The Annals of Statistics
1976, Vol. 4, No. 3, 621-624

AN INVARIANCE PRINCIPLE IN REGRESSION ANALYSIS[1]

BY P. K. BHATTACHARYA

University of Arizona

The sample paths of cumulative sums of induced order statistics obtained from n independent two-dimensional random vectors, when appropriately normalized, converge weakly (as n increases indefinitely) to the sum of a Brownian motion with time change and an integrated Brownian bridge which is independent of the Brownian motion. Applications in regression analysis are given.

1. Introduction. Let (X_i, Y_i), $i = 1, 2, \cdots$ be i.i.d. as (X, Y). We assume the marginal cdf F of X is continuous and let $m(x) = E(Y \mid X = x)$. Let $X_{n1} < \cdots < X_{nn}$ denote the ordered X_i's and define induced order statistics $Y_{n1}, \cdots, Y_{nn}$ as $Y_{ni} = Y_j$ if $X_{ni} = X_j$; thus $E(Y_{n,[nt]})$ will be approximated, under regularity conditions, by $h(t) = m \circ F^{-1}(t)$ for large n. Define $H(t) = \int_0^t h(s)\, ds$. Natural estimates of H are $H_n(t) = n^{-1} \sum_1^{[nt]} Y_{nj}$ and $H_n^*(t) = n^{-1} \sum_{F(X_{nj}) \leq t} Y_{nj}$ when F is known.

If X represents the income of a family and Y its consumption of a particular commodity, then $H(t)/H(1)$ is the proportion of the total national consumption of the commodity consumed by the poorest $100t$ percent of the families. For most commodities m is increasing, so that $H(t)/H(1)$ is a convex function connecting the points (0, 0) and (1, 1). The area below the 45° line and above $H(t)/H(1)$ is typically small for a necessity and large for a luxury. This curve can be of some use in determining taxation policy.

Now $n^{\frac{1}{2}}(H_n - H) = U_n + J_n$ where $U_n(t) = n^{-\frac{1}{2}} \sum_1^{[nt]} \{Y_{nj} - m(X_{nj})\}$ was considered in [1], $V_n(t) = n^{\frac{1}{2}}[G_n(t) - t]$ with F_n the empirical cdf of $X_1, \cdots, X_n$ and $G_n = F_n \circ F^{-1}$ and

$$J_n(t) = n^{\frac{1}{2}}\left[\int_0^{F(X_{n,[nt]})} h(s)\, dG_n(s) - \int_0^t h(s)\, ds\right] = \int_0^t V_n(s)\, dh(s) + R_n(t)\ ,$$

where integration by parts yields an expression for R_n such that $\sup_{a \leq t \leq b} |R_n(t)| \to_p 0$ for all $[a, b] \subset (0, 1)$ provided h is assumed continuous. Likewise, $n^{\frac{1}{2}}(H_n^* - H) = U_n^* + J_n^*$ where

$$U_n^*(t) = n^{-\frac{1}{2}} \sum_{F(X_{nj}) \leq t} \{Y_{nj} - m(X_{nj})\} \qquad \text{and}$$
$$J_n^*(t) = \int_0^t V_n(s)\, dh(s) - V_n(t)h(t)\ .$$

Suppose the following conditions hold.

C1. F is continuous.

Received June 1974; revised August 1975.

[1] Research supported by NSF Grant MPS 74-06952-A01.

AMS 1970 *subject classification*. 62E20.

Key words and phrases. Induced order statistics, regression analysis, weak convergence, Brownian motion.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Statistics*. www.jstor.org

C2. $\beta(x) = E[\{Y - m(x)\}^4 \mid X = x]$ is bounded.
C3. $\sigma^2(x) = \operatorname{Var}(Y \mid X = x)$ is of bounded variation.

In [1] it was shown that under these conditions $U_n \Rightarrow \zeta \circ \psi$ where ζ is a standard Brownian motion and $\psi(t) = \int_{-\infty}^{F^{-1}(t)} \sigma^2(x)\, dF(x)$. Minor modification in that proof shows that under the same conditions we have

$$(1) \qquad (U_n, V_n) \Rightarrow (\zeta \circ \psi, \eta) \quad \text{and} \quad (U_n^*, V_n) \Rightarrow (\zeta \circ \psi, \eta)$$

where η is a Brownian bridge independent of ζ. The key points of such a modification are: (i) Things are done conditionally given $X_1, X_2, \cdots$ in view of the independence of Theorem 1 of [1] (note that Theorem 1(a) is also true conditionally). (ii) Make use of the triangle inequality $\|\hat{\psi}_n - \psi\| \leqq \|\hat{\psi}_n - \psi_n\| + \|\psi_n - \psi\|$ while dealing with the conditional behavior of U_n and an analogous one for U_n^* where $\psi_n(t) = n^{-1} \sum_1^{[nt]} \sigma^2(X_{nj})$, $\hat{\psi}_n(t) = n^{-1} \sum_1^{[nt]} T_{nj}$ and $\{T_{nj}\}$ are the stopping times of Theorem 1. (iii) In the course of the proof of Theorem 2(b), it was actually shown that in the conditional argument $\|\hat{\psi}_n - \psi_n\| \to_p 0$ uniformly in $X_1, X_2, \cdots$ by C2.

Since $\phi(u, v)(t) = u(t) + \int_0^t v(s)\, dh(s)$, $a \leqq t \leqq b$ is a continuous function from $D[0, 1]^2$ to $D[a, b]$ whenever $[a, b] \subset (0, 1)$, we conclude from (1) that under C1—C3 and continuity of h we have

$$(2) \qquad n^{\frac{1}{2}}(H_n(t) - H(t)) \Rightarrow \zeta \circ \psi(t) + \int_0^t \eta(s)\, dh(s) \quad \text{on} \quad [a, b]\,.$$

Actually (2) holds on $[0, 1]$ if h is uniformly continuous. Likewise, under C1—C3

$$(3) \qquad n^{\frac{1}{2}}(H_n^*(t) - H(t)) \Rightarrow \zeta \circ \psi(t) + \int_0^t \eta(s)\, dh(s) - \eta(t)h(t) \quad \text{on} \quad [0, 1]\,.$$

2. Applications in regression analysis.

(i) *Testing for a constant regression.* Assume $\sigma^2(x)$ to be constant and consider the problem of testing H_0: $m(x) \equiv E(Y)$ or equivalently, H_0: $H(t) - tH(1) \equiv 0$ against all alternatives. Note that $H_n(1)$ is simply the sample mean $\bar{Y}$ and the sample variance $s^2 = n^{-1} \sum_1^n (Y_j - \bar{Y})^2$ is a consistent estimator of $\psi(1)$ under H_0. It now follows from (2) that under H_0,

$$n^{\frac{1}{2}} s^{-1}(H_n(t) - t\bar{Y}) \Rightarrow \zeta(t) - t\zeta(1) \quad \text{on} \quad [0, 1]$$

from which we can easily construct a large sample test which rejects H_0 when $\sup_{0 \leqq t \leqq 1} n^{\frac{1}{2}} s^{-1} |H_n(t) - t\bar{Y}|$ is too large where the critical value is obtained from the well-known distribution of the maximum absolute value of a Brownian bridge. If m is increasing under the alternative hypothesis, then a test based on $\sup_{0 \leqq t \leqq 1} n^{\frac{1}{2}} s^{-1}(H_n(t) - t\bar{Y})$ can be used in an analogous manner. These tests could just as well be derived from the resule of [1] which coincides with (2) for constant regression.

(ii) *Confidence interval for H.* For a given t, a confidence interval for $H(t)$ can be constructed in a large sample around $H_n(t)$ or $H_n^*(t)$ when F is known.

By (2) and (3), the statistics $n^{\frac{1}{2}}(H_n(t) - H(t))$ and $n^{\frac{1}{2}}(H_n^*(t) - H(t))$ are asymptotically normal with respective variances $v(t) = D(t) + t(1-t) - 2(1-t)h(t)H(t) - H^2(t)$ and $v^*(t) = D(t) - H^2(t)$ where $D(t) = \int_{-\infty}^{F^{-1}(t)} E(Y^2 \mid X = x)\, dF(x)$. Consistent estimators $v_n(t)$ and $v_n^*(t)$ of these variances are obtained by replacing $D(t)$, $H(t)$ and $h(t)$ by their consistent estimators $D_n(t) = n^{-1} \sum_1^{[nt]} Y_{nj}^2$, $H_n(t)$ and $h_n(t) = m_n(X_{n,[n,t]})$ respectively where m_n is a regression estimator of the type considered by Nadaraya (1964) and Watson (1964). The resulting asymptotic confidence intervals with confidence coefficient $1 - \alpha$ are $H_n(t) \pm \Phi^{-1}(1 - (\alpha/2))(v_n(t)/n)^{\frac{1}{2}}$ and $H_n^*(t) \pm \Phi^{-1}(1 - (\alpha/2))(v_n^*(t)/n)^{\frac{1}{2}}$ where Φ is the standard normal cdf.

In order to obtain a confidence band for the function H around H_n on $[a, b] \subset (0, 1)$, the distribution of $\sup_{a \leq t \leq b} |\zeta \circ \psi(t) + \int_0^t \eta(s)\, dh(s)|$ is needed. This is a hopelessly complicated problem. However, in the course of the proof of (1) it can be seen that conditionally, given $X_1, X_2, \cdots, U_n \Rightarrow \zeta \circ \psi$ and the convergence is uniform in $X_1, X_2, \cdots$. From this we can obtain a conditional confidence band $H_n \pm c_\alpha(\psi_n(1)/n)^{\frac{1}{2}}$ with confidence coefficient $1 - \alpha$ for the function $\hat{H}_n(t) = n^{-1} \sum_1^{[nt]} m(X_{nj})$ on $[0, 1]$ where c_α is the $100(1-\alpha)$ percent point of the distribution of $\sup_{0 \leq t \leq 1} |\zeta(t)|$ and $\psi_n(1) = n^{-1} \sum_1^n \{Y_j - m_n(X_j)\}^2$.

(iii) *Testing the equality of two regression functions.* Suppose (X_i, Y_i), $i = 1, \cdots, n$ and (X_i', Y_i'), $i = 1, \cdots, n'$ are random samples from two bivariate populations with common marginal cdf F of X and X' and $\sigma^2(x) = \operatorname{Var}(Y \mid X = x) = \operatorname{Var}(Y' \mid X' = x)$. The null hypothesis H_0: $m = m'$ is to be tested where m and m' are the regression functions in the two populations. The alternative hypothesis is H_1: $m(x) \geqq m'(x)$ for all x with strict inequality on a set of positive probability. Let X_{nj}, $X'_{n'j}$, Y_{nj} and $Y'_{n'j}$ denote the order statistics and the induced order statististics in the two samples and define $H_n(t)$, $H_n^*(t)$ from the first sample and $H'_{n'}(t)$, $H_{n'}^{*\prime}$ from the second sample as usual. Let R_{nj} be the rank of X_j among $X_1, \cdots, X_j$ and $R'_{n'j}$ the rank of X_j' among $X_1', \cdots, X'_{n'}$. It then follows from (2) and (3) that under H_0 the statistics

$$(4) \qquad \begin{aligned} &(n^{-1} + n'^{-1})^{-\frac{1}{2}} \int_0^1 (H_n(t) - H'_{n'}(t))\, dt \\ &\quad = (n^{-1} + n'^{-1})^{-\frac{1}{2}}[n^{-2} \textstyle\sum_1^n (n - R_{nj})Y_j - n'^{-2} \sum_1^{n'} (n' - R'_{n'j})Y_j'], \end{aligned}$$

$$(5) \qquad \begin{aligned} &(n^{-1} + n'^{-1})^{-\frac{1}{2}} \int_0^1 (H_n^*(t) - H_{n'}^{*\prime}(t))\, dt \\ &\quad = (n^{-1} + n'^{-1})^{-\frac{1}{2}}[n^{-1} \textstyle\sum_1^n \{1 - F(X_{nj})\} Y_{nj} \\ &\qquad - n'^{-1} \textstyle\sum_1^{n'} \{1 - F(X'_{n'j})\} Y'_{n'j}] \end{aligned}$$

are asymptotically normal with mean 0 and variances $w = \int_0^1 \int_0^1 D(s \wedge t)\, ds\, dt - \int_0^1 H^2(t)\, dt + 2H(1) \int_0^1 H(t)\, dt - 4(\int_0^1 H(t)\, dt)^2$ and $w^* = \int_0^1 \int_0^1 D(s \wedge t)\, ds\, dt - (\int_0^1 H(t)\, dt)^2$ respectively where $D(t) = \int_{-\infty}^{F^{-1}(t)} E(Y^2 \mid X = x)\, dF(x)$ as before. It can be verified that

$$\begin{aligned} w_n &= n^{-3} \textstyle\sum_1^n (n - R_{nj})^2 Y_j^2 - n^{-3} \sum_1^n \sum_1^n \{(n - R_{nj}) \wedge (n - R_{nk})\} Y_j Y_k \\ &\quad + 2n^{-3}(\textstyle\sum_1^n Y_j)(\sum_1^n (n - R_{nj})Y_j) - 4\{n^{-2} \sum_1^n (n - R_{nj})Y_j\}^2, \\ w_n^* &= n^{-3} \textstyle\sum_1^n (n - R_{nj})^2 Y_j^2 - \{n^{-2} \sum_1^n (n - R_{nj})Y_j\}^2 \end{aligned}$$

are consistent estimators of w and w^* respectively from the first sample.

Define $w'_{n'}$ and $w^{*\prime}_{n'}$ analogous to w_n and $w_n{}^*$ respectively from the second sample. Then the statistics $Z_{nn'}$ obtained by dividing the RHS of (4) by $\{(nw_n + n'w'_{n'})/(n + n')\}^{\frac{1}{2}}$ and $Z^*_{nn'}$ obtained by dividing the RHS of (5) by $\{(nw_n{}^* + n'w^{*\prime}_{n'})/(n + n')\}^{\frac{1}{2}}$ are both asymptotically standard normal under H_0. Tests for H_0 against H_1 can now be constructed in an obvious manner in which H_0 is rejected for large values of $Z_{nn'}$ or $Z^*_{nn'}$. The statistic $Z^*_{nn'}$ is the simpler of the two but it can be computed only when F is known.

Acknowledgment. Editorial suggestions for rewriting the paper are gratefully acknowledged.

REFERENCES

[1] BHATTACHARYA, P. K. (1974). Convergence of sample paths of normalized sums of induced order statistics. *Ann. Statist.* **2** 1034–1039.

[2] NADARAYA, E. A. (1964). On estimating regression. *Theor. Probability Appl.* **9** 141–142.

[3] WATSON, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A* **26** 359–372.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ARIZONA
TUCSON, ARIZONA 85721