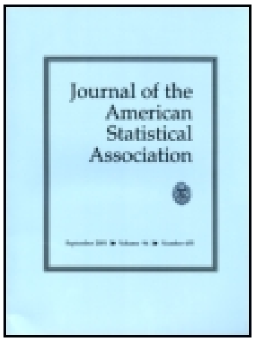

Journal of the American Statistical Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: www.tandfonline.com/journals/uasa20

Estimates of the Regression Coefficient Based on Kendall's Tau

Pranab Kumar Sen

To cite this article: Pranab Kumar Sen (1968) Estimates of the Regression Coefficient Based on Kendall's Tau, Journal of the American Statistical Association, 63:324, 1379-1389, DOI: 10.1080/01621459.1968.10480934

To link to this article: https://doi.org/10.1080/01621459.1968.10480934

Published online: 10 Apr 2012.

ESTIMATES OF THE REGRESSION COEFFICIENT BASED ON KENDALL'S TAU*

PRANAB KUMAR SEN
University of North Carolina, Chapel Hill

The least squares estimator of a regression coefficient β is vulnerable to gross errors and the associated confidence interval is, in addition, sensitive to non-normality of the parent distribution. In this paper, a simple and robust (point as well as interval) estimator of β based on Kendall's [6] rank correlation tau is studied. The point estimator is the median of the set of slopes $(Y_j - Y_i)/(t_j - t_i)$ joining pairs of points with $t_i \neq t_j$, and is unbiased. The confidence interval is also determined by two order statistics of this set of slopes. Various properties of these estimators are studied and compared with those of the least squares and some other nonparametric estimators.

1. INTRODUCTION

LET $Y_1, \cdots, Y_n$ be n independent random variables with distributions

$$P\{Y_i \le x\} = F_i(x) = F(x - \alpha - \beta t_i), \qquad i = 1, \cdots, n, \tag{1.1}$$

where $F(x)$ is a continuous cumulative distribution function (cdf), $t_1, \cdots, t_n$ are known constants (not all equal) and (α, β) are unknown parameters. Our purpose is to consider point as well as interval estimators of the regression coefficient β. If $F(x)$ has a finite variance $\sigma^2(F)$, the best (i.e., minimum variance unbiased) linear estimator of β is provided by the method of least squares. This estimator is vulnerable to gross errors and is also inefficient for distributions with 'heavy tails' (e.g., double exponential or logistic dcf). Moreover, the associated confidence interval for β, being based on the assumed normality of $F(x)$, is sensitive in small samples to any departure from this assumption. Alternative estimators of β based on suitable rank tests are proposed by Mood and Brown [8], Theil [12] and Adichie [1], among others. Mood and Brown propose to estimate α and β simultaneously from the two equations

$$\begin{aligned} \text{Median}(Y_i - \tilde{\alpha} - \tilde{\beta} t_i) &= 0 \qquad \text{for } t_i \le t_M, \\ \text{Median}(Y_i - \tilde{\alpha} - \tilde{\beta} t_i) &= 0 \qquad \text{for } t_i > t_M, \end{aligned} \tag{1.2}$$

where t_M is the median of $t_1, \cdots, t_n$. The point estimate $(\hat{\alpha}, \hat{\beta})$ is to be obtained by a trial and error solution and is subject to some arbitrariness when t_M is not uniquely defined (a case that may arise when $t_1, \cdots, t_n$ are not all distinct). Moreover, $\hat{\beta}$ is usually inefficient as compared to the other estimators (cf. [1]). A general class of point estimators of β (and also of α) is considered by Adichie [1]. However, his basic assumption that $F(x)$ is an absolutely continuous and symmetric distribution function with an absolutely continuous and square integrable density function is more restrictive than what is really needed in this paper. Moreover, his point estimators of β also require trial and error solutions. Such a trial and error procedure may indeed be quite laborious when n is not very small. Finally, Adichie gives no confidence interval for β. When $t_1, \cdots, t_n$

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

are all distinct, Theil [12] proposes a very simple point estimator of β, viz., the median of the $\binom{n}{2}$ slopes $(Y_j-Y_i)/(t_j-t_i)$, $1\leq i<j\leq n$. He also obtains a corresponding confidence interval for β in terms of these slopes. However, the asymptotic properties of the estimators are not studied by him. The procedure to be considered in the present paper is quite analogous to Theil's, but is based on weaker assumptions and do not require $t_1, \cdots, t_n$ to be all distinct. If N be the number of non-zero differences t_j-t_i $(1\leq i<j\leq n)$, the proposed point estimator is the median of the N slopes $(Y_j-Y_i)/(t_j-t_i)$ for which $t_i\neq t_j$. This is shown to be unbiased for β. The confidence interval for β is also obtained in terms of two order statistics of this set of N slopes. It is shown that the point and interval estimators of the location-difference in the two-sample case based on Wilcoxon test, proposed and studied by Hodges and Lehmann [4], Lehmann [7] and Sen [10, 11] are special cases of the estimators considered here. Properties of the estimators such as invariance, unbiasedness and asymptotic distribution are studied, and the asymptotic relative efficiency (A.R.E.) of the proposed procedure with respect to the least squares procedure and Adichie's [1] procedures are discussed. It is shown that for equally spaced values of $t_1, \cdots, t_n$ or for the two-sample problem (i.e., when t_i's can have only two values), the proposed estimator has A.R.E. never less than 0.864 with respect to the least squares estimator, though such a conclusion is not necessarily true when $t_1, \cdots, t_n$ are not equally spaced.

2. FORMULATION OF THE ESTIMATORS

Without any loss of generality we may assume that $t_1\leq t_2\leq \cdots \leq t_n$; they are already assumed to be not all equal. We define $c(u)$ to be 1, 0, or -1 according as u is $>$, $=$ or <0. Let then

$$N = \sum_{1\leq i<j\leq n} c(t_j - t_i), \tag{2.1}$$

i.e., N is the number of positive differences t_j-t_i, so that $N\leq\binom{n}{2}$, where the equality sign holds only when $t_1, \cdots, t_n$ are all distinct. For any real b, define $Z_i(b)=Y_i-bt_i$, $i=1, \cdots, n$. We then consider the following statistic basically related to Kendall's [6] tau between t_i and $Z_i(b)$, $i=1, \cdots, n$.

$$U_n(b) = \left\{N\binom{n}{2}\right\}^{-\frac{1}{2}} \sum_{1\leq i<j\leq n} c(t_j - t_i)c(Z_j(b) - Z_i(b)). \tag{2.2}$$

Thus, $\{N\binom{n}{2}\}^{\frac{1}{2}}\ U_n(b)$ is the difference-sign score that would appear in the numerator of the tau coefficient of correlation between the t_i and the (Y_i-bt_i), for some fixed b. Since $t_j\geq t_i$ for all $i<j$, $Z_j(b)-Z_i(b)$ is non-increasing in b for all $1\leq i<j\leq n$. Hence, from (2.2) it follows that $U_n(b)$ is also non-increasing in b. Now, by definition, $Z_1(\beta), \cdots, Z_n(\beta)$ are n independent and identically distributed random variables having the cdf $F(x-\alpha,)$ independent of $t_n=(t_1, \cdots, t_n)$. Consequently, $U_n(\beta)$ will be an estimator of 0, and will be stochastically small. In fact, $U_n(\beta)$ is a strictly distribution-free statistic having a distribution symmetric about 0 (cf. [6]). Thus one way of estimating β is to make $U_n(b)$ (by a proper choice of the estimator b) as close to zero as possible. Since, $U_n(b)$ is non-increasing in b, there will be an half-open interval

(in b) for which $U_n(b)$ will be equal to zero. The mid-point of this interval suggests itself as a natural estimate of β. Mathematically, we define the estimator as follows. Let

$$\begin{aligned} \beta_1^* &= \operatorname{Sup}\{b\colon U_n(b) > 0\}, \\ \beta_2^* &= \operatorname{Inf}\{b\colon U_n(b) < 0\}. \end{aligned} \tag{2.3}$$

Then, our proposed estimator is

$$\beta^* = \tfrac{1}{2}(\beta_1^* + \beta_2^*). \tag{2.4}$$

It may be noted that if instead of working with Kendall's tau, we work with the sample covariance of $Z_i(b)$ and t_i, $i=1, \cdots, n$, we will obtain the least squares estimator

$$\hat{\beta} = \sum_{i=1}^{n} (Y_i - \overline{Y}_n)(t_i - \bar{t}_n) \Big/ \left\{ \sum_{i=1}^{n} (\bar{t}_i - \bar{t}_n) \right\}^2,$$

where $\overline{Y}_n = (1/n)\sum_{i=1}^{n} Y_i$ and $\bar{t}_n = (1/n)\sum_{i=1}^{n} t_i$. An explicit formula for β^* will be considered in section 3.

To construct a confidence interval for β based on $U_n(b)$, we again note that $U_n(\beta)$ is a distribution-free statistic having a distribution symmetric about 0. Hence, depending on the sample size n, we can always select (U_n^*, ϵ_n) such that

$$P\{-U_n^* \le U_n(\beta) \le U_n^* \mid \beta\} = 1 - \epsilon_n, \tag{2.5}$$

where $0<\epsilon_n<1$. For small values of n (say, $n\le 10$), we may use Table 1 of Kendall [6, p. 171] to find appropriate values of U_n^* and ϵ_n. For large sample sizes, we adopt the following procedure. Let t_n be composed of $a_n(\ge 2)$ distinct sets of elements, where in the ith set there are u_i elements which are all equal, for $i=1, \cdots, a_n$. We define

$$V_n = (1/18)\left\{n(n-1)(2n+5) - \sum_{j=1}^{a_n} u_j(u_j-1)(2u_j+5)\right\}. \tag{2.6}$$

Thus V_n is the variance of $\{N\binom{n}{2}\}^{\frac{1}{2}}\, U_n(\beta)$ with the standard correction for tied observations, in the form that applies when there are ties in only one variable, (viz., t). Also, let τ_ϵ be the upper $100\epsilon\%$ point of a standard normal distribution. Then, from the results of Kendall [6] and Hoeffding [5], we obtain that

$$U_n^* \doteqdot \tau_{\frac{1}{2}\epsilon}\left\{V_n \Big/ \left[\hat{N}\binom{n}{2}\right]\right\}^{\frac{1}{2}}, \quad \text{where} \quad \epsilon_n \to \epsilon \quad \text{as} \quad n \to \infty. \tag{2.7}$$

Let us now define

$$\begin{aligned} \beta_U^* &= \operatorname{Sup}\{b\colon U_n(b) \ge -U_n^*\}, \\ \beta_L^* &= \operatorname{Inf}\{b\colon U_n(b) \le U_n^*\}. \end{aligned} \tag{2.8}$$

From (2.5) and (2.8), we arrive at the following

$$P\{\beta_L^* < \beta < \beta_U^* \mid \beta\} = 1 - \epsilon_n, \tag{2.9}$$

which is our proposed confidence interval for β having the confidence coefficient $1-\epsilon_n(\simeq 1-\epsilon$ for large $n)$. (2.9) provides an exact confidence interval with confidence coefficient $1-\epsilon_n$ for all unknown (but continuous) $F(x)$, no matter whether the normality and the finiteness of the variance of $F(x)$ hold or not. The exact expressions for β_L^* and β_U^* are considered in the next section.

3. EXACT EXPRESSIONS FOR THE ESTIMATORS

We recall that among the $\binom{n}{2}$ values of (t_j-t_i), $1\leq i\leq j\leq n$, only N (defined by (2.1)) values are non-zero, and the corresponding values of $Z_j(b)-Z_i(b)$ only have contributions to $U_n(b)$ in (2.2). We now consider the set S of N distinct pairs (i, j) for which $t_j>t_i$, and define

$$X_{ij} = (Y_j - Y_i)/(t_j - t_i), \qquad (i,j)\epsilon S. \tag{3.1}$$

Thus, the X_{ij}'s are the slopes of the lines connecting each pair of points (t_i, Y_i) and (t_j, Y_j) where $t_i\neq t_j$; the pairs of points for which $t_i=t_j$ are not considered. It will be seen that the N quantities in (3.1) define both the point and interval estimators. To do this, we arrange the N values in (3.1) in ascending order of magnitude and denote the rth smallest value by $X_{(r)}$ for $r=1, \cdots, N$. Then, looking at (2.2), we observe that if we compute the value of $U_n(X_{(r)})$, $(r-1)$ of the differences $Z_j(X_{(r)})-Z_i(X_{(r)})$ (for which $(i, j)\in S$) will be negative, $(N-r)$ will be positive and the remaining one will be exactly equal to 0. As such, $U_n(X_{(r)})$ will be equal to $(N-2r+1)/\{N\binom{n}{2}\}^{\frac{1}{2}}$. Similarly, $\{N\binom{n}{2}\}^{\frac{1}{2}}U_n(X^+_{(r)})$ will be equal to $(N-2r)$, where X^+ (or X^-) indicates that the value is just greater than (or less than) X. Now, we write $N=2M$ or $2M+1$ according as N is even or odd. For $N=2M+1$, we observe that $U_n(X_{(M+1)})=0$, while $U_n(X^-_{(M+1)})>0$ and $U_n(X^+_{(M+1)})<0$. Similarly, for $N=2M$, it follows that for any b in the open interval $(X_{(M)}, X_{(M+1)})$, $U_n(b)=0$, while it is positive or negative according as b is $\leq X_{(M)}$ or $\geq X_{(M+1)}$. Hence, from (2.3) and (2.4), we obtain

$$\beta^* = \begin{cases} X_{(M+1)}, & N = 2M + 1, \\ \frac{1}{2}(X_{(M)} + X_{(M+1)}), & N = 2M. \end{cases} \tag{3.2}$$

Thus β^* is the median of the N numbers $\{X_{ij}: (i, j)\in S\}$. To obtain the expressions for β_L^* and β_U^*, we let

$$N^* = \left\{N\binom{n}{2}\right\}^{\frac{1}{2}}\cdot U_n^* \quad \text{and} \quad M_i = \tfrac{1}{2}(N+(-1)^iN^*) \qquad \text{for } i = 1, 2, \tag{3.3}$$

where U_n^* is defined by (2.5). From (2.8), (3.3) and the observations made above, it follows that $U_n(X_{(M_1)})=(N^*+1)/\{N\binom{n}{2}\}^{\frac{1}{2}}>U_n^*$, but $U_n(X^+_{(M_1)})=N^*/\{N\binom{n}{2}\}^{\frac{1}{2}}=U_n^*$. Hence $\beta_L^*=X^+_{(M_1)}$. Similarly, $\beta_U^*=X^-_{(M_2+1)}$. Hence,

$$P\{X_{(M_1)} < \beta < X_{(M_2+1)} \mid \beta\} = 1 - \epsilon_n. \tag{3.4}$$

It may be noted that the classical least squares estimator $\hat{\beta}$, defined just after (2.4), can also be expressed as a linear function of the slopes $\{X_{ij}: (i, j)\in S\}$. In fact, $\hat{\beta}$ is a weighted mean of the variables X_{ij} with weights equal to $(t_j-t_i)^2$, whereas β^* is the median of the same set of variables. Since the median is less affected by gross errors or outliers than a weighted average, it follows that β^* will be more robust than $\hat{\beta}$.

We also note that the two sample location problem (cf. [4, 7, 10, 11]) is a special case of the general regression problem studied here. In this case, $t_1 = \cdots = t_{n_1} = 0$ and $t_{n_1+1} = \cdots = t_{n_1+n_2} = 1$ (where $n = n_1 + n_2$, $n_1 < n$). Thus, $N = n_1 n_2$ and β^* is the median of the $n_1 n_2$ differences $(Y_j - Y_i)$, $j = n_1+1, \cdots, n_1+n_2$, $i = 1, \cdots, n_1$. Also, β_L^* and β_U^* are defined as the M_1th and (M_2+1)th order statistics of these $n_1 n_2$ differences where M_1 and M_2 are defined by (3.3) and are based on the Wilcoxon two-sample test (cf. [7, 10, 11]).

4. AN ILLUSTRATIVE EXAMPLE

We consider the following data from Graybill [3, pp. 119–120], also considered by Adichie [1].

t_i	1	2	3	4	10	12	18
y_i	9	15	19	20	45	55	78

The least squares estimate of β is 4.02. Since all t_i's are distinct, $N = \binom{7}{2} = 21$. The values of X_{ij} defined by (3.1) are obtained as (in ascending order)

1, 2.5, 2.88, 3.67, 3.71, 3.75, 3.88, 3.93, 3.94, 4, 4, 4, 4,
4.06, 4.14, 4.18, 4.25, 4.75, 5, 5, 6.

Thus the point estimate of β is $X_{(11)} = 4$, which is the same value obtained by Adichie [1]. He has, however, employed a trial and error procedure for the computation of his estimator, as the exact expression in (3.2) is not applicable in his case (cf. [1, section 3)].

Now, from Table 1 of Kendall [6, p. 171], we observe that corresponding to a value of $\epsilon_n = 0.07$, the value of U_n^* in (2.5) is equal to 11/21. Thus, from (3.3), we obtain that $N^* = 11$, $M_1 = 5$ and $M_2 = 16$. Consequently, from (3.4), we obtain that the open interval (3.71), 4.18 provides a 93% confidence interval for β, valid for all continuous $F(x)$.

5. REGULARITY PROPERTIES OF THE ESTIMATORS

I. *Invariance.* We note that if we define $W_i = c_1 + c_2 Y_i$ and $s_i = d_1 + d_2 t_i$, $i = 1, \cdots, n$, (where c_2 and d_2 are different from 0), the regression parameter of W on s will be equal to $(c_2/d_2)\beta$. It is easy to verify that like the least squares estimator $\hat{\beta}$, the point estimator β^* in (2.4) satisfies this relation. The estimators β_L^* and β_U^* in (2.8) also satisfy this condition and as a result, the confidence interval in (2.9) may be regarded as invariant under linear transformations on the variables. Let us denote the point estimator in (2.4) by $\beta^*(Y_n, t_n)$ to denote its dependence on $Y_n = (Y_1, \cdots, Y_n)$ and $t_n = (t_1, \cdots, t_n)$. Then, it readily follows from (2.2), (2.3), and (2.4) that

$$\beta^*(Y_n + at_n, t_n) = \beta^*(Y_n, t_n) + a \qquad \text{for all real } a. \tag{5.1}$$

The same invariance relation is also satisfied by β_L^* and β_U^* in (2.8), and as a result the confidence interval in (2.9) is also invariant in the above sense. Again, by a straightforward generalization of the porof of Theorem 1 of [4], it can be shown that if $F(x)$ is continuous (or absolutely continuous) then so are the cdfs of all the statistics β_1^*, β_2^*, β_L^*, β_u^* and $\beta_u^*-\beta_L^*$.

II. *Unbiasedness.* We have the following theorem establishing this property of β^*.

Theorem 5.1. *The distribution of β^* is symmetric about the true parameter β.*

Proof. By virtue of (5.1), we may assume without any loss of generality that $\beta=0$. Rewriting $U_n(0)$ as $U(\mathbf{Y}_n, \mathbf{t}_n)$, we have from (2.2) that $U(-\mathbf{Y}_n, \mathbf{t}_n) = -U(\mathbf{Y}_n, \mathbf{t}_n)$. Also, for $\beta=0$, $U_n(0)$ has a distribution symmetric about 0 (cf. Kendall [6, p. 68]). Hence, $U(\mathbf{Y}_n, \mathbf{t}_n)$ and, $U(-\mathbf{Y}_n, \mathbf{t}_n)$ have the same distribution. Also from (2.3) and (2.4), we obtain that $\beta^*(\mathbf{Y}_n, \mathbf{t}_n) = -\beta^*(-\mathbf{Y}_n, \mathbf{t}_n)$. Hence, the distribution of $\beta^*(\mathbf{Y}_n, \mathbf{t}_n)$, being the same as of $\beta^*(-\mathbf{Y}_n, \mathbf{t}_n)$, is also symmetrical about 0, the assumed value of β. Q.E.D.

III. *Validity when both variables are subject to errors.* We consider here the more general case, in which $\mathbf{t}_n$ is not observable and the observable (random) variable is $\mathbf{W}_n=(W_1, \cdots, W_n)$, where $W_i=t_i+v_i$, $i=1, \cdots, n$. It is assumed that $Y_i=\alpha+\beta t_i+e_i$, where (e_i, v_i) are stochastically independent, for $i=1, \cdots, n$. Thus, having observed (Y_i, W_i), $i=1, \cdots, n$, we want to estimate β. Theil [11] considered this problem under the assumptions that (*i*) $P\{|v_i|>g_i\}=0$ for some finite $g_i(>0)$, (ii) $|t_j-t_i|>g_i+g_j$ for all $i\neq j$, and (iii) the random variables $\epsilon_i=e_i-\beta v_i$, $i=1, \cdots, n$, are all independent and identically distributed. Under these assumptions, $P\{W_i\neq W_j, \forall i\neq j\}=1$. Thus the W_i's occur in the same order as the t_i's and we can consider (with probability 1) $W_1<W_2< \cdots <W_n$, so that N, defined by (2.1), is equal to $\binom{n}{2}$. Hence, defining $U_n(b)$ as in (2.2), with t_i's replaced by W_i's, we obtain that here with probability 1,

$$\begin{aligned} U_n(\beta) &= \binom{n}{2}^{-1} \sum_{i<j} c(Y_j - Y_i - \beta(W_j - W_i)) \\ &= \binom{n}{2}^{-1} \sum_{i<j} c(\epsilon_j - \epsilon_i), \end{aligned} \tag{5.2}$$

which is symmetrically distributed about 0. Consequently, proceeding as in Theorem 5.1, we may conclude that the estimate β^* in (3.2) (with t_i's replaced by W_i's) is unbiased for β. The invariance property also holds in this case.

Other properties of the estimator are considered in the next section.

6. ASYMPTOTIC PROPERTIES OF THE ESTIMATORS

Here we shall consider (i) the asymptotic normality of the point estimator in (2.4), (ii) asymptotic properties of the confidence interval in (2.9), and (iii) the asymptotic relative efficiencies of the point and interval estimators with respect to the corresponding estimators based on the least squares principle. For this purpose we define $\bar{t}_n$ as in section 2, and let

$$T_n^2 = \sum_{i=1}^{n} (t_i - \bar{t}_n)^2, \quad A_n^2 = (1/12)\left\{n(n^2-1) - \sum_{j=1}^{a_n} u_j(u_j^2 - 1)\right\}, \tag{6.1}$$

where a_n and u_j's are defined just before (2.6), and also let

$$\rho_n = \sum_{i=1}^{n} (i - \tfrac{1}{2}(n+1))(t_i - \bar{t}_n)/(T_n A_n). \tag{6.2}$$

That is, ρ_n is the product moment correlation coefficient between $(t_1, \cdots, t_n)$ and $(1, \cdots, n)$, as adjusted for ties. Finally, we assume that $F(x)$ is absolutely continuous having a continuous density function $f(x)$ satisfying

$$B(F) = \int_{-\infty}^{\infty} f^2(x)dx < \infty. \tag{6.3}$$

Then, we have the following two theorems whose proofs are supplied in the Appendix. (In order to take care of the asymptotic situation we conceive of a sequence of sample sizes and a corresponding sequence of estimators, defined by (2.4), (2.8) and (2.9). We shall attach the suffix n to these estimators to denote such a sequence.)

Theorem 6.1. *If (i) ρ_n is strictly positive and* (ii) $T_n \rightarrow \infty$ *as* $n \rightarrow \infty$, *then* $\rho_n T_n(\beta_n^* - \beta)$ *has asymptotically a normal distribution with zero mean and variance* $1/(12B^2(F))$.

Theorem 6.2. *Under the conditions of Theorem* 6.1, $\rho_n T_n(\beta_{U,n}^* - \beta_{L,n}^*)$ *converges in probability to* $\tau_{\epsilon/2}/(\sqrt{3}\, B(F))$, *where* ϵ *is the limiting value of* ϵ_n, *defined by* (2.9).

We denote the sequence of least squares estimators by $\hat{\beta}_n$ and the allied confidence intervals (corresponding to the same confidence coefficient $1-\epsilon$) by $\hat{\beta}_{L,n} \leq \beta \leq \hat{\beta}_{U,n}$. Then, it is well known that (i) $T_n(\hat{\beta}_n - \beta)$ has asymptotically a normal distribution with 0 mean and variance $\sigma^2(F)$, (where $\sigma^2(F)$ is the variance of the cdf $F(x)$,) and (ii) $T_n(\hat{\beta}_{U,n} - \hat{\beta}_{L,n})$ converges in probability to $2\tau_{\frac{1}{2}\epsilon}\, \sigma(F)$. (In this connection, the reader may be referred to Eicker [2].) Now, to study the asymptotic relative efficiency (A.R.E.) of β_n^* with respect to $\hat{\beta}_n$, we compare the reciprocals of their asymptotic variances, and obtain that

$$\text{A.R.E.}(\beta^*/\hat{\beta}) = 12\sigma^2(F)\rho^2 B^2(F), \tag{6.4}$$

provided ρ_n^2 converges to the limit $\rho^2 (>0)$ as $n \rightarrow \infty$. Similarly, as in [7, 11] we compare the reciprocals of the squares of the limiting values of $T_n(\beta_{U,n}^* - \beta_{L,n}^*)$ and $T_n(\hat{\beta}_{U,n} - \hat{\beta}_{L,n})$ as a measure of their A.R.E., and arrive at (6.4) as the A.R.E. of the confidence interval in (2.9) with respect to the confidence interval derived from the least squares estimators. We shall now study (6.4) in more detail. For this, we recall that t_n is composed of a_n distinct sets of elements, where in the jth set there are u_j elements which are all equal to t_j^*, say, for $j = 1, \cdots, a_n$ (≥ 2), where $t_1^* < \cdots < t_{a_n}^*$. Let $R_j = u_0 + \cdots + u_{j-1} + \frac{1}{2}(u_j + 1)$, for $j = 1, \cdots, a_n$ where $u_0 = 0$. Then, we have the following:

Theorem 6.3. $0 \leq \rho_n \leq 1$, *where the upper bound* 1 *is attained if and only if* $t_j^* = a + bR_j$ *for all* $j = 1, \cdots, a_n$, *where b is positive.*

Proof. Since $t_1 \leq t_2 \leq \cdots \leq t_n$, the numerator on the right hand side of (6.2) is non-negative, and hence, $\rho_n \geq 0$. To prove that $\rho_n \leq 1$, we rewrite $T_n A_n \rho_n$ as $\sum_{j=1}^{a_n} u_j(R_j - \frac{n+1}{2})(t_j^* - \bar{t}_n)$ which by the Cauchy-Schwarz inequality is less than or equal to $T_n A_n$, where the equality sign holds if and only if $(t_j^* - \bar{t}_n) = b(R_j - \frac{n+1}{2})$, for all $j = 1, \cdots, a_n$. This completes the proof of the theorem.

Two particular cases where $\rho_n = 1$ are of special interest. First, the general regression problem with equispaced independent variables where $t_i = t_1 + (i-1)h$, $h > 0$, $i = 1, \cdots, n$. The second case relates to the experimental design where all the observations are placed at the two end-points of an interval for the optimum least squares estimation of the slope, i.e., when $t_1 = \cdots t_{n1} = t_1^*$, $t_{n_1+1} = \cdots = t_n = t_2^* > t_1^*$, where $n_1 < n$. (As has been noted earlier, the second case also resembles the classical two-sample location problem.) In either case, we shall say that the independent variables are *optimally designed* if $\rho_n = 1$. We shall also say that the independent variables are *asymptotically optimally designed* if $\rho_n \to 1$ as $n \to \infty$. As an example, consider the following design:

$$\begin{array}{llllll} t_j^* & -2 & -1 & 1 & 2 & \bar{t}_n = 0 \\ u_j & 1 & m & m & 1 & n = 2m + 2. \end{array} \qquad (6.5)$$

Here, clearly $\rho_n \to 1$ as $n \to \infty$. From theorem 6.3 and the above discussion we readily arrive at the following theorem.

Theorem 6.4. A.R.E. $(\beta^* | \hat{\beta}) \leq 12\sigma^2(F)B^2(F)$, *where the equality sign holds if the independent variables are* (*at least asymptotically*) *optimally designed.*

Thus, for optimal or asymptotically optimal designs, the A.R.E. of β^* relative to $\hat{\beta}$ is the same as that of the Wilcoxon test with respect to the Student's t-test (for the two sample location problem). Thus, as in [9, p. 89], it follows that (i) when $F(x)$ is normal, this A.R.E. is equal to $3/\pi = 0.955$, (ii) when $F(x)$ is logistic or double exponential, it is greater than unity, (iii) for distributions with 'heavy tails' (such as Cauchy etc.), it may be indefinitely large and (iv) for any continuous $F(x)$, it cannot be less than 0.864. On the contrary, if $t_1, \cdots, t_n$ are not optimally designed, so that ρ_n does not tend to 1 as $n \to \infty$, this A.R.E. may not have any lower bound (such as 0.864 or so). In fact, if $\rho_n \to 0$ as $n \to \infty$, so also will this A.R.E. As an example of a bad design, consider the following

$$\begin{array}{llllll} t_j^* & -m & -1 & 1 & m & (m > 1) \\ u_j & 1 & m & m & 1 & n = 2m + 2. \end{array} \qquad (6.6)$$

By straightforward computations it follows that

$$\rho_n = m(3m+1)/\{m(m+1)(m^3 + 4m^2 + 4m + 1)\}^{1/2} = 0(3/m^{1/2}) = 0(n^{-1/2}), \qquad (6.7)$$

and this converges to zero as $n \to \infty$. In spite of such pathological examples, in actual practice, ρ_n is usually well away from 0, and as a result (6.4) can be used to provide a reasonable idea about the efficiency of β^*. However, theorem 6.1, (6.4) and theorem 6.4 clearly indicate that if the choice of t_n is left to the experimenter, he should always try to select t_n in such a way that (i) ρ_n is either exactly or nearly equal to 1 and (ii) T_n^2 is maximum for the practicable range of values of $t_1, \cdots, t_n$.

It is also worth comparing the A.R.E. of β^* with respect to the estimators proposed by Adichie [1]. His estimates are in fact based on a class of 'mixed rank' statistics of the type $\sum_{j=1}^{n}(t_i - \bar{t}_n)\psi_n(R_j/(n+1))$, where R_j refers to the rank of Y_j among $Y_1, \cdots, Y_n$, and ψ_n is some suitable rank score. For general ψ_n, the expression for the A.R.E. of his estimator with respect to $\hat{\beta}$ is given by (6.1) of [1]. Hence, the A.R.E. of β^* with respect to his estimator can be obtained from our (6.4) and his (6.1). A special case considered by him in section 3 [1, pp. 896–897] is the estimator $\hat{\beta}_\omega$ based on the Wilcoxon-scores statistic i.e., on $\sum_{j=1}^{n}(t_j-\bar{t})R_j$, and in this case, the A.R.E. of $\hat{\beta}_\omega$ with respect to $\hat{\beta}$ comes out as $12\sigma^2(F)B^2(F)$. Thus, the A.R.E. of β^* with respect to $\hat{\beta}_\omega$ is equal to the limiting value of ρ_n^2, provided such a limit is different from 0. This means that for optimum or asymptotically optimum designs, β^* and $\hat{\beta}_\omega$ are asymptotically equally efficient, but, unlike β^*, $\hat{\beta}_\omega$ is not affected by bad design of t_n. However, this is not unexpected. $\hat{\beta}_\omega$, like $\hat{\beta}$, utilizes the exact values of $t_1, \cdots, t_n$ in the mixed-rank statistic, whereas β^* only utilizes their ordering. On the other hand, $\hat{\beta}_\omega$ has to be obtained by a trial and error solution, whereas β^* can be obtained simply as the median of the slopes. So in actual practice, if ρ_n is close to unity, it may definitely be of some advantage to consider a (possibly) slightly inefficient but quick estimator rather than a computationally complicated one.

In passing, we may remark that by virtue of theorem 6.2,

$$\hat{\beta}(F) = \tau_{\epsilon/2}/\{\sqrt{3}\rho_n T_n(\overset{*}{\beta}_{U,n}-\overset{*}{\beta}_{L,n})\} \xrightarrow{p} B(F), \qquad \text{as } n \to \infty, \tag{6.8}$$

for all absolutely continuous $F(x)$. This result is an immediate generalization of a similar result (for the two sample location problem) (cf. [7, 11]) to the more general regression problem.

7. APPENDIX

The proofs of theorems 6.1 and 6.2 are based on the following.

Theorem 7.1. *If (i) ρ_n is strictly positive and (ii) $T_n \to \infty$ as $n \to \infty$, then under $H_0:\beta=0$, $[\{N\binom{n}{2}\}^{\frac{1}{2}}U_n(b/T_n)+4bB(F)\rho_n A_n]/V_n^{\frac{1}{2}}$ has asymptotically a normal distribution with zero mean and unit variance, where N, $U_n(b/T_n)$, V_n, T_n and A_n, ρ_n and $B(F)$ are defined by* (2.1), (2.2), (2.6), (6.1), (6.2) and (6.3) *respectively.*

Proof. We note that for large T_n, $E\{c(Z_j(b/T_n)-Z_i(b/T_n))|H_0\}$ $=2P_0(Y_j-Y_i \geq (b/T_n)(t_j-t_i))-1$, (where P_0 indicates that H_0 is assumed to be true), reduces to $-2b(t_j-t_i)B(F)/T_n+o(T_n^{-1})$. Also, we note that $\sum_{i<j}(t_j-t_i)=2\rho_n A_n T_n$. Hence, it follows from (2.2) that $E\{U_n(b/T_n)|H_0\}$ $=-4lB(F)\rho_n A_n/\{N\binom{n}{2}\}^{\frac{1}{2}}+o(1)$. In a similar manner, it can be shown that $\{N\binom{n}{2}\}\mathrm{Var}[U_n(b/T_n)]/V_n$ converges to one as $n \to \infty$. Finally, the asymptotic normality of $U_n(b/T_n)$ follows readily from Theorem 7.1 of Hoeffding [5], after noting that $U_n(b/T_n)$ is a U-statistic for all real b. Q.E.D.

Proof of Theorem 6.1. Here also we assume without any loss of generality that $\beta=0$. Then, it follows from (2.2), (2.3) and (2.4) that for any real a,

$$\begin{aligned}\lim_{n\to\infty} P_0\{\rho_n T_n \overset{*}{\beta}_n \leq a\} &= \lim_{n\to\infty} P_0\{U_n(a/\rho_n T_n) \leq 0\} \\ &= \lim_{n\to\infty} G(4aB(F)A_n/V_n^{\frac{1}{2}}),\end{aligned} \tag{7.1}$$

by theorem 7.1, where $G(x)$ is the standard normal cdf. Now, it follows from (2.6) and (6.1) that $A_n^2/V_n \to 3/4$ as $n \to \infty$. Consequently, $4B(F)A_n/V_n^{\frac{1}{2}}$ tends to $\sqrt{12}\,B(F)$, and this completes the proof.

For the proof of theorem 6.3, we note that for any two real and finite (b, b'), under $H_0:\beta=0$, the covariance of $\{N\binom{n}{2}/V_n\}^{\frac{1}{2}}U_n(b/\{\rho_nT_n\})$ and $\{N\binom{n}{2}/V_n\}^{\frac{1}{2}}U_n(b'/\rho_nT_n\})$ can be shown to be asymptotically equal to unity. Hence, using the results of theorem 7.1, we see that as $n\to\infty$,

$$\left\{N\binom{n}{2}\Big/V_n\right\}^{\frac{1}{2}}\cdot E\{U_n(b/\{\rho_nT_n\}) - U_n(b'/\{\rho_nT_n\})\} \\ - 4(b'-b)B(F)A_n/V_n^{\frac{1}{2}}\to 0, \tag{7.2}$$

$$\left\{N\binom{n}{2}\Big/V_n\right\}\operatorname{Var}\{U_n(b/\{\rho_nT_n\}) - U_n(b'/\{\rho_nT_n\})\}\to 0. \tag{7.3}$$

(7.2) and (7.3) along with the Chebyshev's inequality imply that

$$\left|\left\{N\binom{n}{2}\Big/V_n\right\}^{\frac{1}{2}}\{U_n(b/\{\rho_nT_n\}) - U_n(b'/\{\rho_nT_n\})\} \\ - 4(b'-b)B(F)A_n/V_n^{\frac{1}{2}}\right|\xrightarrow{p}0. \tag{7.4}$$

Now, proceeding as in theorem 7.1, it follows after some manipulations that $\rho_nT_n(\beta^*_{U,n}-\beta)$ has asymptotically a normal distribution with mean $\tau_{\epsilon/2}/\{\sqrt{12}B(F)\}$ and variance $1/\{12B^2(F)\}$. This implies that

$$|\rho_nT_n(\overset{*}{\beta}_{U,n}-\beta) - \tau_{\epsilon/2}/\{\sqrt{12}B(F)\}|\text{ is bounded in probability,} \tag{7.5}$$

and similarly, it can be shown that

$$|\rho_nT_n(\overset{*}{\beta}_{L,n}-\beta) + \tau_{\epsilon/2}/\{\sqrt{12}B(F)\}|\text{ is bounded in probability.} \tag{7.6}$$

From (7.4), (7.5) and (7.6), we may conclude (on noting that by assumption $\beta=0$) that

$$\left\{N\binom{n}{2}\Big/V_n\right\}^{1/2}[U_n(\overset{*}{\beta}_{L\,n}) - U_n(\overset{*}{\beta}_{U,n})] \\ = 4\rho_nT_n(\overset{*}{\beta}_{U,n} - \overset{*}{\beta}_{L,n})B(F)\cdot A_n/V_n^{\frac{1}{2}} + \sigma_p(1). \tag{7.7}$$

Now, by (2.7) and (2.8), the left hand side of (7.7) converges to $2\tau_{\epsilon/2}$, and also, $A_n/V_n^{\frac{1}{2}}\to\sqrt{3}/2$ as $n\to\infty$. Hence, theorem 6.2 follows from (7.7). Q.E.D.

ACKNOWLEDGMENT

The author is grateful to the editor, the associate editor and the referees for their valuable comments on the paper and to Professor Herbert A. David for his careful reading of the manuscript.

REFERENCES

[1] Adichie, J. N., "Estimates of regression parameters based on rank tests," *Annals of Mathematical Statistics*, 38 (1967), 894–904.

[2] Eicker, F., "Asymptotic normality and consistency of least squares estimators for families of linear regressions," *Annals of Mathematical Statistics*, 34 (1963), 447–56.

[3] Graybill, F., *Introduction to linear statistical models, Volume 1.* McGraw-Hill Book Company, New York, 1961.

[4] Hodges, J. L., Jr., and Lehmann, E. L., "Estimates of location based on rank tests," *Annals of Mathematical Statistics*, 34 (1963), 598–611.

[5] Hoeffding, W., "A class of statistics with asymptotically normal distribution," *Annals of Mathematical Statistics*, 19 (1948), 293–325.

[6] Kendell, M. G., *Rank correlation methods.* Charles Griffin and Company: London. Second edition, 1955.

[7] Lehmann, E. L., "Nonparametric confidence intervals for a shift parameter," *Annals of Mathematical Statistics*, 34 (1963), 1507–12.

[8] Mood, A. M., *Introduction to the theory of statistics.* McGraw-Hill Book Company: New York, 1950.

[9] Noether, G., *Elements of nonparametric statistics.* John Wiley: New York, 1967.

[10] Sen, P. K., "On the estimation of relative potency in dilution (-direct) assays by distribution-free methods," *Biometrics*, 19 (1963), 532–52.

[11] Sen, P. K., "On a distribution-free method of estimating asymptotic efficiency of a class of non-parametric tests," *Annals of Mathematical Statistics*, 37 (1966), 1759–70.

[12] Theil, H., "A rank-invariant method of linear and polynomial regression analysis," I, II, and III, *Nederl. Akad. Wetensch. Proc.*, 53 (1950), 386–92, 521–5 and 1397–412.