

ROBUSTNESS OF SOME NONPARAMETRIC PROCEDURES
IN LINEAR MODELS

by

Pranab Kumar Sen

University of North Carolina

Institute of Statistics Mimeo Series No. 547

August 1967

Work supported by the Army Research Office,
Durham, Grant DA-ARO-D-31-124-G746.

DEPARTMENT OF BIostatISTICS

UNIVERSITY OF NORTH CAROLINA

Chapel Hill, N. C.

ROBUSTNESS OF SOME NONPARAMETRIC PROCEDURES IN LINEAR MODELS*

by

Pranab Kumar Sen

University of North Carolina, Chapel Hill

1. Introduction and summary. For the random variables X_{ij} ($i=1, \dots, N$; $j=1, \dots, r$) consider the linear model

$$(1.1) \quad X_{ij} = \mu + \beta_i + \tau_j + Y_{ij} \quad (\sum \beta_i = 0, \sum \tau_j = 0),$$

where τ 's are treatment effects, β 's are nuisance parameters (block effects) and Y_{ij} 's are error components. Nonparametric procedures for estimating and testing contrasts in τ 's, based on the Wilcoxon signed rank statistics, are due to Lehmann (1964) and Doksum (1967), among others, and they rest on the assumption that Y_{ij} 's are independent with a common continuous distribution.

The object of the present paper is to show that these procedures are valid even when Y_{i1}, \dots, Y_{ir} are interchangeable variables, for each $i(=1, \dots, N)$, and moreover, they are robust against possible heterogeneity of the distributions of the error vectors $Y_{\sim i} = (Y_{i1}, \dots, Y_{ir})$, $i=1, \dots, N$.

2. Some fundamental lemmas. Define $\Delta_{jk} = \tau_j - \tau_k$, $j \neq k=1, \dots, r$, and let

$$(2.1) \quad X_{ijk}^* = X_{ij} - X_{ik}, \quad U_{ijk} = Y_{ij} - Y_{ik} \quad \text{for } j \neq k=1, \dots, r \text{ and } i=1, \dots, N.$$

Assume that $Y_{\sim i}$ has a continuous r -variate cumulative distribution function (cdf) $F_{\sim i}(x)$ which is symmetric in its r arguments, for all $i=1, \dots, N$. This interchangeability of Y_{i1}, \dots, Y_{ir} implies that (i) the cdf $G_{\sim i}(x)$ of U_{ijk} is independent of $j \neq k(=1, \dots, r)$ and is symmetric about 0, and (ii) the bivariate

* Work supported by the Army Research Office, Durham, Grant DA-ARO-D-31-124-G746.

cdf $G_i^*(x,y)$ of $U_{ijk}, U_{ijk'}$ is independent of $j \neq k \neq k' (=1, \dots, r)$, for all $i=1, \dots, N$. Let $h(x)$ be a real valued skew-symmetric function, i.e.,

$$(2.2) \quad h(x) + h(-x) = 0 \text{ for all } x.$$

Define

$$(2.3) \quad \zeta_{i,0} = E\{h(U_{ijk})h(U_{ij'k'})\}, j \neq j' \neq k \neq k',$$

$$(2.4) \quad \zeta_{i,1} = E\{h(U_{ijk})h(U_{ijk'})\} \text{ and } \zeta_{i,2} = E\{h^2(U_{ijk})\}, j \neq k \neq k'.$$

Then, we have the following.

LEMMA 2.1. If (i) Y_{i1}, \dots, Y_{ir} are interchangeable random variables and (ii) $h(x)$ satisfies (2.2), then (i) $E\{h(U_{ijk})\} = 0$ and (ii) $\zeta_{i,0} = 0$.

PROOF. (i) follows trivially from the fact that G_i is symmetric about 0 and $h(x)$ satisfies (2.2). To prove (ii), denote by $t_{\sim} = \{t_1 \leq t_2 \leq t_3 \leq t_4\}$ the order statistics corresponding to $Y_{ij}, Y_{ij'}, Y_{ik}$ and $Y_{ik'}$. Since Y_{ij} 's are interchangeable variables, the conditional distribution of $Y_{ij}, Y_{ij'}, Y_{ik}, Y_{ik'}$, given t_{\sim} , will be uniform on the 24 equally likely permutations of t_1, t_2, t_3 and t_4 . Further, $h(x)$ satisfies (2.2). Hence, it is easy to show that for $j \neq k = j' \neq k'$,

$$(2.5) \quad E\{h(U_{ijk})h(U_{ij'k'}) | t_{\sim}\} = E\{h(Y_{ij} - Y_{ik})h(Y_{ij'} - Y_{ik'}) | t_{\sim}\} = 0.$$

Thus, writing $\zeta_{i,0}$ equivalently as

$$E_{t_{\sim}}\{E[h(U_{ijk})h(U_{ij'k'}) | t_{\sim}]\},$$

the proof follows from (2.5), Q.E.D.

LEMMA 2.2. If Y_{i1}, \dots, Y_{ir} are interchangeable variables and $h(x)$ satisfies (2.2), then $\zeta_{i,1} \leq \frac{1}{2}\zeta_{i,2}$, where the equality sign holds only when $h(x) = bx$,

for all x . If, in addition, $h(x)$ is monotonic, $0 \leq \zeta_{i,1} \leq \frac{1}{2}\zeta_{i,2}$.

PROOF. Define $Z_i = h(U_{i12}) + h(U_{i23}) + h(U_{i31})$. Then, by (2.3), (2.4) and lemma 2.1, we obtain that

$$(2.6) \quad V(Z_i) = 3\zeta_{i,2}(1-2\zeta_{i,1}/\zeta_{i,2}) \geq 0,$$

where the equality sign holds only when $Z_i \equiv 0$ a.e.. Now (2.6) implies that $\zeta_{i,1} \leq (\frac{1}{2})\zeta_{i,2}$. Also, by definition $U_{i12} + U_{i23} + U_{i31} = 0$. Hence, $Z_i \equiv 0$ a.e., along with (2.2) implies that

$$(2.7) \quad h(U_{i12}) + h(U_{i23}) = h(U_{i12} + U_{i23}),$$

for all U_{i12}, U_{i23} . (2.7) in turn implies that $h(x) = bx$ for all x , and this completes the first part of the proof. Let now $t_{\sim} = \{t_1 \leq t_2 \leq t_3\}$ be the order statistics corresponding to Y_{ij}, Y_{ik} and $Y_{ik'}$. Using then (2.2) and proceeding as in the proof of lemma 2.1, one obtains that

$$(2.8) \quad \begin{aligned} E\{h(U_{ijk})h(U_{ijk'})|t_{\sim}\} &= E\{h(Y_{ij}-Y_{ik})h(Y_{ij}-Y_{ik'})|t_{\sim}\} \\ &= \frac{1}{3}[h(t_1-t_2)h(t_1-t_3) + h(t_3-t_1)h(t_3-t_2) - h(t_2-t_1)h(t_3-t_2)]. \end{aligned}$$

Assume that $h(x)$ is \uparrow in x (as otherwise, work with $-h(x)$). Then,

$$(2.9) \quad 0 \leq h(t_2-t_1) \leq h(t_3-t_1) \quad \text{and} \quad 0 \leq h(t_3-t_2) \leq h(t_3-t_1).$$

By (2.9), the left hand side of (2.8) is essentially non-negative, and integrating over the distribution of t_{\sim} , it follows that $\zeta_{i,1} \geq 0$. Q.E.D.

Let now

$$(2.10) \quad \lambda(F_1) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} G_1(x)G_1(y)dG_1^*(x,y) \quad \text{for } i=1, \dots, N.$$

LEMMA 2.3. If Y_{i1}, \dots, Y_{ir} are interchangeable random variables, $\frac{1}{4} \leq \lambda(F_1) \leq 7/24$,

where the upper bound $7/24$ is attained only when G_i is a uniform cdf over $(-a, a), a > 0$.

PROOF. We let $h(x) = G_i(x) - \frac{1}{2}$. As G_i is symmetric about 0, (2.2) is satisfied. Straightforward computations yield that $\zeta_{i,2} = 1/12$ and $\zeta_{i,1} = \lambda(F_i) - \frac{1}{4}$. Also $G_i(x)$ is \uparrow in x . Hence, the lemma directly follows from lemma 2.2. Q.E.D.

REMARK. Lemma 2.3 generalizes theorem 2 of Lehmann (1964) to exchangeable random variables and also supplies condition under which the upper bound $7/24$ for $\lambda(F_i)$ may be attained.¹

3. Robustness for interchangeable errors. It may be noted that if $V(Y_{ij}) = \sigma^2$ and $Cov(Y_{ij}, Y_{ik}) = \rho\sigma^2, j \neq k$, the classical ANOVA-test based on the variance-ratio criterion is a valid test for the null hypothesis $H_0: \tau_1 = \dots = \tau_r = 0$, and for the sequence of alternative hypotheses $\{H_N\}$:

$$(3.1) \quad H_N: \Delta_{jk} = N^{-\frac{1}{2}}a_{jk}; a_{jk} = a_j - a_k, 1 \leq j < k \leq r, \sum_{j=1}^r a_j = 0,$$

(where a_1, \dots, a_r are all real and finite), $(r-1)$ times the variance-ratio criterion has asymptotically (under $F_1 = \dots = F_N = F$) a non-central chi-square distribution with $r-1$ degrees of freedom and non-centrality parameter

$$(3.2) \quad \sum_{j=1}^r a_j^2 / \sigma^2 (1-\rho).$$

It is also known that the method of ranking after alignment [cf. Hodges and Lehmann 1962)] allows for the interchangeability of the error components, and it has been shown by Sen (1967a) that the efficiency of the non-parametric procedures based on aligned observations is not affected by the interchangeability

1) The author is grateful to Professor Wassily Hoeffding for pointing out that the existence of F_i for which the corresponding G_i is uniform on $(-a, a), a > 0$, is dubious. Our conjecture is that there exists no such cdf F_i for which the corresponding G_i is uniform on $(-a, a), a > 0$.

of the errors. It will be shown here that the same is true for the procedures considered by Lehmann (1964) and Doksum (1967). For this, define $\phi(u)$ as 1 or 0 according as u is > 0 or not, and let

$$(3.3) \quad W_{N,jk} = \binom{N}{2}^{-1} \sum_{1 \leq i < i' \leq N} \phi(X_{ijk}^* + X_{i'jk}^*), \quad 1 \leq j < k \leq r.$$

Assume that $F_1 = \dots = F_N = F$ ($\implies G_1 = \dots = G_N = G$ and $\lambda(F_1) = \dots = \lambda(F_N) = \lambda(F)$), and that $G(x)$ has a continuous density function $g(x)$ satisfying the conditions of theorem 1 of Lehmann (1964). Then, it is easy to show that

$$(3.4) \quad \lim_{N \rightarrow \infty} \{N^{1/2} E[W_{N,jk}^{-1/2} | H_N]\} = 2a_{jk} \int_{-\infty}^{\infty} g^2(x) dx, \quad \text{for all } 1 \leq j < k \leq r.$$

Using then theorem 7.1 of Hoeffding (1948), our lemmas 2.1 and 2.3 and following some routine steps, we arrive at the following.

THEOREM 3.1. If (i) $F_1 = \dots = F_N = F$, (ii) $F(x)$ is symmetric in its r arguments, and (iii) $\{H_N\}$ in (3.4) holds, then $\{N^{1/2}(W_{N,jk}^{-1/2}) - 2a_{jk} \int_{-\infty}^{\infty} g^2(x) dx, 1 \leq j < k \leq r\}$ has asymptotically a $1/2 r(r-1)$ -variate normal distribution with null mean vector and dispersion matrix $\Gamma_{\sim} = ((\gamma_{jk,j'k'}))$ given by

$$(3.5) \quad \gamma_{jk,j'k'} = \begin{cases} 1/3, & j=j', k=k', j \neq k, \\ 4\lambda(F)-1, & j=j', k \neq k', j \neq k \neq k', \\ 1 - 4\lambda(F), & j \neq j', j=k', j' \neq k, \\ 0 & j \neq j' \neq k \neq k' \end{cases}$$

Theorem 3.1 and lemma 2.3 show that the results derived by Doksum (1967) in his lemmas 2.1 through 2.4 remain true even when the errors are not all independent, but are within block symmetric dependent. Further, the use of theorem 3.1 as in (2.6) of [4] generalizes theorem 1 of Lehmann (1964) to

exchangeable error components. Since the main results of Lehmann (1964) are based on his theorems 1 and 2, and that of Doksum (1967) on his lemmas 2.3 and 2.4, it follows from our lemma 2.3 and theorem 3.1 that the Lehmann-Doksum procedures remain valid for within block exchangeable error components.

Now, we note that the variance of the cdf G is $2 \sigma^2(1-\rho) = \sigma^2(G)$, say. As such, using (3.2) and generalizing theorem 3.1 in the same manner as in lemma 2.3 of Doksum (1967), the asymptotic relative efficiency (A.R.E.) of the Doksum-test with respect to the classical ANOVA test, can again be shown to be equal to e' , defined by (2.11) and (2.12) of Doksum (1967). It is thus clear that the A.R.E. is unaffected by the within block symmetric dependence of the error components. The same conclusion also applies to Lehmann's procedure, as the variance of $(1/N) \sum_{i=1}^N X_{ijk}^*$ is also equal to $\sigma^2(G)$. This shows that theorem 4 of Lehmann (1964) is also true for exchangeable errors.

4. Robustness for heteroscedastic errors. Let us define

$$(4.1) \quad \bar{G}_N(x) = \frac{1}{N} \sum_{i=1}^N G_i(x), \quad \bar{G}_N^*(x,y) = \frac{1}{N} \sum_{i=1}^N G_i(x,y) \quad \text{and} \quad \bar{g}_N(x) = \frac{d}{dx} \{\bar{G}_N(x)\} .$$

Assume that the density functions $g_i(x) = (d/dx)G_i(x)$, $i=1, \dots, N$ satisfy

$$(4.2) \quad \sup_i \int_{-\infty}^{\infty} g_i^2(x) dx < \infty .$$

Then, it follows from the results of Sen (1967b) that under (3.1) and (4.2)

$$(4.3) \quad \lim_{N \rightarrow \infty} \{N^{1/2} E[\{W_{N,jk}^{-1/2}\} | H_N] - 2a_{jk} \int_{-\infty}^{\infty} \bar{g}_N^2(x) dx\} = 0,$$

for all $1 \leq j < k \leq r$. Further, by a direct generalization of theorem 2.1 of Sen (1967b), it follows that under $\{H_N\}$ in (3.1) and (4.2),

$\{[N^{1/2}(W_{N,jk}^{-1/2}) - 2a_{jk} \int_{-\infty}^{\infty} \bar{g}_N^2(x) dx], 1 \leq j < k \leq r\}$ has asymptotically a $\frac{1}{2}r(r-1)$ variate normal distribution with null mean vector and dispersion matrix

$\bar{\Gamma}_N = ((\bar{\gamma}_{N,jk,j'k'}))$, where

$$(4.4) \quad \bar{\gamma}_{N,jk,j'k'} = \begin{cases} 1/3, & j=j', k=k', j \neq k, \\ 4\lambda(\bar{F}_N) - 1, & j=j', k \neq k', j \neq k \neq k', \\ 1 - 4\lambda(\bar{F}_N), & j=k', j' \neq k, j \neq j' \neq k, \\ 0, & j \neq j', k \neq k', \end{cases}$$

and

$$(4.5) \quad \lambda(\bar{F}_N) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \bar{G}_N(x) \bar{G}_N(y) d\bar{G}_N^*(x,y).$$

LEMMA 4.1. $\frac{1}{4} \leq \lambda(\bar{F}_N) \leq 7/24$, uniformly in F_1, \dots, F_N (which are symmetric in their arguments.)

PROOF. We define $h(x) = \bar{G}_N(x) - \frac{1}{2}$. Since G_1, \dots, G_N are all non-decreasing and symmetric about 0, so is \bar{G}_N . Thus, $h(x)$ satisfies (2.2). Using then lemma 2.2, we obtain that

$$(4.6) \quad \frac{1}{4} \leq \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \bar{G}_N(x) \bar{G}_N(y) dG_i^*(x,y) \leq \frac{1}{4} + \frac{1}{2} \int_{-\infty}^{\infty} [\bar{G}_N(x) - \frac{1}{2}]^2 dG_i(x), \text{ for all } i.$$

The proof of the lemma is then completed from (4.1), (4.5) and (4.6), after noting that $\int_{-\infty}^{\infty} [\bar{G}_N(x) - \frac{1}{2}]^2 d\bar{G}_N(x) = 1/12$. Q.E.D.

Let us now denote by $\sigma_i^2 = V(Y_{ij})$, $\rho_i \sigma_i^2 = \text{cov}(Y_{ij}, Y_{ik})$, $j \neq k$, for $i=1, \dots, N$, and let

$$(4.7) \quad \bar{\delta}_N^2 = (1/N) \sum_{i=1}^N \sigma_i^2 (1 - \rho_i).$$

Then, proceeding on the same line as in lemmas 2.3 and 2.4 of Doksum (1967) and theorems 2, 3 and 4 of Lehmann (1964), and noting that $(1/N) \sum_{i=1}^N X_{ijk}^*$ has

the variance $2\bar{\delta}_N^2$, the A.R.E. of the Lehmann-Doksum procedures with respect to the corresponding parametric procedures is obtained as

$$(4.8) \quad e'_N = \bar{e}_N \{r/[2 + 6(r-2)(4\lambda(\bar{F}_N) - 1)]\} ,$$

where

$$(4.9) \quad \bar{e}_N = 12\bar{\delta}_N^2 \left[\int_{-\infty}^{\infty} \bar{g}_N^2(x) dx \right]^2 .$$

By virtue of lemma 4.1, $e'_N \geq \bar{e}_N$. Now, \bar{e}_N is the efficiency (A.R.E.) of the Wilcoxon signed rank test with respect to Student's t-test when F_1, \dots, F_N are not necessarily identical [cf. Sen (1967b)]. It follows from the results of Sen (1967b) that (i) \bar{e}_N is uniformly (in G_1, \dots, G_N) bounded below by 0.864 and (ii) if $G_i(x) = G(x/\sigma_i)$ for all $i=1, \dots, N$, then

$$(4.10) \quad \bar{e}_N \geq 12\sigma^2(G) \left[\int_{-\infty}^{\infty} g^2(x) dx \right]^2 ,$$

where the equality sign holds only when $\sigma_1 = \dots = \sigma_N$. Consequently, the same is also true for e'_N . This clearly indicates the robustness of the Lehmann-Doksum procedures for heteroscedastic errors (i.e., for $F_i(x_1, \dots, x_r) = F(x_1/\sigma_i, \dots, x_r/\sigma_i)$, $i=1, \dots, N$, when $\sigma_1, \dots, \sigma_N$ are not all equal).

REMARK. The estimator of $\lambda(F)$ considered by Lehmann (1964) (and cited by Doksum (1967), (2.7)) will be a consistent estimator of $\lambda(\bar{F}_N)$, even for exchangeable errors. The proof follows by the same technique as in theorem 4.2 of [5]. Consequently, when $F_1 \equiv \dots \equiv F_N \equiv F$, the same will estimate $\lambda(F)$, even for exchangeable errors. This shows that the estimate of $\lambda(F)$, proposed by Lehmann (1964) and required with the Lehmann-Doksum procedures, can also be used for interchangeable errors.

REFERENCES

- [1] DOKSUM, K. (1967). Robust procedures for some linear models with one observation per cell. Ann. Math. Statist. 38, 878-883.
- [2] HODGES, J. L., JR. and LEHMANN, E.L. (1962). Rank methods for combination of independent experiments in analysis of variance. Ann. Math. Statist. 33, 482-497.
- [3] HOEFFDING, W. (1948). On a class of statistics with asymptotically normal distribution. Ann. Math. Statist. 19, 293-325.
- [4] LEHMANN, E. L. (1964). Asymptotically nonparametric inference with one observation per cell. Ann. Math. Statist. 35, 726-734.
- [5] PURI, M. L. and SEN, P. K. (1966). On a class of multivariate multi-sample rank order tests. Sankhyā, Ser. A. 28, 353-376.
- [6] SEN, P. K. (1967a). On a class of aligned rank order tests in two way layouts. Ann. Math. Statist. (Under consideration of publication).
- [7] SEN, P. K. (1967b). On a further robustness property of the test and estimator based on Wilcoxon's signed rank statistic. Ann. Math. Statist. (Under consideration of publication).