ON SELECTING REGRESSORS TO MAXIMIZE THEIR SIGNIFICANCE

Daniel McFadden[1]

Department of Economics, University of California, Berkeley CA 94720-3880
e-mail: mcfadden@econ.berkeley.edu

February 8, 1997 (Revised July 28, 1998)

ABSTRACT: A common problem in applied regression analysis is to select the variables that enter a linear regression. Examples are selection among capital stock series constructed with different depreciation assumptions, or use of variables that depend on unknown parameters, such as Box-Cox transformations, linear splines with parametric knots, and exponential functions with parametric decay rates. It is often computationally convenient to estimate such models by least squares, with variables selected from possible candidates by enumeration, grid search, or Gauss-Newton iteration to maximize their conventional least squares significance level; term this method *Prescreened Least Squares* (PLS). This note shows that PLS is equivalent to direct estimation by non-linear least squares, and thus statistically consistent under mild regularity conditions. However, standard errors and test statistics provided by least squares are biased. When explanatory variables are smooth in the parameters that index the selection alternatives, Gauss-Newton auxiliary regression is a convenient procedure for obtaining consistent covariance matrix estimates. In cases where smoothness is absent or the true index parameter is isolated, covariance matrix estimates obtained by kernel-smoothing or bootstrap methods appear from examples to be reasonably accurate for samples of moderate size.

KEYWORDS: Variable_Selection, Bootstrapping, Structural_Breaks

[1] This research was supported by the E. Morris Cox Endowment. I thank David Freedman, Tom Rothenberg, and James Powell for useful discussion, and two anonymous referees for helpful comments.

ON SELECTING REGRESSORS TO MAXIMIZE THEIR SIGNIFICANCE

Daniel McFadden

1. Introduction

Often in applied linear regression analysis one must select an explanatory variable from a set of candidates. For example, in estimating production functions one must select among alternative measures of capital stock constructed using different depreciation assumptions. Or, in hedonic analysis of housing prices, one may use indicator or ramp variables that measure distance from spatial features such as parks or industrial plants, with cutoffs at distances that are determined as parameters. In the second example, the problem can be cast as one of nonlinear regression. However, when there are many linear parameters in the regression, direct nonlinear regression can be computationally inefficient, with convergence problematic. It is often more practical to approach this as a linear regression problem with variable selection.

This paper shows that selecting variables in a linear regression to maximize their conventional least squares significance level is equivalent to direct application of non-linear least squares. Thus, this method provides a practical computational shortcut that shares the statistical properties of the nonlinear least squares solution. However, standard errors and test statistics produced by least squares are biased by variable selection, and are often inconsistent. I give practical consistent estimators for covariances and test statistics, and show in examples that kernel-smoothing or bootstrap methods appear to give adequate approximations in samples of moderate size.

Stated formally, the problem is to estimate the parameters of the linear model

$$y = X\alpha + Z(\gamma)\beta + u, \quad \gamma \in \Gamma , \tag{1}$$

where y is n×1, X is an n×p array of observations on fixed explanatory variables, Z = Z(γ) is an n×q array of observations on selected explanatory variables, where γ indexes candidates from a set of alternatives Γ, and u is an n×1 vector of disturbances with a scalar covariance matrix. Let k = p + q, and assume $\Gamma \subseteq \mathbb{R}^h$. The set Γ is finite in the traditional problem of variable selection, but will be a continuum for parametric data transformations. Assume the data in (1) are generated by independent random sampling from a model $y = x\cdot\alpha_o + z(\gamma_o,w)\cdot\beta_o + u$, where $(y,x,w) \in \mathbb{R}\times\mathbb{R}^p\times\mathbb{R}^m$ is an observed data vector, $z{:}\Gamma\times\mathbb{R}^m \longrightarrow \mathbb{R}^q$ is a "well-behaved" parametric transformation of the data *w*, $(\alpha_o,\beta_o,\gamma_o)$ denote the true parameter values, and the distribution of *u* is independent of *x* and *w*, and has mean zero and variance σ_o^2. There may be overlapping variables in *w* and *x*. Examples of parametric data transformations are (a) a Box-Cox transformation $z(\gamma,w) = w^{\gamma-1}/(\gamma-1)$ for $\gamma \neq 0$ and $z(0,w) = \log(w)$; (b) a ramp (or linear spline) function $z(\gamma,w) = \text{Max}(\gamma-w,0)$ with a knot at γ; (c) a structural break $z(\gamma,w) = \mathbf{1}(w<\gamma)$ with a break at γ; and (d) an exponential decay $z(\gamma,w) = e^{-\gamma w}$.

One criterion for variable selection is to pick a $c \in \Gamma$ that maximizes the conventional least squares test statistic for the significance of the resulting Z(c) variables, using enumeration, a grid search, or Gauss-Newton iteration, and then pick the least-squares estimates (a,b,s^2) of $(\alpha_o,\beta_o,\sigma_o{}^2)$ in (1) using the selected Z(c). As a shorthand, term this the *Prescreened Least Squares* (PLS) criterion for estimating (1). I will show that PLS is equivalent to selecting Z(γ) to maximize R^2, and is also equivalent to estimating $(\alpha,\beta,\gamma,\sigma^2)$ in (1) jointly by nonlinear least squares. Hence, PLS shares the large-sample statistical properties of nonlinear least squares. However, standard errors and test statistics for a and b that are provided by least squares at the selected Z(c) fail to account for the impact of variable selection, and will usually be biased downward. When Γ is a continuum, a Gauss-Newton auxiliary regression associated with the nonlinear least squares formulation of the problem can be used in many cases to obtain consistent estimates of standard errors and test statistics. When Γ is finite, the effects of variable selection will be asymptotically negligible, but least squares estimates of standard errors will be biased downward in finite samples.

Let $M = I - X(X'X)^{-1}X'$, and rewrite the model (1) in the form

$$(2) \qquad y = X[\alpha + (X'X)^{-1}X'Z(\gamma)\beta] + MZ(\gamma)\beta + u .$$

The explanatory variables X and MZ(γ) in (2) are orthogonal by construction, so that the sum of squared residuals satisfies

$$(3) \qquad SSR(\gamma) = y'My - y'MZ(\gamma)[Z(\gamma)'MZ(\gamma)]^{-1}Z(\gamma)'My .$$

Then, the estimate c that minimizes SSR(γ) for $\gamma \in \Gamma$ also maximizes the expression

$$(4) \qquad S(\gamma) \equiv n^{-1}\cdot y'MZ(\gamma)[Z(\gamma)'MZ(\gamma)]^{-1}Z(\gamma)'My .$$

The nonlinear least squares estimators for α, β, and σ^2 can be obtained by applying least squares to (1) using Z = Z(c); in particular, the estimator of σ^2 is s^2 = SSR(c)/(n-k) and the estimator of β is $b = [Z(c)'MZ(c)]^{-1}Z(c)'My$. Since R^2 is monotone decreasing in SSR(γ), and therefore monotone increasing in S(γ), the estimator c also maximizes R^2. Least squares estimation of (1) also yields an estimator $V_e(b) = s^2[Z(c)'MZ(c)]^{-1}$ of the covariance matrix of the estimator b; however, this estimator does not take account of the impact of estimation of the embedded parameter γ on the distribution of the least squares estimates. The conventional least-squares F-statistic for the null hypothesis that the β coefficients in (1) are zero, treating Z *as if* it were predetermined rather than a function of the embedded estimator c, is

$$(5) \qquad F = b'V_e(b)^{-1}b/q = y'MZ(c)[Z(c)'MZ(c)]^{-1}Z(c)'My/s^2q = \frac{(n-k)\cdot S(c)}{q\cdot(y'My/n - S(c))} .$$

But the nonlinear least squares estimator selects $\gamma \in \Gamma$ to maximize S(γ), and (5) is an increasing function of S(γ). Then *estimation of* $(\alpha,\beta,\gamma,\sigma^2)$ *in* (1) *by nonlinear least squares, with* $\gamma \in \Gamma$, *is equivalent to estimation of this equation by least squares with* $c \in \Gamma$ *selected to* <u>*maximize*</u> *the F-statistic* (5) *for a least squares test of significance for the hypothesis that* $\beta = 0$. When there is a single variable that depends on the embedded parameter γ, the F-statistic equals the square of the

T-statistic for the significance of the coefficient β, and the PLS procedure is equivalent to selecting c to maximize the "significance" of the T-statistic.

I have not found this result stated explicitly in the literature, but it is an easy special case of selection of regressors in nonlinear least squares using Unconditional Mean Square Prediction Error, which in this application where all candidate vectors are of the same dimension coincides with the Mallows Criterion and the Akaike Information Criterion; see Amemiya (1980). Many studies have noted the impact of variable selection or embedded parameters on covariance matrix estimates, and given examples showing that least squares estimates that ignore these impacts can be substantially biased; see Amemiya (1978), Freedman (1983), Freedman-Navidi-Peters (1988), Lovell (1983), Newey-McFadden (1994), and Peters-Freedman (1984).

2. Consistency and Asymptotic Normality

Estimation of (1) by the PLS procedure will be consistent and asymptotically normal whenever nonlinear least squares applied to this equation has these properties. The large-sample statistical properties of nonlinear least squares have been treated in detail by Jenrich (1969), Amemiya (1985, Theorems 4.3.1 and 4.3.2), Tauchen (1985), Gallant (1987), Davidson & MacKinnon (1993), and Newey & McFadden (1994, Theorems 2.6 and 3.4). I give a variant of these results that exploits the semi-linear structure of (1) to limit the required regularity conditions: I avoid requiring a compactness assumption on the linear parameters, and impose only the usual least squares conditions on moments of x and $z(\gamma_o,w)$. For consistency, I require an almost sure continuity condition on $z(\gamma,w)$ that is sufficiently mild to cover ramp and structural break functions. For asymptotic normality, I require almost sure smoothness conditions on $z(\gamma,w)$ that are sufficiently mild to cover ramp functions. These conditions are chosen so that they will cover many applications and are easy to check.

Several definitions will be needed. A function $g:\Gamma\times\mathbb{R}^m \longrightarrow \mathbb{R}^q$ will be termed *well-behaved* if

i. $g(\gamma,w)$ is *measurable* in w for each $\gamma \in \Gamma$,

ii. $g(\gamma,w)$ is *separable*; i.e., Γ contains a countable dense subset Γ_o, and the graph of $g:\Gamma\times\mathbb{R}^m \longrightarrow \mathbb{R}^q$ is contained in the closure of the graph of $g:\Gamma_o\times\mathbb{R}^m \longrightarrow \mathbb{R}^q$.

iii. $g(\gamma,w)$ is *pointwise almost surely continuous* in γ; i.e., for each $\gamma' \in \Gamma$, the set of w for which $\lim_{\gamma\to\gamma'} g(\gamma,w) = g(\gamma',w)$ has probability one.

Condition (iii) allows the exceptional set to vary with γ', so that a function can be pointwise almost surely continuous without requiring that it be continuous on Γ for any w. For example, the structural break function $\mathbf{1}(\gamma<w)$ has a discontinuity for every w, but is nevertheless separable and pointwise almost surely continuous when Γ is compact and the distribution of w has no point mass. A function $g:\Gamma\times\mathbb{R}^m \longrightarrow \mathbb{R}^q$ is *dominated* by a function $r:\mathbb{R}^m \longrightarrow \mathbb{R}$ if $|g(\gamma,w)| \le r(w)$ for all $\gamma \in \Gamma$.

The main result on large sample properties of PLS draws from the following assumptions:

> [A.1] The data in (1) are generated by an independent random sample of size n, with data (y,x,w) from a model $y = x{\cdot}\alpha_o + z(\gamma_o,w){\cdot}\beta_o + u$, with $\gamma_o \in \Gamma$ and $\beta_o \neq 0$. The u are independently identically distributed, independent of x and w, and have mean zero and variance σ_o^2. The vector x has a finite second moment and the function $z(\gamma,w)$ is dominated by a function $r(w)$ that is square integrable. For each $\gamma \in \Gamma$, the array $\mathrm{H}(\gamma) = \mathbf{E}(x,z(\gamma,w))'(x,z(\gamma,w))$ is positive definite.
>
> [A.2] Let $\sigma^2(\gamma) = \mathrm{Min}_{\alpha,\beta}\, \mathbf{E}(y - x\alpha - z(\gamma,w)\beta)^2$. When minimands exist, denote them by $\alpha(\gamma)$ and $\beta(\gamma)$. Then $\alpha(\gamma_o) = \alpha_o$, $\beta(\gamma_o) = \beta_o$, and $\mathrm{Min}_{\gamma\in\Gamma}\, \sigma^2(\gamma) = \sigma^2(\gamma_o) = \sigma_o^2$. For each $\delta > 0$, the identification condition $\mathit{inf}_{\gamma\in\Gamma\,\&|\gamma-\gamma_o|\geq\delta}\ \sigma^2(\gamma)/\sigma_o^2 > 1$ holds.
>
> [A.3] The set Γ is compact and the function $z(\gamma,w)$ is well-behaved.
>
> [A.4] The random disturbance u has a proper moment generating function. There exist an open set Γ_o with $\gamma_o \in \Gamma_o \subseteq \Gamma$ and a bounded function $r(w)$ such that $z(\gamma,w)$ satisfies a Lipschitz property $|z(\gamma,w) - z(\gamma',w)| \leq r(w){\cdot}|\gamma - \gamma'|$ for all $\gamma,\gamma' \in \Gamma_o$. The derivative $d(\gamma,\beta,w) = \beta'\nabla_\gamma z(\gamma,w)$ exists almost surely for $\gamma \in \Gamma_o$ and is well-behaved. Then $\mathbf{E}d(\gamma,\beta,w) = \beta'\nabla_\gamma\mathbf{E}z(\gamma,w)$ and $\mathrm{J}(\gamma,\beta) = \mathbf{E}(x,z(\gamma,w),d(\gamma,\beta,w))'(x,z(\gamma,w),d(\gamma,\beta,w))$ exist and are continuous for $\gamma \in \Gamma_o$. Assume that $\mathrm{J}(\gamma_o,\beta_o)$ is positive definite.

Assumption A.1 guarantees that the regression (1) is almost surely of full rank and satisfies Gauss-Markov conditions. The independence assumption on u is stronger than necessary for large sample results, but nonlinear dependence of $z(\gamma,w)$ on w implies more is needed than just an assumption that u is uncorrelated with x and w. The independence assumption also rules out conditional heteroskedasticity and (spatial or temporal) serial correlation in the disturbances, which may be present in some applications. Independence is not fundamental, and establishing large sample results with weaker assumptions on u merely requires substituting appropriate statistical limit theorems in the proof of Theorem 1 below, with the additional regularity conditions that they require. However, weakening the independence assumption on u requires that bootstrap estimates of variances use a compatible method, such as the conditional wild bootstrap defined in Section 3.

Assumption A.2 is a mild identification condition that requires the asymptotic nonlinear least squares problem have a unique solution at γ_o. One must have $\beta_o \neq 0$ for A.2 to hold. Assumption A.3 holds trivially if Γ is finite. When Γ is not finite, A.3 is needed to guarantee consistency of the nonlinear least squares estimator of (1). It can be relaxed, at a considerable cost in complexity, to a condition that the family of functions $z(\gamma,w)$ for $\gamma \in \Gamma$ can be approximated to specified accuracy by finite subfamilies that are not too numerous; see Pollard (1984, VII.4.15). Pakes and Pollard (1989) characterize some families of functions that satisfy such conditions.

Assumption A.4 is used when Γ is a continuum to guarantee that the nonlinear least squares estimator is asymptotically normal. It will also imply that a Gauss-Newton auxiliary regression can be used to iterate to a solution of the nonlinear least squares problem, and to obtain consistent estimates of standard errors and test statistics for a and b. I have chosen this assumption to cover applications like ramp functions where $z(\gamma,w)$ is not continuously differentiable. The price for generality in this dimension is that I require a strong assumption on the tails of the distribution of the disturbance u (the moment generating function condition), and a strong form of continuity (the uniform Lipschitz condition). The first of these conditions is palatable for most applications, and the second is easy to check. Again, both can be relaxed, using empirical process methods, at the cost of introducing conditions that are more difficult to check. The requirement that $\mathrm{J}(\gamma_o,\beta_o)$ be non-singular is a local identification condition. The following result is proved in the appendix.

Theorem 1. *Suppose A.1-A.3. Then PLS is strongly consistent. In detail, the functions* $(\alpha(\gamma),\beta(\gamma)\sigma^2(\gamma))$, *are continuous on* Γ*; least squares estimators* $(a(\gamma),b(\gamma),s^2(\gamma))$ *from* (1) *satisfy* $\sup_{\gamma\in\Gamma} |(a(\gamma)-\alpha(\gamma),b(\gamma)-\beta(\gamma),s^2(\gamma) - \sigma^2(\gamma))| \longrightarrow_{as} 0$*; and* $c \longrightarrow_{as} \gamma_o$. *Equivalently, nonlinear least squares applied to* (1) *is strongly consistent for the parameters* $(\alpha_o,\beta_o,\gamma_o,\sigma_o^2)$. *If* γ_o *is an isolated point of* Γ, *then* $n^{1/2}(c-\gamma_o) \longrightarrow_p 0$ *and*

$$(6) \qquad n^{1/2} \begin{bmatrix} a-\alpha_o \\ b-\beta_o \end{bmatrix} \longrightarrow_d N(0,\sigma_o^2\mathrm{H}(\gamma_o)^{-1}) .$$

Alternately, if A.4 holds, then

$$(7) \qquad n^{1/2} \begin{bmatrix} a-\alpha_o \\ b-\beta_o \\ c-\gamma_o \end{bmatrix} \longrightarrow_d N(0,\sigma_o^2\mathrm{J}(\gamma_o,\beta_o)^{-1}) .$$

This theorem establishes consistency and asymptotic normality in the case of the classical variable selection problem where Γ is finite, and in the parametric transformation problem where the derivative of the transformation with respect to the index parameters is well-behaved. This covers most applications, including linear spline functions with parametric knots. However, it is possible that γ_o is not isolated and $z(\gamma,w)$ fails to have a well-behaved derivative, so that A.4 fails. The structural break function $z(\gamma,w) = \mathbf{1}(\gamma<w)$ is an example that fails to satisfy the Lipschitz condition. The asymptotic distribution of c, and the properties of various covariance matrix estimators, must be resolved on an individual basis for such cases.

3. Gauss-Newton Auxiliary Regression and Bootstrap Covariance Estimates

When assumptions A.1-A.4 hold, so that γ_o is in the interior of Γ and $z(\gamma,w)$ is almost surely continuously differentiable, Theorem 1 establishes that nonlinear least squares applied to (1) is strongly consistent and asymptotically normal. Consider a least squares regression,

(8) $$y = X\alpha + Z(c^j)\beta + D(c^j,b^j)\Delta\gamma + v ,$$

where (a^j,b^j,c^j) are previous-round estimates of the model parameters and $D(c^j,b^j)$ is the n×h array with rows $d(c^j,b^j,w) = b^{j\prime}\nabla_\gamma z(c^j,w)$. The rule for updating parameter estimates is $a^{j+1} = a$, $b^{j+1} = b$, and $c^{j+1} = c^j + \Delta c$, where a, b, and Δc are the least squares estimates from (8). The regression (8) is called the *Gauss-Newton auxiliary regression* for model (1), and the updating procedure is called *Gauss-Newton iteration*; see Davidson and MacKinnon (1993). Comparing the least squares formulas for $(a,b,\Delta c)$ with the first-order conditions for nonlinear least squares applied to (1), from the proof of Theorem 1, one sees that the non-linear least squares estimators (a,b,c) determine a fixed point of Gauss-Newton iteration: If one starts from (a,b,c) as previous-round estimates, one obtains $\Delta c = 0$, and the linear regressions (1) and (8) will yield identical estimates a and b for the regression coefficients. Comparing the least squares estimate from (8) of the covariance matrix of the estimates $(a,b,\Delta c)$ with the covariance matrix of the nonlinear least squares estimates (a,b,c) from the proof of Theorem 1, one sees that these estimates are asymptotically equivalent. Summarizing, we have the following results:

> (a) *The nonlinear least squares estimates* (a,b,c) *of* (1) *are a fixed point of the Gauss-Newton auxiliary regression* (8), *and the estimated covariance matrix of the regression coefficients from this auxiliary regression at this fixed point is a strongly consistent estimate of the covariance matrix of* (a,b,c).
>
> (b) *For an initial value* c^1, *suppose the initial values* (a^1,b^1) *are obtained from the regression* (1) *using* $Z(c^1)$. *For* c^1 *in some neighborhood of* γ_o, *Gauss-Newton iteration will converge to the fixed point* (a,b,c). *Gauss-Newton iteration from a sufficiently dense grid of initial values for* c^1 *will determine all fixed points of the process, and the fixed point at which* $S(c)$ *is largest is* (a,b,c).

The reason for considering a grid of starting values c^1 is that Gauss-Newton iteration may fail to converge outside a neighborhood of c, or if the fixed point is not unique may converge to a fixed point where $S(c)$ is at a secondary maximum. The assumptions in A.4 that γ_o is an interior point of Γ and that $z(\gamma,w)$ is Lipschitz and its derivative $\nabla_\gamma z(\gamma,w)$ is well-behaved are essential in general for use of the Gauss-Newton auxiliary regression to obtain covariance estimates. For example, the structural break function $\mathbf{1}(w>\gamma)$ fails to satisfy the Lipschitz condition, and (8) fails to give a consistent estimate of the covariance matrix.

In cases where direct application of the Gauss-Newton auxiliary regression fails, it may be possible to use approximate Gauss-Newton auxiliary regression or a bootstrap procedure to obtain covariance estimates that are reasonably accurate for samples of moderate size. The idea is that in a model $y = x\alpha + z(\gamma,w) + u$ satisfying A.1-A.3, but not A.4, one can replace a poorly-behaved $z(\gamma,w)$ by a kernel smoother $z^*(\gamma,w) = \int z(\gamma',w)\cdot k((\gamma-\gamma')/\lambda)\cdot d\gamma'$, where $k(\cdot)$ is a continuously differentiable probability density that has mean zero and variance one, and λ is a small positive bandwidth. For fixed λ, $z^*(\gamma,w)$ satisfies the continuous differentiability condition in A.4. Let $(\alpha^*,\beta^*,\gamma^*)$ be the minimand of $\mathbf{E}(y - x\alpha - z^*(\gamma,w)\beta)^2$. Theorem 1 can be adapted to show that PLS applied to $y = x\alpha + z^*(\gamma,w)\beta + u$ is strongly consistent for $(\alpha^*,\beta^*,\gamma^*)$ and asymptotically normal, and that the auxiliary regression $y = xa + z^*(\gamma^*,w)b + \beta^{*\prime}\nabla_\gamma z^*(\gamma^*,w)\Delta c + u^*$ provides a consistent estimate of the correct covariance matrix for PLS in the smoothed model. The bias in $(\alpha^*,\beta^*,\gamma^*)$ goes to zero with λ. Then, decreasing λ until the coefficients $(\alpha^*,\beta^*,\gamma^*)$ stabilize, and using the Gauss-Newton auxiliary regression at this λ to estimate the covariance matrix, can provide a reasonable approximation when the underlying model is consistent and asymptotically normal.

To illustrate the kernel smoothing approximation, consider the simple structural break model $y = \mathbf{1}(w \le \gamma_o)\beta_o + u$, where A.1-A.3 are satisfied and w has a CDF $F(w)$ with bounded support and a density $f(w)$ that is positive in a neighborhood of γ_o. The asymptotics of structural break problems have been characterized for discrete time series by Andrews (1993) and Stock (1994). The structural break example here is simpler in that w is continuously distributed and the disturbances are i.i.d. The features of this example often appear in spatial models where the span of proximity effects must be estimated. The kernel approximation to $\mathbf{1}(w \le \gamma)$ is $K((\gamma\text{-}w)/\lambda)$, where K is the kernel CDF. The asymptotic covariance matrix obtained from the kernel approximation is J^{-1}, where

$$\mathrm{J} = \begin{bmatrix} \int K((\gamma-w)/\lambda)^2 f(w)dw & \int k(q)K(q)f(\gamma-\lambda q)dq \\ \int k(q)K(q)f(\gamma-\lambda q)dq & \int k(q)^2 f(\gamma-\lambda q)dq/\lambda \end{bmatrix} \longrightarrow \begin{bmatrix} F(\gamma) & f(\gamma)\int k(q)K(q)dq \\ f(\gamma)\int k(q)K(q)dq & +\infty \end{bmatrix}$$

as $\lambda \longrightarrow 0$. This implies that c converges to γ_o at a greater than $n^{1/2}$ rate, and that the asymptotic covariance matrix of b is given by the least squares formula with γ_o known. A Monte Carlo experiment completes the illustration. Let $y = \mathbf{1}(w \le \gamma_o)\beta_o + u$ with w uniformly distributed on [0,1], u standard normal, $\gamma_o = 0.5$, and $\beta_o = 1$. The probability of ties is zero, so the observations can be ordered almost surely with $w_1 < \ldots < w_n$. Let $\bar{y}_k = \sum_{i=1}^{k} y_i/k$. The sum of squares explained, $S(\gamma)$ from (4), is piecewise constant and right-continuous with jumps at w_i, and $S(w_i) = i \cdot \bar{y}_i^2$. Then, c equals the value w_k that maximizes $k\bar{y}_k^2$, and $b = \bar{y}_k$. Table 1 gives exact finite-sample moments, calculated using 10,000 Monte Carlo trials for u, conditioned on one sample of w's. Finite sample bias is small and disappears rapidly as sample size grows. The variance of $n^{1/2}(c - \gamma_o)$ is substantial at n = 100, but drops rapidly with n. Correspondingly, the variance of $n^{1/2}(b - \beta_o)$ at n = 100 is substantially above its asymptotic value of 1.90, but drops rapidly toward this limit as n increases. The probability that the conventional PLS F-statistic exceeds n/2 is given next. If γ_o were known and fixed, then this probability would equal the upper tail probability from the noncentral F-distribution F(n/2,1,n-1,δ), with noncentrality parameter $\delta = \sum \mathbf{1}(w_i \le 0.5)$. This probability and δ are given in the last two rows. The actual finite sample distribution of the F-statistic has a larger upper tail than the corresponding noncentral F-distribution, with the difference disappearing with growing sample size. A kernel-smoothed approximation to the covariance matrix for $n \ge 100$ is accurate to two significant digits for a normal kernel and $\lambda = 15/n$; I omit the detailed results. A disadvantage of kernel-smoothing is lack of a cross-validation method for determining values of λ that give a good empirical approximation.

Table 1. Sample Size n:	100	400	1600	6400
Mean c ($\gamma_o = 0.5$)	0.495	0.496	0.500	0.500
Mean b ($\beta_o = 1$)	1.023	1.005	1.001	1.000
Mean s^2 ($\sigma_o^2 = 1$)	1.004	1.001	1.001	1.000
$\mathrm{Var}(n^{1/2}(c - \gamma_o))$	0.293	0.093	0.027	0.006
$\mathrm{Var}(n^{1/2}(b - \beta_o))$	2.823	2.135	2.024	1.956
$\mathrm{Cov}(n^{1/2}(c - \gamma_o),(n^{1/2}(b - \beta_o))$	-0.469	-0.083	-0.023	-0.007
$\mathrm{Prob}(F > n/2)$	0.317	0.433	0.470	0.682
$1 - F(n/2,1,n-1,\delta)$	0.281	0.415	0.464	0.680
Noncentrality parameter δ	41	193	794	3259

A bootstrap procedure is a possible alternative to a Gauss-Newton auxiliary regression to estimate covariances. The following procedure for estimating the covariances of PLS estimates is called the *conditional bootstrap*; see Peters and Freedman (1984), Brown and Newey (1995), and Veal (1992):

(a) *Given PLS estimates* (a,b,c) *from* (1), *form residuals* $\hat{u} = y - Xa - Z(c)b$.

(b) *Let* $\hat{u}^r$ *be an* $n\times 1$ *vector drawn randomly, with replacement, from elements of* $\hat{u}$. *For bootstrap trials* $r = 1,\ldots,R$, *calculate* $y^r = Xa + Z(c)b + \hat{u}^r$, *and apply PLS to model* (1) *with dependent variable* y^r *to obtain estimates* (a^r,b^r,c^r). *Form bootstrap means* $(a^o,b^o,c^o) = n^{-1} \sum_{r=1}^{R} (a^r,b^r,c^r)$ *and the bootstrap covariance matrix*

$$V^R = n^{-1} \sum_{r=1}^{R} (a^r-a^o,b^r-b^o,c^r-c^o)'(a^r-a^o,b^r-b^o,c^r-c^o).$$

(c) *The vector* $(a^*,b^*,c^*) = 2\cdot(a,b,c) - (a^o,b^o,c^o)$ *is a bias-corrected bootstrap estimate of the parameters of* (1), *and* V^R *is a bootstrap estimate of the covariance matrix of* (a^*,b^*,c^*).

An alternative that can handle conditional heteroskedasticity is the *conditional wild bootstrap*, where $\hat{u}^r = \hat{u}\cdot\zeta^r$, with ζ^r a vector of independent identically distributed draws from a distribution that has $\mathbf{E}\zeta_i = 0$, $\mathbf{E}\zeta_i^2 = 1$, and $\mathbf{E}\zeta_i^3 = 1$. A typical construction is $\zeta_i = \eta_i/2^{1/2} + (\eta_i^2-1)/2$ with η_i standard normal; see Liu (1988) and Davidson and Flachaire (1996).

Suppose A.1-A.3 hold, and PLS is strongly consistent by Theorem 1. Suppose in addition that $n^{1/2}[(a,b,c) - (\alpha_o,\beta_o,\gamma_o)]$ is asymptotically normal, either because A.4 holds or because $n^{1/2}(c - \gamma_o) \longrightarrow_p 0$. Then, the bootstrap procedure with $R \longrightarrow +\infty$ will give consistent estimates of its covariance matrix. The simple bootstrap procedure just described will not guarantee a higher order approximation to the finite sample covariance matrix than conventional first-order asymptotics, although in practice it may capture some sources of variation that appear in finite samples. Various iterated bootstrap schemes are available to get higher order corrections; see Beran (1988) and Martin (1990).

Consider the structural break example $y = \mathbf{1}(w\leq\gamma_o)\beta_o + u$ with w uniformly distributed on [0,1], u standard normal, $\gamma_o = 0.5$, and $\beta_o = 1$. As indicated by the kernel smoothing analysis, this is a case where asymptotic normality holds because $n^{1/2}(c - \gamma_o) \xrightarrow{}_p 0$. For a single sample of *w*'s, we draw 100 samples from the true model, and then apply the conditional bootstrap with R = 100 repetitions to each sample. Table 2 gives exact sample and bootstrap estimates of the means and variances of the estimators of γ_o and β_o. As in Table 1, bias in the estimated coefficients is small, and attenuates rapidly; a bootstrap bias correction has a negligible effect. The bootstrap variance estimates are relatively accurate for the sample sizes larger than 100, and do appear to capture some of the finite-sample effects that disappear asymptotically.

Table 2: Sample Size n	100	400	1600	6400
Exact Mean *c*	0.495	0.496	0.500	0.500
Bootstrap Mean *c*	0.499	0.493	0.500	0.500
Exact Mean *b*	1.023	1.005	1.001	1.000
Bootstrap Mean *b*	1.062	1.007	0.999	1.000
Exact Var($n^{1/2}(c - \gamma_o)$)	0.293	0.093	0.027	0.006
Bootstrap Var($n^{1/2}(c - \gamma_o)$)	0.385	0.090	0.029	0.006
Exact Var($n^{1/2}(b - \beta_o)$)	2.823	2.135	2.024	1.956
Bootstrap Var($n^{1/2}(b - \beta_o)$)	2.893	2.070	2.014	1.926

Summarizing, the conditional bootstrap appears in one example where an explanatory variable is poorly behaved to provide acceptable covariance estimates. It appears to be at least as accurate in samples of moderate size as Gauss-Newton regression using a kernel-smoothing asymptotic approximation, and in addition avoids having to choose a smoothing bandwidth.

4. The Effects of Selection from a Finite Set of Candidates

When Γ is finite, so that γ_o is isolated, Theorem 1 establishes that the asymptotic effects of variable selection are negligible, and one can treat *c* as if it were known and fixed in forming least squares estimates of asymptotic standard errors or test statistics for α and β. However, in finite samples, variable selection will have a non-negligible effect on standard errors and test statistics, and the least squares estimates will generally be biased downward. This problem and several closely related ones have been investigated in the literature. Amemiya (1980) and Breiman and Freedman (1983) deal with the question of how many variables to enter in a regression. Freedman (1983) and Freedman, Navidi, and Peters (1988) show that variable selection leads least squares standard errors to be biased downward. There is a related literature on pre-test estimators; see Judge and Bock (1978). Also closely related is the econometrics literature on the problem of testing for a structural break at an unknown date; see Andrews (1993) and Stock (1994).

Suppose the disturbances in (1) are normal. If γ_o were known and $Z(\gamma_o)$ always selected, then the test statistic (5) for the significance of Z is distributed non-central $F(q,n-k,\delta)$, with noncentrality parameter $\delta = \beta_o'(Z(\gamma_o)'MZ(\gamma_o))\beta_o/\sigma_o^2$. When γ_o is unknown and *c* is selected to maximize (5), then this statistic will always be at least as large as the F-statistic when $Z(\gamma_o)$ is always selected. This implies that the probability in the upper tail of the distribution of (5) exceeds the probability in the corresponding upper tail of the $F(q,n-k,\delta)$ distribution.

An example provides an idea of the magnitude of the bias in the PLS test statistic. Suppose candidate variables (z_o,z_1) take the values (1,1) with probability (1+ρ)/4, (-1,-1) with probability (1+ρ)/4, (1,-1) with probability (1-ρ)/4, and (-1,1) with probability (1-ρ)/4. Suppose $y = z_o\beta_o + u$, with u normal with mean 0 and variance σ_o^2, so that z_o is the correct variable. In a random sample of size n, conditioned on the covariates, the alternative estimates of β_o when z_o or z_1 are selected, respectively, are jointly distributed $\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_o \\ \beta_o\rho_n \end{bmatrix}, \frac{\sigma_o^2}{n}\begin{bmatrix} 1 & \rho_n \\ \rho_n & 1 \end{bmatrix}\right)$, where $\rho_n = \sum z_o z_1/n$. The residuals from these respective regressions are independent of b_0 and b_1, and jointly distributed

$$\begin{bmatrix} w_0 \\ w_1 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ Q_1 Z_o \beta_o \end{bmatrix}, \sigma_o^2 \begin{bmatrix} Q_o & Q_o Q_1 \\ Q_1 Q_o & Q_1 \end{bmatrix}\right) ,$$

where Z_k is the n×1 vector of z_o's for k = 0 and of z_1's for k = 1; $Q_k = I - Z_k Z_k'/n$; and $s_k^2 = w_k' w_k/(n-1)$. The T-statistics are $T_k = n^{1/2} \cdot b_k/s_k$, and Z_o is selected if T_o is larger in magnitude than T_1. The statistic T_o has unconditionally a T-distribution with n-1 degrees of freedom. Conditioned on w_o and w_1, $T_o \sim N(\beta_o n^{1/2}/s_o, \sigma_o^2/s_o^2)$. The distribution of T_1, given W_o, W_1, and T_o = t, is $N\left(\frac{ts_o\rho_n}{s_1}, \frac{\sigma_o^2}{s_1^2}(1-\rho_n^2)\right)$. Let k be the critical level for a conventional T-test of significance level α, and let π denote the conventional power of this test; i.e., $\pi_o = \text{Prob}(F(1,n\text{-}1,\delta) > k)$, with $\delta = \beta_o n/s_o$. Then, the probability that PLS yields a test statistic of magnitude larger than k is

$$\text{Prob}(\max\{|T_o|,|T_1|\} > k) = \pi_o + \int_{-k}^{k} \text{Prob}(|T_1|>k|T_o=t)\cdot\varphi\left(\frac{s_o t - \beta_o n^{1/2}}{\sigma_o}\right) \cdot \frac{s_o}{\sigma_o} \cdot dt$$

$$= \pi_o + \boldsymbol{E}_{w_0,w_1}\int_{-k}^{k} \left[\Phi\left(\frac{-s_1 k - ts_o\rho_n}{\sigma_o(1-\rho_n)^{1/2}}\right) + \Phi\left(\frac{-s_1 k + ts_o\rho_n}{\sigma_o(1-\rho_n)^{1/2}}\right)\right] \cdot\varphi\left(\frac{s_o t - \beta_o \nu^{|||}}{\sigma_o}\right) \cdot \frac{s_o}{\sigma_o} \cdot dt \, .$$

Table 3 gives this probability, denoted π, as well as the probability π_o, for α = 0.05, $\sigma_o = 1$, and selected values of ρ, β_o, and n. When $\beta_o = 0$, the identification condition A.2 fails, and the probability of rejecting the null hypothesis using either z_o or z_1 alone equals the nominal significance level, 0.05. When z_o and z_1 are uncorrelated, $\pi = 1 - (0.95)^2 = 0.0975$; the bias is reduced when z_o and z_1 are correlated. This effect continues to dominate for n and β_o where π_o is low. When π_o is large, the bias is smaller in relative terms, and increases when z_o and z_1 are correlated. Adding more candidates would increase the biases.

The previous section found relatively accurate bootstrap estimates of standard errors in an example where convergence of c to γ_o occurs at a better than $n^{1/2}$ rate, so that $n^{1/2}(c - \gamma_o)$ is asymptotically degenerate normal. The problem of isolated γ_o is similar, so that bootstrap estimates of covariances and asymptotic tests utilizing these bootstrap covariances can also be reasonably accurate in samples of moderate size.

Table 3		ρ:	0.00	0.45	0.90
β_o	n	π_o	π	π	π
0.0	100	0.050	0.097	0.093	0.070
0.0	1000	0.050	0.097	0.093	0.070
0.01	100	0.051	0.093	0.093	0.072
0.01	1000	0.062	0.107	0.104	0.084
0.1	100	0.168	0.208	0.208	0.200
0.1	1000	0.885	0.890	0.893	0.899
0.2	100	0.508	0.532	0.539	0.546
0.2	1000	0.999	0.999	0.999	0.999

5. Conclusions

This paper demonstrates that Prescreened Least Squares, where where explanatory variables are selected to maximize their conventional least squares significance level, or to maximize R^2, are statistically consistent and equivalent to nonlinear least squares under general regularity conditions. However, standard errors and test statistics provided by least squares software are biased. When explanatory variables are smooth in the parameters that index the selection alternatives, a Gauss-Newton auxiliary regression provides a convenient procedure for getting consistent covariance matrix estimates, and asymptotic tests from nonlinear least squares employing these covariance estimates are consistent. In situations where regularity conditions for Gauss- Newton auxiliary regression are not met, such as explanatory variables that are not smooth in the index parameters, or an isolated value for the true index parameter, bootstrap estimates of the covariance matrix appear from examples to be reasonably accurate in samples of moderate size.

APPENDIX: Proof of Theorem 1

I will need three preliminary lemmas, The first is a uniform stong law of large numbers established by Tauchen (1985); see also McFadden (2000), Theorem 4.24:

Lemma 1. *If* w_i *are independently and identically distributed for* $i = 1,...,n$, Γ *is compact,* $g:\Gamma\times\mathbb{R}^m \longrightarrow \mathbb{R}^q$ *is well-behaved and dominated by an integrable function r, then* $\mathbf{E}_w g(\gamma,w)$ *exists and is continuous on* Γ*, and a uniform strong law of large numbers holds; i.e.,*

$$\sup_{\gamma\in\Gamma} |n^{-1} \textstyle\sum_{i=1}^{n} g(\gamma,w_i) - \mathbf{E}_w g(\gamma,w)| \longrightarrow_{as} 0.$$

Let δ_n be a sequence of numbers. A random variable R_n is said to be of order $O_p(\delta_n)$, or $R_n = O_p(\delta_n)$, if for each $\varepsilon > 0$, there exists $M > 0$ such that $P(|R_n| > M\delta_n) < \varepsilon$ for all n sufficiently large. Then convergence in probability, $R_n \longrightarrow_p 0$, is equivalent to $R_n = O_p(\delta_n)$ for a sequence $\delta_n \longrightarrow 0$. For $n^{1/2}$-consistency under the conditions in A.4, I use a result on uniform stochastic boundedness which may be of independent interest.

Lemma 2. *Suppose* $y(\theta,w)$ *is a measurable function of a random vector* w *for each* θ *in an open set* $\Theta \subseteq \mathbb{R}^k$ *that contains* 0, *with* $y(0,w) \equiv 0$ *and* $\mathbf{E}y(\theta,w) = 0$. *Suppose* $y(\theta,w)$ *satisfies a Lipschitz condition* $|y(\theta',w) - y(\theta,w)| \le r(w)\cdot|\theta' - \theta|$ *for* θ *and* θ' *in* Θ, *where* $r(w)$ *is a random variable that has a proper moment generating function. Suppose* w_i *are independent draws. Then,* $\sup_{|\theta|<\delta}|n^{-1/2} \sum_{i=1}^{n} y(\theta,w_i)| = O_p(\delta)$. *In further detail, there exist positive constants* τ *and* A *determined by* $r(w)$ *such that for* $M^2 \ge 16\cdot kA$ *and* $n > (M/2A\tau)^2$,

$$P(\sup_{|\theta|<\delta} |n^{-1/2} \sum_{i=1}^{n} y(\theta,w_i)| > M\cdot\delta) \le 4\cdot\exp(-M^2/16A) .$$

Corollary. *Suppose the assumptions of Lemma* 2, *and suppose that* $y(\theta,w)$ *satisfies the Lipschitz condition* $|y(\theta',w) - y(\theta,w)| \le s(\delta,w)\cdot r(w)\cdot|\theta' - \theta|$ *for* θ *and* θ' *in* Θ *and* $|\theta| < \delta$, *with* $s(\delta,w)$ *a well-behaved function satisfying* $0 \le s(\delta,w) \le 1$ *and* $\mathbf{E}s(\delta,w) \longrightarrow 0$ *when* $\delta \longrightarrow 0$. *Then*

$$\sup_{|\theta|<\delta} |n^{-1/2} \sum_{i=1}^{n} y(\theta,w_i)| = o_p(\delta) .$$

Proof: There exists $\tau > 0$ such that $\mathbf{E}e^{tr(w)} < +\infty$ on an open interval that contains $[0,\tau]$; an implication of this is $A = \mathbf{E}r(w)^2\cdot e^{\tau r(w)} < +\infty$. Since $|y(\theta,w)| \le r(w)\cdot|\theta|$, the moment generating function of $y(\theta,w_i)$ exists for $|t\cdot\theta| \le \tau$ and has a Taylor's expansion

$$\mu(t) = \mathbf{E}e^{t\cdot y(\theta,w)} = 1 + t^2\cdot\mathbf{E}y(\theta,w)^2\cdot e^{\lambda t\cdot y(\theta,w)}/2$$

for some $\lambda \in (0,1)$, implying $\mu(t) \le 1 + t^2\cdot|\theta|^2\cdot\mathbf{E}r(w)^2\cdot e^{|t\theta|\cdot r(w)}/2 \le 1 + A|\theta|^2t^2/2$. The moment generating function of $n^{-1/2} \sum_{i=1}^{n} y(\theta,w_i)$ is then bounded by $[1 + A|\theta|^2t^2/2n]^n \le \exp(A|\theta|^2t^2/2)$ for $|t\cdot\theta|/n^{1/2} \le \tau$. Any random variable X satisfies the bound $P(|X| > m) \le \int_{x>m} e^{-tm}\cdot e^{tx}F(dx) + \int_{x<-m} e^{-tm}\cdot e^{-tx}F(dx) \le e^{-tm}\cdot[\mathbf{E}e^{tX} + \mathbf{E}e^{-tX}]$ for $t > 0$. Apply this inequality to obtain

$$\text{(A-1)} \qquad \sup_{|\theta|<\omega} P(n^{-1/2}|\sum_{i=1}^{n} y(\theta,w_i)| > m) \le 2\cdot\exp(-tm + A\omega^2t^2/2)$$

for $|t\cdot\theta|/n^{1/2} \le \tau$. I use this exponential inequality and a chaining argument to complete the proof. A hypercube centered at 0 with sides of length 2δ contains the θ satisfying $|\theta| < \delta$. Recursively partition this cube into 2^{kj} cubes with sides of length $2\delta\cdot2^{-j}$. Let Θ_j denote the set of centroids of these cubes for partition j, and let $\theta_j(\theta)$ denote the point in Θ_j closest to θ. Note that $|\theta_j(\theta) - \theta_{j-1}(\theta)| \le 2\delta\cdot2^{-j}$, and that

$$n^{-1/2} \sum_{i=1}^{n} y(\theta,w_i) = \sum_{j=1}^{\infty} n^{-1/2} \sum_{i=1}^{n} [y(\theta_j(\theta),w_i) - y(\theta_{j-1}(\theta),w_i)] .$$

Therefore, for $M^2 \ge 16\cdot kA$ and $n > (M/2A\tau)^2$, (A-2) $$P\left(\sup_{|\theta|<\delta} n^{-1/2}\left| \sum_{i=1}^{n} y(\theta,w_i)\right| > M\delta\right)$$

$$\le \sum_{j=1}^{\infty} P\left(\sup_{|\theta|<\delta} n^{-1/2} \sum_{i=1}^{n} \left|[y(\theta_j(\theta),w_i) - y(\theta_{j-1}(\theta),w_i)]\right| > j\cdot 2^{-j-2}\cdot M\delta\right)$$

$$\le \sum_{j=1}^{\infty} 2^{jk} \sum_{i=1}^{n} \max_{\Theta_j} P\left(n^{-1/2}\left| \sum_{i=1}^{n} [y(\theta_j,w_i) - y(\theta_{j-1},w_i)]\right| > j\cdot 2^{-j-2}\cdot M\delta\right)$$

$$\le \sum_{j=1}^{\infty} 2^{jk}\cdot 2\cdot \exp(-j\cdot 2^{-j-2}\cdot M\delta\cdot t_j + A\delta^2 2^{1-2j} t_j^2) \quad \text{for } t_j\delta\cdot 2^{1-j} \le \tau n^{1/2}$$

$$\le \sum_{j=1}^{\infty} 2^{jk}\cdot 2\cdot \exp(-(j-1/2)\cdot M^2/4A) \quad \text{for } t_j = 2^{j-2}M/A\delta \text{ and } n > (M/2A\tau)^2$$

$$\le \frac{2^{k+1}\cdot \exp(-M^2/8A)}{1-2^k\cdot \exp(-M^2/4A)} \le 4\cdot \exp(-M^2/16A) \text{ for } M^2 \ge 16\cdot kA.$$

The first inequality in (A-2) holds because the event on the left is contained in the union of the events on the right. The second inequality holds because the event on the left at partition level j is contained in one of the 2^{jk} events on the right for the partition centroids. The third inequality is obtained by applying the exponential inequality (A-1) to the difference $y(\theta_j,w_i) - y(\theta_{j-1},w_i)$, with m replaced by $M\cdot j\cdot 2^{-j-2}$ and ω replaced by $\delta\cdot 2^{1-j}$. The next inequality is obtained by substituting the specified t_j arguments, and the final inequality is obtained by using the lower bound imposed on M and summing over j.

To prove the corollary, note that the moment generating function of $y(\theta,w_i)$ now satisfies $\mu(t) \le 1 + A|\theta|^2 t^2/2$ for $|t\cdot\theta| \le \tau/2$, with $A = \mathbf{E}s(\delta,w)^2\cdot r(w)^2\cdot e^{\tau\cdot r(w)} \le \{[\mathbf{E}s(\delta,w)^4]\cdot[\mathbf{E}r(w)^4\cdot e^{2\tau r(w)}]\}^{1/2}$ by the Cauchy-Schwartz inequality. The remainder of the proof of the lemma holds for this A, and $\mathbf{E}s(\delta,w)^4 \le \mathbf{E}s(\delta,w) \longrightarrow 0$ implies $A \longrightarrow 0$ as $\delta \longrightarrow 0$, giving the result. ■

Let θ be a finite-dimensional parameter varying in a set Θ. Suppose θ_o minimizes a function $q(\theta)$ defined on Θ. Suppose further that a sample analog $Q_n(\theta)$ is minimized at a point T_n that converges in probability to θ_o. Define $Q_n(\theta) - q(\theta)$ to be uniformly convergent at rate $O_p(\varepsilon_n)$ on $O_p(\delta_n)$ neighborhoods of θ_o if for each sequence of random variables R_n of order $O_p(\delta_n)$ there exists a sequence of random variables S_n of order $O_p(\varepsilon_n)$ such that $sup_{|\theta-\theta_0|\le R_n} |Q_n(\theta) - q(\theta)| \le S_n$. I will employ conditions for $n^{1/2}$-consistency established by Sherman (1993):

Lemma 3. *Suppose T_n minimizes $Q_n(\theta)$ and θ_o minimizes $q(\theta)$ on Θ. If*

(i) $|T_n - \theta_o| = O_p(\delta_n)$ *for a sequence of numbers* $\delta_n \longrightarrow 0$,
(ii) *for some neighborhood* Θ_o *of* θ_o *and constant* $\kappa > 0$, $q(\theta) - q(\theta_o) \ge \kappa|\theta|^2$,
(iii) *uniformly over* $O_p(\delta_n)$ *neighborhoods of* θ_o,

$$Q_n(\theta) - q(\theta) - Q_n(\theta_o) + q(\theta_o) = O_p(n^{-1/2}|\theta-\theta_o|) + o_p(|\theta-\theta_o|^2) + O_p(1/n),$$

then $|T_n - \theta_o| = O_p(n^{-1/2})$. *Suppose, in addition, that uniformly over* $O_p(n^{-1/2})$ *neighborhoods of* θ_o, $Q_n(\theta) - Q_n(\theta_o) = (\theta - \theta_o)'R_n/n^{1/2} + (\theta - \theta_o)J(\theta - \theta_o)/2 + o_p(1/n)$, *where* R_n *is aymptotically normal with mean* 0 *and covariance matrix* K. *Then,* $n^{1/2}(T_n - \theta_o) \longrightarrow_d N(0,J^{-1}KJ^{-1})$.

I turn now to the proof of Theorem 1. Absorb x into $z(\gamma,w)$, so that the model is simply $y = z(\gamma,w)\beta + u$. A.1 and A.3 imply that $z(\gamma,w)'z(\gamma,w)$ is well-behaved and dominated by integrable $r(w)^2$, and $z(\gamma,w)'y$ is well-behaved and dominated by $r(w)y$ which satisfies $\mathbf{E}|r(w)y| \le \mathbf{E}r(w)^2|\beta_o|$. Then, by Lemma 1, $H(\gamma) = \mathbf{E}z(\gamma,w)'z(\gamma,w)$ and $\psi(\gamma) = \mathbf{E}z(\gamma,w)'y$ are continuous in γ, and

$$\sup_{\gamma\in\Gamma} \left|n^{-1} \sum_{i=1}^{n} z(\gamma,w)'z(\gamma,w) - H(\gamma)\right| \longrightarrow_{as} 0 \text{ and } \sup_{\gamma\in\Gamma} \left|n^{-1} \sum_{i=1}^{n} z(\gamma,w)'y - \psi(\gamma)\right| \longrightarrow_{as} 0 .$$

From A.1, $H(\gamma)$ is positive definite. Therefore, $\beta(\gamma) = H(\gamma)^{-1}\psi(\gamma)$ and $s(\gamma) \equiv \beta(\gamma)'H(\gamma)\beta(\gamma)$ are continuous on Γ, $\sup_{\gamma\in\Gamma} |S(\gamma) - s(\gamma)| \longrightarrow_{as} 0$, and $\sigma^2(\gamma) + s(\gamma) = \mathbf{E}y^2$. Given $\varepsilon,\delta > 0$, one has $inf_{\gamma\in\Gamma\&|\gamma-\gamma_o|\ge\delta}\ \sigma^2(\gamma)/\sigma_o^2 = \lambda > 1$ from A.2. Note that $\sigma^2(\gamma) \ge \lambda\cdot\sigma_o^2$ implies $s(\gamma_o) \ge s(\gamma) + (\lambda-1)\sigma_o^2$.

Choose n such that $\text{Prob}(\sup_{n'\ge n} \sup_{\gamma\in\Gamma} |S(\gamma) - s(\gamma)| > (\lambda-1)\sigma_o^2/3) < \varepsilon$. For $n' \ge n$ and $|\gamma-\gamma_o| \ge \delta$, one then has with probability at least $1-\varepsilon$ the inequality $S(\gamma_o) \ge S(\gamma) + (\lambda-1)\sigma_o^2/3$. Then with at least this probability, the maximand c of $S(\gamma)$ is contained within a δ-neighborhood of γ_o for all $n' \ge n$. Therefore $c \longrightarrow_{as} \gamma_o$. Uniform a.s. convergence then implies $s^2 = SSR(c)/(n-k) \longrightarrow_{as} \sigma_o^2$. Let $b(\gamma) = \left[n^{-1}\sum_{i=1}^{n} z(\gamma,w_i)'z(\gamma,w_i)\right]^{-1} n^{-1}\sum_{i=1}^{n} z(\gamma,w_i)'y_i$. Then $\sup_{\gamma\in\Gamma} |b(\gamma) - \beta(\gamma)| \longrightarrow_{as} 0$, and hence $b(c) \longrightarrow_{as} \beta_o$. This establishes strong consistency.

Suppose γ_o is an isolated point of Γ. Then an implication of $c \longrightarrow_{as} \gamma_o$ is that $\text{Prob}(c \ne \gamma_o$ for $n' \ge n) \longrightarrow 0$, and hence $\text{Prob}(n^{1/2}\cdot(c-\gamma_o)\ne 0) \longrightarrow 0$. Hence for $\varepsilon,\delta > 0$, there exists n_o such that $\text{Prob}(n^{1/2}\cdot|\beta(c)-\beta(\gamma_o)| > \delta) \le \text{Prob}(n^{1/2}\cdot(c-\gamma_o)\ne 0) < \varepsilon$ for $n \ge n_o$. But $n^{1/2}\cdot(b(\gamma_o)-\beta_o) \longrightarrow_d N(0,\sigma_o^2H(\gamma_o)^{-1})$ by application of the Lindeberg-Levy central limit theorem to $n^{-1/2} \sum_{i=1}^{n} \left[z(\gamma_o,w_i)'y_i - H(\gamma_o)\beta_o\right]$. Combining these results establishes that $n^{1/2}\cdot(b(c)-\beta_o) \longrightarrow_d N(0,\sigma_oH(\gamma_o)^{-1})$. This establishes asymptotic normality in the case of isolated γ_o.

Suppose A.4 holds. Then $d(\gamma,\beta,w) = \nabla_\gamma z(\gamma,w)\beta$ exists a.s. and is well-behaved, implying that it is piecewise a.s. continuous in γ. Given $\varepsilon > 0$, there a.s. exists $\delta(\varepsilon,\gamma,w) > 0$ such that $|\gamma' - \gamma| < \delta(\varepsilon,\gamma,w)$ implies $|d(\gamma',\beta,w) - d(\gamma,\beta,w)| < \varepsilon|\beta|$ and, by the theorem of the mean, $(z(\gamma',w) - z(\gamma,w))\beta = d(\gamma^*,\beta,w)(\gamma'-\gamma)$ for $\gamma^* \in (\gamma,\gamma')$. Let $W(\varepsilon,\gamma,\nu)$ be the set of w for which $\delta(\varepsilon,\gamma,w) < \nu$. The a.s. continuity of $d(\gamma,\beta,w)$ implies the (outer) probability of $W(\varepsilon,\gamma,\nu)$ approaches zero as $\nu \longrightarrow 0$. Choose ν such that the probability of $W(\varepsilon,\gamma,\nu)$ is less than ε and $\mathbf{E}r(w)\cdot\mathbf{1}(w\in W(\varepsilon,\gamma,\nu)) < \varepsilon$. Then, $\lambda(\gamma,\gamma',\beta,w) \equiv (z(\gamma',w) - z(\gamma,w))\beta - d(\gamma,\beta,w)(\gamma'-\gamma)$ satisfies the bounds $|\lambda(\gamma,\gamma'\beta,w)| \le \varepsilon\cdot|\beta|\cdot|\gamma'-\gamma|$ for $w \notin W(\varepsilon,\gamma,\nu)$ and $|\lambda(\gamma,\gamma'\beta,w)| \le 3\cdot r(w)\cdot|\beta|\cdot|\gamma'-\gamma|$ a.s. Since ε is arbitrary, this establishes that $\lambda(\gamma,\gamma',\beta,w) = o_p(|\beta|\cdot|\gamma'-\gamma|)$ and $\mathbf{E}\lambda(\gamma,\gamma',\beta,w) = o(|\beta|\cdot|\gamma'-\gamma|)$. The same argument establishes the approximations

$$\text{(A-3)} \qquad \mathbf{E}z(\gamma,w)'(z(\gamma,w) - z(\gamma_o,w))\beta = \mathbf{E}z(\gamma,w)'d(\gamma_o,\beta,w)(\gamma-\gamma_o) + o(|\gamma-\gamma_o|)$$

and

(A-4) $\quad \beta'\mathbf{E}(z(\gamma,w) - z(\gamma_o,w))'(z(\gamma,w) - z(\gamma_o,w))\beta = (\gamma-\gamma_o)'\mathbf{E}d(\gamma_o,\beta,w)'d(\gamma_o,\beta,w)(\gamma-\gamma_o) + o(|\gamma-\gamma_o|^2)$.

Define $\theta = (\gamma,\beta)$ and $Q_n(\theta) = n^{-1} \sum_{i=1}^{n} (y_i - z(\gamma,w_i)\beta)^2$. Then $q(\theta_o) = \sigma_o^2$ and $q(\theta) = \mathbf{E}(y - z(\gamma,w)\beta)^2$

$\equiv \sigma_o^2 + \mathbf{E}(z(\gamma_o,w)\beta_o - z(\gamma,w)\beta)^2$. Write $(z(\gamma,w)\beta - z(\gamma_o,w))\beta_o = z(\gamma_o,w)(\beta-\beta_o) + (z(\gamma,w) - z(\gamma_o,w))\beta = z(\gamma_o,w)(\beta-\beta_o) + d(\gamma_o,\beta,w)(\gamma-\gamma_o) + \lambda(\gamma_o,\gamma,\beta,w)$. Then

(A-5) $$q(\theta) - q(\theta_o) = (\beta-\beta_o)'\mathbf{E}z(\gamma_o,w)'z(\gamma_o,w)(\beta-\beta_o) + 2(\beta-\beta_o)'\mathbf{E}z(\gamma_o,w)'(z(\gamma,w) - z(\gamma_o,w))\beta + \beta_o'\mathbf{E}(z(\gamma,w) - z(\gamma_o,w))'(z(\gamma,w) - z(\gamma_o,w))\beta_o$$

$$= {}' \begin{bmatrix}\beta-\beta_o\\ \gamma-\gamma_o\end{bmatrix}' \begin{bmatrix} \boldsymbol{E}z(\gamma,w)'z(\gamma,w) & \boldsymbol{E}z(\gamma,w)'d(\gamma_o,\beta,w) \\ \boldsymbol{E}d(\gamma_o,\beta,w)'z(\gamma,w) & \boldsymbol{E}d(\gamma_o,\beta_o,w)'d(\gamma_o,\beta,w)\end{bmatrix}\begin{bmatrix}\beta-\beta_o\\ \gamma-\gamma_o\end{bmatrix}$$

$$+ 2\cdot\mathbf{E}\lambda(\gamma,\beta,w)[d(\gamma_o,\beta,w)(\gamma-\gamma_o) + z(\gamma_o,w)(\beta-\beta_o)] + \mathbf{E}\lambda(\gamma,\beta,w)^2 .$$

The final terms in this expression are $o(|\theta-\theta_o|^2)$, and the conditions from A.4 that $J(\gamma,\beta)$ is continuous and $J(\gamma_o,\beta_o)$ is positive definite imply that $q(\theta) - q(\theta_o) = (\theta-\theta_o)'J(\gamma_o,\beta_o)(\theta-\theta_o) + o(|\theta-\theta_o|^2) \geq \kappa|\theta-\theta_o|^2$ for some $\kappa > 0$ on a neighborhood of θ_o. Consider the expression

(A-6) $$Q_n(\theta) - q(\theta) - Q_n(\theta_o) + q(\theta_o) = n^{-1} \sum_{i=1}^{n} u_i(z(\gamma,w_i)\beta - z(\gamma_o,w_i))\beta_o)$$

$$+ n^{-1} \sum_{i=1}^{n} \{(z(\gamma,w_i)\beta - z(\gamma_o,w_i))\beta_o)^2 - \mathbf{E}(z(\gamma,w)\beta - z(\gamma_o,w))\beta_o)^2\}.$$

The first term on the right-hand-side of this expression can be written

(A-7) $$n^{-1} \sum_{i=1}^{n} u_i(z(\gamma,w_i)\beta - z(\gamma_o,w_i))\beta_o)$$

$$= n^{-1} \sum_{i=1}^{n} u_i \begin{bmatrix}z(\gamma_o,w_i) & d(\gamma_o,\beta,w_i)\end{bmatrix}\begin{bmatrix}\beta-\beta_o\\ \gamma-\gamma_o\end{bmatrix} + n^{-1} \sum_{i=1}^{n} u_i\cdot\lambda(\gamma_o,\gamma,\beta,w_i).$$

The Lindeberg-Levy central limit theorem implies

(A-8) $$R_n \equiv n^{-1/2} \sum_{i=1}^{n} u_i \begin{bmatrix}z(\gamma_o,w_i) & d(\gamma_o,\beta,w_i)\end{bmatrix} \longrightarrow_d N(0,\sigma_o^2 J(\gamma_o,\beta)) ;$$

hence $n^{-1} \sum_{i=1}^{n} u_i \begin{bmatrix}z(\gamma_o,w_i) & d(\gamma_o,\beta,w_i)\end{bmatrix}\begin{bmatrix}\beta-\beta_o\\ \gamma-\gamma_o\end{bmatrix} = O_p(n^{-1/2}|\theta-\theta_o|)$. The assumptions in A.4 that u has a proper moment generating function and $r(w)$ is bounded implies that $u\cdot\lambda(\gamma_o,\gamma,\beta,w)$ has a proper moment generating function. An application of the Corollary to Lemma 2, utilizing the condition that $\lambda(\gamma,\gamma',\beta,w) = o_p(|\beta|\cdot|\gamma'-\gamma|)$, implies that uniformly $n^{-1} \sum_{i=1}^{n} u_i\lambda(\gamma_o,\gamma,\beta,w) = o_p(n^{-1/2}|\beta|\cdot|\gamma-\gamma_o|)$.

Finally, write

$$\text{(A-9)} \qquad n^{-1} \sum_{i=1}^{n} \{(z(\gamma,w_i)\beta - z(\gamma_o,w_i))\beta_o)^2 - \mathbf{E}(z(\gamma,w)\beta - z(\gamma_o,w))\beta_o)^2\}$$

$$= n^{-1} \sum_{i=1}^{n} \{d(\gamma_o,\beta,w_i)(\gamma-\gamma_o) + z(\gamma_o,w_i)(\beta-\beta_o) + \lambda(\gamma_o,\gamma,\beta,w_i)\}^2$$

$$- \mathbf{E}\{d(\gamma_o,\beta,w)(\gamma-\gamma_o) + z(\gamma_o,w)(\beta-\beta_o) + \lambda(\gamma_o,\gamma,\beta,w)\}^2.$$

The squares and cross-products that do not involve $\lambda(\gamma_o,\gamma,\beta,w)$ will be $O_p(n^{-1/2}|\theta-\theta_o|^2)$ by the Lindeberg-Levy central limit theorem. The terms that involve $\lambda(\gamma_o,\gamma,\beta,w)$ or its square will be uniformly $o_p(n^{-1/2}|\theta-\theta_o|^2)$ by the Corollary to Lemma 2. Conditions (i) to (iii) of Lemma 3 are now established, implying the nonlinear least squares estimates satisfy $T_n - \theta_o = O_p(n^{-1/2})$. Finally, observe from (A-5) to (A-8) that

$$Q_n(\theta) - Q_n(\theta_o) = 2n^{-1} \sum_{i=1}^{n} u_i \begin{bmatrix} z(\gamma_o,w_i) & d(\gamma_o,\beta,w_i) \end{bmatrix} \begin{bmatrix} \beta-\beta_o \\ \gamma-\gamma_o \end{bmatrix} + o_p(n^{-1/2}|\beta|\cdot|\gamma-\gamma_o|)$$

$$+ \begin{bmatrix} \beta-\beta_o \\ \gamma-\gamma_o \end{bmatrix}' \begin{bmatrix} \boldsymbol{E}z(\gamma,w)'z(\gamma,w) & \boldsymbol{E}z(\gamma,w)'d(\gamma_o,\beta,w) \\ \boldsymbol{E}d(\gamma_o,\beta,w)'z(\gamma,w) & \boldsymbol{E}d(\gamma_o,\beta_o,w)'d(\gamma_o,\beta,w) \end{bmatrix} \begin{bmatrix} \beta-\beta_o \\ \gamma-\gamma_o \end{bmatrix} + o(|\theta-\theta_o|^2)$$

$$= 2n^{-1/2}R_n \begin{bmatrix} \beta-\beta_o \\ \gamma-\gamma_o \end{bmatrix} + \begin{bmatrix} \beta-\beta_o \\ \gamma-\gamma_o \end{bmatrix}' J(\theta_o) \begin{bmatrix} \beta-\beta_o \\ \gamma-\gamma_o \end{bmatrix} + o_p(n^{-1/2}|\beta|\cdot|\gamma-\gamma_o|) + o(|\theta-\theta_o|^2) .$$

Then, the final condition of Lemma 3 is satisfied, and the nonlinear least squeares estimator is asymptotically normal with covariance matrix $\sigma_o^2 J(\theta_o)^{-1}$. ■

REFERENCES

Amemiya, T. (1978) "On a Two-Step Estimation of a Multinomial Logit Model," Journal of Econometrics 8, 13-21.

Amemiya, T. (1980) "Selection of Regressors," International Economic Review, 21, 331-354.

Amemiya, T. (1985) Advanced Econometrics, Harvard University Press: Cambridge.

Andrews, D. (1993) "Tests for Parameter Instability and Structural Change with Unknown Change Point," Econometrica, 61, 821-856.

Beran, R. (1988) "Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements," Journal of the American Statistical Association, 83, 687-697.

Breiman, L. and D. Freedman (1983) "How Many Variables Should Enter a Regression Equation," Journal of the American Statistical Association, 78, 131-136.

Brown, B. and W. Newey (1995) "Bootstrapping for GMM," MIT Working Paper.

Brownstone, D. (1990) "Bootstrapping Improved Estimators for Linear Regression Models," Journal of Econometrics, 44, 171- 188.

Brownstone, D. (1992) "Bootstrapping Admissible Model Selection Procedures," in R. LePage and L. Billard (eds.) Exploring the Limits of Bootstrap, New York: Wiley, 327-344.

Davidson, R. and J. MacKinnon (1993) Estimation and Inference in Econometrics, Oxford: New York.

Davidson, R. and E. Flachaire (1996) "Monte Carlo Evidence on the Behavior of the Wild Bootstrap in the Presence of Leverage Points," NSF Symposium on the Bootstrap, University of California, Berkeley.

Freedman, D. (1983) "A Note on Screening Regression Equations," The American Statistician, 37, 152-155.

Freedman, D., W. Navidi, S. Peters (1988) "On the Impact of Variable Selection in Fitting Regression Equations," in T. Dijkstra (ed) On Model Uncertainty and its Statistical Implications, Springer, Berlin.

Gallant, R. (1987) Nonlinear Statistical Models, Wiley: New York.

Jenrich, R.(1969) "Asymptotic Properties of Nonlinear Least Squares Estimators," The Annals of Mathematical Statistics, 40, 633-643.

Judge, G. and M. Bock (1978) The Statistical Implications of Pre- Test and Stein-rule Estimators in Econometrics, North Holland: Amsterdam.

Liu, R. (1988) "Bootstrap Procedure under Some Non-I.I.D. Models," Annals of Statistics, 16, 1696-1708.

Lovell, M. (1983) "Data Mining," Review of Economics and Statistics, 65, 1-12.

Martin, M. (1990) "On Bootstrap Iteration for Coverage Correction in Confidence Intervals," Journal of the American Statistical Association, 85, 1105-1118.

Newey, W. and D. McFadden (1994) "Large Sample Estimation and Hypothesis Testing," in R. Engle and D. McFadden (eds) Handbook of Econometrics, 4, 2111-2245.

Peters, S. and D. Freedman (1984) "Some Notes on the Bootstrap in Regression Problems," Journal of Business and Economic Statistics, 2, 406-409.

Pakes, A. and D. Pollard (1989) "Simulation and the Asymptotics of Optimization Estimation," Econometrica, 57, 1027-1057.

Pollard, D. (1984) Convergence of Stochastic Processes, Springer- Verlag: Berlin.

Sherman, R. (1993) "The Limiting Distribution of the Maximum Rank Correlation Estimator," Econometrica 61, 123-137.

Stock, J. (1994) "Unit Roots, Structural Breaks, and Trends," in R. Engle and D. McFadden (eds) Handbook of Econometrics, 4, 2739-2841.

Tauchen, G. (1985) "Diagnostic Testing and Evaluation of Maximum Likelihood Models," Journal of Econometrics, 30, 415-433.

Veal, M. (1992) "Bootstrapping the Process of Model Selection: An Econometric Example," Journal of Applied Econometrics, 7, 93-99.