The Annals of Statistics
1983, Vol. 11, No. 2, 697–701

JUSTIFICATION FOR A $K - S$ TYPE TEST FOR THE SLOPE OF A TRUNCATED REGRESSION[1]

BY P. K. BHATTACHARYA

University of California, Davis

A $K - S$ type statistic computed from sequential ranks has been proposed in the astrophysics literature for testing the slope of a truncated regression. There is an easy heuristic justification for the test in the nontruncated case, but it fails under truncation. This paper extends the heuristic justification to the truncated case and outlines a more complete proof of the asymptotic property.

1. Introduction. A current controversy in cosmology has led to the problem of nonparametric inference about the slope β of a truncated regression $Y = \beta x + V$ in which only those (x, Y) are observed for which $Y \leq y_0$. Any pair (x, Y) with $Y > y_0$ is not observed, nor do we have any record of such pairs. This is different from censoring where $Y > y_0$ is replaced by y_0. For a more complete discussion of the model see Bhattacharya, Chernoff and Yang (1983), where a modification of a nonparametric estimate of β based on Kendall's tau is developed.

An entirely different approach to inference in the truncated problem was proposed by Turner (1979), using a Kolmogorov-Smirnov test based on sequential ranks. It is easy to present a heuristic justification for the applicability of the Turner procedure in the nontruncated case, but that justification fails in the truncated case. Nevertheless, a Monte Carlo simulation of a special case surprisingly indicated that the test seems to be valid in the sense that the null distribution of the test statistic is approximately that of the $K - S$ statistic D_n^+.

The object of this paper is to explain this apparent anomaly by extending the heuristic justification to the truncated case, and to outline a more complete proof. Though our discussion is focused on a $K - S$ type test, our main result (Theorem 3) justifies the validity of a whole class of tests as pointed out in a remark at the end of Section 4. However, power and efficiency comparisons of these tests are beyond the scope of our discussion.

2. Difference between the nontruncated and truncated cases. First consider the nontruncated case and suppose the data $(x_1, Y_1), \cdots, (x_n, Y_n)$ are arranged such that $x_1 < \cdots < x_n$, (we assume, for simplicity that the x_i are distinct), and the residuals from the true regression are i.i.d. with an unknown continuous cdf F. Then under $H_0: \beta = \beta_0 > 0$, the residuals

$$V_i = Y_i - \beta_0 x_i, \quad 1 \leq i \leq n$$

calculated for the hypothetical regression should exhibit the i.i.d. behavior, which can be tested on the basis of *sequential ranks* R_i of V_i among $(V_1, \cdots, V_{i-1}, V_i)$ using the well-known fact (cf. Bhattacharya and Frierson, 1981) that the sequential ranks $R_1, \cdots, R_n$ from i.i.d. observations are independent of each other and of the order statistics, and that R_i is uniformly distributed on $1, \cdots, i$. Hence the quantities $\xi_i = (2R_i - 1)/2i$ are independent and approximately uniform on $(0, 1)$ for large i, so that if $H_n(t)$ is the empirical cdf of $\xi_1, \cdots, \xi_n$, then for large n, the empirical process $X_n(t) = \sqrt{n}\,\{H_n(t) - t\}$ should behave like the Brownian bridge $B^*(t)$ under H_0. In particular, the asymptotic cdf of T_n

Received March 1981; revised January 1982.

[1] This research was supported by NSF grants MCS-78-01108 and MCS-80-02774.

Key words and phrases. Truncated regression, Kolmogorov-Smirnov test, sequential rank, asymptotic distribution, Brownian bridge.

AMS 1980 *subject classification.* Primary 62G10; secondary 62E20, 62J05.

$= \sup_{0<t<1} \sqrt{n}\{H_n(t) - t\}$ should be the same as that of the $K - S$ statistic D_n^+, viz., $1 - \exp(-2x^2)$. On the other hand, if the true $\beta < \beta_0$, then V_i would tend to be smaller than V_j for $i > j$, the R_i would tend to be small, and $X_n(t)$ would tend to attain larger values than $B^*(t)$, causing T_n to be relatively large. Thus T_n may be used as a test statistic for testing H_0.

The Turner procedure is an adaptation of the test statistic T_n to the truncated case. Now the sequential ranks have to be modified in view of the fact that even under H_0, the independent residuals are not identically distributed, but $V_i = Y_i - \beta_0 x_i$ has cdf $F(v)/F(w_i)$ for $v \le w_i$, where

$$w_i = y_0 - \beta_0 x_i$$

is the *truncation point* for V_i and $w_1 > \cdots > w_n$, i.e., successive observations are progressively truncated. For $j < i$, V_j has a larger range of values than V_i, and a comparison between V_i and V_j is meaningless when $w_i < V_j \le w_j$. Thus we say that for $j \le i$, V_j is *comparable* to V_i only if $V_j \le w_i$ and subject to this condition, V_j and V_i are i.i.d. Hence, Turner uses the *modified sequential ranks* $\{R_i\}$ by ranking each V_i among those V_j for which j is in the set of integers

$$S_i = \{j: V_j \le w_i, \quad 1 \le j \le i\},$$

of which the number of elements N_i is a random variable. Letting $H_n(t)$ denote the empirical cdf of

$$\xi_i = (2R_i - 1)/2N_i, \quad 1 \le i \le n,$$

the test statistic now becomes

$$T_n = \sup_{0<t<1} \sqrt{n}\{H_n(t) - t\}.$$

From the joint behavior of $\{N_i\}$ and $\{R_i\}$ established in Theorem 1, it is seen that for given $\{N_i\}$ each ξ_i has a conditional uniform distribution on $1/2N_i, 3/2N_i, \cdots, (2N_i - 1)/2N_i$, but due to the randomness of the sets S_i, they are no longer independent. This invalidates our heuristic argument for the nontruncated case. Can the almost uniform distribution of ξ_i given $\{N_i\}$ suffice to permit us to extend the result to Turner's statistic T_n? To do so will require, in Section 4, the consideration of:

CONDITION A. $\liminf_{n\to\infty}\{\min_{k_n<i\le n} i^{-(1/2+\delta)} \sum_{j=1}^{i} F(w_i)/F(w_j)\} > 0$ for some δ with $0 < \delta < \frac{1}{2}$ and some sequence $\{k_n\} \to \infty$ such that $n^{-1/2}k_n \to 0$ as $n \to \infty$.

Condition A implies that the rate of progressive truncation is not too severe. This guarantees that $\{N_i\}$ increases sufficiently fast with high probability, so that the granularity of the ξ_i is negligible in the limit. A key theorem on weak convergence to the Brownian bridge in the non-independent case (Theorem 4) is presented in the appendix.

3. Conditional uniformity and independence of $\{\xi_i\}$. We recall that N_i is the number of V_j comparable to V_i for $j \le i$, and that R_i is the rank of V_i among these N_i V's. Theorem 1 states that the R_i, and hence the ξ_i, are conditionally uniform and independent and that the N_i form a relatively simple Markov Chain. Let $U_i = F(V_i)$ and $a_i = F(w_i)$. We shall prove:

THEOREM 1. *The sequence $\{N_i, 1 \le i \le n\}$ is a Markov Chain with $N_1 = 1$ and*

$$P\{N_{i+1} = k \mid N_i = m\} = \binom{m}{k-1}\left(\frac{a_{i+1}}{a_i}\right)^{k-1}\left(1 - \frac{a_{i+1}}{a_i}\right)^{m-k+1}, \quad 1 \le k \le m+1,$$

and for given $\{N_i\}$, the R_i are conditionally independent with R_i uniformly distributed on $\{1, 2, \cdots, N_i\}$.

The proof of this theorem is simplified by use of the following easy lemmas which require some additional notation. For $j = 1, 2, \cdots, N_i$, let U^*_{ij} be the $U = F(V)$ value associated with the jth element of S_i, i.e., the jth V which is comparable to V_i. Let R_{ij} be the *sequential rank* of U^*_{ij} among $U^*_{i1}, U^*_{i2}, \cdots, U^*_{ij}$. Clearly,

$$R_{iN_i} = R_i.$$

Finally let $\mathscr{C}_i$ denote the collection, N_i, S_i, and $\{U_j : j < i, j \notin S_i\}$. Then Lemma 1 is immediate.

LEMMA 1. *Given $\mathscr{C}_i$ the random variables $U^*_{i1}, U^*_{i2}, \cdots, U^*_{iN_i}$ are conditionally* i.i.d. *and uniform on* $[0, a_i]$.

Using the property of sequential ranks and that of conditional expectation, we then have:

LEMMA 2. $P\{R_i = r \mid \mathscr{C}_i, R_{i1}, \cdots R_{i,N_i-1}\} = N_i^{-1} = P\{R_i = r \mid N_i\}, 1 \le r \le N_i$.

To describe the N_i process, we observe that $N_{i+1} - 1$ is the number of the N_i U^*_{ij} which are less than or equal to a_{i+1}. Applying the properties of sequential ranks again, it follows that $N_{i+1} - 1$ has the conditional binomial distribution implied in:

LEMMA 3. $P\{N_{i+1} = k \mid \mathscr{C}_i, R_{i1}, \cdots, R_{i,N_i-1}, R_i\} =$

$$\binom{N_i}{k-1}\left(\frac{a_{i+1}}{a_i}\right)^{k-1}\left(1 - \frac{a_{i+1}}{a_i}\right)^{N_i-k+1} = P\{N_{i+1} = k \mid N_i\}, 1 \le k \le N_i + 1.$$

Lemma 3 implies the Markov chain part of Theorem 1. The proof of Lemma 3 extends to yield:

LEMMA 4. *The conditional distribution of $(N_{i+1}, \cdots, N_n)$ given $(\mathscr{C}_i, R_{i1}, \cdots R_{iN_i-1}, R_i)$ may be expressed in terms of $N_i, a_i, a_{i+1}, \cdots, a_n$.*

PROOF OF THEOREM 1. The crucial point is that $(N_1, R_1, N_2, \cdots, R_{i-1}, N_i)$ is determined by $Z = (\mathscr{C}_i, R_{i1}, \cdots, R_{iN_i-1})$. Hence Lemma 2 implies that given $(N_1, N_2, \cdots, N_i)$, R_i is uniformly distributed on 1 to N_i and independent of $(R_1, R_2, \cdots, R_{i-1})$. It remains to demonstrate this behavior conditional on $\{N_i : 1 \le i \le n\}$.

But Lemma 4 implies that $(N_{i+1}, \cdots N_n)$ and $(R_1, \cdots, R_i)$ are conditionally independent given $(N_1, N_2, \cdots, N_i)$. Then the conditional distribution of $(R_1, \cdots R_i)$ given $(N_1, \cdots, N_n)$ is the same as that given $(N_1, \cdots, N_i)$. Hence, given $(N_1, N_2, \cdots, N_n)$, R_i is uniformly distributed on 1 to N_i and R_i is independent of $R_1, R_2, \cdots, R_{i-1}$. It follows that the R_i are conditionally mutually independent given $\{N_i\}$, and the theorem follows.

4. Granularity and the main result. If it were possible to neglect the granularity of the distributions of the ξ_i, $X_n(t) = n^{1/2}\{H_n(t) - t\}$ would converge weakly to the Brownian bridge on [0, 1] and Turner's test statistic T_n would have the limiting $K - S$ distribution.

In the Appendix, Theorem 4 is established. It implies the convergence of X_n subject to:

CONDITION B. $n^{-1/2}\sum_{i=1}^n N_i^{-1} \to_p 0$ as $n \to \infty$,

which is a restriction on the granularity of the ξ_i. We shall prove Theorem 2 which states that Condition A implies B. The weak convergence of X_n in Theorem 3 and of the $K - S$ type statistic T_n follow immediately, yielding our desired result.

The reader should note that, for simplicity of notation, there has been some abuse of

notation. In stating Condition A, we should have noted that the x_i and hence the w_i and $a_i = F(w_i)$ may depend on n. A more complete notation would use x_{ni}, w_{ni} and a_{ni} for $1 \leq i \leq n$.

THEOREM 2. *Condition* A *implies Condition* B.

PROOF. Since $N_i \geq 1$, the events $\{N_i \geq i^{(1+\delta)/2}, k_n + 1 \leq i \leq n\}$ implies

$$n^{-1/2} \sum_1^n N_i^{-1} < k_n n^{-1/2} + 2(1-\delta)^{-1} n^{-1/2} \{n^{(1-\delta)/2} - k_n^{(1-\delta)/2}\},$$

which can be made less than arbitrary $\varepsilon > 0$ for k_n and δ as in Condition A, making n large. Hence for large n,

$$P\{n^{-1/2} \sum_1^n N_i^{-1} \geq \varepsilon\} \leq \sum_{k_n+1}^n P\{N_i < i^{(1+\delta)/2}\}.$$

Now $N_i = \sum_{j=1}^i 1\{V_j \leq w_i\}$ with $E(N_i) = \sum_{j=1}^i a_i/a_j$, so that by Theorem 1 of Hoeffding (1963), $P\{N_i < i^{(1+\delta)/2}\} \leq \exp(-2b_i^2/i)$, where $b_i = E(N_i) - i^{(1+\delta)/2}$ for $k_n + 1 \leq i \leq n$ can be made larger than $i^{1/2+\delta}\alpha$ for some $\alpha > 0$ by Condition A. Thus as $n \to \infty$,

$$P\{n^{-1/2} \sum_1^n N_i^{-1} \geq \varepsilon\} \leq \sum_{k_n+1}^n \exp(-2\alpha^2 i^{2\delta}) < \int_{k_n}^n \exp(-2\alpha^2 x^{2\delta})\, dx \to 0.$$

Applying Theorem 4 of the Appendix we have:

THEOREM 3. *Condition* B *implies that* $X_n(t) = n^{1/2}\{H_n(t) - t\}$ *converges weakly to the Brownian bridge* $B^*(t)$ *in Skorokhod topology on* $D[0, 1]$ *as* $n \to \infty$.

A proof of Theorem 3 is outlined in the Appendix.

From Theorems 2 and 3 it follows that under Condition A, the limiting distribution of $T_n = \sup_{0 \leq t \leq 1} n^{1/2}\{H_n(t) - t\}$ is that of the Kolmogorov-Smirnov statistic D_n^+.

REMARK. If instead of the $K - S$ statistic T_n, we want to use some other continuous function $g(X_n(\cdot))$ of the empirical process, then such a statistic is also asymptotically distributed as $g(B^*(\cdot))$ by Theorem 3. For example, linear rank statistics are seen to be asymptotically normal.

5. Acknowledgement. The author remains grateful to Herman Chernoff for posing the problem, pointing out the nature of difficulty, and finally suggesting substantial changes in the original version of the paper for improved presentation.

APPENDIX

We first state an extended version of Theorem 15.6 of Billingsley (1968) which is needed for the weak convergence of the empirical process $X_n(t) = n^{1/2}\{H_n(t) - t\}$.

Let $\{Y_n\}$, Y be random functions in $D[0, 1]$ and $\{\eta_n\}$ a sequence of random vectors of arbitrary dimensions. In our case, $\eta_n = (N_1, \cdots N_n)$.

THEOREM 4. *Suppose that the finite-dimensional distributions* (fdd) *of* $\{Y_n\}$ *converge to those of* Y *and that* Y *is left-continuous at* 1 a.s. *Suppose further that*

$$(1) \qquad P\{|Y_n(t) - Y_n(t_1)| \geq \lambda, |Y_n(t_2) - Y_n(t)| \geq \lambda \,|\, \eta_n\} \leq \lambda^{-2\gamma}\{\varphi_n(t_2) - \varphi_n(t_1)\}^{2\alpha}$$

for $t_1 \leq t \leq t_2$ *and* $n \geq 1$, *where* $\gamma \geq 0$, $\alpha > \frac{1}{2}$, *and* φ_n *are* a.s. *nondecreasing random functions (depending on* η_n*) converging pointwise in probability to a continuous function* φ *on* $[0, 1]$. *Then* $Y_n \to_{\mathcal{D}} Y$.

PROOF. Proceed as in Billingsley's proof to arrive at a conditional counterpart of his (15.30) with φ_n in place of F. To complete the proof, use the fact that by the hypothesis of

the theorem, φ_n actually converges uniformly in probability to φ which is uniformly continuous.

REMARK. The probability inequality (1) is implied by the moment inequality

$$E\{|Y_n(t) - Y_n(t_1)|^\gamma |Y_n(t_2) - Y_n(t)|^\gamma | \eta_n\} \leq \{\varphi_n(t_2) - \varphi_n(t_1)\}^{2\alpha}. \tag{2}$$

PROOF OF THEOREM 3. Let $\eta_n = (N_1, \cdots N_n)$, I_t the indicator function of $[0, t]$ and $J_t(\xi_i, \eta_n) = I_t(\xi_i) - P\{\xi_i \leq t | \eta_n\}$. Then $X_n(t) = Y_n(t) + Z_n(t)$, where

$$Y_n(t) = n^{-1/2} \sum_1^n J_t(\xi_i, \eta_n), Z_n(t) = n^{-1/2} \sum_1^n [P\{\xi_i \leq t | \eta_n\} - t].$$

By Theorem 1 and Condition B, $\sup |Z_n(t)| \leq n^{-1/2} \sum_1^n (2N_i)^{-1} \to_p 0$ and it is enough to show that $Y_n(t) \to_w B^*(t)$.

It is moderately routine to show that

$$\sum_1^r \lambda_j Y_n(t_j) = n^{-1/2} \sum_{i=1}^n \sum_{j=1}^r \lambda_j J_{t_j}(\xi_i, \eta_n) = n^{-1/2} \sum_1^n \zeta_i \to_{\mathscr{L}} N(0, \sigma^2),$$

where $\sigma^2 = \operatorname{Var}\{\sum_1^r \lambda_j B^*(t_j)\}$. The main thing is to observe that $\sigma_n^2 = n^{-1} \sum_1^n \operatorname{Var}(\zeta_i | \eta_n) \to_p \sigma^2$ by Condition B and use the Berry-Esseen bound for the convergence of $P\{n^{-1/2} \cdot \sum_1^n \zeta_i \leq \sigma y | \eta_n\}$.

To complete the proof we show Y_n satisfies the conditions of Theorem 4 with $\gamma = 2$, $\alpha = 1$, $\varphi_n(t) = n^{-1} \sum_1^n P\{\xi_i \leq t | \eta_n\}$ and $\varphi(t) = t$. Inequality (2) is established by easy but tedious algebra which we omit, and $\sup |\varphi_n(t) - \varphi(t)| \leq n^{-1} \sum_1^n (2N_i)^{-1} \to_p 0$ by Condition B.

REFERENCES

BHATTACHARYA, P. K., CHERNOFF, HERMAN and YANG, S. S. (1983) Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.* **11** 505–514.

BHATTACHARYA, P. K. and FRIERSON, DARGAN, JR. (1981). A nonparametric control-chart for detecting small disorders. *Ann Statist.* **9** 544–554.

BILLINGSLEY, PATRICK (1968). *Convergence of Probability Measures.* Wiley, New York.

HOEFFDING, WASSILY (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30.

TURNER, EDWIN L. (1979). Statistics of the Hubble diagram. I. Determination of q_0 and luminosity evolution with application to quasars. *Astrophys. J.* **230** 291–303.

DIVISION OF STATISTICS
UNIVERSITY OF CALIFORNIA
DAVIS, CALIFORNIA 95616