Triangular Semiparametric Models Featuring Two Dependent Endogenous Binary Outcomes*

Roger Klein
Rutgers University

Chan Shen
Georgetown University

Francis Vella
Georgetown University

March 9, 2010

Abstract

This paper addresses the estimation of a class of models which features two endogenous and dependent binary outcomes. This class includes the triangular model with a binary outcome and a binary treatment, and several interesting variants on the sample selection model. The structure of the model imposes no distributional assumptions on the disturbances nor does it require that they enter additively. We formulate a quasi maximum likelihood estimator with semiparametric components that incorporate several bias adjustments. Under these adjustments, we establish desirable large sample properties using regular kernels in place of higher order kernels. Simulation evidence confirms that this estimator performs well in finite samples.

*Address correspondence to Chan Shen, Department of Economics, Georgetown University, ICC 580, 37th and O Sts., NW, Washington, DC 20057, USA; e-mail: cs589@georgetown.edu.

1 Introduction

This paper analyses the estimation of a class of semiparametric index models which feature two endogenous binary outcomes. This class incorporates a large range of models that are important for empirical work including binary treatment models with non-additive errors where the outcome of interest is binary. It covers the selection model with non-additive errors where the selection process is captured by an indicator function and the outcome of interest for the selected sample is binary. It also includes models where the outcome of interest for the selected sample is measured by a continuous outcome but the selection process is a function of two binary rules. Earlier papers in the semi parametric literature have focussed on other variants of the binary choice model. For example, Blundell and Powell (2004) and Rothe (2009) have developed estimators for semiparametric binary response models that depend on a continuous endogenous variable, as opposed to an endogenous binary variable as considered here. Hoderlein (2009) formulates an estimator for binary response models when the coefficients are random. However, for the class of index models with joint binary outcomes considered here, to the best of our knowledge, it has not been previously estimated in a semiparametric index framework.

To establish large sample properties for index models, it is necessary to control for the bias in the estimator. There is a literature on multiple and single index models that controls for the bias by selecting a kernel function that is not regular (see, for example, Ichimura and Lee (1991), Klein and Spady (1993), Lee (1995) and Klein and Vella (2009)). However, while it is well known that regular kernels generally perform better than higher order kernels in finite samples, they do not have desirable theoretical properties. For a variant of Semiparametric Least-Squares, Klein and Shen (2009)

provide a bias reducing mechanism that makes it possible to employ regular kernels under a single index assumption. In this paper we relax the single index assumption and propose an estimator under the quasi-likelihood framework. Employing the appropriate bias adjustments, we show that the estimator based on regular kernels has both desirable theoretical properties and finite sample performance.[1] We note that these bias reduction mechanisms can be extended to index models other than the particular class considered here. We also note that in these index models, estimates of marginal effects can be obtained from these index parameters. When a binary response model is fully observed and does depend on a binary exogenous treatment variable, the theory for these marginal effects is immediate. Here we consider joint binary models where either one of the explanatory variables is an endogenous binary treatment or one binary relation is subject to sample selection. In these cases, the theory for estimating marginal effects is substantially different from that for index parameters and is beyond the scope of the present paper.[2]

The following section outlines the general model and highlights some special cases. It also briefly describes the estimation procedure. Sections 3 and 4 provide the assumptions and the details of the estimator. Section 5 provides simulation evidence, and concluding comments are offered in section 6.

[1] There are other alternative methods that control for the bias under regular kernels. For example, Powell and Honore (2005) employ a jackknife approach where the final estimator is a linear combination of estimators using different windows.

[2] In the context of an index model, we address these issues in Klein, Shen, and Vella (2009a-b).

2 Models

The models considered here all contain the following underlying component:

$$Y_{1i} = I\{g(Y_{2i}, X_i\beta_o,\ \epsilon_i) > 0\} \quad (1)$$

$$Y_{2i} = I\{h(Z_i\pi_o,\ u_i) > 0\} \quad (2)$$

where the $Y's$ are the endogenous binary variables generated via the indicator function $I\{.\}$; X and Z are vectors of exogenous variables; ϵ_i and u_i are error terms with a non-zero correlation; $g(.)$ and $h(.)$ are unknown functions; and the β_o and π_o are unknown parameter values.

Notice that linear combinations of exogenous variables, $X_i\beta_o$ and $Z_i\pi_o$, enter each equation. We refer to these linear combinations as indices and assume in (A 3) that probabilities of interest only depend on X and Z through these indices. . We impose this index structure, as opposed to a non-parametric one, to improve the performance of the estimator. As is well known in the literature, the indices are identified up to location and scale. Namely, the $\theta's$ are identified in the following normalized indices:

$$X_i\beta_o = b_{1o}(X_{1i} + X_{2i}\theta_{1o}) + c_{1o}$$

$$Z_i\pi_o = b_{2o}(Z_{1i} + Z_{2i}\theta_{2o}) + c_{2o}$$

where X_{1i} is a continuous variable which belongs to the model, X_{2i} is the vector of all other X-variables. The Z-variables are defined similarly. Note that identifying $\theta's$ will generally make it possible to identify the probabilities and marginal effects of

interest.

The model can be characterized as a triangular system with a binary outcome and a binary endogenous explanatory variable. We allow the index in the reduced form equation, (2), to interact in an unspecified way with the disturbance, while in the main equation, (1), the index freely interacts with both the disturbance and the endogenous explanatory variable. The model can be viewed as a basic component of a number of different models, several of which we discuss below. The models below differ along two dimensions. First, there are endogenous treatment and selection versions. Second, the models differ according to whether or not they contain an additional equation for a continuous outcome.

2.1 Binary Outcomes with Binary Selection Rule

The first model is a semiparametric variant on the Heckman (1974, 1979) selection model where the outcome of interest is binary. More explicitly:

$$Y_{1i} = I\left\{g(V_{1i}, \epsilon_i) > 0\right\} * Y_{2i},\ V_{1i} \equiv V_{1i}\left(\theta_{1o}\right) \equiv X_{1i} + X_{2i}\theta_{1o} \tag{3}$$

$$Y_{2i} = I\left\{h(V_{2i}, u_i) > 0\right\},\ V_{2i} \equiv V_{2i}\left(\theta_{2o}\right) \equiv Z_{1i} + Z_{2i}\theta_{2o} \tag{4}$$

where Y_{1i} is only observed for the subsample for which $Y_{2i} = 1$. When the model is additive, and the joint distribution of the errors is parametrically known, it can be estimated by maximum likelihood (see e.g., Poirier (1980) and Vella (1998)). However, in the present binary context, with neither separability nor known error distributions, the existing available estimators do not apply. We propose a suitable estimator.

2.2 Binary Outcomes with Binary Endogenous Treatment

The second model has an endogenous binary treatment variable without sample selection. Namely:

$$Y_{1i} = I\left\{g(Y_{2i}, V_{1i},\ \epsilon_i) > 0\right\} \tag{5}$$

$$Y_{2i} = I\left\{h(V_{2i},\ u_i) > 0\right\}. \tag{6}$$

If the g and h functions in both equations are additively separable (i.e. threshold-crossing models) and the errors are jointly normal, then maximum likelihood may be employed to estimate the parameters. We provide an estimator when these restrictions do not hold.

Several important extensions of the above models add a continuous outcome equation. With the binary indicator Y_2 not appearing in the Y_1-model, consider the following example of multiple selection:

$$Y_{3i} = (W_i\alpha_o + c_o + e_i) * I\{Y_{1i} = 1, Y_{2i} = 1\}$$

noting that the manner in which the two indicators interact to determine the observability of Y_{3i} determines the applicability of the available procedures. Das et. al. (2003) focus on the above continuous outcome equation and assume that the sample selection correction is a function of the propensity scores from the joint binary model shown above. When the joint binary model is fully observed or does not contain an endogenous variable, this assumption holds.

De Luca and Peracchi (2010) employ the procedure of Gallant and Nychka (1987)

and provide a semiparametric estimator under a threshold crossing structure. Their approach estimates the indices for the selection equations and then uses a Robinson (1988) differencing approach to account for the selection in the Y_{3i} equation. Yavuzoglua and Tunali (2009) consider a similar structure to that of De Luca and Peracchi but impose normality in the selection equations. Normality is relaxed in the Y_{3i} equation by including control functions based on appropriate expansions. We impose neither distributional or threshold crossing assumptions here and note that a Robinson (1988) style can be constructed to estimate the Y_{3i} equation when one has estimates of the indices in the selection equations.

Another extension has Y_2 appearing in the Y_1-model and considers the case where the continuous outcome is subject to sample selection and contains an endogenous binary treatment indicator:

$$Y_{3i} = (W_i\alpha_o + c_o + Y_{2i}\gamma_o + e_i) * I(Y_{1i} = 1)$$

Shen (2009) estimates such a model, where Y_{2i} is a binary insurance decision, Y_{1i} is a binary healthcare utilization decision, and Y_{3i} is a continuous healthcare expenditure variable that is positive for individuals that access healthcare and depends on the insurance decision. In that paper, there is a difficult problem in estimating the marginal treatment effect. Part of the theory for it depends on a $\sqrt{N}$-consistent estimator for the double binary component which we develop here.

3 Estimation

We now develop estimators for both the binary selection case and the binary treatment case. To accommodate both we introduce the following notation: For $\{d_1, d_2\} = \{0, 1\}$, in the binary treatment model, let:

$$Y_i(d_1, d_2) = I\{Y_{1i} = d_1, Y_{2i} = d_2\}.$$

In the binary selection case, let:

$$Y_i(d_1, d_2) = \begin{cases} I\{Y_{1i} = d_1, Y_{2i} = d_2\} & \text{for } d_2 = 1 \\ I\{Y_{2i} = d_2\} & \text{for } d_2 = 0 \end{cases}.$$

Finally, let:

$$\hat{P}_i(d_1, d_2; \theta) \equiv \hat{P}\left(Y_i(d_1, d_2) = 1 | V_i(\theta) = v_i(\theta)\right)$$

where $V_i(\theta) = (V_{1i}(\theta), V_{2i}(\theta))$. The parameter estimates are given by maximizing a quasi or estimated likelihood:

$$\begin{aligned} \hat{\theta} &\equiv \arg\max_{\theta} \hat{L}(\theta), \\ \hat{L}(\theta) &\equiv \sum_{i=1}^{N} \tau_i \sum_{d_1, d_2} Y_i(d_1, d_2) Ln\left(\hat{P}_i(d_1, d_2; \theta)\right). \end{aligned}$$

where τ_i is a trimming function defined below to control for small density denominators.

When Y_{1i} and Y_{2i} are both observed, there are four possible outcomes corresponding to different combinations of $\{d_1, d_2\}$. For the binary model for which Y_{1i}

is observed only if $Y_{2i} = 1$, there are only three possible outcomes because of partial observability.

In maximizing the above likelihood, the properties of the estimates depend on how the probabilities are estimated. If they are based on appropriately chosen higher order or bias-reducing kernels, they have desirable large sample properties, but often do not perform well in finite samples. On the other hand, if they are based on regular kernels without employing any bias reduction mechanisms, the estimates frequently have good finite sample properties but are not asymptotically distributed as normal at a $\sqrt{N}-$rate. Our objective is to avoid this trade-off by providing an estimator that performs well in finite sample while retaining desirable large sample properties. To do so we introduce several bias control mechanisms other than higher-order kernels.

To motivate these mechanisms, we show below that the gradient to the quasi-likelihood is a product of terms, one of which is the derivative of the probability function, $\nabla_\theta \hat{P}_i(d_1, d_2; \theta_0)$, noting that θ_0 denotes the true value. From Theorem 0 below, which is due to Whitney Newey:

$$E\left(\nabla_\theta P_i(d_1, d_2; \theta_0) \mid V_i\left(\theta_0\right)\right) = 0.$$

In an iterated expectations argument, we show that if the trimming function only depended on the index and if the probability derivative could be taken as known, the gradient would have expectation 0.[3] We now need to solve three problems. First, the trimming function must depend on the estimated indices. To this end, in (D8) below, we define a two stage estimation procedure, where the estimated indices are

[3] As shown below, in an iterated expectations argument, conditioning first on X, the indicator has conditional expectation that only depends on indices. The result then follows.

recovered in the first stage. Second, we need to prove that we can take the estimated probability derivative function as known. Employing the adjustment in (D9) below, we are able to resolve this problem. Third, index trimming poses a problem for the consistency argument. We discuss the nature of this problem below and show that it is resolved by employing the adjusted semiparametric probabilities in (D6). With these bias controls, we are able to obtain asymptotic results using regular kernels.

4 Assumptions and Definitions

We now provide the assumptions and definitions that we employ to establish the asymptotic properties for the estimators of the index parameters in the double binary component.

A1. The Data. In the fully observed case, (Y_{1i}, Y_{2i}, S_i), $i = 1, ..., N$, are i.i.d. observations from the model in (1)-(2). With S as the NXK matrix of observations on the explanatory variables and with $\underline{\mathbf{1}}$ as an $NX1$ column vector of ones, the columns of $[S \ \ \underline{\mathbf{1}}]$ are linearly independent with probability 1. In the case of partial observability, Y_{1i} is only observed when $Y_{2i} = 1$, and S_i may or may not be partially observed.

A2. Parameter Space. The vector of true parameter values $\theta_o = (\beta_o, \pi_o)$ for the model in (1)-(2) lies in the interior of a compact parameter space, Θ.

A3. Model. Define the indices for the reduced form and primary equations as V_2 and V_1 respectively, and assume each contains a continuous exogenous variable. Further, V_2 contains at least one continuous variable, which is excluded from

V_1. With $d_k = 0, 1$, assume:

$$\begin{aligned} \Pr\left(Y_{1i} = d_1, Y_{2i} = d_2 | X_i\right) &= \Pr\left(Y_{1i} = d_1, Y_{2i} = d_2 | V_{1i}, V_{2i}\right) \\ \Pr\left(Y_{2i} = d_2 | X_i\right) &= \Pr\left(Y_{2i} = d_2 | V_{1i}, V_{2i}\right). \end{aligned}$$

A4. Densities. Let $g(v_1, v_2|Y_1, Y_2)$ be the indicated conditional density for the indices. Let $\nabla^p g$ be any of the partials or cross partials of g up to order p, with $\nabla^0 g = g$. Assume that $g > 0$ on all fixed compact subsets of the support for the indices. Further, assume that $\nabla^p g$, $\frac{\partial}{\partial\theta}\left(\nabla^p g\right)$, and $\frac{\partial^2}{\partial\theta\partial\theta}\left(\nabla^p g\right)$ are bounded for $p = 0, 1, 2$.

Assumptions (A1) and (A2) are standard. Assumption (A3) imposes a double index structure on the Y_1-model and a single index structure on the Y_2-model. These index assumptions are automatically satisfied when the errors are independent of the index variables. Finally, assumption (A4) provides required smoothness conditions for determining the order of the bias for density estimators. In addition to the above assumptions, we also need a number of definitions for densities, probability functions and estimators.

D1. Unadjusted Densities. Term $K(\cdot)$ as a regular kernel if it is a density symmetric about zero. For $d_2 = 0, 1$, define:

$$\hat{f}_2\left(t_2; d_2\right) \equiv \sum_{j=1}^{N} \frac{Y_{2j}^{d_2}(1 - Y_{2j})^{1-d_2}}{N h_m} K\left[\frac{t_2 - V_{2j}}{h_m}\right],$$

where with σ_2 as the standard deviation for V_2, the window parameter is given as: $h_m \equiv \sigma_2 N^{-r_m}$, $r_m = \frac{1}{6+\xi}$. For regular kernels K_1 and K_2, and with

$d_1, d_2 = 0, 1$, define:

$$\hat{f}\left(t; d_1, d_2\right) \equiv \sum_{j=1}^{N} \frac{Y_{1j}^{d_1}(1-Y_{1j})^{1-d_1} Y_{2j}^{d_2}(1-Y_{2j})^{1-d_2}}{N h_{c1} h_{c2}} K_1(\frac{t_1 - V_{1j}}{h_{c1}}) K_2(\frac{t_2 - V_{2j}}{h_{c2}}),$$

where with σ_k as the standard deviation for V_k, $k = 1, 2$, the window parameters are given as: $h_{c1} \equiv \sigma_1 \; h_c, h_{c2} \equiv \sigma_2 \; h_c, h_c \equiv \; N^{-r_c}, r_c = \frac{1}{8+\xi}$.When the conditioning value t_k, is replaced by the observation V_{ik}, then the above averages are taken over the $(N-1)$ observations for which $j \neq i$.[4]

D2. Unadjusted Probabilities. Let:

$$\begin{aligned} \hat{P}\left(Y_{2i} = d_2 | V_{2i} = t_2\right) &\equiv \hat{f}_2\left(t_2; d_2\right) / \sum_{d_2=0}^{1} \hat{f}_2\left(t_2; d_2\right) \\ \hat{P}\left(Y_{1i} = d_1 | Y_{2i} = d_2, V_i = t\right) &\equiv \hat{f}\left(t; d_1, d_2\right) / \sum_{d_1=0}^{1} \hat{f}\left(t; d_1, d_2\right). \end{aligned}$$

With $d_2 = 1$ for the binary selection model and $d_2 = 0, 1$ for the binary treatment model, define:

$$\hat{P}\left(Y_{1i} = d_1, Y_{2i} = d_2 | V_i = t\right) = \hat{P}\left(Y_{2i} = d_2 | V_{2i} = t_2\right) \hat{P}\left(Y_{1i} = d_1 | Y_{2i} = d_2, V_i = t\right).$$

D3. Smooth Trimming. Define a smooth trimming function as:

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

[4]It can easily be shown that all estimators with windows depending on population standard deviations are asymptotically the same as those based on sample standard deviations. For notational simplicity, we employ population standard deviations throughout.

D4. Interior Index Trimming. Let $\hat{V}_k^U$ and $\hat{V}_k^L$ be the upper and lower sample index quantiles for the indices: $V_k \equiv V_k(\theta)$, $k = 1, 2$; and let V_k^U and V_k^L be the corresponding population quantiles. Then, define smooth interior trimming functions as:

$$\begin{aligned} \hat{\tau}_I(t_k) &\equiv \tau\left(\hat{V}_k^L, t_k\right)\tau\left(t_k, \hat{V}_k^U\right) \\ \tau_I(t_k) &\equiv \tau\left(V_k^L, t_k\right)\tau\left(t_k, V_k^U\right). \end{aligned}$$

D5. Density Adjustment. Let $\hat{q}_2$ be a lower sample quantile for $\hat{f}_2(V_2; d_2)$, and $\hat{q}$ be a lower sample quantile for $\hat{f}(V; d_1, d_2)$, and let q_2 and q be the corresponding population quantiles. Then, define estimated adjusted densities as:

$$\hat{f}_2^*(t_2; d_2) = \hat{f}_2(t_2; d_2) + \hat{\Delta}_2(d_2),\ \hat{\Delta}_2(d_2) \equiv a_{_{2N}}\left[1 - \hat{\tau}_I(t_2)\right]\hat{q}_2$$

$$\hat{f}^*(t; d_1, d_2) = \hat{f}(t; d_1, d_2) + \hat{\Delta}(d_1, d_2),\ \hat{\Delta}(d_1, d_2) \equiv a_{_N}\left[1 - \hat{\tau}_I(t_1)\,\hat{\tau}_I(t_2)\right]\hat{q}.$$

With f_2 and f as the probability limits of $\hat{f}_2$ and $\hat{f}$, define the adjusted densities:

$$f_2^*(t_2; d_2) = f_2(t_2; d_2) + \Delta_2(d_2),\ \Delta_2(d_2) \equiv a_{_{2N}}\left[1 - \tau_I(t_2)\right]q_2$$

$$f^*(t; d_1, d_2) = f(t; d_1, d_2) + \Delta(d_1, d_2),\ \Delta(d_1, d_2) \equiv a_{_N}\left[1 - \tau_I(t_1)\,\tau_I(t_2)\right]q.$$

Referring to the window parameters in (D1), $a_{_N} \equiv N^{-r/2}$ and $a_{_{2N}} \equiv N^{-r_2/2}$.

D6. Adjusted Semiparametric Probability Functions. Let:

$$\begin{aligned}
\hat{P}^*\left(Y_{2i}=d_2|V_{2i}=t_2\right) &\equiv \hat{f}_2^*\left(t_2;d_2\right)/\sum_{d_2=0}^{1}\hat{f}_2^*\left(t_2;d_2\right)\\
\hat{P}^*\left(Y_{1i}=d_1|Y_{2i}=d_2,V_i=t\right) &\equiv \hat{f}^*\left(t;d_1,d_2\right)/\sum_{d_1=0}^{1}\hat{f}^*\left(t;d_1,d_2\right)\\
P^*\left(Y_{2i}=d_2|V_{2i}=t_2\right) &\equiv f_2^*\left(t_2;d_2\right)/\sum_{d_2=0}^{1}f_2^*\left(t_2;d_2\right)\\
P^*\left(Y_{1i}=d_1|Y_{2i}=d_2,V_i=t\right) &\equiv f^*\left(t;d_1,d_2\right)/\sum_{d_1=0}^{1}f^*\left(t;d_1,d_2\right).
\end{aligned}$$

Then, as in (D2), with $d_2=1$ for the binary selection model and $d_2=0,1$, in the case of binary treatment, define adjusted probabilities:

$$\begin{aligned}
\hat{P}^*\left(Y_{1i}=d_1,Y_{2i}=d_2|V_i=t\right) &= \hat{P}^*\left(Y_{2i}=d_2|V_{2i}=t_2\right)\hat{P}^*\left(Y_{1i}=d_1|Y_{2i}=d_2,V_i=t\right)\\
P^*\left(Y_{1i}=d_1,Y_{2i}=d_2|V_i=t\right) &= P^*\left(Y_{2i}=d_2|V_{2i}=t_2\right)P^*\left(Y_{1i}=d_1|Y_{2i}=d_2,V_i=t\right).
\end{aligned}$$

D7. Likelihood Trimming. Define τ_{ix} as an indicator that is one if all of the continuous $X\prime s$ are between their respective lower and upper sample quantiles, and define τ_{iv} as an indicator that is one if the estimated index vector $V\left(\hat{\theta}\right)$ is between lower and upper sample quantiles. Here, $\hat{\theta}$ is a consistent estimator for θ_o that is defined below.

D8. First and Second Stage Estimators. To define estimators for both selection and treatment models we use the definitions stated above and define the first

stage estimator as:

$$\begin{aligned}\hat{\theta} &\equiv \arg\max_{\theta} \hat{L}(\theta), \\ \hat{L}(\theta) &\equiv \sum_{i=1}^{N} \tau_{ix} \sum_{d_1,d_2} Y_i(d_1,d_2) Ln\left(\hat{P}_i(d_1,d_2;\theta)\right)\end{aligned}$$

recalling that τ_{iv} is a trimming function based on the estimated index vector, $V\left(\hat{\theta}\right)$, defined in (D7). In the objective function above, replace $\hat{P}$ with $\hat{P}^*$ defined as in (D6), replace $\hat{\tau}_X$ with $\hat{\tau}_V$, and term the new objective function as $\hat{L}^*(\theta)$. Then, define the second stage estimator:

$$\hat{\theta}^* \equiv \arg\max_{\theta} \hat{L}^*(\theta).$$

D9. The Adjusted Estimator. Letting

$$\begin{aligned}\hat{P}_i^*(d_1,d_2;\theta) &\equiv \hat{P}^*\left(Y_i(d_1,d_2)=1 | V_i(\theta)=v_i(\theta)\right) \\ \hat{\delta}_i^*(d_1,d_2;\theta) &\equiv \nabla_\theta \hat{P}_i^*(d_1,d_2;\theta)/\hat{P}_i^*(d_1,d_2;\theta),\end{aligned}$$

define the bias component of the gradient to $\hat{L}^*(\theta)$ as:

$$\hat{B}^*\left(\hat{\theta}^*\right) \equiv -\sum_{i=1}^{N} \tau_{iv}\left(\hat{\theta}^*\right) \sum_{d_1,d_2} \left[\hat{P}_i^*\left(d_1,d_2;\hat{\theta}^*\right) - P_i\left(d_1,d_2;\hat{\theta}^*\right)\right] \hat{\delta}_i^*\left(d_1,d_2;\hat{\theta}^*\right).$$

Define $\hat{P}^o(d_1,d_2;\theta)$ as an estimated semiparametric probability function where the components are based on optimal window parameters: $r = r^o = 1/6$ and

$r_2 = r_2^o = 1/5$. Define:

$$\hat{B}^o\left(\hat{\theta}^*\right) \equiv -\sum_{i=1}^{N} \tau_{iv}\left(\hat{\theta}^*\right) \sum_{d_1,d_2} \left[\hat{P}_i^o\left(d_1, d_2; \hat{\theta}^*\right) - P_i\left(d_1, d_2; \hat{\theta}^*\right)\right] \hat{\delta}_i^*\left(d_1, d_2; \hat{\theta}^*\right).$$

Then, define a gradient correction as:

$$\hat{C}\left(\hat{\theta}^*\right) \equiv \hat{B}^o\left(\hat{\theta}^*\right) - \hat{B}^*\left(\hat{\theta}^*\right).$$

With $\hat{H}\left(\hat{\theta}^*\right)$ as the estimated hessian, the adjusted estimator is defined as:

$$\hat{\theta}^o \equiv \hat{\theta}^* - \hat{H}\left(\hat{\theta}^*\right)^{-1} \hat{C}\left(\hat{\theta}^*\right).$$

As stated earlier, the proofs exploit a residual-like property of the derivative (with respect to the parameters) of the true semiparametric probability function. Namely, this derivative has conditional expectation of zero when evaluated at the true parameter values. By using this property, which we will define and prove below, we can further control for the bias in the gradient to the objective function, which is essential in establishing asymptotic normality. In so doing, we will not be able to trim on the basis of X and instead must trim on the basis of estimated indices. It is for this reason that we define the two stage estimator in (D8). However, the index trimming is problematic in the consistency argument where density denominators can tend to zero when evaluated away from the truth. Therefore, we employ the $\hat{\Delta}$ adjustment factors in (D5) to keep the density denominators away from zero. By employing these adjustments, together with the bias reducing adjustment in (D9), we are able to establish asymptotic results under regular kernels.

5 Asymptotic Results

We now outline the proof strategy for the various estimator stages in which bias reducing devices are employed in conjunction with regular kernels. The first is based on a result due to Whitney Newey and is given in the following theorem:

Theorem 0: With $V(\theta_0) \equiv V(X;\theta_0)$ as the vector of indices, assume the following index restriction holds:

$$P_i(d_1, d_2; \theta_0) = P(Y_i(d_1, d_2) = 1 | V(\theta_0)) \equiv F(V(\theta_0)).$$

Then:

$$E\{\nabla_\theta P_i(d_1, d_2; \theta_0)\} = 0.$$

Proof: Let $\delta(\theta) \equiv V(\theta_0) - V(\theta)$ and observe that $\delta(\theta_0) = 0$ and that $\nabla_\theta \delta(\theta) = -\nabla_\theta V(\theta)$. Then, employing the index restriction and using iterated expectations:

$$\begin{aligned} P(Y_{1i} = d_1, Y_{2i} = d_2 | V(\theta)) &= E_X[P_i(d_1, d_2; \theta_0) \mid V(\theta)] \\ &\equiv E_X[F[V(\theta_0)] \mid V(\theta)] \\ &\equiv E_X[F[V(\theta) + \delta(\theta)] \mid V(\theta)] \\ &\equiv G(V(\theta), \delta(\theta)). \end{aligned}$$

Let G_k be the partial derivative of G taken w.r.t. θ in the k^{th} argument of G, $k =$

1,2. From the chain rule:

$$\begin{aligned}\nabla_\theta G\left(V\left(\theta\right),\delta\left(\theta\right)\right)|_{\theta=\theta_0} &= G_1\left(V\left(\theta\right),0\right)|_{\theta=\theta_0}+G_2\left(V\left(\theta_0\right),\delta\left(\theta\right)\right)|_{\theta=\theta_0}\\ &= \nabla_\theta F\left(\ V\left(\theta\right)\right)|_{\theta=\theta_0}-E\left[\nabla_\theta F\left(\ V\left(\theta\right)\right)\ \mid V\left(\theta_0\right)\right]_{\theta=\theta_0}.\end{aligned}$$

The theorem now follows.

To take advantage of this result, let

$$\hat{\delta}\left(d_1,d_2;\theta\right)\equiv\frac{\nabla_\theta\hat{P}_i\left(d_1,d_2;\theta\right)}{\hat{P}_i\left(d_1,d_2;\theta\right)}.$$

Then, since indicators and probabilities sum to one over all possible cells, the gradient to the objective function has the form:

$$\begin{aligned}\hat{G} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\frac{Y_i\left(d_1,d_2\right)}{\hat{P}_i\left(d_1,d_2;\theta_o\right)}\right]\nabla_\theta\hat{P}_i\left(d_1,d_2;\theta_o\right)\tau_{ix} \qquad (7)\\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\frac{Y_i\left(d_1,d_2\right)}{\hat{P}_i\left(d_1,d_2;\theta_o\right)}-1\right]\nabla_\theta\hat{P}_i\left(d_1,d_2;\theta_o\right)\tau_{ix}\\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-\hat{P}_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_o\right)\tau_{ix},\end{aligned}$$

where the second line follows because:

$$\sum_{d_1,d_2}\hat{P}_i\left(d_1,d_2;\theta_o\right)=1\Rightarrow\sum_{d_1,d_2}\nabla_\theta\hat{P}_i\left(d_1,d_2;\theta_o\right)=0.$$

With δ as the probability limit of $\hat{\delta}$, from the above theorem, $E\left(\delta|V\right)=0$. Therefore, this multiplicative gradient component can serve as a source of bias reduction

while still employing regular kernels. To exploit this residual-like property of the probability gradient, we need to resolve the three problems discussed earlier. First, even if we could take δ as known, trimming on the basis of X poses a problem. In an iterated expectations argument, conditioning on X,

$$E\left[Y_i\left(d_1, d_2\right) - \hat{P}_i\left(d_1, d_2; \theta_o\right) | X\right] = H(V),$$

a function of index values. If the trimming function were not present or if it depended on the index, the gradient would now have zero expectation. Our solution is to design a two-stage estimator where parameter estimates from the first stage are used to construct the index and then index trimming is employed in the second stage.

Second, while index trimming makes a bias reduction argument possible for the gradient, it poses a problem for the consistency argument. In particular, index trimming provides no protection for small denominators and hence makes uniform convergence difficult to establish. To resolve this problem, we use the adjusted probabilities in (D5, D6) so that denominators are kept away from zero, while the estimated probability still goes rapidly to the truth in gradient expression. In this manner, we are able to establish consistency without X-trimming while at the same time taking advantage of index trimming at the gradient level where bias reduction is important.

Third, it would seem desirable, if possible, to speed up the rate at which estimated probabilities converge to the truth. Indeed, it turns out that faster convergence is critical to the normality argument. The difficulty here, which is typically encountered in estimating semiparametric models using optimization methods, is that a window choice is made before any optimization. This same window must then be employed to show that estimated densities and their derivatives converge appropriately to the

corresponding true values. To accomplish all of these objectives, estimated probabilities are based on a suboptimal window choice. To solve this problem, in (D9), after we obtain the second stage estimator under index trimming, we adjust the resulting estimator. In an argument sketched out below and proved in the Appendix, we show that the adjusted estimator behaves like one with different optimal windows for different components of the problem.

To sketch out the intuition for this adjustment, recalling (D9) we first consider the infeasible estimator:

$$\hat{\theta}_{in}^{o} = \hat{\theta}^{*} - \hat{H}^{*}\left(\theta^{+}\right)^{-1}\hat{C}\left(\theta_{o}\right).$$

With the gradient for $\hat{\theta}^{*}$ given as:

$$\hat{G}^{*}\left(\theta_{0}\right) \equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_{1},d_{2}}\left[Y_{i}\left(d_{1},d_{2}\right) - \hat{P}_{i}^{*}\left(d_{1},d_{2};\theta_{o}\right)\right]\hat{\delta}_{i}^{*}\left(d_{1},d_{2};\theta_{o}\right)\tau_{iv}, \tag{8}$$

the second stage estimator $\hat{\theta}^{*}$ has the standard Taylor series form:

$$\left(\hat{\theta}^{*} - \theta_{o}\right) = -\hat{H}^{*}\left(\theta^{+}\right)^{-1}\hat{G}^{*}\left(\theta_{0}\right),$$

It then follows that:

$$\begin{aligned}\left(\hat{\theta}_{in}^{o} - \theta_{o}\right) &= -\hat{H}^{*}\left(\theta^{+}\right)^{-1}\hat{G}^{o}\left(\theta_{0}\right), \qquad (9)\\ \hat{G}^{o}\left(\theta_{0}\right) &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_{1},d_{2}}\left[Y_{i}\left(d_{1},d_{2}\right) - \hat{P}_{i}^{o}\left(d_{1},d_{2};\theta_{o}\right)\right]\hat{\delta}_{i}^{*}\left(d_{1},d_{2};\theta_{o}\right)\tau_{iv}.\end{aligned}$$

Notice that now the estimated probability is evaluated at an optimal window. The

feasible counterpart of the above estimator is given as:

$$\hat{\theta}^o \equiv \hat{\theta}^* - \hat{H}\left(\hat{\theta}^*\right)^{-1} \hat{C}\left(\hat{\theta}^*\right).$$

In the Appendix, we show that the feasible and infeasible estimators are asymptotically equivalent.

In the remainder of this section, we provide the main asymptotic results in several theorems below. Each theorem will depend on a number of intermediate results, which we state and prove as Lemmas in the Appendix. Theorem 1 below provides consistency and identification results. Theorem 2 provides the normality result using regular kernels throughout.

Theorem 1 (Consistency). For both binary selection and binary treatment models, assume that V_2 contains a continuous variable that is excluded from V_1. In addition, assume that each index satisfies the identifying assumptions required for single index models[5]. Then, under (A1-4) and (D1-9):

$$\hat{\theta} \xrightarrow{p} \theta_o, \hat{\theta}^* \xrightarrow{p} \theta_o, \hat{\theta}^o \xrightarrow{p} \theta_o$$

Proof. We provide the proof for $\hat{\theta}^*$, with the arguments for the other estimators being very similar. Lemmas 2-3 prove that we can replace the $\hat{P}^*$ in the objective function $\hat{L}^*(\theta)$, and obtain $L^*(\theta)$ satisfying:

$$\sup_{\theta} \left| \hat{L}^*(\theta) - L^*(\theta) \right| \xrightarrow{p} 0.$$

[5]See, for example, Ichimura (1993) or Klein and Spady (1993).

From Lemma 4, we may ignore the probability adjustments $\hat{\Delta}'s$ and therefore replace adjusted probabilities P^* in $L^*(\theta)$ with unadjusted ones P. With $L(\theta)$ as the resulting objective function:

$$\sup_{\theta} |L^*(\theta) - L(\theta)| \xrightarrow{p} 0.$$

From conventional uniform convergence arguments:

$$\sup_{\theta} |L(\theta) - E[L(\theta)]| \xrightarrow{p} 0.$$

To complete the argument, we must show that $E[L(\theta)]$ is uniquely maximized at θ_o. From standard arguments, θ_o is a maximum, and the only issue is one of uniqueness. With θ^* as any potential maximizer, it can be shown that:

$$\begin{aligned} 1)\Pr(Y_1 &= 1|Y_2 = 1, V_1(\theta_1^*), V_2(\theta_2^*))\Pr(Y_2 = 1|V_2(\theta_2^*)) = \\ \Pr(Y_1 &= 1|Y_2 = 1, V_1(\theta_{1o}), V_2(\theta_{2o}))\Pr(Y_2 = 1|V_2(\theta_{2o})) \end{aligned}$$

$$\begin{aligned} 2)\Pr(Y_1 &= 0|Y_2 = 1, V_1(\theta_1^*), V_2(\theta_2^*))\Pr(Y_2 = 1|V_2(\theta_2^*)) = \\ \Pr(Y_1 &= 0|Y_2 = 1, V_1(\theta_{1o}), V_2(\theta_{2o}))\Pr(Y_2 = 1|V_2(\theta_{2o})). \end{aligned}$$

Summing (1) and (2):

$$\Pr(Y_2 = 1|V_2(\theta_2^*)) = \Pr(Y_2 = 1|V_2(\theta_{2o})).$$

Under identifying conditions for single index models, $\theta_2^* = \theta_{2o}$. Turning to the remain-

ing index, from (1):

$$\Pr(Y_1 = 1|Y_2 = 1, V_1(\theta_{1o}), V_2(\theta_{2o})) = \Pr(Y_1 = 1|Y_2 = 1, V_1(\theta_1^*), V_2(\theta_{2o})).$$

Solving the first probability function for $V_1(\theta_{1o})$, for some function M we have:

$$V_1(\theta_{1o}) = M(V_1(\theta_1^*), V_2(\theta_{2o})).$$

Since V_2 contains a continuous variable not contained in V_1, differentiating both sides with respect to this variable yields:

$$0 = \nabla_{v_2} M \Rightarrow M(V_1(\theta_1^*), V_2(\theta_{2o})) = G(V_1(\theta_1^*)) = V_1(\theta_{1o}).$$

Identification now follows from conditions that identify single index models.

Theorem 2 (Normality). With $L(\theta)$ as the limiting likelihood defined in Theorem 1 and with H as its hessian matrix, define $H_o \equiv EH(\theta_o)$. Recall that the likelihood components were defined so as to be able to cover both binary selection and binary treatment models. Then, with $\hat{\theta}^o$ as the estimator defined in (D9) for these models and under (A1-4) and (D1-9):

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

Proof. Having established convergence rates for $\hat{\theta}^*$ in Lemma 9, we next show that the adjustment factor in the adjusted estimator defined in (D9) simplifies in

that:

$$\Delta \equiv \hat{H}\left(\hat{\theta}^*\right)\left[\hat{B}^o\left(\hat{\theta}^*\right) - \hat{B}^*\left(\hat{\theta}^*\right)\right] \; - \; \hat{H}\left(\theta^+\right)\left[\hat{B}^o\left(\theta_o\right) - \hat{B}^*\left(\theta_o\right)\right] = o_p(N^{-1/2}).$$

Rewriting the above expression, $\Delta \equiv \Delta_1 + \Delta_2$, where:

$$\begin{aligned}
\Delta_1 &\equiv \left[\hat{H}\left(\hat{\theta}^*\right) - \hat{H}\left(\theta^+\right)\right]\left[\hat{B}^o\left(\hat{\theta}^*\right) - \hat{B}^*\left(\hat{\theta}^*\right)\right] \\
\Delta_2 &\equiv \hat{H}\left(\theta^+\right)\left[\left(\hat{B}^o\left(\hat{\theta}^*\right) - \hat{B}^o\left(\theta_o\right)\right) - \left(\hat{B}^*\left(\hat{\theta}^*\right) - \hat{B}^*\left(\theta_o\right)\right)\right].
\end{aligned}$$

The first term of Δ_1 is $O_p\left(\frac{1}{\sqrt{N}h}\right)$, which follows from the above convergence rate on $\hat{\theta}^*$ and Lemma 1. The second term of Δ_1 is $O(h^2)$ from the convergence rate on $\hat{\theta}^*$ and Lemma 5. Therefore, $\Delta_1 = o_p(1/\sqrt{N})$. For Δ_2, the hessian component is $O_p(1)$ from Lemma 1. Taylor expanding the second term of Δ_2 :

$$\begin{aligned}
& \left(\hat{B}^o\left(\hat{\theta}^*\right) - \hat{B}^o\left(\theta_o\right)\right) - \left(\hat{B}^*\left(\hat{\theta}^*\right) - \hat{B}^*\left(\theta_o\right)\right) \\
= \; & \left[\nabla\hat{B}^o\left(\hat{\theta}^+\right) - \nabla\hat{B}^*\left(\hat{\theta}^+\right)\right]\left(\hat{\theta}^* - \theta_o\right), \; \hat{\theta}^+ \epsilon \left[\hat{\theta}^*, \theta_o\right].
\end{aligned}$$

Both gradient terms are $O_p(\frac{1}{\sqrt{N}h^3})$ from Lemma 1. Since from above: $\left(\hat{\theta}^* - \theta_o\right) = O_p\left(h^4\right)$, we have $\Delta_2 = o_p\left(1/\sqrt{N}\right)$.

Since the estimator based on the feasible adjustment factor is asymptotically equivalent to that based on the infeasible adjustment, we will complete the argu-

ment by analyzing the infeasible estimator. From (9):

$$\begin{aligned}\left(\hat{\theta}_{in}^{o}-\theta_{o}\right) &= -\hat{H}^{*}\left(\theta^{+}\right)^{-1}\left[\hat{A}-\hat{B}^{o}\right] \\ \hat{A} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}_i^{*}\left(d_1,d_2;\theta_o\right)\tau_{iv} \\ \hat{B}^{o} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^{o}\left(d_1,d_2;\theta_o\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}_i^{*}\left(d_1,d_2;\theta_o\right)\tau_{iv}.\end{aligned}$$

From Lemma 6:

$$\hat{A}=\frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv}+o_p(N^{-1/2})$$

where $\delta_i\left(d_1,d_2;\theta_o\right)$ is the probability limit of $\hat{\delta}_i^{*}\left(d_1,d_2;\theta_o\right)$.

It can be shown that:[6]

$$\hat{B}^{o}=B^{o}+o_p(N^{-1/2}),\ B^{o}=\frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^{o}\left(d_1,d_2;\theta_o\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv}.$$

Lemma 8b shows that $\hat{B}_1^{o}$ is a centered U-statistic and proves that $\hat{B}_1^{o}=o_p(1/\sqrt{N})$. The theorem now follows.

[6] From above, with $\hat{P}^{*}$ in place of $\hat{P}^{0}$, we showed that

$$\hat{B}^{*}=B^{*}+O_p\left(N^{-\frac{4}{8+\xi}}\right),\xi>0.$$

For the faster convergence rate on $\hat{P}^{0}$, from a similar argument the result follows.

6 Simulation Evidence

We now consider the finite sample performance of the estimator in four different models. These differ according to whether the model is threshold-crossing or not and according to whether there is binary treatment or sample selection. The first model we consider is a binary treatment model with additive non-normal errors:

$$\begin{aligned} Y_1 &= I\{X_1 + X_3 + Y_2 + \varepsilon_i > c_1\} \\ Y_2 &= I\{X_2 - X_3 + v_i > c_2\} \end{aligned}$$

where the errors are generated as:

$$\begin{aligned} v_i &\sim \chi^2(1) \\ \varepsilon_i &= v_i + z, \;\; z \sim N(0,1). \end{aligned}$$

and rescaled to each have variance 1. The variables X_1 and X_2 are standard normals, while X_3 is a binary variable with probability .5 and support {0,1}. The constants c_1 and c_2 are set so that the marginal probability for each dependent variable is .5.

In a second model, the treatment effect and the errors enter in a non-additive manner giving a non-threshold-crossing structure. More explicitly:

$$\begin{aligned} Y_1 &= I\{(X_1 + X_3) * (1 + mY_2 + s_1\varepsilon_i) > \varsigma_1\} \\ Y_2 &= I\{(X_2 - X_3) * (1 + s_2 v_i) > \varsigma_2\} \end{aligned}$$

where the variables are generated as in the first model. Notice that in the second

equation the error enters as $(X_2 - X_3)s_2 v_i$ which may be viewed as an error component with non-constant variance. We set the scaling constant s_2 so that the average variance of this component is one as in the first model. We set the scaling factor s_1 similarly. In the first equation, the marginal impact of Y_2 is not constant. We set the constant m so that the average marginal impact is the same as that in the first model. Finally ς_1 and ς_2 are set to give probabilities close to .5.

The above two models have a binary endogenous treatment component. As an alternative to this component, each of the above models can have a sample selection structure. In this case, the threshold-crossing model with sample selection is given as:

$$\begin{aligned} Y_1 &= I\{X_1 + X_3 + \varepsilon_i > c_1\} \text{ if } Y_2 = 1 \\ Y_2 &= I\{X_2 - X_3 + v_i > c_2\} \end{aligned}$$

where the variables, errors and constants are set as above.

Similarly, the non-threshold-crossing model with sample selection is given as:

$$\begin{aligned} Y_1 &= I\{(X_1 + X_3) * (1 + s_1\varepsilon_i) > \varsigma_1\} \text{ if } Y_2 = 1 \\ Y_2 &= I\{(X_2 - X_3) * (1 + s_2 v_i) > \varsigma_2\}. \end{aligned}$$

All other aspects of the model are as above.

For all four models, we set $N = 1000$ and conduct 1000 replications. To obtain starting values, we treat both models as if they were linear and then employ an IV estimator appropriate for this case. One might expect such starting values to be better

for the first model than for the second, which was indeed the case. Nevertheless, as discussed below, the final estimates for both models are quite good.

To evaluate the performance of the estimator proposed here we compare it with two alternative estimators. The first is the "MLE estimator" which is based on joint normality and which assumes a threshold structure. The second is an estimation procedure which employs the appropriate higher order kernels to achieve the appropriate bias reduction. We acknowledge that the "MLE estimator" is not appropriate for the non-threshold-crossing model even when the errors are normal. Nevertheless, we report it as this is the procedure frequently employed in situations where one observes binary treatments in models with binary outcomes. To further investigate the performance of bias reducing mechanisms in our estimator, we provide evidence on the bias reductions gained in each step of estimation.

Table 1 presents the estimates for these various estimators for the threshold-crossing model with binary treatment outlined above. For each of the estimates we report the bias, the standard error and the root square mean error. This provides the reader with not only some indication of the bias reduction but also the implications of the various bias reduction methods for the accuracy of the estimators. As the "MLE estimators" identify additional parameters, there is a different number of results for this estimator. Note that the estimates for the constants are not reported for the probit estimator although they are identified and were estimated jointly with the other parameters.

Beginning with the "MLE estimator" results in Table 1, we find that the bias is very large in the Y_2 model with both coefficients displaying a bias in the order of 20 percent. The bias which appears in the main equation is less severe and is in the

order of 10 percent or less. This almost certainly reflects the "smaller" departure from normality in the main equation error. Finally, we note that the variance for the estimated coefficient on Y_2 is very large in both absolute magnitude and relative to all other estimates.

Next we report the estimates obtained by the use of higher order kernels.[7] Given the nature of the estimator, we identify and estimate the ratio of the coefficients. The performance of the estimators is remarkably poor. For the reduced form the bias is 15 percent while for the main equation it is 42 percent. In both cases, the variances are also large. Unlike the other estimators studied here, there are convergence problems specific to this higher order kernel estimator. We therefore use a series of grid searches to obtain a maxima for this particular estimator.

The remainder of this table reports the performance of the various stages of our estimator. Note that the use of regular kernels with X trimming provides estimates with much smaller bias and variance compared to those based on higher kernels. The bias is around 7.5 percent in the reduced form and 12 percent in the main equation. This decreases dramatically however when we re-estimate the model and trim on the basis of the estimated indices. The bias is now 3.3 and 4.1 percent respectively. Finally the bias in the estimates after the smoothing adjustment is even smaller. Also note that this substantial reduction in bias is not associated with an increase in estimator variability. In all stages, the variances are relatively small, being on the order of 7% or less.

Table 2 reports the estimates from the non-threshold-crossing model. As expected the estimates for the "MLE estimator" are extremely poor in terms of bias and

[7] The higher order kernels are from Muller (1984), Table 1 for densities with smoothing parameter $\mu = 3.$ and with kernel parameter k = 4 in the single index case and k = 6 in the double index case.

variance. This highlights the danger in employing MLE with additive structures unless there is some reason to suspect this is indeed the appropriate model for the data. Once again the estimates based on higher kernels are very biased and have large variances. In contrast, the estimates based on the bias adjustments continue to perform well in this non additive setting. The bias, which is never large, is significantly reduced at each stage. The trimming on the estimated indices almost reduces the bias by half, while the smoothing adjustment further reduces bias by a noticeable amount. In all stages, the variances remain small, being lower than 7%.

Table 3 provides results for the threshold-crossing model with sample selection. The MLE estimates are poor in terms of bias and variability and those based on higher order kernels are also problematic. In contrast to the results in Table 1, the estimates based on X-trimming for the main equation are very poor with a bias of 31 percent. However, there is a substantial reduction in bias when we trim on the estimated indices. More explicitly, Table 3 reveals that the index-trimming estimates have biases of 4.3 and 2.7 percent respectively. The final round of bias adjustments further improves the estimates to biases of 3.7 and 1.8 percent respectively. As expected, the variances with sample selection are somewhat larger than without. However, they remain small, being less than 10%.

Finally, Table 4 reports results for the non-threshold-crossing model with sample selection. The results are generally of the same flavor as those above. Namely, the MLE estimators work poorly when their required assumptions are not satisfied. Also, the procedure based on higher order kernels does poorly. As with the earlier tables, the bias reducing methods work well. While the largest reduction in bias is due to the use of index trimming, the final bias corrections significantly reduce the bias further.

In summary, in all cases our estimator does much better than the "MLE estimator" and the higher order kernel based estimator in terms of both bias and variance. In terms of the bias control mechanisms that we employ, index trimming provides the largest bias reduction. The final smoothing adjustment further decreases the bias, most noticeably in the non-threshold-crossing models.

7 Conclusions

In conclusion, in this paper we have examined a class of triangular joint binary models where threshold crossing assumptions need not hold. We propose an estimator based on regular kernels with bias control mechanisms and show that the estimator is consistently and asymptotically distributed as normal. While retaining these desirable large sample properties, the Monte Carlo results show that our estimator performs very well in finite samples

While we have provided these results for a triangular model, **the (a)** double index formulation can be extended to a more general model (under appropriate assumptions) with each binary variable depending on the other.

Threshold Crossing Model					
	Equation	Parameter	Bias	Root Var	RMSE
Probit	Reduced Form	Coef(X1)	0.097	0.078	0.125
		Coef(X3)	0.081	0.127	0.151
	Primary	Coef(Y2)	0.050	0.274	0.278
		Coef(X1)	0.197	0.100	0.221
		Coef(X2)	-0.242	0.083	0.256
		Rho	-0.084	0.136	0.160
Higher-order Kernels	Reduced Form	Ratio_{31}	-0.417	0.300	0.513
	Primary	Ratio_{21}	-0.148	0.292	0.327
X-trimming	Reduced Form	Ratio_{31}	-0.119	0.072	0.140
	Primary	Ratio_{21}	-0.074	0.053	0.091
Index-trimming	Reduced Form	Ratio_{31}	-0.042	0.072	0.083
	Primary	Ratio_{21}	-0.033	0.045	0.055
Adjusted	Reduced Form	Ratio_{31}	-0.038	0.074	0.083
	Primary	Ratio_{21}	-0.021	0.044	0.049

Non-threshold Crossing Model					
	Equation	Parameter	Bias	Root Var	RMSE
Probit	Reduced Form	Coef(X1)	0.211	0.085	0.227
		Coef(X3)	0.227	0.093	0.245
	Primary	Coef(Y2)	0.145	0.177	0.229
		Coef(X1)	2.077	0.224	2.090
		Coef(X2)	-2.075	0.214	2.086
		Rho	-1.280	0.146	1.288
Higher-order Kernels	Reduced Form	Ratio_{31}	-0.317	0.355	0.476
	Primary	Ratio_{21}	-0.178	0.389	0.428
X-trimming	Reduced Form	Ratio_{31}	0.130	0.067	0.146
	Primary	Ratio_{21}	-0.050	0.039	0.064
Index-trimming	Reduced Form	Ratio_{31}	-0.078	0.066	0.102
	Primary	Ratio_{21}	-0.026	0.035	0.043
Adjusted	Reduced Form	Ratio_{31}	-0.061	0.067	0.091
	Primary	Ratio_{21}	-0.015	0.034	0.037

Partial Threshold Crossing Model					
	Equation	Parameter	Bias	Root Var	RMSE
Probit	Reduced Form	Coef(X1)	0.289	0.115	0.311
		Coef(X3)	0.289	0.183	0.342
	Primary	Coef(Y2)	-0.758	0.129	0.769
		Coef(X1)	0.194	0.099	0.218
		Coef(X2)	-0.239	0.083	0.253
		Rho	-0.357	0.230	0.425
Higher-order Kernels	Reduced Form	Ratio_{31}	-0.345	0.358	0.497
	Primary	Ratio_{21}	-0.172	0.336	0.377
X-trimming	Reduced Form	Ratio_{31}	-0.315	0.084	0.326
	Primary	Ratio_{21}	0.062	0.050	0.080
Index-trimming	Reduced Form	Ratio_{31}	-0.044	0.094	0.104
	Primary	Ratio_{21}	-0.027	0.045	0.052
Adjusted	Reduced Form	Ratio_{31}	-0.038	0.095	0.103
	Primary	Ratio_{21}	-0.018	0.044	0.048

Partial Non-Threshold Crossing Model					
	Equation	Parameter	Bias	Root Var	RMSE
Probit	Reduced Form	Coef(X1)	0.672	0.165	0.692
		Coef(X3)	0.648	0.195	0.676
	Primary	Coef(Y2)	-3.460	0.286	3.472
		Coef(X1)	2.051	0.227	2.064
		Coef(X2)	-2.045	0.216	2.056
		Rho	-1.184	0.297	1.221
Higher-order Kernels	Reduced Form	Ratio_{31}	-0.208	0.362	0.417
	Primary	Ratio_{21}	-0.145	0.374	0.401
X-trimming	Reduced Form	Ratio_{31}	-0.129	0.082	0.153
	Primary	Ratio_{21}	-0.056	0.040	0.069
Index-trimming	Reduced Form	Ratio_{31}	-0.080	0.080	0.113
	Primary	Ratio_{21}	-0.031	0.035	0.047
Adjusted	Reduced Form	Ratio_{31}	-0.063	0.082	0.104
	Primary	Ratio_{21}	-0.019	0.034	0.039

References

[1] Blundell, R. and J. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655-679

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

[3] De Luca, G. and F.Peracchi (2010): "Estimating models with unit and item nonresponse from cross-sectional surveys", University of Rome working paper.

[4] Gallant, A. and D. Nychka (1987)" "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica*, 15, 363-390.

[5] Heckman, James (1974): "Shadow Prices, Market Wages and Labor Supply," *Econometrica*, 42(4), 679-94.

[6] Heckman, James (1979): "Sample Selection Bias as a Specification Error," Econometrica, 47(1), 153-61.

[7] Hoderlein, S. (2009): "Endogenous Semiparametric Binary Choice Models with Heteroscedasticity," working paper, Brown University

[8] Honore, B. E. and J. L. Powell (2005): "Pairwise Difference Estimation of Nonlinear Models." *D. W. K. Andrews and J. H. Stock, eds., Identification and Inference in Econometric Models. Essays in Honor of Thomas Rothenberg* (Cambridge: Cambridge University Press), 520–53.

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

[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. and C. Shen (2009): "Bias Corrections in Testing and Estimating Semiparametric, Single Index Models," *Econometric Theory,* forthcoming.

[12] Klein, R., C. Shen and F.Vella (2009a): "Marginal Effects in Continuous Treatment Models with Selection," in progress.

[13] Klein, R., C. Shen and F.Vella (2009b): "Marginal Effects in Binary Treatment and Binary Selection Models," in progress.

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

[15] Klein, R. and F.Vella (2009): "A Semiparametric Model for Binary Response and Continuous Outcomes Under Index Heteroscedasticity," *Journal of Applied Econometrics,* 24, 735-762.

[16] Muller, H. (1984): "Smooth Optimum Kernel Estimators of Densities, Regression Curves and Modes," *The Annals of Statistics,"* V. 12(2), 766-774.

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

[18] Poirier, D. (1980): "Partial Observability in Bivariate Probit Models," *Journal of Econometrics*,12, 209-217.

[19] Robinson, P. (1988): "Root-N consistent semi-parametric regression," *Econometrica,* 56, 931-954.

[20] Rothe, C. (2009): " Semiparametric Estimation of Binary Response Models with Endogenous Regressors," forthcoming, *Journal of Econometrics.*

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

[22] Shen, C. (2008): "Determinants of Healthcare Decisions: Insurance, Utilization, and Expenditures," unpublished manuscript.

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

[24] Vella, F. (1998): "Estimating Models with Sample Selection Bias: A Survey." *The Journal of Human Resources*, 33:1, 127-169.

[25] Yavuzoglua, B. and I.Tunali(2009): "Edgeworth expansion based correction of selectivity bias in models of double selection," working paper

8 Appendix

The appendix provides the intermediate lemmas employed in proving the main theorems. We begin with a basic lemma that provides uniform convergence rates.

With V_2 having conditional density $g_2(v_2|Y = d_2)$ supported on $[a_2(d_2), b_2(d_2)]$, and V having conditional density $g(v|Y_1 = d_1, Y_2 = d_2)$ supported on $[a_k(d_k), b_k(d_k)]$, $k = 1, 2$, $\varepsilon > 0$, define:

$$\mathcal{V}_{2N} = \{v_2 : a_2(d_2) + h_m^{1-\varepsilon} < v_2 < b_2(d_2) - h_m^{1-\varepsilon}\} \tag{10}$$

$$\mathcal{V}_N = \{(v_{1,}v_2) : a_k(d_k) + h_c^{1-\varepsilon} < v_k < b_k(d_k) - h_c^{1-\varepsilon}\} \tag{11}$$

Lemma 1 (Uniform Convergence). For ψ any continuous function of θ, let $\nabla_\theta^p(\psi)$ be the p^{th} partial derivative of ψ with respect to θ, $\nabla_\theta^0(\psi) \equiv \psi$. Let $\hat{f}_2$ and $\hat{f}$ be the estimators in (D!) with respective probability limits f_2 and f. Then, for θ in a compact set, $t_2 \epsilon \mathcal{V}_{2N}$ as defined in 10, $t \epsilon \mathcal{V}_N$ as defined in 11, the following rates

hold for $p = 0, 1, 2$:

$$a) \quad : \quad \sup_{t_2,\theta} \left| \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - \nabla_\theta^p \left(f_2 \left(t_2; d_2 \right) \right) \right| = O_p \left(\min \left[h_m^2, \frac{1}{\sqrt{N} h_m^{p+1}} \right] \right)$$

$$b) \quad : \quad \sup_{t,\theta} \left| \nabla_\theta^p \left(\hat{f} \left(t; d_1, d_2 \right) \right) - \nabla_\theta^p \left(f \left(t; d_1, d_2 \right) \right) \right| = O_p \left(\min \left[h_c^2, \frac{1}{\sqrt{N} h_c^{p+2}} \right] \right).$$

Proof. As the proof is standard for density estimators, we outline it below for case a). Write:

$$\left| \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - \nabla_\theta^p \left(f_2 \left(t_2; d_2 \right) \right) \right| \leq \Delta_1 + \Delta_2,$$

$$\begin{aligned} \Delta_1 &\equiv \left| \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - E \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) \right| \\ \Delta_2 &\equiv \left| E \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - \nabla_\theta^p \left(f_2 \left(t_2; d_2 \right) \right) \right|. \end{aligned}$$

From Klein (1993), $\Delta_1 = O_p \left(\frac{1}{\sqrt{N} h_m^{p+2}} \right)$.

For Δ_2, with $\psi_2(v_2|Y_2 = 1)$ as the conditional density of v_2 conditioned on $Y_2 = 1$, write

$$\begin{aligned} E \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) &= \nabla_\theta^p E \left(\hat{f}_2 \left(t_2; d_2 \right) \right) \\ &= \Pr \left(Y_2 = d_2 \right) \nabla_\theta^p \int_{a_2(d_2)}^{b_2(d_2)} \frac{1}{h_m} K \left(\frac{t_2 - v_2}{h_m} \right) \psi_2(v_2|Y_2 = 1) dv_2 \\ &= \Pr \left(Y_2 = d_2 \right) \nabla_\theta^p \int_{(a_2(d_2) - t_2)/h_m}^{(b_2(d_2) - t_2)/h_m} K \left(z \right) \psi_2(t_2 + h_m z|Y_2 = 1) dz. \end{aligned}$$

Define:

$$C_0(t_2, d_2) = \int_{(a_2(d_2)-t_2)/h_m}^{(b_2(d_2)-t_2)/h_m} K(z)\, dz; \quad C_1(t_2, d_2) = \int_{(a_2(d_2)-t_2)/h_m}^{(b_2(d_2)-t_2)/h_m} zK(z)\, dz.$$

Then, from a Taylor series expansion of $\psi_2(t_2 + h_m z|Y_2 = 1)$ in h_m about $h_m = 0$:

$$\begin{aligned} E\nabla_\theta^p \left(\hat{f}_2(t_2; d_2)\right) &= \nabla_\theta^p \Pr(Y_2 = d_2)\, \psi_2(t_2)\, C_0(t_2, d_2) + h_m \nabla_\theta^p \Pr(Y_2 = d_2)\, \psi_2'(t_2)\, C_1(t_2, d_2) + O\left(h_m^2\right) \\ &= \nabla_\theta^p f_2(t_2; d_2)\, C_0(t_2, d_2) + h_m \nabla_\theta^p \Pr(Y_2 = d_2)\, \psi_2'(t_2)\, C_1(t_2, d_2) + O\left(h_m^2\right) \end{aligned}$$

For $t_2 \epsilon \mathcal{V}_{2N}$, $C_0(t_2, d_2)$ and $C_1(t_2, d_2)$ converge uniformly in t_2 to 1 and 0 respectively faster than h_m^2. The lemma now follows.

The next two lemmas prove that the estimated second-stage objective function $\hat{L}^*(\theta)$ is uniformly close to $L^*(\theta)$. Lemma 2 proves this result when indices are restricted to be smoothly in $\mathcal{V}_N$ while Lemma 3 establishes this result for indices smoothly restricted to be in the complement of $\mathcal{V}_N$.

Lemma 2. Referring (D3), define a smoothed indicator restricting v_i to $\mathcal{V}_N$ in 11 as:

$$l(v_i) \equiv \prod_k \tau[a_k(d_k) + h_{ck}^{1-\varepsilon}, v_{ki}]\tau[v_{ki}, b_k(d_k) - h_{ck}^{1-\varepsilon}].$$

Then:

$$\sup_\theta |\hat{L}_g^*(\theta) - L_g^*(\theta)| = o_p(1),$$

$$\begin{aligned}\hat{L}_g^*(\theta) &= \frac{1}{N}\sum_i \sum_{d_1,d_2} Y_i(d_1,d_2)\, Ln\left[\hat{P}_i^*(d_1,d_2;\theta)\right] l(v_i) \\ L_g^*(\theta) &= \frac{1}{N}\sum_i \sum_{d_1,d_2} Y_i(d_1,d_2)\, Ln\left[P_i^*(d_1,d_2;\theta)\right] l(v_i).\end{aligned}$$

Proof. In $\hat{L}_g^*(\theta)$, Taylor expand $Ln(\hat{P}_i^*)$ about $Ln(P_i^*)$ to obtain:

$$\hat{L}_g^*(\theta) - L_g^*(\theta) = \frac{1}{N}\sum_i \sum_{d_1,d_2} Y_i(d_1,d_2)\frac{1}{\hat{P}_i^+}\left[\hat{P}_i^*(d_1,d_2;\theta) - P_i^*(d_1,d_2;\theta)\right] l(v_i),$$

where $\hat{P}_i^+$ is inside the interval between $\hat{P}_i^*, P_i^*$. If $\hat{P}_i^*(d_1,d_2;\theta) - P_i^*(d_1,d_2;\theta)$ is uniformly close to zero, and $P_i^*(d_1,d_2;\theta)$ is uniformly bounded away from zero, the proof follows. We next show that $\hat{P}_i^*(d_1,d_2;\theta) - P_i^*(d_1,d_2;\theta)$ is uniformly close to zero. Denote:

$$\begin{aligned}\hat{P}_m^* &= \hat{P}^*(Y_{2i} = d_2 | V_{2i} = t_2) \equiv \hat{f}_2^*(t_2;d_2)/\hat{g}_2^*(t_2;d_2) \text{ where } \hat{g}_2^*(t_2;d_2) \equiv \sum_{d_2=0}^{1} \hat{f}_2^*(t_2;d_2); \\ \hat{P}_c^* &= \hat{P}^*(Y_{1i} = d_1 | Y_{2i} = d_2, V_i = t) \equiv \hat{f}^*(t;d_1,d_2)/\hat{g}^*(t;d_1,d_2) \text{ where } \hat{g}^*(t;d_1,d_2) \equiv \sum_{d_1=0}^{1} \hat{f}^*(t;d_1,d_2); \\ P_m^* &\equiv p\lim \hat{P}_m^*;\ P_c^* \equiv p\lim \hat{P}_c^*\end{aligned}$$

Then, in the selection model with $d_2 = 0$ write:

$$\hat{P}_i^*(d_1,d_2;\theta) - P_i^*(d_1,d_2;\theta) = (\hat{P}_m^* - P_m^*)$$

Otherwise:

$$\begin{aligned}\hat{P}_i^*\left(d_1,d_2;\theta\right)-P_i^*\left(d_1,d_2;\theta\right) &= \hat{P}_m^*\hat{P}_c^*-P_m^*P_c^* \\ &= (\hat{P}_m^*-P_m^*)(\hat{P}_c^*-P_c^*)+(\hat{P}_m^*-P_m^*)P_c^*+P_m^*(\hat{P}_c^*-P_c^*)\end{aligned}$$

As the analysis for all of these terms is similar, here we focus on $(\hat{P}_m^*-P_m^*)$.This term itself is comprised of several similar components, one of which from (D5) is given as:

$$\frac{|\hat{f}_2^*\left(t_2;d_2\right)-f_2^*\left(t_2;d_2\right)|}{\hat{g}_2^*\left(t_2;d_2\right)}\leq\frac{|\hat{f}_2\left(t_2;d_2\right)-f_2\left(t_2;d_2\right)|}{\hat{g}_2^*\left(t_2;d_2\right)}+\frac{|\hat{\Delta}_2-\Delta_2|}{\hat{g}_2^*\left(t_2;d_2\right)}$$

Because of the Δ-terms defined in (D5) $\inf\hat{g}_2^*\left(t_2;d_2\right)>h_m^{1/2}$. Therefore, from Lemma 1, the first term above converges in probability to 0. A similar argument applies to the second term.

Lemma 3. With $l\left(v_i\right)$ defined in Lemma 2, then:

$$\sup_\theta|\hat{L}_b^*(\theta)-L_b^*(\theta)|=o_p(1),$$

$$\begin{aligned}\hat{L}_b^*(\theta) &= \frac{1}{N}\sum_i\sum_{d_1,d_2}Y_i\left(d_1,d_2\right)Ln[\hat{P}_i^*\left(d_1,d_2;\theta\right)][1-l\left(v_i\right)] \\ L_b^*(\theta) &= \frac{1}{N}\sum_i\sum_{d_1,d_2}Y_i\left(d_1,d_2\right)Ln\left[P_i^*\left(d_1,d_2;\theta\right)\right]\left[1-l\left(v_i\right)\right]\end{aligned}$$

Proof. Write:

$$|\hat{L}_b^*(\theta)-L_b^*(\theta)|\ \leq|\hat{L}_b^*(\theta)|\ +|L_b^*(\theta)|$$

For the second term:

$$|L_b^*(\theta)| \;\leq\; \sup_{i,\theta}\left|\sum_{d_1,d_2} Y_i\left(d_1,d_2\right) Ln\left[P_i^*\left(d_1,d_2;\theta\right)\right]\right| \sup_{\theta}\frac{1}{N}\sum_i[1-l\left(v_i\right)]$$

From Klein and Spady (1993, footnote 14), $\inf P_i^*\left(d_1,d_2;\theta\right)$ is bounded away from 0. Therefore, the first term above is finite. The second term converges in probability to zero.

For $\hat{L}_b^*(\theta)$,we will show that $\inf \hat{P}_i^*\left(d_1,d_2;\theta\right) > 0$ and then employ the same argument above to complete the proof. With

$$\hat{P}_i^*\left(d_1,d_2;\theta\right) = \begin{cases} \hat{P}_m^* & \text{in the selection model with } d_2 = 0 \\ \hat{P}_m^*\hat{P}_c^* & \text{otherwise} \end{cases},$$

each of these components converges to a finite quantity that is bounded away from zero. With the argument for each component being the same, here we consider $\hat{P}_m^*$.

From the proof of Lemma 1 and employing the notation introduced therein:

$$\hat{P}_m^* - \frac{F\left(t_2;d_2\right)}{\sum\limits_{d_2} F\left(t_2;d_2\right)} = o_p(1), \text{ where } F\left(t_2;d_2\right) \equiv\; f_2\left(t_2;d_2\right) C_0(t_2,d_2) + \Delta_2(d_2).$$

Letting $\bar{C} = \max\limits_{d_2}(C_0(t_2,d_2))$, $\lambda = C_0(t_2,d_2)/\bar{C}$, and $\Delta_2^*(d_2) = \Delta_2(d_2)/\bar{C}$:

$$\begin{aligned} \frac{F\left(t_2;d_2\right)}{\sum\limits_{d_2} F\left(t_2;d_2\right)} \;&>\; \frac{F\left(t_2;d_2\right)}{\sum\limits_{d_2}\left[f_2\left(t_2;d_2\right)\bar{C} + \Delta_2(d_2)\right]} \\ &=\; \frac{\lambda f_2\left(t_2;d_2\right) + \Delta_2^*(d_2)}{\sum\limits_{d_2}\left[f_2\left(t_2;d_2\right) + \Delta_2^*(d_2)\right]}. \end{aligned}$$

Since $f_2(t_2; d_2) = P_m \sum_{d_2} f_2(t_2; d_2)$:

$$\frac{P_m \lambda f_2(t_2; d_2) + \Delta_2^*(d_2)}{f_2(t_2; d_2) + P_m \sum_{d_2} \Delta_2^*(d_2)} > \frac{(P_m \lambda) f_2(t_2; d_2) + \Delta_2^*(d_2)}{f_2(t_2; d_2) + \sum_{d_2} \Delta_2^*(d_2)}.$$

With $0 < P_m \lambda < 1$ behaving as a probability, from Klein and Spady (1993, footnote 14) the above quantity is finite and bounded away from zero. The lemma then follows.

The next lemma proves that we may ignore the probability adjustments $\hat{\Delta}'s$ in the adjusted likelihood, L^*, and therefore replace adjusted probabilities P^* in L^* with unadjusted ones P.

Lemma 4. Referring to (D8), for θ in a compact set:

$$\sup_{\theta} |L^*(\theta) - L(\theta)| \xrightarrow{p} 0$$

Proof. The proof is identical to the argument in Lemmas 2-3 and follows directly by establishing this result on both sets away from support boundaries and "low probability" sets near the boundaries.

To establish asymptotic normality, we require convergence rates for gradient and hessian components of the relevant estimated likelihoods. These rates are provided in Lemma 5 below.

Lemma 5. (Pointwise Convergence). For ψ any pth differentiable function of θ, let $\nabla_\theta^p(\psi)$ be the p^{th} partial derivative of ψ with respect to θ, $\nabla_\theta^0(\psi) \equiv \psi$. Let $\hat{f}_2$ and $\hat{f}$ be the estimators in (D!) with respective probability limits f_2 and f. Then, for θ in a compact set, $t_2 \epsilon \mathcal{V}_{2N}$ as defined in 10, $t \epsilon \mathcal{V}_N$ as defined in 11, the following

rates hold for $p = 0, 1, 2$:

$$a) \quad : \quad \left| \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - \nabla_\theta^p \left(f_2 \left(t_2; d_2 \right) \right) \right| = O_p \left(\min \left[h_m^2, \frac{1}{\sqrt{N h_m^{2p+1}}} \right] \right)$$

$$b) \quad : \quad \left| \nabla_\theta^p \left(\hat{f} \left(t; d_1, d_2 \right) \right) - \nabla_\theta^p \left(f \left(t; d_1, d_2 \right) \right) \right| = O_p \left(\min \left[h_c^2, \frac{1}{\sqrt{N h_c^{2p+2}}} \right] \right)$$

Proof. As the proof is standard, we outline it below for case a). Write:

$$E \left[\nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - \nabla_\theta^p \left(f_2 \left(t_2; d_2 \right) \right) \right]^2 \leq \Delta_1 + \Delta_2,$$

$$\begin{aligned} \Delta_1 &\equiv \left[\nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - E \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) \right]^2 \\ \Delta_2 &\equiv \left[E \nabla_\theta^p \left(\hat{f}_2 \left(t_2; d_2 \right) \right) - \nabla_\theta^p \left(f_2 \left(t_2; d_2 \right) \right) \right]^2 \end{aligned}$$

For Δ_1, this variance calculation is standard (e.g. see Silverman (1986)). For the bias calculation in Δ_2, the argument is identical to that for the uniform case in Lemma 1. The rate is then given by the minimum of the square roots of how fast Δ_1 and Δ_2 converge to zero, which completes the lemma.

To establish asymptotic normality for the adjusted estimator, it is useful to have a rate of convergence for first and second stage estimators. The following lemma is important in this regard as it provides a convergence rate for one of the gradient components of the estimators being studied here.

Lemma 6. For $\hat{\tau} = \hat{\tau}_v$ or $\hat{\tau}_x$, referring to (D9), with

$$\begin{aligned}
\hat{\delta}^{*}\left(d_1, d_2; \theta_o\right) &\equiv \nabla_\theta \hat{P}_i^{*}(d_1, d_2; \theta_o) / \hat{P}_i^{*}(d_1, d_2; \theta_o) \\
\hat{\delta}\left(d_1, d_2; \theta_o\right) &\equiv \nabla_\theta \hat{P}_i(d_1, d_2; \theta_o) / \hat{P}_i(d_1, d_2; \theta_o) \\
\hat{A}^{*} &\equiv \frac{1}{N} \sum_{i=1}^{N} \sum_{d_1, d_2} \left[Y_i\left(d_1, d_2\right) - P_i\left(d_1, d_2; \theta_o\right)\right] \hat{\delta}_i^{*}\left(d_1, d_2; \theta_o\right) \hat{\tau} \\
\hat{A} &\equiv \frac{1}{N} \sum_{i=1}^{N} \sum_{d_1, d_2} \left[Y_i\left(d_1, d_2\right) - P_i\left(d_1, d_2; \theta_o\right)\right] \hat{\delta}_i\left(d_1, d_2; \theta_o\right) \hat{\tau}
\end{aligned}$$

then

$$\begin{aligned}
\hat{A}^{*} - A &= o_p(N^{-1/2}) \\
\hat{A} - A &= o_p(N^{-1/2})
\end{aligned}$$

where

$$A \equiv \frac{1}{N} \sum_{i=1}^{N} \sum_{d_1, d_2} \left[Y_i\left(d_1, d_2\right) - P_i\left(d_1, d_2; \theta_o\right)\right] \delta_i\left(d_1, d_2; \theta_o\right) \tau$$

Proof. Using Lemma 5 and Lemma 2.18 from Pakes and Pollard (1989), Klein and Shen (2009) establishes this result for single index models. The argument extends to double index models.

Using Lemma 6, Lemma 7 provides a useful convergence rate for the initial estimator.

Lemma 7. For $\hat{\theta}$ defined in (D8) and with $h = O(N^{-r}), r = \frac{1}{8+\xi}$:

$$\left(\hat{\theta} - \theta_o\right) = O_p(h^2).$$

Proof. From a Taylor series expansion:

$$\begin{aligned}
\left(\hat{\theta}-\theta_o\right) &= -\hat{H}\left(\theta^+\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-\hat{P}_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_o\right)\tau_{ix} = -\hat{H}\left(\theta^+\right)^{-1}\left[\hat{A}-\hat{B}\right], \\
\hat{A} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_o\right)\tau_{ix}; \\
\hat{B} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i\left(d_1,d_2;\theta_o\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_o\right)\tau_{ix}
\end{aligned}$$

Referring to Lemma 6, since $A = O_p(N^{-1/2})$, $\hat{A} = O_p(N^{-1/2})$. From Lemma 5, $\hat{B} = O_p(h^2)$, which completes the argument.

To obtain a convergence rate for the second-stage estimator and to analyze the final bias-adjusted estimator, Lemma 8 shows that the gradient component which is responsible for the bias in the estimator vanishes in probability.

Lemma 8. Define:

$$\begin{aligned}
a) \quad &: \quad B^* = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^*\left(d_1,d_2;\theta_o\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv} = o_p(1) \\
b) \quad &: \quad B^o = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^o\left(d_1,d_2;\theta_o\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv} = o_p(1)
\end{aligned}$$

Proof. For a), under index trimming the adjustment factors within $\hat{P}_i^*$ vanish exponentially. Therefore:

$$B^* = B + o_p(1), B = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i\left(d_1,d_2;\theta_o\right)-P_i\left(d_1,d_2;\theta_o\right)\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv}$$

Denote:

$$\begin{aligned}
\hat{P}_m &= \hat{P}\left(Y_{2i} = d_2 | V_{2i} = t_2\right) \equiv \hat{f}_2\left(t_2; d_2\right) / \hat{g}_2\left(t_2; d_2\right) \text{ where } \hat{g}_2\left(t_2; d_2\right) \equiv \sum_{d_2=0}^{1} \hat{f}_2\left(t_2; d_2\right); \\
\hat{P}_c &= \hat{P}\left(Y_{1i} = d_1 | Y_{2i} = d_2, V_i = t\right) \equiv \hat{f}\left(t; d_1, d_2\right) / \hat{g}\left(t; d_1, d_2\right) \text{ where } \hat{g}\left(t; d_1, d_2\right) \equiv \sum_{d_1=0}^{1} \hat{f}\left(t; d_1, d_2\right); \\
P_m &\equiv p\lim \hat{P}_m; \; P_c \equiv p\lim \hat{P}_c
\end{aligned}$$

then, in the selection model with $d_2 = 0$ write:

$$\hat{P}_i\left(d_1, d_2; \theta\right) - P_i\left(d_1, d_2; \theta\right) = (\hat{P}_m - P_m)$$

Otherwise:

$$\begin{aligned}
\hat{P}_i\left(d_1, d_2; \theta\right) - P_i\left(d_1, d_2; \theta\right) &= \hat{P}_m \hat{P}_c - P_m P_c \\
&= (\hat{P}_m - P_m)(\hat{P}_c - P_c) + (\hat{P}_m - P_m)P_c + P_m(\hat{P}_c - P_c)
\end{aligned}$$

Since the argument for the first case is similar and easier, here we focus on the second case. In that case, we can rewrite B term as:

$$B = \frac{\sqrt{N}}{N} \sum_{i=1}^{N} \sum_{d_1, d_2} \left[(\hat{P}_m - P_m)(\hat{P}_c - P_c) + (\hat{P}_m - P_m)P_c + P_m(\hat{P}_c - P_c)\right] \delta_i\left(d_1, d_2; \theta_o\right) \tau_{iv}$$

For the first term in B, from Lemma 4, for any given d_1, d_2:

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_m - P_m)^2} = O_p(N^{-2r_m}); \; \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{P}_c - P_c)^2} = O_p(N^{-2r_c})$$

From Cauchy's inequality:

$$\frac{\sqrt{N}}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[(\hat{P}_m - P_m)(\hat{P}_c - P_c)\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv} = \sqrt{N}O_p(N^{-2(r_m+r_c)})$$

Recalling that $r_m = \frac{1}{6+\xi}$ and $r_c = \frac{1}{8+\xi}$, set ξ such that $r_m + r_c > 1/4$. Then, the term above vanishes in probability.

The second term in B is given by:

$$B_2 = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[(\frac{\hat{f}_2\left(t_2;d_2\right)}{\hat{g}_2\left(t_2;d_2\right)} - P_m)P_c\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv}$$

With:

$$U = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[(\frac{\hat{f}_2\left(t_2;d_2\right)}{\hat{g}_2\left(t_2;d_2\right)} - P_m)P_c\right]\left[\frac{\hat{g}_2\left(t_2;d_2\right)}{g_2\left(t_2;d_2\right)}\right]\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv},$$

employing the same "double-convergence" argument used on the first term it can be shown that:

$$B_2 = U + o_p(1).$$

Note that

$$U = \frac{\sqrt{N}}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[(\hat{f}_2\left(t_2;d_2\right) - \hat{g}_2\left(t_2;d_2\right)P_m)P_c\right]\left[\frac{\delta_i\left(d_1,d_2;\theta_o\right)\tau_{iv}}{g_2\left(t_2;d_2\right)}\right]$$

is a centered U-Statistic, which vanishes in probability from standard projection arguments. The third term in B has the same structure as the second and therefore also vanishes in probability, which completes the proof.

Lemma 9. Referring to (D5), for the second stage estimator:

$$\left|\hat{\theta}^* - \theta_o\right| = O_p\left(N^{-\frac{4}{8+\xi}}\right).$$

Proof. From Lemma 7, the initial estimator satisfies: $\left(\hat{\theta} - \theta_o\right) = O_p\left(N^{-2r}\right)$. For the estimator based on index trimming, from a standard Taylor series argument and employing the form for the gradient in (8) with τ_{iv} replacing τ_{ix} :

$$\begin{aligned}
\left(\hat{\theta}^* - \theta_o\right) &= -\hat{H}^*\left(\theta^+\right)^{-1}\left[\hat{A}^* - \hat{B}^*\right], \\
\hat{A}^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right) - P_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}^*\left(d_1,d_2;\theta_o\right)\tau_{iv}; \\
\hat{B}^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^*\left(d_1,d_2;\theta_o\right) - P_i\left(d_1,d_2;\theta_o\right)\right]\hat{\delta}^*\left(d_1,d_2;\theta_o\right)\tau_{iv}
\end{aligned}$$

Referring to Lemma 6, since $A = O_p(N^{-1/2})$, $\hat{A}^* = O_p(N^{-1/2})$.

For the $\hat{B}^*$-term, with $\Delta_{Bi} \equiv [\hat{\delta}^*\left(d_1,d_2;\theta_o\right)\tau_{iv} - \delta\left(d_1,d_2;\theta_o\right)\tau_{iv}]$:

$$\begin{aligned}
\hat{B}^* &= B_1^* + \hat{B}_2^*, \\
B_1^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^*\left(d_1,d_2;\theta_o\right) - P_i\left(d_1,d_2;\theta_o\right)\right]\delta\left(d_1,d_2;\theta_o\right)\tau_{iv} \\
\hat{B}_2^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^*\left(d_1,d_2;\theta_o\right) - P_i\left(d_1,d_2;\theta_o\right)\right]\Delta_{Bi}
\end{aligned}$$

By showing that $\hat{B}_1^*$ is close in probability to a centered U-statistic, Lemma 8, part a) proves that $B_1^* = o_p\left(N^{-1/2}\right)$.From Cauchy's inequality, the convergence rates in

Lemma 5, convergence rates on indicators arbitrarily close to $\sqrt{N}$ (see Klein and Spady (1993)), and with window parameters $r = r^* = \frac{1}{8+\xi}$: $\hat{B}_2^* = O_p\left(N^{-\frac{4}{8+\xi}}\right), \xi > 0$. For these window choices, from the uniform rates in Lemma 1: $\hat{H}^*\left(\theta^+\right) = H_o + o_p(1)$. It now follows that $\left|\hat{\theta}^* - \theta_o\right| = O_p\left(N^{-\frac{4}{8+\xi}}\right)$.