SEMIPARAMETRIC BINARY CHOICE PANEL DATA MODELS WITHOUT STRICTLY EXOGENEOUS REGRESSORS

Bo E. Honoré and Arthur Lewbel
Princeton University and Boston College

November 1998. Revised October 2000, September 2001

Abstract

Most previous studies of binary choice panel data models with fixed effects require strictly exogeneous regressors, and except for the logit model without lagged dependent variables, cannot provide rate root *n* parameter estimates. We assume that one of the explanatory variables is independent of the individual specific effect and of the errors of the model, conditional on the other explanatory variables. Based on Lewbel (2000a), we show how this alternative assumption can be used to identify and root-*n* consistently estimate the parameters of discrete choice panel data models with fixed effects, only requiring predetermined (as opposed to strictly exogeneous) regressors. The estimator is semiparametric in that the error distribution is not specified, and allows for general forms of heteroscedasticity.

Keywords: Panel Data, Fixed Effects, Binary Choice, Semiparametric, Latent Variable, Predetermined Regressors, Lagged Dependent Variable, Instrumental Variable.

This research was supported in part by the National Science Foundation. We are thankful to Debopam Bhattacharya, Luojia Hu, Shakeeb Khan the co-editor and three anonymous referees for providing useful comments.

Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (781)-862-3678, lewbel@bc.edu.

1 Introduction

The contribution of this paper is to provide a set of conditions for identification of the parameters of a binary choice model with individual specific effects and explanatory variables that are predetermined as opposed to strictly exogenous. The identification strategy suggests an estimator which is shown to be root–*n* consistent under appropriate regularity conditions.

Consider the model

$$y_{it} = I(v_{it} + x_{it}'\beta + \alpha_i + \epsilon_{it} > 0) \tag{1}$$

where $i = 1, 2, \ldots, n$, and $t = 1, 2, \ldots, T$. Here $I(\cdot)$ is the indicator function that equals one if $\cdot$ is true and zero otherwise, v_{it} is a regressor having a coefficient that has been normalized to equal one, x_{it} is a J vector of other regressors, β is a J vector of coefficients, α_i is an individual specific ("fixed") effect, and the distribution of the errors ϵ_{it} is unknown. The model (1) was considered by Rasch (1960) and by Andersen (1970) who showed that the parameter β can be estimated by a conditional likelihood approach provided that the errors, $\{\epsilon_{it}\}$, are independent and logistically distributed and independent of the sequence of explanatory variables $\{v_{it}, x_{it}\}$. Manski (1987) generalized this approach by showing that β can be estimated by a conditional maximum score approach as long as the sequence $\{\epsilon_{it}\}$ is stationary conditional on the sequence of explanatory variables $\{v_{it}, x_{it}\}$.

A recent paper by Honoré and Kyriazidou (2000) generalized the approaches of Rasch (1960), Andersen (1970) and Manski (1987) by considering a binary choice model with strictly exogenous explanatory variables as well as lagged dependent variables. The present paper provides an alternative to Honoré and Kyriazidou, which allows for general predetermined explanatory variables (not just lagged dependent variables) and results in a root-*n* consistent estimator, as opposed to the slower rate of Honoré and Kyriazidou's estimator. The cost is that a strong assumption is made on one of the explanatory variables v_{it}. This assumption is not used by Honoré and Kyriazidou. By permitting estimation of β in (1) at rate root–n, this assumption also allows us to overcome a result by Chamberlain (1993) who showed that even if all the explanatory variables are strictly exogenous and the distribution of ϵ_{it} in (1) is known, the logit model is the only version of (1) in which β can be estimated at rate root-n.

The main insight of this paper is to observe that a method due to Lewbel (2000a) can be used to construct a *linear* moment condition from (1). We can then combine this idea with the methods used for linear panel data models. In particular, we can allow for predetermined variables in exactly the same way as can be done in the linear model.

The key assumption used in this paper is that $\alpha_i + \epsilon_{it}$ in (1) is conditionally independent of one of the explanatory variables, v_{it}. This assumption is strong. However, given Chamberlain's result it is clear that some additional assumption is needed in order to construct estimators that are root–n consistent.[1,2] We stress that the requirement is conditional independence. This means that when the value of x_{it} (and instruments z_i) are known, additional knowledge of the one regressor v_{it} does not alter the conditional distribution of $\alpha_i + \epsilon_{it}$. This conditional independence is neither weaker or stronger than unconditional independence. It is possible for v_{it} and α_i to be independent, but still not be conditionally independent, because both may correlate with other regressors. It is also possible that v_{it} and α_i are dependent but still satisfy the required conditional independence. The assumption made here is in the same spirit as the assumption made by Hausman and Taylor (1981), but differs from theirs because their assumption is unconditional.

Whether the assumption made here is reasonable depends on the context. It will naturally arise in applications where $-v_{it}$ is some cost measure and $x'_{it}\beta + \alpha_i$ is some benefit measure, or vice versa. Adams, Berger and Sickles (1999) argue that such an assumption is appropriate in a particular linear model of bank efficiency. In labor supply or consumer demand models, where the errors and fixed effects are interpreted as unobserved ability or preference attributes, the assumption will hold if there exists explanatory variables that are assigned to individuals independently of these unobserved attributes (an example might be government benefits income). Maurin (1999) applies a similar conditional independence as-

[1]In a recent paper, Lee (1999) proposed an estimator based on a different set of assumptions. The advantage of the approach taken here over Lee's is that we only require predetermined regressors and that our assumption is easier to interpret (see Abreveya, 1999, for a discussion of Lee's assumptions).

[2]In some situations it may be more appropriate to take a random effects approach like the one in Chen, Heckman and Vytlacil (1998). Such an approach typically requires assumptions about initial conditions, and about the relationship between the individual specific effect and the explanatory variables, but these additional assumptions often lead to much more precise estimators (if they are satisfied). As pointed out by Wooldridge (2001) such an approach also leads to parameters that are more easily interpretable. Arellano and Carrasco (2000) propose methods for a different panel data discrete choice model that the one considered here. Their model is less general than ours, but their approach captures many of the desirable features of both fixed and random effects. The class of models and parameters considered by Altonji and Matzkin (2000) is in some ways more general than ours, but although endogeneity is permitted, their model cannot accomodate dynamics.

sumption in a model of whether students repeat a grade in elementary school, using date of birth as the special regressor, and Alonso, Fernandez, and Rodriguez-Póo (1999) use age as the independent regressor in a duration model application. Explanatory variables based on experimental design, as in Lewbel, Linton, and McFadden (2001), would also satisfy the assumption. On the other hand, it is clearly not a reasonable assumption in a structural model of the type considered by Heckman and MaCurdy (1980) where the fixed effect is related to all the explanatory variables by construction.

The next section demonstrates identification of β in our framework by expressing it as a function of estimable data densities and expectations. This is the main contribution of the paper. The limiting root n distribution of an estimator based on this identification is then given in the following section. To ease exposition, the results will be presented using a single pair of time periods, r and s, and a corresponding vector of instruments z_i, which will be assumed to be uncorrelated with ϵ_{it} in both periods. z_i would typically consist of predetermined regressors up to period $\min\{r, s\}$, although other instruments could be used (including time–invariant ones).

2 Identification

As discussed in the introduction, identification is obtained by treating one regressor, v_{it}, as special. Assume that the coefficient of v_{it} is positive (otherwise replace v_{it} with $-v_{it}$), and without loss of generality normalize this coefficient to equal one. An estimator of the sign of the coefficient of v_{it} will be provided later.

ASSUMPTION A.1: Equation (1) holds for $i = 1, 2, \ldots, n$, and $t = 1, 2, \ldots, T$. For $t = r$ and $t = s$ the conditional distribution of v_{it} given x_{it} and z_i is absolutely continuous with respect to a Lebesgue measure with nondegenerate Radon-Nikodym conditional density $f_t(v_{it} \mid x_{it}, z_i)$.

ASSUMPTION A.2: For each t, let $e_{it} = \alpha_i + \epsilon_{it}$. Assume e_{it} is conditionally independent of v_{it}, conditioning on x_{it} and z_i. Let $F_{et}(e_{it} \mid x_{it}, z_i)$ denote the conditional distribution of e_{it}, with support denoted by $\Omega_{et}(x_{it}, z_i)$.

ASSUMPTION A.3: For $t = r$ and $t = s$, the conditional distribution of v_{it} given x_{it} and z_i has support $[L_t, K_t]$ for some constants L_t and K_t, $-\infty \leq L_t < 0 < K_t \leq \infty$, and the support of $-x_{it}^T\beta - e_{it}$ is a subset of the interval $[L_t, K_t]$.

ASSUMPTION A.4: Let $\Sigma_{xtz} = E(x_{it}z_i')$ and $\Sigma_{zz} = E(z_i z_i')$. $E(\epsilon_{ir}z_i) =$

0 and $E(\epsilon_{is}z_i) = 0$. $E(\alpha_i z_i)$, Σ_{zz}, Σ_{xrz}, and Σ_{xsz} exist. Σ_{zz} and $(\Sigma_{xrz} - \Sigma_{xsz})\Sigma_{zz}^{-1}(\Sigma_{xrz} - \Sigma_{xsz})'$ are nonsingular.

In the special case of $\alpha_i = 0$ for all i (no fixed effects), for each time period t, these assumptions reduce to the assumptions in Lewbel (2000a), which provided an estimator for β in the corresponding cross section binary choice model. Implications of these assumptions are described at length in that paper, so the discussion below will focus on the additional implications for panels and for fixed effects.

Assumption A.1 says that y_{it} is given by the binary choice model (1) and that v_{it} is drawn from a continuous conditional distribution. Note that $v_{ir} = v_{is} = v_i$ is permitted, that is, the special regressor can be an observed attribute of individual i that does not vary by time. The assumptions allow α_i to be correlated with (and in other ways depend upon) v_{it}, x_{it} or z_i, but as discussed in the introduction, $\alpha_i + \epsilon_{it}$ and v_{it} must be independent given x_{it} and z_i. The assumptions also allow model errors ϵ_{it} to depend on x_{it} and z_i, as long as they are uncorreleted with the instruments z_i. In particular, heteroskedasticity of general form is permitted. Although assumptions are made about the data generating process of the α_i's, we still interpret the model as a "fixed" effects model because the estimator does not make use of any parametric or nonparametric model of the distribution of the α_i's, and in fact differencing will be used to eliminate the contribution of the α_i's, as is done in linear fixed effects models.

Assumption A.3 requires v_{it} to have a large support, and in particular requires that $-v_{it}$ be able to take on any value that the rest of the latent variable $x_{it}^T\beta + e_{it}$ can take on. This implies that for any values of x_{it} and z_t, there are values of v_{it} such that the (conditional) probability that $y_{it} = 1$ is arbitrarily close to 0 or 1. Standard models for the errors like logit or probit would therefore require that v_{it} have support equal to the whole real line. Of course, data and error distribution supports are rarely known in practice. The practical implication of these support assumptions is that the resulting estimator will generally perform better when the spread or variance of observations of v_{it} is large relative to the rest of the latent variable. Assumption A.3 also assumes that zero is in the support of v_{it}. This can be relaxed to assume that there exists some point κ that is known to be in the interior of the support of v_{it}. We may then without loss of generality redefine v_{it} and α_i as $v_{it} - \kappa$ and $\alpha_i + \kappa$, respectively. Finally, the support, $[L_t, K_t]$, can depend on (x_{it}, z_i).

An important feature of Assumptions A.1–3 is that they do not restrict the relationship between the variables over time. They therefore allow for arbitrary feed–back from the current value of y to future values of the explanatory variables. Allowing for this feature is a major contribution of the paper.

Assumption A.4 is identical to the conditions on the instruments z_i that are necessary to identify β from the moment conditions in a linear panel data model. They are basically the conditions on the instruments z_i required for linear two stage least squares estimation on differenced data.

Define y_{it}^* by

$$y_{it}^* = [y_{it} - I(v_{it} > 0)]/f_t(v_{it} \mid x_{it}, z_i) \tag{2}$$

Theorem 1 *If Assumptions A.1, A.2, and A.3 hold then, for* $t = r, s$,

$$E(y_{it}^* \mid x_{it}, z_i) = x_{it}^T\beta + E(\alpha_i + \epsilon_{it} \mid x_{it}, z_i) \tag{3}$$

Proof: Drop the subscripts to ease notation. Also, let $s = s(x, e) = -x^T\beta - e$. Then

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

Theorem 1 above is closely related to results in Lewbel (2000a). The differences are that Lewbel (2000a) has no t subscript, and uses slightly different assumptions so that only $f(v|z)$ is required instead of $f(v|x, z)$ in the definition of y^*. Those alternative assumptions are less plausible in the present context in which the error contains the individual specific effect $\alpha_{i.}$

Define Δ and η_t by

$$\Delta = [(\Sigma_{xrz} - \Sigma_{xsz})\Sigma_{zz}^{-1}(\Sigma_{xrz} - \Sigma_{xsz})']^{-1}(\Sigma_{xrz} - \Sigma_{xsz})\Sigma_{zz}^{-1}$$

$$\eta_t = E(z_i y_{it}^*).$$

Corollary 1: If Assumptions A.1, A.2, A.3 and A.4 hold, then $E(z_i y_{it}^*) = E(z_i x_{it}')'\beta + E(z_i \alpha_i)$ for $t = r, s$, and hence

$$\beta = \Delta(\eta_r - \eta_s)$$

Corollary 1 shows that β is identified, and can be estimated by an ordinary two stage least squares regression of $y_{ir}^* - y_{is}^*$ on $x_{ir} - x_{is}$, using instruments z_i. Alternative GMM estimators can be obtained by replacing Σ_{zz}^{-1} in the definition of Δ with any other nonsingular positive definite matrix.

As mentioned earlier, it is not necessary that v_{it} be time varying. If it is not, and if it is independent of all the other variables, then $y_{ir}^* - y_{is}^*$ simplifies to $(y_{ir} - y_{is}) / f(v_i)$.

3 Root N Estimation

For $t = r, s$, define h_{it} by

$$h_{it} = z_i y_{it}^* = z_i[y_{it} - I(v_{it} > 0)]/f_t(v_{it} \mid x_{it}, z_i) \tag{4}$$

and, given a density estimator $\hat{f}_t$, define

$$\hat{\eta}_t = N^{-1}\sum_{i=1}^{N} \hat{h}_{it} = N^{-1}\sum_{i=1}^{N} z_i[y_{it} - I(v_{it} > 0)]/\hat{f}_t(v_{it} \mid x_{it}, z_i) \tag{5}$$

One choice of conditional density estimator $\hat{f}_t(v_{it} \mid x_{it}, z_i)$ is a kernel estimator of the joint density of v_{it}, x_{it}, and z_i divided by a kernel estimator of the joint density of just x_{it} and z_i (see the Appendix for details). The estimator $\hat{\eta}_t$ is a two step estimator with a nonparametric first step. The limiting root N distribution for two step estimators of this type has been studied by many authors. See, e.g., Sherman (1994), Newey and McFadden (1994), and references therein. Based on these results, the influence function for $\hat{\eta}_t$ is given by

$$q_{it} = h_{it} + E(h_{it} \mid x_{it}, z_i) - E(h_{it} \mid v_{it}, x_{it}, z_i) \tag{6}$$

and therefore

$$\sqrt{N}(\hat{\eta}_t - \eta_t) = N^{-1/2} \sum_{i=1}^{N} \left[q_{it} - E(q_{it})\right] + o_p(1) \tag{7}$$

The Appendix provides one set of regularity conditions that are both sufficient for equations (6) and (7) to hold and are consistent with Assumptions A.1 to A.4.

Define $\hat{\Sigma}_{xtz}$, $\hat{\Sigma}_{zz}$, $\hat{\Delta}$, $\hat{\beta}$, and Q_i by

$$\begin{aligned} \hat{\Sigma}_{xtz} &= N^{-1} \sum_{i=1}^{N} x_{it} z_i', \qquad t = r, s \\ \hat{\Sigma}_{zz} &= N^{-1} \sum_{i=1}^{N} z_i z_i' \end{aligned}$$

$$\hat{\Delta} = [(\hat{\Sigma}_{xrz} - \hat{\Sigma}_{xsz})\hat{\Sigma}_{zz}^{-1}(\hat{\Sigma}_{xrz} - \hat{\Sigma}_{xsz})']^{-1}(\hat{\Sigma}_{xrz} - \hat{\Sigma}_{xsz})\hat{\Sigma}_{zz}^{-1}$$

$$\hat{\beta} = \hat{\Delta}(\hat{\eta}_r - \hat{\eta}_s) \tag{8}$$

$$Q_i = (q_{ir} - q_{is}) - z_i(x_{ir} - x_{is})'\beta$$

It follows immediately from equations (6) and (7) and Corollary 1 that

$$\sqrt{N}(\hat{\beta} - \beta) = N^{-1/2} \sum_{i=1}^{N} \Delta[Q_i - E(Q_i)] + o_p(1) \tag{9}$$

so $\Delta[Q_i - E(Q_i)]$ is the influence function for $\hat{\beta}$, and therefore

$$\sqrt{N}(\hat{\beta} - \beta) \Rightarrow N[0, \Delta var(Q_i)\Delta'] \tag{10}$$

When f is known, equation (6) simplifies to $q_{it} = h_{it} = z_i y_{it}^*$, which makes (10) simplify to ordinary two stage least squares. Otherwise, the density estimation error $\hat{f}_t - f_t$ contributes the term $E(h_{it} \mid x_{it}, z_i) - E(h_{it} \mid v_{it}, x_{it}, z_i)$ in equation (6) to the variance.

The variance of $\hat{\beta}$ can be estimated as $\hat{\Delta}\ \hat{\text{var}}(\hat{Q}_i)\hat{\Delta}'/N$, where vâr denotes the sample variance and $\hat{Q}_i$ is constructed by replacing h_{it}, h_{is}, and β with $\hat{h}_{it}$, $\hat{h}_{is}$, and $\hat{\beta}$, respectively, and replacing the conditional expectations in equation (6) with nonparametric regressions.

The estimator above is based on two time periods, r and s. It can be readily extended to include more time periods as follows. Rewrite Q_i as $Q_{rsi}(\beta)$, where the dependence of the definition of Q_i on β is made explicit, and the rs subscript denotes the pair of time periods used. Then $\hat{\beta}$ in equation (8) and its limiting distribution in equations (9) and (10) are equivalent to applying the standard generalized method of moments (GMM) estimator to the moment conditions $E[Q_{rsi}(\beta)] = 0$. The influence functions q contained in Q appropriately account for the effect of the density estimation error in the resulting limiting distribution.

With more than two time periods, one can stack the moment conditions $E[Q_{rsi}(\beta)] = 0$ for all pairs (r, s), and do standard (optimally weighted) GMM.

4 Additional Comments

Writing the binary choice model as $y_{it} = 1(\beta_0 v_{it} + \beta' x_{it} + e_{it} > 0)$ where $e_{it} = \alpha_i + \epsilon_{it}$, Theorem 1 and the associated estimator, equation (8), assume that $\beta_0 = 1$. The error e_{it} can be arbitrarily scaled, so if $\beta_0 \neq 0$, β_0 can be normalized to equal -1 or 1 without loss of generality. To confirm that β_0 is indeed 1 rather than -1, observe that by Assumption A.2, $E(y_{it} \mid v_{it}, x_{it}, z_i) = 1 - F_{et}[-(\beta_0 v_{it} + \beta' x_{it}) \mid x_{it}, z_i]$, so $\partial E(y_{it} \mid v_{it}, x_{it}, z_i)/\partial v_{it} = \beta_0 f_{et}[-(\beta_0 v_{it} + \beta' x_{it}) \mid x_{it}, z_i]$. Since densities are positive, it follows that β_0 equals the sign[3] of $\partial E(y_{it} \mid v_{it}, x_{it}, z_i)/\partial v_{it}$. Provided that $\partial E(y_{it} \mid v_{it}, x_{it}, z_i)/\partial v_{it}$ is consistently estimated, its sign converges at faster than rate root N. This estimator (or any other consistent estimator of the sign) can be used prior to estimation of β to ensure that v_{it} has the proper sign, without affecting the limiting distribution of $\widehat{\beta}$.

Equation (3) has the same structure as a linear panel data model. All the tools that are available for the linear panel data model can therefore be applied to (3) once y^* has been obtained, and the generalizations to the linear panel data model apply here as well. For example, given $\hat{\beta}$, information regarding the distribution α_i can be recovered. In particular, it follows from Corollary 1 that $N^{-1} \Sigma_{i=1}^N z_i (\hat{h}_{it} - x'_{it} \hat{\beta})$ will be a consistent estimator of $E(z_i \alpha_i)$, so the mean of α_i across individuals, and the correlation of α_i with instruments z_i, can be estimated. It is also possible to allow for a time–varying coefficient on the fixed effect or to replace α_i by a time varying individual specific effect $\alpha_{1i} + \alpha_{2i} u_{it}$ for some observed, strictly exogenous variable u_{it}. For example, u_{it} could be a macroeconomic variable such as interest rates or GDP that affects individuals differentially. On the other hand, it is also clear that the approach discussed here will suffer from

[3]Note that β_0 also equals the sign of a weighted average derivative of $E(y_{it} \mid v_{it}, x_{it}, z_i)$ with respect to v_{it}, which is usually easier to estimate than the derivative at a point. See, for example, Powell, Stock, and Stoker (1989).

many of the problems that makes estimation of linear panel data models with predetermined variables difficult. These include problems associated with many and potentially weak instruments, and an analysis similar to that in Blundell and Bond (1998) might be appropriate.

References

[1] ABREVEYA, J. (1999), "A Root–*n* Consistent Semiparametric Estimator for Related Effect Binary Response Panel Data: A Comment." Unpublished manuscript, University of Chicago.

[2] ADAMS, BERGER AND SICKLES (1999), "Semiparametric Approaches to Stochastic Panel Frontiers with Applications in the banking Industry," Forthcoming, *Journal of Business Economics and Statistics.*

[3] ALONSO, A. A., S. A. FERNÁNDEZ, AND J. RODRIGUEZ-PÓO (1999), "Semiparametric Estimation of a Duration Model, Universidad del País Vasco and Universidad de Cantabria unpublished manuscript.

[4] ALTONJI, J. AND R. MATZKIN (2001), "Panel data Estimators for Nonseperable Models with Endogenous Regressors," unpublished manuscript, Northwestern University.

[5] ANDERSEN, E., (1970), "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of the Royal Statistical Society*, Series B, 32, pp. 283-301.

[6] ARELLANO, M AND R. CARRASCO (2000), "Binary Choice Panel Data Models with Predetermined Variables" unpublished manuscript, CEMFI, Spain.

[7] BLUNDELL, R. W., AND S. BOND, (1998), "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models," *Journal of Econometrics*, 87, 115-143.

[8] CHAMBERLAIN, G., (1993) "Feedback in Panel Data Models," unpublished manuscript, Department of Economics, Harvard University. (April 1993)

[9] CHEN, X., J. HECKMAN AND E. VYTLACIL, (1998), "Semiparametric Identification and Root-N Efficient Estimation of Panel Discrete-Choice

Models with Unobserved Heterogeneity" unpublished manuscript, University of Chicago. Presented at CEME Conference: Recent Developments in Semiparametric Methods. (University of Pittsburgh, 1998)

[10] HAUSMAN, J. A. AND W. E. TAYLOR (1981): "Panel Data an Unobservable Individual Effects", *Econometrica*, 49, 1377-1398.

[11] HÄRDLE, W. AND T. M. STOKER (1989), "Investigating Smooth Multiple Regression by the Method of Average Derivatives," *Journal of the American Statistical Association*, 84, 986–995.

[12] HECKMAN, J. J. AND T. E. MACURDY, (1980), "A Life Cycle Model of Female Labour Supply." *Review of Economic Studies*, 47, pp. 47–74.

[13] HONG, Y. AND H. WHITE, (2000), "Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence" Unpublished Manuscript.

[14] HONORE B. E. AND E. KYRIAZIDOU, (1999), "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839-874.

[15] LEE, M. –J., (1999), "A Root–*n* Consistent Semiparametric Estimator for Related Effect Binary Response Panel Data", *Econometrica*, 60, 533-565.

[16] LEWBEL, A. (2000a), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145-177.

[17] LEWBEL, A. (2000b), "Asymptotic Trimming for Bounded Density Plug-in Estimators," Unpublished Manuscript

[18] LEWBEL, A., O. LINTON, AND D. L. MCFADDEN (2001), "Estimating Features of a Distribution From Binomial Data," Unpublished Manuscript.

[19] MANSKI, C. (1987), "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica*, 55, pp. 357-362

[20] MAURIN, E. (1999), "The Impact of Parental Income on Early Schooling Transitions: A Re-examination Using Data Over Three Generations," CREST-INSEE unpublished manuscript.

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

[22] 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.

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

[24] RASCH, G., (1960), *Probabilistic Models for Some Intelligence and Attainment Tests*, Denmarks Pædagogiske Institut, Copenhagen.

[25] RICE, J., (1986), "Boundary Modification for Kernel Regression," *Communications in Statistics,* 12, pp. 1215-1230.

[26] SHERMAN, R. P. (1994), "U-Processes in the Analysis of a Generalized Semiparametric Regression Estimator," *Econometric Theory*, 10, 372-395.

[27] WOOLDRIDGE, J. M. (2001), "The Initial Conditions Problem in Dynamic, Nonlinear Panel Data Models with Unobserved Heterogeneity." Unpublished manuscript, Michigan State University.

5 Appendix: Root N Convergence

A set of regularity conditions that are sufficient for root N consistent, asymptotically normal convergence of $\hat{\eta}_t$, and hence of $\hat{\beta}$, is provided here as Theorem 2 below. Theorem 2 is a special case of a two step estimator with a nonparametric first step, based on generic results in Newey (1994), Newey and McFadden (1994), and Sherman (1994), with closely related results in numerous other papers.

The regularity conditions provided below are not necessary for identification or consistency. They are merely one possible set of sufficient conditions for root N consistent convergence. Based on Newey (1994), any density estimator that is regular enough to yield root N consistent convergence of $\hat{\eta}_t$ to a normal can be expected to possess the same limiting distribution.

Theorem 2 below provides the limiting distribution for $\hat{\eta}_t$. To ease notation for this Appendix, all time subscripts are dropped and the estimand is denoted η, where

$$h_i = z_i[y_i - I(v_i > 0)]/f(v_i \mid x_i, z_i)$$

$$\eta = E(h_i)$$

The difficulty with applying generic methods like Newey and McFadden (1994) or Sherman (1994) is that those estimators require h_i to vanish on the boundary of the support of v_i, x_i, z_i to avoid boundary effects arising from density estimation (where a kernel or other estimator $\widehat{f}$ is substituted in for f). In our application, this may not hold for x_i or z_i.

We resolve this technicality by bounding f away from zero, and introduce an asymptotic trimming function that sets to zero all terms in the average having data within a distance τ of the boundary. We then let τ go to zero more slowly than the bandwidth to eliminate boundary effects in the kernel estimators, but we also let $N^{1/2}\tau \rightarrow 0$, which sends the volume of the trimmed space to zero at faster than rate root N, which in turn makes the bias from the trimming asymptotically irrelevant. Formally, this trimming requires that the support of the data be known. In practice, trimming might be accomplished by simply dropping out a few of the most extreme observations of the data, e.g., observations where the estimated density is particularly small. In related applications, Hardle and Stoker (1989) and Lewbel (2000a) find that asymptotic trimming has very little impact on estimates and is often unnecessary in practice.

Based on Rice (1986), Hong and White (2000) use jackknife boundary kernels to deal with this same problem of boundary bias for one dimensional densities. Their technique (which also requires known support) could be generalized to higher dimensions as an alternative to the trimming proposed here.

Define t_i to be the vector of variables used to define x_i and z_i, so x_i and z_i can be written as functions of t_i, but no element of t_i equals a function of other elements of t_i. For example, if $x_{1i} = z_{1i}$ and $x_{2i} = z_{1i}^2$, then z_{1i} could be one element of t_i, and x_{1i} and x_{2i} would not also be elements of t_i. By this definition $f(v_i \mid x_i, z_i) = f(v_i \mid t_i)$, and the latter is used in place of the former for estimating η in Theorem 2. The vector t 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.

ASSUMPTION B.1: Each $\omega_i = (y_i, v_i, t_i)$ 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. Let $x_i = x(t_i)$ and $z_i = z(t_i)$ for some known vector valued functions x and z.

ASSUMPTION B.2: Let $t_i = (c_i, d_i)$ for some vectors c_i and d_i. The support of the distribution of c_i is a convex, bounded, subset of $\mathbb{R}^k$ with a nonempty interior. The support of the distribution of d_i 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 finite constants L and K. The underlying measure ν can be written in product form as $\nu = \nu_y \times \nu_v \times \nu_c \times \nu_d$, where ν_c is Lebesgue measure on $\mathbb{R}^k$. c_i is drawn from an absolutely continuous distribution (with respect to a Lebesgue measure with k elements). $f_t(t_i)$ is the product of the (Radon-Nikodym) conditional density of t_i given d_i times the marginal probability mass function of d_i. $f_{vt}(v_i, t_i)$ is the product of the (Radon-Nikodym) conditional density of (v_i, t_i) given d_i times the marginal probability mass function of d_i. Let $\mathbf{\Omega}_{vc}$ and $\mathbf{\Omega}_c$ denote the supports of (v_i, c_i) and c_i, respectively. Let $f(v \mid t) = f_{vt}(v, t)/f_t(t)$.

ASSUMPTION B.3: Assume $f_{vt}(v, t)$, $\mathbf{\Omega}_{vc}$, and the support of h_i are bounded, and that $f_{vt}(v, t)$ is bounded away from zero. Let τ be a trimming parameter. Assume the $\mathbf{\Omega}_{vc}$ is known, and define the trimming function $I_\tau(v, c)$ to equal zero if (v, c) is within a distance τ of the boundary $\mathbf{\Omega}_{vc}$, otherwise, $I_\tau(v, c)$ equals one. Let $h_{\tau i} = h_i I_\tau(v, u)$. The expectations $E[h_\tau^2 f_t(c, d)^{-2} \mid c, d]$ and $E[h_\tau^2 f_{vt}(v, c, d)^{-2} \mid v, c, d]$ exist and are continuous in c and v. Let $\pi_{t\tau}(c, d) = -E(h_\tau \mid c, d)$ and $\pi_{vt\tau}(v, c, d) = -E(h_\tau \mid v, c, d)$. For some v_c and (v_v, v_c) in an open neighborhood of zero there exist some functions $m_t(c, d)$ and $m_{vt}(v, c, d)$ such that the following local Lipschitz conditions hold:

$$
\begin{aligned}
||f_{vt}(v + v_v, c + v_c, d) - f_{vt}(v, c, d)|| &\leq m_{vt}(v, c, d)||(v_v, v_c)|| \\
||\pi_{vt\tau}(v + v_v, c + v_c, d) - \pi_{vt\tau}(v, c, d)|| &\leq m_{vt}(v, c, d)||(v_v, v_c)|| \\
||f_t(c + v_c, d) - f_t(c, d)|| &\leq m_t(c, d)||v_c|| \\
||\pi_{t\tau}(c + v_c, d) - \pi_{t\tau}(c, d)|| &\leq m_t(c, d)||v_c||
\end{aligned}
$$

ASSUMPTION B.4: The following exist for all d in the support of d_i

$$
\begin{aligned}
&\sup_{\tau \geq 0, (v,c) \in \mathbf{\Omega}_{vc}} E[h_\tau^2 f_{vt}(v, c, d)^{-2} \mid v, c, d] \\
&\sup_{\tau \geq 0, c \in \mathbf{\Omega}_c} E[h_\tau^2 f_t(c, d)^{-2} \mid c, d] \\
&\sup_{\tau \geq 0} E\left[([1 + |h_\tau / f_{vt}(v, c, d)|] m_{vt}(v, c, d))^2 \mid d\right] \\
&\sup_{\tau \geq 0} E\left[([1 + |h_\tau / f_t(c, d)|] m_t(c, d))^2 \mid d\right]
\end{aligned}
$$

ASSUMPTION B.5: The kernel function $K_c(c)$ has support $\mathbb{R}^k$. $K_c(c) = 0$ for all c on the boundary of, and outside of, a convex bounded subset of $\mathbb{R}^k$. This subset has a nonempty interior and has the origin as an interior point. $K_c(c)$ is a bounded, differentiable, symmetric function, that satisfies $\int K_c(c)dc = 1$. The kernel function $K_{vc}(v, c)$ satisfies the same properties for (v, c) on the support $\mathbb{R}^{k+1}$.

ASSUMPTION B.6: The kernel $K_c(c)$ has order $p > 1$, that is, $\int c_1^{l_1} \dots c_k^{l_k} K_c(c)dc = 0$ for $0 < l_1 + \dots + l_k < p$, $\int c_1^{l_1} \dots c_k^{l_k} K_c(c)dc \neq 0$ for $l_1 + \dots + l_k = p$ 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 $l_1 + \dots + l_k = \rho$, $\sup_{\tau \geq 0} \int \pi_t(c, d)[\partial^\rho f_t(c, d)/\partial^{l_1} c_1 \dots \partial^{l_k} c_k]dc$ 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.

Define the kernel density estimators:

$$\hat{f}_t(c, d) = (Nh^k)^{-1} \sum_{i=1}^{N} K_c[(c - c_i)/h] I(d = d_i) \tag{A.1}$$

$$\hat{f}_{vt}(v, c, d) = (Nh^{k+1})^{-1} \Sigma_{i=1}^{N} K_{vc}[(v - v_i)/h, (c - c_i)/h] I(d = d_i) \tag{A.2}$$

$$\hat{f}(v \mid x, z)^{-1} = \hat{f}(v \mid t)^{-1} = \frac{\hat{f}_t(c, d) I_\tau(v, c)}{\hat{f}_{vt}(v, c, d)}$$

$$\hat{\eta} = N^{-1} \sum_{i=1}^{N} z_i [y_i - I(v_i > 0)] \hat{f}(v_i \mid x_i, z_i)^{-1}$$

Theorem 2 below also holds if $I(d = d_i)$ in equations (A.1) and (A.2) are replaced by $K_d[(d - d_i)/h]$ 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 (A.1) and (A.2). Equation (A.1) constructs $\hat{f}_t$ separately for each value of d_i and then averages the results.

Define q_i by

$$q_i = h_i + E(h_i \mid x_i, z_i) - E(h_i \mid v_i, x_i, z_i)$$

THEOREM 2: Let Assumptions B.1 to B.6. hold. Let either Assumptions B.7 or B.7' hold. Assume $Nh^{2(k+1)} \to \infty$, $Nh^{2p} \to 0$, $h/\tau \to 0$, and $N\tau^2 \to 0$. Then $\sqrt{N}(\hat{\eta} - \eta) = N^{-1/2} \sum_{i=1}^{N} \left[q_i - E(q_i)\right] + o_p(1)$.

The assumptions of Theorem 2 do not conflict with those of Theorem 1. However, boundedness of $\mathbf{\Omega}_{vc}$ and Assumption A.3 together require that the regressors v and x and the errors e all have bounded support.

Theorem 2 is proved in Lewbel (2000b), which is available on request.