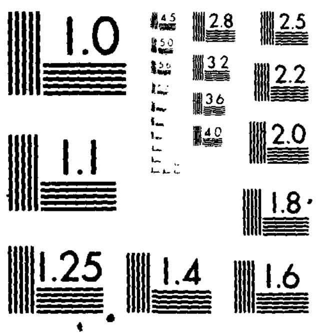
AD-A096 805 MASSACHUSETTS INST OF TECH CAMBRIDGE DEPT OF MATHEMATICS F/G 12/1
NONPARAMETRIC TEST FOR THE SLOPE OF A TRUNCATED REGRESSION. (U)
JAN 81 P K BHATTACHARYA N00014-75-C-0555
UNCLASSIFIED TR-22 NL

1 of 1

END
DATE
FILMED
4-81
DTIC



MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS 1963 A

NONPARAMETRIC TEST FOR THE SLOPE OF A TRUNCATED REGRESSION

by

P. K. BHATTACHARYA
UNIVERSITY OF ARIZONA

TECHNICAL REPORT No. 22

JANUARY 12, 1981

PREPARED UNDER CONTRACT
N00014-75-C-0555 (NR-042-331)
FOR THE OFFICE OF NAVAL RESEARCH

DEPARTMENT OF MATHEMATICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
CAMBRIDGE, MASSACHUSETTS

NONPARAMETRIC TEST FOR THE SLOPE OF A TRUNCATED REGRESSION

(Short Title: TEST FOR A TRUNCATED REGRESSION)

By P. K. Bhattacharya, University of Arizona

ABSTRACT

Consider the hypothesis $H_0: \beta = \beta_0 > 0$ in a linear regression model where the cdf of $Y - \beta x$ is unknown and Y is subject to the truncation $Y \le y_0$. Testing H_0 on the basis of n independent (x_i, Y_i) with $x_1 \le \cdots \le x_n$ is equivalent to testing the underlying homogeneity of the independent $V_i = Y_i - \beta_0 x_i$ subject to progressive truncation $V_i \le w_i = y_0 - \beta_0 x_i$. For analyzing astronomical observations a test has been proposed in the literature, which computes the sequential ranks R_i of V_i among N_i "comparable" V_j, $j \le i$ satisfying $V_j \le w_i$ and compares the empirical cdf $H_n(t)$ of $(2R_i - 1)/2N_i$ with t by a K - S statistic. Since $(2R_i - 1)/2N_i$ are neither indeoendent nor exactly uniform $[0,1]$, the applicability of the usual asymptotic null distribution of the K - S statistic in this context needs justification which is provided in this paper under a sufficient condition requiring that the rate of progressive truncation is not too severe.

---

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

Key Words: Truncated Regression, Hubble Diagram, Testing Homogeneity Under Progressive Truncation, Kolmogorov-Smirnov Test, Sequential Rank, Asymptotic Distribution, Brownian Bridge.

AMS 1980 Subject Classifications: Primary 62G10, Secondary 62E20, 62J05

1. Introduction

Consider a linear regression model in which $Y_j^* - \beta x_j^*$, $j = 1, 2, \ldots$ are iid with cdf F. Suppose we observe (x_j^*, Y_j^*) only if $Y_j^* \leq y_0$ and let $(x^*_{j_i}, Y^*_{j_i}) = (x_i, Y_i)$, $i = 1, 2, \ldots$ denote the observable pairs. Then $Y_1, Y_2, \ldots$ are independent and

$$P[Y_i \leq y] = F(\min(y, y_0) - \beta x_i)/F(y_0 - \beta x_i).$$

F is continuous but otherwise unknown, y_0 is a known constant and we are interested in inference about the regression coefficient β.

In the non-truncated case, Theil (1950) proposed a nonparametric estimate of β based on Kendall's tau and Sen (1968) derived its asymptotic properties. Bhattacharya, Yang and Chernoff (1980) developed a modification of Theil's estimate to suit the truncated case. In an entirely different approach to the truncated regression problem, Turner (1979) used a Kolmogorov-Smirnov test based on sequential ranks for testing $H_0: \beta = \beta_0 > 0$ in analyzing astronomical observations where truncated scatter diagrams called Hubble diagrams are obtained as plots of luminosity distance versus redshift of various celestial objects.

To describe the test statistic in a sample of size n, arrange $x_1, \ldots, x_n$ as $x_{n1} \leq \cdots \leq x_{nn}$ (equal x's being arranged arbitrarily) and define $Y_{ni} = Y_k$ if $x_{ni} = x_k$. Under H_0, the residuals $V_{ni} = Y_{ni} - \beta_0 x_{ni}$ are iid observations subject to progressive truncation $V_{ni} \leq w_{ni} = y_0 - \beta_0 x_{ni}$ with $w_{n1} \geq \cdots \geq w_{nn}$ since β_0 is positive. Each V_{ni}

is comparable to only those V_{nj}, $j < i$ for which $V_{nj} \le w_{ni}$, because the conditional distribution of V_{nj} given $V_{nj} \le w_{ni}$ is the same as the distribution of V_{ni} under H_0. Consequently, under H_0, the rank of each V_{ni} among comparable V_{nj}, $j \le i$, is uniformly distributed over all possibilities. Formally, letting

$$S_{ni} = \{j \le i : V_{nj} \le w_{ni}\}, \quad N_{ni} = \# S_{ni},$$

(1)

$$R_{ni} = \text{rank of } V_{ni} \text{ among } \{V_{nj} : j \in S_{ni}\}, \quad \xi_{ni} = (2R_{ni} - 1)/2N_{ni},$$

Turner heuristically argues that under H_0, the ξ_{ni} are asymptotically unif [0,1], so that their empirical cdf $H_n(t)$ should closely resemble t, and $D_n^+ = \sqrt{n} \sup_{0 \le t \le 1} [H_n(t) - t]$ can be used for a one-sided test of H_0.

This argument can be justified when there is no truncation, because then $N_{ni} = i$ and R_{ni} are independent with $P[R_{ni} = r] = 1/r$, $1 \le r \le i$, as shown by Parent (1965) and Bhattacharya and Frierson (1981). Under truncation, the validity of D_n^+ as a test statistic is not so obvious, because the behavior of N_{ni} and R_{ni} become more complicated and the ξ_{ni} become dependent. Yet a Monte Carlo study supported the fact that the asymptotic null distribution of D_n^+ is what it should be if ξ_{ni} were iid unif [0,1].

In this study, x_{ni} were taken to be i/n, $\beta = 1$ and Y_{ni} were unif [0,c] + βx_{ni} subject to the truncation $Y_{ni} \le y_0$ where $c > y_0 > 1$. This was achieved by generating V_{ni} as unif $[0, w_{ni}]$ with $w_{ni} = y_0 - i/n$. Fifteen cases involving $y_0 = 2.0$, 1.5, 1.1 and $n = 10$, 25, 50, 100, 200 were considered with 100 runs for each case and D_n^+ was computed for each run. The asymptotic cdf of D_n^+ from independent unif [0,1] is $1 - e^{-2z^2}$ having mean $\sqrt{\pi/8} = 0.6267$, st. dev. $\sqrt{(4-\pi)/8} = 0.3276$ and upper 10% and 5% points 1.07298 and 1.22387. The relative frequencies of D_n^+ lying above the theoretical percentage points and their means and standard derivations for the fifteen cases are given in the following table. Though the Monte Carlo experiments tend to give rise to somewhat smaller values then expected from the theoretical distribution due to discreteness of the ξ_{ni}, there is strong indication (especially in respect of the mean) that the asymptotic behavior of D_n^+ is not affected by truncation.

The purpose of this paper is to examine how far Turner's procedure can be justified in the presence of truncation so as to explain the above Monte Carlo results. The main result is a sufficient condition for the applicability of the usual asymptotic null distribution in this context which requires

$$\liminf_{n\to\infty} \min_{k_n < i \le n} i^{-(1/2+\delta)} \sum_{j=1}^{i} F(w_{ni})/F(w_{nj}) > 0$$

for some $0 < \delta \le \frac{1}{2}$ and some sequence $\{k_n\}$ such that $k_n \to \infty$ and $n^{-1/2}k_n \to 0$ as $n \to \infty$. The condition means that the rate of progressive truncation is not too severe.

In the above Monte Carlo experiments, V_{ni} were unif $[0, w_{ni}]$ with $w_{ni} = y_0 - i/n$. Letting $t = i/ny_0$ and $h(t) = t^{1/2-\delta}(1-t)$, we have

$$i^{-(1/2+\delta)} \sum_{j=1}^{i} F(w_{ni})/F(w_{nj}) = i^{-(1/2+\delta)}(1 - i/ny_0) \sum_{j=1}^{i} (1 - j/ny_0)^{-1}$$

$$> (ny_0)^{1/2-\delta} t^{-(1/2+\delta)}(1-t)\log(1-t)^{-1} \ge (ny_0)^{1/2-\delta} h(t).$$

Table. Monte Carlo results on the behavior of D_n^+ from independent observations on unif $[0, y_0 - i/n]$, $1 \leq i \leq n$: Each case based on 100 runs.

n	y_0	Relative Frequency: Above 1.07298	Relative Frequency: Above 1.22387	Mean	St. Dev
10	2.0	.08	.06	.569	.289
	1.5	.02	.02	.564	.251
	1.1	.08	.06	.627	.291
25	2.0	.08	.04	.572	.296
	1.5	.08	.04	.578	.276
	1.1	.07	.06	.589	.294
50	2.0	.07	.03	.601	.297
	1.5	.06	.04	.618	.307
	1.1	.09	.05	.598	.307
100	2.0	.13	.08	.680	.327
	1.5	.15	.07	.684	.327
	1.1	.12	.06	.672	.315
200	2.0	.08	.02	.582	.288
	1.5	.08	.05	.590	.308
	1.1	.12	.03	.614	.309
All Cases		.0873	.0473	.609	.301

Hence for $\delta < \frac{1}{2}$ and $k_n \to \infty$ at a rate slower than $n^{-1/2}$,

$$\min_{k_n < i \leq n} i^{-(1/2+\delta)} \sum_{j=1}^{i} F(w_{ni})/F(w_{nj}) > (ny_0)^{1/2-\delta} \min_{k_n/ny_0 < t \leq 1/y_0} h(t)$$

$$= (ny_0)^{1/2-\delta} \min\{h(k_n/ny_0), h(1/y_0)\}$$

$$= \min\{k_n^{1/2-\delta}(1 - k_n/ny_0), n^{1/2-\delta}(1 - 1/y_0)\}$$

not only stays positive but tends to $+\infty$ as $n \to \infty$. This explains why the effect of truncation appears to be negligible in these experiments.

2. Joint Distribution of $\{N_{ni}, R_{ni}, 1 \leq i \leq n\}$.

Let $U_{ni} = F(V_{ni})$ and $a_{ni} = F(w_{ni})$. Then $1 \geq a_{n1} \geq \dots \geq a_{nn} > 0$ and U_{ni} are independent unif $[0, a_{ni}]$. Formulas (1) for S_{ni}, N_{ni}, R_{ni}, ξ_{ni} are equivalently expressed by substituting U_{ni} and a_{ni} for V_{ni} and w_{ni} respectively. We now obtain the distribution of $\{N_{ni}\}$ and the conditional distribution of $\{R_{ni}\}$ given $\{N_{ni}\}$.

Theorem 1. For each n, $\{N_{ni}\}$ is a Markov chain with

$$P[N_{n,i+1} = k \mid N_{ni} = m] = \binom{m}{k-1}(a_{n,i+1}/a_{ni})^{k-1}(1 - a_{n,i+1}/a_{ni})^{m-k+1}$$

for $1 \leq k \leq m - 1$, starting at $N_{n1} \equiv 1$ and for given $\{N_{ni}\}$, R_{ni} are conditonally uniformly distributed on $\{1, \dots, N_i\}$ and are conditionally independent.

Fix n and suppress n in the subscripts of a_{ni}, U_{ni}, S_{ni}, N_{ni} and R_{ni}. Thus $1 \geq a_1 \geq \dots \geq a_n > 0$, $U_1, \dots, U_n$ are independent unif $[0,a_i]$. $S_i = \{j \leq i : U_j \leq a_i\}$, $N_i = \# S_i$ and R_i is the rank of U_i among $\{U_j : j \in S_i\}$. For $N_i = m$ and $S_i = \{\nu_{i1}, \ldots, \nu_{im}\}$ with $\nu_{i1} < \ldots < \nu_{im} = i$, let $U^*_{ij} = U_{\nu_{ij}}$ identify the random variables $\{U_j : j \in S_i\}$ in their appropriate order in each of the random sets S_i. For brevity of notations, denote the collection N_i, S_i and $\{U_j : j < i, j \notin S_i\}$ by C_i. Let $R_{i1}, \ldots, R_{iN_i}$ denote the sequential ranks and $\hat{R}_{i1}, \ldots, \hat{R}_{iN_i}$ the usual ranks of $U^*_{i1}, \ldots, U^*_{iN_i}$, i.e., R_{ij} and $\hat{R}_{ij}$ are the ranks of U^*_{ij} among $U^*_{i1}, \ldots, U^*_{ij}$ and $U^*_{i1}, \ldots, U^*_{iN_i}$ respectively. Clearly, $R_i = R_{iN_i} = \hat{R}_{iN_i}$. The following property of the U^*_{ij}'s is immediate.

<u>Lemma 1</u>. $U^*_{i1}, \ldots, U^*_{iN_i}$ are conditionally iid unif $[0,a_i]$ given C_i.

From Lemma 1 we draw the following conclusions.

<u>Lemma 2</u>. $P[R_i = r | C_i, R_{i1}, \ldots, R_{i,N_i-1}] = N_i^{-1} = P[R_i = r | N_i]$, $1 \leq r \leq N_i$.

<u>Proof</u>. By Lemma 1, $R_{i1}, \ldots, R_{i,N_i-1}$ and $R_{iN_i} = R_i$ are sequential ranks of random variables which are conditionally iid given C_i and the first equality follows from a known property of sequential ranks of iid rv's (see Bhattacharya and Frierson (1981), Lemma 1). The second equality is obvious because this conditional probability depends only on N_i.

Lemma 3. $P[N_{i+1} = k | C_i, R_{i1}, \ldots, R_{i,N_i-1}, R_i]$

$$= \binom{N_i}{k-1}(a_{i+1}/a_i)^{k-1}(1 - a_{i+1}/a_i)^{N_i-k+1} = P[N_{i+1} = k | N_i] \text{ for } 1 \le k \le N_i + 1.$$

Proof. The order statistics $(U^*_{i(1)}, \ldots, U^*_{i(N_i)})$ of $U^*_{i1}, \ldots, U^*_{iN_i}$ are conditionally independent of their ranks $(\hat{R}_{i1}, \ldots, \hat{R}_{iN_i})$ and hence of their sequential ranks $(R_{i1}, \ldots, R_{i,N_i-1}, R_i)$ given C_i since the two rank vectors are in one-one correspondence. The first equality now follows because $N_{i+1} = k$ if and only if $U^*_{i(k-1)} \le a_{i+1} < U^*_{i(k)}$ and the second equality is obtained as in Lemma 1.

The next lemma follows by standard arguments involving conditional expectations and we omit its proof.

Lemma 4. Let X, Y, Z, T be random variables mapping a probability space into appropriate ranges. Suppose T and Y are conditionally independent given X, and Z is determined by X and Y. Then T and Z are conditionally independent given X.

Proof of Theorem 1. The crucial thing to observe is that $(N_1, R_1), \ldots, (N_{i-1}, R_{i-1})$ are determined by C_i and $R_{i1}, \ldots, R_{i,N_i-1}$. Hence $(N_1, R_1), \ldots, (N_{i-1}, R_{i-1}), N_i$ are determined by C_i, $R_{i1}, \ldots, R_{i,N_i-1}$ and $(N_1, R_1), \ldots, (N_i, R_i)$ are determined by C_i, $R_{i1}, \ldots, R_{i,N_i-1}$, R_i. Using this fact in conjunction with Lemmas 2, 3 and 4, we conclude that for $2 \le i \le n$,

$$P[R_i = r|(N_1,R_1), \ldots, (N_{i-1},R_{i-1}),N_i] = P[R_i = r|N_i] = N_i^{-1}, \quad i \leq r \leq N_i$$

and

$$P[N_{i+1} = k|(N_i,R_1), \ldots, (N_i,R_i)] = P[N_{i+1} = k|N_i]$$

$$= \binom{N_i}{k-1}(a_{i+1}/a_i)^{k-1}(1 - a_{i+1}/a_i)^{N_i-k+1}, \quad 1 \leq k \leq N_i + 1.$$

The theorem now follows because (N_1,R_1) is trivially $(1,1)$ and for $k = 1$ or 2, $P[N_2 = k|N_1,R_1] = P[N_2 = k]$ is given by $P[U_1 > a_2] = 1 - a_2/a_1$ and $P[U_1 \leq a_2] = a_2/a_1$ respectively.

3. Asymptotics

By Theorem 1, $\xi_{ni} = (2R_{ni} - 1)/2N_{ni}$ are conditionally independent given $\eta_n = (N_{n1}, \ldots, N_{nn})$ with conditional cdf $G_{ni}(t|\eta_n)$ increasing by jumps of $1/N_{ni}$ at $(2r - 1)/2N_{ni}$, $1 \leq r \leq N_{ni}$. The question is, how small should the $\varepsilon_{ni}(t,\eta_n) = G_{ni}(t|\eta_n) - t$ be so that the normalized empirical cdf $H_n(t)$ of the ξ_{ni}, viz. $X_n(t) = n^{-1/2}[H_n(t) - t]$ will behave like the Brownian bridge $B^*(t)$ on $[0,1]$? We discuss weak convergence of the empirical cdf of conditionally independent random variables in the Appendix, which may be of some interest in itself, and show that $X_n(t)$ converges weakly to $B^*(t)$ provided that $n^{-1/2}\sum_1^n|\varepsilon_{ni}(t,\eta_n)|$ converges uniformly to 0 in probability (Theorem 4). In the present context, $|\varepsilon_{ni}(t,\eta_n)| \leq (2N_{ni})^{-1}$. Hence a sufficient condition for the desired convergence is $n^{-1/2}\sum_1^n N_{ni}^{-1} \xrightarrow{P} 0$ as $n \to \infty$. We summarize this in the following theorem.

Lemma 3. $P[N_{i+1} = k | C_i, R_{i1}, \ldots, R_{i,N_i-1}, R_i]$

$= \binom{N_i}{k-1}(a_{i+1}/a_i)^{k-1}(1 - a_{i+1}/a_i)^{N_i-k+1} = P[N_{i+1} = k|N_i]$ for $1 \le k \le N_i + 1$.

Proof. The order statistics $(U^*_{i(1)}, \ldots, U^*_{i(N_i)})$ of $U^*_{i1}, \ldots, U^*_{iN_i}$ are conditionally independent of their ranks $(\hat{R}_{i1}, \ldots, \hat{R}_{iN_i})$ and hence of their sequential ranks $(R_{i1}, \ldots, R_{i,N_i-1}, R_i)$ given C_i since the two rank vectors are in one-one correspondence. The first equality now follows because $N_{i+1} = k$ if and only if $U^*_{i(k-1)} \le a_{i+1} < U^*_{i(k)}$ and the second equality is obtained as in Lemma 1.

The next lemma follows by standard arguments involving conditional expectations and we omit its proof.

Lemma 4. Let X, Y, Z, T be random variables mapping a probability space into appropriate ranges. Suppose T and Y are conditionally independent given X, and Z is determined by X and Y. Then T and Z are conditionally independent given X.

Proof of Theorem 1. The crucial thing to observe is that (N_1,R_1), $\ldots$, (N_{i-1},R_{i-1}) are determined by C_i and $R_{i1}, \ldots, R_{i,N_i-1}$. Hence $(N_1,R_1), \ldots, (N_{i-1},R_{i-1}), N_i$ are determined by C_i, $R_{i1}, \ldots, R_{i,N_i-1}$ and $(N_1,R_1), \ldots, (N_i,R_i)$ are determined by C_i, $R_{i1}, \ldots, R_{i,N_i-1}$, R_i. Using this fact in conjunction with Lemmas 2, 3 and 4, we conclude that for $2 \le i \le n$,

$$P[R_i = r|(N_1,R_1), \ldots, (N_{i-1},R_{i-1}),N_i] = P[R_i = r|N_i] = N_i^{-1}, \quad 1 \le r \le N_i$$

and

$$P[N_{i+1} = k|(N_1,R_1), \ldots, (N_i,R_i)] = P[N_{i+1} = k|N_i]$$

$$= \binom{N_i}{k-1}(a_{i+1}/a_i)^{k-1}(1 - a_{i+1}/a_i)^{N_i-k+1}, \quad 1 \le k \le N_i + 1.$$

The theorem now follows because (N_1,R_1) is trivially $(1,1)$ and for $k = 1$ or 2, $P[N_2 = k|N_1,R_1] = P[N_2 = k]$ is given by $P[U_1 > a_2] = 1 - a_2/a_1$ and $P[U_1 \le a_2] = a_2/a_1$ respectively.

3. Asymptotics

By Theorem 1, $\xi_{ni} = (2R_{ni} - 1)/2N_{ni}$ are conditionally independent given $\eta_n = (N_{n1}, \ldots, N_{nn})$ with conditional cdf $G_{ni}(t|\eta_n)$ increasing by jumps of $1/N_{ni}$ at $(2r - 1)/2N_{ni}$, $1 \le r \le N_{ni}$. The question is, how small should the $\varepsilon_{ni}(t,\eta_n) = G_{ni}(t|\eta_n) - t$ be so that the normalized empirical cdf $H_n(t)$ of the ξ_{ni}, viz. $X_n(t) = n^{-1/2}[H_n(t) - t]$ will behave like the Brownian bridge $B^*(t)$ on $[0,1]$? We discuss weak convergence of the empirical cdf of conditionally independent random variables in the Appendix, which may be of some interest in itself, and show that $X_n(t)$ converges weakly to $B^*(t)$ provided that $n^{-1/2} \sum_1^n |\varepsilon_{ni}(t,\eta_n)|$ converges uniformly to 0 in probability (Theorem 4). In the present context, $|\varepsilon_{ni}(t,\eta_n)| \le (2N_{ni})^{-1}$. Hence a sufficient condition for the desired convergence is $n^{-1/2} \sum_1^n N_{ni}^{-1} \xrightarrow{p} 0$ as $n \to \infty$. We summarize this in the following theorem.

Theorem 2. As $N \to \infty$, $X_n(t) = n^{-1/2}[H_n(t) - t]$ converges weakly to the Brownian bridge $B^*(t)$ in Skorokhod topology on $D[0,1]$ provided that $n^{-1/2} \sum_1^n N_{ni}^{-1} \xrightarrow{p} 0$.

We now derive a condition in terms of severity of the rate of progressive truncations, which guarantees the above convergence.

Theorem 3. $n^{-1/2} \sum_1^n N_{ni}^{-1} \xrightarrow{p} 0$ as $n \to \infty$ provided that there is a sequence $k_n \to \infty$ with $k_n n^{-1/2} \to 0$ such that for some $\delta > 0$,

$$\text{(2)} \qquad \liminf_{n\to\infty} \min_{k_n < i \le n} i^{-1/2-\delta} \sum_{j=1}^{i} a_{ni}/a_{nj} > 0.$$

Proof. For arbitrary $\varepsilon > 0$, choose n so large that

$$\text{(3)} \qquad 2(1-\delta)^{-1} n^{-\delta/2}\{1 - (k_n n^{-1})^{(1-\delta)/2}\} < \varepsilon - k_n n^{-1/2}$$

holds for k_n and $\delta < 1$ satisfying the hypothesis of the theorem. Since $N_{ni} \ge 1$, the event $\{N_{ni} \ge i^{(1+\delta)/2}, k_n + 1 \le i \le n\}$ implies

$$n^{-1/2} \sum_1^n N_{ni}^{-1} \le k_n n^{-1/2} + n^{-1/2} \sum_{k_n+1}^{n} i^{-(1+\delta)/2}$$

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

by (3). Hence for sufficiently large n,

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

From the definition of S_{ni} given in (1) it is clear that

$$N_{ni} = \# S_{ni} = \sum_{j=1}^i I_{(-\infty, w_{ni}]}(V_{nj}) = \sum_{j=1}^i Z_{n,ji},$$

where for each i, the indicator random variables $Z_{n,1i}, \ldots, Z_{n,ii}$ are independent with means $F(w_{ni})/F(w_{nj}) = a_{ni}/a_{nj}$, $1 \leq j \leq i$, so that $EN_{ni} = \sum_{j=1}^i a_{ni}/a_{nj}$. By (2), there exists $\alpha > 0$ such that for sufficiently large n,

$$b_{ni} = \sum_{j=1}^i a_{ni}/a_{nj} - i^{(1+\delta)/2} > i^{1/2+\delta}\alpha > 0, \quad k_n + 1 \leq i \leq n.$$

Using Theorem 1 of Hoeffding (1963) we now have

(5) $$P[N_{ni} < i^{(1+\delta)/2}] = P[N_{ni} - EN_{ni} < -b_{ni}]$$

$$\leq \exp[-2b_{ni}^2/i] < \exp[-2\alpha^2 i^{2\delta}]$$

for $k_n + 1 \leq i \leq n$. From (4) and (5) it now follows that

$$P[n^{-1/2}\sum_1^n N_{ni}^{-1} \geq \varepsilon] \leq \sum_{k_n+1}^n \exp[-2\alpha^2 i^{2\delta}] < \int_{k_n}^n \exp[-2\alpha^2 x^{2\delta}]dx$$

which converges to 0 as $n \to \infty$ because $\int_0^\infty \exp[-2\alpha^2 x^{2\delta}]dx < \infty$.

4. Acknowledgement

The Monte Carlo results reported in the Introduction were obtained by Herman Chernoff who kindly made them available to the author and also pointed out the need to explain them.

Appendix

Consider a sequence of random vectors $\{\eta_n\}$ and a triangular array of random variables $\{\xi_{ni}, 1 \leq i \leq n\}$, $n = 1, 2, \ldots$ such that for each n, the ξ_{ni} are conditionally independent given η_n with conditional cdf's $G_{ni}(t|\eta_n)$. We obtain a sufficient condition for weak convergence of appropriately normalized empirical cdf's of $\{\xi_{ni}\}$, viz.

$$X_n(t) = n^{1/2}[n^{-1} \sum_1^n I_{[0,t]}(\xi_{ni}) - t], \quad 0 \leq t \leq 1$$

to the Brownian bridge $\{B^*(t), 0 \leq t \leq 1\}$ in Skorokhod topology on $D[0,1]$. Write

$$X_n(t) = n^{-1/2} \sum_1^n J_{[0,t]}(\xi_{ni}, \eta_n) + n^{-1/2} \sum_1^n \varepsilon_{ni}(t, \eta_n),$$

where $J_{[0,t]}(\xi_{ni}, \eta_n) = I_{[0,t]}(\xi_{ni}) - G_{ni}(t|\eta_n)$, $1 \leq i \leq n$ are conditionally independent given η_n with conditional mean 0 and conditional covariances $G_{ni}(s|\eta_n)\{1 - G_{ni}(t|\eta_n)\}$ for $0 \leq s \leq t \leq 1$, and $\varepsilon_{ni}(t, \eta_n) = G_{ni}(t|\eta_n) - t$. Suppose the following condition holds.

<u>Condition A</u>. $\sup_t n^{-1/2} \sum_1^n |\epsilon_{ni}(t,\eta_n)| \xrightarrow{P} 0$ as $n \to \infty$.

Then $n^{-1/2} \sum_1^n \epsilon_{ni}(t,\eta_n)$ converges to 0 in probability uniformly and hence in Skorokhod topology, so that it is enough to show that

$$Y_n(t) = n^{-1/2} \sum_1^n J_{[0,t]}(\xi_{ni},\eta_n) \xrightarrow{\mathcal{D}} B^*(t) \tag{6}$$

in Skorokhod topology on $D[0,1]$. The following theorem says that we need nothing more than Condition A for the required convergence.

<u>Theorem 4</u>. Under Condition A, $X_n(t) \xrightarrow{\mathcal{D}} B^*(t)$ in Skorokhod topology on $D[0,1]$.

To prove Theorem 4, we need the following extended version of Theorem 15.6 of Billingsley (1968).

<u>Theorem 5</u>. 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

$$E[|Y_n(t) - Y_n(t_1)|^\gamma |Y_n(t_2) - Y_n(t)|^\gamma | \eta_n] \le [\phi_n(t_2) - \phi_n(t_1)]^{2\alpha} \tag{7}$$

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

Proof of Theorem 5. Using Billingsley's notation, we need to show that for given ϵ, η_1 we can find δ such that $P[w''(Y_n,\delta) \geq \epsilon] \leq \eta_1$ for sufficiently large n. Proceeding as in Billingsley's proof we arrve at a counterpart of his (15.29) with ϕ_n in place of F, from which it follows that for $\delta = (2u)^{-1}$, u positive integer,

$$P[w''(Y_n,\delta) \geq \epsilon | \eta_n] \leq 2K\epsilon^{-2\gamma}[\phi_n(1) - \phi_n(0)]V_n^{2\alpha-1},$$

$$V_n = \max\{\max_{0 \leq i \leq u-1} [\phi_n((2i+2)\delta) - \phi_n(2i\delta)], \max_{0 \leq i \leq u-2} [\phi_n((2i+3)\delta) - \phi_n((2i+1)\delta)]\}$$

For given ϵ', let $B_{n\epsilon'} = \{\eta_n : \sup_t |\phi_n(t) - \phi(t)| \leq \epsilon'\}$. By the hypothesis of the theorem, ϕ_n actually converges uniformly in probability to ϕ, so $P[B^c_{n\epsilon'}] < \eta_1/2$ for large n. Now choose δ such that $|\phi(s) - \phi(t)| < \epsilon'$ for $|s - t| \leq 2\delta$. Then for $\eta_n \in B_{n\epsilon'}$, $|\phi_n(s) - \phi_n(t)| \leq 3\epsilon'$ for $|s - t| \leq 2\delta$, so that $V_n \leq 3\epsilon'$ and $\phi_n(1) - \phi_n(0) \leq \phi(1) - \phi(0) + 2\epsilon'$ $\leq 2[\phi(1) - \phi(0)]$ for small ϵ'. Thus for $\eta_n \in B_{n\epsilon'}$,

$$P[w''(Y_n,\delta) \geq \epsilon | \eta_n] \leq 4K\epsilon^{-2\gamma}[\phi(1) - \phi(0)](3\epsilon')^{2\alpha-1}.$$

Choosing ϵ' so that the RHS of the last inequality is $\leq \eta_1/2$, we now have

$$P[w''(Y_n,\delta) \geq \epsilon] \leq 4K\epsilon^{-2\gamma}[\phi(1) - \phi(0)](3\epsilon')^{2\alpha-1} + P[B^c_{n\epsilon'}] < \eta_1,$$

and the thoerem is proved.

Proof of Theorem 4. By Condition A, it suffices to show that (6) holds. We first show the convergence of ffd. Fix a positive integer r, $0 \le t_1 \le \cdots \le t_r \le 1$, $\lambda_1, \ldots, \lambda_r$ real and write $\zeta_{ni} = \sum_{j=1}^{r} \lambda_j J_{[0,t_j]}(\xi_{ni}, \eta_n)$. Then $\zeta_{n1}, \ldots, \zeta_{nn}$ are conditionally independent given η_n with $E(\zeta_{ni} \mid \eta_n) = 0$, $|\zeta_{ni}|$ are bounded by $r^{1/2}\|\lambda\| = (r \sum_1^r \lambda_j^2)^{1/2}$ and $\sum_1^r \lambda_j Y_n(t_j) = n^{-1/2}\sum_1^n \zeta_{ni}$. Using $\mathrm{Cov}[B^*(s), B^*(t)] = s(1-t)$, $0 \le s \le t \le 1$, we have

$$\sigma^2 = \mathrm{Var}\Big[\sum_1^r \lambda_j B^*(t_j)\Big] = \sum_{j=1}^{r} \lambda_j^2 t_j(1-t_j) + 2 \sum_{1 \le j < j' \le r} \lambda_j \lambda_{j'} t_j (1 - t_{j'}),$$

and $\sigma_n^2 = \mathrm{Var}[\sum_1^r \lambda_j Y_n(t_j) \mid \eta_n] = n^{-1}\sum_1^n \mathrm{Var}[\zeta_{ni} \mid \eta_n]$ is obtained by substituting

$$n^{-1}\sum_{i=1}^{n} G_{ni}(t_j \mid \eta_n)\{1 - G_{ni}(t_{j'} \mid \eta_n)\}$$

$$= n^{-1}\sum_{i=1}^{n} \{t_j + \varepsilon_{ni}(t_j, \eta_n)\}\{(1 - t_{j'}) - \varepsilon_{ni}(t_{j'}, \eta_n)\}$$

for $t_j(1 - t_{j'})$ in the formula for σ^2. Since $\sum_1^r \lambda_j B^*(t_j)$ is $N(0, \sigma^2)$, we only need to show that

$$\text{(8)} \qquad \lim_{n\to\infty} P\Big[n^{-1/2}\sum_1^n \zeta_{ni} \le \sigma y\Big] = \Phi(y) \quad \text{for all } y,$$

where Φ is the cdf of $N(0,1)$. For arbitrary $\varepsilon > 0$, let

$A_{n\varepsilon} = \{\eta_n : n^{-1/2}\sum_{i=1}^{n}|\varepsilon_{ni}(t_j,\eta_n)| \le \varepsilon, 1 \le j \le r\}$. For $\eta_n \in A_{n\varepsilon}$, $|\sigma_n^2/\sigma^2 - 1| \le K_1\varepsilon$ with $K_1 = (6r-4)\|\lambda\|^2/\sigma^2$. Hence for small ε, σ/σ_n lies between $1 \pm K_1\varepsilon$, so that

(9) $$|\Phi(\sigma y/\sigma_n) - \Phi(y)| \le \Phi(y + K_1|y|\varepsilon) - \Phi(y - K_1|y|\varepsilon).$$

On the other hand, for $\eta_n \in A_{n\varepsilon}$ with small ε, we use the bound on ζ_{ni} to obtain the Berry-Esseen bound

(10) $$|P[n^{-1/2}\sum_1^n \zeta_{ni} \le \sigma y|\eta_n] - \Phi(\sigma y/\sigma_n)|$$

$$\le n^{-1/2}C\{K_2/(1 - K_1\varepsilon)\}^{3/2} \le 2n^{-1/2}CK_2$$

with $K_2 = r\|\lambda\|^2\sigma^2$ and a universal constant C. Since $\varepsilon > 0$ is arbitrary, (8) now follows from (9), (10) and the fact that $\lim_{n\to\infty} P[A_{n\varepsilon}] = 1$ by Condition A. To complete the proof of Theorem 4, we only need to verify that Y_n satisfies the conditions of Theorem 5 with $\gamma = 2$, $\alpha = 1$, $\phi_n(t) = n^{-1}\sum_1^n G_{ni}(t|\eta_n)$ and $\phi(t) = t$. The conditional moment inequality (7) is established by easy but tedious algebra which we omit and $n^{-1}\sum_1^n G_{ni}(t|\eta_n) - t = n^{-1}\sum_1^n \varepsilon_{ni}(t,\eta_n)$ uniformly converges to 0 in probability by Condition A.

References

[1] Bhattacharya, P. K., Yang, S. S. and Chernoff, Herman. (1980). Nonparametric estimate of the slope of a truncated regression. Tech. Rep. No. 18, Massachusetts Institute of Technology.

[2] Bhattacharya, P. K. and Frierson, Dargan, Jr. (1981). A nonparametric control-chart for detecting small disorders. Ann. Statist. to appear in the July issue.

[3] Billingsley, Patrick (1968). Convergence of Probability Measures. Wiley, New York.

[4] Hoeffding, Wassily (1963). Probability inequalities for sums of bounded random variables. J. Amer. Statist. Assoc. 58, 13-30.

[5] Parent, E. A., Jr. (1965). Sequential ranking procedures. Doctoral Dissertation, Stanford University.

[6] Sen, P. K. (1968). Estimates of the regression coefficient based on Kendall's tau. J. Amer. Statist. Assoc. 63, 1379-1389.

[7] Theil, H. (1950). A rank invariant method of linear and polynomial regression analysis, I, II and III, Nederl. Akad. Wetensch. Proc. 53, 386-392, 521-525, 1397-1412.

[8] Turner, Edwin L. (1979). Statistics of the Hubble diagram. I. Determination of q_0 and luminosity evolution with application to quasars. Astrophys. Jour. 230, 291-303.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER 22	2. GOVT ACCESSION NO. AD-A096 805	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Nonparametric Test for the Slope of a Truncated Regression		5. TYPE OF REPORT & PERIOD COVERED Technical Report 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) P. K. Bhattacharya, University of Arizona		8. CONTRACT OR GRANT NUMBER(s) N00014-75-C-0555
9. PERFORMING ORGANIZATION NAME AND ADDRESS Department of Mathematics Massachusetts Institute of Technology Cambridge, Massachusetts 02139		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS (NR-042-331)
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Statistics & Probability Program Code 436 Arlington, Virginia 22217		12. REPORT DATE January 12, 1981 13. NUMBER OF PAGES 17
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified 15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Truncated Regression, Hubble Diagram, Testing Homogeneity, Under Progressive Truncation, Kolmogorov-Smirnov Test, Sequential Rank, Asymptotic Distribution, Brownian Bridge.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

(see reverse side)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

20.

Consider the hypothesis $H_0: \beta = \beta_0 > 0$ in a linear regression model where the cdf of $Y - \beta x$ is unkown and Y is subject to the truncation $Y \leq y_0$. Testing H_0 on the basis of n independent (x_i, Y_i) with $x_1 \leq \ldots \leq x_n$ is equivalent to testing the underlying homogeneity of the independent $V_i = Y_i - \beta_0 x_i$ subject to progressive truncation $V_i \leq w_i = y_0 - \beta_0 x_i$. For analyzing astronomical observations a test has been proposed in the literature, which computes the sequential ranks R_i of V_i among N_i "comparable" V_j, $j \leq i$ satisfying $V_j \leq w_i$ and compares the empirical cdf $H_n(t)$ of $(2R_i - 1)/2N_i$ with t by a K - S statistic. Since $(2R_i - 1)/2N_i$ are neither independent nor exactly uniform $[0,1]$, the applicability of the usual asymptotic null distribution of the K - S statistic in this context needs justification which is provided in this paper under a sufficient condition requiring that the rate of progressive truncation is not too severe.

DATE

LMED

8