TECHNICAL WORKING PAPER SERIES

PANEL DATA ESTIMATORS FOR NONSEPARABLE MODELS WITH ENDOGENOUS REGRESSORS

Joseph G. Altonji
Rosa L. Matzkin

Technical Working Paper 267
http://www.nber.org/papers/T0267

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2001

We thank Elie Tamer for excellent research assistance and the Institute for Policy Research, Northwestern University and the National Science Foundation (SBR-9512009 (Altonji) and SBR-9410182 (Matzkin)) for research support. We are also grateful to participants in seminars at Northwestern University (May 1997), UCLA (November 1997), UC-Berkeley (November 1997), University of Chicago (April 1998), University of San Andrés (July 1998), University Di Tella (August 2000), and the NSF conference on Semiparametric and Nonparametric Methods (October 1998) for many helpful comments. The views expressed in this paper are those of the authors and not necessarily those of the National Bureau of Economic Research.

Panel Data Estimators for Nonseparable Models with Endogenous Regressors
Joseph G. Altonji and Rosa L. Matzkin
NBER Technical Working Paper No. 267
March 2001
JEL No. C30, C33, C35

ABSTRACT

We propose two new estimators for a wide class of panel data models with nonseparable error terms and endogenous explanatory variables. The first estimator covers qualitative choice models and both estimators cover models with continuous dependent variables. The first estimator requires the existence of a vector z such that the density of the error term does not depend on the explanatory variables once one conditions on z. In some panel data cases we may find z by making the assumption that the distribution of the error term conditional on the vector of the explanatory variables for each "cross-section" unit in the panel is exchangeable in the values of those explanatory variables. This situation may be realistic, in particular, when each unit is a group of individuals, so that the observations are across groups and for different individuals in each group. The basic idea is to first estimate the slope of the mean of the dependent variable conditional on both the explanatory variable and z and then undo the effect of conditioning on z by taking the average of the slope over the distribution of z conditional on a particular value of the explanatory variable. We also extend the procedure to the case in which the explanatory variable is endogenous conditional on z but an instrumental variable is available. The second estimator is based on the assumption that the error distribution is exchangeable in the explanatory variables of each unit. It applies to models that are monotone in the error term. A shift in the value of an explanatory variable for member 1 of a group has both a direct effect on the distribution of the dependent variable for member 1 and an indirect effect through the distribution of the error. A shift in the explanatory variable has an indirect effect on the dependent variable for other members of the panel but no direct effect. We isolate the direct effect by comparing the effect of the explanatory variable on the distribution of the dependent variable for member 1 to its effect on the distribution for the other panel members.

Joseph G. Altonji
Department of Economics
Northwestern University
Evanston, IL 60208
847-491-8218
altonji@northwestern.edu

Rosa L. Matzkin
Department of Economics
Northwestern University
Evanston, IL 60208
847-491-8220
matzkin@northwestern.edu

1 Introduction

In this paper we develop estimators for panel data models with nonseparable error terms and endogenous explanatory variables. Examples of data that can be applied to such models are samples of siblings, observations on individuals or firms through time, and data on individuals who grew up in the same neighborhood or attended the same high school. The methods are 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 (1998)) or the effects of the U.S. pre-school program Head Start on various outcomes (Currie and Thomas (1995)), the use of siblings to study transfers of time and money to and from parents (Altonji, Hayashi and Kotlikoff (2000), Rosenzweig and Wolpin (1994 a and b), and the use of panel data to study the effects of minimum wages on employment (Currie and Fallick (1996). The methods cover a wide range of models, including binary choice models.

To be more specific, consider the model (1.1)

$$(1.1)\ y_{ik} = m(x_{ik}, \varepsilon_i, u_{ik}),\ i = 1...n,\ k = 1...K_i.$$

where y_{ik} is an outcome of person k from group i , x_{ik} is a $1 \times J_1$ vector of observed variables, ε_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}, ε_i, and u_{ik}. The index k may be an element of x_{ik}, which means that the effect of u_{ik} and ε_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. Both ε_i and u_{ik} may be vectors.

To give a simple example, consider Aaronson's (1998) 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 ik 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

$$\begin{aligned} m(x_{ik}, \varepsilon_i, u_{ik}) &= 1 \text{ if } \Psi(x_{ik}, \varepsilon_i, u_{ik}) = x_{ik}\Gamma + \varepsilon_i + u_{ik} > 0 \\ &= 0 \text{ otherwise.} \end{aligned}$$

or

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

where the function $I(.)$ 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 Ψ be separable in x_{ik}, ε_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

$$
\begin{aligned}
(1.2) E(y_{ik}|x_{ik}) &= \int_{\varepsilon_i} \int_{u_{ik}} [m(x_{ik}, \varepsilon_i, u_{ik})] g(u_{ik}, \varepsilon_i | x_{ik}) d\varepsilon_i du_{ik} \\
&= \int_{\varepsilon_i} \int_{-\Gamma x_{ik} - \varepsilon_i}^{\infty} g(u_{ik}, \varepsilon_i | x_{ik}) d\varepsilon_i du_{ik}
\end{aligned}
$$

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 $g(u_{ik}, \varepsilon_i | x_{ik})$ constant. Call this derivative $\beta(x_{ik})$, where

$$
(1.3)\ \beta(x_{ik}) = \int_{\varepsilon_i} \Gamma \cdot g(-\Gamma x_{ik} - \varepsilon_i, \varepsilon_i | x_{ik}) d\varepsilon_i
$$

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 nonparametric estimators for binary choice models, such as the probit and logit, as well as standard nonparametric estimators provide biased estimates of $\beta(x_{ik})$ when ε_i is correlated with x_{ik} The estimate of $\beta(x_{ik})$ will pick up part of the effect of ε_i on y_{ik}.

Aaronson (1998) 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. Our Estimator 1 identifies the parameter of interest and uses data on all groups, even groups for which x_{ik} is observed for more than one member but y_{ik} is observed for only 1 member.

To give a second special case to which our methods apply, consider the model

$$
(1.4)\ y_{ik} = m(x_{ik}, \varepsilon_i, u_{ik}) = x_{ik}\beta_1 + H(x_{ik}, \varepsilon_i) + \varepsilon_i + u_{ik};
$$

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

$$
(1.5)\ \beta(x_{ik}) = \int_{\varepsilon_i} \int_{u_{ik}} [m_{xk}(x_{ik}, \varepsilon_i, u_{ik})] g(u_{ik}, \varepsilon_i | x_{ik}) d\varepsilon_i du_{ik}\ .
$$

There is a huge literature that assumes that $H(x_{ik}, \varepsilon_i)$ is 0 and deals with the correlation between ε_i and x_{ik} by controlling for ε_i with a group specific intercept. These "fixed effects" estimators don't work when the impact of x_{ik} on y_{ik} depends on ε_i. 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 $[x_{i1}, ..., x_{iK}]'$. Estimator 1 requires the existence of a vector z such that $\partial g(u_{ik}, \varepsilon_i | x_{ik}, z)/\partial x_{ik} = 0$. To find one such vector z, we make the assumption that the distribution of u_{ik} and ε_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 $g(u_{ik}, \varepsilon_i | x_{i1}...x_{i_K})$. That is,

$$\begin{aligned} \textit{Assumption A1.1} \quad &: \quad g(u_{ik}, \varepsilon_i | x_{i1}...x_{i_K}) = g(u_{ik}, \varepsilon_i | x_{ik_1}, x_{ik_2}...x_{ik_K}) \text{ for} \\ k_j \quad &\in \quad \{1, 2.., K\}, k_j \neq k_{j'} \end{aligned}$$

For example, the assumption implies that $g(u_{ik}, \varepsilon_i | x_{i1}...x_{iK}) = g(u_{ik}, \varepsilon_i | x_{iK}...x_{i1})$. That is, for any u_{ik}, ε_i, $g(u_{ik}, \varepsilon_i | x_{i1}...x_{iK})$ is a symmetric function of $x_{i1}...x_{iK}$, which in our context means the function is invariant to permutations of the $x_{i1}...x_{iK}$.[1] 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 and/or age is an element of x_{ik}. Exchangeability is essential to our second estimator but not to our first estimator.[2]

Estimator 1 is based on the conditional expectation function $E(y_{ik}|x_i)$, and for this reason we sometimes 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 ε_i and u_{i1}.The variable $[x_{i2}..x_{iK}]$ only have an indirect impact through their effects on the distribution of ε_i and u_{i1}, both of which may be vectors. Exchangeability restricts the distribution of u_{i1} and ε_i to depend on x_i only through a vector of L exchangeable functions $z(x_i)$ of x_i. This implies that $g(u_{ik}, \varepsilon_i | x_{i1}...x_{iK}) = g(u_{ik}, \varepsilon_i | z(x_i))$ and the mean $E(y_{ik}|x_i)$ can be written as $E(y_{ik}|x_{i1}, z(x_i))$. If (1) the $z(.)$ vector of functions is known and (2) x_{ik} and the elements of the vector $z(x_i)$ vary sufficiently, then the partial derivative $\partial E(y_{ik}|x_{i1}, z(x_i))/\partial x_{i1}$ is identified. Since it is possible to estimate the distribution of $z(x_i)$ conditional on x_{i1}, one can recover the parameter of interest $\beta(x_{i1})$ from the estimator of $\partial E(y_{ik}|x_{i1}, z(x_i))/\partial x_{i1}$by integrating this derivative over the distribution of $z(x_i)$ conditional on x_{ik}. Exchangeability is a natural assumption in many panel data situations, but it is not the only condition under which one can find a vector $z(.)$

[1]A sufficient condition for A1.1 is that the joint density of $(u_{ik}, \varepsilon_i, x_{i1}...x_{ik})$ and the marginal density of $(x_{i1}, x_{i2}, ...x_{iK})$ are both exchangeable in $(x_{i1}, x_{i2}, ..., x_{iK})$. However, it is easy to show using an example in which $(u_{ik}, \varepsilon_i, x_{i1}...x_{ik})$ is jointly normal that $g(u_{ik}, \varepsilon_i | x_{i1}...x_{iK})$ can be exchangeable in $(x_{i1}, x_{i2}, ..., x_{iK})$ even if the joint density and the marginal density are not. See de Finetti (1975, Chapter 11) for a brief introduction to the concept of exchangeability.

[2]Note under A1.1 the distributions of both u_{ik} and ε_i may depend on x_{ik}, in contrast to some panel data models in which only the common error component ε_i depends on x_{ik}. The distinction between the elements of u_{ik} and the elements of ε_i plays no essential role in our paper. We maintain it to ease comparisons to the literature.

that satisfies (1) and (2) and has the property that $g(u_{ik}, \varepsilon_i | x_{i1}...x_{iK}) = g(u_{ik}, \varepsilon_i | z(x_i))$; any such vector $z(.)$ can be used with our first estimator. We also extend the procedure to the case in which x_{ik} is endogenous conditional on $z(.)$, but an instrumental variable A_{ik} is available. This case may be handled by augmenting the control variables $z(.)$ with residuals from the regression of x_{ik} on A_{ik} and $z(.)$. This estimator provides an alternative to the frequently used but inconsistent "fixed effects-IV" linear probability model.

In contrast to the first estimator, Estimator 2 relies heavily on A1.1. It also involves a somewhat different set of assumptions. It does not use or require knowledge of the $z(\cdot)$ 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 $m(x_{ik}, e_{ik})$, where $e_{ik} = \Upsilon(\varepsilon_i, u_{ik})$ and $\Upsilon(.,.)$ is a real valued, continuous, not necessarily known, scalar valued function and $m(x_{ik}, e_{ik})$ is strictly monotone in e_{ik}. 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 $\beta(x_{ik})$ 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_{i1} alters the distribution of $y_{i1} = m(x_{i1}, e_{i1})$ by shifting $m(x_{i1}, e_{i1})$ for a given value of e_{i1} and by shifting the distribution e_{i1}. A shift in x_{i1} alters the distribution of $m(x_{i2}, e_{i2})$ only by shifting the distribution of e_{i2}. Exchangeability implies that the x_{i1} and x_{i2} have the same effect on the distribution of $e_{i1}$and e_{i2}. Consequently, one can isolate the direct effect of x_{i1} on the distribution of $m(x_{i1}, e_{i1})$ by comparing the change in the distribution of $m(x_{i1}, e_{i1})$ conditional on (x_{i1}, x_{i2}) as x_{i1} changes to the change in the distribution of $m(x_{i2}, e_{i2})$ conditional on (x_{i2}, x_{i1}) as x_{i1} 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 mean partial effect of x_{ik} on 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).[3] The fixed effects probit model is sometimes used to estimate Γ 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.[4]

Manski (1987) provides a way to estimate Γ up to scale under more general assumptions

[3]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}.

[4]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 ε_i given x_{ik} . However, the MLE estimates of ε_i are unbounded when y is the same for all group members, so one cannot obtain an estimate of the distribution of $\varepsilon_i|x_{ik}$ without making assumptions about this distribution. The same is true in the case of the conditional logit.

than the conditional logit. He places no restrictions on the distribution of ε_i and assumes that the distribution of $u_{i1}|\varepsilon_i, x_{i1}, x_{i2}$, is the same as the distribution of $u_{i2}|\varepsilon_i, x_{i1}, x_{i2}$. He proposes a maximum score estimator that exploits that fact that $sgn(E((y_{i2} - y_{i1})\,|x_{i1}, x_{i2}) = sgn(x_{i2}\Gamma - x_{i1}\Gamma)$ where $sgn(.)$ is -1 if the argument is negative and 1 if it is positive. In contrast to Manski's estimator, our approach requires a-priori information about the distribution of $\varepsilon_i|x_{i1}, x_{i2}$. 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 Γ 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 their current form cannot accommodate dynamics in the model, which are addressed in recent papers by Honoré and Kyriazidou (2000) 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 and applies parametric random effects approaches to estimating Γ up to scale in (1.1a). Assume that u_{ik} is normal, identically distributed across k, and independent of x_i. Also assume that ε_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 Γ up to scale by adding $f(x_i;\theta)$ to a probit model for each k and jointly estimating θ and Γ while imposing cross restrictions across the models for each k. One may also recover the partial effect of x_{ik} on the probability that y_{ik} equals 1. The main disadvantages of this approach relative to ours is that it requires the assumptions that u_{ik} and $\varepsilon_i|x_i$ are normal and additively separable from x_{ik} in $m(x_{ik}, \varepsilon_i, u_{ik})$,as in (1.1a).

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 β_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}.[5]

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 $z(x_i)$ but an instrumental variable is available. In section 5 we derive the second estimator and provide results on its asymptotic properties. In section 6 we present some limited but encouraging Monte Carlo evidence on the performance of the "Regression" estimator. In section 7 we provide some concluding remarks.

[5]Thus 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 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 the property that for all $x_{i1}...x_{iK}$, $g(u_{ik}, \varepsilon_i|x_{i1}...x_{iK}) = g(u_{ik}, e_i|z_i^1, ...,z_i^L)$. In the case that $g(u_{ik}, \varepsilon_i|x_{i1}...x_{iK})$ is exchangeable in $(x_{i1}...x_{iK})$, one might find exchangeable functions $z^1(x_{i1}...x_{iK}), z^2(x_{i1}...x_{iK}), ..., z^L(x_{i1}...x_{iK})$ satisfying this property. 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_i^1 might be the mean of $x_{i1}..x_{iK}$ for family i and z_i^2 might be the average over k of $(x_{ik} - z_{i1})^2$. As we noted in the introduction, assumption (A1.1) that $g(u_{ik}, \varepsilon_i|x_{i1}...x_{iK})$ is exchangeable in $x_{i1}...x_{iK}$ means that without loss of generality we can write $g(u_{ik}, \varepsilon_i|x_{i1}...x_{iK})$ as $g(u_{ik}, e_i|z_i^1, ...,z_i^L)$, where $z_i^1, ...,z_i^L$ are exchangeable functions $z^1(x_{i1}...x_{iK}), z^2(x_{i1}...x_{iK}), ..., z^L(x_{i1}...x_{iK})$ of $(x_{i1}...x_{iK})$. Let z_i be the vector of z_i^ℓ variables for family i. The first estimator requires the following high level assumptions:

Assumption 2.0. $\partial g(u_{ik}, e_i|x_{ik},z_i^1, ...,z_i^L)/\partial x_{ik} = 0$.

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_i^1,...., z_i^L 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 if a-priori information 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

$$(2.1)\ E(y_1|x_1, z) = \int_\varepsilon \int_{u_1} m(x_1, \varepsilon, u_1) g(u_1, \varepsilon|x_1, z) d\varepsilon\, du_1$$

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

$$(2.2)\ E_{x_1}(y_1|x_1, z) \equiv \frac{\partial}{\partial x_1} E(y_1|x_1, z)$$

and $h(z|x_1)$, the conditional distribution of z given x_1. The distribution $h(z|x_1) = h(z, x_1)/h(x_1)$ can be estimated from observations on z and x_1 for the cross section of groups i. When the function m is differentiable, the derivative with respect to x_1 is

$$(2.3)\ E_{x_1}(y_1|x_1, z) = \int_{u_1} \int_\varepsilon m_{x1}(x_1, \varepsilon, u_1) g(u_1, \varepsilon|x_1, z) d\varepsilon\, du_1$$

because

$$\int_{u_1} \int_\varepsilon m(x_1, \varepsilon, u_1)\ g_{x1}(u_1, \varepsilon|x_1, z) d\varepsilon\, du_1 = 0\ .$$

The form of (2.3) is analogous to (1.3) in the binary choice case with a linear index function, although our methods apply to essentially arbitrary forms of the index function in binary choice cases. In this case

$$(2.4)\ E_{x_1}(y_1|x_1,z) \ = \int_{\varepsilon} \ \Gamma \cdot g(-\Gamma x_1 - \varepsilon, \varepsilon|x_1, z)d\varepsilon.$$

Note that $E_{x_1}(y_1|x_1,z)$ differs from $\beta(x_1)$ because the distribution of u_1 and ε 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

$$(2.5)\ g(u_1,\varepsilon|x_1) = \int_z g(u_1,\varepsilon|x_1,z)h(z|x_1)dz$$

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)

$$(2.6)\ \int_z E_{x_1}(y_1|x_1,z)h(z|x_1) = \int_z \int_{u_1} \int_{\varepsilon} m_{x1}(x_1,\varepsilon,u_1)g(u_1,\varepsilon|x_1,z)h(z|x_1)\,d\varepsilon\,du_1\,dz.$$

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,

$$(2.7)\ \beta(x_1) = \int_z E_{x_1}(y_1|x_1,z)h(z|x_1)dz$$

The above equation forms the basis of Estimator 1. The estimator is obtained by substituting parametric or nonparametric estimators of the components of the right hand side of (2.7) into the equation.[6] 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.

The key assumption, aside from the instrumental variable condition 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. In the case in which Assumption 2.0 is satisfied because of exchangeability, 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 which x_k is a scalar and $K=2$ (i.e., there are observations on 2 children per family). Continuous exchangeable functions of the elements of $x_i = \{x_{i1},\ x_{i2}\}$ may be approximated arbitrarily closely by a function of $z_i^1 = x_{i1}+x_{i2}$ and $z_i^2 = |x_{i1}-x_{i2}|$.[7]Thus, conditioning the

[6]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.

[7]A sketch of a proof follows. Let $z_i^3 = x_{i1}x_{i2}$. One may verify that $z_i^3 = .25\{(z_i^1)^2 - (z_i^2)^2\}$. Any continuous function $h(x_1,x_2)$ can be approximated arbitrarily closely by a polynomial in x_1 and x_2. Exchangeability restricts the coefficients on the term of the polynomial that involve $(x_1)^j(x_2)^{j'}$to equal the coefficient on the term involving $(x_1)^{j'}(x_2)^j$. It also implies equality of the coefficients on $(x_1)^j$ and $(x_2)^j$. Consequently for N possibly large
$h(x_1,x_2) \approx a_0 + \sum_{j=1}^{N-1}\sum_{j'=1}^{N-j} a_{1jj'}[(x_1)^j(x_2)^{j'}] + \sum_{j=1}^{N} a_{2j}((x_1)^j + (x_2)^j)$

distribution of $(\varepsilon_i,\ u_{ik})$ on z_{i1} and z_{i2} is general enough. However, $z_i^2 = |2x_{i2} - z_{i1}|$. This dependence among z_i^1, z_i^2, and x_{i1} 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 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 $(\varepsilon, u_k)|x_i$ depends only on $z_{i1} = (x_{i1} + x_{i2} + x_3 + x_k)/4$ and $\sum |x_{ik} - z_{i1}|$. 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 in panel data applications, particularly for rare events, that all group members have the same value for y_{ik}. [8]

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

---

where $a_{1jj'} = a_{1j'j}$ for all j, j' and the approximation is arbitrarily close. It is straightfoward but tedious to show that one may express terms of the form $(x_1)^j + (x_2)^j$as a linear combination of powers of z_i^1, powers of z_i^3, and products of powers of the two. The double sum in the middle term yields $N/2$ terms of the form $a_{jj}(x_1)^j(x_2)^j$, which are equal to $a_{jj}(z_i^3)^j$. Because $a_{1jj'} = a_{1j'j}$, it also yields a bunch of terms of the form $a_{1jj'}((x_1)^j(x_2)^{j'} + (x_1)^{j'}(x_2)^j)$, which are equal to $a_{1jj'}(z_i^3)^{\min(j,j')}[(x_1)^{|j'-j|} + (x_2)^{|j'-j|}]$. Since $(x_1)^{|j'-j|} + (x_2)^{|j'-j|}$ can be expressed in terms of $z_i^1$and z_i^3, and z_i^3 may be expressed in terms of z_i^1 and z_i^2, it follows that any continous exchangeable function of x_1 and x_2 may be approximated arbitrarily closely by a function of z_i^1 and z_i^2. We do not know if there is a useful generalization of this result when K > 2.

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

model with an additive family fixed effect is a special case of our model (1.1). Consider the model

$$(2.8)\ y_{ik} = x_{ik}\beta + \varepsilon_i + u_{ik}.$$

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

$$(2.9)\ y_{ik} = x_{ik}\beta + x_i\lambda + u_{ik} + v_i,\ k = 1...K$$

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

$$(2.10)\ y_{ik} = x_{ik}\beta + f(z_i; \beta_1) + u_{ik} + v_i$$

where z_i is a vector of exchangeable functions of x_i and $f(.; \beta_1)$ is a function with parameter vector β_1, such as a polynomial. In the above model $m_{x_{ik}}(x_{ik}, u_{ik}, \varepsilon_i)$, the partial derivative of y_{ik} with respect to x_{ik} is a constant β, 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 z_i^1, the sum of the elements of x_i and $f(.;.)$ is linear. In this case, (2.10) is equivalent to (2.9) with the restriction that the elements of λ are all equal. With these restrictions one does not need to have data on all of the y_{ik} to identify β 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

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

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.

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

$$\hat{f}_N(t_y, t_x, t_z) = \frac{1}{N\sigma_N^d} \sum_{i=1}^N K(\frac{u_y - t_y}{\sigma}, \frac{u_x - t_x}{\sigma}, \frac{u_z - t_z}{\sigma}) \quad .$$

The following assumptions will be needed:

ASSUMPTION 1: The sequence $\{y_i, x_i, z_i\}$ is i.i.d.

ASSUMPTION 2: $f(\cdot,\cdot,\cdot)$ has compact support $\Theta \subset R^d$ and is continuously differentiable up to the order $g = 1 + s$ for some even $s > 0$.

ASSUMPTION 3: The kernel function $K(\cdot,\cdot,\cdot)$ is continuously differentiable, K vanishes outside a compact set, $\int K(y,x,z)dydxdz = 1$, and K is a kernel of order s.

ASSUMPTION 4: As $N \to \infty$, $\ln(N)/N\sigma_N^{d+2} \to 0$ and $\sigma_N^s \sqrt{N\sigma_N^{Q+2}} \to 0$.

ASSUMPTION 5: $f(x) > 0$ and $\int \frac{1}{f(x,z)^2} dz$ is bounded.

Theorem 1 *: If Assumptions 2.0-2.2 and Assumptions 1-5 are satisfied, then $\hat{\beta}(x)$ is a consistent estimator of $\beta(x)$ and*

$$\sqrt{N}\, \sigma_N^{(K_1/2)+1} \left(\hat{\beta}(x) - \beta(x)\right) \to N(0, V) \text{ in distribution where}$$

$$V = \left\{\int Var\,(y|x,z)\, \frac{f(z|x)}{f(x)}\, dz\right\} \left\{\int \left(\int\int \frac{\partial K(y,x,z)}{\partial x} dydz\right) \left(\int\int \frac{\partial K(y,x,z)}{\partial x} dydz\right)' dx\right\}.$$

PROOF: See Appendix A.

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 ε_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.[9] Here we extend the Regression estimator to handle correlation between u_{ik} and x_{ik} when an instrumental variable A_{ik} is available. Write x_{ik} as $x_{ik} = \Upsilon(A_{ik}, z_i) + \xi_{ik}$, where $\Upsilon(A_{ik}, z_i)$ is $E(x_{ik}|A_{ik}, z_i)$ and ξ_{ik} is defined accordingly. Assume

(A1.1a) A_{ik} is independent of ε_i, u_{ik} conditional on z_i, ξ_{ik}.

(A1.1a) is equivalent to $g(u_{ik}, \varepsilon_i|x_{ik}, z_i, \xi_{ik}) = g(u_{ik}, \varepsilon_i|z_i, \xi_{ik})$. (In some applications it may be necessary to augment z_i to include known exchangeable functions of $A_{i1}, ..., A_{iK}$.). Note that the correlation of x_{ik} with u_{ik} comes from ξ_{ik} and possibly z_i. Since z_i, x_{ik} and A_{ik} are observed, one can consistently estimate $\Upsilon(A_{ik}, z_i)$ and ξ_{ik}, particularly if $\Upsilon(A_{ik}, z_i)$ has a finite number of parameters. Given these facts, we modify the approach to estimation underlying (2.7) by working with the $E_{x_{ik}}(y_{ik}|x_{ik}, z_i, \xi_{ik})$ rather than $E_{x_{ik}}(y_{ik}|x_{ik}, z_i)$.

Suppressing the i subscript and setting k equal to 1.

$$\text{(i) } E(y_1|x_1, z, \xi_1) = \int_{u_1}\int_{\varepsilon} m(x_1, \varepsilon, u_1) g(u_1, \varepsilon|x_1, z, \xi_1) d\varepsilon\, du_1$$

Since A_1 is independent of (ε, u_1) conditional on z, ξ_1,

$$\text{(ii) } \frac{\partial g(u_1, \varepsilon|x_1, z, \xi_1)}{\partial x_1} = \frac{\partial g(u_1, \varepsilon|x_1, z, \xi_1)}{\partial \Upsilon(A_1, z)} = 0$$

Using (i) and (ii) leads to

$$\text{(iii) } E_{x_1}(y_1|x_1, z, \xi_1) = \int_{u_1}\int_{\varepsilon} m_{x_1}(x_1, \varepsilon, u_1) g(u_1, \varepsilon|x_1, z, \xi_1) d\varepsilon\, du_1$$

$E_{x_1}(y_1|x_1, z)$ differs from $\beta(x_1)$ because the distribution of u_1 and ε is conditioned on z and $\xi_1$as well as x_1. However, one may integrate out z and ξ_1 to obtain $\beta(x_1)$ from $E_{x_1}(y_1|x_1, z, \xi_1)$. To see how to do this, note first that

$$\text{(iv) } g(u_1, \varepsilon|x_1) = \int_{z}\int_{\xi_1} g(u_1, \varepsilon|x_1, z, \xi_1) h(z, \xi_1|x_1) d\xi_1 dz$$

where $h(z, \xi_1|x_1)$ is the conditional density of (z, ξ_1) 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 ξ_1. This yields (v)

$$\text{(v) } \int_{z}\int_{\xi_1} E_{x_1}(y_1|x_1, z, \xi_1) h(z, \xi_1|x_1) d\xi dz$$
$$= \int_{z}\int_{\xi_1}\int_{u_1}\int_{\varepsilon} m_{x1}(x_1, \varepsilon, u_1) g(u_1, \varepsilon|x_1, z, \xi_1) h(z, \xi_1|x_1)\, d\varepsilon\, du_1\, d\xi dz.$$

[9] In general, instrumental variables methods are inconsistent when regressor is correlated with a random coefficient. An exception is Heckman and Vytlacil (1998) who provide a consistent IV estimator for a case in which the random coefficient is unobserved but depends on an exogenous observable and an instrument is also available for the endogenous regressor.

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. Basically, the difference between ξ_1 and z is that ξ_1 must be estimated in a first step, as is a number of estimation procedures in the literature in which a residual is introduced as a control variable in a second step.[10] Because of non-separability between the errors and x, one must use (v) to "undo" the effects of conditioning on $z, \xi_1$when estimating the response of y to x.

The key assumption is (A1.1a). Altonji and Ichimura (1997) take a similar approach to treatment of endogenous explanatory variables in the context of nonseparable linear dependent variables models. The condition that A_{ik} is independent of ε_i and u_{ik}, conditional on ξ_{ik} and z_i, is much stronger than the usual conditions for IV estimators in a linear setting. However, it should be kept in mind that IV estimators of partial derivatives are almost always inconsistent in models such as (1.4), where slope coefficients are random and correlated with the endogenous variable, and there are no IV counterparts for binary response models. Consequently, it is not surprising that our approach requires strong conditions. It is important to point out that our approach covers the cross section case in which there is no k subscript or u_{ik} component, z_i variables may or may not be involved, and an instrumental variable A_i that affects the mean of x_i and is independent of ε_i on conditional ξ_i and z_i is available.[11]

5 Estimating the effect of x_{i1} on y_{i1} from the joint distribution of y_i and x_i.

Estimator 2 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 $m(x_{ik}, \varepsilon_i, u_{ik})$ and $g(u_{ik}, \varepsilon_i|x_i)$ from the distribution of y_{ik} conditional on x_i and the distribution of y_{ik} conditional on x_{ik}. Consequently, various functions of $m(x_{ik}, \varepsilon_i, u_{ik})$ and $g(u_{ik}, \varepsilon_i|x_i)$, including average derivatives such as $\beta(x_{i1})$, are identified. Our proof of identification is a constructive proof and, hence, it provides a way of estimating $m(x_{ik}, \varepsilon_i, u_{ik})$ and $g(\varepsilon_i, u_{ik}|x_i)$ from a nonparametric estimator for the joint distribution of y_i and x_i.

To simplify the notation we will 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}. We will also consider only groups of size 2. The model underlying the second approach to estimation is described by the following assumptions:

Assumption A5.1: There exists a real valued, not necessarily known function $\Upsilon(\varepsilon, u)$ such that $y = m(x, e)$, where $e = \Upsilon(\varepsilon, u)$

[10] See, for example, Smith and Blundell (1986) and Rivers and Vuong (1988) in the context the tobit and probit models.

[11] Since the first drafts of this paper were circulated, Blundell and Powell (1999) have proposed a related approach to estimation with endogenous regressors in the context of the binary choice problem with a linear index that we present in the introduction. That is, they consider the model

$y_i = 1\{x_i\Gamma + \varepsilon_i > 0\}$

where we drop the subscript k and the error component u_{ik} because there is only one observation per group and abstract from the possible presence of z_i. They assume that x_i and ε_i are correlated but that an instrumental variable A_i is available, and they make an assumption that is equivalent to (A1.1a).

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

Assumption A5.2: $\forall w, w' \quad q(e|w, w') = q(e|w', w)$.

Assumption A5.3: $\forall x \quad m(x, \cdot)$ is strictly increasing in e.

Assumption 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 $\Upsilon(\varepsilon, u)$ of ε 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. Suppose, for example that (i) $e = \Upsilon(\varepsilon, u) = \varepsilon + u$, (ii) the conditional distribution of ε is exchangeable in x_1 and x_2, and (iii) u is distributed independently of x and ε, then, Assumption A5.2 is satisfied.[12]

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 $g(e_{ik}|x_{ik})$.[13] As we have noted in the introduction, strict monotonicity of m in e rules out qualitative choice models. 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:

Assumption $A5.5: \quad m(0, e) = e$.

Assumption A.5 specifies the values of the function at one value of the vector x. The particular value of x at which this is specified is irrelevant for the identification result. The zero vector was chosen here only for simplicity. When the function m is differentiable, it is easy to prove that this assumptions is innocuous. To see this, note that the assumption of strict monotonicity

[12] To see this, let $g(\varepsilon|w, w')$ and $s(u)$ denote, respectively, the conditional pdf of ε and the pdf of u, and note that

$$(5.2) \quad \forall w, w' \quad q(e|w, w') = \int s(e - \varepsilon|w, w')\, g(\varepsilon|w, w')\, d\varepsilon$$
$$= \int s(e - \varepsilon)\, g(\varepsilon|w, w')\, d\varepsilon$$
$$= \int s(e - \varepsilon|w', w)\, g(\varepsilon|w', w)\, d\varepsilon$$
$$= q(e|w', w).$$

[13] 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.

of m in e implies that, given any function $m(\cdot,\cdot)$, one can define a new function $m'(\cdot,\cdot)$ by $m'(x,\tilde{e}) = m(x, m^{-1}(0,\tilde{e}))$, 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 $\tilde{e}$, $m'(0,\tilde{e}) = \tilde{e}$. Moreover, since for all x and all $\tilde{e}$ and e such that $m'(x,\tilde{e}) = m(x,e)$ it is the case that $\frac{\partial}{\partial x}m'(x,\tilde{e}) = \frac{\partial}{\partial x}m(x,e)$, it follows from Brown and Matzkin (1996) that m' and m are observationally equivalent.

Theorem 2 *:If Assumptions A5.1-A5.5 are satisfied, then $m(x,e)$ is identified on the support of (x,e) and $F_{\varepsilon|x}(x)$ is identified on R, from the distribution of $y|x_1,x_2$.*

Proof. : Assumption A5.2 implies that

$$\forall w, w' \quad \Pr(e \leq \eta | w', w) = \Pr(e \leq \eta | w, w').$$

Hence, it follows by Assumption A5.3 that $\forall w', w$, $\forall e$, and $\forall \eta$,

$$\Pr(m(w',e) \leq m(w',\eta)|w',w) = \Pr(m(w,e) \leq m(w,\eta)|w,w'), \text{ or}$$

$$(5.3)\ \Pr(y \leq m(w',\eta)|w',w) = \Pr(y \leq m(w,\eta)|w,w').$$

In other words,

$$F_{y|w',w}(m(w',\eta)) = F_{y|w,w'}(m(w,\eta))$$

where $F_{y|w',w}(\cdot)$ is the CDF of y conditional on $x_1 = w'$ and $x_2 = w$. In particular,

$$(5.4)\ F_{y|x,0}(m(x,e)) = F_{y|0,x}(m(0,e)) = F_{y|0,x}(e)),$$

where the last equality follows from Assumption A5.5. By Assumption A5.4,

$$(5.5)\ \ m(x,e)) = F^{-1}_{y|x,0}(F_{y|0,x}(e)).$$

Hence, the function m is identified. Next, since $F_{e|x}(e) = F_{y|x}(m(x,e))$,

$$F_{e|x}(e) = F_{y|x}(F^{-1}_{y|x,0}(F_{y|0,x}(e)),$$

which shows that $F_{e|x}(e)$ 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 $m(x,e)$ 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 $F_{y|x,0}(m(x,e)) = F_{y|0,x}(m(0,e))$, 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 two CDFs are equal.

With knowledge of $m(x,e)$ and $F_{e|x}(e)$, it is of course possible to obtain a number of functions summarizing the effect of a change in x on the distribution of y holding the distribution of e constant. Suppose that $F_{e|x}$ has a density $q(e|x)$. Then, $\beta(x_1) \equiv \beta(x) = \int \partial m(x,e)/\partial x \; q(e|x)$ de. To obtain an expression for $\beta(x_1)$ in terms of the pdf's and the derivatives of the pdf's of the observable variables, we note that by the definition of $m(x,e)$,

$$(5.6)\quad \frac{\int^{m(x,e)} f(s,x,0)\, ds}{f(x,0)} = \frac{\int^{e} f(s,0,x)\, ds}{f(0,x)}.$$

Differentiating (5.6) with respect to x, we get

$$\frac{\int^{m(x,e)} \frac{\partial f(s,x,0)}{\partial x}}{f(x,0)} + \frac{f(m(x,e),x,0)}{f(x,0)}\frac{\partial m(x,e)}{\partial x} - \frac{\int^{m(x,e)} f(s,x,0)\, ds}{f(x,0)^2}\frac{\partial f(x,0)}{\partial x}$$
$$= \frac{\int^{e} \frac{\partial f(s,0,x)}{\partial x} ds}{f(0,x)} - \frac{\int^{e} f(s,0,x)\, ds}{f(0,x)^2}\frac{\partial f(0,x)}{\partial x}.$$

Hence,

$$\frac{\partial m(x,e)}{\partial x} = \frac{f(x,0)\int^{e} \frac{\partial f(s,0,x)}{\partial x} ds}{f(m(x,e),x,0)\; f(0,x)} - \frac{f(x,0)\int^{e} f(s,0,x)\, ds}{f(m(x,e),x,0)\; f(0,x)^2}\frac{\partial f(0,x)}{\partial x}$$
$$+ \frac{\int^{m(x,e)} f(s,x,0)\, ds}{f(m(x,e),x,0)\; f(x,0)}\frac{\partial f(x,0)}{\partial x} - \frac{\int^{m(x,e)} \frac{\partial f(s,x,0)}{\partial x}}{f(m(x,e),x,0)}.$$
$$= \frac{F_{y|0,x}(e)}{f_{y|x,0}(m(x,e))}\left[1 - \frac{\frac{\partial f(0,x)}{\partial x}}{f(0,x)}\right] + \frac{F_{y|0,x}(e)}{f_{y|x,0}(m(x,e))}\frac{\frac{\partial f(x,0)}{\partial x}}{f(x,0)} - \frac{\int^{m(x,e)} \frac{\partial f(s,x,0)}{\partial x} ds}{f(m(x,e),x,0)}$$

Differentiating (5.6) again, this time with respect to e, and solving for $\partial m(x,e)/\partial e$, we get

$$\frac{\partial m(x,e)}{\partial e} = \frac{f(x,0)\; f(e,0,x)}{f(m(x,e),x,0)\; f(0,x)} = \frac{f_{y|0,x}(e)}{f_{y|x,0}(m(x,e))}.$$

Hence, since

$$q(e|x) = \frac{\partial F_{e|x}(e)}{\partial e} = \frac{\partial F_{y|x}(m(x,e))}{\partial x} = f_{y|x}(m(x,e))\frac{\partial m(x,e)}{\partial e},$$

it follows that

$$(5.7)\ \beta(x) = \int [\frac{\int^{e} \frac{\partial \widehat{f}(s,0,x)}{\partial x} ds}{\widehat{f}_{y|x,0}(m(x,e))^2\; \widehat{f}(0,x)} - \frac{\widehat{F}_{y|0,x}(e)}{\widehat{f}_{y|x,0}(m(x,e))^2\; \widehat{f}(0,x)}\frac{\partial \widehat{f}(0,x)}{\partial x}$$
$$+ \frac{\widehat{F}_{y|x,0}(\widehat{m}(x,e))}{\widehat{f}_{y|x,0}(m(x,e))^2\; \widehat{f}(x,0)}\frac{\partial \widehat{f}(x,0)}{\partial x} - \frac{\int^{\widehat{m}(x,e)} \frac{\partial \widehat{f}(s,x,0)}{\partial x}}{\widehat{f}_{y|x,0}(m(x,e))^2\; \widehat{f}(x,0)}]\widehat{f}_{y|x}(\widehat{m}(x,e))\; \widehat{f}_{y|0,x}(e)\; de,$$

where

$$\widehat{m}(x,e) = \widehat{F}^{-1}_{y|x,0}(\widehat{F}_{y|0,x}(e)).$$

5.1 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 estimators for the unconditional and conditional pdf's and cdf's are:

$$\hat{f}_N(t_y, t_1, t_2) = \frac{1}{N\sigma_N^d} \sum_{i=1}^N K(\tfrac{u_y - t_y}{\sigma}, \tfrac{u_x - t_1}{\sigma}, \tfrac{u_z - t_2}{\sigma}) \quad \text{for all } (t_y, t_1, t_2) \in R^d,$$

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

$$\hat{F}_{y|x_1,x_2}(y|x_1,x_2) = \frac{\int_{-\infty}^{y} \hat{f}_N(t_y, x_1, x_2)\, dt_y}{\int_{-\infty}^{\infty} \hat{f}_N(t_y, x_1, x_2)\, dt_y}, \text{ and}$$

$$\hat{f}_N(y|x_1,x_2) = \frac{\hat{f}_N(y, x_1, x_2)}{\int_{-\infty}^{\infty} \hat{f}_N(t_y, x_1, x_2)\, dt_y}.$$

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

$$\hat{m}(x,e) = \hat{F}^{-1}_{y|x,0}(\widehat{F}_{y|0,x}(e)).$$

Since we do not restrict $\hat{F}_{y|x,0}$ to be strictly increasing, $\hat{F}^{-1}_{y|x,0}(\widehat{F}_{y|0,x}(e))$ need not be a singleton. If this is the case, we let $\hat{m}(x,e)$ denote any particular element of $\hat{F}^{-1}_{y|x,0}(\widehat{F}_{y|0,x}(e))$. The estimator for $F_{e|x}(e)$ is defined by

$$\widehat{F}_{e|x}(e) = \widehat{F}_{y|x}(\hat{m}(x,e)) = \widehat{F}_{y|x}(\hat{F}^{-1}_{y|x,0}(\widehat{F}_{y|0,x}(e))).$$

To establish the asymptotic properties of these estimators, we make the following assumptions:

ASSUMPTION 1': The sequence $\{y_i, x_{1i}, x_{i2}\}$ is i.i.d.

ASSUMPTION 2': $f(\cdot,\cdot,\cdot)$ has compact support $\Theta \subset R^{d'}$ and is continuously differentiable up to the order s' for some even s'.

ASSUMPTION 3': The kernel function $K(\cdot,\cdot,\cdot)$ is Lipschitz, vanishes outside a compact set, integrates to 1, and is of order s'.

ASSUMPTION 4': As $N \to \infty$, $\ln(N)/N\sigma_N^{d'} \to 0$ and $\sigma_N^{s'}\sqrt{N\sigma_N^{2K}} \to 0$.

ASSUMPTION 5': $0 < f(0,x),\ f(x,0), f(x) < \infty$, and there exist $\delta, \xi > 0$ such that for all $s \in \{s \in R|\ |s - m(x,e)| < \xi\},\ f(s,x,0) \geq \delta$.

Let

$$L = \frac{1}{f(0,x)} F_{y|0,x}(e)(1 - F_{y|0,x}(e)) + \frac{1}{f(x,0)} F_{y|x,0}(m(x,e))(1 - F_{y|x,0}(m(x,e))).$$

Theorem 3 *: Suppose that Assumptions A5.1-A5.5 and 1'-5' are satisfied. Then,*

(i) $\widehat{m}(x,e)$ *converges in probability to* $m(x,e)$,

(ii) $\widehat{F}_{e|x}(e)$ *converges in probability to* $F_{e|x}(e)$,

(iii) $\sqrt{N}\sigma_N^K\left(\hat{m}(x,e) - m(x,e)\right) \to N\left(0, V_m\right)$ *in distribution, where*

$$V_m = \left\{\int \left[\int K(s,z_1,z_2)ds\right]^2 dz_1 dz_2\right\} \left[\frac{1}{f_{y|x,0}(m(x,e))}\right]^2 L \text{ and}$$

(iv) $\sqrt{N}\,\sigma_N^K\left(\widehat{F}_{e|x}(e) - F_{e|x}(e)\right) \to N(0, V_F)$ *in distribution, where*

$$V_F = \left\{\int \left[\int K(s,z_1,z_2)ds\right]^2 dz_1 dz_2\right\} \left[\frac{f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))}\right]^2 L.$$

PROOF: See Appendix A.

After replacing Assumptions 2'-4' with:

ASSUMPTION 2": $f(y,x_1,x_2)$ has compact support $\Theta \subset R^{d'}$ and is continuously differentiable up to the order $1 + s"$ for some even $s" > 0$;

ASSUMPTION 3": The kernel function $K(\cdot,\cdot,\cdot)$ is continuously differentiable, vanishes outside a compact set, integrates to 1, and is of order $s"$;

ASSUMPTION 4": As $N \to \infty$, $\ln(N)/N\sigma_N^{d'+2} \to 0$ and $\sigma^{s"}\sqrt{N\sigma_N^{2K+2}} \to 0$,

we can obtain the asymptotic properties of $\widetilde{\beta}(x)$. Let

$$L_1 = \int \left\{\int \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))^2}\right] \left(1[s \leq e] - F_{y|0,x}(e)\right) de\right\}^2 f_{y|0,x}(s)ds,$$

$$L_2 = \int \left\{\int \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))^2}\right] \left(1[s \leq m(x,e)] - F_{y|0,x}(m(x,e))\right) de\right\}^2 f_{y|x,0}(s)ds,$$

and

$$\overline{\partial K} = \int\int \left(\int \frac{\partial K(s,x,z)}{\partial x} ds\right)\left(\int \frac{\partial K(s,x,z)}{\partial x} ds\right)' dx\, dz.$$

Theorem 4 *: Suppose that Assumptions A5.1-A5.5 and 1',2",3",4", and 5' are satisfied. Then,*

(i) $\widetilde{\beta}(x)$ *converges in probability to* $\beta(x)$*, and*

(ii) $\sqrt{N}\sigma_N^{K+1}\left(\widetilde{\beta}(x)-\beta(x)\right) \rightarrow N\left(0,V_\beta\right)$ *in distribution, where*

$V_\beta = \overline{\partial K}\ \left[L_1 + L_2\right]$

PROOF: See Appendix A.

6 Monte Carlo Evidence

We have performed a small scale Monte Carlo analysis of Estimator 1. We begin with experiments in which y is continuous. The cases are nested in the following model:

$$\begin{aligned} & \text{Model 1} \\ y_{ik} &= m(x_{ik},\varepsilon_i,\eta_i,u_{ik}) = b_0 + b_1 x_{ik} + \gamma x_{ik}\eta_i + \theta_\eta \eta_i + \theta_\varepsilon \varepsilon_i + u_{ik}\ ; \\ k &= 1,2\,;\ i = 1,2,...1500 \\ x_{ik} &= x_i + \tilde{x}_{ik}\ \ ;\ \varepsilon_i = \theta_{\varepsilon x} x_i + \tilde{\varepsilon}_i\ ;\ \eta_i = \theta_{\eta x} x_i + \tilde{\eta}_{ik} \\ x_i &\sim N(0,1);\ \tilde{x}_{ik} \sim N(0,1);\ \tilde{\varepsilon}_i \sim N(0,1);\ \tilde{\eta}_{ik} \sim N(0,1); u_{ik} \sim N(0,1)\ . \end{aligned}$$

The random variables x_i, $\tilde{x}_{ik}$, $\tilde{\varepsilon}_i$, and u_{ik} 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 Estimator 1 we defined z_i as $(x_{i1}+x_{i2})/2$. Because of the linear relationship among the stochastic components and the fact that $\tilde{\varepsilon}_{ik}$, x_i, and $\tilde{x}_{ik}$ are all normally distributed, $g(\eta_i,\varepsilon_i,u_{ik}|x_{i1},x_{i2}) = g(\eta_i,\varepsilon_i,u_{ik}|(x_{i1}+x_{i2})/2)$, so z_i may be restricted to the mean of x_{ik} over 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 to begin with the case in which we have the right conditioning variables, although this case is likely to be the most favorable one for the estimator. We discuss a non-normal case below. For

Model 1, $\beta(x) = b_1 + \gamma E(\eta_i|x_{ik}) = b_1 + \gamma\theta_{\eta x}\theta_{x_i x_{ik}} x_{ik}$ where $\theta_{x_i x_{ik}}$ is the population coefficient of the regression of x_i on x_{ik}. We set $Var(x_i) = 1$ and $Var(\widetilde{x}_{ik}) = 1$, so $\theta_{x_i x_{ik}} = .5$.

We implement Estimator 1 using 2 different approaches. The approach labeled Polynomial/Kernel approximates $E(y_{ik}|x_{ik}, z_i)$ as a fourth order polynomial in x_{ik} and z_i with interactions up to the second order. We use OLS to estimate the coefficients of the polynomial. We use kernel regression to estimate $h(z_i|x_{ik})$. The second, Kernel/Kernel, uses kernel regression to estimate both $E(y_{ik}|x_{ik}, z_i)$ and $h(z_i|x_{ik})$.[14] We also report of $\beta(x)$ based on a OLS regression of y_{ik} on x_{ik} and a constant, both with and without group specific intercepts or "fixed effects". We set K to 2 and n to 1,500 and report results based on 750 replications. The numbers in parentheses are the standard deviations and the mean squared error of the estimators.

We report 3 cases of Model 1. In all cases we set b_0 to 0, b_1 to 2, and evaluate the estimates at x_{ik} = -2, -1, 0, 1, and 2. In Case 1 γ is 0, θ_ε is 0, and $\theta_{\eta x}$ is 0. The fact that $\gamma = 0$ implies that $\beta(x) = b_1 = 2$. In Case 1 both OLS and OLS/Fixed Effects are unbiased.

The results are in Table 1. The mean of the Kernel/Kernel estimator varies between 1.83 and 1.87 across the various values of x, showing some downward bias. This is probably due to the fact that we employed the simple to use normal density for the kernel functions instead of high-order, bias-reducing kernels. The Polynomial/Kernel estimator ranges from 2.01 to 1.989 and thus is very close to the true parameter value of 2.0 for all values of x. Not surprisingly, it has a larger sampling error and mean squared error than OLS or the OLS fixed effects estimator.

Case 2 is the same as Case 1 except that we change γ from 0 to 1 and $\theta_{\eta x}$ from 0 to 1. At these parameter values, $\beta(x)$ is equal to $2 + .5x_{ik}$. It increases from 1 when x is -2 to 3 when x is 2. The means of both versions of Estimator 1 closely track the true value of $\beta(x)$. The OLS estimator has a mean of 2.0, and thus is positively positive biased when x is less than 0 and negatively biased when x is greater than 0. The mean of the OLS/fixed estimator is 2.5. Controlling for group specific intercepts does not control for the $x_{it}\eta_i$.The standard deviation of Estimator 1 is larger than the standard deviation of the OLS and OLS/fixed effects estimator, but the mean squared errors are much smaller at most values of x.

Case 3 is the same as Case 2 except that we set θ_ε to 1 rather than 0, with θ_{ex} set to 0. $\beta(x)$ ranges from 1 when x_{ik} is -2 to 3.0 when x is 2. Once again, the two versions of Estimator 1 track $\beta(x)$ closely. The Kernel/Kernel version suffers from a small downward bias and the Polynomial/Kernel version from a small upward bias. In contrast, OLS is badly biased, with a mean of 3.0. It overstates $\beta(x)$ at all values except $x = 3$. OLS-fixed effects also suffers from a substantial bias at all values of x except $x = 0$. The standard deviation of Estimator 1 is larger than the OLS and the OLS/Fixed Effects estimator, but at most values of x it dominates by a large margin in mean squared error.

6.0.1 A Binary Choice Model

We now turn to an experiment applying Estimator 1 to a binary choice model. The model is of the form

[14] In initial experiments we obtained similar results using local linear regression to estimate one or both of the functions. Details about window width selection are in the footnote to Table 1.

Model 2

$$\begin{aligned}
y_{ik} &= I(m(x_{ik},\eta_i,u_{ik})\geq 0) = I(b_0 + b_1 x_{ik} + \gamma x_{ik}\eta_i + \theta_\eta \eta_i + u_{ik} \geq 0) \ ; \\
k &= 1,2;\ i = 1,2,...n \\
x_{ik} &= x_i + \tilde{x}_{ik} \\
\eta_i &= \theta_{\eta x} x_i + \tilde{\eta} \\
x_i &\sim N(0,\sigma_x^2);\ \ \tilde{x}_{ik} \sim N(0,\sigma_{\tilde{x}}^2);\ \ \tilde{\eta}_i \sim N(0,\sigma_{\tilde{\eta}}^2); u_{ik} \sim N(0,\sigma_u^2)
\end{aligned}$$

For this design we also set z_i to $(x_{i1}+x_{i2})/2$ for reasons discussed above. The row labelled (Probit/Kernel) in Table 2 is based on the assumption 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 second order. We use kernel regression to estimate $h(z_i|x_{ik})$. In the row labelled Kernel/Kernel 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 a base case that we exclude from the table, $\gamma = 0$, $\theta_{\eta x} = 0$, and $b_1 = 0$. We also set $\sigma_{\tilde{x}}^2$ to 1.5. In this case the probit estimator is consistent and $\beta(x) = 0$ for all x. The mean of both versions of Estimator 1 and the probit are essentially 0.

In Case 1 in the Table 2, $b_1 = 2$, $\gamma = 0$, and $\theta_{\eta x}$ is 1. We also set $\theta_{\eta x} = 1$, $\sigma_x^2 = 1$, $\sigma_{\tilde{x}}^2 = 1$, and $\sigma_{\tilde{\eta}}^2 = 1$. This design is a probit model with a group specific error component (η) that is correlated with x_{ik} but no random slope. Not surprisingly, the probit estimator suffers from a substantial positive bias at each value x. In contrast, Estimator 1 does pretty well, particularly in the Kernel/Kernel case, and typically substantially dominates the probit estimator in mean squared error.

Case 2 is similar to Case 1 in that $b_1 = 2$, $\gamma = 0$ and x_{it} is correlated with η_ι, but involves different parameter values. The main difference is that we set $\theta_{\eta x}$ to -1.5. We also set $\sigma_x^2 = 1.5$, $\sigma_{\tilde{x}}^2 = 1.5$, $\sigma_u^2 = 1.5$,and $\sigma_{\tilde{\eta}}^2 = .5$. For this design $\beta(x) = .155$ at all values of x. The Probit/Kernel version of the estimator matches this value at all values of x. The mean of the Kernel/Kernel version ranges between .121 and .123 and so has a modest downward bias. This may be, again, due to the fact that a N(0,1) density was used as a kernel function, instead of a high-order kernel. In this case, $dE(y_{ik}|x_{ik})/dx_{ik} = 0$ because the conditional mean of the index determining y_{ik} is $b_1 x_{ik} + E(\eta|x_{ik})$ = $.75x_{it} - .5 \cdot .1.5x_{ik}$ =0. . This fact is reflected in the probit estimates, which have a mean of 0 at all values of x in the table.

We have experimented with a number of cases in which η_i is correlated with x_{ik} and enters as both a random intercept and a random slope, with $\gamma \neq 0$. The Kernel/Kernel estimator performed very well. In most cases the Probit/Kernel version also performs well, but we present an interesting exception in panel 3 of Table 2. In this case $b_1 = 2$, $\gamma = 1$, $\sigma_x^2 = 1.0$, $\sigma_{\tilde{x}}^2 = 1.0$, $\sigma_u^2 = 1.5$, $\sigma_{\tilde{\eta}}^2 = 1.5$ and $\theta_{\eta x}$ =1.5. The sign of $\beta(x)$ varies with x_{ik} for three reasons. The first is that the mean of the random slope obviously shifts with x_{ik}. The conditional mean of the slope of the index is $b_1 + \gamma E(\eta_i|x_{ik}) = 2 + 1 \cdot .75x_{ik}$. This term changes from positive to negative when $x_{ik} = -2.67$. The sign of $\beta(x)$ does not shift from negative to positive until $x_{ik} = -1.29$. The discrepancy reflects the nonlinearity in the relationship between y_{ik} and the index $b_0 + b_1 x_{ik} + \gamma x_{ik}\eta_i + \theta_\eta \eta_i + u_{ik}$, the fact that the distribution of the random intercept $\theta_\eta \eta_i$ declines with x_{ik}, and the fact that the value of $b_1 x_{ik} + \gamma x_{ik}\eta_i + \theta_\eta \eta_i$ influences the likelihood that a given change in $b_1 x_{ik} + \gamma x_{ik}\eta_i$ will lead the index $b_0 + b_1 x_{ik} + \gamma x_{ik}\eta_\iota + \theta_\eta \eta_i + u_{ik}$ to exceed 0. $\beta(x)$ switches from negative to positive as x increases. $\beta(x)$ is -.1 when x is -2, .107 when x is -1, .393 when x is 0, .080 when $x = 1$, and 0.016 when $x = 2$.

The probit estimator is seriously biased at some values of x, particularly $x = -2$ and $x = 1$. The Probit/Kernel version of Estimator 1 is also biased, particularly when $x = -2$. Adding a 4th order term in x had little effect on these estimates. In contrast, the Kernel/Kernel version of Estimator 1 tracks $\beta(x)$ quite closely.

Overall, the binary choice result are very encouraging. They suggest that the regression based estimator does provide a way to estimate qualitative choice models with random errors that are correlated with and perhaps interact with the explanatory variables in the model.

In addition to the results reported, we have conducted a series of experiments in which x_{ik} is a Chi-square$_3$ random variable normalized to have mean 0 and variance 1. We continue to use $z_i = (x_{i1} + x_{i2})/2$ as the only z variable even though is not the optimal choice in this case. We used the Kernel/Kernel version of the estimator. For the continuous dependent variable design in Table 1 and the binary choice design in Table 2, we found that Estimator 1 tracks $\beta(x)$ reasonably closely. The results for 3 binary choice cases are presented in Table 3.

The Monte Carlo evidence on estimator 1 is obviously limited by the relatively small set of designs and parameter values we have examined. We have not experimented extensively with choice of window width or with more sophisticated kernels in implementing the nonparametric versions of the estimator.. The cases we report illustrate the fact $\beta(x)$ can vary widely over the range of x_{it} and in some circumstances (eg., Table 2, case 3) a low order polynomial may not adequately capture $E(y_{ik}|x_{ik}, z_i)$ leading to bias in Estimator 1. However, we find the results to date to be very encouraging.

7 Concluding Remarks

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 errors and endogenous explanatory variables. One important class of such models are qualitative choice models with group error components that are correlated with the regressors. Estimator 1 covers this case. A version of Estimator 1 also covers a more general case in which a regressor is correlated with the idiosyncratic error components and an instrument is available. Another set of examples consists of random coefficients models in which a group specific random coefficient is correlated with the regressors. The applied econometrician does not have good options in the literature to estimate such models, except in special cases. Thus, Estimators 1 and 2 may prove attractive in a wide range of situations.

Rather than repeat the intuitive discussion of our estimators that is contained in the introduction, we close with a research agenda. Much of our analysis follows from an examination of the implications of the assumption that the distribution of the error components u_{ik} and ε_i conditional on the vector $x_i$of explanatory variables for the members of group i is exchangeable in the elements of x_i. Further research is needed regarding the power of the exchangeability assumption for identification, particularly in the context of Estimator 1. Estimator 1 provides a way to estimate the partial effect of x_{ik} on the expectation of y provided that one has a variable z such the conditional distribution of the error term does not depend on x_{ik} once one conditions on z_i. It is applicable in both cross section applications and in panel data applications, but the problem in applied situations is coming up with a z variable. Exchangeability in the context of panel data provides a place to look. However, we point out in section 2 that although exchangeability restricts the relationship between the distribution of the error components and x_i to exchangeable (symmetric) functions of the elements of x_i, it is not sufficient to identify $E(y_{ik}|x_{ik}, z_i)$ nonparametrically. The conditions for identification involve trade-offs among (a) the size of the panel K, (b) the number L of elements in z_i, and (c) parametric or nonparametric restrictions on $E(y_{ik}|x_{ik}, z_i)$.We conjecture that in actual panel data applications when exchangeability holds, conditioning on 1 or 2 z_i functions capturing the location of x_i (such as the average of the elements of x_i) and the dispersion of the elements of x_i (such as the variance), will be sufficient to eliminate most of the relationship between the error terms and x_{ik}. But further theoretical research and monte carlo studies of the issue in the context of real world data is needed.

Second, additional monte carlo analysis of the estimators is needed. Third, it worth exploring the possibility that Estimator 1 can be adapted to nonseparable dynamic panel data models when the data are stationary and u_{ik} is i.i.d.. The basic idea is that in some situations the distribution of $(\varepsilon_i|y_{i1,...,}y_{ik-1})$ may be exchangeable in the lags of y. Complications will arise when x variables also enter the model even if the x variables are strictly exogenous, because the x variables will influence relationship between the lags of y and ε_i.

As with any new econometric procedure, the greatest needs are for practical experience with the estimators in applications and the development of reliable algorithms and easy to use software to implement them.

8 Appendix A

We present in this Appendix the proofs of Theorems 1, 3, and 4, which present the asymptotic properties of our estimators. All our estimators are functionals of kernel estimators for joint distribution functions. We develop their asymptotic properties by first linearizing the functionals, and then using properties of kernel estimators. (See Newey (1994) and Ait-Sahalia (1994) for such Delta methods techniques.) Appendix B presents known results about kernel estimators which are used in the proofs of the Theorems. (See Lemma 5.3 in Newey (1994) for the first parts of [I] and [II] in the statement of the lemmas.) We include them for the convenience of the reader.

PROOF OF THEOREM 1: To prove the theorem, we will use a Delta method, such as the ones developed in Newey (1994) and Ait-Sahalia (1994). Let $F(y,x,z)$ denote the distribution function (cdf) of the vector of observable variables (y,x,z), $f(y,x,z)$ denote its probability density function (pdf), $f(x,z)$ and $f(x)$ denote, respectively, the marginal pdf's of (x,z) and x, and $f(z|x)$ denote the conditional pdf of z given x. For any function $G : R^{1+K_1+K_2} \to R$, define $g(y,x,z) = \partial^{1+K_1+K_2} G(y,x,z)/\partial y \partial x \partial z$, $g(x,z) = \int g(y,x,z)\, dy$, $g(x) = \int g(y,x,z)\, dy\, dz$, $g(z|x) = g(x,z) \,/\, g(x)$, and $g_x(y,x,z) = \partial g(y,x,z)/\partial x$ when these functions exist. Let $\underline{\mathrm{C}}$ denote a compact set in $R^{1+K_1+K_2}$ that strictly includes Θ. Let D denote the set of all functions $G : R^{1+K_1+K_2} \to R$ such that $g(y,x,z)$ and $\partial g(y,x,z)/\partial x$ exist and vanish outside $\underline{\mathrm{C}}$. Let D denote the set of all functions g_x that are derivatives with respect to x of some g which corresponds to a function G in D. Note that there is a 1-1 relationship between functions in D and functions in D. Hence we can define a functional on D or on D without altering its definition. Define then the functional $\Phi(\cdot)$ by:

$$\Phi(g_x) = \Phi(G) = \int \tfrac{\partial}{\partial x} \int y\, g(y|x,z)\; g(z|x)\, dz.$$

Then,

$$\beta(x) = \Phi(F) \quad \text{and} \quad \hat{\beta}(x) = \Phi(\widehat{F}) = \Phi(\widehat{f_x})$$

where $\widehat{F}$ is the kernel estimator for the cdf F and where, $\widehat{f_x}$, the kernel estimator for the derivative of $f(y,x,z)$ with respect to x, is obtained by differentiating $\widehat{F}$ in the same manner that f_x is obtained from F.

Note that

$$\begin{aligned}\Phi(F) &= \int \tfrac{\partial}{\partial x} E\left(y|x,z\right)\; f(z|x)\, dz \\ &= \int \tfrac{\partial}{\partial x} \left[\tfrac{\int y\, f(y,x,z)\, dy}{f(x,z)}\right] \left[\tfrac{f(x,z)\, dy}{f(x)}\right]\, dz \\ &= \int \tfrac{\int y\, f_x(y,x,z)\, dy}{f(x)}\, dz \;-\; \int \tfrac{f_x(x,z)\;\; \int y\, f(y,x,z)\, dy}{f(x,z)\;\; f(x)}\, dz\end{aligned}$$

where the integrals with respect to z are over the values of z for which $f(x,z) > 0,$ and where $f_x(y,x,z)$ and $f_x(x,z)$ denote, respectively, the gradient with respect to x of the functions $f(y,x,z)$ and $f(x,z)$.

For any $G \in D$ let $\|G\|$ denote the sup norm of $\partial g(y,x,z)/\partial x.$ Then, there exists $\rho > 0$ such that if $\|G\| \le \rho$, then

(1) $f(x) + g(x) > f(x)/2$ and

$\forall(x,z)$ such that $f(x,z) > 0, \quad f(x,z) + g(x,z) > f(x,z)/2.$

For all such G, we have that

$$\Phi(F+G) - \Phi(F) = \Phi(f_x + g_x) - \Phi(f_x)$$

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

$$= \int \frac{\left(\int y\ g_x(y,x,z)\ dy\right) f(x) - \left(\int y\ f_x(y,x,z)\ dy\right) g(x)}{f(x)^2} dz$$
$$- \int \frac{f_x(x,z)\left(\int y\ g(y,x,z) dy\right) f(x,z)\ f(x) + g_x(x,z)\left(\int y f(y,x,z)\ dy\right) f(x,z) f(x)}{f(x,z)^2\ f(x)^2} dz$$
$$+ \int \frac{f_x(x,z)\left(\int y f(y,x,z) dy\right) f(x,z)\ g(x) + f_x(x,z)\left(\int y f(y,x,z) dy\right) g(x,z)\ f(x)}{f(x,z)^2\ f(x)^2} dz$$
$$+ \int \left[\left(\int y\ g_x(y,x,z)\ dy\right) f(x) - \left(\int y\ f_x(y,x,z)\ dy\right) g(x)\right] \left[\frac{1}{f(x)^2 + f(x) g(x)} - \frac{1}{f(x)^2}\right] dz$$
$$- \int \left[f_x(x,z) \int y g(y,x,z) dy\ f(x,z)\ f(x)\right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x)) f(x,z) f(x)} - \frac{1}{f(x,z)^2\ f(x)^2}\right] dz$$
$$- \int \left[g_x(,x,z) \int y f(y,x,z) dy\ f(x,z)\ f(x)\right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x)) f(x,z) f(x)} - \frac{1}{f(x,z)^2\ f(x)^2}\right] dz$$
$$+ \int \left[f_x(x,z) \int y f(y,x,z) dy\ f(x,z) g(x)\right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x)) f(x,z) f(x)} - \frac{1}{f(x,z)^2\ f(x)^2}\right] dz$$
$$+ \int \left[f_x(x,z) \int y f(y,x,z) dy\ g(x,z) f(x)\right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x)) f(x,z) f(x)} - \frac{1}{f(x,z)^2\ f(x)^2}\right] dz$$
$$- \int \frac{g_x(x,z) \int y g(y,x,z) dy\ f(x,z) f(x) - f_x(x,z) \int y f(y,x,z) dy\ g(x,z) g(x)}{(f(x,z)+g(x,z))(f(x)+g(x)) f(x,z) f(x)} dz.$$

Let

$$(2)\quad D\Phi(F,G) = \int \frac{\left(\int y\ g_x(y,x,z)\ dy\right) f(x) - \left(\int y\ f_x(y,x,z)\ dy\right) g(x)}{f(x)^2} dz$$
$$- \int \frac{f_x(x,z) \int y g(y,x,z) dy\ f(x,z)\ f(x) + g_x(x,z) \int y f(y,x,z) dy\ f(x,z)\ f(x)}{f(x,z)^2\ f(x)^2} dz$$

$$+ \int \frac{f_x(x,z) \int y f(y,x,z)dy \ f(x,z) \ g(x) + f_x(x,z) \int y f(y,x,z)dy \ g(x,z) \ f(x)}{f(x,z)^2 \ f(x)^2} \ dz \ ,$$

and

$$R\Phi(F,G) = \int \left[\left(\int y \ g_x(y,x,z)dy \right) f(x) - \left(\int y \ f_x(y,x,z)dy \right) g(x) \right] \left[\frac{1}{f(x)^2 + f(x)g(x)} - \frac{1}{f(x)^2} \right] \ dz$$

$$- \int \left[f_x(x,z) \int y g(y,x,z)dy \ f(x,z) \ f(x) \right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x))f(x,z)f(x)} - \frac{1}{f(x,z)^2 \ f(x)^2} \right] \ dz$$

$$- \int \left[g_x(x,z) \int y f(y,x,z)dy \ f(x,z) \ f(x) \right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x))f(x,z)f(x)} - \frac{1}{f(x,z)^2 \ f(x)^2} \right] \ dz$$

$$+ \int \left[f_x(x,z) \int y f(y,x,z)dy \ f(x,z) \ g(x) \right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x))f(x,z)f(x)} - \frac{1}{f(x,z)^2 \ f(x)^2} \right] \ dz$$

$$+ \int \left[f_x(x,z) \int y f(y,x,z)dy \ g(x,z) \ f(x) \right] \left[\frac{1}{(f(x,z)+g(x,z))(f(x)+g(x))f(x,z)f(x)} - \frac{1}{f(x,z)^2 \ f(x)^2} \right] \ dz$$

$$- \int \left[\frac{g_x(x,z) \int y g(y,x,z)dy \ f(x,z)f(x) - f_x(x,z) \int y f(y,x,z)dy \ g(x,z)g(x)}{(f(x,z)+g(x,z))(f(x)+g(x))f(x,z)f(x)} \right] dz.$$

Then,

(3) $\Phi(F+G) - \Phi(F) = D\Phi(F,G) + R\Phi(F,G).$

Let $R_k(F,G)$ denote the $k-th$ coordinate of $R\Phi(F,G)$, for $k = 1, ..., K_1$, and let g_{x_k} and f_{x_k} denote the $k-th$ coordinate of g_x and f_x, respectively. It follows that

$$|R_k(F,G)| \le \left[\int \int |y| \left| g_{x_k}(y,x,z) \right| dydz \ f(x) + \left| \int \int y \ f_{x_k}(y,x,z)dydz \right| |g(x)| \right] \left[\frac{2|g(x)|}{f(x)^3} \right]$$

$$+ \int \left[4 \left| f_{x_k}(x,z) \right| \int |y| \left| g(y,x,z) \right| dy \ f(x,z) \ f(x) \right] \left[\frac{f(x,z)|g(x)| + |g(x,z)| f(x) + |g(x,z)||g(x)|}{f(x,z)^3 \ f(x)^3} \right] dz$$

$$+ \int \left[4 \left| g_{x_k}(x,z) \right| \left| \int y \ f(y,x,z)dy \right| \ f(x,z) \ f(x) \right] \left[\frac{f(x,z)|g(x)| + |g(x,z)| f(x) + |g(x,z)||g(x)|}{f(x,z)^3 \ f(x)^3} \right] dz$$

$$+ \int \left[4 \left| f_{x_k}(x,z) \right| \left| \int y \ f(y,x,z) \ dy \right| \ f(x,z) \left| g(x) \right| \right] \left[\frac{|f(x,z)g(x)| + |g(x,z)| f(x) + |g(x,z)||g(x)|}{f(x,z)^3 \ f(x)^3} \right] dz$$

$$+ \int \left[4 \left| f_{x_k}(x,z) \right| \left| \int y \ f(y,x,z)dy \right| \left| g(x,z) \right| f(x) \right] \left[\frac{f(x,z)|g(x)| + |g(x,z)| f(x) + |g(x,z)||g(x)|}{f(x,z)^3 f(x)^3} \right] dz$$

$$+ \int \frac{4 \left| g_{x_k}(x,z) \right| \int |y| |g(y,x,z)| dy \ f(x,z) \ f(x) + 4 \left| f_{x_k}(y,x,z) \right| \left| \int y \ f(y,x,z)dy \right| \ |g(x,z)||g(x)|}{f(x,z)^2 \ f(x)^2} \ dz,$$

where all the integrals are over compact sets. Since g vanishes outside $\underline{C}$, if follows by the definition of $\|\cdot\|$ that there exist a, b, c, and d such that $|g(x)| \le a \|G\|$, $|g(x,z)| \le b \|G\|$, $|g(y,x,z)| \le c \|G\|$, and $|g_{x_k}(y,x,z)| \le d \|G\|$ for all (y,z). Hence, the above expression is bounded by $A_k \|G\|$, where

$$A_k = \left[d \ f(x) \int \int |y| \ dydz + a \left| \int \int y \ f_{x_k}(y,x,z)dydz \right| \right] \left[\frac{2 \ a \ \|G\|^2}{f(x)^3} \right]$$

$$+\int\left[4c\left|f_{x_k}(x,z)\right|\int|y|\,dy\ f(x,z)\ f(x)\right]\left[\frac{a\ f(x,z)\|G\|^2+b\ f(x)\|G\|^2+a\ b\ \|G\|^3}{f(x,z)^3\ f(x)^3}\right]dz$$

$$+\int\left[4d\left|\int y\ f(y,x,z)dy\right|\ f(x,z)\ f(x)\right]\left[\frac{f(x,z)a\|G\|^2+b\|G\|^2f(x)+a\ b\ \|G\|^2}{f(x,z)^3\ f(x)^3}\right]dz$$

$$+\int\left[4a\left|f_{x_k}(x,z)\right|\left|\int y\ f(y,x,z)\ dy\right|\ f(x,z)\right]\left[\frac{f(x,z)a\|G\|^2+b\|G\|^2f(x)+a\,b\ \|G\|^3}{f(x,z)^3\ f(x)^3}\right]dz$$

$$+\int\left[4b\left|f_{x_k}(x,z)\right|\left|\int y\ f(y,x,z)dy\right|f(x)\right]\left[\frac{a\ f(x,z)\|G\|^2+b\|G\|^2f(x)+a\ b\ \|G\|^3}{f(x,z)^3f(x)^3}\right]dz$$

$$+\int\frac{4\ c\ d\|G\|^2\ f(x,z)\ f(x)\ \int|y|dy+\ 4\ a\ b\ \|G\|^2\left|f_{x_k}(y,x,z)\right|\ \left|\int y\ f(y,x,z)dy\right|}{f(x,z)^2\ f(x)^2}dz.$$

Since $\|G\|\leq\rho$, and by Assumptions 2 and 5, A_k is bounded, there exists A such that

(4) $\|R(F,G)\|_{R^{K_1}}\leq\ A\ \|G\|^2\,,$

where $\|\cdot\|_{R^{K_1}}$ denotes the Euclidean norm in R^{K_1}.

Let $D\Phi_k(F,G)$ denote the $k-th$ coordinate of $D\Phi(F,G)$. Then,

$$|D\Phi_k(F,G)|\leq\left|\int\frac{\int|y|\ \left|g_{x_k}(y,x,z)\right|\ dy}{f(x)}dz\right|+\left|\int\frac{|g(x)|\ \int\left|y\ f_{x_k}(y,x,z)\right|\ dy}{f(x)^2}dz\right|$$

$$+\int\frac{\left|f_{x_k}(x,z)\right|\ \int|y||g(y,x,z)|dy}{f(x,z)\ f(x)}dz+\int\frac{\left|g_{x_k}(x,z)\right|\left|\int yf(y,x,z)dy\right|}{f(x,z)\ f(x)}dz$$

$$+\int\frac{\left|f_{x_k}(x,z)\right|\left|\int yf(y,x,z)dy\right||g(x)|}{f(x,z)\ f(x)^2}dz+\int\frac{\left|f_{x_k}(x,z)\right|\left|\int yf(y,x,z)dy\right||g(x,z)|}{f(x,z)^2\ f(x)}dz$$

Hence, since by the definition of the metric $\|\cdot\|$, there exist e and f such that for all $k=1,..K_1$, $|g_{x_k}(y,x,z)|\leq e\,\|G\|$ and $|g_{x_k}(x,z)|\leq n\,\|G\|$, it follows that the above expression is bounded by $B_k\,\|G\|$, where

$$B_k=\int\frac{e\int|y|dy}{f(x)}dz+\int\frac{a\int\left|y\ f_{x_k}(y,x,z)\right|\ dy}{f(x)^2}dz$$

$$+\int\frac{c\left|f_{x_k}(x,z)\right|\ \int|y|dy}{f(x,z)\ f(x)}dz+\int\frac{n\left|\int yf(y,x,z)dy\right|}{f(x,z)\ f(x)}dz$$

$$+\int\frac{a\left|f_{x_k}(x,z)\right|\left|\int yf(y,x,z)dy\right|}{f(x,z)\ f(x)^2}dz+\int\frac{b\left|f_{x_k}(x,z)\right|\left|\int yf(y,x,z)dy\right|}{f(x,z)^2\ f(x)}dz$$

Since, by Assumptions 2 and 5, B_k is bounded, it follows that for some $B>0$,

(5) $|D\Phi_k(F,G)|\leq B\ \|G\|\,.$

Hence, the linear map $D\Phi(F,\cdot)$ is continuous.

Let $G=\widehat{F}-F$. By (3)-(5),

$$|\Phi(F+G)-\Phi(F)|\leq B\ \|G\|+A\ \|G\|^2$$

By Assumptions 1-4 and Lemma B.3 in Newey (1994), $\left\|\widehat{F}-F\right\| \to 0$ in probability. Hence,

$$\left|\hat{\beta}(x)-\beta(x)\right| = \left|\Phi(\hat{F})-\Phi(F)\right| \leq B\ \left\|\widehat{F}-F\right\| + A\ \left\|\widehat{F}-F\right\|^2$$

converges in probability to 0. Hence, $\hat{\beta}(x)$ is a consistent estimator for $\beta(x)$.

Also, from (3)-(5), Φ is Hadamard differentiable at F. Hence, by Theorem 3.9.4 in van der Vaart and Wellner (1996) and the Lemma in Appendix B, it follows that

$$\sqrt{N\sigma_N^{K_1+2}}(\Phi(\widehat{f}_x)-\Phi(f_x)) - D\Phi(F,\sqrt{N\sigma_N^{K_1+2}}(\widehat{f}_x-f_x))$$

converges in outer probability to 0. By (2),

$$D\Phi(F,\sqrt{N\sigma_N^{K_1+2}}(\widehat{f}_x-f_x)) = \sqrt{N\sigma_N^{K_1+2}}\frac{\int\int y\ (\widehat{f_x}(y,x,z)-f_x(y,x,z))\ dydz}{f(x)}$$
$$-\sqrt{N\sigma_N^{K_1+2}}(\widehat{f}(x)-f(x))\frac{\int\int y\ f_x(y,x,z)\ dydz}{f(x)^2}$$
$$-\sqrt{N\sigma_N^{K_1+2}}\int\ \frac{f_x(x,z)\int y(\widehat{f}(y,x,z)-f(y,x,z))\ dy}{f(x,z)\ f(x)}\ dz$$
$$-\ \int\ \frac{\sqrt{N\sigma_N^{K_1+2}}\ (\widehat{f}_x(x,z)-f_x(x,z))\ E(y|x,z)}{f(x)}\ dz$$
$$+\ \sqrt{N\sigma_N^{K_1+2}}(\widehat{f}(x)-f(x))\ \frac{\int f_x(x,z)\ E(y|x,z)\ dz}{f(x)^2}$$
$$+\int\ \frac{f_x(x,z)\ E(y|x,z)\ \sqrt{N\sigma_N^{K_1+2}}(\widehat{f}(x,z)-f(x,z))}{f(x,z)\ f(x)}\ dz\ ,$$

By the Lemma in Appendix B, all the terms in the above summation except the first and fourth converge in probability to 0. These two terms are

$$\frac{\sqrt{N\sigma_N^{K_1+2}}\int\int y\ (\widehat{f_x}(y,x,z)-f_x(y,x,z))\ dydz}{f(x)} - \frac{\sqrt{N\sigma_N^{K_1+2}}\int\ (\widehat{f}_x(x,z)-f_x(x,z))\ E(y|x,z)\ dz}{f(x)}$$
$$= \frac{\sqrt{N\sigma_N^{K_1+2}}\int\int(y-E(y|x,z))\ (\widehat{f_x}(y,x,z)-f_x(y,x,z))\ dydz}{f(x)}$$

By the Lemma in Appendix B, this last term converges in distribution to a random vector that possesses a distribution $N(0,V)$ where

$$V = \left\{\int\int\left[y-\frac{\left(\int y\ f(y,x,z)\ dy\right)}{f(x,z)}\right]^2\frac{f(y,x,z)}{f(x)^2}\ dydz\right\}\widetilde{K}$$
$$= \left\{\int Var\ (y|x,z)\ \frac{f(z|x)}{f(x)}\ dz\right\}\ \widetilde{K}$$

and $\widetilde{K} = \left\{\int\left(\int\int\frac{\partial K(y,x,z)}{\partial x}dydz\right)\left(\int\int\frac{\partial K(y,x,z)}{\partial x}dydz\right)'dx\right\}.$

Hence,

$$\sqrt{N}\,\sigma_N^{(K_1/2)+1}\left(\hat{\beta}(x)-\beta(x)\right)=\sqrt{N}\,\sigma_N^{(K_1/2)+1}\left(\Phi(\hat{F})-\Phi(F)\right)\to N(0,V)$$

in distribution.

■

PROOF OF THEOREM 3: As in the proof of Theorem 1, let $F(y,x,z)$ denote the distribution function (cdf) of the vector of observable variables (y,x,z), $f(y,x,z)$ denote its probability density function (pdf), $f(x,z)$ and $f(x)$ denote, respectively, the marginal pdf's of (x,z) and x, and $f(z|x)$ denote the conditional pdf of z given x. For any function $G:R^{d'}\to R$, define $g(y,x,z)=\partial^{d'}G(y,x,z)/\partial y\partial x\partial z$, $g(x,z)=\int g(y,x,z)\ dy$, $g(x)=\int g(y,x,z)\,dy\,dz$, $g(z|x)=g(x,z)\ /\ g(x)$, $G_{y|x',z'}(y')=\int_{-\infty}^{y'}g(y,x',z')ds/g(x',z')$, and $\widetilde{G_Y}(y,x,z)=\int^{y}g(s,x,z)ds$ $=\int 1[s\leq y]\ g(s,x,z)ds$ where $1[\cdot]=1$ if $[\cdot]$ is true, and it equals zero otherwise. Let $\underline{C}$ denote a compact set in $R^{d'}$ that strictly includes Θ. Let D denote the set of all functions $G:R^{d'}\to R$ such that $g(y,x,z)$ vanishes outside $\underline{C}$. Let D denote the set of all functions $\widetilde{G_Y}$ that are derived from some G in D. Since there is a 1-1 relationship between functions in D and functions in D, we can define a functional on D or on D without altering its definition. Define then the functional $\Phi(\cdot)$ by $\Phi(G)=G_{y|x,0}^{-1}\left(G_{y|0,x}(e)\right)$, where $G_{y|x,0}^{-1}$ denotes an arbitrary element of the set $G_{y|x,0}^{-1}$, if $G_{y|x,0}^{-1}$ is not a singleton. Then, $\Phi(F)=\Phi(\widetilde{F_Y})=m(x,e)$ and $\Phi(\widehat{F})=\Phi(\widehat{\widetilde{F_Y}})=\widehat{m}(x,e)$.

Let $\|G\|$ denote the sup norm of $g(y,x,z)$. Then, if $G\in$D, there exists $\rho_1>0$ such that if $\|G\|\leq\rho_1$ then, for some $0<a,b,c<\infty$, all y and all $s\in N(m(x,e),\xi)$,

(1) $|g(x)|\leq a\,\|G\|\,,\ |g(x,0)|\leq a\,\|G\|\,,\ |g(0,x)|\leq a\,\|G\|\,,$

$|g(m(x,e),x)|\leq a\,\|G\|\,,\ \left|\int_{-\infty}^{y}g(s,x)ds\right|\leq a\,\|G\|\,,$

$f(x)+g(x)\geq bf(x),\ f(x,0)+g(x,0)\geq bf(x,0),$

$f(0,x)+h(0,x)\geq bf(0,x)$, and

$f(s,x,0)+g(s,x,0)\geq cf(s,x,0).$

By substituting for x in the proof of Theorem 3 in Matzkin (1999) with $(x,0)$, everywhere except where x appears (in that proof) as an argument of the function m, and by substituting for $\widetilde{x}$ and $\widetilde{e}$, in that same proof, with , respectively, $(0,x)$ and e, it follows that for all H such that $\|H\|\leq\rho_1$,

$\Phi(F+H)-\Phi(F)=D\Phi(F,H)+R\Phi(F,H)$

where

(2) $D\Phi(F,H)=\frac{f(x,0)}{f(0,x)^2 f(\Phi(F),x,0)}A\widetilde{x}+\frac{f(x,0)}{f(x,0)^2 f(\Phi(F),x,0)}Ax,$

$A\widetilde{x}=f(0,x)\int^{e}h(s,0,x)ds-h(0,x)\int^{e}f(s,0,x)ds\,,$

$Ax=f(x,0)\int^{\Phi(F)}h(s,x,0)ds-h(x,0)\int^{\Phi(F)}f(s,x,0)ds,$

and for some A_1,A_2,

$|D\Phi(F,H)|\leq A_1\|H\|$ and $|R\Phi(F,H)|\leq A_2\|H\|^2$.

In particular, the linear map $D\Phi(F,\cdot)$ is continuous.

Let $H=\widehat{F}-F$. Then,

$$\widehat{m}(x,e)-m(x,e)=\Phi(\widehat{F})-\Phi(F)$$
$$=D\Phi(F,\widehat{F}-F)+R\Phi(F,\widehat{F}-F),$$

$\left|D\Phi(F,\widehat{F}-F)\right|\leq A_1\left\|\widehat{F}-F\right\|$, and $\left|R\Phi(F,\widehat{F}-F)\right|\leq A_2\left\|\widehat{F}-F\right\|^2$.

By Assumptions 1'-4' and Lemma B.3 in Newey (1994), $\left\|\widehat{F}-F\right\|\to 0$ in probability. Hence, it follows from above that $\widehat{m}(x,e)\to m(x,e)$ in probability. Moreover, since, as we have shown, Φ is Hadamard differentiable at $F,$ it follows by Theorem 3.9.4 in van der Vaart and Wellner (1996) and the Lemma in Appendix B that

$$\sqrt{N\sigma_N^{2K}}\left(\Phi\left(\widehat{\widetilde{F_Y}}\right)-\Phi\left(\widetilde{F_Y}\right)\right)-D\Phi\left(F,\sqrt{N\sigma_N^{2K}}\left(\widehat{\widetilde{F_Y}}-\widetilde{F_Y}\right)\right)$$

converges in outer probability to 0. By (2),

$$D\Phi\left(F,\sqrt{N\sigma_N^{2K}}\left(\widehat{\widetilde{F_Y}}-\widetilde{F_Y}\right)\right)$$
$$=\frac{1}{f_{y|x,0}^{-1}(m(x,e))}\int\int\frac{\left[1(s\leq e)-F_{y|0,x}(e)\right]}{f(0,x)}\sqrt{N\sigma_N^{2K}}\left(\widehat{f}(s,0,x)-f(s,0,x)\right)ds$$
$$-\frac{1}{f_{y|x,0}(m(x,e))}\int\int\frac{\left[1(s\leq m(x,e))-F_{y|0,x}(m(x,e))\right]}{f(x,0)}\sqrt{N\sigma_N^{2K}}\left(\widehat{f}(s,x,0)-f(s,x,0)\right)ds.$$

Hence, it follows by the Lemma in Appendix B that

$$\sqrt{N}\,\sigma_N^K\,(\hat{m}(x,e)-m(x,e)) = \sqrt{N}\,\sigma_N^K\left(\Phi(\hat{F})-\Phi(F)\right) \to N(0,V_m)$$

where $V_m = \left\{\int\left[\int K(s,z_1,z_2)ds\right]^2 dz_1 dz_2\right\}\left[\frac{1}{f_{y|x,0}(m(x,e))}\right]^2\ L$ and

$$L = \int\left[\frac{1(s<e)}{f(0,x)} - \frac{F_{y|0,x}(e)}{f(0,x)}\right]^2\ f(s,0,x)\ ds\ +\ \int\left[\frac{1(s<m(x,e))}{f(x,0)} - \frac{F_{y|x,0}(m(x,e))}{f(x,0)}\right]^2\ f(s,x,0)\ ds$$

$$= \frac{1}{f(0,x)}F_{y|0,x}(e)(1-F_{y|0,x}(e)) + \frac{1}{f(x,0)}F_{y|x,0}(m(x,e))(1-F_{y|x,0}(m(x,e)))$$

Next, to determine the asymptotic properties of the estimator for $F_{e|x}(e)$, define the functional $\Theta(G)$ by $\Theta(G) = F_{y|x}(\Phi(G))$, where $\Phi(G)$ is as defined above. Then, $\Theta(F) = F_{e|x}(e)$ and $\Theta(\widehat{F}) = \widehat{F}_{e|x}(e)$. Moreover,

$$\Theta(F+H)-\Theta(F) = \frac{\int^{\Phi(F+H)} f(y',x)\ dy' + \int^{\Phi(F+H)} h(y',x)\ dy'}{f(x)+h(x)} - \frac{\int^{\Phi(F)} f(y',x)dy}{f(x)}.$$

By Taylor's Theorem, there exist real numbers R_f, R_h, a_f and a_h such that

$$\int^{\Phi(F+H)} f(y',x)\ dy' = \int^{\Phi(F)} f(y',x)dy' + f(\Phi(F),x)(\Phi(F+H)-\Phi(F)) + R_f$$

and

$$\int^{\Phi(F+H)} h(y',x)\ dy' = \int^{\Phi(F)} h(y',x)dy' + h(\Phi(F),x)(\Phi(F+H)-\Phi(F)) + R_h,$$

where $|R_f| \leq a_f\,\|\Phi(F+H)=\Phi(F)\|^2$ and $|R_h| \leq a_h\,\|\Phi(F+H)=\Phi(F)\|^2$.

Hence,

$$\Theta(F+H)-\Theta(F)$$

$$= \frac{\int^{\Phi(F)} f(y',x)dy' + f(\Phi(F),x)(\Phi(F+H)+R_f-\Phi(F))}{f(x)+h(x)}$$

$$+\frac{\int^{\Phi(F)} h(y',x)dy' + h(\Phi(F),x)(\Phi(F+H)-\Phi(F))+R_h}{f(x)+h(x)}$$

$$-\frac{\int^{\Phi(F)} f(y',x)dy}{f(x)}$$

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

$$+\frac{f(x)\ h(\Phi(F),x)(\Phi(F+H)-\Phi(F)) + f(x)R_h - h(x)\int^{\Phi(F)} f(y',x)dy}{f(x)^2}$$

$$+\left[f(x)\ f(\Phi(F),x)(\Phi(F+H)-\Phi(F)) + f(x)R_f + f(x)\int^{\Phi(F)} h(y',x)dy'\right]$$

$$\left[\frac{1}{f(x)(f(x)+h(x))} - \frac{1}{f(x)^2}\right]$$

$+\left[f(x)\ h(\Phi(F),x)(\Phi(F+H)-\Phi(F))+f(x)R_h-h(x)\int^{\Phi(F)}f(y',x)dy\right]$

$\left[\frac{1}{f(x)(f(x)+h(x))}-\frac{1}{f(x)^2}\right].$

From above,

$$\Phi(F+H)-\Phi(F))=D\Phi(F,H)+R\Phi(F,H),$$

where for some A_1,A_2,

$$|D\Phi(F,H|\leq A_1\|H\| \text{ and } |R\Phi(F,H)|\leq A_2\|H\|.$$

Hence,

$\Theta(F+H)-\Theta(F)=$

$=\frac{\left[f(x)\ f(\Phi(F),x)(D\Phi(F,H)+R\Phi(F,H))+f(x)R_f+f(x)\int^{\Phi(F)}h(y',x)dy'\right]}{f(x)^2}$

$+\frac{f(x)\ h(\Phi(F),x)(D\Phi(F,H)+R\Phi(F,H))+f(x)R_h-h(x)\int^{\Phi(F)}f(y',x)dy}{f(x)^2}$

$+\left[f(x)\ f(\Phi(F),x)(\Phi(F+H)-\Phi(F))+f(x)R_f+f(x)\int^{\Phi(F)}h(y',x)dy'\right]$

$\left[\frac{1}{f(x)(f(x)+h(x))}-\frac{1}{f(x)^2}\right]$

$+\left[f(x)\ h(\Phi(F),x)(\Phi(F+H)-\Phi(F))+f(x)R_h-h(x)\int^{\Phi(F)}f(y',x)dy\right]$

$\left[\frac{1}{f(x)(f(x)+h(x))}-\frac{1}{f(x)^2}\right].$

Let

$D\Theta(F,H)=\frac{\left[f(x)\ f(\Phi(F),x)D\Phi(F,H)+f(x)\int^{\Phi(F)}h(y',x)dy'-h(x)\int^{\Phi(F)}f(y',x)dy\right]}{f(x)^2}$ and

$R\Theta(F,H)$

$=\frac{\left[f(x)\ f(\Phi(F),x)R\Phi(F,H)+f(x)R_f+f(x)\ h(\Phi(F),x)(D\Phi(F,H)+R\Phi(F,H))+f(x)R_h\right]}{f(x)^2}$

$+\left[f(x)\ f(\Phi(F),x)(\Phi(F+H)-\Phi(F))+f(x)R_f+f(x)\int^{\Phi(F)}h(y',x)dy'\right]$

$\left[\frac{-h(x)}{f(x)^2(f(x)+h(x))}\right]$

$+\left[f(x)\ h(\Phi(F),x)(\Phi(F+H)-\Phi(F))+f(x)R_h-h(x)\int^{\Phi(F)}f(y',x)dy\right]$

$\left[\frac{-h(x)}{f(x)^2(f(x)+h(x))}\right].$

Then, for some C and D,

$$|D\Theta(F,H)| \leq C\left\|H\right\| \text{ and } |R\Phi(F,H)| \leq D\left\|H\right\|^2.$$

Hence, Θ is Hadamard differentiable at F. Moreover, since

$$D\Theta(F,H) = \frac{\left[f(x)\ f(\Phi(F),x)D\Phi(F,H)+f(x)\int^{\Phi(F)} h(y',x)dy'-h(x)\int^{\Phi(F)} f(y',x)dy\right]}{f(x)^2}$$

and

$$D\Phi(F,H) = \frac{1}{f_{y|x,0}(m(x,e))}\int\int \frac{\left[1(s\leq e)-F_{y|0,x}(e)\right]}{f(0,x)}\ h(s,0,x)\ ds$$
$$-\frac{1}{f_{y|x,0}(m(x,e))}\int\int \frac{\left[1(s\leq m(x,e))-F_{y|x,0}(m(x,e))\right]}{f(x,0)}\ h(s,x,0)\ ds,$$

it follows that

$$D\Theta\left(F,\sqrt{N\sigma_N^{2K}}\left(\widehat{\widetilde{F}}-\widetilde{F}\right)\right)$$
$$= \frac{f(m(x,e),x)}{f_{y|x,0}(m(x,e))f(x)}\sqrt{N\sigma_N^{2K}}\int \frac{\left[1(s\leq e)-F_{y|0,x}(e)\right]}{f(0,x)}\left(\widehat{f}(s,0,x)-f(s,0,x)\right)ds$$
$$-\frac{f(m(x,e),x)}{f_{y|x,0}(m(x,e))f(x)}\sqrt{N\sigma_N^{2K}}\int \frac{\left[1(s\leq m(x,e))-F_{y|x,0}(m(x,e))\right]}{f(x,0)}\left(\widehat{f}(s,x,0)-f(s,x,0)\right)ds,$$
$$+\frac{1}{f(x)}\sqrt{N\sigma_N^{2K}}\int 1\left[s\leq m(x,e)\right]\left(\widehat{f}(s,x,z_2)-f(s,x,z_2)\right)dsdz_2$$
$$+\frac{\int^{m(x,e)} f(y',x)dy}{f(x)^2}\sqrt{N\sigma_N^{2K}}\int\left(\widehat{f}(s,x,z_2)-f(s,x,z_2)\right)ds\ dz_1\ dz_2.$$

Since, by the Lemma in Appendix B, the last two terms converge in probability to 0, it follows, again, by that Lemma that

$$\sqrt{N}\ \sigma_N^K\left(\widehat{F}_{e|x}(e)-F_{e|x}(e)\right) = \sqrt{N}\ \sigma_N^K\left(\Theta(\hat{F})-\Theta(F)\right)\to N(0,V_F)$$

where $V_F = \left\{\int\left[\int K(s,z_1,z_2)ds\right]^2 dz_1dz_2\right\}\left[\frac{f_{y|x}(m(x,e)}{f_{y|x,0}(m(x,e))}\right]^2\ L$ and

$$L = \int\left[\frac{1(s<e)}{f(0,x)}-\frac{F_{y|0,x}(e)}{f(0,x)}\right]^2\ f(s,0,x)\ ds\ +\ \int\left[\frac{1(s<m(x,e))}{f(x,0)}-\frac{F_{y|x,0}(m(x,e))}{f(x,0)}\right]^2\ f(s,x,0)\ ds$$
$$= \frac{1}{f(0,x)}F_{y|0,x}(e)(1-F_{y|0,x}(e))+\frac{1}{f(x,0)}F_{y|x,0}(m(x,e))(1-F_{y|x,0}(m(x,e))).$$

PROOF OF THEOREM 4: As in the proofs of the previous theorems, let $F(y,x,z)$ denote the distribution function (cdf) of the vector of observable variables (y,x,z), $f(y,x,z)$ denote its probability density function (pdf), $f(x,z)$ and $f(x)$ denote, respectively, the marginal pdf's of (x,z) and x, and $f(z|x)$ denote the conditional pdf of z given x. For any function $G : R^{d'} \to R$, define $g(y,x,z) = \partial^{d'} G(y,x,z)/\partial y \partial x \partial z$, $g(x,z) = \int g(y,x,z)\ dy$, $g(x) = \int g(y,x,z)\, dy\, dz$, $g(z|x) = g(x,z)\ /\ g(x)$, $g_x(y,x,z) = \partial g(y,x,z)/\partial x$, and $\widetilde{G_{Y}}_x(y,x,z) = \int^y \partial g(s,x,z)/\partial x\ ds = \int 1[s \le y]\ \partial g(s,x,z)/\partial x\ ds$. Let D denote the set of all functions $G : R^{d'} \to R$ such that $g(y,x,z)$ and $\partial g(y,x,z)/\partial x$ exist and vanish outside $\underline{C}$, where $\underline{C}$ is a compact set that strictly includes Θ. Let D denote the set of all functions $\widetilde{G_{Y}}_x$ that are generated from a G that belongs to D. Since there is a 1-1 relationship between functions in D and functions in D, we can define a functional on D or on D without altering its definition. Define then the functional $\Phi(\cdot)$ by:

$$\Phi(G) = \Phi\left(\widetilde{G_{Y}}_x\right) = \int \tfrac{\partial}{\partial x}\left(\int y g(y|x,z) dy\right)\ g(z|x)\ dz.$$

Let $\|G\|$ denote the sup norm of $\widetilde{G_{Y}}_x$. Then, there exists $\rho > 0$ such that if $\|G\| \le \rho$, then for some $a, b, d < \infty$

(1) $|g(x,0)| \le a\,\|G\|\,,\ \left|\int_{-\infty}^{y} g(s,x,0)ds\right| \le a\,\|G\|\,,$

$f(x,0) + g(x,0) \ge b\ g(x,0),\quad f(s,x,0) + g(s,x,0) \ge b\ f(s,x,0),$ and

$(F+G)^{-1}_{y|x,0}(F_{y|0,x}(e)) \in N(m(x,e),\xi).$

Define the functionals κ, Ψ_2, and λ by

$$\kappa(G) = \frac{g(x,0)\ \int^{e} \frac{\partial g(s,0,x)}{\partial x} ds}{g(0,x)\ g(\Phi(G),x,0)} - \frac{g(x,0)\ \int^{e} g(s,0,x)\ ds}{g(\Phi(G),x,0)\ g(0,x)^2}\frac{\partial g(0,x)}{\partial x}$$
$$+\frac{\int^{\Phi(G)} g(s,x,0)\ ds}{g(\Phi(G),x,0)\ g(x,0)}\frac{\partial g(x,0)}{\partial x} - \frac{\int^{\Phi(F)} \frac{\partial g(s,x,0)}{\partial x}}{g(\Phi(G),x,0)}$$

$\Psi_2(G) = \frac{g(x,0)\ g(e,0,x)}{g(\Phi(G),x,0)\ g(0,x)}$, and

$\lambda(G) = \frac{g(\Phi(G),x)}{g(x)}$,

where $\Phi(\cdot)$ is as defined in the proof of Theorem 3, $G : R^{d'} \to R$, and for all y, x, z, $G_{y|x,z}(y) = \int_{-\infty}^{y} g(s,x,z)ds/g(x,z)$, $g(s,x,z) = \partial^{L+1}G(s,x,z)/\partial s \partial x \partial z$, $g(x,z) = \int_{-\infty}^{\infty} g(s,x,z)ds$, and $g(x) = \int_{-\infty}^{\infty} g(s,x,z)dsdz$.

Then, $\kappa(F) = \frac{\partial m(x,e)}{\partial x}$, $\Psi_2(F) = \frac{\partial m(x,e)}{\partial e}$, $\lambda(F) = f_{y|x}(m(x,e))$, $\kappa(\widehat{F}) = \frac{\partial \widehat{m}(x,e)}{\partial x}$, $\Psi_2(\widehat{F}) = \frac{\partial \widehat{m}(x,e)}{\partial e}$, and $\lambda(\widehat{F}) = \widehat{f}_{y|x}(\widehat{m}(x,e))$. Define the functional Γ by

$$\Gamma(G) = \int \kappa(G)\ \Psi(G)\ \lambda(G)\ de.$$

Then, $\Gamma(F) = \beta(x)$ and $\Gamma(\widehat{F}) = \widetilde{\beta}(x)$.

As in the proofs of the previous theorems, we will study the asymptotic properties of $\widetilde{\beta}(x)$ by first obtaining a first order Taylor expansion of the nonlinear functional defining $\widetilde{\beta}(x)$. With this aim, we will first obtain first order Taylor expansions of the functionals $\kappa,\ \Psi_2,$ and λ.

To obtain a first order Taylor expansion of $\kappa,$ define the following functionals:

$\mu(G) = \frac{1}{g(\Phi(g),x,0)},\ \beta_1(G) = \frac{\partial g(x,0)}{\partial x},\ \gamma_1(G) = \int^{\Phi(G)} \frac{\partial g(s,x,0)}{\partial x} ds,$

$\Lambda(G) = \frac{\int^e g(s,0,x)\ ds}{g(0,x)},\ \beta_2(G) = \frac{\partial g(0,x)}{\partial x},\ \gamma_2(G) = \int^e \frac{\partial g(s,0,x)}{\partial x} ds,$

$\delta(G) = \int^{\Phi(G)} f(s,x,0)\ ds,\ \eta_1(G) = g(\ x,0),$

$\nu_1(G) = \frac{1}{g(x,0)},$ and $\nu_2(G) = \frac{1}{g(0,x)}.$

Then,

$$\kappa(G) = \mu(G)\,\eta_1(G)\ \gamma_2(G)\,\nu_1(G)\ - \mu(G)\,\eta_1(G)\ \Lambda_1(G)\,\nu_1(G)\,\beta_2(G)$$

$$+\mu(G)\ \nu_2(G)\ \delta(G)\,\beta_1(G)\ - \mu(G)\,\gamma_1.$$

Replacing x by $(x,0)$ in the proof of Theorem 4 in Matzkin(1999) everywhere except where x appears as the argument of the function m, and replacing $\widetilde{e}$ by e and $\widetilde{x}$ by $(0,x)$ in the proof of that same theorem, it follows that there exists $J_1 < \infty$ such that

$$\mu(F+H) - \mu(F) = D\mu(F,H) + R\mu(F,H),$$

$$\beta_1(F+H) - \beta_1(F) = D\beta_1(F,H) + R\beta_1(F,H),$$

$$\gamma_1(F+H) - \gamma_1(F) = D\gamma_1(F,H) + R\gamma_1(F,H),$$

$$\Lambda_1(F+H) - \Lambda_1(F) = D\Lambda_1(F,H) + R\Lambda_1(F,H),$$

$$|D\mu(F,H)| \le J_1\,\|H\|\,,\ |R\mu(F,H)| \le J_1\,\|H\|^2\,,$$

$$|D\beta_1(F,H)| \le J_1\,\|H\|^2\,,$$

$$|D\gamma_1(F,H)| \le J_1\,\|H\|\,,\ \ |R\gamma_1(F,H)| \le J_1\,\|H\|^2\,,$$

$$|D\Lambda_1(F,H)| \le J_1\,\|H\|\,,\ \ |R\Lambda_1(F,H)| \le J_1\,\|H\|^2\,,$$

$$D\mu(F,H) = \frac{-\frac{\partial f(\Phi(F),x,0)}{\partial y} D\Phi(F) - h(\Phi(F),x,0)}{f(\Phi(F),x,0)^2},$$

$$D\beta_1(F,H) = \frac{\partial h(x,0)}{\partial x},$$

$D\gamma_1(F,H) = \frac{\partial f(\Phi(F),x,0)}{\partial x} D\Phi(F) + \int^{\Phi(F)} \frac{\partial h(s,x,0)}{\partial x} ds$, and

$D\Lambda_1(F,H) = \frac{\int^e h(s,0,x)\ ds\ -\ h(0,x)\ F_{y|0x}(e)}{f(0,x)}$.

Making the same substitutions in the proof of Theorem 5 in Matzkin (1999), it follows that there exists $J_2 < \infty$ such that

$$\eta_1(F+H) - \eta_1(F) = D\eta_1(F,H)$$

$$\nu_1(F+H) - \nu_1(F) = D\nu_1(F,H) + R\nu_1(F,H),$$

$$|D\eta_1(F,H)| \leq J_2 \|H\|,\ |D\nu_1(F,H)| \leq J_2 \|H\|^2,\ |R\nu_1(F,H)| \leq J_2 \|H\|^2,$$

$D\eta_1(F,H) = h(x,0)$, and

$$Dv_1(F,H) = -\frac{h(0,x)}{f(0,x)^2}.$$

By making the obvious modifications to the derivations of $D\beta_1$ and $D\nu_1$, it is easy to obtain that there exists $J_3 < \infty$ such that

$$\beta_2(F+H) - \beta_2(F) = D\beta_2(F,H)$$

$$\nu_2(F+H) - \nu_2(F) = D\nu_2(F,H) + R\nu_2(F,H),$$

$$|D\beta_2(F,H)| \leq J_3 \|H\|,\ |D\nu_2(F,H)| \leq J_3 \|H\|^2,\ |R\nu_2(F,H)| \leq J_3 \|H\|^2,$$

$D\beta_2(F,H) = \frac{\partial h(0,x)}{\partial x}$, and

$$Dv_2(F,H) = -\frac{h(x,0)}{f(x,0)^2}.$$

To obtain a first order Taylor expansion for γ_2, we note that

$$\gamma_2(F+H) - \gamma_2(F) = \int^e \frac{\partial f(s,0,x)}{\partial x} ds + \int^e \frac{\partial h(s,0,x)}{\partial x} ds - \int^e \frac{\partial f(s,0,x)}{\partial x} ds.$$

Let $D\gamma_2(F,H) = \int^e \frac{\partial h(s,0,x)}{\partial x} ds.$

Then,

$\gamma_2(F+H) - \gamma_2(F) = D\gamma_2(F,H)$ and, for some $a < \infty$,

$$|D\gamma_2(F,H)| \leq a \|H\|.$$

To obtain a first order expansion for δ, we note that

$$\delta(F+H)-\delta(F)=\int^{\Phi(F+H)} f(s,x,0)ds+\int^{\Phi(F+H)} h(s,x,0)ds-\int^{\Phi(F)} f(s,x,0)ds.$$

By Taylor's Theorem, there exist $e_f, e_h, R_f,$ and R_h such that

$$\int^{\Phi(F+H)} f(s,x,0)ds-\int^{\Phi(F)} f(s,x,0)ds$$

$$=f(\Phi(F),x,0)\left[\Phi(F+H)-\Phi(F)\right]+R_f,$$

and

$$\int^{\Phi(F+H)} h(s,x,0)ds-\int^{\Phi(F)} h(s,x,0)ds$$

$$=h(\Phi(F),x,0)\left[\Phi(F+H)-\Phi(F)\right]+R_h,$$

where $|R_f|\leq e_f\left\|\Phi(F+H)-\Phi(F)\right\|^2$ and $|R_h|\leq e_h\left\|\Phi(F+H)-\Phi(F)\right\|^2$. Hence,

$$\begin{aligned}\delta(F+H)-\delta(F)&=f(\Phi(F),x,0)\left[\Phi(F+H)-\Phi(F)\right]+R_f\\&\quad+\int^{\Phi(F)} h(s,x,0)ds+h(\Phi(F),x,0)\left[\Phi(F+H)-\Phi(F)\right]+R_h.\end{aligned}$$

Let

$$D\delta(F,H)=f(\Phi(F),x,0)\,D\Phi(F,H)+\int^{\Phi(F)} h(s,x,0)ds \quad \text{and}$$

$$R\delta(F,H)=\delta(F+H)-\delta(F)-D\delta(F,H).$$

Then, for some $c_1, c_2<\infty$,

$$|D\delta(F,H)|\leq c_1\left\|H\right\| \text{ and } |R\delta(F,H)|\leq c_2\left\|H\right\|^2.$$

Let $\mu=\mu(F,H)$, $\eta_1=\eta_1(F,H)$, $\gamma_2=\gamma_2(F,H)$, $\nu_1=\nu_1(F,H)$, $\Lambda_1=\Lambda_1(F,H)$, $\beta_2=\beta_2(F,H)$, $\nu_2=\nu_2(F,H)$, $\delta=\delta(F,H)$, $\beta_1=\beta_1(F,H)$, and $\gamma_1=\gamma_1(F,H)$.

Let

$$\begin{aligned}D\kappa(F,H)&=D\mu\,\eta_1\,\gamma_2\,\nu_1+\mu\,D\eta_1\,\gamma_2\,\nu_1+\mu\,\eta_1\,D\gamma_2\,\nu_1+\mu\,\eta_1\,\gamma_2\,D\nu_1\\&\quad-D\mu\,\eta_1\,\Lambda_1\,\nu_1\,\beta_2-\mu\,D\eta_1\,\Lambda_1\,\nu_1\,\beta_2-\mu\,\eta_1\,D\Lambda_1\,\nu_1\,\beta_2\\&\quad-\mu\,\eta_1\,\Lambda_1\,D\nu_1\,\beta_2-D\mu\,\eta_1\,\Lambda_1\,\nu_1 D\beta_2+D\mu\,\nu_2\,\delta\;\beta_1\\&\quad+\mu\,D\nu_2\,\delta\;\beta_1+\mu\,\nu_2\,D\delta\;\beta_1+\mu\,\nu_2\,\delta\,D\beta_1\\&\quad-D\mu\,\gamma_1-\mu\,D\gamma_1.\end{aligned}$$

Let $R\kappa(F,H) = \kappa(F+H) - \kappa(F)$.

Then, it is easy to verify that there exists $J_4 < \infty$ such that

$$|D\kappa(F,H)| \leq J_4 \|H\| \text{ and } |R\kappa(F,H)| \leq J_4 \|H\|^2 .$$

This provides a first order Taylor's expansion for κ.

To obtain a first order Taylor expansion for Ψ_2, we note that by replacing x in the proof of Theorem 5 in Matzkin (1999) with $(x,0)$, everywhere, in that proof, except where x appears as an argument in the function m, and by replacing $\widetilde{x}$ and $\widetilde{e}$ in that proof with $(0,x)$ and e, respectively, it follows that for all H such that $\|H\| \leq \rho$,

$$\Psi_2(F+H) - \Psi_2(F) = D\Psi_2(F,H) + R\Psi_2(F,H),$$

where

$$D\Psi_2(F,H) = -\frac{\frac{\partial f(m(x,e,),x,0)}{\partial y}\, D\Phi(F)\, f(x,0)\, f(e,0,x)}{f(m(x,e),x,0)^2\, f(0,x)} - \frac{h(m(x,e),x,0)\, f(x,0)\, f(e,0,x)}{f(m(x,e),x,0)^2\, f(0,x)}$$
$$+\frac{h(x,0)\, f(e,0,x)}{f(m(x,e),x,0)\, f(0,x)} + \frac{f(x,0)\, h(e,0,x)}{f(m(x,e),x,0)\, f(0,x)}$$
$$-\frac{f(x,0)\, f(e,0,x)\, h(0,x)}{f(m(x,e),x,0)\, f(0,x)^2},$$

$$R\Psi_2(F,H) = \Psi_2(F+H) - \Psi_2(F,H) - D\Psi_2(F,H),$$

and for some $E', F' < \infty$,

$$|D\Psi_2(F,H)| \leq E' \|H\| \text{ and } |R\Psi_2(F,H)| \leq F' \|H\|^2 .$$

To obtain a first order expansion for λ, we note that for any H such that $\|H\| \leq \rho$,

$$\lambda(F+H) - \lambda(F) = \frac{f(\Phi(F+H),x)+h(\Phi(F+H),x)}{f(x)+h(x)} - \frac{f(\Phi(F+H),x)}{f(x)}$$
$$= \frac{f(x)\, f(\Phi(F+H),x)+f(x)\, h(\Phi(F+H),x)-f(x)\, f(\Phi(F),x)-h(x)\, f(\Phi(F),x)}{f(x)\,(f(x)+h(x))}.$$

By Taylor's Theorem, there exist R'_f, R'_h, d_f, and d_h such that

$$f(\Phi(F+H),x) - f(\Phi(F),x) = \frac{\partial f(\Phi(F),x)}{\partial s}(\Phi(F+H) - \Phi(F)) + R'_f, \text{ and}$$

$$h(\Phi(F+H),x) - h(\Phi(F),x) = \frac{\partial h(\Phi(F),x)}{\partial s}(\Phi(F+H) - \Phi(F)) + R'_h,$$

where $R'_f \leq d_f \|\Phi(F+H) - \Phi(F)\|^2$ and $R'_h \leq d_h \|\Phi(F+H) - \Phi(F)\|^2$.

From the proof of Theorem 3, it follows that

$$\Phi(F+H)-\Phi(F)=D\Phi(F,H)+R\Phi(F,H)$$

where

$$D\Phi(F,H)=\frac{f(x,0)}{f(0,x)^2 f(\Phi(F),x,0)}A\widetilde{x}+\frac{f(x,0)}{f(x,0)^2 f(\Phi(F),x,0)}Ax,$$

$$A\widetilde{x}=f(0,x)\int^{e}h(s,0,x)ds-h(0,x)\int^{e}f(s,0,x)ds\,,$$

$$Ax=f(x,0)\int^{\Phi(F)}h(s,x,0)ds-h(x,0)\int^{\Phi(F)}f(s,x,0)ds,$$

and for some A_1, A_2,

$$|D\Phi(F,H)|\leq A_1\|H\| \text{ and } \quad |R\Phi(F,H)|\leq A_2\|H\|^2.$$

Then,

$$\lambda(F+H)-\lambda(F)=\left[\frac{1}{f(x)^2}+\frac{1}{f(x)\,(f(x)+h(x))}-\frac{1}{f(x)^2}\right]\cdot$$
$$\left[f(x)\frac{\partial f(\Phi(F),x)}{\partial s}D\Phi(F,H)+f(x)\,h(\Phi(F),x)-h(x)\,f(\Phi(F),x)\right]$$
$$+\left[\frac{1}{f(x)^2}+\frac{1}{f(x)\,(f(x)+h(x))}-\frac{1}{f(x)^2}\right]\cdot$$
$$\left[f(x)\frac{\partial f(\Phi(F),x)}{\partial s}R\Phi(F,H)+f(x)\,R_f'+f(x)\frac{\partial h(\Phi(F),x)}{\partial s}(D\Phi(F,H)+R\Phi(F,H))+f(x)\,R_h'\right].$$

Let

$$D\lambda(F,H)=\left[\frac{\frac{\partial f(\Phi(F),x)}{\partial s}}{f(x)}D\Phi(F,H)+\frac{h(\Phi(F),x)}{f(x)}-\frac{h(x)\,f(\Phi(F),x)}{f(x)^2}\right]$$

and

$$R\lambda(F,H)=\lambda(F+H)-\lambda(F)-D\lambda(F,H).$$

Then, there exists J_5 such that

$$|D\lambda(F,H)|\leq J_5\|H\| \text{ and } |R\lambda(F,H)|\leq J_5\|H\|^2.$$

We can now obtain a first order Taylor expansion for Γ. Let

$$D\Gamma(F,H)=\int D\kappa(F,H)\;\Psi_2(F)\;\lambda(F)\;de+\int\kappa(F,H)\;D\Psi_2(F)\;\lambda(F)\;de$$
$$+\int\kappa(F,H)\;\Psi_2(F)\;D\lambda(F)\;de,$$

and

$$R\Gamma(F,H) = \Gamma(F+H) - \Gamma(F).$$

Then, for some $\overline{E}, \overline{F} < \infty$,

$$|D\Gamma(F,H)| \leq \overline{E}\ \|H\| \text{ and } |R\Gamma(F,H)| \leq \overline{F}\ \|H\|^2 .$$

In particular, $D\Gamma(F,\cdot)$ is continuous. Letting $H = \widehat{F} - F$, the above Taylor expansion implies that $\widehat{\beta}(x) \to \beta(x)$ in probability, since by Assumptions 1' and 2"-4" and by Lemma B.3 in Newey (1994), $\left\|\widehat{F} - F\right\| \to 0$ in probability. Next, let

$$D_1\Gamma(F,H) = \int \left[\frac{f(x,0)}{f(m(x,e),x,0)} \left(\int^e \frac{\partial h(s,0,x)}{\partial x} ds\right) \frac{1}{f(0,x)}\right] \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e)}{f_{y|x,0}(m(x,e))}\right] de$$
$$- \int \left[\frac{f(x,0)}{f(m(x,e),x,0)} \left(\int^e \frac{f(s,0,x)}{f(0,x)} ds\right) \frac{1}{f(0,x)} \frac{\partial h(0,x)}{\partial x}\right] \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e)}\right] de$$
$$+ \int \left[\frac{1}{f(m(x,e),x,0)} \left(\int^{\Phi(F)} \frac{f(s,x,0)}{f(x,0)} ds\right) \frac{\partial h(0,x)}{\partial x}\right] \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))}\right] de$$
$$- \int \left[\frac{1}{f(m(x,e),x,0)} \left(\int^{\Phi(F)} \frac{\partial h(s,x,0)}{\partial x} ds\right)\right] \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))}\right] de$$
$$= \int \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e)}{f_{y|x,0}(m(x,e))^2 f(0,x)}\right] \left[\left(\int 1[s \leq e] \frac{\partial h(s,0,x)}{\partial x} ds - F_{y|0,x}(e) \frac{\partial h(0,x)}{\partial x}\right)\right]\ de$$
$$- \int \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))^2 f(x,0)}\right] \left[\int 1[s \leq m(x,e)] \frac{\partial h(s,x,0)}{\partial x} ds - F_{y|x,0}(m(x,e)) \frac{\partial h(x,0)}{\partial x}\right] de.$$

By the Lemma in Appendix B, all the terms of $\sqrt{N}\sigma_N^{K+1} D\Gamma(F, \widehat{F} - F)$, except those in $\sqrt{N}\sigma_N^{K+1} D_1\Gamma(F, \widehat{F} - F)$, converge in probability to 0. Hence, by Theorem 3.9.4 in van der Vaart and Wellner (1996) and by the Lemma in Appendix B, it follows that

$$\sqrt{N}\sigma_N^{K+1}(\Gamma(\widehat{F}) - \Gamma(F)) = \sqrt{N}\sigma_N^{K}(\widetilde{\beta}(x) - \beta(x)) \to N(0, V_\beta)$$

where

$$V_\beta = \overline{K}\left[L_1 + L_2\right],$$

$$L_1 = \int \left\{\int \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))^2}\right] \left(1[s \leq e] - F_{y|0,x}(e)\right) de\right\}^2 f_{y|0,x}(s) ds$$

$$L_2 = \int \left\{\int \left[\frac{f_{y|0,x}(e)\, f_{y|x}(m(x,e))}{f_{y|x,0}(m(x,e))^2}\right] \left(1[s \leq m(x,e)] - F_{y|0,x}(m(x,e))\right) de\right\}^2 f_{y|x,0}(s) ds,$$

and

$$\widetilde{K} = \int\int \left(\int \frac{\partial K(s,x,z)}{\partial x} ds\right) \left(\int \frac{\partial K(s,x,z)}{\partial x} ds\right)' dx\, dz.$$

9 Appendix B

LEMMA: Suppose that the following assumptions are satisfied:

(i) $\{y_i, x_i\}$ is iid, y_i has values in R^L and x_i has values in R^Q.

(ii) The joint density $f(y,x)$ has a compact support $\Theta \subset R^{L+Q}$, f is continuously differentiable up to the order $g = t + s$ for some even $s > 0$.

(iii) The kernel function $K(\cdot,\cdot)$ is continuously differentiable, K vanishes outside a compact set, $\int K(y,x)dydx = 1$, and K is a kernel of order s.

(iv) The function $r(y,x)$ is continuous and bounded a.e.

Then,

(I) if $t = 0$, $N\sigma_N^Q \to \infty$ and $\sigma_N^s \sqrt{N\sigma_N^Q} \to 0$, then

$\sqrt{N\sigma_N^Q} \int r(y,x) \left(\widehat{f}(y,x) - f(y,x)\right) dy \to N(0, V_1)$

where $V_1 = \left\{\int [r(y,x)]^2 f(y,x)dy\right\} \left\{\int \left(\int K(y,x)dy\right)^2 dx\right\}$,

and for any two distinct points $x^{(1)}$ and $x^{(2)}$,

$\sqrt{N\sigma_N^Q} \int r(y,x^{(1)}) \left(\widehat{f}(y,x^{(1)}) - f(y,x^{(1)})\right) dy$ and

$\sqrt{N\sigma_N^Q} \int r(y,x^{(1)}) \left(\widehat{f}(y,x^{(1)}) - f(y,x^{(1)})\right) dy$ are asymptotically independent.

(II) if $t = 1$, $N\sigma_N^{Q+2} \to \infty$ and $\sigma_N^s \sqrt{N\sigma_N^{Q+2}} \to 0$ then

$\sqrt{N\sigma_N^{Q+2}} \int r(y,x) \left(\frac{\partial \widehat{f}(y,x)}{\partial x} - \frac{\partial f(y,x)}{\partial x}\right) dy \to N(0, V_2)$

where $V_2 = \left\{\int [r(y,x)]^2 f(y,x)dy\right\} \widetilde{K_2}$

and $\widetilde{K_2} = \left\{\int \left(\int\int \frac{\partial K(y,x)}{\partial x} dy\right) \left(\int\int \frac{\partial K(y,x)}{\partial x} dy\right)' dx\right\}$.

Moreover, for any two distinct points $x^{(1)}$ and $x^{(2)}$,

$\sqrt{N\sigma_N^{Q+2}} \int r(y,x^{(1)}) \left(\frac{\partial \widehat{f}(y,x^{(1)})}{\partial x} - \frac{\partial f(y,x^{(1)})}{\partial x}\right) dy$ and

$\sqrt{N\sigma_N^{Q+2}} \int r(y,x^{(2)}) \left(\frac{\partial \widehat{f}(y,x^{(2)})}{\partial x} - \frac{\partial f(y,x^{(2)})}{\partial x}\right) dy$ are asymptotically independent.

PROOF: We first show (I). By the definition of $\widehat{f}$,

$$\int r(y,x)\,(\widehat{f}(y,x) - f(y,x))\,dy$$

$$= \int r(y,x) \left(\frac{1}{N} \sum_{i=1}^{N} \left[\frac{1}{\sigma^{L+Q}} K(\frac{y_i - y}{\sigma}, \frac{x_i - x}{\sigma},) - f(y,x) \right] \right) dy$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[\int r(y,x) \frac{1}{\sigma^{L+Q}} K(\frac{y_i - y}{\sigma}, \frac{x_i - x}{\sigma}) dy - \int r(y,x) f(y,x) dy \right]$$

Let

$$w_i = \frac{1}{\sigma^{L+Q}} \int r(y,x) K(\frac{y_i - y}{\sigma}, \frac{x_i - x}{\sigma}) dy.$$

Then,

$$\int r(y,x)\,(\widehat{f}(y,x) - f(y,x))\,dy$$

$$= \frac{1}{N} \sum_{i=1}^{N} \left[w_i - \int r(y,x)\; f(y,x)\, dy \right]$$

$$= \frac{1}{N} \sum_{i=1}^{N} [w_i - E(w_i)] + \left[E(w_i) - \int r(y,x)\; f(y,x)\, dy \right]$$

We will show that under the above assumptions, the first term is asymptotically normal and the second converges to 0. For this, we note that

$$E(w_i) = E\left(\frac{1}{\sigma^{L+Q}} \int r(y,x) K(\frac{y_i - y}{\sigma}, \frac{x_i - x}{\sigma}) dy \right)$$

$$= \int \int \left(\frac{1}{\sigma^{L+Q}} \int r(y,x) K(\frac{y_i - y}{\sigma}, \frac{x_i - x}{\sigma}) dy \right) f(y_i, x_i) dy_i dx_i$$

$$= \int \int \left(\int r(y,x) K(\widetilde{y}, \widetilde{x}) dy \right) f(y + \sigma \widetilde{y}, x + \sigma \widetilde{x}) d\widetilde{y} d\widetilde{x}$$

$$= \int r(y,x) \left(\int \int K(\widetilde{y}, \widetilde{x}) f(y + \sigma \widetilde{y}, x + \sigma \widetilde{x}) d\widetilde{y} d\widetilde{x} \right) dy$$

Using a Taylor's expansion of $f(y + \sigma\widetilde{y}, x + \sigma\widetilde{x})$ around $f(y,x)$ and using the assumption that the kernel function K integrates to 1 and is of order s, it follows that

(1) $E(w_i) = \int r(y,x) f(y,x) dy + O(\sigma^s)$.

Next,

$$E(w_i^2) = E\left(\frac{1}{\sigma^{2L+2Q}} \left[\int r(y,x) K(\frac{y_i - y}{\sigma}, \frac{x_i - x}{\sigma}) dy \right]^2 \right)$$

$$= E\left(\frac{1}{\sigma^{2Q}} \left[\int r(y_i - \sigma\widetilde{y}, x) K(\widetilde{y}, \frac{x_i - x}{\sigma}) d\widetilde{y} \right]^2 \right)$$

$$= \int \int \frac{1}{\sigma^{Q}} \left[\int r(y_i - \sigma\widetilde{y}, x) K(\widetilde{y}, \widetilde{x}) d\widetilde{y} \right]^2 f(y_i, x + \sigma\widetilde{x}) dy_i d\widetilde{x}$$

Then, by the continuity and boundedness of f and r, it follows by the Bounded Convergence Theorem that

$$(2)\ \sigma^Q E(w_i^2) \to \left[\int r(y,x)^2 f(y,x)dy\right]\left\{\int\left(\int K(y,x)dy\right)^2 dx\right\}.$$

From (1), (2), and $\sigma \to 0$ it follows that

$$\sigma^Q Var(w_i) \to \left[\int r(y,x)^2 f(y,x)dy\right]\left\{\int\left(\int K(y,x)dy\right)^2 dx\right\}$$

Let $\delta > 0$. Since

$$E\left|\tfrac{1}{N}(w_i - E(w_i)\right|^{2+\delta} \le \tfrac{2^{2+\delta}E|w_i|^{2+\delta}}{N^{2+\delta}}$$

$$E\left|w_i\right|^{2+\delta} = E\left|\tfrac{1}{\sigma^{L+Q}}\int r(y,x)K(\tfrac{y_i-y}{\sigma},\tfrac{x_i-x}{\sigma})dy\right|^{2+\delta}$$

$$= \int\int\left(\tfrac{1}{\sigma^{Q(2+\delta)}}\left|\int r(y_i-\sigma\widetilde{y},x)K(\widetilde{y},\tfrac{x_i-x}{\sigma})d\widetilde{y}\right|^{2+\delta}\right)f(y_i,x_i)dy_idx_i$$

$$= \int\int\left(\tfrac{1}{\sigma^{Q(1+\delta)}}\left|\int r(y_i-\sigma\widetilde{y},x_i-\sigma\widetilde{x})K(\widetilde{y},\widetilde{x})d\widetilde{y}\right|^{2+\delta}\right)f(y_i,x+\sigma\widetilde{x})d\widetilde{y}d\widetilde{x}$$

$$= O\left(\tfrac{1}{\sigma^{Q(1+\delta)}}\right)$$

and

$$\left(\frac{Var(w_i)}{N}\right)^{\frac{2+\delta}{2}} = O\left(\frac{1}{(N\sigma^Q)^{1+\frac{\delta}{2}}}\right),$$

it follows that

$$\frac{\sum_{i=1}^N E\left|(\frac{w_i}{N}-E(\frac{w_i}{N})\right|^{2+\delta}}{\left[\left(Var\sum_{i=1}^N\frac{w_i}{N}\right)^{1/2}\right]^{2+\delta}} = \frac{\sum_{i=1}^N E\left|(\frac{w_i}{N}-E(\frac{w_i}{N})\right|^{2+\delta}}{\left(Var(\frac{w_i}{N})\right)^{1+\frac{\delta}{2}}} = O\left(\tfrac{1}{N^{1+(\delta/2)}\sigma^{Q\delta/2}}\right) \to 0.$$

By Liapounov's Theorem it then follows that

$$\sqrt{N\sigma_N^Q}\left(\tfrac{1}{N}\sum_{i=1}^N w_i - E(w_i)\right) \to N(0,V_1)$$

where

$$V_1 = \left[\int r(y,x)^2 f(y,x)dy\right]\left\{\int\left(\int K(y,x)dy\right)^2 dx\right\}.$$

Since by (1) and by assumption,

$$\sqrt{N\sigma_N^Q}\left(E(w_i) - \int r(y,x)f(y,x)dy\right) = O\left(\sigma_N^s\sqrt{N\sigma_N^Q}\right) \to 0$$

the first part of (I) is proved.

Next, to show the asymptotic independence, we note that by (i) and the definition of $\widehat{f}$ the covariance equals

$$\frac{\sigma^Q}{\sigma^{2(L+Q)}}\left\{E\left[\left(\int r^1K^1\right)\left(\int r^2K^2\right)\right]-E\left(\int r^1K^1\right)E\left(\int r^2K^2\right)\right\}$$

where

$$\left(\int r^kK^k\right)=\int r(y,x^{(k)})K\left(\frac{y_i-y}{\sigma},\frac{x_i-x^{(k)}}{\sigma}\right)dy \quad k=1,2$$

Since

$$E\left[\left(\int r^1K^1\right)\left(\int r^2K^2\right)\right]$$
$$=\sigma^{2L+Q}\int\left(\int\widetilde{r}^1\widetilde{K}^1\right)\left(\int\widetilde{r}^1\widetilde{K}^1\right)f(y_i,x^{(1)}+\sigma\widetilde{x})dy_id\widetilde{x}$$

where $\int\widetilde{r}^1\widetilde{K}^1=\int r(y_i-\sigma\widetilde{y},x^{(1)})K(\widetilde{y},\widetilde{x})d\widetilde{y}$

and $\int\widetilde{r}^2\widetilde{K}^2=\int r(y_i-\sigma\widetilde{y},x^{(2)})K(\widetilde{y},\widetilde{x}+\frac{x^{(1)}-x^{(2)}}{\sigma})d\widetilde{y}$

it follows by bounded convergence, (1), and $\sigma\to 0$ that the covariance converges to 0.

We next show (II). By the definition of $\widehat{f}_x$,

$$\int\int r(y,x)\left(\widehat{f}_x(y,x)-f_x(y,x)\right)dy$$
$$=\int\int r(y,x)\left(\frac{1}{N}\sum_{i=1}^N\left[\frac{(-1)}{\sigma^{L+Q+1}}\frac{\partial K(\frac{y_i-y}{\sigma},\frac{x_i-x}{\sigma})}{\partial x}-\frac{\partial f(y,x)}{\partial x}\right]\right)dy$$
$$=\frac{1}{N}\sum_{i=1}^N\left[\int\int r(y,x)\frac{(-1)}{\sigma^{L+Q+1}}\frac{\partial K(\frac{y_i-y}{\sigma},\frac{x_i-x}{\sigma})}{\partial x}dy-\int\int r(y,x)\frac{\partial f(y,x)}{\partial x}dy\right]$$

Let $w_i=\frac{(-1)}{\sigma^{L+Q+1}}\int\int r(y,x)\frac{\partial K(\frac{y_i-y}{\sigma},\frac{x_i-x}{\sigma})}{\partial x}dy$. Then,

$$\int r(y,x)\left(\widehat{f}_x(y,x,z)-f_x(y,x,z)\right)dy$$
$$=\frac{1}{N}\sum_{i=1}^N\left[w_i-E(w_i)\right]+\left[E(w_i)-\int r(y,x)\frac{\partial f(y,x)}{\partial x}dy\right]$$

We note that

$$E(w_i)=E\left(\frac{(-1)}{\sigma^{L+Q+1}}\int r(y,x)\frac{\partial K(\frac{y_i-y}{\sigma},\frac{x_i-x}{\sigma})}{\partial x}dy\right)$$

$$= \int \int \left(\frac{(-1)}{\sigma^{L+Q+1}} \int r(y,x) \frac{\partial K(\frac{y_i - y}{\sigma}, \frac{x_i - x}{\sigma})}{\partial x} dy \right) f(y_i, x_i) dy_i dx_i$$

$$= \int \int \left(\frac{(-1)}{\sigma} \int r(y,x) \frac{\partial K(\widetilde{y},\widetilde{x})}{\partial x} dy \right) f(y + \sigma\widetilde{y}, x + \sigma\widetilde{x}) d\widetilde{y} d\widetilde{x}$$

$$= \int r(y,x) \left(\int \int K(\widetilde{y},\widetilde{x}) \frac{\partial f(y+\sigma\widetilde{y}, x+\sigma\widetilde{x})}{\partial x} d\widetilde{y} d\widetilde{x} \right) dy$$

where the last inequality follows by integration by parts. Using a Taylor's expansion of $\frac{\partial f(y+\sigma\widetilde{y},x+\sigma\widetilde{x})}{\partial x}$ around $\frac{\partial f(y,x)}{\partial x}$ and using the assumption that the kernel function K integrates to 1 and is of order s, it follows that

(3) $E(w_i) = \int \int r(y,x) \frac{\partial f(y,x)}{\partial x} dy + O(\sigma^s)$.

Next,

$$E(w_i w_i')$$

$$= E\left(\frac{1}{\sigma^{2(L+Q+1)}} \left[\int r(y,x) \frac{\partial K(\frac{y_i-y}{\sigma}, \frac{x_i-x}{\sigma})}{\partial x} dy \right] \left[\int r(y,x) \frac{\partial K(\frac{y_i-y}{\sigma}, \frac{x_i-x}{\sigma})}{\partial x'} dy \right] \right)$$

$$= E\left(\frac{1}{\sigma^{2Q+2}} r_i r_i' \right)$$

$$= \int \int \left(\frac{1}{\sigma^{Q+2}} \overline{r}_i \overline{r}_i' \right) f(y_i, x + \sigma\widetilde{x}) dy_i d\widetilde{x}$$

where $r_i = \int r(y_i - \sigma\widetilde{y}, x) \frac{\partial K(\widetilde{y}, \frac{x_i - x}{\sigma})}{\partial x} d\widetilde{y}$

and $\overline{r}_i = \int r(y_i - \sigma\widetilde{y}, x) \frac{\partial K(\widetilde{y}, \frac{x_i - x}{\sigma})}{\partial x} d\widetilde{y}$.

By the continuity and boundedness of f and r, it follows by the Bounded Convergence Theorem that

(4) $\sigma^{Q+2} E(w_i w_{i'}) \to \left[\int r(y,x)^2 f(y,x) dy \right] \widetilde{K}$

where $\widetilde{K} = \left\{ \int \left(\int \frac{\partial K(y,x)}{\partial x} dy \right) \left(\int \frac{\partial K(y,x)}{\partial x} dy \right)' dx \right\}$.

From (3), (4), and $\sigma \to 0$ it then follows that

$$\sigma^{Q+2} Var(w_i) \to \left[\int r(y,x)^2 f(y,x) dy \right] \widetilde{K}$$

To apply Liapounov's Central Limit Theorem, we note that, for $\delta > 0$,

$$E \left| \frac{1}{N} (w_i - E(w_i) \right|^{2+\delta} \leq \frac{2^{2+\delta} E|w_i|^{2+\delta}}{N^{2+\delta}}$$

where

$$E\left|w_i\right|^{2+\delta} = E\left|\frac{(-1)}{\sigma^{L+Q+1}}\int r(y,x)\frac{\partial K(\frac{y_i-y}{\sigma},\frac{x_i-x}{\sigma})}{\partial x}dy\right|^{2+\delta}$$
$$= \int\int\left(\frac{(-1)^{2+\delta}}{\sigma^{(Q+1)(2+\delta)}}\left|\int r(y_i-\sigma\widetilde{y},x)\frac{\partial K(\widetilde{y},\frac{x_i-x}{\sigma})}{\partial x}d\widetilde{y}\right|^{2+\delta}\right)f(y_i,x_i)dy_i dx_i$$
$$= \int\int\left(\frac{(-1)^{2+\delta}}{\sigma^{Q(1+\delta)+2+\delta}}\left|\int r(y_i-\sigma\widetilde{y},x)\frac{\partial K(\widetilde{y},\widetilde{x})}{\partial x}d\widetilde{y}\right|^{2+\delta}\right)f(y_i,x+\sigma\widetilde{x})dy_i y d\widetilde{x}$$
$$= O\left(\frac{1}{\sigma^{Q(1+\delta)+2+\delta}}\right)$$

and

$$\left(\frac{Var(w_i)}{N}\right)^{\frac{2+\delta}{2}} = O\left(\frac{1}{(N\sigma^{Q+2})^{1+\frac{\delta}{2}}}\right).$$

Hence,

$$\frac{\sum_{i=1}^N E\left|(\frac{w_i}{N}-E(\frac{w_i}{N})\right|^{2+\delta}}{\left[\left(Var\sum_{i=1}^N \frac{w_i}{N}\right)^{1/2}\right]^{2+\delta}} = \frac{\sum_{i=1}^N E\left|(\frac{w_i}{N}-E(\frac{w_i}{N})\right|^{2+\delta}}{\left(Var(\frac{w_i}{N})\right)^{1+\frac{\delta}{2}}} = O\left(\frac{1}{N^{1+\delta/2}\sigma^{Q\delta/2}}\right)\to 0.$$

By Liapounov's Theorem it then follows that

$$\sqrt{N\sigma_N^{Q+2}}\left(\frac{1}{N}\sum_{i=1}^N\left[w_i - E(w_i)\right]\right)\to N(0,V_2)$$

where $V_2 = \left[\int r(y,x)^2 f(y,x)dy\right]\widetilde{K}$

and $\widetilde{K} = \left\{\int\left(\int\frac{\partial K(y,x)}{\partial x}dy\right)\left(\int\frac{\partial K(y,x)}{\partial x}dy\right)' dx\right\}.$

Since, by (3), $\sqrt{N\sigma_N^{Q+2}}\left(E(w_i) - \int\int r(y,x)\frac{\partial f(y,x)}{\partial x})\,dy\right) = O(\sigma_N^s\sqrt{N\sigma_N^{Q+2}})\to 0$. The first part of result (II) is proved. To prove the second part of (II), we note that by (i) and the definition of $\widehat{\frac{\partial f}{\partial x}}$, the covariance equals

$$\frac{\sigma^{Q+2}}{\sigma^{2(L+Q+1)}}\left\{E\left[\left(\int r^1\partial K^1\right)\left(\int r^2\partial K^2\right)\right] - E\left(\int r^1\partial K^1\right)E\left(\int r^2\partial K^2\right)\right\}$$

where

$$\left(\int r^k\partial K^k\right) = \int r(y,x^{(k)})\frac{\partial K\left(\frac{y_i-y}{\sigma},\frac{x_i-x^{(k)}}{\sigma}\right)}{\partial x}dy \quad k=1,2$$

Since

$$E\left[\left(\int r^1\partial K^1\right)\left(\int r^2\partial K^2\right)\right]$$

$$= \sigma^{2L+Q+2} \int \left(\int \widetilde{r}^1 \partial \widetilde{K}^1 \right) \left(\int \widetilde{r}^1 \partial \widetilde{K}^1 \right) f(y_i, x^{(1)} + \sigma \widetilde{x}) dy_i d\widetilde{x}$$

where $\int \widetilde{r}^1 \widetilde{K}^1 = \int r(y_i - \sigma \widetilde{y}, x^{(1)}) \frac{\partial K(\widetilde{y}, \widetilde{x})}{\partial x} d\widetilde{y}$

and $\int \widetilde{r}^2 \widetilde{K}^2 = \int r(y_i - \sigma \widetilde{y}, x^{(2)}) \frac{\partial K(\widetilde{y}, \widetilde{x} + \frac{x^{(1)} - x^{(2)}}{\sigma})}{\partial x} d\widetilde{y}$

it follows by bounded convergence, (3), and $\sigma \to 0$ that the covariance converges to 0.

10 References

Aaronson, D. (1998)"Using Sibling Data to Estimate the Impact of Neighborhoods on Children's Educational Outcomes", *Journal of Human Resources*, (33:4), Fall, p. 915-946.

Abrevaya, J., (1997) "Rank Estimation of a Generalized Fixed Effects Regression Model", Unpublished Paper, University of Chicago.

Ait-Sahalia, Y. (1994) "The Delta and Bootstrap Methods for Nonlinear Functionals of Nonparametric Kernel Estimators Based on Dependent Multivariate Data," mimeo.

Altonji, J.G. and H. Ichimura, (1997) "Estimating Derivatives in Nonseparable Models with Limited Dependent Variables", mimeo.

Altonji, J.G., F. Hayashi, and L. Kotlikoff, (2000) "The Effects of Earnings and Wealth on Time and Money Transfers between Parents and Children" (with F. Hayashi and L. Kotlikoff) in Andrew Masson and Georges Tapinos, Editors, *Sharing the Wealth: Demographic Change and Economic Transfers between Generations*, Oxford University Press.

Blundell, Richard, and James L. Powell, "Endogeneity in Single Index Models". mimeo (August 1999).

Brown, D.J. and R.L. Matzkin, (1996) "Estimation of Nonparametric Functions in Simultaneous Equations Models, with an Application to Consumer Demand," mimeo, Northwestern University.

Chamberlain, G. (1980) "Analysis of Covariance with Qualitative Data", *Review of Economic Studies*, 47 225-238.

Chamberlain, G. (1984), "Panel Data" in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, Volume II, Amsterdam: North Holland.

Currie, J. and B. Fallick, (1996) "The Minimum Wage and Employment of Youth: Evidence from NLSY", *Journal of Human Resources*, Spring.

Currie, J. and D. Thomas, (1995) "Does Head Start Make a Difference", *American Economic Review*, June..

de Finetti, B., (1975) *Theory of Probability*, Volume 2 London, John Wiley and Sons.

Geronimus, A.T. and S. Korenman, (1992) "The socioeconomic consequences of teen childbearing reconsidered", *Quarterly Journal of Economics,* v107 n4 p1187-1228.

Hausman, J.A., and W.E. Taylor, (1982) “Panel Data and Unobservable Individual Effects”, *Econometrica* 49, pp. 1377-1398.

Heckman, J.J. (1981) “The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating A Discrete Time-Discrete Data Stochastic Process and Some Monte Carlo Evidence.”, in C. Manski and D. McFadden (eds.) *Structural Analysis of Discrete Data With Econometric Applications.* M.I.T. Press

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

Heckman, J.J. and E. Vytlacil, (1998) “Instrumental Variables Methods for the Correlated Random Coefficient Model,” *Journal of Human Resources*, 33(4): 974-1002.

Honoré, B. E., (1992) “Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects,” *Econometrica*, 60, pp. 533-565.

Honoré, B. E. and E. Kyriazidou, (2000) “Panel Data Discrete Choice Models with Lagged Dependent Variables”, *Econometrica*, 68(4), July: pp.839-874.

Horowitz, J.L. (1992) “A Smoothed Maximum Score Estimator for the Binary Response Model” *Econometrica,* 60 :505-531.

Jakubson,-G. (1991) “Estimation and Testing of the Union Wage Effect Using Panel Data,” *Review of Economic Studies*, 58(5), October 971-91

Kyriazidou, E. (1997)“Estimation of a Panel Data Sample Selection Model”, *Econometrica, 65(6)* (November)1335-1364

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

Matzkin, R.L. (1999) “Nonparametric Estimation of Nonadditive Random Functions,” mimeo, Northwestern University.

Matzkin, R.L. and W. Newey (1993) “Kernel Estimation of Nonparametric Limited Dependent Variable Models,”, mimeo.

Mundlak, Y. (1978)”On the Pooling of Time Series and Cross Section Data”, *Econometrica* 46 pp. 69-85

W. Newey (1994) “Kernel Estimation of Partial Means and a General Variance Estimator”, *Econometric Theory*, 10, 233-253.

Pagan, A. and A. Ullah (1999), *Nonparametrics Econometrics*: Cambridge University Press: Cambridge.

Powell, J.L, (1994) “Estimation of Semiparametric Models”, in *Handbook of Econometrics*, Volume IV, Edited by R.F. Engle and D.L. McFadden, Elsevier Science B.V.

Rosenzweig, M. R. and K. I. Wolpin, (1994) “Parental and public transfers to young women and their children”, *American Economic Review*, v84 n5 p1195-2018.

Rosenzweig, M. R. and K. I. Wolpin, (1994) “Inequality among young adult siblings, public assistance programs, and intergenerational living arrangements”, *Journal of Human Resources* Fall v29 n4 pp110-25.

Rivers, D. and Q. H. Vuong, (1988), ”Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models”, *Journal of Econometrics* 39, pp. 347-366.

van der Vaart, A. W. and J. A. Wellner (1996) *Weak Convergence and Empirical Processes, with Applications to Statistics*, Springer Series in Statistics.

Table 1: Monte Carlo Simulations of the Estimator:1 The Continuous Dependent Variable Case[1]

$$
\begin{aligned}
Y_{ik} &= b_0 + b_1 X_{ik} + \gamma X_{ik}\eta_i + \theta_\eta \eta_i + \varepsilon_i + u_{ik} \\
X_{ik} &= x_i + \tilde{X}_{ik}; \qquad \tilde{X}_{ik} \sim N(0,1) \\
\eta_i &= \theta_{\eta x} X_i + \tilde{\eta}_{ik}; \qquad \tilde{\eta}_{ik} \sim N(0,1) \\
\varepsilon_i &= \theta_{\varepsilon x} x_i + \widetilde{\varepsilon}_i \\
u_{ik} &\sim N(0,1)
\end{aligned}
$$

Sample Size : $n = 1500$; Group Size: K=2; Replications=750

Case 1: *Parameter Values*

$Y_{ik} = 2X_{ik} + 0X_{ik}\eta_i + 1\eta_i + 0\varepsilon_i + u_{ik}; \quad \theta_{\eta x} = 0$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	2.00	2.00	2.00	2.00	2.00
Estimator of $\beta(x)$ using:					
Kernel/Kernel[1] / *(SE,MSE)*	1.83 / (.17,.27)	1.86 / (.12,10)	1.87 / (.11,.096)	1.87 / (.132,.096)	1.85 / (.19,.27)
Polynomial/Kernel[2] / *(SE,MSE)*	2.01 / (.31,.096)	2.00 / (.13,.016)	1.99 / (.041,1.6$e-3$)	1.995 / (.13,.017)	1.989 / (.032,.10)
OLS / *(SE,MSE)*	1.999 / (.019,1.6$e-4$)	1.999 / (.019,1.6$e-4$)	1.999 / (.019,1.6$e-4$)	1.999 / (.019,1.6$e-4$)	1.999 / (.019,1.6$e-4$)
OLS/Fixed Effects / *(SE,MSE)*	1.999 / (.026,6.8$e-4$)	1.999 / (.026,6.8$e-4$)	1.999 / (.026,6.8$e-4$)	1.999 / (.026,6.8$e-4$)	1.999 / (.026,6.8$e-4$)

Case 2: ParameterValues

$Y_{ik} = 2X_{ik} + 1X_{ik}\eta_i + 1\eta_i + 0\varepsilon_i + u_{ik}; \quad \theta_{\eta x} = 1$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	1	1.5	2.00	2.50	3.00
Estimator of $\beta(x)$ using:					
Kernel/Kernel / *(SE,MSE)*	.852 / (.16,.048)	1.41 / (.11,.020)	1.899 / (.12,.025)	2.44 / (.12,.017)	2.90 / (.25,.073)
Polynomial/Kernel / *(SE,MSE)*	1.03 / (.14,.02)	1.54 / (.106,.012)	2.04 / (.09,9.7$e-3$)	2.55 / (.13,.02)	3.06 / (.2,.043)
OLS / *(SE,MSE)*	2.00 / (.067,1)	2.00 / (.067,.25)	2.00 / (.067,4.5$e-3$)	2.00 / (.067,.25)	2.00 / (.067,1)
OLS/Fixed Effects / *(SE,MSE)*	2.5 / (.073,2.25)	2.5 / (.073,1)	2.5 / (.073,.25)	2.5 / (.073,5.39$e-3$)	2.5 / (.073,.25)

Case 3:	Parameter Values				
	$Y_{ik} = 2X_{ik} + 1X_{ik}\eta_i + 1\eta_i + 1\varepsilon_i + u_{ik}; \quad \theta_{\eta x} = 1$				
X Value:	-2	-1	0	1	2
True Value of $B(x)$:	1	1.5	2.00	2.5	3.00
Estimator of $B(x)$ using:					
Kernel/Kernel (*SE,MSE*)	0.91 (.21,.052)	1.43 (.13,.022)	1.95 (.15,.025)	2.46 (.13,.018)	2.96 (.29,.085)
Polynomial/Kernel (*SE,MSE*)	1.01 (.16,.026)	1.54 (.11,.014)	2.05 (.097,.012)	2.55 (.14,.022)	3.05 (.22,.051)
OLS (*SE,MSE*)	3.0 (.068,4)	3.0 (.068,2.3)	3.0 (.068,1)	3.0 (.068,.25)	3.0 (.068,4.6$e-$3)
OLS/Fixed Effects (*SE,MSE*)	2.0 (.076,1)	2.0 (.076,.25)	2.0 (.076,5.7$e-$3)	2.0 (.076,.25)	2.0 (.076,1)

1. In the Kernel/Kernel we use kernel regression to estimate both $E_x(y_{ik}|x_{ik}, z_i)$ and $h(z_i|x_{ik})$. For both functions we used a gaussian kernel. We chose window widths separately for the kernel estimators of $E_x(y_{ik}|x_{ik}, z_i)$ and for $h(z_i|x_{ik})$ using Silverman's plug in method (See Pagan and Ullah (1999) page 26 for details) for a particular design, and did not re-optimize for the various cases including some that we do not report. We used .35 for $E_x(y_{ik}|x_{ik}, z_i)$ and .15 for $h(z_i|x_{ik})$. In preliminary work we also experimented with window width values that were obtained using cross validation for regression and conditional densities and it did not make much difference. In evaluating $E_x(y_{ik}|x_{ik}, z_i)$ at values of z for a given x we added the constant .0001 to the denominator to avoid extreme values.We evaluated the integral in (2.7) using Simpson's rule after trimming away observations that lie within 1 window width of the boundary of the support of the data because of high bias near the boundary. In the Polynomial/Kernel version we use OLS estimation of a fourth order polynomial (x_{ik}, z_i) with interactions up to the second order as the estimator for $E(y_{ik}|x_{ik}, z_i)$. We use the kernel estimator for $h(z_i|x_{ik})$ with the window width set to .10 . We also report of $\beta(x)$ based on the application of OLS both with and without group specific intercepts or "fixed effects". We set n to 1,500 and the group size K to 2. The Monte Carlo results are based on 750 replications.

Table 2. **Monte Carlo Simulations of the Estimator 1: The Binary Choice Case**[1]

$$
\begin{aligned}
Y_{ik} &= 1(Y_{ik}^* \geq 0);\ Y_{ik}^* = b_0 + b_1 X_{ik} + \gamma X_{ik}\eta_i + \theta_\eta \eta_i + u_{ik} \\
X_{ik} &= X_i + \tilde{X}_{ik}; \qquad \tilde{X}_{ik} \sim N(0,1) \qquad X_i \sim N(0,1) \\
\eta_i &= \theta_{\eta x} X_i + \tilde{\eta}_i; \qquad \tilde{\eta}_i \sim N(0,1) \\
x_i &\sim N(0,\sigma_x^2);\ \tilde{x}_{ik} \sim N(0,\sigma_{\tilde{x}}^2);\ \tilde{\eta}_i \sim N(0,\sigma_{\tilde{\eta}}^2); \tilde{\varepsilon}_i \sim N(0,1);\ u_{ik} \sim N(0,\sigma_u^2)
\end{aligned}
$$

Sample Size : n=1500; Group Size: K=2; Replications=750

Case 1: ParameterValues

$Y_{ik}^* = 2X_{ik} + 0X_{ik}\eta_i + 1\eta_i + u_{ik};\quad \sigma_x^2 = \sigma_{\tilde{x}}^2 = \sigma_{\tilde{\eta}}^2 = \sigma_u^2 = 1,\ \theta_{\eta x} = 1.5$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	.0036	.135	.451	0.135	.0036
Estimator of $\beta(x)$ using:					
Estimator 1 $\frac{Kernel/Kernel}{(SE,MSE)}$	$\frac{.0070}{.0026,6.9e-6}$	$\frac{.14}{.016,2.7e-4}$	$\frac{.41}{.024,3.1e-3}$	$\frac{.14}{.016,2.5e-3}$	$\frac{.0068}{.0026,6.8e-7}$
Estimator 1 $\frac{Probit/Kernel}{(SE,MSE)}$	$\frac{.0040}{.0018,1e-5}$	$\frac{.16}{.025,1.1e-3}$	$\frac{.42}{.037,2.9e-3}$	$\frac{.16}{.024,1.1e-3}$	$\frac{.0045}{.0018,7.7e-3}$
$\frac{Probit}{(SE,MSE)}$	$\frac{.0080}{.004,1.8e-5}$	$\frac{.21}{.022,5.6e-3}$	$\frac{.56}{.031,.01}$	$\frac{.21}{.022,5.6e-3}$	$\frac{.0080}{.0040,1.8e-5}$

Case 2: ParameterValues

$Y_{ik}^* = .75X_{ik} + 0X_{ik}\eta_i + 1\eta_i + 0\varepsilon_i + u_{ik};\quad \sigma_x^2 = \sigma_{\tilde{x}}^2 = \sigma_u^2 = 1.5,\ \sigma_{\tilde{\eta}}^2 = .5,\ \theta_{\eta x} = -1.5$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	.155	.155	.155	.155	.155
Estimator of $\beta(x)$ using:					
Estimator 1. $\frac{Kernel/Kernel}{(SE,MSE)}$	$\frac{.123}{.055,.0040}$	$\frac{.121}{.040,.0027}$	$\frac{121}{.039,.0027}$	$\frac{.121}{.042,.0029}$	$\frac{.123}{.052,.0037}$
Estimator 1. $\frac{Probit/Kernel}{(SE,MSE)}$	$\frac{.155}{.016,.000}$	$\frac{.155}{.012,.000}$	$\frac{.155}{.012,.000}$	$\frac{.155}{.012,.000}$	$\frac{.155}{.016,.000}$
$\frac{Probit}{(SE,MSE)}$	$\frac{.000}{.011,.024}$	$\frac{.000}{.009,.024}$	$\frac{.000}{009,.024}$	$\frac{.000}{.009,.024}$	$\frac{.000}{.011,.024}$

Case 3: ParameterValues

$Y_{ik}^* = 2X_{ik} + 1X_{ik}\eta_i + 1\eta_i + u_{ik};\quad \sigma_x^2 = \sigma_{\tilde{x}}^2 = 1, \sigma_u^2 = 1.5 = \sigma_{\tilde{\eta}}^2 = 1.5,\ \theta_{\eta x} = 1.5$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	-.100	.107	.393	.080	.016
Estimator of $\beta(x)$ using:					
Estimator 1. $\frac{Kernel/Kernel}{(SE,MSE)}$	$\frac{-.086}{(.037,1.7e-3)}$	$\frac{.149}{(.023,2.3e-3)}$	$\frac{.356}{(.037,2.70e-3)}$	$\frac{.111}{(.030,1.63e-3)}$	$\frac{.025}{(.025,7.0e-4)}$
Estimator 1. $\frac{Probit/Kernel}{(SE,MSE)}$	$\frac{.012}{(.012,.013)}$	$\frac{.142}{(.014,1.4e-3)}$	$\frac{.331}{(.024,4.4e-3)}$	$\frac{.149}{(.019,4.5e-3)}$	$\frac{.011}{(.010,1.2e-4)}$
$\frac{Probit}{(SE,MSE)}$	$\frac{-.009}{(.012,8.5e-3)}$	$\frac{.153}{(.009,2.2e-3)}$	$\frac{.438}{(.021,2.5e-3)}$	$\frac{.279}{(.012,.038)}$	$\frac{.043}{(.007,7.7e-4)}$

[1] See Table 1. In the case of the Kernel/Kernel Estimator we use a gaussian Kernel with a window width of .32 for $E(y_{ik}|x_{ik}, z_i)$ and .18 for $h(z_i|x_{ik})$. The Probit/Kernel Estimator uses a MLE-probit with an index consisting of a constant and a third order polynomial in x_{ik} and z_i with a full set of interactions up to the second order to estimate $E(y_{ik}|x_{ik}, z_i)$ and the kernel estimator with a window width of .10 to estimate $h(z_i|x_{ik})$. The Probit estimator is MLE-probit with an index consisting of a constant and third order polynomial in x_{it}.

Table 3: Monte Carlo Simulations of Estimator 1
The Binary Choice Case, Chi-Square—Normal Regressors[1]

$$
\begin{aligned}
Y_{ik} &= 1(Y_{ik}^* \geq 0);\; Y_{ik}^* = b_0 + b_1 X_{ik} + \gamma X_{ik}\eta_i + \theta_\eta \eta_i + \theta_\varepsilon \varepsilon_i + u_{ik}) \\
X_{ik} &= x_i + \tilde{X}_{ik}; \qquad x_i \sim N(0,1);\; \tilde{X}_{ik} \sim (\text{Chi Square}_3 - 3)/(3^{.5}) \\
\eta_i &= \theta_{\eta x} X_i + \tilde{\eta}_{ik}; \qquad \tilde{\eta}_{ik} \sim N(0,1) \\
u_{ik} &\sim N(0,1),\;\; \varepsilon_i \tilde{\ } N(0,1) \\
\text{Sample Size} &: \text{n=1500; Group Size: K=2; Replications=750}
\end{aligned}
$$

Case 1: ParameterValues

$Y_{ik}^* = 2X_{ik} + 0X_{ik}\eta_i + 1\eta_i + 0\varepsilon_i + u_{ik}; \quad \theta_{\eta x} = 1$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	0	.10	.61	.209	0
Estimator of $\beta(x)$ using:					
$\frac{Kernel/Kernel}{(SE,MSE)}$	$\frac{.0018}{.0012,4.68e-6}$	$\frac{.11}{.014,2.96e-4}$	$\frac{.53}{.041,8.1e-3}$	$\frac{.18}{.035,2.07e-3}$	$\frac{.0044}{.0066,6.3e-5}$

Case 2: ParameterValues

$Y_{ik}^* = 2X_{ik} + 1X_{ik}\eta_i + 0\eta_i + 0\varepsilon_i + u_{ik}; \quad \theta_{\eta x} = 1$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	.018	.089	.68	.12	0
Estimator of $\beta(x)$ using:					
$\frac{Kernel/Kernel}{(SE,MSE)}$	$\frac{-.023}{.022,2.17e-3}$	$\frac{.07}{.018,6.85e-4}$	$\frac{.63}{.041,4.18e-3}$	$\frac{.102}{.025,9.40e-4}$	$\frac{.00083}{.018,3.25e-4}$

Case 3: ParameterValues

$Y_{ik}^* = 2X_{ik} + 1X_{ik}\eta_i + 1\eta_i + 1\varepsilon_i + u_{ik}; \quad \theta_{\eta x} = 1, \theta_{\varepsilon x} = 0$

X Value:	-2	-1	0	1	2
True Value of $\beta(x)$:	0	.22	.421	.03	0
Estimator of $\beta(x)$ using:					
$\frac{Kernel/Kernel}{(SE,MSE)}$	$\frac{-.00048}{.012,1.1e-4}$	$\frac{.202}{.020,7.2e-4}$	$\frac{.42}{.037,1.4e-3}$	$\frac{.067}{.031,2.3e-3}$	$\frac{-7.34e-5}{.035,1.2e-3}$

[1] $\tilde{X}_{ik}$ is distributed Chi-square with 3 degrees of freedom, normalized to have mean 0 and variance 1. See Table 1 and 2 for details of the computation of the Kernel-Kernel estimator The window widths are the same as those used in Table 2.