MEDIAN ESTIMATES OF REGRESSION

Roberto N. Padua*
De La Salle University
Taft Avenue, Manila, Philippines

ABSTRACT

An alternative procedure for estimating the unknown parameters in a linear model is proposed in this paper. The procedure consists of splitting the data into smaller blocks, in each of which the least-squares estimate of the unknown parameters are computed. The proposed estimate is the componentwise-medians of the least-squares estimates computed from the blocks. The asymptotic distribution of the proposed estimate is established. Small sample calculations indicate that the proposed estimate behaves very well in the presence of contamination.

Introduction

Consider the linear model:

$$Y = XB + e \qquad (1)$$

where X is an nxp matrix of full rank p, B is a $px1$ vector of unknown parameters and e is an $nx1$ vector of random errors. The usual assumptions about the distribution of the error terms is that they are independent and identically distributed $F(.)$, where $F(.)$ is usually taken to be the normal distribution with mean of 0 and some finite variance. It is well-known that when the underlying distribution $F(.)$ is not exactly normal but rather some contaminants have been introduced, the performance of the least-squares estimates of B is drastically impaired. A single outlying observation can easily distort the estimate value of B if one uses the ordinary least-squares procedure.

It is precisely this problem that motivates the use of alternative estimation techniques in the linear model problem. Robust regression is a collection of techniques and procedures designed to downweight the effect of anomalous data points on the estimated coefficients. Such procedures are effective alternative methods that somehow dampen the effect of corrupt observations.

Among the alternative procedures suggested in the literature are the M estimates of Huber (1973), the L estimates of Koenker and Bassett (1978) and the R estimates of Adichie (1967), Sen (1968), Jureckova (1971) and Maritz (1979).

*Visiting Fulbright Professor of Statistics, University of Nebraska at Lincoln, Lincoln-Nebraska, 68588-0323.

With the exception of Sen's (1968) and Maritz' (1979) estimates, all of the above procedures involved intricate mathematics which, perhaps, explains why public acceptance of these procedures have been slow.

If simplicity and tractability are set as criteria for robust regression, Sen's (1968) and Maritz' (1979) procedures certainly deserve careful attention. In model (1), let $p = 2$, so that

$$Y_i = a + bx_i + \epsilon_i \ , \ i = 1,2,\ldots,n \tag{2}$$

Sen's (1968) estimate of b is:

$$b^* = \underset{i<j}{\text{median}} \ \frac{Y_j - Y_i}{X_j - X_i} \tag{3}$$

while Maritz (1979) estimate of a is:

$$a^* = \underset{R}{\text{median}} \ \frac{X_i Y_j - X_j Y_i}{X_i - X_j}$$

where R is the set of ordered pairs (i, j) chosen from $(1,2,\ldots,n)$ having no components in common.

It is then desired to extend (2) and (3) in the multiparameter case i.e. $p \geqslant 3$. It is of course easy to generalize (2) by simply computing all possible $\binom{n}{p}$ least-squares estimates of β in (1) and then taking the median componentwise. However, this would require a tremendous amount of computing even with today's standards e.g. if $n = 100$, $p = 10$ then approximately 1.73 x 10^{13} least-squares computations are required.

This paper, then, concentrates on the extension of (3) in the multiparameter case.

Formulation

Assume that in model (1), the errors ϵ_i are *iid* $F(.)$ where $F(.)$ belongs to the class of all absolutely continuous functions symmetric about zero:

Consider:

$$Q_J = (X_J{}'X_J)^{-1}X_J{}'Y_J \ , \ J = 1,2,\ldots,m \tag{4}$$

where X_J is a kxp matrix chosen randomly from X, $k < \text{n}$ and $k \geqslant p$. In other words, randomly split the X matrix in m disjoint blocks each of size kxp. Do the same for the Y vector and call the resulting Y vector as Y_J. Since the errors are symmetrically distributed about zero, it follows that

$$Q_J - \beta \text{ for } J = 1,2,\ldots,m$$

are componentwise symmetrically distributed about zero.

Consider

$$T(Q_J) = \sum_J \operatorname{sgn}(Q_J - \beta) = \begin{array}{l} \sum_J \operatorname{sgn}(q_{1J} - \beta_1) \\ \sum_J \operatorname{sgn}(q_{2J} - \beta_2) \\ \cdot \\ \cdot \\ \cdot \\ \sum_J \operatorname{sgn}(q_{pj} - \beta_p) \end{array} \qquad (5)$$

where q_{iJ} is the ith component of Q_J, and:

$$\operatorname{sgn}(u) = \begin{array}{ll} 1, & \text{if } u > 0 \\ -1, & \text{otherwise.} \end{array}$$

It follows that the components of (5) are each symmetrically distributed about zero under the true value of β. An estimate of the rth component of β is obtained by:

$$\beta_r{}^* = \frac{1}{2}\left\{ \beta_r^{(1)} + \beta_r^{(2)} \right\} \qquad (6)$$

where:

$$\beta_r^{(1)} = \inf\left\{ \beta_r : \sum_J \operatorname{sgn}(q_{rJ} - \beta_r) < 0) \right\}$$

$$\beta_r^{(2)} = \sup\left\{ \beta_r : \sum_J \operatorname{sgn}(q_{rJ} - \beta_r) > 0) \right\}$$

$r = 1,2,\ldots,p$. It is easy to verify that the solutions to (6) for $r = 1,2,\ldots,p$ when written in vector form is given by:

$$\beta^* = \operatorname*{median}_J \left\{ \hat{\beta}_J \right\} \quad , \; J = 1,\ldots,m \qquad (7)$$

where $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_m$ are the least-squares estimates of β computed from the m blocks and the median is taken componentwise.

Asymptotic Theory

Let X_i be a sequence of independent p-dimensional random variables. Let X_{ij} denote the jth component of X_i. It is assumed that the distribution of X_{ij} has a positive density at the origin. Let $x = (\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_p)$, where $\tilde{x}_j$ denotes the median of the jth components of $x_1, x_2, \ldots, x_n$. The usual population median $\xi = (\xi_1, \xi_2, \ldots, \xi_p)$, is assumed to be at $\xi = (0, 0, \ldots, 0)'$. The asymptotic distribution for large n is given as follows: Let $z_i = (z_{i1}, \ldots, z_{ip})$, be given by:

$$z_{ij} = \begin{array}{ll} 1, \text{ if } x_{ij} > x_j/\sqrt{n} & \\ 0, \text{ otherwise} & , j = 1, \ldots, p. \end{array}$$

Let F_{ij} and f_{ij} denote the *cdf* and *pdf* respectively of x_{ij} and for $b = (b_1, b_2, \ldots, b_p)$, let $f_i(b) = (f_{i1}(b_1), \ldots, f_{ip}(b_p))$

$F_i(b) = (F_{i1}(b_1), \ldots, F_{ip}(b_p))$ and

$F_{ijk}(b_j, b_k) = P(x_{ij} \leqslant b_j, x_{ik} \leqslant b_k)$.

Let $\Omega_i(b) = (v_{ijk})$ given by

$$v_{ijj} = F_{ij}(b_j)(1 - F_{ij}(b_j)) \tag{8}$$

$$v_{ijk} = F_{ijk}(b_j, b_k) - F_{ij}(b_j) F_{ik}(b_k). \quad \text{Let } S_n = \sum^{n} Z_i. \tag{9}$$

Clearly then,

$$E(Z_{ij}) = 1 - F_{ij}(x_j/\sqrt{n})$$

$$\text{cov}(Z_i) = \Omega(x/\sqrt{n})$$

It follows that

$$\sqrt{n}\, X \leqslant b \quad \text{iff} \quad S_n \leqslant \frac{n-1}{2} \underset{\sim}{e}$$

where $\underset{\sim}{e}$; $= (1, 1, \ldots, 1)$ and $\leqslant$ means componentwise inequality.

Now, $E(n\underset{\sim}{e} - S_n) = \sum^{n} F_i(x/\sqrt{n})$

$$\sum_{1}^{n} F_i(\underset{\sim}{O}) + (1/n)(b^* f_i(\underset{\sim}{O}) + \underset{\sim}{o}(n^{½})$$

where $a^*b = (a_1 b_1, \ldots, a_p b_p)$, and $\underset{\sim}{O}$ denotes the p dimensional null vector.

$$\operatorname{cov}(S_n) = \sum_{1}^{n} \Omega_i\,({}^{c}/\sqrt{n}\,)$$

$$\approx \sum_{1}^{n} \Omega_i(\underset{\sim}{O})$$

To derive the asymptotic distribution of $\tilde{x}$ we make the following assumptions as $n \to \infty$:

Assumption 1. $\frac{1}{n}\sum_{1}^{n} f_i(\underset{\sim}{O}) \to f > \underset{\sim}{O}$, where f is finite.

Assumption 2. $\frac{1}{n}\sum_{i=1}^{n} \Omega_i(\underset{\sim}{O}) \to \Omega$, a non-null matrix

Assumption 3. $\frac{1}{\sqrt{n}}(\sum_{1}^{n} F_i(\underset{\sim}{O}) - \frac{n}{2}\, e) \to O$, the zero vector.

By the multivariate central limit theorem (Rao (1965), Exercise 2 (4.7)) we have:

$$\frac{S_n - E(S_n)}{\sqrt{n}} \xrightarrow{L} N(\underset{\sim}{O}, \Omega)$$

under Assumption 2.

Now:

$$P(\sqrt{n}\,\tilde{x} \leqslant c) = P(X_n \leqslant \frac{n-1}{2}\underset{\sim}{e})$$

$$= P(\frac{S_n - E(S_n)}{\sqrt{n}} \leqslant -\frac{n+1}{2\sqrt{n}} + \frac{1}{\sqrt{n}}\sum_{1}^{n} F_i(\underset{\sim}{O}) + \frac{1}{n}\sum_{1}^{n} c * f_i)$$

$$= P(\frac{S_n - E(S_n)}{\sqrt{n}} \leqslant c * f) \qquad (11)$$

from Assumption 2 and 3. From expressions (8) and (9), it follows that:

Theorem 1 $\sqrt{n}\ \widetilde{x}\ _{*}\ f$ is asymptotically normal with mean $\underset{\sim}{O}$ and covariance Ω, under Assumptions 1, 2 and 3.

Remark 1 The asymptotic distr̃ibution of the sample median in the *iid* case is a special case of this result when $p=1$. See Lehmann (1983).

Remark 2 Mood (1941) derived the asymptotic joint distribution of the sample medians from a multivariate population F. The result in that paper can also be seen as a special case of the above theorem by letting

$$F_1 = F_2 = \ldots = F_n = F.$$

Now consider, $\left\{ \hat{\beta}_i \right\}$ the sequence of least-squares estimates computed according to the scheme in Section 2. As before, let b_{ij} denote the jth component of the ith least-squares estimate $i = 1, 2, \ldots, m,\ j = 1, 2, \ldots, p,\ n = m,\ k$. Let H_{ij} and h_{ij} denote the *cdf* and *pdf* of b_{ij} respectively. Let

$$H_{ijk}\ (C_j, C_k) = P(b_{ij} \leqslant c_j,\ \hat{b}_{ik} \leqslant c_k)$$

the joint *cdf* of any two components of $\hat{\beta}_I$. In Theorem 1, replace each F_i by H_i, each f_i by h_i and so on wherever necessary. We obtain the following main result on the asymptotic distribution of the median estimate β^* given by (6).

Corollary 1 $\sqrt{m}\ \beta^*{}_*\ h$ is symptotically normal with mean β and covariance Ω under Assumptions 1, 2, and 3 where:

$$h = \lim_{m\to\infty} \frac{1}{m} \sum_{1}^{m} h_i(\underset{\sim}{O}) \tag{12}$$

$$\Omega = \lim_{m\to\infty} \frac{1}{m} \sum_{1}^{m} \Omega_i(\underset{\sim}{O}) \tag{13}$$

where $\Omega_i\ (\underset{\sim}{O})$ is defined by equations 8 and 9 replacing F_{ij} by H_{ij} and F_{ijk} by H_{ijk} and so on.

Of special interest is the case $p = 2$, which includes among others, the estimate a^* of Maritz (1979). To this end let $k = 2$ and

$$h_{ij} = \text{density of } X_i\epsilon_j - X_j\epsilon_i \text{ for some } (i, j) \text{ in } R$$

$$g_{ij} = \text{density of } \epsilon_i - \epsilon_j \text{ for some } (i, j) \text{ in } R.$$

Marginally, we obtain the following corollary:

Corollary 2
$$\sqrt{m}\,(a^* - a) \longrightarrow N(0, \frac{1}{4\,h^2\,(0)})$$
$$\sqrt{m}\,(b^* - b) \longrightarrow N(0, \frac{1}{4\,g^2\,(0)})$$

where

$$h\,(0) = \lim_{m \to \infty} \frac{\sum_R \mid x_j - x_i| h_{ij}\,(0)}{m}$$

$$g\,(0) = \lim_{m \to \infty} \quad f^2\,(u)\,du\, \Sigma\; |x_j - x_i| m$$

with

$$f^2\,(u)\,du < \infty$$

As in ordinary least-squares, different objectives may yield different choice of blocking R. Thus, in ordinary least-squares, when the objective is to minimize the variance of the least-squares estimate of a, then the x's are chosen so that $\bar{x} = 0$. On the other hand, if the objective is to minimize the variance of the least-squares estimate of b, then $\sum_{i=1}^{n} (x_i - \bar{x})^2$ is made as large as possible.

In the above corollary, if we wish to minimize the variance of b^*, then we need to maximize $\sum_R |x_j - x_i|/m$. This can be achieved if large values of x are paired with small values of x. It is not clear how the blocking should be done to minimize the variance of a^*.

In the case that the x's are randomly generated independent of the ϵ's, then the blocking technique used for estimating b will yield an optimal blocking structure for estimating a as well in the sense of minimizing the asymptotic variances.

Monte Carlo

We performed a simple Monte Carlo experiment to determine the performance of the proposed estimate for finite sample sizes relative to the usual least-squares estimates of β in model (1).

We took as our model the case p=3 given by:

$$y_j = 2.5 + 3.5x_{1j} - 95x_{2j} + \epsilon_j$$

The sample sizes are n=18 and n=30 to represent small and large sample sizes respectively. The values of x_{1j} and x_{2j} are independently generated from a uniform distribution on the interval [–20,20]. The distribution of the errors, ϵ_j, is given:

$$F = (1-\alpha)\,N(0,1) + \alpha\,N(0.64) \quad . \quad 0\leqslant\alpha<1$$

that is, a contaminated normal distribution contaminated by a normal distribution with a standard deviation of 8. The proportions of contamination, α, are 0.00, 0.01, 0.05 and 0.10.

Both the least-squares estimates of β and the median estimates of β are computed over 1,000 replications. We then estimated the mean-squared errors (MSE) of both these estimates and compared them. The results are given in Table 1.

In this table, the ratios given are:

$$\text{Ratio:} \qquad \frac{\text{MSE Estimate for the Least-Squares}}{\text{MSE Estimate for the Median}}$$

over 1,000 replications. Hence, a ratio less than 1 indicates that the least-squares estimate performed better than the median estimate, otherwise the reverse is true.

Table 1. MSE ratios of the least-squares relative to the median estimates

	Sample Size	
Proportion of Contamination	*18*	*30*
0.00	0.6148	0.4374
0.01	1.3921	1.2452
0.05	1.4125	1.3324
0.10	1.2767	1.2037

Notice that even with the slightest amount of contamination, namely just 1%, the median estimate already out performs the least-squares.

Conclusion

The proposed estimates of multiple regression coefficients are very simple to calculate and are therefore fairly attractive to the average user of statistics. The estimates provide good protection for the effect of corrupt observations in the data.

Finally, we might mention that simple non-parametric tests of hypotheses can be made utilizing the given procedures.

References

Adichie, J. 1969. Estimates of regression coefficients based on Rank Tests. *Annals of Mathematical Statistics* **38**:894-904.

Huber, P.J., 1973. Robust regression: asymptotics, conjectures and Monte-Carlo. *Annals of Statistics* **1**:799-821.

Jureckova, J. 1971. Non-parametric estimates of regression coefficients. *Annals of Mathematical Statistics* Vol. 42.

Koenker, R. and G. Bassett. 1978. Regression quantities. *Econometrical* **46**:33-48.

Maritz, J. 1977. On Theil's method in distribution-free regression. *Australian Journal of Statistics* **21**:30-35.

Sen, P.K. 1968. Estimates of regression coefficients based on Kendall's Tan. *Journal of the American Statistical Ass'n.* **63**: 1379-1389.