Seminario del Departamento de Economía

"Panel Data Estimators for Nonseparable Models with Endogenous Regressors"

Rosa Matzkin
(Northwestern University)

Martes 14 de julio de 1998
9:00 hrs
Aula Chica de Planta Baja
Panel Data Estimators for Nonseparable Models with Endogenous Regressors

Joseph Altonji and Rosa Matzkin
Department of Economics
Northwestern University

First Draft: May 1997
Revised April 1998
(preliminary and incomplete)

Abstract

1 Introduction

In this paper we develop estimators for models with nonseparable error terms and endogenous explanatory variables that can be used when the data are clustered in some way. Examples of such data are samples of siblings, panel data on a set of individuals or firms, and data on a set of individuals who grew up in the same neighborhood or attended the same high school. The methods will be applicable to a wide range of topics. Examples include the use of siblings to analyze the effects of teenage pregnancy on the probabilities of being in poverty, of being on welfare, and of ever marrying (Geronimus and Korenman (1992)), the use of children from the same high school to isolate the effects of family background on educational attainment, the use
of siblings to study the effects of neighborhood characteristics on high school graduation (Aaronson (forthcoming)), and the use of siblings to study transfers of time and money to and from parents (Altonji, Hayashi and Kotlikoff (1996), Rosenzweig and Wolpin (1994a and b). The methods cover a wide range of nonlinear models, including binary choice models.

To be more specific, consider the model (1.1)

\[
y_{ik} = m(x_{ik}, \varepsilon_i, u_{ik}), \quad i = 1 \ldots n, \quad k = 1 \ldots K_i
\]

where \(y_{ik}\) is an outcome of person \(k\) from group \(i\), \(x_{ik}\) is a \(1 \times J_i\) vector of observed variables, \(\varepsilon_i\) is an error component common to observations from group \(i\), \(u_{ik}\) is an error term that is specific to person \(k\) of group \(i\), and \(K_i\) is the number of observations in group \(i\). In some applications, the "group" might be a family. In others, it might be a neighborhood, a school, or a firm. In cross section time series data, \(i\) might refer to an individual and \(k\) to the time period. The function \(m(\cdot, \cdot, \cdot)\) may be nonseparable in \(x_{ik}, \varepsilon_i, \) and \(u_{ik}\). The index \(k\) may be an element of \(x_{ik}\), which means that the effect of \(u_{ik}\) and \(\varepsilon_i\) on \(y_{ik}\) may depend on sibling order in a family context or age or the time period in a cross section time series context.

To give a simple example, consider Aaronson's (forthcoming) analysis of the effects of neighborhood characteristics on college attendance. In this case \(i\) denotes a family and \(k\) a specific child. The outcome \(y_{ik}\) is \(1\) if person \(i, k\) started college and \(0\) otherwise and \(x_{ik}\) is average neighborhood income while person \(ik\) was between the ages of 10 and 16. (We abstract from other elements of \(x_{ik}\) such as the income of family \(i\) when \(k\) is growing up to simplify the exposition). The function \(m(x_{ik}, \varepsilon_i, u_{ik})\) takes on the value \(0\) or \(1\). The form of \(m(x_{ik}, \varepsilon_i, u_{ik})\) might be

\[
m(x_{ik}, \varepsilon_i, u_{ik}) = 1 \text{ if } \Psi(x_{ik}, \varepsilon_i, u_{ik}) = x_{ik}\Gamma + \varepsilon_i + u_{ik} > 0
\]

\[
= 0 \text{ otherwise.}
\]

or

\[
(1.1a) \quad m(x_{ik}, \varepsilon_i, u_{ik}) = I(-x_{ik}\Gamma - \varepsilon_i < u_{ik})
\]

where the function \(I(\cdot)\) is \(1\) if the inequality is true and \(0\) if it is false. (We adopt the linear form for the index function for exposition only—our methods are nonparametric and do not require that the index function \(\Psi\) be separable in \(x_{ik}, \varepsilon_i, \) and \(u_{ik}\).)
Let \( g(u_{ik}, \varepsilon_i|x_{ik}) \) be the density of \((u_{ik}, \varepsilon_i|x_{ik})\). The probability that a person with characteristics \( x_{ik} \) attends college is

\[
E(y_{ik}|x_{ik}) = \int_{\varepsilon_i}^{\infty} \int_{u_{ik}}^{\infty} [m(x_{ik}, \varepsilon_i, u_{ik})]g(u_{ik}, \varepsilon_i|x_{ik})du_{ik}d\varepsilon_i
\]

The objective of the analysis is to estimate the expected value of the partial derivative of the probability of attending college with respect to neighborhood income, holding the distribution \( g(\varepsilon_i, u_{ik}|x_{ik}) \) constant. Call this derivative \( \beta(x_{ik}) \), where

\[
\beta(x_{ik}) = \int_{\varepsilon_i}^{\infty} \Gamma \cdot g(-\Gamma x_{ik} - \varepsilon_i, \varepsilon_i|x_{ik})d\varepsilon_i
\]

in the above binary choice example. The major statistical problem in estimating \( \beta(x_{ik}) \) arises from the fact that neighborhood income, \( x_{ik} \), is likely to be correlated with unobserved characteristics of families who are clustered in the same neighborhood. Standard parametric methods for binary choice models such as the probit, logit, as well as nonparametric estimators provide biased estimates of \( \beta(x_{ik}) \) when \( \varepsilon_i \) is correlated with \( x_{ik} \). In those methods, the estimate of \( \beta(x_{ik}) \) will pick up part of the effect of \( \varepsilon_i \) on \( y_{ik} \).

Aaronson (forthcoming) attempts to get around this problem by comparing the schooling outcomes of siblings who grew up in different neighborhoods. He assumes that most family background characteristics are the same for siblings conditional on observables such as family income and marital status. He used the linear probability model with family fixed effects and the conditional logit model proposed by Gary Chamberlain (1980, 1984) to do this. Many other authors have used one or both of these methods in other contexts. Unfortunately, the linear probability model is biased in almost all circumstances. Chamberlain’s conditional logit model does not estimate the parameter of interest (the population mean at a particular value of \( x_{ik} \) of the effect of \( x_{ik} \) on the mean of \( y \)) and does not use information on groups (e.g., siblings) in which all members have the same value for \( y_{ik} \). There is no suitable estimation method in the literature.
To give a second special case to which our methods apply, consider the model

\[ y_{ik} = m(x_{ik}, \varepsilon_i, u_{ik}) = x_{ik} \beta_1 + H(x_{ik}, \varepsilon_i) + \varepsilon_i + u_{ik} \]

where \( E(\varepsilon_i | x_{ik}) \neq 0 \) and \( H \) is differentiable. Since \( m \) is differentiable in \( x_{ik} \) the parameter of interest may be written as

\[
\beta(x_{ki}) = \int \int [m_{\varepsilon_k}(x_{ki}, \varepsilon_i, u_{ki})] g(u_{ki}, \varepsilon; | x_{ki}) d\varepsilon_d u_{ik}.
\]

There is a huge literature that assumes that \( H(x_{ik}, \varepsilon_i) \) is 0 and deals with the correlation between \( \varepsilon_i \) and \( x_{ik} \) by controlling for \( \varepsilon_i \) with a group specific intercept. These "fixed effects" estimators don't work when the impact of \( x_{ik} \) on \( y_{ik} \) depends on \( \varepsilon_i \), and alternatives in the literature require strong assumptions about the form of \( H \) and the distributions of the error terms.

In this paper we propose two estimators for these problems as well as a wide class of other panel data models involving nonseparable error terms and endogenous regressors. For simplicity let \( K_i = K \) for all groups \( i \), and define \( x_i \) to be the vector \([x_{i1}, ..., x_{iK}] \). Both estimators are based on the assumption that the distribution of \( u_{ik} \) and \( \varepsilon_i \) conditional on \( x_{i1} ... x_{iK} \) is exchangeable in \([x_{i1} ... x_{iK}] \). By "exchangeable" we mean \( g(u_{ik}, \varepsilon_i | x_{i1} ... x_{iK}) \) does not depend on the order in which the \( x_{ik} \) are entered into the function \( g(u_{ik}, \varepsilon_i | x_{i1} ... x_{iK}) \).

That is,

\[ g(u_{ik}, \varepsilon_i | x_{i1} ... x_{iK}) = g(u_{ik}, \varepsilon_i | x_{i1} ... x_{iK}) \]

Assumption \( \text{A1.1} \):

For example, the assumption implies that \( g(u_{ik}, \varepsilon_i | x_{i1} ... x_{iK}) = g(u_{ik}, \varepsilon_i | x_{iK} ... x_{i1}) \).

In neighborhood and sibling applications the assumption that the value of the function \( g(u_{ik}, \varepsilon_i | x_i) \) is the same regardless of the order in which the \( x_{ik} \) are entered into \( x_i \) is a natural one provided that the elements of \( x_{ik} \) are measured at the same age for each child.

Our first estimator is based on the conditional expectation function \( E(y_{ik} | x_i) \), and for this reason we refer to it as the "Regression Estimator". The variable \( x_{i1} \) has a direct impact on \( y_{i1} \) through the function \( m(.) \) and an indirect impact by shifting the distribution of \( \varepsilon_i \) and \( u_{i1} \). \([x_{i2} ... x_{iK}]\) only have an
indirect impact through their effects on the distribution of \( \varepsilon_i \) and \( u_{1i} \). However, exchangeability restricts the distribution of \( u_{1i} \) and \( \varepsilon_i \) to depend on \( x_i \) only through a vector of \( L \) exchangeable functions \( z(x_i) \) of \( x_i \). This implies that \( g(u_{1i}, \varepsilon_i | x_{i1}...x_{iK}) = g(u_{1i}, \varepsilon_i | z(x_i)) \) and that the mean \( E(y_{ik}|x_{i}) \) can be written as \( E(y_{ik}|x_{i1}, z(x_i)) \). If the \( z(.) \) vector of functions is known and \( x_{1i} \) and the elements of the vector \( z(x_i) \) vary sufficiently, then the partial derivative \( \partial E(y_{ik}|x_{i1}, z(x_i))/\partial x_{1i} \) is identified. Since it is possible to estimate the distribution of \( z(x_i) \) conditional on \( x_{1i} \), one can recover the parameter of interest \( \beta(x_{1i}) \) from the estimator of \( \partial E(y_{ik}|x_{i1}, z(x_i))/\partial x_{1i} \) by integrating this derivative over the distribution of \( z(x_i) \) conditional on \( x_{1i} \).

The second estimator also relies heavily on A1.1 but involves a somewhat different set of assumptions. It does not use or require knowledge of the \( z(.) \) functions, but it does require some additional assumptions that we discuss below. The most important is that \( m(x_{ik}, \varepsilon_i, u_{ik}) \) is of the form \( m(x_{ik}, e_{ik}) \) and is strictly monotone in \( e_{ik} \), where \( e_{ik} = \Upsilon(\varepsilon_i, u_{ik}) \) and \( \Upsilon(.,.) \) is a continuous function. The strict monotonicity assumption rules out qualitative choice models but covers many models that take on the form of (1.4). We show that \( m(x_{ik}, e_{ik}) \) and \( g(e_{ik}|x_{ik}) \) are identified under exchangeability of \( g(\varepsilon_i|x_i) \) from knowledge of the joint distribution of \( y_{ik} \) and \( x_i \). With knowledge of \( m(x_{ik}, e_{ik}) \) and \( g(e_{ik}|x_{ik}) \) one can estimate the average response \( \beta(x_{1i}) \) as well as other parameters that characterize the distribution of the response of \( y_{ik} \) to a change in \( x_{ik} \).

The basic intuition underlying the second estimator is as follows. Suppose that \( K = 2 \), with \( k = 1, 2 \). A shift in \( x_{1i} \) alters the distribution of \( y_{ik} = m(x_{i1}, e_{ik}) \) by shifting \( m(x_{i1}, e_{ik}) \) for a given value of \( e_{i1} \) and by shifting the distribution \( e_{ik} \). A shift in \( x_{1i} \) alters the distribution of \( m(x_{i2}, e_{ik}) \) only by shifting the distribution of \( e_{ik} \). Consequently, one can isolate the direct effect of \( x_{1i} \) on the distribution of \( m(x_{i1}, e) \) by comparing the change in the distribution of \( m(x_{i1}, e_{ik})| x_{11}, x_{12} \) as \( x_{11} \) changes to the change in the distribution of \( m(x_{i1}, e_{ik})| x_{12}, x_{11} \) as \( x_{11} \) changes.

The theoretical literature on estimating nonseparable panel data models when the regressors are correlated with the error term is relatively small. (See Powell (1994)). An exception is the recent independent paper by Abrevaya (1997) which deals with generalized regression models with fixed effects. Abrevaya’s approach permits estimation of slope parameters up to scale but, in contrast to our approaches, does not permit estimation of the partial effects of \( x_{ik} \) on mean of \( y_{ik} \). In the case of qualitative response models we
have already mentioned the linear probability model with fixed effects and the conditional logit model. The conditional logit model and the other "fixed effects" approaches that we are aware of are restricted to specifications that take on the additively separable form of (1.1a).\(^1\) The fixed effects probit model is sometimes used to estimate \(\Gamma\) up to scale. It is well known that the fixed effects probit model is inconsistent when the group size is fixed, but Heckman (1981) provides monte carlo evidence suggesting that the bias is small when \(K\) is on the order of 10.\(^2\)

Manski (1987) provides a way to estimate \(\Gamma\) up to scale under more general assumptions than the conditional logit. He places no restrictions on the distribution of \(e_i\) and assumes that the distribution of \(u_{i1}|e_i, x_{i1}, x_{i2}\) is the same as the distribution of \(u_{i2}|e_i, x_{i1}, x_{i2}\). He proposes a maximum score estimator that exploits that fact that \(\text{sgn}(E[(y_{i2} - y_{i1})|x_{i1}, x_{i2}] = \text{sgn}(x_{i2}\Gamma - x_{i1}\Gamma)\) where \(\text{sgn}(\cdot)\) is -1 if the argument is negative and 1 if it is positive. In contrast to Manski's estimator, our approach requires appriori information about the distribution of \(e_i|x_{i1}, x_{i2}\). However, it permits us to estimate the partial effect of \(x_{ik}\) on the probability that \(y_{ik}\) is 1 as well as the parameter vector \(\Gamma\) up to scale. Furthermore, our "regression" approach can handle qualitative choice models cases in which \(x_{ik}\) and the error components interact in arbitrary ways while the other approaches in the literature cannot.

We should point out however that the estimators in there current form cannot accomodate dynamics in the model, which are addressed in recent papers by Honore and Kyriazidou (1997) and Kyriazidou (1997).

The conditional logit and the fixed effects probit estimators may be thought of as parametric "fixed effects" approaches. In addition, Chamberlain (1984) discusses parametric random effects approaches to estimating \(\Gamma\) up to scale in (1.3a). This approach has been used in a number of applications, including Jakubsen (1977). Assume that \(u_{ik}\) is normal, identically

\(^1\)The most common method in empirical studies is the linear probability model with fixed effects, which forces one to maintain that the probability of \(y\) is the sum of \(e_i\) and a function of \(x_{ik}\).

\(^2\)Heckman and MaCurdy (1980) apply this estimator as well as the fixed effects Tobit estimator to the analysis of life cycle labor supply. Note that one can recover an estimate of the partial effect of \(x_{ik}\) on the probability that \(y_{ik}\) is 1 from the probit coefficients and the distribution of \(e_i\) given \(x_{ik}\). However, the MLE estimates of \(e_i\) are unbounded when \(y\) is the same for all group members, so one cannot obtain an estimate of the distribution of \(e_i|x_{ik}\) without making assumptions about this distribution. The same is true in the case of the conditional logit.
distributed across $k$, and independent of $x_i$ and that $\varepsilon_i$ is the sum of a function of $f(x_i; \theta)$ plus a normally distributed error term that is independent of $x_i$. Then one can estimate $\Gamma$ up to scale by adding $f(x_i; \theta)$ to a probit model for each $k$ and jointly estimating $\theta$ and $\Gamma$ while imposing cross restrictions across the models for each $k$. One may also recover the partial effect of $x_{ik}$ on the probability $y_{ik}$ that $1$. The main disadvantages of this approach relative to ours is that it requires the assumption that $u_{ik}$ and $\varepsilon_i|x_i$ are normal and the assumption of additive separability, as in (1.3a).

In the case of continuous variables, the incidental parameters problem limits the utility of parametric “fixed effects” approaches for models such as (1.4). In special cases, parametric random effects approaches may be available. GMM is often used to estimate the parameters of nonseparable models and it may be possible in some cases to estimate elements of $\beta_1$ or some parameters of the $H(x_{ik}, \varepsilon_i; \theta)$ when that function is parametric. However, there are many cases in which this method cannot be used to estimate the partial effect of $x_{ik}$ on the mean of $y_{ik}$.

The paper continues in section 2, where we present the “Regression” estimator based on $E(y_{ik}|x_i)$. In section 3 we discuss a nonparametric version of the estimator and analyze its asymptotic distribution. In section 4 we discuss an extension of the estimator to the case in which $x_{ik}$ is correlated with $u_{ik}$ conditional on $x_i$ but an instrumental variable is available. This estimator provides an alternative to the “fixed effects-IV” linear probability model that is sometimes used in applied studies even though it is inconsistent. In section 5 we derive the second estimator and provide results on its asymptotic properties. In section 6 we present some very encouraging monte carlo evidence on the performance of the “Regression Estimator. In section we provide some concluding remarks.

---

3Thus far, neither of our estimators cover other limited dependent variables models such as the censored regression models or sample selection models. Honore (1992) provides a fixed effects estimator for the limited dependent variables case. Kyriazidou (1997) uses an exchangeability assumption that is similar to ours in her work on panel data sample selection models. The approaches in both of these papers are based on differencing the observations in clever ways and are quite distinct from our approaches.
2 An Estimator Based on $E(y_{ik}|x_i)$

In this section we present our regression based estimator. The estimator uses exchangeable functions $z^1(x_{i1}...x_{iK}), z^2(x_{i1}...x_{iK}), ..., z^L(x_{i1}...x_{iK})$ of $(x_{i1}...x_{iK})$ satisfying that property that for all $x_{i1}...x_{iK}$, $g(u_{ik}, e_i|x_{i1}...x_{iK}) = g(u_{ik}, e_i|z^1_i, ..., z^L_i)$. By exchangeable we mean that the functions are invariant to the order in which the elements of $(x_{i1}...x_{iK})$ enter the function. For example, $z^1_i$ might be the mean of $x_{i1}...x_{iK}$ for family $i$ and $z^2_i$ might be the average over $k$ of $(x_{ik} - z_i)^2$. As we noted in the introduction, assumption (A1.1) that $g(u_{ik}, e_i|x_{i1}...x_{iK})$ is exchangeable in $x_{i1}...x_{iK}$ means that without loss of generality we can $g(u_{ik}, e_i|x_{i1}...x_{iK})$ as $g(u_{ik}, e_i|z^1_i, ..., z^L_i)$, where $z^1_i, ..., z^L_i$ are symmetric functions $z^1(x_{i1}...x_{iK}), z^2(x_{i1}...x_{iK}), ..., z^L(x_{i1}...x_{iK})$ of $(x_{i1}...x_{iK})$. Let $z_i$ be the vector of $z^L_i$ variables for family $i$. The first estimator requires the following assumptions in addition to (A1.1)

**Assumption 2.1.** The functions that define $z_i$ in terms of $x_i$ are known.

**Assumption 2.2.** The distributions of each element of the vector $x_{ik}, z^1_i, ..., z^L_i$ conditional on the other elements of the vector are nondegenerate.

With these assumptions one may estimate $E(y_{ik}|x_{ik}, z_i)$ nonparametrically. (As we discuss below, Assumption 2.2 can be weakened appriori inforamtion about the functional form of $E(y_{ik}|x_{ik}, z_i)$ is available.) The Regression estimator of $\beta(x_{ik})$ is based on the conditional expectation function $E(y_{ik}|x_{ik}, z_i)$. Suppressing the $i$ subscript where it is not needed for clarity and setting $k$ to 1 for concreteness, this function is

$$E(y_1|x_1, z) = \int \int m(x_1, u_1) g(u_1, e|x_1, z) du_1$$

(2)

The idea of the estimator is to recover $\beta(x_1)$ from

$$E_{x_1}(y_1|x_1, z) = \frac{\partial}{\partial x_1} E(y_1|x_1, z)$$

(2.2) and $h(z|x_1)$, the conditional distribution of $z|x_1$, The distribution $h(z|x_1)$ can be estimated from the observations on $x_1$ for the cross section of groups $i$. The derivative with respect to $x_1$ is:

$$E_{x_1}(y_1|x_1, z) = \int \int m_{x_1}(x_1, u_1) g(u_1, e|x_1, z) du_1$$

(2.3)
because
\[ \int_\varepsilon \int m(x_1, \varepsilon, u_1) g_{x_1}(u_1, \varepsilon | x_1, z) \, d\varepsilon \, du_1 = 0 \]

The form of (2.3) is analogous to (1.3a) in the binary choice case. In this case
\[ (2.4) \quad E_{x_1}(y_1 | x_1, z) = \int \Gamma \cdot g(-\Gamma x_1 - \varepsilon, \varepsilon | x_1, z) \, d\varepsilon. \]

Note that \( E_{x_1}(y_1 | x_1, z) \) differs from \( \beta(x_1) \) because the distribution of \( u_1 \) and \( \varepsilon \) is conditioned on both \( x_1 \) and \( z_i \). (See the right hand side of (2.3) or (2.4).) However, one may integrate out \( z \) to obtain \( \beta(x_1) \) from \( E_{x_1}(y_1 | x_1, z) \).

To see how to do this, note first that
\[ (2.5) \quad g(u_1, \varepsilon | x_1) = \int z \cdot g(u_1, \varepsilon | x_1, z) h(z | x_1) \, dz \]

where \( h(z | x_1) \) is the conditional density of \( z \) given \( x_1 \). Multiply both sides of (2.3) by \( h(z | x_1) \) and integrate over the range of \( z \). This yields (2.6)
\[ (2.6) \quad \int z E_{x_1}(y_1 | x_1, z) h(z | x_1) \, dz = \int \int z m_{x_1}(x_1, \varepsilon, u_1) g(u_1, \varepsilon | x_1, z) h(z | x_1) \, dz \, du_1 \, dz \]

Re-arranging the order of integration on the right hand side of the equality and using (2.5) establishes that the right hand side is \( \beta(x_1) \), the function we would like to estimate. That is,
\[ (2.7) \quad \beta(x_1) = \int z E_{x_1}(y_1 | x_1, z) h(z | x_1) \, dz \]

The above equation forms the basis of our first estimator. The estimator is obtained by substituting parametric or nonparametric estimators of the components of the right hand side of (2.7) into the equation.\(^4\) In the next section we provide the asymptotic distribution theory for a nonparametric approach in which kernel estimators of \( E_{x_1}(y_1 | x_1, z) \) and \( h(z | x_1) \) are used.

\(^{4}\)When \( K \), the number of observations per group, differs across \( i \), one could do the estimation for each group size and then combine the estimates.
The key assumption, aside from exchangeability, is that \( E(y_1|x_1, z) \) is identified conditional on prior information about how \( x_1 \) and \( z \) enter the expectation function. The expectation function is identified nonparametrically only if \( x_1 \) varies conditional on \( z \). The conditions for identification involve trade-offs among (a) the size of the panel \( K \), (b) the number \( L \) of elements in \( z \), and (c) parametric or nonparametric restrictions on \( E(y_1|x_1, z) \). For example, consider the case in \( x_1 \) is a scalar and \( K = 2 \) (i.e., there are observations on 2 children per family). In this case, we conjecture that variables that are continuous exchangeable functions of the elements of \( x_i = \{x_{1i}, x_{2i}\} \) may be approximated arbitrarily closely by functions of \( z_i = x_{1i} + x_{2i} \) and \( z_i^2 = |x_{1i} - x_{2i}|^{5} \). If the conjecture is correct, then conditioning the distribution of \((e_i, u_{ik})\) on \( z_1 \) and \( z_2 \) is general enough. However, \( z_i^2 = |2x_{1i} - x_{2i}| \). This dependence among \( z_i, z_i^2, \) and \( x_{1i} \) means that

\[
E(y_1|x_1, z_1, z_2)
\]

is not identified nonparametrically when \( K \) is 2. However, if the function

\[
E(y_1|x_1, z_1, z_2)
\]

is a low order polynomial in the three variables or \( z_2 \) does not enter at all, then it may be identified. It is also possible that the function will be identified over

\[5\](Preliminary) A sketch of a proof of this conjecture follows. Let \( z_i^3 = x_{1i}x_{2i} \). \( z_i^3 = 0.25((z_i^2)^2 - (z_i^3)^2) \). Consider functions of the form \( h_1(\Lambda_1(x_1) + \Lambda_1(x_2)) \). Assuming \( \Lambda_1(x_1) \) is continuous, then it can be approximated arbitrarily closely by a polynomial in \( \Lambda_1(x_1) \). Consequently, \( h_1(\Lambda_1(x_1) + \Lambda_1(x_2)) \approx h_1(a_0 + a_1(x_1 + x_2) + a_2(x_1^2 + x_2^2) + a_3(x_1^3 + x_2^3) + \ldots) \) where we have stopped at three terms for simplicity. It is tedious but straightforward to show that one can express terms \( x_1^2 + x_2^2 \) as a linear combination of powers of \( z_i^2 \), powers of \( z_i^3 \), and products of powers of the two. Thus, functions of the form \( h_1(\Lambda_1(x_1) + \Lambda_1(x_2)) \) can be approximated arbitrarily closely by functions of \( z_i^2 \) and \( z_i^3 \). Functions of the form \( h_2(\Lambda_2(x_1) - \Lambda_2(x_2)) \approx h_2(\Lambda_2(x_1) - \Lambda_2(x_2) + a_1(x_1^2 - x_2^2) + a_2(x_1^3 - x_2^3) + \ldots) \). The sum of the first three terms inside the absolute value sign is equal \((x_1 - x_2)(a_1 + a_2(x_1 + x_2) + a_3(x_1^2 + x_2^2) + a_3(x_1x_2))\). The first term in the product is \( z_i^2 \). The components of the second term may be expressed as functions of \( z_i^2 \), \( z_i^3 \) and powers of \( z_i^3 \). Exchangeable functions of the form \( h_3(\Lambda_3(x_1) \cdot \Lambda_3(x_2)) \) can be expressed as functions of the three \( x \) variables in similar fashion by working with a polynomial approximation to \( \Lambda_3(x_1) \cdot \Lambda_3(x_2) \). We conjecture that any continuous exchangeable function of \( x_1 \) and \( x_2 \) must be a composite function of the terms involving \( h_1(\Lambda_1(x_1) + \Lambda_1(x_2)) \), \( h_2(\Lambda_2(x_1) - \Lambda_2(x_2)) \), and \( h_3(\Lambda_3(x_1) \cdot \Lambda_3(x_2)) \). If this is true, then we conclude that any continuous exchangeable function of \( x_1 \) and \( x_2 \) may be written approximated arbitrarily closely by a function of \( z_i^2 \) and \( z_i^3 \). We do not know if there is a useful generalization of this result when \( K > 2 \).
some ranges of $x$ where there is variation in $x$ conditional on $z$ but not others. Note that when $K$ is greater than 2, it is possible to test restrictions on the dimensionality of $z$. For example, when $K = 4$ one can test the hypothesis that the distribution of $(\varepsilon, u_k)|x_1$ depends only on $z_{11} = (x_{i1} + x_{i2} + x_3 + x_k)/4$ and $\sum |x_{ik} - z_{11}|$. Finally, note that it is necessary to identify the effects of $z_1$ and $z_2$ on the mean of $y_1$ conditional on $x_1$ only over the range in which the conditional density $h(z_1, z_2)$ is positive since it is only these values that enter into (2.7).

2.1 Discussion

One very attractive feature of the estimator in the binary choice case compared to the conditional logit or fixed effects probit estimators is that it can utilize groups in which $y_{ik}$ is either 1 or 0 for all $k$. It is quite common panel data applications, particularly for rare events, that all group members have the same value for $y_{ik}$.  

A second very attractive feature of the regression estimator is that it only requires that data on the dependent variable $y_{ik}$ be available for one member of group $i$, although data on $x_{ik}$ must be available for at least 2 members of group $i$. In contrast, the conditional logit estimator and standard fixed effects estimators require data on $y_{ik}$ and $x_{ik}$ for at least 2 group members. Consequently, data on children as young as 16 can be included in studies of the effects of neighborhood characteristics during childhood on outcomes that occur later in life, such as college graduation or marriage. This will substantially increase the sample sizes for sibling studies. To provide a bit of intuition for why one only needs data on $y_{ik}$ for one member of group $i$ as well as the intuition underlying the Regression estimator, it is helpful to relate it to other panel data estimators for the standard separable case. Note that the standard linear regression model with an additive family fixed effect is a special case of our model (1.1). Consider the model

\[(2.8) \, y_{ik} = x_{ik}\beta + \varepsilon_i + u_{ik}\]

6The regression estimator can also be applied to multinomial models. A simple way to do this is to treat each outcome as a separate 0-1 variable, and estimate $E_{z_{ik}}(y_{ik}|x_{ik}, z_i)$, impose the adding up constraint, and integrate out $z_i$. 

11
Chamberlain (1984) and Mundlak (1978) point out that the parameter \( \beta \), which is the effect of \( x_{ik} \beta y_{ik} \) holding \( \varepsilon_i \) constant, may be estimated by using the decomposition of \( \varepsilon_i \) into its least squares linear projection on the elements of \( x_i \) and the orthogonal error term \( u_i \) to eliminate \( \varepsilon_i \) from the above equation, and using the \( K \) observations on group \( i \) to estimate the system of equations

\[
(2.9) \quad y_{ik} = x_{ik} \beta + x_i \lambda + u_{ik} + u_i, \quad k = 1,...,K
\]

with cross equation restrictions imposed. This does not require the assumption of exchangeability. The assumption of exchangeability places restrictions on the coefficient vector \( \lambda \) summarizing the relationship between \( \varepsilon_i \) and the elements of \( x_i \). In this case our regression based estimation procedure would amount to running the regression

\[
(2.10) \quad y_{ik} = x_{ik} \beta + f(z_i; \beta_1) + u_{ik} + u_i
\]

where \( z_i \) is a vector of exchangeable functions of \( x_i \) and \( f(\cdot; \beta_1) \) is a function with parameter vector \( \beta_1 \), such as a polynomial. In the above model \( m_{z_{ik}}(x_{ik}, u_{ik}, \varepsilon_i) \), the partial derivative of \( y_{ik} \) with respect to \( x_{ik} \) is a constant \( \beta \), so it is not necessary to integrate out over the distribution \( h(z_i|x_{ik}) \). A special case is when \( z_i \) only contains \( x_i \), the sum of the elements of \( x_i \) and \( f(\cdot; \cdot) \) is linear. In this case, (2.10) is equivalent to (2.9) with the restriction that the elements of \( \lambda \) are all equal. With these restrictions one does not need to have data on all of the \( y_{ik} \) to identify \( \beta \) from (2.8).

3 Asymptotic Properties of the Regression Estimator in the Nonparametric Case

The nonparametric version of the estimator introduced in Section 2 is given by

\[
\hat{\beta}(x) = \int \frac{\partial}{\partial \beta} \hat{E}(y|x, z) \hat{h}(z|x) \, dz,
\]

where \( \hat{E}(y|x, z) \) is a kernel estimator of the conditional expectation of \( Y \) given \( (X, Z) \) and \( \hat{h}(z|x) \) is a kernel estimator of the conditional pdf of
Z given X. We suppress the i subscript and k subscripts on \( y_{ik} \) and \( x_{ik} \) and the i subscript on \( z_i \).

To derive the asymptotic properties of \( \hat{\beta}(x) \), we note that \( \hat{\beta}(x) \) is a functional of the kernel estimator, \( \hat{F}(y, x, z) \), for the joint cdf \( F(y, x, z) \) of \((Y, X, Z)\). Hence, a "delta-method" such as the ones described in Newey (1994) or Ait-Sahalia (1992) can be used to derive its asymptotic distribution.

Let \( K_1 \) denote the dimension of \( x \) and \( K_2 \) denote the dimension of \( z \).

Let \( d = 1 + K_1 + K_2 \), and let \( f \) denote the pdf of \((y, x, z)\). Then,

\[
\hat{F}(y, x, z) = \int_{-\infty}^{y} \int_{-\infty}^{x} \int_{-\infty}^{z} \hat{f}_N(t_y, t_x, t_z) \, dt_y \, dt_x \, dt_z
\]

where

\[
\hat{f}_N(t_y, t_x, t_z) = \frac{1}{N\sigma^d} \sum_{i=1}^{N} K\left( \frac{t_x - x_i}{\sigma}, \frac{t_y - y_i}{\sigma}, \frac{t_z - z_i}{\sigma} \right)
\]

for all \((t_y, t_x, t_z) \in R^d\).

The following assumptions will be needed:

ASSUMPTION 1: The sequence \( \{y_i, x_i, z_i\} \) is a strictly stationary \( \beta \)-mixing sequence satisfying \( k^\delta \beta_k \to 0 \) as \( k \to \infty \), for some \( \delta > 1 \).

ASSUMPTION 2: \( f(y, x, z) \) has a compact support \( \Theta \subset R^d \), \( f(\partial \Theta) = 0 \) where \( \partial \Theta \) denotes the boundary of \( \Theta \), and \( f(y, x, z) \) is continuously differentiable up to order \( g \) where \( g \geq 2d \).

ASSUMPTION 3: The kernel function \( K(\cdot, \cdot, \cdot) \) is an even function, integrates to 1, is of order \( r \) where \( r \) is an even integer satisfying \( d/2 < r < g \), is continuously differentiable up to order \( g \) and its derivatives of order up to \( g \) are in \( L^2(R^d) \). I.e.,

(i) \( K(-y, -w, -z) = K(y, w, z) \) for all \((y, w, z) \in R^d\),
(ii) \( \int_0^\infty \int_0^\infty \int_0^\infty K(y, w, z) \, dy \, dw \, dz = 1, \)

(iii) for some even integer \( r \) such that \( d/2 < r < g, \)

(a) \( \forall \lambda \in \mathbb{N}^d \) such that \( \sum_{k=1}^d \lambda_k \in \{1, 2, \ldots, r-1\} \)
\[ \int_0^\infty t_{1}^{\lambda_1} t_{2}^{\lambda_2} \cdots t_{d}^{\lambda_d} K(t) \, dt = 0, \quad \text{where} \ t = (t_1, \ldots, t_d) \]

(b) \( \exists \lambda \in \mathbb{N}^d \) such that \( \sum_{k=1}^d \lambda_k = r \) and
\[ \int_0^\infty t_{1}^{\lambda_1} t_{2}^{\lambda_2} \cdots t_{d}^{\lambda_d} K(t) \, dt \neq 0, \]

(c) \( \int_0^\infty ||t||^r \, |K(t)| \, dt < \infty, \)

(iv) \( K(\cdot, \cdot, \cdot) \) is continuously differentiable on \( \mathbb{R}^d \) up to order \( g \) and its derivatives of order up to \( g \) are in \( L^2(\mathbb{R}^d) \).

ASSUMPTION 4: As \( N \to \infty, \sigma_N \to 0, N\sigma_N^d \to \infty, \sqrt{N}\sigma_N^{-} \to 0, \) and \( \sqrt{N}\sigma_N^{-m} \to \infty \) where \( m \) is an integer such that \( m < r/2 \) and \( m + r \leq g + d. \)

ASSUMPTION 5: \( \infty > f(x) > 0, \) and \( \int f(z) f(x, z) \, dz \)
\[ \int \frac{(\int f(y, z) \, dy)^2}{f(x, z)^3} \, dz, \int \frac{(\int f(y, z) \, dy)^2}{f(x, z)^3} \, dz, \int \frac{1}{f(x, z)^2} \, dz, \int y f(y, x, z) \, dy \, dz, \]
\[ \int \frac{f(y, z)}{f(x, z)^2} \, dz, \int \frac{y f(y, x, z) \, dy}{f(x, z)} \, dz, \int \frac{f(y, x, z)}{f(x, z)} \, dy \, dz, \int \frac{y f(y, x, z) \, dy}{f(x, z)} \, dz, \]
are all bounded.

**Theorem 1.1:** If Assumptions 1-5 are satisfied, then \( \hat{\beta}(x) \) is a consistent estimator of \( \beta(x) \) and
\[ \sqrt{N} \sigma_N^{(K_1/2)+1} \left( \hat{\beta}(x) - \beta(x) \right) \to N(0, V) \] in distribution

where
\[ V = \left\{ \text{Var}(y|x,z) \cdot f(x,z) \, dz \right\} \left\{ \int \left[ \frac{1}{2} \left( \int K(y,x,z) \, dy \right)^2 \right] \, dx \right\} \frac{1}{f(x,z)} \]

PROOF: See the Appendix.

4 An Extension: Correlation between \( u_{ik} \) and \( x_{ik} \) conditional on \( z_i \).

In some applications, \( x_{ik} \) will be correlated with the idiosyncratic error component \( u_{ik} \) even after one conditions on \( z_i \). It is common in panel data applications involving continuous dependent variables with additive error terms such as (2.8) to use an instrumental variable approach to deal with this problem while at the same time adding group specific intercepts to control for \( \varepsilon_i \), or to use the class of estimators discussed in Hausman and Taylor (1982). Unfortunately, this approach is not available in the case of nonseparable models. Here we extend the Regression estimator to handle correlation between \( u_{ik} \) and \( x_{ik} \) when an instrumental variable \( A_{ik} \) is available. We modify the above model by dropping A1.1 and replacing it with the following assumptions.

(A1.1a) \( x_{ik} = Y(A_{ik}) + \xi_{ik} \), where \( Y(A_{ik}) \) is independent \( \xi_{ik} \)

(A1.1b) \( A_{ik} \) is independent of \( \varepsilon_i, u_{ik} \) conditional on \( z_i, \xi_{ik} \).

Note that the correlation of \( x_{ik} \) with \( u_{ik} \) comes from \( \xi_{ik} \). Since \( x_{ik} \) and \( A_{ik} \) are both observed, it possible to consistently estimate \( Y(A_{ik}) \) and \( \xi_{ik} \), particularly if \( Y(A_{ik}) \) has a finite number of parameters. Given these facts, we modify the approach to estimation underlying (2.7) by working with the \( E_{z_{ik}}(y_{ik}|x_{ik}, z_i, \xi_{ik}) \) rather than \( E_{z_{ik}}(y_{ik}|x_{ik}, z_i) \).

Suppressing the \( i \) subscript and set \( k \) to 1.

(i) \( E(y_1|x_1, z_1, \xi_1) = \int \int m(x_1, \varepsilon, u_1)g(u_1, \varepsilon|x_1, z_1, \xi_1) \, d\varepsilon \, du_1 \)

Since \( Y(A_{ik}) \) is independent of \( (\varepsilon, u_{ik}) \),

(ii) \( \frac{\partial g(u_1, \varepsilon|x_1, z_1, \xi_1)}{\partial x_1} = \frac{\partial g(u_1, \varepsilon|x_1, z_1, \xi_1)}{\partial Y(A_{1})} = 0 \)
Using (i) and (ii) leads to

\[(iii) \quad E_{x_1}(y_1|x_1, z, \xi_1) = \int \int m_{x_1}(x_1, \varepsilon, u_1)g(u_1, \varepsilon|x_1, z, \xi_1)dz \, du_1\]

\(E_{x_1}(y_1|x_1, z)\) differs from \(\beta(x_1)\) because the distribution of \(u_1\) and \(\varepsilon\) is conditioned on \(z\) and \(\xi_1\) as well as \(x_1\). However, one may integrate out \(z\) and \(\xi_1\) to obtain \(\beta(x_1)\) from \(E_{x_1}(y_1|x_1, z)\). To see how to do this, note first that

\[(iv) \quad g(u_1, \varepsilon|x_1) = \int \int g(u_1, \varepsilon|x_1, z)h(z, \xi_1|x_1)dz\]

where \(h(z, \xi_1|x_1)\) is the conditional density of \(z\) given \(x_1\). Following the approach above, multiply both sides of (iii) by \(h(z, \xi_1|x_1)\) and integrate over the range of \(z\) and \(\xi_1\). This yields (v)

\[(v) \quad \int \int E_{x_1}(y_1|x_1, z, \xi_1)h(z, \xi_1|x_1)d\xi_1dz\]

\[= \int \int \int m_{x_1}(x_1, \varepsilon, u_1)g(u_1, \varepsilon|x_1, z)h(z|x_1)dz \, du_1 \, d\xi_1dz\]

Using (iv) and re-arranging the order of integration on the right hand side of the equality establishes that the right hand side is \(\beta(x_1)\), the function we would like to estimate.

The key assumption is (A1.1b), but this assumption is very unlikely to hold when \(\xi_{ik}\) is correlated with \(\varepsilon_i\) and (A1.1a) does not hold. Altonji and Ichimura (1997) take a similar approach to treatment of endogenous explanatory variables in the context of nonseparable linear dependent variables models. As they point out, the conditions that \(A_{ik}\) is independent of \(\xi_{ik}\) and independent of \(\varepsilon_i\) and \(\varepsilon_{ik}\) are much stronger than the usual conditions on instrumental variables for IV estimators. However, it should be kept in mind that IV estimators of partial derivatives are inconsistent in models such as (1.4), where slope coefficients are random and correlated with the endogenous variable.
Estimating the effect of $x_{i1}$ on $y_{i1}$ from the joint distribution of $y_i$ and $x_i$.

Our second estimator uses the entire distribution of $y_{i1}$ given $x_i$ rather than just the conditional expectation function. We show that under certain assumptions it is possible to identify $m(x_{ik}, \epsilon_i, u_{ik})$ and $g(u_{ik}, \epsilon_i|x_i)$ from the joint distribution of $y_{ik}|x_i$ and the distribution of $y_{ik}|x_{ik}$. Consequently, various functions of $m(x_{ik}, \epsilon_i, u_{ik})$ and $g(u_{ik}, \epsilon_i|x_i)$, including averages such as $\beta(x_{i1})$, are identified. Our proof of identification is a constructive proof and, hence, it provides a way of estimating $m(x_{ik}, \epsilon_i, u_{ik})$ and $g(u_{ik}, \epsilon_i|x_i)$ from a nonparametric estimator for the joint distribution of $y_i$ and $x_i$.

To simplify the notation we will sometimes suppress both the $k$ and $i$ subscripts and use $y$ to refer to $y_{i1}$, $x$ to refer to $x_{i1}$, and $u$ to refer to $u_{i1}$. The model underlying the second approach to estimation is described by the following assumptions:

(A5.1) There exists a function $Y(\epsilon, u)$ such that $y = m(x, \epsilon)$, where $e = Y(\epsilon_i, u_i)$

Let $q(e|x_1, x_2)$ denote the conditional density of given $(x_1, x_2)$.

(A5.2) $\forall w, w' \in \mathbb{R}$ $q(e|w, w') = q(e|w', w)$.

(A5.3) $\forall x$ $m(x, \cdot)$ is strictly increasing in $\epsilon$.

(A5.4) $q(e|w, w')$ is strictly positive everywhere.

Assumption (A5.1) states that $m(\cdot, \cdot)$ is weakly separable in $x$ and a function $Y(\epsilon, u)$ of $\epsilon$ and $u$. We did not need this restriction for the regression estimator, but it is also assumed to hold in most nonlinear panel data models in the literature, such as the probit and logit binary choice models. Assumption (A5.2) states that the conditional distribution of $e$ is exchangeable in $x_1$ and $x_2$. If, for example, $e = Y(\epsilon, u) = \epsilon + u$, the conditional distribution of $\epsilon$ is exchangeable in $x_1$ and $x_2$, and $u$ is distributed independently of $x$ and
€, then, Assumption (A5.2) is satisfied.

The strict monotonicity assumption (A5.3) is not required for the regression estimator based on (2.7) but it seems to be critical for the identification of $m(x_{ik}, e_{ik})$ and $g(e_{ik}|x_{ik})$. As we have noted in the introduction, strict monotonicity of $m$ in $e$ restricts the analysis rules out qualitative choice models.\(^8\) On the other hand, this second estimator has a number of advantages over the regression estimator. First, it does not require the knowledge of or the use of the $z$ functions, which are required for the regression estimator. Second, it does not require that $x$ and the relevant $z$ functions vary independently for nonparametric identification. Finally, it permits one to estimate $m(x_{ik}, e_{ik})$ and $g(e_{ik}|x_{ik})$ and various functions of them that include but are not limited to $\beta(x_{ik})$. Thus the two approaches have different strengths and weaknesses and are complementary.

We adopt the following normalization:

$$m(0, e) = e.$$ \(^{(A5.5)}\)

This is innocuous because the assumption of strict monotonicity of $m$ in $e$ implies that, given any function $m(\cdot, \cdot)$, one can define a new function $m'(\cdot, \cdot)$ by $m'(x, \bar{e}) = m(x, m^{-1}(0, \bar{e}))$, where $m^{-1}(0, \cdot)$ denotes the inverse function of $m$ with respect to $e$, when $x = 0$. From the definition of $m'$ it follows that for all $\bar{e}$, $m'(0, \bar{e}) = \bar{e}$. Moreover, since for all $x$ and all $\bar{e}$ and $e$ such that $m'(x, \bar{e}) = m(x, e)$ it is the case that $\frac{\partial}{\partial x} m'(x, \bar{e}) = \frac{\partial}{\partial x} m(x, e)$, it follows from Brown and Matzkin (1996) that $m'$ and $m$ are observationally

\(^{7}\)To see this, let $g(e|w, w')$ and $s(u)$ denote, respectively, the conditional pdf of $e$ and the pdf of $u$, and note that

$$\forall w, w' \quad g(e|w, w') = \int s(e - e|w, w') g(e|w, w') \, de$$

$$= \int s(e - \bar{e}) g(e|w, w') \, de$$

$$= \int s(e - e|w, w) g(e|w, w) \, de$$

$$= g(e|w, w).$$

\(^{8}\)The sign of the effect of $e$ on $m$ can depend on $x_1$ provided that the analyst knows the values of $x_1$ at which the sign switches. For example, in the case of the model $m(x_1, e) = \alpha_1 x_1 + \alpha_2 x_1 e$ the sign of the effect of $e$ depends on the sign of $x_1$. 

18
Theorem 2 Under A5.1-A5.5 \( m(x, e) \) is identified from the distribution of \( y|x_1, x_2 \).

Proof. : Equation (5.2) implies that
\[ \forall w, w' \Pr(e \leq \eta|w, w) = \Pr(e \leq \eta|w, w'). \]

It follows by (A5.3) that
\[ \forall w', w, \forall \eta, \Pr(m(w', e) \leq m(\eta|w', w) = \Pr(m(w, e) \leq m(\eta|w, w'), \]
or
\[ (5.3) \Pr(y \leq m(w', \eta)|w', w) = \Pr(y \leq m(w, \eta)|w, w'). \]
In other words,
\[ F_{y|w', w}(m(w', \eta)) = F_{y|w, w}(m(w, \eta)) \text{ where } F_{y|w, w}(. \] is the CDF of \( y \) conditional on \( x_1 = w' \) and \( x_2 = w \). In particular,
\[ (5.4) F_{y|w, 0}(m(x, e)) = F_{y|0, w}(m(0, e)) = F_{y|0, w}(e), \]
where the last equality follows from (A5.5).

Let \( f(y|w', w) \) be the pdf of \( y \) conditional on \( x_1 = w', x_2 = w \). Since \( f(y|w', w) = g(m^{-1}(w', y)) \frac{\partial m^{-1}(w', y)}{\partial y} \), it follows from (A5.3) and (A5.4) that
\[ F_{y|w, w}(. \] is strictly increasing. Hence, by (A5.4)
\[ (5.5) m(x, e) = F_{y|w, 0}^{-1}(F_{y|0, w}(e)). \]
This shows that the function \( m \) is identified. ■

The basic principle underlying identification is quite simple. The assumption of exchangeability implies that the CDF of \( e|x, 0 \) is the same as the CDF
of \( e|0, x \). When one changes \( x \), the distribution of \( m(0, e) \) changes only because of the change in the distribution of \( e|0, x \), while the distribution of \( m(x, e) \) changes both because of the identical change in the distribution of \( e|x, 0 \) and because \( x \) has a direct effect on \( m(x, 0) \) for each value of \( e \). Consequently, one can isolate the direct effect of \( x \) on the distribution of \( m(x, e) \) by comparing the change in the distribution of \( m(x, e) \) \( x, 0 \) as \( x \) changes to the change in the distribution of \( m(0, e) \) \( 0, x \) as \( x \) changes. Our exchangeability and strong monotonicity assumptions imply that \( F_{y|x,0}(m(x, e)) = F_{y|x}(m(0, e)) \), which allows one to pin down \( m(x, e) \) subject to the normalization \( m(0, x) = e \). The mechanics are roughly as follows. Find the CDF of \( y|0, x \) which is the CDF of \( e|0, x \). Then find the CDF of \( y|x, 0 \) For each value of \( e \), \( m(x, e) \) is the value of \( y \) at which the 2 CDFs are equal.

5.1 Identification of \( g(e|x) \)

Next we derive an expression for \( g(e|x) \) in terms of the pdf and CDF of \( y|x, x_2 \) and of \( y|x \), both of which are identified from data on \( y, x_1, \) and \( x_2 \). Note first that

\[
(5.6) \quad g(e|x) = f_{y|x}(m(x, e)) \frac{\partial m(x, e)}{\partial e},
\]

where \( f_{y|x}(m(x, e)) \) is the pdf of \( y|x \) and we recall that \( x \) is \( x_1 \). Since \( g(e|x, 0) = g(e|0, x) \), it follows that

\[
(5.7) \quad f_{y|x,0}(m(x, e)) \frac{\partial m(x, e)}{\partial e} = f_{y|0,x}(m(0, e)) \frac{\partial m(0, e)}{\partial e}, \text{ i.e.}
\]

\[
(5.8) \quad f_{y|x,0}(m(x, e)) \frac{\partial m(x, e)}{\partial e} = f_{y|0,x}(e). \text{ Hence,}
\]

\[
(5.9) \quad \frac{\partial m(x, e)}{\partial e} = \frac{f_{y|0,x}(e)}{f_{y|x,0}(m(x, e))}. \text{ Or, using (5.5),}
\]

Using (5.5), (5.6) and (5.9), we obtain
Equations (5.5) and (5.10) provide a way to estimate \( m(x, e) \) and \( g(e|x) \) provided that one can estimate the pdf and CDF of \( y|x_1, x_2 \) and the pdf of \( y|x \). With knowledge of \( m(x, e) \) and \( g(e|x) \), it is of course possible to obtain a number of functions summarising the effect of a change in \( x \) on the distribution of \( y \) holding the distribution of \( e \) constant. Consider for example, the average partial derivative \( \beta(x) \) defined in (1.7). We now provide an expression that can serve as the basis for an estimator of \( \beta(x) \).

Differentiating (5.5),

\[
(5.10) \quad \frac{\partial m(x, e)}{\partial x} = \frac{\partial F_{y|x=0, \theta}(F_{y|x=0, \theta}(e))}{\partial x} \frac{\partial F_{y|x=0, \theta}(e)}{\partial x} + \frac{\partial F_{y|x=0, \theta}(F_{y|x=0, \theta}(e))}{\partial x} \frac{\partial F_{y|x=0, \theta}(e)}{\partial x}.
\]

Hence,

\[
(5.11) \quad \int \frac{\partial m(x, e)}{\partial x} g(e|x) \, de = \int \left[ \frac{\partial F_{y|x=0, \theta}(e)}{\partial x} + \frac{\partial F_{y|x=0, \theta}(F_{y|x=0, \theta}(e))}{\partial x} \right] f_{y|x}(m(x, e)) \frac{f_{y|x=0}(e)}{f_{y|x=0}(m(x, e))} \, de
\]

\[
= \int \left[ \frac{\partial F_{y|x=0, \theta}(e)}{\partial x} + \frac{\partial F_{y|x=0, \theta}(F_{y|x=0, \theta}(e))}{\partial x} \right] f_{y|x}(F_{y|x=0, \theta}(F_{y|x=0, \theta}(e))) \frac{f_{y|x=0}(e)}{f_{y|x=0}(F_{y|x=0, \theta}(F_{y|x=0, \theta}(e)))} \, de.
\]

### 5.2 Estimation and Asymptotic Properties

We now proceed to define a particular estimator for \( m(x, e) \). Let \( K \) denote the dimension of \( x \). Let \( d = 1 + 2K \), and let \( f \) denote the pdf of \( (y, x_1, x_2) \).
Then, the kernel estimator for the unconditional and conditional pdf’s and cdf’s are:

\[
\hat{f}_N(t_y, t_1, t_2) = \frac{1}{N_{n_y}} \sum_{i=1}^{N} K\left(\frac{t_y - t_{y,i}}{\sigma}, \frac{t_1 - t_{1,i}}{\sigma}, \frac{t_2 - t_{2,i}}{\sigma}\right) \quad \text{for all } (t_y, t_1, t_2) \in \mathbb{R}^d,
\]

\[
\hat{F}(y, x_1, x_2) = \int_{-\infty}^{y} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \hat{f}_N(t_y, t_1, t_2) \, dt_y \, dt_1 \, dt_2,
\]

\[
\hat{F}_{y|x_1,x_2}(y|x_1, x_2) = \int_{-\infty}^{y} \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \hat{f}_N(t_y, t_1, t_2) \, dt_y \, dt_1 \, dt_2,
\]

and

\[
\hat{f}_N(y|x_1, x_2) = \frac{\int_{-\infty}^{y} \hat{F}_{y|x_1,x_2}(y|x_1, x_2) \, dt_y}{\int_{-\infty}^{y} \hat{F}_{y|x_1,x_2}(y|x_1, x_2) \, dt_y}.
\]

The estimator for \( m(x, e) \) can then be defined by

\[
m(x, e) = \hat{m}^{-1}(\hat{F}_{y|x_1,x_2}(y|x_1, x_2)),
\]

where

\[
\hat{F}_{y|x_1,x_2}(y|x_1, x_2) = \left\{ \begin{array}{ll}
t^u & \text{if } \hat{F}_{y|x_1,x_2}(y|x_1, x_2) \geq t^u \\
\hat{F}_{y|x_1,x_2}(y|x_1, x_2) & \text{if } t_l < \hat{F}_{y|x_1,x_2}(y|x_1, x_2) < t^u \\
t_l & \text{if } \hat{F}_{y|x_1,x_2}(y|x_1, x_2) \leq t_l
\end{array} \right.,
\]

\[
t^u = \sup_e \{ \hat{F}_{y|x_1,x_2}(y|x_1, x_2) \} \quad \text{and} \quad t_l = \inf_e \{ \hat{F}_{y|x_1,x_2}(y|x_1, x_2) \}.
\]

Since we do not restrict \( \hat{F}_{y|x_1} \) and \( \hat{F}_{y|x_2} \) to be strictly increasing, \( \hat{m}(x, e) \) need not be a singleton. To measure the distance between \( \hat{m}(x, e) \) and \( m(x, e) \) we will use the metric \( \rho \) defined by

\[
\rho(m, m') = \sup_{x, e} \left\{ \max \left\{ \sup_{n \in m(x, e)} \inf_{n' \in m'(x, e)} |n - n'|, \sup_{n' \in m'(x, e)} \inf_{n \in m(x, e)} |n - n'| \right\} \right\}
\]

for any functions \( m \) and \( m' \). To establish the asymptotic properties of this estimator, we make the following assumptions:

**ASSUMPTION 1**: The sequence \( \{y_i, x_{1i}, x_{2i}\} \) is a strictly stationary \( \beta \)-mixing sequence satisfying \( k^\delta \beta_k \rightarrow 0 \) as \( k \rightarrow \infty \), for some \( \delta > 1 \).
ASSUMPTION 2': \( f(y, x_1, x_2) \) has a compact support \( \Theta \subset R^d \), \( f(\partial \Theta) = 0 \) where \( \partial \Theta \) denotes the boundary of \( \Theta \), and \( f(y, x, z) \) is continuously differentiable up to order \( g \) where \( g \geq 2d \).

ASSUMPTION 3': The kernel function \( K(\cdot, \cdot, \cdot) \) is an even function, integrates to 1, is of order \( r \) where \( r \) is an even integer satisfying \( d/2 < r < g \), is continuously differentiable up to order \( g \) and its derivatives of order up to \( g \) are in \( L^2(R^d) \).

ASSUMPTION 4': As \( N \rightarrow \infty, \sigma_N \rightarrow 0, N\sigma_N^4 \rightarrow \infty, \sqrt{N}\sigma_N^r \rightarrow 0 \), and \( \sqrt{N}\sigma_N^{2m} \rightarrow \infty \) where \( m \) is an integer such that \( m < r/2 \) and \( m + r \leq g + d \).

ASSUMPTION 5': \( f(0, x), f(x, 0), \) and \( f(F^{-1}_{y,0}(F_{y,0,e}(e)), x, 0) \) are bounded from above and, bounded away from zero from below. \( 0 < F_{y,0,e}(e)) < 1 \) and \( 0 < F_{y,0,u}(t)) < 1 \), where \( t = F^{-1}_{y,0}(F_{y,0,e}(e)) \).

**Theorem 3**: If Assumptions (A5.1)-(A5.5) and 1'\(^{5'}\) are satisfied, then \( \hat{m}(x, e) \) is a consistent estimator of \( m(x, e) \), with respect to \( \rho \), and

\[
\sqrt{N} \sigma_N^r (\hat{m}(x, e) - m(x, e)) \rightarrow N(0, \tilde{V}) \text{ in distribution}
\]

where

\[
\tilde{V} = \left\{ \int \int [F^2(y, x, z)dy]^2dz_1dz_2 \right\} \left\{ \left[ F_{y,0,e}(e)(1 - F_{y,0,e}(e)) \right] + \left[ F_{y,0,e}(e)F_{y,0,u}(m(x, e))(1 - F_{y,0,u}(m(x, e))) \right] \right\}.
\]

Proof: See Appendix.

In a future draft we will propose a specific estimator for \( g \) and analyze its asymptotic properties.
6 Monte Carlo Evidence

We have performed a limited monte carlo analysis of our Regression estimator, which we refer to in the tables as AM-1. Our first set of experiments involves cases in which $y$ is a continuous variable. The cases are nested in the following model:

Model 1

$$y_{ikt} = m(x_{ikt}, e_i, u_{ikt}) = b_0 + b_1x_{ikt} + \gamma x_{ikt}\eta_i + \theta_\eta\eta_i + \theta_\varepsilon\varepsilon_i + u_{ikt}; k = 1, 2; i = 1, 2, \ldots n$$

$$x_{ikt} = x_i + \tilde{x}_{ikt}$$

$$e_i = \theta_\varepsilon x_i + \tilde{e}_i$$

$$\eta_i = \theta_\eta x_i + \tilde{\eta}_i$$

$$x_i \sim N(0, 1); \tilde{x}_{ikt} \sim N(0, 1); \tilde{e}_i \sim N(0, 1); \tilde{\eta}_i \sim N(0, 1); u_{ikt} \sim N(0, 1).$$

The random variables $x_i, \tilde{x}_{ikt}, \tilde{e}_i,$ and $u_{ikt}$ are i.i.d. and mutually independent. Model 1 is a special case of the model in (1.4) that we used to motivate that paper. In applying AM-1 we defined $z_i$ to $(x_{i1} + x_{i2})/2$. Because of the linear relationship among the stochastic components and the fact that $\tilde{e}_{ikt}, x_i,$ and $\tilde{x}_{ikt}$ are all normally distributed, $g(\eta_i, e_i, u_{ikt}|x_{i1}, x_{i2}) = g(\eta_i, e_i, u_{ikt}|(x_{i1} + x_{i2})/2)$, so $z_i$ may be restricted to the mean of $x_{ikt}$ group $i$. In practice, the researcher will not know the distribution of the random components, and so it will be necessary to experiment with additional symmetric functions of $x_{i1}$ and $x_{i2}$. It seems sensible to us begin with the case in which we have the right conditioning variables. For this design $\beta(x) = b_1 + \gamma E(\eta_i|x_{ikt}) = b_1 + \gamma \theta_\eta x_i \theta_{x|x_i}x_{ikt}$ where $\theta_{x|x_i}$ is the population coefficient of the regression of $x_i$ on $x_{ikt}$.

We implement AM-1 using 4 different approaches. The first, which we refer to as AM-1(poly/lr) in the table, is to approximate $E(y_{ikt}|x_{ikt}, z_i)$ as a third order polynomial in $x_{ikt}$ and $z_i$ with interactions up to the third order. We use OLS to estimate the coefficients of the polynomial. We use local linear regression to estimate $h(z_i|x_{ikt})$. The second, called A-M(poly/ker) in the table, combines the polynomial approximation to $E(y_{ikt}|x_{ikt}, z_i)$ with a kernel regression estimator of $h(z_i|x_{ikt})$. AM-1(llr/lr) uses local linear regression to estimate both $E(y_{ikt}|x_{ikt}, z_i)$ and $h(z_i|x_{ikt})$, while AM-1(ker/ker) uses kernel regression to estimate both. We also report of $\beta(x)$ based on applying OLS to $y_{ikt} = b_0 + b_1x_{ikt} + \gamma_2x_{ikt}e_i + e_i + u_{ikt}$ both with and without group specific
intercepts or "fixed effects". We set \( n \) to 1,500. The monte carlo results are based on 750 replications. The numbers in parentheses are the standard deviations of the estimators over the 750 replications.

We consider 3 cases of Model 1. In all cases we set \( b_0 \) to 0, \( b_2 \) to 2, and evaluate the estimates at \( x_{it} = 0.5 \). In Case 1 \( \gamma \) is 0, \( \theta_\eta \) is 0, and \( \theta_\varepsilon \) is 1. Case 1 corresponds to the classic case in which the OLS-fixed effects estimator is used in the literature, and it is unbiased. The fact that \( \gamma = 0 \) implies that \( \beta(x) = b_1 = 2 \). In this design \( x_{i0} \) is correlated with \( \eta_i \) so OLS is inconsistent. Given that \( \gamma = 0 \) and \( \theta_\eta = 0 \) the mean of the OLS estimator is \( b_1 + \theta_\eta \theta_\varepsilon x_{i0} = 2.5 \) where \( \theta_\varepsilon x_{i0} \) is the population coefficient of the regression of \( x_i \) on \( x_{i0} \).

The results of the monte carlo simulations are reported in Table 1. The mean of all 4 versions of the AM-1 estimator are very close to the true value of 1, as is the mean of the OLS-fixed effect estimator. Perhaps surprisingly, the sampling error of all 4 versions of the AM-1 estimator are smaller than that of the OLS-fixed estimator. In contrast, OLS is biased, with a mean of 2.5. The results establish that the regression based estimator performs well in the standard case in which OLS-fixed effects can be employed.

In Case 2 we use the same parameter values as Case 1 except that we change \( \gamma \) from 0 to 1. For this design and parameter values, \( \beta(.5) = 2.25 \). Controlling for group specific intercepts does not control for the \( x_{i0} \eta_i \). All versions of the AM-1 estimator have a mean that is very close to 2.25. In contrast, OLS is badly biased with a mean of 3.49, and OLS-fixed effects also suffers from a substantial bias, with a mean of 2.5. The standard deviation of the various versions of the AM-1 estimator range from .127 to .107, which is somewhat larger than the value of .101 for OLS-fixed effects. However, the AM-1 estimator clearly dominates on mean squared error grounds.

Case 3 is the same as Case 2 except \( \theta_\eta \) is set to 0, \( \theta_\varepsilon \) is set to 1, and \( \theta_\varepsilon \) is set to -1. In this design \( x_{i0} \) is correlated with the random slope parameter \( \eta_i \) and is negatively correlated with a separate random intercept term \( \varepsilon_i \), while in Case 2 \( \eta_i \) is both a random slope and a random intercept. The true value of \( \beta(.5) = 2.25 \). OLS is badly downward biased, with a mean of 1.5. The mean of the OLS-fixed effects estimator is 2.00. In contrast, the AM-1 estimator is essentially unbiased regardless of how we estimate the conditional mean of \( y \) and the conditional density of \( z \). Overall, the results for the AM-1 estimator in the continuous dependent variable case are very encouraging.

In a future draft, we will evaluate the performance of our second estimator.
using a similar design.

We now turn to analysis of a binary choice model. The model is of the form

\[
\begin{align*}
Y_{ik} = & \ I(m(x_{ik}, \varepsilon_i, u_{ik})) = I(b_0 + b_1 x_{ik} + \gamma x_{ik} \eta_i + \theta \eta_i + u_{ik}) ; \ k = 1, 2; 
\end{align*}
\]

\[
\begin{align*}
X_{ik} = & \ x_i + \bar{x}_{ik} \\
\eta_i = & \ \theta \eta_i x_i + \tilde{\eta}_{ik} \\
x_i^- N(0, \sigma_{\eta}^2); \ \bar{x}_{ik}^- N(0, \sigma_{\eta}^2); \ \eta_i^- N(0, \sigma_{\eta}^2); \ u_{ik}^- N(0, 1.5)
\end{align*}
\]

For this design we also set \( z_i \) to \((x_{i1} + x_{i2})/2\) for reasons discussed above. We used two approaches to implementing the AM-1 estimator. In the column labelled AM-1(poly/lir) we take a parametric approach and assume that \( E(Y_{ik}|x_{ik}, z_i) \) is well approximated by a probit model with an index consisting of a third order polynomial in \( x_{ik} \) and \( z_i \) with interactions up the third order, and we use local linear regression to estimate \( h(z_i|x_{ik}) \). In the column labelled AM-1(ker/ker) we use kernel regression to estimate both functions. We also report estimates of \( \beta(x) \) based on a probit model involving a third order polynomial in \( x_{ik} \). As before, we set \( n \) to 1,500 and perform 750 replications. In all the experiments, \( b_0 \) is 0. The results are in Table 2.

In Case 1, which we do not report in the table, \( \gamma = 0 \) and \( \eta_{xz} \) is 0, and \( b_1 = 0 \). We also set \( \sigma_{\eta}^2 \) to 1.5. In this case the probit estimator is consistent and \( \beta(x) = 0 \) for all \( x \). The mean of both the AM-1 estimator and the probit are essentially 0. In Case 2, \( \gamma = 0 \), but \( \eta_{xz} \) is 1. We also set \( \eta_{z} = 1, \sigma_{\eta}^2 = 1.5, \sigma_{\eta}^2 = 1.5, \) and \( \sigma_{\eta}^2 = 0.5 \). This design is essentially a probit model with a group specific error component (\( \eta \)) that is correlated with \( x_{ik} \). We report estimates of \( \beta(x) \) at \( x = -2, x = -1, x = 0, x = 1, \) and \( x = 2 \). Not surprisingly, the probit estimator suffers from a strong positive bias at each value of \( x \). The estimates range between .101 and .118. Since the estimated \( \beta(x) \) is almost constant across value of \( x \) the probit estimates imply that an increase of \( x \) from -2 to 2 would increase the probability that \( y \) is positive by about .4, while the true effect is 0. In contrast, the AM-1 estimator does very well. In the case of AM-1(ker/ker) the means of the estimator at the various \( X \) values lie between -.001 to .001. The standard deviations are larger than those for the probit but are still quite small.
In case 3 \( \eta_i \) is correlated with \( x_{ik} \) and enters as both a random intercept and a random slope. The design is the same as Case 2 except that we change the random slope parameter \( \gamma \) from 0 to 1. In this case, \( \beta(x) \) varies with \( x_{ik} \) for two reasons. The first is that the mean \( \gamma E(\eta_i|x_{ik}) = 1 \cdot .75 \cdot x_{ik} \) of the random slope obviously shifts with \( x_{ik} \), and the second is that the value \( \eta \) influences the likelihood that a given change in \( b_1x_{ik} \) will lead the index \( b_0 + b_1x_{ik} + \gamma x_{ik}\eta_i + \theta_i \eta_i + u_{ik} \) to exceed 0. \( \beta(x) \) is \(-2, 0, \) and \( 2 \) when \( x_{ik} \) is \(-2, 0, \) and \( 2 \) respectively. Once again, the probit estimator is seriously biased at some values of \( x \). On the other hand, A-M does quite well, with a mean of \(-2, 0, \) and \( 2 \) when \( x_{ik} \) is \(-2, 0, \) and \( 2 \) respectively. These results are very encouraging. They suggest that the regression based estimator does provide a viable way to estimate qualitative choice models with random errors that interact with and are correlated with the explanatory variables in the model.

The Monte Carlo evidence is obviously limited by the small set of designs and parameter values we have chosen. However, we find the results to date to be quite encouraging.

7 Conclusion (preliminary and incomplete)

There has been an explosion of empirical studies that use variation among members of a panel to try to deal with endogeneity of explanatory variables. In this paper we provide two estimators for models with nonseparable error terms and endogenous explanatory variables that can be used with panel data. One important class of such models are qualitative choice models with group error components that are correlated with the regressors. Another set of examples consists of random coefficients models in which a group specific random coefficient that is correlated with the regressors. The applied econometrician does not have good options in the literature to estimate such models, except in special cases: We believe that our estimators may provide attractive options in a wide range of situations.

There is a long research agenda beyond the obvious tasks of completing the asymptotic theory and providing some initial Monte Carlo evidence for the second estimator. First, detailed Monte Carlo studies of both estimators are needed, particularly in a multivariate context. It is important to try different distributions of the error terms and the explanatory variables. Second, it is important to gain some experience with the estimators using real data.
With this objective in mind we are in the midst of using the regression based estimator to estimate how the probabilities of time and money transfers between married couples and the parents of the husband and of the wife are affected by the relative incomes, health status, and other characteristics factors of the two sets of parents. We will compare our results to those of Altonji, Hayashi and Kotlikoff (1996), who have applied fixed effects linear probability models and conditional logit models to this problem. There are many practical issues that will arise in implementing the estimators that are best worked out in the context of real world applications.
Proof of Theorem 1:

Define the functional \( \Phi(\cdot) \) by:

\[
\Phi(F) = \int \frac{\partial}{\partial x} E(y | x, z) \, h(z | x) \, dz.
\]

Then,

\[
\beta(x) = \Phi(F) \quad \text{and} \quad \hat{\beta}(x) = \Phi(\hat{F})
\]

where \( \hat{F} \) is the kernel estimator for the cdf \( F \).

Note that

\[
\Phi(F) = \int \frac{\partial}{\partial x} E(y | x, z) \, h(z | x) \, dz
\]

\[
= \int \frac{\partial}{\partial x} \left[ \frac{y f(y, x, z)}{f(y, x, z)} \right] \left[ \frac{f(y, x, z) \, dy}{f(y, x, z)} \right] \, dz
\]

\[
= \int \left[ \frac{y f(y, x, z) \, dy}{f(y, x, z)} \right] \, dz - \int \left[ \frac{f(y, x, z) \, dy \, dy}{f(y, x, z) \, f(y, x, z)} \right] \, dz
\]

and for any \( H \),

\[
\Phi(F + H) - \Phi(F) = \left[ \int \left[ \frac{y (f(x + h) + h(z)) \, dy}{f(x + h(z)) + h(z) \, f(x + h(z))} \right] \, dz - \int \left[ \frac{y f(x) \, dy}{f(x)} \right] \, dz \right]
\]

\[
- \left[ \int \frac{f(x + h(z)) \, dy \, f(y, x, z) \, dy}{f(x + h(z)) + h(z) \, f(x + h(z))} \, dz - \int \frac{f(x) \, dy \, f(y, x, z) \, dy}{f(x + h(z)) \, f(x + h(z))} \, dz \right]
\]

\[
= \int \left( \frac{y h(z) \, dy}{f(x)} \right) f(x) - \left( \frac{y f(x) \, dy}{f(x)} \right) h(z) \, dz
\]

\[
- \int \left( \frac{y f(x) \, dy}{f(x)} \right) \left( \frac{y f(x) \, dy}{f(x)} \right) f(x) \, dz
\]

\[
- \int \left( \frac{y h(z) \, dy}{f(x) \, f(x)} \right) \left( \frac{y f(x) \, dy}{f(x)} \right) f(x) \, dz
\]
Let

\[ \Phi'(F, H) = \int \frac{\int y h dy f(x, z) f(z) f(x) - \int y f dy h(x, z) f(x)}{(f(x)^2 + f(z) h(x))} \, dz \]

\[ - \int \frac{\int f dy \int y f dy f(x, z) f(z) f(x) - \int f dy \int y f dy h(x, z) f(z)}{(f(x)^2 + f(z) h(x))} \, dz \]

\[ + \int \frac{\int f dy \int y f dy h(x, z) f(z) f(x) - \int f dy \int y f dy h(x, z) h(x)}{(f(x)^2 + f(z) h(x))} \, dz \]

\[ + \int \frac{\int y h dy f(x, z) f(z) f(x) - \int y f dy h(x, z) h(x)}{f(x)^2 + f(z) h(x)} \, dz \]

and

\[ R(F, H) = \int \frac{\int y h dy f(x, z) f(z) f(x) - \int y f dy h(x, z) h(x)}{(f(x)^2 + f(z) h(x))} \, dz \]

\[ - \int \frac{\int f dy \int y f dy f(x, z) f(z) f(x) - \int f dy \int y f dy h(x, z) f(z)}{(f(x)^2 + f(z) h(x))} \, dz \]

\[ - \int \frac{\int h dy \int y f dy f(x, z) f(z) f(x) - \int h dy \int y f dy h(x, z) h(x)}{(f(x)^2 + f(z) h(x))} \, dz \]

\[ + \int \frac{\int f dy \int y f dy h(x, z) f(z) f(x) - \int f dy \int y f dy h(x, z) h(x)}{(f(x)^2 + f(z) h(x))} \, dz \]
\[
+ \int f \, dy \int f(z) \, dz \left( f(z) h(z, x) f(z) - \int f \, dy \int f(z) \, dz \right)
\]

For any \( H \), let \( \|H\| \) denote the supremum over the values and derivatives up to order \( 2(K_1 + K_2 + 1) \) of \( H \). Then, letting \( R_k(F, H) \) denote the \( k \)-th coordinate of \( R(F, H) \), for \( k = 1, \ldots, K \), and \( h_x \) and \( f_x \) denote the \( k \)-th coordinate of \( h_x \) and \( f_x \), it follows that

\[
|R_k(F, H)| \leq \int f \, dy \int f(z) \, dz \left( f(z) h(z, x) f(z) - \int f \, dy \int f(z) \, dz \right)
\]
\[ \leq \|H\|^2 [A] \]

where \([A] = \left( \int \frac{y^2 \, dy \, dz}{f(z)^2} \right)^{1/2} + \left( \int \frac{\int y^2 \, dy \, dz}{f(z)^2} \right)^{1/2} \]

\[ + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} \]

\[ + \|H\| \left[ \int \left( \int f_s \, dy \, dz \right)^2 \right]^{1/2} + \|H\| \left[ \int \left( \int f_s \, dy \, dz \right)^2 \right]^{1/2} \]

\[ + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} \]

\[ + \|H\| \left[ \int \left( \int f_s \, dy \, dz \right)^2 \right]^{1/2} + \|H\| \left[ \int \left( \int f_s \, dy \, dz \right)^2 \right]^{1/2} \]

\[ + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} \]

\[ + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} + \left[ \int \frac{\int f_s \, dy \, dz}{f(z)^6} \left( \int \frac{y^2 \, dy}{f(z)^6} \right) \right]^{1/2} \]

Since by Assumptions 2 and 5, all the terms in brackets are finite, and \(\|H\|^2\) is bounded on a neighborhood of \(S\), it follows that

\[ [A] \leq C(F) \quad \text{for some constant } C(F). \]

Hence,

\[ (*) \quad \|R(F, E)\| \leq C(F) \|H\|^2. \]

Also, for each coordinate \(k\) of \(\Phi'(F, H)\)

\[ |\Phi_k'(F, H)| \leq \left( \frac{\int y^2 \, dy \, dz}{f(z)} \right)^{1/2} \left( \frac{\|h_2\| \, dy \, dz}{f(z)} \right)^{1/2} + \frac{\|y \, f_s \, dy \, dz\|}{f(z)^2} \]

32
+ \int \left[ \int f(x) \right] \left( \int y^2 \, dy \right)^{1/2} \left( \int h^2 \, dy \right)^{1/2} \, dz + \int \left( \int h^2 \, dy \right)^{1/2} \left( \int y^2 \, dy \right)^{1/2} \, dz

+ \int \left[ \int f(x) \right] \left( \int y^2 \, dy \right)^{1/2} \left( \int h^2 \, dy \right)^{1/2} \, dz + \left( \int \left( \int f(x) \right)^2 \left( \int y^2 \, dy \right)^2 \, dz \right)^{1/2}

\leq \| H \| [B],

where

[B] = \left( \int \int y^2 \, dy \, dz \right)^{1/2} + \left| \int \int f \, dy \, dz \right| \left( \int \int h^2 \, dy \, dz \right)^{1/2}

+ \int \left[ \int f(x) \right] \left( \int y^2 \, dy \right)^{1/2} \left( \int h^2 \, dy \right)^{1/2} \, dz + \int \left( \int h^2 \, dy \right)^{1/2} \left( \int y^2 \, dy \right)^{1/2} \, dz

+ \int \left[ \int f(x) \right] \left( \int y^2 \, dy \right)^{1/2} \left( \int h^2 \, dy \right)^{1/2} \, dz + \left( \int \left( \int f(x) \right)^2 \left( \int y^2 \, dy \right)^2 \, dz \right)^{1/2}

is, by Assumptions 2 and 5, bounded. Let D(F) be such that [B] \leq D(F).

Then,

(**) |\Phi'(H, F)| \leq \| H \| D(F),

From (*) and (**) it follows that \Phi' is the Hadamard derivative of \Phi at F.

Since, as we have just shown,

|\Phi(F + H) - \Phi(F)| \leq D(F) \| H \| + C(F) \| H \|^2

for any bounded H, it follows that

|\Phi(\hat{F}) - \Phi(F)| \leq D(F) \| \hat{F} - F \| + C(F) \| \hat{F} - F \|^2

Hence, since by Assumptions 1-4, \hat{F} converges in probability to F in the supremum norm, it follows that
$$|\hat{\beta}(x) - \beta(x)| = |\Phi(\hat{F}) - \Phi(F)|$$

converges in probability to 0. Hence,

\(\hat{\beta}(x)\) is a consistent estimator for \(\beta(x)\).

Also,

\[
\Phi'(F, H) = \mathbb{E}_{y \sim (y \sim x \sim z) \sim (y \sim w \sim z)} \left( \frac{f(x, z) - f(x, y) f(x, y) f(x, y)}{f(x, y)^2} \right) dH(y, w, z) \\
- \mathbb{E}_{y \sim (y \sim x \sim z) \sim (y \sim w \sim z)} \left( \frac{f(x, z) - f(x, y) f(x, y) f(x, y)}{f(x, y)^2} \right) dH(y, w, z) \\
+ \mathbb{E}_{y \sim (y \sim x \sim z) \sim (y \sim w \sim z)} \left( \frac{f(x, z) - f(x, y) f(x, y) f(x, y)}{f(x, y)^2} \right) dH(y, w, z)
\]
\[ + \int \int \left[ \int \frac{f_{z}(y,z) \, dy}{f(x,z)^{2}} \left( \int f(y,w,z) \, dy \right) \right] \, l_{z}(y,w,z) \, dH(y,w,z). \]

where for any \((y,z)\), \(l_{z}(y,w,z)\) and \(l_{z}^{(1)}(y,w,z)\) denote, respectively, the Dirac function and the derivative of the Dirac function with respect to \(z\) at the point \((y,x,z)\).

It follows by Assumptions 1-4 and Theorem 3 in Ait-Sahalia (1992) that
\[
\sqrt{N} \sigma_{N}^{(K_{1}/2)+1} \left( \hat{\Phi}(F) - \Phi(F) \right) = \sqrt{N} \sigma_{N}^{(K_{1}/2)+1} \left( \hat{\beta}(x) - \beta(x) \right) \rightarrow N(0,V) \text{ in distribution}
\]

where
\[
V = \left\{ \int \int \left[ y - \left( \int \frac{f(y,w,z) \, dy}{f(x,z)} \right) \right]^{2} \, f(y,x,z) \, dydz \right\} \left\{ \int \left[ \frac{\partial}{\partial z} \left( \int \frac{1}{f(x,z)^{2}} \frac{dH}{dx} x \right) \right]^{2} \, dx \right\} \frac{1}{f(x)^{2}}
\]

\[
= \left\{ \int Var (y|x,z) \, f(x,z) \, dz \right\} \left\{ \int \left[ \frac{\partial}{\partial z} \left( \int \frac{1}{f(x,z)^{2}} \frac{dH}{dx} x \right) \right]^{2} \, dx \right\} \frac{1}{f(x)^{2}}
\]

PROOF OF THEOREM 2 (preliminary and incomplete). The proof follows using arguments similar to those used in Matzkin and Newey (1993), Matzkin (1997), and in the proof of Theorem 1. In particular, to derive the asymptotic distribution, one can show that the Hadamard derivative of the functional:

\[ \Phi(F) = F_{y|x=0} ^{-1} \left( F_{y|x=0} (e) \right) \]

is given by

\[ \Phi'(F,H) = \frac{1}{f_{y|x=0} (m(x,e))} \left( L_{1} + L_{2} \right) \]

where

\[ L_{1} = \frac{\int_{-\infty}^{\infty} h(s,0,x) \, ds}{\int_{-\infty}^{\infty} f(s,0,x) \, ds} - \frac{\int_{-\infty}^{\infty} f(s,0,x) \, ds}{(\int_{-\infty}^{\infty} f(s,0,x) \, ds)^{2}} \int_{-\infty}^{\infty} h(s,0,x) \, ds \]

35
\begin{align*}
&= \int \left[ \frac{1[s \leq e] f(0,x) - \int_{-\infty}^{e} f(s,0,x) ds}{f(m(x,e),0,x) f(0,x)^2} \right] 1_{s,0,x}(s, w, z) h(s, w, z) \, ds \, dw \, dz \\
\text{and} \\
L_2 &= \int_{-\infty}^{m(x,e)} \frac{f(s,0,x) ds}{f(m(x,e),0,x) f(0,x)^2} \int_{-\infty}^{\infty} h(s, x, 0) \, ds - \int_{-\infty}^{m(x,e)} \frac{f(x,0,x) ds}{f(m(x,e),0,x) f(0,x)^2} \\
&= \int \left[ \frac{\int_{-\infty}^{m(x,e)} f(s,0,x) ds - 1[s \leq m(x,e)] f(x,0,x)}{f(m(x,e),0,x) f(0,x)^2} \right] 1_{s,0,x}(s, w, z) h(s, w, z) \, ds \, dw \, dz
\end{align*}

By Ait-Sahalia (1992), it then follows that

\[ \sqrt{N} \sigma_{x}^{2} (\hat{m}(x,e) - m(x,e)) = \sqrt{N} \sigma_{x}^{2} (\Phi(F) - \Phi(F)) \rightarrow N(0, \bar{V}) \] in distribution

where

\[ \bar{V} = \left\{ \int \int \int K(y, w, z) dy \, dw \, dz \right\} L \]

and

\begin{align*}
L &= \int \left[ \frac{1[s \leq e, w=0, z=x] f(0,x) - \int_{-\infty}^{e} f(s,0,x) ds}{f(m(x,e),0,x) f(0,x)^2} \right] f(s, 0, x) \, ds \\
&\quad + \int \left[ \frac{\int_{-\infty}^{m(x,e)} f(s,0,x) ds - 1[s \leq m(x,e), w=x, z=0] f(x,0,x)}{f(m(x,e),0,x) f(0,x)^2} \right] f(s, x, 0) \, ds \\
&= \left\{ \frac{F_{y \mid 0,x}(e) (1 - F_{y \mid 0,x}(e)) + F_{y \mid 0,x}(m(x,e)) (1 - F_{y \mid 0,x}(m(x,e)))}{F_{y \mid 0,x}(m(x,e)) f(0,x)^2} \right\}
\end{align*}
9 References (incomplete)


Honoré, B. E., “Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects,” Econometrica,
60, pp. 533-565.


Table 1: Monte Carlo Simulations of AM-1, Continuous Dependent Variable Case
(750 replications). Slopes evaluated at \( x_{ik} = 0.5 \)

\[
y_{it} = b_0 + b_1 x_{it} + \alpha x_{ik} \eta_i \eta_i + \theta_i \xi_i + \theta_{i*} \varepsilon_i + \eta_i (k = 1, 2; i = 1, 2, \ldots 1500)
\]

\[
x_{ik} = x_i + z_{ik}; \quad \varepsilon_i = \theta_{i*} x_i + \zeta_i; \quad \eta_i = L x_i + \tilde{\eta}_{ik};
\]

\[
x_i \sim N(0,1); \quad Z_{ik} \sim N(0,1); \quad \varepsilon_i \sim N(0,1); \quad \eta_i \sim N(0,1); \quad \eta_{ik} \sim N(0,1)
\]

**Case 1:** \( y_{it} = 0 + 2 x_{it} + 0 x_{ik} \eta_i + 1 \eta_i + 0 \varepsilon + \eta_{ik}; \quad \theta_{i*} = 1 \)

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>True value</th>
<th>AM-1 (poly/lr)</th>
<th>AM-1 (poly/ker)</th>
<th>AM-1 (lrr/lrr)</th>
<th>AM-1 (ker/ker)</th>
<th>OLS Fixed Effects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of ( \hat{\beta}(0.5) )</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.00</td>
<td>2.50</td>
<td>2.00</td>
</tr>
<tr>
<td>Mean of ( \hat{\beta}(0.5) )</td>
<td>(0.06)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.04)</td>
<td>(0.07)</td>
<td>(0.09)</td>
</tr>
</tbody>
</table>

**Case 2:** \( y_{it} = 0 + 2 x_{it} + 1 x_{ik} \eta_i + 1 \eta_i + 0 \varepsilon + \eta_{ik} \)

| Mean of \( \hat{\beta}(0.5) \) | 2.25 | 2.25 | 2.25 | 2.25 | 2.50 | 2.00 |
| Mean of \( \hat{\beta}(0.5) \) | (0.12) | (0.12) | (0.10) | (0.10) | (0.07) | (0.09) |

**Case 3:** \( y_{it} = 0 + 2 x_{it} + 1 x_{ik} \eta_i + 1 \eta_i + 0 \varepsilon + \eta_{ik} \)

| Mean of \( \hat{\beta}(0.5) \) | 2.25 | 2.25 | 2.25 | 2.25 | 2.50 | 2.01 |
| Mean of \( \hat{\beta}(0.5) \) | (0.10) | (0.11) | (0.09) | (0.01) | (0.07) | (0.06) |

Notes: AM-1(poly/lr) uses a third order polynomial in \( x_{ik} \) and \( z_i \) to approximate \( E(y_{ik}|x_{ik}, z_i) \) and a local linear regression to approximate \( h(z_i|x_{ik}) \). AM-1(poly/ker) uses a third order polynomial in \( x_{ik} \) and \( z_i \) to approximate \( h(z_i|x_{ik}) \). AM-1(lrr/lrr) uses kernel regression to estimate \( h(z_i|x_{ik}) \). AM-1(lrr/lrr) uses local linear regression to estimate for both parts of the estimation, and AM-1(ker/ker) uses kernel regression. OLS-Fixed Effects is OLS applied to a model that contains a separate intercept for each group \( i \). Technical details are given in footnote ___.
Table 2: Monte Carlo Simulations of AM-1 in the Case of a Binary Dependent Variable (750 replications).

\[ y_{it} = I(b_0 + b_1 x_{it} + \pi_{it} \theta_{it} + \theta_{it} + \eta_{it} + \mu_{it}); \quad k = 1,2; \quad i = 1,2,..,1500 \]

\[ x_{it} = x_i + \varepsilon_{it}; \quad \eta_{it} = \theta_{it} + \varepsilon_{it}; \]

\[ x_i \sim N(0, \sigma^2_x); \quad \varepsilon_{it} \sim N(0, \sigma^2_{\varepsilon}); \quad \eta_{it} \sim N(0, \sigma^2_{\eta}); \quad \mu_{it} \sim N(0, \sigma^2_{\mu}) \]

Case 1: \[ y_{it} = I(0 + 0x_i + 0x_{it} \eta_{it} + 0x_{it} + 0\varepsilon_{it} + 0\mu_{it}); \quad \theta_{it} = 1; \quad \sigma^2_x = 1.5; \quad \sigma^2_{\varepsilon} = 1.5; \quad \sigma^2_{\eta} = 0.5; \]

<table>
<thead>
<tr>
<th>Estimation Method</th>
<th>True value (poly. lin)</th>
<th>AM-1 (pol. lin)</th>
<th>AM-1 (pol. ker)</th>
<th>Probit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (sd) of $\hat{\beta}(2.0)$</td>
<td>0.0</td>
<td>0.002</td>
<td>-0.101</td>
<td>0.114</td>
</tr>
<tr>
<td>(0.023)</td>
<td>(0.015)</td>
<td>(0.022)</td>
<td>(0.009)</td>
<td></td>
</tr>
<tr>
<td>Mean (sd) of $\hat{\beta}(1.0)$</td>
<td>0.0</td>
<td>0.001</td>
<td>-0.001</td>
<td>0.114</td>
</tr>
<tr>
<td>(0.011)</td>
<td>(0.021)</td>
<td>(0.021)</td>
<td>(0.074)</td>
<td></td>
</tr>
<tr>
<td>Mean (sd) of $\hat{\beta}(0.0)$</td>
<td>0.0</td>
<td>0.000</td>
<td>0.000</td>
<td>0.118</td>
</tr>
<tr>
<td>(0.011)</td>
<td>(0.021)</td>
<td>(0.021)</td>
<td>(0.074)</td>
<td></td>
</tr>
<tr>
<td>Mean (sd) of $\hat{\beta}(-1.0)$</td>
<td>0.0</td>
<td>0.000</td>
<td>0.001</td>
<td>0.114</td>
</tr>
<tr>
<td>(0.015)</td>
<td>(0.021)</td>
<td>(0.021)</td>
<td>(0.009)</td>
<td></td>
</tr>
<tr>
<td>Mean (sd) of $\hat{\beta}(-2.0)$</td>
<td>0.0</td>
<td>0.001</td>
<td>-0.001</td>
<td>0.101</td>
</tr>
<tr>
<td>(0.021)</td>
<td>(0.025)</td>
<td>(0.025)</td>
<td>(0.011)</td>
<td></td>
</tr>
</tbody>
</table>