# SIMPLE ENDOGENOUS BINARY CHOICE AND SELECTION PANEL MODEL ESTIMATORS

Arthur Lewbel Boston College

Revised December 2005

**Abstract**

This paper provides numerically trivial estimators for short panels of either binary choices or of linear models that suffer from confounded, missing not at random sample selection. The estimators allow for fixed effects, endogenous regressors, weakly exogenous regressors, and heterokedastic errors with unknown distribution. The estimators, which converge at rate root $n$, are based on variants of the Honoré and Lewbel (2002) panel binary choice model and Lewbel's (2005) cross section sample selection model.

*Keywords:* Panel Data, Fixed Effects, Binary Choice, Binomial Response, Sample Selection, Treatment, Semiparametric, Latent Variable, Predetermined Regressors, Lagged Dependent Variable, Endogeneity, Instrumental Variable. I'd like to thank Ed Vytlacil for helpful comments. Any errors are my own.

Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, http://www2.bc.edu/~lewbel/

# 1 Introduction

This paper starts from the same set of conditions as those used by Honoré and Lewbel (2002) 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. Their associated estimator, which converges at rate root $n$, requires high dimensional nonparametric first step. The present paper first proposes numerically simpler estimators for this model, based on parameterizing a model of one regressor. A related identification concept for cross section selection models proposed by Lewbel (2005) is then extended to panel models, and is similarly simplified by parametrically or semiparametrically modeling one regressor.

First consider the binomial response (binary choice) panel model

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

where $i = 1, 2, \ldots, n$, and $t = 1, 2, \ldots, T$. The asymptotics to be considered is $T$ fixed and $n \to \infty$. 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, $\beta$ is a $J$ vector of coefficients, $\alpha_i$ is an individual specific ("fixed") effect, and the distribution of the errors $\epsilon_{it}$ is unknown. We will speak of $\alpha_i$ as being drawn from some distribution, but it will be treated as a fixed effect in that no attempt is made to model this distribution, and $\alpha_i$ will be differenced out to estimate $\beta$.

The model (1) was considered by Rasch (1960) and by Andersen (1970) who showed that the parameter $\beta$ 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 $\beta$ 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 $\{v_{it}, x_{it}\}$. 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. Honoré and Lewbel (2002) 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 $v_{it}$. By permitting estimation of $\beta$ in (1) at rate root–$n$, this assumption also overcomes a result by Chamberlain (1993) who showed that even if all the explanatory variables are strictly exogenous and the distribution of $\epsilon_{it}$ in (1) is known, the logit model is the only version of (1) in which $\beta$ can be estimated at rate root-$n$. Honoré and Lewbel (2002) works by applying the cross section binary choice estimator of Lewbel (2000) to construct a linear moment condition from (1), and then applies standard methods used for linear panel data models to the resulting linear moment condition. In particular, this allows for predetermined and endogeneous regressors exactly as in linear models.

The key assumption 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] 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 assumption is similar to 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 assumption 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 estimator permits endogeneity or weak exogeneity of $x_{it}$. It also permits $x_{it}$ to contain lagged dependent variables such as $y_{it-1}$, however, in that case the required conditional independence of $v_{it}$ will limit the extent of permitted autocorrelation in $v_{it}$, since $y_{it-1}$ depends on $v_{it-1}$. The moment conditions that the estimator is based on can also be applied separately to each time period, so the model can be immediately extended to replace $\beta$ with $\beta_t$. This fits the model as described where $x_{it}$ includes time dumies, or time dummies interacted with $x_{it}$ regressors.

[1] 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.

The second model to be considered is a panel selection or treatment model

$$y_{it} = I(0 \leq v_{it} + M(x'_{it}, \alpha_i, \epsilon_{it}) \leq A) \tag{2}$$

$$p_{it} = (v_{it}\beta_v + x'_{it}\beta_x + c_i + e_{it})y_{it} \tag{3}$$

So $p_{it}$ is some outcome we wish to model, and $y_{it}$ indexes whether an individual $i$ is selected or treated in time $t$, or more generally indexes if $p_{it}$ is observed. Let $p^*_{it} = v_{it}\beta_v + x'_{it}\beta_x + \alpha_i + e_{it}$. If $p^*_{it}$ is observed then it equals $p_{it}$ and $y_{it} = 1$, otherwise $p_{it} = y_{it} = 0$, and in that case $p^*_{it}$ is what $p_{it}$ would have equaled if it had been observed, which could be a counterfactual. For example, in a classic wage model (Gronau 1974, Heckman 1974, 1976), $y_{it} = 1$ if individual $i$ is employed in time $t$, $p^*_{it}$ is the wage individual $i$ would get if employed in time $t$, and $p_{it}$ is the observed wage, which is zero for the unemployed. Both $p^*_{it}$ and $y_{it}$ depend on observable covariates such as measures of schooling or training, but they may also depend on common unobservables such as ability, so errors $e_{it}$ and $\epsilon_{it}$ and fixed effects $\alpha_i$ and $c_i$ can all be correlated with each other in unknown ways, meaning that the selection or treatment is confounded and nonignorable.

The goal for the selection model is estimation of $\beta = \beta_v,\ \beta_x$. The unknown $M_{it} = M(x'_{it}, c_i, \epsilon_{it})$ will not need to be parameterized or estimated. Common models of selection are special cases of equation (2) in which $A$ is infinite. In a wage model, the typical assumption is that one chooses to work if the gains in utility from working, indexed by the latent $v_{it} + M_{it}$, are sufficiently large. Examples in which $A$ is finite arise in ordered treatment or ordered selection models. For example, if an ordered choice model with latent variable $v_{it}+M_{it}$ determines an individual's years of schooling and $y_{it}$ indexes having exactly 12 years of schooling then individuals with $v_{it}+M_{it} < 0$ choose 11 or fewer years while those with $v_{it} + M_{it} > A$ choose 13 or more years. We might then be interested in modelling the returns $p_{it}$ from having just 12 years of schooling. The lower bound of zero is a free normalization.

The next section summarizes Honoré and Lewbel's theorem for identification of $\beta$ in the binary choice panel model by expressing it as a function of estimable data densities and expectations. Simple root $n$ consistent estimators are then provided. Next the selection model identification result, which is a panel extension of Lewbel (2004), is provided along with its corresponding simple estimators.

## 2 Binary Choice Identification

To ease exposition, for this section the theoretical results will be presented using a single pair of time periods, $r$ and $s$, and a corresponding vector of instruments $z_i$, which is assumed to be uncorrelated with $\epsilon_{it}$ 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). The simple estimators described in later sections will allow for additional time periods

Identification of the panel binary choice model 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.

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 nondegenerate 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 $-x_{it}'\beta - e_{it}$ 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 $E(\epsilon_{is}z_i) = 0$. $E(\alpha_i z_i)$, $\Sigma_{zz}$, $\Sigma_{xrz}$, and $\Sigma_{xsz}$ exist. $\Sigma_{zz}$ 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 (2000), which provided an estimator for $\beta$ in the corresponding cross section binary choice model. 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, $v_{it}$ can be an observed attribute of individual $i$ that does not vary by time. The assumptions allow $\alpha_i$ 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 $\epsilon_{it}$ 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 $\alpha_i$'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 $\alpha_i$'s, and differencing will be used to eliminate the contribution of the $\alpha_i$'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 $x_{it}'\beta + e_{it}$ 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. The parametric estimators described later assume $v_{it}$ has support on the whole real line.

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 $\kappa$ 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 $\alpha_i$ as $v_{it} - \kappa$ and $\alpha_i + \kappa$, respectively. More simply, it will be a good idea to demean $v_{it}$ across observations in each time period, prior to estimation. 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 feedback from the current value of $y$ to future values of the explanatory variables. .

Assumption A.4 is identical to the conditions on the instruments $z_i$ that are necessary to identify $\beta$ 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

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

**Theorem 1** *(Honoré and Lewbel 2002) If Assumptions A.1, A.2, and A.3 hold then, for* $t = r, s$,

$$E(y_{it}^* \mid x_{it}, z_i) = x_{it}'\beta + E(\alpha_i + \epsilon_{it} \mid x_{it}, z_i) \tag{5}$$

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

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

Define $\Delta$ and $\eta_t$ by

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

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

**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

$$\beta = \Delta(\mu_r - \mu_s)$$

Corollary 1 shows that $\beta$ 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 $z_i$.

# 3 Simple Estimators for Binary Choice

Honoré and Lewbel (2002) construct an estimator based on Theorem 1 by first nonparametrically estimating the conditional density $f_t(v_{it} \mid x_{it}, z_i)$ in each time period, then applying the two stage least squares estimator of Corollary 1 using estimates of $y_{it}^*$ that replace $f_t$ with its nonparametric estimate. Simpler estimators are obtained here by parameterizing the density of $v_{it}$.

ASSUMPTION A.5: For each $t$ from 2 to $T$, let Assumptions A.1, A.2, A.3 and A.4 hold for $r = t$ and $s = t-1$, and let the corresponding instrument vector $z_i$ be denoted $z_{it}$.

Given Assumption A.5, an immediate implication of Theorem 1 and Corollary 1 is that, for $t = 2, ..., T$,

$$E\left[z_{it}\left(\frac{y_{it} - I(v_{it} > 0)}{f_t(v_{it} \mid x_{it}, z_{it})} - \frac{y_{it-1} - I(v_{it-1} > 0)}{f_{t-1}(v_{it-1} \mid x_{it-1}, z_{it})} - (x_{it} - x_{it-1})'\beta\right)\right] = 0 \qquad (6)$$

Note that the instruments $z_{it}$ which are suitable as moments for data differenced between time periods $t$ and $t-1$ will in general be dated $t-1$ or earlier, so the density $f_{t-1}$ in equation (6) will not be conditioning time $t-1$ data on time $t$ variables. This also applies to the following assumption.

ASSUMPTION A.6: For each $t$ from 2 to $T$, $f_{t-1}(v_{it-1} \mid x_{it-1}, z_{it}) = f_{t-1}(v_{it-1} \mid x_{it-1}, z_{it-1})$.

Since $z_{it} \subset z_{it+1}$, we can always make Assumption A.6 hold by dropping some instruments in each time period, and thereby losing some efficiency. This assumption is made only to simplify parameterizing the density of $v_{it}$, but if it does not hold and we don't wish to sacrifice efficiency by dropping instruments to make it hold, then estimators like those proposed below can still be constructed, but they will require separate parameterizations of these two conditional distributions for each $t$.

ASSUMPTION A.7: For each $t$ from 2 to $T$, $f_t$ is finitely parameterized as $f_t(v_{it} \mid x_{it}, z_{it}, \lambda_t)$ for a vector of parameters $\lambda_t$. Let $r_t(v_{it}, x_{it}, z_{it}, \lambda_t)$ be any vector valued function having the property that $\lambda_t$ is identified from the moments

$$E[r_t(v_{it}, x_{it}, z_{it}, \lambda_t)] = 0 \tag{7}$$

In general equation (7) can hold defining $r_t$ as $r_t(v_{it}, x_{it}, z_{it}, \lambda_t) = \partial \ln f_t(v_{it} \mid x_{it}, z_{it}, \lambda_t)/\partial\lambda_t$, which makes $r_t$ be the score function associated with maximum likelihood estimation of $\lambda_t$. Given Assumption A.7, define

$$y_{it}^* = y^*(v_{it}, x_{it}, z_{it}, \lambda_t) = \frac{y_{it} - I(v_{it} > 0)}{f_t(v_{it} \mid x_{it}, z_{it}, \lambda_t)}$$

**Corollary 2**: If Assumptions A.5, A.6, and A.7 hold then the parameters $\beta$ are identified from the moment conditions

$$\begin{aligned} E\left[z_{it}\left(y^*(v_{it}, x_{it}, z_{it}, \lambda_t) - y^*(v_{it-1}, x_{it-1}, z_{it-1}, \lambda_{t-1}) - (x_{it} - x_{it-1})'\beta\right)\right] &= 0, \ \ t = 2, ..., T \\ E[r_t(v_{it}, x_{it}, z_{it}, \lambda_t)] &= 0, \ \ t = 1, ..., T \end{aligned}$$

In Corollary 2, identification of each $\lambda_t$ follows from Assumption A.7, and given $\lambda_t$ with Assumption A.6, identification of $\beta$ follows from Corollary 1. Estimation proceeds by applying GMM to the set of moments given in Corollary 2, and standard GMM limiting distribution theory applies (See, e.g., Newey (1984) or Wooldridge (2002), p. 425), assuming $T$ is fixed and $n \to \infty$. . First step, initial consistent estimates for this GMM can be obtained by applying GMM separately in each time period to equation (7) to get $\widehat{\lambda}_t$, then regressing the resulting $\widehat{y}_{it}^* - \widehat{y}_{it-1}^*$ on $x_{it} - x_{it-1}$ using linear two stage least squares with instruments $z_{it}$, for each $t \geq 2$.

As an example, suppose that we can model $v_{it}$ in terms of the other covariates as

$$v_{it} = x_{it}'\gamma_t + z_{it}'\delta_t + \sigma_t\eta_{it}, \qquad \eta_{it} \perp x_{it}, z_{it}, \alpha_i, \epsilon_{it} \tag{8}$$

where the unobserved error term $\eta_{it}$ has some known density function $f_{\eta t}$ with mean zero and variance one, e.g., a standard normal. Then $\lambda_t = \gamma_t, \delta_t, \sigma_t$ and

$$f_t(v_{it} \mid x_{it}, z_{it}, \lambda_t) = \frac{1}{\sigma_t} f_{\eta t}\left(\frac{v_{it} - x_{it}'\gamma_t - z_{it}'\delta_t}{\sigma_t}\right) \tag{9}$$

so

$$y_{it}^* = y^*(v_{it}, x_{it}, z_{it}, \gamma_t, \delta_t, \sigma_t) = \frac{(y_{it} - I(v_{it} > 0))\sigma_t}{f_{\eta t}((v_{it} - x_{it}'\gamma_t - z_{it}'\delta_t)/\sigma_t)} \tag{10}$$

and the moments in Corollary 2 are

$$\begin{aligned} E\left[z_{it}\left(y^*_{it} - y^*_{it-1} - (x_{it} - x_{it-1})'\beta\right)\right] &= 0, \;\; t = 2, ..., T \\ E(v_{it} - x'_{it}\gamma_t - z'_{it}\delta_t) &= 0, \;\; t = 1, ..., T \\ E\left[(v_{it} - x'_{it}\gamma_t - z'_{it}\delta_t)^2 - \sigma^2_t\right] &= 0, \;\; t = 1, ..., T \end{aligned} \tag{11}$$

after substituting equation (10) in for $y^*_{it}$ and $y^*_{it-1}$. In this example, consistent parameter estimates can be obtained just from a sequence of linear least squares estimates, as follows:

1. For each $t$ estimate $\gamma_t$ and $\delta_t$ as the coefficients from linearly regressing $v_{it}$ on $x_{it}$ and $z_{it}$ across observations $i$ in time period $t$.

2. For each $t$ estimate $\sigma^2_t$ as the sample average of $(v_{it} - x'_{it}\widehat{\gamma}_t - z'_{it}\widehat{\delta}_t)^2$ across observations $i$ in time period $t$.

3. For each $t$ construct $\widehat{y}^*_{it} = y^*(v_{it}, x_{it}, z_{it}, \widehat{\gamma}_t, \widehat{\delta}_t, \widehat{\sigma}_t)$ defined by equation (10).

4. For each $t > 1$ regress $\widehat{y}^*_{it} - \widehat{y}^*_{it-1}$ on $x_{it} - x_{it-1}$ using linear two stage least squares with instruments $z_{it}$, call the resulting coefficient estimates $\widehat{\beta}_t$.

5. Construct $\widehat{\beta}$ as the average of $\widehat{\beta}_t$ over $t = 2, ..., T$ Alternatively, three stage least squares could be applied by stacking the regressions in step 4, which will be equivalent to constructing $\widehat{\beta}$ as a weighted average of $\widehat{\beta}_t$, with weights chosen to minimize variance.

This numerically trivial sequence of estimators can be easily bootstrapped to obtain parameter confidence intervals or standard errors, or used as initial consistent estimates for standard GMM estimation of equations (11) to obtain efficient estimates and consistent standard errors. Since $T$ is fixed and the asymptotics have $n \to \infty$, the bootstrap could just draw individuals $i$ from the sample with replacement, thereby preserving any time series dependence in the data.

This estimator assumes the marginal density $f_{\eta t}$ of the scalar random error $v_{it}$ is known. If this error is normal, then moments used above for estimating $\gamma_t, \delta_t, \sigma_t$ correspond to maximum likelihood estimates of these parameters.

This estimator could also be implemented if the marginal density $f_{\eta t}$ is unknown, by replacing $f_{\eta t}(\widehat{\eta}_{it})$ with $\widehat{f}_{\eta t}(\widehat{\eta}_{it})$ where $\widehat{\eta}_{it} = (v_{it} - x'_{it}\widehat{\gamma}_t - z'_{it}\widehat{\delta}_t)/\widehat{\sigma}_t$ and $\widehat{f}_{\eta t}$ is a nonparametric density estimator such as a kernel estimator. A particularly simple estimator $\widehat{f}_{\eta t}$ is the ordered data estimator of Lewbel and Schennach (2003). For this estimator, after step 2 above, for each $t$ sort the observations $\widehat{\eta}_{1t}$,...,$\widehat{\eta}_{nt}$ from lowest to highest. Then, for each $i$ and $t$, let $\widehat{\eta}^+_{it}$ be the value of $\widehat{\eta}$ that, in the sorted data, comes immediately after $\widehat{\eta}_{it}$ and similarly let $\widehat{\eta}^-_{it}$ be the value that comes immediately before $\widehat{\eta}_{it}$. Then

$$f_{\eta t}(\widehat{n}_{it}) = \frac{\partial F_{\eta t}(\widehat{\eta}_{it})}{\partial \eta} \approx \frac{F_{\eta t}(\widehat{\eta}^+_{it}) - F_{\eta t}(\widehat{\eta}^-_{it})}{\widehat{\eta}^+_{it} - \widehat{\eta}^-_{it}} \approx \frac{2/n}{\widehat{\eta}^+_{it} - \widehat{\eta}^-_{it}}$$

where the last step replaces the true distribution function $F_{\eta t}$ with the empirical distribution function. This suggests the estimator $\widehat{f}_{\eta t}(\widehat{\eta}_{it}) = (\widehat{\eta}_{it}^{+} - \widehat{\eta}_{it}^{-})n/2$. Lewbel and Schennach (2003) show that, although this is not a consistent estimator of the density function $f_{\eta t}$, with sufficient regularity sample averages that divide by this estimator are root $n$ consistent. They also show that greater efficiency can be obtained by taking $\widehat{\eta}_{it}^{+}$ and $\widehat{\eta}_{it}^{-}$ to be $k$ values larger and smaller than $\widehat{\eta}_{it}$, where $k$ is larger than one as above (maximum efficiency is obtained by letting $k \rightarrow \infty$ at an arbitrarily slow rate).

In the above application, use of this $\widehat{f}_{\eta t}$ density estimator corresponds to replacing the function $y^{*}$ in step 3 with

$$\begin{aligned}
\widehat{\eta}_{it} &= (v_{it} - x_{it}'\widehat{\gamma}_{t} - z_{it}'\widehat{\delta}_{t})/\widehat{\sigma}_{t} \\
\widehat{y}_{it}^{*} &= \frac{(y_{it} - I(v_{it} > 0))2\widehat{\sigma}_{t}}{(\widehat{\eta}_{it}^{+} - \widehat{\eta}_{it}^{-})n}
\end{aligned}$$

The limiting distribution theory given in Theorem 6 of Lewbel and Schennach (2003), which explicitly allows for estimated data $\widehat{\eta}_{it}$ in the density estimator, could be applied here, though the resulting formulas are rather complicated and depend on conditional expectations that would need to be nonparametrically estimated. Some form of simulation or bootstrap would be simpler and numerically practical given that the estimator is numerically trivial and involves no numerical searches.

# 4 Selection or Treatment Models

As before, the relevant identification theorem will be derived using two time periods, $r$ and $s$. Then the estimator will be given based on all $T$ time periods.

ASSUMPTION B.1: Equations (2) and (3) hold 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 nondegenerate conditional density $f_{t}(v_{it} \mid x_{it}, z_{i})$.

ASSUMPTION B.2: For each $t$, assume $\alpha_{i}$, $\epsilon_{it}$, and $c_{i} + e_{it}$ are conditionally independent of $v_{it}$, conditioning on $x_{it}$ and $z_{i}$.

ASSUMPTION B.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 contains the supports of $-M(x_{it}', \alpha_{i}, \epsilon_{it})$ and of $A - M(x_{it}', \alpha_{i}, \epsilon_{it})$.

ASSUMPTION B.4: Let $\widetilde{x}_{it} = (v_{it}, x_{it}')'/f_{t}(v_{it} \mid x_{it}, z_{i})$, $\widetilde{\Sigma}_{xtz} = E(\widetilde{x}_{it}z_{i}')$, and $\Sigma_{zz} = E(z_{i}z_{i}')$. $E(e_{ir}z_{i}) = 0$ and $E(e_{is}z_{i}) = 0$. $E(\alpha_{i}z_{i})$, $\Sigma_{zz}$, $\widetilde{\Sigma}_{xrz}$, and $\widetilde{\Sigma}_{xsz}$ exist. $\Sigma_{zz}$ and $(\widetilde{\Sigma}_{xrz} - \widetilde{\Sigma}_{xsz})\Sigma_{zz}^{-1}(\widetilde{\Sigma}_{xrz} - \widetilde{\Sigma}_{xsz})'$ are nonsingular.

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

$$E\left(\frac{(p_{it} - v_{it}\beta_v + x'_{it}\beta_x)\, y_{it}}{f_t(v_{it} \mid x_{it}, z_{it})} \mid x_{it}, z_i\right) = E(c_i + e_{it} \mid x_{it}, z_i)A$$

Proof: Dropping subscripts for convenience,

$$\begin{aligned}
& E\left(\frac{(p - v\beta_v + x'\beta_x)\, y}{f(v \mid x, z)} \mid x, z\right) \\
= \; & E\left(\frac{(c+e)\, I(0 \leq v + M \leq A)}{f(v \mid x, z)} \mid x, z\right) \\
= \; & E\left[E\left(\frac{(c+e)\, I(0 \leq v + M \leq A)}{f(v \mid x, z)} \mid c+e, a, \epsilon, x, z\right) \mid x, z\right] \\
= \; & E\left[\int_{supp(v \mid c+e,a,\epsilon,x,z)} \frac{(c+e)\, I(0 \leq v + M \leq A)}{f(v \mid x, z)} f(v \mid c+e, a, \epsilon, x, z) dv \mid x, z\right] \\
= \; & E\left[\int_{supp(v \mid c+e,a,\epsilon,x,z)} (c+e)\, I(0 \leq v + M \leq A) dv \mid x, z\right] \\
= \; & E\left[\int_{-M}^{A-M} (c+e)\, dv \mid x, z\right] \\
= \; & E\left[(c+e)\, (A - M + M) \mid x, z\right] = E(c + e \mid x, z)A
\end{aligned}$$

Theorem 2 is essentially a special case of Theorems 1 and 2 in Lewbel (2004), extended to have both $i$ and $t$ subscripts. That paper considers general cross section GMM models instead of just $p_{it}$ linearity, and also allows $A$ to be random or infinite. The above theorem assumes $A$ is finite, but the same extensions given in Lewbel (2004) to deal with these other cases could be incorporated here. In particular, Lewbel (2004) shows that if $A$ is infinite and the maximum value that $v$ can take on is some finite value $\tau$ then an asymptotic bias term of order $O(\tau^{-1})$ is introduced. Since the support of $v$ can be arbitrarily large, this bias can be arbitrarily small. That result extends to Corollaries 3 and 4 below, so the estimators provided below can be applied without change if $A$ is infinite, at the expense of introducing an arbitrarily small but nonzero bias term. Alternatively, shrinking the bias to zero in that case would require infinite support for $v$ and asymptotic trimming.

In the case of the binary choice estimator, the $y_{it}$ model was converted to a linear model by constructing $y^*_{it}$. In the sample selection case, it is not just the dependent variable $p_{it}$ that is modified, instead both $p_{it}$ and the regressors in the model are weighted by the density of $v_{it}$ to obtain linearity. The general procedure remains the same, which is to difference after weighting.

Define $\widetilde{\Delta}$ and $\widetilde{\mu}_t$ by

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

$$\widetilde{\mu}_t = E\left(\frac{z_i p_{it}}{f_t(v_{it} \mid x_{it}, z_{it})}\right).$$

**Corollary 3**: If Assumptions B.1, B.2, B.3 and B.4 hold, then

$$\begin{pmatrix} \beta_v \\ \beta_x \end{pmatrix} = \widetilde{\Delta}(\widetilde{\mu}_r - \widetilde{\mu}_s)$$

Corollary 3 starts from taking Theorem 2, multiplying both sizes by $z_i$, differencing, and applying ordinary linear two stage least squares to the result. Density weighting the dependent variable and the regressor makes the problem equivalent to a linear panel model without a selection problem, and so it can be estimated in the usual way by differencing out the fixed effect and instrumenting. Note that we are not weighting by a propensity score (instead, the weighting is by the density of a variable that affects the propensity score), and no assumption is made about the joint distribution of errors in the model, other than conditional independence of the one regressor $v$.

ASSUMPTION B.5: For each $t$ from 2 to $T$, let Assumptions B.1, B.2, B.3 and B.4 hold for $r = t$ and $s = t - 1$, and let the corresponding instrument vector $z_i$ be denoted $z_{it}$.

Now let Assumptions A.6 and A.7 apply here, with the same comments as before regarding the dating of instruments $z_{it}$. Then, by essentially the same logic as in Corollary 2, we obtain Corollary 4. Define

$$w_{it} = w(v_{it}, x_{it}, z_{it}, \lambda_t) = \frac{y_{it}}{f_t(v_{it} \mid x_{it}, z_{it}, \lambda_t)} \tag{12}$$

**Corollary 4**: If Assumptions B.5, A.6, and A.7 hold then the parameters $\beta_v, \beta_x$ are identified from the moment conditions

$$\begin{aligned} E\left[z_{it}\left(w_{it}\left(p_{it} - v_{it}\beta_v + x'_{it}\beta_x\right) - w_{it-1}\left(p_{it-1} - v_{it-1}\beta_v + x'_{it-1}\beta_x\right)\right)\right] &= 0, \; t = 2, ..., T \\ E[r_t(v_{it}, x_{it}, z_{it}, \lambda_t)] &= 0, \; t = 1, ..., T \end{aligned}$$

where $w_{it}$ and $w_{it-1}$ are given by equation (12)

Identification of each $\lambda_t$ follows from Assumption A.7, and Corollary 3 shows that $\beta_v, \beta_x$ is identified. Estimates are obtained by applying GMM to the set of moments given in Corollary 4. Similar to Corollary 2. First step, initial consistent estimates for this GMM can be obtained by applying GMM separately in each time period to $E[r_t(v_{it}, x_{it}, z_{it}, \lambda_t)] = 0$, yielding $\widehat{\lambda}_t$, construct $\widehat{w}_{it}$ by putting this $\widehat{\lambda}_t$ into equation (12), and then regress the resulting $\widehat{w}_{it}p_{it} - \widehat{w}_{it-1}p_{it-1}$ on $\widehat{w}_{it}v_{it} - \widehat{w}_{it-1}v_{it-1}$ and $\widehat{w}_{it}x_{it} - \widehat{w}_{it-1}x_{it-1}$ using linear two stage least squares with instruments $z_{it}$, for each $t \geq 2$, to obtain $\beta_v, \beta_x$ estimates.

As before, if we model $f_t$ by equations (8) and (9) then

$$w_{it} = w(v_{it}, x_{it}, z_{it}, \gamma_t, \delta_t, \sigma_t) = \frac{y_{it}\sigma_t}{f_{\eta t}((v_{it} - x_{it}'\gamma_t - z_{it}'\delta_t)/\sigma_t)} \tag{13}$$

and the moments in Corollary 4 are

$$\begin{aligned} E\left[z_{it}\left(w_{it}\left(p_{it} - v_{it}\beta_v + x_{it}'\beta_x\right) - w_{it-1}\left(p_{it-1} - v_{it-1}\beta_v + x_{it-1}'\beta_x\right)\right)\right] &= 0, \;\; t = 2, ..., T \qquad (14) \\ E(v_{it} - x_{it}'\gamma_t - z_{it}'\delta_t) &= 0, \;\; t = 1, ..., T \\ E\left[(v_{it} - x_{it}'\gamma_t - z_{it}'\delta_t)^2 - \sigma_t^2\right] &= 0, \;\; t = 1, ..., T \end{aligned}$$

now substituting equation (13) in for $w_{it}$ and $w_{it-1}$. Once again, consistent parameter estimates can be obtained just from a sequence of linear least squares estimates, as follows:

1. For each $t$ estimate $\gamma_t$ and $\delta_t$ as the coefficients from linearly regressing $v_{it}$ on $x_{it}$ and $z_{it}$ across observations $i$ in time period $t$.

2. For each $t$ estimate $\sigma_t^2$ as the sample average of $(v_{it} - x_{it}'\widehat{\gamma}_t - z_{it}'\widehat{\delta}_t)^2$ across observations $i$ in time period $t$.

3. For each $t$ construct $\widehat{w}_{it} = w(v_{it}, x_{it}, z_{it}, \widehat{\gamma}_t, \widehat{\delta}_t, \widehat{\sigma}_t)$ defined by equation (12).

4. For each $t > 1$ regress $\widehat{w}_{it}p_{it} - \widehat{w}_{it-1}p_{it-1}$ on $\widehat{w}_{it}v_{it} - \widehat{w}_{it-1}v_{it-1}$ and $\widehat{w}_{it}x_{it} - \widehat{w}_{it-1}x_{it-1}$ using linear two stage least squares with instruments $z_{it}$, and call the resulting coefficient estimates $\widehat{\beta}_{vt}, \widehat{\beta}_{xt}$.

5. Construct $\widehat{\beta}_v, \widehat{\beta}_x$ as the average of $\widehat{\beta}_{vt}, \widehat{\beta}_{xt}$ over $t = 2, ..., T$ Alternatively, three stage least squares could be applied by stacking the regressions in step 4.

As before, this is a numerically trivial sequence of steps that can be easily bootstrapped to obtain parameter confidence intervals or standard errors, or used as initial consistent estimates for standard GMM estimation of equation (14) to obtain efficient estimates and consistent standard errors. Each bootstrap replication would consist of $n$ draws of individuals $i$ from the sample with replacement (using all the data from all time periods for each individual drawn), and repeating the above steps with the drawn data.

Also as before, if the marginal density $f_{\eta t}$ of the scalar random error $v_{it}$ is unknown, then it can be estimated using the sorted data estimator, where $\widehat{w}_{it}$ in step 3 is now constructed by

$$\begin{aligned} \widehat{\eta}_{it} &= (v_{it} - x_{it}'\widehat{\gamma}_t - z_{it}'\widehat{\delta}_t)/\widehat{\sigma}_t \\ \widehat{w}_{it} &= \frac{2y_{it}\widehat{\sigma}_t}{(\widehat{\eta}_{it}^+ - \widehat{\eta}_{it}^-)n} \end{aligned}$$

with $\widehat{\eta}_{it}^+$ and $\widehat{\eta}_{it}^-$ defined as before.

# 5 Conclusions

Numerically trivial estimators have been provided for panels of binomial response models, and panels where the dependent variable is sometimes missing not at random, that is, panels suffering from sample selection. In both cases, weighting data by the density of a single regressor $v$ converts these nonlinear models into linear models, as far as moments are concerned, and thereby allows us to remove fixed effects by differencing and lets us deal with endogenous or weakly exogenous regressors by instrumenting, just as we would in linear models. The results are root $n$ consistent, asymptotically normal estimates with limiting distributions that can be obtained by standard GMM, or by bootstrapping. The distributions of the latent binary choice errors, or of the outcome and selection model errors, are jointly unknown and do not need to be estimated. Instead, simple estimators are obtained by having a model for the single regressor $v$.

Since these estimators take the form of GMM, refinements of GMM such as weak instrument GMM, or generalized empirical likelihood to deal with small sample biases can be directly applied. In general, these estimators will suffer from the same problems that can make estimation of linear panel data models with predetermined variables difficult. These include problems associated with many and potentially weak instruments, so an analysis similar to that in Blundell and Bond (1998) might be appropriate. Also, these estimators involve dividing by a density, which can result in extreme observations when the density is small. This suggests that checking the moments for outliers, and perhaps discarding them (corresponding to robust moment estimators) may be advisable. For nonparametric density estimation this would be equivalent to asymptotic trimming, which formally is not required with parametric density estimation, but might still be advisable in small or moderate size data sets.

# 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] Gronau, R. (1974), ”Wage Comparisons - A Selectivity Bias,” *Journal of Political Economy,* 82, 1119-1144.

[11] Heckman, J. (1974), ”Shadow Prices, Market Wages, and Labor Supply,” *Econometrica*, 42, 679-693

[12] Heckman, J. (1976), ”The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models,” *Annals of Economic and Social Measurement,* 5, 475-495

[13] Heckman, J. (1976), ”Sample Selection Bias as a Specification Error,” *Econometrica*, 47, 153-161.

[14] Hausman, J. A. and W. E. Taylor (1981): “Panel Data an Unobservable Individual Effects”, *Econometrica*, 49, 1377-1398.

[15] 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.

[16] Heckman, J. J. and T. E. MaCurdy, (1980), “A Life Cycle Model of Female Labour Supply.” *Review of Economic Studies*, 47, pp. 47–74.

[17] Hong, Y. and H. White, (2000), "Asymptotic Distribution Theory for Nonparametric Entropy Measures of Serial Dependence" Unpublished Manuscript.

[18] Honore B. E. and E. Kyriazidou, (1999), "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839-874.

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

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

[21] Lewbel, A. (2004), "Endogenous Selection or Treatment Model Estimation, Unpublished Manuscript.

[22] Lewbel, A., O. Linton, and D. L. McFadden (2001), "Estimating Features of a Distribution From Binomial Data," Unpublished Manuscript.

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

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

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

[26] 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.

[27] Powell, J. L., J. H. Stock, and T. M. Stoker (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403–1430.

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

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

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

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