The Annals of Statistics
1987, Vol. 15, No. 2, 603–609

A CLASS OF LINEAR REGRESSION PARAMETER ESTIMATORS CONSTRUCTED BY NONPARAMETRIC ESTIMATION

BY J. A. CRISTÓBAL CRISTÓBAL, P. FARALDO ROCA AND W. GONZÁLEZ MANTEIGA,

Universidad de Santiago de Compostela

Given a $(p+1)$-dimensional random vector (X, Y) where f is the unknown density of X, the parameters of the multiple linear regression function $\alpha(x) = E(Y/X = x) = x\beta$ may be estimated from a sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ by minimizing the functional $\hat{\psi}(\beta) = \int(\hat{\alpha}_n(x) - x\beta)^2 \hat{f}_n(x)\,dx$, where $\hat{\alpha}_n$ and $\hat{f}_n$ may be any of a large class of nonparametric estimators of α and f. The strong consistency and asymptotic normality of the estimators so obtained are proved in this article under conditions on (X, Y) that are less restrictive than those assumed by Faraldo Roca and González Manteiga for $p = 1$. This class of estimators includes ordinary and generalized ridge regression estimators as special cases.

1. Introduction and statement of the results. Let (X, Y) be a $(p+1)$-dimensional random variable such that Y is related linearly to X in accordance with the model $Y = X\beta_0 + \varepsilon$, where β_0 is a p-dimensional parameter vector and ε a random error which is independent of X and such that $E[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \sigma^2$. The vector β_0 has the property of being that value of the vector β that minimizes

$$(1.1)\qquad \psi(\beta) = E\left[(Y - X\beta)^2\right] = \int (y - x\beta)^2\,d\mu(x, y) = E[\varepsilon^2],$$

where μ is the probability measure associated with (X, Y). From a theoretical viewpoint, estimation of β_0 from a representative sample $\{(X_1, Y_1), \dots, (X_n, Y_n)\}$ involves first approximating the integral in (1.1) by some function of β constructed using the sample values, and then finding the value of β that minimizes the function so obtained. The least-squares estimator, for example, is obtained by minimizing

$$(1.2)\qquad \psi^*(\beta) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i\beta)^2 = \int (y - x\beta)^2\,d\mu_n^*(x, y),$$

where μ_n^* is the probability measure associated with the empirical distribution function F_n of the sample.

A very general class of estimators may be defined as those obtained by first using the sample to construct a nonparametric estimate $\hat{\alpha}_n(x)$ of the regression function $\alpha(x) = E[Y|X = x]$, and then defining the estimator $\hat{\beta}_n$ itself as the value of β that minimizes

$$(1.3)\qquad \hat{\psi}(\beta) = \int (\hat{\alpha}_n(x) - x\beta)^2\,d\Omega_n(x),$$

Received January 1985; revised September 1986.

AMS 1980 *subject classifications*. Primary 62J05, 62G05.

Key words and phrases. Linear regression, nonparametric estimation, ridge regression.

where Ω_n is a weighting function. Individual members of this class correspond to particular nonparametric estimators $\hat{\alpha}_n$ and weighting functions Ω_n. For example, the least-squares estimator obtained by minimizing (1.2) is the special case in which $\hat{\alpha}_n(x) = \sum_{i=1}^{n} Y_i I_{\{X_i\}}(x)$ and $\Omega_n(x) = (1/n)\sum_{i=1}^{n} I_{(-\infty, x]}(X_i)$, where I_c is the indicator function of C. However, advantages may be expected to derive from the use of nonparametric estimators smoother than the extremely unsmooth function applied in this case [see Titterington (1985) for a general discussion of smoothing techniques]. For instance, we have shown elsewhere [Faraldo Roca and González Manteiga (1985)] the good behavior of the following estimator with respect to least squares when mean-squared error is used for comparison. This class of regression parameter estimators is obtained by restricting $\hat{\alpha}_n$ and Ω_n to the forms

$$\hat{\alpha}_n(x) = \sum_{i=1}^{n} Y_i \delta_m(x, X_i) \Big/ \sum_{i=1}^{n} \delta_m(x, X_i), \quad \text{with } 0/0 \text{ treated as } 0$$

and

$$\Omega_n(x) = \int_{-\infty}^{x} \hat{f}_n(t)\,dt$$
$$= \frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{x} \delta_m(t, X_i)\,dt,$$

where $\{\delta_m\colon R^p \times R^p \to R\}_{m=m(n)\to\infty}$ is a sequence of measurable functions and $\hat{f}_n(t) = \sum_{i=1}^{n} \delta_m(t, X_i)/n$ is a nonparametric estimator of the density of X, f, which is assumed to exist. Almost all nonparametric estimators that have actually been used are, in fact, of this form [Wertz (1978), Susarla and Walter (1981) and Collomb (1985)]. In the rest of this article we shall confine our attention to nonparametric estimators of this class, and we shall refer to the corresponding regression parameter estimators as smooth regression estimators.

The almost sure consistency of smooth regression estimators was proved by Faraldo Roca and González Manteiga (1985) for $p = 1$ when Y and the support of f are bounded and certain fairly unrestrictive assumptions about the sequence δ_m are made. The main purpose of the present article is to extend this result to multiple regression while considerably relaxing the condition on Y.

THEOREM 1.4. *If the support of f is bounded and*

(i) $\sup_u \delta_m(x, u) = O(m^p)$, $\forall x \in R^p$;

(ii) $\delta_m(x, u) = \delta_m(u, x)$, $\forall (x, u) \in R^{2p}$;

(iii) $\int \delta_m(x, u)\,du = 1$ *and* $\delta_m(x, u) = 0$ *if* $\|x - u\| > c\varepsilon(n)$, *where* $c \in R^+$ *and* $\varepsilon(n) \to 0$ *when* $n \to \infty$ $[m = O(1/\varepsilon(n))]$; *then*

(a) *if* $E[Y^4] < \infty$, $\hat{\beta}_n \to \beta_0$ *a.s.*;

(b) *if* $E[\|(X, Y)\|^2] < \infty$ *with* $\gamma > 0$ *such that* $E[|Y|^{2+\gamma}] < \infty$,

$$\sqrt{n}\left(\hat{\beta}_n - \beta_0\right) \to N_p(0, A), \quad \text{as } \sqrt{n}/m \to 0,$$

where $A = \sigma^2(E[X^t X])^{-1}$.

Note that conditions (i)–(iii) on the δ_m are satisfied for a wide class of nonparametric estimators including histogram estimators, kernel estimators whose kernels have compact supports and estimators constructed using the sequence of δ_m described by Susarla and Walter (1981). The proof of part (a) of Theorem 1.4, which is given below in Section 2, is based on the use of Mallows metrics similar to those used by Freedman (1981) and Bickel and Freedman (1981) in the field of bootstrapping regression models. Part (b) is proved by showing that the asymptotic behavior of $\hat{\beta}_n$ is like that of the least-squares estimator.

Hitherto, one of the main attempts to construct regression parameter estimators with mean-squared errors smaller than that of the least-squares estimator has consisted in the development of ridge estimators. These were first introduced to circumvent the purely computational problem posed by the fact that the least-squares estimation matrix $X^t(n)X(n)$ [$X(n)$ indicated in the following theorem] is frequently ill conditioned [Hoerl and Kennard (1970)], but were later given a Bayesian interpretation over such estimators [see Vinod and Ullah (1981)]. In Section 2 below we shall prove the following result:

THEOREM 1.5. *Let K be a symmetric, positive one-dimensional kernel such that $\int K(z)\,dz = 1$, $\int zK(z)\,dz = 0$ and $\int z^2K(z)\,dz < \infty$. If $X(n)$ is the $n \times p$ matrix formed by $(X_1,\dots,X_n)$ and $Y(n)$ the $n \times 1$ matrix $(Y_1,\dots,Y_n)^t$, consider the smooth regression estimator obtained by taking*

$$\delta_m(x,u) = \frac{1}{\varepsilon(n)^p}\prod_{i=1}^{p} K\left(\frac{x_i-u_i}{\varepsilon(n)}\right) = \frac{1}{\varepsilon(n)^p}K*\left(\frac{x-u}{\varepsilon(n)}\right).$$

This estimator is the ordinary ridge regression estimator

$$\beta_k = \left[X^t(n)X(n) + kI_p\right]^{-1}X^t(n)Y(n),$$

with $k = n\varepsilon(n)^2\int z^2K(z)\,dz$.

Furthermore, the smooth regression estimator obtained by taking

$$\delta_m(x,u) = |A_m|K*[A_m(x-u)],$$

where

$$A_m = G\begin{pmatrix} 1/\varepsilon_1 & & 0 \\ & \ddots & \\ 0 & & 1/\varepsilon_p \end{pmatrix}G^t$$

[G being the matrix of the unique transformation such that $X^t(n)X(n) = G\Delta G^t$ with Δ positive and diagonal] is the generalized ridge regression estimator

$$\beta_{KG} = \left[X^t(n)X(n) + G\begin{pmatrix} k_1 & & 0 \\ & \ddots & \\ 0 & & k_p \end{pmatrix}G^t\right]^{-1}X^t(n)Y(n),$$

with $k_i = n\varepsilon_i^2\int z^2K(z)\,dz$.

Note that the problem of choosing ridge factors k or k_i, which optimize the efficiency of the estimator, or at least improve on that of the least-squares estimator, is reformulated as the problem of choosing the appropriate windows $\varepsilon(n)$ or ε_i for the kernel K, whose nature also affects efficiency. For a given kernel, the optimal windows are functions of the sample variances. In the reformulated generalized case (b) the differences between the variances in different directions are thus taken into account when estimating the density of X, and this is done by using a sequence δ_m of the kind introduced by Deheuvels (1977).

Finally, we point out that the class of smooth regression estimators is much wider than that of ridge estimators. It is to be hoped that awareness of this may lead to improved estimators being sought and found among other members of the same class.

2. Proofs.

Proof of Theorem 1.4. (a) The conditions imposed on (X, Y) imply that $E[\|(X,Y)\|^4] < \infty$. Let T_4 be the set of probability measures γ in R^{p+1} for which $\int \|(x,y)\|^4 \, d\gamma(x,y) < \infty$. The Mallows metric

$$d_4(\alpha,\beta) = \inf_{((X,Y),(Z,V))} \left\{E\left[\|(X,Y) - (Z,V)\|^4\right]\right\}^{1/4},$$

where (X,Y) and (Z,V) are random vectors whose associated probability measures are, respectively, α and β, makes T_4 a metric space [see the Appendix of the article by Bickel and Freedman (1981) for more details].

Now (1.1) and the conditions on ε imply that

$$\beta_0 = \beta(\mu) = \{\varepsilon(\mu)\}^{-1} E[X^tY] = \{\varepsilon(\mu)\}^{-1} \int x^t y \, d\mu(x,y),$$

where $\varepsilon(\mu) = \int x^t x \, d\mu(x,y)$. $\hat{\beta}_n$ may likewise be expressed in the form

$$\hat{\beta}_n = \beta(\hat{\mu}_n) = \{\varepsilon(\hat{\mu}_n)\}^{-1} \int x^t y \, d\hat{\mu}_n(x,y),$$

where $\hat{\mu}_n$ is the probability measure associated with the distribution

$$\begin{aligned}\hat{F}_n(x,y) &= \sum_{((X_i,Y_i)|Y_i \le y)} \frac{1}{n} \int_{-\infty}^{x} \delta_m(X_i,t)\,dt \\ &= \int_{(-\infty,-\infty)}^{(\infty,y)} \left(\int_{-\infty}^{x} \delta_m(u,t)\,dt\right) dF_n(u,v).\end{aligned}$$

If $\{\mu_n\}$ is any sequence of probability measures in T_4 such that $d_4(\mu_n,\mu) \to 0$, then $\varepsilon(\mu_n) \to \varepsilon(\mu)$ and $\beta(\mu_n) \to \beta(\mu) = \beta_0$ [see Lemma 3.1 in Freedman (1981)]. In what follows we shall therefore prove part (a) of Theorem 1.4 by showing that $d_4(\hat{\mu}_n,\mu) \to 0$ a.s. Since $d_4(\mu_n,\mu) \to 0$ iff $\mu_n \to \mu$ weakly and $\int \|(x,y)\|^4 \, d\mu_n(x,y) \to \int \|(x,y)\|^4 \, d\mu(x,y)$ [see Lemma 8.3 in Bickel and Freedman (1981)], we shall actually prove (I) that almost surely $\hat{\mu}_n \to \mu$ weakly, and (II) that $\int \|(x,y)\|^4 \, d\hat{\mu}_n(x,y) \to \int \|(x,y)\|^4 \, d\mu(x,y)$ a.s.

(I) Let

$$\Delta_1 = \int_{(-\infty,-\infty)}^{(\infty,y)} \left(\int_{-\infty}^{x} \delta_m(u,t)\,dt\right) d(F_n(u,v) - F(u,v)),$$

$$\Delta_2 = \int_{(x,-\infty)}^{(\infty,y)} \left(\int_{-\infty}^{x} \delta_m(u,t)\,dt\right) dF(u,v),$$

$$\Delta_3 = \int_{(-\infty,-\infty)}^{(x,y)} \left(\int_{-\infty}^{x} \delta_m(u,t)\,dt - 1\right) dF(u,v).$$

Then $\hat{F}_n(x,y) - F(x,y) = \Delta_1 + \Delta_2 + \Delta_3$. With the condition (iii) on the δ_m, it follows that $g_n(u,v) = (\int_{-\infty}^{x}\delta_m(u,t)\,dt - 1) \to 0$ if $u < x$ and $g_n(u,v) + 1 \to 0$ if $u > x$; since $|g_n(u,v)| \le 1$, the dominated convergence theorem yields $\Delta_2 \to 0$ and $\Delta_3 \to 0$.

For Δ_1,

$$\begin{aligned}\Delta_1 &= \int_{(-\infty,-\infty)}^{(x-c\varepsilon_n,y)} d(F_n(u,v) - F(u,v)) \\ &\quad + \int_{(x-c\varepsilon_n,-\infty)}^{(x+c\varepsilon_n,y)} \left(\int_{-\infty}^{x} \delta_m(u,t)\,dt\right) d(F_n(u,v) - F(u,v)) \\ &= F_n(x - c\varepsilon_n, y) - F(x - c\varepsilon_n, y) \\ &\quad + \int_{x-c\varepsilon_n}^{x+c\varepsilon_n} \left(\int_{-\infty}^{x} \delta_m(u,t)\,dt\right) d(F_n(u,y) - F(u,y)).\end{aligned}$$

Using one integration by parts, we finish the proof of (I) with the theorem of Wolfowitz (1960) because

$$|\Delta_1| \le C \sup|F_n(u,v) - F(u,v)| \to 0 \quad \text{a.s. with } C \text{ constant.}$$

(II) Let

$$\Delta_1 = \frac{1}{n}\sum_{i=1}^{n} \int (\|x\|^4 - \|X_i\|^4)\delta_m(x, X_i)\,dx,$$

$$\Delta_2 = \frac{2}{n}\sum_{i=1}^{n} Y_i^2 \int (\|x\|^2 - \|X_i\|^2)\delta_m(x, X_i)\,dx,$$

$$\Delta_3 = \frac{1}{n}\sum_{i=1}^{n} (\|X_i\|^2 + Y_i^2)^2 - E\left[\|(X,Y)\|^4\right].$$

Since

$$\int \|(x,y)\|^4\,d\hat{\mu}_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} \int (\|x\|^2 + Y_i^2)^2 \delta_m(x, X_i)\,dx$$

and $\int \|(x,y)\|^4\,d\mu(x,y)$ is $E[\|(X,Y)\|^4]$, then

$$\int \|(x,y)\|^4\,d\hat{\mu}_n(x,y) - \int \|(x,y)\|^4\,d\mu(x,y) = \Delta_1 + \Delta_2 + \Delta_3.$$

$\Delta_3 \to 0$ a.s. by the strong law of large numbers, and the fact that $\Delta_1 \to 0$ a.s. follows from Bennett's inequality [Bennett (1962)], the compactness of the support of f and the conditions on the δ_m. Finally Schwarz' inequality shows that

$$|\Delta_2| \le 2\left(\frac{1}{n}\sum_{i=1}^{n} Y_i^4\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\left[\int\left|\left(\|x\|^2 - \|X_i\|^2\right)\right|\delta_m(x, X_i)\,dx\right]^2\right)^{1/2}.$$

The strong law of large numbers and reasoning similar to that used for Δ_1 now show that $\Delta_2 \to 0$ a.s. too.

(b) The proof is analogous to that for $p = 1$ [Faraldo Roca and González Manteiga (1985)]. □

Proof of Theorem 1.5. We present only the proof of part (b), since that of part (a) is both similar and simpler.

By (1.3),

$$\begin{aligned}\hat\beta_n &= \left[\int x^t x\, d\Omega_n(x)\right]^{-1}\int x^t\hat\alpha_n(x)\, d\Omega_n(x)\\ &= \left[\int x^t x \hat f_n(x)\, dx\right]^{-1}\int x^t\hat\alpha_n(x)\hat f_n(x)\, dx.\end{aligned}$$

By virtue of the choice of δ_m,

$$\begin{aligned}\hat\beta_n &= \left[\sum_{i=1}^{n}\int x^t x\,\delta_m(x, X_i)\, dx\right]^{-1}\left[\sum_{i=1}^{n}\int x^t\delta_m(x, X_i)\, dx\, Y_i\right]\\ &= \left\{\sum_{i=1}^{n}\int x^t x|A_m|K^*[A_m(x-X_i)]\, dx\right\}^{-1}\left\{\sum_{i=1}^{n}\int x^t|A_m|K^*[A_m(x-X_i)]\, dx\, Y_i\right\}.\end{aligned}$$

Putting $z = A_m(x - X_\tau)$ for $\tau = 1,\dots, n$ now yields

$$\begin{aligned}\hat\beta_n &= \left[\sum_{i=1}^{n}\int\left(A_m^{-1}z + X_i\right)^t\left(A_m^{-1}z + X_i\right)K^*(z)\, dz\right]\sum_{i=1}^{n}\int\left(A_m^{-1}z + X_i\right)^t K^*(z)\, dz\, Y_i\\ &= \left[\sum_{i=1}^{n} X_i^t X_i + nG\begin{pmatrix}\varepsilon_1^2 & & 0\\ & \ddots & \\ 0 & & \varepsilon_p^2\end{pmatrix}G^t\int z^2 K(z)\, dz\right]\sum_{i=1}^{n} X_i^t Y_i,\end{aligned}$$

by virtue of the properties of K and the orthogonality of G. □

Acknowledgments. We wish to thank the Associate Editor and the referees who made possible the best presentation of this paper.

REFERENCES

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *J. Amer. Statist. Assoc.* **57** 33–45.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217.

Collomb, G. (1985). Nonparametric regression: An up-to-date bibliography. *Math. Operationsforsch. Statist. Ser. Statist.* **16** 309–324.

DEHEUVELS, P. (1977). Estimation non parametrique de la densite par histogrammes generalises. II. *Publ. Inst. Statist. Univ. Paris* **22** 1–23.

FARALDO ROCA, P. and GONZÁLEZ MANTEIGA, W. (1985). On efficiency of a new class of linear regression estimates obtained by preliminary non-parametric estimation. In *New Perspectives in Theoretical and Applied Statistics* (M. Puri et al., eds.). Wiley, New York. To appear.

FREEDMAN, D. A. (1981). Bootstrapping regression models. *Ann. Statist.* **9** 1218–1228.

HOERL, A. E. and KENNARD, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12** 55–67.

SUSARLA, V. and WALTER, G. (1981). Estimation of a multivariate density function using delta sequences. *Ann. Statist.* **9** 347–356.

TITTERINGTON, D. M. (1985). Common structure of smoothing techniques in statistics. *Internat. Statist. Rev.* **53** 141–170.

VINOD, H. D. and ULLAH, A. (1981). *Recent Advances in Regression Methods*. Dekker, New York.

WERTZ, W. (1978). *Statistical Density Estimation: A Survey*. Vandenchoeck and Ruprecht, Göttingen.

WOLFOWITZ, J. (1960). Convergence of the empiric distribution function on half-spaces. In *Contributions to Probability and Statistics* (I. Olkin et al., eds.) 504–507. Stanford Univ. Press, Stanford, Calif.

DEPARTAMENTO DE ESTADÍSTICA
FACULTAD DE MATEMÁTICAS
UNIVERSIDAD DE SANTIAGO DE COMPOSTELA
SPAIN