The Annals of Statistics
1992, Vol. 20, No. 4, 1970–1984

ESTIMATING COEFFICIENT DISTRIBUTIONS IN RANDOM COEFFICIENT REGRESSIONS

BY RUDOLF BERAN[1] AND PETER HALL

University of California, Berkeley, and Australian National University

Random coefficient regression models are important in representing linear models with heteroscedastic errors and in unifying the study of classical fixed effects and random effects linear models. For prediction intervals and for bootstrapping in random coefficient regressions, it is necessary to estimate the distributions of the random coefficients consistently. We show that this is often possible and provide practical representative estimators of these distributions.

1. Introduction. The past decade has seen growing interest in the use of random coefficient regression models of the form

$$Y_j = a + (b + b_j)x_j + e_j, \qquad 1 \le j \le n. \tag{1.1}$$

Here a and b are unknown constants, and $\{b_j\}$ and $\{e_j\}$ are each sequences of independent and identically distributed random variables with means zero and unknown distributions and the $\{x_j\}$ typically represent a conditioned version of a random sequence of design points. Such random coefficient regressions model heteroscedastic errors in ordinary linear regression. As special cases, they include some classical random effects models.

Recent surveys of work on random coefficient regression or autoregressive models include Raj and Ullah (1981), Chow (1983), Nicholls and Pagan (1985) and Newbold (1988). The emphasis in the existing theory has been on estimating the constant parameter a, b and on estimating the variances of the random coefficients b_j, e_j. In the case of prediction intervals and for bootstrapping, we would like to know the distributions of b_j and e_j. For example, this would be important if we wished to construct prediction regions for response in random coefficient models for panel data [see Hsiao (1986)]. In general, means and variances alone give little information about these distributions.

We therefore raise and solve two questions in this paper: Given the data (x_j, Y_j), $1 \le j \le n$, when can the distributions of b_j and e_j be estimated consistently? What are practical estimators of these distributions? Section 2.2 provides conditions on the random coefficient model under which the distributions of b_j and e_j are identifiable. Section 2.3 describes several methods for estimating the moments and distributions of b_j and e_j. Section 3 discusses

Received October 1990; revised January 1992.

[1]Work commenced during the tenure of a Visiting Fellowship at the Australian National University and supported in part by an NSF grant.

AMS 1980 *subject classifications*. Primary 62J05; secondary 62G05.

Key words and phrases. Distribution estimation, Hausdorff approximation, identifiability, linear regression, moment estimation, random coefficient.

computational matters. It includes a detailed algorithm for one of the distributional estimates for Section 2.3 and a numerical example.

Work on moment estimation in the context of random coefficient models has points in common with techniques in the econometric literature for estimating residual variances of known functional forms and for testing heteroscedasticity. See for example Hildreth and Houck (1968), Goldfeld and Quandt [(1972), Chapter 3] and Amemiya (1977). Indeed, special cases of some of our estimators [in particular, of those defined at (2.1) in Section 2] have previously been discussed in this context. However, the point at issue in the present paper is not whether we can estimate moments of lower order, such as variance, but whether we can simultaneously estimate a very long sequence of moments, the length of that sequence increasing with increasing sample size. Our main technical result describes just this situation, and enables us to develop theory for distribution estimation. That work is readily extendable to the multivariate case. However, the unwieldy notation required for estimation of multivariate moments of arbitrarily high order and for applying these estimators to produce consistent estimators of multivariate distributions, obscures the main issues. We have therefore chosen to illustrate our results in the simpler, univariate case.

We should comment on our decision to condition on the variables x_i. In our view, regression is intrinsically the study of functional relationships where the design variables are held fixed, that is, are regarded as nonrandom. If the x_i's are not conditioned upon, then the study is one of correlation, not regression. Data for the regression model may be generated by taking an independent sample of independent triples (b_j, e_j, X_j), defining $Y_j = a + (b + b_j)X_j + e_j$ (the correlation model) and conditioning on the values X_j to obtain the regression model. The correlation model admits results very similar to those for the regression model. Indeed, only minor modifications of the techniques used to prove our main result (Theorem 2.2 in Section 2.3 and its implications discussed later as Methods 1–4) are necessary.

2. Main theoretical results.

2.1. *Summary*. We assume that the observed data (x_j, Y_j), $1 \leq j \leq n$, are generated by the model (1.1). Here, a and b are fixed constants, $b_1, e_1, \dots, b_n, e_n$ are totally independent random variables with $b_1, \dots, b_n$ identically distributed and $e_1, \dots, e_n$ identically distributed and $E(b_j) = E(e_j) = 0$. Section 2.2 describes circumstances under which the distributions of b_j and e_j can be identified from an infinite realization $(x_1, Y_1), (x_2, Y_2), \dots$ and Section 2.3 discusses moment-based estimators of those distributions.

2.2. *Identifiability*. It is clear that under very general circumstances, the constants a and b may be estimated consistently by ordinary least squares. However, identifiability of the distributions of b_j and e_j is not so transparent. In addressing the latter problem we shall assume that the design variables $x_1, x_2, \dots$ represent a realization of an independent and identically distributed, nondegenerate sequence $X_1, X_2, \dots$ with cumulative distribution function F. A variety of other cases may be treated similarly. (For example, all of the asymptotic theory which we shall describe applies to the case where the design points $x_i = x_{ni}$, $1 \le i \le n$, are evenly spaced on an interval I, provided the design "density" is taken to be uniform on I.)

We shall say that $s \in [-\infty, \infty]$ is a point of support of the design distribution F if either $|s| < \infty$ and $P(s - \varepsilon < X_1 < s + \varepsilon) > 0$ for all $\varepsilon > 0$, or $s = \pm\infty$ and $P(X_1 > \lambda) > 0$ (when $s = +\infty$) or $P(X_1 < -\lambda) > 0$ (when $s = -\infty$) for all $\lambda > 0$. The following results may be proved.

THEOREM 2.1. *If F has at least one of the points 0, $+\infty$ or $-\infty$ as a point of support, or if the distribution of b_j is uniquely determined by its moments (all assumed finite) and F is nonsingular, then the distributions of b_j and e_j may both be estimated consistently from an infinite realization $(x_1, Y_1), (x_2, Y_2), \dots$.*

The proof is deferred to the Appendix.

If F is singular, then it does not necessarily follow that the b_j and e_j distributions can be identified, even if both distributions are uniquely determined by their moments, or are completely known except for a single parameter. For example, suppose x_j takes only the values ± 1, and the distributions of b_j and e_j are $N(0, \sigma_b^2)$ and $N(0, \sigma_e^2)$, respectively. Then $b_j x_j + e_j$ is normal $N(0, \sigma_b^2 + \sigma_e^2)$ for each j. We can estimate $\sigma_b^2 + \sigma_e^2$, but not σ_b^2 or σ_e^2, from an infinite realization.

2.3. *Estimation by moments.* If both the b_j and e_j distributions are uniquely determined by their moments,

$$\beta_k = E\left(b_j^k\right), \quad \text{and} \quad \gamma_k = E\left(e_j^k\right), \qquad k \ge 1,$$

then sample estimates of β_k and γ_k may be used to estimate those distributions. The first step is to estimate a and b. There is a variety of root-n consistent ways of doing this. Variance is asymptotically minimized by choosing $\hat{a}, \hat{b}$ to be the so-called "efficient" estimators, defined by minimizing

$$S(a, b) = \sum_{j=1}^{n} (Y_j - a - bx_j)^2 \left(x_j^2 \hat{\sigma}_b^2 + \hat{\sigma}_e^2\right)^{-1},$$

where $\hat{\sigma}_b^2$ and $\hat{\sigma}_e^2$ denote consistent estimators of $\operatorname{var}(b_j)$ and $\operatorname{var}(e_j)$, respectively. Ordinary least-squares estimators are given by

$$\hat{b} = \left\{ \sum_{j=1}^{n} Y_j (x_j - \bar{x}) \right\} \left\{ \sum_{j=1}^{n} (x_j - \bar{x})^2 \right\}^{-1} \quad \text{and} \quad \hat{a} = \bar{Y} - \hat{b}\bar{x},$$

where $\bar{x} = n^{-1} \sum_{j \le n} x_j$ and $\bar{Y} = n^{-1} \sum_{j \le n} Y_j$. One way of computing the efficient estimators is to take the ordinary least-squares estimators as pilot estimators of a and b, then calculate estimators of $\operatorname{var}(b_j)$ and $\operatorname{var}(e_j)$ (for example, by using the methods described in the following text) and finally, choose estimators of a and b by minimizing $S(a,b)$. Alternatively, methods suggested by Carroll (1982) and Robinson (1987) could be employed.

However, the matter of efficient estimation of a and b is not a central issue in our analysis. Our aim is to establish uniform consistency of a long sequence of moment estimators, and our main result (Theorem 2.2) is valid for a variety of choices of $\hat{a}$ and $\hat{b}$, including the "efficient" estimators and the ordinary least-squares estimators.

Our estimator of $Z_j = b_j x_j + e_j$ is given by

$$\hat{Z}_j = Y_j - \left(\hat{a} + \hat{b}x_j\right) = b_j x_j + e_j + \varepsilon_j,$$

where $\varepsilon_j = -\{\Delta_1 + \Delta_2(x_j - \bar{x})\}$, $\Delta_1 = n^{-1}\Sigma_{j \le n} b_j x_j + n^{-1}\Sigma_{j \le n} e_j$ and in the special case where $\hat{a}$ and $\hat{b}$ are the ordinary least-squares estimators,

$$\Delta_2 = \left\{\sum_{j=1}^{n} b_j x_j\left(x_j - \bar{x}\right) + \sum_{j=1}^{n} e_j\left(x_j - \bar{x}\right)\right\}\left\{\sum_{j=1}^{n}\left(x_j - \bar{x}\right)^2\right\}^{-1}.$$

If the distributions of b_j, e_j and design are all essentially bounded, then each is uniquely determined by its moments, and also $\max_{j \le n}|\varepsilon_j| = O\{(n^{-1}\log n)^{1/2}\}$ with probability 1. [The latter result follows by Benstein's inequality, using the argument leading to (A.4) in the Appendix]. Therefore, $\max_{j \le n}|\hat{Z}_j - Z_j| \to 0$ with probability 1.

There are many ways of estimating the moments β_k and γ_k root-n consistently, starting from the quantities $\hat{Z}_j$. Observe that

$$\hat{Z}_j^k = \sum_{l=0}^{k}\binom{k}{l} x_j^l \beta_l \gamma_{k-l} + (\text{variable with zero mean}) + o(1), \qquad 1 \le k \le r,$$

and so a variety of techniques based on "regression" may be employed. We suggest two methods. First, the ordinary least-squares method constructs estimators $\hat{\beta}_k$ and $\hat{\gamma}_k$, $1 \le k \le r$, as solutions of the equations

$$\sum_{i=1}^{n} x_j^m\left\{Z_j^k - \sum_{l=0}^{k}\binom{k}{l} x_j^l \hat{\beta}_l \hat{\gamma}_{k-l}\right\} = 0, \qquad 1 \le k, m \le r. \tag{2.1}$$

Typically we would constrain both $\hat{\beta}_1$ and $\hat{\gamma}_1$ to be zero. Naturally, if we should decide to increase the value of r, then this approach requires recomputation of the entire collection $\hat{\beta}_k, \hat{\gamma}_k$. That difficulty may be averted by the second method, based on recursive simple linear regression, as follows. Take $\hat{\beta}_1 = \hat{\gamma}_1 = 0$ and assume that estimates $\hat{\beta}_2, \hat{\gamma}_2, \dots, \hat{\beta}_{k-1}, \hat{\gamma}_{k-1}$ have already been computed. Put $\bar{x}_k = n^{-1}\Sigma_{j \le n} x_j^k$,

$$W_{jk} = \hat{Z}_j^k - \sum_{l=1}^{k-1}\binom{k}{l} x_j^l \hat{\beta}_l \hat{\gamma}_{k-l}, \qquad \overline{W}_{\cdot k} = n^{-1}\sum_{j=1}^{n} W_{jk},$$
$$\hat{\beta}_k = \left\{\sum_{j=1}^{n} W_{jk}\left(x_j^k - \bar{x}_k\right)\right\}\left\{\sum_{j=1}^{n}\left(x_j^k - \bar{x}_k\right)^2\right\}^{-1}, \qquad \hat{\gamma}_k = \overline{W}_{\cdot k} - \hat{\beta}_k \bar{x}_k. \tag{2.2}$$

That is, we estimate β_k and γ_k by simple linear regression on W_{jk}, $1 \le j \le n$, as though the design points were x_j^k.

For each fixed k the estimators $\hat{\beta}_k$ and $\hat{\gamma}_k$ defined by (2.1) and (2.2) are root-n consistent and asymptotically normally distributed. This result is straightforward to prove, and requires only moment conditions on the distributions of b_j, e_j and design. However, it is important to prove that $\hat{\beta}_k$ and $\hat{\gamma}_k$ converge uniformly to β_k and γ_k, respectively, in some sense. The property of uniform convergence is essential to our discussion later in this section of distribution estimation, starting from the moment estimates. When establishing uniform convergence it is simpler to work with the recursive estimators defined by (2.2), rather than the ordinary regression estimators defined by (2.1). For recursive estimators we may obtain the following result. In this instance, $\hat{a}$ and $\hat{b}$ may be taken as either the "efficient" estimators or the ordinary least-squares estimators.

THEOREM 2.2. *Let $\hat{\beta}_k$ and $\hat{\gamma}_k$ be defined as at* (2.2). *Assume that the distributions of b_j, e_j and design are essentially bounded, that $E(b_j) = E(e_j) = 0$ and that the design distribution is nonsingular. Then for each $\delta > 0$ there exists $\eta > 0$ such that with probability* 1,

$$\max_{1 \le k \le \eta(\log n)^{1/2}} \left(|\hat{\beta}_k - \beta_k| + |\hat{\gamma}_k - \gamma_k| \right) = O(n^{-1/2+\delta}) \quad \text{as } n \to \infty.$$

A proof of the theorem is given in the Appendix.

Once the moments have been estimated, the distributions of b_j and e_j may be estimated in a variety of ways. We suggest four methods.

METHOD 1: ORTHOGONAL SERIES. The technique is based on a proposal by Hausdorff (1923) and involves fitting a truncated and rescaled Legendre polynomial expansion to the distribution function (or density) of b_j or e_j.

METHOD 2: PARTIAL FOURIER INVERSION. Assume that the density f_b of b_j exists. Then it may be estimated by

$$(2.3) \qquad \hat{f}_b(y) = \frac{1}{2\pi} \mathscr{R}\ell \int_{-T}^{T} \left\{ 1 + \sum_{k=1}^{r} \frac{(it)^k}{k!} \hat{\beta}_k \right\} e^{-ity}\, dt,$$

where $\mathscr{R}\ell$ denotes "real part". If $r = r(n)$ equals the integer part of $\eta(\log n)^{1/2}$ for a sufficiently small number $\eta > 0$, if $T = T(n)$ diverges to infinity such that $(\log T)/(\log\log n) \to 0$ and if the characteristic function ψ_b of b_j satisfies $\sup_t (1 + |t|)^{1+\varepsilon} |\psi_b(t)| < \infty$ for some $\varepsilon > 0$, then $\sup|\hat{f}_b - f_b| \to 0$ with probability 1. [It will often be found that the truncated Fourier inverse $\hat{f}_b$ has "sidelobes", or oscillations in the tails, caused by the severe zero-one truncation implicit in the integral. These can be substantially reduced by replacing the integral $\int_{[-T,T]} \cdot$ in (2.3) by $\int_{(-\infty,\infty)} \cdot\, \omega_T$, where ω_T is a nonnegative function with bounded support, being equal to 1 over most of $[-T, T]$ and coming down smoothly to zero toward the ends of the interval.]

METHOD 3: DISCRETE APPROXIMATION. Algorithms exist for constructing discrete distributions with m atoms, whose first $2m - 1$ moments match those of a given distribution [e.g., Devroye (1986), page 686ff]. If m equals the integer part of $\eta(\log n)^{1/2}$, for sufficiently small $\eta > 0$, then it follows from our theorem that the $(2m - 1)$-point distribution whose first m moments are $\hat{\beta}_1 = 0, \hat{\beta}_2, \dots, \hat{\beta}_m$, converges as $n \to \infty$ to the distribution of b_j. [With probability converging to 1 as $n \to \infty$, a proper discrete distribution with moments $\hat{\beta}_1, \dots, \hat{\beta}_m$ exists. However, existence is not a crucial matter for the algorithm, which will produce "densities" taking negative values in the tails if the moments do not define a proper $(2m - 1)$-point probability distribution.]

METHOD 4: SMOOTHED DISCRETE APPROXIMATION. Like method 1, this technique is based on a proposal by Hausdorff (1923). It rests on the observation that if G is a distribution function on the interval [0, 1], then

$$G_m(u) = \sum_{j=1}^{[mu]} p_{mj}, \quad \text{where} \quad p_{mj} = \int_0^1 \binom{m}{j} t^j (1-t)^{m-j}\, dG(t)$$

and $[mu]$ denotes the integer part of mu, converges weakly to G as $m \to \infty$. Now, p_{mj} is simply a linear combination of population moments. If these are replaced by their sample counterparts and if m equals the integer part of $\eta(\log n)^{1/2}$ for sufficiently small $\eta > 0$, then the resulting estimator $\hat{G}_m$ converges to G. Shohat and Tamarkin [(1943), page 90ff] have described both of Hausdorff's methods.

In practice, Monte Carlo or bootstrap methods would typically be used when applying estimators of the distributions of b_j and e_j to the problem of constructing prediction intervals. Resampling would be done for the estimated distributions, and in this context the preceding third and fourth techniques would be particularly straightforward to implement. In the case of prediction intervals, nonparametric bootstrap methods have the obvious advantage of consistency when compared with parametric techniques, if the parametric model should be misspecified.

When confidence intervals or hypothesis tests, rather than prediction intervals, are to be the end result of analysis, bootstrap methods offer advantages in terms of coverage accuracy and accuracy of the position of the interval endpoint. However, in this context, an adequate description of a distribution is often obtainable through only the first few moments; see for example Beran (1987) and Hall (1988). Pearson curves are usually fitted to the first four moments, and the second-order accuracy of bootstrap methods relies only on the first three moments. In particular, bootstrap methods for constructing confidence intervals or hypothesis tests about a and b may be based on resampling b_j^* and e_j^* from discrete distributions with moments equal to the estimated values $(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ and $(\hat{\gamma}_1, \hat{\gamma}_2, \hat{\gamma}_3)$, respectively. This is a version of the so-called "wild bootstrap" [Härdle (1989)].

3. Computation of estimators.

3.1. *Overview.* The least-squares methods of Section 2.3 yield consistent estimators of the moments $\{\beta_k\}$ and $\{\gamma_k\}$ of b_j and e_j, respectively. In practice, it is usually reasonable to assume that the distributions of b_j and e_j are supported on compact sets whose endpoints are plus or minus several standard deviations from the respective means. Hausdorff's (1923) second method for approximating a cumulative distribution function (cdf) from its moments (the fourth technique of Section 2.3) is designed for distributions supported on the interval [0, 1]. To use Hausdorff's method in our context requires a one-to-one transformation of the support set and estimated moments to [0, 1].

The basic algebra for such a transformation is as follows. If W is a random variable whose cdf G_W is supported on the compact interval $[-c, c]$, then

$$Z = (2c)^{-1}W + \tfrac{1}{2}$$

has cdf $G_Z(x) = G_W\{2c(x - 1/2)\}$ supported on [0, 1] with moments

$$E(Z^k) = \sum_{j=0}^{k} \binom{k}{j} (2c)^{-j} 2^{-(k-j)} E(W^j), \qquad k \geq 1.$$

These relations, Hausdorff's method and the moment estimators from Section 2.3 are combined in the following algorithm for estimating consistently the distributions of b_j and e_j.

3.2. *Algorithm.* For simpler exposition, we describe only the algorithm that estimates the distribution of b_j. The treatment of e_j is essentially the same. There are four steps:

Step 1. Calculate the least squares estimators $\hat{a}, \hat{b}$, and the moment estimators $(\hat{\beta}_k, \hat{\gamma}_k)$, $1 \leq k \leq m$, by the methods described in Section 2.3.

Step 2. Suppose that the distribution of b_j is supported on the compact interval $[-c, c]$, where $c = \lambda \hat{\beta}_2^{1/2}$ for some finite positive λ. For the reasons discussed in Section 3.1, calculate the transformed estimated moments

$$\tilde{\beta}_k = \sum_{j=0}^{k} \binom{k}{j} (2c)^{-j} 2^{-(k-j)} \hat{\beta}_j, \qquad 1 \leq k \leq m.$$

Step 3. Apply Hausdorff's method by calculating

$$\hat{p}_j = \binom{m}{j} \Delta^{m-j} \tilde{\beta}_j, \qquad 0 \leq j \leq m,$$

where Δ^r is the rth order difference operator defined by

$$\Delta^r \tilde{\beta}_j = \sum_{i=1}^{r} \binom{r}{i} (-1)^i \tilde{\beta}_{j+i}.$$

The discrete distribution that assigns probability $\hat{p}_j$ to the atom j/m for $0 \leq j \leq m$ is the estimated distribution on $[0,1]$ that corresponds to the moments $\hat{\beta}_k$, $1 \leq k \leq m$.

STEP 4. Calculate the polygonal approximant $\tilde{G}$ to the cdf of the discrete distribution from step 3 [cf. Feller (1971), page 540]. Then calculate the estimated cdf of b_j as

$$\hat{G}(x) = \tilde{G}\{(2c)^{-1}x + \tfrac{1}{2}\},$$

with c as in Step 2.

In applying this algorithm, it is important to bear the following four points in mind:

1. All calculations should be done in double precision arithmetic or better. Even so, round-off error can become a problem when m, the number of moments being estimated, exceeds 30.

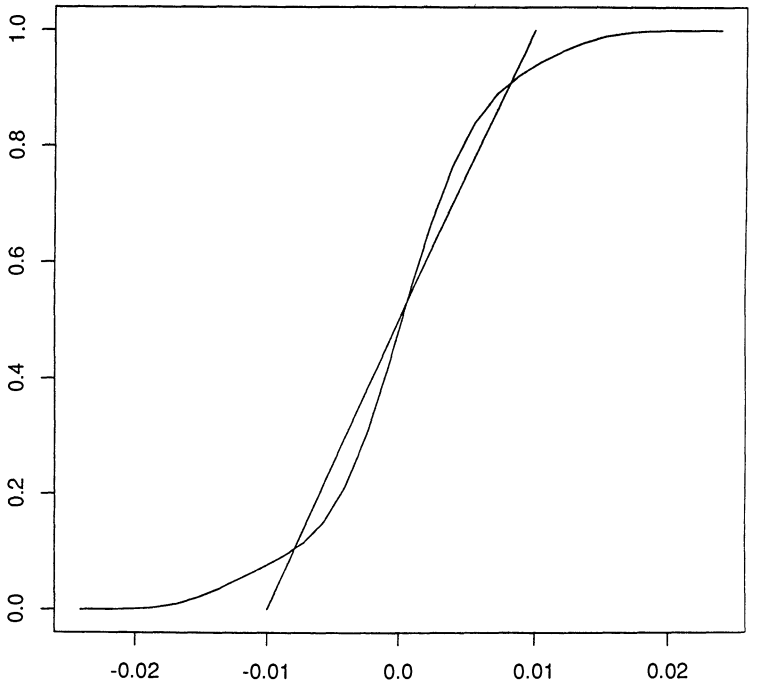

FIG. 1.

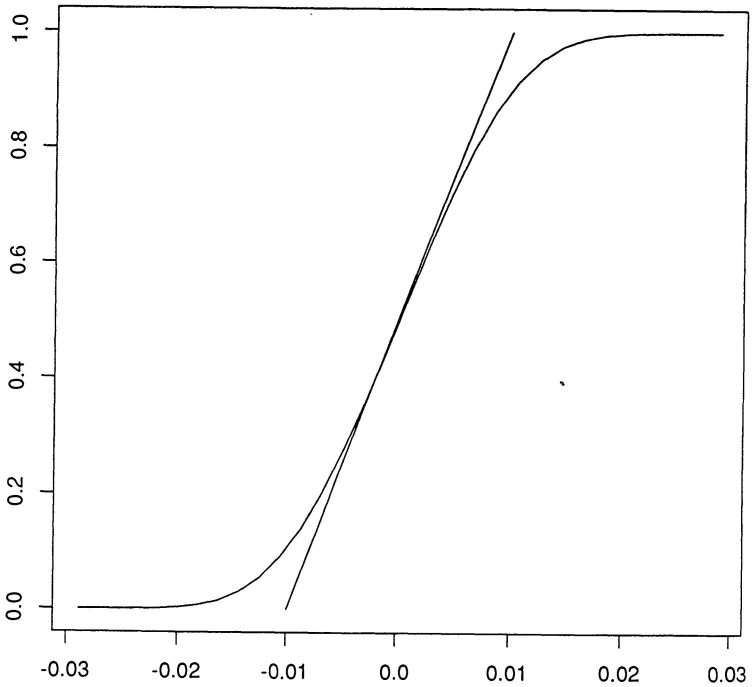

FIG. 2.

2. In a sense, the potential for bias in $\hat{G}$ decreases as m increases, since the polygonal approximation becomes finer. However, the moment estimators are themselves biased and, for fixed n, the effect of this contribution will increase with m.
3. Both round-off error and sampling variability in estimated moments (see Section 2.3) are further reasons for being conservative in selecting m. Trial and error is advisable.
4. A good initial choice of the constant λ in Step 2 is $\lambda = 5$. Too small a choice of λ introduces spurious ripples into the tails of $\hat{G}$. Again, some trial and error is useful.

3.3. *Numerical example.* As an illustration, we applied the preceding algorithm to the random coefficient model in which $a = 0$ and $b = 1$, the design points were $x_i = i/n$ for $1 \leq i \leq n$ and the distributions of b_j and e_j were each uniform on $[-0.01, 0.01]$. An artificial sample of size $n = 200$ from this model served as the data set for this example. We used $m = 30$ estimated moments, taking $\lambda = 5$ in the cdf estimation algorithm. Figures 1 and 2 compare the estimated cdf's of b_j and e_j with the actual uniform cdf's.

The smoothing visible in the tails of both estimated cdf's is the bias that results from using only a finite number of estimated moments in constructing these cdf's. It is remarkable, nevertheless, how well the estimated cdf's approximate the true cdf's in this difficult estimation problem, even at sample size $n = 200$.

APPENDIX: Proof of Theorems

PROOF OF THEOREM 2.1. Observe that since a and b may be estimated consistently, then we may assume, for the purpose of checking identifiability, that $Z_j = b_j x_j + e_j$ is observable. Define $\psi_b(t) = E(e^{itb_j})$, $\psi_e(t) = E(e^{ite_j})$ and

$$\psi(t|s) = E\left(e^{itZ_j}\middle|x_j = s\right) = \psi_b(st)\psi_e(t).$$

If s is a finite support point of F, then, by confining attention to pairs (x_j, Y_j) with x_j close to s, the characteristic function $\psi(t|s)$ may be estimated consistently for all t. The technique involves selecting a small "window" h and computing the empirical characteristic function [Feuerverger and Mureika (1977); Csörgő (1981)] of the sample $\{Z_j\colon s - h < x_j < s + h\}$. This estimator of $\psi(t|s)$ is consistent if, as $n \to \infty$, $h \to 0$ and $nh \to \infty$. Stone (1977) has established consistency in a broad class of regression-type problems of this nature.

If $s = 0$, then $\psi(t|s) = \psi_e(t)$, and so the distribution of e may be estimated consistently. Then, taking s' to be any nonzero support point of F (remembering that F is nondegenerate), we may estimate $\psi_b(t) = \psi(t/s'|s')/\psi_e(t/s')$ consistently for each t and thereby obtain the distribution of b_j. If $+\infty$ is a support point of F, then we may choose a sequence $s_1, s_2, \dots$ of points of support increasing to $+\infty$. The characteristic function $\psi(s_k^{-1}u|s_k) = \psi_b(u)\psi_e(s_k^{-1}u)$ may be estimated consistently for each k and each j. Letting $k \to \infty$ we see that $\psi_b(u)$ can be estimated consistently, as may also be $\psi_e(t) = \psi(t|s')/\psi_e(s't)$ for any support point s'. Therefore, the distributions of both b and e may be identified.

Suppose next that the distribution of b_j is completely determined by its moments and F is nonsingular. Then there exists a nondegenerate open interval I all of whose points are support points of F. We may consistently estimate $\psi(t|s)$ for each $s \in I$ and each t. For fixed t, use difference operators to calculate $\psi_b^{(l)}(st)\psi_e(t)$, $s \in I$ and $l \ge 1$, from $\{\psi(t|s'),\ s' \in I\}$. For example,

$$\psi_b^{(1)}(st)\psi_e(t) = \lim_{s' \to s} \{\psi(t|s') - \psi(t|s)\}/\{(s' - s)t\}.$$

Since $\psi_b^{(l)}(st)\psi_e(t) \to E(b_j^l)/l!$ as $t \to 0$, then all moments of b_j may be estimated consistently. Thus, we may estimate ψ_b, and hence also ψ_e. $\square$

PROOF OF THEOREM 2.2. For the sake of definiteness we take $\hat{a}$ and $\hat{b}$ to be the ordinary least-squares estimators. Let $C_1 \ge 1$ denote an upper bound to each of $\operatorname{ess\,sup}|b_j|$, $\operatorname{ess\,sup}|e_j|$ and $\sup_j |x_j|$.

STEP (i): Lower bound to $\sigma_k^2 = n^{-1}\sum_{j \le n}(x_j^k - \bar{x}_k)^2$. We claim that if $\xi > 0$ is sufficiently small, then there exists a constant $C_2 \ge 1$ such that, with probability 1 among sequences $\{x_j\}$,

$$\min_{1\le k\le \xi \log n} C_2^k\sigma_k^2 \ge 1 \quad \text{for all sufficiently large } n. \tag{A.1}$$

To check this claim, note first that if $U_1,\dots,U_n$ are independent random variables with zero mean and satisfying $\operatorname{ess\,sup}|U_j| \le M$ and $\operatorname{var}(U_j)\le v$ for each j, then by Bernstein's inequality [e.g., Pollard (1984), page 193],

$$P\left(\left|\sum_{j=1}^{n} U_j\right| \ge t\right) \le 2\exp\left\{-\tfrac{1}{2}t^2/\left(nv+\tfrac{1}{3}Mt\right)\right\} \tag{A.2}$$

for each $t>0$. Let $X_1, X_2,\dots$ denote the random sequence of which $x_1, x_2,\dots$ is a realization, and put $\mu_k = E(X_j^k)$, $U_j = n^{-1}(X_j^k-\mu_k)$, $M = n^{-1}2C_1^k$, $v = n^{-2}C_1^{2k}$ and $t = n^{-1/2}C_1^k s$, where $0<s\le n^{1/2}$. By (A.2),

$$P\left(\left|\overline{X}_k-\mu_k\right| \ge n^{-1/2}C_1^k s\right) \le 2e^{-s^2/4}.$$

Define

$$S_k^2 = n^{-1}\sum_{j=1}^{n}\left(X_j^k-\overline{X}_k\right)^2 = \overline{X}_{2k}-\overline{X}_k^2$$

$$= \mu_{2k}-\mu_k^2+\left(\overline{X}_{2k}-\mu_{2k}\right)-\left(\overline{X}_k-\mu_k\right)^2-2\mu_k\left(\overline{X}_k-\mu_k\right).$$

Then,

$$P\left(S_k^2 \le \mu_{2k}-\mu_k^2-n^{-1/2}C_1^{2k}s-n^{-1}C_1^{2k}s^2-2|\mu_k|n^{-1/2}C_1^k s\right) \le 4e^{-s^2/4}.$$

Now, $|\mu_k|\le C_1^k$ and $\mu_{2k}-\mu_k^2 \ge C_3^{2k}$, where $0<C_3\le 1$. Therefore,

$$P\left(S_k^2 \le C_3^{2k}-4n^{-1/2}C_1^{2k}s\right) \le 4e^{-s^2/4}$$

for $0<s\le n^{1/2}$. Put $\xi = \{8\log(C_1/C_3)\}^{-1}$ and $s=n^{1/8}$. Then for $1\le k\le \xi\log n$ and large n,

$$C_3^{2k}-4n^{-1/2}C_1^{2k}s = C_3^{2k}\left\{1-4n^{-1/2}\left(C_1/C_3\right)^{2k}s\right\} \ge \tfrac{1}{2}C_3^{2k}.$$

Hence,

$$P\left\{\inf_{1\le k\le\xi\log n} S_k^2 \le \left(\tfrac{1}{2}C_3\right)^{2k}\right\} \le 4\xi(\log n)\exp\left(-\tfrac{1}{4}n^{1/4}\right).$$

Result (A.1), with $C_2 = (2C_3^{-1})^2$, now follows by the Borel–Cantelli lemma.

STEP (ii): Upper bound to $A_{kl} = n^{-1}\sigma_{k+l}^{-2}\sum_{j=1}^{n}(b_j^k e_j^l-\beta_k\gamma_l)x_j^k(x_j^{k+l}-\bar{x}_{k+l})$. Let ξ be as in (A.1). We claim that, for a constant $C_4\ge 4$,

$$\limsup_{n\to\infty}(n/\log n)^{1/2}\max_{\substack{0\le k,\, l\le\frac{1}{2}\xi\log n,\\ k+l\ge 1}} C_4^{-(k+l)}|A_{kl}| \le C_4 \tag{A.3}$$

with probability 1. We shall establish this result using (A.2), with

$$U_j = n^{-1}\sigma_{k+l}^{-2}\left(b_j^k e_j^l-\beta_k\gamma_l\right)x_j^k\left(x_j^{k+l}-\bar{x}_{k+l}\right).$$

In view of (A.1), ess sup$|U_j|$ and (var $U_j)^{1/2}$ are both dominated by $M = v^{1/2} = 4n^{-1}C_1^{3(k+l)}C_2^{k+l} \le n^{-1}C_4^{k+l}$, where $C_4 = 4C_1^3C_2$. Hence, for large n and all $\lambda > 0$,

$$\max_{\substack{0\le k,\, l\le \frac{1}{2}\xi \log n,\\ k+l\ge 1}} P\Big\{|A_{kl}| > \lambda\big(n^{-1}\log n\big)^{1/2} C_4^{k+l}\Big\} \le \exp\big(-\tfrac{1}{4}\lambda^2 \log n\big),$$

whence

$$P\left\{\max_{\substack{0\le k,\, l\le \frac{1}{2}\xi \log n,\\ k+l\ge 1}} C_4^{-(k+l)}|A_{kl}| > \lambda\big(n^{-1}\log n\big)^{1/2}\right\} \le \big(\tfrac{1}{2}\xi \log n\big)^2 n^{-\lambda^2/4}.$$

It follows via the Borel–Cantelli lemma that if $\lambda > 2$,

$$\limsup_{n\to\infty}\,(n/\log n)^{1/2} \max_{\substack{0\le k,\, l\le \frac{1}{2}\xi \log n,\\ k+l\ge 1}} C_4^{-(k+l)}|A_{kl}| \le \lambda \tag{A.4}$$

with probability 1. Taking $\lambda = C_4$ we deduce (A.3).

STEP (iii): Upper bound to $|\hat{\beta}_k - \beta_k| + |\hat{\gamma}_k - \gamma_k|$. The argument in Step (ii) may be used to show that for a constant $C_5 \ge 1$,

$$\limsup_{n\to\infty}\,(n/\log n)^{1/2} \max_{1\le j\le n} |\varepsilon_j| \le C_5 \tag{A.5}$$

with probability 1. Now, $\hat{Z}_j^k = (b_j x_j + e_j)^k + \Delta_{jk1}$, where, if $|\varepsilon_j| \le 1$,

$$\begin{aligned} |\Delta_{jk1}| &\le \sum_{l=1}^{k} \binom{k}{l} \big|b_j x_j + e_j\big|^{k-l} |\varepsilon_j|^l \\ &\le |\varepsilon_j| \sum_{l=1}^{k} \binom{k}{l} \big|b_j x_j + e_j\big|^{k-l} \le (2C_1)^{2k} |\varepsilon_j|^l. \end{aligned} \tag{A.6}$$

Let $C_6 \ge 1$ be fixed, to be chosen shortly. On the set

$$\mathscr{E}_{k1} = \Big\{\big|\hat{\beta}_l - \beta_l\big| + \big|\hat{\gamma}_l - \gamma_l\big| \le \big(n^{-1}\log n\big)^{1/2} C_6^{\sum_{j=1}^{l} j}, \text{ all } 1 \le l \le k-1\Big\}$$

and assuming that k, n are chosen such that

$$\big(n^{-1}\log n\big) C_6^{\sum_{j=1}^{k} j} \le 1,$$

we have

$$\begin{aligned} &\max_{1\le j\le n} \sum_{l=1}^{k-1} \binom{k}{l} |x_j|^l \big|\hat{\beta}_j \hat{\gamma}_{k-l} - \beta_l \gamma_{k-l}\big| \\ &\le C_1^k (C_1^k + 1) \sum_{l=1}^{k-1} \binom{k}{l} \big(\big|\hat{\beta}_l - \beta_l\big| + \big|\hat{\gamma}_l - \gamma_l\big|\big) \\ &\le (2C_1)^{2k} \big(n^{-1}\log n\big)^{1/2} C_6^{\sum_{j=1}^{k-1} j}. \end{aligned} \tag{A.7}$$

Furthermore,

$$W_{jk} = \hat{Z}_j^k - \sum_{l=1}^{k-1} \binom{k}{l} x_j^l \hat{\beta}_l \hat{\gamma}_{k-l} = \beta_k x_j^k + \gamma_k + \Delta_{jk2} + \Delta_{jk3},$$

where

$$\Delta_{jk2} = \sum_{l=0}^{k} \binom{k}{l} x_j^l \left(b_j^l e_j^{k-l} - \beta_l \gamma_{k-l} \right),$$

$$\Delta_{jk3} = \sum_{l=1}^{k} \binom{k}{l} x_j^l \left(\beta_j \gamma_{k-l} - \hat{\beta}_l \hat{\gamma}_{k-l} \right) + \Delta_{jk1}.$$

On the set

$$\mathscr{E}_{k2} = \mathscr{E}_{k1} \cap \left\{ \max_{1 \le j \le n} |\varepsilon_j| \le 2C_5 \left(n^{-1} \log n \right)^{1/2} \right\},$$

assuming that n is so large that $C_5(n^{-1}\log n)^{1/2} \le 1$, we may deduce from (A.6) and (A.7) that

$$\text{(A.8)} \qquad \max_{1 \le j \le n} |\Delta_{jk3}| \le \left(n^{-1} \log n \right)^{1/2} \left\{ (2C_1)^{2k} 2C_5 + (2C_1)^{2k} C_6^{\sum_{j=1}^{k-1} j} \right\}.$$

The estimator $\hat{\beta}_k$ satisfies

$$\hat{\beta}_k - \beta_k = \sigma_k^{-2} n^{-1} \sum_{j=1}^{n} \left(\Delta_{jk2} + \Delta_{jk3} \right) \left(x_j^k - \bar{x}_k \right).$$

On the set

$$\mathscr{E}_{k3} = \mathscr{E}_{k2} \cap \left\{ \max_{\substack{0 \le l,\, m \le \frac{1}{2}\xi \log n, \\ l+m \ge 1}} C_4^{-(l+m)} |A_{lm}| \le 2C_4 \left(n^{-1} \log n \right)^{1/2} \right\}$$

we have

$$\left| \sigma_k^{-2} n^{-1} \sum_{j=1}^{n} \Delta_{jk2} \left(x_j^k - \bar{x}_k \right) \right| = \left| \sum_{l=0}^{k} \binom{k}{l} A_{l,k-l} \right| \le (2C_4)^{2k} \left(n^{-1} \log n \right)^{1/2}.$$

By (A.8) and Step (i),

$$\left| \sigma_k^{-2} n^{-1} \sum_{j=1}^{n} \left(x_j^k - \bar{x}_k \right) \right| \le \left(n^{-1} \log n \right)^{1/2} (2C_1)^k \left\{ (2C_1)^{2k} 2C_5 + (2C_1)^{2k} C_6^{\sum_{j=1}^{k-1} j} \right\}.$$

Therefore, if $C_6 \ge 128 C_1^3 C_4 C_5$, then

$$\left| \hat{\beta}_k - \beta_k \right| \le \left(n^{-1} \log n \right)^{1/2} \tfrac{1}{2} C_6^{\sum_{j=1}^{k} j}.$$

The estimator $\hat{\gamma}_k$ satisfies

$$\hat{\gamma}_k - \gamma_k = n^{-1} \sum_{j=1}^{n} \left(\Delta_{jk2} + \Delta_{jk3} \right) - \left(\hat{\beta}_k - \beta_k \right) \bar{x}_k$$

and, as in the previous paragraph, it follows that if C_6 exceeds an appropriate function of C_1, C_4, and C_5,

$$|\hat{\gamma}_k - \gamma_k| \le \left(n^{-1} \log n\right)^{1/2} \tfrac{1}{2} C_6^{\sum_{j=1}^k j}.$$

Therefore, for a suitably large C_6 (depending on C_1, C_4 and C_5, but not on k or n) we have

$$\text{(A.9)} \qquad \left|\hat{\beta}_k - \beta_k\right| + |\hat{\gamma}_k - \gamma_k| \le \left(n^{-1} \log n\right)^{1/2} C_6^{\sum_{j=1}^k j}.$$

Recalling the definition of $\mathscr{E}_{k1}$ we see that for this C_6, (A.9) holds for all $1 \le k \le (\xi/2) \log n$, on the set

$$\mathscr{E} = \left\{ \max_{1 \le j \le n} |\varepsilon_j| \le 2C_5\left(n^{-1} \log n\right)^{1/2} \right\}$$

$$\cap \left\{ \max_{\substack{0 \le l,\, m \le \frac{1}{2}\xi \log n,\\ l+m \ge 1}} C_4^{-(l+m)} |A_{lm}| \le 2C_4\left(n^{-1} \log n\right)^{1/2} \right\}.$$

By (A.3) and (A.5), $\mathscr{E}$ holds for all sufficiently large n, with probability 1. Therefore,

$$\max_{1 \le k \le \frac{1}{2}\xi \log n} C_6^{-\sum_{j=1}^k j} \left(\left|\hat{\beta}_k - \beta_k\right| + |\hat{\gamma}_k - \gamma_k| \right) \le \left(n^{-1} \log n\right)^{1/2}$$

for all sufficiently large n. Finally, since $\sum_{j \le k} j \le k^2$, then if η is so small that $n^{-\eta^2 \log C_6} \le \delta$,

$$\max_{1 \le k \le \eta (\log n)^{1/2}} \left(\left|\hat{\beta}_k - \beta_k\right| + |\hat{\gamma}_k - \gamma_k| \right) \le n^{-1/2 + \delta}$$

for all sufficiently large n. □

Acknowledgments. The valuable, constructive comments of two referees and an Associate Editor are gratefully acknowledged.

REFERENCES

AMEMIYA, T. (1977). A note on a heteroskedastic model. *J. Econometrics* **6** 365–370.

BERAN, R. (1987). Prepivoting to reduce level error of confidence sets. *Biometrika* **74** 457–468.

CARROLL, R. J. (1982). Adapting for heteroskedasticity in linear models. *Ann. Statist.* **10** 1233–1244.

CHOW, G. C. (1983). Random and changing coefficient models. In *Handbook of Econometrics* (Z. Griliches and M. D. Intriligator, eds.) **2** 1213–1245. North-Holland, Amsterdam.

CSÖRGŐ, S. (1981). Limit behaviour of the empirical characteristic function. *Ann. Probab.* **9** 130–144.

DEVROYE, L. (1986). *Non-Uniform Random Variate Generation*. Springer, New York.

FELLER, W. (1971). *An Introduction to Probability Theory and Its Applications* **2**, 2nd ed. Wiley, New York.

FEUERVERGER, A. and MUREIKA, R. A. (1977). The empirical characteristic function and its applications. *Ann. Statist.* **5** 88–97.

GOLDFELD, S. M. and QUANDT, R. E. (1972). *Nonlinear Methods in Econometrics*. North-Holland, Amsterdam.

HALL, P. (1988). Theoretical comparison of bootstrap confidence intervals (with discussion). *Ann. Statist.* **16** 927–985.

HÄRDLE, W. (1989). Resampling for inference from curves. In *Proc. 47th Session of the International Statistical Institute, Paris* **3** 53–64. International Statistical Institute, Paris.

HAUSDORFF, F. (1923). Momentprobleme für ein endliches Intervall. *Math. Z.* **16** 220–246.

HILDRETH, C. and HOUCK, J. P. (1968). Some estimators for a linear model with random coefficients. *J. Amer. Statist. Assoc.* **63** 584–595.

HSIAO, C. (1986). *Analysis of Panel Data*. Cambridge Univ. Press.

NEWBOLD, P. (1988). Some recent developments in time series analyses III. *Internat. Statist. Rev.* **56** 17–29.

NICHOLLS, D. F. and PAGAN, A. R. (1985). Varying coefficient regression. In *Handbook of Statistics* (E. J. Hannan, P. R. Krishnaiah and M. M. Rao, eds.) **5** 413–449. North-Holland, Amsterdam.

RAJ, B. and ULLAH, A. (1981). *Econometrics, A Varying Coefficients Approach*. Croom-Helm, London.

ROBINSON, P. M. (1987). Asymptotically efficient estimation in the presence of heteroskedasticity of unknown form. *Econometrica* **55** 875–891.

SHOHAT, J. A. and TAMARKIN, J. D. (1943). *The Problem of Moments*. Amer. Math. Soc., New York.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

DEPARTMENT OF STATISTICS
EVANS HALL
UNIVERSITY OF CALIFORNIA, BERKELEY
BERKELEY, CALIFORNIA 94720

DEPARTMENT OF STATISTICS
AUSTRALIAN NATIONAL UNIVERSITY
GPO BOX 4
CANBERRA ACT 2601
AUSTRALIA