WORKING PAPER
ALFRED P. SLOAN SCHOOL OF MANAGEMENT

CONSISTENT ESTIMATION OF SCALED COEFFICIENTS

by

Thomas M. Stoker

July 1984

WP #1583-84 Revised November 1985

MASSACHUSETTS
INSTITUTE OF TECHNOLOGY
50 MEMORIAL DRIVE
CAMBRIDGE, MASSACHUSETTS 02139

CONSISTENT ESTIMATION OF SCALED COEFFICIENTS

by

Thomas M. Stoker

July 1984

WP #1583-84 Revised November 1985

CONSISTENT ESTIMATION OF SCALED COEFFICIENTS

by

Thomas M. Stoker

July, 1984, revised November 1985

ABSTRACT

This paper studies the estimation of coefficients β in single index models such that $E(y|X)=F(\alpha+X'\beta)$, where the function F is misspecified or unknown. A general connection between behavioral derivatives and covariance estimators is established, which shows how β can be estimated up to scale using information on the marginal distribution of X. A sample covariance estimator and an instrumental variables slope coefficient vector are proposed, which are constructed using appropriately defined score vectors of the X distribution. The framework is illustrated using several common limited dependent variable models, and extended to multiple index models, including models of selection bias and multinomial discrete choice. The asymptotic bias in the OLS coefficients of y regressed on X are analyzed. The asymptotic distribution of the instrumental variables estimator is established, when the X distribution is modeled up to a finite parameterization.

* Thomas M. Stoker is Associate Professor of Applied Economics, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts, 02139. This research was funded by National Science Foundation Grant No. SES-8410030. A large number of colleagues have made valuable comments on this paper and related research, which was presented at numerous seminars. Special thanks go to J. Powell and D. McFadden for ongoing discussions, and G. Chamberlain, A. Deaton, J. Hausmann, J. Heckman, C. Manski, W. Newey, P. Phillips, A. Zellner and the referees for helpful comments.

CONSISTENT ESTIMATION OF SCALED COEFFICIENTS

1. Introduction

This paper considers the generic econometric modeling situation in which a dependent variable y is modeled as a function of a vector of explanatory variables X and stochastic terms, where the conditional expectation of y given X can be written in the single index form $E(y|X) = F(\alpha+X'\beta)$. This situation exists for many standard models of discrete choice, censoring and selection, but is clearly not limited to such models. The question of interest is what can be learned about the coefficients β without specific assumptions on the distribution of unobserved stochastic terms or other functional form aspects; in other words, when the true form of the function F is misspecified or unknown.[1]

For different examples of limited dependent variables models, several researchers have studied the conditions under which ordinary least squares (OLS) regression coefficients and other quasi-maximum likelihood estimators will consistently estimate β up to a scalar multiple. Ruud(1983) points out that a sufficient condition for this property occurs when the conditional expectation of each component of X given $Z = \alpha+X'\beta$ is linear in Z, which is valid when X is multivariate normally distributed, for instance. Chung and Goldberger(1984) and Deaton and Irish(1984) point out the sufficiency of an analogous condition with a more general definition of Z.[2]

An intriguing feature of this work is that it provides special cases where knowledge of the marginal distribution of X is very useful for estimating behavioral effects when certain features of the true model are unknown. The question is immediately raised as to whether more general results of this type can be found, because as Ruud(1983) states, the above sufficient

conditions are "too restrictive to be generally applicable." Results that apply to more general marginal distribution forms are of substantial practical interest because, in general, the marginal distribution of X can be empirically characterized. In this spirit, Ruud(1984) has proposed an estimation technique based on reweighting the data sample so that weighted X distribution is multivariate normal.

This paper proposes an approach for studying β based on estimation of average behavioral derivatives, and shows how information on the marginal distribution of X can be used to estimate average derivatives. In particular, a direct link between average derivatives and covariance estimators is established, which shows how β can be estimated up to a scalar multiple by the sample covariance between y and appropriately defined score vectors of the marginal distribution of X. β is also consistently estimated up to scale by the slope coefficients of the linear equation of y regressed on X using the score vectors as instrumental variables.

The ratio of any two components of either the sample covariance or the instrumental variables coefficient vector will consistently estimate the ratio of the corresponding components of β. These estimates may suffice for many applications, such as judging relative marginal utilities in a discrete choice situation. More broadly, the ratio estimates provide a consistent benchmark for choosing specific modeling assumptions. For instance, in a binary discrete choice situation, separate estimates of β under logit or probit modeling assumptions can be judged in relation to the consistent ratio estimates. This method of assessing specification may be useful in any modeling situation where alternative functional form or stochastic assumptions give rise to substantively different estimates of β.

The exposition begins with notation, examples and formal assumptions in Section 2. Section 3 establishes the connection between average derivatives

and covariance estimators. Section 4 applies the result to the estimation of β up to scale, as well as to estimation of parameters of more general multiple index models. Section 5 studies the asymptotic bias of the vector of OLS coefficients of y regressed on X, as an estimator of the true average derivative, and as an estimator of β up to scale. Section 6 establishes the asymptotic distribution of the proposed instrumental variables estimator, when the distribution of X is modeled up to a finite parameterization. Section 7 contains concluding remarks and topics for future research.

2. Notation, Examples and Basic Assumptions

Consider the situation where data is observed on a dependent variable y_i and an M-vector of explanatory variables X_i for i=1,...,N, where $M \geq 2$. (y_i,X_i), i=1,...,N represent random drawings from a distribution Υ which is absolutely continuous with respect to a σ-finite measure ν, with Radon-Nikodym density $P(y,X)=\partial\Upsilon/\partial\nu$. P(y,X) factors as P(y,X)=q(y|X)p(X), where p(X) is the density of the marginal distribution of X. The conditional density q(y|X) represents the true behavioral econometric model, which we assume permits the conditional expectation E(y|X) to be written in the form

$$(2.1) \qquad E(y|X) = F(\alpha+X'\beta) = F(Z)$$

for some function F, where α is a constant, $\beta=(\beta_1,\ldots,\beta_M)'$ is an M-vector of constants, and Z is defined as $Z=\alpha+X'\beta$. I refer to Z as an index variable, with (2.1) a single index model.

This framework is very general, subsuming many limited dependent variables models, but is not restricted to such models. Before proceeding to specific examples, it is useful to note a generic special case of (2.1). Suppose that Z^* is a general index variable such that $\varepsilon=Z^*-Z$ is independent of X, then if $E(y|X,\varepsilon)=F^*(Z^*)$ for some function F^*, (2.1) is implied. This

includes many models that employ a latent variable $Z^*=\alpha+X'\beta+\varepsilon$, where ε is independent of X. Note also that this implies that behavioral variables can be omitted from X without affecting the results, provided that the omitted variables are independent of the included ones.[3]

I now turn to some specific examples:

Example 1: Binary Discrete Choice

Suppose that y represents a dichotomous random variable modeled as

$$y = 1 \quad \text{if } \varepsilon > -(\alpha+X'\beta)$$
$$= 0 \quad \text{otherwise}$$

Here $E(y|X)=F(\alpha+X'\beta)$ is the probability of y=1 given the value of X, with the true function F determined by the distribution of ε. If ε is distributed normally with mean 0 and variance σ^2, then the familiar probit model results, with $F(\alpha+X'\beta)=\Phi((\alpha+X'\beta)/\sigma)$, where Φ is the cumulative normal distribution function. Logit models, etc., can easily be included.

Example 2: Tobit Models

Suppose that y is equal to an index Z^* only if Z^* is positive, as in the censored tobit specification

$$y = \alpha+X'\beta+\varepsilon \quad \text{if } \varepsilon > -(\alpha+X'\beta)$$
$$= 0 \quad \text{otherwise}$$

Alternatively, if y and X are observed only when $\varepsilon > -(\alpha+X'\beta)$, we have the truncated tobit specification.

Example 3: Dependent Variable Transformations

Suppose there exists a function g(y) such that the true model is of the form

$$g(y) = \alpha+X'\beta+\varepsilon$$

where g(y) is invertible everywhere except for a set of measure 0. A specific example is the familiar Box-Cox transformation where

$$y^{(\lambda)} = \alpha+X'\beta+\varepsilon$$

with $y^{(\lambda)}=[(y^{\lambda}-1)/\lambda]$ for $\lambda\neq 0$, $y^{(\lambda)} = \ln(y)$ for $\lambda=0$.

These examples serve to illustrate the wide spectrum of models covered by the single index form (2.1) with general function F, and many other examples can be found. Multiple index models are considered in Section 4.2.

We now turn to the other required assumptions. X is assumed to be continuously distributed, having support Ω of the following form:[4]

Assumption 1: Ω is a convex subset of R^M with nonempty interior. The underlying measure ν can be written in product form as $\nu=\nu_y\times\nu_X$, where ν_X is Lebegue measure on R^M.

Therefore, no component of X is functionally determined by other components of X, and no two components of X are perfectly correlated.

Denote $\ell(X)$ as the score vector[5] of the marginal density p(X) as:

$$\ell(X) = -\frac{\partial \ln p(X)}{\partial X} \tag{2.2}$$

The main regularity conditions on the marginal density p(X) are

Assumption 2: $p(X)$ is continuously differentiable in the components of X for all X in the interior of Ω. $E(\ell)$ and $E(\ell\ell')$ exist.

Assumption 3: For $X \in d\Omega$, where $d\Omega$ is the boundary of Ω, we have $p(X)=0$.

Assumption 3 allows for unbounded X's, where $\Omega=R^M$ and $d\Omega=\emptyset$. While the majority of the results employ Assumptions 2 and 3, the incorporation of discrete (qualitative) explanatory variables is discussed in Section 4.2.

I will make reference to the following set of regularity conditions on a general random variable $\tilde{y}$ and its conditional expectation $E(\tilde{y}|X) \equiv G(X)$. $(\tilde{y}, G(X))$ satisfies condition A if

Condition A: $G(X)$ is continuously differentiable for all $X \in \bar{\Omega}$, where $\bar{\Omega}$ differs from Ω by a set of measure 0. $E(\tilde{y})$, $E(\partial G/\partial X)$ and $E(\ell\tilde{y})$ exist.

The main regularity condition on the behavioral model (2.1) is contained in

Assumption 4: a) $(y, F(\alpha+X'\beta))$ satisfies condition A. $E(dF/dZ)$ is nonzero.

b) (X_j, X_j) satifies condition A for each $j=1,\ldots,M$.

This completes the list of main assumptions. While somewhat formidable technically, these assumptions are collectively very weak.

The main thrust of the paper concerns how information on the marginal density $p(X)$ can be used to estimate β up to scale. Consequently, the majority of the exposition assumes that the value of $\ell(X)$ at each X_i is known, and denoted $\ell_i=\ell(X_i)$, $i=1,\ldots,N$. Use of empirical characterizations of $p(X)$ is discussed in Section 6.

Finally, sample averages are denoted via overbars as in $\bar{y}= \Sigma y_i/N$, with

the means of y and X denoted as $\mu_y=E(y)$ and $\mu_X=E(X)$. Sample covariances are denoted using S as in $S_{\ell y}= \Sigma(\ell_i-\bar{\ell})(y_i-\bar{y})/N$, with population counterparts denoted using Σ as in $\Sigma_{\ell y}=Cov(\ell,y)$.

3. Behavioral Derivatives and Covariance Estimators

This section presents a fundamental connection between behavioral derivatives and covariance estimators, that is the basis of the consistency results of Section 4. The connection is given in the following theorem, which is interpreted after the proof.

Theorem 1: Given Assumptions 1-3, if $(\tilde{y},G(X))$ satisfies condition A, then

$$(3.1) \qquad E\left[\frac{\partial G}{\partial X}\right] = E(\ell(X)\tilde{y}) = \Sigma_{\ell\tilde{y}}$$

Proof: Let X_1 denote the first component of X, and apply Fubini's Theorem (c.f. Billingsley(1979), among others) to write $E(\partial G/\partial X_1)$ as

$$(3.2) \qquad \int_{\Omega} \frac{\partial G(X)}{\partial X_1}\, p(X)\, d\nu = \int\left[\int_{\omega(X_o)} \frac{\partial G(X)}{\partial X_1}\, p(X) d\nu_1(X_1)\right] d\nu_o(X_o)$$

where X_o represents the other components of X. The result that $E(\partial G/\partial X_1)=E(\ell_1(X)\tilde{y})$ is implied by the validity of the following equation

$$(3.3) \qquad \int_{\omega(X_o)} \frac{\partial G(X)}{\partial X_1}\, p(X) d\nu_1(X_1) = -\int_{\omega(X_o)} G(X)\, \frac{\partial p(X)}{\partial X_1}\, d\nu_1(X_1)$$

since the RHS of (3.3) simplifies to

$$(3.4) \qquad -\int_{\omega(X_o)} G(X)\, \frac{\partial p(X)}{\partial X_1}\, d\nu_1(X_1) = \int_{\omega(X_o)} G(X)\left[-\frac{\partial \ln p(X)}{\partial X_1}\right] p(X) d\nu_1(X_1)$$

By inserting (3.3,4) into (3.2), $E(\partial G(X)/\partial X_1)=E(\ell_1(X)G(X))$ is established, and

by iterated expectation, $E(\ell_1(X)G(X))=E(\ell_1(X)\tilde{y})$.

To establish (3.3), note first that the convexity of Ω implies that $\omega(X_o)$ is either a finite interval $[a,b]$ (where a, b depend on X_o), or an infinite interval of the form $[a,\infty)$, $(-\infty,b]$ or $(-\infty,\infty)$. Supposing first that $\omega(X_o)=[a,b]$, integrate the LHS of (3.3) by parts (c.f. Billingsley(1979)) as

$$(3.5)\qquad \int_a^b \frac{\partial G(X)}{\partial X_1}\, p(X)d\nu_1(X_1) = -\int_a^b G(X)\ \frac{\partial p(X)}{\partial X_1}\, d\nu_1(X_1) + G(b,X_o)p(b,X_o) - G(a,X_o)p(a,X_o)$$

The latter two terms represent $G(X)p(X)$ evaluated at boundary points, which vanish by Assumption 3, so that (3.3) is established for $\omega(X_o)=[a,b]$.

For the unbounded case $\omega(X_o)=[a,\infty)$, note first that the existence of $E(\tilde{y})$, $E(\partial G/\partial X_1)$ and $E(\ell_1(X)\tilde{y})$ respectively imply the existence of $E(G(X)|X_o)$, $E(\partial G/\partial X_1|X_o)$ and $E(\ell_1(X)G(X)|X_o)$ (c.f. Kolmogorov(1950)). Now consider the limit of (3.5) over intervals $[a,b]$, where $b\to\infty$, rewritten as

$$(3.6)\qquad \lim_{b\to\infty} G(b,X_o)p(b,X_o) = G(a,X_o)p(a,X_o) + \lim_{b\to\infty}\int_a^b \frac{\partial G(X)}{\partial X_1}\, p(X)d\nu_1(X_1) + \lim_{b\to\infty}\int_a^b G(X)\ \frac{\partial p(X)}{\partial X_1}\, d\nu_1(X_1)$$

$$= G(a,X_o)p(a,X_o) + p_o(X_o)E\left[\frac{\partial G}{\partial X_1}\middle| X_o\right] - p_o(X_o)E(\ell_1(X)G(X)|X_o)$$

so that $C \equiv \lim G(b,X_o)p(b,X_o)$ exists, where $p_o(X_o)$ is the marginal density of X_o. Now suppose that $C>0$. Then there exists scalars ε and B such that $0<\varepsilon<C$ and for all $b \geq B$, $|G(b,X_o)p(b,X_o)-C|<\varepsilon$. Therefore $G(X_1,X_o)p(X_1,X_o) > (C-\varepsilon)I_{[B,\infty)}$, where $I_{[B,\infty)}$ is the indicator function of $[B,\infty)$. But this implies that $p_o(X_o)E(G(X)|X_o) = \int G(X_1,X_o)p(X_1,X_o)d\nu_1(X_1) > (C-\varepsilon)\int I_{[B,\infty)}d\nu_1(X_o) = \infty$, which contradicts the existence of $E(G(X)|X_o)$. Consequently, $C>0$ is ruled out. $C<0$ similarly contradicts the existence of $E(G(X)|X_o)$.

Since $C \equiv \lim G(b,X_o)p(b,X_o) = 0$, and $G(a,X_o)p(a,X_o)=0$ by Assumption 3, equation (3.3) is valid for $\omega(X_o)=[a,\infty)$. Analogous arguments establish the validity of (3.3) for $\omega(X_o)=(-\infty,a]$ and $\omega(X_o)=(-\infty,\infty)$.

The second equality of (3.1); $E(\ell_1(X)\tilde{y})=Cov(\ell_1(X),\tilde{y})$; is true because the mean of $\ell_1(X)$ is 0.[6] The proof is completed by repeating the same development for derivatives of G(X) with respect to $X_2, \ldots, X_M$. QED

Theorem 1 is of significant theoretical interest. It says that the average behavioral derivative $E(\partial G/\partial X)$ can be written as the covariance between $\tilde{y}$ and a function of X; namely $\ell(X)$. The form of $\ell(X)$ does not depend on the behavioral relation $E(\tilde{y}|X)=G(X)$; $\ell(X)$ is determined only by the marginal density p(X). Thus, Theorem 1 establishes a general link between behavioral derivatives and covariance estimators, that does not depend on assumptions on the form of behavior.[7] The proof is extremely simple, based on integration-by-parts.

A useful intuition for Theorem 1 can be obtained from its connection to results in aggregation theory. In particular, Theorem 1 reflects the local aggregate effects on $E(\tilde{y})$ of translating the base density p(X). To see this connection, consider the unbounded case where $\Omega=R^M$. Suppose that the base density is translated by an M-vector θ; p(X) is altered to $p(X-\theta)$ for all X. The value of $E(\tilde{y})$ after this translation is given as

$$(3.7a) \qquad E(\tilde{y}|\theta) = \int_{\Omega} G(X)p(X-\theta)d\nu$$

By a change-of-variables, $E(\tilde{y}|\theta)$ can also be written as

$$(3.7b) \qquad E(\tilde{y}|\theta) = \int_{\Omega} G(X+\theta)p(X)d\nu$$

The local aggregate effects of the translation are the derivatives $\partial E(\tilde{y}|\theta)/\partial\theta$ evaluated at $\theta=0$.[8] Differentiating (3.7a) under the integral sign

and evaluating at $\theta=0$ gives

$$(3.8a) \qquad \frac{\partial E(\tilde{y}|0)}{\partial \theta} = \int_{\Omega} G(X) \frac{\partial p}{\partial \theta} d\nu = \int_{\Omega} G(X) \frac{\partial \ln p}{\partial \theta} p(X) d\nu = \int_{\Omega} G(X)\, \ell(X)\, p(X) d\nu$$

where the latter equality reflects that $\ell(X)$ equals $\partial \ln p(X-\theta)/\partial\theta$ evaluated at $\theta=0$. Similarly, differentiating (3.7b) gives

$$(3.8b) \qquad \frac{\partial E(\tilde{y}|0)}{\partial \theta} = \int_{\Omega} \frac{\partial G}{\partial \theta} p(X) d\nu = \int_{\Omega} \frac{\partial G}{\partial X} p(X) d\nu$$

Collecting the equalities of (3.8a,b) gives $E(G(X)\ell(X))=E(\partial G/\partial X)$, which underlies equation (3.1) of Theorem 1.

Theorem 1 thus has a simple geometric explanation. For evaluating the mean $E(\tilde{y})$ under translation, one can average $G(X)$ over the distribution $p(X)$ shifted by θ (equation (3.7a)), or one can shift $G(X)$ by $-\theta$ and average over the distribution $p(X)$ (equation (3.7b)). The local effects on $E(\tilde{y})$ can be computed from either perspective (equations (3.8a,b)) to yield the same value. Equation (3.1) just exhibits this equivalence.

4. Consistent Estimation of Scaled Coefficients

This section indicates how to estimate β up to scale for single index models of the form (2.1). Section 4.1 indicates the basic approach and proposes a covariance estimator and an instrumental variables estimator. Section 4.2 discusses immediate extensions of the basic results and Section 4.3 gives some further remarks.

4.1 The Average Derivative Approach to Estimation

Begin by considering a precise empirical implication of the single index model form $E(y|X)=F(\alpha+X'\beta)$. Clearly, the conditional mean of y depends only on X through the value of $X'\beta$. By exploiting differentiability, a precise restriction of the single index form is given as

$$(4.1) \qquad \frac{\partial E(y|X)}{\partial X} = \frac{\partial F(\alpha+X'\beta)}{\partial X} = \left[\frac{dF}{dZ}\right] \beta$$

Thus, $\partial E(y|X)/\partial X$ is proportional to β, although the scale factor dF/dZ will depend on the value of X chosen.

The basic approach in this paper is to focus on the average of the constraint (4.1):

$$(4.2) \qquad E\left[\frac{\partial E(y|X)}{\partial X}\right] = E\left[\frac{\partial F}{\partial X}\right] = E\left[\frac{dF}{dZ}\right] \beta = \gamma\beta$$

where $\gamma=E(dF/dZ)$ exists and is nonzero by Assumption 4. Clearly, any consistent estimate of the average derivative $E(\partial F/\partial X)$ is a consistent estimator of β up to scale.

Two natural consistent estimators are suggested by Theorem 1. First define the estimator $\hat{d}_0$ as the sample covariance between y_i and ℓ_i;

$$(4.3) \qquad \hat{d}_0 = S_{\ell y}$$

The second estimator is more closely related to standard regression estimators, such as the OLS coefficients of y regressed on X. Define $\hat{d}$ as the instrumental variables coefficients of the regression

$$(4.4) \qquad y_i = \hat{c} + X_i' \hat{d} + \hat{u}_i$$

obtained using $(1,\ell_i')$ as the instrumental variable; namely

$$(4.5) \qquad \hat{d} = (S_{\ell X})^{-1} S_{\ell y}$$

The consistency of $\hat{d}_0$ and $\hat{d}$ for $\gamma\beta$ follows immediately from Theorem 1, as in

Theorem 2: Given Assumptions 1-4, $\hat{d}_0$ and $\hat{d}$ are strongly consistent estimators of $\gamma\beta$, where $\gamma=E(dF/dZ)$.

Proof: The Strong Law of Large Numbers (c.f. Rao(1973), Section 2c.3, SLLN2) implies that $\lim S_{\ell y} = \Sigma_{\ell y}$. Theorem 1 and (4.2) imply that $\lim \hat{d}_0 = \gamma\beta$ a.s.. $\lim \hat{d} = \gamma\beta$ a.s. follows if $\lim S_{\ell X} = \Sigma_{\ell X} = I$, an M×M identity matrix. In view of Assumption 4b), Theorem 1 can be applied with $\tilde{y}=X_j$, for each j=1,...,M. Carrying this out gives $\Sigma_{\ell X} = I$. QED

The two estimators $\hat{d}_0$ and $\hat{d}$ appear very similar, however in general they are not first-order ($\sqrt{N}$) equivalent. In particular,

$$\sqrt{N}(\hat{d} - \hat{d}_0) = \sqrt{N}(S_{\ell X}^{-1} - I)\,\hat{d}_0 \tag{4.6}$$

Since $\lim \hat{d}_0 = \gamma\beta \neq 0$, and $\sqrt{N}(S_{\ell X}^{-1} - I)$ in general has a nontrivial limiting distribution, $\sqrt{N}(\hat{d}-\hat{d}_0)$ will not vanish as $N\to\infty$. For expository purposes, I will refer to $\hat{d}$ for the remainder of the exposition, however, all consistency results can be extended to $\hat{d}_0$.[9]

The connection to the aggregate effects of translation permits a further interpretation of the scale factor $\gamma = E(dF/dZ)$. The structure of the single index model (2.1) implies that the local aggregate effects of translation are proportional to β, the parameters of interest. In particular, insert (4.1) into (3.8b), giving

$$\frac{\partial E(y|\theta)}{\partial \theta} = \int_{\Omega} \frac{dF}{dZ}\,\beta\, p(X)d\nu = \gamma\beta \tag{4.7}$$

This appearance of β is due to the correspondence between density translation and the linear form of the index $Z=\alpha+X'\beta$. To interpret γ, note that under translation, the marginal distribution of Z is shifted by the parameter $\eta=\theta'\beta$, with the mean of Z increased by η. (4.7) can be regarded as the chain rule formula $\partial E(y)/\partial\theta = (dE(y)/d\eta)(\partial\eta/\partial\theta)$, where $\partial\eta/\partial\theta$ is equal to β. The scale factor γ is equal to $dE(y)/d\eta$, the effect on E(y) induced by a change in the mean E(Z) of the index variable Z.

4.2 Extraneous Variables and Multiple Index Models

The approach of parameter estimation via average derivatives easily extends to more general models than those relying on a single index. In this section I consider some immediate extensions, namely to models with extraneous variables and multiple index models.

Begin by expanding the notation to consider two sets of explanatory variables; an M_1 vector X_1 and an M_2 vector X_2, distributed with density $p(X_1,X_2)$.[10] Consider first the case where X_2 are extraneous variables, in that the behavioral model for y implies

$$E(y|X_1,X_2) = F(\alpha_1+X_1'\beta_1,X_2) = F(Z_1,X_2) \tag{4.8}$$

for some function F with constant coefficients α_1, β_1, and $Z_1=\alpha_1+X_1'\beta_1$. In this case, β_1 is proportional to the (partial) derivative of F with respect to X_1, as in

$$\frac{\partial E(y|X_1,X_2)}{\partial X_1} = \frac{\partial F}{\partial X_1} = \left[\frac{\partial F}{\partial Z_1}\right]\beta_1 \tag{4.9}$$

so that the average derivative is proportional to β_1:

$$E\left[\frac{\partial F}{\partial X_1}\right] = E\left[\frac{\partial F}{\partial Z_1}\right]\beta_1 = \gamma_1\beta_1 \tag{4.10}$$

Theorems 1-3 can be applied as long as the appropriate analogues of Assumptions 1-4 applied to X_1 are valid. In particular, the proof of Theorem 1 will apply to individual components of X_1 provided that no two components of X_1,X_2 are perfectly correlated, and that the conditional density $p_1^c(X_1|X_2)$ vanishes on the boundary of X_1 values for each value of X_2. Under these conditions, the sample covariance $\hat{d}_{10}=S_{\ell_1 y}$ consistently estimates $\gamma_1\beta_1$, where the partial score $\ell_{1i}=\ell_1(X_{1i},X_{2i})$ is defined via

$$(4.11) \qquad \ell_1(X_1,X_2) = -\frac{\partial \ln p(X_1,X_2)}{\partial X_1} = -\frac{\partial \ln p_1^c(X_1|X_2)}{\partial X_1}$$

Moreover, $\gamma_1\beta_1$ is consistently estimated by the slope coefficients estimates $\hat{d}_1$ of the linear equation

$$(4.12) \qquad y_i = \hat{c}_1 + X_{1i}\hat{d}_1 + \hat{u}_{1i}$$

obtained by instrumenting with $(1,\ell_{1i}')'$. Thus, the extraneous variables X_2 are accomodated in the estimation of β_1 by modification of the appropriate instrumental variables, to reflect the joint distribution of X_1 and X_2. Clearly if X_2 were distributed independently of X_1, then X_2 can be ignored in the estimation of β_1 up to scale.[11]

This extension provides an initial response as to how to accomodate discrete explanatory variables into the analysis. If X_2 is composed of discrete variables, an approach based on average derivatives is not obviously applicable to estimating effects of X_2. However, the coefficients of the remaining continuous variables X_1 can be estimated up to scale by using the score vectors of the conditional density of X_1 given the observed values of X_2 as instrumental variables. Consequently, while the analysis is silent on how to estimate coefficients of discrete variables, their presence does not prohibit the estimation of continuous variable coefficients up to scale.

Putting aside this proviso on discrete variables, I now turn to multiple index models. All relevant points are exhibited by two index models, so assume that $X=(X_1',X_2')'$ is composed entirely of continuous variables with $M_1 \geq 2$ and $M_2 \geq 2$. Suppose that the behavioral model implies the following two index form

$$(4.13) \qquad E(y|X) = F(\alpha_1+X_1'\beta_1,\ \alpha_2+X_2'\beta_2) = F(Z_1,Z_2)$$

where $Z_1=\alpha_1+X_1'\beta_1$ and $Z_2=\alpha_2+X_2'\beta_2$ represent the two index variables. The

derivative of the conditional expectation now takes the form

$$(4.14) \qquad \frac{\partial E(y|X)}{\partial X} = \frac{\partial F}{\partial X} = \begin{bmatrix} \left[\frac{\partial F}{\partial Z_1}\right]\beta_1 \\ \left[\frac{\partial F}{\partial Z_2}\right]\beta_2 \end{bmatrix}$$

so that the average derivative is

$$(4.15) \qquad E\left[\frac{\partial F}{\partial X}\right] = \begin{bmatrix} \gamma_1\beta_1 \\ \gamma_2\beta_2 \end{bmatrix}$$

where $\gamma_1=E(\partial F/\partial Z_1)$ and $\gamma_2=E(\partial F/\partial Z_2)$ are scalar constants. Thus a consistent estimate of the average derivative will estimate β_1 and β_2 up to scale, however the scale factors γ_1 and γ_2 will differ in general.

Such a consistent estimate has already been established, provided that y,F of (4.13) obeys condition A. Namely, the estimator $\hat{d}$ of (4.5) consistently estimates $E(\partial F/\partial X)$, so that its components corresponding to X_1 estimate β_1 up to scale, and its components corresponding to X_2 consistently estimate β_2 up to scale.[12] The main modeling limitation of this result is that no two components of X_1 and X_2 may be functionally related or perfectly correlated. Thus the index variables Z_1 and Z_2 may have no common component variables, an exclusion restriction that is required for estimating both β_1 and β_2 up to scale using average first derivatives.[13] The following example gives a two index model, where $\gamma_1=1$ a priori.

Example 4: Selection Bias

Suppose that the basic behavioral model is $y=\alpha_1+X_1'\beta_1+\varepsilon_1$, but that y, X_1 and X_2 are observed only if $Z_2^*=\alpha_2+X_2'\beta_2+\varepsilon_2 > 0$, where $(\varepsilon_1,\varepsilon_2)$ is distributed independently of (X_1,X_2). This implies that

$$E(y|X_1,X_2,Z_2^* > 0) = \alpha_1 + X_1'\beta_1 + E(\varepsilon_1|\varepsilon_2 > -(\alpha_2 + X_2'\beta_2))$$

$$= F(\alpha_1+X_1'\beta_1,\alpha_2+X_2'\beta_2) = F(Z_1,Z_2)$$

so that $\hat{d}$ will estimate the structural parameters β_1 and the selection parameters β_2 up to scale, without explicit assumptions on the joint distribution of $(\varepsilon_1,\varepsilon_2)$. Note that $\partial F/\partial Z_1=1$, so that $\gamma_1=1$. Thus, the components of $\hat{d}$ corresponding to X_1 will consistently estimate β_1, with no proviso about scale.

By comparing Example 4 and the truncated tobit specification of Example 2, there are two polar cases where selection parameters are estimated up to scale by $\hat{d}$, namely when the selection index Z_2 has no variables in common with the structural index Z_1, or when the selection index Z_2 is equal to the structural index Z_1.

I close with another example, that further illustrates how exclusion and other parameter restrictions bear on the estimation of specific coefficients up to scale.

Example 5: Discrete Choice Among Several Alternatives

Suppose that one is studying the choice between $j=1,\ldots,J$ discrete alternatives. The attractiveness (utility) of the j^{th} alternative is modeled as

$$V_j = \alpha_j + X_{1j}'\beta_{1j} + X_2\delta_j + \varepsilon_j \qquad j=1,\ldots,J$$

where X_{1j} is a set of option specific explanatory variables, with $X_1=(X_{11},\ldots,X_{1J})$ containing no two components that are perfectly correlated. X_2 represents explanatory variables that are observation specific, but bear on the attractiveness of each option. ε_j is a random term, such that $(\varepsilon_1,\ldots,\varepsilon_J)$

is distributed independently of (X_1,X_2).[14] The parameters α_j, β_{1j}, δ_j may vary with option j.

Focus on the J^{th} alternative, and assume that y=1 if J is chosen and y=0 if another alternative is chosen. Define J-1 index variables

$$Z_j = \alpha_J - \alpha_j + X_{1J}'\beta_{1J} - X_{1j}'\beta_{1j} + X_2'\beta_{2j} \qquad j=1,\ldots,J-1$$

where $\beta_{2j}=\delta_J-\delta_j$. Now J is chosen, or y=1, when $V_j \leq V_J$ for j=1,...,J-1. This occurs when

$$\varepsilon_j - \varepsilon_J \leq Z_j \qquad \text{for all } j=1,\ldots,J-1$$

$E(y|X_1,X_2)$ is just the probability of the above event given X_1,X_2; or

$$E(y|X_1,X_2) = F(Z_1,\ldots,Z_{J-1})$$

where F is the cummulative distribution function of $(\varepsilon_1-\varepsilon_J,\ldots,\varepsilon_{J-1}-\varepsilon_J)$. For instance, if $(\varepsilon_1-\varepsilon_J,\ldots,\varepsilon_{J-1}-\varepsilon_J)$ is multivariate normally distributed, F is the multivariate normal distribution function (and this is a multinomial probit model).

Now, what is estimated by $\hat{d}=(\hat{d}_{11}',\ldots,\hat{d}_{1J}',\hat{d}_2')'$, partitioned according to $(X_1,X_2)=(X_{11},\ldots,X_{1J},X_2)$? The coefficients $\hat{d}_{1J}$ of the J^{th} specific attributes X_{1J} will consistently estimate

$$E\left[\frac{\partial F}{\partial X_{1J}}\right] = \left[E\left[\sum_j \frac{\partial F}{\partial Z_j}\right]\right] \beta_{1J} = \gamma_J\beta_J$$

so that $\hat{d}_{1J}$ consistently estimates β_J up to scale. For $j\neq J$, the coefficients $\hat{d}_{1j}$ of X_{1j} will estimate

$$E\left[\frac{\partial F}{\partial X_{1j}}\right] = \left[-E\left[\frac{\partial F}{\partial Z_j}\right]\right] \beta_{1j} = \gamma_j\beta_{1j}$$

so that $\hat{d}_{1j}$ consistently estimates β_j up to scale. Finally, the coefficients $\hat{d}_2$ of X_2 will consistently estimate

$$E\left[\frac{\partial F}{\partial X_2}\right] = \sum_j E\left[\frac{\partial F}{\partial Z_j}\right] \beta_{2j}$$

so that $\hat{d}_2$ estimates a linear combination of the parameters β_{2j}, $j=1,\ldots,J-1$.

Consequently, the respective components of $\hat{d}$ will consistently estimate the option specific parameter vectors β_{1j}, $j=1,\ldots,J$, up to scale values. This occurs because X_{1J} appears in each index with the same coefficients β_{1J}, and for $j\neq J$, X_{1j} only appears in the single index Z_j. β_{2j}, $j=1,\ldots,J-1$, are not separately estimated because X_2 appears with a different coefficient value in each index Z_j.

4.3 Further Remarks

4.3a A Note on Heteroscedastic Disturbances

As indicated in Section 2, the development applies to models where $y=f(\alpha+X'\beta+\varepsilon)$, where ε is distributed independently of X. Often it is desirable to estimate β in situations where ε is heteroscedastic, with the distribution of ε depending on X. Estimators that are robust to heteroscedasticity of ε for specific models are given in Manski's(1975,1985) work on maximum score estimation and Powell's(1984) work on censored least absolute deviations estimation.

It is easy to see in general that $\hat{d}$ will not estimate β up to scale when the distribution of ε depends on X. In this case, the conditional expectation $E(y|X)$ will not be a function of $\alpha+X'\beta$ alone, depending in general on how X alters the distribution of ε over observations. Equations (2.1) and (4.1) will not hold, which breaks the relation between β and the average derivative.[15]

The one special case where heteroscedasticity of ε does not alter the consistency of $\hat{d}$ for β up to scale is when the distribution of ε depends only on the value of the index $Z=\alpha+X'\beta$. Here (2.1) and (4.1) are valid, with the development applying without modification. While this is a strong restriction,

some models obey this restriction, such as truncated Poisson regression models. A good survey of this and related models is given in Manski(1984a).

4.3b: The Statistical Role of the X Distribution

An interesting feature of the estimators suggested by Theorem 1 is their explicit dependence on the density p(X) of the marginal distribution of X. In particular, the consistency of the estimators relies on the fact that the data X_i, i=1,...,N represents a random sample, so that the X's are not taken as ancilliary for estimation.

This raises a rather deep statistical issue concerning the efficiency of the estimators, which is described as follows. The overall object of estimation is to measure the value of β up to scale. β is clearly a parametric feature of the conditional distribution of y given X, and so there is no generic necessity for knowing the marginal distribution of X. The usefulness of the information provided by the density p(X) is more surprising than natural, when viewed in this light.

The role of the density p(X) is built into the particular estimation strategy employed here, namely the estimation of the average derivative $E(\partial F/\partial X)$ of (4.2). The value of the $E(\partial F/\partial X)$ clearly depends on the true marginal density p(X) - altering the marginal density will alter the average derivative. In other words, even if the exact form of the conditional density of y given X were known for all X values, the average $E(\partial F/\partial X)$ could not be consistently estimated without reference to the configuration of the X values in a large sample.

But estimation of the average derivative does not represent the only conceivable method of estimating β up to scale. This can be seen from equation (4.1), which is a derivative constraint on the conditional expectation of y given X, and does not involve the density of X. This suggests that more

efficient estimators of β up to scale could be found, which take X as ancilliary (i.e. which condition on the observed data values X_i, i=1,...,N). No such general, more efficient estimators are known to the author, however the possibility of their existence warrants investigation via future research.[16]

5. Biases in Ordinary Least Squares Coefficients

The linear structure of $\hat{d}$ suggests a natural comparison with the OLS slope coefficient vector $\hat{b}=(S_{XX})^{-1}S_{Xy}$ from the regression of y_i on X_i, i=1,...,N.[17] This section uses the above development to study the asymptotic bias in $\hat{b}$ as an estimator of the true average derivative $E(\partial F/\partial X)$, and as an estimator of β up to scale. The primary focus is on the role of the distribution of X, as the formulae below are applicable regardless of the true form of the function F of (2.1).

Begin by considering the circumstances under which $\ell(X)$ is collinear with X. If so, then $\hat{b}=\hat{d}$, which gives a robust interpretation of $\hat{b}$ as an estimator of β up to scale, or alternately a case where $\hat{d}$ is particularly easy to compute. Now suppose that $\ell(X)=A+BX$, where A is an M vector and B an M×M matrix of constants, and denote $\mu_X=E(X)$. Since $E(\ell(X))=0$, $A=-BE(X)=-B\mu_X$, so that $\ell(X)=B(X-\mu_X)$. By integrating $\ell(X)$, ln p(X) can be written in the form $\ln p(X)=C-(1/2)(X-\mu_X)'B(X-\mu_X)$ for some constant C. Thus p(X) must be of the multivariate normal form, with $B=(\Sigma_{XX})^{-1}$. Consequently $\hat{b}$ and $\hat{d}$ coincide only when X is multivariate normally distributed.

Theorem 1 appears in simple form in this case. $\ell(X)=(\Sigma_{XX})^{-1}(X-\mu_X)$, so that $E(\ell(X)y)=Cov(\ell(X),y)=(\Sigma_{XX})^{-1}\Sigma_{Xy}$. This is clearly the a.s. limit of the OLS coefficient vector $\hat{b}$, namely $b \equiv \lim \hat{b}=(\Sigma_{XX})^{-1}\Sigma_{Xy}$.[18]

To study the asymptotic bias of $\hat{b}$ when X is not normally distributed, first consider the difference between the average derivative $E(\partial F/\partial X)$ and b.

Since $E(\partial F/\partial X)=E(\ell(X)y)$ by Theorem 1,

$$(5.1)\qquad E\left[\frac{\partial F}{\partial X}\right] - b = E([(\ell(X) - (\Sigma_{XX})^{-1}(X-\mu_X)]y) = E(R(X)y) = Cov(R(X),y)$$

where $R(X) \equiv \ell(X)-(\Sigma_{XX})^{-1}(X-\mu_X)$, and the last equality follows from $E(R(X))=0$.

Notice that $R(X)$ can be regarded as a large sample OLS residual vector. The OLS coefficients $\hat{B}$ of the multivariate regression equation $\ell(X_i)=\hat{B}(X_i-\mu_X)+\hat{R}(X_i)$ are such that $\lim \hat{B}=(\Sigma_{XX})^{-1}\Sigma_{X\ell}=(\Sigma_{XX})^{-1}$ a.s., since $\Sigma_{X\ell}=I$ by the proof of Theorem 2. Thus $\lim \hat{R}(X_i) = R(X_i)$ a.s., with $R(X)$ interpreted as the large sample least squares departure of $\ell(X)$ from X. In particular, $R(X)=0$ for all X only if X is normally distributed.

Equation (5.1) says that $\hat{b}$ consistently estimates the average derivative $E(\partial F/\partial X)$ only when y is uncorrelated with the least squares difference $R(X)$ between $\ell(X)$ and X. Thus, unless X is normally distributed, $\hat{b}$ will consistently estimate the average derivative $E(\partial F/\partial X)$ only in certain model-specific cases. To consider the role of the true model in this property, refine equation (5.1) as follows. Since $R(X)$ is a least squares residual, $R(X)$ is uncorrelated with X. Therefore y can be replaced in (5.1) by the large sample residual $\xi=(y-\mu_y)-(X-\mu_X)'b$ from the OLS regression of y on X, as in

$$(5.2)\qquad E\left[\frac{\partial F}{\partial X}\right] - b = Cov(R(X),y) = Cov(R(X),\xi)$$

There are two natural polar cases under which $\hat{b}$ will estimate the average derivative $E(\partial F/\partial X)$; first when $R(X)=0$, or when X is normally distributed, and second when $E(\xi|X)=0$, or when the true model between y and X is a linear regression model.[19] In nonlinear cases, $\hat{b}$ will estimate $E(\partial F/\partial X)$ only if the specific functional form assures that the OLS residual ξ is uncorrelated with $R(X)$. Finally, note that (5.1,2) do not utilize the single index form (2.1), so that y,F could be replaced in the above discussion by any $\tilde{y},G(X)$ obeying condition A.

At face value, this suggests that the conditions under which $\hat{b}$ estimates β up to scale may also be restricted to X normally distributed. However, Ruud(1983,1984), Chung and Goldberger(1984) and Deaton and Irish(1984) have pointed out another sufficient condition on the distribution of X that does not restrict X to be normally distributed. In particular, Ruud(1983a) shows that $\hat{b}$ will consistently estimate β up to scale when $E(X|Z)=G+HZ$, for $Z=\alpha+X'\beta$, or that $E(X|Z)$ is a linear function of Z. Chung and Goldberger(1984) and Deaton and Irish(1984) find the same result using an analogous condition with a generalized definition of Z. Chamberlain(1983) has pointed out that this condition is obeyed when the distribution of X is (elliptically) symmetric, but not necessarily normal (see also Dempster (1969)). Consequently, equations (5.1,2) do not suffice to characterize the asymptotic bias of $\hat{b}$ as an estimator of β up to scale.

A bit more development yields a bias formula that explicitly displays the role of the linearity condition. Assume first that the relationship between y and X can be represented by $y=f(\alpha+X'\beta+\varepsilon)=f(Z+\varepsilon)$ for some (unknown) function f, where ε is distributed independently of X. Denote the marginal distribution of $Z=\alpha+X'\beta$ as $p_Z(Z)$, the mean of Z as $\mu_Z=E(Z)$ and the associated log-density derivative as $\ell_Z(Z)=-\,d \ln p_Z(Z)/dZ$. Define the large sample residual of $\ell_Z(Z)$ regressed on Z as[20]

$$(5.3) \qquad r_n(Z) = \ell_Z(Z) - \sigma_Z^{-2}(Z-\mu_Z)$$

Clearly $r_n(Z)=0$ for all Z if and only if Z is normally distributed. Finally, define the large sample residual vector of $E(X|Z)$ regressed on Z as

$$(5.4) \qquad r_l(Z) = (E(X|Z)-\mu_X) - H(Z-\mu_Z)$$

where $H=\Sigma_{XZ}/\sigma_Z^2$. Clearly $r_l(Z)=0$ for all Z if and only if the linearity

condition holds, or that $E(X|Z)=G+HZ$. The formula to be derived relates the asymptotic bias in $\hat{b}$ to covariances between y and the residuals $r_n(Z)$ and $r_1(Z)$.

Recall that for the single index model (2.1), we have that $E(\partial F/\partial X)=\gamma\beta$, where $\gamma=E(dF/dZ)=\int(dF/dZ)p_Z(Z)d\nu$. By applying Theorem 1, γ can be written as $\gamma=E(\ell_Z(Z)F(Z))=Cov(\ell_Z(Z),y)$. Thus, the average derivative $E(\partial F/\partial X)$ can be written as

$$(5.5) \qquad E\left[\frac{\partial F}{\partial X}\right] = \beta\ Cov(\ell_Z(Z),\ y)$$

To characterize the limit $b=(\Sigma_{XX})^{-1}\Sigma_{Xy}$ of $\hat{b}$, note first that $\Sigma_{Xy}=E[(E(X|Z)-\mu_X)y]$. This is valid because at a given value of Z, the conditional covariance between $X-E(X|Z)$ and $y=f(Z+\varepsilon)$ is zero. Now

$$(5.6) \qquad b = (\Sigma_{XX})^{-1}E[(E(X|Z)-\mu_X)y)$$

$$= E[(\Sigma_{XX})^{-1}H(Z-\mu_Z)y] + (\Sigma_{XX})^{-1}E(r_1(Z)y)$$

by using (5.4). Note that by construction, $\beta=(\Sigma_{XX})^{-1}\Sigma_{XZ}$, so that $(\Sigma_{XX})^{-1}H = (\Sigma_{XX})^{-1}\Sigma_{XZ}/\sigma_Z^{\ 2} = \beta/\sigma_Z^{\ 2}$. Inserting this gives

$$(5.7) \qquad b = \beta\ E(\sigma_Z^{-2}(Z-\mu_Z)y) + (\Sigma_{XX})^{-1}E(r_1(Z)y)$$

$$= \beta\ Cov(\sigma_Z^{-2}(Z-\mu_Z),y) + (\Sigma_{XX})^{-1}Cov(r_1(Z),y)$$

The desired bias formulae is obtained by combining (5.5), (5.7) and (5.3) to yield

$$(5.8) \qquad E\left[\frac{\partial F}{\partial X}\right] - b = \beta\ Cov(r_n(Z),y) + (\Sigma_{XX})^{-1}Cov(r_1(Z),y)$$

Equation (5.8) provides a categorization of the asymptotic bias in the OLS estimator $\hat{b}$ vis-a-vis the X distribution underlying the data. When X is multivariate normally distributed, Z is also normal, $r_n(Z)=0$ and $r_1(Z)=0$ for all Z, and $\hat{b}$ consistently estimates the average derivative $E(\partial F/\partial X)=\gamma\beta$. When X

is not normally distributed but the linearity condition holds, $r_1(Z)=0$, and $\hat{b}$ consistently estimates $(\gamma+Cov(r_n(Z),y))\beta$. The covariance term will not be zero in general, but the asymptotic bias $E(\partial F/\partial X)-b$ is proportional to β, to that $\hat{b}$ still estimates β up to scale. Finally, $\hat{b}$ will not consistently estimate β up to scale in general if $r_1(Z)\neq 0$.

Thus, for a model-free interpretion of $\hat{b}$ as the average derivative $E(\partial F/\partial X)$, multivariate normality of the X distribution is essential. For the question of when $\hat{b}$ estimates β up to scale, the linearity condition $E(X|Z)=G+HZ$ provides a solution that is not directly related to estimation of the average derivative.[21]

6. Distribution Theory with a Parametric Density Form

The above exposition has proposed the estimator $\hat{d}$ as an estimator of β up to scale, that explicitly utilizes information on the marginal density $p(X)$. When $p(X)$ is in the multivariate normal form, $\hat{d}$ can be computed as the OLS slope coefficients, and scale-free inferences on the value of β (as discussed below) can be performed with standard methods.[22] In general, a statistical characterization of the density $p(X)$ will be required to impliment $\hat{d}$. This section establishes the asymptotic distribution of $\hat{d}$ when the density is modeled via a finite parameterization.

Suppose that $p(X)=p^*(X|\Lambda_0)$, where $p^*(X|\Lambda)$ denotes a parametric family, with Λ an L-vector of parameters that characterize the location and shape of $p(X)$; means, variances, skewness, etc. The density score vector is determined by Λ as $\ell(X)=\ell(X|\Lambda_0)$. Assumptions 5 and 6 of the appendix assume that a $\sqrt{N}$ consistent estimator $\hat{\Lambda}$ of $\Lambda=\Lambda_0$ can be computed using the data X_i, $i=1,\ldots,N$, as well as some regularity conditions.

Estimation now proceeds in two steps. First estimate Λ_0 using $\hat{\Lambda}$. Next

compute $\hat{d}^*$ as the instrumental variables estimator using the estimated instrument

$$\hat{\ell}_i = \ell(X_i|\hat{\Lambda}) = -\frac{\partial \ln p^*(X_i|\hat{\Lambda})}{\partial X} \tag{5.1}$$

as in

$$\hat{d}^* = (S_{\hat{\ell}X})^{-1} S_{\hat{\ell}y} \tag{5.2}$$

The strong consistency and asymptotic normality of $\hat{d}^*$ is established in

Theorem 3: Given Assumptions 1-6, (a) $\hat{d}^*$ is a strongly consistent estimator of $\gamma\beta$, and (b) the limiting distribution of $\sqrt{N}(\hat{d}^* - \gamma\beta)$ is multivariate normal with mean 0 and variance-covariance matrix

$$V = \Sigma_{\ell u,\ell u} + D\Sigma'_{\ell u,\lambda} + \Sigma_{\ell u,\lambda}D' + D\Sigma_{\lambda\lambda}D' \tag{5.3}$$

where $u=(y-\mu_y)-(X-\mu_X)'\gamma\beta$, $\ell u=\ell(X)u$, $D=E[u(\partial\ell(X|\Lambda_0)/\partial\Lambda)]$ and λ is the component of $\hat{\Lambda}$ defined in the Appendix.

Proof: (a): Define $\ell_i(\Lambda)=\ell(X_i|\Lambda)$, so that $\ell_i(\Lambda_0)=\ell_i$ and $\ell_i(\hat{\Lambda})=\hat{\ell}_i$. To show consistency of $S_{\hat{\ell}y}$ for $\Sigma_{\ell y}$, define $S_y(\Lambda)=\Sigma\ell_i(\Lambda)(y_i-\bar{y})/N$, so that $S_y(\Lambda_0)=S_{\ell y}$ and $S_y(\hat{\Lambda})=S_{\hat{\ell}y}$. From (A.2a) of Assumption 6, Theorem 2 of Jennrich(1969) implies that $S_y(\Lambda)$ converges uniformly in Λ to $E[\ell(X|\Lambda)(y-\mu_y)]$. Since $\lim \hat{\Lambda}=\Lambda_0$ a.s., by Lemma 4 of Amemiya(1973), $\lim S_{\hat{\ell}y}=\lim S_y(\hat{\Lambda})=E[\ell(X|\Lambda_0)(y-\mu_y)]=\Sigma_{\ell y}$, a.s. A similar argument (using (A2.b) of Assumption 6) shows that $\lim S_{\hat{\ell}X}=\Sigma_{\ell X}$ a.s., so that $\lim \hat{d}^*=(\Sigma_{\ell X})^{-1}\Sigma_{\ell y}=\gamma\beta$ a.s.

(b): Following Newey(1984), define $u_i=(y_i-\bar{y})-(X_i-\bar{X})'\gamma\beta$, and write

$$(5.4)\quad \sqrt{N}(\hat{d}^* - \gamma\beta) = (S_{\hat{\ell}X})^{-1}\left[\frac{\Sigma\,\hat{\ell}_i u_i}{\sqrt{N}}\right]$$

$$= (S_{\hat{\ell}X})^{-1}\left[\frac{\Sigma\,\ell_i u_i}{\sqrt{N}}\right] + (S_{\hat{\ell}X})^{-1}\left[\frac{\Sigma(\hat{\ell}_i-\ell_i)u_i}{\sqrt{N}}\right]$$

A Taylor series expansion of the second term gives

$$(5.5)\quad \sqrt{N}(\hat{d}^*-\gamma\beta) = (S_{\hat{\ell}X})^{-1}\left[\frac{\Sigma\,\ell_i u_i}{\sqrt{N}}\right] + (S_{\hat{\ell}X})^{-1}\left[\frac{\Sigma\,u_i[\partial\ell(X_i|\Lambda_0)/\partial\Lambda]}{N}\right]\sqrt{N}(\hat{\Lambda}-\Lambda_0)+o_p(1)$$

The result follows from $\lim S_{\hat{\ell}X}=\Sigma_{\ell X}=I$, an identity matrix, and $\text{plim}[\Sigma u_i(\partial\ell(X_i|\Lambda_0)/\partial\Lambda)/N]=D$ (Weak Law of Large Numbers). QED

Under the additional regularity conditions (A.2c-g) of Assumption 6 of the appendix, a consistent estimator of the variance-covariance matrix V can be constructed as follows. Define $\hat{u}_i=(y_i-\bar{y})-(X_i-\bar{X})'\hat{d}^*$ as the estimated residual from equation (4.4) using $\hat{d}^*$ as coefficient estimates, define $\hat{\lambda}_i=\lambda(X_i|\hat{\Lambda})$ as the estimated component of $\hat{\Lambda}$, and define the estimator $\hat{D}=\Sigma\hat{u}_i[\partial\ell(X_i|\hat{\Lambda})/\partial\Lambda]/N$ of D. It is easy to verify that $\hat{V}=\Sigma(\hat{\ell}_i\hat{u}_i+\hat{D}\hat{\lambda}_i)(\hat{\ell}_i\hat{u}_i+\hat{D}\hat{\lambda}_i)'/N$ is a consistent estimator of V.

Thus when the density p(X) is modeled up to a finite parameterization, inferences on the value of $\gamma\beta$ can be performed using $\hat{d}^*$ and the consistent estimate of its variance-covariance matrix $\hat{V}$. Of more interest are tests on hypotheses on the value of β available from $\hat{d}^*$; namely hypotheses that are not affected by the true value of γ. The main class of such scale-free hypotheses are homogeneous linear restrictions of the form $\kappa'\beta=0$, where κ is an M-vector of constants. This class includes zero restrictions such as $\beta_j=0$, equality restrictions such as $\beta_\ell=\beta_j$, and ratio restrictions such as $\beta_\ell=\kappa_j\beta_j$ for a constant κ_j. Tests are possible by noting that $\kappa'\beta=0$ is equivalent to $\kappa'(\gamma\beta)=0$, and that $\kappa'\hat{d}^*$ is asymptotically normal with mean $\kappa'(\gamma\beta)$ and variance

$\kappa'V\kappa$. In particular, under the null hypothesis that $\kappa'\beta=0$, the Wald statistic $(\kappa'\hat{d}^*)^2/\kappa'\hat{V}\kappa$ has a limiting $\chi^2(1)$ distribution.

Wald statistics corresponding to joint hypothesis tests of $M' \leq M$ linear homogeneous restrictions can likewise be formulated using $\hat{d}^*$ and $\hat{V}$. Moreover, if $\phi(\beta)$ is any homogeneous M'-vector function of β, tests of $\phi(\beta)=\phi(\gamma\beta)=0$ can be formulated using the "delta method" of Billingsley(1979) and Rao(1973).

7. Concluding Remarks

This paper proposes an approach to parameter estimation based on average behavioral derivatives, and applies the approach to the estimation of β up to scale in single index models. The proposed estimators explicitly utilize information on the marginal distribution of the explanatory variables in the model. The framework is illustrated using several examples of limited dependent variables models, and extended to multiple index models. The asymptotic biases in OLS coefficients are characterized vis-a-vis the distribution of explanatory variables.

There are two major advantages to the proposed estimator $\hat{d}$. First, $\hat{d}$ is nonparametric to the extent that it is robust to many specific functional form and stochastic distribution assumptions. If a particular application requires only estimates of the ratios of components of β, then $\hat{d}$ will suffice. In a general application where different sets of assumptions give rise to different estimates of β, $\hat{d}$ will provide a benchmark estimate for choosing the best specification. Given parametric modeling of the explanatory variable distribution, the precision of the components of $\hat{d}$ can be measured, and tests of scale-free hypotheses on the value of β can be performed.

The other advantage of $\hat{d}$ is computational simplicity. Once the distribution of explanatory variables is characterized, $\hat{d}$ (as well as $\hat{d}_0$) is a

linear estimator, computed entirely from sample covariances. This suggests that implimentation may be particularly easy and inexpensive, especially for large data bases.

There are also two drawbacks, which suggest natural future research topics. First, the results apply only to the estimation of coefficients of continuous variables, but most applications to microeconomic data will require using discrete as well as continuous explanatory variables. While the presence of discrete variables can be accomodated in the estimation of continuous variable coefficients, the question of how to nonparametrically estimate coefficients of discrete variables remains open.

The second drawback involves the empirical characterization of the explanatory variable distribution. While this distribution can in principle always be characterized, I have only established attractive statistical properties for $\hat{d}$ when the distribution is modeled up to a finite parameterization, with the required score vectors computed from the estimated distribution parameters. Of substantial practical importance is the question of whether nonparametric estimators of the score vectors can be utilized in the construction of $\hat{d}$, to give a consistent estimator of $\gamma\beta$ with a straightforward asymptotic distribution theory.[23] Thus, the results of this paper provide further reasons for giving high priority to the application of nonparametric techniques to econometric modeling.

Appendix: Additional Assumptions

The following assumptions are utilized in Section 6.

Assumption 5: $p^*(X|\Lambda)$ is twice differentiable in the components of Λ in an open neighborhood of $\Lambda=\Lambda_0$. The estimator $\hat{\Lambda}$ of $\Lambda=\Lambda_0$ is strongly consistent, and can be written in the form

$$\text{(A.1)} \qquad \hat{\Lambda} = \Lambda_0 + \frac{\Sigma\, \lambda(X_i|\Lambda_0)}{N} + o_p\left[\frac{1}{\sqrt{N}}\right]$$

where $E(\lambda(X|\Lambda_0))=E(\lambda)=0$, and the variance-covariance matrix $E(\lambda\lambda')=\Sigma_{\lambda\lambda}$ exists.

This assumption implies that as $N\to\infty$, $\sqrt{N}(\hat{\Lambda}-\Lambda_0)$ has a limiting normal distribution with mean 0 and variance $\Sigma_{\lambda\lambda}$. If $\hat{\Lambda}$ is a sample average; say $\hat{\Lambda}=\Sigma g(X_i)/N$, then $\lambda=g(X)-\Lambda_0$. If $\hat{\Lambda}$ is the maximum likelihood estimator, then under standard conditions $\lambda=(I_\Lambda)^{-1}\partial \ln p(X|\Lambda_0)/\partial\Lambda$, where $I_\Lambda = -E(\partial^2 \ln p(X|\Lambda_0)/\partial\Lambda^2)$ is the information matrix.

The following regularity condition is also used.

Assumption 6: The covariance matrix of ℓu, and the covariance between any two components of ℓu and λ exist. The mean of $u(\partial\ell_j(X|\Lambda_0)/\partial\Lambda_\ell)$ exists for all $j=1,\ldots,M$ and $\ell=1,\ldots,L$. There exists measurable functions $h^1_j(y,X)$, $h^2_{jj'}(X)$, $h^3_{\ell\ell'}(X)$, $h^4_{\ell j}(y,X)$, $h^5_{j\ell j'}(X)$, $h^6_{j\ell}(y,X)$ and $h^7_{jj'\ell}(X)$ for all $\ell,\ell'=1,\ldots,L$; $j,j'=1,\ldots,M$ such that

$$\text{(A.2a)} \qquad |y\ell_j(X|\Lambda)| \le h^1_j(y,X)$$

$$\text{(A.2b)} \qquad |X_j\ell_{j'}(X|\Lambda)| \le h^2_{jj'}(X)$$

$$\text{(A.2c)} \qquad |\lambda_\ell(X|\Lambda)\lambda_{\ell'}(X|\Lambda)| \le h^3_{\ell\ell'}(X)$$

(A.2d) $|y\lambda_{\ell}(X|\Lambda)\ell_{j}(X|\Lambda)| \leq h^{4}_{\ell j}(y,X)$

(A.2e) $|X_{j}\lambda_{\ell}(X|\Lambda)\ell_{j'}(X|\Lambda)| \leq h^{5}_{j\ell j'}(X)$

(A.2f) $|y(\partial\ell_{j}(X|\Lambda)/\partial\Lambda_{\ell})| \leq h^{6}_{j\ell}(y,X)$

(A.2g) $|X_{j}(\partial\ell_{j'}(X|\Lambda)/\partial\Lambda_{\ell})| \leq h^{7}_{jj'\ell}(X)$

for all Λ in an open neighborhood of $\Lambda=\Lambda_{0}$, where $E(h^{j''})$ exists for $j''=1,\ldots,7$, for all ℓ,ℓ',j,j'.

Notes

1. It is well known that coefficient estimates are sensitive to specific stochastic distribution assumptions in many limited dependent variable contexts. For instance, Heckman and Singer(1984) illustrate the sensitivity of estimates for duration models, and establish an approach based on nonparametrically estimating the stochastic heterogeneity distribution.

2. See also Brillinger(1982), Goldberger(1981), Greene(1981,1983), Lawley(1943) and Singh and Ullah(1985), Stewart(1983), among others.

3. This framework differs slightly from that of Chung and Goldberger(1984) and Deaton and Irish(1984), since those papers only require ε (my notation) to be uncorrelated with X.

4. The support Ω is defined as the closure of the set $\{X \in R^M | p(X)>0\}$.

5. This terminology is due to that fact that $\ell(X)$ is the score vector of p with respect to a translation parameter - see Section 3.

6. This is shown by noting that condition A is satisfied by $\tilde{y}=G(X)=1$, a constant variable, and applying (3.3,4).

7. A similar link is used to establish the consistency of OLS estimators for the standard linear model. Namely, the functional form assumption that $E(\tilde{y}|X)=G(X)=\alpha+X'\beta$ implies $Cov(X,\tilde{y})=\Sigma_{XX}\beta$, or $Cov(\Sigma_{XX}^{-1}X,\tilde{y})=\beta$. By the same assumption, the behavioral effects are $\beta=\partial G(X)/\partial X=E(\partial G(X)/\partial X)$. The latter correspondence underlies the practical usefulness of the standard linear model.

8. Stoker(1986) gives a general development of local aggregate, or macroeconomic effects.

9. Other consistent estimators of $\gamma\beta$ include the "product moment" estimator $\hat{d}_1=\Sigma\ell_i y_i/K$, the "reduced form" OLS estimator of the slope coefficients of $y_i=\hat{c}_2+\hat{d}_2\hat{X}_i+\hat{u}_{2i}$, where $\hat{X}_i=(\Sigma_{\ell\ell}^{-1})^{-1}\ell_i$, and the weighted OLS estimator proposed by Ruud(1984). None of these estimators are first-order equivalent to either $\hat{d}$ or $\hat{d}_0$ in general.

10. This expanded notation is used in this section only.

11. X_2 then takes on the same role as the random term ε of section 2.

12. Notice that $\hat{d}_1$ of (4.12) consistently estimates $\gamma_1\beta_1$, and that the analogous coefficients from the linear equation with X_2 as explanatory variables will estimate $\gamma_2\beta_2$.

13. If X_j appears in both Z_1 and Z_2, its coefficient in $\hat{d}$ will estimate $\gamma_1\beta_{1j}+\gamma_2\beta_{2j}$. Thus common variables will have coefficient estimates that are the sum of the average derivatives induced from Z_1 and Z_2.

14. The framework is subject to the "order independence" property of Tversky(1972); see McFadden(1981). This can be relaxed without changing any of the substantive points of this example by allowing the joint distribution of $\{\varepsilon_j\}$ to depend on X_2, but not X_1.

15. Note that β would be consistently estimated by $\hat{d}$ if the instrument $\ell(X)$ were redefined as the score of the conditional distribution of X given ε, by treating ε as an extraneous variable as above. However, one could never compute these instruments, since the value of ε for each data point must be known as well as the conditional density of X given ε.

16. Nonparametric regression function estimators could be used to estimate β directly from (4.1). See Stone(1977) among many others, and Prakasa-Rao(1983) for a survey of these methods.

17. In this section it is implicitly assumed that the population covariance matrices Σ_{XX} and Σ_{Xy} exist.

18. The translation interpretation of the result is also straightforward within this context. In particular, if X is normally distributed with mean μ_X and variance-covariance matrix Σ_{XX}, then $Z=\alpha+X'\beta$ is normally distributed with mean $\mu_Z=\alpha+\mu_X'\beta$ and variance $\sigma_{ZZ}=\beta'\Sigma_{XX}\beta$. The translated density $p(X|\theta)$ is normal with mean $\mu_X+\theta$ and variance-covariance matrix Σ_{XX}, with the translated density of Z normal with mean $\alpha+\mu_X'\beta+\theta'\beta$ and variance σ_{ZZ}. Thus the mean of y under translation varies only with $\theta'\beta$, so that the aggregate effects $\partial E(y|\theta)/\partial\theta$ are proportional to β.

19. ξ may be heteroscedastic, so that this case includes standard heteroscedastic linear models as well as linear models with random coefficients, where the coefficients are distributed independently of the included X variables.

20. The residual interpretation of $r_n(Z)$ is established along the same line as the interpretation of R(X) above.

21. For example, the linearity condition is valid when the distribution of X is elliptically symmetric, as in $p(X)=p^+[(1/2)(X-\mu_X)'\Sigma_{XX}^{-1}(X-\mu_X)]=p^+(\delta(X))$. Here $\ell(X)=\omega(\delta(X)\Sigma_{XX}^{-1}(X-\mu_X)$, where $\omega(\delta)=-\partial\ln p^+/\partial\delta$. When $\omega(\delta)\geq 0$ for all δ, $\hat{d}$ is the weighted OLS coefficient estimator of (4.4), where the i^{th} observation (y_i,X_i) is weighted by $\sqrt{\omega(\delta(X_i))}$. Note that $\omega(\delta(X_i))=1$ when p(X) is the multivariate normal distribution.

22. The variance of $\hat{d}$ is estimated using the "heteroscedasticity consistent" estimator of White(1980).

23. Nonparametric estimators of the score vectors can be proposed using several methods, as surveyed by Manski(1984b) and Prakasa-Rao(1983). Gallant and Nychka(1985) prove consistency of $\hat{d}$ when score vectors are estimated using Hermite polynomial approximations.

References

Billingsley, P.(1979), Probability and Measure, New York, Wiley.

Brillinger, D.R.(1982), "A Generalized Linear Model with `Gaussian' Regressor Variables, " in P.J. Bickel, K.A. Doksum and J.L. Hodges, eds., A Festschrift for Erich L. Lehmann, Belmont, Woodsworth International Group.

Chamberlain, G.(1983), "A Characterization of the Distributions that Imply Mean-Variance Utility Functions," Journal of Economic Theory, 29, 185-201.

Chung, C-F. and A.S. Goldberger(1984),"Proportional Projections in Limited Dependent Variables Models," Econometrica, 52, 531-534.

Deaton, A.S. and M. Irish(1984), "Statistical Models for Zero Expenditures in Household Budgets," Journal of Public Economics, 23, 59-80.

Dempster, A.P.(1969), Elements of Continuous Multivariate Analysis, Reading, Massachusetts, Addison-Wesley.

Gallant A.R. and D.W. Nychka(1985), "Semi-Nonparametric Maximum Likelihood Estimation," Working Paper, Institute of Statistics, North Carolina State University.

Greene, W.H.(1981), "On the Asymptotic Bias of the Ordinary Least Squares Estimator of the Tobit Model," Econometrica, 49, 505-514.

Greene, W.H.(1983), "Estimation of Limited Dependent Variables Models by Ordinary Least Squares and the Method of Moments," Journal of Econometrics, 21, 195-212.

Goldberger, A.S.(1981), "Linear Regression After Selection," Journal of Econometrics, 15, 357-366.

Heckman, J.(1979), "Sample Selection Bias as a Specification Error," Econometrica, 47, 153-161.

Heckman, J. and B. Singer(1984), "A Method for Minimizing the Impact of Distributional Assumptions on Econometric Models for Duration Data," Econometrica, 52, 271-320.

Jennrich, R.I.(1969), Asymptotic Properties of Non-Linear Least Squares Estimators," Annals of Mathematical Statistics, 40, 633-643.

Kolmogorov, A.N.(1950), Foundations of the Theory of Probability, (German edition, 1933), New York, Chelsea.

Lawley, D.(1943), "A Note on Karl Pearson's Selection Formulae," Proceedings of the Royal Society of Edinburgh, Section A, 62, 28-30.

Manski, C.F.(1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," Journal of Econometrics, 3, 205-228.

Manski, C.F.(1984a), "Recent Work on Estimation of Econometric Models Under Weak Assumptions," SSRI Working Paper No. 8403, University of Wisconsin, January.

Manski, C.F.(1984b), "Adaptive Estimation of Nonlinear Regression Models," draft, Department of Economics, University of Wisconsin.

Manski, C.F.(1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," Journal of Econometrics, 27, 313-334.

McFadden, D.(1981), "Econometric Models of Probabilistic Choice," Chapter 5 in C.F. Manski and D. McFadden, eds., Structural Analysis of Discrete Data With Econometric Applications, Cambridge, Massachusetts, MIT Press.

Newey, W.K.(1984), "A Method of Moments Interpretation of Sequential Estimators," Economics Letters, 14, 201-206.

Powell, J.L.(1984), "Least Absolute Deviations Estimation for the Censored Regression Model," Journal of Econometrics, 25, 303-325.

Prakasa-Rao, B.L.S.(1983), Nonparametric Functional Estimation, Orlando, Florida, Academic Press.

Rao, C.R.(1973), Linear Statistical Inference and Its Applications, 2nd edition, New York, Wiley.

Ruud, P.A.(1983), "Sufficient Conditions for the Consistency of Maximum Likelihood Estimation Despite Misspecification of Distribution in Multinomial Discrete Choice Models," Econometrica, 51, 225-228.

Ruud, P.A.(1984), "Consistent Estimation of Limited Dependent Variables Models Despite Misspecification of Distribution," draft, revised October.

Singh, R.S. and A. Ullah(1985), "Nonparametric Time Series Estimation of Joint DGP, Conditional DGP, and Vector Autoregression," Econometric Theory, 1.

Stewart, M.B.(1983), "On Least Squares Estimation When the Dependent Variable is Grouped," Review of Economic Studies, 50, 737-753.

Stoker, T.M.(1986), "Aggregation, Efficiency and Cross Section Regression," forthcoming in Econometrica.

Stone, C.(1977), "Consistent Nonparametric Estimation," Annals of Statistics, 5, 595-645.

Tversky, A.(1972), "Elimination-by-Aspects: A Theory of Choice," Psychological Review, 79, 281-299.

White, H.(1980), "A Heteroscedasticity-Consistent Covariance Estimator and a Direct Test for Heteroscedasticity," Econometrica, 50, 483-500.