Intercept Estimation in Nonlinear Selection Models*

Wiji Arulampalam† Warwick University

Valentina Corradi ‡ Surrey University

Daniel Gutknecht§ Goethe University Frankfurt

April 13, 2020

Abstract The intercept in endogenous selection models is of fundamental importance for the evaluation of average treatment effects. While various intercept estimators for additive linear selection models exist, there are currently no estimators for nonlinear selection models. This paper introduces estimators for semiparametric nonlinear selection models, where the joint distribution of the error terms remains unspecified. We consider models where the intercept and slope parameters can be separately identified. If the selection equation satisfies an index restriction, the resulting estimator is based on a least squares criterion function with a nonparametric correction term. This estimator is asymptotically normal at a univariate nonparametric rate, even in cases of irregular identification. In a second step, we relax the index restriction in the selection equation and adopt a nonparametric propensity score specification. We suggest a local nonlinear least squares estimator, which only uses observations close but not too close to the boundary. Such an estimator exhibits a slower convergence rate than the first one, but is robust against mis-specification of the propensity score. The empirical illustration studies the effect of private health insurance on health care utilization using count data. We find that our estimates of this effect differ from those of various parametric models (not) controlling for selection.

Key-Words: Irregular Identification, Selection Bias, Local Polynomial, Trimming, Count Data.

JEL Classification: C14, C21, C24.

*We are grateful for comments received by Christoph Breunig, Sarawata Chaudhuri, Xavier D'Haultfoeuille, Prosper Dovonon, Jean-Marie Dufour, Bernd Fitzenberger, Mathieu Marcoux, Jeff Racine, Joao Santos Silva, Victoria Zinde-Walsh, and seminar participants at the ESEM 2018, Kent, Frankfurt, ISNPS 2018, Surrey, Concordia University-Cireq, Humboldt University Berlin, the Econometrics Study Group Meeting in Bristol 2017, and ESEM 2017.

†Department of Economics, University of Warwick, Coventry CV4 7AL, UK. Email: Wiji.Arulampalam@warwick.ac.uk

‡Corresponding Author: Department of Economics, University of Surrey, School of Economics, Guildford GU2 7XH, UK. Email: V.Corradi@surrey.ac.uk

§Department of Business and Economics, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60629 Frankfurt am Main, Germany. Email: Gutknecht@wiwi.uni-frankfurt.de

1 Introduction

In selection models, the outcome equation intercept is of fundamental importance when the aim is to recover average treatment effects (see Heckman, 1979, 1990).[1]

However, while the problem of identification and estimation of the intercept has long been resolved in the parametric case, it is well known that in the absence of parametric assumptions on the joint distribution of outcome and selection equation error, the intercept cannot be separately identified from the selection bias term (Heckman, 1990). Still, as the probability of selection approaches one, the selection bias term converges towards the unconditional mean of the outcome error, which typically satisfies a normalization condition (e.g., zero in the linear case). This is an example of an 'identification at infinity' argument (Chamberlain, 1986; Lewbel, 2007; D'Haultfoeuille and Maurel, 2013), which has been exploited by various authors such as Heckman (1990), Andrews and Schafgans (1998), Schafgans and Zinde-Walsh (2002), and more recently Goh (2018) for the identification of the intercept in linear additive selection models.

Nevertheless, the problem of endogenous selection is not just confined to linear regression set-ups. Count data for instance, which are typically modeled via multiplicative error models, may be subject to non random sampling as well (Terza, 1998; Deb and Trivedi, 2006). As an example consider a count model for the effect of private supplementary health insurance plans on health care utilization, e.g. measured by the number of visits to the physician in a given period of time. Self-selection into private insurance plans may impede the identification of the average treatment effect of private health insurance.

Despite its relevance, nonlinear selection models have so far only been studied in specific parametric settings (e.g., see Terza, 1998), and only recently Jochmans (2015) devised an estimator for the slope coefficients of more flexible semiparametric, nonlinear selection models. However, to the best of our knowledge, intercept identification and estimation in the nonlinear case has not yet been studied. We aim at filling this gap in the literature by introducing simple to use intercept estimators for nonlinear semiparametric selection models.

We start by considering models in which the intercept and slope parameters can be separately identified, and where the selection equation satisfies a monotonic index restriction. Leading examples of separable multiplicative models are count data or accelerated failure time models. Prominent cases of separable additive nonlinear models on the other hand are production functions, which are used for instance in the human capital formation literature typically subject to sample selection (e.g., Olivetti, 2006). Since slope and intercept parameters can be separately identified in these first type of models, we recover the former using an existing $\sqrt{n}$ consistent estimator (Jochmans, 2015) in a preliminary step. This allows us to transform the dependent variable and to isolate the intercept and the selection bias. Using the transformed dependent variable, we then construct a nonparametric estimator of the selection bias through local polynomial estimator, and use this to bias adjust a least squares criterion function. The resulting intercept estimator is consistent, asymptotically normal and attains a univariate nonparametric convergence rate.

To address the issue of irregular identification (Khan and Tamer, 2010), which is often a con-

[1] Examples include, among others, testing for the difference in wages of unionized and non-unionized workers, or estimating the ethnic or gender wage gap (e.g., Schafgans, 2000).

cern when relying on 'identification at infinity' type arguments, a situation where observations in the neighborhood of the evaluation point(s) are 'sparse', our estimation methodology relies on a specific marginalization approach first proposed by Stute (1984) in the context of a symmetrized nearest-neighbours estimator for the conditional mean function. The key identifying assumption in this approach, which was recently also adopted by Goh (2018) for the additive linear model, is that the marginal distribution function of the instrument index in the selection equation is strictly increasing on the support of the selection equation error term. Together with other regularity conditions, this assumption ensures that the upper tail limit point of the marginal distribution function equals one only if the propensity score equals one at the upper limit point. However, while the density of the propensity score may not be bounded away from zero at that limit point (thus representing a case of irregular identification), the density of the marginal distribution function of the instrument index follows a uniform distribution on the entire $[0, 1]$ interval. Hence, the estimator achieves a univariate nonparametric rate regardless of the relative tail behavior of the instrument index and the error term distributions as it circumvents the random denominator problem.

Since the assumption of a monotonic index restriction is often too strong in practice and in fact may be violated, we relax this condition and propose a local nonlinear least squares estimator in a second step. This estimator, which is based on a more flexible nonparametric specification of the propensity score, can be used when the index restriction indeed does not hold. That is, since marginalization based on the index restriction is no longer applicable, but irregular identification remains an issue, we only use observations which are close, but not too close to one (or zero). Formally, this is implemented by introducing a trimming sequence, converging to zero at a sufficiently slow rate. As a result of the trimming, this estimator converges at most at a cubic rate.

In a small Monte Carlo study, we investigate the finite sample performance of our different estimators paying special attention to their sensitivity w.r.t. the bandwidth choices. We find that both estimators behave well under different scenarios. Finally, we investigate the average treatment effect of supplementary private health insurance on the number of annual physician or specialist contacts for a sample of elderly individuals eligible for Medicare (Deb and Trivedi, 2006). Since selection on the basis of health status is a common concern in this type of data, we compare different models (not) controlling for selection into private insurance plans. We find that the estimate of the Average Treatment Effect differs markedly from the estimates of various other parametric estimators (not) controlling for selection. This underlines the importance of more flexible specifications to draw robust conclusions.

The rest of the paper is organized as follows. Section 2 outlines the set-up. Section 3 introduces the estimators for the separable case with linear index restriction in the selection equation, and derives their asymptotic properties. Section 4 studies the non-monotonic case, when the single index restriction in the selection equation is violated and a nonparametric propensity score specification is used instead. Section 5 provides the results of the small scale Monte Carlo simulation, while Section 6 contains our empirical illustration. Finally, Section 7 concludes. All proofs are collected in an Appendix.

2 Set-up

We consider a standard sample selection model setup, and outline the data generating process for the separable case, where slope and intercept parameters can be identified and estimated separately. As it is customary in these models, we postulate that the outcome variable y_i is observed if and only if s_i, a binary selection indicator, equals one, while covariate(s) x_i are observed for all individuals in the sample. For the binary selection indicator s_i, we impose initially the following linear index assumption:

$$s_i = 1\{z_i'\gamma_0 > v_i\},$$

where $1\{A\} = 1$ if the event A holds, and zero otherwise, and z_i is an observed covariate vector that contains at least one element which is not contained in x_i (see Assumption A1(iii) below). This type of index restriction is common in the sample selection literature (e.g., Heckman, 1979; Ahn and Powell, 1993) and will be relaxed in Section 4. By contrast, for the outcome equation, we consider multiplicative as well as additive nonlinear models of the form:

$$\mathrm{E}\left[y_i|x_i,\widetilde{\varepsilon}_i\right] = g_{M1}\left(\theta_{0M}\right)\cdot g_{M2}\left(x_i'\beta_{0M}\right)\widetilde{\varepsilon}_i \tag{1}$$

and

$$\mathrm{E}[y_i|x_i,\varepsilon_i] = g_{A1}\left(\theta_{0A}\right) + g_{A2}\left(x_i'\beta_{0A}\right) + \varepsilon_i, \tag{2}$$

respectively, where $g_{M1}(\cdot)$, $g_{M2}(\cdot)$, $g_{A1}(\cdot)$, $g_{A2}(\cdot)$ are known, real-valued functions. In fact, the standard additive linear model follows as a special case when $g_{A1}(\cdot)$ and $g_{A2}(\cdot)$ are the identity functions. An empirically important example of a separable multiplicative models as in (1) are count data models, where:

$$g_{M1}\left(\theta_{0M}\right)\cdot g_{M2}\left(x_i'\beta_{0M}\right) = \exp\left(\theta_{0,M}\right)\exp\left(x_i'\beta_{0M}\right) \tag{3}$$

and $\widetilde{\varepsilon}_i$ typically plays the role of individual unobserved heterogeneity. Sample selection then arises because s_i depends on $\widetilde{\varepsilon}_i$. For instance, y_i could measure the number of credit card defaults for each individual i in a given period of time, while s_i could record whether person i actually possesses such card(s) or not. Since credit card (non-)holders may differ in terms of their risk attitude $\widetilde{\varepsilon}_i$, which is unobserved and likely to be correlated with v_i, standard estimators for (semi-)parametric count data models do not provide consistent estimators of θ_{0M} and β_{0M}. Another example that fits into the set-up of (3) are Accelerated Failure Time models applied to duration data, where samples are often plagued by the presence of endogenous selection (e.g., Ham and LaLonde, 1996). Other examples of nonlinear additive sample selection models can be found in the human capital formation literature (Olivetti, 2006). We therefore deem the separable case sufficiently relevant to be considered in its own right.

Before introducing sample selection into the above equations formally, we outline two assumptions that ensure identification of the intercept parameters θ_{0A} and θ_{0M}. In fact, given the focus of this paper on estimation, we only provide a set of sufficient high-level assumptions which ensure point identification of the intercept parameters in (1) and (2):

A1: (i) $\mathrm{E}\left[|y_i|\right] < \infty$; (ii) The functions $g_{A1}(\cdot)$, $g_{A2}(\cdot)$, $g_{M1}(\cdot)$, and $g_{M2}(\cdot)$ are known; $g_{M1}(\cdot)$ and

$g_{M2}(\cdot)$ are non-zero almost everywhere; (iii) At least one component of z_i, which has a non-zero coefficient, is not included in x_i, and has everywhere positive continuous density, conditional on the other components; (iv) The slope parameters β_{0M} and β_{0A} are point identified up to a scale normalization; (v) It holds that $\mathrm{E}[\widetilde{\varepsilon}_i] = 1$ and $\mathrm{E}[\varepsilon_i] = 0$.

A2: (i) γ_0 is uniquely identified up to a scale and location normalization; (ii) The marginal distribution function of $z_i'\gamma_0$, $F_{z'\gamma_0}(\cdot)$, is continuously differentiable with non-zero derivatives on $\mathrm{supp}(z_i'\gamma_0)$, the support of $z_i'\gamma_0$; (iii) It holds that $\mathrm{supp}(v_i) \subseteq \mathrm{supp}(z_i'\gamma_0)$; (iv) The joint distribution of $\widetilde{\varepsilon}_i$ (ε_i) and v_i is absolutely continuous and admits a density $f_{\widetilde{\varepsilon},v}(.,.)$ ($f_{\varepsilon,v}(.,.)$) with respect to Lebesgue measure; (v) v_i and $\widetilde{\varepsilon}_i$ (ε_i) are independent of x_i and z_i.

Assumption A1(iii) ensures that at least one continuous instrumental variable exists, which is not included in x_i. This assumption is crucial for point identification and, together with A2(ii), will allow for the 'identification at infinity' argument used in this paper.[2] A1(iv) on the other hand is a high-level condition on the identification of the slope coefficients. Given A1(i) and A2, and the existence of an instrumental variable, A1(iv) is rather innocuous and will be satisfied in the separable case under a set of regularity conditions when x_i exhibits sufficient variation (see Jochmans, 2015, for details). A1(v) is a standard assumption in exponential and linear models with intercept.

Assumption A2(i) is a high-level condition, which is not restrictive as γ_0 can be identified and estimated in a separate step. A2(ii) and A2(iii) on the other hand imply that $F_{z'\gamma_0}(\cdot)$ is strictly increasing and invertible on the support of the continuous random variable v_i. This assumption is crucial for the identification argument in the sequel as it ensures that identification can be achieved at $\lim_{\tau\to 1} F^{-1}_{z'\gamma_0}(\tau)$, where $F^{-1}_{z'\gamma_0}(\tau) = \inf\{w : F_{z'\gamma_0}(w) \geq \tau\}$ denotes the quantile function of $F_{z'\gamma_0}(\cdot)$ and $\tau \in (0,1)$. Note that A2(iii) rules out that $\mathrm{supp}(v_i)$ strictly contains $\mathrm{supp}(z_i'\gamma_0)$, a situation where identification of the intercept fails. Moreover, A2(iv) is a standard identification assumption for semiparametric binary choice models. Finally, note that A2(ii)-(iv) naturally imply that $\Pr(s_i = 1) > 0$, while the independence in A2(v), which is neither required for identification nor for estimation (e.g., Andrews and Schafgans, 1998), will be relaxed in Section 4.

In what follows we will focus on the multiplicative case for illustrative purposes. Given A1(i) and A2(v), we can write:

$$\begin{aligned}
\mathrm{E}[\widetilde{\varepsilon}_i|x_i, z_i, s_i = 1] &= \mathrm{E}[\widetilde{\varepsilon}_i|z_i'\gamma_0 > v_i] \\
&= \mathrm{E}[\widetilde{\varepsilon}_i|F_{z'\gamma_0}(z_i'\gamma_0) > F_{z'\gamma_0}(v_i)] \\
&= \frac{\int_0^{F_{z'\gamma_0}(z_i'\gamma_0)} \int \widetilde{\varepsilon} f_{\widetilde{\varepsilon},F_{z'\gamma_0}(v)}(\widetilde{\varepsilon}, F_{z'\gamma_0}(v)) d\widetilde{\varepsilon} dF_{z'\gamma_0}(v)}{\Pr\left(F_{z'\gamma_0}(v_i) \leq F_{z'\gamma_0}(z_i'\gamma_0)\right)} \\
&\equiv \widetilde{\lambda}(F_{z'\gamma_0}(z_i'\gamma_0)),
\end{aligned}$$

where the first equality uses the fact that $\widetilde{\varepsilon}_i$ is independent of x_i and z_i, and the second equality uses A2(ii)-(iii), which ensure that $F_{z'\gamma_0}(z_i'\gamma_0)$ is strictly increasing on the support of v_i. Hence, by the

[2] Honoré and Hu (2018) have recently examined additive linear sample selection models without such an exclusion restriction, and have derived sharp bounds for the parameters of this type of models.

above arguments we obtain:

$$\mathrm{E}\left[y_i|x_i, z_i, s_i = 1\right] = g_{M1}\left(\theta_{0M}\right) g_{M2}\left(x_i'\beta_{0M}\right) \widetilde{\lambda}(F_{z'\gamma_0}(z_i'\gamma_0)). \tag{4}$$

Likewise, in the additive case an identical argument yields:

$$\begin{aligned} \mathrm{E}\left[y_i|x_i, z_i, s_i = 1\right] &= g_{A1}\left(\theta_{0A}\right) + g_{A2}\left(x_i'\beta_{0A}\right) + E[\varepsilon_i|F_{z'\gamma_0}(v_i) < F_{z'\gamma_0}(z_i'\gamma_0)] \\ &\equiv g_{A1}\left(\theta_{0A}\right) + g_{A2}\left(x_i'\beta_{0A}\right) + \lambda(F_{z'\gamma_0}(z_i'\gamma_0)). \end{aligned} \tag{5}$$

Note also that by A2(ii)-(iii), for $\tau \in (0,1)$, we have invertibility of $F_{z'\gamma_0}$ on $\mathrm{supp}(z_i'\gamma_0)$, so that:

$$\lim_{\tau\to 1} \widetilde{\lambda}\left(F^{-1}_{z'\gamma_0}(\tau)\right) = \lim_{\tau\to 1} \mathrm{E}\left[\widetilde{\varepsilon}_i|F^{-1}_{z'\gamma_0}(\tau) > v_i\right] = \mathrm{E}\left[\widetilde{\varepsilon}_i\right]$$

and

$$\lim_{\tau\to 1} \lambda\left(F^{-1}_{z'\gamma_0}(\tau)\right) = \lim_{\tau\to 1} \mathrm{E}\left[\varepsilon_i|F^{-1}_{z'\gamma_0}(\tau) > v_i\right] = \mathrm{E}\left[\varepsilon_i\right].$$

Given the normalization assumptions $\mathrm{E}\left[\widetilde{\varepsilon}_i\right] = 1$ and $\mathrm{E}\left[\varepsilon_i\right] = 0$ in A1(v), it follows that $\lim_{\tau\to 1} \widetilde{\lambda}\left(F^{-1}_{z'\gamma_0}(\tau)\right) = 1$ and $\lim_{\tau\to 1} \lambda\left(F^{-1}_{z'\gamma_0}(\tau)\right) = 0$. A key difference w.r.t. when the propensity score $\Pr(s_i = 1|z_i) = F_v(z_i'\gamma_0)$ is used to control for sample selection (e.g., Das et al., 2003), is that under A2(ii), $F_{z'\gamma_0}(z_i'\gamma_0)$ is uniformly distributed on $[0,1]$ with marginal density $f_{F_{z'\gamma_0}}(\cdot) = 1$ everywhere on its support. This holds true irrespective of whether $\lim_{p\to 1} f_p(p)$, the density of the propensity score $\Pr(s_i = 1|z_i'\gamma_0)$ in the limit, is actually zero or not.

Given (4) and (5), and A1(i) the auxiliary models for observed y_i are:

$$y_i = g_{M1}\left(\theta_{0M}\right) g_{M2}\left(x_i'\beta_{0M}\right) \widetilde{\lambda}(F_{z'\gamma}(z_i'\gamma_0)) + \widetilde{u}_i \tag{6}$$

and

$$y_i = g_{A1}\left(\theta_{0A}\right) + g_{A2}\left(x_i'\beta_{0A}\right) + \lambda(F_{z'\gamma}(z_i'\gamma_0)) + u_i. \tag{7}$$

In order to disentangle the intercept from the selection bias, in the multiplicative case, we may use A1(ii) to obtain:

$$\begin{aligned} &\lim_{\tau\to 1} \mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\Big|v_i < F^{-1}_{z'\gamma_0}(\tau)\right] \\ \equiv\ & g_{M1}(\theta_{0M})\left(\lim_{\tau\to 1} \lambda\left(F^{-1}_{z'\gamma}(\tau)\right)\right) + \lim_{\tau\to 1} \mathrm{E}\left[\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\Big|v_i < F^{-1}_{z'\gamma_0}(\tau)\right] = g_{M1}(\theta_{0M}), \end{aligned}$$

where $g_{M1}(\cdot)$ is a known function. Note that the focus in this paper lies on the estimation of θ_{0M} and θ_{0A} rather than of $g_{M1}(\theta_{0M})$ and $g_{A1}(\theta_{0A})$. This is so since we want to allow for the possibility to make inference about (generic functions of) θ_{0M} and θ_{0A}, and in this case the construction of standard errors requires direct estimation of θ_{0M} and θ_{0A}. In fact, when only $g_{M1}(\theta_{0M})$ and $g_{A1}(\theta_{0A})$ are of interest, identification and estimation are substantially simplified (see discussion in Remark 3 below).

Remark 1: The above set-up has so far only addressed the selection case where the outcome of individuals are observed iff $s_i = 1$. Our set-up can be straightforwardly generalized to the case of

switching regression, in which we observe different outcomes for selected and non-selected or treated and non-treated individuals. To understand the importance of intercept identification in nonlinear models, consider our illustrative empirical example (Section 6), of a simple data model with $g_M(\cdot) = \exp\left(\theta_{0M}^1\right)\exp\left(x_i'\beta_{0M}\right)$ when $s_i = 1$ and $g_M(\cdot) = \exp\left(\theta_{0M}^0\right)\exp\left(x_i'\beta_{0M}\right)$ when $s_i = 0$. Let y_{1i} and y_{0i} denote the number of physician contacts in a given year for people with $(s_i = 1)$ and without $(s_i = 0)$ a specific insurance plan.[3] In this case, the experimental average treatment effect of being under the insurance plan for randomly assigned individuals is often calculated as the relative change in expected physician contacts:

$$\left(\frac{\exp\left(\theta_{0M}^1\right)}{\exp\left(\theta_{0M}^0\right)} - 1\right) \times 100\%,$$

while, under selection into private insurance, the effect one may be able to recover from the actual sample is given by:

$$\left(\frac{\exp\left(\theta_{0M}^1\right)\mathrm{E}[\widetilde{\varepsilon}_{1i}|x_i'\beta_{0M}, s_i = 1]}{\exp\left(\theta_{0M}^0\right)\mathrm{E}[\widetilde{\varepsilon}_{0i}|x_i'\beta_{0M}, s_i = 0]} - 1\right) \times 100\%.^4$$

provided $\mathrm{E}[\widetilde{\varepsilon}_{0i}|x_i'\beta_{0M}, s_i = 0] > 0$ a.s.. It is again clear that both expressions in general differ except for the special case where $\mathrm{E}[\widetilde{\varepsilon}_{1i}|x_i'\beta_{0M}, s_i = 1] = \mathrm{E}[\widetilde{\varepsilon}_{0i}|x_i'\beta_{0M}, s_i = 0]$. Here, one could for instance test the null hypothesis of no average treatment effect of private insurance, i.e. $H_0 : \theta_{0M}^1 = \theta_{0M}^0$, against the alternative of a positive treatment effect, $H_A : \theta_{0M}^1 > \theta_{0M}^0$. This can be easily done using Theorem 1 in the next section. Finally, note that if the effect of covariates x_i on y_i differs when $s_i = 0$ and $s_i = 1$, the relative change in expected physician contacts displayed above differs, and does also involve x_i and the slope parameters. This implies that testing for (no) average treatment effects becomes more involved and requires knowledge about the joint distribution of the intercept and slope coefficients. Again, inference relies on an estimator of the intercept.

3 Estimation

The objective of this section is to derive estimators of θ_{0M} and θ_{0A}, and to establish their consistency and asymptotic normality. In order to accomplish this, we first require estimators of the unknown quantities in Equations (6) and (7) from the previous section, namely γ_0, $F_{z'\gamma}(\cdot)$ (we omit the subscript "0" in what follows for simplicity), the slope coefficients, and the selection bias term. A $\sqrt{n}$-consistent estimator for the instrument parameter vector γ_0 can be obtained from Klein and Spady (1993). From here onwards, we call this estimator $\widehat{\gamma}$.[5] This allows us to construct an estimator of the cumulative distribution function of $z_i'\gamma_0$ in a straightforward manner:

$$\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}) = \frac{1}{n}\sum_{j=1}^{n} 1\left\{z_j'\widehat{\gamma} \leq z_i'\widehat{\gamma}\right\}.$$

[3] As it is customary in count models, we assume that $\widetilde{\varepsilon}_{1i}, \widetilde{\varepsilon}_{0i} > 0$ almost surely.

[4] The example makes the implicit assumption that β_{0M} can be point identified and individuals with $s_i = 0$ and $s_i = 1$ share a common support in terms of the single index $x_i'\beta_{0M}$.

[5] Since our theoretical results in Theorem 1 and 2 below demonstrate that the estimation error of a $\sqrt{n}$-consistent $\widehat{\gamma}$ does not feature into the limiting distribution of our intercept estimator due to its slower than parametric convergence rate, we do not discuss its estimation further here. For details on the estimation and on the appropriate under-smoothing of the bandwidth see Klein and Spady (1993).

Note that this step is common to both multiplicative and additive models. In a next step, we then obtain estimators for the slope coefficients, say $\widehat{\beta}_M$ and $\widehat{\beta}_A$. Given separability of the models in (6) and (7), we can construct these independently of the intercepts at a parametric $\sqrt{n}$ rate following Jochmans (2015).

3.1 Estimator for Multiplicative Model

We start with the estimator for θ_{0M}. In order to disentangle the intercept from the selection bias term, we require an estimator of the latter, which we will construct next. From here onwards, let

$$\mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}(z_i'\gamma_0)=1^-\right.\right] \equiv \lim_{\tau\to 1}\mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|v_i<F_{z'\gamma}^{-1}(\tau)\right.\right]$$

and

$$\mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}(z_i'\gamma_0)\right.\right] \equiv \mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}(v_i)<F_{z'\gamma}(z_i'\gamma_0)\right.\right].$$

In other words, "1^-" denotes the upper limit point that may or may not be reached. Heuristically, a nonparametric regression of $\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}$ on $F_{z'\gamma}(z_i'\gamma_0)$ gives

$$\begin{aligned} m_M\left(F_{z'\gamma}(z_i'\gamma_0)\right) &\equiv \mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}(z_i'\gamma_0)\right.\right] \\ &= g_{M1}(\theta_{0M})\widetilde{\lambda}(F_{z'\gamma}(z_i'\gamma_0))+\mathrm{E}\left[\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}(z_i'\gamma_0)\right.\right] \\ &= g_{M1}(\theta_{0M})\widetilde{\lambda}(F_{z'\gamma}(z_i'\gamma_0)), \end{aligned}$$

and let

$$\begin{aligned} & m_M\left(1^-\right)=\mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}(z_i'\gamma_0)=1^-\right.\right] \\ \equiv\ & g_{M1}(\theta_{0M})\widetilde{\lambda}(1^-)+\mathrm{E}\left[\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}(z_i'\gamma_0)=1^-\right.\right]=g_{M1}(\theta_{0M})\widetilde{\lambda}(1)=g_{M1}(\theta_{0M}). \end{aligned}$$

This suggests that we can recover the selection term as $\widetilde{\lambda}(F_{z'\gamma}(z_i'\gamma_0))=m_M\left(F_{z'\gamma}(z_i'\gamma_0)\right)/m_M(1^-)$ provided A1(ii) holds. The above reasoning requires an estimator for $m_M\left(F_{z'\gamma_0}(z_i'\gamma)\right)$ and for $m_M\left(1^-\right)$ to construct the selection bias component. In order to account for the boundary issue when estimating $m_M\left(1^-\right)$, we use a local polynomial estimator of odd order, for which the order of the bias is the same in the interior and at the boundary (e.g., Ruppert and Wand, 1994; Fan and Gijbels, 1992).

Thus, define the q-th order local polynomial estimator:

$$\begin{aligned} & \left(\widehat{a}_{M0}\left(1^-\right),...,\widehat{a}_{Mq}\left(1^-\right)\right) \\ =\ & \arg\min_{a_k,k\le q}\frac{1}{nh}\sum_{j=1}^{n}s_j\left(\frac{y_j}{g_{M2}\left(x_i'\widehat{\beta}_M\right)}-\sum_{0\le k\le q}a_k\left(\widehat{F}_{z'\gamma}(z_j'\widehat{\gamma})-1\right)^k\right)^2 \\ & K\left(\frac{\widehat{F}_{z'\gamma}(z_j'\widehat{\gamma})-1}{h}\right) \end{aligned} \tag{8}$$

and let $\widehat{m}_M(1^-) = \widehat{a}_{M0}(1^-)$, where $h \to 0$ as $n \to \infty$ denotes the bandwidth sequence. Analogously, define $\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))$ replacing 1^- with $\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})$ in (8). We can now define the intercept estimator as,

$$\widehat{\theta}_M = \arg \min_{\theta_M \in \Theta_M} \frac{1}{n} \sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{\widehat{m}_M(1^-)}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} - g_{M1}(\theta_M) \right)^2,$$

and its probability limit as

$$\begin{aligned} \theta_{0M} &= \arg \min_{\theta_M \in \Theta_M} \mathrm{E}\left[s_i \left(\frac{y_i}{g_{M2}(x_i'\beta_{0M})} \frac{m_M(1^-)}{m(F_{z'\gamma}(z_i'\gamma_0))} - g_{M1}(\theta_M) \right)^2 \right] \\ &= \arg \min_{\theta_M \in \Theta_M} \mathrm{E}\left[s_i \left(g_{M1}(\theta_{0M}) - g_{M1}(\theta_M) + \frac{u_i}{g_{M2}(x_i'\beta_{0M})} \right)^2 \right]. \end{aligned}$$

3.2 Asymptotics for Multiplicative Model

To derive the asymptotic properties of $\widehat{\theta}_M$, we impose the following conditions in the sequel:

E1: The sample observations $\{y_i, x_i', z_i', s_i\}_{i=1}^n$ are i.i.d. and $\mathrm{E}\left[y_i^2\right] < \infty$.

E2: The known functions $g_{A1}(\cdot)$ and $g_{M1}(\cdot)$ are twice continuously differentiable with non-zero derivatives.

E3: The parameter space of θ_{0M}, Θ_M, is compact and θ_{0M} lies in its interior.

E4: (i) $\widetilde{\lambda}(.)$ is $q+1$ times differentiable on $(0,1)$ with q an odd integer. All derviatives are Lipschitz continuous; (ii) $\widetilde{\lambda}(.)$ and its $q+1$ partial derivatives are left continuous at the upper tail limit point 1.

E5: There exist estimators of (i) γ_0 satisfying $\|\widehat{\gamma} - \gamma_0\| = O_p\left(n^{-1/2}\right)$, and (ii) β_{0M} satisfying $\|\widehat{\beta}_M - \beta_{0M}\| = O_p\left(n^{-1/2}\right)$, respectively, where $\|\cdot\|$ denotes the Euclidean norm.

E6: $\mathrm{E}\left[\frac{s_i \widetilde{u}_i^2}{g_{M2}(x_i'\beta_{0M})^2} \left| F_{z'\gamma}(v_i) < F_{z'\gamma}(z_i'\gamma_0) \right.\right] < \infty$ for all z_i.

E7: The kernel function $K(\cdot)$ is a continuous, symmetric function around zero, with compact support on $[-1,1]$. It satisfies $\int_{-\infty}^{\infty} K(v)dv = 1$, $\int_{-\infty}^{\infty} vK(v)dv = 0$, and $\int_{-\infty}^{\infty} v^2 K(v)dv < \infty$.

While E3 is standard, E4 is an implicit assumption on the smoothness of the joint distribution of $\widetilde{\varepsilon}_i$ and $F_{z'\gamma}(v_i)$. In fact, under the given assumptions, a degree of differentiability of $\widetilde{\lambda}(\cdot)$ up to order $q+1$ can be shown to directly imply the differentiability of the marginal density function of $F_{z'\gamma}(v_i)$ and the joint density function of $(\widetilde{\varepsilon}_i, F_{z'\gamma}(v_i))$ in the direction of $F_{z'\gamma}(v_i)$, up to order q. Assumption E5 on the other hand is a high-level condition on the existence of appropriate estimators for the 'first stage' parameters β_{0M} and γ_0, see e.g. existing estimators such as Klein and Spady (1993) and Jochmans (2015). It naturally requires point identification of β_{0M} and γ_0, respectively, which holds under more primitive normalization conditions and assumptions about the covariate space of x_i and z_i (e.g., Sherman, 1993). Finally, E7 is a standard condition on the kernel function satisfied by most second order kernels. The following theorem establishes the limiting distribution of $\widehat{\theta}_M$.

Theorem 1: Let Assumptions A1-A2, and E1-E7 hold. If as $n \to \infty$, $nh^{2(q+1)+1} \to 0$, $q \geq 1$, and $nh \to \infty$, then

(i)

$$\widehat{\theta}_M \xrightarrow{p} \theta_{0M}$$

(ii)

$$\sqrt{nh}\left(\widehat{\theta}_M - \theta_{0M}\right) \xrightarrow{d} N(0, \Omega_{0M})$$

where

$$\Omega_{0M} = \frac{\mathrm{E}\left[s_i\left(\frac{y_i}{g_{M2}(x_i'\beta_{0M})m_M(F_{z'\gamma}(z_i'\gamma_0))}\right)\right]^2}{\Pr(s_i=1)^2\left(\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)\right)^2}\mathrm{E}\left[\frac{s_i\widetilde{u}_i^2}{g_{M2}^2\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}\left(z_i'\gamma_0\right)=1^-\right.\right]\left[\mathbf{M}_1^{-1}\mathbf{\Gamma}_1\mathbf{M}_1^{-1}\right]_{00},$$

with $[A]_{00}$ denoting the upper left entry of matrix A, and $\mathbf{M}_1$ and $\mathbf{\Gamma}_1$ are defined in the Appendix.

A couple of remarks:

Remark 2: As pointed out before, Theorem 1 establishes the asymptotic normality of $\sqrt{nh}\left(\widehat{\theta}_M - \theta_{0M}\right)$ at a univariate nonparametric rate. This is so because the limiting distribution is driven by

$$\sqrt{nh}\left(\widehat{m}_M(1^-) - m_M(1)\right),$$

which implies that the estimation error of $\widehat{\gamma}$, $\widehat{\beta}_M$, and $\widehat{F}_{z'\gamma}$ do not feature in the limiting distribution of the intercept estimator. Note that this result holds irrespective of whether $\lim_{p\to 1} f_p(p)$ is bounded away from zero or not, due to the fact that we used the transformation $F_{z'\gamma}(\cdot)$. Also, from the discussion above and the bandwidth conditions in Theorem 1, it is clear that a higher degree of smoothness q of the joint distribution of $(\widetilde{\varepsilon}_i, F_{z'\gamma}(v_i))$ directly translates into a faster convergence rate (see Hall and Racine (2015) for the possibility to allow the polynomial order q to grow with the sample size n).

Remark 3: As discussed before, if only $g_{M1}(\theta_{0M})$ is of interest, this quantity can be recovered directly through the relationship $m_M(1^-) = g_M(\theta_{0M})$. In this case, we can estimate $\widehat{g_M(\theta_{0M})}$ through $\widehat{m}_M(1^-)$, and the limiting distribution is given by Lemma 1 in the Appendix, namely:

$$\sqrt{nh}\left(\widehat{m}_M(1^-) - g_{M1}\left(\theta_{0M}\right)\right) \xrightarrow{d} N\left(0, \sigma_1^2(1)\right),$$

where $\sigma_1^2 = \mathrm{E}\left[\frac{s_i\widetilde{u}_i^2}{g_{M1}^2\left(x_i'\beta_{0M}\right)}\left|F_{z'\gamma}\left(z_i'\gamma_0\right)=1^-\right.\right]\left[\mathbf{M}_1^{-1}\mathbf{\Gamma}^1\mathbf{M}_1^{-1}\right]_{00}$. However, as pointed out, to make inference about any other generic function of θ_{0M}, we require the limiting distribution from Theorem 1.

Remark 4: An estimator of Ω_{0M} can be constructed as

$$\begin{aligned}\widehat{\Omega}_M &= \left[\mathbf{M}_1^{-1}\mathbf{\Gamma}_1\mathbf{M}_1^{-1}\right]_{00}\frac{\left(\frac{1}{n}\sum_{i=1}^n\left(s_i\left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))}\right)\right)\right)^2}{\left(\frac{1}{n}\sum_{i=1}^n s_i\right)^2\left(\nabla_{\theta_M}\left(g_{M1}(\widehat{\theta}_M)\right)\right)^2}\\ &\quad\times\frac{1}{nh}\sum_{i=1}^n\left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} - \widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))\right)^2 K\left(\frac{\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})-1}{h}\right),\end{aligned}$$

where the theoretical moments of the kernel function in $\mathbf{M}_1^{-1}\mathbf{\Gamma}_1\mathbf{M}_1^{-1}$ can be computed analytically.

For instance, if an ordinary second order Epanechnikov kernel and a local linear estimator is used, the upper left element of this matrix is approximately given by 4.498.

3.3 The Additive Case

The key difference between the multiplicative and the additive case is that, in the latter, the selection bias enters additively rather than multiplicatively. Therefore, unlike in Equation (8), define

$$\begin{aligned} & \left(\widehat{a}_{A0}\left(1^{-}\right),...,\widehat{a}_{Aq}\left(1^{-}\right)\right) \\ = & \arg\min_{a_k,k\leq q}\frac{1}{nh}\sum_{i=1}^{n}s_i\left(y_i-g_{A2}(x_i'\widehat{\beta}_A)-\sum_{0\leq k\leq q}a_k\left(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})-1\right)^k\right)^2 \\ & K\left(\frac{\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})-1}{h}\right) \end{aligned} \tag{9}$$

and note that $\widehat{m}_A(1^-)=\widehat{a}_{A0}\left(1^-\right)$. Likewise, define $\widehat{m}_A(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))$ replacing again 1^- with $\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})$. As in the multiplicative case we can then write

$$\widehat{\theta}_A=\arg\min_{\theta_A\in\Theta_A}\frac{1}{n}\sum_{i=1}^{n}s_i\left(y_i-g_{A2}(x_i'\widehat{\beta}_A)-g_{A1}(\theta_A)-\left(\widehat{m}_A(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))-\widehat{m}_A(1^-)\right)\right)^2,$$

with the probability limit

$$\begin{aligned} \theta_{0A} &= \arg\min_{\theta_A\in\Theta_A}\mathrm{E}\left[s_i\Big(y_i-g_{A2}(x_i'\beta_{0A})-g_{A1}(\theta_A)-\big(m_A(F_{z'\gamma}(z_i'\gamma_0))-m_A(1^-)\big)\Big)^2\right] \\ &= \arg\min_{\theta_A}\mathrm{E}\left[s_i\Big(g_{A1}(\theta_{0A})-g_{A1}(\theta_A)+u_i\Big)^2\right]. \end{aligned}$$

Here, $m_A(F_{z'\gamma}(z_i'\gamma_0))=\mathrm{E}\left[y_i-g_{A2}(x_i'\beta_{0A})|F_{z'\gamma}(z_i'\gamma_0)\right]$, and $m_A(1^-)=\mathrm{E}\left[y_i-g_{A2}(x_i'\beta_{0A})|F_{z'\gamma}(z_i'\gamma_0)=1^-\right]$.

To establish analogous asymptotic results as in Subsection 3.2, we slightly modify Assumptions E3 to E6 as follows:

E3A: As E3, but θ_{0M} replaced by θ_{0A}, and Θ_M by Θ_A.

E4A: As E4, but $\widetilde{\lambda}(.)$ replaced by $\lambda(.)$.

E5A: As E5, but $\widehat{\beta}_M$ and β_{0M} replaced by $\widehat{\beta}_A$ and β_{0A}, respectively.

E6A: $\mathrm{E}\left[s_iu_i^2|F_{z'\gamma}(z_i'\gamma_0)\right]<\infty$.

We obtain the following results:

Theorem 2: Let Assumptions A1-A2, E1, E3A, E4A, E5A, and E6A. If as $n\to\infty$, $nh^{2(q+1)+1}\to 0$, $q\geq 1$, and $nh\to\infty$, then

(i)

$$\widehat{\theta}_A\xrightarrow{p}\theta_{0A}$$

(ii)

$$\sqrt{nh_n}\left(\widehat{\theta}_A-\theta_{0A}\right)\xrightarrow{d}N(0,\Omega_{0A})$$

where

$$\Omega_{0A} = \frac{\mathrm{E}\left[s_i u_i^2 \left| F_{z'\gamma}\left(z_i'\gamma_0\right) = 1^-\right.\right]\left[\mathbf{M}_1^{-1}\mathbf{\Gamma}_1\mathbf{M}_1^{-1}\right]_{00}}{\Pr\left(s_i=1\right)^2\left(\nabla_{\theta_A} g_{A1}(\theta_{0A})\right)^2}$$

Remark 5: As pointed out in Remark 4, we can estimate Ω_{0A} as

$$\begin{aligned}\widehat{\Omega}_A &= \left[\mathbf{M}_1^{-1}\mathbf{\Gamma}_1\mathbf{M}_1^{-1}\right]_{00} \\ &\quad \frac{\frac{1}{n}\sum_{i=1}^n s_i\left(y_i - g_{A2}\left(x_i'\widehat{\beta}_A\right) - \widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})\right)^2}{\left(\frac{1}{n}\sum_{i=1}^n s_i\right)^2\left(\nabla_{\theta_A}\left(g_{A1}(\widehat{\theta}_A)\right)\right)^2} K\left(\frac{\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})-1}{h}\right)\end{aligned}$$

Remark 6: Note that the procedure outlined in this section can be straightforwardly extended to accommodate (continuous) endogenous regressors (Das et al., 2003). Assume for simplicity that at least one (continuous) element of x_i is endogenous, say $x_{i,2}$, with $x_i = (x_{i,1}', x_{i,2}')'$. Moreover, assume that an additional instrument $z_{i,2}$ exists and that the endogenous variable satisfies

$$x_{i,2} = \mu(x_{i,1}, z_{i,2}) + \xi_i,$$

where $\mathrm{E}[\xi_i | x_{i,1}, z_{i,2}] = 0$. By similar arguments to before, we can show that $\mathrm{E}[\varepsilon_i | x_i, z_i, \xi_i, s_i = 1] = \lambda(F_{z'\gamma}(z_i'\gamma_0), \xi_i)$. This implies that we can construct $\lambda(\cdot,\cdot)$ simply as the difference of

$$\int \mathrm{E}\left[y_i \left| F_{z'\gamma}(z_i'\gamma_0) = 1^-, \xi, s_i = 1\right.\right] f_\xi(\xi) d\xi$$

and $\mathrm{E}\left[y_i \left| F_{z'\gamma}(z_i'\gamma_0), \xi_i, s_i = 1\right.\right]$, where the former can be estimated as

$$\frac{1}{n}\sum_{j=1}^n \widehat{m}_A(1^-, \widehat{\xi}_j).$$

A similar argument can obviously be made for the multiplicative model.

4 Non-Monotonic Case

In the previous section, we assumed that the probability of selection is a monotonic function of the instrument index, i.e. we assumed that z_i was independent of the selection error term v_i. While independence can in principle be relaxed along the lines of Das et al. (2003) or Andrews and Schafgans (1998), mis-specification of the selection equation more generally is a common concern in applied work and can lead to inconsistent estimators of the intercept. We therefore consider a more flexible nonparametric specification of the propensity score using $p(z_i) = \Pr(s_i = 1 | z_i)$ in what follows. That is, we only require that the probability of selection is a smooth function of the instrument(s) z_i. This implies that the marginal distribution function of the propensity score might not necessarily be invertible and so 'marginalization' is no longer possible. Moreover, since we do not impose a functional form of $p(z_i)$, the conditional distribution function $p(z_i)$ needs to be estimated in a nonparametric

manner. Hereafter, for notational simplicity, we assume that all the components of x_i and z_i are continuous. The extension to discrete covariates in boths vectors is immediate at the cost of more complicated notation and more lengthy arguments in the proofs. Moreover, as pointed out by Li and Racine (2008), note that only continuous regressors matter for the convergence rate of estimators of conditional nonparametric distribution functions such as $p(z_i)$.[6] Thus, to construct $p(z_i)$, we simply use a standard Nadaraya-Watson estimator of the form:

$$\widehat{p}(z_i) = \frac{\sum_{j=1}^{n} s_i \mathbf{K}\left(\frac{z_i - z_j}{h_1}\right)}{\sum_{j=1}^{n} \mathbf{K}\left(\frac{z_i - z_j}{h_1}\right)}, \tag{10}$$

where $\mathbf{K}(\cdot)$ denotes the product of d_z univariate kernel functions $K(\cdot)$. Moreover, define

$$C_{n,x_0'\beta_0} = \left(x_0'\beta_{0M} - \xi_n, x_0'\beta_{0M} - \xi_n\right)$$

for some $x_0'\beta_{0M} \in \text{supp}(x_i'\beta_{0M})$, where ξ_n is a deterministic sequence that may converge to 0, to some $\xi_0 > 0$, or even diverge (if $\text{supp}(x_i'\beta_{0M})$ is unbounded) as $n \to 0$. In the sequel, we require the following additional assumptions:

E8: (i) For the estimated $\widehat{p}(z_i)$, it holds that $\sup_z |\widehat{p}(z) - p(z)| = o_p(1)$; (ii) The estimated $\widehat{p}(z)$ admits the following representation:

$$\widehat{p}(z) - p(z) = \frac{1}{nh_1^{d_z}} \sum_{j=1}^{n} \frac{\mathbf{K}\left(\frac{z - z_j}{h_1}\right)}{f_z(z_i)} \psi_j + \Xi_n(z) + o_p\left(\frac{1}{\sqrt{nh_1^{d_z}}} + h_1^s\right)$$

for some $s \geq 4$ and $d_z \leq 3$. Moreover, it holds that $\Xi_n(\cdot)$ is continuously differentiable with bounded derivative and $\sup_z |\Xi_n(z)| = O_p(h_1^s)$, and $\mathrm{E}[|\omega(z_i, z_j)|^2] = o(n)$ with $\omega(z_i, z_j) = h_1^{-d_z} \mathbf{K}\left(\frac{z_i - z_j}{h_1}\right) / f_z(z_i)$. Finally:

$$\frac{1}{\sqrt{nh_1^{d_z}}} \sum_{j=1}^{n} \frac{\mathbf{K}\left(\frac{z_i - z_j}{h_1}\right)}{f_z(z_i)} \psi_j$$

satisfies a central limit theorem for triangular arrays of i.i.d. random variables.

E9: There exists at least one point $\mathcal{C}_{n,x'\beta_0}$ such that for some z in $\text{supp}(z_i)$ it holds that $p(z) = 1$. Moreover, there exists a strictly positive, continuous, and integrable function $v_{\widetilde{u},x'\beta_0,1}(\widetilde{u}, t, 1)$ such that for all $\widetilde{u} \in \text{supp}(\widetilde{u})$ and $t \in \mathcal{C}_{n,x'\beta_0}$:

$$\sup_{\widetilde{u} \in \text{supp}(\widetilde{u}), t \in t \in \mathcal{C}_{n,x'\beta_0}} \left(\left| \frac{f_{\widetilde{u},x'\beta_0,p}(\widetilde{u}, t, 1 - H)}{v_{\widetilde{u},x'\beta_0,p}(\widetilde{u}, t, 1) H^{\eta}} - 1 \right| \right) \to 0$$

as $n \to \infty$ for some $0 \leq \eta < \overline{\eta} < 1$, and there exists a C such that:

$$\sup_{\widetilde{u} \in \text{supp}(\widetilde{u}), t \in \mathcal{C}_{n,x'\beta_0}} |\Pr(s_i = 1 | \widetilde{u}, t, 1 - H) - 1| \leq C H^{1-\eta}.$$

[6] In fact, in our empirical illustration in Section 6, apart from two continuous instruments, all of our first stage covariates are either discrete or categorical and so the curse of dimensionality does not 'bite'.

E10: Let $p_i = p(z_i)$. The joint distribution of $(\widetilde{u}_i, p_i, x_i'\beta_{0M})$ is absolutely continuous, and admits a density $f_{\widetilde{u},p,x'\beta_{0M}}(\widetilde{u}_i, p_i, x_i'\beta_{0M})$ with respect to the Lebesgue measure, which is continuously differentiable in p_i everywhere on (0,1) with bounded partial derivative.

E11: (i) it holds that $\mathrm{E}[\widetilde{\varepsilon}_i|x_i, z_i, s_i = 1] = \mathrm{E}[\widetilde{\varepsilon}_i|p_i] \equiv \widetilde{\lambda}(p_i)$; (ii) the function $\widetilde{\lambda}(\cdot)$ is continuously differentiable on $(0, 1)$ with bounded derivative, (iii) there exist positive constants C such that:

$$\left|\widetilde{\lambda}(1-H) - 1\right| \leq CH^{1-\eta}.$$

E8 represents a high level condition on the form of the propensity score. It requires the use of a higher order kernel function, though as long as the number of continuous instruments does not exceed three, a quartic kernel is sufficient. Moreover, note that in the selection case, where more data is available at the first stage than at the second (selected) stage, the required under-smoothing appears less problematic than in other contexts. Assumption E9 requires identification at infinity, for at least some value of $x_i'\beta_{0M}$. The case where $\xi_n \to \xi_0 > 0$ so that $(x_0'\beta_{0M} - \xi_0, x_0'\beta_{0M} - \xi_0) \subset \mathrm{supp}(x_i'\beta_{0M})$ corresponds to the case of strong support, as we require the propensity score to approach one for all $x_i'\beta_{0M}$ in that set. On the other hand, when $\xi_n \to 0$, only observations over an interval shrinking to a singleton in the support of $x'\beta_{0M}$ contribute to the estimation of the intercept. Furthermore, in all cases we allow for so called irregular support, in the sense that $f_{\widetilde{u},x'\beta_0,p}(\widetilde{u}, x'\beta_{0M}, 1)$ is not necessarily bounded away from zero at $p = 1$. In fact, when $\eta = 0$, $\lim_{H\to 0} f_{\widetilde{u},x'\beta_0,p}(\widetilde{u}, x'\beta_{0M}, 1)$ is bounded away from zero for all $x_i'\beta_{0M} \in C_{n,x_0'\beta_0}$, while $\eta > 0$ corresponds to the case of irregular support with a larger value of η representing thinner tails. That is, if $\eta > 0$, we allow for a thin set of observations with a propensity score close to one. Assumption E11(i) is a conditional mean independence assumption which is weaker than A2(v) and requires that the conditional expectation of $\widetilde{\varepsilon}_i$ is only a function of the propensity score. Assumption E11(ii) on the other hand imposes smoothness on the joint distribution of $\widetilde{\varepsilon}_i$ and v_i in the same way as Assumption E4 did. Finally, E11(iii) imposes smoothness in proximity to the boundary point 1, and allows to handle the case of irregular support where the density of p_i at 1 is not bounded away from zero.

The key difference with respect to the previous case is that, while a positive density at $\lim_{\tau\to 1} F_{z'\gamma}(\tau)$ was ensured by the marginalization, $\lim_{\tau\to 1} f_p(\tau)$ may indeed not be bounded away from zero. Heuristically, when identification at infinity holds and $p(z_i)$ reaches the value of one in the limit, it may still be possible that observations are very sparse in the neighborhood of one ('thin density set'), and so convergence occurs at an irregular rate (Khan and Tamer, 2010).

In order to overcome this irregular identification issue, we suggest a local nonlinear least squares estimator which makes use of observations with propensity score close but not too close to one. This is implemented by introducing a trimming sequence, which approaches zero at a sufficiently slow rate. In a nutshell, instead of using observations with a propensity score $p_i \in (1 - h_2, 1)$ we use observations with $p_i \in (1 - h_2 - H, 1 + h_2 - H)$, where $H > h_2$, and both h_2 and H go to zero as the sample size increases, but H approaches zero at a slower rate. In fact, the degree of trimming is controlled by the rate at which H goes to zero. The slower the rate, the higher the degree of trimming. Intuitively, we are discarding all observations with $p_i \in (1 - H, 1)$. By noting that $\widetilde{\lambda}(1^-) - \widetilde{\lambda}(1 - H) = O(H)$, and

recalling that the bias of the intercept estimator depends on the selection bias, it is immediate to see that such bias cannot approach zero at a rate faster than H. Consequently, the intercept estimator converges at a rate, which is at most cubic. Importantly, this cubic rate is not due to the boundary, but to the trimming sequence.[7] In the sequel, we shall treat the separable case formally, and only outline the non-separable one. Moreover, we restrict ourselves to the multiplicative model as the additive case follows by analogous arguments.

4.1 Separable Models

We begin by estimating the propensity score using Equation (10) in a first stage. Moreover, as before, we can recover the slope parameters at a parametric rate using e.g. Jochmans (2015). The propensity score can also be estimated in a separate preliminary stage. We then obtain the transformed dependent variable to construct a local nonlinear least squares estimator, which uses observations with propensity score close to but not too close to one.

Next, let $\delta = 1 - H$, and define:

$$\widehat{\theta}_M^p = \arg \min_{\theta_M^p \in \Theta_M} \frac{1}{nh_2} \sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}\left(x_i'\widehat{\beta}_M\right)} - g_{M1}\left(\theta_M^p\right) \right)^2 K\left(\frac{\widehat{p}(z_i) - \delta}{h_2} \right).$$

Moreover, let $n_0 = n\text{plim}_{n\to\infty} \frac{1}{n} \sum_{=1}^{n} 1\left\{ x_i'\beta_{0M} \in C_{n,x_0'\beta_0} \right\}$ the number of observations that fall into the set $C_{n,x_0'\beta_0}$ and note that whenever $(x_0'\beta_{0M} - \xi_n, x_0'\beta_{0M} - \xi_n)$ converges to a subset of non-zero Lebesgue measure in $\text{supp}(x_0'\beta_{0M})$:

$$\text{plim}_{n\to\infty} \frac{1}{n} \sum_{=1}^{n} 1\left\{ x_i'\beta_{0M} \in C_{n,x_0'\beta_0} \right\} = c > 0$$

for some positive constant $c \in (0, 1]$, while when $\xi_n \to 0$ we have that $\text{plim}_{n\to\infty} \frac{1}{n} \sum_{=1}^{n} 1\left\{ x_i'\beta_{0M} \in C_{n,x_0'\beta_0} \right\} = 0$.

Theorem 3 Let Assumptions A1, E1-E3, E5(ii), E6-E11 hold. If as $n \to \infty$, $h_1, h_2, H \to 0$ and $H/h_2 \to \infty$, (i) $n_0 h_2 H^{2-\eta} \to 0$, (ii) $n_0 h_2 h_1^{2s} H^{-\eta} \to 0$ and (iii) $n_0 h_2^2 h_1^{d_z} H^{\eta} \to \infty$, then
(i)

$$\widehat{\theta}_M^p \xrightarrow{p} \theta_{0M}$$

(ii)

$$\widehat{\Omega}_{M,p}^{-1/2} \left(\widehat{\theta}_M^p - \theta_{0M} \right) \xrightarrow{d} N(0, 1)$$

where

$$\widehat{\Omega}_{M,p} = \frac{\int K(v)^2 dv}{\nabla_{\theta_M} g_{1M}\left(\widehat{\theta}_M^p\right)^2} \sum_{i=1}^{n} \widehat{\widetilde{u}}_i^2 s_i K\left(\frac{\widehat{p}_i - \delta}{h_2} \right),$$

[7] Note that our trimming approach requires a 'well behaved' $f_p(\cdot)$ in areas which lie to the left of $1 - H$, and thus implicitly imposes certain restrictions on the tail behavior of $f_p(\cdot)$. This type of tail conditions are common to estimation problems which use trimming sequences (e.g. Escanciano et al., 2014).

$$\widehat{\widetilde{u}}_i = \frac{y_i}{g_{M2}\left(x_i'\widehat{\beta}_M\right)} - g_{M_1}\left(\widehat{\theta}_M^p\right),$$

Remark 7: Theorem 3 establishes the limiting distribution of the studentized statistic. Note that the convergence rate can be at most $\sqrt{nh_2}$, which is at most a cubic rate. However, the rate can be slower if the observations with $p_i \in (1-h_2-H, 1+h_2-H)$ grow at a rate slower than nh_2, i.e. if $\eta > 0$, and/or $n_0/n \to 0$. In both cases, $\left(\widehat{\theta}_M^p - \theta_{0M}\right)$ and $\widehat{\Omega}_{M,p}^{1/2}$ will diverge to infinity at the same rate, and so the studentized statistic still remains bounded and converges to a standard normal.

Remark 8: As the theoretical results crucially hinge on the tuning parameters, whose rates depend in turn on the unknown ξ_n and η, a discussion of their choice in practice is warranted: For $s = 4$, we set $h_1 = n^{-\frac{1}{8+d_z}}$. Now, (i) and (ii) become identical if we set $H = h_1^4$, and $h_2 = h_1^{4(1+\varepsilon)}$. Then, (i) and (ii) are both satisfied if

$$\frac{12 + 4\varepsilon - 4\eta}{8 + d_z} > 1$$

which implies $\eta < \frac{4-d_z+4\varepsilon}{8+d_z}$. Note that if $n_0/n \to 0$, (i)-(ii) are satisfied also for larger η. Suppose that $n_0 = n^\delta$ with $\underline{\delta} \leq \delta \leq 1$, then (iii) is satisfied if

$$\eta < \frac{\delta}{4}(8 + d_z) - (1 + \varepsilon) - \frac{d_z}{4}$$

Thus, we need

$$\min\left\{\frac{4 - d_z + 4\varepsilon}{8 + d_z}, \frac{\delta}{4}(8 + d_z) - (1 + \varepsilon) - \frac{d_z}{4}\right\} > 0$$

and

$$\eta < \min\left\{\frac{4 - d_z + 4\varepsilon}{8 + d_z}, \frac{\delta}{4}(8 + d_z) - (1 + \varepsilon) - \frac{d_z}{4}\right\}. \tag{11}$$

4.2 Non-Separable

We now outline the case in which neither monotonicity nor separability hold. In this case, in fact, intercept and slope parameters can no longer be separately identified, and hence have to be estimated jointly. Letting $\delta = 1 - H$, we define the estimator as

$$\widehat{b}_M^p = \arg\min_{b_M^p \in \mathcal{B}_M} \frac{1}{nh_2}\sum_{i=1}^n s_i \left(y_i - g_M\left(x_i' b_M^p\right)\right)^2 K\left(\frac{\widehat{p}(z_i) - \delta_n}{h_2}\right).$$

Here, with an abuse of notation, x_i includes also the intercept.

In principle, we can establish consistency of $\widehat{b}_M^p$ for $b_{0,M}^p$, the probability limit, and establish asymptotic normality of a properly studentized version of $\left(\widehat{b}_M^p - b_{0,M}^p\right)$, along the lines of Theorem 3. The key difference is that now we need a much stronger version of Assumption E9. More precisely, assume for simplicity that the joint support of x_i is given as the Cartesian products of the marginal support. Then, $C_{n,x_0'\beta_0}$ should be replaced by

$$\mathcal{C}_{0,n} \equiv \otimes_{j=1}^{d_x}[x_{0,j} - \xi_{j,n}, x_{0,j} + \xi_{j,n}],$$

where for $j = 1, ..., d_x$, $\xi_{j,n}$ may either converge to zero to some non-zero constant. The case where $\otimes_{j=1}^{d_x}[x_{0,j} - \xi_{j,n}, x_{0,j} + \xi_{j,n}] = \otimes_{j=1}^{d_x}\text{supp}\,(x_j)$ for all $j \in \{1, \ldots, d_x\}$, corresponds to the case of strong support, as we require the propensity score to approach one for all x_s, $s = 1, \ldots, d_x$. On the other hand, in the case where instead $\xi_{j,n} \to 0$ for some but not all j, we require identification at infinity over a subset of non-zero Lebesgue of $\otimes_{j=1}^{d'_x}\text{supp}\,(x_j)$, with $d'_x < d_x$. Finally, if $\xi_{j,n} \to 0$ as $n \to \infty$ for all j, we have identification at infinity only over an interval shrinking to a singleton. Thus, the set of observation for which identification at infinity hold is no longer n_0, as defined in the previous sybsection, but n_{d_x} which is defined as

$$n_{d_x} = n\text{plim}_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n} 1\left\{x_i \in \mathcal{C}_{0,n}\right\}.$$

Now in the case of large support it still holds for all $j \in \{1, \ldots, d_x\}$ that $n_{d_x} = n$, however if $\xi_{j,n} \to 0$ for all $j = 1, ..., d_x$, then $n_{d_x} = n\xi_n^{d_x}$. Hence, the rate of convergence can be as slow as $\sqrt{n\xi_n^{d_x}h_2H^{\eta}}$. Concluding, in the non monotonic, non-separable case, provided all the rate conditions in Theorem 3 hold with n_0 replaced by n_{d_x}, consistency and asymptotic normality of intercept and slope parameters follow.

5 Monte Carlo

In this section we evaluate the finite sample performance of the estimators proposed in Sections 2 to 4. In particular, we assess their robustness w.r.t. the choice of the main tuning parameter(s), the bandwidth(s), and different degrees of selection.

We start by outlining the Monte Carlo design, which is similar to that of Jochmans (2015). As in Section 2, we assume a selection equation of the form:

$$s_i = 1\{z_i'\gamma_0 > v_i\},$$

where $z_i = (z_{1i}, z_{2i})'$ with:

$$\begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} .75 & \rho_z\sqrt{0.5} \\ \rho_z\sqrt{0.5} & .75 \end{pmatrix}\right)$$

and $\gamma_0 = (1, 1)'$ as well as $\rho_z = -.5$. The outcome equation is given by:

$$\text{E}[y_i|\widetilde{\varepsilon}_i, s_i = 1] = \exp(\theta_{0M})\widetilde{\varepsilon}_i.$$

Selection is modeled in this set-up through the correlation of $w_i = \log(\widetilde{\varepsilon}_i)$ and v_i, which are distributed as follows:

$$\begin{pmatrix} w_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_w^2 & \rho\sigma_w \\ \rho\sigma_w & 1 \end{pmatrix}\right)$$

where $|\rho| < 1$. Note that the unconditional mean of $\widetilde{\varepsilon}_i$ in the above equation is given by $\exp(\sigma_w^2/2)$. We therefore set θ_{0M} equal to $\exp(-\sigma_w^2/2)$, so that the unconditional mean of the outcome equation

equals one. We consider two sample sizes $n = \{600; 1,000\}$. Given that the probability of selection is approximately .5 in our set-up, this implies an effective sample size for the outcome equation of around 300 to 500 observations, respectively.

In what follows, we assess the performance of our estimators under three different designs, namely $\rho = 0$ (no selection), $\rho = -.5$, and $\rho = +.5$. In a first step, we compare the estimator of Sections 2 and 3, which we label 'Bias Corrected Least Squares' (LS Cr.) estimator, with a standard Nonlinear Least Squares (NLS) estimator, which ignores selection altogether. To implement the bias corrected version proposed in this paper, we estimate the coefficient vector γ_0 using the Klein and Spady (1993) estimator, while $\widehat{F}_{z'\gamma}(\cdot)$ is constructed using the empirical distribution function.[8] For the estimation of the outcome equation, we consider two different designs: one with a fixed bandwidth grid $[0.75, 1.00, 1.25, 1.50]$ and one where we estimate the bandwidth as a parameter $(\widehat{h})$ jointly with the intercept. Estimating the bandwidth as a parameter is similar to Härdle et al. (1993) who demonstrated that this procedure possessed certain optimality properties in the context of a least squares estimator for semiparametric single index models. Although we do not claim that the same type of optimality carries over to our set-up, we find that this procedure works well in the simulations, in particular w.r.t. the Root Mean Squared Error (RMSE).

Tables 1 and 2 below provide the first set of simulations, where we compare average mean and median bias as well as RMSE for both estimators over 1,500 replications. As expected, in the absence of selection, the NLS estimator clearly outperforms our bias corrected estimator in all three measures. By contrast, when examining $\rho = -.5$ and $\rho = +.5$, this picture is reversed and we observe a substantially larger mean and median bias, which does not diminish when the sample size increases. Moreover, note that the performance of the corrected least squares estimator when the bandwidth is treated as a parameter $(\widehat{h})$ is overall comparable to the fixed bandwidth case in terms of mean and median bias, and interestingly even outperforms the latter in terms of RMSE. This feature may be an interesting subject for future research.

Next we turn to the non-monotonic case discussed in Section 4, where we estimate the intercept using a local nonlinear least squares estimator with a fully nonparametric propensity score in place of the marginal distribution function of $z_i'\gamma_0$. As before, we consider the three cases $\rho = 0$, $\rho = -.5$ and $\rho = +.5$, but this time focus on the performance of the estimator w.r.t. the trimming procedure: changing the evaluation point δ from .925 to .950 and .975, we match this with three different bandwidth choices, namely $h_2 = 0.05$, $h_2 = 0.075$, and $h_2 = 0.1$. Thus, by construction, no observations from the boundary are used when $\delta = .925$ or $\delta = .95$ and $h_2 = .05$ (or when $h_2 = .75$ and $\delta = .925$). For comparison reasons, we again also assess the performance of the local nonlinear least squares estimator when the bandwidth is treated as a parameter to be estimated $(\widehat{h})$. Finally, note that in all scenarios, the nonparametric propensity score is estimated using a standard Nadaraya-Watson estimator with a cross-validated bandwidth.

Turning to the results in Tables 3 and 4, observe that, when there is no selection, choosing a larger δ and moving closer to the boundary leads to an increase in terms of mean and median bias as well as RMSE. This does of course not come at a surprise, and in fact we do no longer observe this pattern

[8] As starting values for the Klein and Spady (1993) estimator we use Probit estimates, while the bandwidth is chosen via the 'rule of thumb' bandwidth proposed in the original paper, i.e. $C \cdot n^{-1/6.02}$ (the scaling constant C is chosen over a grid $[0.5, 0.75, 1.0]$, where the value that maximizes the likelihood is retained).

when selection is indeed present. Moreover, note that, choosing a larger bandwidth and thus giving more weight to observations at the boundary (and in the interior), does in general not lead to very different results in any of the aforementioned measures. By contrast, we do however note a somewhat less pronounced reduction in terms of mean and median bias compared to Tables 1 and 2 when moving from $n = 600$ to $n = 1,000$. This may be related to the slower convergence rate of the local nonlinear least squares estimator relative to the bias corrected estimator from Sections 2 and 3. Finally, note that jointly estimating the bandwidth parameter alongside θ_{0M} does in general yield much poorer (relative) results compared to the fixed bandwidth choice. This is in contrast to the previous case, where we found estimation of the bandwidth to work well.

Table 1: Bias Corrected Least Squares (Sample Size: $n = 600$)

		Mean Bias					Median Bias					RMSE				
		0.075	0.1	0.125	0.15	$\widehat{h}$	0.075	0.1	0.125	0.15	$\widehat{h}$	0.075	0.1	0.125	0.15	$\widehat{h}$
$\rho = 0$	LS Cr.	−0.1596	−0.1021	−0.0734	−0.0579	−0.0754	−0.0573	−0.0453	−0.036	−0.0262	−0.0603	0.7689	0.5505	0.437	0.3269	0.2322
	NLS			−0.0044					−0.0032					0.0743		
$\rho = -.5$	LS Cr.	−0.2088	−0.1426	−0.117	−0.1051	−0.1307	−0.0917	−0.0743	−0.0659	−0.0688	−0.1056	0.9144	0.5779	0.42	0.3467	0.2645
	NLS			−0.2378					−0.2386					0.2505		
$\rho = +.5$	LS Cr.	−0.1176	−0.0677	−0.0462	−0.0361	−0.0449	−0.0337	−0.0155	−0.0091	−0.0033	−0.0374	0.7113	0.4883	0.4354	0.4088	0.2315
	NLS			0.1839					0.1840					0.1965		

Notes: (1) Number of Monte Carlo replications: 1,500; (2) $h = 0.075, 0.1, 0.125, 0.15$ correspond to different fixed bandwidth sizes; (3) $\rho = 0$ corresponds to the no selection case.

Table 2: Bias Corrected Least Squares (Sample Size: $n = 1,000$)

		Mean Bias					Median Bias					RMSE				
		0.075	0.1	0.125	0.15	$\widehat{h}$	0.075	0.1	0.125	0.15	$\widehat{h}$	0.075	0.1	0.125	0.15	$\widehat{h}$
$\rho = 0$	LS Cr.	−0.0735	−0.053	−0.0406	−0.0339	−0.0466	−0.0434	−0.0335	−0.0241	−0.0185	−0.0336	0.3377	0.2828	0.2513	0.2314	0.1704
	NLS			0.0004					0.0000					0.0588		
$\rho = -.5$	LS Cr.	−0.0982	−0.0823	−0.0749	−0.0719	−0.0959	−0.0633	−0.0566	−0.0562	−0.0592	−0.0834	0.3485	0.2907	0.2612	0.2404	0.1901
	NLS			−0.2350					−0.2333					0.2428		
$\rho = +.5$	LS Cr.	−0.0555	−0.0383	−0.0272	−0.0199	−0.0324	−0.0287	−0.0166	−0.0098	−0.0061	−0.0203	0.3406	0.2881	0.2566	0.2332	0.1749
	NLS			0.1886					0.1877					0.1961		

Notes: (1) Number of Monte Carlo replications: 1,500; (2) $h = 0.075, 0.1, 0.125, 0.15$ correspond to different fixed bandwidth sizes; (3) $\rho = 0$ corresponds to the no selection case.

Table 3: Local Nonlinear Least Squares (Sample Size: $n = 600$)

		Mean Bias				Median Bias				RMSE			
		0.05	0.075	0.1	$\widehat{h}$	0.05	0.075	0.1	$\widehat{h}$	0.05	0.075	0.1	$\widehat{h}$
$\rho = 0$	$\delta = 0.925$	−0.0525	−0.0308	−0.0191	−0.1242	−0.0268	−0.0186	−0.0128	−0.0649	0.3417	0.2528	0.2009	0.4799
	$\delta = 0.950$	−0.0798	−0.0380	−0.0273	−0.1546	−0.0417	−0.0171	−0.0167	−0.0829	0.4749	0.2783	0.2328	0.5530
	$\delta = 0.975$	−0.1082	−0.0655	−0.0401	−0.1946	−0.0424	−0.0260	−0.0229	−0.0729	0.6038	0.4207	0.2842	0.8551
$\rho = -.5$	$\delta = 0.925$	−0.1231	−0.0997	−0.0897	−0.1924	−0.0933	−0.0897	−0.0873	−0.1419	0.3495	0.2679	0.2201	0.4865
	$\delta = 0.950$	−0.1403	−0.0968	−0.0893	−0.2323	−0.1046	−0.0790	−0.0802	−0.1425	0.4862	0.2888	0.2472	0.6734
	$\delta = 0.975$	−0.1442	−0.1154	−0.0959	−0.2034	−0.0791	−0.0808	−0.0766	−0.1127	0.5703	0.4122	0.2945	0.6484
$\rho = +.5$	$\delta = 0.925$	−0.0106	0.0073	0.0203	−0.0692	0.0077	0.0153	0.0222	−0.0363	0.3442	0.2423	0.1959	0.3910
	$\delta = 0.950$	−0.0482	0.0006	0.0099	−0.1193	−0.0042	0.0070	0.0118	−0.0511	0.5180	0.2698	0.2277	0.5745
	$\delta = 0.975$	−0.0594	−0.0246	−0.0016	−0.1098	−0.0023	0.0054	0.0065	−0.0353	0.5465	0.4064	0.2775	0.6172

Notes: (1) Number of Monte Carlo replications: 1,500; (2) $h = 0.05, 0.75, 0.1$ correspond to different fixed bandwidth sizes; (3) $\rho = 0$ corresponds to the no selection case.

Table 4: Local Nonlinear Least Squares (Sample Size: $n = 1,000$)

		Mean Bias				Median Bias				RMSE			
		0.05	0.075	0.1	$\widehat{h}$	0.05	0.075	0.1	$\widehat{h}$	0.05	0.075	0.1	$\widehat{h}$
$\rho = 0$	$\delta = 0.925$	−0.0227	−0.0115	−0.0076	−0.0659	−0.0171	−0.0137	−0.0079	−0.0487	0.2303	0.1766	0.1456	0.2407
	$\delta = 0.950$	−0.0290	−0.0164	−0.0109	−0.0828	−0.0135	−0.0063	−0.0103	−0.0536	0.2659	0.1949	0.1662	0.2955
	$\delta = 0.975$	−0.0370	−0.0245	−0.0177	−0.0718	−0.0169	−0.0136	−0.0075	−0.0415	0.2879	0.2371	0.1993	0.2852
$\rho = -.5$	$\delta = 0.925$	−0.0906	−0.0817	−0.0790	−0.1370	−0.0813	−0.0727	−0.0746	−0.1171	0.2530	0.2031	0.1729	0.2744
	$\delta = 0.950$	−0.0812	−0.0734	−0.0741	−0.1325	−0.0692	−0.0676	−0.0655	−0.1028	0.2718	0.2117	0.1893	0.2993
	$\delta = 0.975$	−0.0805	−0.0737	−0.0720	−0.1198	−0.0585	−0.0636	−0.0633	−0.0915	0.2999	0.2499	0.2152	0.2999
$\rho = +.5$	$\delta = 0.925$	0.0064	0.0238	0.0318	−0.0339	0.0260	0.0286	0.0383	−0.0067	0.2691	0.1735	0.1452	0.2733
	$\delta = 0.950$	−0.0065	0.0109	0.0208	−0.0578	0.0136	0.0206	0.0260	−0.0232	0.2571	0.1888	0.1631	0.3112
	$\delta = 0.975$	−0.0185	−0.0015	0.0089	−0.0525	0.0131	0.0196	0.0204	−0.0143	0.3152	0.2258	0.1922	0.3185

Notes: (1) Number of Monte Carlo replications: 1,500; (2) $h = 0.05, 0.75, 0.1$ correspond to different fixed bandwidth sizes; (3) $\rho = 0$ corresponds to the no selection case.

6 Emprirical Illustration

We provide an empirical illustration using a data set from the 2003 Medical Expenditure Panel Survey (MEPS) used in Cameron and Trivedi (2013). The MEPS is a representative U.S. survey containing information on demographic characteristics, health status, earnings, and a range of measures of health care utilization. For this illustration, we focus on a specific measure, namely the number of annual doctor visits (DOCVIS), which includes visits to a primary care physician or specialist in an office setting, but excludes hospital stays or emergency room visits. We limit ourselves to a sample of elderly aged 65 and higher, who are eligible for Medicare. Medicare offers basic health insurance to individuals who have worked and paid into the program through payroll tax. Our aim is to study the average treatment effect of additional private health insurance (PRIVATE) that individuals obtain to supplement Medicare coverage on the number of physician contacts in a given year.[9] Since individuals are likely to purchase this additional health insurance on the grounds of their private health status, and insurance coverage may by contrast only be offered to individuals satisfying certain health criteria by the insurance companies, self selection into private insurance is likely to be an issue in this context. We use total household income (INCOME) and Ratio of Social Security income to total income (SSIRATIO) as instruments. The first variable is likely to be positively correlated with PRIVATE as higher household income may make private insurance more accessible, while a higher value for SSIRATIO may indicate an income constraint and is expected to be negatively related with private insurance coverage. Validity of these instrumental variables relies on the assumption that both variables do not directly impact the number of doctor visits conditional on other covariates, which could in principle be tested formally using the nonparametric test of Kitagawa (2010).

The overall sample size is 3,628 individuals, and the unconditional mean of DOCVIS (6.74) is less than its variance (45.58). This feature, typically labeled over-dispersion, may be an indication that a standard Poission model, unable to capture this, may in principle not be well suited to model the outcome. Finally, the additional control variables we consider in the outcome as well as selection equation (in addition to INCOME and SSIRATIO) are age (AGE), age^2 (AGE2), number of years of education (EDUCYR), activity limitation (ACTLIM), no of chronic conditions (TOTCHR).

We implement our intercept estimator for the treatment (with private insurance) and control (without private insurance) group as described in the previous sections:

- We recover γ_0 using Klein and Spady (1993) with INCOME and SSIRATIO as instruments, and the aforementioned covariates as additional regressors.[10]
- We replace $F_{z'\gamma}(\cdot)$ by its empirical counterpart.
- Using the standard deviation of a nonparametric estimate $\widehat{p}(z_i'\widehat{\gamma})$ as ad-hoc bandwidth choice, $\widehat{\beta}_M$ is estimated in a subsequent step from the subsample of individuals with private health insurance along the lines of Jochmans (2015).

[9] PRIVATE is a dummy variable, which indicates that the person is enrolled in some supplementary private insurance plan. For this illustration, we do not differentiate between different insurance plans, nor between the degree of coverage which insurance plans might offer.

[10] The bandwidth is chosen to be $\text{std}(z_i'\widehat{\gamma}_{\text{prob}})n^{-\frac{1}{6.02}}$ as recommended by Klein and Spady (1993), where std$(\cdot)$ denotes the standard deviation of the estimated instrument index using a Probit specification (we found that different norming constants did not have a large impact on the resulting estimates $\widehat{\gamma}$).

- We construct the average treatment effect of PRIVATE as the relative change in the number of physician contacts using the estimated intercepts from treatment and control group (covariate effects are taken to be the same for both groups). Without claiming optimality of any kind, we estimate the bandwidth alongside the intercept as a parameter. This procedure, which is used also in other parts of the semiparametric literature (e.g., Härdle et al., 1993), was shown to work well in the simulations.

All estimates are presented in Table 5, and standard errors are constructed using the delta method. For simplicity, we only present results for the estimated relative change in doctor visits due to PRIVATE, which we calculated as in Remark 1 for our estimator. Columns (I) and (II) contain estimates for a standard poisson and a negative binomial specification not accounting for selection. While gamma unobserved heterogeneity is found to be significant, both effects are rather similar and indicate a significant increase in the number of doctor of visits of around 15%. By contrast, columns (III) and (IV) contain estimates from control function estimators, where the conditional mean has the same parametric form as in columns (I) and (II), respectively, but selection is controlled for through a linear control function argument (Terza et al., 2008; Wooldridge, 1997).[11] As can be inferred from Table 5, while (III) suggests an increase in the effect to about 27%, the effect is no longer significant at any conventional level. By contrast, the negative binomial model paired with the linear control function argument in (IV) yields an increase to about 42%, which is roughly significant at the 5-10% level. The last estimator based on a parametric specification in (V) is the nonlinear least squares estimator proposed by Terza (1998), where the joint distribution of the unobservables of the outcome and selection equation is modeled as a multivariate normal distribution. As before, we obtain estimates which are no longer significant, although the magnitudes are similar to the original level of columns (I) and (II). Finally, we turn to (VI), the bias corrected estimator outlined in Section 2. In line with but different from columns (III), (IV), and (V) we find that the effect is not significantly different from zero, and in fact its magnitude is very close to zero as well.

Examining Figure 1 below, we can observe that a specification based on $z_i'\widehat{\gamma}$ may be problematic in that the estimated propensity score does not reach one and zero, respectively. More importantly even, it appears that, especially in the tails, the estimated propensity score is not a monotonic function of $z_i'\widehat{\gamma}$, which could indicate mis-specification. In a second step, we therefore estimate a fully nonparametric specification of the propensity score using the `np` package in `R` introduced by Hayfield and Racine (2008) with a completely data-driven bandwidth choice. As can be inferred from Figure 2, allowing for this more flexible specification clearly increases the number of observations in the tails underlining the strength of more flexible modeling. Turning to the last column (VII) of Table 5, we see that the estimated effect of private supplementary health insurance increases marginally w.r.t. (VI), but is still insignificant and far from the other estimates.[12] We interpret this closeness of the estimates in (VI) and (VII) as evidence for robustness of our estimators. That is, given the wide range of results obtained from the different parametric models even among those controlling for selection, we believe that this illustration highlights the need for a more flexible modeling approach to draw robust

[11] Standard errors are bootstrapped with 400 replications to account for estimation error in the first stage (cf. Wooldridge, 1997).

[12] We have chosen the evaluation points $\delta_1 = 0.95$ and $\delta_2 = .05$ for the treatment and control group, respectively.

conclusions.

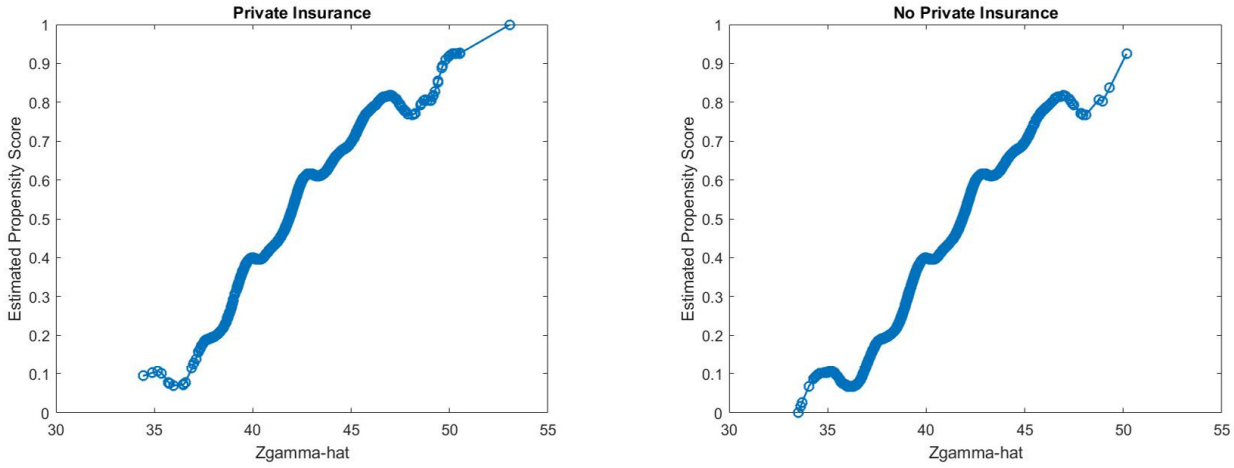
From an economic perspective, one reason for the lack of significance of private supplementary health insurance on the number of annual doctor contacts when controlling for selection may be the types of insurance plans offered as well as the inability to differentiate between them.[13] Finally, while supplementary private health insurance does not appear to increase the overall number of doctor visits significantly, it may of course lead to other (undetected) effects: for instance, it may have reduced waiting times or led to substitution effects in the type of doctor visits (private vs. Medicare doctors).

Table 5: Estimated Relative Change in Doctor Visits

Model	(I)	(II)	(III)	(IV)	(V)	(VI)	(VII)
Est.Rel.Change	0.1572	0.1782	0.3061	0.5170	0.1664	0.0057	0.0121
(S.E.)	(0.0381)	(0.0402)	(0.2678)	(0.3161)	(0.2045)	(0.0686)	(0.0130)

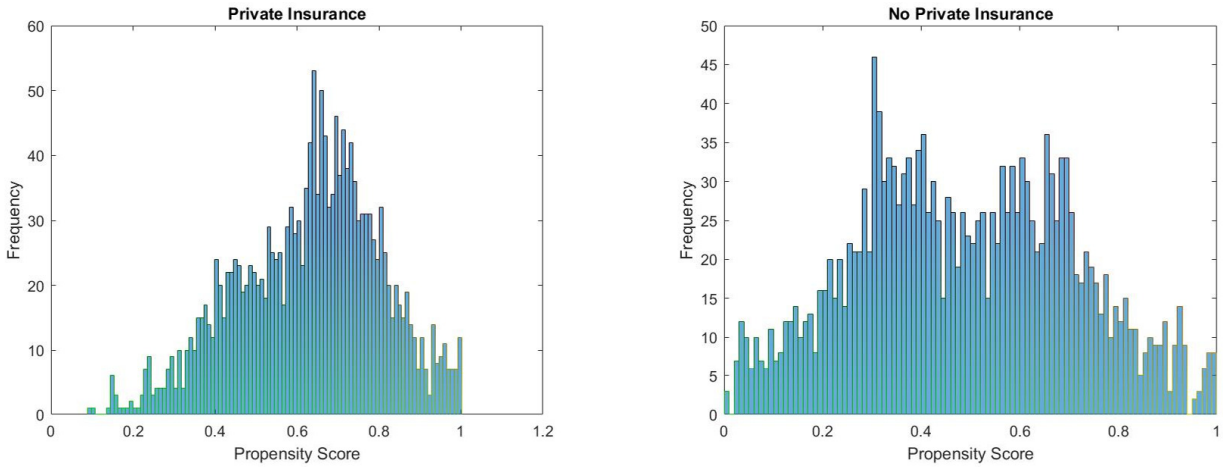
Notes: (1) Sample Size: 3,628 individuals; (2) (VI) - Individuals used for $\widehat{m}_M(1^-)$: 92 priv.ins. vs. 103 w/o priv.ins.; (3) (VII) - Individuals used for $\widehat{m}_M(1^-)$: 45 priv.ins. vs. 45 w/o priv. ins.;

Figure 1: Estimated Propensity Score with Index



[13]That is, since many insurance plans may still require co-payment or the payment of a certain fee per doctor visit, the effect on physician contacts may still be negligible as payments relative to standard payments under Medicare may still be high. This fact is likely to be aggravated by the inability to differentiate between different types of insurance plans.

Figure 2: Estimated Nonparametric Propensity Score



7 Conclusion

Identification and estimation of the intercept is crucial for the evaluation of average treatment effects in non-experimental settings, which are often subject to selection on unobservables (Heckman, 1990). While various estimators for linear additive sample selection models exist, many other data types, which are also affected by endogenous selection, are modeled nonlinearly. This paper introduces the first estimators of the intercept in general nonlinear semiparametric selection models, where the joint distribution of the error terms remains unknown. We consider multiplicative and general non-additive models and propose two different types of estimators depending on whether the selection equation satisfies a linear index restriction or not: in the first case where the index restriction holds, our estimator is based on a least squares criterion function with a nonparametric correction for the selection bias term. This estimator is shown to converge at a univariate nonparametric rate even in cases of irregular identification. In the second case, we relax the index restriction in the selection equation and base our estimator on a more flexible nonparametric specification of the propensity score. The resulting estimator is a local nonlinear least squares estimator, which uses observations close but not too close to the boundary. While robust against mis-specification of the first stage, this estimator requires trimming and is shown to converge generally at a slower rate than the first one. Finally, we investigate the problem of self-selection into health insurance using U.S. survey data. While our estimators confirm the conclusion that supplementary private health insurance does not increase the number of physician contacts when controlling for selection, we obtain different results from those of various parametric models (not) controlling for selection. This highlights the need for more flexible model specifications as advocated in this paper.

8 Appendix

In the following, we use ∇_F^i to denote the i-th order derivative(s) w.r.t. the argument F. Moreover, for $0 \leq t \leq 2q$, let:

$$\mu_{1,t}(K) = \int_{-\infty}^{0} \nu^t K(\nu) d\nu$$

as well as

$$\gamma_t(K) = \int_{-\infty}^{0} \nu^t K^2(\nu) d\nu.$$

Also, define the $q \times q$ dimensional matrix:

$$\mathbf{M}_1^1 = \begin{bmatrix} \mu_{1,0}(K) & \dots & \mu_{1,q}(K) \\ \vdots & \ddots & \vdots \\ \mu_{1,q}(K) & \dots & \mu_{1,2q}(K) \end{bmatrix} \tag{12}$$

The matrix $\mathbf{\Gamma}^1$ is defined accordingly, but contains elements $\gamma_j(k)$ instead of $\mu_{1,j}(k)$.

To prove Theorem 1, we require Lemma 1 and Lemma 2 below.

Lemma 1: Let Assumption A1-A2, E1-E7 hold. If as $n \to \infty$, $nh^{2(q+1)+1} \to 0$ and $nh \to \infty$, then

$$\sqrt{nh}\left(\widehat{m}_M(1^-) - g_{M1}\left(\theta_{0M}\right)\right) \xrightarrow{d} N\left(0, \sigma^2(1)\right)$$

where $\sigma_1^2 = \mathrm{E}\left[\frac{s_j \widetilde{u}_j^2}{g_{M1}^2\left(x_j'\beta_{0M}\right)} \left| F_{z'\gamma}\left(z_j'\gamma_0\right) = 1^-\right.\right] \left[\mathbf{M}_1^{-1}\mathbf{\Gamma}^1\mathbf{M}_1^{-1}\right]_{00}$, with $[A]_{00}$ denoting the upper left entry of matrix A, $\mathbf{M}_1^1$ and $\mathbf{\Gamma}^1$ are defined above.

Proof of Lemma 1: Recall that $\mathrm{E}\left[s_i = 1 | F_{z'\gamma_0}\left(z_i'\gamma_0\right) = 1^-\right] = 1$. Given Assumptions A2(ii)-(iii), $\lim_{\tau \to 1} F_{z'\gamma}^{-1}(\tau)$ implies $\lim_{\tau \to 1} F_v(\tau)$, and so $\mathrm{E}\left[s_i = 1 | F_v\left(z_i'\gamma_0\right) = 1^-\right] = 1$. Let $F_i = F_{z'\gamma_0}\left(z_i'\gamma_0\right)$ and $\widehat{F}_i = \frac{1}{n}\sum_{j=1}^n 1\left\{z_j'\widehat{\gamma}_0 \leq z_i'\widehat{\gamma}_0\right\}$. Moreover, let $\widetilde{m}(1^-)$ be defined as $\widehat{m}(1^-)$ in the text, with $\widehat{F}_j$ replaced by F_j. Finally, we write $\widehat{K}_j(1) = K((\widehat{F}_j - 1^-)/h)$, $\widehat{\mathcal{P}}_j(1) = \left(1, (\widehat{F}_j - 1)\frac{1}{h}, \dots, (\widehat{F}_j - 1)^q \frac{1}{q!h^q}\right)'$, and $\widehat{\mathcal{Y}}_j = y_j / g_{M2}(x_j'\widehat{\beta}_M)$, and let $K_j(1)$, $\mathcal{P}_j(1)$, and $\mathcal{Y}_j$ be defined accordingly with $\widehat{F}_j$ and $\widehat{\beta}_M$ replaced again by F_j and β_{0M}.
First note that $\widehat{m}_M(1^-)$ is defined as the first element of the $(q \times 1)$ vector

$$\left(\frac{1}{nh}\sum_{i=1}^n s_i \widehat{\mathcal{P}}_i(1)\widehat{K}_i(1)\widehat{\mathcal{P}}_i(1)'\right)^{-1}\left(\frac{1}{nh}\sum_{i=1}^n s_i \widehat{\mathcal{P}}_i(1)\widehat{K}_i(1)\widehat{\mathcal{Y}}_i(1)\right),$$

while $\widetilde{m}_M(1^-)$ is the first element of the corresponding $(q \times 1)$ vector

$$\left(\frac{1}{nh}\sum_{i=1}^n s_i \mathcal{P}_i(1)K_i(1)\mathcal{P}_i(1)'\right)^{-1}\left(\frac{1}{nh}\sum_{i=1}^n s_i \mathcal{P}_i(1)K_i(1)\mathcal{Y}_i(1)\right).$$

Also note that $g_{M1}\left(\theta_{0M}\right)$ is the probability limit of $\widetilde{m}_M(1^-)$. Given Assumption E5 and recalling that because of A1 and E1, the empirical process

$$\frac{1}{\sqrt{n}}\sum_{j=1}^n \left(1\left\{z_j'\gamma_0 \leq z_i'\gamma_0\right\} - F_{z'\gamma}(z_i'\gamma_0)\right)$$

satisfies a central limit for i.i.d. random variables. Thus, standard mean value expansion arguments (joint with the fact that for any two symmetric nonsingular matrices A_1 and A_2 it holds that $A_1^{-1} - A_2^{-1} = A_2^{-1}(A_2 - A_1)A_1^{-1}$) yield that:

$$\sqrt{nh}\left(\widetilde{m}_M(1^-) - \widehat{m}_M(1^-)\right) = o_p(1).$$

Moreover, recalling that the density of $F_{z'\gamma_0}\left(z_i'\gamma_0\right)$ is uniform on $[0,1]$, note that by E7:

$$\frac{1}{nh}\sum_{i=1}^n s_i \mathcal{P}_i(1)K_i(1)\mathcal{P}_i(1)' \xrightarrow{p} \mathbf{M}_1.$$

Thus:

$$\begin{aligned}
& \sqrt{nh}\left(\widetilde{m}_M(1^-) - g_{M1}\left(\theta_{0M}\right)\right) \\
= \quad & \mathbf{M}_1{}^{-1}\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\left(s_j\mathcal{P}_j(1)K_j\left(1\right)\frac{y_j}{g_{M2}(x_j'\beta_{0M})} - g_{M1}\left(\theta_{0M}\right)\right) + o_p(1) \\
= \quad & o_p(1) + \mathbf{M}_1{}^{-1}\frac{1}{\sqrt{nh}}\sum_{j=1}^{n} s_j K_j\left(1\right)\frac{\widetilde{u}_j}{g_{M2}(x_j'\beta_{0M})} \\
& + \mathbf{M}_1^{-1} g_{M1}\left(\theta_{0M}\right)\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\left(s_j\mathcal{P}_j(1)K_j\left(1\right)\left(\left(\widetilde{\lambda}\left(F_j\right)-1\right) + \sum_{k=1}^{q}\frac{1}{\mathbf{k}}\nabla_F^{\mathbf{k}}\lambda\left(F_j\right)|_{F_j=1}(F_j-1)^{\mathbf{k}}\right)\right) \\
= \quad & o_p(1) + I_{n,h} + II_{n,h}
\end{aligned}$$

Given Assumption E4 and E7, $II_{n,h} = O_p\left(nh^{2(q+1)+1}\right)$. Finally, given E1 and E2, $I_{n,h}$ converges in distribution to a zero mean normal with the variance as in the statement of the Lemma.

Lemma 2: Let Assumption A1–A2, E3–E7 hold. If as $n \to \infty$, $nh^{2(q+1)+1} \to 0$ and $nh \to \infty$, then

$$\frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widehat{m}_M\left(\widehat{F}_{z'\gamma_0}\left(z_i'\widehat{\gamma}\right)\right) - g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)\right)\nabla_{\theta_M} g_{M1}\left(\theta_{0M}\right) = o_p(1)$$

Proof of Lemma 2: Let $\widetilde{m}(F_{z'\gamma_0}\left(z_i'\gamma\right))$ be defined as $\widehat{m}_M\left(F_{z'\gamma_0}\left(z_i'\gamma\right)\right)$, but with $\widehat{F}_j$ replaced by F_j. Given Assumption E5,

$$\begin{aligned}
& \frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widehat{m}_M\left(\widehat{F}_{z'\gamma_0}\left(z_i'\widehat{\gamma}\right)\right) - g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)\right)\nabla_{\theta_M} g_{M1}\left(\theta_{0M}\right) \\
= \quad & \frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widehat{m}_M\left(\widehat{F}_{z'\gamma_0}\left(z_i'\widehat{\gamma}\right)\right) - \widetilde{m}_M\left(F_{z'\gamma_0}\left(z_i'\gamma\right)\right)\right)\nabla_{\theta_M} g_{M1}\left(\theta_{0M}\right) \\
& + \frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widetilde{m}_M\left(F_{z'\gamma_0}\left(z_i'\gamma\right)\right) - g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)\right)\nabla_{\theta_M} g_{M1}\left(\theta_{0M}\right) \\
= \quad & o_p(1) + \frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widetilde{m}_M\left(F_{z'\gamma_0}\left(z_i'\gamma\right)\right) - g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)\right)\nabla_{\theta_M} g_{M1}\left(\theta_{0M}\right)
\end{aligned}$$

The statement follows if we show that

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widetilde{m}_M\left(F_{z'\gamma_0}\left(z_i'\gamma\right)\right) - g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)\right)\nabla_{\theta_M} g_{M1}(\theta_{0M}) = O_p(1). \tag{13}$$

This however follows once we show that the left hand side of (13) converges in distribution to a zero mean normal random variable.
First, note that:

$$\begin{aligned}
& E\left[s_j K_h\left(\frac{F_{z'\gamma}\left(z_j'\gamma\right) - F_{z'\gamma}\left(z_i'\gamma\right)}{h}\right)\right] \\
= \quad & f_{z'\gamma_0}\left(z'\gamma\right)\Pr\left(s=1|F_{z'\gamma}\left(z_i'\gamma\right)\right) + o(1) = \Pr\left(s=1|F_{z'\gamma}\left(z_i'\gamma\right)\right) + o(1).
\end{aligned}$$

For notational simplicity, hereafter we omit the constant term $\nabla_{\theta_M} g_{M1}(\theta_{0M})$. Then, given that A2 implies that

$\Pr\left(s=1|F_{z'\gamma}\left(z_i'\gamma\right)\right)>0$ a.s.,

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widetilde{m}_M\left(F_i\right)-g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(F_i\right)\right) \\
= \quad & \mathbf{M}_1^{-1}\frac{1}{n^{\frac{3}{2}}h}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{s_i}{\Pr\left(s=1|F_{z'\gamma}\left(z_i'\gamma\right)\right)}s_j\mathcal{P}_j(F_i)K_j\left(F_i\right)\frac{\widetilde{u}_j}{g_{M2}(x_j'\beta_{0M})} \\
& +\mathbf{M}_1^{-1}\frac{1}{n^{\frac{3}{2}}h}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{s_i g_{M1}\left(\theta_{0M}\right)}{\Pr\left(s=1|F_{z'\gamma}\left(z_i'\gamma\right)\right)} \\
& \times s_j\mathcal{P}_j(F_i)K_j\left(F_i\right)\left(\widetilde{\lambda}\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)-\sum_{k=1}^{q}\frac{1}{\mathbf{k}}\nabla_F^{\mathbf{k}}\lambda\left(F_i\right)\left(F_j-F_i\right)^{\mathbf{k}}\right)+o_p(1) \\
= \quad & I_{n,h}+II_{n,h}
\end{aligned}$$

Given Assumption E4 and E7, $II_{n,h}=O_p\left(nh^{2(q+1)+1}\right)$. Let $\phi_i=\mathbf{M}_1^{-1}\frac{1}{\Pr(s=1|F_i)}$ define,

$$\Phi_{i,j}=\frac{1}{h}s_i\phi_i s_j\mathcal{P}_j(F_i)K_j\left(F_i\right)\frac{\widetilde{u}_j}{g_{M2}(x_j'\beta_{0M})}+s_i\phi_j s_j\mathcal{P}_i(F_j)K_i\left(F_j\right)\frac{1}{h}\frac{\widetilde{u}_i}{g_{M2}(x_i'\beta_{0M})}$$

and so

$$\begin{aligned}
I_{n,h} &= \frac{2}{n^{3/2}}\sum_{i=1}^{n}\sum_{j>i}\Phi_{i,j}+\frac{1}{n^{3/2}}\sum_{i=1}^{n}s_i\phi_i\mathcal{P}(0)K\left(0\right)\frac{\widetilde{u}_i}{g_{M2}(x_i'\beta_{0M})} \\
&= \frac{2}{(n-1)n^{1/2}}\sum_{i=1}^{n}\sum_{j>i}\Phi_{i,j}+o_p(1)
\end{aligned}$$

Finally, observe that $\mathrm{E}\left[\Phi_{i,j}^2\right]=O\left(h^{-1}\right)=o(n)$ for $nh\to\infty$. Therefore, applying standard Hoeffding decomposition arguments to this second order U-statistic (e.g., Powell et al., 1989), and noting that

$$\mathrm{E}\left[\Phi_{i,j}|\overline{\omega}_j\right]=s_j\frac{\widetilde{u}_j}{g_{M2}(x_j'\beta_{0M})}\mathrm{E}\left[\phi_i s_i\frac{1}{h}\mathcal{P}_i(F_j)K_i\left(F_j\right)|\overline{\omega}_i\right],$$

where $\overline{\omega}_i=(F_i,s_i)'$, by the Lindeberg-Levy CLT we obtain that $\frac{2}{\sqrt{n}}\sum_{i=1}^{n}\mathrm{E}\left[\Phi_{i,j}|\overline{\omega}_j\right]$ converges to a zero mean normal random variable. This in turn establishes (13).

Proof of Theorem 1:
(i) In order to show consistency, we first need to show that θ_{0M} is uniquely identified. Define:

$$\begin{aligned}
\widetilde{\theta}_M &= \arg\min_{\theta_M\in\Theta_M}\widetilde{S}_n\left(\theta_M\right) \\
&= \arg\min_{\theta_M\in\Theta_M}\frac{1}{n}\sum_{i=1}^{n}s_i\left(\frac{y_i}{g_{M2}(x_i'\beta_0)}\frac{m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma_0))}-g_{M1}(\theta_M)\right)^2 \\
&= \arg\min_{\theta_M\in\Theta_M}\frac{1}{n}\sum_{i=1}^{n}s_i\left(\frac{\widetilde{u}_i}{g_{M2}(x_i'\beta_0)}+g_{M1}(\theta_{0M})-g_{M1}(\theta_M)\right)^2
\end{aligned}$$

Given Assumptions A1-A2 and E1 as well as E6, by a uniform law of large numbers:

$$\begin{aligned}
\sup_{\theta_M\in\Theta_M}\Bigg|&\frac{1}{n}\sum_{i=1}^{n}s_i\left(\frac{\widetilde{u}_i}{g_{M2}(x_i'\beta_0)}+g_{M1}(\theta_{0M})-g_{M1}(\theta_M)\right)^2 \\
&-\mathrm{E}\left[s_i\left(\frac{\widetilde{u}_i}{g_{M2}(x_i'\beta_0)}+g_{M1}(\theta_{0M})-g_{M1}(\theta_M)\right)^2\right]\Bigg|\xrightarrow{p}0
\end{aligned}$$

Given the invertibility of g_{M1}, Assumption E2, and by the first order conditions, it follows that $\theta_M=\theta_{0M}$ is the unique minimizer of

$$\mathrm{E}\left[s_i\left(\frac{\widetilde{u}_i}{g_{M2}(x_i'\beta_0)}+g_{M1}(\theta_{0M})-g_{M1}(\theta_M)\right)^2\right].$$

This establishes unique identification of θ_{0M}. Hence, $\widetilde{\theta}_M=\theta_{0M}+o_p(1)$. It remains to show that $\widehat{\theta}_M-\widetilde{\theta}_M=o_p(1)$.

In what follows, define:

$$\widehat{S}_n(\theta_M) = \frac{1}{n}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{\widehat{m}_M(1^-)}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} - g_{M1}(\theta_M) \right)^2$$

and let $\widehat{S}_n(\theta_M)$ be defined accordingly, but with $\widehat{\beta}_M$, $\widehat{F}_i$, and $\widehat{F}_j$ replaced by β_{0M}, F_i, and F_j. Now, simple arithmetic gives

$$\begin{aligned}
&\widehat{S}_n(\theta_M) \\
= \;& \frac{1}{n}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma_0))} - g_{M1}(\theta_M) \right)^2 \\
&+ \frac{1}{n}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma))} \left(\frac{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - m_M(F_{z'\gamma}(z_i'\gamma_0))}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} \right) \right)^2 \\
&+ \frac{1}{n}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \left(\frac{\widehat{m}_M(1^-) - m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma_0))} \right) \right)^2 \\
&+ \frac{1}{n}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \left(\frac{\widehat{m}_M(1^-) - m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma))} \right) \left(\frac{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - m_M(F_{z'\gamma}(z_i'\gamma_0))}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} \right) \right)^2. \qquad (14)
\end{aligned}$$

Given Lemma 1 and Lemma 2, the second, third and fourth term on the RHS of (14) are $o_p(1)$, which holds uniformly in Θ_M. Also, given Assumptions A1-A2, E1-E3, and E5,

$$\sup_{\theta_M \in \Theta_M} \left| \frac{1}{n}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma))} - g_{M1}(\theta_M) \right)^2 - \widetilde{S}_n(\theta_M) \right| \xrightarrow{p} 0$$

This establishes consistency.

(ii) By first order conditions, for $\overline{\theta}_M \in \left(\widehat{\theta}_M, \theta_{0M}\right)$, and given (14),

$$\begin{aligned}
0 &= \sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{\widehat{m}_M(1^-)}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} - g_{M1}(\widehat{\theta}_M) \right) \nabla_{\theta_M} g_{M1}(\widehat{\theta}_M) \\
&= \sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{\widehat{m}_M(1^-)}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} - g_{M1}(\theta_{0M}) \right) \nabla_{\theta_M} g_{M1}(\theta_{0M}) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} s_i \nabla_{\theta_M} g_{M1}(\overline{\theta}_M) \nabla_{\theta_M} g_{M1}(\widehat{\theta}_M) \sqrt{nh} \left(\widehat{\theta}_M - \theta_{0M} \right)
\end{aligned}$$

and so

$$\begin{aligned}
&\sqrt{nh}\left(\widehat{\theta}_M - \theta_{0M} \right) \\
= \;& \frac{1}{\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right) \Pr(s_i = 1)} \sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{\widehat{m}_M(1^-)}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} - g_{M1}(\theta_{0M}) \right) + o_p(1) \\
= \;& o_p(1) + \frac{1}{\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right) \Pr(s_i = 1)} \left(\sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma_0))} - g_{M1}(\theta_M) \right) \right. \\
&+ \sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \frac{m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma_0))} \left(\frac{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - m_M(F_{z'\gamma}(z_i'\gamma_0))}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} \right) \right) \\
&+ \sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \left(\frac{\widehat{m}_M(1^-) - m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma_0))} \right) \right) \\
&\left. + \sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} \left(\frac{\widehat{m}_M(1^-) - m_M(1^-)}{m_M(F_{z'\gamma}(z_i'\gamma_n))} \right) \left(\frac{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - m_M(F_{z'\gamma}(z_i'\gamma))}{\widehat{m}_M(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma}))} \right) \right) \right) \\
= \;& \frac{1}{\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right) \Pr(s_i = 1)} \left(A_{1,n,h} + A_{2,n,h} + A_{3,n,h} + A_{4,n,h} \right)
\end{aligned}$$

Now, given E5, $A_{1,n,h} = O_p\left(\sqrt{h}\right)$, and given Lemma 2, $A_{2,n,h} = O_p\left(\sqrt{h}\right)$. Also,

$$A_{3,n,h} = \frac{1}{n}\sum_{i=1}^{n} s_i \left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)m_M(F_{z'\gamma}(z_i'\gamma))}\right)\sqrt{nh}\left(\widehat{m}_M(1^-) - m_M(1^-)\right)$$

and thus

$$A_{3,n,h} \xrightarrow{d} N\left(0, \mathrm{E}\left[s_i\left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)m_M(F_{z'\gamma}(z_i'\gamma))}\right)\right]^2 \mathrm{var}\left(\sqrt{nh}\left(\widehat{m}_M(1^-) - m_M(1^-)\right)\right)\right)$$

Finally, $A_{4,n,h} = o_p\left(\sqrt{h}\right)$. Hence, by Lemma 1,

$$\begin{aligned}
& \mathrm{avar}\left(\sqrt{nh}\left(\widehat{\theta}_M - \theta_{0M}\right)\right) \\
= & \frac{\mathrm{E}\left[s_i\left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)m_M(F_{z'\gamma}(z_i'\gamma))}\right)\right]^2}{\left(\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)\right)^2\left(\mathrm{E}\left[s_i\right]\right)^2} \\
& \mathrm{E}\left[\frac{s_j\widetilde{u}_j^2}{g_{M1}^2\left(x_j'\beta_{0M}\right)}\left|F_{z'\gamma}\left(z_j'\gamma_0\right) = 1^-\right.\right]\left[\mathbf{M}_1^{-1}\mathbf{\Gamma}^1\mathbf{M}_1^{-1}\right]_{00}
\end{aligned}$$

The statement in the theorem then follows.

Proof of Theorem 2
(i) By a similar argument as in the proof of Lemma 1,

$$\sqrt{nh}\left(\widehat{m}_A(1^-) - g_{A1}\left(\theta_{0A}\right)\right) \xrightarrow{d} N\left(0, \sigma_A^2(1)\right) \tag{15}$$

where $\sigma_A^2 = \mathrm{E}\left[s_j u_j^2\left|F_{z'\gamma}\left(z_j'\gamma_0\right) = 1^-\right.\right]\left[\mathbf{M}_1^{-1}\mathbf{\Gamma}^1\mathbf{M}_1^{-1}\right]_{00}$. Also, by a similar argument as in the proof of Lemma 2,

$$\frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n} s_i\left(\widehat{m}_A\left(\widehat{F}_{z'\gamma_0}\left(z_i'\widehat{\gamma}\right)\right) - g_{A1}\left(\theta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)\right)\nabla_{\theta_A} g_{A1}\left(\theta_{0A}\right) = o_p(1). \tag{16}$$

In order to show consistency, we first need to show that θ_{0A} is uniquely identified. Define:

$$\begin{aligned}
\widetilde{\theta}_A &= \arg\min_{\theta_A\in\Theta_A} \widetilde{S}_n\left(\theta_A\right) \\
&= \arg\min_{\theta_A\in\Theta_A} \frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\beta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right) - g_{A1}(\theta_A)\right)^2 \\
&= \arg\min_{\theta_A\in\Theta_A} \frac{1}{n}\sum_{i=1}^{n} s_i\left(u_i + g_{A1}(\theta_{0A}) - g_{M1}(\theta_M)\right)^2
\end{aligned}$$

Given Assumptions A1, A2, E1, and E3A–E6A, by a uniform law of large numbers

$$\begin{aligned}
&\sup_{\theta_A\in\Theta_A}\left|\frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\beta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right) - g_{A1}(\theta_A)\right)^2\right. \\
&\left.-\mathrm{E}\left[s_i\left(y_i - g_{A2}\left(x_i'\beta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right) - g_{A1}(\theta_A)\right)^2\right]\right| \xrightarrow{p} 0
\end{aligned}$$

Given the invertibility of g_{A1}, Assumption A1(ii), and by the first order conditions, it follows that $\theta_A = \theta_{0A}$ is the unique minimizer of $\mathrm{E}\left[s_i\left(y_i - g_{A2}\left(x_i'\beta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right) - g_{A1}(\theta_A)\right)^2\right]$. This establishes unique identification of θ_{0A}. Hence, $\widetilde{\theta}_A = \theta_{0A} + o_p(1)$. It remains to show that $\widehat{\theta}_A - \widetilde{\theta}_A = o_p(1)$.

Now,

$$\begin{aligned}
&\widehat{S}_n(\theta_A) \\
= \;& \frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\beta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right) - g_{A1}(\theta_A)\right)^2 + \frac{1}{n}\sum_{i=1}^{n} s_i\left(\left(\widehat{m}_M(1^-) - g_{A1}\left(\theta_{0A}\right)\right)\right)^2 \\
& + \frac{1}{n}\sum_{i=1}^{n} s_i\left(\widehat{m}_A\left(\widehat{F}_{z'\gamma_0}\left(z_i'\widehat{\gamma}\right)\right) - g_{A1}\left(\theta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right)\right)^2 + \text{cross products} \\
= \;& \frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\beta_{0A}\right) - \lambda\left(F_{z'\gamma_0}\left(z_i'\gamma_0\right)\right) - g_{A1}(\theta_A)\right)^2 + o_p(1) \\
= \;& \frac{1}{n}\sum_{i=1}^{n} s_i\left(u_i + g_{A1}(\theta_{A0}) - g_{A1}(\theta_A)\right)^2 + o_p(1)
\end{aligned}$$

given (15) and (16), and noting that the cross terms cannot be of larger order. Finally, given Assumption E5,

$$\sup_{\theta_A\in\Theta_A}\left|\frac{1}{n}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\widehat{\beta}_A\right) - \lambda\left(\widehat{F}_{z'\gamma_0}\left(z_i'\widehat{\gamma}\right)\right) - g_{A1}(\theta_A)\right)^2 - \widetilde{S}_n(\theta_A)\right| \xrightarrow{p} 0.$$

This establishes consistency.
(ii) By first order conditions, for $\overline{\theta}_A \in \left(\widehat{\theta}_A, \theta_{0A}\right)$,

$$\begin{aligned}
0 = \;& \sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\widehat{\beta}_A\right) - \left(\widehat{m}_A(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - \widehat{m}_A(1^-)\right) - g_{A1}(\widehat{\theta}_A)\right)\nabla_{\theta_A}\left(g_{M1}(\widehat{\theta}_A)\right) \\
= \;& \nabla_{\theta_A}\left(g_{M1}(\widehat{\theta}_A)\right)\sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\widehat{\beta}_A\right) - \left(\widehat{m}_A(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - \widehat{m}_A(1^-)\right) - g_{A1}(\theta_{0A})\right) \\
& + \nabla_{\theta_A}\left(g_{M1}(\widehat{\theta}_A)\right)\nabla_{\theta_A}\left(g_{M1}(\overline{\theta}_A)\right)\frac{1}{n}\sum_{i=1}^{n} s_i\sqrt{nh}\left(\widehat{\theta}_A - \theta_{0A}\right)
\end{aligned}$$

$$\begin{aligned}
& \sqrt{nh}\left(\widehat{\theta}_A - \theta_{0A}\right) \\
= \;& \frac{1}{\nabla_{\theta A}\left(g_{A1}(\theta_{0A})\right)\frac{1}{n}\sum_{i=1}^{n} s_i} \\
& \times\sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\widehat{\beta}_A\right) - \left(\widehat{m}_A(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - \widehat{m}_A(1^-)\right) - g_{A1}(\theta_{0A})\right) + o_p(1) \\
= \;& \frac{1}{\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)\frac{1}{n}\sum_{i=1}^{n} s_i}\sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i\left(y_i - g_{A2}\left(x_i'\widehat{\beta}_A\right) - m_A(F_{z'\gamma}(z_i'\gamma_0) - g_{A1}(\theta_{0A})\right) \\
& - \frac{1}{\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)\frac{1}{n}\sum_{i=1}^{n} s_i}\sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i\left(\widehat{m}_A(\widehat{F}_{z'\gamma}(z_i'\widehat{\gamma})) - m_A(F_{z'\gamma}(z_i'\gamma_0))\right) \\
& + \frac{1}{\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)\frac{1}{n}\sum_{i=1}^{n} s_i}\sqrt{\frac{h}{n}}\sum_{i=1}^{n} s_i\left(\widehat{m}_A(1^-) - m_A(1^-)\right) + o_p(1) \\
= \;& I_{1,n,h} + II_{2,n,h} + III_{3,n,h} + o_p(1)
\end{aligned}$$

Now, given E4, $I_{1,n,h} = O_p\left(\sqrt{h}\right)$, and given (16), $II_{2,n,h} = O_p\left(\sqrt{h}\right)$. Now,

$$III_{3,n,h} = \frac{1}{\left(\frac{1}{n}\sum_{i=1}^{n} s_i\right)\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)}\sqrt{nh}\left(\widehat{m}_A(1^-) - m_A(1^-)\right)$$

and thus given (15),

$$III_{3,n,h} \xrightarrow{d} N\left(0, \frac{\mathrm{E}\left[s_j u_j^2 \left| F_{z'\gamma}\left(z_j'\gamma_0\right) = 1^-\right.\right]\left[\mathbf{M}_1^{-1}\mathbf{\Gamma}^1\mathbf{M}_1^{-1}\right]_{00}}{\Pr(s_i=1)^2\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)^2}\right).$$

The statement in the theorem then follows.

Proof of Theorem 3 (i) Define,

$$\widetilde{\theta}_M^p = \arg\min_{\theta_M^p \in \Theta_M} \frac{1}{n_0 H^\eta h_2} \sum_{i=1}^n s_i \left(\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)} - g_{M1}\left(\theta_M^p\right) \right)^2 K\left(\frac{p(z_i)-\delta}{h_2} \right)$$

We first show that $\widetilde{\theta}_M^p - \theta_{0M} = o_p(1)$. Given E5, E7, E9-E10 and E11(ii), and taking the lower boundary of the support to be p_l, we have uniformly in Θ_M that

$$\begin{aligned}
& \mathrm{E}\left[\frac{1}{H^\eta \left(\frac{n_0}{n}\right) h_2} s_i \left(\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)} - g_{M1}\left(\theta_M^p\right) \right)^2 K\left(\frac{p(z_i)-\delta}{h_2} \right) \right] \\
= & \frac{1}{H^\eta \left(\frac{n_0}{n}\right) h_2} \int_{\mathcal{C}_{n,x'\beta_0}} \int_{p_l}^1 \int \Pr(s_i = 1|\widetilde{u}, p, t) \left(g_{M1}\left(\theta_{0M}\right) \widetilde{\lambda}\left(p\right) - g_{M1}\left(\theta_M^p\right) + \frac{\widetilde{u}}{g_{M2}\left(t\right)} \right)^2 K\left(\frac{p_i - \delta}{h_2} \right) f_{\widetilde{u},p,x'\beta_0}(\widetilde{u}, p, t) \mathrm{d}\widetilde{u}\mathrm{d}p\mathrm{d}t \\
= & \frac{1}{H^\eta \left(\frac{n_0}{n}\right)} \int_{\mathcal{C}_{n,x'\beta_0}} \int_{\frac{p_l+H-1}{h_2}}^{\frac{H}{h_2}} \int \Pr(s_i = 1|\widetilde{u}, 1-H+vh_2, t) \left(g_{M1}\left(\theta_{0M}\right) \widetilde{\lambda}\left(1-H+vh_2\right) - g_{M1}\left(\theta_M^p\right) + \frac{\widetilde{u}}{g_{M2}\left(t\right)} \right)^2 \\
& f_{\widetilde{u},p,x'\beta_0}(\widetilde{u}, 1-H+vh_2, t) \mathrm{d}\widetilde{u} K\left(v\right) \mathrm{d}v\mathrm{d}t \\
= & \frac{1}{\left(\frac{n_0}{n}\right)} \int_{\mathcal{C}_{n,x'\beta_0}} \int \left(\left(g_{M1}\left(\theta_{0M}\right) - g_{M1}\left(\theta_M^p\right)\right)^2 + \left(\frac{\widetilde{u}}{g_{M2}\left(t\right)} \right)^2 \right) v_{\widetilde{u},x'\beta_0,p}(\widetilde{u}, t, 1) \mathrm{d}\widetilde{u}\mathrm{d}t \int_{-1}^1 K(v)\mathrm{d}v + O\left(\xi_n H^{1-\eta}\right) + O(\xi_n H^{-\eta} h_2^2) \\
= & \frac{1}{\left(\frac{n_0}{n}\right)} \int\int 1\{t \in \mathcal{C}_{n,x'\beta_0}\} \left(\left(g_{M1}\left(\theta_{0M}\right) - g_{M1}\left(\theta_M^p\right)\right)^2 + \frac{\widetilde{u}^2}{g_{M2}\left(x'\beta_{0M}\right)^2} \right) v_{\widetilde{u},x'\beta_0,p}(\widetilde{u}, t, 1) \mathrm{d}\widetilde{u}\mathrm{d}t + O\left(\xi_n H^{1-\eta}\right) + O(\xi_n H^{-\eta} h_2^2)
\end{aligned}$$

where the second equality follows from change of variables, and the third equality from standard mean value expansion arguments, assumptions E9 and E12 together with the fact that $H/h_2 \longrightarrow \infty$ and $H > h_2$ by construction as well as $(p_l - 1)/h_2 < -1$ for large enough n, and that:

$$\frac{n}{n_0} \int_{\mathcal{C}_{x'\beta_0}} \int 2 \frac{\left(g_{M1}\left(\theta_{0M}\right) - g_{M1}\left(\theta_M^p\right)\right) \widetilde{u}}{g_{M2}\left(t\right)} f_{\widetilde{u},p,x'\beta_0}(\widetilde{u}, 1-H, t) \mathrm{d}\widetilde{u}\mathrm{d}t = 0$$

by iterated expectations uniformly in H. Thus, since $\widetilde{u}^2 / g_{M2}\left(x'\beta_{0M}\right)^2$ does not depend on θ_{0M}, it is clear that the above expression is uniquely minimized at $\widetilde{\theta}_M^p = \theta_{0M}$. Moreover, given that $\frac{n_0}{n} = c_1 + o_p(1)$, where $n_0 = \sum_{i=1}^n 1\left\{ x_i'\beta_{0M} \in C_{n,x_0'\beta_0} \right\}$, and by a uniform law of large numbers for i.i.d. random variables,

$$\begin{aligned}
& \sup_{\theta_M^p \in \Theta_M} \left| \frac{1}{n_0 h_2 H^\eta} \sum_{i=1}^n s_i \left(\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)} - g_{M1}\left(\theta_M^p\right) \right)^2 K\left(\frac{p(z_i)-\delta}{h_2} \right) \right. \\
& \left. - \frac{1}{\left(\frac{n_0}{n}\right) H^\eta} \int_{\mathcal{C}_{x'\beta_0}} \int \left(\left(g_{M1}\left(\theta_{0M}\right) - g_{M1}\left(\theta_M^p\right)\right)^2 + \frac{\widetilde{u}^2}{g_{M2}\left(x'\beta_{0M}\right)^2} \right) f_{\widetilde{u},p,x'\beta_0}(\widetilde{u}, 1-H, t) \mathrm{d}\widetilde{u}\mathrm{d}t \right| \\
= & o_p(1).
\end{aligned}$$

Hence, it holds that $\widetilde{\theta}_M^p - \theta_{0M} = o_p(1)$.
Given Assumption E5(ii) and E8, uniformly in Θ_M,

$$\begin{aligned}
& \frac{1}{n_0 H^\eta h_2} \sum_{i=1}^n \left(s_i \left(\frac{y_i}{g_{M2}\left(x_i'\widehat{\beta}_M\right)} - g_{M1}\left(\theta_M^p\right) \right)^2 K\left(\frac{\widehat{p}(z_i)-\delta}{h_2} \right) \right. \\
& \left. - s_i \left(\frac{y_i}{g_{M2}\left(x_i'\widehat{\beta}_M\right)} - g_{M1}\left(\theta_M^p\right) \right)^2 K\left(\frac{\widehat{p}(z_i)-\delta}{h_2} \right) \right) \\
= & o_p(1),
\end{aligned}$$

As the argmin is a continuous function the statement in (i) follows.
(ii) Let $\widetilde{\Omega}_{M,p} = \frac{1}{n_0 H^\eta h_2} \widehat{\Omega}_{M,p}$ and note that

$$\widehat{\Omega}_{M,p}^{-1/2} \left(\widetilde{\theta}_M^p - \theta_{0M} \right) = \widetilde{\Omega}_{M,p}^{-1/2} \sqrt{n_0 h_2 H^\eta} \left(\widetilde{\theta}_M^p - \theta_{0M} \right)$$

The statement follows, once we establish that

$$\widetilde{\Omega}_{M,p}^{-1/2} \sqrt{n_0 h_2 H^\eta} \left(\widetilde{\theta}_M^p - \theta_{0M} \right) \xrightarrow{d} N\left(0, 1\right) \tag{17}$$

and

$$\widetilde{\Omega}_{M,p}^{-1/2}\sqrt{n_0h_2H^\eta}\left(\widehat{\theta}_M^p-\widetilde{\theta}_M^p\right)=o_p(1) \tag{18}$$

We begin by showing (17). By FOC, and recalling the definition of $\widehat{\Omega}_{M,p}$ in the statement of the theorem,

$$\begin{aligned}
&\sqrt{n_0h_2H^\eta}\widetilde{\Omega}_{M,p}^{-1/2}\left(\widetilde{\theta}_M^p-\theta_{0M}\right)\\
=&\frac{\widetilde{\Omega}_{M,p}^{-1/2}\frac{1}{\sqrt{n_0h_2H^\eta}}\sum_{i=1}^n s_i\left(g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(p(z_i)\right)-g_{M1}\left(\theta_{0M}\right)+\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\right)K\left(\frac{p(z_i)-\delta}{h_2}\right)}{\nabla_\theta g_{M1}\left(\widetilde{\theta}_M^p\right)}\\
=&\frac{\frac{1}{\sqrt{n_0h_2H^\eta}}\sum_{i=1}^n s_i\left(g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(p(z_i)\right)-g_{M1}\left(\theta_{0M}\right)+\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\right)K\left(\frac{p(z_i)-\delta}{h_2}\right)}{\sqrt{\frac{1}{n_0h_2H^\eta}\sum_{i=1}^n\widetilde{\widetilde{u}}_i^2 s_iK\left(\frac{\widehat{p}_i-\delta}{h_2}\right)}}
\end{aligned}$$

By the same argument used in part (i),

$$\begin{aligned}
&\frac{1}{\sqrt{n_0h_2H^\eta}}\sum_{i=1}^n\mathrm{E}\left[s_i\left(g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(p(z_i)\right)-g_{M1}\left(\theta_{0M}\right)+\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\right)K\left(\frac{p(z_i)-\delta}{h_2}\right)\right]\\
=&O\left(\sqrt{n_0h_2}H^{1-\frac{\eta}{2}}\right)=o(1)
\end{aligned}$$

for $nh_2H^{2-\eta}\to 0$. Also, by straihgthforward calculations, we have that

$$\begin{aligned}
&\mathrm{var}\left(\frac{1}{\sqrt{H^\eta\left(\frac{n_0}{n}\right)h_2}}s_i\left(g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(p\right)-g_{M1}\left(\theta_{0M}\right)+\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\right)K\left(\frac{p(z_i)-\delta}{h_2}\right)\right)\\
=&\int\int\frac{\widetilde{u}_i^2}{g_{M2}\left(x_i'\beta_{0M}\right)^2}v_{\widetilde{u},x'\beta_0,p}(\widetilde{u},t,1)\mathrm{d}\widetilde{u}\mathrm{d}t\int K(v)^2dv+O(H^{1-\frac{\eta}{2}})
\end{aligned}$$

Given A1, E9-E12, the statement in (17) then follows from law of large and central limit theorem for iid random variables. As for (18), it suffices to show that

$$\frac{1}{\sqrt{n_0h_2H^\eta}}\sum_{i=1}^n s_i\left(g_{M1}\left(\theta_{0M}\right)\widetilde{\lambda}\left(p(z_i)\right)-g_{M1}\left(\theta_{0M}\right)+\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\right)\left(K\left(\frac{\widehat{p}(z_i)-\delta}{h_2}\right)-K\left(\frac{p(z_i)-\delta}{h_2}\right)\right)=o_p(1),$$

and given $n_0h_2H^{2-\eta}\to 0$, it is enough to show that

$$\frac{1}{\sqrt{n_0h_2H^\eta}}\sum_{i=1}^n s_i\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left(K\left(\frac{\widehat{p}(z_i)-\delta}{h_2}\right)-K\left(\frac{p(z_i)-\delta}{h_2}\right)\right)=o_p(1)$$

Now, given Assumption E8(ii),

$$\begin{aligned}
&\frac{1}{\sqrt{n_0h_2H^\eta}}\sum_{i=1}^n s_i\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\left(K\left(\frac{\widehat{p}(z_i)-\delta}{h_2}\right)-K\left(\frac{p(z_i)-\delta}{h_2}\right)\right)\\
=&\frac{1}{\sqrt{n_0h_2H^\eta}h_2}\sum_{i=1}^n s_i\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_2}\right)\left(\widehat{p}(z_i)-p(z_i)\right)\\
=&\frac{1}{\sqrt{n_0h_2H^\eta}h_2}\sum_{i=1}^n s_i\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_2}\right)\Xi_n(z_i)\\
&+\frac{1}{n_0^{3/2}h_2^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}}\sum_{i=1}^n\sum_{j=1}^n s_i\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_2}\right)\frac{\mathbf{K}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_i)}\psi_j\\
=&o_p(1)+\underbrace{\frac{1}{n_0^{3/2}h_2^{3/2}h_2^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}}\sum_{i=1}^n\sum_{j=1}^n s_i\frac{\widetilde{u}_i}{g_{M2}\left(x_i'\beta_{0M}\right)}\nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_2}\right)\frac{\mathbf{K}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_i)}\psi_j}_{I_{n,h_1,h_2}}
\end{aligned}$$

where the $o_p(1)$ term comes because $\sup_z |\Xi_n(z)| = O(h_1^s)$, $s \geq 4$, and $n_0 h_2 h_1^{2s} H^{-\eta} \to 0$. Now,

$$\begin{aligned}
& I_{n,h_1,h_2} \\
= \; & \frac{1}{n_0^{3/2} h_2^{3/2} h_1^{d_z} H^{\frac{\eta}{2}}} \sum_{i=1}^{n} s_i \frac{\widetilde{u}_i}{g_{M2}\left(x_i' \beta_{0M}\right)} \nabla K\left(\frac{\overline{p}(z_i) - \delta}{h_2}\right) \frac{\mathbf{K}(0)}{f(z_i)} \psi_i \\
& + \frac{1}{n_0^{3/2} h_2^{3/2} h_1^{d_z} H^{\frac{\eta}{2}}} \sum_{i=1}^{n} \sum_{j>i}^{n} s_i \frac{\widetilde{u}_i}{g_{M2}\left(x_i' \beta_{0M}\right)} \nabla K\left(\frac{\overline{p}(z_i) - \delta}{h_2}\right) \frac{\mathbf{K}\left(\frac{z_i - z_j}{h_1}\right)}{f(z_i)} \psi_j \\
& + \frac{1}{n_0^{3/2} h_2^{3/2} h_1^{d_z} H^{\frac{\eta}{2}}} \sum_{i=1}^{n} \sum_{j>i}^{n} s_j \frac{\widetilde{u}_j}{g_{M2}\left(x_j' \beta_{0M}\right)} \nabla K\left(\frac{\overline{p}(z_j) - \delta}{h_2}\right) \frac{\mathbf{K}\left(\frac{z_i - z_j}{h_1}\right)}{f(z_j)} \psi_i \\
= \; & \left(\frac{1}{n_0^{3/2} h_2^{3/2} h_1^{d_z} H^{\frac{\eta}{2}}} \sum_{i=1}^{n} \sum_{j>i}^{n} \Psi_{1,i,j} + \frac{1}{n_0^{3/2} h_2^{3/2} h_1^{d_z} H^{\frac{\eta}{2}}} \sum_{i=1}^{n} \sum_{j>i}^{n} \Psi_{2,i,j} \right) (1 + o_p(1))
\end{aligned}$$

Given that $\widetilde{u}_i$ and ψ_i are zero mean i.i.d., it follows that $\mathrm{E}\left[\Psi_{1,i,j} | s_i, x_i, z_i, p_i\right] = \mathrm{E}\left[\Psi_{2,i,j} | s_i, x_i, z_i, p_i\right] = 0$. Also, by change of variables and integration by parts, for $k = 1, 2$

$$\frac{1}{h_2^2 h_1^{d_z} H^{\eta}} \mathrm{E}\left[\frac{1}{h_1^{d_z} h_2} \Psi_{k,i,j}^2\right] = o(n_0)$$

for $n_0 H^{\eta} h_1^{d_z} h_2^2 \to \infty$. Hence, (18) follows.

References

Ahn, H. and J. Powell (1993). Semiparametric estimation of censored selection models with a nonparametric selection mechanism. *Journal of Econometrics 58*, 3–29.

Andrews, D. and M. Schafgans (1998). Semiparametric estimation of the intercept of a sample selection model. *Review of Economic Studies 65*, 497–517.

Cameron, A. and P. Trivedi (2013). *Regression Analysis of Count Data* (2nd ed.). Number 53 in Econometric Society Monographs. Cambridge University Press.

Chamberlain, G. (1986). Asymptotic efficiency in semi-parametric models with censoring. *Journal of Econometrics 32*, 189–218.

Das, M., W. K. Newey, and F. Vella (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies 70*(1), 33–58.

Deb, P. and P. Trivedi (2006). Specification and simulated likelihood estimation of a non-normal treatment-outcome model with selection: Application to health care utilization. *Econometrics Journal 9*, 307–331.

D'Haultfoeuille, X. and A. Maurel (2013). Another look at identification at infinity of sample selection models. *Econometric Theory 29*, 213–224.

Escanciano, J. C., D. Jacho-Chavez, and A. Lewbel (2014). Uniform convergence of weighted sums of non- and semiparametric residuals for estimation and testing. *Journal of Econometrics 178*, 426–443.

Fan, J. and I. Gijbels (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics 20*(4), 2008–2036.

Goh, C. (2018). Rate-optimal estimation of the intercept in a semiparametric sample-selection model. Unpublished manuscript, University of Wisconsin-Milwaukee.

Hall, P. and J. Racine (2015). Infinite order cross-validated local polynomial regression. *Journal of Econometrics 185*, 510–525.

Ham, J. and R. LaLonde (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica 64*, 175–205.

Härdle, W., P. Hall, and H. Ichimura (1993). Optimal smoothing in single index models. *Annals of Statistics 21*, 157–178.

Hayfield, T. and J. Racine (2008). Nonparametric econometrics: The np package. *Journal of Statistical Software 27*, 1–32.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica 47*, 153–161.

Heckman, J. (1990). Variety of selection bias. *American Economic Review 80*(2), 679–694.

Honoré, B. and L. Hu (2018). Selection without exlusion. Working Paper 2018-10, Federal Reserve Bank of Chicago.

Jochmans, K. (2015). Multiplicative-error models with sample selection. *Journal of Econometrics 184*, 315–327.

Khan, S. and E. Tamer (2010). Irregular identification, support conditions, and inverse weight estimation. *Econometrica 78*(6), 2021–2042.

Kitagawa, T. (2010). Testing for instrument independence in the selection model. Unpublished manuscript, UCL.

Klein, R. and R. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*(2), 387–421.

Lewbel, A. (2007). Endogenous selection or treatment model estimation. *Econometric Theory 13*, 32–51.

Li, Q. and J. Racine (2008). Nonparametric estimation of conditional cdf and quantile functions with mixed categorical and continuous data. *Journal of Business and Economic Statistics 26*(4), 423–434.

Olivetti, C. (2006). Changes in women's hours of market work: The role of returns to experience. *Review of Economic Dynamics 9*, 557–587.

Powell, J., J. Stock, and T. Stoker (1989). Semiparametric estimation of index coefficients. *Econometrica 57*, 1403–1430.

Ruppert, D. and M. Wand (1994). Multivariate locally weighted least squares regression. *The Annals of Statistics 22*(3), 1346–1370.

Schafgans, M. (2000). Gender wage difference in malaysia: Parametric and semiparametric estimation. *Journal of Develoment Economics 63*, 351–368.

Schafgans, M. and V. Zinde-Walsh (2002). On intercept estimation in the sample selection model. *Econometric Theory 18*, 40–50.

Sherman, B. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica 61*(1), 123–137.

Stute, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Annals of Statistics 12*, 917–926.

Terza, J. (1998). Estimating count data models with endogenous switching: Sample selection and endogenous treatment effects. *Journal of Econometrics 84*, 129–154.

Terza, J., A. Basu, and P. Rathouz (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics 27*(3), 531–543.

Wooldridge, J. (1997). Quasi-likelihood methods for count data. In M. Pesaran and P. Schmidt (Eds.), *Handbook of Applied Econometrics*, Volume 2, Chapter 8, pp. 352–406. Blackwell.