SEMIPARAMETRIC QUALITATIVE RESPONSE MODEL ESTIMATION WITH UNKNOWN HETEROSCEDASTICITY OR INSTRUMENTAL VARIABLES

Arthur Lewbel Boston College

original: Oct. 1996
revised: Aug. 1999

Abstract

This paper provides estimators of discrete choice models, including binary, ordered, and multinomial response (choice) models. The estimators closely resemble ordinary and two stage least squares. The distribution of the model's latent variable error is unknown and may be related to the regressors, e.g., the model could have errors that are heteroscedastic or correlated with regressors. The estimator does not require numerical searches, even for multinomial choice. For ordered and binary choice models the estimator is root N consistent and asymptotically normal. A consistent estimator of the conditional error distribution is also provided.

JEL Codes: C14, C25, C13. Keywords: Semiparametric, Discrete Choice, Qualitative response, heteroscedasticity, Binomial response, Multinomial response, Measurement Error, Instrumental variables, Binary choice, Ordered Choice, Latent Variable Models.

* This research was supported in part by the National Science Foundation through grant SBR-9514977. The author wishes to thank some anonymous referees, Dan McFadden, Bo Honoré, Jim Heckman, Xiaohong Chen, Richard Blundell, Andrew Chesher, Whitney Newey, Don Andrews, and participants in seminars at Berkeley, Bristol, Chicago, Harvard, MIT, Northwestern, Princeton, and Yale for helpful comments. Any errors are my own.

Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu.

1 Introduction

A new estimator for qualitative response models is proposed, including binary, multinomial, ordered response, and other discrete choice models. The distribution of the latent variable error is unknown and may be related to the regressors, e.g., the model can suffer from conditional heteroscedasticity of unknown form. An instrumental variables version of the estimator can deal with some forms of endogeneous or mismeasured regressors. The estimator does not require numerical searches, even for multinomial choice models. For binary and ordered choice the estimator is root N consistent and asymptotically normal. A consistent estimator of the conditional error distribution is also provided.

This paper's estimator treats one regressor as special. Let v denote the special regressor, and let x be a J vector of other regressors. Assume that the coefficient of v is positive (otherwise replace v with $-v$. An estimator of the sign of v is provided). Without loss of generality normalize this coefficient to equal one. With this normalization, the standard latent variable binary choice or binomial response model (see, e.g., Maddala 1983 or McFadden 1984 for surveys) is

$$(1.1) \qquad y_i = I(v_i + {x_i}^T\beta + e_i > 0)$$

where i indexes observations, y is the dependent variable, e is the unobserved latent variable error, β is a J vector of coefficients, ^{T} denotes transpose, and $I(\cdot)$ is the indicator function that equals one if $\cdot$ is true and zero otherwise. The data are N observations of y, v, and x.

Let $f(v|x)$ denote the conditional probability density function of v given an observation x, which can be estimated from the data. This paper's main result is that, under fairly general conditions,

$$\beta = E(xx^T)^{-1}E\left(x\frac{y - I(v > 0)}{f(v|x)}\right)$$

so β can be estimated by an ordinary least squares linear regression of $\widetilde{y}$ on x, where $\widetilde{y}_i = [y_i - I(v_i > 0)]/f(v_i|x_i)$.

If x is correlated with e, as in a model with mismeasured regressors, β can be estimated by an analogous two stage least squares linear regression based on

$$E(zx^T)\beta = E\left(z\frac{y - I(v > 0)}{f(v|z)}\right)$$

where z is a vector of instruments that are uncorrelated with e.

Let $F_e(e|\cdot)$ denote the conditional distribution of an observation of e given data $\cdot$. The minimal uncorrelated error assumption for linear models,

$$(1.2) \qquad E(ex) = 0$$

(or equation 1.4) is not generally sufficient to identify β in the binary choice model (1.1). An additional assumption that is made for identification and estimation here is that the distribution of e be conditionally independent of the one regressor v, or equivalently,

$$(1.3) \qquad F_e(e|v, x) = F_e(e|x).$$

The distribution of v will also be assumed to have a large support. Equations (1.2) and (1.3) require that the error distribution F_e not depend on v, but permit virtually any form of heteroscedasticity with respect to the vector of other regressors x.

The goal is estimation of the entire model, both the vector β and the conditional error distribution $F_e(e|x)$. Linearly regressing an estimate of $\widetilde{y}$ on x yields a root N consistent and asymptotically normal estimator for β. A uniformly consistent estimator of the distribution function $F_e(e|x)$, and of its density function $f_e(e|x)$ is also provided.

Lewbel (1998) uses similar assumptions to those here to identify general latent variable models of the form $y = L(v + x^T\beta + e)$ for some functions L, but the two papers differ in some fundamental ways. The identification in Lewbel (1998) comes from the change of variables argument $\int v\theta(y)dv = \int v\theta[L(v+x_i{}^T\beta+e)]dv = \int (w-x^T\beta-e)\theta[L(w)]dw = \int w\theta[L(w)]dw - (x^T\beta - e)\int \theta[L(w)]dw$, which is linear in $x^T\beta - e$, and hence permits estimation of β from the conditional expectation of the above object. This estimator works for many L functions but not for binary choice, and requires selection of a "tuning" function θ that possesses special properties.

The present paper's identification method is quite different, resulting in a simpler estimator. This paper's estimator is based on the observation that $y = I(v > -x^T\beta - e)$ and that $\int_L^K I(v > -x^T\beta - e) - I(v > 0)dv = x^T\beta + e$ (as long as L and K are large enough). Equation (1.3) makes $E(\widetilde{y}|x)$ equal the expection of this integral given x, and the least squares estimator then follows directly.

A perhaps more intuitive estimator can be derived by observing that, given equation (1.3), $E(y|v,x)$ equals the conditional distribution function of $-x^T\beta - e$, evaluated at v. From this distribution a direct estimate of the conditional mean of $-x^T\beta - e$ given x can be constructed, and β could then be estimated by regressing that conditional mean estimate on x. An integration by parts argument is later used to show that this alternative estimator is equivalent in expectation to regressing $\widetilde{y}$ on x.

Many estimators exist for binary choice models. Standard maximum likelihood estimation requires a finitely parameterized conditional distribution $F_e(e|x)$. Semiparametric and quasimaximum likelihood estimators of linear latent variable qualitative response models include Cosslett (1983), Ruud (1983), Powell, Stock and Stoker (1989), Ichimura (1993), Klein and Spady (1993), Newey and Ruud (1994), and Härdle and Horowitz (1996). To estimate (1.1), these estimators require either that $F_e(e|v,x) = F_e(e)$ or that $F_e(e|v,x) = F_e(e|v + x^T\beta)$, and hence they impose far more restrictions on e than does equation (1.3) or (1.5). The present paper shares a feature with the Ruud papers of density weighting, though instead of making all of the regressors appear normal as Ruud requires, the present paper's weighting makes the single regressor v appear uniform.

Other semiparametric estimators of binary choice models assume the conditional median or other known quantile of e given x is zero. Examples are Manski (1975), (1985), and Horowitz (1992), (1993). Like the present paper's estimator, these quantile estimators permit quite general forms of conditional heteroscedasticity, but they converge at slower than root N rates. Let $Q_\theta(e|\cdot)$ denote the θ'th quantile of e, conditional on the data $\cdot$. The quantile estimators assume $Q_\theta(e|v,x) = Q_\theta(e)$ for some known θ. In contrast, by equa-

tion (1.3) this paper assumes that $Q_\theta(e|v,x) = Q_\theta(e|x)$ holds for all quantiles θ. Instead of assuming that one of the quantiles does not depend on all of the regressors, this paper assumes that all of the quantiles do not depend on one of the regressors.

The estimator proposed here includes estimation of the constant term in the latent variable, unlike some other estimators like Powell, Stock and Stoker (1989) or Klein and Spady (1993), which fail to estimate the constant term.

The present paper's estimator can be extended to handle certain types of endogeneous or mismeasured regressor models. For this extension, equations (1.2) and (1.3) are replaced with

$$E(ez) = 0 \tag{1.4}$$

$$F_{ex}(e,x|v,z) = F_{ex}(e,x|z) \tag{1.5}$$

where z is a vector of instruments and $F_{ex}(e,x|\cdot)$ denotes the distribution of an observation of (e,x) conditional on data $\cdot$. Equation (1.5) limits the range of applications of this extension (see section 3 for details), but very few semiparametric estimators exist for any kind of endogeneous or mismeasured regressor binary choice models. One example is Newey (1985).

Section 2 below provides general examples of models that satisfy the identification equations (1.2) and (1.3), and describes the basic ordinary least squares identification result. Section 3 gives example models that satisfy the instrumental variables identification equations (1.4) and (1.5), and provides the corresponding two stage least squares identification of β. An alternative estimator that gives some intuition behind the identification is also provided.

When the distribution of v is unknown, the estimation of β employs a first stage nonparametric density estimator. Section 4 gives the limiting root N distribution of $\widehat{\beta}$, taking into account the estimation error in the required density estimate. A simpler estimator that does not require a nonparametric first stage is also provided in section 4, though more stringent conditions are needed for its consistency. The only computations this alternative, ordered data based estimator uses are a sorting of the data and two linear regressions.

Section 5 describes estimation of the distribution of the errors, including conditional and unconditional moments of e, and the sign of the coefficient of v.

Section 6 provides extensions of the estimator, including estimation of ordered choice, multinomial choice, partly linear latent variable, threshold, and censored regression models.

Section 7 discusses selection of the kernel and bandwidth for the required nonparametric density estimation, and provides the results of a Monte Carlo analysis of the estimator. The Monte Carlo includes an analysis of the behavior of estimated asymptotic standard errors, which is always relevant for empirical applications but is often ignored in Monte Carlo studies of estimators. Section 8 provides concluding remarks.

2 The Basic Model and Estimator

2.1 Example Models

Each of the following conditions and models is sufficient to make the identifying equations (1.2) and (1.3) hold. Example models that satisfy the instrumental variables assumptions (1.4) and (1.5) are discussed later.

1. Equations (1.2) and (1.3) hold when the errors e are mean zero and independent of the regressors (v, x). This is the traditional latent variable error assumption satisfied by standard models like logit and probit.

2. Equation (1.3) holds when $e = \nu(e^*, x)$ for any function ν and any random vector e^*, where the distribution of e^* is independent of (v, x). The function ν, the vector e^* and its distribution do not need to be known, observed, or estimated. This permits virtually any form of conditional heteroscedasticity involving x, but not v. For example, the standard random coefficients model $y = I[v + x^T(\beta + e^*) + e_0^* > 0]$ with $E(e^*) = 0$ and $E(e_0^*) = 0$ equals equation (1.1) with $e = x^T e^* + e_0^*$, and satisfies equations (1.2) and (1.3). Every regressor except v could have a random coefficient.

3. Equation (1.3) holds when v is independent of (e, x). This case is most likely to occur when v is determined by experimental design, such as the "Willingness-to-Pay" models used to determine the perceived value of a public good. In these models each agent i is asked if he or she would be willing to pay some randomly drawn (log) price $-v_i$ for a public good like a road or a park. Then y_i is one for a yes answer and zero for no, and x_i consists of observables that affect agent i's decision. See McFadden (1993) and Lewbel and McFadden (1997), who study this application in detail. In cases like these where the density of v is known, the proposed estimator of β has no nonparametric estimation component.

A common application of discrete choice models is where y_i denotes whether a consumer i purchases a particular good or service. In these applications, as in the willingness to pay models, the error e is interpreted as preference heterogeneity, and hence may depend on demographic and taste attributes of the consumer. The error e will therefore be independent of prices when these prices are determined solely from the supply side of the economy (e.g., prices that equal the marginal cost of production under perfect competition and constant returns to scale). This makes equation (1.2) hold with v equaling the negative logged price.

The estimator requires that the latent variable $v_i + {x_i}^T\beta + e_i$ be linear in v_i. This too will be generally satisfied in purchase decision applications, because such models are identical to willingness-to-pay models, except that the price is determined by the market instead of by experimental design. The consumer is assumed to buy if the (log) actual price charged, $-v_i$, is less than the consumer's (log) reservation price, ${x_i}^T\beta + e_i$, resulting in a latent variable linear in v. However, unless purchases are observed over many regimes (or are determined by experimental design, as in the willingness to pay applications), the large support assumption may not hold when v is a price. Of course, other variables may

be suitable choices for v as well.

2.2 Identification

The ordinary least squares identification of β is described here. The extension to instrumental variables is provided in the next section.

ASSUMPTION A.1: Equation (1.1) holds. The conditional distribution of v given x is absolutely continuous with respect to a Lebesgue measure with nondegenerate Radon-Nikodym conditional density $f(v|x)$.

ASSUMPTION A.2: Let Ω denote the support of the distribution of an observation of (v, x). Let $F_e(e|v, x)$ denote the conditional distribution of an observation of e given an observation of (v, x), with support denoted $\Omega_e(v, x)$. Assume $F_e(e|v, x) = F_e(e|x)$ and $\Omega_e(v, x) = \Omega_e(x)$ for all $(v, x) \in \Omega$.

ASSUMPTION A.3: The conditional distribution of v given x has support $[L, K]$ for some constants L and K, $-\infty \leq L < 0 < K \leq \infty$. The support of $-x^T\beta - e$ is a subset of the interval $[L, K]$.

ASSUMPTION A.4: $E(ex) = 0$. $E(xx^T)$ exists and is nonsingular.

These assumptions do not require independent observations, though the root N estimator provided later will assume independence. The identification result in Theorem 1 below only requires that the expectation of a certain function of f be identified.

Assumption A.1 says that y is given by the binary choice model (1.1) and that v has a continuous distribution. Assumptions A.2 and A.4 were discussed in sections 1 and 2.1 in terms of equations (1.2) and (1.3).

For estimation of β, the distribution of e is not required to be continuous, e.g., it can be discrete or contain mass points. However, later estimation of the error distribution function F_e will assume continuity.

The vector of regressors x can include dummy variables. Squares and interaction terms, e.g., $x_{3i} = x_{2i}^2$, are also permitted. In addition, x can be related to (e.g., correlated with) v, though Assumption A.1 rules out having elements of x be deterministic functions of v.

Assumption A.3 requires v to have a large support. Standard models like logit or probit have errors that can take on any value, which would by Assumption A.3 require v to have support equal to the whole real line. This assumption implies that the estimator is likely to perform best when the spread of observations of v is large relative to the spread of $x^T\beta + e$ (since if the observed spread of v values were not large, then the observed data would resemble data drawn from a process that violated A.3). Assumption A.3 assumes zero is in the support of v. To make this hold, for any constant κ in the support of v (e.g., the mean or median of v) we can redefine v as $v - \kappa$ and correspondingly add κ to the estimated constant in $x^T\beta$.

Theorem 1: Define $\widetilde{y}$ by

$$(2.1) \qquad \widetilde{y} = \frac{y - I(v > 0)}{f(v|x)}$$

If Assumptions A.1, A.2, and A.3 hold then $E(\widetilde{y}|x) = x^T\beta + E(e|x)$. If Assumption A.4 also holds then

$$(2.2) \qquad \beta = E(xx^T)^{-1}E(x\widetilde{y})$$

Theorem 1 shows that β is identified, and can be estimated by an ordinary least squares regression of $\widetilde{y}$ on x. Theorem 1 is proved in Appendix A.

3 Instrumental Variables

This section describes models in which $E(ex) \neq 0$, including endogeneous regressors and measurement error models, that can be estimated using an instrumental variables extension of Theorem 1. These models satisfy equations (1.4) and (1.5) instead of (1.2) and (1.3). It should be recalled throughout this section that x and z can overlap, that is, one or more elements of the vector of regressors x can also be in the vector of instruments z.

3.1 Example Instrumental Variables Models

One way equation (1.5) arises is in structural models of the form $(e, x) = \varsigma(\tilde{e}, z)$ where $\tilde{e}$ is a random vector that is independent of (v, z) and ς is some vector valued function. The function ς, the variables $\tilde{e}$, and the distribution function of $\tilde{e}$ do not need to be known, observed, or estimated. This structure has the empirically testable implication that the conditional distribution of x given v and z is independent of v. As with many semiparametric estimators, the distribution of the regressors is not ancillary, and hence restrictions like this one on v can arise.

A system of equations that satisfies (1.1), (1.4) and (1.5) is, for any function τ,

$$(3.1) \qquad y = I(v + x^T\beta + e > 0)$$
$$(3.2) \qquad x = \tau(z) + \epsilon$$

where e and ϵ are unconditionally mean zero and the distribution of (e, ϵ) is independent of (v, z). Here e and ϵ can be correlated, and hence e can be correlated with x. Dependence of (e, ϵ) on z would also be permitted in this model, provided that (e, ϵ) is conditionally independent of v, conditioning on z. The proposed estimator for β does not require specifying or estimating the function τ.

Equations (3.1) and (3.2) are a standard example of a system of equations that would be recursive, except for the fact that the errors are correlated across equations. More fully endogeneous systems, in which x depends on v or y, are ruled out by equation (1.5).

Equations (3.1) and (3.2) can arise in a standard mismeasured regressors framework. Consider the model

$$
\begin{aligned}
(3.3) \qquad y &= I(v + x^{*T}\beta + e^* > 0) \\
(3.4) \qquad x &= x^* + \tilde{\epsilon} \\
(3.5) \qquad x^* &= \tau(z) + \epsilon^*
\end{aligned}
$$

with $E(e^*, \epsilon^*, \tilde{\epsilon}) = 0$. Here v is a regressor that is observed without error, x is a mismeasured observation of the unobserved x^*, e^* is the latent model error and $\tilde{\epsilon}$ is a vector of measurement errors. Each regressor x_j that is not mismeasured has $z_j = x_j^* = x_j$ and $\epsilon_j^* = \tilde{\epsilon}_j = 0$. Equations (3.3), (3.4), and (3.5) are equivalent to (3.1) and (3.2) with $e = -\tilde{\epsilon}^T\beta + e^*$ and $\epsilon = \epsilon^* + \tilde{\epsilon}$. Once again, estimation of β will not require specification or estimation of the function τ.

Let $w = v + x^{*T}\beta + e^*$. If the latent variable w were observed instead of y, then β would be estimated in this measurement error model by a two stage least squares regression of w on x, assuming that $(e^*, \epsilon^*, \tilde{\epsilon})$ is uncorrelated with (v, z). When y is observed instead of w, equation (1.5) will hold and the estimator described in the next two sections can be used to estimate β, as long as $(e^*, \epsilon^*, \tilde{\epsilon})$ is independent of (v, z), or more generally if $(e^*, \epsilon^*, \tilde{\epsilon})$ is uncorrelated with z and its conditional distribution given v and z does not depend on v. Having these errors be independent of v and z is stronger than necessary.

The only way equations (3.3), (3.4), and (3.5) differ from standard linear mismeasured regressors models (other than y being observed instead of the latent w) is that standard models would permit x^* to depend on v. That is possible in the current framework only if we let $x^* = \rho(v, u) + \epsilon^*$ for some instruments u, and define $z = \rho(v, u)$, which makes (3.5) hold with $\tau(z) = z$. The drawback of permitting x^* to depend on v in this way is that the estimation of β would then first require estimation of $z = \rho(v, u) = E(x|v, u)$ (as the fits from parametrically or nonparametrically regressing the observed x on v and u).

Equation (1.5) says that e and v are conditionally independent, conditioned on x and z, but does not rule out correlations of v with x and z. For example, the structure $x = az + \epsilon_1$, $v = bz + \epsilon_2$, and $e = c\epsilon_1 + \epsilon_3$ for any mutually independent variables z, ϵ_1, ϵ_2, and ϵ_3 (with ϵ_1 and ϵ_3 being mean zero) and any constants a, b, and c, makes (1.4) and (1.5) hold while making v correlate with x and z, and x correlate with e. This structure will be used later in a Monte Carlo design.

3.2 Identification With Instrumental Variables

Theorem 1' below describes the instrumental variables identification of β, as would be used for the models described in the previous section, that is, where equations (1.2) and (1.3) are replaced with (1.4) and (1.5). Note that Assumption A.4' below is the standard assumption about instruments in two stage least squares regressions.

ASSUMPTION A.1': Equation (1.1) holds. The conditional distribution of v given z is absolutely continuous with respect to a Lebesgue measure with nondegenerate Radon-Nikodym conditional density $f(v|z)$.

ASSUMPTION A.2': Let Ω denote the support of the distribution of an observation of (v, z). Let $F_{ex}(e, x|v, z)$ denote the conditional distribution of an observation of (e, x) given an observation of (v, z), with support denoted $\Omega_{ex}(v, z)$. Assume $F_{ex}(e, x|v, z) = F_{ex}(e, x|z)$ and $\Omega_{ex}(v, z) = \Omega_{ex}(z)$ for all $(v, z) \in \Omega$.

ASSUMPTION A.3': The conditional distribution of v given z has support $[L, K]$ for some constants L and K, $-\infty \leq L < 0 < K \leq \infty$. The support of $-x^T\beta - e$ is a subset of the interval $[L, K]$.

ASSUMPTION A.4': $E(ez) = 0$, $E(zz^T)$ exists and is nonsingular, and the rank of $E(xz^T)$ is J (the dimension of x).

Define Σ_{xz}, Σ_{zz}, Δ, and $\tilde{y}^*$ by $\Sigma_{xz} = E(xz^T)$, $\Sigma_{zz} = E(zz^T)$,

$$\Delta = (\Sigma_{xz}\Sigma_{zz}^{-1}\Sigma_{xz}^T)^{-1}\Sigma_{xz}\Sigma_{zz}^{-1} \tag{3.6}$$

$$\tilde{y}^* = \frac{y - I(v > 0)}{f(v|z)} \tag{3.7}$$

Theorem 1': If Assumptions A.1', A.2', and A.3' hold then $E(\tilde{y}^*|z) = E(x|z)^T\beta + E(e|z)$. If Assumption A.4' also hold then

$$\beta = \Delta E(z\tilde{y}^*) \tag{3.8}$$

The earlier Theorem 1 is the special case of Theorem 1' in which $z = x$. Theorem 1' shows that β is identified, and can be estimated by an ordinary linear two stage least squares regression of $\tilde{y}^*$ on x, using instruments z. The proof of Theorem 1' is in Appendix A.

Corollary 1: If Assumptions A.1', A.2', A.3' and A.4' hold then

$$\beta = \Delta E(z\int_L^K E[y - I(v > 0)|v, z]dv) \tag{3.9}$$

The two stage least squares estimator for β based on equation (3.8) entails averaging $z\tilde{y}^*$, where $\tilde{y}^*$ has the conditional density of v in the denominator, and hence may be adversely affected by extreme observations of v. Corollary 1 suggests an alternative based on numerically integrating an estimate of the conditional expectation $E[y - I(v > 0)|v, z]$. While more complicated, this alternative might be more robust to outliers in v.

3.3 Another Estimator and a Little Intuition

This section describes an alternative estimator closely related to Theorem 1', which provides some insight into the identification. This alternative will later be extended to yield an

estimator for multinomial choice models. Here again one can take $z = x$ for applications that don't require separate instruments.

Let $G^*(v,z) = E(y|v,z)$, $s = s(x,e) = -x^T\beta - e$ and let $F_s(s|v,z)$ and $f_s(s|v,z)$ denote the conditional distribution and probability density functions of s, assuming that e (and hence s) conditional on x and z is continuously distributed. It follows from equation (1.5) that $F_s(s|v,z) = F_s(s|z)$, so $E[I(s < v)|v,x,z] = E[I(s < v)|x,z]$. Therefore

$$F_s(v|z) = E[I(s < v)|z] = E(y|v,z) = G^*(v,z) \tag{3.10}$$

If s were observed, then β could be estimated by a two stage least squares regression of $-s$ on x using instruments z. This regression would depend on s only through $E(zs)$. The key insight is that this regression does not require that s be observed. Only an estimate of $E(zs)$ is needed, and that can be obtained because, by equation (3.10), the distribution of s is available. In particular,

$$\beta = \Delta E[-zE(s|z)] = \Delta E[-z\int_L^K sf_s(s|x,z)ds = \Delta E[-z\int_L^K v\frac{\partial G^*(v,z)}{\partial v}dv] \tag{3.11}$$

$$\begin{aligned}\beta &= \Delta E[-z\int_L^K v\frac{\partial G^*(v,z)/\partial v}{f(v|z)}f(v|z)dv] = \Delta E\left(-zE[v\frac{\partial G^*(v,z)/\partial v}{f(v|z)}|z]\right) \\ &= \Delta E[-zv\frac{\partial G^*(v,z)/\partial v}{f(v|z)}] = \Delta E(-zy^{**})\end{aligned} \tag{3.12}$$

where y^{**} is defined by

$$y^{**} = -v\frac{\partial G^*(v,z)/\partial v}{f(v|z)} \tag{3.13}$$

Equations (3.12) and (3.13) show that Theorem 1' holds using y^{**} in place of $\tilde{y}^*$, so β can be estimated as a linear two stage least squares regression of y^{**} on x using instruments z. Unlike Theorem 1', this estimator would require a preliminary estimate of G^* as well as f. Kernel estimators could be used for both.

The above analysis also provides an alternative derivation (and hence interpretation) of Theorem 1'. Applying an integration by parts to the last integral in equation (3.11) yields $\beta = \Delta E[z\int_L^K G^*(v,z) - I(v > 0)dv]$. Proceeding as in equation (3.12) with this expression then yields $\beta = \Delta E\,(z[G^*(v,z) - I(v > 0)]/f(v|z)) = \Delta E\,(z[y - I(v > 0)]/f(v|z))$, which is Theorem 1'.

4 Root N Consistent Estimation

This section gives the root N consistent, asymptotically normal limiting distribution for a two stage least squares estimator based on Theorem 1', using instruments z. All of the equations here simplify to the ordinary least squares estimator in Theorem 1 when $z = x$.

Let $u = u(z)$ be any vector of variables such that the conditional density of v given z equals the conditional density of v given u, that is, $f(v|u) = f(v|z)$, where no element of u equals a deterministic function of other elements of u. For example, if $z = (1, \widetilde{z}, \widetilde{z}^2)$, we could take $u = \widetilde{z}$. This construction of u is employed because $f(v_i|z_i)$ will be estimated as $f(v_i|u_i)$ using a kernel density estimator (see Appendix B for details). Also, if v were known to be independent of some elements of z, then u could exclude those elements. For example, in the willingness to pay application discussed earlier, v is determined by experimental design and so could by construction be independent of some or all the elements of z.

Define y_i^* and η by

$$(4.1) \qquad y_i^* = \frac{y_i - I(v_i > 0)}{f(v_i|u_i)}$$

$$(4.2) \qquad \eta = E(zy^*)$$

Given the definition of u, $y_i^* = \tilde{y}_i^*$ from equation (3.7), and if $x = z$ then $y_i^* = \tilde{y}_i$ from equation (2.1). The estimator is based on the expression $\beta = \Delta\eta$, which holds by Theorem 1'.

Assume we have an estimator $\hat{f}(v|u)$ of the conditional density $f(v|u)$. One choice is a kernel estimate of the joint density of v and u divided by a kernel estimate of the density of u (Equation B.3 in Appendix B). A simpler alternative is given in the next section.

Define $\hat{\Sigma}_{xz}$, $\hat{\Sigma}_{zz}$, $\widehat{y_i^*}$, $\hat{\eta}$, and $\hat{\beta}$, by

$$(4.3) \qquad \hat{\Sigma}_{xz} = N^{-1}\Sigma_{i=1}^N x_i z_i^T$$

$$(4.4) \qquad \hat{\Sigma}_{zz} = N^{-1}\Sigma_{i=1}^N z_i z_i^T$$

$$(4.5) \qquad \hat{\Delta} = (\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}\hat{\Sigma}_{xz}^T)^{-1}\hat{\Sigma}_{xz}\hat{\Sigma}_{zz}^{-1}$$

$$(4.6) \qquad \widehat{y_i^*} = [y_i - I(v_i > 0)]/\hat{f}(v_i|u_i)$$

$$(4.7) \qquad \hat{\eta} = \Sigma_{i=1}^N z_i \widehat{y_i^*}/N$$

$$(4.8) \qquad \hat{\beta} = \hat{\Delta}\hat{\eta}$$

The $\hat{\eta}$ in equation (4.7) is a two step estimator with a nonparametric first step. The limiting distribution of estimators of these forms has been studied by many authors. See e.g. Newey and McFadden (1994) and references therein. Appendix B provides one set of regularity conditions that do not violate Assumptions A.1' to A.4' and are sufficient for root N consistent, asymptotically normal convergence of $\hat{\eta}$ and hence of $\hat{\beta}$. To summarize the result, define q_i by

$$(4.9) \qquad q_i = z_i y_i^* + E(z_i y_i^*|u_i) - E(z_i y_i^*|v_i, u_i)$$

Note that

$$(4.10) \qquad E(q) = E(zy^*) = \eta$$

Appendix B then provides conditions such that

$$(4.11) \qquad \sqrt{N}\hat{\eta} = N^{-1/2}\Sigma_{i=1}^{N} q_i + o_p(1)$$

so q_i is the influence function for $\hat{\eta}$. It follows from (4.8) and (4.11) that $\sqrt{N}\hat{\beta} = \sqrt{N}\hat{\Delta}\hat{\eta} = N^{-1/2}\hat{\Delta}\Sigma_{i=1}^{N}[z_i x_i^T\beta + (q_i - z_i x_i^T\beta)] + o_p(1) = \sqrt{N}\beta + N^{-1/2}\hat{\Delta}\Sigma_{i=1}^{N}(q_i - z_i x_i^T\beta) + o_p(1)$, and therefore

$$(4.12) \qquad \sqrt{N}(\hat{\beta} - \beta) \Rightarrow N\left(0, \Delta var(q - zx^T\beta)\Delta^T\right)$$

The variance of $\hat{\beta}$ can be estimated as $\hat{\Delta}\widehat{var}(\hat{q} - zx^T\hat{\beta})\hat{\Delta}^T/N$, where $\widehat{var}(\hat{q} - zx^T\hat{\beta})$ is the sample variance of $\hat{q}_i - z_i x_i^T\hat{\beta}$ for $i = 1, ..., N$, and $\hat{q}_i$ is constructed by replacing y_i^* with $\widehat{y}_i^*$ in equation (4.9) and replacing the conditional expectations in that equation with nonparametric regressions.

If $E(z|u) = z$ (as will often be the case given the definition of u) then $q_i = z_i[y_i^* + E(y_i^*|u_i) - E(y_i^*|v_i, u_i)]$, and, the limiting distribution of $\hat{\beta}$ becomes identical to the limiting distribution of a two stage least squares regression of $[y_i^* + E(y_i^*|u_i) - E(y_i^*|v_i, u_i)]$ on x using instruments z. When f is a known function that does not need to be estimated, the above expressions simplify further to $q_i = z_i y_i^*$, a two stage least squares regression of y_i^* on x using instruments z.

Theorem 1' can be generalized to permit nonindependent and nonidentically distributed observations, essentially by adding i subscripts to Ω, F, Ω_{ex}, and the expectation operator. Only the conditional density function $f(v|u)$ must be assumed to be constant (or its variation finitely parameterized) across observations, so f and hence β can be estimated. Many results exist providing limiting distribution theory for semiparametric estimators when observations are not independently or identically distributed. See, e.g., Andrews (1995).

In place of kernel estimators, a consistent $\hat{\beta}$ could be obtained using a series expansion based density estimator of f in equation (4.6), as in Gallant and Nychka (1987). Alternatively, parametric specifications of f may be used in some applications, e.g., when v is income, which is known to be well approximated by a lognormal distribution with a Pareto tail. A consistent $\hat{\beta}$ could also be obtained from Corollary 1, using either kernel or series based estimators of the conditional expectation in equation (3.9). The next section provides another estimator, one that is computationally extremely simple.

4.1 A Very Simple Ordered Data Estimator

This section describes an alternative two stage least squares estimator based on Theorem 1' that does not require a first stage kernel estimator.

ASSUMPTION A.5': Assume that for some vector γ, equation (4.13) holds, where w_i is continuously distributed with bounded support, has mean zero, and is independent of z_i.

$$(4.13) \qquad v_i = z_i^T\gamma + w_i$$

Let f_w denote the unconditional density function of w. If Assumption A.5' holds then $f_w(w_i) = f(v_i|z_i)$. Define $\hat{w}$ as the residuals from linearly regressing v on z, so

$$(4.14) \qquad \hat{w}_i = v_i - z_i(\Sigma_{i=1}^N z_i z_i^T)^{-1}\Sigma_{i=1}^N z_i v_i$$

Let w_i^+ denote the smallest element of $\{\hat{w}_1, ..., \hat{w}_N\}$ that is greater than $\hat{w}_i$, and let w_i^- denote the largest element of $\{\hat{w}_1, ..., \hat{w}_N\}$ that is less than $\hat{w}_i$. In other words, if the data $\hat{w}_1, ..., \hat{w}_N$ were sorted in ascending order, the number immediately preceeding $\hat{w}_i$ would be w_i^-, and the number immediately following $\hat{w}_i$ would be w_i^+. Now i/N is an estimate of the distribution of w evaluated at $\hat{w}_i$, so $2/[(w_i^+ - w_i^-)N] \approx f_w(\hat{w}_i) \approx f(v_i|z_i)$. Define $\widehat{\widehat{y}}_i^*$ by

$$(4.15) \qquad \widehat{\widehat{y}}_i^* = [y_i - I(v_i > 0)](w_i^+ - w_i^-)N/2$$

It follows from Theorem 3 of Lewbel and McFadden (1997) that $N^{-1}\Sigma_{i=1}^N z_i \widehat{\widehat{y}}_i^*$ is a consistent estimator of $E(z[y - I(v > 0)]/f_w(w))$. Therefore, if Assumptions A.1' to A.5' hold, then

$$(4.16) \qquad \hat{\beta} = \hat{\Delta}\Sigma_{i=1}^N z_i \widehat{\widehat{y}}_i^*/N$$

is a consistent estimator of β.

To summarize, the ordered data estimator (4.16) consists of 1. define $\hat{w}$ as the residuals from regressing v on z using ordinary least squares, 2. sort the $\hat{w}$ data from smallest to largest, to find w_i^+ and w_i^- for each observation i, 3. construct $\widehat{\widehat{y}}_i^*$ in equation (4.15) for $i = 1, \ldots, N$, and 4. let $\hat{\beta}$ be the estimated coefficients from regressing this $\widehat{\widehat{y}}_i^*$ on x_i using two stage least squares, with instruments z_i. As with the other estimators, (4.16) reduces to an ordinary least squares regression when $z = x$.

This estimator is convenient for its numerical simplicity, but it requires the extra Assumption A.5' for consistency. This assumption limits the permitted dependence of v on z. White's test for homoscedasticity or more general tests of independence of errors from regressors can be applied to $\hat{w}$ versus z to test Assumption A.5'.

5 Other Model Components

5.1 The Sign of the Coefficient of v

Without loss of generality, the linear latent variable binary choice model can be written as

$$(5.1) \qquad y_i = I(v_i\alpha + x_i{}^T\beta + e_i > 0)$$

where $\alpha = -1$, 0, or $+1$. The estimator of β assumed $\alpha = 1$. If α is not known a priori, then an estimator of α is required. Let $G(v, x, z) = E(y|v, x, z)$. Assume $F_e(e|v, x, z) = F_e(e|x, z)$, which is a little stronger than Assumption A.2' when $x \neq z$. Then

$$(5.2) \qquad E(y|v, x, z) = G(v, x, z) = 1 - F_e(-v\alpha - x^T\beta|x, z)$$

Let $f_e(e|x,z) = \partial F_e(e|x,z)/\partial e$ denote the probability density function of e, assuming e has a continuous distribution. Then G will be differentiable in v, so let $g(v,x,z) = \partial G(v,x,z)/\partial v$. It then follows from (5.2) that.

$$(5.3) \qquad \partial E(y|v,x,z)/\partial v = g(v,x,z) = \alpha f_e(-v\alpha - x^T\beta|x,z)$$

Let δ_v denote the density weighted average of $g(v,x,z)$, and let $\hat{\delta}_v$ equal Powell, Stock, and Stoker's (1989) estimator of the weighted average derivative δ_v. By (5.3) and the definition of δ_v, $\alpha = sign(\delta_v)$. This suggests the estimator $\hat{\alpha} = sign(\hat{\delta}_v)$. Since $\hat{\delta}_v$ converges at rate root N, $sign(\hat{\delta}_v)$ will converge at a rate faster than root N, so estimating α first in this way will not affect the limiting root N distribution of $\hat{\beta}$.

More generally, $\alpha = sign[g(v,x,z)]$ for all v, x, and z. Therefore a test of the hypothesis $g(v,x,z) = 0$ is a test of whether $\alpha = 0$. Applicable nonparametric consistent tests of $g(v,x,z) = 0$ include Lewbel (1995) and Ait-Sahalia, Bickel, and Stoker (1997). Also, $G(v,x,z)$ must be nondecreasing in v if $\alpha = 1$, nonincreasing in v if $\alpha = -1$, and is independent of v if $\alpha = 0$. The function G can be estimated as a nonparametric regression of y on v, x, and v. This $\hat{G}$ could also be used as a check on the model specification, in that nonmonotonicity of G as a function of v would indicate that the model is misspecified. Equivalently, having the nonparametric regression derivative $\hat{g}(v,x,z)$ be significantly positive for some values of (v,x,z) and significantly negative for others would indicate model misspecification.

Suppose we failed to check $\hat{\alpha}$ as above, and erroneously applied the estimator of Theorem 1 assuming $\alpha = 1$, when in fact $\alpha = -1$? It is straightforward to verify, following the steps of the proof of Theorem 1, that if $\alpha = -1$ then $E(\tilde{y}^*|z)$ will equal $E(x^T\beta + e - L - K|z)$, so $p\lim\hat{\beta}$ will equal $\beta - \Delta E(z)(L+K)$, where Δ is given by equation (3.6) and L and K are the lower and upper bounds of the support of v. If the support of v is symmetric, so $L = -K$, then $\hat{\beta}$ will consistently estimate β regardless of whether α is plus one or minus one. However, this dependence on the support of v implies that the estimate $\hat{\beta}$ is likely to be quite erratic (i.e., very sensitive to the few largest observations of $|v|$) if $\alpha = -1$.

5.2 The Distribution of the Latent Error

Let $\hat{G}(v,x,z)$ be a kernel or other nonparametric regression of y on v, x, and z, and let $\hat{g}(v,x,z) = \partial\hat{G}(v,x,z)/\partial v$, so $\hat{g}$ is a nonparametric regression derivative. Assuming again that $\alpha = 1$, it follows from equation (5.2) that

$$(5.4) \qquad F_e(e|x,z) = 1 - G(-e - x^T\beta, x, z)$$

and therefore $\hat{F}_e(e|x,z) = 1 - \hat{G}(-e - x^T\hat{\beta}, x, z)$ and $\hat{f}_e(e|x,z) = \hat{g}(-e - x^T\hat{\beta}, x, z)$ are consistent estimators of F_e and f_e. The limiting distributions of these functions will be the same as if $\hat{\beta}$ were replaced by the true β, because $\hat{\beta}$ converges to its limit at rate root N.

Attributes of F of interest include moments such as $E[\psi(e)|x,z]$ and $E[\psi(e)]$ for given functions ψ. For example, $\psi(e) = e$ and $\psi(e) = e^2$ provide conditional and unconditional means and variances of e. Estimates of such moments could be recovered from $\hat{g}$ since $E[\psi(e)|x,z] = \int \psi(e)g(-e - x^T\beta, x, z)de$. The following theorem, which is closely related to McFadden (1993) and Lewbel (1997), provides a simpler expression for $E[\psi(e)|x,z]$, which can be estimated without first nonparametrically estimating the derivative g.

Theorem 2: Let Assumptions A.1', A.2', and A.3' hold. Assume e has a continuous conditional distribution with density function $f_e(e|x,z) = f_e(e|v,x,z)$ and support equal to the whole real line. Assume that $\psi(e)$ is differentiable, $\psi(0) = 0$, and $\lim_{e\to\infty} \psi(e)[F_e(e|x,z) - 1] = \lim_{e\to-\infty} \psi(e)F_e(e|x,z) = 0$. Let $\psi'(e) = \partial\psi(e)/\partial e$. Then equation (5.5) holds.

$$E[\psi(e)|x,z] = E[\psi'(-v - x^T\beta)\frac{y + I(v + x^T\beta < 0) - 1}{f(v|x,z)}|x,z] \tag{5.5}$$

The assumption that $\psi(0) = 0$ is made without loss of generality, since if it does not hold then ψ could be replaced with $\psi(e) - \psi(0)$. Define

$$\hat{H}_{\psi i} = \psi'(-v_i - x_i^T\hat{\beta})[y_i + I(v_i + x_i^T\hat{\beta} < 0) - 1]/\hat{f}(v_i|x_i,z_i) \tag{5.6}$$

By Theorem 2, $E[\psi(e)|x,z]$ can be estimated by a nonparametric regression of $\hat{H}_{\psi i}$ on x_i and z_i, and unconditional moments $E[\psi(e)]$ can be estimated as the sample average of $\hat{H}_{\psi i}$. In particular, σ_e^2, the unconditional variance of e, has

$$\hat{\sigma}_e^2 = N^{-1}\Sigma_{i=1}^N 2(v_i + x_i^T\hat{\beta})[1 - y_i - I(v_i + x_i^T\hat{\beta})]/\hat{f}(v_i|x_i,z_i) \tag{5.7}$$

Given this estimator we can, if desired, let $\tilde{\alpha} = 1/\sigma_e$ and $\tilde{\beta} = \beta/\sigma_e$ to rewrite the model as $y_i = I(\tilde{\alpha}v_i + x_i^T\tilde{\beta} + \tilde{e}_i > 0)$ where, as in the standard probit form, $\tilde{e}_i$ has unconditional mean zero and variance one.

6 Extensions

6.1 Ordered Choice

Ordered response models are summarized in, e.g., Maddala (1983), pp 46-49. The ordered choice model with K choices is defined as

$$y_i = \Sigma_{k=0}^{K-1} kI(\alpha_k < v_i + x_i{}^T\beta + e_i \leq \alpha_{k+1}) \tag{6.1}$$

where $\alpha_0 = -\infty$ and $\alpha_K = \infty$. The choices are $y_i = 0, 1, 2, \ldots K-1$, and $y_i = k$ is chosen if the latent variable $v_i + x_i{}^T\beta + e_i$ lies between α_k and α_{k+1}. Let $x_{1i} = 1$ (the

constant) and let $\beta_1 = 0$. This assumption is made without loss of generality, since if $\beta_1 \neq 0$ then each α_k could be redefined as $\alpha_k - \beta_1$.

Let $y_{ki} = I(y_i \geq k)$ for $k = 1, \ldots, K-1$, and let Δ_j equal the j'th row of Δ from equation (3.6). Then

$$(6.2) \qquad \alpha_k = -\Delta_1 E\left(z\frac{y_k - I(v > 0)}{f(v|z)}\right) \qquad k = 1, \ldots, K-1$$

$$(6.3) \qquad \beta_j = \Delta_j E\left(z\frac{[\Sigma_{k=1}^{K-1} y_k/(K-1)] - I(v > 0)}{f(v|z)}\right) \qquad j = 2, \ldots, J$$

To prove this result let η_k equal the expectation in (6.2). Now $y_{ki} = I(v_i - \alpha_k x_{1i} + \Sigma_{j=2}^{J}\beta_j x_{ji} + e_i > 0)$ is in the form of equation (1.1), so by Theorem 1' equation (6.2) holds and $\beta_j = \Delta_j \eta_k$ for $j = 2, \ldots, J, k = 1, \ldots, K-1$. It then follows that $\beta_j = \Delta_j \Sigma_{k=1}^{K-1}\eta_k/(K-1)$, which equals equation (6.3).

Equations (6.2) and (6.3) are in the form of the two stage least squares estimator, and so have the corresponding root N consistent, asymptotically normal distribution. In particular, each $\widehat{\alpha}_k$ equals $-\widehat{\beta}_1$ in equations (4.1) to (4.12), replacing y_i with y_{k_i} in equations (4.1) and (4.6), and each $\widehat{\beta}_j$ based on (6.3) equals the corresponding $\widehat{\beta}_j$ in (4.12), replacing y_i with $\Sigma_{k=1}^{K-1} y_k/(K-1)$ in equations (4.1) to (4.6).

6.2 Multinomial Choice

The estimator given by equations (3.10) to (3.13) is here extended to multinomial choice model estimation. Multinomial (also called polychotomous) response model estimators are discussed in, e.g., Maddala (1983) and McFadden (1984). A convenient feature of the multinomial choice estimator described below is that it requires no numerical integrations, simulations, or searches.

Let the latent variable w_{ki}^*, sometimes interpreted as the utility associated with choice k, be given by $w_{ki}^* = \alpha v_{ki}^* + \beta_k^{*T} x_i + e_{ki}^*$ for $k = 0, \ldots K$. It is assumed that the special regressors v_{ki}^* are all distinct, so v_{ki}^* does not equal a deterministic function of the other v_{mi}^*'s, and that the special regressors have a joint continuous distribution, conditional on x_i and a vector of instruments z_i. It is also assumed here that α, the coefficient of v_{ki}^*, is negative and is the same for every choice k. The coefficient α is then without loss of generality normalized to equal -1. This change in normalization simplifies the multinomial estimator.

Define $v_{ki} = v_{ki}^* - v_{0i}^*$, $\beta_k = (\beta_k^* - \beta_0^*)$, $e_{ki} = (e_{ki}^* - e_{0i}^*)$, and

$$(6.4) \qquad w_{ki} = -v_{ki} + x_i^T \beta_k + e_{ki} \qquad k = 0, \ldots K.$$

Equation (6.4) makes $w_{0i} = 0$. The multinomial choice model with $K+1$ choices is defined as $y_{ki} = I(w_{ki}^* > \max_{m \neq k} w_{mi}^*)$, which is equivalent to

$$(6.5) \qquad y_{ki} = I(w_{ki} > \max_{m \neq k} w_{mi}) \qquad k = 0, \ldots, K.$$

An estimator of $\beta_k = (\beta_k^* - \beta_0^*)$ for $k = 1, \ldots, K$ will be provided, using data on y_{0i}, x_i and (if necessary) a vector of instruments z_i. The estimator can be repeated K additional times, renumbering the choices so that each time a different choice is designated by 0, to yield estimates of $\beta_k^* - \beta_m^*$ for all pairs of choices k and m. Let $s_{ki} = x_i^T \beta_k + e_{ki}$, and let e_i, s_i and v_i be the K vectors of elements e_{ki}, s_{ki}, and v_{ki}, for $k = 1, \ldots, K$. Assume as before that $F_{ex}(e, x|v, z) = F_{ex}(e, x|z)$ and $E(ze) = 0$, where F_{ex} is now the conditional distribution of the vector of errors e and the vector of regressors x. It follows that $F_s(s|v, z) = F_s(s|z)$ where F_s is the conditional distribution of the vector s, and

$$(6.6) \qquad E(y_0|v, z) = E[I(s_k < v_k, k = 1, \ldots, K)|v, z] = F_s(v|z)$$

Assume that e, and hence s, has a continuous conditional distribution. Let $f_s(s|z)$ denote the conditional probability density function of s. Let Ω_v denote the support of v, a compact subset $\mathbb{R}^K$, and assume that the support of s is a compact subset of Ω_v. Then, using (6.6) and paralleling the derivation of (3.10 to 3.12),

$$\begin{aligned} (6.7) \qquad E(x|z)^T \beta_k + E(e_k|z) &= E(s_k|z) = \int_{\Omega_v} v_k f_s(v|z) dv \\ &= E[v_k f_s(v|z)/f(v|z) \mid z] = E[v_k \frac{\partial^K E(y_0|v, z)}{\partial v_1 \partial v_2 \ldots \partial v_K} / f(v|z) \mid z] \\ (6.8) \qquad \beta_k &= \Delta E[z v_k \frac{\partial^K E(y_0|v, z)}{\partial v_1 \partial v_2 \ldots \partial v_K} / f(v|z)] \qquad k = 1, \ldots, K \end{aligned}$$

Equation (6.8) shows the identification of β_k for $k = 1, \ldots, K$. Letting $\hat{G}^*(v, z)$ be a nonparametric regression of y_{0i} on v_i and z_i, and $\hat{f}(v|z)$ be a kernel estimator of the conditional density $f(v|z)$, the estimator corresponding to equation (6.8) is

$$(6.9) \qquad \hat{\beta}_k = \hat{\Delta} N^{-1} \Sigma_{i=1}^N z_i v_{ki} \frac{\partial^K \hat{G}(v_i, z_i)}{\partial v_{1i} \partial v_{2i} \ldots \partial v_{Ki}} / \hat{f}(v|z) \qquad k = 1, \ldots, K$$

For $K = 1$, equation (6.8) reduces (after sign changes) to equation (3.12). We could apply integration by parts K times to the integral in (6.7) in an attempt to simplify the estimator analogous to obtaining Theorem 1' from equation (3.11). Unfortunately, applying integration by parts to the derivatives of $E(y_0|v, z)$ with respect to v_m for each $m \neq k$ yields expressions involving $E(y_0|v, z)$ evaluated at the boundary of the support of v_m, and hence the corresponding estimator would require evaluating a nonparametric regression at the boundary of the support of the data. A similar transformation is required to derive the influence function for equation (6.9) (see, e.g., Stoker 1991 for an analogous calculation), and hence equation (6.9), while consistent, will in general converge at a slower than root N rate.

6.3 Partly Linear Latent Variable Choice Models

Theorem 1' can be extended to estimate $m(x)$ in the model

$$(6.10) \qquad y_i = I[v_i + m(x_i) + e_i > 0]$$

where $m(x)$ is an unknown smooth function. An example is where y_i is a purchase decision, $-v_i$ is a log price (or a willingness to pay bid), and $m(x_i) + e_i$ is agent i's log reservation price (or willingness to pay) function. Let Assumptions A.1, A.2, and A.3 hold with equation (6.10) instead of equation (1.1), and assume $E(e|x) = 0$. The proof of Theorem 1 will then yield $E(y^*|x) = m(x) + E(e|x) = m(x)$, so a consistent estimator of $m(x)$ is any nonparametric regression (e.g., a kernel regression) of $y - I(v > 0)/\widehat{f}(v|x)$ on x.

Under weaker conditions regarding the error, one can estimate the vector β in models of the partly linear form

$$y_i = I[v_i + \varphi(x, \beta) + e_i > 0] \tag{6.11}$$

where $\varphi(x, \beta)$ is a known function of the unknown parameter vector β. Let Assumptions A.1', A.2', A.3' and A.4' hold with equation (6.11) instead of equation (1.1). The proof of Theorem 1' will then yield $E(\tilde{y}^*|z) = E[\varphi(x, \beta) + e|z]$, so

$$E\left(z[\frac{y - I(v > 0)}{f(v|z)} - \varphi(x, \beta)]\right) = 0 \tag{6.12}$$

Therefore, β can be consistenly estimated if it can be identified from the moments (6.12). Given the limiting distributions derived in section 4, the resulting estimator will have the same limiting distribution as GMM using the moment conditions $E\,(z[q - \varphi(x, \beta)]) = 0$, where q is defined in equation (4.9).

6.4 Threshold and Censored Regression Models

For any scalar κ, define $y_i(\kappa)$ by

$$y_i(\kappa) = I(v_i\alpha + {x_i}^T\beta + e_i > \kappa) \tag{6.13}$$

so $y_i(\kappa)$ is one when the latent variable exceeds the threshold κ, and zero otherwise. Continue to assume that e_i satisfies either (1.2) and (1.3) or (1.4) and (1.5). The binary choice model (1.1) is equivalent to this model when only $y_i = y_i(0)$ is observed. Assume now that $y_i(\kappa)$ is observed for at least two values of κ. Without loss of generality, let the observed data be $y_i(0)$ and $y_i(\kappa)$ for some known $\kappa > 0$. This will be sufficient to identify both β and the scale parameter α.

Let $x_{1i} = 1$ for all i, so β_1 is the constant term in the latent variable. Maintaining the assumption $\alpha > 0$, (6.13) can be written as

$$y_i(\kappa) = I[v_i + {x_i}^T\beta(\kappa) + e_i(\kappa) > 0] \tag{6.14}$$

where $\beta_1(\kappa) = (\beta_1 - \kappa)/\alpha$, $\beta_j(\kappa) = \beta_j/\alpha$ for $j = 2, \ldots, J$, and $e_i(\kappa) = e_i/\alpha$. Since equation (6.14) is in the form of equation (1.1), for each observed threshold κ we can apply

Theorem 1' to estimate the vector $\beta(\kappa)$ by linearly regressing $y^*(\kappa)$ on x using instruments z, where

$$(6.15) \qquad y_i^*(\kappa) = [y_i(\kappa) - I(v_i > 0)]/f(v_i|u_i)$$

The limiting root N distribution of the resulting $\widehat{\beta}(\kappa)$ is given by equations (4.1) to (4.12) with $y_i(\kappa)$ in place of y_i. Given $\widehat{\beta}(0)$ and $\widehat{\beta}(\kappa)$ for some $\kappa \neq 0$, root N consistent, asymptotically normal estimates of the original α and β in equation (6.13) are then given by $\widehat{\alpha} = \kappa/[\widehat{\beta}_1(0) - \widehat{\beta}_1(\kappa)]$ and, for all κ, $\widehat{\beta} = \widehat{\beta}(\kappa)\widehat{\alpha}$. If $y_i(\kappa)$ is observed for more values of κ, then additional estimates of α and β are obtained, and these estimates could be efficiently combined using a minimum chi squared procedure.

This same estimator can also be used for censored regression models. Suppose $\widetilde{w}_i$ is observed, where

$$(6.16) \qquad \widetilde{w}_i = \max(v_i\alpha + x_i{}^T\beta + e_i, 0)$$

This is the standard fixed censoring point censored regression model. In particular, if e were normal and independent of v and x then this would be a tobit model. Assume that the distribution of e_i is unknown, but satisfies (1.2) and (1.3), or (1.4) and (1.5). Then for any $\kappa \geq 0$, we can construct $y_i(\kappa) = I(w_i > \kappa)$, and so the threshold model estimator of the previous paragraph can be used to estimate α and β.

Some more efficient, related estimators of censored and truncated regression models based on weighting by f are given in Khan and Lewbel (1999).

7 A Simulation Study

7.1 Kernel and Bandwidth Selection

When equations (B.1), (B.2) and (B.3) are used to estimate $f(v|u)$, a kernel and bandwidth must be selected. While Theorem 3 in the appendix calls for higher order kernels to eliminate asymptotic bias, it is commonly recognized that the greater variance of high order kernels makes them inferior to low order kernels unless the sample size is very large. For equation (B.1), the simulations here use the quartic product kernel

$$(7.1) \qquad K_c(c) = \Pi_{\ell=1}^{k} .9375 I(|c_\ell/\sigma_\ell| < 1)[1 - (c_\ell/\sigma_\ell)^2]^2/\sigma_\ell$$

where σ_ℓ is the estimated standard deviation of the component c_ℓ. The analogous quartic product kernel is also used for K_{vc} in (B.2). This kernel is chosen because it is computationally quick, it is generally well behaved (see Härdle 1990), and is optimal in a related kernel averaging context (see Härdle, Hart, Marron, and Tsybakov 1992).

Ordinary cross validation methods are not suitable for choosing the bandwidth because, typical of root N semiparametrics, Theorem 3 requires undersmoothing relative to optimal pointwise convergence. Härdle, Hart, Marron, and Tsybakov (1992) and Powell and Stoker (1996) calculate optimal bandwidths for some root N estimators of density functionals. While it is probably possible to adapt methods similar to theirs to the present context, a simpler procedure was used here.

If y_i equaled $I(v_i + \delta > 0)$ for any scalar constant δ, then by Theorem 1

$$(7.2) \qquad \delta = E(\frac{I(v > -\delta) - I(v > 0)}{f(v|u)}$$

So, analogous to $\hat{\eta}$ in (4.6), define $\hat{\delta}$ by

$$(7.3) \qquad \hat{\delta} = N^{-1}\Sigma_{i=1}^{N}[I(v_i > -\delta) - I(v_i > 0)]/\hat{f}(v_i|u_i)$$

The functional in equation (7.3) is very similar to the one defining $\hat{\beta}$, and by Theorem 3 the rate at which the bandwidth must shrink for root N convergence of $\hat{\delta}$ to δ is comparable to that required for $\hat{\beta}$. This suggests choosing the bandwidth that maximizes the accuracy of $\hat{\delta}$.

This procedure is used in the Monte Carlo's. In each replication, the following steps are performed. First, $\hat{f}$ in equation (B.3) is estimated using each of the bandwidths .5, 1, 1.5, 2, 2.5, 3, 3.5, and 4, resulting in eight different estimates of $\hat{f}$. Next, letting δ equal twice the estimated standard deviation of v, $\hat{\delta}$ in equation (7.3) is estimated using each of the eight estimates of $\hat{f}$, and $(\hat{\delta} - \delta)^2$ is calculated for each. Among these eight, the $\hat{f}$ that minimizes $(\hat{\delta} - \delta)^2$ is then used for that replication to estimate $\hat{\beta}$ and its asymptotic variance.

No optimality properties for this method of bandwidth selection are claimed here. It is proposed only as a simple, reasonable procedure that seems to work moderately well in the Monte Carlo's.

7.2 Monte Carlo Analysis

The simulated model is

$$(7.4) \qquad y_i = I(v_i + \beta_1 + \beta_2 x_{2i} + e_i > 0)$$

for scalar random variables y_i, v_i, x_{2i}, and e_i ($x_{1i} = 1$ is the constant). Let ϵ_{i1}, ϵ_{i2}, ϵ_{i3}, and ϵ_{i4} be independent random variables, each having mean zero and variance one, where ϵ_{i1} is uniformly distributed, ϵ_{i2} and ϵ_{i3} are standard normals, and ϵ_{i4} is a mixture of normals defined below.

The "clean probit" design defines regressors, errors, and instruments by

$$(7.5) \qquad x_{2i} = \epsilon_{i1},\ v_i = 2\epsilon_{i2},\ e_i = \epsilon_{i3},\ u_i = x_{2i},\ z_i = (1, u_i)^T$$

making v_i and e_i be independent normals, and x_{2i} be an independent uniformly distributed regressor. In this clean design $z_i = x_i$, so the two stage least squares estimator reduces to ordinary least squares.

Let ϵ_{i4} be $N(-.3, .91)$ with probability .75 and $N(.9, .19)$ with probability .25. This mixture of normals is designed to yield a distribution that is both skewed and bimodal, but

still has mean zero and variance one. The "messy" design defines regressors, errors, and instruments by

$$x_{2i} = \epsilon_{i1} + \epsilon_{i4},\ \upsilon_i = 2\epsilon_{i2} + \epsilon_{i4},\ e_i = \epsilon_{i1} + \epsilon_{i3},\ u_i = \epsilon_{i4},\ z_i = (1, u_i)^T \tag{7.6}$$

The result is that this messy design has regressors and instruments that are all skewed, multimodal, and cross correlated with each other, and further has errors e_i that are correlated with x_i with correlation .5.

The true values of the coefficients are $\beta_1 = \beta_2 = 1$. The models are estimated with sample size $N = 100$. The number of Monte Carlo replications is 10,000. The results are in Tables 1 and 2.

The two rows in each block of numbers in the tables correspond to the estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. The summary statistics reported for the Monte Carlo distribution of each $\hat{\beta}_j$ are the mean (MEAN), the standard deviation (SD), the lower quartile (LQ), the median (M), the upper quartile (UQ), the root mean squared error (RMSE) the mean absolute error (MAE), and the median absolute error (MDAE).

The replications include calculation of the estimated standard errors of the β's, as described after equation (4.12). The mean of these estimated standard error estimates (MESE) is reported, as is the percent of replications in which each $\hat{\beta}_j$ is within two estimated standard errors of the true β_j (%2SE).

Four different estimators are compared. The first is maximum likelihood probit. This is the efficient estimator in the clean probit design (Table 1). This ML probit is inconsistent in the messy design, but is still provided there to serve as a benchmark.

The second estimator is the proposed semiparametric estimator in equations (4.3) to (4.8), but using the true conditional density $f(\upsilon|u)$ that the υ_i observations are drawn from.

The third estimator is the same as the second, except that it uses the kernel estimated density $\hat{f}(\upsilon|u)$ instead of the true $f(\upsilon|u)$. There are two sources of error in $\hat{\beta}$. One source is the least squares estimation error from regressing y_i^* on x_i, and the other source of error is the use of the estimated conditional density function $\hat{f}$ in place of the true density f in constructing $\widehat{y}_i^*$. The second and third blocks of numbers in the tables can be compared to assess the separate contributions of these sources of error. The estimated density $\hat{f}$ in the third estimator is given by equations (B.1), (B.2), and (B.3) with d_i empty and $c_i = u_i$. The kernels and bandwidths are chosen as described in the previous section. Although υ_i and e_i have supports equal to the whole real line, no trimming is done, so $\tau_N = 0$ in equation (B.3).

The fourth reported estimator is the simple ordered data estimator, equation (4.16).

As would be hoped, the semiparametric estimator using the true f appears close to mean unbiased in both the clean and messy designs. The semiparametric estimator using $\hat{f}$ has a 12% to 14% mean and median bias in the clean design, and a more substantial 20% to 57% bias in the messy design. For comparison, ML probit appears unbiased in the clean design, but is much worse than the semiparametric estimator in the messy design (where it is inconsistent), having a 46% to 91% bias. The observation that the semiparametric estimator using the true f is virtually mean unbiased in both designs implies that the estimator

will be unbiased even in small samples when the density f is known, and should be close to unbiased whenever samples are large enough to accurately estimate f.

The simple ordered data estimator performed comparably to the kernel based estimator, having greater variance and less mean bias, leading overall to a modestly higher root mean squared error.

Variances and other measures of error magnitudes are similar across the estimators, with differences being primarily due to the above mean and median biases. The exceptions are that the root mean squared errors of ML are about 30% smaller than the semiparametric estimator in the clean probit design (where ML is efficient), and the variance of $\hat{\beta}$ is quite large using the true f in the messy design. This latter result is likely due to a few outliers arising from very tiny f values, and might be mitigated with trimming. A similar result is found in Lewbel (1997). It is also possible that the asymptotic variance using $\widehat{f}$ is smaller than the asymptotic variance using the true f. This would be consistent with the density f being ancillary to $\hat{\beta}$.

In applications, confidence intervals and test statistics are often as important as point estimates, yet many Monte Carlo analyses fail to evaluate the quality of standard error estimates. Standard errors are estimated here exactly as they would be with real data. In each replication, q_i is estimated by replacing y_i^* with $\widehat{y}_i^*$ in equation (4.9) and replacing each conditional expectation in that equation with a kernel regression. Standard errors are then given by the square root of the diagonal of $\hat{\Delta}\ \hat{var}(\hat{q}_i - z_i x_i^T \hat{\beta})\hat{\Delta}^T]/N$. For the probit model, White corrected standard errors are reported. In keeping with the spirit of the simple ordered data estimator, for that estimator ordinary two stage least squares White corrected standard errors (which ignore estimation error in the construction of $\widehat{\widehat{y}}_i^*$) are reported.

Standard errors are of course intended to estimate the standard deviations of $\hat{\beta}$. In almost all of the estimators, the mean estimated standard errors of the estimators are close to the Monte Carlo standard deviations, as desired. The exception is the estimator using the true f in the messy design, which as discussed above suffers from outliers. The coverage probabilities of ± 2 standard error confidence intervals are close to (but a little smaller than) 95%, except for $\hat{\beta}_2$ in the messy design which suffered from substantial mean bias.

As discussed earlier, based on the support assumptions it is expected that the larger the observed spread of v, the better the performance of the estimator. This is demonstrated in Table 3. In Tables 1 and 2, v was constructed to have the same standard deviation as $x^T\beta + e$. The design in Table 3 is identical to the messy design in Table 2, except that each v was doubled, so in Table 3 the standard deviation of v is twice the standard deviation of $x^T\beta + e$. Comparing Tables 2 and 3 shows that this increased spread in v reduces the mean and median bias considerably.

These results are encouraging. Limited experiments with other sample sizes indicate that the mean and median biases, where present, do diminish as N increases, though rather slowly.

8 Conclusions

This paper used instruments (or regressors) uncorrelated with latent variable errors and a "special regressor" to obtain semiparametric identification of discrete choice model coefficients in the presence of conditional heteroscedasticity of unknown form, and for some endogeneous and mismeasured regressor models. Simple root N consistent, asymptotically normal estimators for the binary choice model were constructed based on this identification. The estimators closely resemble ordinary least squares and two stage least squares regressions. A Monte Carlo analysis shows that the estimators work reasonably well.

The Monte Carlo also indicates that the asymptotic standard error estimates perform well. It should be possible to construct bootstrap confidence intervals instead (though the consistency of doing so is not proved here). The estimator requires at most order N^2 calculations, so bootstrapping would be computationally practical.

Successful application of the estimator hinges crucially on appropriate selection and construction of the special regressor v. Theoretical and simulation results suggest some guidelines, for example, the regressor v should be demeaned or otherwise centered close to zero, $var(v|x)$ should be large relative to $var(x^T\beta)$, and v should have a strong positive effect on y.

Further study is called for on ways to improve the finite sample properties of the estimator. For example, accuracy might be improved by replacing v and $x^T\beta$ with $v-\kappa$ and $x^T\beta+\kappa$ for one or more values of κ, or more generally replacing v and $x^T\beta$ with $v-x^T\theta$ and $x^T(\beta+\theta)$ for some vector θ. Different values of κ or θ will result in different numbers of observations in which $y_i - I(v_i > 0)$, and hence y_i^*, is nonzero. Possibilities might include letting κ be the sample mean or median of v, or letting θ be the coefficients from regressing v on x, analogous to the ordered data estimator. Note that κ or θ would need to be chosen in some way that does not violate the assumed conditional independence of e with the transformed v.

It would also be useful to study the more complicated estimators based on equations (3.9), (3.12), and (4.16), to see if they perform better in finite samples. Estimation accuracy might also be improved by better bandwidth choice procedures.

The estimators in this paper can be applied and extended in a variety of ways. Honoré and Lewbel (1999) extend them to derive binary choice panel estimators with weakly exogeneous regressors. Khan and Lewbel (1999) construct related estimators for censored regression models. Lewbel and McFadden (1997) apply the present paper's estimator as the first stage in the estimation of features of the distribution of the latent variable in applications where v is determined by experimental design.

Efficiency of this paper's estimator might be increased by use of some weighted least squares procedure, or by estimating β and the density of v jointly instead of in two stages. Lewbel and McFadden (1997) report comparisons of the ordinary least squares ($z = x$) estimator derived here to Klein and Spady (1993). They find that this paper's least squares estimator is very inefficient relative to Klein and Spady when the errors are independent of the regressors (i.e., when Klein and Spady is consistent and semiparametrically efficient.).

It is not known what the efficiency bound is under the present paper's assumptions, where general forms of heteroscedasticity are permitted.

9 Appendix A: Proofs

Proof of Theorem 1': Let $s = s(x,e) = -x^T\beta - e$. Then

$$\begin{aligned}
E(\tilde{y}^* \mid z) &= E\left(\frac{E[y - I(v>0) \mid v,z]}{f(v \mid z)} \mid z\right) \\
&= \int_L^K \frac{E[y - I(v>0) \mid v,z]}{f(v \mid z)} f(v \mid z)dv \\
&= \int_L^K E[I(v + x^T\beta + e > 0) - I(v>0) \mid v,z]dv \\
&= \int_L^K \int_{\Omega_{ex}} I(v + x^T\beta + e > 0) - I(v>0)dF_{ex}(e,x \mid v,z)dv \\
&= \int_{\Omega_{ex}} \int_L^K I(v > s) - I(v>0)dv\, dF_{ex}(e,x \mid z) \\
&= \int_{\Omega_{ex}} \int_L^K I(s \le v < 0)I(s \le 0) - I(0 < v \le s)I(s>0)dv\, dF_{ex}(e,x \mid z) \\
&= \int_{\Omega_{ex}} \left(I(s \le 0)\int_s^0 1dv - I(s>0)\int_0^s 1dv\right) dF_{ex}(e,x \mid z) \\
&= \int_{\Omega_{ex}} -s\, dF_{ex}(e,x \mid z) = \int_{\Omega_{ex}} (x^T\beta + e)dF_{ex}(e,x \mid z) = E(x \mid z)^T\beta + E(e \mid z)
\end{aligned}$$

Equation (3.8) then follows from the law of iterated expectations.

Proof of Theorem 2: Let $F^*(e,x,z) = F_e(e|x,z) - I(e>0)$. Then

$$\begin{aligned}
E[\psi(e)|x,z] &= \int_{-\infty}^{\infty} \psi(e) f_e(e|x,z)de \\
&= \int_{-\infty}^{0} \psi(e)[\partial F_e(e|x,z)/\partial e]de + \int_0^{\infty} \psi(e)\{\partial[F_e(e|x,z) - 1]/\partial e\}de \\
&= \int_{-\infty}^{\infty} \psi(e)[\partial F^*(e,x,z)/\partial e]de = -\int_{-\infty}^{\infty} \psi'(e)F^*(e,x,z)de
\end{aligned}$$

Note that $\psi(e) = 0$ at $e = 0$, and that the last equality above is an integration by parts, which uses the assumption that $[\psi(e)F^*(e,x,z)]|_{-\infty}^{\infty} = 0$. By the above equation, the definition of F^*, and equation (5.4),

$$E[\psi(e)|x,z] = \int_{-\infty}^{\infty} \psi'(e)[G(-e - x^T\beta, x, z) + I(e>0) - 1]de$$

Next do the change of variables $v = -e - x^T\beta$ to get

$$\begin{aligned} E[\psi(e)|x,z] &= \int_{-\infty}^{\infty} \psi'(-v - x^T\beta)[G(v,x,z) + I(v + x^T\beta < 0) - 1]dv \\ &= \int_{-\infty}^{\infty} \psi'(-v - x^T\beta)[E(y|v,x,z) + I(v + x^T\beta < 0) - 1]dv \\ &= E[\psi'(-v - x^T\beta)[E(y|v,x,z) + I(v + x^T\beta < 0) - 1]/f(v|x,z) \mid x,z] \end{aligned}$$

Equation (5.5) then follows from the law of iterated expectations.

10 Appendix B: Root N Convergence

A set of regularity conditions that are sufficient for root N consistent, asymptotically normal convergence of $\hat{\eta}$, and hence of $\hat{\beta}$, is provided here as Theorem 3 below. Theorem 3 is a special case of a two step estimator with a kernel estimated nonparametric first step. The general theory of such estimators is described in Newey and McFadden (1994) and Newey (1994). The proof of Theorem 3 is omitted, because the result is a special case of the theorems provided in Appendix B of Lewbel (1998), and is little more than an application of Theorem 8.11 in Newey and McFadden (1994). See also Robinson (1988), Powell, Stock, and Stoker (1989), Lewbel (1995), and Andrews (1995).

As discussed in section 4, define u to be the vector of variables such that $f(v|u) = f(v|z)$ and no element of u equals a function of other elements of u. In theorem 3, $f(v|u)$ is used for estimating η. The vector u below is divided into a vector of continuously distributed elements c and a vector of discretely distributed elements d, to permit regressors and instruments of both types.

Theorem 3 gives the limiting distribution for an estimate of $E[w/f(v|u)]$ for arbitrary w.

ASSUMPTION B.1: Each $\omega_i = (w_i, v_i, u_i)$ is an independently, identically distributed draw from some joint data generating process, for $i = 1, \ldots, N$. Let Ω be the support of the distribution each ω_i is drawn from.

ASSUMPTION B.2: Let $u_i = (c_i, d_i)$ for some vectors c_i and d_i. The support of the distribution of c_i is a compact subset of $\mathbb{R}^k$. The support of the distribution of d is a finite number of real points. The support of the distribution of v_i is some interval $[L, K]$ on the real line $\mathbb{R}$, for some constants L and K, $-\infty \leq L < K \leq \infty$. The underlying measure ν can be written in product form as $\nu = \nu_w \times \nu_v \times \nu_c \times \nu_d$, where ν_c is Lebesgue measure on $\mathbb{R}^k$. Each c_i is drawn from an absolutely continuous distribution (with respect to a Lebesgue measure with k elements). Define $f_u(u_i)$ to be the (Radon-Nikodym) conditional density of u_i given d_i, multiplied by the marginal probability mass function of d_i, and define $f_{vu}(v_i, u_i)$ to be the (Radon-Nikodym) conditional density of (v_i, u_i) given d_i, multiplied by the marginal probability mass function of d_i.

ASSUMPTION B.3: Let $f(v|u) = f_{vu}(v,u)/f_u(u)$ and let $h_i = w_i/f(v_i|u_i)$. Assume $f(v|u)$ is bounded away from zero except possibly on the boundary of the support of v, and assume that for any positive constant τ, $\sup_{\omega\in\Omega} I(|v| \leq \tau^{-1})|w|/f(v|u)$ exists. Assume there exists a positive constant ϑ such that $w = 0$ whenever c is within a distance ϑ of the boundary of the support of c.

ASSUMPTION B.4: Let $I_{\tau i} = I(|v_i| \leq \tau^{-1})$, $\pi_u(u_i) = -E(I_{\tau i}h_i|u_i)$ and $\pi_{vu}(v_i,u_i) = -E(I_{\tau i}h_i|v_i,u_i)$. Assume $E(I_{\tau i}h_i^2|u_i)$ and $E(I_{\tau i}h_i^2|v_i,u_i)$ exist and are continuous in c_i and v_i. For some v_c and (v_v, v_c) in an open neighborhood of zero there exist some functions $m_u(c_i,d_i)$ and $m_{vu}(v_i,c_i,d_i)$ such that the following local Lipschitz conditions hold:

$||f_{vu}(v+v_v, c+v_c, d) - f_{vu}(v,c,d)|| \leq m_{vu}(v_i,c_i,d_i)||(v_v, v_c)||$

$||\pi_{vu}(v+v_v, c+v_c, d) - \pi_{vu}(v,c,d)|| \leq m_{vu}(v_i,c_i,d_i)||(v_v, v_c)||$

$||f_u(c+v_c,d) - f_u(c,d)|| \leq m_u(c_i,d_i)||v_c||$

$||\pi_u(c+v_c,d) - \pi_u(c,d)|| \leq m_u(c_i,d_i)||v_c||$

Also, $E\{[m_{vu}(v,c,d)(1+|I_\tau h|)]^2 \mid d\}$ and $E\{[m_u(c,d)(1+|I_\tau h|)]^2 \mid d\}$ exist for all d in the support of d_i.

ASSUMPTION B.5: The kernel functions $K_c(c)$ and $K_{vc}(v,c)$ have supports that are convex, possibly unbounded subsets of $\mathbb{R}^k$ and $\mathbb{R}^{k+1}$, respectively, with nonempty interiors, with the origin as an interior point. The kernel functions are bounded differentiable functions having absolutely integrable fourier transforms. $K_c(c)$ satisfies $\int K_c(c)dc = 1$, $\int cK_c(c)dc = 0$, and $K_c(c) = K_c(-c)$ for all c, and similarly for $K_{vc}(v,c)$.

ASSUMPTION B.6: The kernel $K_c(c)$ has order p, that is,

$\int c_1^{l_1}\ldots c_k^{l_k}K_c(c)dc = 0$ for $0 < l_1 + \ldots + l_k < p$,

$\int c_1^{l_1}\ldots c_k^{l_k}K_c(c)dc \neq 0$ for $l_1 + \ldots + l_k = p$ and

All partial derivatives of $f_t(c,d)$ with respect to c of order p exist, and for all $0 \leq \rho \leq p$ and all d on the support of d_i, for $l_1 + \ldots + l_k = \rho$,

$\int \pi_u(c,d)[\partial^\rho f_t(c,d)/\partial^{l_1}c_1\ldots\partial^{l_k}c_k]dc$ exists, where the integral is over the support of c. All of the conditions in this assumption also hold for K_{vc} and f_{vc}, replacing c with (v,c) everywhere above.

ASSUMPTION B.7: There exists a positive constant τ^* such that $L < -1/\tau^*, 1/\tau^* < K$, and $y - I(v > 0) = 0$ for all $|v| > 1/\tau^*$. For all N greater than some constant, τ_N satisfies $\tau_N < \tau^*$.

Define the kernel density estimators:

$$(B.1) \qquad \hat{f}_u(c,d) = (Nb_N^k)^{-1}\Sigma_{i=1}^N K_c[(c-c_i)/b_N]I(d=d_i)$$

$$(B.2) \qquad \hat{f}_{vu}(v,c,d) = (Nb_N^{k+1})^{-1}\Sigma_{i=1}^N K_{vc}[(v-v_i)/b_N,(c-c_i)/b_N]I(d=d_i)$$

$$(B.3) \qquad \hat{f}(v|u) = I(|v| \leq \tau_N^{-1})\hat{f}_{vu}(v,c,d)/\hat{f}_u(c,d)$$

where b_N is the bandwidth and τ_N is a trimming parameter.

Equations (B.1) and (B.2) construct $\hat{f}_u$ and $\hat{f}_{vu}$ separately for each value of d_i, and then averages the results. Theorem 3 below also holds if $I(d = d_i)$ in equations (B.1) and (B.2) is replaced by $K_d[(d-d_i)/b_N]$ for some kernel function K_d, which results in

smoothing data across discrete d "cells" at small sample sizes, and at large sample sizes becomes equal to (B.1) and (B.2).

With $h_i = w_i / f(v_i|u_i)$ in Assumption B.3, define q_i, η, and $\hat{\eta}$ by

$$(B.4) \qquad q_i = h_i + E(h_i|u_i) - E(h_i|v_i, u_i)$$
$$(B.5) \qquad \eta = E(h) = E(q)$$
$$(B.6) \qquad \hat{\eta} = N^{-1} \Sigma_{i=1}^{N} w_i / \hat{f}(v_i|u_i)$$

Theorem 3: Assume equations (B.1) to (B.6) and Assumptions B.1 to B.7 hold. If $N b_N^{2k+2} \to \infty$ and $N b_N^{2p} \to 0$ then $\sqrt{N}(\hat{\eta} - \eta) = [N^{-1/2} \Sigma_{i=1}^{N} q_i - E(q)] + o_p(1)$.

Theorem 3 implies that $\sqrt{N}(\hat{\eta} - \eta) \Rightarrow N[0, var(q)]$. Equations (4.9) to (4.11) follow from Theorem 3 by taking $w_i = z_i[y_i - I(v_i > 0)]$ in Assumptions B.1 and B.3, and equation (B.6).

In general, kernel plug-in estimators employ some form of trimming to deal with boundary issues that otherwise interfere with root N convergence. Assumption B.3 assumes $w_i = 0$ when c_i is a near a boundary, which automatically trims out those observations. Let I_ϑ equal zero when c is within ϑ of the boundary of the support of c, otherwise $I_\vartheta = 1$. The assumption that $w = 0$ when c is near a boundary can be satisfied by replacing instruments z with zI_ϑ, however, this can conflict with Assumption A.4, i.e., if instruments z satisfy A.4, instruments zI_ϑ might not.

This is not likely to be serious problem in practice, for a variety of reasons. By taking ϑ to be arbitrarily small, the difference between moments in assumption A.4 using zI_ϑ in place of z, and hence the biases introduced by using zI_ϑ in place of z, can be made arbitrarily small. These biases could be made asymptotically negligible by asymptotic trimming, which would send ϑ to zero as N goes to infinity. Finally, this trimming is only needed to deal with the effects on $\hat{\eta}$ of bias in the estimates of the conditional density of v resulting from observations c_i that lie in a neighborhood of the boundary of the support of c. As long as the probability of observing a c near the boundary of the support is small, the effect that such terms have on the estimate $\hat{\eta}$ will be small. On the other hand, trimming based on c will reduce variation in $x^T\beta + e$ relative to variation in v, which could improve the small sample behavior of the estimator given the earlier discussion of Assumption A.3.

Next consider trimming of v. The estimator in Theorem 3 is asymptotically equivalent to a fixed "trimming" of all observations having $|v| > 1/\tau^*$, which is inside the support of v. By Assumption B.7, $h_i = 0$ for all $|v_i| > 1/\tau^*$, so the estimand is unaffected by this trimming. Given equation (1.1), this assumption requires that $x^T\beta + e$ have bounded support, and hence rules out normal or logistic errors. This is not as serious a limitation as it might seem, since for example an e distribution that was a normal, truncated at plus and minus 100 standard deviations would not be ruled out, and would be indistinguishable from a normal at any feasible sample size.

Alternatively, from Lewbel (1997) (based on Robinson 1988), Theorem 3 will hold replacing Assumption B.7 with the following asymptotic trimming assumption B.7', which

permits both e and v to have support equal to the whole real line, but requires the distribution of v to have relatively thick tails.

ASSUMPTION B.7': For all (u, v) on their support, $E\left([|h| + E(|h|) + E(|h||u) + E(|h||u, v)]^{\nu}\right)$ exists for some $\nu > 2$. The support of c is compact. Let $f_v(v)$ be the marginal density of an observation of v and let $\zeta(v) = E[hf_v(v)|v]$. There exist constants $\delta > 0$ and $\delta^* > 0$ such that as $|v| \to \infty$, $\zeta(v) = O(|v|^{-1-\delta})$ and $E[||q||^2 f_v(v)|v) = O(|v|^{-1-\delta^*})$. $N\tau_N^{2\delta} \to 0$ as $N \to \infty$.

References

[1] AIT-SAHALIA, Y., P. J. BICKEL, AND T. M. STOKER (1997), "Goodness of Fit Tests For Regression Using Kernel Methods," Unpublished Manuscript.

[2] ANDREWS, D. W. K., (1995), "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

[3] GALLANT, A. R. AND D. W. NYCHKA (1987), "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica* 55, 363-390.

[4] HARDLE, W., J. HART, J. S. MARRON, AND A. B. TSYBAKOV, (1992) "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association*, 87, 218-226.

[5] HARDLE, W. AND J. L. HOROWITZ (1996), "Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632-1640.

[6] HONORÉ, B. E. AND A. LEWBEL (1999), "Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors," unpublished manuscript.

[7] HOROWITZ, J. L. (1992), "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505-532.

[8] HOROWITZ, J. L. (1993), "Semiparametric Estimation of a Work-Trip Mode Choice Model," *Journal of Econometrics*, 58, 49-70.

[9] HOROWITZ, J. L., (1998), "Nonparametric estimation of a generalized additive model with an unknown link function," Iowa City Manuscript.

[10] ICHIMURA, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS estimation of Single-index Models," *Journal of Econometrics*, 58, 71–120.

[11] KHAN, S. AND A. LEWBEL (1999), "Weighted and Two Stage Least Squares Estimation of Semiparametric Censored and Truncated Regressions," unpublished manuscript.

[12] KLEIN, R. AND R. H. SPADY (1993), "An efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61, 387-421

[13] LEWBEL, A. (1995), "Consistent Nonparametric Tests With An Application to Slutsky Symmetry," *Journal of Econometrics*, 67, 379–401.

[14] LEWBEL, A. (1997), "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory, 13, 32-51.*

[15] LEWBEL, A. (1998a), "Semiparametric Latent Variable Model Estimation With Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121.

[16] LEWBEL, A. AND D. L. MCFADDEN (1997), "Estimating Features of a Distribution From Binomial Data," unpublished manuscript.

[17] MADDALA, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Econometric Society Monograph No. 3, Cambridge: Cambridge University Press.

[18] MANSKI, C. (1975), "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205-228.

[19] MANSKI, C. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313-334.

[20] MCFADDEN, D. L. (1984), "Econometric Analysis of Qualitative Response Models," Handbook of Econometrics, vol. 2, ed. by Z. Griliches and M. D. Intriligator, pp. 1395-1457, Amsterdam: Elsevior.

[21] MCFADDEN, D. L. (1993), "Estimation of Social Value From Willingness-To-Pay Data," Unpublished Manuscript.

[22] NEWEY, W. K. (1985), "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," Annales de l'INSEE, n59-60, 219-237.

[23] NEWEY, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

[24] NEWEY, W. K. AND D. MCFADDEN (1994), "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[25] NEWEY, W. K. AND P. A. RUUD (1994), "Density Weighted Linear Least Squares," University of California at Berkeley working paper.

[26] POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403–1430.

[27] POWELL, J. L., AND T. M. STOKER (1996), "Optimal Bandwidth Choice For Density-Weighted Averages," Journal of Econometrics, 75, 291-316.

[28] ROBINSON, PETER M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

[29] 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.

[30] STOKER, THOMAS M. (1991), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, W. A. Barnett, J. Powell, and G. Tauchen, Eds., Cambridge University Press.

TABLE 1: CLEAN PROBIT DESIGN

	j	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE	MESE	%2SE
Probit	1	1.00	0.21	0.86	0.99	1.14	0.21	0.17	0.14	0.20	0.94
ML	2	1.01	0.22	0.86	1.00	1.15	0.22	0.17	0.15	0.21	0.94
Semipar	1	1.00	0.28	0.81	0.99	1.17	0.28	0.22	0.18	0.27	0.94
true f	2	1.00	0.30	0.80	0.98	1.19	0.30	0.24	0.20	0.28	0.94
Semipar	1	1.13	0.27	0.95	1.13	1.31	0.30	0.24	0.20	0.29	0.94
kernel $\hat{f}$	2	1.14	0.32	0.92	1.12	1.33	0.35	0.27	0.21	0.32	0.92
Simple	1	1.00	0.30	0.80	0.98	1.19	0.30	0.24	0.19	0.34	0.97
ordered	2	1.00	0.36	0.76	0.98	1.20	0.36	0.28	0.23	0.36	0.94

The two rows in each block correspond to $\hat{\beta}_1$ and $\hat{\beta}_2$, respectively. Each block is a different estimator: probit maximum likelihood, equations (4.3) to (4.8) using the true density f, equations (4.3) to (4.8) using the kernel estimated $\hat{f}$, and the ordered data estimator (4.16). The reported summary statistics are the mean (MEAN), the standard deviation (SD), the lower quartile (LQ), the median (M), the upper quartile (UQ), the root mean squared error (RMSE) the mean absolute error (MAE), the median absolute error (MDAE), the mean estimated standard error estimate (MESE) and the percent of replications in which $\hat{\beta}_j$ is within two estimated standard errors of the true β_j (%2SE).

TABLE 2: MESSY DESIGN

	j	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE	MESE	%2SE
Probit	1	1.46	0.35	1.22	1.43	1.67	0.57	0.48	0.44	0.33	0.76
ML	2	1.91	0.39	1.64	1.85	2.13	0.99	0.91	0.85	0.37	0.24
Semipar	1	1.01	2.10	0.53	0.85	1.23	2.10	0.55	0.38	0.60	0.90
true f	2	0.99	2.64	0.38	0.73	1.20	2.64	0.70	0.49	0.69	0.80
Semipar	1	0.80	0.38	0.55	0.81	1.06	0.43	0.34	0.29	0.37	0.92
kernel $\hat{f}$	2	0.43	0.40	0.18	0.44	0.69	0.69	0.59	0.57	0.34	0.59
Simple	1	0.87	0.60	0.47	0.82	1.19	0.61	0.47	0.39	0.57	0.91
ordered	2	0.77	0.69	0.32	0.67	1.09	0.73	0.57	0.49	0.60	0.80

TABLE 3: MESSY DESIGN, V DOUBLED

	j	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE	MESE	%2SE
Probit	1	1.34	0.36	1.10	1.33	1.57	0.49	0.40	0.35	0.33	0.81
ML	2	1.73	0.32	1.51	1.71	1.93	0.80	0.73	0.71	0.31	0.33
Semipar	1	1.00	0.69	0.55	0.95	1.37	0.69	0.51	0.42	0.60	0.93
true f	2	0.97	0.87	0.42	0.85	1.38	0.87	0.63	0.50	0.68	0.88
Semipar	1	0.99	0.58	0.61	1.00	1.38	0.58	0.46	0.38	0.56	0.94
kernel $\hat{f}$	2	0.71	0.59	0.32	0.70	1.09	0.66	0.53	0.45	0.53	0.87
Simple	1	0.98	0.79	0.47	0.90	1.40	0.79	0.59	0.47	0.71	0.92
ordered	2	0.94	0.96	0.35	0.81	1.36	0.96	0.68	0.53	0.77	0.86