Klein, Roger; Vella, Francis

Working Paper

A semiparametric model for binary response and continuous outcomes under index heteroscedasticity

IZA Discussion Papers, No. 2383

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Klein, Roger; Vella, Francis (2006) : A semiparametric model for binary response and continuous outcomes under index heteroscedasticity, IZA Discussion Papers, No. 2383, Institute for the Study of Labor (IZA), Bonn, https://nbn-resolving.de/urn:nbn:de:101:1-20090406149

This Version is available at:
https://hdl.handle.net/10419/33942

DISCUSSION PAPER SERIES

IZA DP No. 2383

A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity

Roger Klein
Francis Vella

October 2006

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity

Roger Klein
Rutgers University

Francis Vella
Georgetown University
and IZA Bonn

Discussion Paper No. 2383
October 2006

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 2383
October 2006

ABSTRACT

A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity*

This paper formulates a likelihood-based estimator for a double index, semiparametric binary response equation. A novel feature of this estimator is that it is based on density estimation under local smoothing. While the proofs differ from those based on alternative density estimators, the finite sample performance of the estimator is significantly improved. As binary responses often appear as endogenous regressors in continuous outcome equations, we also develop an optimal instrumental variables estimator in this context. For this purpose, we specialize the double index model for binary response to one with heteroscedasticity that depends on an index different from that underlying the "mean-response". We show that such (multiplicative) heteroscedasticity, whose form is not parametrically specified, effectively induces exclusion restrictions on the outcomes equation. The estimator developed below exploits such identifying information. We provide simulation evidence on the favorable performance of the estimators and illustrate their use through an empirical application on the determinants, and affect, of attendance at a government financed school.

JEL Classification: C35, C14

Keywords: binary response, semiparametric, heteroscedasticity, endogenous treatment

Corresponding author:

Francis Vella
Department of Economics
Georgetown University
Washington, DC 20057
USA
E-mail: fgv@georgetown.edu

* Financial support from the Research Council at Rutgers University is acknowledged. Any errors are the sole responsibility of the authors.

1 Introduction

The last thirty years have witnessed the introduction of several estimators for the semiparametric binary response model under minimal distributional assumptions on the disturbance terms (see for example, Manski (1975, 1985), Horowitz (1992), Powell et al (1989), Ichimura (1993), and Klein and Spady (1993)). Much of the focus on relaxing distributional assumptions in the binary response model was motivated by the fact that maximum likelihood estimation of discrete choice models would generally lead to inconsistent estimates if the underlying distribution was incorrectly chosen.

In addition to the "shape" of the error distribution, it may also be misspecified in the manner in which it depends on the explanatory variables. For example, if the error exhibits multiplicative heteroscedasticity that is not a function of the "mean" response, then only the above mentioned estimators of Manski and Horowitz are consistent. However, these estimators will not recover binary response probabilities. By estimating binary quantile models, Kordas (2000) obtains interval estimates of the probabilities under general conditions. One of the main objectives of the present paper is to obtain these probabilities. We model a binary response probability as depending on a double index where the distribution of the error may depend on the explanatory variables through one or both of the indices. For example, this specification allows for, but is not restricted to, multiplicative heteroscedasticity that depends on one or both indices.

To estimate the binary response model described above, we extend the estimator in Klein and Spady (1993). The estimator in Klein and Spady depends on a single index assumption, which in the present context would imply that it can handle heteroscedasticity only if the "error" distribution depends on the same index that determines the "mean response". Here we allow a double index formulation in which the index underlying the "mean response" may differ from that upon which heteroscedasticity depends. Such an index formulation is particularly important in view of a result due to Chen and Khan (1998). They consider a binary response model where the heteroscedasticity depends on an unknown function of the explanatory variables and does not have an index structure. In this case, they show there does not exist a $\sqrt{N}$-consistent estimator for the model's parameters. Here, we will obtain a $\sqrt{N}$-consistent estimator under an index specification.

It should be emphasized that the estimator developed here depends on density estimators obtained under estimated local smoothing, where under-

lying density estimators are based on windows that vary for each observation in the sample. This is analogous to characterizing a distribution with a histogram in which the bin interval is allowed to vary depending on whether one is in the tails of the unknown density (where observations are sparse) or in regions where the true density is "high". With such local smoothing, the proofs for the asymptotic properties of the estimator formulated here substantially differ from those in the literature that employ bias-reducing kernels. We pursue this strategy first because density estimators under local smoothing have mean-squared-error optimally properties (Abramson (1982)). Second and most importantly, in the present context we have found that the finite sample performance of the estimator for the binary response model is much improved under local smoothing in contrast to bias-reducing kernels. We also found further improvements in the finite sample performance of the estimators by employing dependent kernels that depend on an estimated sample covariance matrix as advocated by Fukunaga (1972). Accordingly, all proofs in this paper are for estimation under local smoothing and dependent bivariate kernels.

In adopting the above smoothing strategy, we have found it necessary to employ a property of the derivative of semiparametric probability function due to Whitney Newey. Namely, when this derivative is taken with respect to index parameters and then evaluated at the true parameter values, it coincides with the corresponding parametric derivative minus its conditional expectation (conditioned on the indices). This "residual-type" property of this derivative function is important below in controlling the bias in gradient terms in the asymptotic normality argument. As is typical for many semiparametric estimators, we will need to downweight (trim) observations where density denominators become "too small". To exploit the residual property of the semiparametric derivative, we will employ a trimming strategy that depends on estimated indices as opposed to the explanatory variables.

The estimator developed here for the binary response model is also related to those of Ichimura and Lee (1991) and Lee (1995) who examine alternative multiple index models. While the present paper makes use of several key identification results of the Ichimura and Lee paper, it differs from both in several important respects. First, and most important, we have formulated the estimator and all proofs for the case of estimated local smoothing rather than bias-reducing kernels. Second, we make use of identification results in Ichimura and Lee without imposing exclusion restrictions on the indices. We emphasize that we are not concerned here with recovering the original para-

meters in the binary response model (which even in the presence of exclusion restrictions are still only obtained up to location and scale). Rather, we are interested in estimating those identifiable functions of the parameters that suffice to identify the semiparametric probability function. It can be argued that with binary response models, one is generally not concerned with the parameters themselves but rather with the response probability and marginal effects. Such marginal effects, which examine how the probability function changes as the explanatory variables change, are identified once the probability function is identified. Moreover, while the entire probability function converges pointwise and uniformly to the true function at a rate below the parametric rate of $\sqrt{N}$, averaged marginal effects converge at the parametric rate. The original parameter values of the model are not required for such identification. In part, for this reason we focus on identifying the probability function itself rather than index parameters.

While one of our primary objectives is to provide an estimator for this double index binary choice model,[1] we note that applied researchers have become increasingly interested in larger systems in which the choice appears in another equation as an endogenous regressor. This type of model, frequently referred to as an endogenous binary treatment model, is at best poorly identified without an exclusion restriction. The well-known problem here is that the treatment probability, which would serve as an instrument for estimating the continuous outcomes equation, is often approximately linear in its argument. In the absence of an exclusion restriction on the continuous outcome equation, the instrument is then very close to being linearly related to the same exogenous variables in the continuous equation of interest. To resolve this problem here, we consider the case of multiplicative heteroscedasticity in the binary response equation, which is some function of the explanatory variables X. Write this function as $S(X)$. In the next section we show that such heteroscedasticity may be viewed as inducing exclusion restrictions on the continuous outcomes equation. With no parametric assumptions on $S(X)$ (other than that it depends on one or two indices) and with no parametric assumptions on the distribution of the error term in the binary response model, below we will develop an estimator that exploits such identifying information. We will then show that such information is useful both in theory and in practice (as indicated in a series of monte-carlo experiments and in

[1] Virtually all of the technical difficulties in this paper arise from estimating a double index specification for the binary response model under estimated local smoothing.

an empirical application).

For continuous simultaneous equations models, other authors have exploited heteroscedasticity as an identification strategy. For example, in a semiparametric formulation, Klein and Vella (2006) exploit such information to identify and estimate triangular simultaneous equations models without exclusion restrictions. In parametric formulations, Vella and Verbeek (1997), Rummery et al (1999), Rigobon (2003) and Lewbel (2004) also exploit heteroscedasticity as an identification strategy for simultaneous equations. From the structure of the problem considered here, there is information in higher order powers of the $X's$ that could be exploited to construct instruments for the outcomes equation. Dagenais and Dagenais (1997) and Lewbel (1997) exploit such information in models with measurement error. In this paper, since the nature of the heteroscedastic function in the treatment equation is unknown, it is unclear which higher orders of the $X\prime s$ should be used as instruments. Consequently, we pursue an alternative strategy here that involves direct estimation of a double index binary response model. One could attempt to by-pass estimation of this equation and determine the appropriate higher orders of $X\prime s$ to use as instruments by extending Donald and Newey (2001) to the model considered here. However, as the treatment probability is itself of direct interest, we pursue an alternative strategy that employs the estimated treatment probability in estimating the continuous outcomes equation. In the present context, the conditional treatment probability is an optimal instrument (Amemiya (1975)).

The next section outlines the model and the estimation methods. In Section 3 we provide and discuss the assumptions required to establish asymptotic results. When estimating the treatment effect, we note that our procedure is of particular value when there are no exclusion restrictions which provide instruments. Accordingly we focus on identification in the absence of conventional exclusion restrictions. In Section 4 we establish the asymptotic properties of the estimators for both the binary response and outcome models. In so doing, we sketch out the proofs, and provide complete technical details in the Appendix. The proof strategy differs from other arguments in the literature as it relies on estimated local smoothing. Section 5 provides simulation evidence. In Section 6 we provide an empirical application where an individual's total education level (the outcome) depends in part on whether or not the individual attended a State financed high school in Australia (the treatment). Section 7 concludes.

2 Model and Motivation for Estimators

Consider the following model:

$$\begin{aligned} Y_{1i} &= X_i\beta_0 + \theta_0 Y_{2i} + u_i \quad (1) \\ Y_{2i} &= \{X_i\pi_0 + v_i > 0\}, \quad (2) \end{aligned}$$

where Y_{1i} is the outcome variable and Y_{2i} is a dummy endogenous variable defined through the indicator function $\{\bullet\}$; X_i is a vector of exogenous variables; β_0, π_0 and θ_0 are unknown true parameter values; and u_i and v_i are random disturbances. While the treatment effect, θ_0, is invariant across individuals, this assumption can be relaxed as in the empirical application. The disturbances can be characterized as:

$$v_i = S(X_i\gamma_0)v_i^* \quad (3)$$

$$E(u_i|X_i) = 0, \quad (4)$$

where $S(\bullet)$ is an unknown (positive and non-constant) function; γ_0 is an unknown parameter vector, and v_i^* is a homoscedastic random disturbance which is independent of the elements of X_i but dependent on u_i. The model allows heteroscedasticity in each equation, though we only model it explicitly in index form for the binary response model. Note that there may or may not be known restrictions on the parameters in the above model. For example, suppose $X \equiv \left[X_{[1]}, X_{[2]}\right]$, where $X_{[2]}$ contains powers and cross products of the "basis" elements in $X_{[1]}$. Then, in some formulations it will be reasonable to restrict the elements of β_0 and π_0 so that the "mean effects" only depend on $X_{[1]}$. In contrast, one may want to let heteroscedasticity, S, depend on the basis elements $X_{[1]}$ and the higher order terms $X_{[2]}$. Alternatively, we could interpret X itself as containing the "basis variables" for the model and impose no exclusion restrictions on β_0, π_0, or γ_0. Because of the aspects of the above model in which we are interested, we permit and indeed focus on this second case of no exclusion restrictions. The estimator developed here is for a model more general than above, but we will specialize to the above case for expositional convenience.

For the model in (1-4), the treatment probability has the form:

$$\begin{aligned} \Pr(Y_{2i} &= 1|X_i) = \Pr(Y_{2i} = 1|X_i\pi_0;\ X_i\gamma_0) \quad (5) \\ &\equiv P(Z_i\pi_0), \quad Z \equiv [X_i/S(X_i\gamma_0)], \end{aligned}$$

where $P(.)$ is the distribution function for v_i^*. We estimate this probability function in a double index formulation based on local smoothing. The estimator will depend neither on the functional form for S nor on the distribution of the disturbances.

We can also employ this probability function as an (optimal) instrument for estimating the continuous outcomes equation. Here we make several observations. First, if there is no heteroscedasticity in the above model, then effectively $Z = X$, in which case the model can be poorly identified because P is often approximately linear in its argument. When the argument of P is $X\pi_0$ (i.e. $Z = X$), it is still possible to identify the model provided that P is not linear in $X\pi_0$. However, this form of non-linearity in the function P itself will typically occur in the tails of the $X_i's$ and thus relies on a small fraction of the sample for identification. In contrast, in the presence of heteroscedasticity, Z no longer coincides with X and indeed will typically be linearly independent of the columns of X. Consequently, the Z-variables are effectively excluded from the continuous outcomes equation. Such induced exclusion restrictions serve to identify the model even in the region of the data for which P is linear in Z.

3 Assumptions, Identification, and Definitions

We now provide the assumptions and definitions that we employ to establish the asymptotic properties for the estimator.

A1. The Data. The data : (Y_{1i}, Y_{2i}, X_i), $i = 1, ..., N$, are i.i.d. observations from the model in (1)-(4). With X as the NxK matrix of observations on the explanatory variables and with $\underline{\mathbf{1}}$ as an $Nx1$ column vector of ones, the columns of $[X \quad \underline{\mathbf{1}}]$ are linearly independent with probability 1.

A2. Errors. The error in the continuous outcomes equation (1), u_i, is independent over i with $E(u_i|\ X_i) = 0$ and with $E[u_i^2 \mid X_i]$ uniformly bounded. The error in the binary response model (2) is given as:

$$v_i \equiv S(X_i\gamma_0)\, v_i^*,$$

where the unscaled error, v_i^*, is i.i.d. with finite variance. The scaling function $S(\bullet)$ is finite, bounded away from zero, and is not constant. The vector X_i is independent of the unscaled error v_i^*.

A3. Parameter Space. The vector of true parameters values for the model in (1-4) lies in the interior of a compact parameter space, Θ.

A4. Index Assumptions. Assume that the vector of indices, I, depends on two distinct (functionally independent) continuous variables, X_1 and X_2 . With X_3 containing all other explanatory variables, write:

$$I \equiv [I_1,\ I_2] \equiv [X_1, X_2,\ X_3] \begin{bmatrix} & \Gamma_c & \\ \Gamma_{31} & & \Gamma_{32} \end{bmatrix},\ \Gamma_c \equiv \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix},$$

and assume that the 2x2 submatrix Γ_c has rank 2.

A5. Reparameterized Model. With $\eta \equiv (\eta_{31},\ \eta_{32})$, define:

$$W \equiv I * \Gamma_c^{-1} \equiv [W_1, W_2] \equiv [X_1, X_2,\ X_3] \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \eta_{31} & \eta_{32} \end{bmatrix} \equiv X\beta(\eta).$$

Under this reparameterization, notice that $P(Y = 1|I) = P(Y = 1|W)$. Define W^* by replacing η above with η^*. With x a realized value of X, write $w \equiv x\beta(\eta)$ and $w^* \equiv x\beta(\eta^*)$. Assume:

$$P(w) \equiv P(Y = 1|W = w) = P(Y = 1|W^* = w^*) \equiv P^*(w^*).$$

Following Ichimura and Lee (1991), let $t \equiv w^* = (x_1, x_2)+x_3\eta_3^*$. Then, write:

$$P(w) \equiv P(t + x_3(\eta - \eta^*)) = P^*(t).$$

Assume that there exists a set of positive probability on which the above equality may be differentiated with respect to the continuous elements of x_3 with t held fixed. Further assume that condition (4) of Ichimura and Lee (1991, Lemma 3) holds.

A6. Densities. Assume that all continuous variables have compact support. To provide required smoothness conditions, let $X_c \equiv (X_1, X_2)$ be the vector of continuous variables in (A5). Then, with $f(x_c|X_3, Y_2)$ as the indicated conditional density for X_c, denote $\nabla_1^i\nabla_2^j f(\bullet|\bullet)$ as the i^{th} and j^{th} cross-partial with respect to the elements of $x_c \equiv [x_1, x_2]$. Then, with $\nabla_1^0\nabla_2^0 f(x_c|\bullet) \equiv f(\bullet|\bullet)$, assume that $f(w|\bullet)$ has positive support on a compact set $\mathcal{A}$, is bounded away from 0 on any compact subset of its support, and that on $\mathcal{A}$ $\left|\nabla_1^i\nabla_2^j f(\bullet|\bullet)\right|$ is bounded above by a positive finite constant for $i + j \leq 4$.

Assumptions **A1-3** define the index model that we propose to estimate. An index formulation of low dimension is and important for obtaining reasonable results in finite samples. Notice that this index assumption permits a more general error structure than that shown in **(A2)**. Namely, we require that the binary response probability depend on two indices, but do not otherwise restrict the manner in which the probability depends on the indices. The particular double index structure implicit in (**A2)** provides a convenient motivating case.

With the possible exception of assumptions **A4-5,** the above assumptions are somewhat standard in index models. Assumptions **A4-5** essentially provides identification conditions. To motivate these assumptions**,** note that the W-parameterization in **(A5)** is equivalent to the I-parameterization in **(A4)** as both yield the same conditional probability function in x. We employ the W-parameterization to allow for the possibility that there may not be exclusion restrictions in the original I-parameterization. In this lower dimensional parameterization, we then seek to identify the (nuisance) parameters η. Before proceeding, we note that these parameters have no natural interpretation as they are linear functions of the model's original parameters. However, if these parameters are identified, we can easily recover the binary response probability function and identify the marginal effects which measure how the response probability changes in response to changes in x. Moreover, asymptotic properties for these estimated marginal effects will readily follow from those for $\hat{\eta}$. Finally, as elaborated below, the probability function is of interest in estimating a continuous outcomes equation that depends on the binary response variable.

Having reparameterized the model in (A5) we then assume that the W-parameterization satisfies the identification conditions in Ichimura and Lee (1991).[2] The condition on discrete variables is that given by Ichimura and

[2]Ichimura and Lee use this differentiability condition in their proof. We have explicitly stated it as an assumption, because it can fail when all continuous variables of an index are functionally related. For example, suppose $x_3 = x_1^2$ and write

$$t_1 = x_1 + x_1^2\eta_1.$$

The derivative condition in (A5) does not hold for this case. Moreover, the model is not identified as

$$\Pr(Y = 1|W = w) = \Pr(Y = 1|X_1 = x_1, W_2 = w_2).$$

In other words, we are unable to distinguish η_{31} from $\eta_{31}^* = 0$ as both yield the same binary response probability.

Lee to identify their coefficients. Note that these identification conditions are based on the underlying assumption of a double index model. In presenting simulation results, we will present results both for double and single index models. If a single index model generates the data, it will not be possible to identify all of the parameters of a double index specification. However, it is still possible to identify the probability function of interest. As the focus of this paper is on a double index specification we defer further discussion of this issue to the simulation section.

Assumption **(A6)** provides smoothness conditions. These conditions and densities satisfying them are discussed in Klein and Spady (1993, p. 393). It is possible to relax the compact support assumption at some technical expense in the proofs.[3]

In addition to the above assumptions, we also need a number of conditions or definitions that define the densities and probability functions of interest. Throughout, we employ kernel density estimators to estimate the semiparametric probability function entering a quasi likelihood. As is standard in this literature, such density estimators need to have an appropriately low order of bias. Here, we obtain bias reduction first by employing local smoothing as developed by Abramson (1982) and discussed in Silverman (1986). Such local smoothing requires that the windows in the final kernel density estimator vary by observation and depend on a pilot density estimator. Not surprisingly, these windows satisfy the intuitive requirement that they be smaller in the center of the distribution than in the tails. As a second source of bias reduction, we exploit a property of expected semiparametric probability derivatives. Namely, such derivatives have expected value zero when conditioned on the true indices. As will also be discussed below, to improve the finite sample performance of the estimators, we estimate the density for the vector of indices, W, using kernels that depend on the sample covariance

A similar issue arises in the case of single index models. The identification argument in Klein and Spady (1993) requires that there is a continuous variable that is functionally independent of other continuous variables in the model. Ichimura (1993) provides weaker identification conditions by relaxing this assumption. It remains the case, however, that when the index is a linear in "basis" functions of the same continuous variable, Z, then the index is not identified.

[3]This assumption is used in two types of arguments. First, in conjunction with the parameters being in a compact set, it implies that probabilities are bounded away from one and zero. In the absence of this simplifying condition, one would need to make a tail assumption on how fast the probability function tends to one or zero. Second, this compact support assumption simplifies various uniform convergence arguments.

matrix for W. Below, we will first define these estimators and then discuss their properties.

D1. Density Estimators Under Local Smoothing. Let K be a symmetric, smooth univariate kernel function satisfying condition C8 in Klein and Spady (1993, 394). The normal kernel, which is employed in the simulations and the empirical example, satisfies this condition. Let T be a matrix such that $T'T = \hat{\Sigma}_s^{-1}$, the inverse sample covariance matrix for W given that $Y_2 = s, s = 0, 1$. Partitioning $T = [T_1\ T_2]'$ conformably with the i^{th} observation on W: $W_i = [W_{1i}\ W_{2i}]'$, define:

$$k_j^s(w; h, \lambda) \equiv \frac{\det\left(\hat{\Sigma}_s\right)^{-1/2}}{[\lambda_j h]^2} K\left(T_1\left[w - W_j\right] / [\lambda_j h]\right) K\left(T_2\left[w - W_j\right] / [\lambda_j h]\right).$$

With $g_s(w)$ as joint density for $W \equiv [W_1, W_2]$ conditioned on $Y_2 = s$, $s = 0, 1$, and with P_s as the unconditional probability that $Y_2 = s$, define an estimator for $f_s(w) \equiv P_s g_s(w)$ as:

$$\hat{f}_1(w; h, \lambda) \equiv \frac{1}{N} \sum_j Y_{2j} k_j^1(w; h, \lambda); \quad \hat{f}_0(w; h, \lambda) \equiv \frac{1}{N} \sum_j (1 - Y_{2j}) k_j^0(w; h, \lambda).$$

For $w = W_i$, the above averages are taken over the N-1 observations for which $j \neq i$.

D2: Smooth Trimming Functions. Define a smooth trimming functions as:

$$\tau(z; a) \equiv [\ 1 + \exp\left(N^a[z]\right)]^{-1}.$$

D3. Estimated Local Smoothing Parameters. Referring to (D1), denote $\hat{m}_s$ as the geometric mean of the $\hat{f}_s(w; h, \lambda)'s$ and let $\hat{\gamma}_{sj} \equiv \left[\hat{f}_s(w_j; h, \lambda) / \hat{m}_s\right]$. Then, for $j = 1, ..., N$, define estimated local smoothing parameters as:

$$\begin{aligned}\hat{\lambda}_{sj} &\equiv \hat{\lambda}_s\left(\hat{f}_s\left(w_j;h,\lambda\right)\right) = \left[\hat{d}_{sj}\hat{\gamma}_{sj} + \left(1-\hat{d}_{sj}\right)/Ln(N)\right]^{-1/2}, \\ \hat{d}_{sj} &\equiv \tau\left(\frac{1}{Ln\left(N\right)} - \hat{\gamma}_{sj}; .01\right),\end{aligned}$$

where the parameter a in (D2) is set here to .01.[4]

D4. Multi-Stage Local Smoothing. Employing (D3), the estimator for $f_s\left(w\right)$ is defined under several stages of local smoothing as:

$$\begin{aligned}\hat{f}_s(w) &\equiv \frac{1}{N}\sum_{j\neq i} Y_{2j}k_{ij}^s\left(h_3, \hat{\lambda}_s^*\right),\ s=1,0, \\ \hat{\lambda}_{sj}^* &\equiv \hat{\lambda}_{sj}\left(\hat{f}_1\left(w_j;h_2,\hat{\lambda}_{sj}\left(\hat{f}_1\left(w_j;h_1,\mathbf{1}\right)\right)\right)\right),\end{aligned}$$

where $\mathbf{1}$ is a vector of ones. With $h_i = O(N^{-r_i})$, set $r_3 = 1/11$ and $0 < \delta < r_3/2$. Then, set $r_2 = \left(r_3 - \delta/2\right)/2$, and $r_1 = \left(r_3 - \delta\right)/4$.[5]

D5. Semiparametric Probability Function. Define:

$$\hat{P}\left(\eta\right) \equiv \hat{f}_1^*(w)/\hat{g}^*\left(w\right) \equiv \left[\hat{f}_1(w) + \hat{\Delta}_{1N}\right] / \left[\hat{g}\left(w\right) + \hat{\Delta}_N\right],$$

where $\hat{g}\left(w\right) \equiv \hat{f}_1(w) + \hat{f}_0(w)$ estimates the unconditional density for W. To define the Δ adjustment-factors, first define the smoothed indicator :

$$\Delta_{sN} \equiv \hat{c}_s N^{\varepsilon'}\left[\ 1 + \exp\left(N^{a_1}\left[\hat{f}_s(w) - N^{-a_2}\right]\right)\right]^{-1},$$

where $a_1 \equiv \varepsilon' r_3/4$, $a_2 \equiv \varepsilon' r_3/5$, $\hat{c}_s = O_p(1)$, and $\Delta_N \equiv \Delta_{1N} + \Delta_{0N}$.

[4] In taking fourth-order Taylor series expansions to examine bias terms, the fourth derivative will involve N^{4a}. Consequently, it is important that a be "small". Here, $a = (r_3 - \varepsilon_a)/8$, with ε_a positive and arbitrarily small. The value $a = 1/100$ satisfies the required constraint and is employed in the monte-carlo study.

[5] In proving Lemma 8, we require $r_3 < 1/10$, $4\left(r_1 + r_2 + r_3\right) > 1/2$, and $0 < \delta < r_3/2$. In making a bias calculation (Lemma 3A-C), we will require $0 < r_1 < r_2 - 2a$ and $0 < r_1 + r_2 < r_3 - 2a$. These conditions are satisfied with the parameter a set as in (D4), r_i as in (D5), and δ as in (D5).

D6. Pilot Estimator. Let $\underline{x}_k$ be the lower α^{th} sample quantile for the continuous variable X_k (e.g. $\alpha = .01$) and let $\bar{x}_k$ be the upper $(1-\alpha)^{th}$ sample quantile. For the K_c continuous variables, define the indicators: $\hat{t}_{ik} \equiv \{\underline{x}_k < x_{ik} < \bar{x}_k\}, k = 1, ..., K_c$. In the notation of D1, define a pilot probability estimator as:

$$\hat{P}^*(\eta) \equiv \hat{f}_1(w_j; 1/11, \mathbf{1}) / \left[\hat{f}_1(w_j; 1/11, \mathbf{1}) + \hat{f}_0(w_j; 1/11, \mathbf{1})\right].$$

Then, with $\hat{\tau}_{xi} \equiv \Pi \hat{t}_{ik}$, the pilot estimator for η_0 is defined as:

$$\begin{aligned} \hat{\eta}_p &\equiv \arg\max_{\eta} \hat{l}_p(\eta), \\ \hat{Q}_p(\eta) &\equiv \sum \hat{\tau}_{xi} \left[Y_{2i} Ln\left(\hat{P}_i^*\right) + (1 - Y_{2i}) Ln\left(1 - \hat{P}_i^*\right)\right]. \end{aligned}$$

D7: Final Estimator. With $\hat{\eta}_p$ defined in (D6), let $\hat{w}_j \equiv \left[w_{1j}\left(\hat{\eta}_p\right) \quad w_{2j}\left(\hat{\eta}_p\right)\right]$ denote the vector of estimated indices. Denote $\underline{w}_1\left(\hat{\eta}_p\right)$ as the lower β^{th} sample quantile for the $w_{1j}\left(\hat{\eta}_p\right)'s$ and let $\bar{w}_1\left(\hat{\eta}_p\right)$ be the corresponding upper $(1-\beta)^{th}$ quantile. With τ as the trimming function in (D2), the index trimming function is defined as:

$$\begin{aligned} \hat{\tau}_{wi} &\equiv \hat{\tau}_{1i}\, \hat{\tau}_{2i}, \\ \hat{\tau}_{ki} &\equiv \tau\left(\underline{w}_1\left(\hat{\eta}_p\right) - \underline{w}_k\left(\hat{\eta}_p\right); 1/12\right)\ \tau\left(, w_{ki}\left(\hat{\eta}_p\right) - \overline{w}_k\left(\hat{\eta}_p\right); 1/12\right) \text{ for } k = 1, 2. \end{aligned}$$

Then, with probabilities defined in (D5), the final estimator for η_0 is defined as:

$$\begin{aligned} \hat{\eta} &\equiv \arg\max_{\eta} \hat{l}(\eta), \\ \hat{Q}(\eta) &\equiv \sum \hat{\tau}_i \left[Y_{2i} Ln\left(\hat{P}_j\right) + (1 - Y_{2i}) Ln\left(1 - \hat{P}_j\right)\right]. \end{aligned}$$

Before discussing the role of the above definitions, as an overview note that there are two general aspects that need to be addressed in estimating semiparametric models. First, it is necessary to control the bias in the underlying density estimators. As discussed below, here we control this bias by employing local smoothing and exploiting a "residual" property of semiparametric probability functions. Second, it is necessary to downweight or trim those observations for which densities become too small. For reasons

discussed below, we employ a trimming strategy outlined in (D4-6) that is quite similar to that in Klein and Spady (1993).

In explaining why we have defined various estimators as above, turn first to (D1). As discussed by Silverman (1986) and advocated by Fukunaga (1972) we have employed bivariate Kernels based on a sample covariance matrix. We "match" this feature of the data as follows. Following Fukunaga (1972) we specify a density estimate for the vector W by first constructing the standardized vector $W^* \equiv TW$. With the covariance matrix for W^* being the identity matrix, the density estimator for W^* is then somewhat naturally based on a product of independent kernels. The implied density estimator for W is then that given above. Fukunaga (1972) documents the performance of this estimator in a monte-carlo study. Here, we have found that we obtain "better" estimates of the parameters of interest when we select a density estimator in this manner.

For known local smoothing parameters (bounded away from zero), Abramson showed that the locally-smoothed density estimator is optimal in a mean-squared error sense. This estimator also has the desired bias reducing properties. As the local smoothing parameters are not known, they must be estimated. In using the estimates, we are able to prove that the resulting density estimators have desired bias reducing properties when estimated in several stages. Namely, first employ a regular kernel density estimator ($\lambda = \mathbf{1}$ in the above notation) to construct estimated local smoothing parameters. Second, obtain a density estimator using these estimated local smoothing parameters. Third, and finally, use this second stage estimator to reconstruct estimated local smoothing parameters and obtain the final density estimator shown in (D4). We have been able to show "essentially" that the bias is reduced at each stage. At the third stage, the order of the bias is $O\left(h_3^2 h_2^2 h_1^2\right)$. This order is sufficiently small to obtain the asymptotic results below. We note that this order is larger than that would be the case if the local smoothing parameters were known . For the windows required to obtain the above order for the bias, $h_3^2 h_2^2 h_1^2 > h_3^4$, the order that Abramson and Silverman establish for known local smoothing parameters (bounded away from zero). For technical reasons, we smoothly trim in (D2) so as to keep the local smoothing parameters above $1/ln(N)$.[6]

[6]For bias reasons, it is important to let the densities that define these parameters be closer to zero than the densities upon which the semiparametric probability function is based. As the likelihood trimming insures that densities have a lower bound of the form

The proofs exploit a residual-like property of the derivative (with respect to the parameters) of the true semiparametric probability function , with this derivative having conditional expectation of zero when evaluated at the true parameter values By using this property, we can further control for the bias in the gradient to the objective function, which is essential to establishing asymptotic normality. To this end, we first estimate the model under X-trimming. The resulting parameter estimates, which we do not require to be $\sqrt{N} - convergent$, are employed to obtain estimated indices or index-densities. The model is then re-estimated with trimming based on estimated indices or their corresponding estimated densities. Such trimming affords "protection" against small denominators when analyzing the gradient as it will be evaluated at the true parameter values. However, this type of trimming is problematic for analyzing the averaged Log-likelihood and the Hessian matrix as we need to examine these components away from the truth. As in Klein and Spady (1993), we employ the Δ adjustment factors in (D5) above for this purpose. These factors will vanish exponentially provided the density is not "too small". In this manner, such factors will quickly vanish from the gradient where they are not needed, but will serve to control density denominators when analyzing likelihood and Hessian components.

4 Asymptotic Results

In this section we provide and discuss the asymptotic properties for the estimator for both equations in the endogenous treatment model defined above. The Appendix contains formal proofs for all required intermediate lemmas and the main theorems given below. For expositional and notational purposes we will consider the more difficult case in which every index in the model depends on a linear combination of variables in X. In practice there will certainly be cases in which exclusion restrictions for the various indices are justified. In what follows, we begin with the binary response model and establish consistency using standard uniform convergence arguments. We then turn to the proofs for asymptotic normality.

$B > 0$, we permit the local smoothing parameters to slowly tend to zero.

4.1 Binary Response

To show that the proposed estimator for the binary response model is consistent, denote $\eta \equiv [\eta_1,\ \eta_2]$ as the (nuisance) "reduced form" parameters entering W_1 and W_2 as above. Then, with the quasi likelihood given by $\hat{Q}$ above, the estimator for this binary response model is given as:

$$\hat{\eta} = \underset{\eta}{\arg\sup} \hat{Q}(\eta).$$

With the semiparametric probability function given as $\hat{P}_i(\eta)$ in (D5):

$$\hat{Q}(\eta) \equiv \frac{1}{N}\sum_{i=1}^{N} \hat{\tau}_i \left(Y_{2i} Ln\left[\hat{P}_i(\eta)\right] + [1 - Y_{2i}]\, Ln\left[1 - \hat{P}_i(\eta)\right] \right),$$

where $\hat{\tau}_i$ is a trimming function that is defined and discussed in the Appendix.

Obtain $Q(\eta)$ from $\hat{Q}(\eta)$ by replacing $\hat{P}_i(\eta)$ with its uniform probability limit, $P_i(\eta)$. It can then be shown (see the Appendix) that

$$\hat{Q}(\eta) - Q(\eta) \xrightarrow{p} 0, \text{ uniformly in } \eta.$$

From standard uniform convergence arguments, $Q(\eta)$ converges in probability and uniformly in η to its expectation, $E[Q(\eta)]$. Under conditions for identification given above, $E[Q(\eta)]$ is uniquely maximized at η_0. Therefore, we have established:

Theorem 1 *Under (A1-6) and (D1-7):*

$$\hat{\eta} = \eta_0 + o_p(1).$$

Note that consistency of $\hat{\eta}$ will imply that the probability function is also consistently estimated. It is then also possible to establish consistency for estimated marginal effects. The asymptotic distribution for marginal effects readily follows from that for the estimated nuisance parameters. In the remainder of this section, we outline the normality argument, with the Appendix providing detailed proofs.

As discussed in the Appendix, we first employ fixed trimming on the basis of the X-variables to obtain a convergence rate for a pilot estimator for the parameters:

$$\left[\hat{\eta}_p - \eta_0\right] = O_p\left(N^{-r_p}\right),\ r_p > r_3.$$

Using $\hat{\eta}_p$, we then estimate the two W-indices and construct a smooth trimming function based on these indices. With $\hat{w}_i$ as the estimator for the indices, denote $\hat{\tau}_{wi}$ as the estimated trimming function under Index-trimming (D7). Employing this trimming function , write the objective function as:

$$\hat{Q}(\eta) \equiv \frac{1}{N} \sum \hat{\tau}_{wi} \left(Y_{2i} Ln \left[\hat{P}_i(\eta) \right] + (1 - Y_{2i}) Ln \left[1 - \hat{P}_i(\eta) \right] \right).$$

Denote $\hat{G}(\eta)$ and $\hat{H}(\eta)$ as the Gradient and Hessian for this objective function. Let $Q(\eta)$ be the objective function obtained by replacing estimated with true probability functions (to which the estimated functions uniformly tend) and denote $H(\eta)$ as the Hessian for $Q(\eta)$. Then, with the Appendix containing the details, from a standard Taylor series expansion of $\hat{G}(\hat{\eta})$ about η_o:

$$\begin{aligned} \sqrt{N}(\hat{\eta} - \eta_o) &= -H\left(\eta^+\right)^{-1} \sqrt{N} \hat{G}(\eta_o) + o_p(1), \ \eta^+ \ \epsilon \ [\hat{\eta}, \eta_o] \\ &\equiv -H_o^{-1} \sqrt{N} \hat{G}(\eta_o) + o_p(1), \ H_o \equiv E[H(\eta_o)]. \end{aligned}$$

The normality result then follows from an analysis of the gradient term.

To outline the argument for the gradient, define an estimated weight function

$$\hat{\rho}_i = \left[\frac{\partial}{\partial \eta} \hat{P}_i(\eta_0) \right] / \left[\ \hat{P}_i(\eta_0) \left[\ 1 - \hat{P}_i(\eta_0) \right] \ \right].$$

The gradient to the objective function is then of the form:

$$N^{1/2} \hat{G} = [A_1 + A_2 - B_1 + B_2],$$

$$\begin{aligned} A_1 &= N^{-1/2} \sum \tau_{wi} [\ Y_i - P_i] \hat{\rho}_i; \quad A_2 = N^{-1/2} \sum [\hat{\tau}_{wi} - \tau_{wi}] [\ Y_i - P_i] \hat{\rho}_i \\ B_1 &= N^{-1/2} \sum \tau_{wi} \left[\hat{P}_i - P_i \right] \hat{\rho}_i; \quad B_2 = N^{-1/2} \sum [\hat{\tau}_{wi} - \tau_{wi}] \left[\hat{P}_i - P_i \right] \hat{\rho}_i. \end{aligned}$$

To simplify A_1, we first show that the estimated weight may be taken as given by showing:

$$N^{-1/2} \sum \tau_{wi} [Y_i - P_i] (\hat{\rho}_i - \rho_i) / N = o_p(1) \ .$$

Since $Y - P$ has conditional expectation of zero, a natural strategy would be to establish the above result by showing that $E(A_1^2)$ converges to zero.

After simplifying A_1, in the Appendix we provide the required mean-square convergence argument. It then follows that:

$$A_1 = N^{-1/2} \sum \tau_{wi} \left[Y_i - P_i \right] \rho_i / N + o_p(1).$$

Employing Lemma 8 in the Appendix, we are able to show that A_2 converges to zero in probability.

Turning to the term in B_1, in the Appendix we establish uniform convergence rates under multi-stage local smoothing (to reduce the bias) for estimated probability functions and their derivatives. In the first stage, estimated local smoothing parameters are constructed as functions of a regular kernel density estimator. These estimated local smoothing parameters are then employed (as variable windows) to re-estimate the density, which is in turn used to reconstruct estimated local smoothing parameters, which in the final stage are used to re-estimate the density. When local smoothing parameters are unknown, we show in the Appendix that this multi-stage approach results in increased convergence rates by reducing the order of the bias in the density estimator. Using these convergence rates, in the Appendix we show:

$$\sup \left| \hat{P}_i - P_i \right| = o_p \left(N^{-r} \right); \ \sup \left| \hat{\rho}_i - \rho_i \right| = o \left(N^{-s} \right), \ r + s > 1/2.$$

We then have:

$$\begin{aligned} \left| N^{1/2} \sum \tau_{ip}^2 \left[\hat{P}_i - P_i \right] \left[\hat{\rho}_i - \rho_i \right] \right| / N & \leq \\ N^{1/2} \sup \left[\tau_{ip} \left| \hat{P}_i - P_i \right| \right] \sup \left[\tau_{ip} \left| \hat{\rho}_i - \rho_i \right| \right] & = o_p(1). \end{aligned}$$

From above:

$$B_1 = N^{-1/2} \sum \tau_{wi} \left[\hat{P}_i - P_i \right] \rho_i + o_p(1).$$

Recall that for technical reasons the estimated probability function was defined as a ratio of adjusted, estimated densities. With the adjustment factors vanishing exponentially under trimming, we may ignore these adjustments and replace $\hat{P}_i$ with $\hat{f}_i / \hat{g}_i$. In the Appendix, (using a uniform convergence argument similar to that above), it is shown that:

$$N^{-1/2} \sum \tau_{ip} \left[\hat{f}_i / \hat{g}_i - P_i \right] \left[(\hat{g}_i / g_i - 1) \right] \rho_i = o_p(1).$$

It now follows that B_1 simplifies to:

$$\begin{aligned} B_1 &= N^{-1/2} \sum \tau_{ip} \left[\hat{f}_i / \hat{g}_i - P_i \right] \left[(\hat{g}_i / g_i) \right] \rho_i + o_p(1) \\ &= N^{-1/2} \sum R_i + o_p(1), \; R_i \equiv \tau_{ip} \left[\hat{f}_i - P_i \hat{g}_i \right] \left[\rho_i / g_i \right]. \end{aligned}$$

To further analyze B_1 above, it is important to show that it is "nearly" unbiased: $E(R) = o\left(N^{-1/2}\right)$. With biased reducing kernels:

$$\begin{aligned} E\left(R_i\right) &\equiv E\left(\left[\hat{f}_i - P_i \hat{g}_i \right] \rho_i / g_i \right) \\ &= E\left(\rho_i / g_i E\left(\left[\hat{f}_i - P_i \hat{g}_i \right] | X_i \right) \right) = o\left(N^{-1/2}\right), \end{aligned}$$

because the density estimators are "nearly" unbiased. However, once we have shown that the gradient has the above form (under locally smoothed kernels), we can control the bias by exploiting a property of semiparametric probability derivatives. Let

$$\delta\left(W_i\right) \equiv E\left(\left[\hat{f}_i - P_i \hat{g}_i \right] | X_i \right),$$

where δ only depends on index values. Then:

$$E\left(R_i\right) = E\left[\delta\left(w_i\right) \rho_i / g_i \right] = E\left[\delta\left(w_i\right) / g_i E\left(\rho_i | W_i \right) \right] = 0$$

from the residual property of the semiparametric probability derivative. By exploiting this property, the Appendix employs a mean-square convergence argument to show that B_1 converges to zero in probability (as does the comparable term in single index models). Similar to the analysis for A_2, using Lemma (8) in the Appendix, it can be shown that B_2 also vanishes in probability.

Employing the above results:

$$\sqrt{N}\left(\hat{\eta} - \eta_o\right) = -H_o^{-1} \sqrt{N} \sum \tau_{iw} \left[Y_i - P_i \right] \rho_i / N + o_p(1).$$

Noting that an information equality holds for this problem, a standard central limit theorem then gives asymptotic normality in the theorem below.

Theorem 2 *Under (A1-6) and (D1-7):*

$$\sqrt{N}\left[\hat{\eta} - \eta_0\right] \xrightarrow{d} Z \sim N\left(0, -H_o^{-1}\right).$$

4.2 The Outcomes Equation

With $\theta_o \equiv [\beta_o,\ \mu_o]$ and $Z \equiv [X,\ Y_2]$, recall that this equation is given as:

$$Y_1 = Z\theta_o + u,$$

Then, letting $\hat{Z}^*(\eta) \equiv \left[X,\ \hat{P}(\eta)\right]$ be an instrument for Z, the IV estimator is given as: :

$$\begin{aligned} \hat{\theta}_{IV} &= \left[\hat{Z}^*(\hat{\eta})' Z\right]^{-1} \hat{Z}^*(\hat{\eta})' Y_1 \Rightarrow \\ \sqrt{N}\left[\hat{\theta}_{IV} - \theta_o\right] &\equiv \left[\left(\hat{Z}^*(\hat{\eta})' Z\right)/N\right]^{-1} \sqrt{N}\hat{Z}^*(\hat{\eta})' u/N. \end{aligned}$$

From Lemma 9 in the Appendix, with $Z^* \equiv [X,\ P(\eta_0)]$:

$$\begin{aligned} \left[\hat{Z}^*(\hat{\eta})' Z - Z^* Z^*\right]/N &= o_p(1), \\ \sqrt{N}\left[\hat{Z}^*(\hat{\eta})' u - Z^* u\right]/N &= o_p(1). \end{aligned}$$

We can now immediately establish that the estimator is consistent and that it is asymptotically distributed as normal with a covariance matrix having the standard White heteroscedastic corrected form.

Theorem 3 *Defining* $\hat{u} \equiv \left(Y_1 - Z\hat{\theta}_{IV}\right)$, $\hat{D} \equiv Diag(\hat{u}^2)$, *and* $\hat{M} \equiv \left(\hat{Z}^*(\hat{\eta})' Z\right)/N$ *let:*

$$\hat{\Omega} \equiv \hat{M}^{-1}\left[\left(\hat{Z}^*(\hat{\eta})' \hat{D}\hat{Z}^*(\hat{\eta})\right)/N\right]\hat{M}^{-1}.$$

With $\hat{\Omega} \xrightarrow{p} \Omega$, *under (A1-5) and (D1-4):*

$$\sqrt{N}\left[\hat{\theta}_{IV} - \theta_o\right] \xrightarrow{d} Z \sim N(0,\ \Omega)$$

Note that we have assumed that $E(u|X) = 0$. If we assume further that u is independent of these conditioning vectors and let trimming vanish as the sample size increases, then $\hat{\theta}_{IV}$ is an optimal IV estimator (see Amemiya (1975)).

5 Simulation Evidence

To investigate the performance of the above estimator in a controlled setting, we conducted a monte-carlo study.[7] As the focus of this paper is on a double index binary response equation, with heteroscedasticity providing the main motivation, one of the designs below is of this form. It is also of interest to examine the consequences of a double index specification when the true binary response model is generated by a single index. Accordingly, we also present results for this case along with a related discussion of identification issues.

In formulating a design for the double index case, note that the number of factors determining the nature of the simulation is very large, precluding an exhaustive examination of the estimator under all possible conditions. Accordingly we adopt the following strategy. We consider the worse case situation where we are unwilling to make any restrictions on which variables enter the means or the variances. That is, the same variables affect the means and the variances. With all exogenous variables distributed as standard normal, the true model with heteroscedastic errors is given as:

$$\begin{aligned} v_i &= S_v(x)v_i^*, S_v(x) = \left[1 + (1 * x_{1i} + 2 * x_{2i} + 3 * x_{3i})^2\right] && (6) \\ Y_{2i} &= \{x_{1i} + x_{2i} + x_{3i} > 2v_i\} && (7) \\ u_i &= S_u\left(x\right)u_i^*, S_u\left(x\right) = \left[5 + Ln\left(1 + \left(x_{1i} + x_{2i} + x_{3i}\right)^2\right)\right]u_i^* && (8) \\ Y_{1i} &= 1 + x_{1i} + x_{2i} + x_{3i} + Y_{2i} + 6u_i. && (9) \end{aligned}$$

The unscaled errors, v_i^* and u_i^*, were generated as normal with expectation zero. Their variances were selected to insure that the scaled errors, v_i and u_i, each had unconditional variance of one. Finally, the unscaled errors were generated so as to have correlation of approximately .25 with each other. For the case in which the binary response is generated by a single index, we set S_v to a constant such that v has the same unconditional variance in both designs.

Turning to the double-index data generating process, we first examine our ability to recover the reduced form parameters in the binary choice model. Second, we examine the ability of the IV estimator to estimate the outcome equation parameters.

[7] We set trimming and smoothing parameters as follows: $a = .01$, $r_3 = 1/11$, $\delta = 1/25$ (require $\delta < r_3/8$), $r_2 = (r_3 - \delta/2)/2$, and $r_2 = (r_3 - \delta)/4$.

In the first experiment we conduct simulations with a sample size of 1000 and with 500 replications. Under the W-parameterization discussed earlier, x_2 is excluded from the first index and x_1 is excluded from the second index. The true values for the nuisance parameters (the coefficients on x_3 in each index after re-parameterization) are 2 and -1 respectively.[8] In estimating these parameters we obtained starting values from a coarse grid search. The average of the estimates for these two parameters are 2.031 and -1.037 with standard deviations of .469 and .508. Thus the estimates appear to be unbiased and they are reasonably precisely estimated. In addition to computing the double index parameters we also estimated a probit model which does not account for the presence of heteroscedasticity.

We also compared probit, semiparametric, and true probability functions. As an overall summary comparison, we estimated the correlation between the true probability that Y_{2i} is equal to 1, given the x_i vector, and that from the double index and probit models. The correlation between the probit probability and the true probability over the 500 replications was .726 with a standard deviation of .018. In contrast, the correlation between the true probability and that from the estimated double index model was .907 with a standard deviation of .010. In a more detailed comparison of probability functions, in Tables 1a-b we report the predicted probabilities for each of 5 quantiles.[9] These tables not only highlight the superior performance of the double index model, relative to the probit model, but also suggest that the estimator is performing very well in estimating the predicted probability.

Using the first step estimates we now employ these implied probabilities which we employ as an instrument for Y_{2i} in estimating the second equation. In Table 2 we report the second step IV and OLS estimates for the Y_1 equation. We report the estimates for each of the second step variables

[8] Write the transformed matrix of index coefficients as:

$$\begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 2 \end{bmatrix}.$$

[9] In constructing this table, probabilities were sorted on the basis of the calculated true probabilities in each sample. Then, for the first N/5 observations, average probabilities were calculated for the true probabilities, probit probabilities, and double index probabilities. These average probabilities were then averaged over replications (with minimal monte-carlo sampling error). Similar calculations were made for the each of the other reported quantiles.

as each contributes differently in the heteroscedasticity index. When the semiparametric probability function is employed as an instrument, we refer to the resulting estimator as SPIV.

Column 1 reports the average value of the OLS estimates from the second step. Recall that the true value for each coefficient is 1. Each of the coefficients for the exogenous variables displays a level of bias in the range of 3.3 to 8.7 percent. The standard errors for the estimates, given below the estimates in parentheses, indicates the degree of precision of the estimates. We report these for comparison sake with the adjusted coefficients which follow. The average estimate for the intercept is 1.205 revealing that the bias is greatly influencing this coefficient. Finally, focus on the estimate of the treatment effect. The average OLS point estimate is .590 which reflects a bias in excess of 40 percent. Clearly the design employed is generating a substantial degree of endogeneity.

In Column 2 we present the estimates in which we employ arbitrary functions of the explanatory variables as instruments. These included quadratic and cubic terms and all interactions between the variables, including the linear terms. Throughout, we use all of the variables in this available set. Column 2 indicates that this IV procedure reduces the bias on the coefficients on the exogenous variables and the intercept. The bias for the estimated treatment effect, however, is still on the order of 12.2 percent although this represents a marked improvement over the OLS eliminates.

Column 3 presents the estimates from the SPIV procedure. For each of the parameters on the exogenous variables there is a large reduction in the bias in comparison to the OLS estimates. The procedure is successfully eliminating the bias from the endogeneity of the treatment effect. This is also true for the treatment effect itself which now only displays 2 percent bias. Note, importantly, that the standard deviation for the treatment effect is smaller for this estimator than that shown in Column 2.

We now repeat the same exercises after increasing the sample size to 2000. The first step estimates are now 1.986 and -.988 with standard deviations of .241 and .249 respectively. Thus the estimates continue to be very accurate and we also see a large decrease in the level of variability. Once again we compute the correlations described above and we now find that the probit estimate is .727, with a standard deviation of .013, while the correlation between the truth and the probability from the estimated double index model is .915 with a standard deviation of .007. In the lower panel of Table 1 we report the quantiles for the various probabilities. Again the double index

model not only dominates the probit model, but also produces an excellent performance in absolute terms.

We now focus on the estimation of the binary treatment model and this is reported in Table 2b. The SPIV estimator formulated here continues to dominate the alternative estimators. The estimator using the higher orders and the cross products of the $x's$ continues to eliminate some of the bias but even doubling the sample size has not produced a notable decrease in the degree of bias. Once again, the SPIV estimator is remarkably accurate with the estimates seemingly unbiased for all coefficients. Perhaps the most remarkable feature of Table 2b is the increase in efficiency for this estimator as it now displays a standard deviation significantly lower than that for the alternative IV procedure.[10]

Turn now to the single-index data generating process noting that with constant S_v the binary response model becomes a probit model. However, suppose that the single index restriction is not imposed and that we continue to estimate the binary response in double index form. For this purpose, it is expositionally convenient to rewrite the model in an equivalent but more revealing form. Letting C and A be appropriately dimensioned nonsingular matrices, return to the original parameterization and write the binary response as:

$$\begin{aligned} E\left[Y_{2i}|X_i\right] &= E\left(Y_{2i} \mid \left[X_{1i},\ X_{2i},\ X_{3i}\right] CC^{-1} \begin{bmatrix} \alpha_{10} & \beta_1 \\ \alpha_{20} & \beta_2 \\ \alpha_{30} & \beta_3 \end{bmatrix} A\right) \\ &= E\left(\ Y_{2i} \mid \left[X_{1i},\ X_{2i},\ X_{3i}\right]\theta_0\ \right),\ \theta_0 \equiv \begin{bmatrix} 1 \\ \alpha_{20}/\alpha_{10} \\ \alpha_{30}/\alpha_{10} \end{bmatrix}. \end{aligned}$$

The first characterization is the double-index form, while the second follows from a single index restriction under a conventional normalization . With $\hat{\theta} \equiv \left[1,\ \hat{\theta}_2,\ \hat{\theta}_3\right]'$ obtained by imposing the single index restriction (e.g. as in

[10] It may be possible to improve this alternative IV procedure by developing a method for selecting the degree of the approximating polynomial. We have not pursued this strategy here primarily because the semiparametric probability function is of direct interest and secondly because this probability function is an optimal instrument.

Klein and Spady, 1993), define the non-singular matrix C as

$$C = \begin{bmatrix} 1 & 0 & 0 \\ \hat{\theta}_2 & 1 & 0 \\ \hat{\theta}_3 & 0 & 1 \end{bmatrix}.$$

Notice that the transformed variables are given as:

$$XC \equiv [X_1^*,\ X_{2,}\ X_3],$$

where $X_1^* = X\hat{\theta}$ is the estimated index under a single index restriction.

The transformed parameters corresponding to the above transformed variables are given as:

$$C^{-1} \begin{bmatrix} \alpha_{10} & \beta_1 \\ \alpha_{20} & \beta_2 \\ \alpha_{30} & \beta_3 \end{bmatrix} A \equiv \begin{bmatrix} \alpha_{10}^* & \beta_1^* \\ \alpha_{20}^* & \beta_2^* \\ \alpha_{30}^* & \beta_3^* \end{bmatrix} A.$$

With β_k^* not identified, consider the set of β_k^* values such that the upper block of the transformed parameter matrix is non-singular and, as earlier, set: A as the inverse of this block.. The following double-index form now follows::

$$E[Y_{2i}|X_i] = E\left[Y_{2i} \mid [X_{1i}^*,\ X_{2i},\ X_{3i}] \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \eta_{1o}^* & \eta_2^* \end{bmatrix}\right].$$

When the model is generated by a single index, $\eta_{1o}^* = 0$ is identified. However, once we condition on the single index, X_1^*, any additional "information" is irrelevant. Namely:

$$E[Y_{2i}|X_i] = E[Y_{2i}|X_{1i}^*] = E[Y_{2i}|X_{1i}^*, (X_{2i} + X_{3i}\eta_2^*)]$$

for all η_2^*. As a result, while the above expectation (probability) is identified, η_2^* is not identified. Consequently, when the binary response equation is estimated in double-index form, we expect the estimator for η_{1o}^* to be close to zero and the estimator for η_2^* to have a "large" variance. For N = 1000 observations, Table 3 provides results when the binary response model is estimated in both single and double index forms. Under the single index constraint, the estimated coefficients have small biases and low variances.

Furthermore, the distribution of the estimator is such that the mean components are close to the corresponding medians. In contrast, the bottom portion of this table provides results when the model is estimated in double-index form. On average, the estimator for η_{1o}^* (.055) is small as one would expect. The corresponding standard error of .65 is relatively large, which is misleading as there were a small number of very large outliers. Notice that the median of the estimator (.0003) is much smaller than the mean and is consistent with the true value for the coefficient being 0. The other parameter is not identified, as is reflected in an extremely large sampling variance.

Table 4 provides results for sample size equal to 2000. Other than there being less of an outlier issue, these results are similar to those above. Namely, as one would expect the estimator for the identified parameter is close to 0 and is much more precisely estimated than when the sample size is 1000. Note that the smaller standard error is due largely to a much better estimated binary response probability. The sampling variance for the unidentified parameter is relatively large.

Turning to the outcomes equation, shown in Table 5, the results are as expected. Note that the estimated probability function converges (pointwise and uniformly) slowly to the truth in double-index form. As a result, and not surprisingly, there is only a slight advantage to the SPIV estimator over the IV estimator. As found earlier, the bias for the OLS estimator is substantial, ranging up to almost 50% for the treatment effect. At the larger sample size (N = 2000), the semiparametric probability is better estimated, which is reflected in a noticeable improvement of SPIV over IV. In particular, the standard error for the estimated treatment effect is approximately 20% lower for the SPIV estimator relative to the IV estimator.

It is also instructive to compare the above results across designs in the case of the outcomes equation. When a double index really generates the data, the SPIV estimator has small biases and standard errors. However, now turn to the case where the model is still estimated in double index form, but where a single index actually generates the data. In this case, the biases and standard errors are noticeably larger.

6 Empirical Example

We now employ the estimators formulated here to study two questions of interest. There is a large recent literature on the effect of attendance at private

schools on educational attainment and subsequent labor market performance (for recent examples see Evans and Schwab 1995, Neal 1997 and Vella 1999). This has become an increasingly well studied area due to the common finding that attending private and catholic schools increases the number of years of school acquired and the level of post schooling qualifications. Unlike previous papers which examine the effect of catholic schools on education , we examine the effect of attending a government or state financed school. We begin first by estimating the marginal effects of particular variables on the probability of attendance at a government financed school. This allows us to identify the determinants of the school choice while allowing for general forms of heteroscedasticity and without making distributional assumptions. Second, we examine the impact of attendance at a government financed school on educational attainment. The issue of endogeneity of school type and education level needs little motivation. Schooling represents a form of human capital investment and the investment can differ in terms of duration and quality. However, as both decisions reflect human capital investments, albeit on different margins, each should be influenced by similar factors. As the unobservable factors are likely to be similar this highlights the endogeneity. Moreover, as both decisions are likely to be influenced by the same observable factors the absence of reasonable exclusion restrictions is immediately apparent. Despite the simultaneity the triangular structure is reasonable as the school type is chosen first and then the number of years follows from the individual's schooling success and the cost of the investment.

We employ data from the Australian Longitudinal Survey for 1985. The data comprises 5353 observations on youth who have completed their schooling. The binary response variable is the school type of the individual which we denote as *Govt* and which is a binary indicator function indicating that the individual attended a government run high school. The mean of this variable is .808. The outcome variable is the number of years of schooling which has a mean of 11.639. The model is the following

$$\begin{aligned} Schooling \;=\; & \alpha_o + \alpha_1 * Age + \alpha_2 * Australian\ Born + \alpha_3 * Both\ Parents \\ & \Pr esent\ in\ Household\ at\ Age\ 14 + \alpha_4 * Mother\ with\ Degree + \\ & \alpha_5 * Father\ with\ Degree + \alpha_6 * Siblings + \\ & \alpha_7 * Roman\ Catholic + \alpha_8 * Male + \alpha_9 * Attitudes + \\ & \alpha_{10} * Govt + u \end{aligned} \tag{10}$$

$$
\begin{aligned}
Govt &= \begin{cases} 1: & I_1 > v \\ 0: & Otherwise \end{cases}, \qquad (11) \\
I_1 &= \beta_o + \beta_1 * Age + \beta_2 * Australian\ Born + \beta_3 * Both\ Parents \\
&\quad \Pr esent\ in\ Household\ at\ Age\ 14 + \beta_4 * Mother\ with\ Degree + \\
&\quad \beta_5 * Father\ with\ Degree + \beta_6 * Siblings + \\
&\quad \beta_7 * Roman\ Catholic + \beta_8 * Male + \\
&\quad \beta_9 * Attitudes.
\end{aligned}
$$

The explanatory variables are those one would expect to influence human capital investment. With three exceptions the variables are indicator functions. For these indicator functions the variable name reflects what it measures. The variable *Age* is measured in years and *Siblings* denotes the number of siblings in the family. The one explanatory variable which requires some explanation is *Attitudes.* This variable is constructed from each individual's responses to a series of questions which aim to elicit the individual's view of the roles of females in the labor market. Vella (1994) investigates the role of this variable in the human capital investment for Australian youth and concludes that the variable captures family forces which influence educational attainment. An important issue in that study, which is equally of relevance here, is whether this variable can be treated as exogenous to human capital investment. While Vella (1994) starts with the conjecture that the attitudes variable is endogenous to human capital investment, that study is unable to provide any evidence that the attitudes variable is endogenous to schooling. Employing the same data set, we proceed on the assumption that the *Attitudes* is exogenous. The variable takes discrete values from 5 to 35, where a low score reflects a very traditional role for females while a higher score reflects an attitude of gender equality. We treat this variable and age as continuous for identification purposes..

Before focussing on the estimates, it is useful to consider why the schooling choice equation might exhibit heteroscedasticity. Many of the explanatory variables are indicator functions and their inclusion is meant to capture their average effect on the schooling choice. However, the direction, and magnitude, of these effects might be expected to vary across individuals. For example, consider the indicator function capturing that the individual is Australian Born. This captures the contrast with non Australian born individuals and for many reasons one might expect that there may be a difference across groups. However, just as it is likely that those comprising the

Australian born are very different in various ways, such as family attitudes towards education and scholastic abilities, it also true that those comprising the non Australian born are also heterogeneous. Accordingly, while the inclusion of the indicator function captures the mean difference across the two groups there is likely to be a large variance in the effect depending on which individuals from the respective groups are compared. Moreover, this difference may not be correlated with the other explanatory variables and thus it is not easily taken into account. The same type of argument is true for many of the other explanatory variables. Allowing the explanatory variables to effect the variance is an attempt to more accurately capture this affect.

We begin by estimating the schooling type decision. In column 1 of Table 6 we present the estimated parameters obtained by probit. In columns 2 and 3 of Table 6 we report the estimates from estimating the double index binary choice model. The standard error for each estimate is shown in parentheses under the estimate. Recall that we are able to transform the model to an equivalent one under a nonsingular linear transformation so as to induce exclusion restrictions for purposes of estimating probabilities. Further, we obtain an equivalent model by normalizing the constant term to zero and one of the coefficients in each index to one. In view of these normalizations, it is difficult to interpret the coefficients other than to note that many of the variables have a statistically significant impact. Accordingly, we perform the following exercise using both parametric and semi-parametric models. We use the estimates to evaluate the probability of each individual attending a government school with and without each of the characteristics. We then, with the exception of age, the attitudes variable and the number of siblings, compute the average effect of each individual acquiring the characteristic. For age and attitudes variables, evaluate the impact of a one standard deviation change while for siblings we increase the variable by one. These are all reported in Table 7. Without exception, the partial effect for each of the variables have the same sign across estimation procedures. Perhaps the most striking difference across the two procedures is the magnitude of the effect of the variable denoting that the individual is Catholic. In the probit model the estimated effect is over 50 percentage points while for the double index model the effect is around 33 percentage points. Thus, while overall the partial effects are quite similar across models the large difference in the Catholic effect illustrates the value of the double index approach.

While there are some important differences between the estimated marginal effects from the probit and double index model, it is valuable to test

the probit model of Government school attendance for the presence of heteroscedasticity and non-normality by employing the conditional moment tests outlined in Pagan and Vella (1989). The tests are implemented via artificial regressions whereby one regresses the product of the generalized residual and the single index from the probit model with the explanatory variable potentially causing the heteroscedasticity against the scores from the probit model and intercept. The test against the null of no heteroscedasticity is a t-test on the null that the intercept is equal to zero. We conducted this test for each of the variables which appear in the conditional mean of the *Govt* equation and report the results in Table 8. The tests indicated the presence of heteroscedasticity operating through several of the variables. More precisely there was a rejection at the 5 percent level for the *Age, Aust* and *Both Parents Present* variables and *Attitudes* at the 10 percent level. Moreover, the test for the imposed distributional assumptions strongly rejected normality. Note that the presence of both forms of misspecification makes it difficult to fully understand the cause of the rejections. Nevertheless, the evidence suggests that heteroscedasticity is present.

We now examine how the presence of heteroscedasticity can help detect the effect of exogenous effect of attendance at a government high school. Before we do so, we report the OLS estimates and also employ two alternative approaches for accounting for the simultaneity. In Column 1 of Table 9 we report the ordinary least squares (OLS) estimates of equation (10). They indicate that attending a Government schooling appears to decrease the years of educational investment by .559 years. The standard error is small indicating the effect is relatively precisely estimated. This effect is not particularly large given the large premium associated with attending a private institution when at high school. For example, in this sample only 47.8 percent of the individuals attending government schools obtained at least twelve years of schooling in comparison to 68.3 percent of the non-government students. Also, while only 2.9 percent of government students obtained a college degree the corresponding number for the non-government students is 7.3 percent. The remaining coefficients are also generally statistically significantly different from zero and are all of a reasonable magnitude although it is difficult to have strong expectations. The variables capturing the presence of both parents in the household and the level of each parent's education capture the effect of role models as well as higher incomes. The variable reflecting the number of siblings has the expected negative sign and is reasonable in magnitude. As found in Vella (1994) the *Attitudes* variable has a strong positive

effect on years of education acquired.

From above, the OLS estimated impact of attending a Government school appears to be too small. Accordingly, we are motivated to consider a model that incorporates the schooling decision, and does so in a general specification. However, first we employ two procedures which do not directly exploit the heteroscedasticity. First we perform IV by using the predicted probability from the probit model as an instrument for the Government indicator. The second is to include the Inverse Mills ratio, from this parametric estimation of the Government equation, as an additional regressor in the years of education equation. Note that the first of these estimates is consistent in the absence of normality while the latter is not. To implement these procedures, it is necessary to employ the probability that the individual attends a government school from the estimates are reported in Column 1 of Table 6.

The second column of Table 9 presents the estimates of the education equation when we conduct IV by instrumenting the *Govt* dummy with the predicted probabilities from the probit model. As the same variables appear in the *Govt* equation and the schooling equation the model is identified from the non-linear mapping from the explanatory variables. In general the coefficients are similar to those in column 1 although there is a difference with respect to the school and religion variables. The coefficient on the attendance at a government school variable is now unreasonable in that it indicates those who attend a government school, ceteris paribus, will obtain only .05 years of education less than those at private schools. This is in complete contrast to the conventional understanding of the affect of attendance at state financed schools. Note, however, that this coefficient is not statistically different from zero at the 10 percent significance level. When we adopt the plug in version of this model we obtain an estimate of the government school effect of -.071 with a standard error of .891.

In Column 3 we report the alternative procedure whereby one includes the inverse mills ratio from the model in Column 1 of Table 6 as an additional regressor in the education equation. These results are generally reasonable in magnitude, in that they are similar to the OLS estimates, although the government variable's coefficient is now less than half the OLS estimate in absolute terms. However, the coefficient on this variable is very imprecisely estimated.[11] Overall the evidence in Columns 2 and 3 confirms our suspicion

[11]Note that the standard errors for this column are underestimated as they have not been corrected to account for the estimation of the inverse mills ratio.

that there appears to be inadequate non-linearity in the transformations performed to enable accurate estimation of the model. Also note that as the t-statistic associated with the inverse mills ratio is low there is no evidence to support the conjecture that school type is endogenous to years of education. One suspects that the test has relatively low power given the inaccurate manner in which the parameters are estimated and the associated collinearity.

In the fourth column of Table 9 we report the estimates from the schooling equation when we instrument the *Govt* variable with the estimated probability from the semi-parametric binary choice model. The estimates are generally similar to those in the first column. The most striking change is the increase in the magnitude of the *Govt* school coefficient which now indicates that the effect is .99 years and is statistically significantly different from zero at the 10 percent level. This estimate seems far more reasonable given the educational behavior of those attending non-government schools. In order to explore the role of the double index structure in this result we also estimate the model where we first semi parametrically estimated the probability to employ as an instrument via the single index approach of Klein and Spady (1993). For this approach we found that the point estimate for the *Govt* coefficient was -.852 with a large standard error of .723. While the point estimate is similar to the double index approach, the increased identifying power of the double index model provides a different conclusion regarding whether the effect is statistically different from zero at conventional levels of testing.

Finally we explore the possibility that the treatment effect is not constant. To this end, denote $X_i : 1xK$ as the i^{th} observation on the K exogenous variables. Let the treatment variable enter as: $Govt_i * [c_o + X_i\theta_o]$. In this form, the *Govt* variable interacts with the individual's characteristics. We estimated the resulting model by IV, where we used the predicted probability from our double index model interacted with the individual's characteristics as instruments for these interaction variables. To examine overall whether or not there is a treatment effect, we considered a Wald test for the joint null hypothesis: $c_o = 0$ and $\theta_o = 0$. With a P-value of .0058, we reject the null hypothesis at conventional significance levels. We also calculated the average treatment effect (at the mean values of the $X\imath s$) to be -2.975 with an associated standard error of 1.162. Accordingly, there would seem to be a treatment effect whose magnitude is much larger than the average OLS effect previously reported. Not surprisingly given the above results, we also reject the null hypothesis of a constant treatment effect ($\theta_o = 0$) with an associated

P-value of .0114. [12]

7 Conclusions

The primary objective of this paper is to develop a semiparametric estimator for the binary choice model under the presence of heteroscedasticity. To do so we present a double index model where the indices capture the conditional mean and conditional variance respectively. We then estimate the parameters by maximizing a quasi likelihood function that depends on these two indices. We note that this procedure that is applicable for any discrete choice models which is a function of two indices. We also highlight that in providing the asymptotic properties of our procedure we develop a theoretical argument which justifies the use of local smoothing as bias reducing device in discrete choice models with a double index structure.

The interest in binary response models often follows from the appearance of the response as an endogenous explanatory variable of some interest in an other equation. An additional difficulty is that it is frequently difficult to identify variables which determine the response but which do not enter directly into equation in which the response appears as a regressor. We illustrate how the presence of heteroscedasticity in the model can provide identification in such models in such instances. Using the predicted probability from the binary response model as an instrument for the treatment variable, we show that one can consistently estimate the treatment effect. We show that the estimators for both models are consistent and asymptotically $(\sqrt{N})$ distributed as normal. We provide simulation evidence that illustrates that both procedures formulated here work well even in the case where the same variables are driving the conditional means and variances of both the treatment and outcome equations.

In an empirical investigation we illustrate the utility of both of our proposed estimators. In the first step we examine the determinants of the probability to undertake education at a Government financed school. In the second step we use this probability as an instrument to estimate the impact of attending such a school on the level of education. The evidence suggests that the estimated first step probability is quite different than that generated by

[12] While several of the individual interactions were significant, a number were not. Thus, it would seem reasonable, but beyond the scope of this paper, to further explore variable treatment effects.

a probit model assuming homoskedasticity. The second step estimates are suggestive that the heteroscedasticity in the schooling choice equation may be an effective means of identifying the effect of the school type on level of schooling.

References

[1] Abramson, I.S. (1982): "Bandwidth Variation in Kernel Estimates- A Square Root Law," *The Annals of Statistics,* 10, 1217-1223.

[2] Amemiya, T. (1975): "The Nonlinear Limited-Information Maximum-Likelihood Estimator and the Modified Nonlinear Two-Stage Least-Squares Estimator," *Journal of Econometrics*, 3, 375-386.

[3] Chamberlain, G. (1986): "Asymptotic Efficiency in Semi-Parametric Models with Censoring," *Journal of Econometrics,* 32, 189-218.

[4] Chen, S., and S.Khan (2003): "Rates of Convergence for Estimating Regression Coefficients in Heteroscedastic Discrete Response Models," *Journal of Econometrics*, 117, 245-278.

[5] Dagenais, M., and D.Dagenais (1997): "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics, 76 (1-2),* 193-222.

[6] Donald, S., and W.Newey (2001):"Choosing the Number of Instruments," *Econometrica*, 69, 1161-1191.

[7] Evans, W., and R.Schwab (1995): "Finishing High School and Starting College: Do Catholic Schools Make a Difference?" *Quarterly Journal Of Economics,* 60, 941-74.

[8] Fukunaga, K. (1972): *Introduction to Statistical Pattern Recognition,* New York Academic Press.

[9] Heckman, J.J. (1978): "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica,* 46, 931-959.

[10] ———- (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.

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

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

[13] Ichimura, H., and L.F.Lee (1991): "Semiparametric least squares (SLS) and weighted SLS estimation of multiple index models: Single equation estimation," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics,* ed. W.Barnett, J.Powell and G.Tauchen, Cambridge University Press.

[14] Klein, R, (1993): "Specification Tests for Binary Choice Models Based on Index Quantiles," *Journal of Econometrics*, 59, 343-375.

[15] Klein, R. and R. Spady (1993): "An Efficient Semiparametric Estimator for the Binary Response Model," *Econometrica, 61, 387-421.*

[16] Klein, R. and F.Vella (2006):"Estimating a Class of Triangular Simultaneous Equations Models Without Exclusion Restrictions" unpublished manuscript.

[17] Kordas, G. (2000). ""Smoothed Binary Regression Quantiles," forthcoming *Journal of Applied Econometrics.*

[18] Lee, L.F. (1995): "Semi-Parametric Estimation of Polychotomous and Sequential Choice Models", *Journal of Econometrics,* 65, 381-428.

[19] Lewbel, A. (1997): "Constructing Instruments for Regressions with Measurement Error when No Additional Data are Available, With an Application to Patents and R&D," *Econometrica*, 65, 1201-1213.

[20] Lewbel, A. (2004): "Identification of Heteroskedastic Endogenous Models or Mismeasured Regressor Models," unpublished manuscript.

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

[22] Manski, C. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of Maximum Score Estimation," *Journal of Econometrics,* 27, 313-334.

[23] Newey, W. and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," *Handbook of Econometrics*, v. 4, Chapter 36, Amsterdam, North Holland.

[24] Neal, D. (1997): "The Effects of Catholic Secondary Schooling on Educational Attainment," *Journal of Labor Economics,* 15, 98-123.

[25] Pagan, A. and F.Vella (1989): "Diagnostic Tests for Models Based on Unit Record Data: A Survey" *Journal of Applied Econometrics, 4*, S29-S60.

[26] Pakes, A. and D. Pollard (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57, 1027-1058.

[27] Powell, J., Stock,J., and T. Stoker (1989): "Semiparametric Estimation of Weighted Average Derivatives," *Econometrica,* 57, 1403-1430.

[28] Rigobon, R, (2003): "Identification through heteroscedasticity" *Review of Economics and Statistics,* 85, 777-792.

[29] Rummery, S., F.Vella and M.Verbeek (1999):"Estimating the Returns to Education for Australian Youth via Rank-Order Instrumental Variables," *Labour Economics,* 6, 491-507.

[30] Serfling, R.S. (1980) : *Approximation Theorems of Mathematical Statistics.* New York; Wiley.

[31] Silverman, P. (1986): *Density Estimation.* New York; Chapman and Hall.

[32] Vella, F. (1994), "Gender Roles and Human Capital Investment: The Relationship between Traditional Attitudes and Female Labor Market Performance," *Economica,* 61, 191-211.

[33] Vella, F. (1999), "Do Catholic Schools make a Difference? Evidence from Australia," *Journal of Human Resources,* 34, 208-224.

[34] Vella, F. and M.Verbeek (1997): "Rank Order as an Instrumental Variable" unpublished manuscript.

Table 1a: Probability Quantiles, N=1000

True	**Probit**	**Double Index**
.499	.449	.449
.593	.549	.578
.693	.626	.683
.790	.690	.775
.874	.701	.815

Table 1b: Probability Quantiles, N=2000

True	**Probit**	**Double Index**
.500	.500	.500
.593	.550	.580
.694	.626	.686
.791	.690	.779
.875	.700	.822

Table 2a: Simulation Results, N = 1000

Variable	**OLS**	**IV**	**SPIV**
Intercept	1.205	1.061	1.010
	(.063)	(.150)	(.125)
x_1	1.087	1.024	1.003
	(.051)	(.074)	(.071)
x_2	1.061	1.016	1.004
	(.051)	(.062)	(.063)
x_3	1.033	1.010	1.004
	(.046)	(.047)	(.050)
Y_2	.590	.878	.980
	(.097)	(.289)	(.233)

Table 2b: Simulation Results, N = 2000

Variable	**OLS**	**IV**	**SPIV**
Intercept	1.206	1.047	1.009
	(.046)	(.111)	(.088)
x_1	1.088	1.019	1.003
	(.036)	(.055)	(.050)
x_2	1.057	1.011	1.001
	(.035)	(.046)	(.045)
x_3	1.032	1.008	1.003
	(.033)	(.034)	(.036)
Y_2	0.592	.908	.987
	(.061)	(.219)	(.168)

Table 3: Single-Index Binary Response
N = 1000
Single Index Constraint

Coef.	True	$Avg\left(\hat{\beta}\right)$	$Med\left(\hat{\beta}\right)$
X_1	1	$--$	$--$
X_2	2/3	.6636 (.04225)	.6653
X_3	1/3	.3262 (.0368)	.3278

Double Index Constraint

	I1				I2		
Coef	True	$Avg\left(\hat{\beta}\right)$	$Med\left(\hat{\beta}\right)$	Coef	True	$Avg\left(\hat{\beta}\right)$	$Med\left(\hat{\beta}\right)$
X_1^*	1	$--$	$--$	X_1^*	0	$--$	$--$
X_2	0	$--$	$--$	X_2	1	$--$	$--$
X_3	0	.0551 (.6437)	.0003	X_3	$--$	-12.2801 (117)	.1456

Table 4: Single-Index Binary Response
N = 2000
Single Index Constraint

Coef.	True	$Avg\left(\hat{\beta}\right)$	$Med\left(\hat{\beta}\right)$
X_1	1	$--$	$--$
X_2	2/3	.6630 (0331)	.6611
X_3	1/3	.3315 (0295)	.3296

Double Index Constraint

	I1				I2		
Coef	True	$Avg\left(\hat{\beta}\right)$	$Med\left(\hat{\beta}\right)$	Coef	True	$Avg\left(\hat{\beta}\right)$	$Med\left(\hat{\beta}\right)$
X_1^*	1	$--$	$--$	X_1^*	0	$--$	$--$
X_2	0	$--$	$--$	X_2	1	$--$	$--$
X_3	0	-0.0026 (.0513)	.0016	X_3	$--$	$-.5424$ (7)	$-.2791$

Table 5: Outcomes Equation
Single Index Treatment, Double Index Constraint

	N = 1000		
	OLS	IV	$SPIV$
Intercept	1.2362	1.0393	1.0410
	(.0657)	(.2008)	(.1833)
x_1	1.1344	1.0123	1.0242
	(.0508)	.1223	(.1164)
x_2	1.0859	1.0125	1.0201
	(.0525)	.0949	(.0972)
x_3	1.0474	1.0108	1.0120
	.0436	.0543	(.0604)
Y_2	.5296	1.0393	.9203
	.1166	.3954	(.3594)

	N = 2000		
	OLS	IV	$SPIV$
Intercept	1.2342	1.0298	1.0427
	(.0443)	(.1520)	(.1217)
x_1	1.1332	1.018	1.0259
	(.0347)	(.0927)	(.0793)
x_2	1.0879	1.0116	1.0208
	(.0303)	(.0725)	(.0656)
x_3	1.0495	1.0112	1.0149
	(.0299)	(.0418)	(.0405)
Y_2	.5352	.9439	.9176.
	(.5351)	(.3018)	(.2397)

Table 6: Determinants of Attending a Government School

	PROBIT	S-P	S-P
	Govt School	Govt School	Govt School
Constant	2.726		
	(.232)		
Age	-0.017		1
	(.008)		
Attitudes	-0.022	1	
	(.005)		
Both Parents	-0.094	-0.294	1.382
	(.064)	(0.423)	(0.831)
Mother/Degree	-0.583	5.662	-2.451
	(.101)	(1.455)	(3.039)
Father/Degree	-0.549	0.865	-0.345
	(.078)	(0.241)	(0.511)
Siblings	0.020	0.165	-0.721
	(.011)	(0.248)	(0.496)
Roman Cath	-1.270	3.567	-2.339
	(.044)	(0.879)	(2.021)
Males	-0.032	2.961	-6.320
	(.045)	(1.702)	(3.750)
Aust	-0.296	-0.740	2.697
	(.074)	(0.810)	(1.518)

Table 7: Partial Effects

	PROBIT	S-P
Age	-.010	-.007
Attitudes	-.034	-.027
Both Parents	-.059	-.052
Mother/Degree	-.150	-.162
Father/Degree	-.164	-.099
Siblings	.004	.002
Roman Cath	-.530	-.326
Male	-.009	-.020
Aust	-.020	-.084

Table 8: Test Values for Heteroscedasticity

Variable	**Test Value**
Age	2.160
Aust	3.801
Both Parents Present	3.313
Mother with Degree	1.398
Father with Degree	0.365
Siblings	0.100
Roman Catholic	1.288
Male	0.820
Attitudes	1.695

Table 9: The Impact of Government School Attendance on Years of Education

	OLS	**IV**	**CF**	**SPIV**
	School	**School**	**School**	**School**
Constant	6.025	5.408	5.597	6.897
	(0.238)	(0.795)	(0.578)	(0.729)
Age	0.193	0.195	0.195	0.171
	(0.008)	(0.009)	(0.008)	(0.009)
Aust	0.030	0.063	0.053	0.002
	(0.069)	(0.081)	(0.075)	(0.083)
Both Parents	0.294	0.306	0.303	0.310
	(0.062)	(0.064)	(0.063)	(0.070)
Mother/Degree	0.283	0.365	0.340	0.240
	(0.119)	(0.156)	(0.138)	(0.162)
Father/Degree	0.659	0.734	0.711	0.600
	(0.090)	(0.128)	(0.110)	(0.129)
Siblings	-0.117	-0.118	-0.118	-0.120
	(0.011)	(0.011)	(0.011)	(0.012)
Roman Cath	-0.045	0.129	0.075	-0.202
	(0.052)	(0.220)	(0.158)	(0.213)
Male	0.215	0.218	0.218	0.236
	(0.045)	(0.045)	(0.045)	(0.048)
Attitudes	0.081	0.084	0.083	0.082
	(0.005)	(0.005)	(0.005)	(0.006)
Govt	-0.559	-0.050	-.206	-0.986
	(0.062)	(.626)	(.439)	(0.591)
Mills Ratio			-.200	
			(.247)	

8 Appendix

The Appendix is organized into two subsections, with the first stating and proving all intermediate lemmas that we require to establish the asymptotic properties of the estimators. The second subsection employs these lemmas to prove the main results in the paper.

8.1 Intermediate Results

From (D1-D7) of the Assumptions and Definitions section, recall that $\hat{f}_1(\bullet)$ estimates $\Pr(Y_2 = 1)g_1(w)$, where $g_1(w)$ is the density for W conditioned on $Y_2 = 1$. Similarly, $\hat{f}_0(\bullet)$ estimates $\Pr(Y_2 = 0)g_0(w)$, where $g_0(w)$ is the density for W conditioned on $Y_2 = 0$. Throughout, all lemmas apply to both $\hat{f}_1(\bullet)$ and $\hat{f}_0(\bullet)$. Accordingly, for notational convenience, we will simply write $\hat{f}(\bullet)$ to refer to either of these estimators. In so doing, we will refer to the local smoothing parameters as λ without subscripting. Throughout, we will write $\nabla_\eta^k f$ to mean the k^{th} partial derivative of f with respect to η, with $\nabla_\eta^0 f \equiv f$. Finally, in terms of notation, denote X_c and X_d as the vectors of continuous and discrete variables respectively, with realizations $x_c \, \epsilon \mathcal{X}_c$ and $x_d \, \epsilon \mathcal{X}_d$. With $\mathcal{X}_{c1}$ as the subset of $\mathcal{X}_c$ on which $\tau_x = 1$ (see D6), define $\mathcal{X}_1 \equiv \{x : x_c \, \epsilon \mathcal{X}_{c1}, \; x_d \, \epsilon \mathcal{X}_d\}$. Recalling that $w = [x_1 + x_3\eta_{31}, \; x_2 + x_3\eta_{32}] \equiv [w_1(\eta), w_2(\eta)]$, let $\mathcal{X}_2 = \{x : \; \underline{w}_k < w_k(\eta) < \bar{w}_k, \;\; k = 1, 2\}$. Finally, let $\mathcal{A} \equiv \{x : x \, \epsilon \, \mathcal{X}_1 \cup \mathcal{X}_2\}$. All uniform results will be on $\mathcal{A}$, and, when appropriate, the compact parameter space. Though not stated explicitly, for all of the results below, we employ all assumptions in (A1-6) and (D1-7).

The estimated conditional densities above depend on the sample covariance matrix for W. As W depends on the index parameters, η, we denote this covariance matrix as $\hat{\Sigma}(\eta)$. With $\Sigma(\eta)$ as the uniform (in η) probability limit of $\hat{\Sigma}(\eta)$, Lemma 1 below will enable us to treat this estimated matrix as if it were known.

Lemma 1: Denote $\hat{f}\left(w; \hat{\Sigma}(\eta)\right)$ as the estimator defined in (D1-3) and denote $\hat{f}(w; \Sigma(\eta))$ as the corresponding estimator with $\Sigma(\eta)$ replacing $\hat{\Sigma}(\eta)$.

Define $\widehat{f}_0(\bullet)$ analogously. Then:

$$\sup_{\eta,\bar{x}} \left| \nabla_\eta^k \widehat{f}\left(\bar{w}; \hat{\Sigma}(\eta)\right) - \nabla_\eta^k \hat{f}\left(\bar{w}; \Sigma(\eta)\right) \right| = o_p\left(N^{-1/2}\right), \quad k = 0, 1, 2,$$

where uniformity is over the sets described above.

Proof of Lemma 1: From a Taylor series expansion:

$$\left| \nabla_\eta^k \hat{f}\left(w; \hat{\Sigma}(\eta)\right) - \nabla_\eta^k \hat{f}_m\left(w; \Sigma(\eta)\right) \right| \leq \sup_{\eta,x} \left| \nabla_\Sigma \nabla_\eta^k \hat{f}\left(w; \hat{\Sigma}(\eta)\right) \right| \sup_{\eta} \left| \hat{\Sigma}(\eta) - \Sigma(\eta) \right|.$$

Since $\hat{f}$ converges to f even under an inconsistent estimator for Σ, it can be shown that the first term above is $o_p(1)$. As the second term is $O_p\left(N^{-1/2}\right)$, the result follows.

Employing Lemma 1, we will proceed with $\Sigma(\eta)$ replacing $\hat{\Sigma}(\eta)$ throughout. To simplify the argument further, it is also convenient to replace all estimated components in local smoothing parameters with their expectations. From (D2-3) estimated smoothing parameters are given as:

$$\hat{\lambda}_j = \left[\hat{d}_j \hat{\gamma}_j + \left(1 - \hat{d}_j\right) / Ln(N) \right]^{-1/2} \equiv \lambda\left(\hat{\gamma}_j\right),$$

where $\hat{\gamma}_j \equiv [\hat{\pi}_j / \hat{m}]$ and $\hat{d}$ is the smoothed indicator:

$$\hat{d}_j \equiv \left\{ 1 + \exp\left(-N^{\varepsilon} \left[\hat{\gamma}_j - \frac{1}{Ln(N)} \right] \right) \right\}^{-1} \equiv d\left(\hat{\gamma}_j\right).$$

Define $\bar{\gamma}_j \equiv [E(\hat{\pi}_j)/m]$, $\bar{d}_j \equiv d\left(\gamma_j\right)$, and

$$\bar{\lambda}_j \equiv \left[\bar{\gamma}_j \bar{d}_j + \left(1 - \bar{d}\right) / Ln(N) \right]^{-1/2} = \lambda\left(\bar{\gamma}_j\right).$$

Write $\hat{f}\left(\bar{w}; \hat{\lambda}\right)$ as the estimator of f at $\overline{w} = \bar{x}\eta$ and let $\hat{f}\left(\overline{w}; \bar{\lambda}\right)$ be the corresponding estimator with $\bar{\lambda}$ replacing $\hat{\lambda}$. In the next three lemmas, we examine convergence rates under multi-stage local smoothing. For estimated densities and first derivatives Lemmas $2A - B$ provide the required intermediate results needed to establish convergence rates in the third stage of local smoothing (Lemma 2C). Throughout, $\overline{w} \equiv \bar{x}\eta$.

Lemma 2A: Stage 1 (No local smoothing). Let $\hat{\boldsymbol{\lambda}}_1 = \mathbf{1}$ Then, for $\bar{x} \epsilon \mathcal{A}$ η in a compact set, and $k = 0, 1, 2$:

$$a) \quad : \quad \sup_{\bar{x},\eta} \left| \nabla_\eta^k \hat{f}\left(\overline{w}; \mathbf{1}, h_1\right) - E\nabla_\eta^k \hat{f}\left(\overline{w}; \mathbf{1}, h_1\right) \right| = O_p\left(1/\left[N^{1/2} h_1^{k+2}\right]\right)$$

$$b) \quad : \quad \sup_{\bar{x},\eta} \left| E\nabla_\eta^k \hat{f}\left(\overline{w}; \mathbf{1}, h_1\right) - \nabla_\eta^k f\left(\overline{w}\right) \right| = O_p\left(h_1^2\right).$$

Proof of Lemma 2A: Standard bias and uniform convergence results provide the proof (see Klein and Spady(1993)).

Employing the above results without local smoothing, Lemma 2B below examines convergence rates in which local smoothing is based on the density estimator in Lemma 2A.

Lemma 2B: Stage 2 (Local Smoothing). Let $\hat{\boldsymbol{\lambda}}_2 \equiv \lambda\left(\hat{f}\left(w; h_1, \mathbf{1}\right)\right)$, $\bar{\boldsymbol{\lambda}}_2 \equiv \lambda\left(E\left[\hat{f}\left(w; h_1, \mathbf{1}\right)\right]\right)$, and $h_i = O(N^{-r_i})$, $i = 1, 2$. Assuming $r_1 < r_2$, for $k = 0, 1, 2$:

$$a) \quad : \quad \sup_{\bar{x},\eta} \left| \nabla_\eta^k \hat{f}\left(\overline{w}; \hat{\boldsymbol{\lambda}}_2, h_2\right) - \nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_2, h_2\right) \right| = O_p\left(1/\left[N^{1/2} h_2^{k+2}\right]\right)$$

$$b) \quad : \quad \sup_{\bar{x},\eta} \left| \nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_2, h_2\right) - E\nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_2, h_2\right) \right| = O_p\left(1/\left[N^{1/2} h_2^{k+2}\right]\right)$$

$$c) \quad : \quad \sup_{\bar{x},\eta} \left| E\nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_2, h_2\right) - \nabla_\eta^k f\left(\overline{w}\right) \right| = O_p\left(h_2^2 h_1^2\right).$$

Proof of Lemma 2B. Employing a Taylor series approximation, the proof for (a) follows from the uniform convergence rate of $\hat{\boldsymbol{\lambda}}_{2i}$ to $\bar{\boldsymbol{\lambda}}_{2i}$ (Klein and Spady, 1993, Lemma 1) and the window condition: $r_1 < r_2$. The proof for (b) is essentially the same as that for (a). To establish (c), write (employing a dominance condition to differentiate under an integral):

$$E\left(\nabla_\eta^k \hat{f}\left(\overline{w}_{;2}, h_2, \bar{\lambda}_2\right)\right) = \nabla_\eta^k E\left(\hat{f}\left(\overline{w}_{;2}, h_2, \bar{\lambda}_2\right)\right) \equiv \nabla_\eta^k \Delta_2,$$

where the second expectation is taken with respect to the density for w. Taylor expand Δ_2 in h_2 about $h_2 = 0$ and use the symmetry in K about 0 to obtain:

$$\Delta_2 = \nabla_\eta^k h_2^2 \left[\hat{C}_2 - C_2\right] + \nabla_\eta^k h_2^2 C_2 + h_2^4 \hat{C}_4.$$

Here, $\hat{C}_2 \xrightarrow{p} C_2$, where C_2 contains terms (densities and density derivatives) that would follow from a Taylor series expansion using local smoothing parameters based on true densities.[13] For the first term in Δ_2, it consists of estimated densities and density derivatives. From the rate at which the expectation of an estimator (density or higher order derivatives) converges to the truth:

$$h_2^2\left[\hat{C}_2 - \bar{C}_2\right] = O_p\left[h_2^2 h_1^2\right].$$

From Abramson and Silverman, the second term vanishes as $C_2 = 0$. The argument now follows because in the final term: $\mathrm{h}_2^2\hat{C}_4 = o_p\left(h_1^2\right)$.[14] Referring to (D2-4), $C_4 = O\left(N^{4a}\right)$, where $a = .01$ is a local smoothing parameter. Under assumptions on smoothing parameters, the final term is of smaller order than the first, which completes the argument.[15]

Lemma 2C: Stage 3 (Local Smoothing). Let $\hat{\boldsymbol{\lambda}}_3 \equiv \lambda\left(\hat{f}\left(w;\ \hat{\boldsymbol{\lambda}}_2, h_2\right)\right)$ and $h_i = O(N^{-r_i})$, $i = 1, 2, 3$. With $r_i > 0$, assume $r_1 < r_2$ and that $r_1 + r_2 < r_3$. With $\bar{\boldsymbol{\lambda}}_2$ given as above, define $\bar{\boldsymbol{\lambda}}_3 \equiv \lambda\left(\hat{f}\left(w; \bar{\boldsymbol{\lambda}}_2\ , h_2\right)\right)$.

---

[13] Local smoothing parameters employ separate trimming to keep local smoothing parameters from becoming smaller than $\mathrm{O}_p\left(1/Ln(N)\right)$. In taking a Taylor series expansion about $\mathrm{h}_2 = 0$, derivatives of Local-smoothing trimming will appear. However, with densities evaluated at a "target" point for which they are bounded from below by $c > 0$, then such derivatives will vanish exponentially (and can therefore be ignored). This derivative can not be ignored in the final term of such an expansion as it is evaluated at an intermediate point.

[14] A typical term of $\hat{\mathrm{C}}_4$ depends on the integral of the product of a term involving the inverse of a density estimator raised to a power below 4 (T1), the fourth derivative of a density estimator (T2), the fourth derivative of the smooth trimming function (T3), the kernel, and the true density. Based on the smooth trimming of local smoothing, uniformly:

$$|T1T3| = o_p\left(N^{.04}Ln(N)\right)$$

Given the uniform rate at which the fourth derivative of a density estimator converges to the truth and the fact that $\mathrm{h}_2^2 N^{.04}Ln(N) = o\left(h_1^2\right)$, the result follows.

[15] With $\varepsilon_a > 0$ and arbitrarily small, set $a = \left(r_3 - \varepsilon_a\right)/8$. Here, a = .01 and $\mathrm{r}_3 = 1/11$. For $\delta < r_3/2$, set:

$$r_1 = \left(r_3 - \delta\right)/4 \text{ and } r_2 = \left(r_3 - \delta/2\right)/2.$$

For these settings, $r_1 < r_2 - 2a$.

Then, for $k = 0, 1, 2$:

$$a) \quad : \quad \sup_{\bar{x},\eta} \left| \nabla_\eta^k \hat{f}\left(\overline{w}; \hat{\boldsymbol{\lambda}}_3, h_3\right) - \nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_3, h_3\right) \right| = O_p\left(1/\left[N^{1/2} h_3^{k+2}\right]\right)$$

$$b) \quad : \quad \sup_{\bar{x},\eta} \left| \nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_3, h_3\right) - E\nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_3, h_3\right) \right| = O_p\left(1/\left[N^{1/2} h_3^{k+2}\right]\right)$$

$$c) \quad : \quad \sup_{\bar{x},\eta} \left| E\nabla_\eta^k \hat{f}\left(\overline{w}; \bar{\boldsymbol{\lambda}}_3, h_3\right) - \nabla_\eta^k f\left(\overline{w}\right) \right| = O_p\left(h_3^2 h_2^2 h_1^2\right).$$

Proof of Lemma 2C. The proof of (a-b) is the same as that in the previous lemma. For (c), define Δ_3 as in the previous lemma with $\bar{\boldsymbol{\lambda}}_3$ replacing $\bar{\boldsymbol{\lambda}}_2$. Then from the same type of Taylor expansion as in the previous theorem:

$$\Delta_3 \ = \nabla_\eta^k h_3^2 \left[\hat{C}_2^* - C_2\right] + h_3^4 \hat{C}_3^*.$$

From Lemma 2C, the first term above has order $h_3^2 h_2^2 h_1^2$. With $\hat{C}_3^* = O\left(N^{4a}\right)$, similar to the previous lemma, this last term is of smaller order than the first under the assumptions on smoothing parameters, which completes the proof.

Employing the above results, it is now possible to establish uniform rates of convergence (on compact sets) for estimated probability functions and derivatives.

Lemma 3 (Estimated Probability Functions).

$$\sup_{\bar{x},\eta} \left| \nabla_\eta^k \hat{P}\left(\overline{w}; \eta\right) - \nabla_\eta^k P\left(\overline{w}\right) \right| = O_p\left(\max\left\{1/\left[N^{1/2} h_3^{k+2}\right],\ h_3^2 h_2^2 h_1^2\right\}\right).$$

Proof of Lemma 3. The proof immediately follows from the lemmas above.

Below we will establish asymptotic normality by exploiting a "residual" property of semiparametric probability derivatives. The following lemma provides this property.

Lemma 4. Let $P\left(\eta\right)$ be the semiparametric probability function, where $P\left(\eta_0\right) = \Pr\left(Y_2 = 1 |\ X\right)$. Then, with $\nabla_\eta = \nabla_\eta^1$ as the first partial operator:

$$E\left[\nabla_\eta P\left(\eta\right)\ |\ W_1\left(\eta_1\right),\ W_2\left(\eta_2\right)\right]_{\eta\ =\ \eta_0} = 0.$$

Proof of Lemma 4. The proof of this result for the single index case is due to Whitney Newey and is contained in Klein and Spady (1993) and Klein and Sherman (2002). The extension to the double index case immediately follows from the same type of argument employed for the single index case.

As a final set of intermediate lemmas, we require results to deal with trimming. As discussed earlier, one trimming strategy below is based on a trimming sequence defined on the $X\prime s$. In particular, recall from (D6) that $\hat{\tau}_{xi}$ is a trimming indicator that is 1 on the set where each of the continuous variables is in a region defined by sample quantiles (e.g. the lower 1% and upper 99% sample quantiles). We refer to this trimming indicator as being estimated. Denote τ_{xi} as the corresponding trimming indicator with all sample quantiles replaced by their population counterparts. Lemma 5 provides a useful result relating estimated to known trimming. As such trimming occurs in normalized sums, the result below is written in this form to facilitate its subsequent use below.

Lemma 5: X-Trimming. Let r_i be random variables with $E|r_i|$ bounded. Then, under X-trimming, for any $\varepsilon > 0$:

$$\left| \frac{1}{N} \sum_{i=1}^{N} \left[\hat{\tau}_{xi} - \tau_{xi} \right] r_i \right| \leq \sum_{m=1}^{M} R_m \sum_{i=1}^{N} b_{im} \left| r_i \right| / N + o_p\left(N^{-1/2}\right) = O_p\left(N^{-(1/2)+\varepsilon}\right),$$

where M is finite, $R_m = O_p\left(N^{-(1/2)+\varepsilon}\right)$, and b_{im} is i.i.d., non-negative, and bounded.

Proof of Lemma 5. The proof for this lemma is based on an inequality due to Jim Powell for bounding $|(\hat{\tau}_{xi} - \tau_{xi})|$ from above by a "smoothed" indicator and is contained in Klein (1993, Lemmas 1-2, and the proof for Lemma 2). Once the indicator is smoothed, standard Taylor series arguments yield the above result. Here, ε is the "penalty" for approximating an indicator with a smooth function.

We will also be employing a trimming strategy based on the indices. Denote $\hat{\eta}_{kp}, k = 1, 2,$ as a matrix pilot estimates of nuisance parameters (obtained below under X-trimming) and define estimated indices as:

$$\hat{W}_1 \equiv X_1 + X_3\hat{\eta}_{1p};\ \hat{W}_2 \equiv X_2 + X_3\hat{\eta}_{2p}$$

Recall that the smoothed trimming function in (D7) depends on $\hat{\eta}_{kp}$ and estimated sample quantiles. From (D2), we defined an underlying smooth trimming function as:

$$\tau(z;a) \equiv [\ 1+\exp(N^a[z])]^{-1}.$$

The estimated trimming function then applies this smooth trimming function to each of the $k=1,2$ indices to insure that each indices stays (asymptotically) between lower and upper sample quantiles . Namely, from (D7),

$$\begin{aligned}
\hat{\tau}_{wi} &\equiv \hat{\tau}_{1i}\,\hat{\tau}_{2i},\ \ \hat{\tau}_{ki} \equiv \hat{L}_{ki}\hat{U}_{ki}, \\
\hat{L}_{ki} &\equiv \tau\left(\ \underline{w}_k\left(\hat{\eta}_p\right) - w_{ki}\left(\hat{\eta}_p\right);1/12\right) \\
\hat{U}_{ki} &\equiv \tau\left(w_{ki}\left(\hat{\eta}_p\right) - \overline{w}_k\left(\hat{\eta}_p\right);1/12\right) \quad for\ k=1,2.
\end{aligned}$$

Here, $\underline{w}_k\left(\hat{\eta}_p\right)$ is a lower sample quantile of the $w_{ki}\left(\hat{\eta}_p\right)'s$ while $\overline{w}_k\left(\hat{\eta}_p\right)$ is the corresponding upper sample quantile. Letting η_0, λ_{kL}, and λ_{kU} be the probability limits for $\hat{\eta}_p, \underline{w}_k\left(\hat{\eta}_p\right), and\ \overline{w}_k\left(\hat{\eta}_p\right), \tau_i$ is obtained from $\hat{\tau}_i$ by replacing all estimates with their probability limits. Analogously, L and U are defined by replacing all estimators by their population counterparts. To examine estimated trimming, we require a rate at which estimated quantiles $(\underline{w}_k\left(\hat{\eta}_p\right), \overline{w}_k\left(\hat{\eta}_p\right))$ converge to the corresponding true quantiles. With virtually any rate sufficing, Lemma 6 below provides a rate that is subsequently employed in Lemma 7 in arguing that estimated trimming can be treated as known.

Lemma 6: Estimated Quantiles. Assuming $\left(\hat{\eta}_p - \eta_0\right) = o_p(N^{-r})$, $r < 1/2$:

$$\begin{aligned}
\underline{w}_k\left(\hat{\eta}_p\right) - \lambda_{kL} &= o_p\left(N^{-r+\varepsilon}\right) \\
\overline{w}_k\left(\hat{\eta}_p\right) - \lambda_{kU} &= o_p\left(N^{-r+\varepsilon}\right).
\end{aligned}$$

Proof of Lemma 6. It suffices to consider the lower α^{th} quantile with $k=1$. With $\{\bullet\}$ is an indicator on the indicated event, for this case:

$$\sum \left\{w_{1i}\left(\hat{\eta}_p\right) < \underline{w}_1\left(\hat{\eta}_p\right)\right\}/N = \alpha.$$

Employing the same type of smooth approximation argument used in the proof of Lemma 5 and with $\varepsilon > 0$:

$$\sum \left[\left\{w_{1i}\left(\hat{\eta}_p\right) < \underline{w}_1\left(\hat{\eta}_p\right)\right\} - \left\{w_{1i}\left(\eta_0\right) < \underline{w}_1\left(\hat{\eta}_p\right)\right\}\right]/N = O_p\left(N^{-(r-\varepsilon)}\right).$$

Define $\underline{w}_{10}$ such that:

$$\sum \left\{ w_{1i}\left(\eta_0\right) < \underline{w}_{10} \right\} / N = \alpha.$$

Then it follows from above that:

$$\sum \left[\left\{ w_{1i}\left(\eta_0\right) < \underline{w}_1\left(\hat{\eta}_p\right) \right\} - \left\{ w_{1i}\left(\eta_0\right) < \underline{w}_{10} \right\} \right] / N = O_p\left(N^{-(r-\varepsilon)}\right).$$

Letting F_N be the empirical distribution for the $w_{1i}\left(\eta_0\right)$'$s$:

$$F_N\left(\underline{w}_1\left(\hat{\eta}_p\right)\right) - F_N\left(\underline{w}_{10}\right) = O_p\left(N^{-(r-\varepsilon)}\right).$$

From the uniform convergence of the empirical distribution function to the the true distribution function, F:

$$\begin{aligned} F\left(\underline{w}_1\left(\hat{\eta}_p\right)\right) - F\left(\underline{w}_{10}\right) &= O_p\left(N^{-(r-\varepsilon)}\right) \Rightarrow \\ \underline{w}_1\left(\hat{\eta}_p\right) - \underline{w}_{10} &= O_p\left(N^{-(r-\varepsilon)}\right). \end{aligned}$$

The lemma now follows since:

$$\left|\underline{w}_1\left(\hat{\eta}_p\right) - \lambda_{1L}\right| \leq \left|\underline{w}_1\left(\hat{\eta}_p\right) - \underline{w}_{10}\right| + \left|\underline{w}_{10} - \lambda_{1L}\right|.$$

For the case of index-trimming, recall that the trimming function is a smooth exponential function. In employing Taylor series arguments to analyze this function, it is important that trimming function derivatives behave as trimming functions themselves in that they severely downweight the same observations as the initial trimming function. This follows, because derivatives have the structure of being a bounded function multiplied by the initial trimming function. For example:

$$\frac{\partial}{\partial z}\tau\left(z\right) = \left[\tau - 1\right]\tau; \;\; \frac{\partial^2}{\partial z \partial z}\tau\left(z\right) = \left[\left(2\tau - 1\right)\left(\tau - 1\right)\right]\tau.$$

The proof of Lemma 7 below, which is essentially the same as that in Klein and Spady[1993], exploits this replicative property.

Lemma 7: Index-Trimming. Let $\left(\hat{\eta}_p - \eta_o\right) = O\left(N^{-r_p}\right)$, and assume $r_p > r_3$, with $h_3 = O_p\left(N^{-r_3}\right)$ as specified in (D4). Then, for $R_m = o_p\left(1\right)$,

M is finite, and b_{im} is i.i.d. and bounded over i for each m.

$$a) \quad : \quad N^{-1/2}\sum\left[\hat{\tau}_{wi}-\tau_{wi}\right]\left[Y_i-P_i\right]\hat{\rho}=\sum_{m=1}^{M}R_m\sqrt{N}\sum b_{im}\tau_{wi}\left[Y_i-P_i\right]\hat{\rho}/N+o_p\left(1\right)$$

$$b) \quad : \quad N^{-1/2}\sum\left[\hat{\tau}_{wi}-\tau_{wi}\right]\left[\hat{P}_i-P_i\right]\hat{\rho}_i=o_p\left(1.\right)$$

Proof of Lemma 7. To establish (a), expand the components of $\hat{\tau}_{wi}$ in a Taylor series expansion, to obtain

$$\begin{aligned}
\sqrt{N}\sum_i\left[\hat{\tau}_{wi}-\tau_{wi}\right]\left[Y_i-P_i\right]\hat{\rho}_i/N &= \sqrt{N}\sum_{d=1}^{D}T_d/N\ , \\
T_d \quad &\equiv \quad \sum_{s_d=1}^{S_d}R_{s_d}\sum_i b_{is_d}\tau_{wi}\left[Y_i-P_i\right]\hat{\rho}_i,\ d=1,...,D-1 \\
\left|T_D\right| \quad &\leq \quad \sum_{s_D=1}^{S_D}R_{s_D}\sum_i b_{is_D}\left|\hat{\rho}_i\right|,
\end{aligned}$$

where S_d is finite, $d=1,...,D$ and b_{is_D} is i.i.d. over i and bounded. With D selected such that $D\left(r-a\right)>1/2+2r_3$, d and D are both finite. The R-terms satisfy:

$$\begin{aligned}
R_{s_d} \quad &= \quad O_p\left(N^{-d(r-a)}\right),\ d=1,...,D-1 \\
R_{sD} \quad &= \quad O_p\left(N^{-D(r-a)}\right),\ D\left(r-a\right)>1/2+3r_3.
\end{aligned}$$

The result now follows.

$$\begin{aligned}
N^{1/2}N^{-d(r-a)}\sup\tau_i^{1/2}\left|\hat{P}_i-P_i\right|\sum_i\tau_i^{1/2}b_{isd}\left|\hat{\rho}_i\right| \quad &= \quad o_p\left(1\right) \\
N^{1/2}N^{-D(r-a)}\sup\left|\hat{P}_i-P_i\right|N^{2r3} \quad &= \quad o_p\left(1\right).
\end{aligned}$$

The argument for (b) is similar.

To establish asymptotic normality in the next section, we will need to analyze several components that comprise the gradient. To simplify the exposition, we examine these components in Lemmas 8A-B below. In providing these results, recall that we use the notation τ_x and τ_w to refer respectively

to X-trimming and Index-trimming. We employ the notation τ without x or w subscript for results that hold under either form of trimming. These gradient components have a standard form and depend on an estimated weight involving probability derivatives (see D5). Denoting this estimated weight as $\hat{\rho}_i^*$:

$$\begin{aligned}\hat{\rho}_i^* &= \nabla_\eta \hat{P}_i(\eta_0) / \left[\hat{P}_i(1-\hat{P}_i)\right] = \left[\nabla_\eta \left(\hat{f}_i^*(\eta_0) / \hat{g}_i^*(\eta_0)\right)\right] / \hat{P}_i(1-\hat{P}_i) \\ &= \frac{\hat{g}_i^*(\eta_0)\nabla_\eta \hat{f}_i^*(\eta_0) - \hat{f}_i^*(\eta_0)\nabla_\eta \hat{g}_i^*(\eta_0)}{\hat{g}_i^{*2}(\eta_0)\hat{P}_i(1-\hat{P}_i)} \equiv \frac{\hat{r}_i^*}{\hat{s}_i^*},\end{aligned}$$

where from (D5):

$$\hat{P} \equiv \left[\hat{f}_1 + \hat{\Delta}_1\right] / \left[\hat{g} + \hat{\Delta}\right] \equiv \hat{f}_1^* / \hat{g}^*.$$

Denote $\hat{\rho}_i \equiv \frac{\hat{r}_i}{\hat{s}_i}$ as the corresponding quantity without $\Delta-$adjustment factors (i.e. replace $\hat{P}$ with $\hat{f}_1/\hat{g}$). Note that these adjustment factors vanish in probability even in the absence of trimming, but vanish exponentially under the trimming employed below.

Lemma 8A: Primary Gradient Components. Define:

$$A_1 = \sum \tau_i [\, Y_i - P_i] \, \hat{\rho}_i / N; \qquad A_2 = \sum [\hat{\tau}_i - \tau_i] [\, Y_i - P_i] \, \hat{\rho}_i / N$$

Then:

$$\begin{aligned}&1) \;:\; N^{1/2} A_1 = N^{-1/2} \sum \tau_i [\, Y_i - P_i] \, \rho_i + o_p(1) \\ &2) \;:\; N^{1/2} A_2 = o_p(1), \text{ for } \hat{\tau}_i - \tau_i = \hat{\tau}_{wi} - \tau_{wi} \\ &3) \;:\; N^{r_p} A_2 = o_p(1), \text{ for } \hat{\tau}_i - \tau_i = \hat{\tau}_{xi} - \tau_{xi} \text{ and } r_p > r_3.\end{aligned}$$

Proof of Lemma 8A. Beginning with A_1, in (1), we show that the estimated weight may be taken as given by showing:

$$\delta \equiv N^{-1/2} \sum_i \tau_i [Y_i - P_i] [\hat{\rho}_i^* - \rho_i] = o_p(1).$$

Write $\delta \equiv \delta_1 + \delta_2$, where

$$\begin{aligned}\delta_1 &= N^{-1/2} \sum_i \tau_i \left[Y_i - P_i\right] \left[\hat{\rho}_i - \rho_i\right] \left(\hat{s}_i / s_i\right) \\ \delta_2 &= N^{-1/2} \sum_i \tau_i \left[Y_i - P_i\right] \left[\hat{\rho}_i - \rho_i\right] \left[\left(\hat{s}_i / s_i\right) - 1\right].\end{aligned}$$

Here,

$$|\delta_2| \leq N^{-1/2} \sup \left| \tau_i^{1/2} \left[\hat{\rho}_i - \rho_i\right] \right| \sup \left| \tau_i^{1/2} \left[\left(\hat{s}_i / s_i\right) - 1\right] \right|,$$

which is $o_p(1)$ from Lemma 2C. Therefore, to show that $\delta = o_p(1)$, it suffices to show $\delta_1 = o_p(1)$. We have:

$$\delta_1 = N^{-1/2} \sum_i \tau_i \left[Y_i - P_i\right] \varepsilon_i, \;\; \varepsilon_i \equiv \left[s_i \left(\hat{r}_i - r_i\right) - r_i \left(\hat{s} - s_i\right)\right] / s_i^2.$$

Exploiting the fact that $[Y_i - P_i]$ has 0 conditional expectation, we show that $\delta_1 = o_p(1)$ by showing that its expected square converges to zero. We have:

$$\begin{aligned}E\left(\delta_{11}^2\right) &= E\left[\sum_i \tau_i^2 \left[Y_i - P_i\right]^2 \varepsilon_i^2 / N\right] + \\ &\quad \sum_{i \neq j} \sum E\left[\left(\tau_i \left[Y_i - P_i\right] \varepsilon_i\right) \left(\tau_j \left[Y_j - P_j\right] \varepsilon_j\right)\right] / N.\end{aligned}$$

The first term is bounded from above by:

$$\sum_i E\left(\tau_i^2 \varepsilon_i^2\right) / N,$$

which converges to zero. Employing the fact that $E\left[Y_i - P_i \mid \mathrm{X}_i\right] = 0$, it can be shown that the second term also converges to zero. The result now follows.

Turning to (2), the argument for the smooth index-trimming function is based on a Taylor expansion of $\hat{\tau}_{wi}$, the observation that the derivative of a trimming function behaves as a trimming function, and the proof for A_1 above. Lemma 7 contains the details of this argument from which (2) follows. For (3), the argument is based on a characterization result for indicator X-trimming in Lemma 5.

Lemma 8B: Secondary Gradient Components. Define

$$B_1 = \sum \tau_i \left[\hat{P}_i - P_i\right] \hat{\rho}_i / N; \quad B_2 = \sum \left[\hat{\tau}_i - \tau_i\right] \left[\hat{P}_i - P_i\right] \hat{\rho} / N.$$

Then:

1) : $N^{1/2} B_1 = o_p(1)$ and $N^{1/2} B_2 = o_p(1)$ for $\hat{\tau}_i - \tau_i = \hat{\tau}_{wi} - \tau_{wi}$
2) : $N^{r_p} B_1 = o_p(1)$, $N^{r_p} B_2 = o_p(1)$ for: $\hat{\tau}_i - \tau_i = \hat{\tau}_{xi} - \tau_{xi}$, $r_p > r_3$.

Proof of Lemma 8B. Beginning with B_1 in (1) , we first simplify this term by showing:

$$\Delta \equiv N^{-1/2} \sum_i \tau_{wi} \left[\hat{P}_i - P_i\right] \left[\hat{\rho}_i - \rho_i\right] = o_p(1).$$

Bounding this term:

$$\begin{aligned} |\Delta| &= N^{-1/2} \left| \sum_i \tau_{wi} \left[\hat{P}_i - P_i\right] \left[\hat{\rho}_i - \rho_i\right] \right| \\ &\leq N^{1/2} \sup \left| \tau_{wi}^{1/2} \left[\hat{P}_i - P_i\right] \right| \sup \left| \tau_{wi}^{1/2} \left[\hat{\rho}_i - \rho_i\right] \right|, \end{aligned}$$

which is $o_p(1)$ from Lemma 3. Therefore:

$$N^{1/2} \mathbf{B}_1 = N^{-1/2} \sum_i \tau_{wi} \left[\hat{P}_i - P_i\right] \rho_i + o_p(1).$$

To further simplify $\mathbf{B}_1$ and show that it is $o_p(1)$, note that under an argument similar to that above:

$$N^{-1/2} \sum_i \tau_{wi} \left[\hat{P}_i - P_i\right] \rho_i \left[(\hat{g}_i / g_i) - 1\right] = o_p(1),$$

which implies:

$$N^{1/2} \mathbf{B}_1 = N^{-1/2} \sum_i \tau_{wi} \left[\hat{P}_i - P_i\right] \rho_i (\hat{g}_i / g_i) + o_p(1).$$

Next, recall that $\hat{P}_i = \left[\hat{f}_i + \hat{\Delta}_{1N}\right] / \left[\hat{g}_i + \hat{\Delta}_N\right]$. Under τ-trimming, the Δ-adjustment factors and their derivatives vanish exponentially when evaluated

at the true densities. Accordingly, under a Taylor series argument, we may replace $\hat{P}_i$ with $\hat{f}_i/\hat{g}_i$ to obtain:

$$N^{1/2}\mathbf{B}_1 = N^{-1/2}\sum_i \tau_{wi}\left[\hat{f}_i - P_i\hat{g}_i\right] r_i + o_p(1), r_i \equiv \rho_i/g_i.$$

Noting that r_i has expectation conditioned on the indices of 0 (Lemma 4), employ the same type of mean-square convergence argument used to analyze A. We have:

$$\begin{aligned} E\left[\left(N^{1/2}\mathbf{B}\right)^2\right] &= \frac{1}{N}E\left[\sum_i \tau_{wi}^2\left[\hat{f}_i - P_i\hat{g}_i\right]^2 r_i^2\right] + C, \\ C &= \sum_{i\neq j}\sum E\left[\tau_{wi}\left(\hat{f}_i - P_i\hat{g}\right)\tau_{wj}\left(\hat{f}_j - P_j\hat{g}_j\right) r_i r_j\right]/N. \end{aligned}$$

It can readily be shown directly that the first term above vanishes for large N. Turning to the more complicated cross-product terms in C, from iterated expectations:

$$\begin{aligned} C &= EE\left[\tau_{wi}\left(\hat{f}_i - P_i\hat{g}\right)\left(\tau_{wj}\left[\left(\hat{f}_j - P_j\hat{g}_j\right)\right] r_i r_j\right) \mid X\right] \\ &= E\left[E\left[\tau_{wi}\left(\hat{f}_i - P_i\hat{g}\right)\left(\tau_{wj}\left[\left(\hat{f}_j - P_j\hat{g}_j\right)\right]\right) \mid X\right] r_i r_j\right]. \end{aligned}$$

As the inner expectation only depends on the indices $W : Nx2$, denote this inner expectation as $H(W)$ and write:

$$C = E\left[H(W) r_i r_j\right] = E\left[H(W) E\left[r_i r_j\right] \mid W\right] = 0,$$

with the last result directly following from Lemma 4. The proof for B_1 in (1) now follows.

The proof for B_2 in (1), which is analogous to that for A_2 in Lemma 8A, part (2), readily follows from Lemma 7. To establish (2), we need to analyze B_1 and B_2 under X-trimming, The argument here, which is essentially the same as that for A_2 in Lemma 8A, part 3, follows from Lemma 5.

8.2 Main Results

As in the previous section, throughout this section, all results are provided under Assumptions (A1-6) and Definitions (D1-7).

Theorem 1. With $\hat{\tau}_i = \hat{\tau}_{xi}$ or $\hat{\tau}_{wi}$, define the quasi-likelihood as in Section 4.1:

$$\hat{Q}(\eta) \equiv \frac{1}{N}\sum_{i=1}^{N} \hat{\tau}_i \left(Y_{2i} Ln\left[\hat{P}_i(\eta)\right] + [1 - Y_{2i}]\, Ln\left[1 - \hat{P}_i(\eta)\right] \right)$$

and define $\hat{\eta} \equiv \arg\sup \hat{Q}(\eta)$. Then : $\hat{\eta} \xrightarrow{p} \eta_0$, the vector of true parameter values.

Proof of Theorem 1. Employing (D5) and deleting the i subscript for notational simplicity, define the probability functions:

$$\begin{aligned} \hat{P}(\eta) &\equiv \left[\hat{f}_1 + \hat{\Delta}_1\right] / \left[\hat{g} + \hat{\Delta}\right] \\ P_N(\eta) &\equiv [f_1 + \Delta_N] / [g + \Delta_N] \\ P(\eta) &\equiv f_1/g. \end{aligned}$$

With $P_N(\eta)$ replacing $\hat{P}(\eta)$ throughout, denote $Q_N(\eta)$ as the corresponding objective function. Finally, denote $Q(\eta)$ as the objective function obtained by replacing $\hat{P}(\eta)$ with $P(\eta)$ throughout. Then:

$$\left|\hat{Q}(\eta) - Q(\eta)\right| \leq \left|\hat{Q}(\eta) - Q_N(\eta)\right| + \left|Q_N(\eta) - Q(\eta)\right|.$$

Employing arguments similar to those in Klein and Spady (1993) and Lemma 4, it can be shown that each of these terms vanish in probability, uniformly

in η.[16] Next, write:

$$\begin{aligned} Q(\eta) &\equiv \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i\left[\, Y_{2i}Ln\left[P_i(\eta)\right] + \left[1-Y_{2i}\right]Ln\left[1-P_i(\eta)\right]\right] \equiv \bar{Q}(\eta) + R, \\ \bar{Q}(\eta) &\equiv \frac{1}{N}\sum_{i=1}^{N}\tau_i\left[\, Y_{2i}Ln\left[P_i(\eta)\right] + \left[1-Y_{2i}\right]Ln\left[1-P_i(\eta)\right]\right] \\ R &\equiv \frac{1}{N}\sum_{i=1}^{N}\left[\,\hat{\tau}_i-\tau_i\right]\left[\, Y_{2i}Ln\left[P_i(\eta)\right] + \left[1-Y_{2i}\right]Ln\left[1-P_i(\eta)\right]\right]. \end{aligned}$$

It can be shown that R vanishes in probability, uniformly in η. From standard uniform convergence arguments:

$$\sup_{\eta}\left|\bar{Q}(\eta) - E\left[\bar{Q}(\eta)\right]\right| \xrightarrow{p} 0.$$

Employing the identification condition in (A5), $E\left[\bar{Q}(\eta)\right]$ is uniquely maximized at η_0, which completes the argument.

Theorem 2. Defining $H_0 \equiv \nabla_\eta^2 E\left(L(\eta_0)\right)$:

$$\sqrt{N}\left[\hat{\eta}-\eta_0\right] \xrightarrow{d} N\left(0, -H_0^{-1}\right).$$

Proof of Theorem 2. With the quasi-likelihood defined under X-trimming (see D6) and with $\eta^+ \in [\hat{\eta},\ \eta_0]$, from a standard Taylor series expansion:

$$\begin{aligned} N^{r_p}\left[\hat{\eta}_p-\eta_0\right] &= -\hat{H}\left(\eta^+\right)^{-1}N^{r_p}\hat{G}(\eta_0), \\ \hat{H}\left(\eta^+\right) &= \nabla_\eta^2\hat{L}\left(\eta^+\right), \quad \hat{G}(\eta_0) = \nabla_\eta^1\hat{L}(\eta_0), \end{aligned}$$

[16] In analyzing the first term, it is important to exploit the fact that the δ-adjustment factors behaving as trimming functions in that they control the rate which denominators in various expressions tend to zero [see Klein and Spady (1993, proof of lemma 4, p. 414)]

To analyze the second term, it is important to note that from the assumption of bounded $X\prime s$, it follows that $P(\eta_o)$ is strictly bounded away from one and zero. It then follows that $P(\eta)$, a conditional expectation of $P(\eta_o)$, is also strictly bounded away from one and zero. While the assumption of bounded $X\prime s$ could be replaced by tail conditions, this assumption considerably simplifies the argument for the second term. [see Klein and Spady (1993, Proof of Theorem 3, p. 415)].

where we have employed X-trimming. Beginning with the Hessian component, as in the previous theorem define the probability functions: $\hat{P}(\eta)$, $P_N(\eta)$, and $P(\eta)$. From Lemma 3 and arguments very similar to those employed to analyze the averaged likelihood in Theorem 1, it can be shown that:

$$\sup_{\eta} \left| \hat{H}(\eta) - H(\eta) \right| \xrightarrow{p} 0.$$

From standard uniform convergence arguments, $H(\eta)$ converges in probability and uniformly in η to its expectation. It follows that $\hat{H}(\eta^+)^{-1} = H_0^{-1}(\eta_0) + o_p(1)$. Therefore, a convergence rate for the pilot estimator, $\hat{\eta}_p$, will follow from the rate at which the gradient converges to zero.

In the notation of Lemmas 8A and 8B:

$$N^{r_p}\hat{G}(\eta_0) = N^{r_p}[A_1 + A_2] + N^{r_p}[B_1 + B_2]$$

From Lemmas 8A and 8B, it now follows that:

$$N^{r_p}\left[\hat{\eta}_p - \eta_0\right] = o_p(1), \; r_p > r_3.$$

Employing the $\hat{\eta}_p$ to construct a smooth Index-trimming function, employ the quasi-likelihood under Index-trimming (D7) and a Taylor series expansion to obtain:

$$N^{1/2}[\hat{\eta} - \eta_0] = -\hat{H}\left(\eta^+\right)^{-1} N^{1/2}\hat{G}(\eta_0).$$

As above, $\hat{H}(\eta^+)^{-1} = H_0^{-1}(\eta_0) + o_p(1)$. From Lemmas 8A and 8B:

$$\begin{aligned} N^{1/2}\hat{G}(\eta_0) &= N^{-1/2}\sum \tau_{wi}[Y_i - P_i]\rho_i + o_p(1) \\ &\equiv N^{1/2}G(\eta_0) + o_p(1), \end{aligned}$$

where $G(\eta_0)$ is the gradient term with all estimated functions replaced by their (uniform) probability limits. The theorem now follows from a standard central limit theorem.

Turning to the outcomes equation, recall that it is given as:

$$Y_1 = Z\theta_o + u,$$

$\theta_o \equiv [\beta_o,\ \mu_o]$ and $Z \equiv [X,\ Y_2]$. Then, the IV estimator is given as :

$$\hat{\alpha}_{IV} = \left[\hat{Z}^*(\hat{\eta})' Z\right]^{-1} \hat{Z}^*(\hat{\eta})' Y_1,\ \hat{Z}^*(\eta) \equiv \left[X,\ \hat{P}(\eta)\right]$$

Consistency and asymptotic normality (Theorem 3 of Section 4.2) will now be immediate if the conditions given in the next lemma hold.

Lemma 8: With $Z^* \equiv [X,\ P(\eta_0)]$, under Assumptions (A1-4) and Definitions (D1-5):

$$\begin{aligned} 1) \ &: \ \left[\hat{Z}^*(\hat{\eta})' Z - Z^{*\prime} Z\right]/N = o_p(1), \\ 2) \ &: \ \sqrt{N}\left[\hat{Z}^*(\hat{\eta})' u - Z^* u\right]/N = o_p(1). \end{aligned}$$

Proof of Lemma 8. The first condition follows from Theorem 2 and Lemma 3. The second condition follows from a standard U-Statistics argument and is to be expected from Newey and McFadden (Handbook of Econometrics, vol. 4, Chapter 36, section 6.2 and Theorem 6.2).