Jour. Ind. Soc. Ag. Statistics
53(1), 2000 : 20–34

Outlier Robust Finite Population Estimation under a Linear Regression Model

A.K. Srivastava and V. Geethalakshmi[1]
IASRI, New Delhi-110012
(Received : May, 1999)

SUMMARY

When outliers appear in survey data, the use of conventional estimators to estimate the population mean may not be appropriate. For instance, the simple mean as an estimator of the population mean may give a distorted picture of the finite population under consideration as equal weights will be given to all sampling units resulting in over-estimation or under-estimation thus affecting the subsequent inference. Moreover, when an underlying model is assumed the estimation procedure may further get affected by possible violations of the model assumptions. Some estimators have been proposed to take the twin problems of outliers and model violations. Also a comparative empirical study of the proposed estimators and some standard outlier-robust estimators was carried out through simulation. It was observed that the proposed estimators perform fairly well in a scenario where the underlying model assumptions are not satisfied and the finite population under consideration is outlier-prone.

Key words: Outliers, Robust estimation, Regression, Sample surveys.

1. Introduction

Sample surveys are often conducted with aim of estimating the finite population total or mean. The very nature of the finite population mean to take any desired value by changing just a single observation makes it sensitive especially when outliers are present in the data. The usual estimation procedure in such situations may lead to a very distorted picture of the finite population under consideration. Although much has been written about outliers in the past years, some amount of subjectivity appears in the available definitions of outliers in the literature. Because of the emphasis on modelling in recent years, "outlier" now seems to indicate any observation that does not come from the target population, as is also viewed by many authors. Throughout this article "outliers" are discussed in accordance with this visualisation.

1 CSWRI, Avikanagar-304501

Chambers [2] classifies survey sample outliers as representative and nonrepresentative. The former type of outlier is a sample unit which is correctly measured and can not be assumed to be unique. The unsampled part of the population may contain similar units which are markedly different in value from other sampled units. The latter type of outlier is a sample unit which is either unique *i.e.* has a value characteristic to that particular unit or which is incorrectly recorded. In this article, only representative outliers are considered and some estimators of the population total have been proposed which prove to lower the impact of these outliers under a linear regression model.

2. *Outliers in Sample Survey Data and Their Treatment*

When the sample is from a population known to contain outliers, one or more large units may appear in it and when they are assigned small selection probabilities they receive large weights and have strong effects on the sample mean and its variance. This point is highlighted in the famous Basu's elephant example (Basu [1]) of circus elephants which constitutes a population of size 50 with a single large unit. It emphasises the inappropriateness of the Horvitz-Thompson estimator when large units are present in the population.

Outliers may crop up into the data at any time of the survey. The sources and detection coupled with the removal of the outliers can be explained algorithmically (see Fig. 1). Let the finite population, of which we wish to infer, contain N units u_1, u_2..., u_N denoted by the vector $\mathbf{U}_0$ whose values are based on the characteristic of interest Y. This population, popularly known as the 'target population', may be notated as $(\mathbf{U}_0, \mathbf{Y}_0)$.

The general scheme of sampling is to draw a sample from the target population and infer about the latter using the former, with the help of tools like sampling frame, statistical methods, etc.

This paves way for multistage contamination. The sampling frame not based on the target population may include or exclude units, resulting in a new unit vector $\mathbf{U}_1$. The information pertaining to the characteristic of interest may get distorted resulting in a new observation vector $\mathbf{Y}_1$, at the survey population level.

In the next step, instead of the ideal sample, $s_0 \subset \mathbf{U}_1$, a new sample s_1 may be got owing to non-response compulsions. While recording the actual data from the survey sample, aberrations such as coding errors may occur resulting in $\mathbf{Y}_2$ instead of $\mathbf{Y}_1$. These erroneous observations get detected while the processing of data is done, giving a clean data $(s_2, \mathbf{Y}_3)$ from which inference

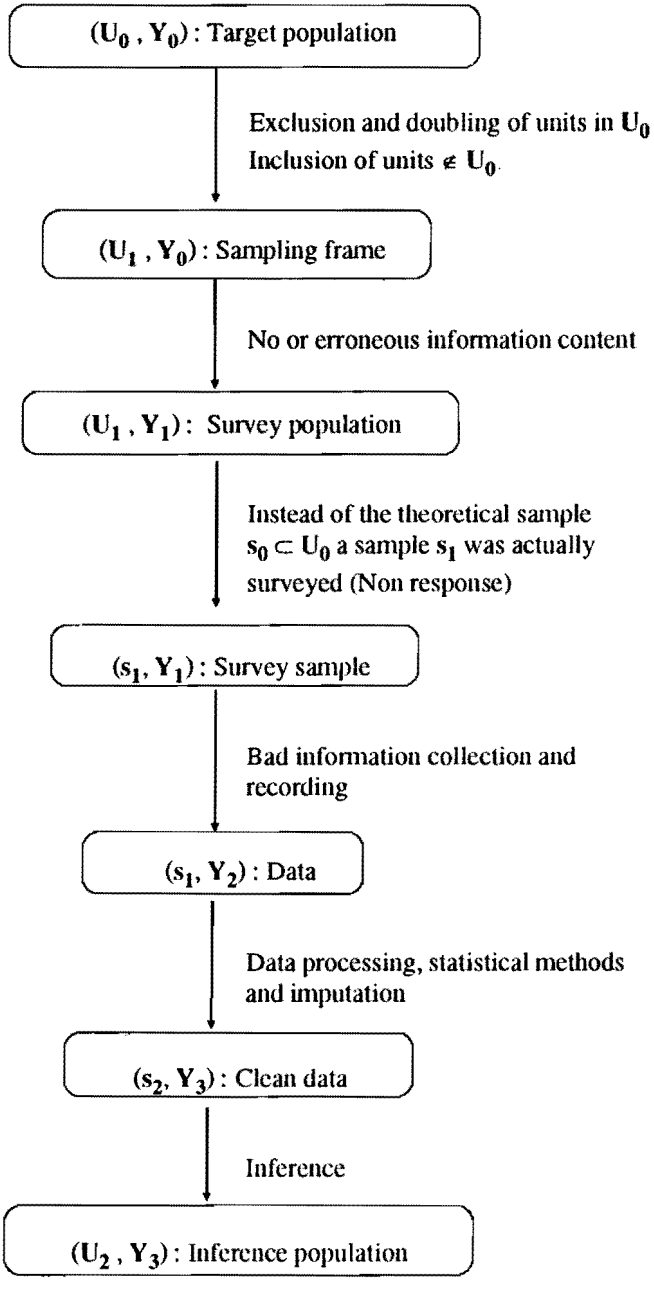

Fig 1. Emergence of outliers in survey data (Source: Hulliger [7])

has to be made which gives the inference population, (U_2, Y_3). A sound estimation procedure is one which will ensure that the inference population is closer to the target population.

In survey sampling literature, the problem of outliers or aberrant values has often been treated under the heading "skewed populations". Judicious selection of the sampling scheme may give some sort of control over the skewedness of the population. If the prior information about large values is available to the sampler, then an obvious strategy is to stratify the population, trying to put all the large values in the population in a separate stratum and enumerate it completely, thus effectively removing them from the population for the purpose of statistical inference (Glasser [4]).

If the prior information about the outliers is not perfect, or is totally absent, the previous procedure will fail to take off. Kish [8] suggests the construction of a stratum of surprises where all the suspected outliers are placed. The problem is then to estimate the size of the stratum. As a convention, over all these years the sample mean $\bar{y}$ is used as an estimator of the population mean $\bar{Y}$ unless precise form of the distribution of the study variable Y is known. When outliers appear in the data, the experimenter is concerned over the fact that one or more observations are unduly influencing the estimate of the mean. Insisting on a close estimate of $\bar{Y}$, usually he suggests that the offending observations be discarded. Thus the practice of "discarding" outliers from the sample and replacing the offending observations by more reasonable values took shape. Trimming, Winsorization procedures, utilising censored samples are a few procedures to name, which follow this practice. Hidiroglou and Srinath [6] have developed estimators which use corrected weights to outlying units in order to deflate their values at the estimation stage once they have been sampled and identified as unusually large units.

Smith [10] has given an estimator of the population total using inclusion probabilities based on post-stratification. Chambers [2] developed an outlier-robustification of the prediction approach using M-estimation assuming a regression through the origin model with known variance σ_i which is a function of the auxiliary variable X. Let

$$\xi_0: \text{the random variables } r_i = (y_i - \beta x_i)\,\sigma_i^{-1} \text{ are identically and independently distributed with mean 0 and variance 1} \qquad (2.1)$$

where $\sigma_i^2 = \sigma^2 V(x_i)$ and β, σ are generally unknown parameters. Under ξ_0, an outlier-robust estimator of the population total, using an outlier robust estimator b_n of β and ψ function which downweights outliers, is given by

$$T_n = \sum_1 y_i + b_n \sum_2 x_i + \sum_1 u_i \psi\{(y_i - b_n x_i)/\sigma_i\} \qquad (2.2)$$

where $\sum_1$ and $\sum_2$ denote the summations over the sampled and non-sampled parts of the population respectively and

$$u_i = \frac{x_i}{\sigma_i} \sum_2 x_j \left(\sum \frac{x_j^2}{\sigma_j^2} \right)^{-1}$$

He compares various finite population strategies using T_n and those using conventional estimators through simulation studies done on a real population. Four versions of the ψ-function used in this study where

(i) $\psi(t) = 0$ (outlier rejection)

(ii) $\psi(t) = t$ (ξ_0 BLUE under $V(x) = x^2$)

(iii) $\psi(t) = te^{-a/2(|t|-b)^2}$ with $a = 0.5$, $b = 6$

and

(iv) $\psi(t) = te^{-a/2(|t|-b)^2}$ where a, b are estimated from the sample data by minimising

$$\int_0^\infty \psi^2(t,a,b) e^{-t^2/2} dt$$

subject to a bound on the absolute value of an estimate of the asymptotic bias of T_n under the outlier-prone alternative ξ_M. All strategies based on the estimator T_n, with bounded ψ, perform extremely well with respect to root mean square error.

3. Robust Estimation under a Linear Regression Model

Often outliers can be thought of as having a bivariate (Gwet and Rivest [5]) or even a multivariate dimension, i.e., a sampling unit may be influential with respect to an auxiliary variable X or with respect to a set of auxiliary variables X_1, X_2,... X_p. Hence model-based approach to outlier robust finite population estimation gains more weight at the outset.

The issue of robustness also arises when model-dependent methods are used for both sample selection and estimation. Consider a simple situation when a design variable Z is known at the design stage for each member of the finite population of size N. The regression approach makes use of an auxiliary variable

X and observations are made upon this variable along with the study variable Y for the n units selected in the sample. While formulating a model one can either include or exclude the design variable Z in the regression model. In a large scale survey, Z may be used for stratification but not for estimation purpose because the values of this variable may not be available at the unit level for its inclusion in the model. In such situations, the usual OLS estimator will underestimate the parameter β_{yx} due to the use of this model. This portrays a 'mis-specification' of the model under specific situations and it is essential that a robust estimation procedure be adopted for a reliable estimator.

Considering the twin problems of outlier-robust finite population estimation and robust estimation under a classical linear regression model, it is worthwhile to think of an approach which will provide estimators that are robust against both outliers and model violations. In this article some estimators of the finite population have been proposed which are both outlier-robust and robust against model violations. They have been constructed on the basis of model-free estimators of the parameter β which are due to Nathan and Holt [9].

3.1 Weighted Estimators due to Nathan and Holt [9]

Let there be a finite population of size N arisen from a super - population such that the observed values of the design variable Z are identically and independently distributed with mean μ_3 and variance σ_3^2. Let $\hat{\mu}_3 = \frac{1}{N}\sum_{\alpha=1}^{N} z_\alpha$ and $\sigma_3^2 = \frac{1}{N-1}\sum (z_\alpha - \hat{\mu}_3)^2$. And let X denote the auxiliary variable and Y the study variable used in the survey. The maximum likelihood estimator under a trivariate normal distribution for (X, Y, Z) proposed by Demets and Halperin [3] which is a asymptotically unbiased estimator of β_{12}, the regression coefficient of Y on X is given by

$$\hat{\beta}_{12} = \frac{s_{12} + \frac{s_{13}s_{23}}{s_3^2}\left(\frac{\hat{\sigma}_3^2}{s_3^2} - 1\right)}{s_2^2 + \frac{s_{23}^2}{s_3^2}\left(\frac{\hat{\sigma}_3^2}{s_3^2} - 1\right)} \qquad (3.1.1)$$

where

$$s_{12} = \frac{1}{N-1}\sum_i \sum_j (x_i - \bar{x})(y_j - \bar{y})$$

$$s_{13} = \frac{1}{N-1}\sum_i\sum_j (z_i - \bar{z})(y_j - \bar{y})$$

$$s_{23} = \frac{1}{N-1}\sum_i\sum_j (z_i - \bar{z})(x_j - \bar{x})$$

$$s_2^2 = \frac{1}{N-1}\sum_i (x_i - \bar{x})^2 \qquad s_3^2 = \frac{1}{N-1}\sum_i (z_i - \bar{z})^2$$

and $$\bar{x} = \frac{1}{n}\sum_i x_i\,,\ \bar{y} = \frac{1}{n}\sum_i y_i,\ \bar{z} = \frac{1}{n}\sum_i z_i$$

Under a weaker set of linear model assumptions given below, Nathan and Holt [9] have shown the above estimator $\hat{\beta}_{12}$ to be an asymptotically unbiased estimator (unconditionally).

$$\left.\begin{aligned} X_{1\alpha} &= \mu_1 + \beta_{13}(X_{3\alpha} - \mu_3) + e_{1\alpha} \\ X_{2\alpha} &= \mu_2 + \beta_{23}(X_{3\alpha} - \mu_3) + e_{2\alpha} \\ e_{1\alpha} &= \beta_{12.3} e_{2\alpha} + \eta_{1\alpha} \end{aligned}\right\} \qquad (3.1.2)$$

where

$$E(e_{2\alpha} \mid X_{3\alpha}) = E(\eta_{1\alpha} \mid X_{3\alpha}) = E(e_{2\alpha}\eta_{1\alpha} \mid X_{3\alpha}) = 0$$

$$E(e_{2\alpha}^2 \mid X_{3\alpha}) = \sigma_{2.3}^2;\ E(\eta_{1\alpha}^2 \mid X_{3\alpha}) = \sigma_{1.23}^2$$

They have developed estimators b_{12}^* and $\hat{\beta}_{12}^*$ which are π- weighted versions of estimators $b_{12} = \frac{s_{12}}{s_2^2}$ and the MLE $\hat{\beta}_{12}$ for the probability sampling designs *i.e.* for designs that, with probability one, having first order inclusion probabilities $\{\pi_\alpha : p(\alpha \in S \mid Z) > 0, \alpha = 1, 2, ..., N\}$. Let the weighted sample means, variances and covariance be defined as

$$\bar{x}^* = \frac{1}{n}\sum_{\alpha \in s}\frac{x_\alpha}{N\pi_\alpha}\,,\ \bar{y}^* = \frac{1}{n}\sum_{\alpha \in s}\frac{y_\alpha}{N\pi_\alpha}\,,\ \bar{z}^* = \frac{1}{n}\sum_{\alpha \in s}\frac{z_\alpha}{N\pi_\alpha}$$

$$s_{12}^* = \sum_{\alpha \in s}\frac{x_\alpha y_\alpha}{N\pi_\alpha} - \frac{\bar{x}^*\bar{y}^*}{\sum_{\alpha \in s}\left(\frac{1}{N\pi_\alpha}\right)} \qquad s_{13}^* = \sum_{\alpha \in s}\frac{z_\alpha y_\alpha}{N\pi_\alpha} - \frac{\bar{z}^*\bar{y}^*}{\sum_{\alpha \in s}\left(\frac{1}{N\pi_\alpha}\right)}$$

$$s_{23}^{*} = \sum_{\alpha \in s} \frac{x_\alpha z_\alpha}{N\pi_\alpha} - \frac{\bar{x}^{*}\bar{z}^{*}}{\sum_{\alpha \in s}\left(\frac{1}{N\pi_\alpha}\right)} \qquad s_1^{*2} = \sum_{\alpha \in s} \frac{y_\alpha^2}{N\pi_\alpha} - \frac{\bar{y}^{*2}}{\sum_{\alpha \in s}\left(\frac{1}{N\pi_\alpha}\right)}$$

$$s_2^{*2} = \sum_{\alpha \in s} \frac{x_\alpha^2}{N\pi_\alpha} - \frac{\bar{x}^{*2}}{\sum_{\alpha \in s}\left(\frac{1}{N\pi_\alpha}\right)} \qquad s_3^{*2} = \sum_{\alpha \in s} \frac{z_\alpha^2}{N\pi_\alpha} - \frac{\bar{z}^{*2}}{\sum_{\alpha \in s}\left(\frac{1}{N\pi_\alpha}\right)}$$

Nathan and Hot [9] gave the following π-weighted estimators of β

$$b_{12}^{*} = \frac{s_{12}^{*}}{s_2^{*2}} \tag{3.1.3}$$

$$\hat{\beta}_{12}^{*} = \frac{s_{12}^{*} + \frac{s_{13}^{*}s_{23}^{*}}{s_3^{*2}}\left(\frac{\hat{\sigma}_3^2}{s_3^{*2}} - 1\right)}{s_2^{*2} + \frac{s_{23}^{*2}}{s_3^{*2}}\left(\frac{\hat{\sigma}_3^2}{s_3^{*2}} - 1\right)} \tag{3.1.4}$$

These estimators have been proved as asymptotically unbiased (unconditionally) to $0(n^{-1})$ and also the following relations hold good

$$V(\hat{\beta}_{12}) \leq V(\hat{\beta}_{12}^{*})$$

As these estimators $\hat{\beta}_{12}^{*}$, b_{12}^{*} have been established to perform very well even when the assumption of trivariate normality of X, Y, Z does not hold good and since they are essentially model-free, these estimators were considered for outlier-robustification studies along with the MLE $\hat{\beta}_{12}$.

3.2 Proposed Estimators

The finite population total T can be expressed as

$$T = \sum_{\alpha=1}^{N} y_\alpha = \sum_{\alpha \in s} y_\alpha + \sum_{\alpha \notin s} y_\alpha$$

where s is a sample of size n selected from a population of size N using a probability sampling design p(s). The first sum in this expression for T is known and the second must be estimated from the sample. Using the MLE $\hat{\beta}_{12}$ due to Demets and Halperin [3] and the weighted estimators b_{12}^{*} and $\hat{\beta}_{12}^{*}$ due to Nathan and Holt [9], three estimators were constructed for estimation of the population total as given below

$$T_{DH} = \sum_1 y_i + \hat{\beta}_{12} \sum_2 x_i \tag{3.2.1}$$

$$T_{NH1} = \sum_1 y_i + b^*_{12} \sum_2 x_i \tag{3.2.2}$$

$$T_{NH2} = \sum_1 y_i + \hat{\beta}^*_{12} \sum_2 x_i \tag{3.2.3}$$

where $\sum_1$ denotes the summation over sampled units and $\sum_2$ the summation over unsampled part of the population. With the intention of making these estimators further robust against outliers, outlier-robustification was done using Chambers [2] approach. A real-valued function ψ, was utilised for this purpose and the model given by (2.1) was used with $V(x_i) = 1$. The outlier-robustified versions of these estimators T_{DH}, T_{NH1} and T_{NH2} are given as under

$$T'_{DH} = \sum_1 y_i + \hat{\beta}_{12} \sum_2 x_i + \sum_1 p_i \psi\left|(y_i - \hat{\beta}_{12} x_i) / \sigma_i\right| \tag{3.2.4}$$

where $$p_i = \frac{\frac{1}{n}\frac{(x_i - \bar{x})}{\sigma_i} + \frac{1}{n}\frac{(z_i - \bar{z})}{\sigma_i}\frac{s_{23}}{s_3^2}\left(\frac{\hat{\sigma}_3^2}{s_3^2} - 1\right)}{s_2^2 + \frac{s_{23}^2}{s_3^2}\left(\frac{\hat{\sigma}_3^2}{s_3^2} - 1\right)}$$

$$T'_{NH1} = \sum_1 y_i + b^*_{12} \sum_2 x_i + \sum_1 q_i \psi\left|(y_i - b^*_{12} x_i) / \sigma_i\right| \tag{3.2.5}$$

where $$q_i = \frac{\left(\frac{x_i}{N\pi_i}\right) - \left(\frac{\bar{x}^*}{N\pi_i}\right)\sum_1\left(\frac{1}{N\pi_i}\right)}{s_2^{*2}}$$

$$T'_{NH2} = \sum_1 y_i + \hat{\beta}^*_{12} \sum_2 x_i + \sum_1 r_i \psi\left|(y_i - \hat{\beta}^*_{12} x_i) / \sigma_i\right| \tag{3.2.6}$$

where $$r_i = \frac{\left(\frac{x_i}{N\pi_i}\right) - \left(\frac{\bar{x}^*}{N\pi_i}\right)\sum_1\left(\frac{1}{N\pi_i}\right) + \left\{\frac{z_i}{N\pi_i} - \left(\frac{\bar{z}^*}{N\pi_i}\right)\sum_1\left(\frac{1}{N\pi_i}\right)\right\}\frac{s_{23}^{*2}}{s_3^{*2}}\left(\frac{\hat{\sigma}_3^2}{s_3^{*2}} - 1\right)}{s_2^{*2} + \frac{s_{23}^{*2}}{s_3^{*2}}\left(\frac{\hat{\sigma}_3^2}{s_3^{*2}} - 1\right)}$$

4. Simulation Study

With a view of evaluating the performance of the proposed robust estimators, a simulation study was done, comparing the proposed estimators with outlier-robust estimators T_n given by Chambers [2]. For this a background was needed in the form of a finite population, such that

(1) a small proportion p of the population is contaminated by the presence of outliers.

(2) the trivariate normality assumption of (Y, X, Z) does not hold necessarily, but when this assumption does not hold atleast the variables Y, X, Z satisfy the relations given by (3.1.2).

For the generation of data, following parameters were used

(i) the population mean vector was

$$\mu' = [\mu_x, \mu_y, \mu_z] = [50, 60, 70]$$

(ii) Variances $[\sigma_x^2, \sigma_y^2, \sigma_z^2] = (310.64^2, 763.99^2, 100^2)$

and

(iii) $\rho_{yx} = 0.76$, $\rho_{yz} = 0.61$, $\rho_{xz} = 0.63$, $\rho_{yz.x} = 0.27$ and $\beta_{yx} = 0.309$

The variance of Z was assumed as 100^2 first and then the other variances were calculated according to parametric relationships. The population parameters were selected for the study because the real population considered by Nathan and Holt [9] satisfy these parameters and since the proposed estimators are based on Nathan and Holt estimators of β, it was thought that a simulated population with these parameters would be appropriate for a comparative simulation study.

A finite population of 180 units was generated with observations on Y, X, Z using the above mentioned parametric specifications and satisfying the relationships given by the model set-up (2.1). The finite population thus obtained was further contaminated by deliberately replacing k = 4 units by previously generated outlier values. These outliers were generated in a similar fashion as that of the remaining units but with an inflated variance. With k = 4, it was observed that on an average, nearly 50% of total 1000 samples generated from the population for the purpose of simulation study gets contaminated with outliers. In an agricultural set-up, such a situation may correspond to a study from a large number of farm holdings where data on each farm are available for crop yield (Y), total acreage (X) and the total value of products sold in the previous year (Z).

In order to provide a basis for comparing the sampling distributions of a variety of finite population strategies under repeated sampling from the

above-simulated study population, the study was conducted in the following pattern.

Denoting a strategy as a pair (π, E) where π is a method used to select a sample and E is an associated estimator of T, the comparisons made were in terms of average of errors, E – T, denoted by BIAS and the square root of the average of squared errors $(E - T)^2$ denoted by RMSE. In all cases these averages were over 1,000 independent samples selected from the population according to a sampling method π. When two strategies are based on the same sampling method, the errors associated with two different estimators defining these strategies were based on the same set of samples selected from the study population. The seven sampling methods considered for the study are given by Table 1.

Table 1. Sampling methods used in the study

Code	Description
1.	Simple random sampling without replacement
2.	Systematic sampling, with sample π_i's proportional to Z carried out on a population list ordered by increasing Z (random start)
3.	Probability proportional to size sampling with replacement
4.	Stratified random sampling with 3 strata
5.	Stratified random sampling with 6 strata
6.	Stratified sampling with pps (with 3 strata)
7.	Stratified sampling with pps (with 6 strata)

Table 2 gives the estimators used in the study. From the finite population of 180 units, various sampling designs were used to select 1000 independent samples of size n = 20 units based on the Z variable. The BIAS and RMSE were calculated for all the sampling strategies based on these data sets.

In general, the proposed estimators, particularly T'_{NH2} and T'_{DH} perform well in most to the cases since for equal probability selection methods (EPSEM) designs, the π-weighted MLE $\hat{\beta}^*_{12}$ of β reduces to MLE $\hat{\beta}_{12}$, it can be seen that in the case of *srswor* the estimators T'_{DH} and T'_{NH2} register the same values for BIAS and RMSE (see Figures 2 and 3). For designs involving pps sampling the proposed robust estimator T'_{NH2} performs extremely well which shows that the incorporation of design information (through π-weights) into the estimator robustifies it against model violations. Although for designs *ppswr* and systematic sampling with *pps*, the estimator T'_{NH2} performs very well with respect to RMSE, it records the highest value for BIAS in comparison with other estimators. But this is not at all surprising given the well known feature of bias/variance trade off in the presence of outliers. Also for *ppswr* sampling

Table 2. Estimators used in the study

Code	Description
1.	T_n with $\Psi(t) = 0$
2.	T'_n i.e. T_n with $\Psi(t) = t.e^{-a/2(\|t\|-b)}$, $a = 0.5$ and $b = 6$
3.	T_{DH} given by (3.2.1) based on MLE $\hat{\beta}_{12}$
4.	T'_{DH} given by (3.2.4) which is the robustified version of T_{DH}
5.	T_{NH1} given by (3.2.2) based on π-weighted OLS estimator b^*_{12}
6.	T'_{NH1} given by (3.2.5) which is the robustified version of T_{NH1}
7.	T_{NH2} given by (3.2.3) based on π-weighted MLE β^*_{12}
8.	T'_{NH2} given by (3.2.6) which is the robustified version of T_{NH2}

Table 3. Results of the proposed simulation study
Total is 48044.09

	Estimator	BIAS	RMSE
	(1) Simple random sampling without replacement		
1.	T_n	–6974.04	12996.70
2.	T'_n	–6941.32	12956.72
3.	T_{DH}	–6812.12	12933.78
4.	T'_{DH}	–6716.16	12797.87
5.	T_{NH1}	–7135.03	13064.05
6.	T'_{NH1}	–7030.95	12921.45
7.	T_{NH2}	–6812.11	12933.78
8.	T'_{NH2}	–6716.17	12797.87
	(2) Systematic sampling with pps		
1.	T_n	–4968.92	8384.40
2.	T'_n	–4969.01	8364.47
3.	T_{DH}	–5209.15	8337.24
4.	T'_{DH}	–5199.04	8325.62
5.	T_{NH1}	–5877.23	8353.73
6.	T'_{NH1}	–5827.12	8350.36
7.	T_{NH2}	–6004.89	8264.73
8.	T'_{NH2}	–5961.74	8222.86

	(3) pps sampling with replacement		
1.	T_n	–6388.41	13044.84
2.	T'_n	–6338.44	12969.30
3.	T_{DH}	–6206.33	12809.57
4.	T'_{DH}	–6048.08	12589.68
5.	T_{NH1}	–6238.85	12977.89
6.	T'_{NH1}	–5838.30	12727.04
7.	T_{NH2}	–8506.00	12901.04
8.	T'_{NH2}	–7934.20	12558.12
	(4) Stratified random sampling with 3 strata		
1.	T_n	–6574.40	12321.65
2.	T'_n	–6370.88	12035.07
3.	T_{DH}	–6747.46	12468.71
4.	T'_{DH}	–6528.23	12165.52
5.	T_{NH1}	–7095.63	12600.27
6.	T'_{NH1}	–6859.50	12268.00
7.	T_{NH2}	–6960.50	12399.91
8.	T'_{NH2}	–6727.56	12148.40
	(5) Stratified random sampling with 6 strata		
1.	T_n	–6467.46	12313.05
2.	T'_n	–6299.23	12084.72
3.	T_{DH}	–6596.47	12397.75
4.	T'_{DH}	–6409.88	12158.90
5.	T_{NH1}	–6632.71	12290.76
6.	T'_{NH1}	–6430.68	12035.82
7.	T_{NH2}	–6592.00	12285.59
8.	T'_{NH2}	–6393.55	12038.25
	(6) Stratified sampling with pps (with 3 strata)		
1.	T_n	–6822.91	12918.57
2.	T'_n	–6649.43	12691.09
3.	T_{DH}	–6829.80	12978.47
4.	T'_{DH}	–6656.57	12747.29
5.	T_{NH1}	–7022.11	13243.24
6.	T'_{NH1}	–6811.44	12945.25
7.	T_{NH2}	604.78	12991.11
8.	T'_{NH2}	549.10	12816.51

(7) Stratified sampling with pps (with 6 strata)			
1.	T_n	–7540.15	13415.28
2.	T'_n	–7368.98	13226.46
3.	T_{DH}	–7491.78	13395.81
4.	T'_{DH}	–7327.29	13223.07
5.	T_{NH1}	–7436.35	13878.24
6.	T'_{NH1}	–7226.66	13590.43
7.	T_{NH2}	774.16	13761.16
8.	T'_{NH2}	614.16	13339.75

For further illustration, the results of strategy - simple random sampling with various estimators is presented graphically as follows.

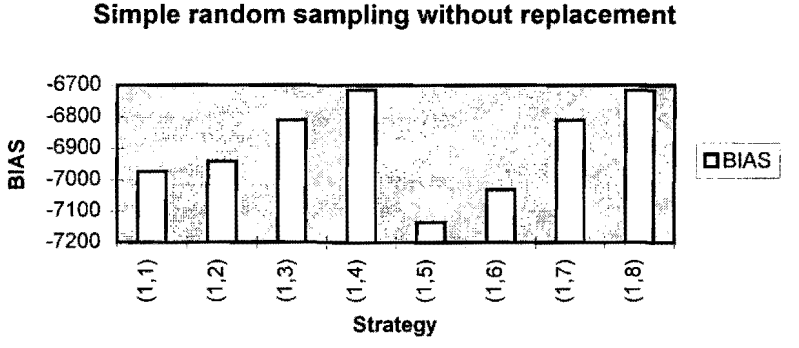

Fig. 2. BIAS performance of the proposed estimators (srswor)

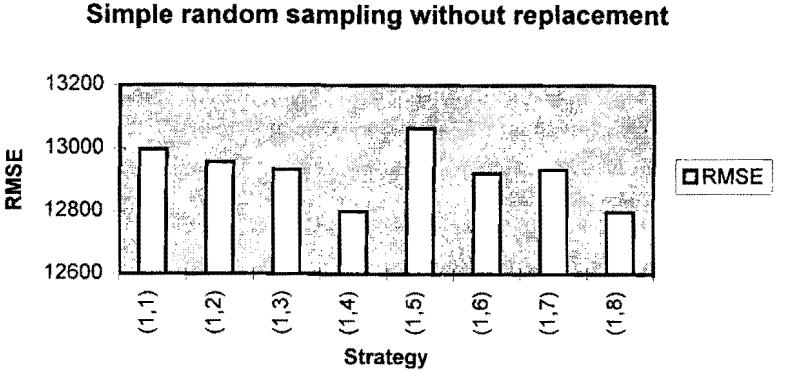

Fig. 3. RMSE performance of the proposed estimators

design, among the proposed robust estimators T'_{DH}, T'_{NH2} the estimator T'_{DH} is superior in performance with respect to RMSE as well as BIAS. This is again due to the theoretical result that the MLE $\hat{\beta}_{12}$ is better than the π-weighted estimator $\hat{\beta}^*_{12}$ when *ppswr* sampling design is employed.

5. *Conclusions*

In this article some estimators have been proposed which are robust against violations in model assumptions and the outliers occuring in survey data. These estimators are based on the model-free estimators of the regression coefficient β. A simulation study comparing the proposed estimators with the standard outlier-robust estimators due to Chambers [2] reveal the superiority of these estimators in terms of root mean square error. Particularly, when probability proportional to size sampling design is employed the proposed estimators are robust against the anomalies to a great extent.

REFERENCES

[1] Basu, D., 1971. An essay on the logical foundations of survey sampling, Part I. In: V.P. Godambe and D.A. Sprott (Eds.), *Foundation of Statistical Inference* (pp 203-242) (Montreal, Holt, Rinehart and Winson).

[2] Chambers, R.L., 1986. Outlier robust finite population estimation, *JASA*, **81**, 1063-1069.

[3] Demets, D. and Halperin, M., 1977. Estimation of simple regression coefficient in samples arising from a subsampling procedure. *Biometrics*, **33**, 47-50.

[4] Glasser, G.J., 1962. On the complete coverage of large units in a statistical study. *Review of the International Statistical Institute*, **30**, 28-32.

[5] Gwet, J.P. and Rivest, L.P., 1992. Outlier resistant alternatives to the ratio estimator. *JASA*, **87**, 1174-1182.

[6] Hidiroglou, M.A. and Srinath, K.P., 1981. Some estimators of a population total from simple random samples containing large units. *JASA*, **69**, 383-393.

[7] Hulliger, B., 1986. A small simulation study on the outlier sensitivity of survey sample strategies. *Research Report No. 46* (Eidger versische Technische Hochschule, Zurich, Semminar fur Statistik).

[8] Kish, L., 1965. *Survey Sampling*, Wiley, New York.

[9] Nathan, G. and Holt, D., 1980. The effect of survey design on regression analysis. *Journal of the Royal Statistical Society*, **B42**, 377-386.

[10] Smith, T.M.F., 1987. Influential observations in survey sampling. *Journal of Applied Statistics*, **14**, 143-152.