The Semiparametric Binary Treatment Model Under Heteroscedasticity

Roger Klein
Rutgers University

Francis Vella
Rutgers University and N.Y.U.

May 17, 2001

Abstract

This paper provides an estimation procedure for semiparametric models with binary endogenous treatments and continuous outcomes. We discuss how heteroscedasticity of a general index form can provide identifying information even in the absence of exclusion restrictions. The estimators developed here exploit such information. We first formulate a semiparametric binary response model for the treatment. To capture both mean and heteroscedastic effects, this model is specified to have a double index form in which the treatment probability is an unknown function of two parametrically specified indices. By maximizing a quasi likelihood, we show that the resulting estimator is $\sqrt{N}-$consistent and asymptotically distributed as normal. We then employ the estimated treatment probability as an instrument for estimating a continuous outcomes equation. This estimator is also $\sqrt{N}-$consistent and asymptotically distributed as normal. We obtain these results by employing bivariate density estimators that depend on a sample covariance matrix and on data dependent windows that vary by observation (local smoothing). Simulation evidence suggests that this procedure works well even when the same variables drive conditional means and variances. We illustrate these estimators through an application based on estimating the impact of attending a government financed school on total schooling years acquired.

1 Introduction

A frequent objective of empirical microeconomics is the estimation of a treatment effect where undergoing the treatment is indicated by a dichotomous dummy variable. Examples in labor economics include the impact of union membership on wages and the effect of training programs on labor market performance. In estimating treatment effects it is often necessary to account for the endogeneity of the treatment decision. For example, when examining the impact of union membership on wages one should incorporate that a worker is only likely to join a union if he/she perceives he/she will be "better off", in terms of wages. To overcome this simultaneity, and its implication for parameter estimation, a variety of closely related instrumental variables (IV) type estimators exist. First, there are conventional IV procedures, including the natural experiment approach, which require a variable(s), or external shock, which affects the treatment decision but not the outcome. Second, there are the control function procedures developed by Heckman (1978, 1979) where one obtains an estimate of the unobserved heterogeneity underlying the endogeneity of the treatment to include as an additional regressor in the primary equation. These estimators generally rely on either on exclusion restrictions and/or distributional assumptions to obtain identification. As such information will not always be available, and the cost of incorrectly imposing it can be high, this paper suggests an approach to estimating treatment effects in which heteroscedasticity is exploited to obtain identification.

The intuition behind our approach is simple. In a binary choice treatment model with heteroscedasticity, the variables determining the heteroscedasticity directly affect the probability that the treatment is undertaken. Accordingly specific functions of these variables, depending on the form of the heteroscedasticity, can be employed as instruments for the treatment providing they do not appear in the conditional mean of the outcome equation. However, without knowledge of the form of the heteroscedasticity one cannot implement such a procedure. We bypass the estimation of the form of the heteroscedasticity by specifying it as some unknown function of a single index. By then estimating the probability of treatment as a function of two indices, one for the conditional mean and one for the conditional variance, we obtain an estimate of the probability which is necessarily non-linear in the exogenous variables. This non-linearity, which arises naturally in the presence of heteroscedasticity, provides identification of the treatment effect. We obtain such identification even in the extreme case in which there are no

exclusion restrictions in the linear indices upon which the treatment probability depends. Throughout this discussion, we make no distributional form assumptions.

We propose a two-step procedure whereby we first estimate the probability of treatment and then perform IV estimation using the estimated probability as an instrument for the treatment. The first step extends the binary choice model in Klein and Spady (1993) to allow for index heteroscedasticity. We show that this first step estimator is $\sqrt{N}$ consistent and is asymptotically distributed as normal. Using the first step estimates we show how one can consistently estimate the outcome equation via IV. We show that the IV estimator is also $\sqrt{N}$ consistent and asymptotically distributed as normal. Under certain conditions (models) these estimators will be efficient.

The objective of this paper is to provide an estimator for the treatment model where the equation explaining the treatment decision is contaminated with heteroscedasticity. Accordingly, in discussing and deriving the appropriate estimation procedure for this model we focus on a double index model where the indices capture the conditional mean and conditional variance respectively. However, with the exception of our identification strategy, the results below are applicable to any discrete choice model that is characterized by two indices.

The next section outlines the model and the estimation method. In Section 3 we provide and discuss the assumptions required to establish asymptotic results. As we are concerned with identification even in the absence of exclusion restrictions, we also discuss and provide identification results in this section. Here, it is important to emphasize that we are not concerned with identifying the index parameters upon which the treatment probability depends. Rather, we are concerned with identifying the treatment probability itself. Conditions for its identification are weaker than those that would be required for index parameters (See Ichimura and Lee (1991) and Lee (1995) for a related discussions). In Section 4 we establish the asymptotic properties of the estimators. In so doing, we sketch out the proofs, and provide complete technical details in the Appendix. Section 5 provides some simulation evidence and in Section 6 we provide an empirical example focussing on the impact of attendance at State financed high schools in Australia on the total level of an individual's education. Section 7 concludes.

2 Model and Motivation for Estimator

Consider the following model

$$\begin{aligned} Y_{1i} &= X_i\beta_0 + \mu_0 Y_{2i} + u_i \quad (1) \\ Y_{2i} &= \{X_i\pi_0 + v_i > 0\}, \quad (2) \end{aligned}$$

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 and μ_0 are unknown parameters with the latter being the treatment effect; and u_i and v_i are random disturbances. 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; $[X_i]$ is a vector of variables; γ_0 is an unknown parameter vector, and v_i^* is a homoscedastic random disturbances 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. It should be remarked 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 reasonable to restrict the elements of β_0 and π_0 so that the "mean effects" are linear in that they 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 do permit this second case of no exclusion restrictions.

With u_i depending on v_i^*, it is well-known that OLS estimates of the treatment parameters (1) will be inconsistent. As a result, a number of alternative procedures have been developed. One may estimate the probability that $Y_{2i} = 1$ to use as an instrument for Y_{2i} or simply insert it directly in (1) in place of Y_{2i}. This would identify the model provided $\Pr(Y_{2i} = 1|X_i)$ is not strictly linear in the $X_i's$. However, this form of linearity will typically occur in the tails of the $X_i's$ and thus relies on a small fraction of the sample for

identification. Note that while the "plug-in" approach requires that the first step be estimated consistently, the instrumental variables approach does not. If one made distributional assumptions, and modelled the heteroscedasticity, it would be possible to employ the control function methodology. For example, under normality and homoscedasticity, estimate the treatment equation by probit and the generalized residual $\lambda_i = \frac{\{Y_{2i} - \Phi(X_i\pi)\}\varphi(X_i\pi))}{\Phi(X_i\pi)\{1-\Phi(X_i\pi)\}}$, where φ and Φ are the probability density and cumulative distribution functions of the standard normal distribution respectively. Then, add λ_i as a regressor in (1). This procedure relies on λ_i being nonlinear in the $X_i's$.[1] The added complications of heteroscedasticity and unknown error distributions make this procedure more difficult to employ.

The structure of the model implies that $\Pr(Y_{2i} = 1|X_i)$ is non-linear in the $X's$. However, as the nature of the heteroscedastic function in the treatment equation is unknown, the exact nature of the non-linearity is unknown. Nevertheless, this suggests that higher orders of the $X's$ could be used as instruments for the treatment variable. One could choose the appropriate higher orders, or functions, by extending Donald and Newey (2001) to the setting with heteroscedasticity. 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)).

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

$$\begin{aligned} \Pr(Y_{2i} &= 1|X_i) = E[Y_{2i}|X_i\ ] = \Pr(Y_{2i} = 1|X_i\pi_0;\ X_i\gamma_0) \\ &= P([X_i\pi_0/S(X_i\gamma_0)]) \end{aligned} \tag{5}$$

where $P(.)$ is the distribution function for v_i^*. We estimate this probability by adapting the Klein and Spady (1993) estimator to account for the presence of heteroscedasticity. To implement this approach, we construct a probability which is a function of two linear indices. The first characterizes the conditional mean while the second captures the conditional variance. We then estimate those identifiable functions of the index parameters that

[1]The reliance on normality can be relaxed and replaced by some alternative parametric assumption or estimated non-parametrically. Nevertheless, the model remains only identified if the mapping from the index $X_i\pi$ to $Pr(Y_{2i} = 1|X_i)$ is non linear in the index and/or the vector explaining the probability includes at least one variable which does not explain Y_{1i}.

make it possible to estimate the treatment probability: $\Pr(Y_{2i} = 1|X_i) = \Pr(Y_{2i} = 1|X_i\pi_0; X_i\gamma_0)$. We remark again that this probability function can be identified without identifying the index parameters up to location and scale. We then propose estimating the second step by instrumenting the treatment indicator with its estimated probability.

In developing estimators that are quite different to those presented here, other papers have also employed higher order moments of the error distribution. Klein and Vella (2001a, 2001b) discuss how such information can be employed in estimating triangular simultaneous equations and sample selection models. Dagenais and Dagenais (1997) and Lewbel (1997) show that in models with measurement error in the regressors it is possible to exploit higher order information. Rigobon (1999) develops an estimation method for simultaneous equations models that also relies on heteroscedasticity for information. That paper assumes a specific form for heteroscedasticity and then uses the implied additional moments in a GMM framework to identify the parameters of the model. This, and the other approaches above, are fundamentally different from the present paper.

In terms of estimating the discrete choice model, there are several related papers. First, Ichimura and Lee (1991) also examine multiple index models. However, they focus on semiparametric least-squares while here we focus on likelihood estimation. Second, Lee (1995) adopts a likelihood approach for estimating multinomial choice models in a multiple index approach. While this strategy is similar to ours, there are several differences. First, with different models being considered (binary treatment vs. multinomial choice), the identification arguments differ. Second, the present paper differs from the others above in terms of the density estimators that are employed. In establishing that extremum estimators are asymptotically distributed as normal at the $\sqrt{N}$ parametric rate, there is typically a step in the argument where it is necessary to show that the (normalized) gradient to the estimator's objective function is asymptotically distributed as normal. To this end, it is important to show that the gradient is close in probability to a random variable whose expectation is $o\left(N^{-1/2}\right)$. At this point, bias reducing kernels are often employed to control for the bias. While we have obtained reasonable simulation results with such kernels in the case of single index models, we have not obtained satisfactory results with such kernels for the double index binary response model being studied. As an alternative, here we control for the bias by employing (mean-square optimal) kernels based on local smoothing (Abramson (1982)) and by exploiting a property of semipara-

metric probability functions due to Whitney Newey.[2] Though this approach involves different theoretical arguments (e.g. to handle $O(N)$ estimated local smoothing parameters and to exploit the property of semiparametric probability functions mentioned above), we have found that it performs well in monte-carlo simulations. We have found further that we could improve the finite sample performance of the estimators by employing dependent kernels that depend on an estimated sample covariance matrix for the indices, as advocated by Fukunaga (1972).

3 Assumptions and Definitions

In this section we first provide the assumptions and definitions that we employ to establish the asymptotic properties for the estimator. Second, we will discuss the relevance of these assumptions and consider several illustrative examples.

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). The vector X_i has a positive definite covariance matrix and is independent of the unscaled error v_i^*, in (3). With X_d as the discrete subvector of X, denote $X_d(k)$ as the kth discrete variable, k = 1,...,K$_d$. Without loss of generality, re-normalize the discrete variables so that zero is a support point for each X$_d(j)$ and assume:

$$\Pr\left(\cap_{i=1}^{k}\left[X_d(i) = 0\right]\right) > 0, \ k = 1, ..., K_d.$$

A2. Errors. For the error u_i in the continuous outcomes equation (1), assume that u_i is independent over i with $E(u_i|\ X_i) = 0$ and with

[2]Denote η_0 as a vector of true parameter values, $\Pr(\eta_0)$ as the semiparametric probability function, and $W(\eta_0)$ be a vector of indices upon which the model depends. Then:

$$E\left[\frac{\partial}{\partial \eta}\Pr\ \mid W(\eta)\right]_{\eta\ =\ \eta_0} = 0.$$

As the objective function's gradient depends on an estimate of this derivative, the above result turns out to be useful in controlling the bias in this gradient.

$E\left[|u_i|^2 \mid X_i\right]$ uniformly bounded. The error in the binary response model (2) is given as:

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

where the scaled error, v_i^*, is i.i.d. with finite variance and scaling function $S(\bullet)$, that is finite, bounded away from zero, and is not constant.

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

A4. Index Assumptions. Assume that the vector of indices, I, depends on two distinct continuous variables, X_1 and X_2 . Let $\bar{X}$ be the other (discrete and continuous) variables upon which the indices depend and write::

$$I \equiv [I_1,\ I_2] \equiv \left[X_1, X_2,\ \bar{X}\right] \Gamma.$$

Assume that all 2x2 submatrices of Γ have rank 2.

A5. Densities. Referring to (A4), let X_c denote the vector of continuous variables: (X_1, X_2) . Then, with $f(x_c|\bar{X}, Y_2)$ 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 $f(w|\bullet)$ is continuous in $\mathcal{R}^2$, is bounded away from 0 on any compact subset of its support, and that $\left|\nabla_1^i \nabla_2^j f(\bullet|\bullet)\right|$ 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. As we employ an estimate of the treatment probability to estimate the continuous equation in which the treatment enters, we will require certain convergence conditions for this estimated probability function. An index formulation of low dimension is important in this context.

With the possible exception of assumption **A1** and **A4,** the above assumptions are somewhat standard in index models. The restricition on probabilities for discrete variables in **A1** is not necessary, but simplifies the identification argument. Assumption **A4** essentially provides identification

conditions. This assumption differs from that for general multiple index models (see Lee and Ichimura (1991)) for two reasons. First, we are concerned with a double index model under multiplicative heteroscedasticity. Second, and perhaps most importantly, it should be emphasized that we are not concerned with identifying parameters in each of the two indices upon which the treatment probability depends. Rather, we are only concerned with estimating the treatment probability itself. In otherwords, we only seek to isolate the identifiable functions of the index parameters upon which the treatment probability depends.

To motivate assumption **A4,** and to relate it to other identification conditions in the literature, suppose that the model contains three continuous variables. If it is known that each index excludes one of these continuous variables and we know which variables are excluded, then it will follow from an argument in the appendix that the model is identified. We have not made such an assumption as we want to allow for the possibility of no exclusion restrictions. For this case, suppose that we know that Γ_1, a particular 2x2 submatrix of Γ, has full rank In this case, as will be shown directly below, there will exist a nonsingular transformation $\left(\Gamma_1^{-1}\right)$of the indices that will induce exclusion restrictions. As we are concerned with being able to identify probability functions, not index parameters, such a transformation will serve to isolate the identifiable functions of the index parameters from which we can identify the treatment probability. To illustrate this point, recall that the treatment probabilities are given as:

$$\Pr\left(Y_2 = 1 \mid X\right) = \Pr\left(Y_2 = 1 \mid [I_1,\ I_2]\right) = \Pr\left[Y_2 = 1 \mid \mathrm{X}\Gamma\right].$$

With $A = \Gamma_1^{-1}$:

$$\Pr\left[Y_2 = 1 \mid \mathrm{X}\Gamma\right] = \Pr\left[Y_2 = 1 \mid \mathrm{X}\Gamma A,\right]$$

Where the transformed indicies are given as:

$$W \equiv \mathrm{X}\Gamma A = [X_1,\ X_2,\ X_3] \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \eta_{10} & \eta_{20} \end{bmatrix}.$$

We then argue below that it is possible to identify the nuisance parameter vector $\eta_o \equiv [\eta_{1o},\ \eta_{2o}]$ and in so doing can identify the treatment probability.

The above argument can be shown to yield identification in this example, but it requires that we know which 2x2 submatrix is nonsingular. In the case

of known exclusion restrictions, one of the submatrices must be nonsingular and we will know which one it is. However, when there are no exclusion restrictions, in the absence of a very complicated estimation strategy, it will not suffice to know that some 2x2 submatrix is nonsingular.[3] Accordingly, to avoid a computationally intensive estimation strategy, here we assume that all 2x2 submatrices are non-singular. In this case we will then know that any particular submatrix is nonsingular.

To complete the identification discussion, we now need to discuss identification for the "reduced form" nuisance parameters , η_o. For expositional purposes, we will consider the above case when there are no exclusion restrictions. A similar and somewhat less complicated argument holds when the initial indices satisfy exclusion restrictions. The following theorem, which is proved in the Appendix, provides an identification result based on the above discussion.

Theorem 1 *Write the initial indicies (prior to linear transforms) as:*

$$I \equiv [I_1|I_2] = X\Gamma,$$

where $\Gamma : Kx2$ *has rank 2,* $K \geq 2$. *Assume that all 2x2 submatrices of* Γ *have rank 2 and that* X *contains two distinct continuous variables. For* $K > 2$ *(otherwise identification trivially holds*[4]*), there exists an identified parameterization with associated identified index vectors,* $W \equiv [W_1, W_2]$, *such that*

$$Pr(Y_2 = 1|X) = Pr(Y_2 = 1|W)$$

Having established identification for the semiparametric probability function, we employ kernel density estimators to estimate it. As is standard in

[3]It may be that all submatrices are not nonsingular, and it may also be that one does not know which submatrix has full rank. For this case, it is possible to develop a consistent estimator by separately maximizing a quasi-likelihood over all submatrices. Namely, proceed as if any given submatrix has full rank and parameterize the model accordingly as above. It can be shown that the consistent estimator is given as the maximum over all submatrices. As this method is computationally intensive, here we proceed under the assumption that a given selected submatrix has full rank.

[4]If K = 2, the binary response probabilitiy is immediately identified with one index being X_1 and the other being X_2.

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 indicies. 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 matrix for W. In what follows, we will first define these estimators and then discuss their properties.

D1. Pilot Density Estimators. 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. For fixed (small) $\delta : 0 < \delta < 1/12$, select $\alpha : 1/12 < \alpha < 1/(10+\delta)$. The pilot window is then given as $h_p \equiv N^{-(2/3)\alpha}$. 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:

$$k_{ij}^s(h,\lambda) \equiv \frac{\det\left(\hat{\Sigma}_s\right)^{-1/2}}{[\lambda h]^2} K\left(T_1\left[W_i - W_j\right]/[\lambda h]\right) K\left(T_2\left[W_i - W_j\right]/[\lambda h]\right)$$

With $g_1(w)$ as joint density for $W \equiv [W_1, W_2]$ conditioned on $Y_2 = s$ and P as the unconditional probability that $Y_2 = 1$, define the pilot estimator for $f_s(w) \equiv P g_1(w)$ as:

$$\hat{\pi}_{1i} \equiv \frac{1}{N} \sum_{j \neq i} Y_{2j} k_{ij}^s(h_p, 1), \; s = 1, 0.$$

Similarly, define $\hat{\pi}_{0j}$ with $(1 - Y_{2j})$ replacing Y_{2j} throughout.

D2. Locally Smoothing Parameters. Referring to (D1), denote $\hat{m}_1$ as the geometric mean of the $\hat{\pi}$'s and let $\hat{\gamma}_{1j} \equiv [\hat{\pi}_{1j}/\hat{m}_1]$. Then, for $j = 1, ..., N$, define local smoothing parameters as:

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

To define the smoothed indicator, $\hat{d}$, with α and δ given in (D1), select $\varepsilon : 0 < \varepsilon < (1/4)(\alpha - \delta)$. Then:

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

Define $\hat{\lambda}_{0j}$ in a similar manner.

D3. Second Stage (Locally Smoothed) Density Estimators. For α given in (D1), define a global window component: $h \equiv N^{-\alpha}$. With $\hat{\lambda}_s$ as the vector of local smoothing parameters in (D2) and with $f_s(s)$ and k_{ij}^s defined as in (D1), define a locally-smoothed estimator for f_s as:

$$\hat{f}_s(w) \equiv \frac{1}{N}\sum_{j \neq i} Y_{2j} k_{ij}^s \left(h, \hat{\lambda}_s\right), \; s = 1, 0.$$

D4. Semiparametric Probability Function. Define:

$$\hat{P}(\eta) \equiv \hat{f}_1(w)/\hat{g}(w),$$

where $\hat{g}(w) \equiv \hat{f}_1(w) + \hat{f}_0(w)$ estimates the unconditional density for W.

As stated above, for known local smoothing parameters (bounded away from zero), Abramson showed that the locally-smoothed density estimator in (D3) is optimal in a mean-squared error sense. This estimator also has the desired bias reducing properties that are required to establish the asymptotic

results below.[5] Throughout, as density estimators appear in various denominators, we will trim out observations for which the exogenous variables are outside of a compact set. To avoid boundary bias problems, we need to insure that local smoothing parameters are evaluated on a set. containing that on which probability functions are evaluated. Simultaneously, we need to insure that local smoothing parameters are not "too small". Accordingly, we do permit local smoothing parameters to tend to zero, but in (D2) keep them above $1/Ln(N)$. With global window sizes being a reciprocal power of the sample size, $1/Ln(N)$ is suitable for our purposes.

Finally, it is important to point out that 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.

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. Throughout, 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 be cases in which exclusion restrictions for the various indices are justified.

[5]The proof strategy here is based on a frequently employed U-statistic argument that requires that the bias of the statistic being studied is sufficiently small. Here, we satisfy this condition by employing local smoothing in conjunction with an expectations property of semiparametric probability functions.

4.1 Binary Response

We begin with the binary response model as these results are of interest in themselves and are here required to estimate a continuous outcomes equation that depends on the treatment. Recall that this equation is given as:

$$Y_{2i} = \{X_i\pi_o + v_i > 0\}, v_i \equiv S(X_i\gamma_o)v_i^*$$

With $\eta \equiv [\eta_1,\ \eta_2]$ and the "reduced form" indices W_1 and W_2 defined as above, the estimator for this binary response model is given as:

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

Here, with the semiparametric probability function given as $\hat{P}_i(\eta)$ in (D4):

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

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

With the details given in the Appendix, it can be shown that with $L(\eta)$ obtained from $\hat{L}(\eta)$ by replacing $\hat{P}_i(\eta)$ with its uniform probability limit $P_i(\eta)$:

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

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

Theorem 2 *Under (A1-5) and (D1-4):*

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

To establish normality for the estimator, it will be important to characterize the gradient to the objective function as being close in probability to a random variable to which a central limit theorem will apply. With the Appendix providing the complete argument, here we briefly sketch out the

proof strategy. In so doing, it will become clear how the argument depends on an important property of semiparametric probability functions that makes it possible to appropriately control the bias in the estimator.[6] With an estimated weight function defined (to incorporate a trimming function) as:

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

the gradient has the form:

$$\sqrt{N} \sum \left[Y_i - \hat{P}_i \right] \hat{\rho}_i / N = N^{-1/2} \sum \left[Y_i - P_i \right] \hat{\rho}_i - N^{-1/2} \sum \left[\hat{P}_i - P_i \right] \hat{\rho}_i$$

In the Appendix, we show that the estimated weight function may be replaced by its uniform probability limit, ρ_i. Accordingly, if the last term vanishes in probability, it will readily follow that the above normalized gradient is asymptotically distributed as normal. Recalling that $\hat{P}_i = \hat{f}_i / \hat{g}_i$, in the Appendix it is shown that

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

As this last term is a linear combination of kernel density estimators, it is a U-statistic to which a standard project argument applies. A critical part of this argument is to show that R is "nearly" unbiased: $E(R) = o\left(N^{-1/2}\right)$. With biased reducing kernels, this result follows by showing that

$$\delta_i \equiv E\left(\left[\hat{f}_i - P_i \hat{g}_i \mid X_i \right] \mid X_i \right) / g_i = o\left(N^{-1/2}\right)$$

However, once we have shown that the gradient has the above structure (an argument for which local smoothing is used here), then bias reducing kernels are not required to control the bias. To explain why this is the case, with R_i as the argument of the sum defining R above:

$$E\left(R_i | X_i\right) = \delta_i \rho_i,$$

[6] We developed the current approach because we did not obtain satisfactory results in finite samples using bias reducing kernels.

Noting that δ_i only depends on index values, $\delta_i \rho_i$ will have expectation zero if $E[\rho_i | W_i] = 0$. As elaborated on in the Appendix, this latter result follows from a property of semiparametric probability derivatives.[7]

Once it is shown that R has expectation of zero, in the Appendix, we employ projection arguments to show that R converges in probability to zero. Having established that R vanishes in probability, from a conventional Taylor series and a standard uniform convergence argument:

$$\begin{aligned} \sqrt{N}[\hat{\eta} - \eta_0] &= -H_o^{-1}\sqrt{N}\sum_{i=1}^{N} \tau_i \frac{[Y_{2i} - P_i(\eta_0)]}{P_i(\eta_0)[1 - P_i(\eta_0)]} \frac{\partial}{\partial \eta} P_i(\eta_0), \\ H_o &\equiv E\left[\frac{\partial^2}{\partial \eta \partial \eta'} L(\eta_0)\right]. \end{aligned}$$

¿From a standard central limit theorem, we now have:

Theorem 3 *Under (A1-5) and (D1-4):*

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

It should be noted that we make no efficiency claims for this estimator. For a general double index model, provided that identification holds, it appears from Chamberlain (1986) that this estimator attains a semiparametric efficiency bound. However, under the additional restrictions imposed by multiplicative index heteroscedasticity, it does not appear to be efficient. It remains an open question as to whether or not it is possible to develop an efficient estimator for this model. In the next section, we employ the above results to analyze the continuous outcomes equation.

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:

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

[7] Recall that

$$\rho_i \equiv (\tau_i / [P_i(1 - P_i)]) \frac{\partial}{\partial \eta} P_i(\eta_0),$$

where all terms other than the probability derivativie depend only on the indices. Therefore, ρ_i will have conditional expectation of zero if $\frac{\partial}{\partial \eta} P_i(\eta_0)$ has conditional expectation of zero. It is this property of semiparametric probability derivatives (further examined in the Appendix) that we exploit here.

Then, letting $\hat{Z}^*(\eta) \equiv \left[X,\ \hat{P}(\eta)\right]$ be an instrument for Z, the IV estimator, $\hat{\theta}_{IV}$ 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 6 in the Appendix, section 8.2, 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 4 *Under (A1-5) and (D1-4):*

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

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

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

Then: $\hat{\Omega} = \Omega + o_p(1)$.

Before proceeding, it should be noted 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 some simulation exercises. As the number of factors determining the nature of the simulation is very large an exhaustive examination of

the estimator under various conditions is not feasible. Accordingly we adopt the following strategy. We examine the performance of our procedure in 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 is given as:

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

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.

In the first experiment we conduct simulations with a sample size of 1000 and with 500 replications. In the first step we estimate the binary choice model and, given the normalization involved, we omit one variable from each index and set the coefficient on one of the others to 1. Given the design the true values are 2 and -1, and to reduce computation time these are the starting values employed.[8] The average estimate for these two parameters are 2.156 and -1.137 with standard deviations of .809 and .728. Thus the estimates are reasonably unbiased although they are not precisely estimated. As one would expect, this precision is substantially increased (and the bias is also decreased) in simulations for sample size 2000 reported below. Using the first step estimates we compute the implied probability which we then employ as an instrument for Y_{2i} in estimating the second equation. In table 1 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 as each contributes differently in the heteroscedasticity index.

Column 1 reports the average value of the OLS estimates from the second step. Recall from the design that the true value for each coefficient is 1. Each of the exogenous variables displays a level of bias in the range of 3.5 to 10 percent. The standard errors for the estimates, given below the estimates

[8]To insure that the final estimates were not sensitive to starting values, for a subset of the simulations we experimented with very different starting values and obtained the same estimates.

and shown 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 .587 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 13.5 percent although this represents a marked improvement over the OLS eliminates.

Column 3 presents the estimates from our procedure. For each of the parameters on the exogenous variables there is a remarkable reduction in the bias in comparison to the OLS estimates. In fact, it is quite clear that 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 1.7 percent bias. Note, importantly, that the standard deviation for the treatment effect is smaller for this estimator than that shown in Column 2.

In Table 1b we repeat the simulation exercise although we now increase the sample size to 2000. In the first step the average estimates are now 2.059 and -1.046 with standard deviations .336 and .304. Thus the increase in the sample size has resulted in a significant increase in the degree of accuracy, in terms of bias and precision. A number of points are worth noting from Table 1b. First, the IV estimator formulated here continues to dominate the alternative estimators for this model. 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 estimator developed here is remarkably accurate with the estimates seemingly unbiased for all coefficients. Perhaps the most remarkable feature of Table 1b is the increase in efficiency for this estimator as it now displays a standard deviation significantly lower than that for the alternative IV procedure.

6 Empirical Example

We now apply our approach to examine the effect of school type on the number of years of schooling attained. 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 (for recent examples see Evans and Schwab 1995, Neal 1997 and Vella 1999). Unlike previous papers which examine the effect of catholic schools on education we examine the effect of attending a government or state financed school. 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 endogenous treatment 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 would 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. Moreover, the variable appears to have a effect for both males and females. 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 hypothesis that attitudes are exogenous to educational attainment should be rejected. 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. Note that we treat the variables as continuous and, accordingly, the model satisfies the assumptions required to ensure the first step probabilities are identified.

In Column 1 of Table 2 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, shown in parenthesis under the estimate, is small indicating the effect is relatively precisely estimated. This effect is not particularly large given the large premium associated with attending a private institu-

tion 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 *Attitude* variable has a strong positive effect on years of education acquired.

Before employing the estimators formulated here, we employ two alternative approaches for accounting for the simultaneity. The first is to 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 estimate the probability that the individual attends a government school. We first estimate such a model by probit and the estimates are reported in Column 1 of Table 3. As this represents an investment in relatively lower quality human capital, given that private schools are the comparison group, one suspects that these coefficients should have the opposite signs to those in Column 1. With the exception of the Roman Catholic and gender dummies this is found to be the case although there is some inconsistency with respect to which variables across the two columns are statistically significant. Note that the coefficient on the Roman Catholic dummy is large in magnitude reflecting the tendency of Catholic individuals to attend private Catholic Schools.

The second column of Table 2 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. Note that 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 3 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. Overall the evidence in Columns 2 and 3 confirms our suspicion 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.

Before presenting results based on index heteroscedasticity, we first test the above Government school attendance model 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. 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.

In columns 2 and 3 of Table 3 we report the estimates from estimating the double index binary choice model. Due to the normalizations required we exclude one continuous variable from one index and one indicator function from the other. We also constrain one of the variables in each index to have coefficient 1. The variables involved in the normalizations can be inferred from the Table. Before returning to a discussion of the estimates of the primary equation it is useful to consider the implications for the probability of attendance at a government school from these estimates. Due to the nature of the model it is not straightforward to interpret the coefficients. 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 5. Without exception, the partial effect for each of the variables have the same sign across estimation procedure. Perhaps the most striking difference is the magnitude of the effect of the catholic variable. In the probit model the estimated effect is over 50 percentage points while for the double index model the effect is around 30 percentage points. In general the similarity of the estimates are reassuring while the difference in magnitude highlights some potential problems with the probit model.

In the fourth column of Table 2 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 with the exception of the coefficient on the *Roman Catholic* and *Govt* variables and the two variables capturing the influence of parent's education. However, despite the change in the values the magnitudes are all reasonable. The most striking increase is in the magnitude of the *Govt* school which now indicates that the effect is 1.3 years. This estimate seems far more reasonable given the educational behavior of those attending non-government schools. The coefficients on the parents education variables indicate that the direct effects apparent in column 1 appear to be now working through the *Govt* variable. The opposite, however, is true for the *Roman Catholic* variable. That is, Catholics appear

to obtain approximately .3 years less education although this is largely offset by their tendency to attend non government schools. Also, the effect is not statistically significant at the 10 percent level.

7 Conclusions

This paper describes and illustrates an estimation procedure for models with binary endogenous treatments when there are no available exclusion restrictions to enable instrumental variables estimation. Rather than rely on inherent non-linearities existing in the tails of the cumulative distribution function to obtain identification we illustrate how the presence of heteroscedasticity in the model can provide identification. We first develop an estimator for the semiparametric binary choice model with heteroscedastic errors. Using the predicted probability from this model as an instrument for the treatment variable, we then 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 the procedure formulated here works well even in the case where the same variables are driving the conditional means and variances of both the treatment and outcome equations. An empirical example, based on estimating the impact of attending a government financed school, on total schooling years acquired illustrates the use of our approach.

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] 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.

[5] Donald, S., and W.Newey (2001):"Choosing the Number of Instruments," *Econometrica*, forthcoming

[6] 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.

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

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

[9] ———- (1979): "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.

[10] 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.

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

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

[13] Klein, R. and F.Vella (2001a):"Employing heteroscedasticity to Identify and Estimate Triangular Semiparametric Models" *in progress.*

[14] —— (2001b): "Semi-Parametric Estimation of the Heteroscedastic Sample Selection Model in the Absence of Exclusion Restrictions", *in progress*

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

[16] Lewbell, 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.

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

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

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

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

[21] Rigobon, R, (1999): "Identification through heteroscedasticity" working paper, Sloan School MIT.

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

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

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

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

Table 1a: Simulation Results

Variable	**OLS**	**IV**	**SP-IV**
n=1000			
Intercept	1.205	1.065	1.005
	(.065)	(.147)	(.119)
x_1	1.097	1.028	1.004
	(.053)	(.075)	(.074)
x_2	1.068	1.019	1.005
	(.051)	(.060)	(.058)
x_3	1.035	1.009	1.004
	(.050)	(.047)	(.052)
Y_{2i}	.587	.865	.983
	(.099)	(.279)	(.224)

Table 1b: Simulation Results

Variable	**OLS**	**IV**	**SP-IV**
n=2000			
Intercept	1.203	1.044	1.001
	(.047)	(.114	(.089)
x_1	1.094	1.018	1.001
	(.038)	(.053)	(.052)
x_2	1.066	1.013	1.002
	(.038)	(.043)	(.043)
x_3	1.034	1.008	1.002
	(.034)	(.034)	(.039)
Y_{2i}	0.592	.909	.993
	(.072)	(.215)	(.163)

Table 2: Estimating the Impact of Attendance of A Government School on Years of Education

	OLS	**IV**	**CF**	**S-P.IV**
	School	**School**	**School**	**School**
Constant	6.025	5.408	5.597	7.329
	(0.238)	(0.795)	(0.578)	(0.745))
Age	0.193	0.195	0.195	0.184
	(0.008)	(0.009)	(0.008)	(0.009)
Aust	0.030	0.063	0.053	-0.010
	(0.069)	(0.081)	(0.075)	(0.077)
Both Parents	0.294	0.306	0.303	0.297
	(0.062)	(0.064)	(0.063)	(0.068)
Mother/Degree	0.283	0.365	0.340	0.193
	(0.119)	(0.156)	(0.138)	(0.173)
Father/Degree	0.659	0.734	0.711	0.508
	(0.090)	(0.128)	(0.110)	(0.126)
Siblings	-0.117	-0.118	-0.118	-0.112
	(0.011)	(0.011)	(0.011)	(0.011)
Roman Cath	-0.045	0.129	0.075	-0.294
	(0.052)	(0.220)	(0.158)	(0.211)
Male	0.215	0.218	0.218	0.216
	(0.045)	(0.045)	(0.045)	(0.046))
Attitudes	0.081	0.084	0.083	0.074
	(0.005)	(0.005)	(0.005)	(0.006)
Govt	-0.559	-0.050	-.206	-1.315
	(0.062)	(.626)	(.439)	(0.615))
Mills Ratio			-.200	
			(.247)	

Table 3: 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)		
Aust	-0.296	1	
	(.074)		
Both Parents	-0.094	0.950	-5.366
	(.064)	(0.291)	(0.766)
Mother/Degree	-0.583	9.102	-1.035
	(.101)	(7.765)	(11.682)
Father/Degree	-0.549	0.816	9.378
	(.078)	(0.541)	(0.822)
Siblings	0.020	-0.033	-0.906
	(.011)	(0.068)	(0.149)
Roman Cath	-1.270	2.341	3.326
	(.044)	(0.273)	(0.590)
Males	-0.032	-0.019	1.398
	(.045)	(0.287)	(0.578)
Attitudes	-0.022	0.040	0.978
	(.005)	(0.048)	(0.071)

Table 4: 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 5: Partial Effects

	PROBIT	**S-P**
Age	-.010	-.008
Aust	-.020	-.059
Both Parents	-.059	-.078
Mother/Degree	-.150	-.152
Father/Degree	-.164	-.207
Siblings	.004	.005
Roman Cath	-.530	-.329
Male	-.009	-.003
Attitudes	-.034	-.028

8 Appendix

The Appendix is organized into two subsections, with the first stating and proving all intermediate lemmas that we will 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 (D2-D4) 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 $\Pr(Y_2 = 0)g_1(w)$, where $g_1(w)$ 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 kth partial derivative of f with respect to η, with $\nabla_\eta^0 f \equiv f$.

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 (D3) and denote $\hat{f}(w; \Sigma(\eta))$ as the corresponding estimator with $\Sigma(\eta)$ replacing $\hat{\Sigma}(\eta)$. Define $\hat{f}_0(\bullet)$ analogously. Then under (A1-A5), for η and w in compact sets:

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

Proof of Lemma 1: From a Taylor series expansion:

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

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 (D1-2) 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 $\gamma_j \equiv [E(\hat{\pi}_j)/m], \bar{d}_j \equiv d\left(\gamma_j\right)$, 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 $\hat{f}\left(w;\hat{\lambda}\right)$ as the estimator of f and let $\hat{f}\left(w;\bar{\lambda}\right)$ be the corresponding estimator with $\bar{\lambda}$ replacing $\hat{\lambda}$. Then: $\Delta_k \equiv \left|\nabla_\eta^k\left[\hat{f}\left(w;\hat{\lambda}_1\right) - f(w)\right]\right|$ is bounded above by:

$$\begin{array}{ll}
\mathbf{A}_k: & \left|\nabla_\eta^k\left[\hat{f}\left(w;\hat{\lambda}\right) - \hat{f}\left(w;\bar{\lambda}\right)\right]\right| + \\
\mathbf{B}_k: & \left|\nabla_\eta^k\left[\hat{f}\left(w;\bar{\lambda}\right) - E\left(\hat{f}\left(w;\bar{\lambda}\right)\right)\right]\right| + \\
\mathbf{C}_k: & \left|\nabla_\eta^k\left[E\left(\hat{f}\left(w;\bar{\lambda}\right)\right) - f(w)\right]\right|.
\end{array}$$

Lemma 2 below analyzes these components so as to obtain uniform convergence rates for Δ_k, k = 0,1,2.

Lemma 2 Under (A1-4) and (D1-5), for η and w in compact sets:

$$\sup_{w,\eta}\left|\nabla_\eta^k\left[\hat{f}\left(w;\hat{\lambda}\right) - f(w)\right]\right| = \begin{cases} o_p\left(h^3\right), & k=0 \\ o_p\left(h^2\right), & k=1 \\ O_p\left(h^2\right), & k=2 \end{cases}$$

Proof of Lemma 2. In what follows, we will establish Lemma 2 by obtaining convergence rates for the terms in **A, B**, and **C** above. We begin with **C**, as this term converges slower than the other two and therefore will establish the desired convergence rates.

C

We will establish that:

$$\mathbf{C}_k = \begin{cases} o_p\left(h^3\right), & k=0 \\ o_p\left(h^2\right), & k=1 \\ O_p\left(h^2\right), & k=2 \end{cases}$$

Beginning with k = 0, let $\bar{w}$ be a fixed value of w and define $P \equiv \Pr\left(Y_2 = 1\right)$. Then, the expectation $E\left[\hat{f}\left(\bar{w}\right)\right]$ is given as:

$$P \int\int \frac{K\left(T_1\left[\bar{w}-w\right]/\left[\bar{\lambda}\left(w\right)h\right]\right) K\left(T_2\left[\bar{w}-w\right]/\left[\bar{\lambda}\left(w\right)h\right]\right) g\left(w\right) dw}{\det\left(\Sigma\right)^{1/2}\left[\bar{\lambda}\left(w\right)h\right]^2}$$

With $z \equiv T\left(w-\bar{w}\right)/h$, the above expression becomes:

$$P \int\int \frac{K\left(z_2/\bar{\lambda}_1\left(\bar{w}+hT^{-1}z\right)\right) K\left(z_2/\bar{\lambda}_1\left(\bar{w}+hT^{-1}z\right)\right) g\left(\bar{w}+hT^{-1}z\right)}{\bar{\lambda}\left(\bar{w}+hT^{-1}z\right)^2} dw$$

Then, from a Taylor series expansion about $h = 0$:

$$\begin{aligned} E\left[\hat{f}\left(\bar{w}\right)\right] &= Pg_1\left(\bar{w}\right) + h^2\hat{C}_2 + h^4\hat{C}_4 \Rightarrow \\ \left|E\left[f\left(\bar{w}\right)\right] - f\left(\bar{w}\right)\right| &\leq h^2\left|\hat{C}_2\right| + h^4\left|\hat{C}_4\right| \\ &\leq h^2\left|C_2\right| + h^2\left|\hat{C}_2 - C_2\right| + h^4\left|\hat{C}_4\right| \end{aligned}$$

where the h and h^3 terms vanish from the symmetry of the kernel, K.

¿From Abramson (1982), with C_2 as the probability limit of $\hat{C}_2$, $C_2 = 0$ from local smoothing. Proceeding to the second term, it can be shown that $\left|\hat{C}_2 - C_2\right| = o_p\left(h\right)$ uniformly in $\bar{w}$, η. For example, with $\hat{\pi}_1\left(\bar{w}\right)$ as the pilot estimator and h_p as the pilot window in (D1), one of the components of

$\left|\hat{C}_2 - C_2\right|$ will uniformly have the order of the term $\left|\nabla_\eta^2 \hat{\pi}_1(\bar{w}) - \nabla_\eta^2 f_1(\bar{w})\right| Ln(N)$, which is bounded above by:

$$\begin{aligned} & Ln(N)\left[\left|\nabla_\eta^2 \hat{\pi}_1(\bar{w}) - \nabla_\eta^2 E\hat{\pi}_1(\bar{w})\right| + \left|\nabla_\eta^2 E\hat{\pi}_1(\bar{w}) - \nabla_\eta^2 f_1(\bar{w})\right|\right] \\ = \; & Ln(N)\left[O_p\left(\left[N^{-1/2}h_p^4\right]^{-1}\right) + O_p\left(h_p^2\right)\right] \end{aligned}$$

Recall from (D1) that $h = N^{-\alpha}$ *and* $h_p = N^{-(2/3)\alpha}$, where $0 < \alpha < 1/(10+\delta)$ and $0 < \delta < 1/12$. Then:

$$Ln(N)\left[N^{-1/2}h_p^4\right]^{-1} = N^{-\alpha} Ln(N) N^{-1/2} N^{5\alpha(2/3)} < N^{-\alpha} = h$$

It also follows from the window restrictions that $h_p^2 \leq h$. With the analysis for other components being similar, $h^2\left|\hat{C}_2 - C_2\right| = o_p(h^3)$ uniformly.

Proceeding to the final term, it can be shown that $h^4\left|\hat{C}_4\right| = o_p(h^3)$, which completes the argument for k = 0.

For the case in which $k = 1$, it can be shown that the appropriate dominance condition holds making it possible to move the derivative operator outside an expectation to yield:

$$E\left[\nabla_\eta^1 \hat{f}\left(w;\bar{\lambda}_1\right)\right] = \nabla_\eta^1 E\left[\hat{f}\left(w;\bar{\lambda}_1\right)\right] = \nabla_\eta^1 f(w) + \nabla_\eta^1\left[h^2\hat{C}_2(\eta) + h^4\hat{C}_4(\eta)\right],$$

where the last equality follows from the argument for $k = 0$ above. We now have $\left|E\left[\nabla_\eta^1 \hat{f}\left(w;\bar{\lambda}_1\right)\right] - \nabla_\eta^1 f(w)\right|$ bounded above by:

$$h^2\left|\nabla_\eta^1 C_2(\eta)\right| + h^2\left|\nabla_\eta^1 \hat{C}_2(\eta) - \nabla_\eta^1 C_2(\eta)\right| + h^4\left|\nabla_\eta^1 \hat{C}_4(\eta)\right|$$

As above, the first term vanishes as $C_2(\eta) = 0$. For the second and third terms it can be shown uniformly that

$$\left|\nabla_\eta^1 \hat{C}_2(\eta) - \nabla_\eta^1 C_2(\eta)\right| = o_p(1) \text{ and } h^4\left|\nabla_\eta^1 \hat{C}_4(\eta)\right| = o_p\left(h^2\right),$$

which completes the argument for k = 1.

For $k = 2$, the argument is similar to that above. Having established rates for **C,** turn to the terms in **B**.

B

¿From Klein and Spady [Lemma 1]

$$\mathbf{B}_k \equiv \sup_{w,\eta} \left| \nabla_\eta^k \hat{f}\left(w;\bar{\lambda}_1\right) - E\left(\nabla_\eta^k \hat{f}\left(w;\bar{\lambda}_1\right)\right)\right| = O_p\left(N^{-1/2}h^{2+k}\right), k = 0,1,2..$$

A

Employing a standard Taylor series argument, it can be shown that $\mathbf{A}_k$ converges to zero faster than $\mathbf{C}_k$ for k = 0,1,2. For example, for k = 0:

$$\begin{aligned} \left|\hat{f}\left(w;\hat{\lambda}\right) - \hat{f}\left(w;\bar{\lambda}\right)\right| &\leq \sup_{\eta,w}\left|\frac{\partial}{\partial\lambda}\hat{f}\left(w;\bar{\lambda}\right)\right|\Delta_\lambda + o_p\left(N^{-1/2}\right), \\ \Delta_\lambda &\equiv \sup_{w}\left|\hat{\lambda}\left(w\right) - \bar{\lambda}\left(w\right)\right| = O_p\left(N^{\varepsilon}N^{-1/2}/h_p^2\right). \end{aligned}$$

In obtaining the uniform rate on Δ_λ , recall that $\hat{\lambda} = \lambda\left(\hat{\gamma}_j\right)$, where $\hat{\gamma}_j$ is essentially a density estimator. Then:

$$\lambda\left(\hat{\gamma}_j\right) - \lambda\left(\bar{\gamma}_j\right) = \left[\frac{\partial}{\partial\gamma}\lambda\right]\left(\hat{\gamma}_j - \bar{\gamma}_j\right)$$

The first component is $\mathrm{O}_p\left(N^{\varepsilon}\right)$, while the second (see, e.g. Klein and Spady, Lemma 1) is $\mathrm{O}_p\left(N^{-1/2}/h_p^2\right)$. Since $h/h_p = o_p(N^{-\varepsilon})$:

$$\sup\left|\hat{f}\left(w;\hat{\lambda}\right) - \hat{f}\left(w;\bar{\lambda}\right)\right| = o_p(N^{-1/2})$$

if $\sup_{\eta,w}\left|\frac{\partial}{\partial\lambda}\hat{f}\left(w;\bar{\lambda}\right)\right| = O_p\left(h^2\right)$. Bound $\left|\frac{\partial}{\partial\lambda}\hat{f}\left(w;\bar{\lambda}\right)\right|$ from above by:

$$\left|\frac{\partial}{\partial\lambda}\hat{f}\left(w;\bar{\lambda}\right) - E\left(\frac{\partial}{\partial\lambda}\hat{f}\left(w;\bar{\lambda}\right)\right)\right| + \left|E\left(\frac{\partial}{\partial\lambda}\hat{f}\left(w;\bar{\lambda}\right)\right)\right|$$

¿From Klein and Spady [1993, Lemma 1], the first component above is $\mathrm{O}_p\left(N^{-1/2}/h^2\right)$, which is $\mathrm{O}_p\left(h^2\right)$ under the window restrictions. Since density estimates converge to their expectations for a wide range of local windows, it can be shown that the second component is $\mathrm{O}_p\left(h^2\right)$.

Under the restrictions on pilot windows, h_p, and on the second-stage window, h, the convergence rate for $\mathbf{B}_k$ is faster than for $\mathbf{C}_k$, k = 0,1,2. The lemma now follows from the convergence rates for $\mathbf{C}_k$ as established above.

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). Under assumptions (A1-5) and definitions (D1-5), for η and w in compact sets:

$$\left|\nabla_\eta^k \hat{P}(w;\eta) - \nabla_\eta^k P(w)\right| = \begin{cases} o_p(h^3), & k=0 \\ o_p(h^2), & k=1 \\ O_p(h^2), & k=2 \end{cases}$$

Proof of Lemma 3. The proof immediately follows from Lemmas 2.

Below we exploit a property of semiparametric probability derivatives to obtain asymptotic normality for the estimator for the binary response model.

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

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

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). The extension to the double index case immediately follows from the same type of argument employed for the single index case. Namely, with F as the distribution function for $-u^*$, it can be shown that:

$$\begin{aligned} \nabla_{\eta_2} P(\eta)_{\eta\,=\,\eta_0} &= T_1 - E(T_1|W_{10},\ W_{20}), \\ T_1 &\equiv \nabla_{\eta_2} F(W_{10}/s(\eta_2))_{\eta\,=\,\eta_0}. \end{aligned}$$

which has conditional expectation of zero as claimed.

As a final set of intermediate lemmas, we require results that make it possible to deal with the estimated trimming indicator. Recall that this indicator is given as:

$$\hat{\tau}_i \equiv \left\{\hat{W}_1^2 + \hat{W}_2^2 \leq \hat{c}_w\right\},$$

where $\hat{c}_w$ is a quantile (e.g. .95) of $\|W\|$. With

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

$\hat{\eta}_{ip}$ contains pilot estimates of nuisance parameters and $\hat{c}$ estimates $c > 0$ and finite with

$$\begin{aligned} \hat{\eta}_{ip} &= \eta_{i0} + O_p\left(h^3\right) \\ \hat{c}_w &= c_w + O_p\left(h^3\right) \end{aligned}$$

The $\hat{\eta}_{ip}$ will be defined below by trimming on the basis of the norm of X rather than $\hat{W}$.[9]

Lemma 5: Estimated vs. Known Trimming. With $\hat{\tau}_i$ defined as above, let:

$$\tau_i \equiv \left\{W_1^2 + W_2^2 \leq c_w\right\}.$$

Then, assuming that $\left|\hat{\eta}_{1p} - \eta_{10}\right|, .\left|\hat{\eta}_{2p} - \eta_{20}\right|$, and $|\hat{c}_w - c_w|$ are each $\mathrm{O}_p\left(N^{-r}\right)$, for $\varepsilon > 0$:

$$N^{1/2}\sum_i |(\hat{\tau}_i - \tau_i)|\,|r_i|/N \leq N^{1/2}T\sup|r_i| + o_p\left(N^{-1/2}\right),\ T = O_p\left(N^{-(r-\varepsilon)}\right)$$

Proof of Lemma 5. The proof for this lemma is based on an inequality due to Jim Powell for bounding $|(\hat{\tau}_i - \tau_i)|$ from above by a "smoothed" indicator and is contained in Klein (1993, Lemmas 1-2, and the proof for Lemma 2).

8.2 Main Results

Throughout this section, all results are provided under Assumptions (A1-5) and Definitions (D1-4). In obtaining these results, from Lemmas 1 and

[9] We could simplify the theoretical arguments considerably by employing higher order kernels and trimming on the basis of X. However, we have obtained much better results in finite samples by employing local smoothing rather than higher order kernels. In showing that it is still possible control the bias in the gradient to this problem, it will be shown below that it is important to trim on the basis of the indices rather than the X's. As these indices are not known, we require preliminary estimates of them.

2, we may and do proceed with the covariance matrix Σ known.Recalling that Theorem 1 in Section 3 provides identification conditions for the binary response model, we begin with the proof of this theorem.

Proof of Theorem1(Identification) For any candidate index, W*:

$$Pr(Y_2 = 1|X) = P\left(Y_2 = 1 \mid W_1,\ W_2\right) = Pr(Y_2 = 1|W^*)$$

To establish identification, it suffices to show that for any candidate index, $W^* : W = W^*$ on a set of positive probability. Without loss of generality, first normalize every discrete variable so that each takes on the values 0,1 with positive probability (there may or may not be other support points).

Coefficients on Continuous Variables

We then begin by setting all discrete variables to zero and show that the coefficients for all remaining continuous variables are identified. Let $X \equiv [X_1, X_2, X_3]$ be a vector of continuous variables.[10] With the initial indicies given as $[I_1,\ I_2] \equiv X\Gamma, \Gamma \equiv [\Gamma_1'\ \Gamma_2']'$ and $\Gamma_1 : 2x2$, select $\mathrm{A} \equiv \Gamma_1^{-1}$ and define:

$$W \equiv X\Gamma A \equiv [W_1\ W_2] \equiv X \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \eta_{10} & \eta_{20} \end{bmatrix},$$

where $\eta_0 = \left[\eta_{10}\ , \eta_{20}\right]$ denotes the true values for the nuisance parameters. With $\eta^* = [\eta_1^*, \eta_{2*}^*]$ as any other matrix of nuisance parameter values, define $W^* \equiv [W_1^*\ W_2^*]$ by replacing η_0 with η^* above. Then, if W^* is a candidate satisfying the condition on probabilities above, from the implicit function theorem:[11]

$$a) : W_1 = H\left(W_1^*, W_2^*, W_2\right);\quad b) : W_2 = \bar{H}\left(W_1^*, W_2^*, W_2\right)$$

[10] We require that there be at least two continuous variables. Since there may not be any other continuous variables, we allow for this possibility below in discussing identification of the coefficients of discrete variables.

[11] With $\bar{A} \equiv A^{-1}$, rewrite the probability function to read as:

$$\begin{gathered} P\left(Y_2 = 1|W_1, W_2\right) \equiv G(W_1, W_2) \\ = P\left(Y_2 = 1|W\bar{A}\right) \\ = P\left(Y_2 = 1|W_1\bar{A}_{11} + W_2\bar{A}_{12},\ W_1\bar{A}_{21} + W_2\bar{A}_{22}\right) \\ = P(Y_2 = 1|I_1, I_2) = F\left(I_1/S\left(I_2\right)\right), \\ I_1 \equiv W_1\bar{A}_{11} + W_2\bar{A}_{12};\ I_2 \equiv W_1\bar{A}_{21} + W_2\bar{A}_{22} \end{gathered}$$

Differentiating (a) with respect to X_1, which is excluded from W_2 and W_2^* :

$$c) : 1 = \frac{\partial}{\partial X_1} H = \frac{\partial}{\partial W_1^*} H \Rightarrow W_1 = W_1^* + F\left(W_2^*, W_2\right)$$

Similarly, differentiating (b) with respect to X_2, which is excluded from W_1 and W_1^*:

$$d) : W_2 = W_2^* + \bar{F}\left(W_1^*, W_1\right)$$

To simplify a subsequent argument, we first show that $W_2 = W_2^* \Leftrightarrow W_1 = W_1^*$. To this end, if $W_2 = W_2^*$, from (c) above:

$$W_1 = W_1^* + F\left(W_2, W_2\right) \equiv W_1^* + F^*\left(W_2\right)$$

Differentiating with respect to X_2 :

$$\frac{\partial}{\partial X_2} F^*\left(W_2\right) = F'^* = 0 \Rightarrow F^*\left(W_2\right) = 0 \Rightarrow W_1 = W_1^*.$$

In an analogous argument based on differentiating (d) with respect to X_1, it can be shown that $W_1 = W_1^* \Rightarrow W_2 = W_2^*$.

Differentiating the distribution function F with respect to W_1:

$$\frac{\partial}{\partial W_1}\left[G(W_1, W_2) - G(W_1^*, W_2^*)\right] = \frac{\partial}{\partial W_1} G(W_1, W_2)$$

$$= F'\left[S A_{11} - I_1 S' A_{21}\right] / S^2$$

If this derivative is zero everywhere and $A_{21} \neq 0$, then I_1 is a function of I_2, contradicting the double index assumption. If the above derivative is zero everywhere and $A_{21} = 0$, then since $A_{11} \neq 0$ ($\bar{A}$ nonsingular), it would follow that $S = 0$ on a set of positive probability. This is a contradiction as we have assumed that S is positive everywhere. Therefore, it follows that the above derivative must be non-zero on a set of positive probability. Similarly, it follows that:

$$\frac{\partial}{\partial W_2}\left[G(W_1, W_2) - G(W_1^*, W_2^*)\right] \neq 0$$

on a set of positive probability.

Returning to the separable form in (c), let F_i denote the derivative of F with respect to its ith argument, i = 1,2. Then, differentiating (c) with respect to X_2 yields:

$$0 = F_1 + F_2 \Rightarrow F_1 = -F_2.$$

Differentiating (c) with respect to X_3 :

$$\begin{aligned} \eta_1 &= \eta_1^* + F_1\eta_2^* + F_2\eta_2 = \eta_1^* + F_2\left[\eta_2 - \eta_2^*\right] \\ &\Rightarrow \eta_1 - \eta_1^* = F_2\left[\eta_2 - \eta_2^*\right] \end{aligned}$$

If $\eta_2 \neq \eta_2^*$ (otherwise, from above, $\eta_1 = \eta_1^*$), it follows that $\eta_1 \neq \eta_1^*$ and we may write:

$$F_2 = \frac{\eta_1 - \eta_1^*}{\eta_2 - \eta_2^*} \equiv m \neq 0$$

Since $X_3\left(\eta_1 - \eta_1^*\right) = mX_3\left[\eta_2 - \eta_2^*\right]$:

$$W_1 - W_1^* = mW_2^* - mW_2 \Rightarrow W_1 + mW_2 = W_1^* + mW_2^*$$

For m $\neq$ 0, $[W_1^* + mW_2^*, W_1^*]$ and $[W_1 + mW_2, W_1]$ are obtained from 1-1 transformations of $[W_1^*, W_2^*]$ and $[W_1\ W_2]$ respectively. Therefore, with

$$\begin{aligned} K\left(W_1^* + mW_2^*, W_1^*\right) &\equiv \Pr\left(Y_2 = 1 | W_1^* + mW_2^*, W_1^*\right) = \\ L(W_1 + mW_2, W_1) &\equiv \Pr\left(Y_2 = 1 | W_1 + mW_2, W_1\right), \end{aligned}$$

$$K\left(W_1^* + mW_2^*, W_1^*\right) = K\left(W_1 + mW_2, W_1^*\right) = L(W_1 + mW_2, W_1)$$

Differentiating the second two equalities with respect to X_1 :

$$K_1 + K_2 = L_1 + L_2$$

Differentiating the second two equalities with respect to X_2 :

$$mK_1 = mL_1 \Rightarrow K_1 = L_1 \text{ for m} \neq 0 \Rightarrow K_2 = L_2$$

Differentiating with respect to X_3 :

$$K_1\left(\eta_1 + m\eta_2\right) + K_2\eta_1^* = L_1\left(\eta_1 + m\eta_2\right) + L_2\eta_1 \Rightarrow \eta_1^* = \eta_1 \Rightarrow \eta_2^* = \eta_2.$$

The result now immediately follows by contradiction.

Coefficients on Discrete Variables

To identify coefficients on discrete variables, set all of the discrete variables but one (X_d) to zero and define:

$$Z \equiv [Z_1, Z_2, Z_3] \equiv [X_1 + X_3\eta_{31},\ X_d,\ X_2 + \eta_{32}X_3]$$

Note that if X_1 and X_2 are the only continuous variables in the model (which is permissible), then $\eta_{31} = \eta_{32} = 0$. Write the indicies as:

$$W = Z \begin{bmatrix} 1 & 0 \\ \eta_{1d} & \eta_{2d} \\ 0 & 1 \end{bmatrix}$$

Define W^* analogously with η^*_{id} replacing η_{id}. Letting:

$$C(\eta_d) \equiv \begin{bmatrix} 1 & 0 \\ \eta_{1d} & \eta_{2d} \end{bmatrix}^{-1},\ WC = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ \lambda_1 & \lambda_2 \end{bmatrix}.$$

Then, with $\alpha \equiv 1/\lambda_2$ ($\lambda_2 \neq 0$ for nonsingular submatrices), define:

$$V \equiv WC \begin{bmatrix} 1 & 0 \\ 0 & \alpha \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \alpha \\ \lambda_1 & 1 \end{bmatrix}$$

Similarly, with $C^* \equiv C(\eta^*_d)$ and with α^* defined analogously to α :

$$V^* \equiv W^*C^* \begin{bmatrix} 1 & 0 \\ 0 & \alpha^* \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \alpha^* \\ \lambda^*_1 & 1 \end{bmatrix}$$

To establish that there exists an identifiable parameterization, it suffices to show that $V = V^*$.

As in the identification argument for coefficients on continuous variables, from the implicit function theorem:

$$\begin{aligned} V_1 &= V^*_1 + G(V_2, V^*_2) \\ &= G(Z_2\alpha + Z_3, Z_2\alpha^* + Z_3) \\ &= G(Z_3, Z_3) \text{ at } Z_2 = 0. \end{aligned}$$

As above, it can be shown that $V_2 = V_2^* \Leftrightarrow V_1 = V_1^*$. Therefore, $G(Z_3, Z_3) = 0$, which implies that $V_1 = V_1^*$. Since $V_1 = V_1^*$, it now follows that $V_2 = V_2^*$. The argument for the other discrete variables is identical, which completes the proof.

Theorem 2. Define the quasi-likelihood as in Section 4.1:

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

and define $\hat{\eta} \equiv \arg\sup \hat{L}(\eta)$. Then, under (A1-5) and (D1-4): $\hat{\eta} \xrightarrow{p} \eta_0$, the vector of true parameter values.

Proof of Theorem 2. From Lemma 3, $\hat{P}_i(\eta) \equiv \hat{P}(w_i; \eta)$ converges uniformly in w_i and η to $P_i(\eta) \equiv P(w_i; \eta)$ on compact sets. From Lemma 5, we may replace $\hat{\tau}_i$ with τ_i. Therefore, with $L(\eta)$ defined by replacing $\hat{P}_i(\eta)$ with $P_i(\eta)$ and $\hat{\tau}_i$ with τ_i.in $\hat{L}(\eta)$, it can then be shown that

$$\left| \hat{L}(\eta) - L(\eta) \right| \xrightarrow{p} 0, \text{ uniformly in } \eta.$$

Next, from standard uniform convergence arguments:

$$L(\eta) \xrightarrow{p} E(L(\eta)), \text{ uniformly in } \eta.$$

Employing Theorem 1 (Identification), from Section 3 it can be shown that $E(L(\eta))$ is uniquely maximized at η_0, which completes the argument.

Theorem 3. Defining $H_0 \equiv \nabla_\eta^2 E(L(\eta_0))$:

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

Proof of Theorem 3. With $\eta^+ \in [\hat{\eta}, \eta_0]$, from a standard Taylor series expansion:

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

Beginning with the Hessian component, from Lemmas 2-3:

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

¿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, the theorem will follow if the gradient term is asymptotically distributed as normal with zero expectation and covariance matrix $-H_0$.

Write the gradient component as:

$$\sqrt{N}\hat{G}(\eta_0) = \underset{\mathbf{A}}{N^{-1/2}\sum \hat{\tau}_i \left[Y_i - P_i\right]\hat{\rho}_i} \quad - \quad \underset{\mathbf{R}}{N^{-1/2}\sum \hat{\tau}_i \left[\hat{P}_i - P_i\right]\hat{\rho}_i},$$

where the estimated weight has the form:

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

In what follows, we will show that the remainder term, **R**, vanishes in probability and that **A** has the required normal form.

Before analyzing these terms, it is important discuss several aspects of trimming. If we trim so as to restrict X to a compact set, then as the argument below will show, term **A** has the form:

$$N^{-1/2}\sum \tau_i \left[Y_i - P_i\right]\rho + o_p(1).$$

However, the second term , **R**, does not vanish in probability. Instead, with the estimator, $\hat{\eta}_p$, obtained by trimming on the basis of X, can be shown that:

$$\left(\hat{\eta}_p - \eta_0\right) = O_p\left(h^3\right),\ h^3 > N^{-1/2},$$

which follows from the uniform rate at which **R** vanishes from Lemmas 2-3. With this preliminary estimator, it is now possible to construct estimated indices and to define, as given above in Lemma 4, a trimming indicator on the estimated indices. With such trimming, below we establish asymptotic normality at the usual $\sqrt{N}$ rate.

A

Beginning with the term **A,** write it as:

$$\begin{array}{ll}
\mathbf{A_1}: & N^{-1/2}\sum \tau_i \left[Y_i - P_i\right]\rho + \\
\mathbf{A_2}: & N^{-1/2}\sum \left(\hat{\tau}_i - \tau_i\right)\left[Y_i - P_i\right]\rho + \\
\mathbf{A_3}: & N^{-1/2}\sum \tau_i \left[Y_i - P_i\right]\left[\hat{\rho}_i - \rho_i\right]\left(\hat{s}_i/s_i\right) + \\
\mathbf{A_4}: & N^{-1/2}\sum \left(\hat{\tau}_i - \tau_i\right)\left[Y_i - P_i\right]\left[\hat{\rho}_i - \rho_i\right]\left(\hat{s}_i/s_i\right) + \\
\mathbf{A_5}: & N^{-1/2}\sum \tau_i \left[Y_i - P_i\right]\left[\hat{\rho}_i - \rho_i\right]\left[s_i - \hat{s}_i\right]/s_i. + \\
\mathbf{A_6}: & N^{-1/2}\sum \left(\hat{\tau}_i - \tau_i\right)\left[Y_i - P_i\right]\left[\hat{\rho}_i - \rho_i\right]\left[s_i - \hat{s}_i\right]/s_i.
\end{array}$$

We will show that all terms other than $\mathbf{A_1}$ vanish in probability. Beginning with $\mathbf{A}_2$, "linearize" the argument of the trimming function as in Klein (1993, p. 371, Footnote 30). Namely, with $\mathrm{C}_i(\eta) \equiv W_{1i}^2(\eta) + W_{2i}^2(\eta)$ and with $\left[\hat{\eta}_p - \eta_0\right] = O_p\left(N^{-r}\right)$

$$\hat{\tau}_i \equiv \left\{C_i(\eta_0) + \nabla^1 C_i(\eta_0)\left[\hat{\eta}_p - \eta_0\right] + O_p\left(N^{-2r}\right)\right\}$$

With $\hat{\tau}_i^* \equiv \left\{C_i(\eta_0) + \nabla^1 C_i(\eta_0)\left[\hat{\eta}_p - \eta_0\right] \leq \hat{c}_p\right\}$, from Lemma 5:

$$N^{-1/2}\sum \left(\hat{\tau}_i - \tau_i\right)\left[Y_i - P_i\right]\rho = N^{-1/2}\sum \left(\hat{\tau}_i^* - \tau_i\right)\left[Y_i - P_i\right]\rho + o_p(1).$$

¿From Pakes and Pollard (1989, Lemma 2.17), the second term above vanishes in probability.

¿From Lemmas2-3 and 5, the terms $\mathbf{A_4}$, $\mathbf{A_5}$, and $\mathbf{A_6}$ all vanish in probability. For the final term, $\mathbf{A_3}$, with $\varepsilon_i \equiv \tau_i\left[\hat{\rho}_i - \rho_i\right]\left(\hat{s}_i/s_i\right)$:

$$E\left(\mathbf{A_3^2}\right) < 4E\left(\sum \varepsilon_i^2/N\right) + E\sum\sum\left(\left[Y_i - P_i\right]\varepsilon_j\right)\left(\left[Y_j - P_j\right]\varepsilon_i\right)/N.$$

Since ε_i^2 is uniformly $o_p(1)$, the argument of the expectation in the first term converges to zero in probability. As ε_i^2 also satisfies a uniform integrability condition, the first term above converges to zero. For the second term, note that ε_j depends on Y_i but not on Y_j. Similarly, ε_i depends on Y_j but not on Y_i. Therefore,

$$E\sum\sum \varepsilon_j\left[Y_i - P_i\right]\left[Y_j - P_j\right]\varepsilon_i/N = \sum\sum\left[E\left(\left[Y_i - P_i\right]\varepsilon_j\right)\right]\left[E\left(\left[Y_j - P_j\right]\varepsilon_i\right)\right]/N$$

Next, note that ε_i is a sum of terms, only one of which depends on Y_j. From the definition ε_i, it can be shown that:

$$\varepsilon_i = \bar{\varepsilon}_i + \Delta_i,$$

where $\bar{\varepsilon}_i$does not depend on either Y_i or Y_j. Furthermore, from the dependence of ε_i on estimated density derivatives and the form of the kernel, it can be shown that $\Delta_i = O\left(N^{-1}h^3\right)$. A similar characterization holds for ε_j. Next, note that from iterated expectations:

$$E\left(\left[Y_i - P_i\right]\bar{\varepsilon}_j\right) = EE\left[\left(Y_i - P_i\right)\bar{\varepsilon}_j \mid X_i\right] = E\left[\left(E(Y_i|X_i) - P_i\right)\bar{\varepsilon}_j\right] = 0.$$

Consequently, with $r_{ij} \equiv:\left[Y_i - P_i\right]\varepsilon_j$:

$$\begin{aligned} \sum\sum E\left(r_{ij}\right)E\left(r_{ji}\right)/N &= \sum\sum E\left(\left[Y_i - P_i\right]\Delta_j\right)E\left(\left[Y_j - P_j\right]\Delta_i\right)/N = \\ O(N^2)O(N^{-1}h^3)O(N^{-1}h^3)O(N) &= O(N^{-1}h^6) = o(1). \end{aligned}$$

Since $\mathbf{A_2}$ converges to zero in mean-square, it converges in probability to zero.

R

¿From above, the normalized gradient has the form:

$$\sqrt{N}\hat{G}(\eta_0) = N^{-1/2}\sum\left[Y_i - P_i\right]\rho_i + \mathbf{R} + \mathbf{o_p(1)},\ \mathbf{R} = N^{-1/2}\sum\hat{\tau}_i\left[\hat{P}_i - P_i\right]\hat{\rho}_i$$

The theorem will now follow if the remainder term $\mathbf{R}$ converges to zero in probability. In evaluating this term, recall from the proof for Lemma 2, part **A:**

$$\left|\hat{f}\left(w;\hat{\lambda}\right) - \hat{f}\left(w;\bar{\lambda}\right)\right| = o_p\left(N^{-1/2}\right).$$

Accordingly, in what follows we evaluate all densities at $\bar{\lambda}$. With this simplification, write:

$$\begin{aligned} \mathbf{R} \equiv\ & N^{-1/2}\sum\tau_i\left[\hat{P}_i - P_i\right]\rho_i + N^{-1/2}\sum\left(\hat{\tau}_i - \tau_i\right)\left[\hat{P}_i - P_i\right]\rho_i \\ & N^{-1/2}\sum\tau_i\left[\hat{P}_i - P_i\right]\left[\hat{\rho}_i - \rho_i\right] + N^{-1/2}\sum\left(\hat{\tau}_i - \tau_i\right)\left[\hat{P}_i - P_i\right]\left[\hat{\rho}_i - \rho_i\right]. \end{aligned}$$

Employing an argument similar to that for $\mathbf{A_{22}}$ above, from Lemmas 2-3, each of the last three components of $\mathbf{R}$ is $o_p\left(1\right)$. Therefore,:

$$\mathbf{R} = \mathbf{R_1} + o_p(1),\ \mathbf{R_1} = N^{-1/2}\sum\tau_i\left[\hat{P}_i - P_i\right]\rho_i$$

The proof will now follow if the remainder term, $\mathbf{R_1}$, converges to zero in probability. With $\hat{P}_i \equiv \hat{f}_i/\hat{g}_i$

$$\begin{aligned}\mathbf{R_1} &\equiv N^{-1/2}\sum \tau_i\left[\left(\hat{f}_i/\hat{g}_i\right)-P_i\right]\rho_i = N^{-1/2}\sum \tau_i\left[\left(\hat{f}_i/\hat{g}_i\right)-P_i\right]\rho_i\left(\hat{g}_i/g_i\right)\\ &+N^{-1/2}\sum \tau_i\left[\hat{P}_i-P_i\right]\rho_i\left[1-\left(\hat{g}_i/g_i\right)\right]\end{aligned}$$

¿From Lemmas 2-3 , the second term converges to zero in probability as it is bounded in absolute value by:

$$N^{1/2}\sup_i\left[\tau_i\left|\hat{P}_i-P_i\right|\right]\sup_i\left[\tau_i\left|1-\left(\hat{g}_i/g_i\right)\right|\right]\sum\left|\tau_i\rho_i\right|/N = o_p(1).$$

Consequently, the remainder term $\mathbf{R} = \mathbf{R_2} + o_p(1)$,

$$\mathbf{R_2} = N^{-1/2}\sum \tau_i\left[\left(\hat{f}_i/\hat{g}_i\right)-P_i\right]\rho_i\left(\hat{g}_i/g_i\right) = N^{-1/2}\sum \tau_i\left[\hat{f}_i-P_i\hat{g}_i\right]\left(\rho_i/g_i\right)$$

As $\mathbf{R_2}$ is a linear combination of kernel density estimators, it is a U-statistic to which we can apply standard projection arguments. From (D3), write as:

$$\begin{aligned}\mathbf{R_2} &\equiv N^{-1/2}\sum\sum \tau_i\left[\frac{1}{N-1}\left(Y_{2j}k^1_{ij}\right)-P_i\left(Y_{2j}k^1_{ij}+\left(1-Y_{2j}\right)k^0_{ij}\right)\right]\left(\rho_i/g_i\right)\equiv\\ &\frac{N^{1/2}}{N\left(N-1\right)}\sum\sum\left[\varepsilon_{ij}\right]\left(\rho_i/g_i\right),\end{aligned}$$

where ε_{ij} is implicitly defined above and the double sum is taken over $(i,j) : i \neq j$. Employing a standard symmetrization argument, write:

$$\mathbf{R_2} = N^{1/2}\left[\begin{array}{c}N\\2\end{array}\right]^{-1}\sum\sum\left[\varepsilon_{ij}\left(\rho_i/g_i\right)+\varepsilon_{ji}\left(\rho_j/g_j\right)\right]/2,$$

where the double sum is now taken over observation $(i,j) : i < j$. With $\delta_{ij} \equiv \left[\varepsilon_{ij}\left(\rho_i/g_i\right)+\varepsilon_{ji}\left(\rho_j/g_j\right)\right]/2$, $\mathbf{R_2}$ is a U-statistic. Before employing standard projection arguments, it is important to note several properties of δ_{ij} and its components. First, from Lemma 4:

$$E\left(\rho_j|Y_{2i},W_i,W_j\right) = E\left(\rho_j|W_j\right) = 0. \qquad \text{(P1)}$$

Since (ε_{ji}/g_j) only depends on the random variables (Y_{2i}, W_i, W_j), from P1:

$$E\left(\varepsilon_{ji}\left(\rho_j/g_j\right) \mid Y_{2i},\ X_i\right) = E\left[E\left(\rho_j\left(\varepsilon_{ji}/g_j\right) \mid Y_{2i}, W_i,\ W_j\right)\right] = 0$$

Therefore, since $E\left(\varepsilon_{ji}\left(\rho_j/g_j\right)\right) = E\left(\varepsilon_{ij}\left(\rho_i/g_i\right)\right)$:

$$E\left(\delta_{ij}\right) = 0 \tag{P3}$$

It can also be shown that δ_{ij} satisfies:

$$E\left(\left\|\delta_{ij}\right\|^2\right) = o(N) \tag{P4}$$

Therefore from Serfling (1980) and Powell, Stock, and Stoker (1989), from a projection argument:

$$\mathbf{R_2} = N^{-1/2}\sum E\left(\left[\varepsilon_{ij}\left(\rho_i/g_i\right) + \varepsilon_{ji}\left(\rho_j/g_j\right)\right]/2 \mid Y_i,\ X_i\right) + o_p(1)$$

¿From (P1), the second component has zero expectation. For the first component, from Lemma 5:

$$E\left(\varepsilon_{ij}|Y_i, X_i\right) = E\left(\varepsilon_{ij}|W_i\right) = h^2 d_i,$$

where $|d_i| = O(1)$ uniformly. Therefore:

$$E\left(\varepsilon_{ij}\left(\rho_i/g_i\right)|Y_i, X_i\right) = E\left(\varepsilon_{ij}|W_i\right)\left(\rho_i/g_i\right) = h^2 d_i\left(\rho_i/g_i\right),$$

which has unconditional expectation of zero (P2). We have now established that:

$$\mathbf{R_2} = N^{-1/2}\sum h^2 d_i\left(\rho_i/g_i\right) + o_p(1).$$

As the h^2 term has expectation zero and variance $O(h^2)$, $\mathbf{R_2} = o_p(1)$, which completes the proof.

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

$$Y_1 = Z\theta_o + u,\ u \equiv s_1(X^*)u^*,$$

$\theta_o \equiv [\beta_o,\ \mu_o]$ 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 will now be immediate if the conditions given in the next lemma hold.

Lemma 6: With $Z^* \equiv [X,\ P(\eta_0)]$, under Assumptions (A1-4) and Definitions (D1-5):

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

Proof of Lemma 6. The first condition follows from Theorem 2 and Lemma 3. To establish the second condition, note that the non-zero component in condition (2) is given as:

$$\begin{aligned} \sqrt{N}\left[\hat{P}(\hat{\eta})' u - P(\eta_0)' u\right]/N &\equiv \mathbf{R_1^*} + \mathbf{R_2^*}\ , \\ \mathbf{R_1^*} &\equiv \sqrt{N}\left[\hat{P}(\eta_0)' u - P(\eta_0)' u\right]/N \\ \mathbf{R_2^*} &\equiv \sqrt{N}\left[\hat{P}(\hat{\eta})' u - \hat{P}(\eta_0)' u\right]/N. \end{aligned}$$

¿From a standard Taylor series argument, Lemma 3, and the $\sqrt{N}$ convergence of $\hat{\eta}$ in Theorem 3, $\mathbf{R_2^*}$ converges to zero in probability. Turning to $\mathbf{R_1^*}$, recall that in the Proof to Theorem 3, we considered a remainder term of the form:

$$\mathbf{R_1} = N^{-1/2} \sum \left[\hat{P}_i - P_i\right] \rho_i,$$

where ρ_i has expectation conditioned on the index of zero. In the Proof to Theorem 3, we showed that $\mathbf{R_1}$ was appropriately close in probability to a U-statistic that was $o_p(1)$. As the structure of $\mathbf{R_1^*}$ is identical to that for $\mathbf{R_1}$, the same argument (with u_i replacing ρ_i) shows that $\mathbf{R_1^*} = o_p(1)$.