The Annals of Statistics
1987, Vol. 15, No. 1, 443–448

THE PENALTY FOR ASSUMING THAT A MONOTONE REGRESSION IS LINEAR[1]

BY DAVID FAIRLEY, DENNIS K. PEARL AND JOSEPH S. VERDUCCI

Ohio State University

For jointly distributed random variables (X, Y) having marginal distributions F and G with finite second moments and F continuous, the proportion of $\mathrm{Var}(Y)$ explained by linear regression is $[\mathrm{Corr}(X, Y)]^2$ while the proportion explained by $E(Y|X)$ can be arbitrarily near 1. However, if the true regression, $E(Y|X)$, is monotone, then the proportion of $\mathrm{Var}(Y)$ it explains is at most $\mathrm{Corr}[Y, G^{-1}(F(X))]$.

1. Introduction. Interest is often focused on the problem of finding a function of a random variable $X \sim F$ that closely predicts another variable $Y \sim G$. If we use expected squared-error loss as a criterion, the best such function is $\phi(X) = E(Y|X)$. This paper discusses the penalty associated with assuming that $\phi(X)$ has a linear form. When F is continuous, bounds on this penalty are obtained in terms of $\rho = \mathrm{Corr}[Y, X]$ or $\rho_0 = \mathrm{Corr}[Y, G^{-1}(F(X))]$.

Taking $E(X) = E(Y) = 0$ and $\mathrm{Var}(X) = \mathrm{Var}(Y) = 1$, we partition the expected unexplained variation left from linear prediction:

$$E(Y - \rho X)^2 = E(Y - \phi(X))^2 + E(\phi(X) - \rho X)^2. \tag{1.1}$$

The left-hand side of (1.1) $= 1 - \rho^2$ while the right-hand side is the sum of an "intrinsic variation" component $\eta = E(Y - \phi(X))^2$ and an "extra-linear variation" component $\psi = E(\phi(X) - \rho X)^2$. For random variables with arbitrary means and finite nonzero variances, we define intrinsic and extra-linear variation as the values of η and ψ for the corresponding standardized linear transforms. This definition makes η the proportion of $\mathrm{Var}(Y)$ unexplained by ϕ, and ψ the penalty for assuming that ϕ is linear. Breiman and Meisel (1976) discuss the difficulties in estimating η.

Since each term in (1.1) is nonnegative we quickly see that η is bounded between 0 and $1 - \rho^2$. The upper bound is obtained, for example, when (X, Y) follows the standard bivariate normal distribution since here $\phi(X) = \rho X$. For arbitrary marginal distributions, it may be impossible for ϕ to be linear. The expected squared distance between the linear regression ρX and the set of possible regressions thus becomes a lower bound for ψ. Vitale (1979) provides a complete characterization of the form of the functions ϕ which are allowable when the marginal distributions of X and Y are fixed.

Received September 1985; revised May 1986.

[1]The information in this paper has been funded wholly or in part by the U.S. Environmental Protection Agency through contract #69-01-6721 through the Office of Toxic Substances. The views expressed in this paper are those of the authors and do not necessarily reflect the views and policies of the Agency.

AMS 1980 *subject classifications.* Primary 62J02; secondary 62E99, 62J05.

Key words and phrases. Fixed margins, inequalities, intrinsic variation, isotonic regression, monotone regression.

The focus of this paper concerns lower bounds for η or, equivalently, upper bounds for ψ. Theorem 1 below shows that if F is continuous we can create a sequence of joint distributions approaching $\eta = 0$ and $\psi = 1 - \rho^2$. However, if we restrict our attention to regression functions ϕ which are monotone, Theorem 2 shows that η and ψ generally take values in a much narrower range, depending on F and G. A consequence of this theorem is that $\eta \geq \psi$ when both margins are normal so that the extra-linear variation cannot be greater than the intrinsic variation in this case. We conclude by suggesting a simple way to estimate the bounds for η and ψ.

2. Lower bounds for intrinsic variation. In general, the marginal distributions F and G by themselves may restrict the range of η for any fixed correlation ρ. For example, if F is concentrated on two points then all regressions are linear and $\eta = 1 - \rho^2$. If both X and Y are discrete with marginal probability functions given by

$$P(X = x_i) = p_i > 0 \quad \text{and} \quad P(Y = y_i) = q_i > 0 \quad \text{for } i = 1, \ldots, k,$$

then η can achieve its lower bound of zero only when the values q_i are rearrangements of the p_i, that is, only when F and G have the same size steps. However, if F is continuous, then for any possible correlation ρ, zero is a tight bound for η, as we now prove. Note that the range of ρ over the class $D(F, G)$ of all joint distributions with fixed margins F and G is given by Hoeffding (1940).

THEOREM 1. *Let F and G be distribution functions with finite second moments and F continuous, and let $H \in D(F, G)$ have correlation ρ. Then there exists a joint distribution function H_N with marginal distributions uniformly close to F and G over their domains, and correlation arbitrarily close to ρ, for which η is arbitrarily close to 0.*

PROOF. The proof proceeds by constructing a joint distribution H_N based on points sampled from H.

Without loss of generality assume that X and Y are standardized and that H has compact support C. If H does not have compact support, approximating the conditional distribution of H given C for a suitably large region C leads to the conclusions of the theorem for ρ and η. Taking the conditional distribution of H_N outside of C equal to that of H leads to the stated convergence of marginal distributions. Let

$$x_0 = \inf\{x\}, \quad y_0 = \inf\{y\}, \quad x_n = \sup\{x\}, \quad y_n = \sup\{y\},$$

where inf and sup are taken over all $(x, y) \in C$. Sample $n - 1$ points (x_i, y_i), $i = 1, \ldots, n - 1$, from H and let $S = \{(x_i, y_i) | i = 0, \ldots, n\}$. Let

$$x_i^+ = \inf\{x | x > x_i \text{ and either } (x, y) \in S \text{ or } x \text{ is a flat point of } F\},$$

$$y_i^+ = \inf\{y | y > y_i \text{ and either } (x, y) \in S \text{ or } y \text{ is a flat point of } G\},$$

$$R_i \text{ be the "rectangle" } [(x_i, y_i), (x_i, y_i^+), (x_i^+, y_i), (x_i^+, y_i^+)],$$

$$i = 0, \ldots, n - 1,$$

and let H_n be the joint distribution corresponding to the uniform mixture of uniform variates on the R_i. [Note: x is a flat point of F if there is some $\varepsilon > 0$ such that $F(x + \varepsilon) = F(x)$.] Then the two marginal distributions of H_n agree with the sample margins at the sampled coordinates and are linearly interpolated between these points over intervals where the marginal distribution is strictly increasing. By the Glivenko–Cantelli theorem it follows that the marginal distributions of H_n converge uniformly to those of H.

To show the convergence of $\rho_n = E[XY|H_n]$ to ρ, let $r_n = n^{-1}\Sigma x_i y_i$, so that

$$|\rho_n - \rho| \leq |\rho_n - r_n| + |r_n - \rho|.$$

Set $\delta_i = (x_i^+ - x_i)/2$, $\varepsilon_i = (y_i^+ - y_i)/2$, and notice that the bounded supports of F and G imply that $(\max_i \delta_i + \max_i \varepsilon_i) \to 0$, almost surely. Thus for μ_i equal to the uniform distribution on R_i, and for almost every realization

$$\begin{aligned}\rho_n &= \frac{1}{n}\sum \int_{R_i} xy\, d\mu_i \\ &= \frac{1}{n}\sum (x_i + \delta_i)(y_i + \varepsilon_i) \\ &= r_n + o(1),\end{aligned}$$

because the x_i and y_i are bounded. Thus $|\rho_n - r_n|$ converges almost surely to 0. Since $r_n \to \rho$ almost surely, it follows that $\rho_n \to \rho$ a.s.

Finally, since under H_n $\operatorname{Var}(Y|X = x_i) = (y_i^+ - y_i)^2/12 \leq (\max_i \varepsilon_i)^2/3$ converges almost surely to 0, so does $\eta_n = E[\operatorname{Var}(Y|X)]$. The proof is concluded by choosing $N = n$ sufficiently large. □

When the form of ϕ is restricted only by the marginal distributions and the value of ρ, Theorem 1 shows that we can have η approach 0 through the construction of a sequence of distributions H_n. However, the function $E[Y|X = x]$ under H_n is very irregular, suggesting that further restricting the allowable regression functions may provide a more meaningful lower bound on η. Specifically, as in isotonic regression, we now consider only functions $\phi(x)$ that are monotone. Heuristically, the monotone regression "farthest" from linear is a step function. This suggests that such functions may produce lower bounds for η. The following example is instructive.

EXAMPLE. Let $-\infty = z_0 < z_1 < \cdots < z_{K-1} < z_K = \infty$ and define the regions $S_i = \{(x, y)\colon z_{i-1} < x \leq z_i \text{ and } z_{i-1} < y \leq z_i\}$ for $i = 1, \ldots, K$. Let the random variables X and Y both have cdf F and density f with mean 0 and variance 1. Take their joint density, h, to be

$$h(x, y) = \begin{cases} f(x)f(y)/[F(z_i) - F(z_{i-1})] & \text{for } (x, y) \in S_i, \\ 0 & \text{otherwise.} \end{cases}$$

Notice that X and Y are conditionally independent given the region S_i and that densities of this form can give $\operatorname{Corr}(X, Y)$ with any value in $[0, 1)$ by adjusting the size of K and the relative sizes of the S_i's. Also for this joint density,

$\phi(x) = E(Y|z_{i-1} < Y \le z_i)$ if $z_{i-1} < x \le z_i$, $i = 1, \ldots, K$. Since $E[\phi(X)^2] = \rho$ in this case, $\eta = 1 - \rho$.

The value of η given by the step function ϕ in this example suggests the lower bound on η among all monotone ϕ given by the following theorem:

THEOREM 2. *Suppose* $X \sim F$ *and* $Y \sim G$, *with* F *continuous*, $E(Y) = 0$, $\mathrm{Var}(Y) = 1$ *and* $\mathrm{Var}(X)$ *finite. Suppose that there is a version of* $\phi(x) = E(Y|X = x)$ *which is increasing in* x. *Then*

$$\eta = E\left[(Y - \phi(X))^2\right] \ge 1 - \rho_0, \tag{2.1}$$

where $\rho_0 = \mathrm{Corr}[Y, G^{-1}(F(X))]$ *and* $G^{-1}(u) = \inf\{x: G(x) > u\}$.

PROOF. Let $Z = G^{-1}(F(X))$. Then by Lemma 2.4 in Whitt (1976), Z and Y are identically distributed with cdf G. Also let $W = \phi(X)$ and let F_W be its cdf. We can rewrite (2.1) as

$$\eta = 1 - E(Y\phi(X)) \ge 1 - E(YZ) = 1 - E(Z\phi(X)),$$

or

$$\begin{aligned} \mathrm{Corr}[Z, W] &= E(Z\phi(X))/\sqrt{\mathrm{Var}(\phi(X))} \\ &\ge E(Y\phi(X))/\sqrt{\mathrm{Var}(\phi(X))} = \mathrm{Corr}[Y, W]. \end{aligned} \tag{2.2}$$

Inequality (2.2) follows from a slight generalization of Theorem 2.5 in Whitt (1976): Let $U \sim \mathrm{Uniform}(0,1)$, and let $W = f(U)$ and $Z = g(U)$, where f and g are increasing functions with $W \sim F_W$ and $Z \sim G$. Then the random variables W and Z have the maximal correlation among all random variables with distributions in $D(F_W, G)$.

It is straightforward to show that the joint cdf of W and Z is $H^*(w, z) = \min(F_W(w), G(z))$. The generalization follows by applying the maximal correlation result in Hoeffding (1940).

To apply the generalization for showing (2.2), take $U = F(X)$, $f(u) = \phi(F^{-1}(u))$ and $g(u) = G^{-1}(u)$. □

REMARKS.

1. If $\phi(x)$ is decreasing in x, substitute $-X$ for X in the proof, yielding $\eta \ge 1 - \rho_0$, where $\rho_0 = \mathrm{Corr}[Y, G^{-1}(1 - F(X))]$.

2. If Y is not standardized, the theorem implies that

$$E\left[(Y - \phi(X))^2\right]/\mathrm{Var}(Y) \ge 1 - \rho_0.$$

Thus it provides a bound on the proportion of variance explained by $\phi(X)$.

3. The theorem can be extended to cases where both X and Y are discrete provided there exists a monotone function h so that $h(X)$ has the same distribution as Y. In particular, this includes the case where X and Y are identically distributed.

4. Another method of proof proceeds by first assuming $W = \phi(X)$ is discrete, say $W = w_i$ for $x_i \le X < x_{i+1}$, $i = 0, \ldots, n - 1$, where $x_0 = -\infty$ and $x_n = +\infty$.

Then, noting that $E(Y) = E(Z) = 0$, we have by Abel's partial summation formula

$$E(YW) = \sum p_i w_i E(Y|x_i \le X < x_{i+1})$$
$$= \sum (w_{i+1} - w_i) P(X \ge x_{i+1}) E(Y|X \ge x_{i+1}),$$

and

$$E(ZW) = \sum p_i w_i E(Z|x_i \le X < x_{i+1})$$
$$= \sum (w_{i+1} - w_i) P(X \ge x_{i+1}) E(Z|X \ge x_{i+1}),$$

where $p_i = P(W = w_i)$.

Now $(w_{i+1} - w_i)P(X \ge x_i) \ge 0$, and it is easy to show that $E(Y|X \ge x_{i+1}) \le E(Z|X \ge x_{i+1})$ by using the fact that Y and Z are identically distributed and that Z is an increasing function of X. Thus the $E(ZW)$ sum dominates the $E(YW)$ sum termwise.

The proof for arbitrary $\phi(X)$ is obtained by taking a sequence of discrete $\phi_n(X)$ converging to $\phi(X)$.

This method of proof makes it clear that the example before Theorem 2 hits the bound, since there $Z = X$ and taking $x_i = z_i$,

$$E(Y|x_i \le X < x_{i+1}) = E(Y|x_i \le Y < x_{i+1}).$$

Thus the bound cannot be improved without more restrictive assumptions.

3. Discussion. In this paper we have developed new sharp bounds for the intrinsic and extra-linear variation in regression.

Among functions $f(X)$ with finite variance, Corr$[Y, f(X)]$ is maximized by $\rho_m = \text{Corr}[Y, \phi(X)]$ [see, for example, Brillinger (1966)]. In particular, ρ_m cannot be smaller than either ρ_0 or ρ. Additionally, the intrinsic variation may be written as $\eta = 1 - \rho_m^2$. These facts, together with Theorem 2, demonstrate that

$$1 - \rho_0 \le \eta \le 1 - \max(\rho_0^2, \rho^2),$$

when ϕ is monotone and X is continuous. The equivalent bounds for the extra-linear variation are then given by

$$(\rho_0^2 - \rho^2)^+ \le \psi \le \rho_0 - \rho^2.$$

Furthermore, since ϕ minimizes the expected squared error, we see that $E(Y - h(X))^2 \ge 1 - \rho_0$ no matter what transformation $h(X)$ is chosen.

An interesting special case arises when the two marginal distributions are known to be in the same location/scale family—for example, both margins normal. Here $\rho_0 = \rho$ so that an increasing regression implies

$$\eta\rho \ge \psi.$$

Since $0 \le \rho \le 1$ in this case, the extra-linear variation cannot be greater than the intrinsic variation.

The quantity ρ_0 can be easily estimated from data. For example, if $(x_1, y_1), \ldots, (x_n, y_n)$ are observations where the y_i's have been scaled to make

$\Sigma y_i = 0$ and $n^{-1}\Sigma y_i^2 = 1$, then we may take $\hat{\rho}_0 = n^{-1}\Sigma y_i y_{[\pi(i)]}$ where $\pi(i)$ is the rank of x_i in the x sample and $y_{[j]}$ denotes the jth largest y value. This estimate has the form of a generalized correlation coefficient (Kendall, 1970).

REFERENCES

BREIMAN, L. and MEISEL, W. S. (1976). General estimates of the intrinsic variability of data in nonlinear regression models. *J. Amer. Statist. Assoc.* **71** 301–307.

BRILLINGER, D. R. (1966). An extremal property of the conditional expectation. *Biometrika* **53** 594–595.

HOEFFDING, W. (1940). Masstabinvariante korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts fur Angewandte Mathematik der Universität Berlin* **5** 179–233.

KENDALL, M. G. (1970). *Rank Correlation Methods.* Griffin, London.

VITALE, R. A. (1979). Regression with given marginals. *Ann. Statist.* **7** 653–658.

WHITT, W. (1976). Bivariate distributions with given marginals. *Ann. Statist.* **4** 1280–1289.

DEPARTMENT OF STATISTICS
OHIO STATE UNIVERSITY
1958 NEIL AVENUE
COLUMBUS, OHIO 43210