cemmap
centre for microdata methods and practice

Semiparametric selection models with binary outcomes

Roger Klein
Chan Shen
Francis Vella

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP30/11

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

Semiparametric Selection Models with Binary Outcomes

Roger Klein
Rutgers University

Chan Shen
Georgetown University

Francis Vella
Georgetown University

September 22, 2011

Abstract

This paper addresses the estimation of a semiparametric sample selection index model where both the selection rule and the outcome variable are binary. Since the marginal effects are often of primary interest and are difficult to recover in a semiparametric setting, we develop estimators for both the marginal effects and the underlying model parameters. The marginal effect estimator only uses observations which are members of a high probability set in which the selection problem is not present. A key innovation is that this high probability set is data dependent. The model parameter estimator is a quasi-likelihood estimator based on regular kernels with bias corrections. We establish their large sample properties and provide simulation evidence confirming that these estimators perform well in finite samples.

1 Introduction

Despite the substantial literature extending the sample selection model of Heckman (1974, 1979) there is no detailed semiparametric treatment of the model in which both the outcome variable and the selection rule are binary.[1] This represents a significant void as important empirical examples exist in many areas of micro economics. In the fully parametric setting both the model parameters and the marginal effects, the objects which are generally of primary interest, are easily obtainable. Less is known for the semiparametric index model considered here. In fact the literature does not address estimation of marginal effects in such a model. Even the index parameter estimator has not been developed explicitly, though it would be possible to develop such an estimator within various frameworks (see, e.g. Gallant and Nychka (1987), Klein and Spady (1993), Ichimura and Lee (1991), Lee (1995), and Klein and Vella (2009)). Issues related to the identification of the model we consider are explicitly treated in Newey (2007) although that paper does not address estimation. In a related literature Chesher (2005), Vytlacil and Yildiz (2007), and Shaikh and Vytlacil

[1] For a survey see Vella (1998).

(2011) discuss identification of the marginal impact of a discrete endogenous variable, but do not consider the case of sample selection.

This paper develops semiparametric estimators for both the index parameters and the marginal effects. We make no distributional assumptions and allow a model structure more general than threshold-crossing. Our primary focus is upon the marginal effects as they have not been addressed in this setting. In the fully parametric case the marginal effects can be retrieved as known functions of the parameter estimates. In the semiparametric case, these effects cannot be directly derived from parameter estimates because the error distributions are unknown. Moreover, the relevant distribution is difficult to estimate because the outcome equation is only observed for the selected sample. We propose to estimate the relevant distribution by focusing on those observations in an estimated high probability set where the selection probability tends to one. The framework of this approach is developed in Heckman (1990) and Andrews and Schafgans (1998) for a known high probability set. This set depends on the tail behavior of index and error distributions. Therefore, in practice it is important to study the empirical tail behavior so as to find the appropriate high probability set. In this paper we characterize the high probability set as one where the probability exceeds a cutoff that approaches one as the sample size increases. We propose and establish the theoretical properties for an estimator of this cutoff that depends on empirical tail behavior. Based on the estimated high probability set, we formulate a marginal effect estimator and provide the theory for it which takes the estimation of this set into account. In monte-carlo simulations, we find that the estimator performs very well in finite samples.

For our semiparametric model, estimation of the marginal effects requires estimates of the index parameters and we propose a likelihood-based procedure employing a double index formulation. The resulting estimator employs bias adjustment mechanisms similar to those developed for single index regression models in Klein and Shen (2010). Employing these likelihood-based bias adjustments we show that our estimator based on regular kernels has both desirable theoretical properties and good finite sample performance.[2]

Section 2 describes the model while Section 3 discusses estimators for the marginal effects and the index parameters. Section 4 provides the assumptions and the details of the estimators. Section 5 provides simulation evidence and concluding comments are offered in Section 6. The Appendix contains all proofs.

[2]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 Model

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

$$Y_{1i} = Y_{2i} * I\{g(X\beta_0, \epsilon_i) > 0\} \quad (1)$$
$$Y_{2i} = I\{h(Z\pi_0) + u_i > 0\}, \quad (2)$$

where Y_{1i} is only observed for the subsample for which $Y_{2i} = 1$. Here $I\{.\}$ is an indicator function; 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 with $h(.)$ being increasing; and β_0 and π_0 are unknown parameter values. When the model is additive, and the joint distribution of the errors is parametrically known, it can be estimated by maximum likelihood. However, without separability or known error distributions, the existing available estimators do not apply. Our proposed estimator for index parameters also applies to the case where Y_2 does not have a threshold-crossing structure. However, for the marginal effect estimator, the theory in this paper requires that the outcome equation (Y_2) has a threshold-crossing structure. Without loss of generality, we simplify the Y_2-model by replacing h with the identity function for notational purposes.

As in most semiparametric models the parameters are identified up to location and scale. Writing

$$X\beta_0 = b_1(X_{1i} + X_{2i}\theta_{10}) + c_1 \equiv b_1 V_{10} + c_1$$
$$Z\pi_0 = b_2(Z_{1i} + Z_{2i}\theta_{20}) + c_2 \equiv b_2 V_{20} + c_2,$$

the $\theta_0' s$ are identified, while the $b\prime s$ and $c\prime s$ are not identified. We refer to V_{10} and V_{20} as indices and assume that the model satisfies index restrictions:

$$\Pr(Y_{1i} = d_1, Y_{2i} = d_2 | X_i) = \Pr(Y_{1i} = d_1, Y_{2i} = d_2 | V_{1i}, V_{2i}) \quad (3)$$
$$\Pr(Y_{2i} = d_2 | X_i) = \Pr(Y_{2i} = d_2 | V_{2i}). \quad (4)$$

We impose this index structure, as opposed to a non-parametric one, to improve the performance of the estimators.

3 Estimation

3.1 Marginal Effects

Our marginal effect of interest is the change in $\Pr(Y_1 = 1|V_1)$ as V_1 responds to a change in one of the explanatory X-variables. To motivate this marginal effect, let Y_2 denote whether or not an individual decides to be screened for a particular illness and let Y_1 denote whether or not an individual has that disease. For simplicity, assume that the screen is completely accurate and necessary for diagnosis. We would like to know how a change in one of the X-variables affects the probability of having the disease for the entire population and not just the subgroup that are screened. In the fully parametric case (e.g. bivariate probit with selection) the probability of having the disease $\Pr(Y_1 = 1|V_1)$ is a known function, and the corresponding marginal effect of interest can be directly calculated once the parameters of the model are estimated.

Now consider the semiparametric case where the functional form of this probability function is not known. The probability of interest can be written as:

$$\begin{aligned}\Pr(Y_1 = 1|V_1) &= P(Y_1 = 1|Y_2 = 1, V_1, V_2) P_2 \\ &\quad + P(Y_1 = 1|Y_2 = 0, V_1, V_2)(1 - P_2).\end{aligned}$$

where $P_2 = P(Y_2 = 1|V_2)$. We can recover the first argument on the right hand side semiparametrically. That is, we can estimate the probability of having the disease given that the individual is screened and the probability that an individual elects to be screened. The question then becomes how to recover the second part: $P(Y_1 = 1|Y_2 = 0, V_1, V_2)(1 - P_2)$. In general, this probability is not estimable because we do not observe Y_1 (disease) when $Y_2 = 0$ (no screening). However, if $P_2 = 1$ this second term disappears and we can estimate the marginal effect of interest based only on the first term. In an approach related to that in Heckman (1990) and Andrews and Schafgans (1998, hereafter referred to as A&S), we estimate the marginal effect by only using those observations for which $P_2 > 1 - N^{-a}$, where $a > 0$ defines the high probability set of interest. The probability of being in this high probability set is given by $P_h = \Pr(V_2 > F^{-1}(1 - N^{-a})) = 1 - G\left(F^{-1}(1 - N^{-a})\right)$, where F and G are the distribution functions for the selection error and index respectively. For example, when the index has a Weibull distribution $G = 1 - exp(-v_2)$, and the error follows a Weibull distribution with thinner tail $F = 1 - exp(-u^c), c > 1$, $P_h = \exp(-[-Ln(N^{-a})]^{1/c})$. As the error tails become thinner (c increases), P_h increases. This example demonstrates that the appropriate value for a depends on the thickness of index tails relative to that for the error. As these tails are unknown, we employ and establish asymptotic results for a data-dependent value for a.

To describe our estimation strategy, we introduce $\zeta_0(\bar{v}) \equiv \Pr(Y_1 = 1|V_1 = \bar{v},)$ and write the true marginal effect as:

$$ME = \zeta_0(ve) - \zeta_0(vb),$$

where vb refers to a base or initial level of the index and ve refers to an index evaluated at a new level for the explanatory variable of interest. Then, with $\bar{v}$ referring to either ve or vb :

$$\hat{\zeta}(\bar{v}) \equiv \hat{f}_M / \hat{g}_M,$$

where with S as a smoothed indicator of the form in A&S that is one on a high probability set (and also controls for small density denominators):

$$\begin{aligned} \hat{f}_M &\equiv \sum_j \frac{1}{Nh} Y_{1j} Y_{2j} K\left[\left(\bar{v} - V_{1j}\right)/h\right] S_j \\ \hat{g}_M &\equiv \sum_j \frac{1}{Nh} Y_{2j} K\left[\left(\bar{v} - V_{1j}\right)/h\right] S_j. \end{aligned}$$

To motivate this estimator, notice that in a threshold-crossing model context, $\zeta(\cdot)$ is the distribution function for ε. Beginning with $\hat{f}_M$, it converges to its expectation given as:

$$\begin{aligned} E\left(\hat{f}_M\right) &= E\left[\Pr(Y_1 = 1|Y_2 = 1, V_1, V_2) \Pr(Y_2 = 1|, V_2) \frac{1}{h} K\left[\left(\bar{v} - V_{1j}\right)/h\right] S_j\right] \\ &\simeq E\left[\zeta(V_1) \frac{1}{h} K\left[\left(\bar{v} - V_{1j}\right)/h\right] S_j\right] \simeq \zeta(\bar{v}) \left[g\left(\bar{V}_1\right) E\left[S|V_1 = \bar{v}\right]\right]. \end{aligned}$$

It can be shown that $\hat{g}_M$ also converges to its expectation, which is approximately $g(\bar{v}) E\left[S|V_1 = \bar{v}\right]$, on a high probability set. The ratio then converges to $\zeta(\bar{v})$.

3.2 Index Parameters

While our proposed marginal effect estimator is the primary focus, it depends on estimated index parameters. These are obtained 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

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

and

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

Here $V_i\left(\theta\right) \equiv \left(V_{1i}\left(\theta\right), V_{2i}\left(\theta\right)\right)$, and τ_i is a trimming function defined below to control for small density denominators.

The properties of the estimates depend on how the probabilities entering the likelihood function are estimated. We employ adjusted probabilities with regular kernels (D8) and several bias-reducing mechanisms in (D10-11) to ensure that the estimator has desirable large sample properties and also performs well in finite samples. We discuss the details of these features when we present asymptotic results below.

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. Subject to some issues that we address below, the key to our bias reduction mechanisms is the result due to Whitney Newey (see Klein and Shen (2010) for a statement of Newey's result) that:

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

4 Assumptions and Definitions

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

A1. The Data. The observations are i.i.d. from the model in (1)-(2). The matrices X and Z have full rank with probability 1.

A2. Parameter Space. The vector of true parameter values $(\theta_{10}, \theta_{20})$ lies in the interior of a compact parameter space, Θ.

A3. Model. The indices V_1 and V_2 each contains a continuous exogenous variable. Further, V_2 contains at least one continuous variable, which is excluded from V_1. The model satisfies index restrictions as in (3) and (4).

A4. Densities. Let $g(v_1, v_2|Y_1, Y_2)$ be the conditional density for the indices. Letting $\nabla^p g$ be any of the partials or cross partials of g up to order p, with $\nabla^0 g = g$, assume $g > 0$ on all fixed compact subsets of the support for the indices, and $\nabla^p g$, $\frac{\partial}{\partial \theta}\left(\nabla^p g\right)$, and $\frac{\partial^2}{\partial \theta \partial \theta}\left(\nabla^p g\right)$ are bounded for $p = 0, 1, 2, 3, 4$.

A5. Let F be the distribution for the selection error, G the distribution function for the selection index, and G_c be the conditional distribution of $v_2|V_1 = \bar{v}$. For all $t > T$, assume that $1 - G(t) > 1 - F(t)$ and $1 - G_c(t) > 1 - F(t)$.

A6. Let $g(v_2)$ be the marginal density for V_2 and $g(v_2|\bar{v})$ the density for V_2 conditioned on $V_1 = \bar{v}$. For all $t > T$ assume that $O(g(t)) \geq O(g(t|\bar{v}))$.

A7. Assume $P(Y_1|Y_2, V_1 = v_1, V_2 = v_2)$ and $g(v_1, v_2)$ have up to four bounded derivatives with respect to v_1 at $\bar{v}$.

The first three assumptions are standard in index models. Assumption A4 provides required smoothness conditions for determining the order of the bias for density estimators. Similar to A&S, assumption A5 is needed to develop the large sample distribution for the estimator of marginal effects in the outcome equation. As is well known in the literature (see e.g. Kahn and Tamer (2010)) support conditions are needed for the consistency of these types of estimators. Assumptions A6-7 are used in Lemma 3 to derive the order of the bias in estimating marginal effect components. In addition to the above assumptions, we also need a number of definitions for densities, probability functions and estimators. The next section discusses how these definitions relate to the asymptotic results. One of these definitions (D4) provides moment conditions for selecting the high probability set.

D1. Unadjusted Probabilities. Let $K(\cdot)$ be a density symmetric about zero, σ_k be the standard deviation for V_k, $k = 1, 2$, and ξ be a small positive value. For the Y_2-model, let:

$$\begin{aligned}
\hat{P}(Y_{2i} = d_2|V_{2i} = t_2) &\equiv \hat{f}_2(t_2; d_2) / \sum_{d_2=0}^{1} \hat{f}_2(t_2; d_2), \\
\hat{f}_2(t_2; d_2) &\equiv \sum_{j=1}^{N} \frac{Y_{2j}^{d_2}(1 - Y_{2j})^{1-d_2}}{N h_m} K\left[\frac{t_2 - V_{2j}}{h_m}\right].
\end{aligned}$$

where $h_m \equiv \sigma_2 N^{-r_m}$, $r_m = \frac{1}{6+\xi}$.

For the Y_1-model, conditioned on $Y_2 = 1$, let:

$$\begin{aligned}
\hat{P}(Y_{1i} = d_1|Y_{2i} = 1, V_i = t) &\equiv \hat{f}(t; d_1) / \sum_{d_1=0}^{1} \hat{f}(t; d_1) \\
\hat{f}(t; d_1) &\equiv \sum_{j=1}^{N} \frac{Y_{1j}^{d_1}(1 - Y_{1j})^{1-d_1} Y_{2j}}{N h_{c1} h_{c2}} K(\frac{t_1 - V_{1j}}{h_{c1}}) K(\frac{t_2 - V_{2j}}{h_{c2}})
\end{aligned}$$

where $h_{c1} \equiv \sigma_1 h_c, h_{c2} \equiv \sigma_2 h_c, h_c \equiv N^{-r_c}, r_c = \frac{1}{8+\xi}$.

When the conditioning value t_k, is replaced by the observation V_{ik}, the above averages are taken over the $(N-1)$ observations for which $j \neq i$.[3]

[3]It can easily be shown that all estimators with windows depending on population standard deviations

D2. Smooth Trimming. Define a smooth trimming function as:

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

As we employ an index estimator that is $\sqrt{N}$-convergent, which is faster than the rate at which the estimated marginal effect converges, it can be shown that estimated index parameters can be taken as known. Accordingly, in the next three definitions related to marginal effects all quantities are defined in terms of known indices for expositional purposes.

D3. The S-function. With $b>0$, k a large integer, the S-function (adapted from A&S) is given as:

$$S(x) = \begin{cases} 0, & R1: x \leq 0 \\ 1-\exp\frac{-x^k}{b^k-x^k}, & R2: 0<x<b \\ 1, & R3: x \geq b. \end{cases}$$

With τ as an indicator restricting the density for V_2 to be above $O(N^{-\iota})$ where ι is a small positive number[4],

$$x \equiv \tau\left[Ln\left(\frac{1}{1-P}\right) - Ln\left(N^a\right)\right].$$

With $\hat{P}$ as an estimator for $P \equiv \Pr(Y_2=1|V_2)$, let $\hat{S} = S(x(\hat{a},\hat{P})) \equiv S(\hat{a},\hat{P})$, and $S_0 = S(x(a_0,P)) \equiv S(a_0,P)$.

D4. True and estimated high probability parameters a_0 and $\hat{a}$. With K_2 a normal twicing kernel (Newey et al. (2004)), $h_2 = O(N^{-.1})$,

$$\begin{aligned} \hat{E}_2\left(\hat{S}\right) &\equiv \frac{1}{N}\sum_j S(\hat{a},\hat{P}_{aj}) \equiv \hat{E}_2\left(S(\hat{a},\hat{P}_{aj})\right) \\ \hat{E}_2\left(\hat{S}^\kappa|\bar{v}\right) &\equiv \frac{\sum_j S^\kappa(\hat{a},\hat{P}_{aj})K_2\left[(\bar{v}-V_{1j})/h_2\right]}{\sum_j K_2\left[(\bar{v}-V_{1j})/h_2\right]} \equiv \hat{E}_2\left(S^\kappa(\hat{a},\hat{P}_{aj})\right) \\ \text{where } \hat{P}_{aj} &\equiv \sum_j Y_{2j}K_2\left[(\bar{v}-V_{1j})/h_2\right] / \sum_j K_2\left[(\bar{v}-V_{1j})/h_2\right]. \end{aligned}$$

are asymptotically the same as those based on sample standard deviations. For notational simplicity, we employ population standard deviations throughout.

[4]It can be shown that trimming based on a density estimator is asymptotically equivalent to trimming on the true density.

Then

$$a_0 = \arg\min_{a \epsilon \mathcal{A}} \left[\frac{[E(S_0)]^2}{E(S_0^2|\bar{v})} - N^{2(a_0-.4)+\varepsilon-\eta} \right]^2$$

$$\hat{a} = \arg\min_{a \epsilon \mathcal{A}} \left[\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} - N^{2(\hat{a}-.4)+\varepsilon-\eta} \right]^2,$$

$\mathcal{A} = \left\{a : 0 < a < .4 - \frac{\varepsilon-\eta}{2}\right\}$ where small positives numbers ε, η satisfy $\varepsilon > \eta$; $\frac{[E(S)]^2}{E(S^2|\bar{v})}$is an increasing function of N^{-a} for N sufficiently large.

The S-function in (D3) smoothly restricts observations to a high probability set where $P_2 > 1 - N^{-a}$. The moment conditions in (D4) reflect the bias/variance trade-off in estimating the marginal effect. To maximize the rate at which the mean-squared error for the marginal effect estimator converges to zero, the order of the squared bias should be the same as that for the variance. However, to establish normality, we need to send the squared bias to zero a little faster than the variance, which is achieved by setting $\eta > 0$ in the moment condition above. Detailed arguments are shown in Lemmas 3-4. We note that a_0 satisfies a moment condition in terms of expectations that depend on the sample size, N. Accordingly, while a_0 will not depend on the actual data, it may not be a fixed parameter value. More specifically, it will depend not only on the tails of index and error distributions but possibly also on the sample size. To illustrate the solution to the moment condition, recall the Weibull tail example in Section 3.1 and assume that the indices are independent and that b in (D3) is sufficiently small that the middle region R2 is almost empty. Ignoring density trimming for simplicity, it can be shown that a_0 is an increasing function of N with limiting value $a_0 = .4 - \frac{\varepsilon-\eta}{2}$ as N goes to infinity. Replacing expectations with semiparametric estimators, we define the data-dependent value $\hat{a}$ to satisfy a similar moment condition.

D5. The estimator for marginal effects.

$$\hat{\zeta}(\bar{v}) \equiv \frac{\sum_j \frac{1}{Nh} Y_{1j} Y_{2j} K\left[\left(\bar{v} - V_{1j}\right)/h\right] S\left(\hat{a}, \hat{P}_{aj}\right)}{\sum_j \frac{1}{Nh} Y_{2j} K\left[\left(\bar{v} - V_{1j}\right)/h\right] S\left(\hat{a}, \hat{P}_{aj}\right)},$$

where K is a regular kernel with window $h = O(N^{-.2-\varepsilon}), \varepsilon$ is the small positive value in (D4).

D6. Interior Index Trimming. Let $\hat{V}_k^U$ and $\hat{V}_k^L$ be the upper and lower sample index quantiles for the indices: $V_k \equiv V_k(\theta)$, $k = 1, 2$. Referring to (D2), define smooth

interior trimming functions as:

$$\hat{\tau}_I(t_k) \equiv \tau\left(\hat{V}_k^L, t_k\right)\tau\left(t_k, \hat{V}_k^U\right).$$

D7. Density Adjustment. Referring to (D1), 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 $\hat{f}(V; d_1, d_2)$. Then, define adjusted estimates as:

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

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

where the window parameters are set to be $a_{_{2N}} \equiv N^{-r_m/2}$ and $a_{_N} \equiv N^{-r_c/2}$.

D8. Adjusted Semiparametric Probability Functions. Let:

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

D9. Likelihood Trimming. Define τ_{ix} as an indicator that is equal to one if all of the continuous $X\prime s$ are between their respective lower and upper quantiles, and define τ_{iv} as an indicator that is equal to one if the index vector V_{0i} is between its lower and upper quantiles.

D10. First and Second Stage Estimators. We define the first stage estimator as:

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

Recall that τ_{ix} is a trimming function based on X while τ_{iv} is based on the index vector.[5] In the objective function above, replace $\hat{P}$ with $\hat{P}^*$ defined as in (D8), replace τ_{ix} with τ_{iv}, and term the new objective function as $\hat{L}^*(\theta)$. Then, define the second

[5] Define $\hat{\tau}_{ix}$ and $\hat{\tau}_{iv}$ as estimated trimming functions based on sample quantiles of X and the estimated index respectively. It can be shown using Lemma 2.18 in Pakes and Pollard(1989) that the estimators based on known trimming functions are asymptotically equivalent to those based on the true ones. For expositional simplicity, we take these trimming functions as known throughout.

stage estimator:

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

D11. The Adjusted Estimator. Letting

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

define $\hat{P}^o(d_1, d_2; \theta)$ as an estimated semiparametric probability function where the components are based on optimal window parameters: $r_m^o = 1/5$ and $r_c^o = 1/6$. Then, define a gradient correction as:

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

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

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

5 Asymptotic Results

5.1 Marginal Effects

We now provide the asymptotic results for the marginal effect estimator. We begin with a characterization theorem underlying consistency and normality. This result allows us to take the estimated high probability set in (D4) as given and provides a linear characterization of the estimator.

Theorem 1: Referring to (D5), define

$$\gamma_N(\bar{v}) \equiv \frac{\sum_j \frac{1}{Nh}\left[Y_{1j} - \zeta_0(\bar{v})\right] Y_{2j} K\left[\left(\bar{v} - V_{1j}\right)/h\right] S_j}{E\left(S | V_1 = \bar{v}\right) g(\bar{v})},$$

where γ_N depends on true selection probabilities and a known high probability set. Then:

$$C_N(\bar{v})\left[\hat{\zeta}(\bar{v}) - \zeta_0(\bar{v})\right] = C_N(\bar{v})\gamma_N(\bar{v}) + o_p(1)$$

where $C_N(\bar{v}) \equiv \frac{\sqrt{Nh}E(S|\bar{v})}{\sqrt{E(S^2|\bar{v})}}$ satisfies $C_N^2(\bar{v}) = O(\frac{1}{Var(\gamma_N(\bar{v}))})$.

The consistency and normality results now follow:

Theorem 2 (Consistency): Select the high probability set as in (D4) and assume

that

$$NhE(S|V_1 = \bar{v}) \to \infty$$

as N increases. Then, for the estimator defined in (D5):

$$\hat{\zeta}(\bar{v}) \xrightarrow{p} \zeta_0(\bar{v}).$$

As shown in the Appendix, this result follows from Theorem 1, because the bias and variance of $\gamma_N(\bar{v})$ both tend to zero as N increases.

Theorem 3 (Normality): Let

$$\begin{aligned} \hat{V} &= \widehat{Var}(\gamma_N(ve)) + \widehat{Var}(\gamma_N(vb)) \\ where\ \ \widehat{Var}(\gamma_N(\bar{v})) &= \frac{\hat{\zeta}(\bar{v})\left[1 - \hat{\zeta}(\bar{v})\right]\sum_j \frac{1}{Nh} K^2\left[(\bar{v} - V_{1j})/h\right]\hat{S}_j^2}{Nh\hat{E}_2^2\left(\hat{S}|\bar{v}\right)\hat{g}_M^2}. \end{aligned}$$

Then

$$\frac{\widehat{ME} - ME}{\sqrt{\hat{V}}} \xrightarrow{d} Z \tilde{} N(0,1).$$

To see that this result follows from Theorem 1, we need to show that the covariance between $\gamma_N(vb)$ and $\gamma_N(ve)$ tends to 0 as N increases and we need to establish the relevant Lindberg condition. These results are established in the Appendix.

5.2 Index Parameters

To provide an overview of the theoretical arguments, we note that the consistency argument is rather standard except that we need to accommodate the bias controls used in the normality arguments. Hence we start by giving an overview of the normality arguments.

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

$$\hat{G} = \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2} \hat{\varepsilon}_i(d_1,d_2)\,\hat{\delta}_i(d_1,d_2;\theta_0)\,\tau_i, \tag{5}$$

where $\hat{\varepsilon}_i(d_1,d_2) \equiv Y_i(d_1,d_2) - \hat{P}_i(d_1,d_2;\theta_0)$, $\hat{\delta}(d_1,d_2;\theta_0) \equiv \nabla_\theta \hat{P}_i(d_1,d_2;\theta_0)/\hat{P}_i(d_1,d_2;\theta_0)$, and we have taken the trimming function as known for expositional purposes. As is standard, the key part of the normality argument is to show that the normalized gradient converges to a normal distribution.

Denoting $\varepsilon_i(d_1,d_2) \equiv Y_i(d_1,d_2) - P_i(d_1,d_2;\theta_0)$, $\delta(d_1,d_2;\theta_0) \equiv \nabla_\theta P_i(d_1,d_2;\theta_0)/P_i(d_1,d_2;\theta_0)$, and suppressing the $(d_1,d_2;\theta_0)$notation for simplicity, for each cell we may write the nor-

malized gradient as:

$$\frac{1}{\sqrt{N}}\left[\sum_{i=1}^{N}\varepsilon_i\delta_i\tau_i+\sum_{i=1}^{N}\varepsilon_i\left(\hat{\delta}_i-\delta_i\right)\tau_i+\sum_{i=1}^{N}\left(\hat{\varepsilon}_i-\varepsilon_i\right)\left(\hat{\delta}_i-\delta_i\right)\tau_i+\sum_{i=1}^{N}\left(\hat{\varepsilon}_i-\varepsilon_i\right)\delta_i\tau_i\right].$$

We establish normality by showing that every term vanishes except the first.

The second term above readily vanishes from a mean-square convergence argument. For the third term, a Cauchy-Schwartz argument would enable us to separate the individual components and take advantage of the known convergence rate of each. However, the rates are not fast enough; hence we employ the adjustment in (D11) to speed up the convergence rates. It can be shown that for the adjusted estimator $\hat{\theta}^o$, the gradient for each cell will have the following form:

$$\frac{1}{\sqrt{N}}\left[\sum_{i=1}^{N}\varepsilon_i\delta_i\tau_i+\sum_{i=1}^{N}\varepsilon_i\left(\hat{\delta}_i-\delta_i\right)\tau_i+\sum_{i=1}^{N}\left(\hat{\varepsilon}_i^o-\varepsilon_i\right)\left(\hat{\delta}_i-\delta_i\right)\tau_i+\sum_{i=1}^{N}\left(\hat{\varepsilon}_i^o-\varepsilon_i\right)\delta_i\tau_i\right]$$

where $\hat{\varepsilon}_i^o\left(d_1,d_2\right)\equiv Y_i\left(d_1,d_2\right)-\hat{P}_i^o\left(d_1,d_2;\theta_0\right)$. With $\hat{P}_i^o$ based on an optimal window, the rate of convergence for $\hat{\varepsilon}_i^o-\varepsilon_i$ is now fast enough so that the third term vanishes.

For the final term, we rely on a result due to Whitney Newey (see Klein and Shen (2010), Theorem 0 for a statement of Newey's result). Namely, that with $F\left(V_i\left(\theta_0\right)\right)\equiv E(Y(d_1,d_2)|V\left(\theta_0\right))$:

$$\nabla_\theta P_i\left(d_1,d_2;\theta_0\right)=\nabla_\theta F\left(V_i\left(\theta_0\right)\right)-E\left[\nabla_\theta F\left(V_i\left(\theta_0\right)\right)|V_i\left(\theta_0\right)\right].$$

From the above theorem, $E\left(\delta_i|V_i\right)=0$. Therefore, this multiplicative gradient component can serve as a source of bias reduction. To exploit this residual-like property of the probability gradient, denote V_0 as the matrix of observations on the indices and define $H(V_0)\equiv E\left[\left(\hat{\varepsilon}_i^o-\varepsilon_i\right)|X\right]$. Then from an iterated expectations argument, conditioning on X :

$$EE\left[\left(\hat{\varepsilon}_i^o-\varepsilon_i\right)\delta_i\tau_i|X\right]=E\left[H(V_0)\delta_i\tau_i\right]=E\left(H(V_0)\,E\left[\delta_i\tau_i|V_0\right]\right.$$

If the trimming function depended on the index, this gradient component would now have zero expectation. We 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. We then show that this fourth term is equivalent to a centered U-statistic that converges in probability to zero. To achieve consistency with index trimming, we use the adjusted probabilities in (D7, D8) so that denominators are kept away from zero, while the estimated probability still goes rapidly to the truth.

The remainder of this section provides the main asymptotic results in several theorems.

Each theorem will depend on a number of intermediate results, which we state and prove as Lemmas in the Appendix. Theorem 4 below provides consistency and identification results. Theorem 5 provides the normality result using regular kernels throughout.

Theorem 4 (Consistency): Under (A1-4) and (D6-11):

$$\hat{\theta} \xrightarrow{p} \theta_0, \hat{\theta}^* \xrightarrow{p} \theta_0, \hat{\theta}^o \xrightarrow{p} \theta_0.$$

In double index models it is usually necessary to impose continuous exclusion restrictions on each index. Because we impose a single index restriction in estimating the Y_2-model, we are able to show that we do not require the exclusion restriction on V_2.

Theorem 5 (Normality): With $L(\theta)$ as the limiting likelihood of $\hat{L}^*(\theta)$ defined in (D10) and with H as its hessian matrix, define $H_0 \equiv EH(\theta_0)$. Then, with $\hat{\theta}^o$ as the estimator defined in (D11) and under (A1-4) and (D6-11):

$$\sqrt{N}\left[\hat{\theta}^o - \theta_0\right] \xrightarrow{d} Z \tilde{} N(0, -H_0^{-1}).$$

6 Simulation Evidence

We now consider the finite sample performance of the estimator in four different models. These differ according to: i) whether or not the model is threshold-crossing; and ii) whether or not the errors are normal. The first two models we consider have threshold-crossing structures. The first model (TNorm design) has normal errors and is given as:

$$\begin{aligned} Y_1^* &= I\left\{\sqrt{2}(X_1 + X_3) > \varepsilon\right\} \\ Y_2 &= I\left\{\sqrt{2}(X_2 - X_3) > v\right\}, \end{aligned}$$

where $Y_1 = Y_1^*$ is observed when $Y_2 = 1$. The errors and the continuous $X\imath s$ (X_1, X_2) are generated as:

$$\begin{aligned} v, X_2 &\sim N(0,1) \\ \varepsilon &= 2v + z, \;\; z \sim N(0,1) \\ X_1 &= X_2 + 2z_1, \;\; z_1 \sim N(0,1). \end{aligned}$$

and re-scaled to each have variance 1, while X_3 is a binary variable independent of the errors and the continuous $X\imath s$ above with probability .5 at each of its support points: -1,1. Notice that the indices have bigger variance than the errors. For the second index, this ensures that the index has fatter tails than the error, which is theoretically needed in estimating the marginal effect.

In a second model (TWeibull design), the selection error is non-normal while the model structure stays the same. The error v follows a Weibull (1,1.5) giving a right tail probability of $exp(-v^{1.5})$. We set the X_2 to follow Weibull (1,1) so that the tail comparison condition is satisfied. As above, all the variables and errors are rescaled to have zero mean and variance one.

In the third (NTNorm design) and fourth (NTWeibull design) models, the Y_1^* equation has a non-threshold-crossing structure:

$$Y_1^* = I\left\{X_1 + X_3 > s\left[1 + (X_1 + X_3)^2/4\right]\varepsilon\right\}$$

where the variables are generated as in the previous models. Note that s is chosen to ensure the right-hand-side of the inequality is rescaled to have zero mean and variance one as above. Similar to the first two models above, here the third and fourth models differ according to whether Normal or Weibull distributions are employed.

For all models, we set $N = 2000$ and conduct 1000 replications. We compare the finite sample performance of our semiparametric marginal effect estimator and the bivariate probit with selection counterpart. We also compare the parameter estimates upon which these marginal effects are based. Finally, we also provide results for the estimation of the high probability set. Results for the marginal effects are shown in Table 1. Notice that there are an infinite number of marginal effects because there are an infinite number of base levels and evaluation levels. Here we report the marginal effect of moving X_1 from its median level to one unit above while keeping the binary variable X_3 at zero. Overall, the semiparametric estimator performs well with a small bias and standard deviation over all designs. In contrast, the bivariate probit counterpart does not perform well outside of the TNorm design where bivariate probit is correct. In the TNorm case, where bivariate probit is the correct specification, it does indeed have a small bias and standard deviation. However, the advantage over the semiparametric marginal effect is minimal. The RMSE of bivariate probit is .06 compared with .07 from the semiparametric counterpart. In the TWeibull case, the semiparametric method shows significant advantage in terms of the bias. The bias of the semiparametric marginal effect estimator is almost zero, while the bivariate probit counterpart has a bias of .08, which is almost 30% of the truth (.29). When we move on to the non-threshold-crossing designs, we continue to see the semiparametric estimator performing significantly better. In the NTNorm case, the semiparametric estimator has both smaller bias (.02 vs .10) and smaller standard deviation (.05 vs .07). In the NTWeibull case, the semiparametric estimator still performs much better than the bivariate probit in terms of RMSE (.08 vs .21). Most of the advantage comes from the standard deviation (.06 vs .20).

The direct comparison between parametric and semiparametric estimators is best done in terms of marginal effects as was done above. Nevertheless, we also provide the index parameter estimation results in Table 2. For semiparametric estimation, the parameters are identified up to location and scale, hence we report $\text{Ratio}_{31}=\frac{coef(X_3)}{coef(X_1)}$ in the outcome equation and $\text{Ratio}_{32}=\frac{coef(X_3)}{coef(X_2)}$ in the selection equation. Notice that for the non-threshold-crossing designs, we report the median and median absolute deviation (MAD) for the bivariate probit estimators because there were a number of replications where bivariate probit performed extremely poorly. The semiparametric estimator, however, does not have this issue, hence we report not only median and MAD but also mean, standard deviation, and RMSE. For the selection equation, over all designs, both parametric and semiparametric estimators perform quite well. Turning to the outcome equation, both estimators perform better in normal than in non-normal designs and also better in threshold-crossing than in non-threshold-crossing designs. The non-threshold-crossing model with Weibull distributions poses the most challenge for both estimators. It is noteworthy that for all other designs,the bias and the standard deviation for the semiparametric estimator are quite small. Finally, we also investigated the performance of higher order kernels for estimating index parameters as an alternative to the bias controls implemented here.[6] Due to convergence problems, we found it necessary to calculate this estimator on a two-dimensional grid, which was quite time-consuming. Accordingly, we only examined 100 replications for each design (at which point the estimator seemed quite stable). For the selection equation, the RMSE's were close with the exception of the TWeibull design where the RMSE using higher order kernels was 2.5 times larger. For the output equation, in all designs the RMSE under higher order kernels was approximately 3 times larger.

Lastly, we provide the estimation results for the high probability set parameters. The means of $\hat{a}$ with standard deviations in parentheses are as follows: .31(.004), 28(.006),31(.004), and .28(.005) for TNorm,TWeibull, NTNorm, and NTWeibull respectively. While the variances for all of the estimates are quite small, it is difficult to evaluate the performance of the estimator without knowing a_0. Accordingly, we examined the performance of the estimator for the example given earlier in section 4. Following the discussion in (D4), the moment condition is equivalent to:

$$\left[2 + (a_0 \ln N)^{\frac{1}{c}-1}\right] a_0 = .8 - (\varepsilon - \eta)\,.$$

Since a_0 depends on the sample size, we examined three different sample sizes: $N = 500$, 1000, and 2000. At each of these sample sizes, we solved the above equation for a_0 and conducted a monte-carlo with 100 replications to evaluate the performance of $\hat{a}$ at the base

[6] In our monte-carlo studies, the higher order kernel we use is the twicing kernel for both index parameter estimation and estimation of the high probability set parameter.

level of the index. The results are as follows:

$SAMPLE\ SIZE$	a_0	$\|BIAS\|$	SD	$RMSE$
500	.279	.037	.027	.046
1000	.280	.025	.025	.035
2000	.283	.019	.009	.020

where bias, standard deviation(SD) and RMSE are standardized by the truth a_0. This table shows that our $\hat{a}$ performs very well in terms of absolute bias, standard deviation and RMSE. It also confirms that the absolute bias, standard deviation and RMSE all decline as the sample size increases. As expected, a_0 increases slowly with the sample size.

7 Conclusions

This paper studies the binary outcome model with sample selection in a semiparametric framework. As marginal effects are often of primary interest in this type of model, we propose a semiparametric marginal effect estimator. This marginal effect estimator is based on observations in a high probability set where the selection probabilities are above a cutoff. We propose an estimator for this cutoff and establish its large sample properties. Based on that, we establish the large sample properties for our marginal effect estimator, which takes into account that the cutoff and the selection probability are estimated. In a monte-carlo study we find that our marginal effect estimator based on an estimated high probability set performs quite well in finite samples.

This marginal effect estimator is developed under an index framework so as to achieve good performance in finite samples. Accordingly, it depends on an estimator for index parameters. In this paper, we propose an index parameter estimator based on regular kernels with bias control mechanisms and show that the estimator is consistent and asymptotically distributed as normal. While retaining these desirable large sample properties, the monte-carlo results show that this estimator performs very well in finite samples.

Marginal Effect Estimators				
	Truth		Bivariate Probit	Semiparametric
TNorm	.34	mean	.33	.31
		std	.06	.06
		RMSE	.06	.07
TWeibull	.29	mean	.37	.29
		std	.06	.06
		RMSE	.10	.06
NTNorm	.46	mean	.36	.48
		std	.07	.05
		RMSE	.13	.05
NTWeibull	.59	mean	.63	.63
		std	.20	.06
		RMSE	.21	.08

Index Parameters						
	Bivariate Probit				Semiparametric	
	Outcome		Selection		Outcome	Selection
	Coef(X1)	Coef(X3)	Coef(X2)	Coef(X3)	$Ratio_{31}$	$Ratio_{32}$
TNorm						
mean	.98	1.02	1.01	-1.01	.95	-1.04
std	.05	.23	.06	.04	.07	.05
RMSE	.05	.22	.06	.04	.08	.06
TWeibull						
mean	1.23	1.23	1.02	-1.06	.93	-1.04
std	.09	.21	.06	.04	.06	.04
RMSE	.25	.31	.06	.08	.10	.06
NTNorm						
mean					.96	-1.02
median	1.07	1.13	1.00	-1.00	.96	-1.01
std					.05	.04
MAD	.12	.15	.03	.03	.04	.03
RMSE					.06	.04
NTWeibull						
mean					.84	-1.04
median	2.30	2.52	1.03	-1.07	.84	-1.04
std					.04	.04
MAD	1.30	1.52	.05	.07	.16	.05
RMSE					.16	.05

References

[1] Andrews, D. and M. Schafgans (1998): "Semiparametric Estimation of the Intercept of a Sample Selection Model" *Review of Economic Studies*, 65, 497-517.

[2] Chesher, A. (2005): "Nonparametric Identification under Discrete Variation", *Econometrica,* 73, 1525-1550.

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

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

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

[6] Heckman, J. (1990): "Varieties of Selection Bias," *American Economic Review*, 80, 313-18.

[7] 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.

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

[9] 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.

[10] Khan, S. and E.Tamer (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation, *Econometrica*, 6, 2021-2042.

[11] Klein, R. and C. Shen (2010): "Bias Corrections in Testing and Estimating Semiparametric, Single Index Models," *Econometric Theory,* 1683-1718.

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

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

[15] Newey, W., F. Hsieh, and J. Robins (2004): "Twicing Kernels and a Small Bias Property of Semiparametric Estimators," *Econometrica*, 72, 947-962.

[16] Newey, W, (2007): "Nonparametric continuous/discrete choice models", *International Economic Review*, 48: 1429–1439.

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

[18] Shaikh, A. M. and Vytlacil, E. J. (2011): "Partial Identification in Triangular Systems of Equations With Binary Dependent Variables," *Econometrica* 79(3), 949–955.

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

[20] Vytlacil, E. and N. Yildiz (2007): "Dummy Endogenous Variables in Weakly Separable Models," *Econometrica*, 75, 757-779.

8 Appendix

8.1 Main Results

8.1.1 Marginal Effects

Proof of Theorem 1: By definition,

$$C_N(\bar{v})\left[\hat{\zeta}\left(\bar{v}\right)-\zeta_0\left(\bar{v}\right)\right]=C_N(\bar{v})\frac{\sum_j \frac{1}{Nh}\left[Y_{1j}-\zeta_0\left(\bar{v}\right)\right]Y_{2j}K\left[\left(\bar{v}-V_{1j}\right)/h\right]\hat{S}_j}{\sum_j \frac{1}{Nh}Y_{2j}K\left[\left(\bar{v}-V_{1j}\right)/h\right]\hat{S}_j}.$$

Lemma 8 enables us to replace (up to $o_p(1)$) the denominator with $E\left(S|V_1=\bar{v}\right)g(\bar{v})$, while Lemma 9 continues to show that the numerator has the desired form.

Proof of Theorem 2: By Theorem 1:

$$C_N(\bar{v})\left[\hat{\zeta}\left(\bar{v}\right)-\zeta_0\left(\bar{v}\right)\right]=C_N(\bar{v})\gamma_N(\bar{v})+o_p(1).$$

Lemma 3 characterizes the order of the bias of the estimator. Recalling the definition of high probability parameter in (D4), the bias in the estimator vanishes. From Lemma 4, the reciprocal of the estimator variance has the order:

$$NhE\left(S|\bar{v}\right)^2/E\left(S^2|\bar{v}\right)>NhE\left(S|\bar{v}\right)$$

which completes the proof.

Proof of Theorem 3: By definition, $\widehat{ME}-ME=\left[\hat{\zeta}(ve)-\zeta_0(ve)\right]-\left[\hat{\zeta}(vb)-\zeta_0(vb)\right]$. We begin by showing that the covariance between these two components vanishes. Notice that $\hat{\zeta}(\bar{v})-\zeta_0(\bar{v})$ is close to $\gamma_N(\bar{v})$ which we can write as a sample average $\sum_j \frac{1}{N}t_j\left(\bar{v}\right)$. The covariance is then of the form $E\left[t_j\left(ve\right)t_k\left(vb\right)\right]$ For $j\neq k$, from independence and the vanishing bias of the expectation of each term, this expectation vanishes. For $j=k$, the kernel function ensures that this expectation also vanishes as V_{1j} cannot be close to both ve and vb. Therefore, we can study the sum of the variances of $\gamma_N(ve)$ and $\gamma_N(vb)$.

To prove normality, we first establish a Lindberg condition for $C_N(\bar{v})\left[\hat{\zeta}\left(\bar{v}\right)-\zeta_0\left(\bar{v}\right)\right]$. Namely, for $\varepsilon>0$, we must show that the following expectation converges to 0:

$$E\left\{\frac{(Y_{1i}-\theta_o)^2Y_{2i}k_i^2S_i^2/h}{E(S^2|\bar{v})}1\left\{(Y_{1i}-\theta_o)^2Y_{2i}k_i^2S_i^2>Nh^2E(S^2|\bar{v})\right\}\right\}.$$

Since $(Y_{1i}-\theta_o)^2 Y_{2i} k_i^2 S_i^2$ is bounded, it suffices to show that $Nh^2 E(S^2|v) \to \infty$:

$$\begin{aligned} Nh^2 E(S^2|v) &> Nh^2 \Pr(R3|\bar{v}) \\ \text{where } \Pr(R3|\bar{v}) &= \Pr(F(V_1) > 1 - N^{-a_0}/\exp(b)), \\ &= 1 - G_c(F^{-1}(1 - N^{-a_0}/\exp(b))) \\ &> 1 - F(F^{-1}(1 - N^{-a_0}/\exp(b))) \text{ from (A5).} \end{aligned}$$

Since $h = O(N^{-.2})$ and $0 < a_0 < .4$, the normality of $C_N(\bar{v}) \left[\hat{\zeta}(\bar{v}) - \zeta_0(\bar{v})\right]$ follows.

Turning to the marginal effects, for expositional purposes, suppose $O(Var(\gamma_N(ve))) > O(Var(\gamma_N(vb)))$, then

$$\begin{aligned} \frac{1}{\sqrt{Var(\gamma_N(ve)) + Var(\gamma_N(vb))}} &= O\left(\frac{1}{\sqrt{Var(\gamma_N(ve))}}\right) \\ &= O\left(C_N(ve)\right). \end{aligned}$$

Therefore, the characterization results in Theorem 1 apply to yield:

$$\frac{\widehat{ME} - ME}{\sqrt{Var\left(\widehat{ME}\right)}} = O\left(C_N(ve)\right)\left[\hat{\zeta}(ve) - \zeta_0(ve)\right] + o_p(1).$$

Now asymptotic normality follows from the above Lindberg condition. A symmetric argument holds for the case where $O(Var(\gamma_N(ve))) < O(Var(\gamma_N(vb)))$. For the case where $O(Var(\gamma_N(ve))) = O(Var(\gamma_N(vb)))$ a Lindberg condition similar to the above applies. Therefore, $\frac{\widehat{ME}-ME}{\sqrt{Var(\widehat{ME})}} \xrightarrow{d} Z \tilde{} N(0,1)$. Employing similar arguments as in Lemma 8, it can be shown that $\frac{Var(\widehat{ME})-\hat{V}}{Var(\widehat{ME})} \xrightarrow{p} 0$. Hence the theorem follows.

8.1.2 Index Parameters

Proof of Theorem 4: We provide the proof for $\hat{\theta}^*$, with the arguments for the other estimators being very similar. Lemma 10 proves that we can replace the $\hat{P}^*$ with P^* in the objective function $\hat{L}^*(\theta)$, and obtain $L^*(\theta)$ satisfying:

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

From Lemma 11, 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_0. From standard arguments, θ_0 is a maximum, and the only issue is one of uniqueness. With θ^* as any potential maximizer, it can be shown that any candidate for a maximum must give correct probabilities for all three cells: $(Y_1 = 1, Y_2 = 1), (Y_1 = 0, Y_2 = 1)$, and $Y_2 = 0$. It then follows that for the $Y_2 = 0$ cell:

$$\Pr(Y_2 = 0|V_2(\theta_2^*)) = \Pr(Y_2 = 0|X) = \Pr(Y_2 = 0|V_2(\theta_{20})).$$

Under identifying conditions for single index models, $\theta_2^* = \theta_{20}$. For the $(Y1 = 1, Y2 = 1)$ cell:

$$\begin{aligned} \Pr(Y_1 &= 1|Y_2 = 1, V_1(\theta_1^*), V_2(\theta_2^*)) \Pr(Y_2 = 1|V_2(\theta_2^*)) = \\ \Pr(Y_1 &= 1|Y_2 = 1, V_1(\theta_{10}), V_2(\theta_{20})) \Pr(Y_2 = 1|V_2(\theta_{20})). \end{aligned}$$

Since $\theta_2^* = \theta_{20}$:

$$\Pr(Y_1 = 1|Y_2 = 1, V_1(\theta_{10}), V_2(\theta_{20})) = \Pr(Y_1 = 1|Y_2 = 1, V_1(\theta_1^*), V_2(\theta_{20})).$$

Solving the first probability function for $V_1(\theta_{10})$, for some function Υ we have:

$$V_1(\theta_{10}) = \Upsilon(V_1(\theta_1^*), V_2(\theta_{20})).$$

Since V_2 contains a continuous variable not contained in V_1, differentiating both sides with respect to this variable yields $\nabla_{v_2}\Upsilon = 0$. Therefore, Υ must only be a function of the first index. Calling this function G:

$$G(V_1(\theta_1^*)) = V_1(\theta_{10}).$$

Identification now follows from conditions that identify single index models.

Proof of Theorem 5: From a Taylor expansion, the unadjusted estimator has the

form:

$$\left(\hat{\theta}^* - \theta_0\right) = -\hat{H}\left(\theta^+\right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[Y_i\left(d_1,d_2\right) - \hat{P}_i^*\left(d_1,d_2;\theta_0\right)\right] \hat{\delta}_i^*\left(d_1,d_2;\theta_0\right) \tau_{iv}.$$

where θ^+ is an intermediate point. To simplify the adjustment to this estimator, referring to (D11) we will show below:

$$\Delta \equiv \hat{H}\left(\hat{\theta}^*\right)^{-1} \hat{C}\left(\hat{\theta}^*\right) - \hat{H}\left(\theta^+\right)^{-1} \hat{C}\left(\theta_0\right) = o_p(N^{-1/2}).$$

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

$$\begin{aligned} \Delta_1 &\equiv \hat{H}\left(\theta^+\right)^{-1} \hat{H}\left(\hat{\theta}^*\right)^{-1} \left[\hat{H}\left(\theta^+\right) - \hat{H}\left(\hat{\theta}^*\right)\right] \hat{C}\left(\hat{\theta}^*\right) \\ \Delta_2 &\equiv \hat{H}\left(\theta^+\right)^{-1} \left[\hat{C}\left(\hat{\theta}^*\right) - \hat{C}\left(\theta_0\right)\right]. \end{aligned}$$

To study Δ_1, note that lemma 15 gives a convergence rate for $\hat{\theta}^* - \theta^+$. Then, using a Taylor series expansion on $\left[\hat{H}\left(\theta^+\right) - \hat{H}\left(\hat{\theta}^*\right)\right]$ and Lemmas 1, 14 and 15, it can be shown that $\left[\hat{H}\left(\theta^+\right) - \hat{H}\left(\hat{\theta}^*\right)\right]$ and $\hat{C}\left(\hat{\theta}^*\right)$ converge to zero sufficiently fast that $\Delta_1 = o_p(1/\sqrt{N})$.

For Δ_2, Taylor expanding the second component:

$$\Delta_2 = \hat{H}\left(\theta^+\right)^{-1} \nabla\hat{C}\left(\hat{\theta}^* - \theta_0\right),$$

where $\nabla\hat{C}$ is evaluated at an intermediate point. The first component is $O_p(1)$ from Lemma 1; the second component is $O_p(\frac{1}{\sqrt{N}h^3})$ from Lemma 1; and the third component is $O_p\left(h^4\right)$, hence we have $\Delta_2 = o_p\left(1/\sqrt{N}\right)$.

From the definition of $\hat{\theta}^o$ in (D11) and employing the result above:

$$\begin{aligned} \sqrt{N}\left(\hat{\theta}^o - \theta_0\right) &= \sqrt{N}\left(\hat{\theta}^o - \theta_0 + \Delta\right) + o_p(1) = -\hat{H}\left(\theta^+\right)^{-1} \sqrt{N}\left(\hat{A}^* - \hat{B}^o\right) + o_p(1), \\ \hat{A}^* &= \frac{1}{N} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[Y_i\left(d_1,d_2\right) - P_i\left(d_1,d_2;\theta_0\right)\right] \hat{\delta}_i^*\left(d_1,d_2;\theta_0\right) \tau_{iv} \\ \hat{B}^o &= \frac{1}{N} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[\hat{P}_i^o\left(d_1,d_2;\theta_0\right) - P_i\left(d_1,d_2;\theta_0\right)\right] \hat{\delta}_i^*\left(d_1,d_2;\theta_0\right) \tau_{iv}. \end{aligned}$$

From Lemma 12:

$$\hat{A}^* = \frac{1}{N} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[Y_i\left(d_1,d_2\right) - P_i\left(d_1,d_2;\theta_0\right)\right] \delta_i\left(d_1,d_2;\theta_0\right) \tau_{iv} + o_p(N^{-1/2})$$

where $\delta_i(d_1, d_2; \theta_0)$ is the probability limit of $\hat{\delta}_i^*(d_1, d_2; \theta_0)$. It can also be shown that:

$$\hat{B}^o = \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^o(d_1, d_2; \theta_0) - P_i(d_1, d_2; \theta_0)\right]\delta_i(d_1, d_2; \theta_0)\tau_{iv} + o_p(N^{-1/2}).$$

Lemma 14b shows that $\hat{B}^o$ is of order $o_p(1/\sqrt{N})$. The theorem now follows.

8.2 Intermediate Lemmas

This section provides three types of lemmas: 1) basic lemmas required by all estimators, 2) lemmas required to analyze the marginal effects estimator, and finally 3) lemmas relevant for the index estimator.

8.2.1 Basic Lemmas

With V_2 having conditional density $g_2(v_2|Y_2 = d_2)$ 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:

$$\mathcal{V}_{2N} = \{v_2 : a_2(d_2) + h_m^{1-\varepsilon} < v_2 < b_2(d_2) - h_m^{1-\varepsilon}\} \tag{6}$$

$$\mathcal{V}_N = \{(v_1, v_2) : a_k(d_k) + h_c^{1-\varepsilon} < v_k < b_k(d_k) - h_c^{1-\varepsilon}\}. \tag{7}$$

We begin with two basic lemmas on uniform and pointwise convergence rates. As the proofs of these lemmas are standard in the literature, they are not provided here but are available upon request.

Lemma 1 (Uniform 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 (D1) with respective probability limits f_2 and f. Then, for θ in a compact set, $t_2 \epsilon \mathcal{V}_{2N}$ as defined in 6, $t \epsilon \mathcal{V}_N$ as defined in 7, the following rates hold for $p = 0, 1, 2$:

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

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

Lemma 2 (Pointwise Convergence): Using the same notation as in Lemma 1:

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

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

8.2.2 Marginal Effects Lemmas

Lemma 3: Under (A4,A6,A7), with $\zeta_0(\bar{v}) \equiv \Pr(Y_1 = 1|V_1 = \bar{v})$ and $\gamma_N \equiv \frac{\sum_j \frac{1}{Nh}[Y_{1j} - \zeta_0(\bar{v})]Y_{2j}K[(\bar{v}-V_{1j})/h]S_j}{E(S|V_1=\bar{v})g(\bar{v})}$,

$$|E(\gamma_N)| = O(N^{-a_0}E(S)/E(S|\bar{v})).$$

Proof: With $P_2 = \Pr(Y_2 = 1|V_2)$ and $\mu_d(V_1, V_2) \equiv E[Y_1 - \zeta_0(\bar{v})|Y_2 = d, V_1, V_2]$, and γ_{1N} as the numerator of γ_N :

$$\begin{aligned} E(\gamma_{1N}) &= E(\frac{1}{h}\mu_1(V_1, V_2)K[(\bar{v} - V_1)/h]S)P_2 \\ &= \iint \mu_1(\bar{v} + hz, v_2)K(z)SP_2g(\bar{v} + hz, v_2)dzdv_2. \end{aligned}$$

Using a Taylor series expansion,

$$|E(\gamma_{1N})| \leq \left| \int \mu_1(\bar{v}, v_2)P_2Sg(\bar{v}, v_2)dv_2 \right| + |RES|.$$

Note that $\mu_1(\bar{v}, v_2)P_2 + \mu_0(\bar{v}, v_2)(1 - P_2) = E[Y_1 - \zeta_0(\bar{v})|V_1 = \bar{v}, V_2] = 0$, hence for the first term on the right-hand-side:

$$\begin{aligned} \left| \int \mu_1(\bar{v}, v_2)P_2Sg(\bar{v}, v_2)dv_2 \right| &= \left| \int \mu_0(\bar{v}, v_2)(1 - P_2)Sg(\bar{v}, v_2)dv_2 \right| \\ &\leq O\left(N^{-a_0} \int Sg(\bar{v}, v_2)dv_2\right) \\ &= O\left(N^{-a_0}g(\bar{v}) \int Sg(v_2|\bar{v})dv_2\right) \\ &= O\left(N^{-a_0}E(S|\bar{v})\right). \end{aligned}$$

The second term on the right-hand-side ($|RES|$) is a residual term from the Taylor series expansion, which is $O\left(h^2E(S)\right)$. Therefore, combining those two terms, the slowest rate would be $|E(\gamma_{1N})| = O(N^{-a_0}E(S))$ since $O(h^2) < O(N^{-a_0})$ and $O(E(S|\bar{v})) \leq O(E(S))$ from (A6).

Lemma 4: For γ_N defined in Lemma 3 and a_0 defined in (D4),

$$\frac{1}{\sqrt{Var(\gamma_N)}} = O\left(\frac{\sqrt{Nh}E(S|\bar{v})}{\sqrt{E(S^2|\bar{v})}}\right).$$

Proof: For a_0 set as in (D4), $\frac{(\gamma_N - E(\gamma_N))^2}{Var(\gamma_N)} \to 0$; hence

$$\begin{aligned} Var(\gamma_N) &= O\left(\frac{E([Y_1 - \zeta_0(\bar{v})]^2 Y_2 K^2[(\bar{v} - V_1)/h] S^2)}{Nh^2 (E(S|V_1 = \bar{v}))^2 g^2(\bar{v})}\right) \\ &= O\left(\frac{E(K^2[(\bar{v} - V_1)/h] S^2)}{Nh^2 (E(S|V_1 = \bar{v}))^2}\right). \end{aligned}$$

Letting $z = (V_1 - \bar{v})/h$, the result follows from a Taylor series expansion about $h = 0$.

To obtain the convergence rate of $\hat{a}$ to a_0, Lemma 5 below shows that the moment condition for them are close.

Lemma 5: Let

$$\begin{aligned} c_N(a) &\equiv N^{-2(a-.4)+\varepsilon-\eta} \\ M_1(a) &\equiv E(S)^2 \\ M_2(a) &\equiv E(S^2|\bar{v}) \end{aligned}$$

and recall (D4), then:

$$c_N(\hat{a}) \left(\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} - \frac{M_1(\hat{a})}{M_2(\hat{a})} \right) = O_p\left(N^{-\delta}\right) \text{ where } \delta > 0.$$

Proof: To prove the result, we need to show that:

$$\begin{aligned} c_N(\hat{a}) \left(\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} - \frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{M_2(\hat{a})} \right) &= O_p\left(N^{-\delta}\right) \\ \text{and } c_N(\hat{a}) \left(\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{M_2(\hat{a})} - \frac{M_1(\hat{a})}{M_2(\hat{a})} \right) &= O_p\left(N^{-\delta}\right). \end{aligned}$$

Here, we provide the proof for the first equation, as the proof of the second is very similar.

Employing the definition of $\hat{a}$, we have $c_N(\hat{a}) \frac{\left[\hat{E}_2(\hat{S})\right]^2}{\hat{E}_2(\hat{S}^2|\bar{v})} = O(1)$. Hence the left-hand-side

has order:

$$\begin{aligned}
& \frac{M_2(\hat{a}) - \hat{E}_2\left(\hat{S}^2|\bar{v}\right)}{M_2(\hat{a})} \\
= & \frac{\left[\hat{E}_2\left(S^2\left[\hat{a},P\right]|\bar{v}\right) - \hat{E}_2\left(\hat{S}^2|\bar{v}\right)\right]}{M_2(\hat{a})} + \frac{\left[M_2(\hat{a}) - \hat{E}_2\left(S^2\left[\hat{a},P\right]|\bar{v}\right)\right]}{M_2(\hat{a})} \\
\equiv & A + B.
\end{aligned}$$

Beginning with term A, from a Taylor series expansion:

$$\begin{aligned}
S^2\left(\hat{a},\hat{P}_a\right) - S^2(\hat{a},P) &= \sum_{k=1}^{m-1}\left[S^2\right]'^{(k)}(\hat{a},P)\left[\frac{\hat{P}_a - P}{1-P}\right]^k/k! \\
&\quad + \left[S^2\right]'^{(m)}\left(\hat{a},P^+\right)\left[\frac{\hat{P}_a - P}{1-P^+}\right]^m/m!
\end{aligned}$$

where $\left[S^2\right]'^{(m)}$ is the mth derivative of the function S^2 w.r.t x. Hence we have to show both of the following terms vanish in probability

$$\begin{aligned}
A_k &\equiv \left|\sum_{i=1}^{N}\frac{1}{N}\left[S^2\right]'^{(k)}(\hat{a},P)\left[\frac{\hat{P}_a - P}{1-P}\right]^k K_2\left(\frac{\bar{v}-v_i}{h_2}\right)/M_2(\hat{a})h_2\right| \\
A_m &\equiv \left|\sum_{i=1}^{N}\frac{1}{N}\left[S^2\right]'^{(m)}\left(\hat{a},P^+\right)\left[\frac{\hat{P}_a - P}{1-P^+}\right]^m K_2\left(\frac{\bar{v}-v_i}{h_2}\right)/M_2(\hat{a})h_2\right|.
\end{aligned}$$

For A_k, setting $k=1$ for expositional purposes, the term will be bounded above by

$$2\sup\left|\left(\frac{\hat{P}_{ai}-P_i}{1-P_i}\right)s^1(\hat{a},P)\right|\sum_{i=1}^{N}\frac{1}{N}S(\hat{a},P)K_2\left(\frac{\bar{v}-v_i}{h_2}\right)/M_2(\hat{a})h_2$$

where s^m is the m^{th} derivative of S w.r.t x. The $\sup\left|\left(\frac{\hat{P}_{ai}-P_i}{1-P_i}\right)s^1(\hat{a},P)\right|$ vanishes in probability because $\hat{P}_a$ is based on higher order kernels and converges to P faster than $N^{-\hat{a}}$, which is the order of the denominator $1-P$.[7] Turning our attention to the second part of

[7] Note that $\hat{a} < .4$, and $s^1(\hat{a},P))$ restricts $1-P$ to a middle region $R2$ where:

$$\frac{N^{-\hat{a}}}{\exp(b)} < 1 - P < N^{-\hat{a}}.$$

the above expression, it is bounded above by:

$$\sup_a \left| \sum_{i=1}^{N} \frac{1}{N} S(a,P) K_2 \left(\frac{\bar{v} - v_i}{h_2} \right) / h_2 - E(S(a,P)|\bar{v}) \right| / M_2(\hat{a}) + \frac{E(S(a,P)|\bar{v})}{M_2(\hat{a})}.$$

For the first term, from uniform convergence, the numerator is converging to zero at a rate arbitrarily close to $N^{-.4}$. For the denominator, referring to (D3), notice that S^2 function is bounded below by an indicator function set to be zero when x is in either $R1$ or $R2$, and one when x is in $R3$. Hence $M_2(a)$ is bounded below by the conditional expectation of that indicator function, which is a probability that is of order N^{-a} provided that the tail of $v_2|\bar{v}$ is fatter than the error tail (A5). Therefore, $M_2(\hat{a}) \geqslant O(N^{-\hat{a}}) \geqslant O(N^{-\left(.4-\frac{\varepsilon-\eta}{2}\right)})$ (see D4). Hence it suffices to show that $\sup_a E(S(a,P)|\bar{v})/M_2(a) = O(1)$. Referring to (D3), notice that

$$\frac{E(S(a,P)|\bar{v})}{M_2(a)} = \frac{c_1 \Pr(R2, \tau = 1|\bar{v}) + \Pr(R3, \tau = 1|\bar{v})}{c_2 \Pr(R2, \tau = 1|\bar{v}) + \Pr(R3, \tau = 1|\bar{v})}$$

where

$$\begin{aligned} c_1 &\equiv E\left[1 - \exp \frac{-x^k}{b^k - x^k} | R2, \tau = 1, \bar{v}\right] \\ c_2 &\equiv E\left[\left(1 - \exp \frac{-x^k}{b^k - x^k}\right)^2 | R2, \tau = 1, \bar{v}\right]. \end{aligned}$$

The above ratio converges to some constant irrespective of which of the regional probabilities converges faster to zero.

For the remainder term A_m, it vanishes for m sufficiently large. For term B, the argument is very similar to that above.

Lemma 6: Referring to Lemma 5, define

$$\begin{gathered} z(a) \equiv N^{-a} \\ R\left(z\left(a\right)\right) \equiv \tfrac{M_1(a)}{M_2(a)} \\ z_0 \equiv z(a_0), \hat{z} \equiv z\left(\hat{a}\right), z^+ \equiv z\left(a^+\right) \end{gathered}$$

then, for $\delta > 0 : |\hat{a} - a_0| = o_p\left(N^{-\delta}\right)$.

Proof: Lemma 5 shows that:

$$\begin{aligned} c_N(\hat{a})\frac{\left[\hat{E}_2\left(\hat{S}\right)\right]^2}{\hat{E}_2\left(\hat{S}^2|\bar{v}\right)} &= c_N(\hat{a})R(z(\hat{a})) + O_p\left(N^{-\delta}\right) \\ &= c_N(a_0)R(z(a_0)) + O_p\left(N^{-\delta}\right) - \\ &\quad Ln(N)c_N\left(z^+\right)\left[2R\left(z^+\right) + z^+R'\left(z^+\right)\right][\hat{a}-a_0]. \end{aligned}$$

Therefore, since $R = \frac{[E(S)]^2}{E(S^2|\bar{v})}$ is increasing as in (D4), $z^+R'(z^+) > 0$; hence

$$[\hat{a}-a_0] = O_p\left(N^{-\delta}/Ln(N)c_N\left(z^+\right)R\left(z^+\right)\right).$$

Suppose $\hat{a} < a_0$ (the argument when $\hat{a} \geqslant a_0$ is the same), then $\hat{a} < a^+ < a_0$, and $z_0 < z^+ < \hat{z}$. Since R is increasing, $R(z_0) < R(z^+) < R(\hat{z})$. Therefore, $c_N(z_0)R(z_0) < c_N(z^+)R(z^+) < c_N(\hat{z})R(\hat{z})$. The proof now follows from Lemma 5.

Lemma 7.(Expectations of Kernel Products): Let $\{\varepsilon_{1j},\varepsilon_{2j},\varepsilon_{3j}\}$ be i.i.d. over j with properties:

$$\begin{aligned} a) &: E(\varepsilon_{\gamma j}) = O(h^{2p}) \\ b) &: E\left[\varepsilon_{\gamma j,}^{\rho}\right] = \frac{1}{h^{\varrho-1}}, \rho > 1. \end{aligned}$$

Set $h^{4p} = O(\frac{1}{Mh})$ and denote $\bar{\varepsilon}_\gamma = \frac{1}{M}\sum_{j=1}^{M}\varepsilon_{1j}$, $\gamma = 1,2,3$,then

$$\begin{aligned} E\left\{[\bar{\varepsilon}_1]^r[\bar{\varepsilon}_2]^s[\bar{\varepsilon}_3]^t\right\} &= O\left(h^{2p4r/4}\right)O\left(h^{2p4s/4}\right)O\left(h^{2p2t/2}\right) \\ &= O\left(h^{2p(r+s+t)}\right). \end{aligned}$$

Proof: From the Cauchy–Schwartz inequality

$$E\left\{[\bar{\varepsilon}_1]^r[\bar{\varepsilon}_2]^s[\bar{\varepsilon}_3]^t\right\} \leq \left\{E[\bar{\varepsilon}_1]^{4r}\right\}^{1/4}\left\{E[\bar{\varepsilon}_2]^{4s}\right\}^{1/4}\left\{E[\bar{\varepsilon}_3]^{2t}\right\}^{1/2}.$$

It suffices to order one of the three terms, hence we can study a general term: $E[\bar{\varepsilon}]^q$. This general term has q types of terms, with kth $(k = 1,...,q)$ type:

$$\begin{aligned} &\frac{1}{M^{q-k}}\underbrace{\sum\cdots\sum}_{k}\frac{1}{M^k}\varepsilon_{j_1}^{i_1}\cdots\varepsilon_{j_k}^{i_k} \\ \text{where } i_1+...+i_k &= q \text{ and } j_1 \neq \cdots \neq j_k. \end{aligned}$$

From i.i.d. property of the $\varepsilon's$, the expectation of this term is given as:

$$\frac{1}{M^{q-k}} E\left[\varepsilon_{j_1}^{i_1}\right] \cdots E\left[\varepsilon_{j_k}^{i_k}\right].$$

Suppose we study a term $E\left[\varepsilon_{j_t}^{i_t}\right]$, where $1 \leqslant t \leqslant k$. There are two types of expectations: single power of ε and multiple powers of ε. For the single power case, from property (a):

$$E\left[\varepsilon_{j_t}^{i_t}\right] = O(h^{2p}) \text{ for } i_t = 1.$$

For the multiple power case, from property (b):

$$E\left[\varepsilon_{j_t}^{i_t}\right] = O\left(\left[\frac{1}{h}\right]^{(i_t-1)}\right) \text{ for } i_t > 1$$

There are different combinations of $i_1, \cdots, i_k$ for a given q and k. To order $\frac{1}{M^{q-k}} E\left[\varepsilon_{j_1}^{i_1}\right] \cdot \cdot \cdot E\left[\varepsilon_{j_k}^{i_k}\right]$, we next need to find the combination which yields the slowest convergence rate. One observation we make here is that the slowest term is the one with the least number of single power $\varepsilon's$.

When $k \leqslant \frac{q}{2}$, the slowest term would have no single power of ε in it (see below for an example). Therefore, from property (b) the convergence rate will be:

$$O\left((\frac{1}{Mh})^{q-k}\right).$$

When $k > \frac{q}{2}$, the slowest term would include at least one single power ε in it, hence from property (a) and (b) the rate will be:

$$O\left(\ (h^{2p})^{2k-q}\right) O\left((\frac{1}{Mh})^{q-k}\right).$$

Suppose q is an even number (the odd number case is very similar), for example $q = 6$. If we denote the type by the powers of the elements, for example 1122 would mean the

$\varepsilon^1_{j_1}\varepsilon^1_{j_2}\varepsilon^2_{j_3}\varepsilon^2_{j_4}$ term, then we have the following table:

$$\begin{bmatrix} k & Slowest\ Type & Rate \\ 1 & 6 & O\left[(\frac{1}{Mh})^5\right] \\ 2 & 33 & O\left[(\frac{1}{Mh})^4\right] \\ 3 & 222 & O\left[(\frac{1}{Mh})^3\right] \\ 4 & 1122 & O\left[(h^{2p})^2(\frac{1}{Mh})^2\right] \\ 5 & 11112 & O\left[(h^{2p})^4(\frac{1}{Mh})\right] \\ 6 & 111111 & O\left[(h^{2p})^6\right] \end{bmatrix}$$

For $k = \frac{q}{2}$, the slowest term would be the one with k squared terms

$$\frac{1}{M^{q-k}} E\left[\varepsilon^2_{j_1}\right] \cdots E\left[\varepsilon^2_{j_k}\right]$$

(examples of faster terms: 114 and 123 types) hence the rate would be

$$O\left(\frac{1}{M^{q-k}}(\frac{1}{h})^k\right) = O\left((\frac{1}{Mh})^{q-k}\right).$$

It can be shown that this same expression holds for all smaller k. For $k = \frac{q}{2}+1$, the slowest term would have two single power $\varepsilon' s$ and the rest are squared terms, e.g.

$$\frac{1}{M^{q-k}} E\left[\varepsilon_{j_1}\right] E\left[\varepsilon_{j_2}\right] E\left[\varepsilon^2_{j_3}\right] \cdots E\left[\varepsilon^2_{j_k}\right]$$

hence the rate would be

$$\frac{1}{M^{q-k}} O\left((h^{2p})^2(\frac{1}{h})^{k-2}\right) = O\left((h^{2p})^{2k-q}(\frac{1}{Mh})^{q-k}\right).$$

This same expression holds for all larger k.

We now need to find the k^{th} term with the slowest convergence rate. Set h optimally, i.e. $h^{4p} = O(\frac{1}{Mh})$ and substitute it in each term above, we have

$$\begin{array}{ll} O\left(\ (h^{4p})^{q-k}\right) = O\left(\ (h^{2p})^{2q-2k}\right) & \text{when } k \leqslant \frac{q}{2}, \\ O\left(\ (h^{2p})^{2k-q}(h^{4p})^{q-k}\right) = O(h^{2pq}) & \text{when } k > \frac{q}{2}. \end{array}$$

hence the slowest convergence rate is $O(h^{2pq})$. The lemma follows.

Lemma 8:. Define

$$\begin{aligned}\hat{E}\left(Y_2 S\left(a,P\right)|\bar{v}\right) &\equiv \sum_j \frac{1}{Nh} Y_{2j} K\left[\left(\bar{v}-V_{1j}\right)/h\right] S(a,P_j) \\ M_3(a) &\equiv E(\sum_j \frac{1}{Nh} Y_{2j} K\left[\left(\bar{v}-V_{1j}\right)/h\right] S(a,P_j)).\end{aligned}$$

Then

$$\frac{M_3(a_0)-\hat{E}\left(Y_2 S\left(\hat{a},\hat{P}\right)|\bar{v}\right)}{M_3(a_0)} \xrightarrow{p} 0.$$

Proof: The above term can be decomposed into the following as in Lemma 5:

$$\frac{\left[\hat{E}\left(Y_2 S\left[a_0,P\right]|\bar{v}\right)-\hat{E}\left(Y_2 S\left[\hat{a},\hat{P}\right]|\bar{v}\right)\right]}{M_3(a_0)}+\frac{\left[M_3(a_0)-\hat{E}\left(Y_2 S\left[a_0,P\right]|\bar{v}\right)\right]}{M_3(a_0)}.$$

The above two terms are similar to terms A and B in Lemma 5. Employing a Taylor series argument and utilizing the result from Lemma 6, it can be shown that the first term goes to zero in probability. The second term only requires pointwise convergence instead of the uniform convergence arguments in Lemma 5.

Lemma 9: For notational simplicity, we denote $C_N = C_N\left(\bar{v}\right)$ as in Theorem 1. Letting $\vartheta_j \equiv \left(Y_{1j}-\zeta_0\left(\bar{v}\right)\right) Y_{2j} K\left[\left(\bar{v}-V_{1j}\right)/h\right]/h$,

$$C_N\left[\hat{\zeta}\left(\bar{v}\right)-\zeta_0\left(\bar{v}\right)\right] = C_N \frac{\sum_j \frac{1}{N}\vartheta_j S(a_0,P_j)}{M_3(a_0)}+o_p(1).$$

Proof: From Lemma 8:

$$\begin{aligned}C_N\left[\hat{\zeta}\left(\bar{v}\right)-\zeta_0\left(\bar{v}\right)\right] &= C_N \frac{\sum_j \frac{1}{N}\vartheta_j S\left(\hat{a},\hat{P}_j\right)}{\sum_j \frac{1}{Nh} Y_{2j} K\left[\left(\bar{v}-V_{1j}\right)/h\right] S\left(\hat{a},\hat{P}_j\right)} \\ &= C_N \frac{\sum_j \frac{1}{N}\vartheta_j S\left(\hat{a},\hat{P}_j\right)}{M_3(a_0)}+o_p(1).\end{aligned}$$

It remains to be shown that

$$C_N \frac{\sum_j \frac{1}{N}\vartheta_j \left[S\left(\hat{a},\hat{P}_j\right)-S(a,P_j)\right]}{M_3(a_0)} \xrightarrow{p} 0.$$

With s^m as the m^{th} derivative of S w.r.t x, for $P_j^+ \in \left[\hat{P}_j,P_j\right]$, $a^+ \in \left[\hat{a},a_0\right]$, a Taylor

expansion provides:

$$\begin{aligned} S\left(\hat{a},\hat{P}_j\right) - S\left(a_0,P_j\right) &= \sum_{k=1}^{K}\sum_{l=1}^{L}\frac{1}{k!l!}T_{kl} \\ T_{kl} &\equiv s^{k+l}\left(\bar{a},\bar{P}_j\right)\left[Ln(N)\right]^k\left[\hat{a}-a_0\right]^k\left[\frac{\hat{P}_j-P_j}{1-\bar{P}_j}\right]^l \end{aligned}$$

where

$$\bar{a}\equiv\begin{cases} a_0 & k<K \\ a^+ & k=K \end{cases} \quad \bar{P}_j\equiv\begin{cases} P_j & l<L \\ P_j^+ & l=L \end{cases}.$$

Substituting the Taylor series expansion, and noting that $\left[Ln(N)\right]^k\left[\hat{a}-a_0\right]^k$ is converging to zero, we now need to show that terms of the following form converge in probability to 0:

$$C_N\frac{\sum_j\frac{1}{N}\vartheta_j s^{k+l}\left(\bar{a},\bar{P}_j\right)\left[\frac{\hat{P}_j-P_j}{1-\bar{P}_j}\right]^l}{M_3(a_0)}.$$

To show that, we study the expectation of the square of the above term. Squaring the above term yields two types of elements:

$$\frac{1}{N}\frac{C_N^2}{NM_3(a_0)^2}\sum_j\left\{\vartheta_j s^{k+l}\left(\bar{a},\bar{P}_j\right)\left[\frac{\hat{P}_j-P_j}{1-\bar{P}_j}\right]^l\right\}^2 \; and \tag{8}$$

$$\frac{C_N^2}{N^2M_3(a_0)^2}\sum_j\sum_j\left\{\vartheta_i s^{k+l}\left(\bar{a},\bar{P}_i\right)\left[\frac{\hat{P}_i-P_i}{1-\bar{P}_i}\right]^l\right\}\left\{\vartheta_j s^{k+l}\left(\bar{a},\bar{P}_j\right)\left[\frac{\hat{P}_j-P_j}{1-\bar{P}_j}\right]^l\right\}. \tag{9}$$

For the first type, write:

$$\left[\frac{\hat{P}_j-P_j}{1-\bar{P}_j}\right]^l=\left[\frac{\bar{\varepsilon}_{1j}}{1-\bar{P}_j}\right]^l\left[\frac{g_j}{\hat{g}_j}\right]^l \text{ where } \bar{\varepsilon}_{1j}\equiv\left(\hat{P}_j-P_j\right)\frac{\hat{g}_j}{g_j}.$$

From a Taylor series expansion with $\bar{\varepsilon}_{2j}\equiv\left[\hat{g}_j-g_j\right]$:

$$\begin{aligned} \left[\frac{g_j}{\hat{g}_j}\right]^l &= \sum_{t=0}^{m}T_{tj} \\ T_{tj} &= (-1)^t\frac{1}{t!}\frac{(l+t-1)!}{(l-1)!}\left[\frac{\bar{\varepsilon}_{2j}}{g_j}\right]^t\left((\frac{\hat{g}_j}{g_j})^+\right)^{(-l-m)I\{t=m\}} \end{aligned}$$

Substituting a typical term T_{tj} into (8) we must show that the following expression converges

in probability to 0:

$$\frac{1}{N}\frac{C_N^2}{NM_3(a_0)^2}\sum_j\left\{\vartheta_j s^{k+l}\left(\bar{a},\bar{P}_j\right)\left[\frac{\bar{\varepsilon}_{1j}}{1-\bar{P}_j}\right]^l T_{tj}\right\}^2.$$

For non-remainder terms in the above Taylor series expansion ($k<K, l<L, t<m$) , the expectation of the above expression has the form:

$$\begin{aligned}
&\frac{C_N^2}{NM_3(a_0)^2}E\frac{1}{N}\sum_j\vartheta_j^2\left[s^{k+l}\left(a_0,P_j\right)\right]^2\left[\frac{\bar{\varepsilon}_{1j}}{1-P_j}\right]^{2l}\left[\frac{\bar{\varepsilon}_{2j}}{g_j}\right]^{2t}\\
=\;&\frac{C_N^2}{NM_3(a_0)^2}E\left\{E\left[\vartheta_j^2|X_j\right]\left[s^{k+l}\left(a_0,P_j\right)\right]^2\left[\frac{\bar{\varepsilon}_{1j}}{1-P_j}\right]^{2l}\left[\frac{\bar{\varepsilon}_{2j}}{g_j}\right]^{2t}\right\}\\
=\;&\frac{C_N^2}{NM_3(a_0)^2}E\left\{E\left[\vartheta_j^2|X_j\right]\left[s^{k+l}\left(a_0,P_j\right)\right]^2E\left(\left[\frac{\bar{\varepsilon}_{1j}}{1-P_j}\right]^{2l}\left[\frac{\bar{\varepsilon}_{2j}}{g_j}\right]^{2t}|V_j\right)\right\},
\end{aligned}$$

where $E\left[\vartheta_j^2|X_j\right]$ only depends on the indices V_j. From Lemma 7, the expectation conditioned on the indices is $o(1)$. Therefore, since C_N^2/N is converging to 0, the result follows as:

$$\frac{1}{M_3(a_0)^2}E\left\{E\left[\vartheta_j^2|X_j\right]\left[s^{k+l}\left(a_0,P_j\right)\right]^2\right\}=O(1).$$

Turning to the expectation of the cross-product terms (9), define $\hat{P}_j[i]$ by removing from $\hat{P}_j$ its dependence on Y_{2i} Similarly, define $\hat{P}_i[j]$. Notice that $\hat{P}_j[i]$ and do not depend on Y_{2i} and Y_{2j}. Finally, denote:

$$\hat{P}_i\equiv\hat{P}_i[j]+\lambda_j(1-P_i);\quad\hat{P}_j\equiv\hat{P}_j[i]+\lambda_i(1-P_j),$$

then the non-remainder terms in (9) have the following form:

$$\frac{C_N^2}{M_3(a_0)^2}E\left\{\vartheta_i s^{k+l}\left(a_0,P_i\right)\left[\frac{\left(\hat{P}_i[j]-P_i\right)}{1-P_i}+\lambda_j\right]^l\right\}\left\{\vartheta_j s^{k+l}\left(a_0,P_j\right)\left[\frac{\left(\hat{P}_j[i]-P_j\right)}{1-P_j}+\lambda_i\right]^l\right\}.$$

Performing the binomial expansion on $\left[\frac{\left(\hat{P}_i[j]-P_i\right)}{1-P_i}+\lambda_j\right]^l$ and $\left[\frac{\left(\hat{P}_j[i]-P_j\right)}{1-P_j}+\lambda_i\right]^l$, the slowest

converging term has the following form:

$$\begin{aligned}
&\frac{C_N^2}{M_3(a_0)^2}E\left\{\vartheta_i s^{k+l}(a_0,P_i)\,\vartheta_j s^{k+l}(a_0,P_j)\,\lambda_j\lambda_i\left(\frac{\hat{P}_i[j]-P_i}{1-P_i}\right)^{l-1}\left(\frac{\hat{P}_j[i]-P_j}{1-P_j}\right)^{l-1}\right\}\\
=\;&\frac{C_N^2}{M_3(a_0)^2}E\left\{\vartheta_i s^{k+l}(a_0,P_i)\,\vartheta_j s^{k+l}(a_0,P_j)\,\lambda_j\lambda_i E\left[\left(\frac{\hat{P}_i[j]-P_i}{1-P_i}\right)^{l-1}\left(\frac{\hat{P}_j[i]-P_j}{1-P_j}\right)^{l-1}|V_i,V_j\right]\right\}.
\end{aligned}$$

Applying Lemma 7 in a similar manner as above yields:

$$E\left[\left(\frac{\hat{P}_i[j]-P_i}{1-P_i}\right)^{l-1}\left(\frac{\hat{P}_j[i]-P_j}{1-P_j}\right)^{l-1}|V_i,V_j\right]=\frac{B(V_i,V_j)}{[(1-P_i)(1-P_j)]^{l-1}}o(N^{-2a(l-1)})$$

where $B(V_i,V_j)$ is a bounded function. Since $\lambda_j\lambda_i=\frac{1}{(1-P_i)(1-P_j)h_2^2N^2}Y_iY_jK_2\left([V_{2i}-V_{2j}]/h_2\right)^2$, the absolute value of (9) is bounded above by

$$\frac{C_N^2}{N^2h_2^2M_3(a_0)^2}E\left\{\left|\vartheta_i s^{k+l}(a_0,P_i)\,\vartheta_j s^{k+l}(a_0,P_j)\right|Y_iY_jK_2\left([V_{2i}-V_{2j}]/h_2\right)^2\left|\frac{B(V_i,V_j)}{[(1-P_i)(1-P_j)]^l}\right|o(N^{-2a(l-1)})\right\}.$$

Notice that the s derivatives restricts us to the middle region where $1-P>N^{-a}/\exp(b)$ and hence $\left|\frac{B(V_i,V_j)}{[(1-P_i)(1-P_j)]^l}\right|=O(N^{2al})$. Since $h_2=N^{-.1}$, we have $C_N^2/h_2N=o(1)$ and $o(N^{2a})/h_2N=o(1)$. Hence it suffices to show $E\left|\vartheta_i s^{k+l}(a_0,P_i)\,\vartheta_j s^{k+l}(a_0,P_j)\right|/M_3(a_0)^2=O(1)$.

Recall that $\vartheta_j\equiv(Y_{1j}-\zeta_0(\bar{v}))\,Y_{2j}K[(\bar{v}-V_{1j})/h]/h$. Further s^m is the m^{th} derivative of S w.r.t x, and it is zero except in region $R2$:

$$\frac{E\left|\vartheta_i s^{k+l}(a_0,P_i)\,\vartheta_j s^{k+l}(a_0,P_j)\right|}{M_3(a_0)^2}=\left(\frac{c_1\Pr(R2,\tau=1|\bar{v})}{c_2\Pr(R2,\tau=1|\bar{v})+c_3\Pr(R3,\tau=1|\bar{v})}\right)^2$$

where

$$\begin{aligned}
c_1 &\equiv \left\{E\left[\vartheta s^{k+l}(a_0,P)\,|R2,\tau=1,\bar{v}\right]\right\}\\
c_2 &\equiv E\left[\left(1-\exp\frac{-x^k}{b^k-x^k}\right)\left(\frac{1}{h}Y_2K\left[\frac{\bar{v}-V_1}{h}\right]\right)|R2,\tau=1,\bar{v}\right]\\
c_3 &= E\left[\left(\frac{1}{h}Y_2K\left[\frac{\bar{v}-V_1}{h}\right]\right)|R3,\tau=1,\bar{v}\right].
\end{aligned}$$

Then, similar to Lemma 5, it follows that

$$E\left|\vartheta_i s^{k+l}(a_0,P_i)\,\vartheta_j s^{k+l}(a_0,P_j)\right|/M_3(a_0)^2=O(1).$$

8.2.3 Index Lemmas

The next lemma proves that the estimated second-stage objective function $\hat{L}^*(\theta)$ is uniformly close to $L^*(\theta) \equiv \frac{1}{N}\sum_i \sum_{d_1,d_2} Y_i(d_1,d_2)\, Ln\left[P_i^*(d_1,d_2;\theta)\right]$.

Lemma 10: With $D \equiv \frac{1}{N}\sum_i \sum_{d_1,d_2} D_i$,where $D_i \equiv Y_i(d_1,d_2)\, Ln\left[\hat{P}_i^*(d_1,d_2;\theta)/P_i^*(d_1,d_2;\theta)\right]$,

$$\sup_\theta |D| = o_p(1)$$

Proof: We prove this result when indices are restricted to be smoothly in $\mathcal{V}_N$ and its complement. Referring (D2), define a smoothed indicator restricting v_i to $\mathcal{V}_N$ in (7) as:

$$\begin{aligned} l(v_i) &\equiv \prod_k \tau[a_k(d_k) + h_{ck}^{1-\varepsilon}, v_{ki}]\tau[v_{ki}, b_k(d_k) - h_{ck}^{1-\varepsilon}], \\ D &= \frac{1}{N}\sum_i \sum_{d_1,d_2} D_i l(v_i) + \frac{1}{N}\sum_i \sum_{d_1,d_2} D_i(1-l(v_i)). \end{aligned}$$

Taylor expanding the first term on the right, Lemma 1 proves that it converges to zero in probability. For the second term,

$$\sup_\theta |\frac{1}{N}\sum_i \sum_{d_1,d_2} D_i(1-l(v_i))| \;\le\; \sup_{i,\theta}\left|\sum_{d_1,d_2} Y_i(d_1,d_2)\, Ln\left[\frac{\hat{P}_i^*(d_1,d_2;\theta)}{P_i^*(d_1,d_2;\theta)}\right]\right| \sup_\theta \frac{1}{N}\sum_i [1-l(v_i)].$$

It can be shown that $\inf P_i^*(d_1,d_2;\theta)$ is bounded away from 0 and $\inf \hat{P}_i^*(d_1,d_2;\theta)$ converges to a term bounded away from zero. Therefore, the first term above is finite. The second term converges in probability to zero.

The next lemma proves that $L^*(\theta)$ which is the probability limit of $\hat{L}^*(\theta)$(defined in D10) is uniformly close to $L(\theta) \equiv \frac{1}{N}\sum_i \sum_{d_1,d_2} Y_i(d_1,d_2)\, Ln\left[P_i(d_1,d_2;\theta)\right]$. Therefore, we may ignore the probability adjustments $\Delta' s$ in the adjusted likelihood, L^*.

Lemma 11: For θ in a compact set:

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

Proof: The proof is identical to the argument in Lemma 10 and follows directly by establishing this result on both sets away from support boundaries and "low probability" sets near the boundaries.

The following lemma shows that the trimming and the $\delta' s$ can be taken as known.

Lemma 12: For $\tau = \tau_v$ or τ_x, referring to (D10), with

$$\begin{aligned}\hat{A}^* &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\hat{\delta}_i^*\left(d_1,d_2;\theta_0\right)\tau \\ A &\equiv \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\delta_i\left(d_1,d_2;\theta_0\right)\tau\end{aligned}$$

then

$$\hat{A}^* - A = o_p(N^{-1/2}).$$

Proof: Klein and Shen (2010) establish this result in a semiparametric least squares context for single index models. The argument extends to double index likelihood-based models.

Using Lemma 12, Lemma 13 provides a useful convergence rate for the initial estimator in (D10).

Lemma 13: For $\hat{\theta}$ defined in (D10) and with $h = O(N^{-r}), r = \frac{1}{8+\xi}$:

$$\left(\hat{\theta}-\theta_0\right) = O_p(h^2).$$

Proof: From a Taylor series expansion:

$$\begin{aligned}\left(\hat{\theta}-\theta_0\right) &= -\hat{H}\left(\theta^+\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-\hat{P}_i\left(d_1,d_2;\theta_0\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_0\right)\tau_{ix} = -\hat{H}\left(\theta^+\right)^{-1}\left[\hat{A}-\hat{B}\right], \\ \hat{A} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i\left(d_1,d_2\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_0\right)\tau_{ix}; \\ \hat{B} &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i\left(d_1,d_2;\theta_0\right)-P_i\left(d_1,d_2;\theta_0\right)\right]\hat{\delta}_i\left(d_1,d_2;\theta_0\right)\tau_{ix}.\end{aligned}$$

Employing the same argument as in Lemma 12, $\hat{A} - A = o_p(N^{-1/2})$, and since $A = O_p(N^{-1/2})$ we have $\hat{A} = O_p(N^{-1/2})$. From Lemma 2, $\hat{B} = O_p(h^2)$, which completes the argument.

To obtain a convergence rate for the second-stage estimator and to analyze the final bias-adjusted estimator, Lemma 14 shows that the gradient component which is responsible for the bias in the estimator vanishes in probability.

Lemma 14:

$$a) \quad : \quad B^* = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[\hat{P}_i^* \left(d_1, d_2; \theta_0 \right) - P_i \left(d_1, d_2; \theta_0 \right) \right] \delta_i \left(d_1, d_2; \theta_0 \right) \tau_{iv} = o_p(1)$$

$$b) \quad : \quad B^o = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[\hat{P}_i^o \left(d_1, d_2; \theta_0 \right) - P_i \left(d_1, d_2; \theta_0 \right) \right] \delta_i \left(d_1, d_2; \theta_0 \right) \tau_{iv} = o_p(1).$$

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

$$B^* = B + o_p(1), B = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[\hat{P}_i \left(d_1, d_2; \theta_0 \right) - P_i \left(d_1, d_2; \theta_0 \right) \right] \delta_i \left(d_1, d_2; \theta_0 \right) \tau_{iv}.$$

Denote:

$$\begin{aligned} \hat{P}_m &= \hat{P} \left(Y_{2i} = d_2 | V_{2i} = t_2 \right), \; P_m \equiv p \lim \hat{P}_m; \\ \hat{P}_c &= \hat{P} \left(Y_{1i} = d_1 | Y_{2i} = d_2, V_i = t \right), \; P_c \equiv p \lim \hat{P}_c. \end{aligned}$$

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:

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

For the second case (the first is similar and easier), we can rewrite B term as:

$$B = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{d_1,d_2} \left[(\hat{P}_m - P_m)(\hat{P}_c - P_c) + (\hat{P}_m - P_m) P_c + P_m (\hat{P}_c - P_c) \right] \delta_i \left(d_1, d_2; \theta_0 \right) \tau_{iv}.$$

For the first term in B, we may employ Cauchy's inequality and Lemma 2 to show that it vanishes in probability.

The difference between the second term in B and the following U-statistic converges in

probability to zero:

$$\begin{aligned} U &\equiv \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[(\frac{\hat{f}_2(t_2;d_2)}{\hat{g}_2(t_2;d_2)}-P_m)P_c\right]\left[\frac{\hat{g}_2(t_2;d_2)}{g_2(t_2;d_2)}\right]\delta_i(d_1,d_2;\theta_0)\tau_{iv} \\ &= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[(\hat{f}_2(t_2;d_2)-\hat{g}_2(t_2;d_2)P_m)P_c\right]\left[\frac{\delta_i(d_1,d_2;\theta_0)\tau_{iv}}{g_2(t_2;d_2)}\right]. \end{aligned}$$

Notice that the U-statistic vanishes in probability from standard projection arguments, hence the second term in B vanishes. The third term in B has the same structure as the second and therefore also vanishes in probability, which completes the proof for a). The proof for b) is very similar.

Lemma 15: Referring to (D10), for the second stage estimator:

$$\left|\hat{\theta}^*-\theta_0\right| = O_p\left(N^{-\frac{4}{8+\xi}}\right).$$

Proof: From Lemma 14, the initial estimator satisfies: $\left(\hat{\theta}-\theta_0\right) = O_p\left(N^{-2r}\right)$. For the estimator based on index trimming, from a standard Taylor series argument with τ_{iv} replacing τ_{ix} :

$$\begin{aligned} \left(\hat{\theta}^*-\theta_0\right) &= -\hat{H}^*\left(\theta^+\right)^{-1}\left[\hat{A}^*-\hat{B}^*\right], \\ \hat{A}^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[Y_i(d_1,d_2)-P_i(d_1,d_2;\theta_0)\right]\hat{\delta}^*(d_1,d_2;\theta_0)\tau_{iv}; \\ \hat{B}^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^*(d_1,d_2;\theta_0)-P_i(d_1,d_2;\theta_0)\right]\hat{\delta}^*(d_1,d_2;\theta_0)\tau_{iv}. \end{aligned}$$

Referring to Lemma 13, since $A = O_p(N^{-1/2})$, $\hat{A}^* = O_p(N^{-1/2})$.

For the $\hat{B}^*$-term, with $\Delta_{Bi} \equiv [\hat{\delta}^*(d_1,d_2;\theta_0)\tau_{iv} - \delta(d_1,d_2;\theta_0)\tau_{iv}]$:

$$\begin{aligned} \hat{B}^* &= B_1^*+\hat{B}_2^*, \\ B_1^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^*(d_1,d_2;\theta_0)-P_i(d_1,d_2;\theta_0)\right]\delta(d_1,d_2;\theta_0)\tau_{iv} \\ \hat{B}_2^* &= \frac{1}{N}\sum_{i=1}^{N}\sum_{d_1,d_2}\left[\hat{P}_i^*(d_1,d_2;\theta_0)-P_i(d_1,d_2;\theta_0)\right]\Delta_{Bi} \end{aligned}$$

By showing that $\hat{B}_1^*$ is close in probability to a centered U-statistic, Lemma 14, part a) proves that $B_1^* = o_p\left(N^{-1/2}\right)$. From Cauchy's inequality, the convergence rates in Lemma

2, and with window parameters $r = r^* = \frac{1}{8+\xi}$, it follows that $\hat{B}_2^* = O_p\left(N^{-\frac{4}{8+\xi}}\right), \xi > 0$. For these window choices, from the uniform rates in Lemma 1: $\hat{H}^*\left(\theta^+\right) = H_0 + o_p(1)$. It now follows that $\left|\hat{\theta}^* - \theta_0\right| = O_p\left(N^{-\frac{4}{8+\xi}}\right)$.