ON SOME ROBUST PROPERTIES OF ESTIMATES OF REGRESSION BASED ON RANK TESTS

J. N. ADICHIE

(Received Jan. 20, 1972; revised July 24, 1972)

Introduction and summary

In the linear regression model $Y_j=\alpha+\beta x_j+Z_j$ many point estimates of α and β, other than the classical least squares estimates, have been proposed see [8], [12], and [1] and variants of [12] proposed in [3], [11] and [9] among others. All these estimates are generally called "robust", a description which has a rather vague interpretation. In a recent paper, Huber [7], has drawn attention to the various forms that "robustness" can take and to the necessity for specifying what aspect of robustness one is discussing.

In this paper, we shall study the asymptotic behaviour, in the sense of the relative efficiency, of the "rank score" estimates defined by Adichie [1], under small variations of, and departures from the underlying distribution of the observations. The main results are presented in Section 2.

1. Preliminaries

Let $Y_1, Y_2, \cdots, Y_n$ be independent random variables with distributions

$$\mathrm{P}_\beta[Y_j\leqq y]=G_j(y)=F_j(y-\beta x_j) \qquad j=1,\cdots,n \tag{1.1}$$

where P_β denotes the probability computed for the parameter value β, and F_j's belong to a class of continuous distribution functions, with bounded derivatives. "Rank Score" estimates $\hat{\beta}(\phi)$ of β can be defined as in [1] without extra conditions on the F_j's. Let

$$\hat{\beta}(\phi)=\frac{1}{2}(\beta^*+\beta^{**}) \tag{1.2}$$

where

$$\begin{aligned}&\beta^*=\inf\{b: T_n(Y-bx)<0\} \quad \text{and}\\ &\beta^{**}=\sup\{b: T_n(Y-bx)>0\}\end{aligned} \tag{1.3}$$

with

$$T_n(Y)=\frac{1}{n}\sum_j (x_j-\bar{x})\phi\left(\frac{R_j}{n+1}\right). \tag{1.4}$$

R_j is the rank of Y_j, while ϕ is a smooth non decreasing weight function (score function) defined on $[0, 1]$; $T_n(Y-bx)$ denotes the statistics (1.4) when the observations Y_j are replaced by Y_j-bx_j, where b is a real number.

Assume that the regression constants satisfy the boundedness condition:

$$\lim_n [n^{-1}\sum (x_j-\bar{x})^2]<\infty\,, \qquad \sum_j x_j^2 \leqq M\sum_j (x_j-\bar{x})^2 \tag{1.5a}$$

for some M and Noether condition

$$\lim_n \max_j \left[\frac{(x_j-\bar{x})^2}{\sum_j (x_j-\bar{x})^2}\right]=0\,. \tag{1.5b}$$

While the distribution functions are such that

$$\max_j |F_j(y)-H_n(y)|\leqq w(y) \tag{1.6}$$

where

$$\left|\int w(y)\phi'(H_n(y))dH_n(y)\right|<\left[\sum_j (x_j-\bar{x})^2\right]^{-1}$$

and

$$H_n(y)=\frac{1}{n}\sum_j F_j(y)\,.$$

Also let $H_n(y)\to H(y)$ a distribution function such that

$$|H_n(y)-H(y)|<\frac{K(y)}{n} \tag{1.7}$$

where

$$\int K(y)\phi'(H(y))dH(y)<\infty\,.$$

Observe that Assumption (1.6) is fundamental as it expresses the "smallness" of the variations of F_j's from some underlying distribution H. In [5], Hajek expressed a similar condition by

$$\max_{i,j,y} |F_i(y)-F_j(y)|<\delta$$

where also

$$\max_j \left[\frac{(x_j-\bar{x})^2}{\sum_j (x_j-\bar{x})^2}\right] < \delta .$$

2. The main results

THEOREM 2.1. *Let* $\hat{\beta}(\psi)$ *be as defined in* (1.2), *and let*

$$|\psi''(u)| < c . \tag{2.1}$$

Then under (1.5), (1.6) *and* (1.7),

$$\lim_n \mathrm{P}\left[\sqrt{n}\,(\hat{\beta}(\psi)-\beta) \leqq b\right] = \Phi\left(\frac{bB}{A}\right) \tag{2.2}$$

with

$$B_n = \frac{1}{n}\sum_j (x_j-\bar{x})^2 \int \frac{d}{dy}\psi(H(y))dH(y) ; \qquad B = \lim_n B_n \tag{2.3}$$

$$A_n^2 = \frac{1}{n}\sum_j (x_j-\bar{x})^2 \left[\int \psi^2(u)du - \left(\int \psi(u)du\right)^2\right] ; \qquad A^2 = \lim_n A_n^2 \tag{2.4}$$

and Φ *stands for the distribution function of the standard Normal.*

PROOF. Hajek (Theorem 4.2 of [5]) has in fact proved that under the assumptions (1.4), (1.5a) and (2.1),

$$\lim_n \mathrm{P}\left[\sqrt{n}\,(T_n-\mu_n) \leqq y\tau_n\right] = \Phi(y) \tag{2.5}$$

where

$$\mu_n = \frac{1}{n}\sum_j (x_j-\bar{x}) \int \psi(\bar{G}(y))dG_j(y) \tag{2.6}$$

and

$$\tau_n^2 = 2n^{-3}\sum_i \sum_j (x_j-x_i)^2 \iint_{y<z} G_i(y)(1-G_i(z))\psi'(\bar{G}(y))\psi'(\bar{G}(z))dG_j(y)dG_j(z) \tag{2.7}$$

with

$$\bar{G}(y) = n^{-1}\sum_j G_j(y) .$$

To prove our theorem we need to find the limiting distribution

$$\mathrm{P}_n\left[\frac{\sqrt{n}\,(T_n-\mu_n)}{\tau_n} \leqq y\right]$$

where P_n denotes the probability computed for the parameter value

$$\Delta_n = -bn^{-1/2} .$$

On writing

$$G_i = H_n + (G_i - F_i) + (F_i - H_n)$$

and noting that under Δ_n,

$$\bar{G}(y) = H_n(y) - \frac{b \sum x_j f_j(y+\theta)}{\sqrt{n^3}}; \quad \text{and} \quad G_i(y) - F_i(y) = -\frac{b x_i f_i(y+\theta)}{\sqrt{n}},$$

we have that under Δ_n

$$\int G_i \phi'(\bar{G}) dG_j = \int H_n \phi'(H_n) dH_n + C_{1i} + C_{1j} + C_{2ij} + D_{1i} + D_{1j} + D_{2ij} + D_{3ij} + D_{4ij} + O(n^{-1})$$

where

$$C_{1i} = \int (F_i - H_n) \phi'(H_n) dH_n$$

$$C_{1j} = \int H_n \phi'(H_n) d(F_j - H_n)$$

$$C_{2ij} = \int (F_i - H_n) \phi'(H_n) d(F_j - H_n)$$

$$D_{1i} = \int (G_i - F_i) \phi'(H_n) dH_n$$

$$D_{1j} = \int H_n \phi'(H_n) d(G_j - F_j)$$

$$D_{2ij} = \int (G_i - F_i) \phi'(H_n) d(F_j - H_n)$$

$$D_{3ij} = \int (G_i - F_i) \phi'(H_n) d(G_j - F_j)$$

$$D_{4ij} = \int (F_i - H_n) \phi'(H_n) d(G_j - F_j).$$

Integrating the C and D terms by parts and making use of (1.5a) and (1.5b) on the D terms, and (1.6) on the C terms, we get that under Δ_n,

$$\int G_i \phi'(\bar{G}) dG_j = \int H_n \phi'(H_n) dH_n + R_{nij}$$

where

$$|R_{nij}| < M \left[\sum_j (x_j - \bar{x})^2 \right]^{-1}.$$

It follows then that

$$(2.8)\quad \tau_n^2(\Delta_n)=2n^{-3}\sum_i\sum_j (x_i-x_j)^2 \cdot\iint_{y<z} H_n(y)(1-H_n(y))\phi'(H_n(y))\phi'(H_n(z))dH_n(y)dH_n(z)+O(n^{-1})$$
$$=n^{-1}\sum_j (x_j-\bar{x})^2 2\iint_{u<v} u(1-v)\phi'(u)\phi'(v)du\,dv+O(n^{-1})$$
$$=\left[\int \phi^2(u)du-\left(\int \phi(u)du\right)^2\right]n^{-1}\sum_j (x_j-\bar{x})^2+O(n^{-1})$$
$$=A_n^2+O(n^{-1}).$$

Now under Δ_n, the expression $\sqrt{n}(T_n-\mu_n)$ can be written as

$$\sqrt{n}(T_n(\Delta_n)-\mu_n(0))-\sqrt{n}(\mu_n(\Delta_n)-\mu_n(0))$$

where

$$\sqrt{n}(\mu_n(\Delta_n)-\mu_n(0))=n^{-1/2}\sum (x_j-\bar{x})\int \phi(H_n(y))d(G_j(y)-F_j(y))+O(n^{-1/2})$$

and on integrating by parts, this reduces to

$$n^{-1}\sum_j (x_j-\bar{x})^2 b\int f_j(y)\phi'(H_n(y))dH_n(y)+O(n^{-1/2}).$$

Furthermore, on writing

$$f_j(y)=h(y)+(f_j(y)-h(y)) \quad \text{where} \quad \frac{d}{dy}H(y)=h(y)$$

and making use of (1.6) and (1.7), it is found that under Δ_n,

$$\lim_n \sqrt{n}(\mu_n(\Delta_n)-\mu_n(0))=\lim_n bn^{-1}\sum_j (x_j-\bar{x})^2\int \frac{d}{dy}\phi(H(y))dH(y)=bB.$$

On making use of Slutsky's Theorem, it then follows that under Δ_n,

$$\lim_n \mathrm{P}_n[\sqrt{n}(T_n-\mu_n)\leqq y]=\Phi\left(\frac{y+bB}{A}\right)$$

where A and B are given in (2.3) and (2.4). The rest of the proof follows from the result of Hodges and Lehmann (Theorem 4 of [6]).

The asymptotic variance of the estimates $\hat{\beta}$ is thus deduced from Theorem 2.1, and is given by

$$(2.9)\quad \text{Asymptotic Var}\,(\sqrt{n}\,\hat{\beta})$$
$$=\frac{A^2}{B^2}=\int \phi^2(u)-\left(\int \phi(u)du\right)^2\Big/\left\{\int \frac{d}{dy}\phi(H(y))dH(y)\right\}^2.$$

The least squares estimate $\hat{\beta}$ is obtained from (1.2) and (1.3) but with

$$T'_n(Y)=n^{-1}\sum_j (x_j-\bar{x})Y_j$$

for details see [1]. It is also known that under very general conditions, (see e.g. Eicker [4]),

$$\lim_n \mathrm{P}\left[\frac{\sqrt{n}\,(T'_n-\nu_n)}{\tau'_n}\leqq y\right]=\Phi(y)$$

where

$$\nu_n=n^{-1}\sum_j (x_j-\bar{x})\int y dG_j(y)\,; \quad \text{and} \quad (\sqrt{n}\,\tau'_n)^2=n^{-1}\sum_j (x_j-\bar{x})^2 \operatorname{Var} Y_j\,,$$

and in particular

$$\lim_n \mathrm{P}_n[\sqrt{n}\,(T'_n-\nu_n)\leqq y]=\left(\frac{y+bB'}{A'}\right) \tag{2.10}$$

where

$$\begin{aligned} B' &=\lim_n n^{-1}\sum_j (x_j-\bar{x})^2\,; \quad \text{and} \\ A'^2&=\lim_n n^{-1}\sum_j (x_j-\bar{x})^2[\bar{\sigma}_n^2+(\sigma_j^2-\bar{\sigma}_n^2)] \end{aligned} \tag{2.11}$$

with

$$\sigma_j^2=\int y^2 dF_j(y)-\left(\int y dF_j(y)\right)^2$$

and

$$\bar{\sigma}_n^2=\int y^2 dH_n(y)-\left(\int y dH_n(y)\right)^2. \tag{2.12}$$

On using (1.6) on $(\sigma_j^2-\bar{\sigma}_n^2)$, we obtain

$$A'^2_n=n^{-1}\sum_j (x_j-\bar{x})^2\bar{\sigma}_n^2+O(n^{-1})\,, \qquad \lim_n A'^2_n=A'^2\,. \tag{2.13}$$

From (2.10), it follows that under the conditions of Theorem 2.1, the limiting distribution of the least squares estimate is given by

$$\lim_n \mathrm{P}_\beta[\sqrt{n}\,(\tilde{\beta}-\beta)\leqq b]=\Phi\left(\frac{bB'}{A'}\right) \tag{2.14}$$

where B' and A' are given by (2.11) and (2.13). From (2.14), it is clear that the asymptotic variance of the least squares estimate $\tilde{\beta}$, is given by

$$\text{Asymptotic Var}\,(\sqrt{n}\,\tilde{\beta})=\bar{\sigma}^2=\lim_n \bar{\sigma}_n^2\,. \tag{2.15}$$

We have therefore obtained the following result:

THEOREM 2.2. *Under the assumptions of Theorem* 2.1, *the asymptotic efficiency of the "rank score" estimates relative to the least squares estimates is given by*

$$(2.16)\quad e\hat{\beta},\ \tilde{\beta}(\phi(H))=\bar{\sigma}_n^2\left\{\int \frac{d}{dy}\phi(H(y))dH(y)\right\}^2\Big/\left[\int \phi^2(u)du-\left(\int \phi(u)du\right)^2\right].$$

In his study of the one sample problem, Sen [10] obtained the same result for the special case $\phi(u)=u$, and showed that for a class $\mathcal{F}_0$ of distribution functions containing the normal and the exponential, among others,

$$(2.17)\qquad e\hat{\beta},\ \tilde{\beta}(H)\geqq e\hat{\beta},\ \tilde{\beta}(F)\,.$$

It is easy to see that the same result holds for the efficiency expression (2.16). More precisely one can show following Sen, that for convex $\phi(u)$ or for $\phi(u)$ defined through a strongly unimodal distribution,

$$(2.18)\qquad e\hat{\beta},\ \tilde{\beta}(\phi(H))\geqq e\hat{\beta},\ \tilde{\beta}(\phi(F))$$

where the right-hand side of the inequalities (2.17) and (2.18) are efficiencies computed on the assumption that $F_1=F_2=\cdots=F_n=F$.

Remark: Observe that neither in the definition nor in the limiting distribution of $\hat{\beta}$ is the symmetry of F_j's necessary. Its use is, however, in the small sample property of unbiasedness of $\hat{\beta}$ (see Lemma 4.2 of [1]).

In the next two sections, we give a few examples of the common models of departures from the underlying distributions.

3. Robustness against shifts

Model I. Let $F_j(y)=F(y-\alpha_j)$.

This could happen, for example if the outcome Y_j of the observations, taken under different experimental conditions, depends not only on the x_j's but also on the time α_j when taking the jth observation. In this set up, the estimates of β naturally depend on the unknown α_j. Neither the least squares nor the rank scores estimates are unbiased. In fact,

$$\mathrm{E}\,(\tilde{\beta})=\beta+\frac{\sum_j (x_j-\bar{x})\alpha_j}{\sum_j (x_j-\bar{x})^2}$$

and the estimates of α_j are not available, unless we take replicated observations for each value of x_j. If, however, the α_j's differ only slight-

ly, then the efficiency expression (2.16) is attained with

$$\bar{\sigma}_n^2=\sigma^2=\text{var}\,(Y_j)\,.$$

To see this, write

$$|F_j(y)-H_n(y)|=\left|(\bar{\alpha}-\alpha_j)f(y)+\frac{1}{2}\left(\alpha_j^2-\frac{1}{n}\sum_j\alpha_j^2\right)f'(y+\theta)\right|$$

$$\leqq|\bar{\alpha}-\alpha_j|f(y)+\frac{1}{2}f'(y+\theta)|(\alpha_j^2-\bar{\alpha}^2)|\,.$$

Suppose $\alpha_j=j^{-2}$, then

$$\max_j|\bar{\alpha}-\alpha_j|=\left|\frac{1}{n}\left(\sum_k\frac{1}{k^2}-\frac{1}{n}\right)\right|\leqq\left|n^{-1}\sum_k\frac{1}{k^2}\right|$$

$$\leqq\left|n^{-1}\sum_k(k(k+1))^{-2}\right|=n^{-1}[1-(n+1)^{-1}]$$

$$=O(n^{-1})$$

and the rest follows.

Model II. The case of non linear regression.

In Model I, put $\alpha_j=\gamma x_j^2$ and obtain the polynomial regression model. Ordinarily one estimates both β and γ by the method of least squares. The rank score estimates of γ are also available, for example, through the use of the test statistic defined in [2]. For this model to come within our frame work, γ will be small; for example, if

$$\max_j|nx_j^2-\sum x_i^2|\leqq\gamma^{-1}$$

then assumption (1.6) is satisfied.

4. Robustness against changes in scale

Model III. Let $F_j(y)=F(y/\sigma_j)$.

This corresponds to the important case of non uniformity of experimental conditions, where the effects of such non uniformity are multiplicative rather than additive (Model I). In this model,

$$|F_j(y)-H_n(y)|=\left|yf(\theta)\left[\sigma_j^{-1}-n^{-1}\sum_j\sigma_j^{-1}\right]+\frac{1}{2}y^2f'(\theta)\left[\sigma_j^{-2}-n^{-1}\sum_j\sigma_j^{-2}\right]\right|$$

$$\leqq\left|\left[\sigma_j^{-1}-n^{-1}\sum_j\sigma_j^{-1}\right]\left[yf(\theta)+\frac{1}{2}y^2f'(\theta)\left(\sigma_j^{-2}+n^{-1}\sum_j\sigma_j^{-2}\right)\right]\right|.$$

If we put either $\sigma_j=\sigma+\epsilon_j$ or $\sigma_j=\sigma\epsilon_j^{-1}$ where ϵ_j is small, such that $|(\epsilon_j-\bar{\epsilon})|=O(n^{-1})$, then it can easily be shown that assumption (1.6) is satisfied:

Model IV. In the Gross error model

$$F_j(y)=(1-\epsilon_j)F(y)+\epsilon_j Q(H)$$

if $|(\epsilon_j-\bar{\epsilon})|=O(n^{-1})$, (1.6) is also satisfied.

Acknowledgement

The author is grateful to the referee for indicating several points which were not clear in the first version of this paper.

UNIVERSITY OF NIGERIA

REFERENCES

[1] Adichie, J. N. (1967). Estimates of regression based on rank tests, *Ann. Math. Statist.*, **38**, 894-904.
[2] Adichie, J. N. (1971). Rank score tests for linearity of regression, Submitted to *Ann. Math. Statist.*
[3] Bhattacharyya, G. K. (1968). Robust estimates of linear trend in multivariate time series, *Ann. Inst. Statist. Math.*, **20**, 299-310.
[4] Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions, *Ann. Math. Statist.*, **34**, 447-456.
[5] Hajek, J. (1968). Asymptotic normality of simple linear rank statistics under alternatives, *Ann. Math. Statist.*, **39**, 325-346.
[6] Hodges, J. L., Jr., and Lehmann, E. L. (1963). Estimates of location based on rank tests, *Ann. Math. Statist.*, **34**, 598-611.
[7] Huber, P. J. (1972). Robust Statistics, *Ann. Math. Statist.*, **43**, 1041-1067.
[8] Mood, A. M. (1950). *Introduction to the theory of Statistics*, MacGraw Hill, N.Y.
[9] Rao, P. V. and Thornby, J. I. (1969). A robust point estimate in a generalised regression model, *Ann. Math. Statist.*, **40**, 1784-1790.
[10] Sen, P. K. (1968). On a further robust property of the test and estimator based on Wilcoxon signed rank statistic, *Ann. Math. Statist.*, **39**, 282-285.
[11] Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau, *Journal Amer. Stat. Assoc.*, **63**, 1379-1389.
[12] Theil, H. (1950). A rank invariant method of linear and polynomial regression analysis, *Indag. Math.*, **12** Fasc. **2**, 85-91, 173-177.