Econometric Theory, **40**, 2024, 1311–1363.
doi:10.1017/S0266466623000105

INTERCEPT ESTIMATION IN NONLINEAR SELECTION MODELS

WIJI ARULAMPALAM
University of Warwick

VALENTINA CORRADI
University of Surrey

DANIEL GUTKNECHT
Goethe University Frankfurt

We propose various semiparametric estimators for nonlinear selection models, where slope and intercept can be separately identified. When the selection equation satisfies a monotonic index restriction, we suggest a local polynomial estimator, using only observations for which the marginal cumulative distribution function of the instrument index is close to one. Data-driven procedures such as cross-validation may be used to select the bandwidth for this estimator. We then consider the case in which the monotonic index restriction does not hold and/or the set of observations with a propensity score close to one is thin so that convergence occurs at a rate that is arbitrarily close to the cubic rate. We explore the finite sample behavior in a Monte Carlo study and illustrate the use of our estimator using a model for count data with multiplicative unobserved heterogeneity.

1. INTRODUCTION

The outcome equation intercept is of fundamental importance in selection models, when the aim is to recover average treatment effects (ATE; see Heckman, 1979, 1990).[1] However, even though 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 the errors in the outcome and selection equation, the intercept cannot be separately identified from the

We are grateful to the Editor (Peter Phillips), the Co-Editor (Simon Lee), and three anonymous referees for their very useful and constructive comments. We also thank 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 for useful comments and suggestions. Address correspondence to Valentina Corradi, Department of Economics, School of Economics, University of Surrey, Guildford GU2 7XH, UK; e-mail: V.Corradi@surrey.ac.uk.

[1]Examples include, among others, testing for the difference in wages of unionized and nonunionized workers, or estimating the ethnic (e.g., Schafgans, 1998) or gender wage gap (e.g., Schafgans, 2000).

selection bias term (Heckman, 1990). On the other hand, 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 Andrews and Schafgans (1998), Schafgans and Zinde-Walsh (2002), Heckman (1990), and more recently by 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 setups. Count data, for instance, which are typically modeled via multiplicative error models, may be subject to nonrandom sampling as well. A popular example is a count model that looks at the effect of private medical insurance (Terza, 1998; Deb and Trivedi, 2006), or self-reported health status (Windmeijer and Santos Silva, 1998), on the number of doctor visits.

Despite its relevance, nonlinear selection models have so far only been studied in specific parametric settings (see, e.g., Terza, 1998), and only recently Jochmans (2015) devised an estimator for the slope coefficients of more flexible semi-parametric, 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 focus on models in which the intercept and slope parameters can be separately identified, and have a separable error term which is either multiplicative or additive. Leading examples of separable multiplicative error models are count data and accelerated failure time models. A prominent case of a separable additive error in nonlinear models, on the other hand, is the production function which is used in the human capital formation models, and is typically subject to nonrandom sample selection (e.g., Olivetti, 2006).

We start with the case where the selection equation satisfies a monotonic index restriction. Since slope and intercept parameters can be separately identified in these 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 isolate the intercept and the selection bias. Using the transformed dependent variable, we then construct a nonparametric estimator of the intercept, which is consistent, asymptotically normal, and attains a univariate nonparametric convergence rate. Nevertheless, such a rate may vary from "close to" cubic to an "almost" parametric rate, depending on the relative thickness of the tails of the distributions of the instrument index and selection error. In the linear additive case, the key difference with respect to Andrews and Schafgans (1998), Schafgans and Zinde-Walsh (2002), and Heckman (1990) is that these papers construct the estimator by giving positive weight only to observations for which the index value from the selection equation is above a given threshold. By contrast, our first estimator uses observations for which the marginal cumulative distribution function (cdf) of that index variable is close to one.[2] Since our approach is implemented through a standard local polynomial estimator, the main advantage of this approach is that the bandwidth can be chosen in a data-driven manner, e.g., through cross-validation. However, it should be mentioned that in the additive case, our approach implicitly requires that the propensity score has an unbounded density in the neighborhood of one, and bounded away from zero in the multiplicative case.

We then turn to the case in which the monotonic index restriction does not hold and/or the density of the propensity score is not bounded above zero in the proximity of one. In this case, we can no longer rely on the marginal cdf of the instrument index. Instead, we first obtain an estimator of the nonparametric propensity score, and then estimate the intercept via a nonparametric regression using only those observations having a propensity score close but not too close to one. Formally, this is implemented by introducing a trimming sequence that converges to zero at a sufficiently slow rate. While we still require the propensity score to reach one in the limit, we no longer require that its density at that point be bounded away from zero. Thus, we can also accommodate the possibility that observations are rather sparse in the proximity of one (thin density set), so that convergence occurs at an irregular rate (see Khan and Tamer, 2010). As a result of the trimming, this latter estimator converges at a rate that can be arbitrarily close to the cubic rate.

We provide an extensive Monte Carlo study of the properties of our estimators in terms of mean and median bias as well as root mean squared error (RMSE). In particular, when the monotonic index restriction holds, we compare the finite sample properties of our estimator in the linear additive error case with the estimator introduced by Heckman (1990) and formally developed by Schafgans and Zinde-Walsh (2002), and with the estimator of Andrews and Schafgans (1998). Overall, when the bandwidth is chosen via cross-validation, our estimator performs at least on par with both of these estimators. Importantly, the estimator appears to be relatively robust against a violation of the assumption about the tail behavior of the propensity score density, at least for the chosen design. We also study our estimator in the multiplicative error case. Generally, we find that estimators based on adaptive (cross-validated) bandwidth perform at least as well as those based on fixed bandwidth choice, in terms of Integrated Mean Squared Error. This is not surprising, since cross-validated bandwidth minimizes the mean square error. Finally, we also assess the performance of the estimator when the monotonicity assumption is violated and an estimator of the nonparametric propensity score is used. Also, in this case, we find that the estimator exhibits good finite sample properties in terms of RMSE. Moreover, an ad hoc data-driven procedure to select the tuning parameters appears to work well at least for the chosen design.

[2]For linear additive error models, Goh (2018) provides a set of sufficient conditions under which the upper tail limit point of the marginal cdf of the index variable equals one only if the propensity score equals one at that limit point. He develops an estimator for this case, but does not consider multiplicative error models or models where the monotonicity of the index restriction may actually be violated (see our Section 4).

Finally, we provide an empirical illustration using a dataset similar to Windmeijer and Santos Silva (1998). The outcome variable (number of recent doctor visits) is modeled as a multiplicative function of a binary observed (self-reported) health status variable, unobserved multiplicative heterogeneity, and other observed covariates. We allow for endogenous selection into the status of health, as this self-reported status may not be independent of the error in the outcome equation. The results indicate that for the particular sample used, the estimates of the effect of self-reported health from using our estimators are very similar to that from a fully parametric model estimator that treats self-reported health status as exogenous.

The rest of the paper is organized as follows: Section 2 outlines the setup. 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, whereas Section 6 contains our empirical illustration. Finally, Section 7 concludes. All proofs as well as some extra figures and tables are collected in the Appendix. The Supplementary Material contains additional results from the Monte Carlo simulation.

2. SETUP AND IDENTIFICATION

We motivate our estimator using the standard sample selection model setup. The data generating process for the separable case, where the slope and the intercept parameters can be separately identified and estimated, is discussed next.

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, whereas covariate(s) x_i are observed for all individuals in the sample. We initially impose the following linear index assumption for s_i:

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

where $1\{A\} = 1$ if the event A holds, and zero otherwise, and z_i is a vector of observed covariates. 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. For the outcome equation, we consider additive as well as multiplicative error nonlinear models of the form:

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

and

$$\mathrm{E}[y_i|x_i,\widetilde{\varepsilon}_i] = g_{M1}(\theta_{0M}) \cdot g_{M2}\left(x_i'\beta_{0M}\right)\widetilde{\varepsilon}_i, \tag{3}$$

respectively, where $g_{A1}(\cdot)$, $g_{A2}(\cdot)$, $g_{M1}(\cdot)$, and $g_{M2}(\cdot)$ 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 model as in (3) is the count data model, where

$$g_{M1}(\theta_{0M}) \cdot g_{M2}\left(x_i'\beta_{0M}\right) = \exp(\theta_{0M}) \exp\left(x_i'\beta_{0M}\right) \tag{4}$$

and $\widetilde{\varepsilon}_i$ typically plays the role of unobserved individual heterogeneity. Sample selection issues can arise if $\widetilde{\varepsilon}_i$ (or ε_i, respectively) are not independent of s_i. For instance, y_i could measure the number of credit card defaults for each individual i in a given period of time, whereas 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 nonindependent of 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 within the setup of (4) is the Accelerated Failure Time model applied to duration data, where samples are often plagued by the presence of endogenous selection (e.g., Ham and LaLonde, 1996). An example of a nonlinear additive sample selection model can be found in the human capital formation literature (Olivetti, 2006). We, therefore, deem the separable multiplicative case sufficiently relevant to be considered in its own right. Moreover, note that the above setup can easily be generalized to the case of endogenous covariates (as illustrated by our empirical example; cf. Section 6), and also to endogenous switching regressions.

We now provide a set of sufficient high-level assumptions which ensure point identification of the intercept parameters in (2) and (3):

A1: (i) $\mathrm{E}[|y_i|] < \infty$. (ii) The functions $g_{A1}(\cdot)$, $g_{A2}(\cdot)$, $g_{M1}(\cdot)$, and $g_{M2}(\cdot)$ are known; $g_{A1}(\cdot)$ as well as $g_{M1}(\cdot)$ are invertible almost everywhere, and $g_{M1}(\cdot)$ and $g_{M2}(\cdot)$ are nonzero almost everywhere. (iii) The slope parameters β_{0A} and β_{0M} are point identified up to a scale normalization. (iv) $\widetilde{\varepsilon}_i$ (ε_i) are independent of x_i and z_i. (v) $\mathrm{E}[\widetilde{\varepsilon}_i] = 1$ and $\mathrm{E}[\varepsilon_i] = 0$.

A2: (i) γ_0 is uniquely identified up to a scale and location normalization. (ii) The marginal cdf of $z_i'\gamma_0$, $F_{z'\gamma_0}(\cdot)$, is continuously differentiable at least once, with nonzero derivative on supp$(z_i'\gamma_0)$, the support of $z_i'\gamma_0$. (iii) It holds that supp$(v_i) \subseteq$ supp$(z_i'\gamma_0)$. (iv) v_i is independent of x_i and z_i.

The invertibility of $g_{A1}(\cdot)$ and $g_{M1}(\cdot)$ will be crucial for the identification of the intercept parameters θ_{0M} and θ_{0A}, respectively. Assumption A1(iii), on the other hand, is a high-level condition on the identification of the slope coefficients. In fact, the point identification (and estimation) of the slope parameters will require sufficient variation in x_i and the existence of at least one component in z_i which is not in x_i (cf. Jochmans, 2015).[3] On the other hand, the identification and estimation of θ_{0M} and θ_{0A}, respectively, only rely implicitly on such an excluded variable in z_i through the identification of the slope parameters β_{0A} and β_{0M} (cf. also Andrews and Schafgans, 1998; Schafgans and Zinde-Walsh, 2002). A1(iv) is a

[3] Honoré and Hu (2020) have recently examined semiparametric additive linear sample selection models without such an exclusion restriction, and have derived sharp bounds for the parameters of this type of model.

standard assumption, which can be restrictive and will be relaxed in Section 4, whereas A1(v) is a normalization assumption in exponential and linear models with intercept.

Turning to A2, Assumption A2(i) is also a high-level condition, which is not restrictive as γ_0 can be identified and estimated in a separate step. A sufficient condition for point identification of γ_0 (cf. Klein and Spady, 1993, Thm. 1) is that the marginal cdf of v_i is strictly increasing on the support of v_i and that z_i contains at least one element with a nonzero coefficient that has continuous density everywhere (cf. Klein and Spady, 1993, Assumption C.3b). A2(ii) and (iii), on the other hand, implies that $F_{z'\gamma_0}(\cdot)$, the marginal cdf of $z_i'\gamma_0$, 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 infinity," that is, as $F_{z'\gamma_0}(z'\gamma_0) \to 1$. Note that A2(iii) rules out that $\text{supp}(v_i)$ strictly contains $\text{supp}(z_i'\gamma_0)$, a situation where identification of the intercept fails. Finally, A2(iv) is a standard identification assumption for semiparametric binary choice models. In addition, note that A2(ii)–(iv) naturally implies that $\Pr(s_i = 1) > 0$, whereas the independence in A1(iv) and A2(iv) will be relaxed in Section 4 to accommodate, for instance, some specific forms of conditional heteroskedasticity in the selection error variance. The following theorem establishes the identification of the intercept parameters.

THEOREM 1. *Under Assumptions A1 and A2, the intercept parameters θ_{0A} and θ_{0M} from (2) and (3), respectively, are (point) identified.*

Similar to Goh (2018), and in contrast to Andrews and Schafgans (1998) and Schafgans and Zinde-Walsh (2002), identification is not achieved using the index $z_i'\gamma_0$ but its marginal cdf. Under the aforementioned conditions, the following is established in the proof of Theorem 1 for some values $x_i = x$ and $z_i = z$ in their respective supports. Letting $w_i \equiv z_i'\gamma_0$, under A1 and A2, we have that

$$\lambda(F_w(w)) \equiv \mathrm{E}[\varepsilon_i|x_i = x, z_i = z, s_i = 1] = \mathrm{E}[\varepsilon_i|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)] \tag{5}$$

and

$$\mathrm{E}\left[(y_i - g_{A2}\left(x_i'\beta_{0A}\right))|x_i = x, z_i = z, s_i = 1\right] = g_{A1}\left(\theta_{0A}\right) + \lambda(F_w(w))$$

for the additive model. Similarly, for the multiplicative model, we obtain

$$\widetilde{\lambda}(F_w(w)) \equiv \mathrm{E}[\widetilde{\varepsilon}_i|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)] \tag{6}$$

and

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

The key insight of the proof of Theorem 1 is that under Assumptions A1 and A2, it holds that

$$\lim_{F_w(w)\to 1} \lambda(F_w(w)) = 0 \quad \text{and} \quad \lim_{F_w(w)\to 1} \widetilde{\lambda}(F_w(w)) = 1. \tag{7}$$

As a result, the intercept parameters of the additive and of the multiplicative model can be (point) "identified at infinity." That is, recalling that β_{0A} and β_{0M} are point identified by A1(iii),

$$\lim_{F_w(w)\to 1} \mathrm{E}[(y_i - g_{\mathrm{A2}}(x_i'\beta_{0\mathrm{A}}))|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)] = g_{\mathrm{A1}}(\theta_{0\mathrm{A}})$$

and

$$\lim_{F_w(w)\to 1} \mathrm{E}\left[\frac{y_i}{g_{\mathrm{M2}}(x_i'\beta_{0\mathrm{M}})}\,|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)\right] = g_{\mathrm{M1}}(\theta_{0\mathrm{M}}).$$

This, in turn, implies point identification of the intercepts since $g_{A1}(\cdot)$ and $g_{M1}(\cdot)$ are known and invertible everywhere by A1(ii).

On the other hand, if the marginal distribution of v_i is assumed to be continuous with a density that is nonzero on $\mathrm{supp}(v_i)$, the support of v_i, a sufficient condition for A2(i), then an alternative identification argument could have relied on the propensity score $\Pr(s_i = 1|z_i) = F_v(z_i'\gamma_0) \equiv F_v(w_i)$. Using the propensity score instead of the marginal cdf $F_w(\cdot)$ is typically the more common way to control for sample selection (e.g., Das, Newey, and Vella, 2003). However, the key difference w.r.t. the use of the propensity score is that under A2(ii), $F_w(w_i)$ is uniformly distributed on $[0,1]$ with the marginal density equal to one. In fact, it is immediate to see that whenever $w \to \infty$, both $F_w(w)$ and the propensity score $p = F_v(w)$ approach one, thus ensuring identification at infinity. The advantage of relying on $F_w(w_i)$ rather than on $\Pr(s_i = 1|w_i) = F_v(w_i)$ is that the former has marginal density equal to one regardless of whether $\lim_{p\to 1} f_p(p)$ is zero, bounded or unbounded.

3. ESTIMATION

Given Theorem 1, the next step is to derive the estimators of θ_{0A} and θ_{0M}, and to establish their consistency and asymptotic normality. In order to accomplish this, we first require estimators of the unknown quantities γ_0, $F_w(\cdot)$, and the corresponding slope coefficients β_{0A} and β_{0M}, respectively. A $\sqrt{n}$-consistent estimator for the instrument parameter vector γ_0 can be obtained from Klein and Spady (1993). Henceforth, we call this estimator $\widehat{\gamma}$.[4] This allows us to construct

[4]Since our theoretical results in Theorems 2 and 3 demonstrate that the estimation error of a $\sqrt{n}$-consistent $\widehat{\gamma}$ does not feature in the limiting distribution of our intercept estimator due to its slower than the parametric convergence rate, we do not discuss its estimation further here. See Klein and Spady (1993) for details on the estimation and the appropriate under-smoothing of the bandwidth.

an estimator of the cdf of $z_i'\gamma_0$ in a straightforward manner:

$$\widehat{F}_{z'\gamma_0}(z_i'\widehat{\gamma}) = \widehat{F}_w(\widehat{w}_i) = \frac{1}{n}\sum_{j=1}^{n} 1\left\{\widehat{w}_j \leq \widehat{w}_i\right\}.$$

Note that this step is common to both additive and multiplicative models. In the next step, we obtain estimators for the slope coefficients, say $\widehat{\beta}_A$ and $\widehat{\beta}_M$. Given the separability of the models in (2) and (3), we can construct these independently of the intercepts at a parametric $\sqrt{n}$ rate following Jochmans (2015). As noted in the previous section, this will require at least one element from z_i to be excluded from x_i. Next, we outline how to construct the estimators of the intercept parameters θ_{0A} and θ_{0M}, starting with the additive model.

3.1. The Additive Model

Recall that the identification argument for this model (equation (2)) exploited the fact that

$$\begin{aligned}&\lim_{F_w(w)\to 1} \mathrm{E}[(y_i - g_{\mathrm{A2}}(x_i'\beta_{0\mathrm{A}}))|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)]\\ &= g_{A1}(\theta_{0\mathrm{A}}) + \lim_{F_w(w)\to 1} \lambda(F_w(w)) = g_{A1}(\theta_{0\mathrm{A}}).\end{aligned}$$

Heuristically, since $g_{A1}(\cdot)$ is known and invertible almost everywhere by A1(ii), we may estimate $g_{A1}(\theta_{0\mathrm{A}})$ through a nonparametric regression of $(y_i - g_{\mathrm{A2}}(x_i'\widehat{\beta}_A))$ on $\widehat{F}_w(\widehat{w})$ at the upper limit point one in the first place, and then recover $\theta_{0\mathrm{A}}$ through a simple inversion using the Delta method. That is, denote the conditional expectation:

$$m_A(1) \equiv \lim_{F_w(w)\to 1} \mathrm{E}[(y_i - g_{\mathrm{A2}}(x_i'\beta_{0\mathrm{A}}))|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)],$$

which is the probability limit of the aforementioned nonparametric regression. In order to account for the boundary issue when estimating $m_A(1)$, 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., Fan and Gijbels, 1992; Ruppert and Wand, 1994). More specifically, define the local polynomial estimator of order q as

$$\begin{aligned}&\left(\widehat{a}_{A0}(1),\ldots,\widehat{a}_{Aq}(1)\right)\\ &\quad = \arg\min_{a_k, k\leq q} \frac{1}{nh}\sum_{i=1}^{n} s_i \left(y_i - g_{A2}(x_i'\widehat{\beta}_A) - \sum_{0\leq k\leq q} a_k\left(\widehat{F}_w(\widehat{w}_i) - 1\right)^k\right)^2\\ &\quad K\left(\frac{\widehat{F}_w(\widehat{w}_i) - 1}{h}\right), \qquad \textbf{(8)}\end{aligned}$$

where $K(\cdot)$ denotes a kernel function defined in E6, and h is a bandwidth parameter satisfying $h \to 0$ as $n \to \infty$. Setting $\widehat{m}_A(1) = \widehat{a}_{A0}(1)$, and given A1(ii), we obtain

$$\widehat{\theta}_A = g_{A1}^{-1}\left(\widehat{m}_A(1)\right) \tag{9}$$

as an estimator of the intercept parameter θ_{0A}. To derive the asymptotic properties of $\widehat{\theta}_A$, note that under A1 and A2, we may, without loss of generality, write

$$y_i - g_{A2}(x_i'\beta_{0\mathrm{A}}) = g_{A1}(\theta_{0\mathrm{A}}) + \lambda(F_w(w_i)) + u_i,$$

where $\mathrm{E}[u_i|F_w(w_i) = F_w(w)] = 0$ by construction. 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 parameter space of θ_{0A}, Θ_A, is compact and θ_{0A} lies in its interior.

E3: (i) $\lambda(.)$ is r times differentiable on $(0,1)$ with $r \geq 1$ with Lipschitz continuous derivatives. (ii) $\lambda(.)$ and the r derivatives are left continuous at the upper boundary point 1.

E4: There exist estimators of (i) γ_0 satisfying $\|\widehat{\gamma} - \gamma_0\| = O_p\left(n^{-1/2}\right)$, and (ii) β_{0A} satisfying $\|\widehat{\beta}_A - \beta_{0A}\| = O_p\left(n^{-1/2}\right)$, respectively, where $\|\cdot\|$ denotes the euclidean norm.

E5: $\lim_{F_w(w)\to 1} \mathrm{E}\left[s_i u_i^2 \,|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)\right] < \infty$.

E6: The kernel function $K(\cdot)$ is a continuously differentiable (with Lipschitz continuous derivative), nonnegative, symmetric function around zero, with compact support on $[-1,1]$ and satisfies $\int_{-\infty}^{\infty} K(v)dv = 1$.

Assumptions E1, E2, E5, and E6 are standard and warrant no further discussion. E4 is a high-level condition on the existence of appropriate estimators for the "first-stage" parameters β_{0A} and γ_0 (see, e.g., existing estimators such as Klein and Spady (1993) and Jochmans (2015)). E4 naturally requires point identification of β_{0A} 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, E3 requires that the selection bias $\lambda(.)$ is r times differentiable on $(0,1)$, with $r \geq 1$. As outlined in the previous section, it is important to note that the definition of the selection bias in E3 is as a function of the marginal cdf of w, $F_w(w)$, which is different from the definition in the standard sample selection literature where the selection bias is typically written as a function of the propensity score $p = F_v(w)$. In fact, as discussed in Remark 1, E3 implicitly imposes conditions on the relative tail behavior of the instrument index $z_i'\gamma_0$ and of the selection error. In particular, when the conditional expectation function $\mathrm{E}[\varepsilon_i|v_i]$ is linear in v_i, it implies that the density of the propensity score p becomes unbounded as $p \to 1$ and so does, in the joint normal case, the derivative of the selection bias as a function of p (see below). On the other hand, in the latter case of joint normality, E3 will, for instance, hold whenever $\mathrm{var}(w_i) < \mathrm{var}(v_i)$. Moreover, following Fan and Guerre (2016), note that we can allow for $r \leq q$, where q is the polynomial order used for estimation in (8).

Remark 1. Here, we consider an example where the outcome error is linearly related to the selection error such that $\mathrm{E}[\varepsilon_i|v_i] = \rho v_i$. Suppose that the marginal cdf of v_i, $F_v(\cdot)$, is strictly increasing and differentiable everywhere (a sufficient condition for A2(i)) and that the first moment of v_i exists and is finite. Then, using integration by parts and the fact that $wF_v(w) \to 0$ as $w \to -\infty$ by the existence and finiteness of $\mathrm{E}[v_i]$, we obtain

$$\begin{aligned}\mathrm{E}[\varepsilon_i|v_i < w] &= \rho \int_{-\infty}^{w} \frac{v f_v(v)}{F_v(w)} \mathrm{d}v \\ &= \rho F_w^{-1}(F_w(w)) - \frac{\rho}{F_v(F_w^{-1}(F_w(w)))} \int_{-\infty}^{F_w^{-1}(F_w(w))} F_v(v)\mathrm{d}v \\ &= \lambda(F_w(w)),\end{aligned}$$

where we have used that $w = F_w^{-1}(F_w(w))$ by Assumption A2(ii) and (iii) with $F_w^{-1}(\cdot)$ denoting the inverse function of $F_w(\cdot)$. Then, letting $\nabla_{F_w(w)}\lambda(F_w(w))$ denote the derivative of $\lambda(\cdot)$, note that

$$\begin{aligned}\nabla_{F_w(w)}\lambda(F_w(w)) &= \rho \frac{f_v(F_w^{-1}(F_w(w)))}{F_v(F_w^{-1}(F_w(w)))^2 f_w(F_w^{-1}(F_w(w)))} \int_{-\infty}^{F_w^{-1}(F_w(w))} F_v(v)\mathrm{d}v \\ &= \rho \frac{f_v(w)}{F_v(w)^2 f_w(w)} \int_{-\infty}^{w} F_v(v)\mathrm{d}v \simeq \rho \frac{f_v(w) w}{F_v(w)^2 f_w(w)},\end{aligned}$$

where $f(w) \simeq g(w)$ as $w \to \infty$ is defined as $\lim_{w\to\infty} \frac{f(w)}{g(w)} = 1$. The last term in the above display exists and is finite provided $f_v(w)w$ goes to zero at least as fast as $f_w(w)$. This, in turn, implies that the density of the propensity score $f_p(p) = \nabla_p F_p(p) = \frac{f_w(F_v^{-1}(p))}{f_v(F_v^{-1}(p))} = \frac{f_w(w)}{f_v(w)}$ tends to infinity as $p = F_v(w) \to 1$, where $F_v^{-1}(\cdot)$ denotes again the inverse function of $F_v(\cdot)$. Similarly, one can show that if ε_i and v_i are jointly normal with mean zero, the former with variance σ_ε and the latter with unit variance for simplicity, the derivative of the selection bias as a function of the propensity score $p = \Phi_v(w)$:

$$\frac{\rho\sigma_\varepsilon \Phi_v^{-1}(p)}{p} - \frac{\ddot{\lambda}(p)}{p}$$

becomes unbounded as $p \to 1$, where we used $\phi(\cdot)$ and $\Phi(\cdot)$ to denote the marginal density and distribution function of the standard normal, respectively, and

$$\ddot{\lambda}(p) \equiv \frac{-\rho\sigma_\varepsilon \phi_v(\Phi_v^{-1}(p))}{p}.$$

On the other hand, for a given w, the selection bias as a function of $F_w(w)$ is given in this case by

$$\mathrm{E}[\varepsilon_i|v_i < w] = -\rho\sigma_\varepsilon \frac{\phi_v(w)}{\Phi_v(w)} = -\rho\sigma_\varepsilon \frac{\phi_v(F_w^{-1}(F_w(w)))}{\Phi_v(F_w^{-1}(F_w(w)))} = \lambda(F_w(w)).$$

Using the fact that $\nabla_w \phi_v(w)/\phi_v(w) = -w$ with $\nabla_w \phi_v(w)$ denoting the derivative of $\phi_v(\cdot)$, we obtain

$$\nabla_{F_w(w)}\lambda(F_w(w)) = \rho\sigma_\varepsilon \frac{\phi_v(w)}{\Phi_v(w)} \left(\frac{w}{f_w(w)} - \frac{\lambda(F_w(w))}{f_w(w)} \right).$$

Hence, for $\nabla_{F_w(w)}\lambda(F_w(w))$ to exist and be finite as $w \to \infty$, we need that $\frac{\phi_v(w)}{f_w(w)} \to 0$ as $w \to \infty$. Indeed, it is immediate to see that if w_i is normal, then provided $\mathrm{var}(w_i) > \mathrm{var}(v_i)$, all derivatives exist and are finite. On the other hand, if $\mathrm{var}(w_i) \leq \mathrm{var}(v_i)$, then Assumption E3 violated. Nevertheless, at least for the case where v_i and w_i are normally distributed as reported in Section 5, our estimator has good and comparable finite sample properties to existing estimators from the literature like Andrews and Schafgans (1998) or Schafgans and Zinde-Walsh (2002) even when E3 is violated.

THEOREM 2. *Let Assumptions A1, A2, and E1–E6 hold. If as* $n \to \infty$, $nh^{2\min\{r,q+1\}+1} \to 0$, $q \geq 1$ *odd, and* $nh \to \infty$, *then*

$$\sqrt{nh_n}\left(\widehat{\theta}_A - \theta_{0A}\right) \xrightarrow{d} N\left(0, \frac{\sigma_A^2(1)}{\nabla_{\theta_A} g_{A1}(\theta_{0A})^2}\right),$$

where $\nabla_{\theta_A} g_{A1}(\cdot)$ *denotes the derivative of* $g_{A1}(\cdot)$ *and*

$$\sigma_A^2(1) = \lim_{F_w(w)\to 1} \mathrm{E}\left[s_i u_i^2 | F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)\right]\left[\mathbf{M}_1^{-1}\Gamma_1\mathbf{M}_1^{-1}\right]_{00},$$

where $[\mathbf{A}]_{00}$ *denotes the upper-left entry of the matrix* $\mathbf{A}$, *and* $\mathbf{M}_1$ *as well as* Γ_1 *are theoretical moments of the kernel function defined at the beginning of the Appendix.*

A consistent estimator of the asymptotic variance $\frac{\sigma_A^2(1)}{\nabla_{\theta_A} g_{A1}(\theta_{0A})^2}$ is given by $\frac{\widehat{\sigma}_A^2(1)}{\nabla_{\theta_A} g_{A1}(\widehat{\theta}_A)^2}$, where

$$\widehat{\sigma}_A^2(1) = \left[\mathbf{M}_1^{-1}\Gamma_1\mathbf{M}_1^{-1}\right]_{00} \times \frac{1}{nh_{v1}} \sum_{i=1}^{n} s_i \left(y_i - g_{A2}\left(x_i'\widehat{\beta}_A\right) - \widehat{m}(\widehat{F}_w(\widehat{w}_i))\right)^2 K\left(\frac{\widehat{F}_w(\widehat{w}_i) - 1}{h_{v1}}\right)$$

with $h_{v1} \to 0$ as $n \to \infty$ satisfying $nh_{v1} \to \infty$. Moreover, note that the theoretical moments of the kernel function in $\mathbf{M}_1^{-1}\Gamma_1\mathbf{M}_1^{-1}$ can be computed analytically. For instance, if an ordinary second-order Epanechnikov kernel and a local linear estimator are used, the upper-left element of this matrix is approximately given by 4.498.

In addition, note that the rate of convergence of $\widehat{\theta}_A$ depends on $r \leq q$ or $r > q$, i.e., whether the number of (left) derivatives r of $\lambda(\cdot)$ is larger than the polynomial order q used for estimation or not. For example, if $r = 1$, regardless of the value of q, we obtain a convergence rate that is arbitrarily close to the cubic rate (Fan and Guerre, 2016). On the other hand, if $r > q$, we may improve the rate by choosing a polynomial order closer to or as large as r. Thus, if we have a function with finite r derivatives, with $r \to \infty$, then we may obtain a convergence rate arbitrarily close to $\sqrt{n}$ by setting $q = q_n$ with $q_n \to \infty$ as $n \to \infty$ (see Hall and Racine (2015)). Hence, under Assumption E3, we get a rate that is arbitrarily close to cubic if $r = 1$ and a rate which can be arbitrarily close to $\sqrt{n}$ if r and q are sufficiently large. Thus, in the additive case, the convergence rates that can be obtained are akin to the ones from the Weibull examples in Andrews and Schafgans (1998, p. 504), though, for instance, the case where w_i and v_i have the same Weibull upper tail density is ruled out by our assumptions formally. In fact, the key advantage of our approach does not consist in improved convergence rates as also pointed out in the discussion of Remark 1, but in the fact that we may use standard adaptive methods like cross-validation to choose the bandwidth. That is, there are no adaptive selection procedures on how to choose the threshold parameters from existing estimators such as Andrews and Schafgans (1998) or Heckman (1990) and Schafgans and Zinde-Walsh (2002). In fact, the Monte Carlo findings below show that, when the bandwidth is chosen via cross-validation, our estimator performs at least on par with these estimators in terms of RMSE and bias. The fact that the cross-validated bandwidth behaves well in terms of RMSE is not surprising at all, since by construction it minimizes the Integrated Mean Squared Error. On the other hand, if our goal is to construct confidence intervals for the intercept parameter, then it is well known that confidence intervals (CIs) based on a cross-validated bandwidth, say $h_{n,CV}$, can be severely distorted, since squared bias and variance are of the same order. Calonico, Cattaneo, and Titiunic (2014), therefore, suggest not only to bias-correct the conditional mean estimator by recentering around the estimated bias, but also to adjust the estimator of the variance in a suitable manner. That is, suppose that we use a bandwidth b_n to estimate the bias. If $h_{n,CV}/b_n \to \pi > 0$ as $n \to \infty$, then one has to consider bias estimation error in the construction of the variance estimator. Hence, the authors advocate the use of a variance estimator that takes account of the possibly nonvanishing bias estimation error. Armstrong and Kolesar (2020), on the other hand, suggest an alternative approach to the construction of CIs based on cross-validated bandwidths. Instead of direct bias correction, they propose the use of critical values from a "folded" normal distribution with variance 1 and mean equal to a standardized version of the "worst case" bias for a given smoothness class of functions to which the target functional (e.g., the conditional mean function) belongs. This maximum bias, which depends on the specific kernel and bandwidth used, can be computed and straightforwardly tabulated in practice.

Finally, when reporting empirical results, a common practice is to try different bandwidths, and to report CIs for each of them. Armstrong and Kolesar (2018) proposed to construct CIs which hold uniformly over such a bandwidth set. Suppose that we select a bandwidth h such that $\underline{h} \leq h \leq \overline{h}$. Let $\widehat{\theta}_A(h) = g_{A1}^{-1}(\widehat{m}_A(1-h))$ be the estimator defined in equation (9) evaluated at a given h, and let $\theta_A(h)$

be its population counterpart for that fixed h. Then, Armstrong and Kolesar (2018) suggest to construct a CI that is valid uniformly over $[\underline{h},\overline{h}]$ using critical values based on the distribution of $\sup_{\underline{h}\leq h\leq\overline{h}}\left|\frac{\sqrt{nh}\left(\widehat{\theta}_A(h)-\theta_A(h)\right)}{\widehat{\sigma}(h)}\right|$, rather than based on asymptotically standard normal critical values. We outline how this approach can be applied to the intercept estimator in Theorem 2 in Remark A.1 in the Appendix.

3.2. The Multiplicative Model

We now move to the multiplicative case (equation (3)). The key difference between the multiplicative case and the additive case is that in the former, the sample selection bias enters multiplicatively rather than additively. In fact, as outlined in the discussion of Theorem 1,

$$m_M(1) \equiv \lim_{F_w(w)\to 1} \mathrm{E}[\frac{y_i}{x_i'\beta_{0\mathrm{M}}}|F_w(w_i)=F_w(w), F_w(v_i)<F_w(w)] = g_{\mathrm{M}1}(\theta_{0\mathrm{M}}).$$

Thus, similarly to the additive case, we may construct an estimator of this conditional expectation in the first step, and then invert again $g_{M1}(\cdot)$ to obtain an estimate of θ_{0M} in the second step. That is, given the invertibility of g_{M1} by A1(ii), $\theta_{0M}=g_{M1}^{-1}\left(m_M(1)\right)$ and, thus, it suffices to have a consistent estimator of $m_M(1)$. Therefore, as with the additive case, we use a local polynomial estimator of odd order defined as

$$\begin{aligned}&\left(\widehat{a}_{M0}(1),\ldots,\widehat{a}_{Mq}(1)\right)\\ &=\arg\min_{a_k,k\leq q}\frac{1}{nh}\sum_{j=1}^{n}s_j\left(\frac{y_j}{g_{M2}\left(x_i'\widehat{\beta}_M\right)}-\sum_{0\leq k\leq q}a_k\left(\widehat{F}_w(\widehat{w}_j)-1\right)^k\right)^2\\ &\quad K\left(\frac{\widehat{F}_w(\widehat{w}_j)-1}{h}\right),\end{aligned} \quad \textbf{(10)}$$

and let $\widehat{m}_M(1)=\widehat{a}_{M0}(1)$, where $h\to 0$ as $n\to\infty$ denotes again the bandwidth sequence. Given A1(i), we can define

$$\widehat{\theta}_M=g_{M1}^{-1}\left(\widehat{m}_M(1^-)\right) \text{ and } \theta_{0M}=g_{M1}^{-1}\left(m_M\left(1^-\right)\right).$$

As before, to derive the asymptotic properties of $\widehat{\theta}_M$, note that under A1 and A2, we write, without loss of generality, y_i as

$$\frac{y_i}{g_{M2}(x_i'\beta_{0M})}=g_{M1}(\theta_{0M})\widetilde{\lambda}(F_{w_0}(w_{0i}))+\frac{\widetilde{u}_i}{g_{M2}(x_i'\beta_{0M})},$$

where $\mathrm{E}[\widetilde{u}_i|x_i=x,F_w(w_i)=F_w(w)]=0$ by construction. Moreover, we impose the following conditions in the sequel:

E1M: The same as E1.

E2M: As E2, but θ_{0A} replaced by θ_{0M}, and Θ_A by Θ_M.

E3M: As E3, but $\lambda(.)$ replaced by $\widetilde{\lambda}(.)$.

E4M: As E4, but $\widehat{\beta}_A$ and β_{0A} replaced by $\widehat{\beta}_M$ and β_{0M}, respectively.

E5M:

$$\lim_{F_w(w)\to 1} \mathrm{E}\left[\frac{s_i \widetilde{u}_i^2}{g_{M2}^2\left(x_i'\beta_{0M}\right)} \,|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w_i) \right] < \infty.$$

E6M: The same as E6.

Assumption E3M is the multiplicative analog of E3, and it is discussed in Remark 2 for the case of the standard normal distribution. Moreover, Assumption E4M is again a high-level condition on the existence of appropriate estimators for the "first-stage" parameters β_{0M} and γ_0. In fact, identification and estimation of β_{0M} is also treated in Jochmans (2015) and requires more primitive normalization conditions and assumptions about the covariate space of x_i and z_i as outlined before.

Remark 2. To understand the implications of E3M, we look at a specific example using the normal distribution. As we cannot simply assume joint normality of $\widetilde{\varepsilon}_i$ and v_i in the multiplicative case, let $\widetilde{\varepsilon}_i = \exp(e_i)$ in the following. Then, if e_i and v_i are jointly normal (where v_i has variance one and e_i has variance σ_e^2) so that $e_i = \rho v_i + \xi_i$ with $\mathrm{E}[\xi_i|v_i] = 0$, we have that

$$\mathrm{E}\left[\widetilde{\varepsilon}_i|v_i < w\right] = \mathrm{E}\left[\exp(e_i)\,|v_i < w\right] = \exp\left(\frac{\sigma_e^2}{2}\right)\frac{\Phi_v\left(w-\rho\sigma_e\right)}{\Phi_v\left(w\right)}$$

and, thus,

$$\widetilde{\lambda}(F_w(w)) = \exp\left(\frac{\sigma_e^2}{2}\right)\frac{\Phi_v\left(F_w^{-1}\left(F_w(w)\right)-\rho\sigma_e\right)}{\Phi_v\left(F_w^{-1}\left(F_w(w)\right)\right)}.$$

Then, by the same argument used for the additive case,

$$\nabla_{F_w}\widetilde{\lambda}(F_x(x)) = \frac{\exp(\frac{\sigma_e^2}{2})}{\Phi_v(w)}\left(\frac{\phi_v(w-\rho\sigma_e)}{f_w(w)} - \frac{(\widetilde{\lambda}(F_w(w))\phi_v(w)}{f_w(w)}\right).$$

For this derivative to exist and be finite, it has to be the case that the lead term, $\frac{\phi_v(w-\rho\sigma_e)}{f_w(w)}$, exists and is finite as $w \to \infty$. This is a weaker condition than in the additive case and allows, for instance, for setups where $f_w(w)$ goes to zero as fast as $\phi_v(w-\rho\sigma_e)$.

The following theorem establishes the limiting distribution of $\widehat{\theta}_M = g_{1M}^{-1}\left(\widehat{m}(1)\right)$.

THEOREM 3. *Let Assumptions A1, A2, and E1M–E6M hold. If as* $nh^{2\min\{r,q+1\}+1} \to 0$, $q \geq 1$ *odd, and* $nh \to \infty$, *then*

$$\sqrt{nh}\left(\widehat{\theta}_M - \theta_{0M}\right) \xrightarrow{d} N(0,\sigma_{0M}^2),$$

where

$$\sigma_{0M}^2 = \frac{1}{\left(\nabla_{\theta_M}\left(g_{M1}(\theta_{0M})\right)\right)^2} \times \lim_{F_w(w)\to 1} \mathrm{E}\left[\frac{s_i \widetilde{u}_i^2}{g_{M2}^2\left(x_i'\beta_{0M}\right)} |F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)\right]\left[\mathbf{M}_1^{-1}\Gamma_1\mathbf{M}_1^{-1}\right]_{00},$$

with $\nabla_{\theta_M} g_{M1}(\cdot)$ *denoting the derivative of* $g_{M1}(\cdot)$*,* $[A]_{00}$ *denoting the upper left entry of matrix A, and* $\mathbf{M}_1$ *and* Γ_1 *being defined in the Appendix.*

As before, a consistent estimator of σ_{0M}^2 can be constructed as

$$\widehat{\sigma}_M^2 = \frac{1}{\left(\nabla_{\theta_M}\left(g_{M1}(\widehat{\theta}_M)\right)\right)^2} \times \frac{1}{nh_{v2}}\sum_{i=1}^{n}\left(\frac{y_i}{g_{M2}(x_i'\widehat{\beta}_M)} - \widehat{m}_M(\widehat{F}_w(\widehat{w}_i))\right)^2 K\left(\frac{\widehat{F}_w(\widehat{w}_i)-1}{h_{v2}}\right)\left[\mathbf{M}_1^{-1}\Gamma_1\mathbf{M}_1^{-1}\right]_{00},$$

for some $h_{v2} \to 0$ as $n \to \infty$ satisfying $nh_{v2} \to \infty$, where $\left[\mathbf{M}_1^{-1}\Gamma_1\mathbf{M}_1^{-1}\right]_{00}$ may again be computed as in the previous section.

4. NON-MONOTONICITY AND IRREGULAR SUPPORT

In the previous section, we assumed that the probability of selection is a monotonic function of the instrument index. Furthermore, Assumption E3 implicitly required (at least in the case where $\mathrm{E}[\varepsilon_i|v_i]$ is linear in v_i) that the density of the propensity score is unbounded as $p \to 1$ in the additive case, whereas Assumption E3M in the multiplicative case imposed that it is bounded away from zero as $p \to 1$.

In the sequel, we discuss how the estimation of the intercept may still be carried out under weaker conditions on the propensity score density in the neighborhood of one. We focus, for brevity reasons, only on the multiplicative case. Since the misspecification of the selection equation is a common concern in applied work as this can lead to inconsistent estimators of the intercept, in what follows (cf. Vytlacil, 2002; Jochmans, 2015), we also consider a more flexible nonparametric specification of the propensity score using $p(z_i) = \Pr(s_i = 1|z_i)$, where the selection indicator defined as

$$s_i = 1\{p(z_i) > \widetilde{v}_i\} \tag{11}$$

with $\widetilde{v}_i$ distributed uniformly on $(0,1)$. As a consequence, the marginal cdf of the propensity score might not necessarily be invertible in z_i and so "marginalization" as in the previous section is no longer possible. Before we turn to the estimation, a comment on the identification of θ_{0M} in this context is warranted for. That is, recalling that $p(z_i) = p_i$, we replace the identification assumption A2 by

A2*: (i) Assume that $\mathrm{E}[\widetilde{\varepsilon}_i|x_i, z_i, s_i = 1] = \mathrm{E}[\widetilde{\varepsilon}_i|p_i]$; (ii) $\lim_{p\to 1}\mathrm{E}[\widetilde{\varepsilon}_i|p] = 1$.

Assumption A2*(i) is equivalent to Assumption 2.1(i) in Das et al. (2003, p. 35) for a multiplicative model, whereas Assumption A2*(ii) is a high-level condition, which ensures that "identification at infinity" holds. In particular, note that when the index restriction $z_i'\gamma_0$ of Section 3 is indeed satisfied, Assumptions A1 and A2 from before imply A2*(i) and (ii). In addition, as in the setup of Section 3, observe that A2* does not explicitly require that z_i contains an element that is not in x_i, and so identification will again only rely on such an exclusion restriction implicitly through A1(iii). On the other hand, note that, for estimation purposes, we will require the existence of a continuous variable in z_i, which is not in x_i (cf. the discussion of E8M further below). Finally, observe that A2*(i) together with the selection equation in (11) is less restrictive than the full independence assumption of observables and unobservables in A1(iv) and A2(iv), respectively. In fact, as in Andrews and Schafgans (1998, Sect. 5, p. 505), the present setup allows, for instance, situations where $\widetilde{v}_i$ is conditionally heteroskedastic with the conditional variance of $\widetilde{v}_i$ determined by an index function of z_i.

Thus, under Assumption A2*(i), it holds that

$$\overline{\lambda}(p_i) \equiv \mathrm{E}[\widetilde{\varepsilon}_i | z_i, s_i = 1] = \mathrm{E}[\widetilde{\varepsilon}_i | p_i],$$

and so by A1(iii),

$$\mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)} | x_i = x, z_i = z, s_i = 1\right] = g_{M1}(\theta_{0M})\mathrm{E}[\widetilde{\varepsilon}_i | p_i = p] = g_{M1}(\theta_{0M})\overline{\lambda}(p).$$

Thus, using also A2*(ii), we have that

$$\lim_{\delta \to 1} m_M^p(\delta) = \lim_{\delta \to 1} \mathrm{E}\left[\frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)} | p_i = \delta\right] = g_{M1}(\theta_{0M}).$$

By A1(ii), this gives

$$\theta_{0,M} = g_{M1}^{-1}\left(\lim_{\delta \to 1} m_M^p(\delta)\right),$$

which establishes the identification of θ_{0M}.

Turning to the estimation, note that we will work again with the following auxiliary equation:

$$y_i = g_{M1}(\theta_{0M}) g_{M2}\left(x_i'\beta_{0M}\right)\overline{\lambda}(p_i) + \overline{u}_i, \tag{12}$$

where $\mathrm{E}[\overline{u}_i | x_i = x, p_i = p] = 0$ by construction. As 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. Thus, for notational simplicity, hereafter we assume that all the components of x_i and z_i are continuous. The extension to discrete covariates in both vectors is immediate at the cost of more complicated notation and more lengthy arguments in the proofs. In fact, as pointed out by Li and Racine (2008), note that typically only continuous regressors matter for the convergence rate of estimators of conditional nonparametric distribution functions such as $p(z_i)$.

We begin by estimating the propensity score $p(z_i)$ using a standard local constant Nadaraya–Watson (NW) estimator of the form:

$$\widehat{p}(z_i) = \frac{\sum_{j=1}^{n} s_i \overline{\mathbf{K}}\left(\frac{z_i - z_j}{h_1}\right)}{\sum_{j=1}^{n} \overline{\mathbf{K}}\left(\frac{z_i - z_j}{h_1}\right)}, \tag{13}$$

where $\overline{\mathbf{K}}(\cdot)$ denotes the product of d_z univariate higher-order kernel functions $\overline{K}(\cdot)$, and h_1 is the corresponding bandwidth sequence satisfying $h_1 \to 0$ as $n \to \infty$. As s_i is assumed to be observed for every i in the sample, we can obtain this estimator in a separate preliminary first stage. Moreover, note that, as before, we can estimate the slope parameters in β_{0M} at a parametric $\sqrt{n}$ rate using, e.g., Jochmans (2015). We then obtain the transformed dependent variable as in the previous section to construct an estimator of

$$m_M^p(\delta) = \mathrm{E}\left[s_i \frac{y_i}{g_{M2}\left(x_i'\beta_{0M}\right)} | p_i = \delta \right],$$

where δ is a trimming sequence defined as $\delta = 1 - H$ with H representing a deterministic sequence $H \to 0$ as $n \to \infty$ (see below for a discussion). Formally, define the local constant NW estimator as

$$\widehat{m}_M^p(\delta) = \frac{\sum_{i=1}^{n} s_i \frac{y_i}{g_{M2}\left(x_i'\widehat{\beta}_M\right)} K\left(\frac{\widehat{p}(z_i) - \delta}{h_p}\right)}{\sum_{i=1}^{n} K\left(\frac{\widehat{p}(z_i) - \delta}{h_p}\right)}, \tag{14}$$

where $h_p \to 0$ as $n \to \infty$. In fact, three remarks are noteworthy about this estimator: first, as noted above,

$$\theta_{0M} = \lim_{\delta \to 1} g_{M1}^{-1}\left(m_M^p(\delta)\right),$$

which suggests that we may construct an estimator of $\theta_{0,M}$ as

$$\widehat{\theta}_M^p(\delta) = g_{M1}^{-1}\left(\widehat{m}_M^p(\delta)^{-1}\right).$$

Second, note that we use a local constant rather than a local polynomial estimator in (14) since estimation may be carried out under weaker assumptions than the differentiability of the selection bias in E3 and E3M from the previous section (see below). Third, observe that we will assume that $h_p = o(H)$ in Theorem 4. In a nutshell, this is so since $\lim_{p \to 1} f_p(p)$ may indeed not be bounded away from zero. That is, heuristically, even if identification at infinity holds and p_i converges to one, it may still often be the case 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). To overcome this irregular identification issue, we suggest the above local constant estimator which makes use of observations with propensity scores close but not too close to one. This is implemented by introducing a trimming sequence, which approaches zero at a sufficiently slow rate. That is, instead of using observations with a propensity score $\widehat{p}_i \in (1-h_p, 1)$, we use observations with $\widehat{p}_i \in \left(1-H-h_p, 1-H+h_p\right)$, where $H > h_p$, and both h_p and H go to zero as the sample size increases, but H approaches zero at a slower rate. This allows, in fact, to accommodate cases where the marginal density of $p_i, f_p(\cdot)$, is not bounded away from zero as $p \to 1$. On the downside, this construction will not allow us to choose a data-driven bandwidth through cross-validation. Heuristically, this is because, as shown in the proof of Theorem 4, the bias depends only on H, whereas the order of the variance is $1/(nh_p)$ in the case where the upper tail of the propensity score density is strictly bounded away from zero and $1/(nh_pH^\eta)$ for some $0 < \eta < 1$ in the case of so-called irregular support where the density of the propensity score may not be bounded away from zero as $p \to 1$ (with η determining the "thickness" of the density tail; see below). Thus, even if we fix H, and we search over all $h_p < H$, the value of h_p which minimizes the integrated mean squared error is always the largest possible value of h_p. We make the following additional assumptions.

E7M: (i) $\sup_{z\in\text{supp}(z_i)} |\widehat{p}(z) - p(z)| = o_p(1)$. (ii) The estimated $\widehat{p}(z)$ admits the following representation:

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

for some $\bar{r} \geq \max\{d_z, 2\}$, where ψ_j is the influence function satisfying $\mathrm{E}[\psi_j|z_j] = 0$ and $\mathrm{E}[\psi_j^2|z_j] < \infty$, whereas $\overline{\mathbf{K}}(\cdot)$ denotes the product of d_z univariate kernel functions $\overline{K}(\cdot)$ with uniformly bounded derivative satisfying $\int \overline{K}(t)dt = 1$, $\int t^l \overline{K}(t)dt = 0$, for any positive integer l with $l \leq \bar{r}$, and $\int t^{\bar{r}+1} \overline{K}(t)dt < \infty$. Moreover, $\sup_{z\in\text{supp}(z_i)} |\Xi_n(z)| = O_p(h_1^{\bar{r}})$, and

$$\mathrm{E}\left[\left|\frac{s_i \overline{u}_{x,i} \psi_i}{f_z(z)}\right|^2\right] < \infty, \quad \text{and} \quad \mathrm{E}\left[\left|s_i \overline{u}_{x,i}\, \Xi_n(z_i)\right|^2\right] < \infty,$$

where $\overline{u}_{x,i} = \overline{u}_i / g_{M2}(x_i'\beta_{0M})$.

E8M: (i) There exist constants $C_1, C_2 > 0$ and $\varepsilon_1, \varepsilon_2 > 0$ such that

$$\sup_{u\in(0,1)} \left|f_p(uh_p + 1 - H) - f_p(1-H)\right| \leq C_1 h_p^{\varepsilon_1+\eta},$$

$$\sup_{u\in(0,1)} \left|\Pr\left(s = 1|p = uh_p + 1 - H\right) - \Pr\left(s = 1|p = 1 - H\right)\right| \leq C_2 h_p^{\varepsilon_2+\eta},$$

for some $0 \leq \eta < 1$.

(ii) The density function $f_p(\cdot)$ is continuous on $(0,1)$, and there exists a constant $c(1) > 0$ such that

$$\lim_{H\to 0}\left|\frac{f_p(1-H)}{c(1)H^{\eta}}-1\right| = 0,$$

for some $0 \le \eta < 1$.

E9M: There exists a strictly positive, continuous function $w_{\overline{u}_x,p}(\overline{u}_x,1)$ satisfying $\int \overline{u}_x^2 w_{\overline{u}_x,p}(\overline{u}_x,1)\mathrm{d}\overline{u}_x < \infty$ such that, for some $0 \le \eta < 1$,

$$\sup_{\overline{u}_x\in \operatorname{supp}(\overline{u}_x)}\left|\frac{f_{\overline{u}_x,p}(\overline{u}_x,1-H)}{w_{\overline{u}_x,p}(\overline{u}_x,1)H^{\eta}}-1\right| \to 0 \text{ as } H\to 0,$$

$$\sup_{\overline{u}_x\in \operatorname{supp}(\overline{u}_x)}\left|\Pr(s=1|\overline{u}_x,p=1-H)-1\right| \to 0 \text{ as } H\to 0,$$

where $\overline{u}_x$ was defined in E7M.

E10M: There exist positive constants C such that

$$\sup_{p\in(1-H-h_p,\,1-H+h_p)}\left|\overline{\lambda}(p)-1\right| \le CH^{1-\eta},$$

for some $0 \le \eta < 1$, and $h_p < H$.

E7M represents a high-level condition in the form of the propensity score. It requires the use of a higher-order kernel function, though as long as the number of continuous elements in z_i does not exceed 3, a quartic kernel function is sufficient. Assumption E8M allows for irregular support, in the sense that the density of the propensity score may not necessarily be bounded away from zero as $p \to 1$. More specifically, E8M(i) and (ii) regulate the behavior of the propensity score density as $p \to 1$. E8M(i) is a Lipschitz-type condition tied to the fact that $h_p = o(H)$. The first part of it will, for instance, be satisfied by construction if the marginal density function $f_p(\cdot)$ is continuously differentiable everywhere and $\epsilon_1 + \eta < 1$. E8M(ii), on the other hand, directly imposes conditions on the tail behavior of the propensity score density in the neighborhood of one: when $\eta = 0$, $\lim_{H\to 0} f_p(1-H)$ is bounded away from zero, whereas $\eta > 0$ 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. Thus, note that E8M(ii) restricts the speed at which $f_p(p)$ may converge to zero as $p \to 1$ to be of order H^{η}. For instance, for the case $\eta > 0$, going back to the working example with the index restriction in Remark 2, suppose that $v_i \sim N(0,1)$, and that $w_i \sim N(0,\sigma_w^2)$ with $\sigma_w^2 = 1-\epsilon$ for some $0 < \epsilon < 1/2$ and $\frac{\epsilon}{1-\epsilon} \le \eta$. It then follows that

$$f_p(p) = \frac{f_w(w)}{f_v(w)} = \frac{1}{\sqrt{1-\epsilon}} \exp\left(-\frac{1}{2}\left(\left(\frac{\epsilon}{1-\epsilon}\right) w^2\right)\right) \to 0,$$

as $w \to \infty$. Moreover, setting $c(1) = 1$ for simplicity, note that

$$\begin{aligned} & \frac{f_w(\Phi_v^{-1}(1-H))}{f_v(\Phi_v^{-1}(1-H))} \geq H^\eta \\ \Leftrightarrow \quad & -\frac{\left(\Phi^{-1}(1-H)\right)^2}{2} \frac{(1-\sigma_w^2)}{\sigma_w^2} \geq \eta \ln(H) + \ln(\sigma_w). \end{aligned}$$

Now, using the fact that $\Phi^{-1}(1-H) \simeq \sqrt{-2\ln(H)}$ for $H \to 0$ (e.g., Blair, Edwards, and Johnson, 1976), the last inequality may be approximately written as

$$\frac{\epsilon}{1-\epsilon} \ln(H) \geq \eta \ln(H) + \ln(\sqrt{1-\epsilon}),$$

which holds as strict inequality whenever $\frac{\epsilon}{1-\epsilon} < \eta$, and as equality when $\frac{\epsilon}{1-\epsilon} = \eta$ and $c(1) = \exp(-\ln(\sqrt{1-\epsilon}))$.

Assumption E9M, on the other hand, imposes smoothness on the joint density of the propensity score p_i and $\overline{u}_{x,i} = \overline{u}_i / g_{M2}(x_i' \beta_{0M})$ in proximity of the boundary point 1. Note that it requires that p_i exhibits continuous variation independently of $\overline{u}_{x,i}$. This, in turn, requires that z_i includes at least one variable which is not in x_i and which has continuous density (conditional on the other elements) such that the partial derivative of p_i w.r.t. that element is nonzero with probability one. Finally, Assumption E10M imposes another high-level Lipschitz condition on the behavior of the selection bias term $\overline{\lambda}(\cdot)$ in proximity to one. Using the same example as for E8M(i), note that the condition is satisfied if we assume again that the index $w_i \sim N(0, \sigma_w^2)$ with $\sigma_w^2 < 1$, whereas v_i and e_i are jointly normally distributed with variance one and σ_e^2, respectively. That is, recall that in this case $f_p(p) = \frac{\frac{1}{\sigma_w} \phi\left(\frac{w}{\sigma_w}\right)}{\phi(w)} \to 0$ as $w \to \infty$ whenever $\sigma_w < 1$. In addition, similar calculations to the ones used in Remark 2 yield that

$$\overline{\lambda}(1-H) = \exp\left(\frac{\sigma_e^2}{2}\right) \frac{\Phi_v\left(\Phi_v^{-1}(1-H) - \rho\sigma_e\right)}{1-H}.$$

Then, using the fact that (e.g., Feller, 1968)

$$1 - \Phi(w) \simeq \frac{\phi(w)}{w}$$

as $w \to \infty$ and noting that $\mathrm{E}[\widetilde{\varepsilon}_i] = \mathrm{E}[\exp(e_i)] = \exp\left(\frac{\sigma_e^2}{2}\right)$, we obtain that

$$\left|\exp\left(\frac{\sigma_e^2}{2}\right)\frac{1-\frac{\phi\left(\Phi^{-1}(1-H)-\rho\sigma_e\right)}{\left(\Phi^{-1}(1-H)-\rho\sigma_e\right)}}{1-H}-\exp\left(\frac{\sigma_e^2}{2}\right)\right|$$
$$=\left|\exp\left(\frac{\sigma_e^2}{2}\right)\left(\frac{H}{1-H}-\left(\frac{1}{1-H}\right)\frac{\phi\left(\Phi^{-1}(1-H)-\rho\sigma_e\right)}{\left(\Phi^{-1}(1-H)-\rho\sigma_e\right)}\right)\right|$$
$$\leq CH^{1-\eta}.$$

As mentioned before, 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 as we are discarding all observations with $\widehat{p}_i \in (1-H+h_p, 1]$. Given A1, $\overline{\lambda}(p)-1 = O_p\left(H^{1-\eta}\right)$ for $p \in \left(1-H-h_p, 1-H+h_p\right)$, and so the bias of the intercept estimator cannot approach zero at a rate faster than H. We now establish the limiting distribution of $\widehat{\theta}_M^p$.

THEOREM 4. *Let Assumptions A1, A2*, E1M–E6M, and E7M–E10M hold. If as* $n \to \infty$, $h_1, h_p, H \to 0$, *and* $H/h_p \to \infty$, *(i)* $nh_pH^{2-\eta} \to 0$, *(ii)* $nh_1^{2\overline{r}}h_pH^{\eta} \to 0$, *and (iii)* $nh_1^{d_z}h_p^2H^{\eta} \to \infty$, *then,* $0 \leq \eta < 1$,

$$\widehat{\omega}_{M,p}^{-1}\sqrt{nh_p}\left(\widehat{\theta}_M^p-\theta_{0M}\right) \xrightarrow{d} N(0,1),$$

where

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

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

Theorem 4 establishes the limiting distribution of the Studentized statistic. Note that the convergence rate can be at most $\sqrt{nh_p}$, which given rate condition (i) means that the rate is strictly slower than a cubic rate. Importantly, this rate is not due to the boundary, but to the trimming sequence outlined before. However, the rate can be slower if the observations with $\widehat{p}_i \in \left(1-h_p-H, 1+h_p-H\right)$ grow at a rate slower than nh_p, which occurs if $\eta > 0$. In this case, both $\sqrt{nh_p}\left(\widehat{\theta}_M^p-\theta_{0M}\right)$ and $\widehat{\omega}_{M,p}^{-1}$ will diverge to infinity at the same rate, and so the Studentized statistic still remains bounded and converges to a standard normal.

Remark 3. As outlined in the proof of Theorem 4, asymptotic normality for the infeasible statistic, based on the unknown propensity score requires rate conditions (i), i.e., $nh_pH^{2-\eta} \to 0$, and $nh_pH^{\eta} \to \infty$, which is implied by (iii). If the latter is violated and $nh_pH^{\eta} \to c$ for some $0 < c < \infty$, then the denominator cannot converge in probability to its mean and will remain random in the limit. Hence, we can no longer establish asymptotic normality. On the other hand, rate conditions (ii) and (iii) ensure that both the bias component and the variance component of the propensity score estimation error vanishes. If (iii) is violated and $nh_1^{d_z}h_p^2H^\eta \to c'$ for some $0 \le c' < \infty$ (but $nh_pH^\eta \to \infty$), propensity score estimation error contributes to the limiting distribution of the estimator.

Since rate conditions (i)–(iii) in Theorem 4 hinge on the unknown "tuning parameter" η, a discussion of its choice in practice is warranted. Setting $H = h_p^{\frac{1}{1+\epsilon}}$ for some $\epsilon > 0$ and letting $\overline{\eta}$ denote the maximum admissible value of η, from (i) we obtain $h_p = n^{-\frac{1+\epsilon}{3-\overline{\eta}+\epsilon}}$ and $H = n^{-\frac{1}{3-\overline{\eta}+\epsilon}}$ after some simple calculations, which in turn implies that (ii) and (iii) can be restated as

(ii) $n^{\frac{2-2\overline{\eta}}{3-\overline{\eta}+\epsilon}}h_1^{2\overline{r}} \to 0$ and (iii) $n^{\frac{1-2\overline{\eta}+\epsilon}{3-\overline{\eta}+\epsilon}}h_1^{d_z} \to \infty$.

We first consider the case where the propensity score density is strictly bounded away from zero as $\eta = 0$, and we may set $\overline{\eta} = \epsilon = 0$ so that (ii) $n^{\frac{2}{3}}h_1^{2\overline{r}} \to 0$ and for (iii) $n^{\frac{1}{3}}h_1^{d_z} \to \infty$. In this case, for (ii) to be satisfied, we require that h_1 is of order smaller than $n^{-\frac{1}{3\overline{r}}}$, whereas (iii) is satisfied for $d_z \le 3$ if, for instance, $\overline{r} = 4$ and $h_1 = O(n^{-\frac{1}{11}})$. In fact, (ii) holds for any value of $\overline{\eta} < 1$, whereas for $\epsilon = 0.05$ (iii) holds for $\overline{\eta} = 0.25$ when $d_z = 1$, for $\overline{\eta} = 0.15$ when $d_z = 2$, and for $\overline{\eta} = 0.05$ when $d_z = 3$. On the other hand, if $\overline{\eta} = 0.25$ and $d_z = 1$ as in the simulations or the empirical application, one can verify that the above conditions are also satisfied when $h_1 = O(n^{-\frac{1}{5}})$, suggesting that the first-stage bandwidth maybe chosen through cross-validation. Finally, in practice, we may choose $h_p(\eta)$ and $H(\eta) = h_p(\eta)^{\frac{1}{1+\epsilon}}$ in an ad hoc, data-driven manner from a grid of values satisfying $\{0.05, 0.1, 0.15, \ldots\}$ such that $\widehat{f}_p(1-H)$ lies, for instance, above some threshold value, say 0.1. We explore this data-driven choice further in the simulations of the next section.

5. MONTE CARLO

In this section, we evaluate the finite sample performance of the estimators proposed in Sections 3 and 4. In particular, we assess their robustness w.r.t. the choice of the main tuning parameter(s), and different degrees of selection, and compare their performance with other estimators available in the literature.

We start by outlining the Monte Carlo design, which shares some features with Jochmans (2015). We consider (i) a standard linear design (CASE I) as well as (ii) a multiplicative Poisson design (CASE II) and a multiplicative model with non-monotonic propensity score design (CASE III). For CASE I and CASE II, we assume that the selection equation takes the form

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

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

$$\begin{pmatrix} z_{1i} \\ z_{2i} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_z^2 & -.25 \\ -.25 & \sigma_z^2 \end{pmatrix} \right)$$

and $\gamma_0 = (1,1)'$. The outcome equation for CASE I, on the other hand, is given by

$$y_i = \theta_{0A} + \varepsilon_i, \tag{15}$$

whereas in the multiplicative design of CASE II, we consider

$$y_i = \exp(\theta_{0M})\widetilde{\varepsilon}_i. \tag{16}$$

Selection is modeled in this setup through the correlation between v_i and ε_i in the additive design, and v_i and e_i ($e_i = \log(\widetilde{\varepsilon}_i)$) in the multiplicative design. Specifically, we model the joint distribution as

$$\begin{pmatrix} \varepsilon_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon^2 & \rho\sigma_\varepsilon \\ \rho\sigma_\varepsilon & 1 \end{pmatrix} \right) \quad \text{and} \quad \begin{pmatrix} e_i \\ v_i \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_e^2 & \rho\sigma_e \\ \rho\sigma_e & 1 \end{pmatrix} \right), \tag{17}$$

where $0 \le |\rho| < 1$ and set $\sigma_\varepsilon = \sigma_e = \sqrt{0.5}$. Note that the unconditional mean of $\widetilde{\varepsilon}_i$ in (16) is given by $\exp(\sigma_e^2/2)$. We, therefore, set θ_{0M} equal to $\exp(-\sigma_e^2/2)$, so that the unconditional mean of the outcome equation equals one, whereas θ_{0A} is set to one.

We consider two sample sizes $n = \{600; 1,000\}$, which, given an (unconditional) probability of selection of approximately 0.5 in our designs, implies an effective sample size for the outcome equation of around 300 and 500 observations, respectively. In what follows, we assess the performance of our and other estimators under three different sample selection designs, namely $\rho = 0$ (no sample selection), $\rho = -0.5$ (negative sample selection), and $\rho = +0.5$ (positive sample selection).

We start with CASE I, and assess the finite sample performance of the estimator from Section 3.1 under fixed and data-driven bandwidth schemes in terms of RMSE, whereas results for the Mean Bias (MBIAS) and Median Bias (MDBIAS) can be found in the Supplementary Material. We use the distribution function estimator from Section 3 and subsequently estimate θ_{0A} through a local linear estimator with second-order Epanechnikov kernel evaluated $F_w(w_i) = 1$. Since γ_0 may be estimated at rate $\sqrt{n}$ using the method of Klein and Spady (1993), we set the index $z_i'\gamma$ either equal to the "oracle" index $z_i'\gamma_0$ or estimate it as $z_i'\widehat{\gamma}$, using Klein and Spady (1993). The estimator of Klein and Spady (1993) as well as the local linear estimator are constructed using routines from the `np package` in `R` of Hayfield and Racine (2008). This package allows to provide the program with a fixed bandwidth for which we choose the values $h = 0.15$, $h = 0.10$, and $h = 0.05$ (corresponding to giving a positive weight to observations $F_w(\widehat{w}_i)$ larger 0.85, 0.90, and 0.95, respectively). Alternatively, we use the automated cross-validation procedures implemented in the `np package` and outlined in Li and Racine (2004, 2008), respectively. Likewise, the bandwidth for the estimator of

Klein and Spady (1993) is also routinely chosen by the `np package` through cross-validation.

We compare these estimators with the naïve OLS estimator, which ignores sample selection altogether,[5] as well as with the estimator first suggested by Heckman (1990) and formally developed by Schafgans and Zinde-Walsh (2002):

$$\mathrm{HSZ}(\delta_n) = \frac{\sum_{i=1}^{n} s_i y_i 1\{\widehat{w}_i > \delta_n\}}{\sum_{i=1}^{n} s_i 1\{\widehat{w}_i > \delta_n\}}. \tag{18}$$

In addition, we also consider the estimator suggested by Andrews and Schafgans (1998):

$$\mathrm{AS}(\delta_n, b) = \frac{\sum_{i=1}^{n} s_i y_i \kappa_b(\widehat{w}_i > \delta_n)}{\sum_{i=1}^{n} s_i \kappa_b(\widehat{w}_i > \delta_n)}, \tag{19}$$

where

$$\kappa_b(x) = \begin{cases} 1 - \exp\left(-\frac{x}{b-x}\right), & \text{for } x \in (0, b), \\ 0, & \text{for } x \leq 0, \\ 1, & \text{for } x \geq b. \end{cases}$$

For the tuning parameter b, which determines the weight given to observations with $\widehat{w}_i > \delta_n$, we choose $b = 0.5$ and $b = 1$ (e.g., Schafgans, 1998).[6] Moreover, for the threshold parameter δ_n, we use the 85%, 90%, and 95% unconditional quantiles of $\widehat{w}_i$ from the selected sample, which correspond to the bandwidth choices $h = 0.15$, $h = 0.10$, and $h = 0.05$, respectively.

Turning to the results in Tables 1–3, note first that results are presented through five panels in each table (for a graphical representation of these results, see Figures A.1–A.3 in Section A.2 of the Appendix). Panels A–D use the "oracle" index w_i, and consider different ratios of the unconditional variance of $z_i'\gamma_0$ and v_i. In particular, Panels A and B use a setup where $\mathrm{var}(w_i) \leq \mathrm{var}(v_i) = 1$, which, when the conditional mean is additive, violates the conditions of the estimator for $\widehat{\theta}_A$ outlined in Section 3 (cf. discussion of Remark 1). On the other hand, Panels C and D are compatible with the conditions of Section 3 since in this case $\nabla_{F_w(w)} \lambda(F_w(w))$ exists and is finite as $w \to \infty$. Finally, Panel E is like Panel B, but replacing w_i by the estimator of Klein and Spady (1993), $\widehat{w}_i$.

Examining the finite sample performance of $\widehat{\theta}_A$ for the fixed bandwidths $h = 0.15$, $h = 0.10$, and $h = 0.05$ across Tables 1–3, we see little difference in the finite sample behavior relative to the competing estimators HZS(0.85) (AS$(0.85, \cdot)$), HZS(0.90) (AS$(0.90, \cdot)$), and HZS(0.95), respectively.[7] Unsurprisingly, the RMSE increases in a similar manner for all estimators as we decrease the number of observations for each estimator. In addition, when $\rho = 0$ (Table 1), one

[5] For the multiplicative design of CASE II, we estimate θ_{0M} as $\ln(\widehat{\theta}_{\mathrm{OLS}})$.

[6] We also experimented with $b = 1.5$ as value, but found the performance of the Andrews and Schafgans (1998) estimator to be uniformly dominated by the versions with $b = 0.5$ and $b = 1$ (results available upon request).

[7] Please refer to the Supplementary Material for the results on AS$(0.95, \cdot)$.

TABLE 1. Additive error model (CASE I)—$\rho = 0$, RMSE

n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
Panel A: Oracle index (w_i) − $\mathrm{var}(w_i) = 0.5 < \mathrm{var}(v_i) = 1$													
600	RMSE	0.040	0.111	0.136	0.189	0.122	0.104	0.127	0.182	0.122	0.136	0.154	0.171
1,000	RMSE	0.031	0.090	0.108	0.149	0.099	0.082	0.101	0.144	0.097	0.107	0.121	0.132
Panel B: Oracle index (w_i) − $\mathrm{var}(w_i) = \mathrm{var}(v_i) = 1$													
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.041	0.109	0.132	0.184	0.120	0.105	0.128	0.184	0.119	0.131	0.148	0.164
1,000	RMSE	0.031	0.088	0.105	0.145	0.099	0.082	0.102	0.143	0.094	0.103	0.117	0.127
Panel C: Oracle index (w_i) − $\mathrm{var}(w_i) = 1.25 > \mathrm{var}(v_i) = 1$													
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.040	0.108	0.131	0.184	0.122	0.106	0.128	0.184	0.118	0.129	0.145	0.161
1,000	RMSE	0.031	0.087	0.104	0.144	0.095	0.082	0.102	0.143	0.092	0.101	0.116	0.125
Panel D: Oracle index (w_i) − $\mathrm{var}(w_i) = 1.5 > \mathrm{var}(v_i) = 1$													
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.041	0.107	0.130	0.183	0.125	0.105	0.129	0.183	0.117	0.127	0.144	0.158
1,000	RMSE	0.031	0.086	0.104	0.144	0.095	0.082	0.101	0.144	0.092	0.100	0.115	0.124
Panel E: Klein–Spady index $(\widehat{w}_i)$ − $\mathrm{var}(w_i) = \mathrm{var}(v_i) = 1$													
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.040	0.113	0.136	0.189	0.126	0.108	0.131	0.188	0.123	0.135	0.153	0.167
1,000	RMSE	0.031	0.087	0.107	0.148	0.101	0.084	0.104	0.144	0.096	0.105	0.118	0.128

Note: (1) Number of Monte Carlo replications: 1,500. (2) Columns $h = 0.15, 0.10$, and 0.05 correspond to the estimator $\widehat{\theta}_A$ with a fixed bandwidth size, whereas $\widehat{h}$ denotes the same estimator with a data-driven bandwidth. (3) HSZ$(\cdot)$ corresponds to the estimator (18), with δ_n set to the 85%, 90%, and 95% (unconditional) quantiles of $z_i'\widehat{\gamma}$. (4) AS$(\cdot,\cdot)$ corresponds to the estimator in (19), with δ_n again set to the 85% and 90% quantiles and $b \in \{0.5, 1\}$ (the RMSE results for $AS(0.95, 0.5)$ and $AS(0.95, 1)$ can be found in the tables in the Supplementary Material).

TABLE 2. Additive error model (CASE I)—$\rho = +0.5$, RMSE.

		Panel A: Oracle index (w_i) $-\mathrm{var}(w_i) = 0.5 < \mathrm{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.234	0.118	0.138	0.187	0.136	0.121	0.135	0.181	0.131	0.142	0.156	0.171
1,000	RMSE	0.232	0.098	0.111	0.146	0.108	0.106	0.111	0.145	0.108	0.114	0.125	0.134
		Panel B: Oracle index (w_i) $-$ $\mathrm{var}(w_i) = \mathrm{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.203	0.107	0.130	0.184	0.126	0.108	0.127	0.182	0.120	0.131	0.148	0.162
1,000	RMSE	0.202	0.083	0.101	0.141	0.102	0.086	0.099	0.142	0.092	0.101	0.114	0.125
		Panel C: Oracle index (w_i) $-$ $\mathrm{var}(w_i) = 1.25 > \mathrm{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.192	0.106	0.129	0.184	0.132	0.105	0.127	0.182	0.118	0.129	0.146	0.159
1,000	RMSE	0.191	0.082	0.100	0.140	0.103	0.084	0.099	0.143	0.090	0.099	0.112	0.123
		Panel D: Oracle index (w_i) $-$ $\mathrm{var}(w_i) = 1.5 > \mathrm{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.183	0.107	0.129	0.184	0.136	0.105	0.127	0.183	0.117	0.127	0.145	0.158
1,000	RMSE	0.181	0.083	0.100	0.141	0.103	0.083	0.098	0.143	0.089	0.097	0.111	0.122
		Panel E: Klein–Spady index $(\widehat{w}_i)$ $-$ $\mathrm{var}(w_i) = \mathrm{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.204	0.114	0.135	0.185	0.128	0.111	0.132	0.182	0.123	0.134	0.151	0.165
1,000	RMSE	0.202	0.083	0.103	0.142	0.097	0.087	0.099	0.143	0.094	0.102	0.114	0.125

Note: (1) Number of Monte Carlo replications: 1,500. (2) Columns $h = 0.15, 0.10$, and 0.05 correspond to the estimator $\widehat{\theta}_A$ with a fixed bandwidth size, whereas $\widehat{h}$ denotes the same estimator with a data-driven bandwidth. (3) HSZ$(\cdot)$ corresponds to the estimator (18), with δ_n set to the 85%, 90%, and 95% (unconditional) quantiles of $z_i'\widehat{\gamma}$. (4) AS$(\cdot,\cdot)$ corresponds to the estimator in (19), with δ_n again set to the 85% and 90% quantiles and $b \in \{0.5, 1\}$ (the RMSE results for $AS(0.95, 0.5)$ and $AS(0.95, 1)$ can be found in the tables in the Supplementary Material).

TABLE 3. Additive error model (CASE I)—$\rho = -0.5$, RMSE.

		Panel A: Oracle index (w_i) − $\text{var}(w_i) = 0.5 < \text{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, .5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.234	0.121	0.139	0.188	0.130	0.124	0.137	0.184	0.134	0.144	0.157	0.173
1,000	RMSE	0.231	0.098	0.112	0.149	0.111	0.105	0.113	0.146	0.109	0.116	0.127	0.137
		Panel B: Oracle index (w_i) − $\text{var}(w_i) = \text{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, .5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.204	0.107	0.129	0.180	0.117	0.107	0.126	0.180	0.119	0.129	0.145	0.160
1,000	RMSE	0.200	0.084	0.102	0.143	0.100	0.084	0.101	0.141	0.093	0.102	0.116	0.126
		Panel C: Oracle index (w_i) − $\text{var}(w_i) = 1.25 > \text{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.193	0.106	0.128	0.179	0.118	0.104	0.125	0.180	0.117	0.127	0.144	0.158
1,000	RMSE	0.189	0.082	0.102	0.142	0.099	0.081	0.099	0.142	0.091	0.100	0.114	0.124
		Panel D: Oracle index (w_i) − $\text{var}(w_i) = 1.5 > \text{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.183	0.105	0.128	0.180	0.121	0.103	0.126	0.180	0.116	0.125	0.143	0.156
1,000	RMSE	0.180	0.082	0.101	0.142	0.097	0.080	0.099	0.141	0.089	0.098	0.112	0.122
		Panel E: Klein–Spady index $(\widehat{w}_i)$ − $\text{var}(w_i) = \text{var}(v_i) = 1$											
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	HSZ(0.85)	HSZ(0.90)	HSZ(0.95)	AS(0.85, 0.5)	AS(0.85, 1)	AS(0.90, 0.5)	AS(0.90, 1)
600	RMSE	0.204	0.109	0.129	0.178	0.118	0.108	0.129	0.175	0.120	0.129	0.145	0.157
1,000	RMSE	0.202	0.085	0.105	0.147	0.102	0.086	0.105	0.144	0.097	0.105	0.118	0.128

Note: (1) Number of Monte Carlo replications: 1,500. (2) Columns $h = 0.15, 0.10$, and 0.05 correspond to the estimator $\widehat{\theta}_A$ with a fixed bandwidth size, whereas $\widehat{h}$ denotes the same estimator with a data-driven bandwidth. (3) HSZ($\cdot$) corresponds to the estimator (18), with δ_n set to the 85%, 90%, and 95% (unconditional) quantiles of $z_i'\widehat{\gamma}$. (4) AS($\cdot,\cdot$) corresponds to the estimator in (19), with δ_n again set to the 85% and 90% quantiles and $b \in \{0.5, 1\}$ (the RMSE results for $AS(0.95, 0.5)$ and $AS(0.95, 1)$ can be found in the tables in the Supplementary Material).

can observe that all estimators perform similarly in terms of RMSE. Interestingly and contrary to theoretical predictions, the performance $\widehat{\theta}_A$ does not seem to depend much on the relationship of $\mathrm{var}(w_i)$ and $\mathrm{var}(v_i)$ in this design built on joint normality. That is, observe that results in Panels A and B for each of the tables change only marginally in terms of RMSE relative to Panels C and D. Interestingly, in the case of sample selection ($\rho = +0.5$ or $\rho = -0.5$), we instead see that *all* estimators slightly deteriorate in terms of RMSE when $\mathrm{var}(v_i) > \mathrm{var}(w_i)$ relative to the case when $\mathrm{var}(v_i) < \mathrm{var}(w_i)$. This deterioration can also be observed for the mean and median bias, but holds again across all estimators (see the Supplementary Material for details). This suggests that, at least in the normal case, the discussion from Section 3 concerning the requirements of the estimator presented there may not play a crucial role in finite sample considerations, at least in the present setup. Overall, for $h = 0.15$, $\widehat{\theta}_A$ behaves very similarly to HSZ(0.85), AS(0.85,0.5), and AS(0.85,1) in terms of both RMSE. Additional results in the Supplementary Material show that this is also true for the performance in terms of Mean and Median Bias, where $\widehat{\theta}_A$ does a slightly better job in terms of achieving a smaller average mean or median bias relative to HZS$(\cdot)$, though not necessarily w.r.t. AS$(\cdot,\cdot)$. The same applies for $h = 0.1$ and $h = 0.05$. On the other hand, when we use $\widehat{h}$, $\widehat{\theta}_A$ performs at least on par with HSZ and AS in terms of RMSE and in most cases delivers a smaller mean and median bias.

Next, we move to the multiplicative design with a separable Poisson model (CASE II), whose results can be found in Table 4. Panel A contains the results from the estimations using the "oracle" index w_i, and Panel B from using the index w_i. Moreover, since in the multiplicative case Assumption E3M is actually compatible with $\phi_w(w) = \phi_v(w)$ as $w \to \infty$, we only consider this design throughout. Turning to the results, note that, as expected, the variance of the estimator measured by the RMSE is generally higher than in the additive case. Moreover, as expected by Theorem 3, the first step estimation of $\widehat{\gamma}$ does not appear to contribute to this variance. Turning to the estimates of θ_{0M} using a cross-validated bandwidth $(\widehat{h})$, we see that the estimator generally performs well in terms of RMSE relative to the case where a fixed bandwidth is used. In a final step, we compare the latter estimator also with an estimator where the propensity score is used instead of $\widehat{F}_w(\widehat{w})$ $(\widehat{p})$, and its bandwidth is determined by cross-validation. As can be seen in Table 4, the RMSE is generally larger than when we use $\widehat{F}_w(\cdot)$. This does, of course, not come as a surprise given the nonparametric nature of the propensity score estimator, and further underscores the advantage of using the estimator in Section 3 when its assumptions are satisfied.

Finally, in Table 5, we explore the finite sample behavior of the estimator proposed in Section 4 for the non-monotonic multiplicative model (CASE III). More specifically, letting $z_i \sim N(0,1)$, while the joint distribution of e_i and v_i is left as in (17) before, we set

$$\begin{aligned} s_i &= 1\{2 \cdot \sin(1.5 z_i) > v_i\} \\ &= 1\{\Phi(2 \cdot \sin(1.5 z_i)) > \widetilde{v}_i\}, \end{aligned}$$

TABLE 4. Multiplicative error model

$\rho = 0$							
n		Panel A: Oracle index (w_i)					
		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	$\widehat{p}$
$n = 600$	RMSE	0.072	0.209	0.255	0.397	0.242	0.254
$n = 1,000$	RMSE	0.058	0.157	0.187	0.265	0.187	0.198
n		Panel B: Klein–Spady index ($\widehat{w}_i$)					
		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	$\widehat{p}$
$n = 600$	RMSE	0.074	0.209	0.254	0.407	0.239	0.387
$n = 1,000$	RMSE	0.058	0.156	0.186	0.264	0.168	0.212
$\rho = +0.5$							
n		Panel A: Oracle index (w_i)					
		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	$\widehat{p}$
$n = 600$	RMSE	0.235	0.204	0.254	0.382	0.260	0.308
$n = 1,000$	RMSE	0.230	0.159	0.189	0.278	0.195	0.261
		Panel B: Klein–Spady index ($z_i'\widehat{\gamma}$)					
		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	$\widehat{p}$
$n = 600$	RMSE	0.234	0.205	0.253	0.382	0.284	0.325
$n = 1,000$	RMSE	0.232	0.159	0.189	0.276	0.193	0.229
$\rho = -0.5$							
n		Panel A: Oracle index (w_i)					
		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	$\widehat{p}$
$n = 600$	RMSE	0.192	0.213	0.255	0.378	0.225	0.268
$n = 1,000$	RMSE	0.186	0.158	0.190	0.270	0.198	0.229
		Panel B: Klein–Spady index ($\widehat{w}_i$)					
n		OLS	$h = 0.15$	$h = 0.10$	$h = 0.05$	$\widehat{h}$	$\widehat{p}$
$n = 600$	RMSE	0.194	0.214	0.254	0.386	0.243	0.314
$n = 1,000$	RMSE	0.184	0.158	0.190	0.269	0.197	0.242

Note: (1) Number of Monte Carlo replications: 1,500. (2) Columns $h = 0.15, 0.10$, and 0.05 correspond to the estimator $\widehat{\theta}_M$ using a fixed bandwidth size. (3) $\widehat{h}$ and $\widehat{p}$ correspond to the estimator $\widehat{\theta}_M$ with cross-validated bandwidth choice ($\widehat{h}$) or the nonparametric propensity score ($\widehat{p}$).

where $\widetilde{v}_i = \Phi(v_i)$ is uniformly distributed on $(0,1)$. The propensity score $p(z_i) = \Phi(2 \cdot \sin(1.5z_i))$ exhibits a highly non-monotonic pattern in z_i and, thus, violates the conditions of the estimator $\widehat{\theta}_M$ from Theorem 3. To construct $\widehat{\theta}_M^p(\delta)$, we proceed as follows: first, we estimate the propensity score $p(z_i)$ via a local constant estimator with second-order Epanechnikov kernel and cross-validated bandwidth from the `np package`. Next, we construct the estimator outlined in Section 4.

TABLE 5. Non-monotonic model

	$\rho = 0$				
	(δ, h_p)	$(0.925, 0.075)$	$(0.95, 0.05)$	$(0.975, 0.025)$	$(\widehat{\delta}, \widehat{h}_p)$
$n = 600$	RMSE	0.097	0.110	0.212	0.112
$n = 1,000$	RMSE	0.073	0.081	0.147	0.103
	$\rho = +0.5$				
	(δ, h_p)	$(0.925, 0.075)$	$(0.95, 0.05)$	$(0.975, 0.025)$	$(\widehat{\delta}, \widehat{h}_p)$
$n = 600$	RMSE	0.114	0.121	0.218	0.123
$n = 1,000$	RMSE	0.095	0.093	0.156	0.108
	$\rho = -0.5$				
	(δ, h_p)	$(0.925, 0.075)$	$(0.95, 0.05)$	$(0.975, 0.025)$	$(\widehat{\delta}, \widehat{h}_p)$
$n = 600$	RMSE	0.100	0.110	0.206	0.109
$n = 1,000$	RMSE	0.079	0.083	0.142	0.105

Note: (1) Number of Monte Carlo replications: 1,500. (2) Columns $(\delta, h_p) = (0.925, 0.075)$, $(\delta, h_p) = (0.95, 0.05)$, and $(\delta, h_p) = (0.975, 0.025)$ correspond to the estimator $\widehat{\theta}_M^p(\delta)$ using a fixed δ-h_p combination. (3) $\widehat{\delta}$ and $\widehat{h}_p$ correspond to the estimator $\widehat{\theta}_M^p(\delta)$ using the data-driven choice of the tuning parameters.

Specifically, the first three columns of Table 5 display the estimator's performance for fixed choices of δ and h_p, namely $(\delta, h_p) = (0.925, 0.075)$, $(\delta, h_p) = (0.95, 0.05)$, and $(\delta, h_p) = (0.975, 0.025)$.[8] Finally, in the last column, we use the data-driven method suggested at the end of the last section (we set the threshold to 0.1 and use $\{0.05, 0.1, \ldots, 0.45, 0.5\}$ as a grid for η). Turning to the results, we find that the estimator has an RMSE that is rather low and that increases as δ increases and h_p decreases, respectively.[9] Interestingly, observe that the data-driven choice of δ and h_p generally leads to good and comparable bias and variance results, which is encouraging for practical applications.

6. EMPIRICAL ILLUSTRATION

We now turn to the empirical illustration of the use of the estimator outlined in Section 3. The sample for the analysis is drawn from the second round of the British Health and Lifestyle Survey 1991–92 (HALS2), which was used in the illustration

[8]We have also experimented with setting h_p slightly smaller than H, specifically $(\delta, h_p) = (0.925, 0.070)$, $(\delta, h_p) = (0.95, 0.045)$, and $(\delta, h_p) = (0.975, 0.020)$. However, results do not vary qualitatively and so we only present the specifications from the main text.

[9]Similarly and as expected, results available upon request demonstrate that the estimator generally decreases in terms of MBias and MEDBias when δ increases and h_p decreases, which is particularly pronounced for the case of $\rho = -0.5$.

provided in Windmeijer and Santos Silva (1998).[10] As we were not able to find all the relevant survey reports in order to reconstruct the exact sample used by Windmeijer and Santos Silva (1998), we have created an almost identical sample except for a minor difference in the number of observations used in the estimation. We have 4,820 individuals in our estimation sample compared to 4,814 used in Windmeijer and Santos Silva (1998). The descriptive statistics of our variables match those provided in Table 1 of Windmeijer and Santos Silva (1998), to the first or second decimal place.

The outcome variable of interest is the number of visits to or by a doctor (general practitioner), in the last month prior to the interview, *DOCVIS*. The objective is to model the demand for medical care as a function of individual's health status and to estimate the effect of the latter on the outcome. We follow Windmeijer and Santos Silva (1998) and use a binary self-reported health status variable *HS* as a measure of this unobserved health status and allow this to be dependent on other unobserved individual characteristics in the outcome equation. That is, we treat HS_i as an endogenous regressor in the outcome equation, which takes the value of 1 if health is reported to be poor or fair, and 0 if good or excellent. Adopting the potential outcome framework, we write

$$DOCVIS_i = HS_i \times DOCVIS_i(1) + (1 - HS_i) \times DOCVIS_i(0),$$

where $DOCVIS_i(HS)$, $HS \in \{0,1\}$, denotes the potential number of doctor visits under poor or fair ($HS = 0$), and good or excellent health ($HS = 1$), respectively. We model the conditional mean functions of the potential outcomes in the following exponential regression framework:

$$\mathrm{E}[DOCVIS_i(1)|x_i, \widetilde{\varepsilon}_i(1)] = \exp(\theta_{1M} + x_i'\beta_{0M})\widetilde{\varepsilon}_i(1), \quad \mathrm{E}[\widetilde{\varepsilon}_i(1)|x_i] = 1$$

and

$$\mathrm{E}[DOCVIS_i(0)|x_i, \widetilde{\varepsilon}_i(0)] = \exp(\theta_{0M} + x_i'\beta_{0M})\widetilde{\varepsilon}_i(0), \quad \mathrm{E}[\widetilde{\varepsilon}_i(0)|x_i] = 1,$$

where $\widetilde{\varepsilon}_i(0)$ and $\widetilde{\varepsilon}_i(1)$ represent the multiplicative unobserved heterogeneity under $HS_i = 0$ and $HS_i = 1$, respectively.[11] The relative ATE of a positive health status (relative to poor or fair health) in this multiplicative setup can therefore be calculated as

$$\left(\frac{\mathrm{E}[DOCVIS_i(1)]}{\mathrm{E}[DOCVIS_i(0)]} - 1\right) \times 100\% = (\exp(\alpha_0) - 1) \times 100\%, \tag{20}$$

[10]The data and accompanying documents are available for free download for academic users, from the website of the UK Data Service: www.ukdataservice.ac.uk (accessed on November 22, 2020).

[11]We model unobserved heterogeneity explicitly by adopting a multiplicative model where the observed and unobserved heterogeneity enter the conditional mean of the outcome variable in the same way. This is in contrast to Windmeijer and Santos Silva (1998), who consider an additive model without unobserved heterogeneity.

where $\alpha_0 \equiv \theta_{1M} - \theta_{0M}$. By contrast, due to the potential endogeneity of HS_i, the data only allow to identify (provided $\mathrm{E}[\widetilde{\varepsilon}_i(0)|x_i, HS_i = 0] \neq 0$ a.s.)

$$\left(\frac{\mathrm{E}[DOCVIS_i|x_i, HS_i = 1]}{\mathrm{E}[DOCVIS_i|x_i, HS_i = 0]} - 1\right) \times 100\% = \left(\exp(\alpha_0)\frac{\mathrm{E}[\widetilde{\varepsilon}_i(1)|x_i, HS_i = 1]}{\mathrm{E}[\widetilde{\varepsilon}_i(0)|x_i, HS_i = 0]} - 1\right) \times 100\%,$$

which is not equal to the relative ATE from (20) unless $\mathrm{E}[\widetilde{\varepsilon}_i(1)|x_i, HS_i = 1] = \mathrm{E}[\widetilde{\varepsilon}_i(0)|x_i, HS_i = 0]$ a.s. In what follows, we thus write the conditional expectation function of $DOCVIS_i$ as

$$\mathrm{E}[DOCVIS_i|x_i, \widetilde{\varepsilon}_i] = \exp(\theta_{0M} + x_i'\beta_{0M} + \alpha_0 HS_i)\widetilde{\varepsilon}_i, \tag{21}$$

where $\widetilde{\varepsilon}_i = HS_i \cdot \widetilde{\varepsilon}_i(1) + (1 - HS_i) \cdot \widetilde{\varepsilon}_i(0)$, and assume that endogenous health status HS_i is determined by the following threshold model:

$$HS_i = 1\left\{z_i'\gamma_0 > v_i\right\}. \tag{22}$$

As a robustness check, we also estimate the model in (21), but with a nonparametric propensity score,

$$HS_i = 1\{p(z_i) > \widetilde{v}_i\}, \tag{23}$$

where $\widetilde{v}_i$ is distributed uniformly on $(0, 1)$. The choice of variables to include in z and x are based on Windmeijer and Santos Silva (1998). However, for computational reasons (in particular, for the estimation of the single index coefficient vector using Klein and Spady (1993)), we drop those variables with estimated coefficients that were always insignificant in the models estimated while keeping to the spirit of the discussions provided by the authors for the choice of variables to act as instruments. The variables included in x_i are sex, education, income, and short-term health status. The instrumental variables in z_i are, in addition to those from x_i, variables that explain an individual's health but are likely to affect the demand for doctor services only via the health status. These variables are current work status, alcohol consumption, and binary indicators for smoking behavior, social class, and accommodation, as well as long-term disability or infirmity. Of course, since health status is only coarsely measured through a binary variable, concerns about these variables affecting the number of doctor visits even after conditioning on the binary health status might arise. To address this limitation, we carried out checks for the sensitivity of our results on the choice of these instruments. The estimated effect of interest, α, however, remained qualitatively similar. The definitions and the summary statistics for the variables are provided in Table A.1 in the Appendix. A more detailed discussion of these variables is provided in Windmeijer and Santos Silva (1998).

We estimate the following four models:

- **Model 1**: A standard Poisson (P) specification where HS_i is treated as exogenous. This model does not contain unobserved heterogeneity $\widetilde{\varepsilon}_i$.

- **Model 2**: A negative binomial (NB2) model with HS_i treated as exogenous. In addition, the unobserved heterogeneity $\widetilde{\varepsilon}_i$ is assumed to follow a Gamma distribution with Gamma $(1, \frac{1}{\tau})$, and also to be distributed independently of x.
- **Model 3**: A general exponential model (ACG-1) where HS_i is treated as endogenous according to (22) and the joint distribution of $\widetilde{\varepsilon}_i$ and v_i are left unspecified.
- **Model 4**: A general exponential model (ACG-2) where HS_i is treated as endogenous according to (23) and the marginal distribution of $\widetilde{\varepsilon}_i$ as well as the dependence of $\widetilde{\varepsilon}_i$ and $\widetilde{v}_i$ are left unspecified.

The parameter τ in Model 2 is sometimes called the *over-dispersion parameter*. This particular model is commonly used in the case of over-dispersed count variable data as it is the case with our variable DOCVIS, which has an unconditional mean of 0.402 and a variance of 0.634. However, it does impose independence between $\widetilde{\varepsilon}_i$ and variables in x_i. Model 3, on the other hand, allows for dependence between unobserved heterogeneity $\widetilde{\varepsilon}_i$ and the health status variable HS_i, whereas Model 4 relaxes in addition to the index restriction $z_i'\gamma_0$ and is thus robust against misspecification of the propensity score.

The estimation steps for Model 3 are as follows:

- **Step 1**: Estimate γ_0 using the estimator of Klein and Spady (1993) from the `np package` of Hayfield and Racine (2008). The bandwidth parameter is chosen via a built-in cross-validation procedure, and $\widehat{F}_{z'\widehat{\gamma}}(z_i'\widehat{\gamma})$ is constructed subsequently using the distribution function estimator outlined in Section 3. In addition, we also estimate the propensity score $\widehat{p}(z_i'\widehat{\gamma})$ via the local constant estimator from the `np package` with second-order Epanechnikov kernel and cross-validated bandwidth.
- **Step 2**: Estimate β_{0M} for the entire sample using the one-step estimator proposed in Jochmans (2015) with the author's recommended plug-in bandwidth and a second-order Gaussian kernel.[12]
- **Step 3**: As outlined in Section 3.2, estimate the intercept θ_{0M} and $(\theta_{0M} + \alpha_0)$ separately for the subsample with $HS_i = 0$ and $HS_i = 1$, respectively, using a local linear estimator from the `np package` with second-order Epanechnikov kernel and cross-validated bandwidth.[13]
- **Step 4**: Compute the standard errors for the intercept estimators of Step 3 with the estimator outlined after Theorem 3 and bandwidth choice $h_{v2} = 0.25$ (we also experimented with slightly different choices for h_{v2}, but results remain qualitatively similar).

[12] Changing the order of the kernel as well as the bandwidth did not alter results substantially.

[13] More specifically, we construct $\widehat{\theta}_{0M} = \ln\left(\widehat{E}\left[\frac{y_i}{\exp(x_i'\widehat{\beta}_M)}\,\middle|\widehat{F}_w(\widehat{w}_i) = 0\right]\right)$ using the subsample with $HS_i = 0$, and $\widehat{\alpha} = \ln\left(\widehat{E}\left[\frac{y_i}{\exp(x_i'\widehat{\beta}_M)}\,\middle|\widehat{F}_w(\widehat{w}_i) = 1\right]\right) - \ln\left(\widehat{E}\left[\frac{y_i}{\exp(x_i'\widehat{\beta}_M)}\,\middle|\widehat{F}_w(\widehat{w}_i) = 0\right]\right)$, where the first term is only estimated for $HS_i = 1$. The cross-validated bandwidth chosen for the estimation of θ_{0M} using the subsample $HS_i = 0$ was 0.052 (211 observations with a positive weight), whereas the cross-validated bandwidth for the estimation of $\theta_{0M} + \alpha$ using the subsample $HS_i = 1$ was 0.074 (274 observations with positive weight).

TABLE 6. Estimation results

	P	NB2	ACG-1	ACG-2		
				$(\widehat{H}, \widehat{h}_p)$	$(0.05, 0.05)$	$(0.025, 0.025)$
$\widehat{\alpha}$	0.534	0.549	0.727	0.560	0.664	0.849
s.e. (robust)	(0.064)	(0.062)				
s.e. ($h_{v2} = 0.25$)			(0.351)			
s.e. (H, h_p)				(0.294)	(0.182)	(0.285)
$\widehat{\theta}_{0M}$	−1.111	−1.102	−1.468	−1.116	−1.094	−1.194
s.e. (robust)	(0.053)	(0.052)				
s.e. ($h_{v2} = 0.25$)			(0.305)			
s.e. (H, h_p)				(0.251)	(0.054)	(0.179)

Note: (1) Columns P and NB2 represent the output for Models 1 and 2, respectively, with robust standard errors. (2) Column ACG-1 provides the estimates of Model 3 with cross-validated bandwidth choice (cf. footnote 13). Standard errors are computed as in Step 4. (3) Columns ACG-2 provide the estimates of Model 4 using ad hoc, data-driven $(\widehat{H}, \widehat{h}_p)$ or fixed ($(0.05, 0.05)$ and $(0.025, 0.025)$) tuning parameters. Standard errors are computed as in Step 4.

On the other hand, Model 4 parameters are estimated by first estimating the nonparametric propensity score $p(z_i)$ via the local constant estimator from the `np package` with fourth-order Epanechnikov kernel.[14] We then follow Steps 3 and 4 as outlined above, but replace the estimators from Section 3.2 with the ones of Section 4. The standard errors for the intercept estimator of Model 4 are determined as outlined in Section 4, and the threshold value is set to 0.1 when $H(\eta)$ and $h_p(\eta)$ are determined in an ad hoc data-driven manner as outlined at the end of Section 4.

Table 6 reports the estimates of θ_{0M} and α_0, which are our main parameters of interest, for all four models. The standard errors reported for Models 1 and 2 are robust standard errors based on the pseudo-maximum likelihood estimator (Gourieroux, Monfort, and Trognon, 1984).

As expected, the estimated $\widehat{\alpha}$ does not differ much between Models 1 and 2. Note that not accounting for unobserved heterogeneity does not affect the consistency property of the Poisson MLE estimator. On the other hand, when we relax some of the parametric assumptions and also account for possible endogeneity of HS_i due to dependence between $\widetilde{\varepsilon}_i$ and v_i as in Model 3 (ACG-1), the estimate of α increases to 0.727. Note that this estimate (as the one of ACG-2) cannot be directly compared to the ones of Models 1 and 2 since the latter does not allow for endogeneity. Also, it

[14]Note that except for the variable "Wine" (number of units of wine consumption last week), all variables in z_i are binary, and hence the rate conditions for continuous covariates of Section 4 apply for the case $d_z = 1$ (recall that discrete covariates do not matter for the convergence rate of estimators of conditional nonparametric distribution functions such as $p(z_i)$ (Li and Racine, 2008)). Also, to reduce computational complexity, we follow the method outlined in Racine (1993) and conduct cross-validation on random subsets of the data (size $n = 500$), to select the median values over 50 replications.

is worthwhile noting that we cannot compare the t-ratio to standard normal critical values since CIs are based on a cross-validated bandwidth are biased. As remarked earlier, Armstrong and Kolesar (2020) suggest using wider CIs which account for the largest possible bias, for a given function class. For instance, following Table 1 of their paper, if E3M holds with $r = 1$ and the bias was to equal half of the standard deviation, one should use ± 2.36 instead of the common ± 1.96 for the common 5% significance level, whereas if $r = 3$, we may use ± 2.11 instead. Hence, when using "honest" CIs that account for the worst case bias, the ATE parameter α_0 is no longer significant at the 5% significance level.

We next turn to Figure A.4, which plots the estimated propensity score $\widehat{p}(\widehat{w}_i)$ as a function of the estimated index $\widehat{w}_i$ for observations with $HS_i = 0$ in Figure A.4a, and with $HS_i = 1$ in Figure A.4b, respectively. These figures cast some doubt that monotonicity in the index $\widehat{w}_i$, an assumption required for ACG-1, may be violated for $\widehat{p}(\widehat{w}_i)$ close to one for observations with $HS_i = 1$, whereas this issue does not seem to arise for $\widehat{p}(\widehat{w}_i)$ close to zero for observations with $HS_i = 0$. We, therefore, move to the results of Model 4 next, which is robust against violations of monotonicity due to the nonparametric nature of the propensity score.

Turning to Figure A.5, we observe that $\widehat{p}(z_i)$ appears to exhibit, in fact, some (empirical) support in proximity to 0 for the subsample with $HS_i = 0$ and to 1 for the subsample with $HS_i = 1$, which suggests that the estimator may be applied. We consider the sensitivity of the results to two different ways of choosing the tuning parameters $\delta = 1 - H$ and h_p: the first uses the data-driven ad hoc procedure described at the end of Section 4 setting $\epsilon = 0.1$, which yields $(\widehat{H}_0, \widehat{h}_{p0}) = (0.028, 0.020)$ for $HS_i = 0$ and $(\widehat{H}_1, \widehat{h}_{p1}) = (0.073, 0.056)$ for $HS_i = 1$, respectively. The second uses fixed choices for the tuning parameters, which are identical across health status, namely $(H, h_p) = (0.05, 0.05)$ and $(H, h_p) = (0.025, 0.025)$, respectively.[15]

As results in Table 6 show, using the data-driven ad-hoc choice for the tuning parameters yields an estimate of α_0 of 0.560, which is very similar to Models 1 and 2 estimates where the health status variable HS_i is treated as exogenous in the outcome equation. On the contrary, when fixed values for the tuning parameters are used, the estimated $\widehat{\alpha}$ is higher at 0.664 and 0.849, respectively. Since the data-driven choice of $(\widehat{H}_1, \widehat{h}_{p1})$ is larger than of $(\widehat{H}_0, \widehat{h}_{p0})$, and of the fixed choices, the sensitivity of the point estimates may primarily be due to the sparsity of observations with propensity score value close to one for $HS_i = 1$.

We next link the estimates of Model 4 to the implied number of extra visits to the doctor. The raw difference in the average number of doctor visits between individuals with $HS_i = 0$ and $HS_i = 1$ is 0.43 $(0.73 - 0.29)$. However, the estimated extra doctor visits across the specifications of Model 4 are 1.8, 1.9, and 2.3, translating into relative ATEs of sizes 80%, 90%, and 130%, respectively. Since the predicted number of extra visits for Models 1 and 2, on the other hand, are

[15] As in the previous section, note that setting h_p slightly smaller than H did not really affect the results qualitatively.

1.7 (or 70%), we conclude that, for this particular sample, the numbers are very similar across the different estimations.

7. CONCLUSION

Identification and estimation of the intercept is crucial for the evaluation of ATEs in nonexperimental settings where the treatment selection is often dependent 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 estimators of the intercept in nonlinear semiparametric selection models, where the joint distribution of the error terms remains unknown and the intercept and slope parameters can be separately identified. We consider multiplicative and general nonadditive 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 a standard local polynomial estimator, and the bandwidth may be selected through cross-validation. 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, that does not require that the marginal density function of the propensity score is bounded away from zero at the upper limit point. The resulting estimator is a local constant estimator, which uses observations close but not too close to the boundary. This estimator is robust against misspecification of the first stage and converges at a rate that can be arbitrarily close to a cubic rate. Finally, we investigate the effect of self-reported health on the number of recent doctor visits modeling doctor visits as a multiplicative function of a binary (self-reported) health status variable, unobserved heterogeneity, and other observed covariates. Our findings suggest that for the particular sample used, the estimates of the effect of self-reported health from using our estimators are very similar to that from a fully parametric model estimator that treats self-reported health status as exogenous.

APPENDIX

A.1. Proofs

In the following, for $0 \le t \le 2q$, let

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

as well as

$$\gamma_t(K) = \int_{-1}^{0} \nu^t K^2(\nu) d\nu.$$

Also, define the $(q+1)\times(q+1)$ dimensional matrix

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

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

Proof of Theorem 1. We start with the identification of $\theta_{0\mathrm{M}}$, and then comment on the identification of $\theta_{0\mathrm{A}}$. First, recall that $w_i = z_i'\gamma_0$, and note that

$$\mathrm{E}[\widetilde{\varepsilon}_i|x_i = x, z_i = z, s_i = 1] = \mathrm{E}[\widetilde{\varepsilon}_i|w_i = w, v_i < w] = \mathrm{E}[\widetilde{\varepsilon}_i|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)],$$

where the first equality follows from A1(iv) and A2(i) and (iv) and the selection model in (1), whereas the second equality follows from A2(ii)–(iv). In addition, using Assumption A1(iii), we obtain

$$\begin{aligned}\mathrm{E}[y_i|x_i = x, z_i = z, s_i = 1] &= \mathrm{E}[y_i|x_i'\beta_{0\mathrm{M}} = x'\beta_{0\mathrm{M}}, F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)] \\ &= g_{\mathrm{M}1}(\theta_{0\mathrm{M}})g_{\mathrm{M}2}(x'\beta_{0\mathrm{M}})\mathrm{E}[\widetilde{\varepsilon}_i|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)] \\ &= g_{\mathrm{M}1}(\theta_{0\mathrm{M}})g_{\mathrm{M}2}(x'\beta_{0\mathrm{M}})\widetilde{\lambda}(F_w(w)).\end{aligned}$$

Thus, without loss of generality, we may write

$$y_i = g_{\mathrm{M}1}(\theta_{0\mathrm{M}})g_{\mathrm{M}2}(x'\beta_{0\mathrm{M}})\widetilde{\lambda}(F_w(w_i)) + \widetilde{u}_i,$$

where $\mathrm{E}[\widetilde{u}_i|x_i = x, F_w(w_i) = F_w(w)] = 0$ by construction. Moreover, by A1(ii), it holds that

$$\mathrm{E}\left[\frac{y_i}{g_{\mathrm{M}2}(x_i'\beta_{0\mathrm{M}})}|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)\right] = g_{\mathrm{M}1}(\theta_{0\mathrm{M}})\widetilde{\lambda}(F_w(w)).$$

Now, observe that under A2(ii) and (iii),

$$\lim_{F_w(w)\to 1}\left(g_{\mathrm{M}1}(\theta_{0\mathrm{M}})\widetilde{\lambda}(F_w(w))\right) = g_{\mathrm{M}1}(\theta_{0\mathrm{M}})\mathrm{E}[\widetilde{\varepsilon}_i] = g_{\mathrm{M}1}(\theta_{0\mathrm{M}}),$$

where the last equality follows from $\mathrm{E}[\widetilde{\varepsilon}_i] = 1$ in A1(v). Finally, since $g_{\mathrm{M}1}(\cdot)$ is known and invertible by A1(ii), this establishes the unique identification of $\theta_{0\mathrm{M}}$.

For the additive case, note that by A1(ii) and (iii), it holds similarly that

$$\mathrm{E}[(y_i - g_{\mathrm{A}2}(x_i'\beta_{0\mathrm{A}}))|z_i, s_i = 1] = g_{\mathrm{A}1}(\theta_{0\mathrm{A}}) + \mathrm{E}[\varepsilon_i|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w_i)]$$

and therefore

$$\begin{aligned}\lim_{F_w(w)\to 1}&\mathrm{E}[(y_i - g_{\mathrm{A}2}(x_i'\beta_{0\mathrm{A}}))|F_w(w_i) = F_w(w), F_w(v_i) < F_w(w)] \\ &= g_{\mathrm{A}1}(\theta_{0\mathrm{A}}) + \mathrm{E}[\varepsilon_i] = g_{\mathrm{A}1}(\theta_{0\mathrm{A}}),\end{aligned}$$

where the last equality follows from $\mathrm{E}[\varepsilon_i] = 0$ in A1(v). Finally, since $g_{\mathrm{A}1}(\cdot)$ is known and invertible by A1(ii), this establishes the unique identification of $\theta_{0\mathrm{A}}$. □

Proof of Theorem 2. We first show that under A1 and A2, E1–E6, and the rate conditions in the statement of the theorem,

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

where

$$\sigma_A^2(1) = \lim_{F_w \to 1} \mathrm{E}\left[s_i u_i^2 \,|F_w(F_w), F_w(v_i) < F_w(w)\right]\left[\mathbf{M}_1^{-1}\Gamma_1\mathbf{M}_1^{-1}\right]_{00},$$

with $[A]_{00}$ denoting the upper-left entry of matrix A, and $\mathbf{M}_1$ and Γ_1 are defined above.

Given Assumption A2(i)–(iii), $\lim_{F_w \to 1} \mathrm{E}[s_i\,|F_w(F_w), F_w(v_i) < F_w(w)] = 1$. Moreover, let $\widetilde{m}_A(1)$ be defined as $\widehat{m}_A(1)$ in the text, with $\widehat{F}_w(w_j)$ replaced by $F_w(w_j)$, which we will abbreviate by $\widehat{F}_j$ replaced by F_j in what follows. Finally, we write $\widehat{K}_j(1) = K((\widehat{F}_j - 1)/h)$, $\widehat{\mathcal{P}}_j(1) = \left(1, (\widehat{F}_j - 1), \ldots, (\widehat{F}_j - 1)^q \frac{1}{q!}\right)'$, and $\widehat{\mathcal{Y}}_j = y_j - g_{A2}(x_j'\widehat{\beta}_A)$, and let $K_j(1)$, $\mathcal{P}_j(1)$, and $\mathcal{Y}_j$ be defined accordingly with $\widehat{F}_j$ and $\widehat{\beta}_A$ replaced again by F_j and β_{0A}.

First, letting $e' = (1, 0, \ldots, 0)'$ denote a vector of dimension $((q+1) \times 1)$, note that $\widehat{m}_A(1)$ is defined as the first element of the $((q+1) \times 1)$ vector

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

whereas $\widetilde{m}_A(1)$ is the first element of the corresponding $((q+1) \times 1)$ vector, i.e.,

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

Also, note that $g_{A1}(\theta_{0A})$ is the probability limit of $\widetilde{m}_A(1)$. Given Assumption E5 and recalling Assumptions A2(i) and E1, the empirical process

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n}\left(1\left\{w_j \le w_i\right\} - F_w(w_i)\right)$$

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

$$\sqrt{nh}\,(\widetilde{m}_A(1) - \widehat{m}_A(1)) = o_p(1).$$

Then, recalling that the density of $F_w(w_i)$ is uniform on $(0,1)$, note that by E1, E6, and a Law of Large Numbers for triangular arrays,

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

where $\mathbf{M}_1$ was defined before, and the $((q+1) \times (q+1))$ diagonal matrix is given by

$$\mathbf{G}_h = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & h & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \ldots & \ldots & h^q \end{pmatrix}.$$

Next, using the fact that $\min\{r, q+1\}$ (left) derivatives of $\lambda(\cdot)$ exist and are finite by Assumption E3, we obtain after standard arguments for local polynomial estimators

$$\begin{aligned}
&\sqrt{nh}\left(\widetilde{m}_A(1) - g_{A1}(\theta_{0A})\right) \\
&= e'\mathbf{M}_1{}^{-1}(1+o_p(1))\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\mathbf{G}_h^{-1}s_j\mathcal{P}_j(1)K_j(1)u_j \\
&\quad + e'\mathbf{M}_1{}^{-1}(1+o_p(1))\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\mathbf{G}_h^{-1}s_j\mathcal{P}(1)K_j(1)\left(\frac{1}{\min\{r,q+1\}!}\left(\nabla_{-}^{\min\{r,q+1\}}\lambda(F_j)\left|_{F_j=1}\right.\right)\right. \\
&\quad \left.\times\left(F_j-1\right)^{\min\{r,q+1\}}\right) \\
&\quad + e'\mathbf{M}_1{}^{-1}(1+o_p(1))\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\mathbf{G}_h^{-1}s_j\mathcal{P}(1)K_j(1)\varepsilon_n(1) \\
&= I_{n,h} + II_{n,h} + III_{n,h},
\end{aligned}$$

where $\nabla_{-}^{\min\{r,q+1\}}\lambda(F_j)|_{F_j=1}$ denotes the $\min\{r,q+1\}$th left derivative of $\lambda(\cdot)$ evaluated at $F_j = 1$, whereas (see, e.g., Masry, 1996, p. 575)

$$\begin{aligned}
\varepsilon_n(1) = \left(F_j-1\right)^{\min\{r,q+1\}}\int_0^1&\left(\frac{1}{\min\{r,q+1\}!}\nabla_{-}^{\min\{r,q+1\}}\lambda(F_j)\left|_{F_j=1-\tau(F_j-1)}\right.\right. \\
&\left.-\frac{1}{\min\{r,q+1\}!}\nabla_{-}^{\min\{r,q+1\}}\lambda(F_j)\left|_{F_j=1}\right.\right)(1-\tau)\mathrm{d}\tau.
\end{aligned}$$

Now, given E1, E5, and E6, by a CLT for triangular arrays, we have that

$$I_{n,h} \xrightarrow{d} N\left(0, \sigma_A^2(1)\right), \tag{A.3}$$

where $\sigma_A^2(1)$ was defined in Theorem 2. Note that $II_{n,h}$ and $III_{n,h}$, on the other hand, characterize the bias term. In particular, note that our estimator is computed at the boundary, but that for local polynomial estimators of odd order, the bias is of the same order in the interior and on the boundary (see, e.g., Fan and Gijbels, 1996). Thus, starting with the case of $r \geq q+1$, and using similar arguments to the ones used for Proposition 2 and Theorem 4 of Masry (1996), it follows that

$$\mathrm{E}\left[\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\mathbf{G}_h^{-1}s_j\mathcal{P}(1)K_j(1)\varepsilon_n(1)\right] = o(h^{q+1})$$

and

$$\begin{aligned}
&\left|\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\mathbf{G}_h^{-1}s_j\mathcal{P}(1)K_j(1)\varepsilon_n(1) - \mathrm{E}\left[\frac{1}{\sqrt{nh}}\sum_{j=1}^{n}\mathbf{G}_h^{-1}s_j\mathcal{P}(1)K_j(1)\varepsilon_n(1)\right]\right| \\
&\quad = h^{q+1}O_p\left(\frac{1}{n^{\frac{1}{2}}h^{\frac{1}{2}}}\right) = o_p(h^{q+1}),
\end{aligned}$$

where we note that the last term does not involve a $\ln(n)$ term as in Masry (1996) since we are dealing with the pointwise (and not the uniform) case. Moreover, note that for $II_{n,h}$

$$\left|\frac{h^{-(q+1)}}{\sqrt{nh}}\sum_{j=1}^{n}\mathbf{G}_h^{-1}s_j\mathcal{P}(1)K_j(1)\left(\frac{1}{(q+1)!}\left(\nabla_{-}^{(q+1)}\lambda(F_j)\Big|_{F_j=1}\right)\right.\right.$$
$$\left.\left.\times\left(F_j-1\right)^{(q+1)}\right)-\mathbf{B}_{(q+1)}\right|$$
$$=O_p\left(\frac{1}{n^{\frac{1}{2}}h^{\frac{1}{2}}}\right),$$

where the $((q+1)\times 1)$ vector $\mathbf{B}_{(q+1)}$ is defined as

$$\mathbf{B}_{(q+1)}=\begin{bmatrix}\int_{-1}^{0}\nu^{q+1}K(\nu)d\nu\\ \vdots\\ \int_{-1}^{0}\nu^{2q+1}K(\nu)d\nu\end{bmatrix}.$$

For the case of $r\le q$, we follow Fan and Guerre (2016). Define

$$\left(\overline{a}_{A0}(1),\ldots,\overline{a}_{Aq}(1)\right)$$
$$=\arg\min_{a_k,k\le q}\mathrm{E}\left[s_i\left(y_i-g_{A2}(x_i'\beta_{0A})-\sum_{0\le k\le q}a_k\left(F_w(w_i)-1\right)^k\right)^2K\left(\frac{F_w(w_i)-1}{h}\right)\right],$$

where $\overline{m}_A(1)=\overline{a}_{A0}(1)$. Now,

$$\sqrt{nh}\left(\widehat{m}_A(1)-g_{A1}(\theta_{0A})\right)=\sqrt{nh}\left(\widehat{m}_A(1)-\overline{a}_{A0}(1)\right)+\sqrt{nh}\left(\overline{a}_{A0}(1)-g_{A1}(\theta_{0A})\right),\qquad\textbf{(A.4)}$$

where the first term on the RHS of (A.4) has the same limiting distribution as in (A.3) regardless of r being larger than q or not. Hence, it suffices to consider the second term on (A.4). Our Assumption E3 is equivalent to Assumption S2 in Fan and Guerre (2016) (which in turn implies their S1), whereas our Assumption E6 corresponds to their Assumption K. Finally, their Assumption X holds since $F_w(w_i)$ has marginal density equal to one everywhere on (0,1). Thus, it follows from Theorem 1 in Fan and Guerre (2016) that

$$|\overline{a}_{A0}(1)-g_{A1}(\theta_{0A})|\le Ch^r,$$

and so the bias is of order h^r whenever $r\le q$.

Finally, to complete the proof, recall that $\widehat{m}_A(1)=\widehat{a}_A(1)$ and $m_{0A}(1)=a_{A0}(1)$, Given A1(ii), we can define

$$\widehat{\theta}_A=g_{A1}^{-1}(\widehat{a}_A(1))\text{ and }\theta_{0A}=g_{A1}^{-1}(a_{A0}(1)),$$

and by a mean value expansion, for $\overline{\theta}_A\in\left(\widehat{\theta}_A,\theta_{0A}\right)$,

$$\widehat{a}_{A0}(1)-a_{A0}(1)=g\left(\widehat{\theta}_A\right)-g\left(\theta_{0A}\right)=\nabla_{\theta_A}g\left(\overline{\theta}_A\right)\left(\widehat{\theta}_A-\theta_{0A}\right).$$

Hence,

$$\sqrt{nh}\left(\left(\widehat{\theta}_A - \theta_{0A}\right)\right)$$
$$= \frac{1}{\nabla_{\theta_A} g\left(\overline{\theta}_A\right)} \sqrt{nh}\left(\widehat{a}_{A0}(1) - a_{A0}(1)\right)$$

and given (A.2), it follows that

$$\sqrt{nh}\left(\widehat{\theta}_A - \theta_{0A}\right) \xrightarrow{d} N\left(0, \frac{\sigma_A^2(1)}{\nabla_{\theta_A} g\left(\theta_{0A}\right)^2}\right). \qquad \square$$

Remark A.1. In this remark, we sketch how the procedure of Armstrong and Kolesar (2018) can be used in the context of this paper to construct CIs that are valid uniformly in some range $\underline{h} \le h \le \overline{h}$. Now, since the $\widehat{F}_W(\cdot)$ and $\widehat{\beta}_A$ do not depend on h and converge at a parametric rate, note that, uniformly in h, it holds that

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

Next, recalling the definitions of $\mathbf{M}_1$, $\mathcal{P}_j(1-h)$, and u_j, using similar arguments as in the supplement of Armstrong and Kolesar (2018), we have that

$$\sqrt{nh}\left(\widetilde{m}_A(1-h) - g_{A1}\left(\theta_{0A}(h)\right)\right)$$
$$= e'\mathbf{M}_1 \frac{1}{\sqrt{nh}} \sum_{j=1}^{n} \mathbf{G}_h^{-1} s_j \mathcal{P}_j(1-h) K\left(\frac{F_j - 1}{h}\right) u_j 1\left\{F_j < 1\right\} + o_p\left(\frac{1}{\sqrt{\ln\ln\left(\frac{\overline{h}}{\underline{h}}\right)}}\right).$$

Analogously, one can obtain that

$$\frac{\sqrt{nh}\left(\widehat{\theta}_A(h) - \theta_{0A}(h)\right)}{\widehat{\sigma}_A(h)}$$
$$= \frac{1}{\nabla g_{A1}\left(\theta_{0A}(h)\right)\widehat{\sigma}_{A,h}(1)} e'\mathbf{M}_1 \frac{1}{\sqrt{nh}} \sum_{j=1}^{n} \mathbf{G}_h^{-1} s_j \mathcal{P}_j(1-h) K\left(\frac{F_j - 1}{h}\right) u_j 1\left\{F_j < 1\right\}$$
$$+ o_p\left(\frac{1}{\sqrt{\ln\ln\left(\frac{\overline{h}}{\underline{h}}\right)}}\right),$$

with $\widehat{\sigma}_{A,h}(1)$ defined as in the main text. Now, let

$$K^{\dagger}(z;1) = e'\mathbf{M}_1^{-1}\mathcal{P}(z)K(z).$$

From Section 2.1 of the Supplementary Material of Armstrong and Kolesar (2018), $\sup_{\underline{h} \le h \le \overline{h}} \frac{\sqrt{nh}\left|\widehat{\theta}_A(h) - \theta_{0A}(h)\right|}{\widehat{\sigma}_{A,h}(1)}$ has the same limiting distribution as $\sup_{1 \le t \le \overline{h}/\underline{h}} |H(t)|$, where $H(t)$ is a zero mean Gaussian process with $\mathrm{Cov}(H(t), H(s)) = \rho(s,t;1)$, where

$$\rho(s,t;1) = \frac{\int_{-1}^{0} K^{\dagger}(z/s;1/s) K^{\dagger}(z/t;1/t)\mathrm{d}z}{\sqrt{\int_{-1}^{0} K^{\dagger}(z/s;1/s)^2 \mathrm{d}z \int_{-1}^{0} K^{\dagger}(z/t;1/t)^2 \mathrm{d}z}}.$$

This suggests that we may simulate from a mean-zero multivariate Gaussian process with covariance kernel $\rho(s,t;1)$ from a grid $s,t \in [1,\overline{h}/\underline{h}]$, and compute the $1-\alpha/2$ critical value using the maximum value from each draw, say $c_{1-\alpha/2}(\overline{h}/\underline{h},r,1),$ where r is the order of the local polynomial and 1 the boundary point. The $1-\alpha$ uniformly valid confidence interval (CI) is then given by

$$\widehat{\theta}_A(h) \pm \widehat{\sigma}_{A,h}(1) c_{1-\alpha/2}(\overline{h}/\underline{h},r,1)/\sqrt{nh}.$$

As established in Corollary 3.1 in Armstrong and Kolesar (2018), such a CI is uniformly valid in $\underline{h} \le h \le \overline{h}$. Note that this approach requires that the bias goes to zero sufficiently fast uniformly over $\underline{h} \le h \le \overline{h}$.

Proof of Theorem 3. By a similar argument as in the proof of Theorem 2,

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

where $\sigma_M^2(1)$ was defined in Theorem 3. The statement in the theorem then follows by an application of a standard delta method argument as in the proof of Theorem 2. □

Proof of Theorem 4. First, note that the auxiliary model writes as

$$y_i = g_{M1}(\theta_{0M})\, g_{M2}\left(x_i'\beta_{0M}\right) \overline{\lambda}(p_i) + \overline{u}_i,$$

and let

$$\widetilde{m}_M^p(\delta) = \frac{\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_i \frac{y_i}{g_{M2}(x_i'\beta_M)} K\left(\frac{p(z_i)-\delta}{h_p}\right)}{\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_i K\left(\frac{p(z_i)-\delta}{h_p}\right)}.$$

We first show that

$$\frac{\widetilde{m}_M^p(\delta) - g_{M1}(\theta_{0M})}{\sqrt{\widehat{\operatorname{var}}\left(\widetilde{m}_M^p(\delta) - g_{M1}(\theta_{0M})\right)}} \xrightarrow{d} N(0,1),$$

where $\widehat{\operatorname{var}}(.)$ denotes the estimated variance. Then, by a standard delta method argument,

$$\frac{g_{M1}^{-1}\left(\widetilde{m}_M^p(\delta)\right) - \theta_{0M}}{\sqrt{\nabla_{\theta_M} g_{M1}(\theta_{0M})^2\, \widehat{\operatorname{var}}\left(\widetilde{m}_M^p(\delta) - g_{M1}(\theta_{0M})\right)}} \xrightarrow{d} N(0,1),$$

where $\theta_{0M} = m_M^p(1)$. Now,

$$\begin{aligned}
&\frac{\frac{1}{\sqrt{nh_pH^\eta}}\sum_{i=1}^n s_i \frac{y_i}{g_{M2}(x_i'\beta_M)} K\left(\frac{p(z_i)-\delta}{h_p}\right)}{\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_i K\left(\frac{p(z_i)-\delta}{h_p}\right)} \\
&= \frac{\frac{1}{\sqrt{nh_pH^\eta}}\sum_{i=1}^n s_i g_{M1}(\theta_{0M})\overline{\lambda}(p_i) K\left(\frac{p(z_i)-\delta}{h_p}\right)}{\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_i K\left(\frac{p(z_i)-\delta}{h_p}\right)} + \frac{\frac{1}{\sqrt{nh_pH^\eta}}\sum_{i=1}^n s_i \frac{\overline{u}_i}{g_{M2}(x_i'\beta_M)} K\left(\frac{p(z_i)-\delta}{h_p}\right)}{\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_i K\left(\frac{p(z_i)-\delta}{h_p}\right)} \\
&= I_{n,h_p,H} + II_{n,h_p,H}.
\end{aligned}$$

Note that

$$I_{n,h_p,H} = g_{M1}(\theta_{0M}) + \frac{\frac{1}{\sqrt{nh_pH^\eta}}\sum_{i=1}^n s_i\left(\overline{\lambda}(p_i)-1\right)K\left(\frac{p(z_i)-\delta}{h_p}\right)}{\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_iK\left(\frac{p(z_i)-\delta}{h_p}\right)},$$

and because of Assumption E10M and the nonnegativity of the kernel function,

$$\begin{aligned}&\left|\frac{\frac{1}{\sqrt{nh_pH^\eta}}\sum_{i=1}^n s_i\left(\overline{\lambda}(p_i)-1\right)K\left(\frac{p(z_i)-\delta}{h_p}\right)}{\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_iK\left(\frac{p(z_i)-\delta}{h_p}\right)}\right|\\&\quad\le\sqrt{nh_pH^\eta}\sup_{p\in(1-h_p-H,1+h_p-H)}\left|\overline{\lambda}(p)-1\right|\le C\sqrt{nh_pH^\eta}H^{1-\eta}=o(1)\end{aligned}$$

by rate condition (i). As for the denominator of $II_{n,h_p,H}$, letting $u=\frac{p-\delta}{h_p}$,

$$\begin{aligned}&\mathrm{E}\left[\frac{1}{nh_pH^\eta}\sum_{i=1}^n s_iK\left(\frac{p(z_i)-\delta}{h_p}\right)\right]\\&\quad=\mathrm{E}\left[\frac{1}{h_pH^\eta}s_iK\left(\frac{p(z_i)-\delta}{h_p}\right)\right]=\frac{1}{h_pH^\eta}\int_0^1\Pr(s=1|p)K\left(\frac{p-\delta}{h_p}\right)f_p(p)\mathrm{d}p\\&\quad=\frac{1}{H^\eta}\int_0^1\Pr\left(s=1|p=uh_p+\delta\right)K(u)f_p\left(uh_p+\delta\right)\mathrm{d}u\\&\quad=H^{-\eta}\Pr(s=1|p=1-H)f_p(1-H)+o(1)=c(1)+o(1),\end{aligned}$$

where the second last equality follows from Assumption E8M(i), and the last equality from E8M(ii). Also, recall that $c(1)=f_p(1)$ when $\eta=0$.

As for the limiting distribution of the numerator in $II_{n,h_p,H}$, recalling that $\overline{u}_{x,i}=\frac{\overline{u}_i}{g_{M2}(x_i'\beta_M)}$ and $v=\frac{p-\delta}{h_p}$,

$$\begin{aligned}&\mathrm{var}\left(\frac{1}{\sqrt{nh_pH^\eta}}\sum_{i=1}^n s_i\overline{u}_{x,i}K\left(\frac{p(z_i)-\delta}{h_p}\right)\right)\\&\quad=\frac{1}{h_pH^\eta}\mathrm{var}\left(s_i\overline{u}_{x,i}K\left(\frac{p(z_i)-\delta}{h_p}\right)\right)\\&\quad=\frac{1}{h_pH^\eta}\int_0^1\int_{\mathrm{supp}(u_x)}\Pr(s=1|p,\overline{u}_x)\,\overline{u}_x^2K^2\left(\frac{p-\delta}{h_p}\right)f_{p,\overline{u}}(p,\overline{u}_x)\mathrm{d}\overline{u}_x\mathrm{d}p\\&\quad=\frac{1}{H^\eta}\int_{-1}^1\int_{\mathrm{supp}(u_x)}\Pr\left(s=1|vh_p+\delta,\overline{u}_x\right)\overline{u}_x^2K^2(v)f_{p,\overline{u}_x}(vh_p+\delta,\overline{u}_x)\mathrm{d}\overline{u}_x\mathrm{d}v\\&\quad=\int_{-1}^1K^2(v)\,\mathrm{d}v\int_{\mathrm{supp}(u_x)}\overline{u}_x^2w_{\overline{u},p}(\overline{u}_x,1)\mathrm{d}\overline{u}_x+o(1),\end{aligned}$$

where the last equality follows from Assumption E9M. Hence, as $nh_pH^\eta \to \infty$, using standard arguments together with E7M, it follows that

$$\frac{\sum_{i=1}^{n} s_i \bar{u}_{x,i} K\left(\frac{p(z_i)-\delta}{h_p}\right)}{\sqrt{\int K(v)^2 dv \sum_{i=1}^{n} \bar{u}_{x,i}^2 s_i K\left(\frac{\widehat{p}_i-\delta}{h_p}\right)}} \xrightarrow{d} N(0,1),$$

which gives the limiting distribution of the re-scaled $II_{n,h_p,H}$. It remains to show that

$$\sqrt{nh_pH^\eta}\left(\widehat{m}_M^p(\delta) - \widetilde{m}_M^p(\delta)\right) = o_p(1),$$

as this implies that

$$\sqrt{nh_pH^\eta}\left(g_{M1}^{-1}\left(\widehat{m}_M^p(\delta)\right) - g_{M1}^{-1}\left(\widetilde{m}_M^p(\delta)\right)\right) = o_p(1).$$

Given Assumption E4M, and recalling that $nh_pH^{2-\eta} \to 0,$ to this end, it suffices to show that

$$\frac{1}{\sqrt{nh_pH^\eta}} \sum_{i=1}^{n} \bar{u}_{x,i} s_i \left(K\left(\frac{\widehat{p}_i-\delta}{h_p}\right) - K\left(\frac{p_i-\delta}{h_p}\right)\right) = o_p(1).$$

Given Assumption E7M,

$$\begin{aligned}
&\frac{1}{\sqrt{nh_pH^\eta}} \sum_{i=1}^{n} s_i \bar{u}_{x,i} \left(K\left(\frac{\widehat{p}(z_i)-\delta}{h_p}\right) - K\left(\frac{p(z_i)-\delta}{h_p}\right)\right) \\
&= \frac{1}{\sqrt{nh_pH^\eta}h_p} \sum_{i=1}^{n} s_i \bar{u}_{x,i} \nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_p}\right)(\widehat{p}(z_i) - p(z_i)) \\
&= \frac{1}{\sqrt{nh_pH^\eta}h_p} \sum_{i=1}^{n} s_i \bar{u}_{x,i} \nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_p}\right)\Xi_n(z_i) \\
&\quad + \frac{1}{n^{3/2}h_p^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \bar{u}_{x,i} \nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_i)}\psi_j \\
&= o_p(1) + \underbrace{\frac{1}{n^{3/2}h_p^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}} \sum_{i=1}^{n}\sum_{j=1}^{n} s_i \bar{u}_{x,i} \nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_i)}\psi_j}_{I_{n,h_1,h_p}},
\end{aligned}$$

where the $o_p(1)$ term follows because of E7M, E8M, $\sup_z|\Xi_n(z)| = O(h_1^{\bar{r}})$, $\bar{r} \geq \max\{2, d_z\}$, $nh_ph_1^{2\bar{r}}H^\eta \to 0$, as well as standard change of variables and integration by parts arguments. Now,

$$\begin{aligned}
&I_{n,h_1,h_p} \\
&= \frac{1}{n^{3/2}h_p^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}} \sum_{i=1}^{n} s_i \bar{u}_{x,i} \nabla K\left(\frac{\overline{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}(0)}{f(z_i)}\psi_i
\end{aligned}$$

$$+\frac{1}{n^{3/2}h_p^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}}\sum_{i=1}^{n}\sum_{j>i}^{n}\left(s_i\bar{u}_{x,i}\nabla K\left(\frac{\bar{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_i)}\psi_j\right.$$
$$\left.+s_j\bar{u}_{x,j}\nabla K\left(\frac{\bar{p}(z_j)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_j)}\psi_i\right)$$
$$=I^A_{n,h_1,h_p}+I^B_{n,h_1,h_p}.$$

For the first term I^A_{n,h_1,h_p}, note that

$$\begin{aligned}
I^A_{n,h_1,h_p}&=\frac{1}{n^{3/2}h_p^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}}\sum_{i=1}^{n}s_i\bar{u}_{x,i}\nabla K\left(\frac{\bar{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}(0)}{f(z_i)}\psi_i\\
&=\frac{\sqrt{h_pH^\eta}}{\sqrt{n}h_1^{d_z}}\left(\frac{1}{nH^\eta h_p^2}\sum_{i=1}^{n}s_i\bar{u}_{x,i}\nabla K\left(\frac{\bar{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}(0)}{f(z_i)}\psi_i\right)\\
&=\frac{\sqrt{h_pH^\eta}}{\sqrt{n}h_1^{d_z}}O_p(1)
\end{aligned}$$

by Assumptions E6M and E7M, and bandwidth condition (iii). For the second term on the RHS of I_{n,h_1,h_p}, I^B_{n,h_1,h_p} can be written as a second-order U-statistic:

$$\begin{aligned}
I^B_{n,h_1,h_p}&=\frac{1}{n^{3/2}h_p^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}}\sum_{i=1}^{n}\sum_{j>i}^{n}\left(s_i\bar{u}_{x,i}\nabla K\left(\frac{\bar{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_i)}\psi_j\right.\\
&\quad\left.+s_j\bar{u}_{x,j}\nabla K\left(\frac{\bar{p}(z_j)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_j)}\psi_i\right)\\
&\cong\frac{2\sqrt{n}}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n}\left(\Psi_{i,j,n}+\Psi_{j,i,n}\right),
\end{aligned}$$

where

$$\Psi_{i,j,n}=\frac{1}{h_p^{3/2}h_1^{d_z}H^{\frac{\eta}{2}}}s_i\bar{u}_{x,i}\nabla K\left(\frac{\bar{p}(z_i)-\delta}{h_p}\right)\frac{\overline{\mathbf{K}}\left(\frac{z_i-z_j}{h_1}\right)}{f(z_i)}\psi_j.$$

Given that $\bar{u}_{x,i}$ has conditional mean zero, it follows that $\mathrm{E}\left[\Psi_{i,j,n}|s_i,x_i,z_i,p_i\right]=0$, and hence that the U-statistic is degenerate. Also, by change of variables and standard arguments, from $nh_1^{d_z}h_p^2H^\eta\to\infty$ and E7M, we have that

$$\mathrm{E}\left[\Psi^2_{i,j,n}\right]=o(n).$$

Hence, by Lemma 3.1 in Powell, Stock, and Stoker (1989) and the degeneracy of the U-statistic, we can conclude that the second term is

$$\left(\frac{2\sqrt{n}}{n(n-1)} \sum_{i=1}^{n} \sum_{j>i}^{n} \left(\Psi_{i,j,n} + \Psi_{j,i,n} \right) \right) = o_p(1),$$

which completes the proof. □

A.2. Additional Figures and Tables

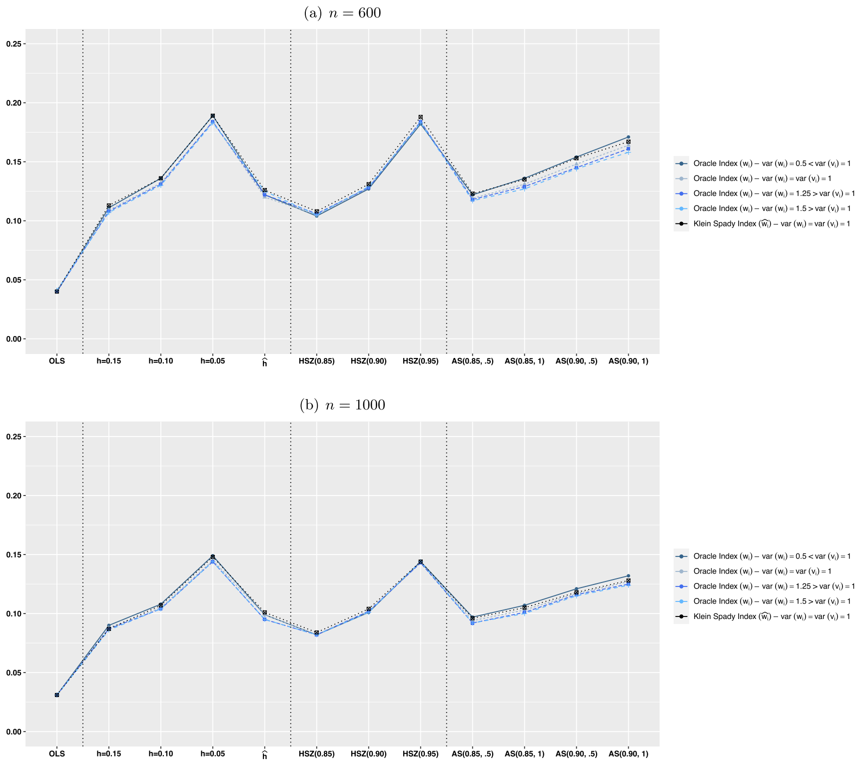
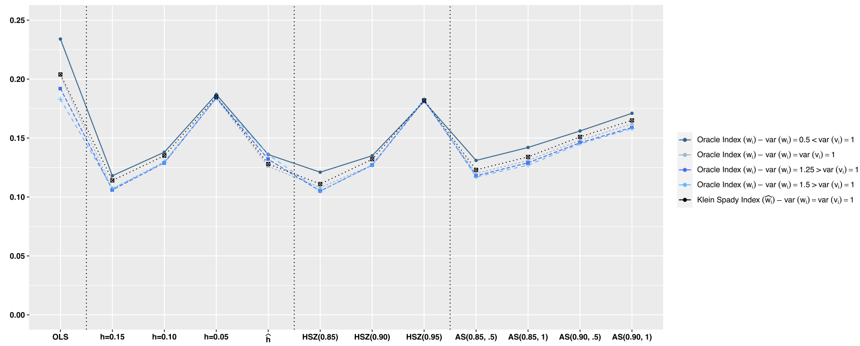


FIGURE A.1. RMSE comparison—additive error model (CASE I)—$\rho = 0$.
Note: (1) Number of Monte Carlo replications: 1,500. (2) $h = 0.15, 0.10$, and 0.05 correspond to the estimator $\widehat{\theta}_A$ with a fixed bandwidth size, whereas $\widehat{h}$ denotes the same estimator with a data-driven bandwidth. (3) HSZ$(\cdot)$ corresponds to the estimator (18), with δ_n set to the 85%, 90%, and 95% (unconditional) quantiles of $z_i'\widehat{\gamma}$. (4) AS$(\cdot,\cdot)$ corresponds to the estimator in (19), with δ_n again set to the 85% and 90% quantiles and $b \in \{0.5, 1\}$ (the RMSE results for $AS(0.95, 0.5)$ and $AS(0.95, 1)$ can be found in the tables in the Supplementary Material).

(a) $n = 600$



(b) $n = 1000$

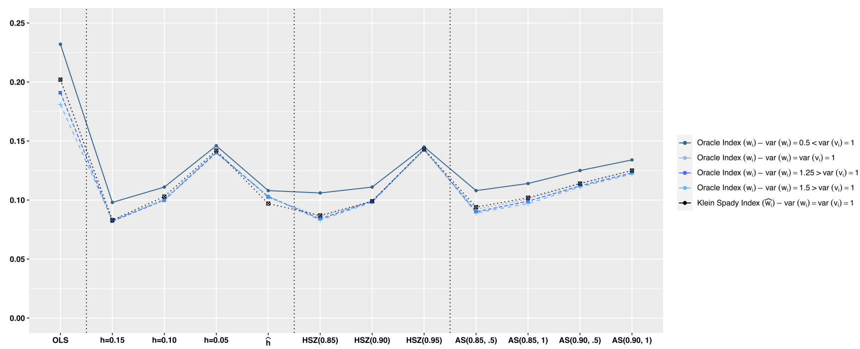


FIGURE A.2. RMSE comparison—additive error model (CASE I)—$\rho = +0.5$.

Note: (1) Number of Monte Carlo replications: 1,500. (2) $h = 0.15, 0.10$, and 0.05 correspond to the estimator $\widehat{\theta}_A$ with a fixed bandwidth size, whereas $\widehat{h}$ denotes the same estimator with a data-driven bandwidth. (3) HSZ$(\cdot)$ corresponds to the estimator (18), with δ_n set to the 85%, 90%, and 95% (unconditional) quantiles of $z_i'\widehat{\gamma}$. (4) AS$(\cdot,\cdot)$ corresponds to the estimator in (19), with δ_n again set to the 85% and 90% quantiles and $b \in \{0.5, 1\}$ (the RMSE results for $AS(0.95, 0.5)$ and $AS(0.95, 1)$ can be found in the tables in the Supplementary Material).

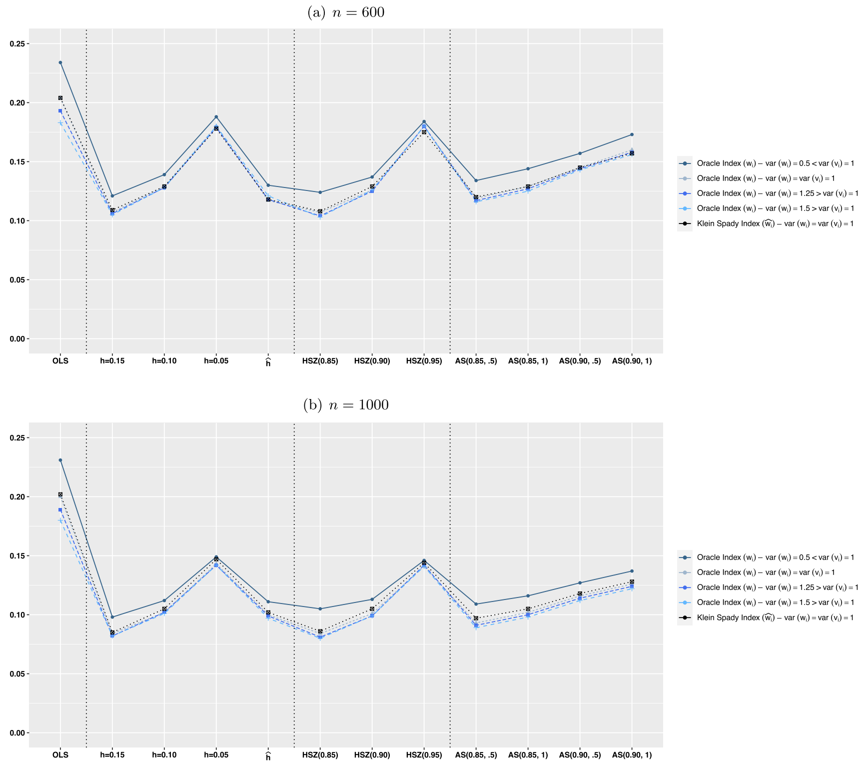
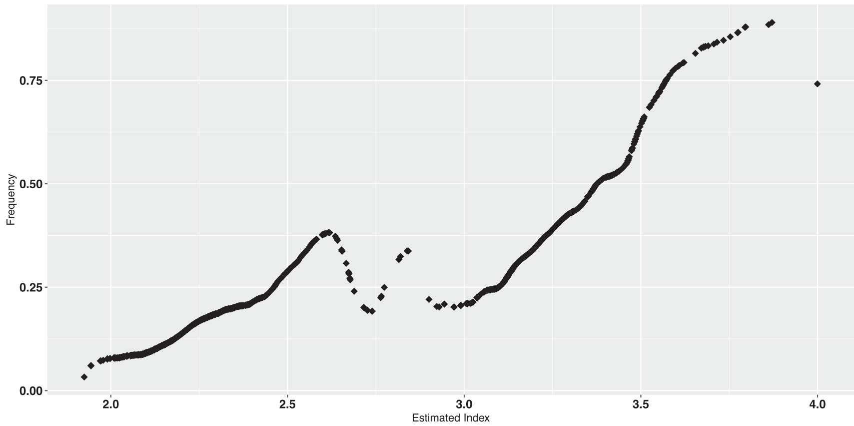


Figure A.3. RMSE comparison—additive error model (CASE I)—$\rho = -0.5$.

Note: (1) Number of Monte Carlo replications: 1,500. (2) $h = 0.15, 0.10$, and 0.05 correspond to the estimator $\widehat{\theta}_A$ with a fixed bandwidth size, whereas $\widehat{h}$ denotes the same estimator with a data-driven bandwidth. (3) HSZ($\cdot$) corresponds to the estimator (18), with δ_n set to the 85%, 90%, and 95% (unconditional) quantiles of $z_i'\widehat{\gamma}$. (4) AS($\cdot,\cdot$) corresponds to the estimator in (19), with δ_n again set to the 85% and 90% quantiles and $b \in \{0.5, 1\}$ (the RMSE results for $AS(0.95, 0.5)$ and $AS(0.95, 1)$ can be found in the tables in the Supplementary Material).

(a) Estimated Propensity Score $\widehat{p}(\widehat{w}_i)$ for Observations with $\text{HS}_i = 0$



(b) Estimated Propensity Score $\widehat{p}(\widehat{w}_i)$ for Observations with $\text{HS}_i = 1$

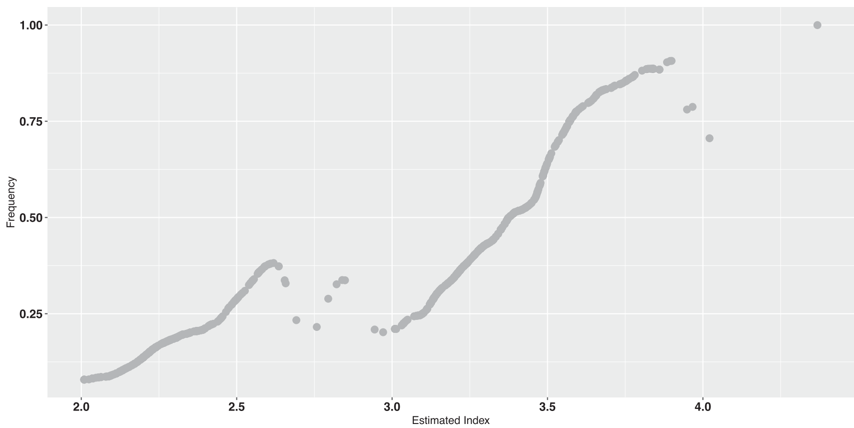
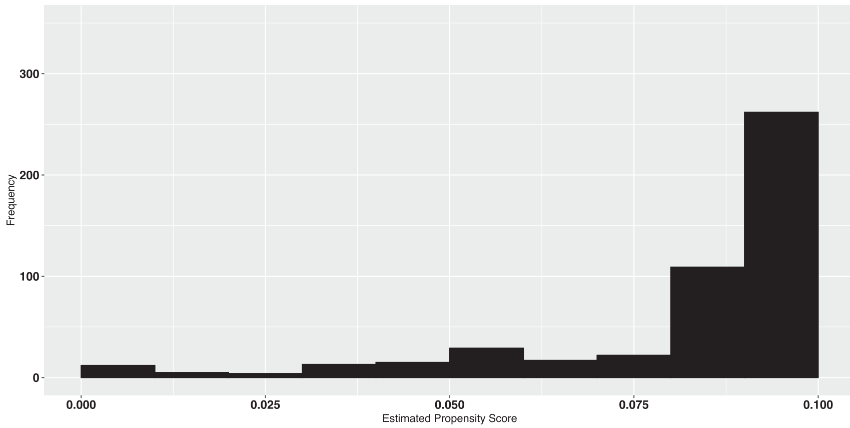


FIGURE A.4. Empirical distribution of $\widehat{p}(\widehat{w}_i)$ with index restriction $\widehat{w}_i = z_i'\widehat{\gamma}$.

(a) $\widehat{p}(z_i) \in (0.0, 0.1)$ for Observations with $\mathrm{HS}_i = 0$



(b) $\widehat{p}(z_i) \in (0.9, 1.0)$ for Observations with $\mathrm{HS}_i = 1$

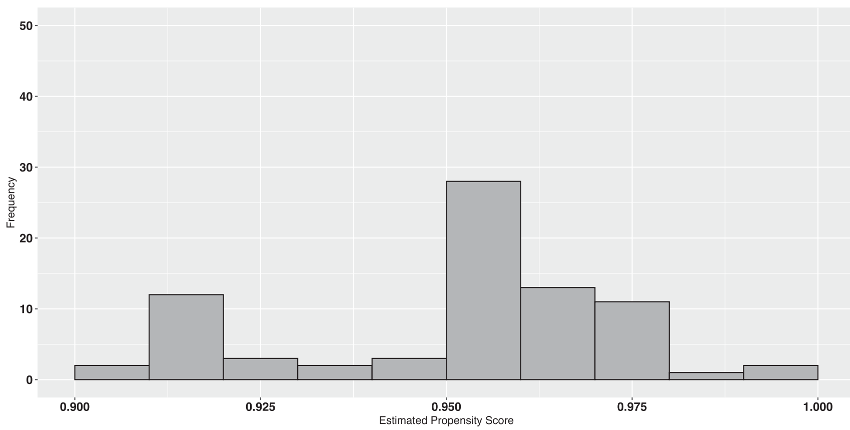


FIGURE A.5. Empirical distribution of estimated nonparametric propensity score.

TABLE A.1. Descriptive statistics

Variable	Description	Mean	Std. Dev.	Min	Max
Dependent variable: DOCVIS	Number of doctor visits/seen in the last month	0.401	0.796	0	9
Endogeneous covariate: H	Self-reported health is poor or fair	0.252	0.434	0	1
Male	Sex=male	0.437	0.496	0	1
Edu	Highest educational qualification: GCSE OL or higher	0.557	0.497	0	1
Inc	Post-tax weekly personal income is at least £250	0.185	0.389	0	1
Tempsick	Out of work as temporarily sick	0.004	0.064	0	1
Hlim	Activities in last month limited by health	0.109	0.312	0	1
Excluded covariates (not in x_i)					
Perm_Sick	Current work status—permanently sick	0.029	0.167	0	1
Retired	Current work status—retired	0.276	0.447	0	1
Soc3	Social class—other nonmanual	0.192	0.394	0	1
Soc4	–Skilled manual	0.336	0.472	0	1
Soc5	–Semi skilled manual and personal services	0.150	0.357	0	1
Soc6	–Unskilled	0.049	0.216	0	1
Accom	Accommodation—Bungalow	0.106	0.308	0	1
Wine	Number of units of wine consumption last week	1.483	3.677	0	53
Winesq	Wine squared/100	15.718	87.877	0	2809
Crntsmkr	Current smoker	0.286	0.452	0	1
Disab	Has long-standing disability	0.343	0.475	0	1

SUPPLEMENTARY MATERIAL

Arulampalam, Wiji, Valentina Corradi, and Daniel Gutknecht (2023). Supplement to "Intercept Estimation in Nonlinear Selection Models," Econometric Theory Supplementary Material. To view, please visit: https://doi.org/10.1017/S0266466623000105

REFERENCES

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

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

Armstrong, T. & M. Kolesar (2018) A simple adjustment for bandwidth snooping. *Review of Economic Studies* 85, 732–765.

Armstrong, T. & M. Kolesar (2020) Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics* 11(1), 1–39.

Blair, J., C. Edwards, & J. Johnson (1976) Rational "Chebyshev" approximations for the inverse of the error function. *Mathematics of Computation* 30(136), 827–830.

Calonico, S., M. Cattaneo, & R. Titiunic (2014) Robust nonparametric confidence interval for regression discontinuity design. *Econometrica* 82, 2295–2326.

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

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

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

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

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

Fan, J. and I. Gijbels (1996) *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability, 66. Chapman and Hall/CRC.

Fan, Y. and E. Guerre (2016) Multivariate local polynomial estimators: Uniform boundary properties and asymptotic linear representation. In G. Gozalez-Rivera, R. Hill, and T.-H. Lee (eds.), *Essays in Honor of Aman Ullah*. Advances in Econometrics, 36, pp. 489–538. Emerald.

Feller, W. (1968) *An Introduction to Probability Theory and Its Applications*, vol. 1. Wiley.

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

Gourieroux, C., A. Monfort, & A. Trognon (1984) Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica* 52(3), 701–720.

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

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

Hayfield, T. & J. Racine (2008) Nonparametric econometrics: The np package. *Journal of Statistical Software* 27(5), 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. & L. Hu (2020) Selection without exclusion. *Econometrica* 88(3), 1007–1029.

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

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

Klein, R. & 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. & J. Racine (2004) Nonparametric estimation of regression functions with both categorical and continuous data. *Journal of Econometrics* 119(1), 99–130.

Li, Q. & 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.

Masry, E. (1996) Multivariate local polynomial regression for time series: Uniform strong consistency and rates. *Journal of Time Series Analysis* 17(6), 571–599.

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, & T. Stoker (1989) Semiparametric estimation of index coefficients. *Econometrica* 57, 1403–1430.

Racine, J. (1993) An efficient cross-validation algorithm for window width selection for nonparametric kernel regression. *Communications in Statistics* 22(4), 1107–1114.

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

Schafgans, M. (1998) Ethnic wage differences in Malaysia: Parametric and semiparametric estimation of the Chinese–Malay wage gap. *Journal of Applied Econometrics* 13, 481–504.

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

Schafgans, M. & 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.

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

Vytlacil, E. (2002) Independence, monotonicity, and latent index model: An equivalent result. *Econometrica* 70, 331–341.

Windmeijer, F. & J. Santos Silva (1998) Endogeneity in count data models: An application to demand for health care. *Journal of Applied Econometrics* 12(3), 281–294.