Semiparametric Estimation of Binary Response Models with Endogenous Regressors

Christoph Rothe*

Department of Economics, University of Mannheim

First Version: October 8, 2007

This Version: April 27, 2009

Abstract

In this paper, we propose a two-step semiparametric maximum likelihood (SML) estimator for the coefficients of a single index binary choice model with endogenous regressors when identification is achieved via a control function approach. The first step consists of estimating a reduced form equation for the endogenous regressors and extracting the corresponding residuals. In the second step, the latter are added as control variates to the outcome equation, which is in turn estimated by SML. We establish the estimator's $\sqrt{n}$-consistency and asymptotic normality. In a simulation study, we compare the properties of our estimator with those of existing alternatives, highlighting the advantages of our approach.

JEL Classification: C14, C31, C35,

Keywords: *Binary Choice Model, Semiparametric Maximum Likelihood, Endogenous Regressors, Instrumental Variables, Control Function*

*Address: Department of Economics, University of Mannheim, D-68131 Mannheim, Germany. E-Mail: crothe@rumms.uni-mannheim.de. Website: http://webrum.uni-mannheim.de/vwl/crothe. This paper is part of my PhD thesis at the University of Mannheim. I am grateful to Richard Blundell, Xiaohong Chen, Andrew Chesher, Stefan Hoderlein, Arthur Lewbel, Oliver Linton, Enno Mammen, Whitney Newey, an Associate Editor, three anonymous referees and seminar participants at Bielefeld, Kaiserslautern, Mannheim, the EEA-ESEM 2007 meeting in Budapest and the workshop on "Semiparametric and Nonparametric Methods in Econometrics" in Oberwolfach for numerous helpful comments. The usual disclaimer applies.

1 Introduction

This paper is concerned with the semiparametric estimation of the coefficients of a single index binary response model with endogenous regressors when identification is achieved via the control function approach put forward by Blundell and Powell (2004). The type of model we consider is of the form

$$Y = \begin{cases} 1 & \text{if } Y^* = X'\theta_o - U > 0 \\ 0 & \text{else,} \end{cases}$$

where Y is an indicator of the sign of a latent variable Y^* generated through a linear model with regressors X, vector of parameters θ_o and error term U. Our interest is in the estimation of the (normalized) coefficients θ_o, which is a semiparametric problem in the sense that the distribution of the unobservable variables is not assumed to belong to some parametric family. Furthermore, we do not assume that the error is independent of the regressors since we want to allow some components of X to be endogenous and thus correlated with U. To account for endogeneity, a control function approach introduces additional control variables, such as residuals from a reduced form of the endogenous variables for example, as covariates into the outcome equation. Within this class of models, the only estimator that has been suggested so far is the one proposed by Blundell and Powell (2004), which is an extension of the Ahn, Ichimura, and Powell (1996) "matching" estimator.

This paper contributes to the literature by proposing a new two-step semiparametric maximum likelihood (SML) estimator. The procedure, which is also suggested but not further developed in Blundell and Powell (2004), is an extension of the Klein and Spady (1993) estimator, which achieves the semiparametric efficiency bound in the exogenous case. The first step consists of estimating the control variables through an auxiliary regression, which can either be fully nonparametric, or incorporate some parametric restrictions. In the second step, these are added nonparametrically to the equation of interest, which is in turn estimated by semiparametric maximum likelihood. Compared with the Blundell-Powell estimator, our procedure exploits the restrictions implied by the model more effectively, and does not require high-dimensional smoothing. The estimator possesses the classic asymptotic properties of $\sqrt{n}$-consistency and asymptotic normality, and valid standard errors and test statistics can be obtained via a nonparametric bootstrap procedure. Through a simulation study, we show that using our SML approach yields a considerable gain in terms of finite sample performance over other existing semiparametric estimators for binary choice models with endogenous regressors in many

empirically relevant settings. The procedure should thus be appealing to applied researchers.

Binary response models play a prominent role in microeconometrics and are therefore the focus of an extensive literature. Estimation is typically carried out using standard Logit or Probit procedures, assuming that the distribution of the error term follows some parametric law and that X and U are independent. Having an estimator like ours that relies on neither of these two assumptions is of considerable practical importance since both might be inappropriate for many empirical applications.

First, economic theory usually provides no guidance about the functional form of the distribution of the error term, but misspecifications will generally result in inconsistent estimates for likelihood-based approaches. A number of semiparametric estimators have therefore been proposed which do not impose parametric restrictions on the distribution of U. Such estimators include Semiparametric Least Squares (Ichimura 1993), Semiparametric Maximum Likelihood (Klein and Spady (1993), Ai (1997)), Average Derivative estimators (Stoker (1986), Powell, Stock, and Stoker (1989)), the Maximum Score estimator (Manski 1975) and the semiparametric estimator for discrete regressors of Horowitz and Härdle (1996), to mention a few.

Second, when the binary choice model arises in the context of a system of triangular or fully simultaneous equations, or certain measurement error models, some components of X will typically be endogenous, violating the independence assumption. Although neglecting this problem will again render the usual estimates inconsistent, it has received much less attention in literature. If one has access to an instrumental variable, an *ad-hoc* solution often recommended in econometrics textbooks would be to estimate a linear probability model by two-stage least squares (2SLS), although this procedure is generally inconsistent and might imply choice probabilities that are not between 0 and 1. More adequate estimators that are widely used have been proposed by Smith and Blundell (1986), Rivers and Vuong (1988) and Newey (1987), but they require fairly strong parametric distributional assumptions.

A semiparametric way of recovering the index coefficients that does not assume the unobservables to follow any parametric law is provided by Newey (1985). The approach requires a correctly specified parametric reduced form with homoskedastic error terms, where in particular the latter condition can be restrictive in practice. More recently, Lewbel (2000) proposed a simple to implement semiparametric procedure for estimating θ_o when X contains a continuously distributed, strictly exogenous "special regressor" that satisfies a large support condition. While this approach has the advantage that it allows the endogenous variable to be discrete or even binary, in many applications there might be no exogenous variable which qualifies as a

"special regressor".

The control function approach that we use in this paper was proposed by Blundell and Powell (2004). The general idea of using residuals from a reduced form of the regressors to account for endogeneity is well established in parametric econometrics and has recently been used in the identification and estimation of various non- and semiparametric models with endogenous regressors (e.g. Newey, Powell, and Vella (1999), Blundell and Powell (2003), Chesher (2003), Das, Newey, and Vella (2003), Florens, Heckman, Meghir, and Vytlacil (2008), Imbens and Newey (2009), Blundell and Powell (2007), Lee (2007)). It has the drawback that it requires the endogenous regressor to be continuously distributed, but other variables, including the instruments, can well be discrete.

The plan of this paper is as follows. In the next section, we specify the model being used. In Section 3, we show how identification is achieved and describe our SML approach to estimation. Asymptotic properties of our estimator are analyzed in Section 4. In Section 5, we discuss a number of extensions of our setup, while Section 6 deals with implementation issues and presents the results of our simulation study. The application of our procedure is illustrated via an empirical example in Section 7. Finally, Section 8 concludes.

2 The Model

The setup we consider in this paper is a linear single-index binary response model with an arbitrary large number of endogenous regressors, similar to the one of Blundell and Powell (2004). It is given by:

$$Y \quad = \quad \mathbb{I}\{X'\theta_o - U \geq 0\}, \tag{2.1}$$

where Y is the binary dependent variable, X is d_x-dimensional vector of regressors, U is an unobserved random error term, and $\mathbb{I}\{A\}$ is the indicator function that equals 1 when A is true and 0 otherwise. Furthermore, there is a d_e-dimensional subvector X^e of X that contains the endogenous variables, in the sense that these are potentially correlated with U. We think of (2.1) as a structural equation, describing the causal relationship between the right-hand and left-hand side variables, and refer to it in the following as the outcome equation.

Since it is clear from the exogenous case that we can only hope to identify the index coefficients θ_o up to a multiplicative constant, we normalize the coefficient on the first component of X to unity, i.e. we assume that $\theta_o = (1, \beta_o)$.[1] The object of interest that we want to estimate

[1]This choice is of course totally arbitrary. In general, we could normalize the coefficient on any of the regressors

is the remaining vector of coefficients β_o. Also, for notational convenience, we use $X\beta_o$ as a shorthand for $(1, \beta_o)'X$.

Without making further assumptions, it is generally not possible to identify β_o in (2.1). To this end, we assume the existence of a control variable. That is, we assume that U and X are independent conditional on some (unobserved) random d_v-vector V, that can be written as an identified function of X^e and some vector of exogenous instruments Z, which may include some of the exogenous components of X:

$$U \perp X|V \text{ for some } V = v_o(X^e, Z). \tag{2.2}$$

Such a control variable can be available under various circumstances, but the specific source is not important for the construction and analysis of our estimator. We only require that the function v_o is identified and can be estimated by some $\hat{v}$ satisfying a "high-level" condition given below, which can be easily verified under very general circumstances.

The leading case in which such a control variable will typically be available is when the endogenous regressors are generated through a second equation as

$$X^e = m_o(Z) + V, \quad \mathbb{E}(V|Z) = 0, \tag{2.3}$$

where m_o is a conditional mean function. This function can either be left unspecified, in which case (2.3) is the standard nonparametric regression model, or assumed to satisfy some parametric or semiparametric restrictions. For example, it is possible to specify (2.3) as a single-index model, with $m_o(Z) = \tilde{m}_o(Z'\alpha_o)$ for some unknown function $\tilde{m}_o$ and an unknown vector of parameters α_o, or as a fully parametric nonlinear regression model, with $m_o(Z) = \tilde{m}(Z, \alpha_o)$ for a function $\tilde{m}$ that is known up to a finite dimensional parameter α_o.

It has been shown by Blundell and Powell (2004) that under the distributional exclusion restriction that

$$\Pr(U < c|X, Z) = \Pr(U < c|V), \tag{2.4}$$

for all c, the error term $V = X^e - m_o(Z) \equiv v_o(X^e, Z)$ is a control variable that satisfies condition (2.2). This restriction is more flexible than a "full independence" condition like $(U, V) \perp Z$, since it allows for example the variance of V to be a function of the instruments. However, it retains the general drawback of the control function approach that one has to correctly specify the relevant instrumental variables Z in (2.3), and that the endogenous regressor has to be

as long we can be sure that its true value is different from zero.

continuous, since otherwise the distribution of V and thus its relation with U will in general depend upon Z, which violates (2.4).

A specification like (2.3)–(2.4) is plausible in a number of contexts. For example, equations (2.1) and (2.3) could be seen as a triangular system of structural equations, with (2.3) describing the causal mechanism that determines the values of the endogenous regressor. Alternatively, such a specification could also arise when the latent variable Y^* and X^e are jointly determined through a system of simultaneous equations. In this case, equation (2.3) would be a reduced form equation resulting from an equilibrium condition. Another option would be a classical measurement error framework such as

$$\begin{aligned} Y &= \mathbb{I}\{\tilde{X}^{e'}\theta_{o1} + Z_1'\theta_{o2} - \epsilon_1 \geq 0\} \\ X^e &= \tilde{X}^e + \epsilon_2 \\ \tilde{X}^e &= m_o(Z) + \epsilon_3, \end{aligned}$$

where X^e is a noisy version of the unobserved regressor $\tilde{X}^e$ measured with error ϵ_2. This model is equivalent to (2.1) and (2.3) with $U = \epsilon_1 + \epsilon_2$ and $V = \epsilon_2 + \epsilon_3$.

While in this paper we will focus on control variables emerging from a structure like the one in (2.3), they might also appear under different circumstances, as pointed out by Imbens and Newey (2009). For example, as shown in Newey (2007), in a sample selection model where Y is only observed conditional on a selection variable $S = \mathbb{I}\{m(Z) > U^*\}$ being equal to one, and (U, U^*) is independent of Z, the selection probability $P = \Pr(S = 1|Z)$ is a control variable in the sense of condition (2.2). Such models can hence be treated in our framework as well.

3 Identification and Estimation Approach

3.1 Identification

The most important consequence of the restriction (2.2) is that the conditional expectation of the dependent variable Y given the observable variables X and V can be written as a function of the linear index $X\beta_o$ and the control variables V. Denoting the conditional distribution function of U given V by G_o, we can write

$$\mathbb{E}(Y|X,V) = \mathbb{E}(\mathbb{I}\{U \leq X\beta_o\}|X,V) = \mathbb{E}(\mathbb{I}\{U \leq X\beta_o\}|V) = G_o(X\beta_o, V), \tag{3.1}$$

and thus reduce the dimension from $d_x + d_v$ to $1 + d_v$.

This restriction is also useful for identifying β_o. In particular, it is clear that our parameter of interest is identified by the data if the following condition holds:

Identification Condition (IC). *There exists a unique interior point $\beta_o \in \mathcal{B}$ such that the relationship $\mathbb{E}(Y|X,V) = \mathbb{E}(Y|X\beta_o, V)$ holds for $(X,Z) \in \mathcal{A}$, a set with positive probability.*

Thus, what remains to establish identification of β_o is to give conditions on the primitives of the model under which IC is fulfilled. It turns out that for this purpose, in addition to requiring that v_o is identified, only the standard regularity conditions for identification of single-index binary response models are needed. The reason is that we are not dealing with an actual multiple-index model: although the function G_o has $1+d_v$ arguments, only the first one contains index parameters to be identified. We therefore have the following theorem.

Theorem 1 (Identification)**.** *The parameter β_o in the model (2.1)–(2.2) is identified in the sense that the identification condition IC holds, if the following conditions are satisfied:*

i) The function G_o is differentiable and strictly increasing in its first argument on a set $\mathcal{A}$ with positive probability under the distribution of X.

ii) Conditional on the control variable V, the vector X contains at least one continuously distributed component $X^{(1)}$ with nonzero coefficient.

iii) The span of the remaining components $X^{(-1)}$ contains no proper linear subspace which has probability 1 under the distribution of X.

The proof, which is analogous to the argument in Manski (1988), is given in the appendix. Note that when the control variables emerge from a structure like (2.3)–(2.4), the fact that the endogenous regressors are continuously distributed is not sufficient for condition (ii) to be fulfilled. Instead, it is required that additionally either one of the exogenous regressors or the "fitted value" $m_o(Z)$ from the reduced form is continuously distributed as well[2]. To see this, assume that all exogenous regressors and instruments are discrete. Then $X = (X^e, X^{(-e)}) = (m_o(Z) + V, X^{(-e)})$ is discretely distributed conditional on V, which violates condition (ii).

3.2 The Estimator

To motivate the estimator, assume for the moment that the function G_o was known and that V was observable. If observations are stochastically independent, it would then be straightforward to estimate β_o by maximizing the log-likelihood function

$$\frac{1}{n}\sum_{i=1}^{n} Y_i \log(G_o(X_i\beta, V_i)) + (1 - Y_i)\log(1 - G_o(X_i\beta, V_i)) \tag{3.2}$$

[2]I would like to thank an anonymous referee for pointing this out.

with respect to β. When G_o and V are unknown, this approach is clearly not feasible. However, generalizing the idea of Klein and Spady (1993), we can approximate the objective function by replacing all unknown quantities with appropriate estimates.

To make this idea more precise, we have to introduce some notation. For any candidate value of β and some function v, define $W(\beta, v) = (X\beta, v(X^e, Z))$, and set

$$G(w|\beta, v) = \mathbb{E}(Y|W(\beta, v) = w).$$

Furthermore, we use the convention that arguments indexing a function are dropped when they are evaluated at their true value, i.e. $G(w|\beta) = G(w|\beta, v_o)$, $G(w) = G(w|\beta_o)$, $W(\beta) = W(\beta, v_o)$, $W = W(\beta_o)$ etc. Using this notation, we have that $G_o(X\beta_o, V) = G(W(\beta_o, v_o)|\beta_o, v_o) \equiv G(W)$. The idea is now to replace the term $G_o(X_i\beta, V_i)$ in (3.2) by a nonparametric kernel estimate $\hat{G}(\hat{W}_i(\beta)|\beta, \hat{v})$, where $\hat{W}(\beta) = W(\beta, \hat{v})$ and $\hat{v}$ is itself a (possibly nonparametric) estimate of v_o from a preliminary estimation stage. Note that the function $G(W(\beta, v)|\beta, v)$ and its estimate depend on β both through its first argument, which determines the point of evaluation, and its second one, which influences the shape of the function.

Since we have made no assumptions about the structural form generating the control variates, we also do not impose a specific estimation procedure. Instead, we simply assume the existence of an estimator $\hat{v}$ of v_o satisfying some high-level conditions given below. Then for any β and v, a nonparametric kernel estimate of $G(\cdot|\beta, v)$ can be obtained as

$$\hat{G}(w|\beta, v) = \hat{N}(w|\beta, v)/\hat{D}(w|\beta, v)$$

where

$$\begin{aligned} \hat{N}(w|\beta, v) &= \frac{1}{n}\sum_{j=1}^{n} K_h\left(W_i(\beta, v) - w\right) Y_j, \\ \hat{D}(w|\beta, v) &= \frac{1}{n}\sum_{j=1}^{n} K_h\left(W_i(\beta, v) - w\right). \end{aligned}$$

Here $K_h(\cdot) = K(\cdot/h)/h$ is a kernel function on $\mathbb{R}^{1+d_v}$ and h is a bandwidth sequence that goes to zero as n goes to infinity. The exact specifications are given below. Substituting this estimate into equation (3.2) we obtain the semiparametric likelihood function

$$L_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau_i(Y_i \log(\hat{G}(\hat{W}_i(\beta)|\beta, \hat{v})) + (1 - Y_i)\log(1 - \hat{G}(\hat{W}_i(\beta)|\beta, \hat{v}))),$$

and define our estimator $\hat{\beta}$ of β_o as the maximizer of this objective function:

$$\hat{\beta} = \underset{\beta \in \mathcal{B}}{\operatorname{argmax}}\, L_n(\beta). \tag{3.3}$$

Here $\tau_i = \mathbb{I}\{(X_i, Z_i) \in \mathcal{X}\}$ is a trimming term that equals 1 whenever the values of (X_i, Z_i) lie within an appropriately chosen compact set $\mathcal{X}$ and 0 otherwise. In particular, the set is chosen such that the probability limit of $\hat{G}$ is bounded away from zero and one on $\mathcal{X}$.

While the maximization in (3.3) can be carried out using standard numerical optimization procedures, it is certainly computationally expensive, since we have to run n nonparametric regressions for *every* iteration step. A further complication is the possible presence of local maxima in the objective function. We discuss these issues in more detail in the simulation study.

4 Asymptotic Properties

In this section, we establish the asymptotic properties of our estimator. We start with stating the assumptions and then give results on consistency, asymptotic normality and variance estimation. Here we only sketch our proofs and delegate rigorous arguments to the Appendix.

4.1 Assumptions and Preliminaries

Before we present our framework, we have to introduce some more notation. For μ a k-vector of nonnegative integers, we define (i) $|\mu| = \sum_{i=1}^{n} \mu_i$, (ii) for any function $f(x)$ on $\mathbb{R}^k$, $\partial_x^\mu f(x) = \partial^{|\mu|}/(\partial^{\mu_1} x_1, \ldots, \partial^{\mu_k} x_k) f(x)$ and (iii) $x^\mu = \prod_{i=1}^{n} x_i^{\mu_i}$. Furthermore, we write ∂_k as a shorthand for ∂_{w_k} for $k = 1, 2$. We can now state the assumptions for our asymptotic analysis.

Assumption 1. *The sample observations $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ are a sequence of independent and identically distributed random vectors generated according to the model defined in equation* (2.1) – (2.2)*. The model is identified in the sense that IC holds.*

Assumption 2. *The parameter space $\mathcal{B}$ is a compact subset of $\mathbb{R}^{d_x - 1}$ and β_o is an element of its interior.*

These are standard regularity conditions in the semiparametrics literature.

Assumption 3. *i) For all $\beta \in \mathcal{B}$, the distribution of the random vector $W(\beta)$ admits a density function $D(w|\beta)$ with respect to the Lebesque measure.*

ii) For all $\beta \in \mathcal{B}$, $D(w|\beta)$ is r times continuously differentiable in w, and the derivatives are uniformly bounded: $\sup_{w,\beta} |\partial_w^\mu D(w|\beta)| < \infty \; \forall \mu$ with $|\mu| \leq r$.

iii) For all $\beta \in \mathcal{B}$, $G(w|\beta)$ is r times continuously differentiable in w, and the derivatives are uniformly bounded: $\sup_{w,\beta} |\partial_w^\mu G(w|\beta)| < \infty \ \forall \mu$ *with* $|\mu| \leq r$.

iv) $D(w|\beta)$ and $G(w|\beta)$ are twice continuously differentiable in β.

Assumption 3 collects some conventional smoothness restrictions on the functions being estimated through kernel methods. The higher-order differentiability conditions are needed to obtain certain uniform convergence rates on the estimates of $G(\cdot|\beta)$ and its derivatives.

Assumption 4. *For $\mathcal{X}$ a compact subset of the support of (X, Z), define $W(\mathcal{X}) = \{w \in \mathbb{R}^{1+d_v} : \exists (x, z) \in \mathcal{X}, \beta \in \mathcal{B}$ s.t. $w = (x\beta, v_o(x^e, z))\}$. Then $\mathcal{X}$ is chosen such that:*

i) $\inf_{w \in W(\mathcal{X}), \beta \in \mathcal{B}} D(w|\beta) > 0$

ii) $\inf_{w \in W(\mathcal{X}), \beta \in \mathcal{B}} G(w|\beta) > 0$ *and* $\sup_{w \in W(\mathcal{X}), \beta \in \mathcal{B}} G(w|\beta) < 1$.

Assumption 4 prescribes a fixed trimming procedure, which significantly simplifies the derivation of the asymptotic properties. Since trimming is generally considered to be of minor practical importance and thus is often disregarded in empirical applications, this seems to be a mild restriction. However, at the cost of a more complicated proof it would be possible to replace the fixed trimming function $\tau_i = \mathbb{I}\{(X_i, Z_i) \in \mathcal{X}\}$ with a random, data dependent one that tends to one as the sample size increases. Using results from e.g. Pakes and Pollard (1989), one could for example implement a trimming procedure on the basis of the upper and lower sample quantiles of the data, as in Lee (1995).

Assumption 5. *The matrix*

$$\Sigma = \mathbb{E}\left[\tau \partial_\beta G(W) \partial_{\beta'} G(W) \left(G(W)(1 - G(W))\right)^{-1}\right]$$

is positive definite.

Assumption 5 ensures the non-singularity of the asymptotic covariance matrix of our final estimator. Note that here and in the following the notation $\partial_\beta G(W)$ is understood to denote the derivative of $G(W(\beta)|\beta)$ with respect to both occurrences of β, evaluated at $\beta = \beta_o$.

Assumption 6. *The kernel functions $K : \mathbb{R}^{d_v+1} \to \mathbb{R}$ satisfies (i) $\int K(z)dz = 1$, (ii) $\int K(z)z^\mu dz = 0$ for all $|\mu| = 1, \ldots, r-1$, (iii) $\int |K(z)z^\mu| dz < \infty$ for $|\mu| = r$, (iv) $K(z) = 0$ if $|z| > 1$ (v) $K(z)$ is r times continuously differentiable.*

Assumption 7. *The bandwidth vector $h = (h_1, \ldots, h_{d_v+1})$ satisfies $h_i = c_i n^{-\delta}$, $i = 1, \ldots, d_v+1$, for some constants $c_i > 0$ and δ such that $1/2r < \delta_i < 1/(2 + 6d_v)$.*

The last two assumptions define a standard bias-reducing kernel of order r, which is used for reducing asymptotic bias in the estimates of G and its derivatives, and determine the rate at which the bandwidth sequences go to zero as $n \to \infty$. In order to ensure that the set of possible values for δ is not empty, a sufficient condition is that $r > 1 + 3d_v$.

Assumption 8. *i) The estimate $\hat{v}$ of v_o satisfies*

$$\hat{v}(X_i^e, Z_i) - v_o(X_i^e, Z_i) \equiv \hat{V}_i - V_i = \frac{1}{n}\sum_{j=1}^{n} \omega_n(Z_i, Z_j)\psi_j + r_{in},$$

with

$$\max_i \tau_i \|r_{in}\| = o_p(n^{-1/2}) \quad \text{and} \quad \max_i \tau_i \|\hat{V}_i - V_i\| = o_p(n^{-1/4}),$$

where $\psi_j = \psi(X_j^e, Z_j)$ is an influence function with $\mathbb{E}(\psi_j|Z_j) = 0$ and $\mathbb{E}(\psi_j^2|Z_j) < \infty$, and the weights $\omega_n(Z_i, Z_j)$ satisfy $\mathbb{E}(\|\omega_n(Z_i, Z_j)\|^2) = o(n)$.

ii) There exists a space $\mathcal{V}$, such that $\Pr(\hat{v} \in \mathcal{V}) \to 1$, and that $\int_0^\infty \sqrt{\log N(\lambda, \mathcal{V}, \|\cdot\|_\infty)}d\lambda < \infty$, where $N(\lambda, \mathcal{V}, \|\cdot\|_\infty)$ is the covering number with respect to the L_∞-norm of the class of functions $\mathcal{V}$, i.e. the minimal number of balls with $\|\cdot\|_\infty$-radius λ needed to cover $\mathcal{V}$.

This assumption is a high-level condition on the estimator of the control variables. The first part states that the estimator admits a certain asymptotic expansion, whereas the second part requires the estimator to take values in some well-behaved function space with probability approaching 1.

These conditions can be shown to be fulfilled for various scenarios discussed in Section 2. For example, assume that $X^e = m_o(Z) + V$ with $E(V|Z) = 0$, $\hat{V}_i = \hat{v}(X_i^e, Z_i) = X_i^e - \hat{m}(Z_i)$, $\hat{m}$ is the usual Nadaraya-Watson estimator, and $\mathcal{V}$ is the class of all functions f taking the form $f(x^e, z) = x^e - g(z)$ for some function g whose partial derivatives up to order p exist and are uniformly bounded. Then, under certain assumptions on the kernel and the bandwidth, the first part of Assumption 8 is fulfilled with

$$\omega_n(Z_i, Z_j) = \kappa_b(Z_i - Z_j)f_Z(Z_i)^{-1} \text{ and } \psi_j = -(X_j^e - m_o(Z_j)) = -V_j,$$

where f_Z is the density function of the vector of instruments Z, κ is a kernel function and b is the bandwidth. Also, $\Pr(\hat{v} \in \mathcal{V}) \to 1$ in this case if the kernel function has uniformly bounded partial derivatives up to order p. The remaining requirement then follows from Corollary 2.7.4 in van der Vaart and Wellner (1996) if $p > d_z/2$. Similar arguments can also be used when m_o is specified in a semiparametric way, for example as a single-index or partially linear model,

or when other nonparametric smoothers, such as local polynomials are used (see e.g. Kong, Linton, and Xia (2009)).

On the other hand, when $m_o(z) = m(z, \alpha_o)$ is known up to some vector of parameters α_o, under standard regularity conditions for nonlinear regression models we obtain that part (i) is fulfilled with

$$\omega_n(Z_i, Z_j) = \partial_\alpha m(Z_i, \alpha_o)\mathbb{E}(\partial_\alpha m(Z, \alpha_o)\partial_\alpha m(Z, \alpha_o)')^{-1}\partial_\alpha m(Z_j, \alpha_o)' \text{ and } \psi_j = -V_j,$$

whereas part (ii) is true when m satisfies a Lipschitz conditions with respect to α, as shown van der Vaart and Wellner (1996, Theorem 2.7.11).

4.2 Consistency and Asymptotic Normality

To establish consistency, we take the usual route and first show that the estimated likelihood function $L_n(\beta)$ converges uniformly to a nonrandom limit function $L(\beta)$. Secondly, we show that this function attains a unique maximum at β_o, which implies both that β_o is identified and that $\hat{\beta}$ is consistent. This is formally stated in the following theorem:

Theorem 2 (Consistency)**.** *Under Assumptions 1 – 8, it holds that* $\hat{\beta} = \beta_o + o_p(1)$ *as* $n \to \infty$.

Showing that $\hat{\beta}$ is also asymptotically normal requires a somewhat more involved argument. Our strategy is to use general results on semiparametric estimation procedures given in Chen, Linton, and Van Keilegom (2003). As shown in the Appendix, this requires deriving uniform rates of convergence for the nonparametric estimates of the link function $G(\cdot|\beta)$ and its derivatives. This constitutes the main difficulty for the proof, since the estimates of $G(\cdot|\beta)$ are in turn based on possibly non- or semiparametrically generated regressors $\hat{V}$.

Intuitively, the asymptotic normality result follows from the following argument. From a standard Taylor expansion of the semiparametric score function $S_n(\beta) = \partial_\beta L_n(\beta)$ around the true parameter values β_o we obtain, after rearranging terms,

$$\sqrt{n}(\hat{\beta} - \beta_o) = -(\partial_{\beta,\beta} L_n(\bar{\beta}))^{-1}\sqrt{n}\partial_\beta L_n(\beta_o), \tag{4.1}$$

where $\bar{\beta}$ is some intermediate value between $\hat{\beta}$ and β_o. Starting with the first term on the right-hand-side of (4.1), it follows from the uniform convergence results on $\hat{G}(\cdot|\beta, \hat{v})$ and its derivatives, and the consistency of $\hat{\beta}$ and $\hat{v}$, that it converges in probability to some matrix, i.e.

$$\partial_{\beta\beta} L_n(\bar{\beta}) \xrightarrow{p} \Sigma,$$

where the limit is positive definite by Assumption 5. Continuing with the second term in (4.1), it is shown in the Appendix that

$$\begin{aligned}\sqrt{n}\partial_\beta L_n(\beta_o) = &\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{Y_i - G(W_i)}{G(W_i)(1-G(W_i))}(\tau_i\partial_\beta G(W_i) - \mathbb{E}(\tau_i\partial_\beta G(W_i)|W_i)) \\ &+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\xi_{1i} - \xi_{2i})\psi_i + o_p(1),\end{aligned}$$

where ψ_i is the influence function from Assumption 8, and

$$\begin{aligned}\xi_{1i} &= \mathbb{E}(\tau\partial_\beta G(W)\partial_2 G(W)(G(W)(1-G(W)))^{-1}\omega_n(Z,Z_i)|Z_i), \\ \xi_{2i} &= \mathbb{E}(\mathbb{E}(\tau\partial_\beta G(W)|W)\partial_2 G(W)(G(W)(1-G(W)))^{-1}\omega_n(Z,Z_i)|Z_i).\end{aligned}$$

Taken together, and applying a Central Limit Theorem, we obtain the following result:

Theorem 3 (Asymptotic Normality)**.** *Under Assumptions 1–8, it holds that*

$$\sqrt{n}(\hat{\beta} - \beta_o) \xrightarrow{d} N(0, \Omega)$$

where

$$\Omega = \Sigma^{-1}(\Psi_1 + \Psi_2)\Sigma^{-1}$$

and

$$\begin{aligned}\Sigma &= \mathbb{E}\left[\frac{\tau\partial_\beta G(W_i)\partial_{\beta'} G(W_i)}{G(W_i)(1-G(W_i))}\right], \\ \Psi_1 &= \mathbb{E}\left[\frac{(\tau\partial_\beta G(W) - \mathbb{E}(\tau\partial_\beta G(W)|W))(\tau\partial_\beta G(W) - \mathbb{E}(\tau\partial_\beta G(W)|W))'}{G(W_i)(1-G(W_i))}\right] \\ \Psi_2 &= \mathbb{E}\left[(\xi_{1i} - \xi_{2i})\psi_i\psi_i'(\xi_{1i} - \xi_{2i})'\right].\end{aligned}$$

It is instructive to compare our asymptotic variance matrix to that of an infeasible maximum likelihood estimator using the true functions $G(\cdot|\beta)$ and v_o. If we define $\tilde{\Sigma}$ be equal to Σ with $\tau \equiv 1$, the variance of such an estimator would be $\tilde{\Sigma}^{-1}$. In general, our matrix Ω will be larger for two reasons. First, due to the fixed trimming procedure our estimator does not use all available observations, which obviously results in a loss of efficiency. Second, there is an additional penalty in terms of asymptotic variance for only using an estimate of the function v_o. However, there is no penalty term for estimating the unknown link function $G(\cdot|\beta)$, which is also the case when all regressors are exogenous.

To see this, let $\tilde{\Omega}$ be equal to Ω with $\tau \equiv 1$, and define $\tilde{\Psi}_1$, $\tilde{\Psi}_2$, $\tilde{\xi}_{1i}$ and $\tilde{\xi}_{2i}$ analogously. Then it follows from the fact that $\mathbb{E}(\partial_\beta G(W)|W) = 0$ (see Klein and Spady (1993, p. 403))

that $\tilde{\Sigma} = \tilde{\Psi}_1$ and $\tilde{\Psi}_2 = \mathbb{E}[\tilde{\xi}_{1i}\psi_i\psi_i'\tilde{\xi}_{1i}']$. Thus, if we neglect the effect of trimming, the asymptotic covariance matrix of our estimator would be $\tilde{\Omega} = \tilde{\Sigma}^{-1} + \tilde{\Sigma}^{-1}\tilde{\Psi}_2\tilde{\Sigma}^{-1}$, where the presence of the second term $\tilde{\Sigma}^{-1}\tilde{\Psi}_2\tilde{\Sigma}^{-1}$ is due to using an estimate of v_o. Since this term is generally nonnegative definite, the variance will be larger than it would be if v_o was known and thus the control variable V was observed.

4.3 Variance estimation

In order to be able to conduct inference on $\hat{\beta}$, an estimate of the asymptotic variance matrix is needed, but since Ω depends on a number of unknown functions in a relatively complicated way, a direct sample moment estimator would be hard to implement. However, the results in Chen, Linton, and Van Keilegom (2003) justify the use of an ordinary nonparametric bootstrap procedure to calculate confidence regions for the unknown parameters. To be specific, let $\{(Y_i^*, X_i^*, Z_i^*)\}_{i=1}^n$ be the bootstrap sample drawn randomly with replacement from the original data $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, and let $\hat{v}^*$ and $\hat{G}^*(\cdot|\beta, v)$ be the same estimators as $\hat{v}$ and $\hat{G}(\cdot|\beta, v)$ but based on the bootstrap data. Also, define the bootstrap estimator $\hat{\beta}^*$ as

$$\hat{\beta}^* = \underset{\beta\in\mathcal{B}}{\operatorname{argmax}} \frac{1}{n}\sum_{i=1}^n \tau_i^*(Y_i^* \log(\hat{G}^*(W_i(\beta, \hat{v}^*)|\beta, \hat{v}^*)) + (1 - Y_i^*)\log(1 - \hat{G}^*(W_i(\beta, \hat{v}^*)|\beta, \hat{v}^*))).$$

Then it can be shown using Theorem B in Chen, Linton, and Van Keilegom (2003) and similar arguments as in the proof of our Theorem 3, that $\sqrt{n}(\hat{\beta}^* - \hat{\beta})$ has the same asymptotic limiting distribution as $\sqrt{n}(\hat{\beta} - \beta_o)$.

A general disadvantage of using such resampling techniques for a semiparametric optimization estimator like ours is that they can be extremely costly from a computational point of view. For practical applications, the following approximation might thus be useful. Note that the complicated functional form of Ω is mainly an effect of the fixed trimming procedure. Yet when only a small amount of observations is trimmed, this effect should be small. In particular, when $\tau = 1$ for most observations, then $\mathbb{E}(\tau\partial_\beta G(W)|W)) \approx 0$ and by continuity the matrix Ω can be well approximated by $\bar{\Omega} = \Sigma^{-1} + \Sigma^{-1}\bar{\Psi}_2\Sigma^{-1}$, where $\bar{\Psi}_2 = \mathbb{E}[\xi_{1i}\psi_i\psi_i'\xi_{1i}']$. Under our assumptions stated above, the matrix Σ can be consistently estimated by

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \tau_i \frac{\partial_\beta \hat{G}(W_i(\hat{\beta}|\hat{\beta}, \hat{v})\partial_{\beta'}\hat{G}(W_i(\hat{\beta})|\hat{\beta}, \hat{v})}{\hat{G}(W_i(\hat{\beta}|\hat{\beta}, \hat{v})(1 - \hat{G}(W_i(\hat{\beta})|\hat{\beta}, \hat{v}))}.$$

Estimating the matrix $\bar{\Psi}_2$ is more difficult when using only the "high-level" condition on the control function from Assumption 8. However, when imposing more structure on the estimates of the control variables, the shape of the terms ξ_i and ψ_i can usually be made more explicit,

and thus suggest a potential estimator. Consider for example the case where $\hat{V}_i = X_i^e - \hat{m}(Z_i)$ is the residual from a nonparametric reduced for equation estimated by some kernel method, as in (2.3). Then $\psi_i = -V_i$, and it is easy to show that $\xi_{1i} = \mathbb{E}(\tau_i \partial_\beta G(W_i) \partial_2 G(W_i)(G(W_i)(1 - G(W_i)))^{-1} | Z_i)$. Accordingly, one could estimate $\bar{\Psi}_2$ by

$$\hat{\bar{\Psi}}_2 = \frac{1}{n} \sum_{i=1}^{n} \hat{\xi}_{1i} \hat{\psi}_i \hat{\psi}_i' \hat{\xi}_{1i}',$$

where $\hat{\psi}_i = -\hat{V}_i$, and $\hat{\xi}_{1i}$ is defined as the fitted value of some nonparametric kernel regression of $\tau_i \partial_\beta \hat{G}(W) \partial_2 \hat{G}(W)(\hat{G}(W)(1 - \hat{G}(W)))^{-1}$ on Z. Under suitable regularity conditions, one can verify that a Law of Large Numbers holds for $\hat{\bar{\Psi}}_2$ in this case.

5 Some Extensions of the Structure of the Model

For the ease of exposition, we have chosen a formulation our model in Section 2 that is simple but also restrictive in many ways, yet various aspects can easily be generalized. First, the linear relationship in the outcome equation (2.1) could be replaced with a nonlinear one, such as

$$Y = \mathbb{I}\{g(X, \theta_o) - U > 0\}$$

for some known function g, at the cost of a slightly more complicated normalization of the parameters (see Ichimura 1993, Klein and Spady 1993).

Second, we could replace the conditional independence restriction in (2.2) with the alternative, slightly weaker version

$$U \perp X | (V, X\beta_o). \tag{2.2b}$$

This would allow for a limited degree of dependence between X and U even when conditioning on the control variable V, as long as this dependence is restricted to run through the index values (as would be the case under index heteroskedasticity, for example). In particular, it would still be possible to write $\mathbb{E}(Y|X, V)$ as some function G_o of $X\beta_o$ and V, but now this function would not be confined to be monotone in its first argument. As our estimator (in contrast to the one of Blundell and Powell) does not explicitly use the properties of a distribution, it automatically works under (2.2b) as well. We illustrate this point in more detail in our simulation study.

Finally, in this paper we focus on the estimation of the (normalized) index coefficients β_o. Another object of practical interest one could consider would be the choice probability for some exogenously determined value of the regressors $X = \bar{x}$. Blundell and Powell (2004) call this the

average structural function (ASF), and show that it is identified as the partial mean of G_o with respect to the distribution of the control variable V,

$$ASF(\bar{x}) = \int G_o(\bar{x}\beta_o, V) dF_V, \tag{5.1}$$

provided that the support of V does not vary with $\bar{x}$. The estimation of this object is discussed in more detail in Imbens and Newey (2009).

6 Simulation Study

6.1 Setup

In order to demonstrate the usefulness of our proposed estimator for applications to finite samples, we report the results of three simulation experiments in this section. Apart from our SML procedure, we also consider Blundell and Powell's (2004) semiparametric "matching" estimator, the "Two-Stage-Probit" estimator of Smith and Blundell (1986) or Rivers and Vuong (1988), and Two-Stage Least Squares (2SLS) estimation of a linear probability model, which is frequently used in applied work. These are intended to serve as points of reference.

For the three simulations, we always use the same specification for the regressors and instruments, but change the properties of the joint distribution of the error terms (U, V). The dependent variable is generated by a binary response model with two covariates in the outcome equation, of which one is endogenous, and two additional instruments in a linear reduced form equation:

$$\begin{aligned} Y_1 &= \mathbb{I}\{X^e + Z_1\beta_o > U\}, \\ X^e &= \alpha_{o0} + Z_1\alpha_{o1} + Z_{21}\alpha_{o2} + Z_{22}\alpha_{o3} + V. \end{aligned}$$

The true parameter values $\beta_o = 1$ and $\alpha_o = (1, 2/3, 2/3, 1/3)'$ are held constant across simulations. The exogenous variables are independent, with Z_1 being exponentially distributed, truncated from above at 3, and standardized to have mean zero and variance two, and Z_{21}, Z_{22} are standard normal. In order to ensure a sensible comparison, all estimators are based on the OLS residuals from the reduced form equation. For the error distributions, we simulate V as $N(0, 1)$ and $U = U^* + V$, where we use the following specifications for U^*:

- Design I: $U^* \sim N(0, 5)$
- Design II: $U^* \sim 0.8N(-1, .6) + 0.2N(4, 2)$

- Design III: $U^* \sim N(0, \exp(0.1 + 0.5X\beta_o))$

Design I implies a jointly normal distribution of (U, V) and is the one under which a Probit should give the best results. The second design is a mixture of two normal distributions, resulting in a right-skewed and bimodal density of U. It is constructed such that the Probit estimator should be markedly biased, and we thus expect a comparatively better performance of the semiparametric procedures. For the third design, the variance of U conditional on V is a function of the linear index. It is included to show that our estimator also works when the restriction in (2.2) is replaced with its weaker version (2.2b) (see section 5).

While these designs correspond to very different distributions, they are chosen such that some features are approximately the same. In particular, it holds that $\text{Var}(U) \approx 6, \text{Var}(Y_2+Z_1) \approx 4.5$, $\text{Cor}(U, V) \approx 0.4$ and $\text{Cor}(U, Y_2) \approx 0.25$. With the multiple R^2 in the reduced form regression being about 0.6, we are in a situation with relatively strong instruments. In all three cases, we consider the sample sizes $n = 250, 500, 1000$, and set the number of replications to 1000.

6.2 Implementation Issues

In order to implement our SML estimator, we have to select a kernel function and the bandwidth parameters. In particular, our assumptions require the use of higher-order kernels to eliminate asymptotic bias. However, when using higher-order kernels to calculate $\hat{G}(\cdot|\beta, \hat{v})$, some observations will be given a negative weight and the result is not confined to be between zero and one, which of course causes problems when taking logarithms. For our simulations, we therefore consider two approaches to circumvent this this problem. The first one employs an idea from Klein and Spady (1993) and consists of minimizing a modified criterion function $\tilde{L}_n$, where

$$\tilde{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau_i(Y_i \log(\hat{G}(\hat{W}_i(\beta)|\beta)^2) + (1 - Y_i)\log((1 - \hat{G}(\hat{W}_i(\beta, \hat{v})|\beta, \hat{v}))^2)).$$

The corresponding estimator $\tilde{\beta}$ can easily be shown to be consistent and having the same limiting distribution as our SML estimator $\hat{\beta}$. In particular, note that both are solutions of the same first-order condition. We refer to this estimator as SML-1 below.

As a second possibility, we simply compute our estimator as described above, but without the use of higher-order kernel functions. This is motivated by the frequently made observation that while higher-order kernel might be required from a theoretical point of view in many semiparametric applications, the resulting estimators often tend to have inferior finite sample properties compared to those based on standard kernels (see Marron (1994) or Jones and Sig-

norini (1997)). Thus, although strictly speaking not compatible with our asymptotic analysis, we also consider this approach for our simulations. It is referred to as SML-2 below.

Regarding the choice of the bandwidth parameters $h = (h_1, h_2, \ldots, h_{d_v+1})$, for our simulation study we follow Härdle, Hall, and Ichimura (1993) and Delecroix, Hristache, and Patilea (2005), and consider the following pragmatic approach: we treat the bandwidths as additional parameters of the estimated likelihood and perform the maximization with respect to both β and h. That is, we use the first component of

$$\left(\hat{\beta}, \hat{h}\right) = \underset{(\beta,h)\in B\times\mathbb{R}_+^{d_v+1}}{\operatorname{argmax}} L_n(\beta, h)$$

as our estimator. While we do not claim any optimality of this approach for our problem at hand, the method seems to perform well in applications to finite samples, as shown by our simulation study. A further advantage is that it can also serve as an informal test for endogeneity: when X^e is actually exogenous, typically a large value will be chosen for the bandwidth, because in this case $G_o(X\beta, V)$ does not vary with V. As an alternative, one could also experiment with various multiples of $n^{-\delta}$, but practitioners are generally reluctant to do so because it involves a large degree of subjectivity.

In line with most of the literature in this field, no trimming is used. We investigated various forms of trimming, but found no substantial effect on the performance of the estimator in our simulation scenarios. This result is common when evaluating the finite sample properties of semiparametric estimators of single-index models in general. However, the use of trimming might be beneficial in practice if the data contains some extreme outliers, as they can have a substantial impact on the estimate of the link function and the chosen bandwidth. In this case, a trimming procedure could for example be implemented on the basis of the upper and lower sample quantiles of the data, as mentioned above[3].

The numerical optimization is carried out using a Gauss-Newton type algorithm as implemented in the software package *R* 2.5.1. We use the Probit results as starting values for the index coefficients and .4 for the bandwidths. To guard against the algorithm converging to possible local maxima, we also use half and twice the starting values values to compute the estimator, and retain the result that gives the highest value of the objective function. However, it turns out that in our simple setup the values coincide in most runs.

All other estimators were implemented as described in the respective literature. For the

[3]In the presence of extreme outliers, the use of trimming should of course be beneficial even for correctly specified parametric estimators

Table 1: Simulation Results Design I

		MEAN	SD	RMSE	MAD	25%	50%	75%	CR
$n = 250$	SML-1	1.073	0.443	0.449	0.366	0.722	1.000	1.336	0.997
	SML-2	1.008	0.248	0.248	0.199	0.854	1.012	1.177	0.895
	Probit	1.011	0.187	0.188	0.149	0.889	1.007	1.122	0.897
	2SLS	1.089	0.186	0.206	0.165	0.956	1.090	1.204	0.844
	BP	0.735	0.389	0.471	0.391	0.459	0.722	0.969	0.787
$n = 500$	SML-1	1.061	0.463	0.467	0.371	0.672	0.999	1.333	0.975
	SML-2	1.001	0.163	0.163	0.131	0.895	1.002	1.116	0.913
	Probit	0.999	0.137	0.137	0.111	0.901	1.000	1.093	0.883
	2SLS	1.079	0.138	0.159	0.128	0.979	1.084	1.172	0.789
	BP	0.812	0.287	0.343	0.279	0.602	0.810	1.027	0.675
$n = 1000$	SML-1	1.010	0.444	0.444	0.356	0.667	0.983	1.332	0.945
	SML-2	1.002	0.120	0.120	0.095	0.926	0.999	1.080	0.901
	Probit	1.003	0.094	0.094	0.077	0.936	1.009	1.070	0.904
	2SLS	1.082	0.094	0.125	0.103	1.018	1.088	1.147	0.745
	BP	0.857	0.195	0.242	0.197	0.733	0.851	0.981	0.582

Blundell-Powell estimator, we use Least Squares Cross Validation to determine the bandwidth for the nonparametric regression part, and $1.06\sigma_w n^{-1/5}$ for the "matching" part, which corresponds to the specification in their empirical application.

6.3 Results

To facilitate comparison of our SML estimator with the other procedures, we make use of a different normalization than the one described in Section 2: instead of setting the coefficient of the endogenous variable to one, we rescale the estimates of the coefficients such that the sum of their absolute values is equal to 2, which corresponds to the sum of the magnitude of the true coefficients. The reason for this change is that using ratios of estimated coefficients results in a number of extreme outliers for the Blundell-Powell estimator that corrupt the analysis. With the new normalization, the estimates are much more well behaved.

The results of the simulation experiments are given in Tables 1– 3. For each estimator of $\beta_o = 1$, we report the mean value (MEAN), standard deviation (SD), root mean squared error (RMSE), median absolute deviation (MAD), the 25%, 50% and 75% sample quantiles, and the coverage rate (CR) of a bootstrap confidence interval with nominal level of 90%, obtained via the percentile method from 200 bootstrap replications.

Some general conclusions can be drawn from these results. First, although the SML-1

Table 2: Simulation Results Design II

		MEAN	SD	RMSE	MAD	25%	50%	75%	CR
$n = 250$	SML-1	1.053	0.459	0.462	0.371	0.667	0.999	1.334	0.995
	SML-2	1.122	0.311	0.334	0.264	0.911	1.107	1.317	0.913
	Probit	1.209	0.267	0.339	0.265	1.038	1.200	1.368	0.784
	2SLS	1.286	0.246	0.377	0.312	1.120	1.282	1.437	0.672
	BP	1.019	0.576	0.576	0.497	0.546	0.990	1.558	0.915
$n = 500$	SML-1	1.084	0.449	0.457	0.366	0.695	1.000	1.344	1.000
	SML-2	1.088	0.229	0.245	0.194	0.924	1.073	1.258	0.890
	Probit	1.204	0.178	0.271	0.220	1.081	1.199	1.313	0.724
	2SLS	1.285	0.170	0.332	0.288	1.169	1.278	1.383	0.523
	BP	1.061	0.509	0.512	0.429	0.699	1.022	1.459	0.928
$n = 1000$	SML-1	1.026	0.412	0.413	0.333	0.668	0.980	1.328	0.989
	SML-2	1.054	0.165	0.173	0.136	0.941	1.045	1.157	0.897
	Probit	1.200	0.135	0.241	0.207	1.104	1.205	1.281	0.506
	2SLS	1.277	0.128	0.305	0.278	1.188	1.283	1.355	0.272
	BP	1.065	0.355	0.360	0.293	0.817	1.067	1.341	0.913

Table 3: Simulation Results Design III

		MEAN	SD	RMSE	MAD	25%	50%	75%	CR
$n = 250$	SML-1	1.052	0.428	0.431	0.349	0.672	1.000	1.333	1.000
	SML-2	1.101	0.234	0.255	0.195	0.945	1.080	1.246	0.921
	Probit	1.203	0.216	0.296	0.236	1.050	1.191	1.337	0.767
	2SLS	1.242	0.204	0.317	0.260	1.114	1.235	1.358	0.677
	BP	1.016	0.548	0.548	0.472	0.546	1.008	1.508	0.918
$n = 500$	SML-1	1.006	0.396	0.397	0.316	0.671	0.986	1.329	0.985
	SML-2	1.068	0.170	0.183	0.145	0.941	1.059	1.193	0.896
	Probit	1.200	0.144	0.247	0.210	1.105	1.189	1.291	0.587
	2SLS	1.238	0.136	0.274	0.241	1.152	1.231	1.324	0.431
	BP	1.094	0.464	0.473	0.396	0.731	1.112	1.453	0.922
$n = 1000$	SML-1	0.973	0.372	0.373	0.287	0.672	0.970	1.243	1.000
	SML-2	1.036	0.109	0.114	0.090	0.961	1.032	1.108	0.911
	Probit	1.188	0.103	0.215	0.190	1.117	1.182	1.257	0.371
	2SLS	1.226	0.097	0.246	0.227	1.159	1.223	1.292	0.193
	BP	1.098	0.329	0.343	0.273	0.865	1.074	1.329	0.881

estimator has slightly better bias properties than SML-2, it also has a substantially higher variability in all three designs. Thus, in terms of RMSE or MAD, the SML estimator based on standard kernels uniformly dominates the one using higher-order kernels. Secondly, the SML-2 estimator compares favourably with the other alternatives and performs well uniformly over the different models we consider. It has the lowest RMSE under all designs but the first, where it exceeds the RMSE of the correctly specified Probit by about 20%. In addition, the confidence intervals' coverage rates are remarkably close to the nominal level for all sample sizes and designs in the study. Third, the Probit estimator performs best when the parametric model is correctly specified, as is the case in Design I, and least well when the deviations from this model are most extreme. In general, its variance tends to be somewhat smaller than that of the SML estimator, but the bias is higher. Thus, when the bias induced by the misspecification is not too large, it tends to give reasonably good estimates. The bootstrap confidence intervals on the other hand have coverage rates far below their nominal level in the misspecified cases, and can thus be misleading in practice. Fourth, the Blundell-Powell estimators' performance is generally inferior to our that of our SML-2 procedure. For the relatively small sample sizes we consider, its RMSE and MAD also exceed the ones of the misspecified parametric estimators. For larger samples however, one would expect this relation to revert, since, at least for the second and third design, the Blundell-Powell estimator has a relatively small bias. Since the bootstrap confidence intervals perform satisfactory as well, this procedure could then be a useful alternative to SML in large samples, since then the latter is hard to compute. Moreover, it should be possible to improve the performance of the Blundell-Powell estimator through more effective rules for selecting the smoothing parameters, which is an important topic for future research. Finally, the 2SLS estimator turns out to have a low variance, but it is markedly biased in the second and third design. Consequently, the confidence intervals' coverage rates are far below their nominal values in this case. Although this estimator is applied frequently in empirical applications, one should thus be very careful when interpreting the results.

7 An Empirical Application: Home-ownership and Income in Germany

As an empirical application, we study the role of household income on the decision to rent an apartment or house versus owning it. The data we use are taken from the 2004 wave of the German Socioeconomic Panel (GSOEP), an extensive longitudinal survey of households in

Table 4: Descriptive Statistics

Variable	Mean	Std.Dev.	Min	Max
Homeowner	0.599	0.490	0	1
ln(total income)	7.853	0.324	6.397	9.473
Age	40.613	5.374	30	50
Children in HH	0.848	0.359	0	1
Education of wife				
Low degree	0.482	0.498	0	1
Intermediate degree	0.415	0.493	0	1
High degree	0.103	0.304	0	1
Wife Working	0.699	0.459	0	1

Notes: Sample size is $n = 981$. Education dummys indicate the highest of the three main secondary school tracks in Germany completed by the wife: *Hauptschulabschluss* ("low degree"), *Realschulabschluss* ("intermediate degree") or *Abitur* ("high degree"; university entry qualification)). "Wife Working" is an indicator that takes the value 1 when the wife has done any for-pay work in 2004.

Germany similar to the Panel Study of Income Dynamics (PSID) in the United States. The sample we use consists of 981 married men aged 30 to 50 that are working full time and have completed at most the lowest secondary school track of the German education system. Our dependent variable Y is an indicator that takes the value of 1 if a person owns its residence, and 0 if it is rented. The covariates X we are controlling for are the 2004 average total monthly income of the corresponding household (X^e), the person's age in years (Z_{11}) and an indicator for the presence of children younger than 16 in the household (Z_{12}). Generally speaking, home ownership should be determined by the permanent component of the income stream, of which monthly income is only a noisy measure. Therefore, we treat income as a mismeasured and thus potentially endogenous variable and employ dummy variables for the wife's education level (Z_{21}) and employment status (Z_{22}) as instruments. These human capital variables should be strongly related to the household income but have no direct influence on the housing decision. Some descriptive statistics for these variables are given in Table 4.

A priori, we would expect that all three regressors are positively related with home-ownership for the following reasons: First, buying a house is associated with high financial costs including down payments, mortgage interests and repayments, maintenance costs and transaction costs such as notary fees and transfer taxes. Particularly in the first few years after buying a home, these costs can exceed the costs of renting an equivalent place considerably. Thus, a higher level of income is needed to acquire a house in the first place. Second, the transition from renting to home-ownership is usually a one-time, non-reversible event associated with the family lifecycle. Thus, the proportion of home-owners should increase, other things equal, with age. Finally, it

is well known that parenthood is a trigger for buying a home, and hence families with children should be more likely to own their residence.

For our application, we normalize the coefficient on the indicator for the presence of children to unity. Hence the model is given by

$$\begin{aligned} Y &= \mathbb{I}\{X^e\beta_{o1} + Z_{11}\beta_{o2} + Z_{12} \geq U\}, \\ X^e &= m_o(Z) + V. \end{aligned}$$

We consider specifying the reduced form for the endogenous regressor both parametrically as a linear model and in a fully nonparametric way. Since the resulting residuals are relatively similar, in Table 4 only report OLS estimates of α_o when the mean function is specified as $m_o(z) = z'\alpha_o$.

We then estimate the unknown parameter vector β_o by SML. Following the results from our simulations, we consider only the SML-2 estimator. The estimated coefficients $\hat{\beta}$ and their corresponding standard errors are given in the second column of Table 5. For the purpose of comparison, we also estimate the outcome equation by SML without taking the potential endogeneity into account, i.e. we use the ordinary Klein-Spady estimator with kernel and bandwidth specification analogous to the ones described in the preceding section. Finally, we also report results from applying the Blundell-Smith estimator and a standard probit model in the fourth and firth column of Table 5, respectively.

We can see that under all specifications the general tendencies we described above are confirmed. However, the difference between the estimates of β_o with and without controlling for endogeneity are quite substantial. Consider for example the estimates obtained by SML. After accounting for endogeneity, the coefficient on income is about twice as large as before (relative to the coefficient on the child indicator). Using a fully parametric approach leads to a quantitatively similar conclusion.

To illustrate the impact of such a change in coefficients, we consider the implications for the Average Structural Function (ASF), which gives the choice probabilities when the value of the regressors X is fixed at some exogenously determined value $\bar{x}$. As mentioned in Section 5, this object is identified as a partial mean of the link function G_o with respect to V. Following the advice of Imbens and Newey (2009), we estimate the ASF by

$$\hat{ASF}(\bar{x}) = \frac{1}{n}\sum_{i=1}^{n} \hat{G}_{LL}(\bar{x}\hat{\beta}, \hat{V}_i|\hat{\beta}, \hat{v}), \tag{7.1}$$

where $\hat{G}_{LL}(\bar{x}\hat{\beta}, \hat{V}_i|\hat{\beta}, \hat{v})$ is the local linear estimator of $E(Y|X\beta, V)$ evaluated at $\bar{x}\hat{\beta}$ and $\hat{V}_i$, and

Table 5: Estimation Results from Semiparametric and Fully Parametric Procedures

Variable	Reduced Form	SML estimates		Probit estimates	
	X^e	$\Pr(Y\|V)$	$\Pr(Y)$	$\Pr(Y\|V)$	$\Pr(Y)$
log(Total Income)	—	3.8533	1.9118	4.7923	2.1343
		(1.3338)	(.7310)	(1.5135)	(.5571)
Age	.0117	.0982	.1916	0.0863	0.2076
	(.0017)	(.0889)	(.0439)	(0.0209)	(.0257)
Children in HH	.0911	1.0000	1.0000	1.0000	1.0000
	(.0194)	—	—	—	—
$\hat{V}$ (Control variable)	—	—	—	−3.0348	—
				(1.3048)	
Education of wife					
Intermediate degree	.0642	—	—	—	—
	(.0185)				
High degree	.1291	—	—	—	—
	(.0298)				
Wife Working	.0911	—	—	—	—
	(.0194)				
R^2	.1072	—	—	—	—
F-statistic	23.42 (5, 975 df)	—	—	—	—
Bandwidth	—	$h = (0.04, .21)$	$h = .03$	—	—

Notes: Standard errors (based on bootstrap 500 bootstrap replications for SML and the usual formulas otherwise) in parentheses. Baseline category for Education of wife is "low degree".

the bandwidth is chosen by least squares cross-validation.

In Figure 1, the estimated ASF is plotted from the 5% to the 95% quantile of the income distribution for a man aged 40 with children. We can see that the two models imply vastly different probabilities of home-ownership, particularly in the lower half of the income distribution. For a monthly household net income of 1800 EUR (corresponding to a log income of about 7.5), the probability of owning the residence reduces from 50% to roughly 20% when controlling for endogeneity. This difference diminishes as we move up the income distribution, and for values of income larger than 2500 EUR (which corresponds to a log income of about 7.8), the predictions from the two models are qualitatively similar.

8 Concluding Remarks

This paper presents a semiparametric maximum likelihood procedure for the estimation of the coefficients of a single index binary choice model with endogenous regressors. We discuss how identification is achieved via a control function approach, and derive the asymptotic properties of the new estimator. In our Monte Carlo experiments, the new estimator performs well in

Figure 7.1: Estimated probability of owning the residence for a man aged 40 with children.

comparison with other related procedures.

One of the major issues of our estimator is its computational complexity when applied in settings with many regressors and/or observations. In this case, even evaluating the likelihood function at a specific point is very time consuming, and the function might have several local maxima. However, these problems are not specific to our SML estimator but are encountered in general when computing semiparametric optimization estimators such as the ones by Ichimura (1993) or Klein and Spady (1993). For these estimators, a number of suggestions have been made to improve their numerical properties, such as e.g. the use of Fast Fourier Transforms or binning techniques (see Ichimura and Todd (2007) for a comprehensive overview). All of these approaches could in general be adapted to our estimator as well.

It might also be possible to avoid the use of numerical optimization routines altogether. In a recent paper, Xia (2006) shows that the computationally much simpler rMAVE procedure of Xia, Tong, Li, and Zhu (2002) achieves the same asymptotic variance as the Klein and Spady estimator when applied to a standard binary choice model without endogenous regressors. Again, it should be possible to adapt this technique to our problem and thus reduce the computational complexity.

A Appendix: Proofs

Proof of Theorem 1. The proof of the theorem is analogous to the argument in Manski (1988): First, note that that $V = v_o(X^e, Z)$ is identified by assumption. Now assume that there exists a $\tilde{\beta} \in \mathcal{B}$ such that $\mathbb{E}(Y|X,V) = \mathbb{E}(Y|X\tilde{\beta},V) = \mathbb{E}(Y|X\beta_o,V) \equiv G_o(X\beta_o,V)$. Then there must exist a function $H(\cdot,V)$ that is strictly monotone for all V, such that $X^{(1)} + X^{(-1)'}\tilde{\beta} = H(X^{(1)} + X^{(-1)'}\beta_o, V)$. Differentiating both sides of this equation with respect to $X^{(1)}$ for $X \in \mathcal{A}$, we see that $H(\cdot,V)$ must be the identity function since $X^{(1)}$ is continuously distributed conditional on V, and thus $X^{(-1)'}\tilde{\beta} = X^{(-1)'}\beta_o$. By condition (iii), this relation can hold with probability one only if $\tilde{\beta} = \beta_o$. □

We now turn to the proofs of the consistency and asymptotic normality result. First, we give some useful preliminary results on uniform rates of convergence for nonparametric estimators based on generated regressors. Second, we show consistency via a classical, direct argument. Third, we prove asymptotic normality of our estimator by showing that our problem fits the framework of Chen, Linton, and Van Keilegom (2003).

Lemma 1. *Under Assumption 1-8, we have that uniformly in $w \in \mathcal{W}$ and $\beta \in \mathcal{B}$, respectively, i) $\hat{D}(w|\beta,\hat{v}) - \hat{D}(w|\beta) = o_p(n^{-1/4})$, ii) $\partial_\beta \hat{D}(w|\beta,\hat{v}) - \partial_\beta \hat{D}(w|\beta) = o_p(n^{-1/4})$, iii) $\partial_k \hat{D}(w|\beta,\hat{v}) - \partial_k \hat{D}(w|\beta) = o_p(n^{-1/4})$ for $k = 1,2$, iv) $\hat{N}(w|\beta,\hat{v}) - \hat{N}(w|\beta) = o_p(n^{-1/4})$, v) $\partial_\beta \hat{N}(w|\beta,\hat{v}) - \partial_\beta \hat{N}(w|\beta) = o_p(n^{-1/4})$, vi) $\partial_k \hat{N}(w|\beta,\hat{v}) - \partial_k \hat{N}(w|\beta) = o_p(n^{-1/4})$ for $k = 1,2$.*

Proof. We only proof the first result, as the remaining ones can by shown analogously. Using the definition of $\hat{D}$ and Hölder's inequality, it follows that

$$\begin{aligned}
|\hat{D}(w|\beta,\hat{v}) - \hat{D}(w|\beta)| &= \left| \frac{1}{nh^{d_v}} \sum_{i=1}^{n} \partial_2 K_h(X_i\beta - w_1, \tilde{V}_i - w_2)(\hat{V}_i - V_i) \right| \\
&\leq h^{-d_v} \left(\frac{1}{n} \sum_{i=1}^{n} (\partial_2 K_h(X_i\beta - w_1, \tilde{V}_i - w_2))^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} (\hat{V}_i - V_i)^2 \right)^{1/2} \\
&= h^{-d_v} T_1 \times T_2,
\end{aligned}$$

where $\tilde{V}_i$ is some value between V_i and $\hat{V}_i$. It is then easy to show that $T_1 = O_p(1)$ uniformly in β and w. Now consider T_2. Substituting the "high-level" representation for $\hat{V}_i - V_i$ from Assumption 8 into the expression, and applying Jensen's inequality and the usual projection arguments for U-Statistics, we obtain that

$$\begin{aligned}
T_2^2 &= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{1}{n} \sum_{j=1}^{n} \omega_n(Z_i, Z_j)\psi_j + o_p(n^{-1/2}) \right)^2 \\
&\leq \frac{1}{n^3} \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_n(Z_i, Z_j)^2 \psi_j^2 + o_p(n^{-1}) \\
&= O_p(n^{-1})
\end{aligned}$$

It then follows together with Assumption 7 that $h^{-d_v} T_1 T_2 = O_p(h^{-d_v} n^{-1/2}) = o_p(n^{-1/4})$, as claimed. □

Lemma 2. *Under Assumption 1–8, (i)*

$$\sup_{w\in\mathcal{W},\beta\in\mathcal{B}} |\hat{G}(w|\beta,\hat{v}) - G(w|\beta)| = o_p(1)$$

and (ii)

$$\sup_{w\in\mathcal{W},\|\beta-\beta_o\|\leq\delta_n} |\hat{G}(w|\beta,\hat{v}) - G(w|\beta)| = o_p(n^{-1/4})$$

$$\sup_{w\in\mathcal{W},\|\beta-\beta_o\|\leq\delta_n} |\partial_\beta\hat{G}(w|\beta,\hat{v}) - \partial_\beta G(w|\beta)| = o_p(n^{-1/4})$$

$$\sup_{w\in\mathcal{W},\|\beta-\beta_o\|\leq\delta_n} |\partial_1\hat{G}(w|\beta,\hat{v}) - \partial_1 G(w|\beta)| = o_p(n^{-1/4})$$

for all $\delta_n = o(1)$.

Proof. This follows from standard kernel smoothing theory together with Lemma 1. □

Proof of Theorem 2. To show that $\hat{\beta}$ is consistent, we first define an infeasible version of the semi-parametric likelihood function, with $\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v})$ replaced with its probability limit $G(W_i(\beta)|\beta)$:

$$\tilde{L}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \tau_i(Y_i \log(G(W_i(\beta)|\beta)) + (1-Y_i)\log(1-G_i(W(\beta)|\beta))).$$

The difference between $\tilde{L}_n(\beta)$ and $L_n(\beta)$ goes to zero uniformly in β, for $n\to\infty$, because

$$\begin{aligned}
\sup_{\beta\in B}|L_n(\beta)-\tilde{L}_n(\beta)| &\leq \left(\inf_{\beta\in B}\min_i \left\{\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v}), 1-\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v}), G(W_i(\beta), 1-G(W_i(\beta))\right\}\right) \\
&\quad \times \sup_{\beta\in B}\max_i \tau_i|\hat{G}(\hat{W}_i(\beta)|\beta,\hat{v}) - G(W_i(\beta)|\beta)| \\
&= O_p(1)\sup_{\beta\in B}\max_i \tau_i|\hat{G}(W_i(\beta)|\beta,\hat{v}) - G(W_i(\beta)) + \partial_2\hat{G}(\tilde{W}_i(\beta)|\beta,\hat{v})|\beta)(\hat{V}_i - V_i))| \\
&= o_p(1)
\end{aligned}$$

by Lemma 2 and Assumption 8, where $\tilde{W}_i(\beta)$ is some value between $\hat{W}_i(\beta)$ and $W_i(\beta)$. Furthermore, since $\tilde{L}_n(\beta)$ is an ordinary parametric likelihood function, by a standard uniform law of large numbers, e.g. Lemma 2.4 in Newey and McFadden (1994), it converges uniformly in β to its expectation, i.e. we have

$$\sup_{\beta\in B}|\tilde{L}_n(\beta) - L(\beta)| = o_p(1),$$

where

$$L(\beta) = \mathbb{E}(L_n(\beta)) = \mathbb{E}(\tau_i\left[Y_i\log(G(W(\beta)|\beta)) + (1-Y_i)\log(1-G(W(\beta)|\beta))\right])$$

is a non-random function that is continuous in β. Taken together, it follows from the triangle inequality that

$$\sup_{\beta\in B}|L_n(\beta) - L(\beta)| = o_p(1),$$

which implies that $\hat{\beta}$ is consistent whenever $L(\beta)$ attains a unique maximum at β_o. By the law of iterated expectations,

$$L(\beta) = \mathbb{E}(\tau\left[G_o(X\beta_o, V)\log(G(W(\beta)|\beta)) + (1 - G_o(X\beta_o, V))\log(1 - G(W(\beta)|\beta))\right]),$$

and the term in square brackets attains its maximum whenever the relation $G(W(\beta)|\beta) = G_o(X\beta_o, V)$ holds. By Assumption 1, this is the case if and only if $\beta = \beta_o$. The statement of the Theorem then follows from the usual consistency argument, e.g. Theorem 2.1 in Newey and McFadden (1994). □

We now turn to the proof of asymptotic normality of our estimator $\hat{\beta}$. This is done by verifiying the conditions of Theorem 2 in Chen, Linton, and Van Keilegom (2003) in Lemma 3–8. Similar arguments are used by Linton, Sperlich, and Van Keilegom (2008), who consider semiparametric estimation of a transformation model. Their problem is technically related to ours since they also consider a semiparametric maximum likelihood estimator based on nonparametrically generated regressors, but the actual model is very different.

We start with introducing some further notation. First, we have to define a criterion function depending on β and some unknown nuisance function, whose population value is equal to zero at the true parameter values. To this end, write $\gamma = (\gamma_1, \ldots, \gamma_4)$ for a generic collection of nuissance functions, and define $\gamma_\beta = (\partial_1 G(\cdot|\beta), \partial_\beta G(\cdot|\beta), G(\cdot|\beta), v_o)$, $\gamma_o = \gamma_{\beta_o}$, and $\hat{\gamma}_\beta = (\partial_1 \hat{G}(\cdot|\beta, \hat{v}), \partial_\beta \hat{G}(\cdot|\beta, \hat{v}), \hat{G}(\cdot|\beta, \hat{v}), \hat{v}_o)$, and $\hat{\gamma}_o = \hat{\gamma}_{\beta_o}$. Then, for any γ, let

$$S_n(\beta, \gamma) = \frac{1}{n}\sum_{i=1}^{n} s(Y_i, X_i, Z_i, \beta, \gamma)$$

where

$$\begin{aligned} s(Y_i, X_i, Z_i, \beta, \gamma) = {} & (\gamma_1(X_i\beta, \gamma_4(X_i, Z_i))\tilde{X}_i + \gamma_2(X_i\beta, \gamma_4(X_i, Z_i))) \\ & \times \frac{Y_i - \gamma_3(X_i\beta, \gamma_4(X_i, Z_i))}{\gamma_3(X_i\beta, \gamma_4(X_i, Z_i))(1 - \gamma_3(X_i\beta, \gamma_4(X_i, Z_i)))} \end{aligned}$$

and note that

$$S_n(\beta, \hat{\gamma}_\beta) = \partial_\beta L_n(\beta),$$

i.e. $S_n(\beta, \hat{\gamma}_\beta)$ is the score corresponding to our likelihood function $L_n(\beta)$. Furthermore, define the population version of the criterion function as

$$S(\beta, \gamma) = \mathbb{E}(S_n(\beta, \gamma)).$$

Finally, we have to define an appropriate space for the nuissance functions γ. Denote this space by $\Gamma = \Gamma_1 \times \mathcal{V}$, where $\mathcal{V}$ is defined in Assumption 8 and Γ_1 is the class of all functions $f : \mathbb{R}^{1+d_v} \to \mathbb{R}$ whose partial derivatives up to order $\alpha > (1 + d_v)/2$ exist and are uniformly bounded by some constant M. This class of functions is typically denoted by $C_M^\alpha(\mathbb{R}^2)$ in the literature (see e.g. van der Vaart and

Wellner (1996, p. 154)). A norm $\|\cdot\|_\Gamma$ on the space Γ that satisfies the requirements of Chen, Linton, and Van Keilegom (2003) can be defined as

$$\|\gamma\|_\Gamma = \sup_{\beta\in\mathcal{B}} \max\{\|\gamma_1\|_\infty, \ldots, \|\gamma_4\|_\infty\}.$$

Note that our Assumption 3 and 8 are sufficient to ensure that $\gamma_o \in \Gamma$.

We can now prove the Lemmas needed to verify the conditions of Theorem 2 in Chen, Linton, and Van Keilegom (2003).

Lemma 3 (Condition (2.1))**.** $\|S_n(\hat{\beta}, \hat{\gamma}_\beta)\| = \inf_{\beta\in\mathcal{B}} \|S_n(\beta, \hat{\gamma}_\beta)\| + o_p(n^{-1/2})$

Proof. This is trivially satisfied since $\|S_n(\hat{\beta}, \hat{\gamma}_\beta)\| = 0$ by construction. □

Lemma 4 (Condition (2.2))**.** *The ordinary derivative $S_\beta(\beta, \gamma_\beta) = \partial S(\beta, \gamma_\beta)/\partial\beta$ exists in a neighborhood of β_o, is continuous at $\beta = \beta_o$, and the matrix $S_\beta(\beta_o, \gamma_{\beta_o})$ is of full rank.*

Proof. This follows directly from Assumptions 3 and 5. □

Lemma 5 (Condition (2.3))**.** *The pathwise derivative $\dot{S}(\beta, \gamma_\beta)$ of $S(\beta, \gamma_\beta)$ exists in all directions $\gamma - \gamma_\beta$, and satisfies: (i)*

$$\|S(\beta, \gamma) - S(\beta, \gamma_\beta) - \dot{S}(\beta, \gamma_\beta)[\gamma - \gamma_\beta]\| \leq c\|\gamma - \gamma_\beta\|_\Gamma^2$$

for all $\beta \in \mathcal{B}$ with $\|\beta - \beta_o\| \leq \delta_n$, all γ with $\|\gamma - \gamma_o\|_\Gamma \leq \delta_n$, some positive sequence $\delta_n = o(1)$, and some constant $c < \infty$; and (ii)

$$\|\dot{S}(\beta, \gamma_\beta)[\gamma - \gamma_\beta] - \dot{S}(\beta_o, \gamma_o)[\gamma - \gamma_o]\| \leq o(1)\delta_n.$$

Proof. Using standard rules for calculating pathwise derivatives, we obtain after some calculations that

$$\begin{aligned}
\dot{S}(\beta, \gamma_\beta)[\gamma] = \mathbb{E}\Bigg[& \tau \frac{Y - G(W(\beta)|\beta)}{G(W(\beta)|\beta)(1 - G(W(\beta)|\beta))}(\gamma_1(W(\beta))\tilde{X} + \gamma_2(W(\beta))) \\
& - \tau\partial_\beta G(W(\beta)|\beta)\frac{1}{G(W(\beta)|\beta)(1 - G(W(\beta)|\beta))}\gamma_3(W(\beta)) \\
& - \tau\partial_\beta G(W(\beta)|\beta)\frac{(Y - G(W(\beta)|\beta))(1 - 2G(W(\beta)|\beta))}{G(W(\beta)|\beta)(1 - G(W(\beta)|\beta))}\gamma_3(W(\beta)) \\
& - \tau\partial_\beta G(W(\beta)|\beta)\frac{1}{G(W(\beta)|\beta)(1 - G(W(\beta)|\beta))}\partial_2 G(W(\beta)|\beta)\gamma_4(X^e, Z) \\
& - \tau\partial_\beta G(W(\beta)|\beta)\frac{(Y - G(W(\beta)|\beta))(1 - 2G(W(\beta)|\beta))}{G(W(\beta)|\beta)(1 - G(W(\beta)|\beta))}\partial_2 G(W(\beta)|\beta)\gamma_4(X^e, Z) \\
& + \tau \frac{Y - G(W(\beta)|\beta)}{G(W(\beta)|\beta)(1 - G(W(\beta)|\beta))}\partial_{\beta,2} G(W(\beta)|\beta)\gamma_4(X^e, Z)\Bigg].
\end{aligned}$$

Furthermore, since $\mathbb{E}(Y - G(W)) = 0$, it follows from the Law of Iterated Expectations that

$$\dot{S}(\beta_o, \gamma_o)[\gamma] = \mathbb{E}\left[-\frac{\tau\partial_\beta G(W)}{G(W)(1 - G(W))}\gamma_3(W) - \frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1 - G(W))}\gamma_4(X^e, Z)\right].$$

The two inequalities then follow immediately by using that under our assumptions the functions involved satisfy a Lipschitz property. □

Lemma 6 (Condition (2.4))**.** *$\hat{\gamma} \in \Gamma$ with probability tending to one; and $\|\hat{\gamma} - \gamma_o\|_\Gamma = o_p(n^{-1/4})$.*

Proof. The first part follows directly from the definition of the estimators and the smoothness conditions imposed on the kernel function, whereas the second part is a consequence of Lemma 2 and Assumption 8. □

Lemma 7 (Condition (2.5'))**.** *For all sequences of positive numbers $\{\delta_n\}$ with $\delta_n = o(1)$,*

$$\sup_{\|\beta-\beta_o\|\leq\delta_n, \|\gamma-\gamma_o\|\leq\delta_n} \|S_n(\beta,\gamma) - S(\beta,\gamma) - S_n(\beta_o,\gamma_o)\| = o_p(n^{-1/2})$$

Proof. This statement follows from Theorem 3 in Chen, Linton, and Van Keilegom (2003). To verify the conditions of that theorem one first has to show that

$$\mathbb{E}\left(\sup_{\|\bar{\beta}-\beta\|<\delta, \|\bar{\gamma}-\gamma\|_\Gamma<\delta} \left|s(Y,X,Z,\bar{\beta},\bar{\gamma}) - s(Y,X,Z,\beta,\gamma)\right|^2\right) \leq K\delta^2$$

for all $(\beta,\gamma) \in \mathcal{B}\times\Gamma$, all $\delta > 0$ and some constant $K > 0$. This follows from the differentiability of the functions of which s is composed and the mean value theorem.

Secondly, one has to show that

$$\int_0^\infty \sqrt{\log N(\lambda, \Gamma, \|\cdot\|_\Gamma)}d\lambda < \infty,$$

where $N(\lambda,\Gamma,\|\cdot\|_\Gamma)$ is the minimal number of balls with $\|\cdot\|_\Gamma$-radius λ needed to cover Γ. This is a consequence of a result in van der Vaart and Wellner (1996, Corollary 2.7.4) and Assumption 8. □

Lemma 8 (Condition (2.6))**.**

$$\sqrt{n}(S_n(\beta_o,\gamma_o) + \dot{S}(\beta_o,\gamma_o)[\hat{\gamma}-\gamma_o]) \xrightarrow{d} N(0,\Omega)$$

Proof. Note that as shown in the proof of Lemma 5, we have that

$$\dot{S}(\beta_o,\gamma_o)[\gamma] = -\mathbb{E}\left(\frac{\mathbb{E}(\tau\partial_\beta G(W)|W)}{G(W)(1-G(W))}\gamma_3(W)\right) - \mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}\gamma_4(X^e,Z)\right),$$

and hence

$$\begin{aligned}\dot{S}(\beta_o,\gamma_o)[\hat{\gamma}_o-\gamma_o] &= -\mathbb{E}\left(\frac{\mathbb{E}(\tau\partial_\beta G(W)|W)}{G(W)(1-G(W))}(\hat{G}(W|\hat{v}) - G(W))\right) \\ &\quad - \mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}(\hat{V}-V)\right) \\ &\equiv -A_1 - A_2.\end{aligned}$$

To simplify the notation, let $t(w) = \mathbb{E}(\tau \partial_\beta G(W)|W = w)/(G(w)(1 - G(w)))$. Then we have that

$$\begin{aligned}
A_1 &= \int t(w)(\hat{G}(w|\hat{v}) - G(w))D(w)dw \\
&= \int t(w)((\hat{N}(w|\hat{v}) - N(w)) - \frac{N(w)}{D(w)}(\hat{D}(w|\hat{v}) - D(w)))dw + o_p(n^{-1/2}) \\
&= \int t(w)(\hat{N}(w) - N(w))dw && \text{(A.1)} \\
&- \int t(w)G(w)(\hat{D}(w) - D(w))dw && \text{(A.2)} \\
&+ \int t(w)(\hat{N}(w|\hat{v}) - \hat{N}(w))dw && \text{(A.3)} \\
&- \int t(w)G(w)(\hat{D}(w|\hat{v}) - \hat{D}(w))dw + o_p(n^{-1/2}) && \text{(A.4)}
\end{aligned}$$

Now consider the term in (A.1). Due to the use of higher-order kernels, the difference between $N(w)$ and $\mathbb{E}(\hat{N}(w))$ is of the order $o(n^{-1/2})$ uniformly in w. Hence

$$\begin{aligned}
\int t(w)(\hat{N}(w) - N(w))dw &= \int t(w)(\hat{N}(w) - \mathbb{E}(\hat{N}(w)))dw + o(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \int t(w)(Y_i K_h(W_i - w) - \mathbb{E}(Y_i K_h(W_i - w))dw + o(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} t(W_i)Y_i - \mathbb{E}(t(W)G(W)) + o_p(^{-1/2})
\end{aligned}$$

where the last equality follows from standard change-of-variables and Taylor-expansion arguments. Similarly, one obtains for the term in (A.2) that

$$\int t(w)G(w)(\hat{D}(w) - D(w))dw = \frac{1}{n}\sum_{i=1}^{n} t(W_i)G(W_i) - \mathbb{E}(t(W)G(W)) + o_p(^{-1/2}).$$

Next, inserting the definition of the respective estimators, we obtain for the term in (A.3) that

$$\begin{aligned}
\int t(w)(\hat{N}(w|\hat{v}) - \hat{N}(w))dw &= \frac{1}{n}\sum_{i=1}^{n} Y_i \int t(w)(K_h(X_i\beta_o - w_1, \hat{V}_i - w_2) - K_h(X_i\beta_o - w_1, V_i - w_2))dw \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i \int t(X_i\beta_o - rh, V_i - sh)(K(r, s + (\hat{V}_i - V_i)/h) - K(r, s))drds \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{V}_i - V_i)/h \int t(X_i\beta_o - rh, V_i - sh)\partial_s K(r, s)drds + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{V}_i - V_i) \int \partial_2 t(X_i\beta_o - rh, V_i - sh)K(r, s)drds + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i(\hat{V}_i - V_i)\partial_2 t(W_i) + o_p(n^{-1/2}),
\end{aligned}$$

where the 2nd to 5th line follow by substitution, a Taylor expansion of the kernel, partial integration, and the higher order property of the kernel, respectively. The last expression is then equal to

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \omega_n(Z_i, Z_j)\psi_j Y_i \partial_2 t(W_i) + o_p(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(\omega_n(Z, Z_i)\mathbb{E}(Y|X, Z)\partial_2 t(W)|Z_i)\psi_i + o_p(n^{-1/2})$$

using Assumption 8 and common projection arguments for U-statistics. Finally, one can use similar arguments to show that the term in (A.4) is equal to

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\omega_n(Z,Z_i)\partial_2 l(W)|Z_i)\psi_i + o_p(n^{-1/2}),$$

where $l(w) = G(w)t(w)$, and thus the terms in (A.3)–(A.4) are equal to $n^{-1}\sum_{i=1}^{n}\xi_{2i}\psi_i + o_p(n^{-1/2})$ since $\mathbb{E}(Y|X,Z) = G(W)$.

Now consider the term A_2. It follows directly from Assumption 8 that

$$\begin{aligned}
A_2 &= \mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}(\hat{v}(X^e,Z) - v(X^e,Z)\right) \\
&= \int \frac{\tau\partial_\beta G(x\beta_o, v(x,z))\partial_2 G(x\beta_o, v(x,z))}{G(x\beta_o, v(x,z))(1-G(x\beta_o, v(x,z)))}\frac{1}{n}\sum_{i=1}^{n}\omega_n(z,Z_i)\psi_i dF_{X,Z}(x,z) + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{\tau\partial_\beta G(W)\partial_2 G(W)}{G(W)(1-G(W))}\omega_n(Z,Z_i)|Z_i\right)\psi_i + o_p(n^{-1/2}) \\
&= \frac{1}{n}\sum_{i=1}^{n}\xi_{1i}\psi_i + o_p(n^{-1/2}),
\end{aligned}$$

where $F_{X,Z}$ is the joint CDF of (X,Z). Taken together, we have shown so far that

$$\begin{aligned}
&\sqrt{n}(S_n(\beta_o,\gamma_o) + \dot{S}(\beta_o,\gamma_o)[\hat{\gamma}-\gamma_o]) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{Y_i - G(W_i)}{G(W_i)(1-G(W_i))}(\tau_i\partial_\beta G(W_i) - \mathbb{E}(\tau_i\partial_\beta G(W_i)|W_i)) \\
&+ \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\xi_{1i}-\xi_{2i})\psi_i + o_p(1).
\end{aligned}$$

The statement of the Lemma then follows from applying an ordinary CLT, since ψ_i and $Y_i - G(W_i)$ are orthogonal. □

Proof of Theorem 3. The results in Theorem 2 and Lemma 3 – 8 imply that the conditions of Theorem 2 in Chen, Linton, and Van Keilegom (2003) are fulfilled, which in turn implies the statement of the theorem. □

References

Ahn, H., H. Ichimura, and J. Powell (1996): "Simple Estimators for Monotone Index Models," *manuscript, Department of Economics, UC Berkeley.*

Ai, C. (1997): "A Semiparametric Maximum Likelihood Estimator," *Econometrica*, 65(4), 933–963.

Blundell, R., and J. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, 2.

——— (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71(3), 655–679.

Blundell, R., and J. Powell (2007): "Censored Regression Quantiles with Endogenous Regressors," *Journal of Econometrics*, 141(1), 65–83.

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

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71(5), 1405–1441.

Das, M., W. Newey, and F. Vella (2003): "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70(1), 33–58.

Delecroix, M., M. Hristache, and V. Patilea (2005): "On Semiparametric M-estimation in Single-Index Regression," *Journal of Statistical Planning and Inference*, 136(3), 730–769.

Florens, J., J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76(5), 1191–1206.

Horowitz, J., and W. Härdle (1996): "Direct Semiparametric Estimation of Single-Index Models with Discrete Covariates.," *Journal of the American Statistical Association*, 91(436), 1632–1640.

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

Ichimura, H. (1993): "Semiparametric least squares(SLS) and weighted SLS estimation of single-index models," *Journal of Econometrics*, 58(1-2), 71–120.

Ichimura, H., and P. Todd (2007): "Implementing Nonparametric and Semiparametric Estimators," *Handbook of Econometrics*, 6.

Imbens, G., and W. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, fortcoming.

Jones, M., and D. Signorini (1997): "A Comparison of Higher-Order Bias Kernel Density Estimators.," *Journal of the American Statistical Association*, 92(439).

Klein, R., and R. Spady (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica*, 61(2), 387–421.

Kong, E., O. Linton, and Y. Xia (2009): "Uniform Bahadur Representation for Local Polynomial Estimates of M-regression and its Application to the Additive Model," *Econometric Theory*, in press.

Lee, L. (1995): "Semiparametric Maximum Likelihood Estimation of Polychotomous and Sequential Choice Models," *Journal of Econometrics*, 65(2), 381–428.

Lee, S. (2007): "Endogeneity in quantile regression models: A control function approach," *Journal of Econometrics*, 141(2), 1131–1158.

Lewbel, A. (2000): "Semiparametric Qualitative Response Model Estimation with Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97(1), 145–177.

Linton, O., S. Sperlich, and I. Van Keilegom (2008): "Estimation of a Semiparametric Transformation Model," *Annals of Statistics*, 36(2), 686–718.

Manski, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3(3), 205–228.

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

Newey, W. (1985): "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Annales de l'INSEE*, 59(60), 219–235.

——— (1987): "Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," *Journal of Econometrics*, 36(3), 231–250.

——— (2007): "Nonparametric Continuous/Discrete Choice Models," *International Economic Review*, 48(4), 1429–1439.

Newey, W., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W., J. Powell, and F. Vella (1999): "Nonparametric Estimation of Triangular Simulataneous Equations Models," *Econometrica*, 67(3), 565–603.

Pakes, A., and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027–1057.

Powell, J., J. Stock, and T. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57(6), 1403–1430.

Rivers, D., and Q. Vuong (1988): "Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models," *Journal of Econometrics*, 39(3), 347–366.

Smith, R., and R. Blundell (1986): "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply," *Econometrica*, 54(3), 679–686.

Stoker, T. (1986): "Consistent Estimation of Scaled Coefficients," *Econometrica*, 54(6), 1461–1481.

van der Vaart, A., and J. Wellner (1996): *Weak Convergence and Empirical Processes*. Springer.

Xia, Y. (2006): "Asymptotic Distributions For Two Estimators Of The Single-Index Model," *Econometric Theory*, 22(06), 1112–1137.

Xia, Y., H. Tong, W. Li, and L. Zhu (2002): "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society Series B(Statistical Methodology)*, 64(3), 363–410.