Performance of Robust Estimators: Sampling, Variables and Dimensions

Elizabeth Reis[1,2] and Maria do Carmo Botelho[1]
[1]Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
[2]Corresponding author: Elizabeth Reis, e-mail: ear@iscte.pt

Abstract

The use of data collected from market research and opinion surveys is common in social and business areas. Probability samples are usually the first option for data collection though they are quite often excluded due to the non existence of a suitable sampling frame. In addition to the lack of randomness of sample data, further problems are caused by inadequate sample representativeness for some population characteristics, limitations of the measuring instrument and occurrence of measurement errors. In these situations, robust statistical techniques can be a valid option for estimation purposes as they are not sensitive to sample biases. The main aim of this study is to evaluate the performance of robust estimators, particularly Huber M-estimator, Tukey's biweight and Least Trimmed Squares (LTS) estimators, when compared to the sample mean and median, and applied to different types of variables, diverse sampling methods and dimensions. Quantitative and qualitative ordinal Likert type variables with 4, 6 and 10 point were used. Samples were generated by stratified and quota methods, both with dimensions 50, 100 and 300. Results show the best behavior of the Huber and the Tukey's biweight estimators in most situations, particularly for quantitative variables, for both sample methods. The LTS estimator performed worse than any other estimator, being better solely in the case of ordinal variables with a 4 point scale, sharp skewness and high kurtosis.

Keywords: estimation, M-estimators, robustness, sampling methods.

1. Introduction

In studies based on survey sampling the use of inferential techniques is a sensitive field. Surveys are frequently used in social and business research to study values, beliefs, attitudes and behaviors. Many times it is difficult to get a sampling frame for a certain target population, so nonrandom methods such as quota sampling are often used. In survey sampling, independent and identically distributed (i.i.d.) observations is a rare situation. Even when a random selection is possible, based on a defined finite population, missing data occurs and randomness is not completely guaranteed. Without random data, the application of classical inferential methods is obviously compromised (Hampel, 2000), and robust techniques need to be applied.

The empirical knowledge accumulated in many years of survey and opinion polls practice shows that results from nonrandom sampling methods, such as quota sampling, are close to those obtained by probabilistic methods, particularly when a careful data collection process is guaranteed (Cochran, 1977).

Hoaglin et al. (1983: 1) refers that “robust and resistant methods, instead of being the best possible in a narrowly defined situation, are “best” compromises for a broad range of situations and, surprisingly often, are close to “best” for each situation alone. Whereas distribution-free methods treat all distributions equally, robust and resistant methods discriminate between those that are more and less plausible”. The robust theory and applications are not only suited for small deviations caused by marginal values or gross errors changing the normal distribution, frequent in sampling studies, but also when deviations occur in usually assumed assumptions, for any parametric model, such as independence.

Cuevas (1994) argues that, after four decades of studies and applications, the resistance to the widespread use of robust techniques still remains, probably due to the lack of communication and dialogue between theorists and users (statisticians). The increasing availability of electronic means and high-efficiency in the treatment of large quantities of information caused a stimulus on statistical analysis by an audience mostly of non-statistical professionals. Any theory that cannot communicate with this audience tends to be away from several fields of application.

This investigation arises from the interest in robust statistical methods and its applications to surveys and market research real data, in social and business sciences. It proposes an evaluation of robust techniques applications when models and assumptions fail, thus contributing to the dissemination of these methods. The aim of this study is to compare the performance of three robust location estimators, the Huber M-estimator, the Tukey's biweight and the Least Trimmed Squares (LTS), to estimate a location parameter, in presence of different variable types, sampling processes and sample dimensions.

2. Robust estimation

The robust theory was introduced by Huber (1964) to solve a location problem for the parameter μ when the *F(x)* distribution is approximately known. The author considered that the observations were normally distributed with variance equal to 1, partly affected by gross errors with *H(x)* distribution, according to the contaminated model $F(x) = (1-\varepsilon)\Phi(x) + \varepsilon.H(x)$, with $0 \leq \varepsilon < 1$ a known value, with standard normal distribution $\Phi(x)$ and where H is arbitrary. With this model, the concept of "neighborhood" of a strictly parametric model arises for the first time and robust methods can be assumed as a natural development from the classical parametric models.

The location M-estimators are employed in the same study to solve a location problem when the distribution is not exactly normal. Based on the maximum likelihood method, an estimator is defined as $T_n = T_n(X_1; \ldots; X_n)$ that maximizes$\prod_{i=1}^{n} f(X_i, T_n)$. According to Huber, an M-estimator minimizes a more general odd function, $\sum_{i=1}^{n} \rho(X_i, T_n)$, where ρ is a defined function in $\mathcal{F}X\Theta$, $\mathcal{F}$ represents a function family and Θ is a parameter space. Consider $\psi(x,\theta) = (\partial/\partial\theta)\rho(x,\theta)$, so T_n satisfies the implicit equation $\sum \psi(X_i; T_n) = 0$. The mean and the median are solutions of this equation, so they belong to this group of estimators. The Huber M-estimators family shows a quadratic odd function in the center and linear in the tail, with *c* as a constant

$$\rho(u) = \begin{cases} \frac{1}{2}u^2 & se\ |u| \leq c \\ c|u| - \frac{1}{2}c^2 & se\ |u| > c \end{cases}$$

Consider $u_i = (X_i - T_n)/S_n$, where S_n is a scale measure, frequently the normalized median of absolute deviation, (normalized MAD). The corresponding ψ and the weight functions $\varpi = \psi(u)/u$ are

$$\psi(u) = \begin{cases} u & se\ |u| \leq c \\ c.sgn(u) & se\ |u| > c \end{cases}, \qquad \varpi(u) = \begin{cases} 1 & se\ |u| \leq c \\ \frac{c}{u}.sgn(u) & se\ |u| > c \end{cases}$$

With this estimator, if the standardized distance to the location estimate is lower than the tuning constant *c*, the observation has a weight equal to 1. For values larger than this tuning constant the weight decreases, but all observations are used to produce the estimate. However, this estimator might still be sensitive to extreme observations.

A family of estimators, less sensitive to the presence of extreme values, is the redescendent family. These estimators have a finite point rejection, so after a certain value, the observations are excluded from the estimate (Hampel, 2001). The most popular robust family is the Tukey's biweight, with good overall performance, robustness of efficiency and resistant (Hoaglin et al., 1983). The odd function, with tuning constant *c* is

$$\rho(u) = \begin{cases} \frac{1}{6}[1-(1-u^2)^3] & se\ |u| \le 1 \\ \frac{1}{6} & se\ |u| > c \end{cases}$$

with $c > 0$ and $u_i = (X_i - M)/cMAD$. The corresponding ψ and weight functions $\varpi = \psi(u)/u$ are

$$\psi(u) = \begin{cases} u(1-u^2)^2 & se\ |u| \le 1 \\ 0 & se\ |u| > 1 \end{cases}, \qquad \varpi(u) = \begin{cases} (1-u^2)^2 & se\ |u| \le 1 \\ 0 & se\ |u| > c \end{cases}$$

Other estimators have been proposed, such as the Least Trimmed Squares (LTS) widely used in regression analysis (Rousseeuw and Leroy, 1987). In order to estimate a parameter θ, in the univariate case, the LTS estimator is obtained by the minimization of $\sum_{i=1}^{h}(r^2)_{i:n}$, where $h = (n/2) + 1$ and $(r^2)_{i:n}$ are the ordered square residuals. From the ordered observations, $n - h + 1$ subsamples are drawn with h observations each. The subsamples mean is calculated $(\bar{x}^{(j)}, j = 1, \dots, (n-h+1))$ and the correspondent sum of squares$\left(SQ^{(j)}, j = 1, \dots (n-h+1)\right)$. The LTS estimate is the mean of the subsample with smallest sum of squares. The LTS algorithm is simpler and easier to calculate then the M-estimators.

Previous studies with M-estimators and LTS estimators are mainly theoretical and only a few applications to real data are known. The Princeton study, published by Andrews et al. (1972), is the first systematic and exhaustive investigation comparing robust estimators. It has been followed by applications with biomedical data and in the chemical area (Gross, 1973; Stigler, 1977; Rocke et al., 1982; Hill and Dixon, 1982), using point estimation and confidence intervals with robust estimators of location and scale (Gross, 1976; Boos, 1980). In most recent studies robust estimation has been related to multivariate problems and methods such as principal components, discriminant analysis or multivariate regression (Marona, 1976; Portnoy and He, 2000). Stigler (1977) proposed a comparison between robust estimators applied to real data and evaluated their performance with the mean square error (MSE), the index of relative error (RE) and its corresponding variation (SE). Considering the comparison of p estimators ($\hat{\theta}$), for k groups of X_j elements, the mean absolute error is $s_j = p^{-1}\sum_{i=1}^{p}\left|\hat{\theta}_{ij} - \theta_j\right|$, where $\hat{\theta}_{ij}$ is the value of the estimator $\hat{\theta}_i$ for X_j and θ_j is the "true value" for the *j*th data set. The relative error can be calculated by $e_{ij} = \left|\hat{\theta}_{ij} - \theta_j\right|/s_j$. If $e_{ij} < 1$ it means that the estimator has a smaller error than the average error of all estimators under analysis. The index of relative error $RE_{(i)}$ for the estimator $\hat{\theta}_i$ is calculated by the average of the estimators relative error across k data sets, $RE_{(i)} = k^{-1}\sum_{j=1}^{k} e_{ij}$. The performance variation of estimator $\hat{\theta}_i$ through the several data sets is $SE_{(i)} = \left\{(k-1)^{-1}\sum_{j=1}^{k}\left(e_{ij} - RE_{(i)}\right)^2\right\}^{1/2}$. A small value of $SE_{(i)}$reflects consistent performance.

3. Methodology

The need for application and disclosure of robust estimators is steel necessary in social and business areas, using sampling studies real data. The present research arises from the difficulty of analyzing data obtained by sampling surveys, frequently with nonrandom sampling processes. It aims to provide statisticians and survey and market research professionals with a performance evaluation of alternative techniques of location estimation, using robust statistics, different variable types, sampling methods and sample dimensions. Hampel states that "statistics is most alive when reacts to the needs of applications" (1997: 3) and supports the existence of some areas where the application of robust techniques can still be further explored. This is the case of information collected on people's behavior, beliefs, perception or motivations where Likert-type items are the most used form of measuring scales, with a finite number of values anchored at each point or only at the end. Statistical analysis considering this

type of variables as metric is frequent and based on the underlying assumption of equal distances between adjacent categories implicit in the usual quantification (assigning consecutive integers). But this is not a consensual assumption.
The present study uses quantitative variables as well as Likert-type items with ten, six and four points, with different levels of skewness and kurtosis. The identification of the 7 variables under analysis and their distributional characteristics are presented in Table 1.

Table 1 – Population distribution characteristics by variable type

Quantitative	Likert-type items					
	10 points	6 points			4 points	
age	*p9*	*p4.7*	*p4.3*	*p4.6*	*inc7*	*inc9*
Slightly asymmetric	Slightly asymmetric	Slightly asymmetric	Asymmetric	Sharp asymmetry Extreme values Hight kurtosis	Slightly asymmetric	Sharp asymmetry

Population data were obtained from two surveys A and B (N_A=1800 and N_B=1500). Samples with 50, 100 and 300 observations were drawn from the two populations. Several earlier studies used samples with dimensions from 10 up to 50 cases (Hoaglin et al., 1983). The "large sample" definition is associated with more than 30 cases, some authors referring the need to use 50 or more cases, when the population structure is unknown (Chernick, 1999; Lohr, 1999).
The sampling methods are considered a central issue in survey sampling. Stratified and quota samples were selected. To construct quota samples some observations were intentionally excluded, to generate a group of samples with a lack of randomness. 50 samples were generated for each dimension of 50, 100 and 300 cases.
Huber and Tukey´s biweight were the initially selected M-estimators, but they need large data variability for correct calculation, difficult to find on four points Lykert-type items; so the LTS estimator was also included in the analysis. For comparison purposes, the usual mean and median estimators were also used. The five estimators were applied to all generated samples across sampling methods, sample sizes and variable types in a total of 1500 point estimates calculated for each variable. All samples were generated by Complex Samples, an add-in on the IBM SPSS Statistics and by created routines for nonstandard situations. The point and interval estimates were calculated with IBM SPSS Statistics and S-Plus.
In order to analyze the estimators' behavior, estimates distributions were compared for each variable, sampling method and sample size. Confidence intervals were also calculated to estimate location, for each sampling method and sample size for quantitative data and 10 points Likert-type item. This is an opinion variable, different from quantitative demographic data. The large number of points allows a certain data variability and enables a greater detail in the discrete structure of opinion, as referred by Bryman and Cramer (2003). Furthermore, the variable distribution is close to normal in skewness and kurtosis. The simplicity of use, when the population distribution is unknown and data are based on a sample, is a good argument for the choice of bootstrap confidence intervals. The stratified bootstrap and percentile method were used to construct the interval estimates. This method is the most suitable for M-estimators (Wilcox, 2003). The mean square error (MSE), the index of relative error (RE) and corresponding variation (SE) were used to compare the estimators' performance, as referred by Stigler (1977).

4. Results

Applications on quantitative data reveal differences on estimates distributions with a greater dispersion for those obtained by quota samples. Huber and Tukey´s biweigth are the only two estimators with identical estimates distributions, on samples with 100 and 300 cases, for both sampling methods. This is an important result showing that

these M-estimators generate similar results with random and nonrandom quantitative data, when the sample dimension is large.
The 10 points Likert-type item shows identical estimates distribution for Huber and Tukey´s biweigh estimators among the three stratified sample sizes. In quota samples the estimates distributions show again greater dispersion. The distribution of the median is the one with lowest dispersion. The 6 point Likert-type items produce different results according to the sampling method and distribution shape. For the slightly asymmetric variable distribution and stratified samples, the LTS estimates distribution is the most dissimilar; for quota samples almost all distributions are different with the median showing the smallest dispersion in all sample sizes. The estimates variation is highest when quota samples are associated to highly skewed distributions.
The sample size effects show decreasing variance on estimates distributions in the majority of the created scenarios. The exception occurs with the quantitative variable and the 10 point Likert-type item, on quota samples: the increasing sample size does not reduce the variance. The 6 point Likert-type items with minor asymmetry increase the LTS bias; sharp asymmetry and high kurtosis associated with large sample size reduce LTS and the variance of mean, but decrease bias for all estimates distributions. The 4 points Likert-type items associated with increasing sample size reduce bias of the median in stratified samples and the bias of LTS bias in quota samples. The decrease of Huber and Tukey´s biweight variance distributions do not occurs with 4 point Likert-type items characterized by sharp asymmetry.
The bootstrap confidence intervals for quantitative data are similar for all estimators and all sample sizes, on stratified and quota samples, the only exception being the quota sample size 300 where the interval does not include the parameter. The bootstrap confidence intervals for the 10 points Likert-type item are identical in both sampling methods and reveal similarity between the median and the Tukey´s biweight estimators, the later showing smaller dispersion for samples with 50 and 100 cases.
When applied to quantitative data, the Huber and the Tukey´s biweight estimators show the best performance. Tukey´s biweight still performs best when applied to the 10 point Likert-type item, in stratified samples. The LTS performance shows high values of the relative error index for both sampling methods, frequently increasing as sample size increases; the performance of LTS is less distant from the others estimators' performance for stratified samples.

5. Conclusions

This paper studies the performance of Huber and Tukey´s biweight M-estimators and LTS, either individually or by comparison to the mean and the median estimators. The performance was evaluated for different sample sizes, variable types, for stratified and quota samples. The use of robust techniques is justified because real data do not follows the theoretical models, showing deviations and errors or marginal values. In these situations robust procedures should be a preferential choice when confronted with classical procedures.
The Huber and Tukey´s estimators presented a good performance, followed by the median and the mean in almost all studied scenarios. In fact, when applied to quantitative real data, the former estimators show a similar performance in each sampling method, with a relative error index and corresponding variation mostly often smaller than median's and always better than the mean's. M-estimators generate similar estimates results with random and nonrandom quantitative data, with large samples. In addition, its bootstrap confidence intervals have lower range, particularly in stratified samples.
The LTS estimator reveals the poorest performance in both sampling methods and sample sizes. Only two exceptions are noticed: for the 6 point Likert-type item with a sharp asymmetry and high kurtosis in both sampling methods, and in the 4 point Likert-type item with minor skewness.

In general, results point out the better performance of the Tukey´s biweight and also the Huber M-estimators to estimate location for real quantitative data, with large stratified or quota samples. For a 10 points Likert-type item the choice should still be the Tukey´s biweight, applied to stratified samples. For data obtained from Likert-type items with smaller number of points, the median is still is the best choice, followed by M-estimators, except for sharp skewed and high kurtosis distributions.

References

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H. and Tukey, J. W. (1972) *Robust Estimates of Location: Survey and Advances*, Princeton University Press, New Jersey.

Boos, D. D. (1980) "A new method for constructing approximate confidence intervals from M-estimates," *Journal of the American Statistical Association,* 75(369), 142-145.

Bryman, A. and Cramer, D. (2003), *Análise de Dados em Ciências Sociais: Introdução às Técnicas Utilizando o* SPSS *para Windows*, Celta Editora, Lisboa.

Chernick, M. R. (1999) *Bootstrap Methods: A practitioner's guide*, Wiley, NY.

Cochran, W. G. (1977) *Sampling Techniques*, 3ª ed, Wiley, New York.

Cuevas, A. (1994) "Estimación robusta," *Estadistica Española,* 36(137), 351-355. Discussion to Zamar´s paper.

Gross, A. M. (1973) "A Monte Carlo swindle for estimators of location," *Applied Statistics,* 22(3), 347-353.

Gross, A. M. (1976) "Confidence interval robustness with long-tailed symmetric distributions," *Journal of the American Statistical Association,* 71, 409-416.

Hampel, F. (1997) "Is statistics too dificult?," *Research Report nº 81*, Seminar für Statistik, Eidgenossische Technische Hochschule (ETH), Zürich.

Hampel, F. (2000) "Robust inference," *Research Report nº 9,*. Seminar für Statistik, Eidgenossische Technische Hochschule (ETH), Zürich.

Hampel, F. (2001) "Robust statistics: A brief introduction and overview," *Research Report nº 94*, Seminar für Statistik, Eidgenossische Technische Hochschule (ETH), Zürich.

Hill, M. and Dixon, W. J. (1982) "Robustness in real life: A study of clinical laboratory data," *Biometrics,* 38, 377-396.

Hoaglin, D. C., Mosteller, F. and Tukey, J. W. (1983) *Understanding robust and exploratory data analysis*, John Wiley & Sons, New York.

Huber, P. J. (1964) "Robust estimation of a location parameter," *The Annals of Mathematical Statistics,* 35, 73-101.

Lohr, S. L. (1999) *Sampling: Design and Analysis*, Duxbury Press, New York.

Marona, R. A. (1976) "Robust m-estimators of multivariate location and scatter," *The Annals of Statistics,* 4(1), 51-67.

Portnoy, S. and He, X. (2000) "A robust journey to the new millennium," *Journal of the American Statistical Association,* 95(452), 1331-1335.

Rocke, D. M., Downs, G. W. and Rocke, A. J. (1982) "Are robust estimators really necessary?," *Technometrics,* 24(2), 95-101.

Rousseeuw, P. J. and Leroy, A. M. (1987) *Robust regression and outlier detection*, Wiley, New York.

Stigler, S. M. (1977) "Do estimators work with real data?," *The Annals of Statistics,* 5 (6), 1055-1098.

Wilcox, R. R. (2003) *Applying Contemporary Statistical Techniques*, Academic Press.