Identification and Estimation of Semiparametric Two Step Models*

Juan Carlos Escanciano† Indiana University

David Jacho-Chávez‡ Emory University

Arthur Lewbel§ Boston College

First Draft: May 2010
This Draft: January 2015

Abstract

Let $H_0(X)$ be a function that can be nonparametrically estimated. Suppose $E[Y|X] = F_0[X^\top \beta_0, H_0(X)]$. Many models fit this framework, including latent index models with an endogenous regressor, and nonlinear models with sample selection. We show that the vector β_0 and unknown function F_0 are generally point identified without exclusion restrictions or instruments, in contrast to the usual assumption that identification without instruments requires fully specified functional forms. We propose an estimator with asymptotic properties allowing for data dependent bandwidths and random trimming. A Monte Carlo experiment and an empirical application to migration decisions are also included.

Keywords: Identification by functional form; double index models; two step estimators; semiparametric regression; control function estimators; sample selection models; empirical process theory; limited dependent variables; migration.

JEL classification: C13; C14; C21; D24

*We would like to thank the Co-Editor, Petra Todd, as well as two anonymous referees for their helpful comments and suggestions. We especially want to thank Yingying Dong, for providing data and other assistance on our empirical application, and Jeffrey Racine for his advice on how to use the `np` package in our reported estimation results. We would also like to thank Hidehiko Ichimura, Simon Lee, Philip Shaw, Jörg Stoye, Ingrid Van Keilegom, Adonis Yatchew, and participants of many conferences and seminars at various institutions for many helpful comments. We acknowledge the usage of the Big Red high performance cluster at Indiana University where part of the computations were performed. All errors are our own.

†Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/ Research funded by the Spanish Plan Nacional de I+D+I, reference number ECO2012-33053.

‡Department of Economics, Emory University, Rich Building 3rd Floor, 1602 Fishburne Dr., Atlanta, GA 30322-2240, USA. E-mail: djachocha@emory.edu. Web Page: https://sites.google.com/site/djachocha/.

§Corresponding Author: Department of Economics, Boston College, 140 Commonwealth Avenue, Chesnut Hill, MA 02467, USA. E-mail: lewbel@bc.edu. Web Page: http://www2.bc.edu/~lewbel/.

1 Introduction

We provide new identification and estimation results for two step estimators with a nonparametric first step. Given an observable vector X, suppose $H_0(X)$ is some identified function that can be nonparametrically estimated, e.g. $H_0(X)$ could be a conditional mean, quantile, distribution, or density function. This paper considers identification and estimation of the function F_0 and the vector β_0 where

$$E[Y|X] = M(X) = F_0[X^\top \beta_0, H_0(X)], \tag{1.1}$$

for some observed outcome Y. Equation (1.1) is a double index model, with one linear index $X^\top \beta_0$ and one general index function $H_0(X)$. As described below, many common econometric models involving either selection or an endogenous regressor can be written in the form of equation (1.1).

Identification of models like these is generally obtained by exclusion or instrument assumptions, such as knowing that some elements of β_0 are zero, thereby making the corresponding elements of X be valid instruments for H_0. This implies having some variable that affects the selection or treatment model or the endogenous regressor, but does not affect the outcome. The gold standard for such an identifying variable would be random assignment, satisfying exclusion conditions by construction.

However, randomization or other sources of exclusion restrictions or instruments, while preferable, are not always available. In such cases point identification is generally known to be obtainable only by parametric functional form restrictions. What we show here is that identification can in fact be obtained, without exclusion restrictions or instruments, under far weaker conditions than fully parameterizing the model. Specifically, we show that identification is generally possible in equation (1.1) where nothing more than linearity of $X^\top \beta_0$ is parameterized. Identification is obtained despite having the parameters β_0 be unknown and the functions H_0 and F_0 be unknown.

We illustrate our identification results by showing how they apply to a couple of common econometric models. One is a semiparametric double hurdle model, in which a latent outcome is observed only if it and a separate binary variable are both positive. This is equivalent to combining both Tobit estimation and a Heckman selection model, except that in our case the error distributions and the propensity score function for the selection are all unknown and dealt with nonparametrically. The second example illustration is a control function estimator of a binary choice model with an endogenous regressor, as in Blundell and Powell (2004). Our results show that both these classes of models are identified without instruments or exclusion assumptions.

While our primary contribution is identification, we also provide some new estimation results. Given identification of equation (1.1), either by our theorems or by standard exclusion restrictions and instruments, we propose an estimator for F_0 and β_0 based on minimizing a weighted least squares criterion similar to Ichimura (1993) and Ichimura and Lee (2010). However, in the latter the first stage plug in estimate is parametric, satisfying an index restriction that we do not impose on our estimate, i.e. $\widehat{H}$, which can be nonparametric.

To establish the asymptotic normality of our proposed estimator, we employ a new uniform-in-bandwidth expansion for sample means of weighted semiparametric residuals recently obtained in Escanciano, Jacho-Chávez and Lewbel (2014). We obtain the same limiting distributions that would

be found using more well-known approaches such as Newey and McFadden (1994, p. 2197), Chen, Linton, and van Keilegom (2003) and Ichimura and Lee (2010), but under weaker conditions. These conditions simultaneously permit data-driven bandwidths, random trimming and estimated observation weights. Some of these conditions could in principle be shown to satisfy the higher level assumptions provided in Chen, Linton, and van Keilegom (2003), but establishing that those assumptions hold is difficult (see, e.g., Rothe, 2009). Other assumptions, such as the random trimming, does not fit the standard framework at all.

We apply these identification and estimation results in an empirical application. We estimate a migration decision model, which is a control function specification of a binary choice model with an endogenous regressor. Our new identification result is required because exclusions are not economically plausible in this application. Our new limiting distribution theory for this application accounts for data dependent bandwidth choice and random trimming.

Overall, our results suggest that identification by functional form need not be as fragile or unreliable as the name suggests, since at least in our large class of models identification without instruments does not require parameterizing the distribution of errors or the functions F_0 or H_0. Even when instruments are available, having identification not depend on these means that their validity as instruments can be tested, by comparing equality of estimates with and without imposing associated exclusion assumptions.

The rest of the paper is organized as follows: Section 2 provides some background information, including examples of models that fit our framework. Section 3 gives our general identification theorems, while Section 4 describes two examples that fit our framework. Section 5 describes the proposed estimator and establishes its limiting distribution. A Monte Carlo experiment and an empirical application to migration data is presented in Section 6. Section 7 concludes. The main proofs are gathered into the Appendix A, while the conditions for the provided Asymptotic Theory are listed and discussed in Appendix B.

2 Background

Models of the form $M(X) = F_0\left[r_0(X), H_0\left(X\right)\right]$ where $M(X)$ can be estimated are called double index models. Examples of estimators of double or multiple index models include Ichimura and Lee (1991), Pinkse (2001), and Lewbel and Linton (2007). Additional models where both $r_0(X)$ and $H_0\left(X\right)$ are linear indices include Klein, Shen and Vella (2014), the Sliced Inverse Regression models of Li (1991) and artificial neural networks. This paper focuses on the case where just one of the two indices is linear, and hence parameterized as $X^\top\beta_0$. This is an appropriate assumption in contexts where $X^\top\beta_0$ arises as part of a structural model, while $H_0(X)$ is a nuisance function associated with selection or endogeneity of a regressor.

In the semiparametric literature, when one or more of the functions H_0, r_0, F_0 are not parameterized, identification is generally obtained by exclusion restrictions, that is, some element of X is assumed to drop out of either $r_0(X)$ or $H_0(X)$. In models where all of the unknown functions are parameterized, exclusion restrictions are generally not required for identification, and we may instead obtain identification by functional form. For example, Heckman's (1979) sample selection model does

not require an exclusion restriction for identification, since identification is obtained by parameterization of the joint error distribution in the selection and outcome equations. Empirically, identification by exclusion restrictions is generally considered more reliable than identification based on functional form, but it is often the case that exclusions are hard to find or to plausibly impose. Identification by functional form also provides a way to test exclusion restrictions, since it nests models with exclusions.

As noted in the Introduction, this paper shows that 'identification by functional form' extends to the semiparametric model (1.1). In particular, it is shown that β_0 and F_0 can be identified without exclusion restrictions under some relatively mild regularity conditions (essentially, nonlinearity in H_0 and some inequalities suffice). So for example in Heckman's model, identification by functional form does not actually require a parameterized functional form for the error distributions or for the selection index $H_0(X)$.

One large class of models that fits in this paper's framework are endogenous regressor models. Suppose $Y = L(X_1^\top \alpha_0 + X^e \delta_0, e)$ for some possibly unknown function L, where X^e is an endogenous regressor with $X^e = g_0(X_1) + u$, e and u are unobserved correlated error terms, and g_0 represents a generally unknown conditional mean function. Let $E[u|X_1] = 0$ and assume the endogeneity takes the 'control function' form of $e = h_0(u, v)$ for some function h_0, where v is an unobserved error that is independent of X^e and X_1. Another way to describe the endogeneity is to say $e|X_1, u \sim e|u$. Define $X^\top \beta_0 := X_1^\top \alpha_0 + X^e \delta_0$, and $H_0(X) := X^e - g_0(X_1)$. Then equation (1.1) holds with F_0 defined by $E[Y|X] = E[L(X^\top \beta_0, h_0(H_0(X), v))|X] =: F_0\left[X^\top \beta_0, H_0(X)\right]$.

An example of this type of endogenous regressor model is Rivers and Vuong (1988). More generally, for Y binary this is Blundell and Powell's (2004) semiparametric binary choice model with an endogenous regressor, so this paper's identification results show that Blundell and Powell's control function model is generally identified (and so could be estimated using their estimator or Rothe's 2009 estimator), without the exclusion restrictions they impose for identification. Our empirical application to estimation of a migration equation is an example of this model.

Another large class of models that fit this paper's framework are limited dependent variable models with selection. Suppose $Y^* = L(X^\top \beta_0 + e)$ for some function L, e.g., L could be the identity function so $Y^* = X^\top \beta_0 + e$ is a linear model, or Y^* could be a binomial response with $Y^* = \mathbb{I}(X^\top \beta_0 + e > 0)$ for $\mathbb{I}$ being the indicator function that equals one if its argument is true and zero otherwise, or L could be a censored regression such as $Y^* = \max(0, X^\top \beta_0 + e)$. Suppose in addition to this possibly limited dependent variable we also have non-random selection, so we do not observe Y^* but instead observe $(Y, D, X^\top)$ where $Y = Y^* D = L(X^\top \beta_0 + e)D$, implying that Y^* is only observed when $D = 1$. Suppose $D = \mathbb{I}[s_0(X) + u > 0]$ where the conditional distribution u given e is continuous. Non-random selection arises because the unobserved errors e and u, though independent of X, are correlated with each other. It then follows that $H_0(X) = E[D|X] = F_u[s_0(X)]$ where F_u is marginal distribution function of $-u$, and equation (1.1) holds with F_0 defined by $E[Y|X] = E[L\left(X^\top \beta_0 + e\right)\mathbb{I}(F_u^{-1}[H_0(X)] + u > 0)|X] =: F_0\left[X^\top \beta_0, H_0(X)\right]$. This includes a wide class of selection models, including standard Heckman-type selection models, extensions of Tobit models like double hurdle models, and censored binary choice models, among others.

3 Identification

Here we provide sufficient conditions for identifying the function F_0 and the parameter vector β_0 in the semiparametric double index model $M(X) = F_0\left[X^\top\beta_0, H_0(X)\right]$, where $M(X)$ and $H_0(X)$ are assumed to be identified. The estimators and example applications we discuss later are based on defining $M(X)$ and $H_0(X)$ in terms of conditional expectations, but this is not necessary for our identification Theorems. For example, identification would hold in the same way regardless of whether $M(X)$ was a nonparametrically identified density, distribution, or quantile function rather than a conditional mean, and similarly for $H_0(X)$.

To obtain identification under general conditions, we provide two different sets of identifying assumptions, each of which can be applied to different sets of regressors in the same model. We therefore let $X^\top\beta_0 = V^\top\alpha_0 + Z^\top\delta_0$ and write the model as

$$M(V,Z) = F_0[V^\top\alpha_0 + Z^\top\delta_0, H_0(V,Z)]$$

Here V is a vector of one or more continuous regressors, while Z is a vector that can include covariates which are discrete, continuous, continuous with mass points, etc. Theorem 3.1 below identifies α_0 using differentiability and inequality constraints, while Theorem 3.2 below identifies δ_0 exploiting support and invertibility restrictions instead. Z could be empty so only Theorem 3.1 would be needed, or V could just have one element with a coefficient normalized to equal one, in which case only Theorem 3.2 would be needed, or both Theorems can be applied to identify models where different sets of regressors satisfy different regularity assumptions.

3.1 Continuous Regressors

Assumption 1 *Assume V is a $K \times 1$-vector and assume there exists a function F_0 and a vector α_0 such that $m(V) = F_0\left(V^\top\alpha_0, H_0(V)\right)$. Assume functions $m(V)$ and $H_0(V)$ are identified. Let $\alpha_{0;k}$ and V_k denote the kth element of α_0 and V respectively, for $k = 1,\ldots,K$. Assume $\alpha_{0;1} = 1$.*

Assumption 1 applies to the general double index model $M(V,Z) = F_0\left[V^\top\alpha_0 + Z^\top\delta_0, H_0(V,Z)\right]$ defining $m(V) := M(V,0)$ and $H(V) := H_0(V,0)$, and provided the $J \times 1$ vector of zeroes is in the support of Z. If Z is empty then Theorem 3.1 based on Assumptions 1 and 2 below will identify the entire model, otherwise it will identify just the α_0 coefficients, and F_0 only on the supports of $V^\top\alpha_0$ and $H_0(V,0)$. If $K = 1$, then α_0 just equals $\alpha_{0;1} = 1$ and so is known by the scale normalization assumption in that case.

The scaling of α_0 is arbitrary, since changes in scaling can be freely absorbed into F_0. Assumption 1 imposes the convenient scale normalization that V_1, the first element of V, has a coefficient of $\alpha_{0;1} = 1$. This is a free normalization if one knows that this regressor has a positive effect on F_0 through the first index. This is also the most natural normalization in some contexts, e.g., if in a binary choice model Y is a purchase decision and $-V_1$ is the price, then with the normalization $\alpha_{0;1} = 1$ the remainder of the index $V^\top\alpha_0$, that is, $\sum_{k=2}^K V_k\alpha_{0;k}$, equals (up to location) the willingness to pay for the product (see e.g. Lewbel, Linton, and McFadden, 2011).

Assumption 2 *Assume H and F_0 are differentiable and define $F_{0;1}(r,H) := \partial F_0(r,H)/\partial r$, the partial derivative of F_0 with respect to its first element. Define $H_{0;k}(V) := \partial H_0(V)/\partial V_k$ and $\partial m_k(V) := \partial m(V)/\partial V_k$ for $k = 1,...,K$. Assume there exists two vectors v and $\widetilde{v}$ on the support V such that the derivatives $H_{0;k}(V)$ and $m_k(V)$ are identified at $V = v$ and $V = \widetilde{v}$ for $k = 1,\ldots,K$, and assume there exist two elements k and j of the set $\{1,...,K\}$ such that the following inequalities hold*

$$
\begin{aligned}
&F_{0;1}(\widetilde{v}^\top \alpha_0, H_0(\widetilde{v})) \neq 0 \text{ and } F_{0;1}(v^\top \alpha_0, H_0(v)) \neq 0,\\
&H_{0;j}(\widetilde{v}) \neq H_{0;1}(\widetilde{v})\,\alpha_{0;j} \text{ and } H_{0;k}(\widetilde{v}) \neq H_{0;1}(\widetilde{v})\,\alpha_{0;k},\\
&H_{0;\ell}(v) \neq H_{0;1}(v)\,\alpha_{0;\ell} \text{ for } \ell = 2,\ldots,K,\\
H_{0;j}(v)\,H_{0;k}(\widetilde{v}) - H_{0;j}(\widetilde{v})\,H_{0;k}(v) &\neq [H_{0;1}(v)\,H_{0;k}(\widetilde{v}) - H_{0;1}(\widetilde{v})\,H_{0;k}(v)]\,\alpha_{0;j}\\
&\quad - [H_{0;1}(v)\,H_{0;j}(\widetilde{v}) - H_{0;1}(\widetilde{v})\,H_{0;j}(v)]\,\alpha_{0;k}.
\end{aligned}
$$

The inequalities in Assumption 2 essentially require F_0 to depend on $V^\top\alpha_0$, and require some variation in $H_0(V)$ that distinguishes it from $V^\top\alpha_0$. These inequalities will not hold if $H_0(V)$ equals a transformation of a single index in V for example (see Chamberlain, 1986), but otherwise it is very difficult to construct examples that violate the inequalities of Assumption 2. This assumption can be interpreted generically as a rank condition, which is a common feature of many identification theorems. See, e.g., Komunjer (2012), Heckman, Matzkin, and Nesheim (2010), Lewbel (2007), Matzkin (2007), and references therein.

The identification of derivatives of m and H_0 at the two points v and $\widetilde{v}$ generally requires that V be continuously distributed at those points, so Theorem 3.1 below cannot be applied to discrete regressors. Theorem 3.2 later will provide identification for other regressor distributions, including discrete regressors.

Theorem 3.1 *If Assumptions 1 and 2 hold, then the vector α_0 and the function F_0 (on the supports of $V^\top\alpha_0$ and $H_0(V,0)$) are identified.*

Theorem 3.1 obtains identification without an exclusion assumption, that is, all of the covariates V can appear in both the index $V^\top\alpha_0$ and in the function $H_0(V)$. Identification can also be obtained from Theorem 3.1 using exclusions restrictions to satisfy Assumption 2. However, with an exclusion restriction, identification can be obtained more simply as follows: Let Assumption 1 hold with $\alpha_{0;K} = 0$ and assume $H_0(V)$ varies with V_K (so V_K is in $H_0(V)$ but not in the linear index $V^\top\alpha_0$ and hence V_K is the excluded regressor). Then if F_0 is differentiable in its first element it follows from equation (A-1) that α_0 is identified by

$$
\alpha_{0;k} = \frac{\partial E[m(V)\,|V_1,...,V_{K-1},H_0(V)]}{\partial V_k} \Big/ \frac{\partial E[m(V)\,|V_1,...,V_{K-1},H_0(V)]}{\partial V_1}, \text{ for } k = 2,\ldots,K-1,
$$

evaluated at any value of V that makes the derivative in the denominator of the above expression nonzero.

3.2 Discrete Regressors

Theorem 3.1 does not identify the coefficients of discrete regressors, but it can identify the coefficients of the continuous regressors when both continuous and discrete regressors are present. Given both continuous and discrete regressors, Theorem 3.2 below can be combined with Theorem 3.1 to identify the coefficients of the remaining discrete regressors, or of other regressors that satisfy the alternative regularity conditions provided in Assumptions 3 and 4. If all regressors satisfy these alternative regularity conditions, then Theorem 3.2 alone can be used for identification with both types of regressors. Assumption 4 imposes support restrictions and local invertibility of F_0, instead of the differentiability and inequality constraints in Assumption 2.

Assumption 3 *Let $M(V,Z) = F_0[V^\top\alpha_0 + Z^\top\delta_0, H_0(V,Z)]$. Assume $H_0(V,Z)$ and $M(V,Z)$ are identified. Assume Z is a $J\times 1$ vector and that α_0 is identified.*

If V is a scalar, then α_0 is identified by the free (up to sign) normalization $\alpha_{0;1} = 1$ as before. Alternatively, if V is a vector then α_0 is identified by Theorem 3.1 as long as Assumptions 1 and 2 hold with $m(V) = M(V,0)$ and $H_0(V) = H_0(V,0)$. Having V be a vector instead of a scalar makes the support requirements in Assumption 4 below less restrictive. These support conditions are particularly mild when Z consists only of discrete regressors, which would then require V to contain all the continuous covariates in the model.

Assumption 4 *Assume the $J\times 1$ vector of zeroes is in the support of Z. Let $\widetilde{z}_j$ denote the $J\times 1$ vector that has element j equal to z_j and all other elements equal to zero. Assume for some $z_j \neq 0$ in the support of Z_j, there exists $v(z_j)$ in the support of V such that $v(z_j)^\top\alpha_0 + z_j\gamma_j$ is in the support of $V^\top\alpha_0$ and $H_0(v(z_j), z_j)$ is in the support of $H_0(V,0)$. Assume $F_0[r,\widetilde{H}]$ is invertible on its first element at the point $r = v(z_j)^\top\alpha_0 + z_j\gamma_j$, $\widetilde{H} = H_0(v(z_j), z_j)$.*

A sufficient condition for Assumption 4 to hold is if $V^\top\alpha_0$ and $H_0(V,0)$ have support on the entire real line and if F_0 is strictly monotonic in its first element. Alternatively, if Z is discrete, then only a limited range of values of $V^\top\alpha_0$ and $H_0(V,0)$ are required, e.g., if there is a v such that $v^\top\alpha_0$ can take on the value $-z_j\gamma_j$, then letting $v(z_j)$ equal that v makes $v(z_j)^\top\alpha_0 + z_j\gamma_j$ lie in the support of $V^\top\alpha_0$, and a similar analysis applies to H_0.

Theorem 3.2 *If Assumptions 3 and 4 hold for $j = 1,...,J$ then α_0, δ_0, and F_0 at all points on the support of $V^\top\alpha_0 + Z^\top\delta_0$, and $H_0(V,Z)$ are identified.*

4 Examples of Model Identification

We now illustrate the identification results of Theorems 3.1 and 3.2 by applying them to two examples, a double hurdle model and a binary choice control function model with an endogenous regressor.

4.1 A Semiparametric Double Hurdle Model

Suppose a latent binary variable Y^* satisfies the standard fixed (at zero) censored regression model $Y^* = \left(X^\top \beta_0 - e\right) \mathbb{I}\left(X^\top \beta_0 - e \geq 0\right) = \max\left(0, X^\top \beta_0 - e\right)$ with e independent of X and the distribution function of e, F_e, may be unknown. Suppose we only observe Y^* for some subset of the population, indexed by a binary variable D, i.e. we only observe $Y = Y^* D$. This is a sample selection model with a censored regression outcome. For example, Y^* could indicate the quantity of a good an individual might want to purchase, D indicates whether the good is available for purchase where the individual lives, and Y indicates the quantity of the good the individual purchases, which is nonzero only when both $X^\top \beta_0 - e > 0$ and $D = 1$.

We apply our previous results to identify this censored regression with selection model without exclusion assumptions. We also do not assume selection on observables, so that D and Y^* remain correlated even after conditioning on observables X, as in various Heckman-type selection models. The practitioner is assumed to know relatively little about selection D other than that it is binary, so we assume D is given by a nonparametric threshold crossing model $D = \mathbb{I}\left[H_0\left(X\right) - u \geq 0\right]$ where $u \perp X$ and both the function $H_0(X)$ and the distribution of u are unknown. Based on Matzkin (1992), we may without loss of generality assume $H_0\left(X\right) = E[D|X]$ and u has a uniform distribution, since then $\Pr(D = 1|X) = \Pr\left[u \leq H_0\left(X\right)\right] = H_0\left(X\right)$.

We then have the model

$$D = \mathbb{I}\left[H_0\left(X\right) - u \geq 0\right], \tag{4.1}$$

$$Y = \max(0, X^\top \beta_0 - e) D. \tag{4.2}$$

The latent error terms e and u are not independent of each other, so the model does not have selection on observables. When H_0 and the joint distribution of e and u are parameterized, Cragg (1971) and later authors call this a double hurdle model, because two hurdles must be crossed, $H_0\left(X\right) \geq u$ and $X^\top \beta_0 \geq e$, before a positive quantity of Y can be observed. We therefore call equations (4.1) and (4.2) the semiparametric double hurdle model.

Let $F_{e,u}\left(e, u\right)$ denote the unknown joint distribution function of $(e, u)^\top$, and define the functions M and $\widetilde{F}_0$ by

$$M\left(X\right) := E[Y|X] = \int_{\text{support } (e,u)} \max(0, X^\top \beta_0 - e) \mathbb{I}(H_0\left(X\right) - u \geq 0) dF_{e,u}\left(e, u\right) =: \widetilde{F}_0[X^\top \beta_0, H_0\left(X\right)].$$

We now give one set of conditions that suffice to identify this model based on Theorem 3.2. Define the function F_0 by $F_0[V + Z^\top \delta_0, H_0\left(V, Z\right)] = \widetilde{F}_0[X^\top \beta_0, H_0\left(X\right)]$, where V is the first element of X, Z is the vector of remaining elements of X, and δ_0 is the corresponding vector of elements of β_0 scaled by the first element of β_0. This construction is made without loss of generality because it just absorbs a scale normalization of the function $X^\top \beta_0$ into F_0. Note that in this construction, V is a scalar, and so is a special case of the more general framework given in the previous Theorems.

Assumption 5 *Assume equations (4.1) and (4.2) hold and that $(e, u)^\top$ are continuously distributed with unknown joint distribution function $F_{e,u}\left(e, u\right)$, and are independent of X. Assume V is continuously distributed with support $\mathbb{R}$, conditional on Z. Assume the $J \times 1$ vector of zeroes is in the support*

of Z. Let $\widetilde{z}_j$ denote the $J \times 1$ vector that has element j equal to z_j, and all other elements equal to zero. For $j = 1, ..., J$, assume for some $z_j \neq 0$ in the support of Z_j, there exists a $v(z_j)$ such that $H_0(v(z_j), z_j)$ is in the support of $H_0(V, 0)$.

Corollary 4.1 *Let Assumption 5 hold, then δ_0, and the functions F_0 and H_0 are identified.*

Note that Corollary 4.1 identifies δ_0, and so only identifies the original β_0 up to a scale normalization. If desired, the scaling factor (corresponding to the coefficient of V) could be identified in a variety of ways, e.g., if the probability that $D = 1$ goes to one as V goes to infinity, then the scaling factor will equal $\lim_{v \to \infty} E[\partial M(v, Z)/\partial v]$.

Based on our Theorems, Assumption 5 is stronger than necessary for Corollary 4.1. For example, the Corollary bases identification on Theorem 3.2 but not 3.1, and assumes infinite support for V, which simplifies for Theorem 3.2 but is not required. We provide this particular corollary as one example of easy to interpret conditions that suffice for identification without exclusion, or parametric functional form restrictions (beyond linearity of one index). Note also that many of these conditions are testable, e.g., we can estimate $H_0(V, Z) = E[D|V, Z]$, and then check whether a given $v(z_j)$ yields an estimate $H_0(v(z_j), z_j)$ that equals the estimate of $H_0(v, 0)$ for some value of v.

4.2 Binary Choice With an Endogenous Regressor Control Function Without Instruments

Suppose we have a threshold crossing binary choice model

$$Y = \mathbb{I}(X_1^\top \alpha_0 + X^e \gamma_0 - e \geq 0), \tag{4.3}$$

where X^e is an endogenous regressor with

$$X^e = g_0(X_1) + u, \tag{4.4}$$

where e and u are unobserved possibly correlated error terms. In our empirical application, Y will indicate whether an individual moves (migrates) from one state to another in the United States, and X^e will be logged income. People often move to find better jobs, and unobservables that affect the willingness to relocate are likely to be related to unobservables affecting income, so X^e will generally be an endogenous regressor.

We assume to have the semiparametric model of equation (4.3) for migration Y (semiparametric because the distribution of e is both unknown and correlated with X^e), but the model for our endogenous regressor X^e is nonparametric. One could argue that much is known about the determinants of income, so perhaps the model for X^e could be parameterized as well, but we wish to consider the case where covariates X_1 have been collected that focus on migration, and just nonparametrically define $g_0(X_1) := E[X^e|X_1]$. It then follows that $E[u|X_1] = 0$. We assume endogeneity of X^e in the Y equation takes the 'control function' form $e|X_1, u \sim e|u$. Define $X^\top \beta_0 := X_1^\top \alpha_0 + X^e \gamma_0$ and $H_0(X) := X^e - g_0(X_1)$. Let V be the first element of the vector X_1, and let Z be the vector consisting

of the remaining elements of X_1 and X^e (as in the previous subsection, here we simplify application of our theorems to the case where V is a scalar). Then, assuming the first element of α_0 is positive, we may without loss of generality scale all the coefficients and scale e to set the first element of α_0 equal to one ($\alpha_{0;1} = 1$), and thereby rewrite equation (4.3) as

$$Y = \mathbb{I}(X^\top \beta_0 - e \geq 0) = \mathbb{I}(V + Z^\top \delta_0 - e \geq 0), \tag{4.5}$$

where $\beta_0 = (1, \delta_0^\top)^\top$. Exploiting the condition $e|X_1, u \sim e|u$ and $u = H_0(X)$, define the functions M and F_0 by

$$\begin{aligned} M(X) &:= E[Y|X] = E[E[\mathbb{I}(X^\top \beta_0 - e \geq 0)|X, u]|X], \\ &= E[E[\mathbb{I}(X^\top \beta_0 - e \geq 0)|X^\top \beta_0, H_0(X)]|X] =: F_0[X^\top \beta_0, H_0(X)]. \end{aligned}$$

Assumption 6 *Equations (4.4) and (4.5) hold with $g_0(X_1) := E[X^e|X_1]$ and $e|X_1, u \sim e|u$. Also V is continuously distributed conditional on Z.*

Corollary 4.2 *Let Assumptions 4 and 6 hold for $j = 1, \ldots, J$, then β_0, and the functions F_0 and H_0 are identified.*

In our empirical implementation of this model we will have some discrete regressors, and a continuous exogenous regressor, an individual's age, which we take to be the regressor V. We use the more general support restrictions of Assumption 4 in Corollary 4.2 rather than the simpler conditions of Corollary 4.1, because age does not have full real line support, though it is continuous over an interval, which facilitates satisfying the support restrictions in Assumption 4. We have evidence that the other required restrictions will similarly hold. In particular, among working age individual's migration probabilities decrease steadily with age (since the gains in lifetime expected earnings from migrating to a better paying job decrease linearly with age). Also, empirically income is highly non-linear in age, providing the required nonlinearity in $H_0(X)$. To satisfy the remaining conditions, we have that F_0 is monotonic in its first element since it equals a conditional distribution function, and migration probabilities vary greatly with age as required by the support assumptions.

5 Two-Step Semiparametric Least Squares Estimation

Assume $E[Y|X] = F_0[X^\top \beta_0, H_0(X)]$, and define $R(\beta, X) := E[Y \mid X^\top \beta, H_0(X)]$ for β in the parameter space $\Theta \subset \mathbb{R}^d$. The identification results above provide sufficient conditions for $R(\beta_0, X) \neq R(\beta, X)$ for all $\beta \in \Theta$, $\beta \neq \beta_0$ (i.e. $E[Y|X] \neq R(\beta, X)$ for all $\beta \neq \beta_0$). In turn, this implies by standard least squares arguments that $\beta_0 = \arg\min_{\beta \in \Theta} E[(Y - R(\beta, X))^2]$, i.e. the nonlinear least squares criteria uniquely identifies the parameter β_0. In this section, we propose an estimator for β_0 based on a sample analog of this minimization, in which the functions F_0 and H_0 that comprise R are estimated using kernel regressions. We describe this estimator in detail for the binary choice model with an endogenous regressor without instruments. Identification in the binary choice model follows under Assumptions 4 and 6, by Corollary 4.2 above. Although we focus on this binary choice model here, the same estimation

approach can be applied to the more general model described in equation (1.1). The only difference for other applications is the form of the plug-in estimator for $H_0(X)$.

To derive the asymptotic properties of our estimator we apply generic limiting distribution results in Escanciano, Jacho-Chávez and Lewbel (2014). One of the example applications in that paper applies to a similar binary choice model, however, the estimator we propose here differs substantially from the one proposed there, and has a different limiting distribution with different asymptotic properties. For example, the semiparametric least squares estimator we propose here has an oracle property regarding estimation of the function F_0 that is not possessed by the semiparametric maximum likelihood based estimator proposed in Escanciano, Jacho-Chávez and Lewbel (2014). This oracle property follows from remark 3.1 in Escanciano, Jacho-Chávez and Lewbel (2014, p. 430), and the fact that the gradient of F_0 has zero conditional mean as in Ichimura (1993). Therefore, the proposed estimator can be interpreted as a generalization of Ichimura (1993), that allows for additional generated conditional covariates while preserving the oracle property, and as such is applicable to a wider range of problems beyond those with dichotomous responses.

Recall that $E[Y|X] = E[Y|\, w(X, \beta_0, g_0(X_1))]$ a.s., where $w(X, \beta_0, g_0(X_1)) := [X^\top \beta_0, X^e - g_0(X_1)]^\top$, $X := [X_1^\top, X^e]^\top$ may contain both continuous and discrete regressors, $\beta_0 := \left[\alpha_0^\top, \gamma_0\right]^\top$ – recall that we impose the normalization restriction $\alpha_{0;1} = 1$, and with some abuse of notation we denote the remainder parameters by β_0 in what follows. Similarly, we let $g_0(x_1) = E[X^e|X_1 = x_1]$ where X^e is a scalar random variable and $X_1 \subset X$. We assume that a random sample $\{Y_i, X_{1i}^\top, X_i^e\}_{i=1}^n$ is observed from the joint distribution of $(Y, X_1^\top, X^e)^\top$ taking values in $\mathcal{X}_Y \times \mathcal{X}_{X_1} \times \mathcal{X}_{X^e} \in \mathbb{R}^{d+2}$. For any candidate value of β and conditional mean function g, let $W(\beta, g) := [X^\top \beta, X^e - g(X_1)]^\top$, $W_i(\beta, g) := [X_i^\top \beta, X_i^e - g(X_{1i})]^\top$, $W_0 := W(\beta_0, g_0)$, $W_{0i} := W_i(\beta_0, g_0)$, and set

$$F(w|\, \beta, g) := E[Y|\, W(\beta, g) = w], \qquad w \in \mathbb{R}^2.$$

The regression function $F(w|\, \beta, g)$ can be consistently estimated by the nonparametric Nadaraya-Watson kernel estimator

$$\begin{aligned}
\widehat{F}(w|\, \beta, g) &:= \widehat{T}(w|\, \beta, g) / \widehat{f}(w|\, \beta, g), \text{ where} \\
\widehat{T}(w|\, \beta, g) &:= \frac{1}{n} \sum_{j=1}^n Y_j K_{\widehat{h}_n}(W_j(\beta, g) - w), \\
\widehat{f}(w|\, \beta, g) &:= \frac{1}{n} \sum_{j=1}^n K_{\widehat{h}_n}(W_j(\beta, g) - w),
\end{aligned}$$

$K_h(w) = k_h(w_1) k_h(w_2)$, $w = (w_1, w_2)$, $k_h(u) = h^{-1} k(u/h)$, $k(\cdot)$ is a kernel function and $\widehat{h}_n$ denotes a possibly data-dependent bandwidth parameter; see Assumptions B.4 and B.5 in Appendix B. Our two step semiparametric least squares estimator (SLS) is

$$\widehat{\beta} := \arg\min_{\beta \in \Theta} \mathcal{S}_n(\beta, \widehat{g}) \equiv \frac{1}{n} \sum_{i=1}^n \left[Y_i - \widehat{F}(W_i(\beta, \widehat{g})\,|\beta, \widehat{g})\right]^2 \widehat{a}_i, \tag{5.1}$$

where $\widehat{g}$ represents the first-stage Nadaraya-Watson estimator of g_0 satisfying Assumption B.9 in Appendix B, and $\widehat{a}_i := \mathbb{I}(\widehat{f}_i^* \geq \tau_n)$, with $\widehat{f}_i^* \equiv \widehat{f}(W_i(\widehat{\beta}^*, \widehat{g})|\widehat{\beta}^*, \widehat{g})$ for $i = 1, \ldots, n$, is a trimming function

introduced here to keep $\widehat{f}\left(W_i\left(\beta,\widehat{g}\right)|\,\beta,\widehat{g}\right)$ away from zero, with $\tau_n \to 0$, as $n \to \infty$ at a suitable rate. The preliminary consistent estimator, $\widehat{\beta}^*$, of β_0 can be obtained by semiparametric M-estimation with fixed trimming, i.e. $\widehat{a}_i := \mathbb{I}(X_i \in A)$ for a compact set $A \subset \mathcal{X}_X$, see e.g. Delecroix, Hristache and Patilea (2006).

Notice that the estimator defined by (5.1) is like Ichimura's (1993) estimator after plugging in $\widehat{g}$, and is also considered in Ichimura and Lee (2010), except that here g_0 is not assumed to have an index structure, and we are allowing for a random trimming function and a possibly data-dependent bandwidth. We establish the asymptotic properties of (5.1) by repeated applications of Escanciano, Jacho-Chávez and Lewbel's (2014) uniform-in-bandwidth result.

After defining

$$\Psi_i := [Y_i - F_0(W_{0i}|W_0)]\partial_\beta F_0(W_{0i}|W_0) - [X_i^e - g_0(X_{1i})]E[\partial_{\overline{g}} F_0(W_{0i})\partial_\beta F_0(W_{0i}|W_0)|X_{1i}],$$

with $\partial_{\overline{g}} F_0(W_{0i}) := \partial F_0(w(X_i,\beta_0,\overline{g})|W_0)/\partial\overline{g}|_{\overline{g}=g_{0i}}$, the following theorem establishes the consistency and asymptotic normality of the proposed estimator.

Theorem 5.1 *Let Assumptions 4 and 6 above, and Assumptions B.1 – B.10 in Appendix B hold. Then, $\widehat{\beta}$ is consistent and asymptotically normal*

$$\sqrt{n}(\widehat{\beta} - \beta_0) \longrightarrow_d N(0, \Delta_0^{-1}\Omega_0\Delta_0^{-1}),$$

where $\Omega_0 := E[\Psi_i\Psi_i^\top]$ and $\Delta_0 := E[\partial_\beta F_0(W_{0i}|W_0)\partial_\beta^\top F_0(W_{0i}|W_0)]$.

The asymptotic variance of $\widehat{\beta}$ in this theorem can be readily estimated by the analogue principle or by bootstrap methods. Estimation of g_0 has an impact in the asymptotic variance of $\widehat{\beta}$ through the term $[X_i^e - g_0(X_{1i})] \times E[\partial_{\overline{g}} F_0(W_{0i})\partial_\beta F_0(W_{0i}|W_0)|X_{1i}]$, however, the estimator of $\widehat{\beta}$ has an *oracle* property with respect to F_0, as its asymptotic properties are not affected by the lack of knowledge of F_0. Ichimura and Lee (2010) show this oracle property for the special case where the first stage estimate g_0 has an index structure (and assumed identification by exclusion assumptions), while our results show this oracle property holds more generally for g_0 estimated nonparametrically, and under our more general conditions regarding trimming and bandwidth selection.

6 Numerical Results

In this section we discuss the numerical implementation of our estimator (5.1) in the context of a small Monte Carlo experiment and then an empirical implementation.

We make use of the `np` package by Hayfield and Racine (2008) in the statistical computing environment `R`. In particular, under its General Public License (GPL) we modify its estimating function called `npindex(...,method="ichimura",...)` to allow the inclusion of a second conditioning variable when calculating Ichimura's (1993) estimator and user-specified weights. We use the option `optim.method="BFGS"` and `optim.method="Nelder-Mead"` in each of the 20 times (option `nmulti=20`) we randomly restarted the optimization algorithm to avoid finding local minima in both the Monte

Carlo and the empirical application respectively. We find our numerical calculations to be stable using a simple 2nd order Gaussian kernel. We estimate our bandwidths jointly with the unknown β_0, i.e.

$$(\widehat{\beta}^\top, \widehat{h}^\top) = \underset{(\beta^\top, h^\top)^\top \in \Theta \times \mathbb{R}_+^2}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} [Y_i - \widehat{F}_i(W_i(\beta, \widehat{g}) | \beta, \widehat{g})]^2 \widehat{a}_i, \tag{6.1}$$

making use of the fact that our asymptotic theory permits data dependent bandwidth choice, and where $\widehat{F}_i$ denotes a leave-one-out version of $\widehat{F}$. See also Marron (1994), Härdle, Hall, and Ichimura (1993), and Rothe (2009) for related results regarding kernel and bandwidth choice. Random starting values in each restart and trimming function, $\widehat{a}_i$, follow the original implementation by the np package developers and maintainers.

6.1 Monte Carlo Experiments

We assess the performance of our identification strategy and our estimator with a Monte Carlo simulation of a binary choice model with an endogenous regressor control function, where no outside instruments are available, as described in Section 4.2. In particular, we generate 1000 samples of pseudo-random numbers, $\{Y_i, X_{1i}, X_i^e\}_{i=1}^n$, with $n \in \{250, 500, 1000\}$, from (4.3)–(4.4), where X_1 is univariate and has a centered beta marginal distribution with shape parameters $(2, 2)$, $g_0(u) = 2\phi(u)$ where ϕ represents the probability density function of a standard normal random variable, and the error terms, ε and u, were generated independently of $X = [X_1, X^e]^\top$ from marginal standard normal distributions with correlation coefficient $\rho = \{0, -1/2, -3/4\}$.

As required for identification, we set $\alpha_0 = 1$, and the parameter of interest becomes $\beta_0 = [1, \gamma_0]^\top$, so γ_0 is to be estimated, with its true value equal one. We calculated 4 sets of estimators of our parameter of interest, $\gamma_0 = 1$: The standard Probit [1], Ichimura's (1993) estimator [2], the infeasible semiparametric estimator that uses the true g_0 [3] and the proposed (feasible) semiparametric least squares with g_0 estimated by the Nadaraya-Watson estimator with bandwidths chosen in each replication by least squares cross-validation. Results are presented in Table 1. They show the simulated median bias (Bias), standard deviation (Std. Dev.), root mean square error (RMSE), and the mean absolute deviation (MAE). Both [1] and [2] are only consistent when $\rho = 0$, because they are misspecified otherwise. They are useful as benchmarks to check if errors in our estimator are smaller than errors due to (sometimes) misspecified existing estimators. Notice that when $\rho = 0$, the Probit model is correctly specified and the most efficient. Model [3] provides an infeasible standard in all scenarios against which the proposed estimator can be compared to measure the extent of the first-step impact on the estimation precision of the proposed estimator, [4].

In all scenarios and sample sizes, estimators [3] and [4] performs well in terms of median bias and RMSE. The Monte Carlo variance of [4] is larger than [3]'s as expected by the asymptotic theory for small to medium sample sizes, but they become comparable to [3]'s for $n = 1000$. The relatively small RMSE of estimators [3] and [4], and the rate at which it shrinks as the sample size grows, suggests that our identification without instruments is not weak.

When $\rho = 0$, the simulated variance of the infeasible version of the proposed estimator, [3], is slightly larger than [2]'s indicating a loss of efficiency when trying to control for non-existent endogeneity in

the model. When the endogeneity problem is severe, i.e. $\rho = -3/4$, both versions of the proposed estimator perform very well in comparison to the standard Probit and Ichimura's (1993) estimators, which become substantially biased.

6.2 Empirical Application

As previously discussed, we now estimate a binary choice model for workers' migration decisions based on a sample of 22-69 years old male household heads who had completed education by the time of interview and who reported positive labor income during 1989-90. The sample is drawn from the 1990 wave of the Panel Study of Income Dynamics (PSID). The top 1% highest earning individuals are dropped to reduce the impact of outliers. Details regarding this data construction are in Dong (2010). Table 2 shows descriptive statistics of the resulting 4582 observations in the sample.

Let Y indicate if an individual moves (migrates, 1 or 0) from one state to another in the United States in the years 1991-93, and let X^e be the logarithm of their average labor income in 1989 and 1990. Exogenous covariates X_1 are State (number of states ever lived in, 1-8), Edu (dummy indicating college or above education, 1 or 0), Size (logarithm of family size, 1-17), and Age (22-69). The model is $Y = \mathbb{I}(X^\top \beta_0 - e \geq 0) = \mathbb{I}(X_1^\top \alpha_0 + X^e \gamma_0 - e \geq 0)$ where X^e is an endogenous regressor with $X^e = g_0(X_1) + u$, and e and u are unobserved error terms, correlated with each other but independent of X_1. Then as shown earlier, $E[Y|X] = F_0\left[w(X, \beta_0, g_0)\right]$ where $w(X, \beta_0, g_0) := [X^\top \beta_0, X^e - g_0(X_1)]^\top$.

Note that by construction $E\left(u|X_1\right) = 0$, so u is mean independent of X_1. This means that u is not unobserved ability, rather, it is unobserved ability after conditioning on education level as well as other covariates. The standard control function assumption regarding u is therefore that this mean independence extends to full independence of X_1. We include Edu in the list of covariates X_1 since it is relevant for wages and hence for the migration decision, and is predetermined at the time of the migration decision.

If we had an exclusion, i.e., a covariate that affected Y but not X^e, then this model would be identified by, e.g., Blundell and Powell (2004). However, it is not plausible in this application that any of observed determinants of Y would not also affect, X^e, since the utility from migrating depends at least in part on potential changes in labor income. We therefore assume identification without exclusion assumptions based on Theorem (3.2) (or more specifically, based on Corollary 4.2). We take age to be the continuous regressor V. This requires that the latent variable driving the probability that Y equals one be linear in age (which as noted earlier is supported by the human capital theory of migration) and that this probability varies over a reasonably wide range with age. The nonlinearity of $X^e - g_0(X_1) =: H_0(X)$ required for identification will hold if $g_0(X_1)$ is nonlinear in age.

Figure 1 shows Nadaraya-Watson Kernel regression estimates of $E\left[\text{Y}|\text{Age}\right]$ and $E\left[\log(\text{Income})|\text{Age}\right]$ with bandwidths chosen by Least-Squares cross-validation and 95% pointwise confidence intervals based on 399 bootstrapped replications. This figure provides empirical evidence supporting our above identifying assumptions. For example, $E\left[\text{Y}|\text{Age}\right]$ takes on a reasonably wide range of values, considering that the expected probability that a randomly chosen individual migrates in a three year period would not generally be high. Also the estimated migration probability is close to linear and certainly plausibly

monotonic in age, while log(Income) is highly nonlinear and non-monotonic in age. Similar results are reported in Dong (2010), who estimates a restricted version of a model similar to ours with more parametric restrictions, but still exploiting nonlinearity in $E[\log(\text{Income})|\text{Age}]$ to aid in identification.

We estimate the model three ways. Estimator [I] uses a Probit which ignores endogeneity of labor income and assumes e is normal, estimator [II] is Ichimura's (1993) which ignores endogeneity but does not assume a parameterized distribution for e, and estimator [III] is our proposed 2-Steps SLS. The coefficient of State is normalized to 1 in all specifications for comparison purposes. This is a free normalization. We also report marginal effects. Bandwidths in [II] and [III] were estimated jointly with the unknown β_0 coefficients to minimize least squares, and then used throughout to calculate related quantities such as asymptotic standard errors and marginal effects.

For our proposed estimator, [III], in the first stage we nonparametrically regress the endogenous covariate log(Income) on State, Edu, Size, and Age using generalized kernels as suggested in Racine and Li (2004) to handle the combination of discrete and continuous regressors, with smoothing parameters chosen by Least-Squares cross-validation. The resulting data-dependent bandwidths are 1 for State, 0.0442 for Edu, 0.1348 for Size and 2.1121 for Age. We then estimate β in the second step as in (6.1) where $X_i^e - \widehat{g}(X_{1i})$ equals the residuals from the first-step nonparametric regression. Table 3 shows the resulting estimates $\widehat{\beta}$. Define $\partial_{w_1} F_0(W_{0i}) := \partial F_0(w_1, w_2|W_0)/\partial w_1|_{w_1 = X_i^\top \beta_0}$. Table 3 also reports $\partial_{w_1}\widehat{F}$ evaluated at the sample mean of the estimated index in [II], and the sample mean of the estimated indexes corresponding to [III]. Marginal effects for all models are shown in Table 4. The marginal effects marked as (*) were obtained by multiplying the reported $\partial_{w_1}\widehat{F}$ in Table 3 by their corresponding estimated coefficient. Using results in Sperlich (2009), the asymptotic standard errors for model [II] and [III] are approximated as the square root of $\widehat{\text{var}}(\partial_{w_1}\widehat{F})\widehat{\beta}_j^2 + \partial_{w_1}\widehat{F}^2\widehat{\text{var}}(\widehat{\beta}_j)$. For discrete covariates Edu and Size, Table 4 also reports for each $t - s$ the corresponding change, $\widehat{F}_t - \widehat{F}_s$, where $\widehat{F}_l$ represents the value of $\widehat{F}$ evaluated at the implied index where a particular discrete variable is set equal to l while keeping the remaining part of the index equal to its sample mean. Results in Schuster (1972) also allow us to approximate their standard errors as the square root of $\widehat{\text{var}}(\widehat{F}_t) + \widehat{\text{var}}(\widehat{F}_s)$.

Although β_0 has the same normalization in all the estimators, estimated marginal effects were still generally closer across specifications than estimates of β_0, suggesting that small sample biases in estimating β_0 and F_0 may be offsetting to some extent. Moreover, marginal effects are more directly economically interpretable as the impacts of regressors on the probability of migrating. We therefore focus on summarizing marginal effects.

In all the estimates age has a negative impact on the probability of moving as expected by theory, however, the effect of age is more than 50% larger in our estimators that take endogeneity of income into account, which suggests the importance of controlling for endogeneity. The endogenous regressor log income has a negative effect in all the specifications, consistent with higher wage individuals having less income incentive to move. It is not clear how education should affect migration probabilities, and the estimates of this effect vary inconclusively across models. The effects of family size were not statistically significant, but our estimates controlling for income endogeneity suggest that larger families are more likely to move.

7 Conclusions

The new identification and estimation results in this paper are applicable to a wide variety of common econometric models including latent index models with an endogenous regressor, and nonlinear models with sample selection. The estimator we propose for this class of models allows for data-driven bandwidths, which we exploit by selecting bandwidths that minimize the same objective function that is used to estimate model parameters. This numerically effective procedure performs well in our Monte Carlo experiments and in an empirical application to a migration decision model.

Regarding identification, we show that identification based on functional form, without exclusion restrictions or instruments, extends to a semiparametric setting where error distributions are unknown, first stage regression functions are unknown, and the only parameterization is that one index in the model is linear. The somewhat surprising result, based on our theorems, simulations, and empirical results, is that parameters can be strongly identified and accurately estimated in this setting. While identification based on valid (and ideally randomized) instruments is of course preferable, it is important to know the extent to which reasonably precise identification and inference results can be obtained without instruments.

References

Andrews, D. W. K. (1995): "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

Blundell, R. W., and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655–679.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semi-parametric Models with Censoring," *Journal of Econometrics*, 32(2), 189–218.

Chen, X., O. B. Linton, and I. van Keilegom (2003): "Estimation of Semiparametric Models when the Criterion Function Is Not Smooth," *Econometrica*, 71(5), 1591–1608.

Cragg, J. G. (1971): "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods," *Econometrica*, 39(5), 829–44.

Delecroix, M., M. Hristache and V. Patilea (2006): "On Semiparametric M-Estimation in Single-Index Regression," *Journal of Statistical Planning and Inference*, 136(3), 730–769.

Dong, Y. (2010): "Endogenous Regressor Binary Choice Models Without Instruments, With an Application to Migration," *Economics Letters*, 107(1), 33-35.

Escanciano, J. C., D. T. Jacho-Chávez and A. Lewbel (2014): "Uniform Convergence of Weighted Sums of Non- and Semi-parametric Residuals for Estimation and Testing," Journal of Econometrics, 178, 426-443.

Härdle, W., P. Hall, and H. Ichimura (1993): "Optimal Smoothing in Single-Index Models," *The Annals of Statistics*, 21(1), 157–178.

Hayfield, T., and J. S. Racine (2008): "Nonparametric Econometrics: The np Package," *Journal of Statistical Software*, 27(5), 1–32.

Heckman, J. J. (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47(1), 153–161.

Heckman, J. J., R. L. Matzkin and L. Nehseim (2010): "Nonparametric Identification and Estimation of Nonadditive Hedonic Models," *Econometrica*, 78(5), 1569–1591.

Ichimura, H., and L. Lee (1991): "Semiparametric least squares estimation of multiple index models: single equation estimation"," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, ed. by W. A. Barnett, J. Powell, and G. Tauchen, pp. 3–49. Cambridge University Press.

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58, 71–120.

Ichimura, H., and S. Lee (2010): "Characterization of the Asymptotic Distribution of Semiparametric M-estimators," *Journal of Econometrics*, 159(2), 252–266.

Klein, R., C. Shen, and F. Vella (2014) "Semiparametric Selection Models with Binary Outcomes," Unpublished manuscript.

Komunjer, I. (2012): "Global Identification in Nonlinear Models with Moment Restrictions," *Econometric Theory*, 28(4), 719–729.

Lewbel, A. (2007): "Estimation of Average Treatment Effects With Misclassification," *Econometrica*, 75(2), 537–551.

Lewbel, A., and O. B. Linton (2007): "Nonparametric Matching and Efficient Estimators of Homothetically Separable Functions," *Econometrica*, 75(4), 1209–1227.

Lewbel, A., O. B. Linton, and D. McFadden (2011): "Estimating Features of a Distribution from Binomial Data," *Journal of Econometrics*, 162(2), 170–188.

Li, K.-C. (1991): "Sliced Inverse Regression for Dimension Reduction," *Journal of the American Statistical Association*, 86(414), 316–327.

Marron, J. S. (1994): "Visual Understanding of Higher-Order Kernels," *Journal of Computational and Graphical Statistics*, 3(4), 447–458.

Matzkin, R. L. (1992): "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and the Binary Choice Models," *Econometrica*, 60, 239–270.

Matzkin, R. L. (2007): "Nonparametric Identification,' in *Handbook of Econometrics*, ed. by J. J. Heckman, and E. E. Leamer, vol. VIB, pp. 5307–5368. Elsevier, North-Holland, Amsterdam.

Neumeyer, N., and Van Keilegom, I. (2010): "Estimating the Error Distribution in Nonparametric Multiple Regression with Applications to Model Testing," *Journal of Multivariate Analysis*, 101, 1067–1078.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by D. McFadden, and R. F. Engle, vol. IV, pp. 2111–2245. Elsevier, North-Holland, Amsterdam.

Pinkse, J. (2001): "Nonparametric Regression Estimation using Weak Separability," Unpublished manuscript.

Racine, J. S., and Q. Li (2004): "Nonparametric Estimation of Regression Functions With Both Catagorical and Continuous Data," *Journal of Econometrics*, 119(1), 99–130.

Rivers, D., and Q. H. Vuong (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39(3), 347–366.

Robinson, P.M. (1988): "Root-n-consistent Semiparametric Regression," *Econometrica*, 56(4), 931–954.

Rothe, C. (2009): "Semiparametric Estimation of Binary Response Models with Endogenous Regressors," *Journal of Econometrics*, 153(1), 51–64.

Schuster, E. F. (1972): "Joint Asymptotic Distribution of the Estimated Regression Function at a Finite Number of Distinct Points," *Annals of Mathematical Statistics*, 43(1), 84–88.

Sperlich, S. (2009): "A Note on Non-parametric Estimation with Predicted Variables," *The Econometrics Journal*, 12(2), 382–395.

Appendix A Proofs of Main Results

Proof of Theorem 3.1: Dropping V for now for clarity, define $\partial_r F_0 := \partial F_0(r,H)/\partial r$ and $\partial_H F_0 := \partial F_0(r,H)/\partial H$. Equating m with F_0 and taking derivatives shows that

$$\partial_{V_k} m = \partial_r F_0 \alpha_{0;k} + \partial_H F_0 \partial_{V_k} H. \tag{A-1}$$

Since $\alpha_{0;1} = 1$ we have for each $k = 2, \ldots, K$

$$\begin{pmatrix} \partial_{V_1} m \\ \partial_{V_k} m \end{pmatrix} = \begin{pmatrix} 1 & \partial_{V_1} H \\ \alpha_{0;k} & \partial_{V_k} H \end{pmatrix} \begin{pmatrix} \partial_r F_0 \\ \partial_H F_0 \end{pmatrix}.$$

By Assumption 2, $\partial_{V_k} g \neq \partial_{V_1} H \alpha_{0;k}$ so the matrix in the above equation is nonsingular. Inverting to solve for $\partial_r F_0$ and $\partial_H F_0$ gives:

$$\partial_r F_0 = \frac{\partial_{V_k} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_k} m}{\partial_{V_k} H - \partial_{V_1} H \alpha_{0;k}}$$
$$\partial_H F_0 = \frac{\partial_{V_k} m - \partial_{V_1} m \alpha_{0;k}}{\partial_{V_k} H - \partial_{V_1} H \alpha_{0;k}}$$

Equating the above expression for $\partial_r F_0$ based on coefficients indexed by k with the same expression evaluated at some other index i gives

$$\frac{\partial_{V_k} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_k} m}{\partial_{V_k} H - \partial_{V_1} H \alpha_{0;k}} = \frac{\partial_{V_i} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_i} m}{\partial_{V_i} H - \partial_{V_1} H \alpha_{0;i}}$$

so

$$\left(\partial_{V_k} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_k} m\right)\left(\partial_{V_i} H - \partial_{V_1} H \alpha_{0;i}\right) = \left(\partial_{V_i} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_i} m\right)\left(\partial_{V_k} H - \partial_{V_1} H \alpha_{0;k}\right)$$

which simplifies to

$$\left(\partial_{V_k} H \partial_{V_i} m - \partial_{V_i} H \partial_{V_k} m\right) = \left(\partial_{V_k} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_k} m\right) \alpha_{0;i} - \left(\partial_{V_i} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_i} m\right) \alpha_{0;k} \quad \text{(A-2)}$$

which is linear in $\alpha_{0;k}$ and $\alpha_{0;i}$. The same equation is obtained if one equates the expression for $\partial_H F_0$ based on two indices k and i.

Recalling the dependence of the functions above on V, equation (A-2) evaluated at $V = v$ and at $V = \widetilde{v}$ with $i = j$ can be written as

$$\begin{pmatrix} \partial_{V_k} H\left(\widetilde{v}\right) \partial_{V_j} m\left(\widetilde{v}\right) - \partial_{V_j} H\left(\widetilde{v}\right) \partial_{V_k} m\left(\widetilde{v}\right) \\ \partial_{V_k} H\left(v\right) \partial_{V_j} m\left(v\right) - \partial_{V_j} H\left(v\right) \partial_{V_k} m\left(v\right) \end{pmatrix} = \Psi \begin{pmatrix} \alpha_{0;j} \\ \alpha_{0;k} \end{pmatrix} \quad \text{(A-3)}$$

where the matrix Ψ is given by

$$\Psi = \begin{pmatrix} \partial_{V_k} H\left(\widetilde{v}\right) \partial_{V_1} m\left(\widetilde{v}\right) - \partial_{V_1} H\left(\widetilde{v}\right) \partial_{V_k} m\left(\widetilde{v}\right) & \partial_{V_j} H\left(\widetilde{v}\right) \partial_{V_1} m\left(\widetilde{v}\right) - \partial_{V_1} H\left(\widetilde{v}\right) \partial_{V_j} m\left(\widetilde{v}\right) \\ \partial_{V_k} H\left(v\right) \partial_{V_1} m\left(v\right) - \partial_{V_1} H\left(v\right) \partial_{V_k} m\left(v\right) & \partial_{V_j} H\left(v\right) \partial_{V_1} m\left(v\right) - \partial_{V_1} H\left(v\right) \partial_{V_j} m\left(v\right) \end{pmatrix}$$

Using equation (A-1), each term in the matrix Ψ has the form

$$\begin{aligned} &\partial_{V_k} H \partial_{V_1} m - \partial_{V_1} H \partial_{V_k} m \\ &= \partial_{V_k} H\left(\partial_r F_0 \alpha_{0;1} + \partial_g F_0 \partial_{V_1} H\right) - \partial_{V_1} H\left(\partial_r F_0 \alpha_{0;k} + \partial_H F_0 \partial_{V_k} H\right) \\ &= \left(\partial_{V_k} H - \alpha_{0;k} \partial_{V_1} H\right) \partial_r F_0 \end{aligned} \quad \text{(A-4)}$$

so

$$\Psi = \begin{pmatrix} \partial_r F_0(\widetilde{v}'\alpha, H\left(\widetilde{v}\right)) & 0 \\ 0 & \partial_r F_0(v'\alpha, H\left(v\right)) \end{pmatrix} \begin{pmatrix} \partial_{V_k} H\left(\widetilde{v}\right) - \alpha_{0;k} \partial_{V_1} H\left(\widetilde{v}\right) & \partial_{V_j} H\left(\widetilde{v}\right) - \alpha_{0;j} \partial_{V_1} H\left(\widetilde{v}\right) \\ \partial_{V_k} H\left(v\right) - \alpha_{0;k} \partial_{V_1} H\left(v\right) & \partial_{V_j} H\left(v\right) - \alpha_{0;j} \partial_{V_1} H\left(v\right) \end{pmatrix}.$$

Assumption 2 imposes sufficient conditions to ensure that the determinants of each of the two matrices on the right above are nonzero, which shows that Ψ is nonsingular.

Since Ψ is nonsingular, equation (A-3) can be solved for $\alpha_{0;k}$ and $\alpha_{0;j}$ by inverting Ψ, thereby identifying $\alpha_{0;k}$ and $\alpha_{0;j}$. Given identification of $\alpha_{0;k}$, we then may identify all other coefficients $\alpha_{0;i}$ by solving equation (A-2) (evaluated at $V = v$) for $\alpha_{0;i}$, which gives

$$\alpha_{0;i} = \frac{\partial_{V_k} H(v)\,\partial_{V_i} m(v) - \partial_{V_i} H(v)\,\partial_{V_k} m(v) + [\partial_{V_i} H(v)\,\partial_{V_1} m(v) - H_1(v)\,\partial_{V_i} m(v)]\,\alpha_{0;k}}{\partial_{V_k} H(v)\,\partial_{V_1} m(v) - \partial_{V_1} H(v)\,\partial_{V_k} m(v)},$$

Noting that the denominator in this expression is nonzero by Assumption 2 and equation (A-4). Finally, given identification of α_0, the function F_0 is identified by $F_0\left(V^\top \alpha_0, H(V)\right) = E\left[m(V) \,|V^\top \alpha_0, H(V)\right]$. *Q.E.D.*

Proof of Theorem 3.2: F_0 is identified on the support of $V^\top \alpha_0$, $H_0(V,0)$ by $F_0(r,H) = E[M(V,0) \,|\, V^\top \alpha_0 = r, H_0(V,0) = H]$. Then for each j, $\delta_{0;j}$ solves $M\left[v\left(z_j\right), \widetilde{z}_j\right] = F_0[v\left(z_j\right)^\top \alpha_0 + z_j \delta_{0;j}, H_0\left(v\left(z_j\right), z_j\right)]$. Invertibility of F_0 on its first element ensures that this solution is unique, and the support assumptions ensure that F_0 is identified at this point. Then, given this identification of each $\delta_{0;j}$, F_0 is identified by $F_0\left(V^\top \alpha_0 + Z^\top \delta_0, H_0(V,Z)\right) = E\left[M(V,Z) \,|V^\top \alpha_0 + Z^\top \delta_0, H_0(V,Z)\right]$. *Q.E.D.*

Proof of Corollary 4.1: The functions M and H_0 are identified by $M(V,Z) = E[Y|V,Z]$ and $H_0(V,Z) = E[D|V,Z]$. Given these results and the construction of F_0, Assumption 1 holds with $\alpha_0 = 1$. By construction, differentiability of F_0 follows from e and u being jointly continuously distributed, and similarly by construction F_0 is strictly monotonic and hence invertible in $V + Z^\top \delta_0$. The remaining conditions of Assumption 4 for $j = 1, ..., J$ hold by Assumption 5 with V having infinite support. Therefore the conditions of Theorem 3.2 hold, so Corollary 4.1 holds. *Q.E.D.*

Proof of Corollary 4.2: The functions M and H_0 are identified by $M(V,Z) = E[Y|V,Z]$ and $H_0(V,Z) := X^e - E[X^e|X_1]$. Given these results and the construction of F_0, Assumption 1 holds with $\alpha_0 = 1$. Given Assumptions 1 and 4 we have Theorem 3.2 holding, so Corollary 4.2 holds. *Q.E.D.*

Proof of Theorem 5.1: To prove the consistency of $\widehat{\beta}$, we need to prove the uniform convergence of $\mathcal{S}_n(\beta, \widehat{g})$ to $\mathcal{S}(\beta, g_0) \equiv E[\overline{\mathcal{S}_n}(\beta, g_0)]$, uniformly in β, where

$$\overline{\mathcal{S}_n}(\beta, g_0) \equiv \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - F_0(W_i(\beta, g_0)\,|\beta, g_0)\right]^2.$$

Firstly, it follows from the Triangle inequality that

$$\sup_{\beta\in\Theta} |\mathcal{S}_n(\beta, \widehat{g}) - \mathcal{S}(\beta, g_0)\,| \leq \sup_{\beta\in\Theta} |\mathcal{S}_n(\beta, \widehat{g}) - \overline{\mathcal{S}_n}(\beta, g_0)\,| + \sup_{\beta\in\Theta} |\overline{\mathcal{S}_n}(\beta, g_0) - \mathcal{S}(\beta, g_0)\,|.$$

Notice that

$$\begin{aligned}
\sup_{\beta\in\Theta} |\mathcal{S}_n(\beta, \widehat{g}) - \overline{\mathcal{S}_n}(\beta, g_0)\,| \leq & \max_{1\leq i\leq n} \sup_{\beta\in\Theta; x\in\mathcal{X}} \left|[\widehat{F}(W_i(\beta,\widehat{g})|\beta,\widehat{g}) - F_0(W_i(\beta,g_0)|\beta,g_0)]\right| \\
& \times \sup_{\beta\in\Theta; x\in\mathcal{X}} \left|\frac{1}{n}\sum_{i=1}^{n} \widehat{a}_i[\widehat{F}(W_i(\beta,\widehat{g})|\beta,\widehat{g}) + F_0(W_i(\beta,g_0)|\beta,g_0) - 2Y_i]\right| \\
& + \sup_{\beta\in\Theta} \left|\frac{1}{n}\sum_{i=1}^{n} \left[Y_i - F_0(W_i(\beta,g_0)\,|\beta,g_0)\right]^2 (\widehat{a}_i - 1)\right|.
\end{aligned}$$

By the uniform-in-bandwidth results of Escanciano, Jacho-Chávez and Lewbel (2014), the simple inequality

$$\mathbb{I}(\widehat{f}^* < \tau_n) \leq \mathbb{I}(f\left(W_{0i}|\, W_0\right) < 2\tau_n) + \mathbb{I}(|\, \widehat{f}^* - f\left(W_{0i}|\, W_0\right)\, | > \tau_n),$$

and Assumptions B.7, B.9 and B.3, we obtain $\max_{1\leq i\leq n} |\widehat{a}_i - 1| = o_P(1)$. From a Uniform Law of Large Numbers (ULLN), it then follows that

$$\sup_{\beta\in\Theta} \left| \frac{1}{n}\sum_{i=1}^{n} \left[Y_i - F_0(W_i\left(\beta, g_0\right)|\beta, g_0)\right]^2 \left(\widehat{a}_i - 1\right) \right| = o_P(1).$$

Now, by the uniform-in-bandwidth results of Escanciano, Jacho-Chávez and Lewbel (2014) and Assumption B.3

$$\max_{1\leq i\leq n} \sup_{\beta\in\Theta; x\in\mathcal{X}} \left| [\widehat{F}(W_i(\beta,\widehat{g})|\beta,\widehat{g}) - F_0(W_i(\beta,g_0)|\beta,g_0)] \right| = o_P(1). \tag{A-5}$$

Using the last equality and the ULLN we conclude that

$$\sup_{\beta\in\Theta; x\in\mathcal{X}} \left| \frac{1}{n}\sum_{i=1}^{n} \widehat{a}_i [\widehat{F}(W_i(\beta,\widehat{g})|\beta,\widehat{g}(x)) + F_0(W_i(\beta,g_0)|\beta,g_0) - 2Y_i] \right| = O_P(1).$$

The ULLN is justified since $F_0(W_i\left(\beta, g_0\right)|\beta, g_0)$ is continuous in β and uniformly bounded under our assumptions. The same arguments imply that $\sup_{\beta\in\Theta} |\overline{\mathcal{S}_n}\left(\beta, g_0\right) - \mathcal{S}\left(\beta, g_0\right)| = o_P(1)$.

We now show identification of the nonlinear least squares criteria. By Corollary 4.2, under Assumptions 4 and 6, it holds that if $\beta_0 \neq \beta$ then $F_0(W_i\left(\beta, g_0\right)|\beta, g_0) \neq F_0(W_i\left(\beta_0, g_0\right)|\beta_0, g_0)$ with positive probability. Then, by $E[Y|X] = F_0(W_i\left(\beta_0, g_0\right)|\beta_0, g_0)$, we have

$$\begin{aligned} \mathcal{S}\left(\beta, g_0\right) &= \mathcal{S}\left(\beta_0, g_0\right) + E\left[\left(F_0(W_i\left(\beta, g_0\right)|\beta, g_0) - F_0(W_i\left(\beta_0, g_0\right)|\beta_0, g_0)\right)^2\right] \\ &> \mathcal{S}\left(\beta_0, g_0\right). \end{aligned}$$

Thus, the nonlinear least squares criteria uniquely identifies the parameter β_0. Hence we conclude by Theorem 2.1 in Newey and McFadden (1994, p. 2121) that $\widehat{\beta} = \beta_0 + o_P(1)$.

We now prove the asymptotic normality of the proposed estimator by standard methods coupled with the general result in Escanciano, Jacho-Chávez and Lewbel (2014). By the first order conditions

$$0 \equiv -\sqrt{n}\partial_\beta \mathcal{S}_n(\widehat{\beta},\widehat{g}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [Y_i - \widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g}(x))]\widehat{\psi}(X_i,\widehat{\beta}),$$

where $\widehat{\psi}(X_i,\widehat{\beta}) = \widehat{a}_i\partial_\beta^\top \widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g})$. Now by a standard Taylor expansion,

$$-\sqrt{n}\partial_\beta \mathcal{S}_n(\widehat{\beta},\widehat{g}) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} [Y_i - \widehat{F}(W_i(\beta_0,\widehat{g})|\beta_0,\widehat{g}(x))]\widehat{\psi}(X_i,\widehat{\beta}) + H_n(\widetilde{\beta},\widehat{g})\sqrt{n}(\widehat{\beta} - \beta_0),$$

where $\widetilde{\beta}$ is an intermediate point between $\widehat{\beta}$ and β_0 and

$$H_n(\widetilde{\beta},\widehat{g}) = \frac{1}{n}\sum_{i=1}^{n} \widehat{a}_i\partial_\beta \widehat{F}(W_i(\widetilde{\beta},\widehat{g})|\widetilde{\beta},\widehat{g})\partial_\beta^\top \widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g}).$$

Note that,

$$\partial_\beta \widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g}) = X_i \partial_{w1}\widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g}) + \partial_{\beta 2}\widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g}),$$

where $\partial_{w1}\widehat{F}$ and $\partial_{\beta 2}\widehat{F}$ denote the derivatives of $\widehat{F}(w_1, w_2|\beta, g)$ with respect to w_1 and β respectively. From Escanciano, Jacho-Chávez and Lewbel (2014) it then follows that

$$\left|\partial_\beta \widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g}) - \partial_\beta^\top F_0(W_i|\beta_0, g_0)\right| = o_P(1).$$

Hence, applying Theorem 3.2 in Escanciano, Jacho-Chávez and Lewbel (2014, p. 430) with the class $\Phi = \mathcal{T}^{\eta_F}$, and using the uniform consistency of $\widehat{\psi}(X_i, \widehat{\beta})$ we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - \widehat{F}(W_i(\widehat{g})|\widehat{g}(x))]\widehat{\psi}(X_i,\widehat{\beta}) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - F_0(W_{0i}|W_0)]\psi^\perp(X_i,\beta_0) - [X_i^e - g_0(X_{1i})]E[\partial_{\overline{g}} F_0(W_{0i})\psi^\perp(X_i,\beta_0)|X_{1i}] + o_P(1),
\end{aligned}$$

where

$$\psi^\perp(X_i,\beta_0) = \partial_\beta^\top F_0(W_{0i}|W_0) - E[\partial_\beta^\top F_0(W_{0i}|W_0)|W_0].$$

Using Ichimura's (1993) arguments one can show that

$$E[\partial_\beta^\top F_0(W_{0i}|W_0)|W_0] = 0.$$

On the other hand, by the uniform consistency of $\partial_\beta \widehat{F}(W_i(\widetilde{\beta},\widehat{g})|\widetilde{\beta},\widehat{g})$ it follows

$$H_n(\widetilde{\beta},\widehat{g}) \to_P \Delta_0 = E[\partial_\beta F_0(W_0)\partial_\beta^\top F_0(W_0)].$$

Hence, we conclude that

$$\begin{aligned}
&\sqrt{n}(\widehat{\beta} - \beta_0) \\
&= -\Delta_0^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^n [Y_i - F_0(W_{0i}|W_0)]\partial_\beta F_0(W_{0i}|W_0) - [X_i^e - g_0(X_{1i})]E[\partial_{\overline{g}} F_0(W_{0i})\partial_\beta F_0(W_{0i}|W_0)|X_{1i}] + o_P(1).
\end{aligned}$$

The result then follows from an application of the Linderberg-Lévy CLT. *Q.E.D.*

Appendix B Asymptotic Theory - Conditions

Define for any vector $(a_1, ..., a_d)$ of d integers the differential operator $\partial_x^a := \partial^{|a|}/\partial x_1^{a_1} \dots \partial x_d^{a_d}$, where $|a| := \sum_{i=1}^d a_i$. Let $\mathcal{X}_X$ be the support of X, and let $\mathcal{X}_{X_1}$ be the support of X_1. Assume that $\mathcal{X}_{X_1}$ is a convex, bounded subset of $\mathbb{R}^d$, with non-empty interior. For any smooth function $h : \mathcal{X}_{X_1} \subset \mathbb{R}^d \to \mathbb{R}$ and some $\eta > 0$, let $\underline{\eta}$ be the largest integer smaller than η, and

$$\|h\|_{\infty,\eta} := \underset{|a|\leq\underline{\eta} x\in\mathcal{X}}{\text{maxsup}} |\partial_x^a h(x)| + \max_{|a|=\underline{\eta}} \sup_{x_1\neq x_2} \frac{|\partial_x^a h(x_1) - \partial_x^a h(x_2)|}{\|x_1 - x_2\|^{\eta-\underline{\eta}}}.$$

Further, let $C_M^\eta(\mathcal{X}_{X_1})$ be the set of all continuous functions $h : \mathcal{X}_{X_1} \subset \mathbb{R}^d \to \mathbb{R}$ with $\|h\|_{\infty,\eta} \leq M$. Since the constant M is irrelevant for our results, we drop the dependence on M and denote $C^\eta(\mathcal{X}_{X_1})$. Define the sup-norm $\|h\|_\infty := \sup_{x\in\mathcal{X}_{X_1}} |h(x)|$.

Let $x := [x_1^\top, x^e]^\top$, $f_X(x|w, W)$ be the density, with respect to a σ-finite measure $\mu_W(\cdot)$, of X conditional on $W(\beta, g) = w$ and evaluated at $x \in \mathcal{X}_X$. Similarly, define $W_0 := W(\beta_0, g_0)$ and $W_{0i} := W_i(\beta_0, g_0)$. Let $\mathcal{G}$ from $\mathbb{R}^d$ to $\mathbb{R}$ be the class of functions where g belongs. Define the class of functions $\mathcal{W}$ as

$$\mathcal{W} := \{x \to (v_1 + x_2^\top\beta, x^e - g(x_1)) : \beta \in \Theta \subset \mathbb{R}^{d+1},\ g \in \mathcal{G} \subset C^{\eta_g}(\mathcal{X}_{X_1}),\ \|g - g_0\|_\infty < \delta\},$$

for an arbitrarily small $\delta > 0$ and $\eta_g > d$. The following assumptions will be needed in our subsequent analysis. Escanciano, Jacho-Chávez and Lewbel (2014) provide sufficient primitive conditions for some of our high-level assumptions in terms of the density of X and the class $\mathcal{G}$.

Assumption B.1 *The sample observations $\{Y_i, X_i^\top\}_{i=1}^n$ are a sequence of independent and identically distributed (iid) variables, distributed as $(Y, X^\top)$ and satisfying $E[|Y|^p | X = x] < \infty$ a.s., for some $p > 2$, and $E[Y|X] = E[Y|W_0]$ a.s.*

Assumption B.2 *The parameter space Θ is a compact subset of $\mathbb{R}^d$ and β_0 is an element of its interior. The class $\mathcal{G} \subset C^{\eta_g}(\mathcal{X}_{X_1})$, for some $\eta_g > d$.*

Assumption B.3 *For all $W \in \mathcal{W}$ and $x \in \mathcal{X}_X$: $f(w|W)$, $F_0(w|W)$ and $f_X(x|w, W)$ are r-times continuously differentiable in w, with uniformly (in w, W and x) bounded derivatives (including zero derivatives) where r is as in Assumption B.4 below.*

Assumption B.4 *The kernel function $k(t) : \mathbb{R} \to \mathbb{R}$ is bounded, symmetric, continuously differentiable, and satisfies the following conditions: $\int k(t)\,dt = 1$, $\int t^l k(t)\,dt = 0$ for $0 < l < r$, and $\int |t^r k(t)|\,dt < \infty$, for some $r \geq 2$; $|\partial k(t)/\partial t| \leq C$ and for some $v > 1$, $|\partial k(t)/\partial t| \leq C|t|^{-v}$ for $|t| > L$, $0 < L < \infty$.*

Assumption B.5 *The possibly data-dependent bandwidth $\widehat{h}_n$ satisfies $P(a_n \leq \widehat{h}_n \leq b_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers a_n and b_n such that: (i) $b_n \to 0$ and $a_n^4 n/\log n \to \infty$; (ii) $nb_n^{4r} \to 0$.*

Conditions B.1–B.3 are standard in the literature. Assumptions B.3 and B.7 below are needed for establishing uniform convergence rates for kernel estimators, and for a uniform expansion of nonparametric residual-marked empirical processes. Note that by consistency of $\widehat{g}$ we can take $\mathcal{G}$ to be contained in an arbitrary neighborhood of g_0.

Assumption B.4 is standard in the literature of nonparametric kernel estimation, while Assumption B.5 permits data-dependent bandwidths, as in Andrews (1995). In particular, our theory allows for plug-in bandwidths of the form $\widehat{h}_n = \widehat{c}h_n$ with $\widehat{c}$ stochastic and h_n a suitable deterministic sequence converging to zero as $n \to \infty$. Andrews (1995) points out that this condition holds in many cases for other common data-dependent bandwidth selection procedures, such as cross-validation, and generalized cross-validation. Obviously, our results also apply to deterministic sequences. In particular if $\widehat{h}_n$ is of the form $\widehat{h}_n = cn^{-\delta}$, for some constant $c > 0$, then Assumption B.5(ii) requires that $1/2r < \delta < 1/2$, so r needs to be greater than 1.

We now introduce two classes of functions that will serve as parameter spaces for the functions $F(w(x,\beta,g)|\beta,g)$ and $f(w(x,\beta,g)|\beta,g)$, respectively, where henceforth $w(x,\beta,g) := [x^\top\beta, x^e - g(x_1)]^\top$. Let $\mathcal{T}^\eta$ be a class of uniformly bounded and measurable functions

$$\{x \to q(w(x,\beta,g)|\beta,g) : \beta \in \Theta,\ g \in \mathcal{G}, q \in \mathcal{T}\} \tag{B-1}$$

such that for a universal constant C_L, all $g_j \in \mathcal{G}$, $\beta_j \in \Theta$, $j = 1, 2$, and all $q \in \mathcal{T}$,

$$|q(w|\beta_1, g_1) - q(w|\beta_2, g_2)| \leq C_L\{|\beta_1 - \beta_2| + \|g_1 - g_2\|_\infty\}. \tag{B-2}$$

Moreover, we assume that there exist a convex, bounded subset of $\mathbb{R}^2$, say $\mathcal{C}_\mathcal{W}$, with non-empty interior such that for each $g \in \mathcal{G}$, $\beta \in \Theta$, it holds that $q(\cdot|\beta, g) \in C^\eta(\mathcal{C}_\mathcal{W})$, for some $\eta > 1$.

Define the rates $p_n := \Pr(f(W_0|W_0) \leq 2\tau_n)$, $w_n := \|\widehat{g} - g_0\|_\infty$,

$$d_n := \sqrt{\frac{\log a_n^{-2} \vee \log\log n}{na_n^2}} + b_n^r,$$

and $q_n := \tau_n^{-1}d_n + w_n$. Use the short notation $\partial_\beta\widehat{F}_i \equiv \partial_\beta\widehat{F}(W_i(\widehat{\beta},\widehat{g})|\widehat{\beta},\widehat{g})$.

Assumption B.6 *(i) $F_0 \in \mathcal{T}^{\eta_F}$, (ii) For all $i = 1, ..., n$, $P(\widehat{F} \in \mathcal{T}^{\eta_F}) \to 1$ and $P(\partial_\beta\widehat{F} \in \mathcal{T}^{\eta_F}) \to 1$, for some $\eta_F > 1$.*

Assumption B.7 *τ_n is a sequence of positive numbers satisfying $\tau_n \to 0$, $n\tau_n^{-6}d_n^4 \to 0$ and $n(\tau_n^{-l}q_n^l + p_n^2) \to 0$, for some $l \geq 2$.*

Assumption B.6 is a high level condition and Escanciano, Jacho-Chávez and Lewbel (2014) provides sufficient low-level conditions for it to hold in various set-ups. Similarly, Assumption B.7 is used to handle the random trimming as in Escanciano, Jacho-Chávez and Lewbel (2014).

Assumption B.8 *The function $F_0(w_1, w_2|W_0)$ is twice continuously differentiable in w_2 with bounded derivatives, for all w_2.*

Assumption B.9 *(i) The regression g_0 is estimated by a NW kernel estimator with a kernel function satisfying Assumption B.4 with $r = \rho$ and a possibly stochastic bandwidth $\widehat{h}_{gn}$ satisfying $P(l_n \leq \widehat{h}_{gn} \leq u_n) \to 1$ as $n \to \infty$, for deterministic sequences of positive numbers l_n and u_n such that: $u_n \to 0$ and $nl_n^d / \log n \to \infty$; $nu_n^{2\rho} \to 0$; (ii) the function g_0 and the density $f_{X_1}(\cdot)$ of X_1 are ρ-times continuously differentiable in x, with bounded derivatives. The density $f_{X_1}(\cdot)$ is bounded away from zero. Furthermore $g_0 \in \mathcal{G} \subset C^{\eta_g}(\mathcal{X}_{X_1})$, $P(\widehat{g} \in \mathcal{G}) \to 1$ for some $\eta_g > 2$.*

Assumption B.10 *The matrix $\Delta_0 := E[\partial_\beta F_0(W_{0i}|W_0)\partial_\beta^\top F_0(W_{0i}|W_0)]$ is positive definite.*

Assumptions B.8 and B.9 are standard in the literature. Examples of random bandwidths that satisfy our assumptions are plug-in bandwidths of the form $\widehat{h}_{gn} = \widehat{c}h_{gn}$ with $\widehat{c}$ is bounded in probability and h_n a suitable deterministic sequence. Assumption B.9 has been verified to hold under primitive regularity assumptions for the local polynomial estimator in Neumeyer and van Keilegom (2010) when all the X_1's are continuous. See also Escanciano, Jacho-Chávez and Lewbel (2014). Finally, Assumption B.10 is also standard and it ensures the non-singularity of the asymptotic covariance matrix of the final estimator.

Table 1: Monte Carlo Results

	$\rho = 0$				$\rho = -1/2$				$\rho = -3/4$			
Est.	Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE
$n = 250$												
[1]	0.033	0.154	0.161	0.119	0.014	0.114	0.118	0.087	-0.343	0.063	0.337	0.331
[2]	-0.032	0.395	0.399	0.266	0.295	0.366	0.513	0.388	0.224	0.520	0.615	0.423
[3]	-0.025	0.460	0.492	0.334	0.004	0.438	0.438	0.267	-0.003	0.481	0.514	0.317
[4]	-0.015	0.476	0.502	0.344	0.004	0.501	0.501	0.294	-0.002	0.619	0.639	0.338
$n = 500$												
[1]	0.011	0.104	0.107	0.082	0.007	0.077	0.079	0.060	-0.333	0.040	0.332	0.329
[2]	-0.022	0.234	0.234	0.169	0.365	0.273	0.484	0.405	0.123	0.277	0.321	0.230
[3]	-0.008	0.348	0.420	0.267	0.000	0.315	0.333	0.173	-0.003	0.313	0.354	0.191
[4]	-0.009	0.363	0.432	0.282	0.000	0.347	0.357	0.197	-0.003	0.322	0.354	0.200
$n = 1000$												
[1]	0.009	0.072	0.073	0.057	-0.064	0.045	0.077	0.067	-0.372	0.024	0.371	0.371
[2]	-0.015	0.156	0.156	0.120	0.361	0.196	0.433	0.388	0.142	0.190	0.249	0.188
[3]	-0.004	0.329	0.413	0.254	0.000	0.262	0.285	0.124	-0.001	0.282	0.319	0.152
[4]	-0.005	0.328	0.412	0.256	0.000	0.264	0.282	0.125	-0.001	0.287	0.323	0.159

Note: Table displays Monte Carlo median bias (Bias), standard deviation (Std. Dev), root mean squared error (RMSE) and mean absolute error (MAE) of standard Probit [1], Ichimura's (1993) estimator [2], Infeasible 2-Steps SLS [3] and Feasible 2-Steps SLS [4].

Table 2: Descriptive Statistics

Name	Description	Mean	Std. Dev.	25% Q.	Median	75% Q.	Min.	Max.
Y	Dummy variable of whether and individual changes state of residence during 1991-1993 or not	0.174	0.379	0	0	0	0	1
Edu	Dummy variable of whether and individual has a college education or not	0.418	0.493	0	0	1	0	1
State	The number of US states individual ever lived in	2.18	1.478	1	2	3	1	8
Size	Family size	3.28	1.584	2	3	4	1	17
Age	Age in years	38.444	10.816	30	36	44	22	69
log(Income)	Natural logarithm of average annual labor income	10.012	0.931	9.651	10.178	10.601	4.068	11.68

Note: Sample consists of 4582 observations from the 1990 wave of the Panel Study of Income Dynamics (PSID). The sample consists of male households head who are not students and reported positive labor income between 1989-1990.

Table 3: Migration Binary Choice Model Estimation Results

Variable	[I]		[II]		[III]	
Edu	-0.2535	(0.5802)	-0.037	(0.0145)	1.5257	(0.2924)
Size	0.1397	(0.1737)	-0.0042	(0.0076)	0.5309	(0.1471)
Age	-0.1897	(0.0403)	-0.0084	(0.0009)	-0.3138	(0.0191)
log(Income)	-1.5803	(0.4089)	-0.1408	(0.0044)	-2.6427	(0.0892)
Const.	8.7084	(3.5356)				
σ^2	12.334	(2.2075)				
Bandwidths			0.0731		(0.9325,1.4135)	
Obj. Func.	-2063.0		0.1399		0.1398	
$\partial_{\partial_{w_1}} \widehat{F}$	0.2510	(0.0622)	0.4378	(0.1912)	0.0216	(0.0056)

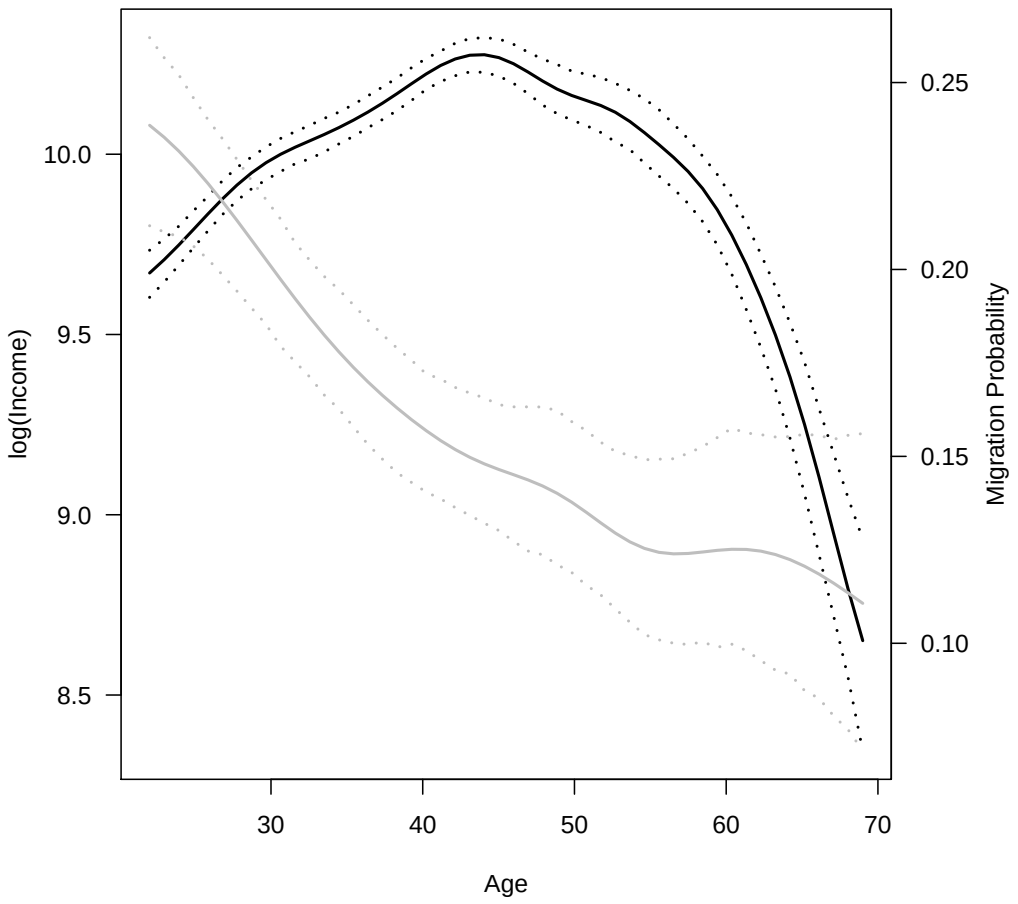
Note: Results for Probit [I], Ichimura's (1993) estimator [II], Feasible 2-Steps SLS [III]. Coefficient of State was normalized to 1 for all estimators. Asymptotic standard errors are in parenthesis. Nonparametric residuals in the first step were obtained using the generalized kernel estimator of Racine and Li (2004) with bandwidth chosen by Least-Squares cross-validation.

Table 4: Estimated Marginal Effects

Variable		[I]		[II]		[III]	
Edu*		-0.0052	(0.0119)	-0.0162	(0.0095)	0.0329	(0.0106)
	0 - 1	-0.0052	(0.0119)	-0.0166	(0.0296)	0.0355	(0.0114)
Size*		0.0028	(0.0035)	-0.0018	(0.0034)	0.0115	(0.0044)
	1 - 2	0.0028	(0.0033)	-0.0017	(0.0305)	0.0036	(0.0110)
	2 - 3	0.0028	(0.0034)	-0.0018	(0.0301)	0.0081	(0.0109)
	3 - 4	0.0028	(0.0035)	-0.0018	(0.0296)	0.0124	(0.0112)
Age*		-0.0039	(0.0005)	-0.0037	(0.0017)	-0.0067	(0.0018)
log(Income)*		-0.0322	(0.0060)	-0.0617	(0.0270)	-0.057	(0.0149)

Note: (*) marginal effects for variable j is calculated as $\partial_{w_1}\widehat{F} \times \widehat{\beta}_j$. All other marginal effects are calculated as increments in $\widehat{F}$ from changing the value of the discrete covariate from t to $t+1$ while keeping the remaining part of the index at their respective sample mean. Asymptotic standard errors are in parenthesis.

Figure 1: Nonparametric Age Effects on Labor Income and Migration Probabilities



Note: Nadaraya-Watson estimates of E[Y|Age] (solid gray line) and E[log(Income)|Age] (solid black line) with bandwidths chosen by Least-Squares cross-validation and 95% pointwise confidence intervals based on 399 bootstrapped replications (dotted lines).