean square error decreases.

Now we assumed that the error term e is from the skewed family and $x$is from $U(0,1)$. We assumed the model to be $GL(3,1)$ and determine our super population as

(8) True model: $GL(3,1)$

(9) Dixon outlier model: $N - N_o$observations from $GL(3,1)$ and N_o (we don't know which) from $GL(3, 2)$, where $(N_o = |0.5 + 0.1 * N| = 50)$.

(10) Mis-specified model: $GL(5,1)$.

The model (8) is assumed as a super population model and all other models (9) and(10) are taken as its plausible alternatives. The generated $e_i's, (i = 1,2, \dots, N)$ were standardized in the models (9) and (10) to have the same variance as that of $GL(3,1)$ i.e. $V(e_i) = \{\Psi'(4) + \Psi'(1)\}$,where $\Psi'(x)$ is the derivative of the psi function. The simulated values of the MSEs and relative efficiencies of the estimators and relative efficiency under models (8) to(10) are given in the table 4. Also, from the table 4, we see that the proposed estimator T_{lr} is more efficient than the regression estimator t_{lr} since the theoretical condition is satisfied. Also, it is clear that when sample size increases, mean square error decreases.

Table 4 : Mean square errors and efficiencies under super-populations (5)–(7) for LTS family and under super-populations (8)–(10) for skewed family

Est.	n			n		
	11	15	21	5	11	15
	$True\ Model(5)$: $x \sim Uni(0,1)$ and $e \sim LTS(3.5,1)$			$Dixon\ outlier\ Model(6)$: $x \sim Uni(0,1)$ and $(N - N_0)\ e \sim LTS(3.5,1 + N_0 e \sim LTS(3.5,2)$		
$\bar{y}^*$	0.2189 (100.00)*	0.1742 (100.00)	0.1157 (100.00)*	0.2385 (100.00)	0.1684 (100.00)	0.1228 (100.00)
t_{lr}	0.1627 (134.54)	0.1081 (161.15)	0.0695 (166.48)	0.1521 (156.81)	0.1061 (158.72)	0.0629 (195.23)
T_{lr}	0.1609 (136.05)	0.1074 (162.20)	0.0691 (167.44)	0.1438 (165.86)	0.1048 (160.69)	0.0587 (209.20)
	$Mis - specified\ Model(7)$: $x \sim Uni(0,1)$ and $e \sim LTS(4.0,1)$			$True\ Model(8)$: $x \sim Uni(0,1)$ and $e \sim GL(3.0,1)$		
$\bar{y}^*$	0.7036 (100.00)	0.5112 (100.00)	0.2919 (100.00)	0.4989 (100.00)	0.3117 (100.00)	0.2416 (100.00)
t_{lr}	0.3866 (182.00)	0.2594 (197.07)	0.1653 (176.59)	0.2975 (167.70)	0.2167 (143.84)	0.1429 (169.07)
T_{lr}	0.3798 (185.26)	0.2566 (199.22)	0.1632 (178.86)	0.2878 (173.35)	0.2138 (145.79)	0.1416 (170.62)
	$Dixon\ outlier\ Model(9)$: $x \sim Uni(0,1)$ and $(N - N_0)\ e \sim GL(3.0,1 + N_0 e \sim GL(3.0,2)$			$Mis - specified\ Model(10)$: $x \sim Uni(0,1)$ and $e \sim GL(5.0,1$		
$\bar{y}^*$	0.6942 (100.00)	0.8314 (100.00)	0.7949 (100.00)	0.4725 (100.00)	0.2977 (100.00)	0.2232 (100.00)
t_{lr}	0.4563	0.6873	0.5410	0.3069	0.1970	0.1324

	(152.14)	(120.97)	(146.93)	(153.96)	(151.12)	(168.58)
T_{lr}	0.4373 (158.75)	0.6767 (122.86)	0.5302 (149.93)	0.3014 (156.77)	0.1949 (152.75)	0.1300 (171.69)

*(*Efficiencies are in the parenthesis)*

VI. Determination of Shape Parameter

In order to determine whether when a particular density is appropriate for the error term, a Q-Q plot of the ordered estimated residuals which are calculated using the LSEs (Least Square Estimation), $e_i = y_i + \tilde{\theta}_L x_i, (i = 1,2,\dots,n)$ are plotted against population quantiles for that density.

The population quantiles t_i are determined from the equation $\int_{-\infty}^{t_{(i)}} t(u)du = \frac{i}{n+1}$; $1 \le i \le n$, where n is the sample size.

The Q-Q plot that closely approximates a straight line would be assumed to be the most appropriate.

VII. Conclusion

In this study, we show that when the error term is not normal which is applicable in most areas, MML integrated regression estimator(T_{lr})in the presence of non-response can improve the efficiency of regression estimatort_{lr}. We also show that the MML integrated regression estimator (T_{lr})(modified regression estimators) is robust to outliers as well as other data anomalies.

References Références Referencias

1. Chambers, R. L. (1986), Outlier robust finite population estimation. Journal of the American Statistical Association, 81, 1063–1069.
2. Cochran, W.G. (1977). Sampling Techniques. John Wiley & Sons, New York.
3. Farrell, P. J., & Barrera, M. S. (2006). A comparison of several robust estimators for a finite population mean. Journal of Statistical Studies, 26, 29–43.
4. Farrell, P.J., Barrera, S.M. (2007). A comparison of several robust estimators for a finite population mean. J. Stat. Stud. 26, 29–43.
5. Gwet, J. P., & Rivest, L. P. (1992). Outlier resistant alternatives to the ratio estimator. Journal of the American Statistical Association, 87, 1174–1182.
6. Hamilton, L. C. (1992). Regression with graphics. California: Brooks, Cole Publishing Company.
7. Hansen, M. H. and Hurwitz, W. N. (1946). "The problem of nonresponse in sample surveys", J. Amer. Stat. Assoc.,41, 517-529.
8. Islam, M. Q., Tiku, M.L. and Yildirim, F. (2001). Non-normal regression I skew distributions. Commun. in Statist.: Theo. and Meth., 30, 993-1020.
9. Jenkins, O. C., Ringer, L.J., Hartley, H.O. (1977). Root estimators. Journal of the American Statistical Association, 68, 414–419.
10. Rao, J. N. K., & Beegle, L. D. (1967). A Monte Carlo study of some ratio estimators. Sankhyfia Ser. B, 47–56.
11. Rao, P. S. R. S. (1986). "Ratio estimation with subsampling the Non-respondents", Survey Methodology, 12(2), 217-230.
12. Rao, P. S. R. S. (1987). "Ratio and regression estimates with subsampling the non-respondents", Paper presented at a special contributed session of the International Statistical Association Meetings, Sept,, 2-16, 1987,Tokyo, Japan.

Notes

13. Rao, P. S. R. S. (1990). Regression estimators with subsampling of non-respondents. Data Quality Control. Theory and Pragmatics (Gunar e. Liepins and V.R.R. Uppuluri, Eds. Marcel Dekker, New York, 191-208.

Notes

This page is intentionally left blank