Rate-Optimal Estimation of the Intercept in a Semiparametric Sample-Selection Model*

Chuan Goh†

September 2018

Abstract

This paper presents a new estimator of the intercept of a linear regression model in cases where the outcome variable is observed subject to a selection rule. The intercept is often in this context of inherent interest; for example, in the evaluation of social programs, the difference between the intercepts in outcome equations for participants and non-participants can be interpreted as the difference in average outcomes of the self-selected participants and their counterfactual average outcomes if these participants had in fact chosen not to participate. The new estimator can exhibit a rate of convergence in probability equal to $n^{-p/(2p+1)}$, where $p \geq 2$ is an integer that indexes the strength of certain smoothness assumptions. This rate of convergence is in this context the optimal rate of convergence for estimation of the intercept parameter in terms of a minimax criterion. The new estimator, unlike other proposals in the literature, is under mild conditions consistent and asymptotically normal with a rate of convergence that is the same regardless of the degree to which selection depends on unobservables in the outcome equation. Simulation evidence and an empirical example are included.

JEL *Classification:* C14, C31, C34
Keywords: sample selection, rate optimality, minimax efficiency, nearest neighbours

---

*The author is grateful to Shakeeb Khan for several helpful conversations and to seminar participants at Louisiana State University and The University of Wisconsin-Madison for their comments. The usual disclaimer applies. The author also thanks Daniel Siercks for computing support.

†Department of Economics and Finance, University of Guelph, 50 Stone Road East, Guelph, ON, Canada N1G 2W1. gohc@uoguelph.ca, http://www.chuangoh.org.

1 Introduction

This paper considers a regression model in which selectivity bias is potentially a problem, i.e., a model in which the observability of the outcome variable depends on a non-random selection mechanism. In particular, we consider the model given by

$$\begin{aligned} Y^* &= \theta_0 + \boldsymbol{X}^\top \boldsymbol{\beta}_0 + U, & (1)\\ D &= 1\left\{\boldsymbol{Z}^\top \boldsymbol{\gamma}_0 \geq V\right\}, & (2)\\ Y &= DY^*, & (3) \end{aligned}$$

where $[\ D \quad \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top \quad Y\ ]$ is an observed random vector, and where $[\ U \quad V\ ]$ is an unobserved random vector such that $E\left[U^2\right] < \infty$ and $E[U|\boldsymbol{X}] = 0$ and $E[U|\boldsymbol{Z}] = E[U]$ almost surely. Models of this form apparently originated in the literature in labour economics on wage comparisons between socioeconomic groups (Gronau, 1974; Heckman, 1974; Lewis, 1974) and have come to be referred to as "sample-selection" (e.g., Heckman, 1976, 1979) or "type 2" Tobit models (Amemiya, 1984). Sample selection models involve data-generating processes in which the outcome variable is not independent of its observability to the researcher; for example, the volume of goods produced for export by a firm is expected to depend on the firm's decision to enter a particular export market (Helpman et al., 2008). More generally, the dependence of an outcome variable on its observability is a feature of many applied problems in sociology, political science, economics, criminology, finance and other disciplines (see e.g., Winship and Mare, 1992; Collier and Mahoney, 1996; Vella, 1998; Bushway et al., 2007; Lennox et al., 2012, and references cited for discipline-specific applications). Relatively recent economic applications include those of Helpman et al. (2008), Mulligan and Rubinstein (2008) and Jiménez et al. (2014).

This paper focuses on statistical inference regarding the intercept θ_0 appearing in the outcome equation (1). The intercept in the outcome equation is often of inherent interest in various applications of the sample-selection model. For example, suppose that (2) accurately describes the selection of individuals into some treatment group. In this case, the difference between the intercepts in the outcome equations for treated and non-treated individuals may be interpreted as the causal effect of treatment when selection to treatment is mean independent of the unobservable U in the outcome equation (e.g., Andrews and Schafgans, 1998, p. 500). This interpretive framework would apply for example to the analysis of the wage effects of migration between geographical regions; in particular, the difference between the intercepts in log-wage equations for individuals who choose to migrate and those who choose to remain in their region of birth can be interpreted as the counterfactual difference in average log wages between workers who chose to migrate and otherwise identical workers who did not migrate.

Early applied work generally proceeded from the assumption that the unobservables $[\ U \quad V\ ]$ appearing above in (1)–(2) are bivariate normal mean-zero with an unknown covariance matrix and independent of $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$. This assumption in turn allowed for the estimation of the parameters appearing in (1)–(2) via the method of maximum likelihood or the related two-step procedure of Heckman (1976, 1979). These estimates, however, are generally inconsistent under departures from the assumed bivariate normality of $[\ U \quad V\ ]$ (e.g., Arabmazar and Schmidt, 1982; Little, 1982; Olsen, 1982; Goldberger, 1983; Lee, 1983; Paarsch, 1984; Zuehlke and Zeman, 1991; Schafgans, 2004). The desirability of not imposing a parametric specification on the joint distribution of the unobservables in the outcome and selection equations has led in turn to the development of distribution-free methods of estimating the parameters appearing in (1)–(2). Distribution-free methods of estimating the intercept θ_0 in (1) include the proposals of Gallant and Nychka (1987); Heckman (1990); Andrews and Schafgans (1998) and Lewbel (2007).

Estimators of the intercept of the outcome equation implemented by distribution-free procedures have to date been characterized by large-sample behaviours that vary depending on the extent of endogeneity in the selection mechanism, i.e., on the nature and extent of any dependence between the random variables U and V appearing in (1) and (2), respectively. Given that these features of the joint distribution of $[\ U \quad V\ ]$ are typically unknown in empirical practice, the dependence of the asymptotic behaviour of intercept estimators on these features potentially complicates statistical inference regarding θ_0. This issue is easily and starkly illustrated in the case of ordinary least squares (OLS). In particular, suppose that selection in the model is based strictly on observables, which is equivalent to assuming that the unobservable U appearing in (1) and the selection indicator D are conditionally mean independent given $\boldsymbol{X}$ and $\boldsymbol{Z}$, i.e., $P\left[E[U|D=1,\boldsymbol{X},\boldsymbol{Z}]=E[U|D=0,\boldsymbol{X},\boldsymbol{Z}]\right]=1$. In this case θ_0 can be consistently estimated at a parametric rate with no additional assumptions imposed on the joint distribution of $[\ U \quad V \quad \boldsymbol{Z}^\top\boldsymbol{\gamma}_0\ ]$ by applying OLS to the outcome equation using only those observations for which $D=1$. On the other hand, the OLS estimate of θ_0 obtained in this way is inconsistent if the difference $1-P\left[E[U|D=1,\boldsymbol{X},\boldsymbol{Z}]=E[U|D=0,\boldsymbol{X},\boldsymbol{Z}]\right]$ is positive, even if arbitrarily small. It follows that OLS generates inferences regarding θ_0 that vary drastically with respect to the degree to which $E[U|D=1,\boldsymbol{X},\boldsymbol{Z}]$ may differ from $E[U|D=0,\boldsymbol{X},\boldsymbol{Z}]$.

This paper develops a distribution-free estimator of the intercept θ_0 in the outcome equation that is consistent and asymptotically normal with a rate of convergence that is the same regardless of the joint distribution of $[\ U \quad V \quad \boldsymbol{Z}^\top\boldsymbol{\gamma}_0\ ]$. I show that there exists an implementation of the proposed estimator of θ_0 that converges uniformly at the rate $n^{-p/(2p+1)}$, where n denotes the sample size and where $p \geq 2$ is an integer that indexes the strength of certain smoothness assumptions described below. The uniformity of this convergence involves uniformity over a class of joint distributions of $[\ U \quad V \quad \boldsymbol{Z}^\top\boldsymbol{\gamma}_0\ ]$

satisfying necessary conditions for the identification of θ_0. In other words, the estimator developed below is *adaptive* to the nature of selection in the model.

This paper also shows that the uniform $n^{-p/(2p+1)}$-rate attainable by the proposed estimator is in fact the optimal rate of convergence of an estimator of θ_0 in terms of a minimax criterion. It follows that the proposed estimator may be implemented in such a way as to converge in probability to θ_0 at the fastest possible minimax rate.

The estimator developed below differs from earlier proposals of Heckman (1990) and Andrews and Schafgans (1998) that involve the *rate-adaptive* estimation of the intercept in the outcome equation. These proposals also involve estimators that are consistent and asymptotically normal regardless of the extent to which selection is endogenous, but converge to the limiting normal distribution at unknown rates; in particular, see Schafgans and Zinde-Walsh (2002, Theorems 1–2) for the estimator of Heckman (1990) and Andrews and Schafgans (1998, Theorems 2, 3, 5 and 5*).

The estimator developed below also differs from estimators of θ_0 that take the form of averages weighted by the reciprocal of an estimate of the density of the selection index $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ appearing above in (2). Such estimators (e.g., Lewbel, 2007), in common with the estimators of Heckman (1990) and Andrews and Schafgans (1998), are known to converge generically at unknown rates. In addition, estimators taking the form of inverse density-weighted averages may have sampling distributions that are not even asymptotically normal (Khan and Tamer, 2010). In general, the estimators of Heckman (1990), Andrews and Schafgans (1998) and those in the class of inverse density-weighted averages converge at rates that depend critically on conditions involving the relative tail thicknesses of the distributions of the selection index $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ and of the latent selection variable V appearing in (2) (Khan and Tamer, 2010). These conditions may be difficult to verify in applications. The estimator developed below converges by contrast at a known rate under conditions implied by the identification of θ_0 to a normal distribution uniformly over the underlying parameter space regardless of the relative tail behaviours of $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ and V. This facilitates statistical inference regarding θ_0.

The remainder of this paper proceeds as follows. The following section discusses identification of the intercept parameter in the outcome equation and presents the new estimator along with its first-order asymptotic properties. Section 3 derives the minimax rate optimality of the new estimator. Section 4 presents the results of simulation experiments that investigate the behaviour in finite samples of the proposed estimator in relation to other methods. Section 5 considers an application of the new estimator to the analysis of gender wage gaps in Malaysia. Section 6 concludes. Proofs of all theoretical results are collected in the appendix.

2 The New Estimator

This section presents the new estimator of the intercept θ_0 appearing in (1) and describes its asymptotic behaviour to first order. Let $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ denote $\sqrt{n}$-consistent estimators of the parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ appearing above in (1) and (2), respectively, where $\boldsymbol{\gamma}_0$ is assumed to be identified up to a location and scale normalization. The existence of such estimators has long been established; see e.g., the proposals of Han (1987); Robinson (1988); Powell et al. (1989); Andrews (1991); Ichimura and Lee (1991); Ichimura (1993); Klein and Spady (1993); Powell (2001) or Newey (2009). In addition, suppose that $\left\{[\ D_i \quad \boldsymbol{X}_i^\top \quad Y_i \quad \boldsymbol{Z}_i^\top\ ] : i = 1, \ldots, n\right\}$ are iid copies of the random vector $[\ D \quad \boldsymbol{X}^\top \quad Y \quad \boldsymbol{Z}^\top\ ]$. Let $z \in \mathbb{R}^l$ be an arbitrary vector, and define

$$\hat{\eta}_n(z) \equiv \frac{1}{n} \sum_{i=1}^{n} 1\left\{(\boldsymbol{Z}_i - z)^\top \hat{\boldsymbol{\gamma}}_n \leq 0\right\} \tag{4}$$

Let

$$\hat{W}_i \equiv D_i \left(Y_i - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}_n\right) \tag{5}$$

for each $i \in \{1, \ldots, n\}$. This paper proposes to estimate θ_0 via a locally linear smoother of the form

$$\hat{\theta}_n \equiv \boldsymbol{e}_1^\top \left(\sum_{i=1}^{n} \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^{n} \boldsymbol{S}_i K_i \hat{W}_i, \tag{6}$$

where $\boldsymbol{e}_1 = [\ 1 \quad 0\ ]^\top$ and where

$$\boldsymbol{S}_i = [\ 1 \quad \hat{\eta}_n(\boldsymbol{Z}_i) - 1\ ]^\top \tag{7}$$

and

$$K_i = K\left(h_n^{-1}\left(\hat{\eta}_n(\boldsymbol{Z}_i) - 1\right)\right) \tag{8}$$

for $i = 1, \ldots, n$. The quantity h_n appearing in each K_i denotes a bandwidth $h_n > 0$ such that $h_n \to 0$ with $nh_n^3 \to \infty$ as $n \to \infty$, while for some $p \geq 2$, $K(\cdot)$ denotes a smoothing kernel of order p, i.e., one where $\int_{-\infty}^{\infty} K(u)du = 1$, $\int_{-\infty}^{\infty} u^r K(u) = 0$ for all $r \in \{1, \ldots, p-1\}$ and $\int_{-\infty}^{\infty} u^p K(u) < \infty$.

Assume that the disturbance U in the outcome equation (1) satisfies $E[|U|] < \infty$ with $P\left[E[U|\boldsymbol{Z}] = 0\right] = 1$. In addition, let the selection index $\boldsymbol{Z}^\top \boldsymbol{\gamma}_0$ be distributed with distribution F_0, assumed to be absolutely continuous. The estimator of θ_0 given in (6) exploits the fact that identification of θ_0 depends crucially on the selection index $\boldsymbol{Z}^\top \boldsymbol{\gamma}_0$ being able to take values sufficiently large so that the corresponding conditional probabilities of selection take values close to one. In particular, θ_0 is characterized by the equalities

$$\begin{aligned} & E\left[D\left(Y - \boldsymbol{X}^\top \boldsymbol{\beta}_0\right) \middle| F_0\left(\boldsymbol{Z}^\top \boldsymbol{\gamma}_0\right) = 1\right] \\ = \; & E\left[1\left\{F_0(V) \leq 1\right\}\left(\theta_0 + U\right) \middle| F_0\left(\boldsymbol{Z}^\top \boldsymbol{\gamma}_0\right) = 1\right] \\ = \; & \theta_0 \end{aligned} \tag{9}$$

The proposed estimator of θ_0 exploits the representation of the estimand in (9), which suggests the estimation of θ_0 by direct estimation of the quantity $E\left[D\left(Y - \boldsymbol{X}^\top \boldsymbol{\beta}_0\right) \middle| F_0\left(\boldsymbol{Z}^\top \boldsymbol{\gamma}_0\right) = 1\right]$.

One can view the estimator given in (6) as an extension of the Yang–Stute symmetrized nearest-neighbours (SNN) estimator of a conditional mean (Yang, 1981; Stute, 1984) to the problem of estimating the intercept in the outcome equation (1). SNN estimators are characterized by asymptotic behaviours that are asymptotically "design adaptive" in the sense that their asymptotic normality can be established without technical conditions on the probability of the design variable taking values in regions of low density (Stute, 1984). In the present context, one can show that the estimator $\hat{\theta}_n$ in (6) is asymptotically normal with a rate of convergence that depends neither on the extent to which the unobservables U and V are dependent, nor on the relationship between the behaviours of the selection index $\boldsymbol{Z}^\top \boldsymbol{\gamma}_0$ and the unobservable V in the right tails of their respective marginal distributions. This is essentially accomplished by transforming the estimated indices $\boldsymbol{Z}_i^\top \hat{\boldsymbol{\gamma}}_n$ into random variables $\hat{\eta}_n(\boldsymbol{Z}_i)$ that are approximately uniformly distributed on $[0, 1]$.

Estimators of θ_0 implemented by distribution-free procedures have to date been characterized by rates of convergence that vary with the extent to which U and V are dependent, as well as with the relative right-tail behaviours of $\boldsymbol{Z}^\top \boldsymbol{\gamma}_0$ and V. This is particularly true for those estimators of θ_0 that take the form of inverse density-weighted averages (e.g., Lewbel, 2007; Khan and Tamer, 2010). These estimators rely on an alternative representation of (9) and are consistent under additional regularity conditions. In this connection suppose $m_0\left(\boldsymbol{z}^\top \boldsymbol{\gamma}_0\right) \equiv E\left[D\left(Y - \boldsymbol{X}^\top \boldsymbol{\beta}_0\right) \middle| \boldsymbol{Z}^\top \boldsymbol{\gamma}_0 = \boldsymbol{z}^\top \boldsymbol{\gamma}_0\right]$ is everywhere differentiable in $\boldsymbol{z}^\top \boldsymbol{\gamma}_0$ with derivative given by $m_0^{(1)}\left(\boldsymbol{z}^\top \boldsymbol{\gamma}_0\right) \equiv \partial m_0\left(\boldsymbol{z}^\top \boldsymbol{\gamma}_0\right) / \partial \boldsymbol{z}^\top \boldsymbol{\gamma}_0$. Suppose in addition that the distribution of $\boldsymbol{Z}^\top \boldsymbol{\gamma}_0$ is absolutely continuous with density $f_0(\cdot)$ such that $E\left[\left| m_0^{(1)}\left(\boldsymbol{Z}^\top \boldsymbol{\gamma}_0\right) / f_0\left(\boldsymbol{Z}^\top \boldsymbol{\gamma}_0\right)\right|\right] < \infty$. One can then write

$$\begin{aligned} \theta_0 &= \lim_{F_0(\boldsymbol{z}^\top \boldsymbol{\gamma}_0) \to 1} E\left[D\left(Y - \boldsymbol{X}^\top \boldsymbol{\beta}_0\right) \middle| \boldsymbol{Z} = \boldsymbol{z}\right] \\ &= \int_{-\infty}^{\infty} m_0^{(1)}\left(\boldsymbol{z}^\top \boldsymbol{\gamma}_0\right) d\boldsymbol{z}^\top \boldsymbol{\gamma}_0 \\ &= \int_{-\infty}^{\infty} \frac{m_0^{(1)}\left(\boldsymbol{z}^\top \boldsymbol{\gamma}_0\right)}{f_0\left(\boldsymbol{z}^\top \boldsymbol{\gamma}_0\right)} \cdot f_0\left(\boldsymbol{z}^\top \boldsymbol{\gamma}_0\right) d\boldsymbol{z}^\top \boldsymbol{\gamma}_0 \end{aligned}$$

$$= \quad E\left[\frac{m_0^{(1)}\left(\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right)}{f_0\left(\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right)}\right], \tag{10}$$

which suggests estimating θ_0 via its approximate sample analogue in which $m_0^{(1)}(\cdot)$ and $f_0(\cdot)$ are replaced by suitable preliminary estimates, and in which the systematic trimming of observations corresponding to small values of $f_0\left(\boldsymbol{Z}_i^\top\boldsymbol{\gamma}_0\right)$ may be required. This approach to estimating θ_0 follows that proposed by Lewbel (1997) for estimating a binary choice model arising from a latent linear model in which a mean restriction is imposed on the latent error term, and is the approach to estimating θ_0 considered in more recent work by Lewbel (2007); Khan and Tamer (2010) and Khan and Nekipelov (2013). Consistent estimators of θ_0 that exploit (10) in this way naturally depend critically on the assumed finiteness of $E\left[\left|m_0^{(1)}\left(\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right)/f_0\left(\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right)\right|\right]$, an assumption that implies that the rates at which they converge to their limiting distributions depend on the relative tail behaviours of the variables in the selection equation or the extent to which selection is endogenous (Khan and Tamer, 2010). The non-uniformity in the rate of convergence as one varies the relative tail behaviours of the determinants of selection or the dependence between the disturbance terms U and V is a feature that is also shared by estimators of θ_0 that involve locally constant or polynomial regressions of $\hat{W}_i$ on the untransformed estimated selection indices $\boldsymbol{Z}_i^\top\hat{\boldsymbol{\gamma}}_n$ (e.g., Heckman, 1990; Andrews and Schafgans, 1998). Results of this nature significantly complicate the task of statistical inference regarding θ_0. By way of contrast, the transformations to $\hat{\eta}_n(\boldsymbol{Z}_i)$ of the selection indices used in the locally linear SNN estimator given in (6) permit the locally linear SNN estimator to enjoy asymptotic normality with a rate of convergence that varies neither with the endogeneity of selection nor with the relative tail behaviours of the variables appearing in the selection equation.

It should also be noted that the result given above in (9), in which the estimand is identified as $\theta_0 = E\left[D\left(Y - \boldsymbol{X}^\top\boldsymbol{\beta}_0\right)\middle| F_0(\boldsymbol{Z}^\top\boldsymbol{\gamma}_0) = 1\right]$, motivates the formulation of the estimator in (6) as the intercept in a locally linear regression. One could as easily in this context use (9) to motivate an estimator of θ_0 as the corresponding variant of a Nadaraya–Watson (i.e., locally constant regression) estimator; see in particular the approach taken in Stute and Zhu (2005). The focus on a locally linear regression estimator of θ_0 is purely to improve the rate at which the bias of the proposed estimator vanishes in large samples, given that locally linear regression estimators have biases that converge at the same rate regardless of whether the conditioning variable is evaluated at an interior or at a limit point of its support (e.g., Fan and Gijbels, 1992). Nadaraya–Watson estimators, on the other hand, have biases that converge relatively slowly when the conditioning variable is evaluated at a limit point of its support.

Assumptions underlying the first-order asymptotic behaviour of the estimator given by

$\hat{\theta}_n$ in (6) are given as follows.

Assumption 1. *1.* *(a)* $\boldsymbol{X}$ *is k-variate, with support not contained in any proper linear subspace of* $\mathbb{R}^k$*;*

(b) $E[\|\boldsymbol{X}\|] < \infty$.

2. *(a)* $\boldsymbol{Z}$ *is l-variate;*

(b) the support of $\boldsymbol{Z}$ *is not contained in any proper linear subspace of* $\mathbb{R}^l$*,;*

(c) the first component of $\boldsymbol{\gamma}_0$ *is equal to one;*

(d) $\boldsymbol{Z}$ *does not contain a non-stochastic component;*

(e) the distribution F_0 *of the selection index* $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ *is absolutely continuous, with a density function* f_0 *that is differentiable on the support of* $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$.

3. The set $\left\{(D_i, \boldsymbol{X}_i^\top, \boldsymbol{Z}_i^\top, Y_i) : i = 1, \ldots, n\right\}$ *consists of independent observations each with the same distribution as the random vector* $(D, \boldsymbol{X}^\top, \boldsymbol{Z}^\top, Y)$*, which is generated according to the model given above in (1)–(3), and where*

$$\begin{aligned} E\left[U^2\right] &< \infty; \\ P\left[E[U|\boldsymbol{X}] = 0\right] &= 1; \\ P\left[E[U|\boldsymbol{Z}] = E[U]\right] &= 1. \end{aligned}$$

4. *(a) The joint conditional distribution given* $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ] = [\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ *of the disturbances* $[\ U \quad V\ ]$ *appearing in (1) and (2) is absolutely continuous for all* $[\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ *in the support of* $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$*; moreover the corresponding joint density* $g_{U,V|\boldsymbol{x},\boldsymbol{z}}(\cdot,\cdot)$ *is continuously differentiable in both arguments almost everywhere on* $\mathbb{R}^2$.

(b) The conditional distribution given $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ] = [\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ *of* V *is absolutely continuous for all points* $[\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ *in the support of* $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$*; the corresponding density* $g_{V|\boldsymbol{x},\boldsymbol{z}}(\cdot)$ *is differentiable almost everywhere on* $\mathbb{R}$.

Assumption 2. *1. There exist estimators* $\hat{\boldsymbol{\beta}}_n$ *and* $\hat{\boldsymbol{\gamma}}_n$ *such that* $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right\| = O_p\left(n^{-1/2}\right)$ *and* $\|\hat{\boldsymbol{\gamma}}_n - \boldsymbol{\gamma}_0\| = O_p\left(n^{-1/2}\right)$.

2. The smoothing kernel $K(\cdot)$ *is bounded and twice continuously differentiable with* $K(u) > 0$ *on* $[0,1]$*,* $K(u) = 0$ *for all* $u \notin [0,1]$*, with* $\int K(u)du = 1$ *and where* $\int K^2(u)du < \infty$.

3. *The bandwidth sequence* $\{h_n\}$ *satisfies* $h_n > 0$ *with* $h_n \to 0$ *as* $n \to \infty$*, and* $nh_n^3 \to \infty$.

4. (a) *There exists* $p \geq 2$ *such that* $\int u^j K(u)du = 0$ *for all* $j \in \{1, \ldots, p-1\}$ *and* $\int u^p K(u)du < \infty$.

   (b) *The following hold for* p^* *equal to the smallest odd integer greater than or equal to the constant* $p + 1$ *specified in part 4a of this assumption:*

      i. *The joint conditional density* $g_{U,V|\boldsymbol{x},\boldsymbol{z}}(\cdot,\cdot)$ *specified in Assumption 1.4 is* p^**-times continuously differentiable in both arguments almost everywhere on* $\mathbb{R}$ *for all* $[\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ *in the support of* $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$.

      ii. *Similarly, the conditional density* $g_{V|\boldsymbol{x},\boldsymbol{z}}(\cdot)$ *specified in Assumption 1.4 is* p^**-times continuously differentiable almost everywhere on* $\mathbb{R}$ *for all* $[\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ *in the support of* $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$.

The conditions of Assumption 1 are largely standard and notably suffice for the selection parameter $\boldsymbol{\gamma}_0$ to be identified up to the particular location and scale normalization imposed by Assumption 1.2. Assumption 1 also does not restrict the components V and $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ of the selection equation to be independent.

Assumption 1 plays a crucial role in controlling the asymptotic bias of the proposed estimator $\hat{\theta}_n$. In particular, identification of $\boldsymbol{\gamma}_0$ subject to the conditions of Assumption 1, along with the differentiability conditions of Assumption 2.4(b)i–2.4(b)ii, imply a smoothness restriction on the conditional mean function

$$m_{F_0}(q) \equiv E\left[D\left(Y - \boldsymbol{X}^\top\boldsymbol{\beta}_0\right)\middle|F_0\left(\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right) = q\right], \tag{11}$$

This smoothness restriction takes the form of differentiability of $m_{F_0}(q)$ for $q \in (0,1)$ up to order no less than p, where $p \geq 2$ is the constant specified in Assumption 2.4, along with finiteness of the left-hand limit of $(d^p/dq^p)\, E\left[D\left(Y - \boldsymbol{X}^\top\boldsymbol{\beta}_0\right)\middle|F_0\left(\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right) = q\right]$ at $q = 1$. This smoothness restriction, in other words, corresponds to a standard assumption in the literature on kernel estimation of conditional mean functions. On the other hand, the differentiability to p^*-order in the second argument of $g_{U,V|\boldsymbol{x},\boldsymbol{z}}(\cdot,\cdot)$ is slightly stronger than the usual assumption of differentiability to order p. This slight strengthening of the standard differentiability condition is used in the rate optimality arguments developed below in Section 3. Details are contained in the proof of Theorem 2 below.

The smoothness restriction on $m_{F_0}(q)$ given in (11) can also be seen to be implied by the identification of $\boldsymbol{\gamma}_0$ up to a location and scale normalization and by the smoothness conditions imposed in Assumption 2.4 on the conditional densities $g_{U,V|\boldsymbol{x},\boldsymbol{z}}(\cdot,\cdot)$ and $g_{V|\boldsymbol{x},\boldsymbol{z}}(\cdot,\cdot)$ for any $[\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ in the support of $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$. In particular, one can show that under the conditions of Assumption 1 and 2.4, U has a conditional distribution given $F_0(V) = q$ for any $q \in [0, 1]$ that is absolutely continuous with density

$$r_{U|Q}(u|q) \equiv \frac{g_{UV}\left(u, F_0^{-1}(q)\right)}{g_V\left(F_0^{-1}(q)\right)}, \tag{12}$$

where $g_{UV}(\cdot, \cdot)$ and $g_V(\cdot)$ are respectively the joint density of $[\ U \quad V\ ]$ and the marginal density of V. The conditional density $r_{U|Q}(u|q)$ is, given the absolute continuity of F_0 and the differentiability conditions on g_{UV} and g_V implied by Assumption 2.4, $(p + 1)$-times differentiable in q on $(0, 1)$ for any $u \in \mathbb{R}$. The $(p + 1)$-times differentiability of $r_{U|Q}(u|q)$ in q on $(0, 1)$ in turn implies the finiteness of $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1}$ for any $u \in \mathbb{R}$. It is the finiteness of $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1}$ that implies the smoothness restrictions on $m_{F_0}(q)$ mentioned above. Further details are supplied below in Appendix A.1.

Let $\sigma^2_{U|F_0(\mathbf{Z}^\top\boldsymbol{\gamma}_0)}(q) \equiv E\left[U^2 \,\middle|\, F_0\left(\mathbf{Z}^\top\boldsymbol{\gamma}_0\right) = q\right]$, where U is the disturbance in the outcome equation (1). The following result summarizes the large-sample behaviour to first order of the proposed estimator:

Theorem 1. *Under the conditions of Assumptions 1 and 2, the estimator $\hat{\theta}_n$ given above in (6) satisfies*

$$\sqrt{nh_n}\left(\hat{\theta}_n - \theta_0 - \frac{h_n^p}{p!}\int u^p K(u)du \cdot m_{F_0}^{(p)}(1)\right) \xrightarrow{d} N\left(0, \sigma^2_{U|F_0(\mathbf{Z}^\top\boldsymbol{\gamma}_0)}(1)\int K^2(u)du\right)$$

as $n \to \infty$, where $m_{F_0}^{(p)}(1) = \lim_{q\uparrow 1}\, (d^p/(dq'^p)\, m_{F_0}(q')\big|_{q'=q}$ for $m_{F_0}(\cdot)$ as given above in (11).

It follows from Theorem 1 that the rate of convergence of $\hat{\theta}_n$ to its limiting normal distribution is unaffected by the dependence, if any, between the disturbance terms U and V in (1) and (2), respectively. The rate of convergence of $\hat{\theta}_n$ is also unaffected by the relative upper tail thicknesses of the distributions of V and of the selection index $\mathbf{Z}^\top\boldsymbol{\gamma}_0$.

We also have from the statement of Theorem 1 that a necessary condition for the consistency of $\hat{\theta}_n$ is the finiteness of the derivative $m_{F_0}^{(p)}(1)$. The finiteness of $m_{F_0}^{(p)}(1)$, as discussed above, is implied by the identification of $\boldsymbol{\gamma}_0$ up to a location and scale normalization as well as by the differentiability conditions specified in Assumptions 2.4(b)i–2.4(b)ii. From this it follows that the consistency of the proposed estimator is implied by natural restrictions on the joint distribution of $[\ U \quad V \quad \mathbf{Z}^\top\boldsymbol{\gamma}_0\ ]$. These distributional restrictions correspond collectively to a standard smoothness restriction in the literature on kernel estimation of conditional mean functions.

The presence of $m_{F_0}^{(p)}(1)$ in the bias term appearing in Theorem 1, however, indicates that the approximate large-sample bias of $\hat{\theta}_n$ does depend on the extent to which U is mean dependent on V. In particular, the conditional mean derivative $m_{F_0}^{(p)}(1)$ depends on the smoothness of the conditional mean $E\left[U\,|F_0(V) = q\right]$ as a function of q for values of q near one; Appendix A.1 below contains further discussion. It is worth noting in this connection that $m_{F_0}^{(p)}(1) = 0$ when the selection mechanism is exogenous to the extent that U is mean independent of V, i.e., when $P\left[E[U|V] = E[U]\right] = 1$.

The dependence of the approximate large-sample bias of $\hat{\theta}_n$ on the joint distribution of $[\ U \quad V \quad \mathbf{Z}^\top\boldsymbol{\gamma}_0\ ]$ through the conditional mean derivative $m_{F_0}^{(p)}(1)$ can be ameliorated in practice by a judicious choice of variable bandwidth; see Corollary 1 below and the corresponding discussion and simulation evidence presented in Section 4. Theorem 1 in any case indicates that the approximate large-sample bias, but not the variance, of the proposed estimator depends on the joint distribution of $[\ U \quad V \quad \mathbf{Z}^\top\boldsymbol{\gamma}_0\ ]$. Theorem 1 as such distinguishes the asymptotic behaviour of $\hat{\theta}_n$ from those of existing estimators of θ_0 (e.g., Heckman, 1990; Lewbel, 2007) whose biases and variances both depend on the joint distribution of $[\ U \quad V \quad \mathbf{Z}^\top\boldsymbol{\gamma}_0\ ]$.

The following corollary is immediate from Theorem 1:

Corollary 1. *The following hold under the conditions of Theorem 1:*

1. *If the additional condition that $nh_n^{2p+1} \to 0$ holds, we have*

$$\sqrt{nh_n}\left(\hat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \sigma^2_{U|F_0(\mathbf{Z}^\top\boldsymbol{\gamma}_0)}(1)\int K^2(u)du\right)$$

   as $n \to \infty$.

2. *The theoretical bandwidth h_n^* minimizing the asymptotic mean-squared error of $\hat{\theta}_n$ is given by*

$$h_n^* = \left[\frac{(p!)^2\sigma^2_{U|F_0(\mathbf{Z}^\top\boldsymbol{\gamma}_0)}(1)\int K^2(u)du}{2p\left(\int u^p K(u)du\right)^2\left(m_{F_0}^{(p)}(1)\right)^2\cdot n}\right]^{\frac{1}{2p+1}}.$$

3 Rate Optimality

Continue to let $p \geq 2$ be as specified above in Assumption 2 in the previous section, and n the sample size. This section shows that under the conditions of Assumptions 1 and 2, the rate $n^{-p/(2p+1)}$ is the fastest achievable, or *optimal*, rate of convergence of an estimator of the intercept θ_0 in (1). The optimality in question is relative to the convergence rates of all other estimators of θ_0 and excludes by definition those estimators that are asymptotically superefficient at particular points in the underlying parameter space. The exclusion of superefficient estimators in this context notably rules out estimators that converge at the parametric rate of $n^{-1/2}$ under conditions stronger than those given above in Assumptions 1 and 2. For example, the OLS estimator of θ_0 based on observations for which $D_i = 1$ is superefficient for specifications of (1)–(3) over submodels in which the disturbance U and the selection indicator D are conditionally mean independent given $\boldsymbol{X}$ and $\boldsymbol{Z}$, i.e., models where $P\left[E[U|D = 1, \boldsymbol{X}, \boldsymbol{Z}] = E[U|D = 0, \boldsymbol{X}, \boldsymbol{Z}] = 0\right] = 1$. As noted in the Introduction, the OLS estimator of θ_0 converges at the standard rate of $n^{-1/2}$ when U and D are conditionally mean independent given $\boldsymbol{X}$ and $\boldsymbol{Z}$ but is otherwise inconsistent.

The approach to optimality taken here follows that of Horowitz (1993), which was in turn based on the approach of Stone (1980). In particular, let $\{\Psi_n : n = 1, 2, 3, \ldots\}$ denote a sequence of sets of the form

$$\Psi_n = \{\psi : \psi = (\boldsymbol{\psi}_1, g)\}, \tag{13}$$

where $\boldsymbol{\psi}_1 = (\theta, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top \in \mathbb{R}^{1+k+l}$ and where g denotes the joint conditional density given $\boldsymbol{X}$ and $\boldsymbol{Z}$ of the disturbances U and V appearing above in (1) and (2), respectively. The quantity ψ may depend generically on n.

Consider the observable random variables D, Y, $\boldsymbol{X}$ and $\boldsymbol{Z}$ appearing above in (1)–(3). Suppose that for each n, the joint conditional distribution of the vector (D, Y) given $\boldsymbol{X}$ and $\boldsymbol{Z}$ is indexed by some $\psi \in \Psi_n$. Let $P_\psi[\cdot] \equiv P_{(\boldsymbol{\psi}_1, g)}[\cdot]$ denote the corresponding probability measure. Following Stone (1980), one may in this context define a constant $\rho > 0$ to be an *upper bound* on the rate of convergence of estimators of the intercept parameter θ_0 if for every estimator sequence $\{\theta_n\}$,

$$\liminf_{n\to\infty} \sup_{\psi\in\Psi_n} P_\psi\left[|\theta_n - \theta| > sn^{-\rho}\right] > 0 \tag{14}$$

for all $s > 0$, and if

$$\lim_{s\to 0} \liminf_{n\to\infty} \sup_{\psi\in\Psi_n} P_\psi\left[|\theta_n - \theta| > sn^{-\rho}\right] = 1, \tag{15}$$

where θ as it appears in (14) and (15) refers to the first component of the finite-dimensional component $\boldsymbol{\psi}_1$ of ψ.

In addition, define $\rho > 0$ to be an *achievable* rate of convergence for the intercept parameter if there exists an estimator sequence $\{\theta_n\}$ such that

$$\lim_{s\to\infty} \limsup_{n\to\infty} \sup_{\psi\in\Psi_n} P_\psi\left[|\theta_n - \theta| > sn^{-\rho}\right] = 0. \tag{16}$$

One calls $\rho > 0$ the *optimal* rate of convergence for estimation of the intercept parameter if it is both an upper bound on the rate of convergence and achievable. In what follows, I first show that for large n, $p/(2p+1)$ is an upper bound on the rate of convergence. I then show that there exists an implementation of the estimator given in (6) that attains the $n^{-p/(2p+1)}$-rate of convergence uniformly over Ψ_n as $n \to \infty$.

The approach taken first involves the specification for each n of a subset Ψ_n^* of the parameter set Ψ_n in which the finite-dimensional component $\boldsymbol{\psi}_1 \equiv \boldsymbol{\psi}_{1n}$ lies in a shrinking neighbourhood Ψ_{1n}^* of some point $[\ \theta_0 \quad \boldsymbol{\beta}_0^\top \quad \boldsymbol{\gamma}_0^\top\ ]^\top \in \mathbb{R}^{1+k+l}$. In addition, the infinite-dimensional component g is embedded in a curve (i.e., parametrization) indexed by a scalar ψ_{2n} on a shrinking neighbourhood Ψ_{2n}^* of a bivariate density function g_0 satisfying all relevant conditions of Assumptions 1 and 2 for a conditional density of U and V given $\boldsymbol{X}$ and $\boldsymbol{Z}$.

In particular, consider a parametrization of the conditional joint density g of (U, V) given $\boldsymbol{X}$ and $\boldsymbol{Z}$ given by $g_{\psi_{2n}}$ for $\psi_2 \in \Psi_{2n}^*$, where for some $\psi_{2n0} \in \Psi_{2n}^*$, we have $g_{\psi_{2n0}}(u, v|\boldsymbol{x}, \boldsymbol{z}) = g_0(u, v|\boldsymbol{x}, \boldsymbol{z})$ for each $[\ u \quad v \quad \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ] \in \mathbb{R}^{2+k+l}$; i.e., the curve on Ψ_{2n}^* given by $\psi_{2n} \to g_{\psi_{2n}}$ passes through the true conditional joint density g_0 at some point $\psi_{2n0} \in \Psi_{2n}^*$.

Now let $\Psi_n^* \equiv \Psi_{1n}^* \times \Psi_{2n}^*$. Let $s > 0$ be arbitrary, and let $\{\theta_n\}$ denote an arbitrary sequence of estimators of θ_0. Consider that if

$$\liminf_{n\to\infty} \sup_{\psi_n \in \Psi_n^*} P_{\psi_n}\left[n^{\frac{p}{2p+1}}\,|\theta_n - \theta| > s\right] > 0, \tag{17}$$

then (14) holds with $\rho = p/(2p+1)$. This is because the set Ψ_n in (14) contains the set over which the supremum is taken in (17). Similarly, if

$$\lim_{s\to 0} \liminf_{n\to\infty} \sup_{\psi_n \in \Psi_n^*} P_{\psi_n}\left[n^{\frac{p}{2p+1}}\,|\theta_n - \theta| > s\right] = 1 \tag{18}$$

holds, then so does (15).

It follows that proving (17) and (18) suffices to show that $p/(2p+1)$ is an upper bound on the rate of convergence; the key step in the proof is the specification of a suitable parametrization $\psi_{2n} \to g_{\psi_{2n}}$ for $\psi_{2n} \in \Psi_{2n}^*$. This is in fact the approach taken in Appendix A.3, which contains a proof of the following result:

Theorem 2. *Under the conditions of Assumptions 1 and 2, (17) and (18) hold.*

Theorem 2 implies that $p/(2p+1)$ is an upper bound on the rate of convergence of an estimator sequence $\{\theta_n\}$ in the minimax sense of (14) and (15) above.

Next, it is shown that $p/(2p+1)$ is an achievable rate of convergence in the sense of (16) by exhibiting an estimator sequence $\{\theta_n\}$ such that (16) holds with $\rho = p/(2p+1)$. In this connection, let $\hat{\theta}_n^*$ denote the proposed estimator given above in (6) implemented with a bandwidth $h_n^* = cn^{-1/(2p+1)}$ for some constant $c > 0$. In this case, (16) is satisfied with $\theta_n = \hat{\theta}_n^*$ and $\rho = p/(2p+1)$:

Theorem 3. *Suppose Assumptions 1 and 2 hold. Then (16) holds with $\theta_n = \hat{\theta}_n^*$ and $\rho = p/(2p+1)$, where $\hat{\theta}_n^*$ denotes the estimator given above in (6) implemented with a bandwidth $h_n^* = cn^{-1/(2p+1)}$ for some constant $c > 0$.*

Theorems 2 and 3 jointly imply that $p/(2p+1)$ is the optimal rate of convergence for estimation of θ_0.

4 Numerical Evidence

This section reports the results of simulation experiments that compare the finite-sample behaviour of the estimator in (6) to the behaviours of alternative estimators. The simulations involved:

- variation in the correlation between the unobservable terms in the outcome and selection equations;
- variation in the relative upper tail thicknesses of the selection index and the unobservable term in the selection equation, thus implying variation in the degree to which the parameter of interest is identified;
- and the imposition of two different parametric families for the joint distribution of $[\ U \quad V \quad \boldsymbol{Z}^\top \boldsymbol{\gamma}_0\ ]$, where U is the unobservable term in the outcome equation, V is the unobservable term in the selection equation and $\boldsymbol{Z}^\top \boldsymbol{\gamma}_0$ is the selection index.

Each simulation experiment involved 1000 replicated samples of sizes $n \in \{100, 400\}$ from the model given above in (1)–(3), where the parameter of interest was fixed at $\theta_0 = 1$, the variance of the unobservable term in the selection equation was fixed at $Var[V] = 1$ and where for some constant $\rho \in [-1, 1]$, the unobservable term in the outcome equation was specified as $U = \rho V + E$ for a random variable E independent of V, where $E \sim N\left(0, 1 - \rho^2\right)$. The parameter ρ in this case is by construction the correlation coefficient between U and V. The simulations considered the settings $\rho \in \{0, .25, .50, .75, .95\}$.

In addition, the vector $\boldsymbol{Z}$ of observable predictors of selection was taken to be l-variate with $\boldsymbol{Z} = [\ Z_1 \quad \cdots \quad Z_l\ ]^\top$, while the vector $\boldsymbol{X}$ of outcome predictors was specified to be

k-variate with $k < l$ and $\boldsymbol{X} = [\ Z_1 \ \cdots \ Z_k\ ]^\top$. The coefficient vector attached to $\boldsymbol{X}$ was set to $\boldsymbol{\beta}_0 = \boldsymbol{\iota}_k$, i.e., the k-dimensional unit vector. The simulations imposed the settings $l = 7$, $k = 4$, which were intentionally set to equal the dimensions of the corresponding vectors appearing in the empirical example used below in Section 5.

The selection index $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ and the selection-equation disturbance term were simulated from two data-generating processes (DGPs), considered in turn. The distributions of $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ under both DGPs, as required by Assumption 1, are absolutely continuous. In addition, the parameter $\alpha > 0$ used in the specification of both models is defined so as to index the degree to which the parameter of interest θ_0 is identified. In particular, $\alpha \geq 1$ can be seen in this context to be a necessary condition for the identification of θ_0, with $\alpha \geq 1$ corresponding to the case where the transformation $F_0(V)$ has an absolutely continuous distribution with density given by the ratio $r_Q(q)$ defined in (28) below. Values of $\alpha < 1$, on the other hand, correspond to the non-identifiability of θ_0, with $\alpha \in (0, 1)$ in the context of either of the following two DGPs implying failure of the condition that $F_0(V)$ have an absolutely continuous distribution supported on $[0, 1]$. In particular, $\alpha \in (0, 1)$ in the following two DGPs implies that the quantity $r_Q(q)$ given in (28) below has the property that $r_Q(1) = \infty$:

- (DGP1) I take
$$\begin{bmatrix} \boldsymbol{Z} \\ V \end{bmatrix} \sim N\left(\boldsymbol{0}, \boldsymbol{I}_{l+1}\right).$$
In addition, the selection parameter $\boldsymbol{\gamma}_0$ is set to $\boldsymbol{\gamma}_0 = [\ \sqrt{\alpha/l} \ \cdots \ \sqrt{\alpha/l}\ ]^\top$ for a constant $\alpha > 0$. In this way we have
$$\begin{bmatrix} \boldsymbol{Z}^\top\boldsymbol{\gamma}_0 \\ V \end{bmatrix} \sim N\left(\boldsymbol{0}, \begin{bmatrix} \alpha & 0 \\ 0 & 1 \end{bmatrix}\right).$$

- (DGP2) $Z_1, \ldots, Z_l$ are iid standard Cauchy and jointly mutually independent of V, while V is absolutely continuous on $[1, \infty)$ with Pareto type-I density given by
$$g_{0V}(v) = \alpha v^{-\alpha-1},$$
where $\alpha > 0$ is a constant. In addition, the selection parameter $\boldsymbol{\gamma}_0$ is set to $\boldsymbol{\gamma}_0 = [\ \boldsymbol{0}_{l-1}^\top \ 1\ ]^\top$. In this way the selection index $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ is standard Cauchy and independent of V.

The simulations under both DGPs considered the settings $\alpha \in \{2.00, 1.50, 1.25, 1.00\}$. The effect of variation in the correlation coefficient ρ and the parameter α on estimation of the intercept was a primary focus of these simulations. The outcome-equation nuisance parameter $\boldsymbol{\beta}_0$ and the selection parameter $\boldsymbol{\gamma}_0$ were accordingly fixed at their true values in these simulations in order to provide a clearer picture of the effects of variation in ρ and α on the behaviour of the various intercept estimators considered.

The proposed intercept estimator given above in (6) was implemented with a standard (i.e., second-order) Epanechnikov kernel. The bandwidth used to implement $\hat{\theta}_n$ was initially set to the sample analogue of the theoretical asymptotic MSE-optimal bandwidth h_n^* specified above in Corollary 1. In particular, the simulations involved the bandwidth $\hat{h}_n^*$, where $\hat{h}_n^*$ was taken to be the sample analogue of h_n^*. As such, $\hat{h}_n^*$ was set to decay at the MSE-optimal rate of $n^{-1/5}$ corresponding to the order of kernel employed (i.e., $p = 2$), while its leading constant was specified as the sample analogue of the leading constant appearing in h_n^*. The unknown parameters appearing in the leading constant of h_n^* were estimated via auxiliary locally cubic regressions as described in Fan and Gijbels (1996, §4.3). The sensitivity of the proposed estimator's sampling behaviour to the choice of bandwidth was also assessed by considering implementations using the bandwidth settings $h_n = (2/3)\hat{h}_n^*$ and $h_n = (3/2)\hat{h}_n^*$.

Comparisons of the corresponding sampling behaviours in terms of squared bias, standard deviation and root mean-squared error (RMSE) over 1000 Monte Carlo replications for values of $(\rho, \alpha) \in \{0, .25, .50, .75, .95\} \times \{2.00, 1.50, 1.25, 1.00\}$ are presented below for samples of size $n = 100$ in Tables 1 and 2 for DGP1 and DGP2, respectively. The corresponding results for samples of size $n = 400$ appear in Tables 3 and 4. The RMSE figures displayed in these tables are multiplied by $\sqrt{n}$ in order to provide a clearer indication of the rate of convergence of the proposed estimator. The increases in $\sqrt{n}\times$ RMSE as one moves from simulated samples of size $n = 100$ to those of $n = 400$ indicate the slower-than-parametric rate of convergence of the proposed estimator regardless of the precise setting of (ρ, α). It is also clear that the rate of convergence of the proposed estimator is slower for settings of (ρ, α) with one or both of ρ and α close to one. In addition, and as predicted by Theorem 1 above, one can see that the effect of variation in (ρ, α) on the squared bias of the proposed estimator is more pronounced than the corresponding effect on the variance; indeed the standard deviation of the proposed estimator tends to be relatively stable over the various settings of (ρ, α) used in the simulations, particularly for values of $(\rho, \alpha) \in \{.25, .50, .75\} \times \{2.00, 1.50, 1.25\}$. Finally, the tabulated results indicate that the sampling performance of $\hat{\theta}_n$ is not sensitive to moderate variations in bandwidth.

Table 1: DGP1 (bivariate normal), $n = 100$, 1000 replications. Proposed estimator with second-order Epanechnikov kernel ($p = 2$). (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(optimal bandwidth)												
0.0000	0.0081	0.1859	2.0645	0.0018	0.1872	1.9196	0.0008	0.1951	1.9723	0.0006	0.1900	1.9154
0.2500	0.0003	0.1890	1.8985	0.0004	0.1857	1.8674	0.0023	0.1911	1.9695	0.0064	0.1802	1.9724
0.5000	0.0014	0.1812	1.8503	0.0071	0.1861	2.0432	0.0127	0.1802	2.1253	0.0227	0.1779	2.3324
0.7500	0.0092	0.1768	2.0125	0.0216	0.1784	2.3125	0.0353	0.1761	2.5750	0.0459	0.1687	2.7272
0.9500	0.0215	0.1821	2.3373	0.0404	0.1665	2.6105	0.0548	0.1730	2.9114	0.0812	0.1626	3.2811
(2/3× optimal bandwidth)												
0.0000	0.0081	0.1860	2.0657	0.0018	0.1873	1.9206	0.0008	0.1952	1.9733	0.0006	0.1901	1.9164
0.2500	0.0003	0.1891	1.8996	0.0004	0.1858	1.8682	0.0023	0.1912	1.9702	0.0064	0.1803	1.9727
0.5000	0.0014	0.1813	1.8507	0.0071	0.1862	2.0430	0.0126	0.1803	2.1243	0.0227	0.1780	2.3314
0.7500	0.0091	0.1770	2.0109	0.0215	0.1786	2.3100	0.0351	0.1762	2.5724	0.0457	0.1689	2.7242
0.9500	0.0212	0.1822	2.3336	0.0401	0.1666	2.6058	0.0545	0.1732	2.9068	0.0809	0.1628	3.2764
(3/2× optimal bandwidth)												
0.0000	0.0081	0.1858	2.0640	0.0018	0.1872	1.9192	0.0008	0.1951	1.9719	0.0006	0.1900	1.9150
0.2500	0.0003	0.1890	1.8980	0.0004	0.1857	1.8670	0.0023	0.1910	1.9692	0.0064	0.1802	1.9723
0.5000	0.0014	0.1811	1.8502	0.0071	0.1861	2.0433	0.0127	0.1802	2.1257	0.0228	0.1779	2.3328
0.7500	0.0093	0.1768	2.0132	0.0217	0.1784	2.3136	0.0354	0.1760	2.5762	0.0460	0.1687	2.7285
0.9500	0.0216	0.1820	2.3389	0.0406	0.1664	2.6126	0.0550	0.1729	2.9134	0.0814	0.1625	3.2831

Table 2: DGP2 (non-normal), $n = 100$, 1000 replications. Proposed estimator with second-order Epanechnikov kernel ($p = 2$) and optimal bandwidth. (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(optimal bandwidth)												
0.0000	0.1541	0.1719	4.2854	0.1892	0.1805	4.7097	0.2268	0.1708	5.0593	0.2714	0.1612	5.4535
0.2500	0.0202	0.2119	2.5523	0.0394	0.2070	2.8669	0.0445	0.2229	3.0684	0.0636	0.2280	3.3994
0.5000	0.0079	0.2391	2.5503	0.0071	0.2741	2.8687	0.0006	0.2942	2.9521	0.0000	0.3010	3.0102
0.7500	0.1113	0.2796	4.3536	0.0972	0.3386	4.6030	0.0735	0.3288	4.2615	0.0585	0.4767	5.3448
0.9500	0.2747	0.3174	6.1273	0.2510	0.3605	6.1721	0.2317	0.4509	6.5956	0.1950	0.4976	6.6528
(2/3× optimal bandwidth)												
0.0000	0.1522	0.1725	4.2652	0.1871	0.1813	4.6906	0.2247	0.1715	5.0405	0.2693	0.1619	5.4357
0.2500	0.0190	0.2125	2.5336	0.0376	0.2078	2.8422	0.0426	0.2242	3.0466	0.0613	0.2293	3.3752
0.5000	0.0091	0.2396	2.5796	0.0084	0.2754	2.9025	0.0010	0.2960	2.9769	0.0000	0.3032	3.0319
0.7500	0.1178	0.2799	4.4284	0.1036	0.3404	4.6850	0.0789	0.3307	4.3392	0.0632	0.4805	5.4233
0.9500	0.2874	0.3178	6.2324	0.2633	0.3623	6.2816	0.2435	0.4536	6.7023	0.2055	0.5015	6.7598
(3/2× optimal bandwidth)												
0.0000	0.1550	0.1716	4.2942	0.1902	0.1801	4.7180	0.2277	0.1705	5.0675	0.2724	0.1609	5.4612
0.2500	0.0208	0.2116	2.5607	0.0401	0.2066	2.8777	0.0453	0.2224	3.0780	0.0646	0.2274	3.4100
0.5000	0.0073	0.2389	2.5382	0.0066	0.2736	2.8547	0.0005	0.2934	2.9420	0.0001	0.3000	3.0015
0.7500	0.1086	0.2795	4.3217	0.0945	0.3379	4.5683	0.0712	0.3280	4.2285	0.0565	0.4750	5.3115
0.9500	0.2693	0.3173	6.0825	0.2458	0.3598	6.1255	0.2268	0.4497	6.5502	0.1906	0.4960	6.6072

Table 3: DGP1 (bivariate normal), $n = 400$, 1000 replications. Proposed estimator with second-order Epanechnikov kernel ($p = 2$). (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(optimal bandwidth)												
0.0000	0.0073	0.0972	2.5875	0.0021	0.0917	2.0524	0.0007	0.0942	1.9547	0.0000	0.0933	1.8683
0.2500	0.0007	0.0924	1.9180	0.0001	0.0966	1.9371	0.0017	0.0916	2.0082	0.0052	0.0920	2.3367
0.5000	0.0011	0.0962	2.0330	0.0057	0.0908	2.3585	0.0112	0.0959	2.8557	0.0199	0.0917	3.3640
0.7500	0.0082	0.0919	2.5766	0.0194	0.0924	3.3416	0.0293	0.0894	3.8594	0.0460	0.0889	4.6434
0.9500	0.0178	0.0922	3.2404	0.0345	0.0878	4.1101	0.0505	0.0840	4.7995	0.0731	0.0845	5.6663
(2/3× optimal bandwidth)												
0.0000	0.0073	0.0972	2.5890	0.0021	0.0918	2.0535	0.0007	0.0942	1.9557	0.0000	0.0934	1.8694
0.2500	0.0007	0.0924	1.9204	0.0001	0.0966	1.9382	0.0017	0.0916	2.0077	0.0052	0.0921	2.3354
0.5000	0.0011	0.0963	2.0315	0.0056	0.0908	2.3546	0.0111	0.0960	2.8519	0.0198	0.0918	3.3594
0.7500	0.0080	0.0920	2.5692	0.0192	0.0924	3.3330	0.0291	0.0894	3.8510	0.0458	0.0891	4.6352
0.9500	0.0175	0.0923	3.2283	0.0342	0.0879	4.0973	0.0502	0.0841	4.7867	0.0727	0.0846	5.6532
(3/2× optimal bandwidth)												
0.0000	0.0073	0.0971	2.5869	0.0021	0.0917	2.0520	0.0007	0.0941	1.9542	0.0000	0.0933	1.8678
0.2500	0.0007	0.0924	1.9170	0.0001	0.0965	1.9367	0.0017	0.0916	2.0084	0.0052	0.0920	2.3372
0.5000	0.0011	0.0962	2.0337	0.0057	0.0908	2.3602	0.0112	0.0959	2.8574	0.0199	0.0917	3.3660
0.7500	0.0082	0.0919	2.5799	0.0195	0.0923	3.3453	0.0293	0.0893	3.8632	0.0461	0.0889	4.6471
0.9500	0.0179	0.0921	3.2458	0.0346	0.0878	4.1158	0.0507	0.0840	4.8052	0.0733	0.0844	5.6721

Table 4: DGP2 (non-normal), $n = 400$, 1000 replications. Proposed estimator with second-order Epanechnikov kernel ($p = 2$). (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(optimal bandwidth)												
0.0000	0.1460	0.0908	7.8539	0.1864	0.0884	8.8133	0.2179	0.0881	9.5014	0.2658	0.0831	10.4445
0.2500	0.0184	0.1045	3.4218	0.0317	0.1053	4.1369	0.0426	0.1153	4.7277	0.0651	0.1158	5.6037
0.5000	0.0095	0.1255	3.1783	0.0048	0.1261	2.8790	0.0025	0.1498	3.1582	0.0004	0.1898	3.8157
0.7500	0.1296	0.1433	7.7487	0.1058	0.1769	7.4044	0.0912	0.1956	7.1956	0.0686	0.2002	6.5943
0.9500	0.2897	0.1551	11.2027	0.2707	0.1828	11.0287	0.2632	0.2245	11.1992	0.2507	0.2913	11.5847
(2/3× optimal bandwidth)												
0.0000	0.1428	0.0913	7.7763	0.1829	0.0891	8.7369	0.2143	0.0888	9.4263	0.2621	0.0837	10.3754
0.2500	0.0163	0.1050	3.3076	0.0289	0.1061	4.0066	0.0392	0.1164	4.5947	0.0611	0.1170	5.4683
0.5000	0.0120	0.1259	3.3398	0.0068	0.1271	3.0284	0.0040	0.1515	3.2815	0.0010	0.1918	3.8896
0.7500	0.1426	0.1434	8.0783	0.1180	0.1784	7.7409	0.1022	0.1968	7.5083	0.0780	0.2022	6.8948
0.9500	0.3146	0.1552	11.6389	0.2946	0.1837	11.4600	0.2865	0.2260	11.6209	0.2723	0.2939	11.9783
(3/2× optimal bandwidth)												
0.0000	0.1473	0.0905	7.8874	0.1879	0.0882	8.8462	0.2195	0.0878	9.5336	0.2674	0.0828	10.4742
0.2500	0.0193	0.1042	3.4714	0.0329	0.1049	4.1929	0.0440	0.1149	4.7847	0.0668	0.1153	5.6615
0.5000	0.0085	0.1254	3.1136	0.0041	0.1256	2.8209	0.0020	0.1491	3.1118	0.0002	0.1890	3.7895
0.7500	0.1243	0.1432	7.6116	0.1009	0.1763	7.2650	0.0868	0.1950	7.0661	0.0649	0.1993	6.4701
0.9500	0.2797	0.1551	11.0222	0.2610	0.1824	10.8502	0.2537	0.2239	11.0246	0.2419	0.2902	11.4212

I next consider the simulated performances over 1000 Monte Carlo replications across DGPs, sample sizes and settings of (ρ, α) of several alternative estimators of the intercept θ_0. The results for samples of size $n = 100$ are summarized below in Tables 5 and 6 for

DGPs 1 and 2, respectively. The corresponding results for samples of size $n = 400$ appear in Tables 7 and 8. The standard Heckman 2-step estimator is found under DGP1 to have a performance in terms of RMSE that is comparable to that of the proposed estimator in (6). The proposed estimator under DGP2, on the other hand, is found to dominate in terms of RMSE the performance of the following alternative estimators under all combinations of (ρ, α) considered:

- (OLS) The ordinary least squares estimator of the intercept parameter using only those observations for which $D = 1$.

  These results are consistent with well established theory. In particular, Tables 5–6 indicate the good performance of OLS when $\rho = 0$ and the poor performance of OLS when $\rho > 0$. In addition, the decrease in $\sqrt{n}\times$ RMSE for the OLS estimator when $\rho = 0$ as one moves from $n = 100$ to $n = 400$ is suggestive of superefficiency, while at the same time the increases in $\sqrt{n}\times$RMSE when $\rho > 0$ is consistent with OLS being inconsistent under $\rho > 0$.

- (2-step) The estimator of the intercept based on the well known procedure of Heckman (1976, 1979), which is known to be $\sqrt{n}$-consistent if $[\ U \quad V\ ]$ is bivariate normal (i.e., generated according to DGP1).

  The results for DGP1 given in Table 5 below are consistent with expectations; in particular, the 2-step procedure exhibits an RMSE that is stable across the various configurations of (ρ, α) that were tried. In addition, a comparison of the relevant sections of Table 5 and Table 7 highlights the stability of $\sqrt{n}\times$ RMSE as one moves from $n = 100$ to $n = 400$, which is consistent with the $\sqrt{n}$-consistency of the procedure under DGP1.

  The results for DGP2 appearing in Table 6, on the other hand, show that the performance of the 2-step procedure can vary dramatically with (ρ, α). A comparison of the $\sqrt{n}\times$RMSE figures in Table 8 with those in Table 6 also suggests that the 2-step procedure under DGP2 is superefficient at $\rho = 0$ and converges at a slower-than-parametric rate for model specifications with $\rho > 0$.

- (H90) The intercept estimator suggested by Heckman (1990), which in the context of the model specified in (1)–(3) has the form

$$\hat{\theta}_{H90} \equiv \frac{\sum_{i=1}^{n} D_i \left(Y_i - \boldsymbol{X}_i^{\top}\hat{\boldsymbol{\beta}}\right) 1\left\{\boldsymbol{Z}_i\hat{\boldsymbol{\gamma}} > b_n\right\}}{\sum_{i=1}^{n} D_i 1\left\{\boldsymbol{Z}_i\hat{\boldsymbol{\gamma}} > b_n\right\}} \tag{19}$$

for some sequence of positive constants $\{b_n\}$ with $b_n \to \infty$ as $n \to \infty$. I present the results of simulations in which the nuisance-parameter estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are fixed at the true values of the corresponding estimands. These results appear below in Tables 5, 6, 7 and 8 for b_n equal to the sample .95-quantile of $\boldsymbol{Z}_i^\top \boldsymbol{\gamma}_0$.

Table 5 below indicates that the performance of $\hat{\theta}_{H90}$ is comparable to that of the 2-step procedure under DGP1 in that its RMSE is stable over changes in (ρ, α). The stability in the $\sqrt{n}\times$RMSE figures in these tables as one moves from $n = 100$ to $n = 400$, which is evident from a comparison of Table 5 with Table 7, also suggests that $\hat{\theta}_{H90}$ may be $\sqrt{n}$-consistent under DGP1.

Table 6, on the other hand, shows that the performance of $\hat{\theta}_{H90}$ can deteriorate dramatically as ρ moves away from zero, although its performance under DGP2 appears to be unaffected by variation in α for any given value of ρ. The $\sqrt{n}\times$RMSE figures in Table 6 and Table 8 indicate that $\hat{\theta}_{H90}$ has a slower-than-parametric rate of convergence under DGP2 that is highly sensitive to variation in ρ but relatively insensitive to variation in α.

- (AS98) The intercept estimator developed by Andrews and Schafgans (1998) as a generalization of the procedure of Heckman (1990). The AS98 estimator in the context of the model given above in (1)–(3) has the form

$$\hat{\theta}_{AS98} \equiv \frac{\sum_{i=1}^{n} D_i \left(Y_i - \boldsymbol{X}_i^\top \hat{\boldsymbol{\beta}}\right) s\left(\boldsymbol{Z}_i^\top \hat{\boldsymbol{\gamma}} - b_n\right)}{\sum_{i=1}^{n} D_i s\left(\boldsymbol{Z}_i^\top \hat{\boldsymbol{\gamma}}_n - b_n\right)}, \tag{20}$$

where, following Andrews and Schafgans (1998, eq. (4.1)), I set

$$s(u) = \begin{cases} 1 - \exp\left(-\frac{u}{\tau - u}\right) & , \quad x \in (0, \tau) \\ 0 & , \quad x \leq 0 \\ 1 & , \quad x \geq \tau \end{cases}. \tag{21}$$

Note that the setting $\tau = 0$ reduces $\hat{\theta}_{AS98}$ to $\hat{\theta}_{H90}$ as given earlier in (19). In addition, the tuning parameter b_n in (20), as it does for $\hat{\theta}_{H90}$ in (19) above, refers to a sequence of positive constants with $b_n \to \infty$ as $n \to \infty$.

I present, in common with other simulations reported here, results for $\hat{\theta}_{AS98}$ in which the nuisance-parameter estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are fixed at the true values of the corresponding estimands. These simulations also involve setting the nuisance parameter τ in (21) to the sample median of $\boldsymbol{Z}_i^\top \boldsymbol{\gamma}_0$ and the smoothing parameter b_n in (20) to the sample .95-quantile of $\boldsymbol{Z}_i^\top \boldsymbol{\gamma}_0$. The corresponding results appear below in Tables 5–6 and also in Tables 7–8.

It is clear from Tables 5–6 below that $\hat{\theta}_{AS}$ is numerically unstable under DGP1 but numerically stable under DGP2. Table 6 also indicates the sensitivity of the performance of $\hat{\theta}_{AS}$ to variation in (ρ, α). A comparison of Table 6 with Table 8 also underscores the slower-than-parametric rate of convergence of $\hat{\theta}_{AS98}$.

Table 5: DGP1 (bivariate normal), $n = 100$, 1000 replications. Alternative estimators. (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(OLS)												
0.0000	0.0000	0.1788	1.7882	0.0000	0.1693	1.6936	0.0000	0.1728	1.7280	0.0001	0.1746	1.7495
0.2500	0.0219	0.1727	2.2745	0.0256	0.1698	2.3326	0.0265	0.1795	2.4239	0.0306	0.1651	2.4053
0.5000	0.0936	0.1656	3.4796	0.1029	0.1593	3.5820	0.1152	0.1605	3.7550	0.1218	0.1565	3.8253
0.7500	0.2162	0.1582	4.9117	0.2378	0.1609	5.1352	0.2559	0.1464	5.2658	0.2712	0.1431	5.4004
0.9500	0.3443	0.1458	6.0461	0.3887	0.1362	6.3815	0.4124	0.1362	6.5651	0.4361	0.1325	6.7356
(Heckman 2-step)												
0.0000	0.0000	0.3007	3.0068	0.0001	0.3299	3.3007	0.0000	0.3503	3.5025	0.0003	0.3886	3.8898
0.2500	0.0011	0.3143	3.1596	0.0023	0.3265	3.2994	0.0009	0.3564	3.5757	0.0016	0.3650	3.6722
0.5000	0.0031	0.2963	3.0150	0.0046	0.3320	3.3892	0.0058	0.3338	3.4248	0.0064	0.3712	3.7973
0.7500	0.0083	0.2969	3.1056	0.0106	0.3148	3.3116	0.0111	0.3270	3.4354	0.0111	0.3509	3.6635
0.9500	0.0171	0.2678	2.9806	0.0169	0.2952	3.2264	0.0131	0.3221	3.4179	0.0207	0.3052	3.3747
(Heckman (1990) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top\boldsymbol{\gamma}_0}(.95)$))												
0.0000	0.0003	0.4509	4.5124	0.0002	0.4524	4.5261	0.0004	0.4557	4.5611	0.0001	0.4689	4.6895
0.2500	0.0004	0.4488	4.4929	0.0006	0.4507	4.5135	0.0003	0.4488	4.4918	0.0005	0.4513	4.5180
0.5000	0.0000	0.4481	4.4815	0.0011	0.4454	4.4661	0.0015	0.4423	4.4403	0.0002	0.4342	4.3443
0.7500	0.0011	0.4516	4.5283	0.0001	0.4372	4.3732	0.0025	0.4379	4.4076	0.0040	0.4315	4.3615
0.9500	0.0030	0.4531	4.5641	0.0020	0.4363	4.3863	0.0030	0.4326	4.3604	0.0060	0.4327	4.3953
(Andrews and Schafgans (1998) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top\boldsymbol{\gamma}_0}(.95)$))												
0.0000	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
0.2500	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf
0.5000	0.0000	0.5650	5.6499	Inf	Inf	Inf	Inf	Inf	Inf	0.0000	0.5579	5.5796
0.7500	0.0019	0.5757	5.7739	Inf	Inf	Inf	0.0025	0.5465	5.4883	Inf	Inf	Inf
0.9500	Inf	Inf	Inf	0.0017	0.5772	5.7865	0.0007	0.5559	5.5656	0.0036	0.5365	5.3982

Table 6: DGP2 (non-normal), $n = 100$, 1000 replications. Alternative estimators. (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(OLS)												
0.0000	0.0001	0.4821	4.8221	0.0000	0.3909	3.9091	0.0014	1.0831	10.8368	0.0005	0.6892	6.8959
0.2500	0.1510	0.3288	5.0909	0.1736	0.4121	5.8603	0.2394	0.7670	9.0977	0.2633	2.1317	21.9257
0.5000	0.5682	0.3763	8.4250	0.7799	0.7149	11.3620	0.8475	0.4896	10.4266	1.1118	1.1903	15.9015
0.7500	1.2924	0.2869	11.7248	1.7087	0.4380	13.7860	1.8913	1.8977	23.4365	2.3993	1.4258	21.0529
0.9500	2.1300	0.3085	14.9169	2.6775	0.5903	17.3954	3.2066	1.2800	22.0111	3.6936	4.8277	51.9616
(Heckman 2-step)												
0.0000	0.0000	0.5162	5.1620	0.0002	0.5989	5.9913	0.0020	1.6850	16.8557	0.0055	1.1861	11.8844
0.2500	0.1849	0.5086	6.6596	0.2433	0.6251	7.9627	0.3461	0.7657	9.6563	0.4278	1.4736	16.1227
0.5000	0.7718	0.6221	10.7649	1.2359	1.9152	22.1445	1.3865	0.9240	14.9673	2.3721	2.9882	33.6172
0.7500	1.6186	0.5040	13.6844	1.0515	16.7408	167.7214	3.2286	2.1692	28.1672	4.6267	2.7391	34.8276
0.9500	2.7113	0.5127	17.2457	3.8929	0.9294	21.8101	5.4307	3.6028	42.9077	8.1794	4.7219	55.2045
(Heckman (1990) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top \boldsymbol{\gamma}_0}(.95)$))												
0.0000	0.0001	0.4388	4.3889	0.0001	0.4741	4.7418	0.0004	0.4621	4.6254	0.0000	0.4743	4.7428
0.2500	0.2121	0.4677	6.5640	0.3145	0.5099	7.5793	0.4265	0.6050	8.9023	0.5743	0.6945	10.2790
0.5000	0.8606	0.4819	10.4540	1.3574	0.8578	14.4677	1.5614	0.8493	15.1087	2.1035	0.9589	17.3869
0.7500	1.8488	0.5264	14.5806	2.9232	1.0078	19.8465	3.4719	0.9811	21.0580	5.0749	1.8515	29.1596
0.9500	3.0237	0.7279	18.8508	4.3979	1.0647	23.5192	5.6603	1.6073	28.7121	7.8382	1.9172	33.9323
(Andrews and Schafgans (1998) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top \boldsymbol{\gamma}_0}(.95)$))												
0.0000	0.0000	0.9234	9.2336	0.0015	1.0125	10.1326	0.0017	1.0037	10.0452	0.0001	0.9807	9.8079
0.2500	0.2035	1.0138	11.0963	0.3754	1.1914	13.3967	0.6937	1.8209	20.0236	1.0490	2.1058	23.4170
0.5000	0.8978	1.0810	14.3748	2.0336	2.9706	32.9515	2.3156	3.0879	34.4245	3.2494	2.6734	32.2432
0.7500	1.8968	1.1576	17.9912	Inf	Inf	Inf	4.9692	2.9053	36.6200	10.1655	7.9701	85.8419
0.9500	3.3558	2.1412	28.1792	5.0293	2.8442	36.2197	7.8162	3.3294	43.4752	13.5506	7.1477	80.3995

Table 7: DGP1 (bivariate normal), $n = 400$, 1000 replications. Alternative estimators. (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(OLS)												
0.0000	0.0000	0.0866	1.7318	0.0000	0.0809	1.6194	0.0000	0.0833	1.6676	0.0000	0.0756	1.5130
0.2500	0.0232	0.0840	3.4805	0.0258	0.0823	3.6111	0.0289	0.0829	3.7818	0.0295	0.0779	3.7712
0.5000	0.0935	0.0805	6.3227	0.1071	0.0772	6.7232	0.1149	0.0771	6.9534	0.1211	0.0750	7.1190
0.7500	0.2163	0.0758	9.4248	0.2390	0.0712	9.8808	0.2530	0.0712	10.1604	0.2719	0.0684	10.5179
0.9500	0.3417	0.0695	11.7732	0.3830	0.0664	12.4488	0.4087	0.0634	12.8487	0.4366	0.0624	13.2739
(Heckman 2-step)												
0.0000	0.0000	0.1552	3.1042	0.0000	0.1640	3.2810	0.0000	0.1784	3.5672	0.0000	0.1832	3.6645
0.2500	0.0000	0.1522	3.0456	0.0001	0.1591	3.1876	0.0001	0.1720	3.4472	0.0001	0.1797	3.6009
0.5000	0.0002	0.1525	3.0627	0.0005	0.1646	3.3198	0.0003	0.1709	3.4357	0.0002	0.1770	3.5502
0.7500	0.0009	0.1435	2.9301	0.0006	0.1559	3.1585	0.0003	0.1641	3.2977	0.0011	0.1708	3.4785
0.9500	0.0007	0.1368	2.7857	0.0008	0.1468	2.9915	0.0012	0.1558	3.1939	0.0010	0.1614	3.2917
(Heckman (1990) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top \boldsymbol{\gamma}_0}(.95)$))												
0.0000	0.0000	0.1623	3.2457	0.0000	0.1627	3.2548	0.0000	0.1653	3.3058	0.0000	0.1608	3.2164
0.2500	0.0001	0.1585	3.1769	0.0001	0.1594	3.1950	0.0002	0.1566	3.1470	0.0005	0.1557	3.1449
0.5000	0.0005	0.1582	3.1931	0.0010	0.1556	3.1755	0.0016	0.1647	3.3916	0.0028	0.1696	3.5550
0.7500	0.0007	0.1569	3.1826	0.0016	0.1607	3.3140	0.0032	0.1557	3.3125	0.0070	0.1502	3.4401
0.9500	0.0009	0.1539	3.1347	0.0027	0.1458	3.0967	0.0052	0.1487	3.3033	0.0098	0.1515	3.6168
(Andrews and Schafgans (1998) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top \boldsymbol{\gamma}_0}(.95)$))												
0.0000	Inf	Inf	Inf	0.0000	0.3118	6.2369	Inf	Inf	Inf	0.0000	0.2991	5.9817
0.2500	0.0000	0.3043	6.0860	0.0000	0.3056	6.1114	Inf	Inf	Inf	Inf	Inf	Inf
0.5000	Inf	Inf	Inf	0.0001	0.2973	5.9497	Inf	Inf	Inf	0.0004	0.3065	6.1438
0.7500	0.0001	0.2888	5.7792	0.0000	0.3041	6.0817	Inf	Inf	Inf	Inf	Inf	Inf
0.9500	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	Inf	0.0015	0.3021	6.0908

Table 8: DGP2 (non-normal), $n = 400$, 1000 replications. Alternative estimators. (RMSE is multiplied by $\sqrt{n}$.)

ρ	$\alpha = 2.00$			$\alpha = 1.50$			$\alpha = 1.25$			$\alpha = 1.00$		
	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE	sq bias	sd	RMSE
(OLS)												
0.0000	0.0000	0.1225	2.4496	0.0000	0.1295	2.5902	0.0000	0.1356	2.7151	0.0000	0.1423	2.8469
0.2500	0.1464	0.1253	8.0533	0.1855	0.1377	9.0428	0.2246	0.1549	9.9707	0.2653	0.1651	10.8174
0.5000	0.5746	0.1254	15.3660	0.7346	0.1398	17.3680	0.8895	0.2137	19.3406	1.0931	0.2617	21.5551
0.7500	1.3338	0.1182	23.2184	1.6652	0.1752	26.0457	1.9547	0.2239	28.3186	2.3994	0.2910	31.5215
0.9500	2.1328	0.1128	29.2955	2.6396	0.1762	32.6839	3.1089	0.2503	35.6178	4.0168	0.4762	41.1997
(Heckman 2-step)												
0.0000	0.0000	0.1900	3.8004	0.0000	0.2107	4.2133	0.0002	0.2368	4.7430	0.0000	0.2641	5.2828
0.2500	0.1966	0.1881	9.6325	0.2873	0.2342	11.6982	0.4031	0.3386	14.3913	0.6158	0.4610	18.2023
0.5000	0.7901	0.1930	18.1916	1.1622	0.2667	22.2114	1.6124	0.5121	27.3834	2.4130	0.6840	33.9460
0.7500	1.8144	0.2124	27.2728	2.6411	0.4291	33.6168	3.4999	0.5283	38.8794	5.4115	0.8450	49.4998
0.9500	2.8707	0.2172	34.1632	4.1789	0.4178	41.7300	5.6244	0.6344	49.0992	9.1834	1.2123	65.2785
(Heckman (1990) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top \boldsymbol{\gamma}_0}(.95)$))												
0.0000	0.0000	0.1625	3.2499	0.0000	0.1668	3.3373	0.0001	0.1698	3.4017	0.0000	0.1703	3.4069
0.2500	0.1881	0.1647	9.2789	0.2496	0.1838	10.6470	0.3165	0.2183	12.0686	0.4023	0.2231	13.4475
0.5000	0.7417	0.1732	17.5695	0.9928	0.2013	20.3307	1.2631	0.3301	23.4275	1.6212	0.4082	26.7415
0.7500	1.7112	0.1877	26.4303	2.2915	0.3216	30.9514	2.7843	0.3655	34.1634	3.5582	0.4415	38.7456
0.9500	2.7126	0.1885	33.1547	3.6019	0.2978	38.4220	4.4380	0.4261	42.9862	6.0276	0.7106	51.1176
(Andrews and Schafgans (1998) ($b_n = \hat{F}^{-1}_{\mathbf{Z}^\top \boldsymbol{\gamma}_0}(.95)$))												
0.0000	0.0000	0.4867	9.7345	0.0001	0.5337	10.6766	0.0001	0.5335	10.6730	0.0004	0.5566	11.1378
0.2500	0.2272	0.5277	14.2224	0.4256	0.7807	20.3484	0.7890	1.5216	35.2385	1.1932	1.4466	36.2534
0.5000	0.9464	0.7039	24.0155	1.6530	1.0790	33.5688	2.9148	2.8347	66.1824	5.3105	3.1842	78.6128
0.7500	2.2403	0.9842	35.8267	3.9629	2.3050	60.9133	6.0330	2.7449	73.6684	10.2806	3.3502	92.7458
0.9500	3.5332	0.9977	42.5610	5.7880	1.8579	60.7943	9.2207	3.1294	87.2098	21.6148	5.9786	151.4702

In summary, the simulations presented here show the potential of the proposed estimator to exhibit good performance in terms of RMSE across two different parametric families of data-generating process, across variation in the degree to which the errors U and V are dependent and across variation in the extent to which the parameter of interest is identified. This assessment is unaffected by moderate variation in the estimated MSE-optimal bandwidth used to implement the proposed estimator. Tables 1 and 2 also support the conclusion of Theorem 1 in indicating the sensitivity of the bias of the proposed estimator to variation in the parameter (ρ, α) under both DGP1 and DGP2. These results also show that the estimated MSE-optimal bandwidth used to implement the proposed estimator was effective in limiting the extent to which the RMSE of the proposed estimator was sensitive to variation in (ρ, α). In particular, the RMSE of the proposed estimator was found under DGP2 to dominate those of the other estimators considered.

5 Empirical Example

This section reconsiders individual labour-market data from Malaysia that were originally analyzed by Schafgans (2000). The estimator developed above is applied to the problem of estimating the extent of plausible gender wage discrimination in Malaysia using data from the Second Malaysian Family Life Survey (MFLS2) conducted between August 1988 and January 1989. Inferences available from the proposed estimator are compared with those obtained via the same alternative estimators considered in Section 4. The proposed estimator was found to generate inferences that differ significantly from those obtainable via the alternative estimators used in the simulation experiments described in Section 4. All estimators applied to the MFLS2 data considered in this section were implemented in precisely the same way in which they were implemented in the simulation experiments presented earlier.

I consider a decomposition of the female–male log-wage difference for ethnic Malay workers. The consideration of gender wage gaps for Malaysian workers of the same ethnicity is potentially important because of the differential treatment of Malays in the labour force after 1970 (see e.g., Schafgans, 2000, and references cited). I follow Schafgans (2000) by analyzing gender wage gaps in the MFLS2 using the basic decomposition technique of Oaxaca (1973). In particular, suppose that the generic model given above in (1)–(3) holds for both men and women, i.e.,

$$
\begin{aligned}
Y_j^* &= \theta_j + \boldsymbol{X}_j^\top \boldsymbol{\beta}_j + U_j, && (22)\\
D_j &= 1\left\{\boldsymbol{Z}_j^\top \boldsymbol{\gamma}_j \geq V_j\right\}, && (23)\\
Y_j &= DY_j^* && (24)
\end{aligned}
$$

where Y_j^* is the natural logarithm of the offered average hourly wage, and where the index $j \in \{0, 1\}$ denotes a given gender. For $j \in \{0, 1\}$ let $\bar{Y}_j \equiv E\left[Y_j \middle| D_j = 1\right]$, and let $\bar{\boldsymbol{X}}_j$ denote the average "endowments" of wage-determining attributes for workers of gender j. The observed log-wage gap $\bar{Y}_1 - \bar{Y}_0$ between the two genders can then be decomposed as

$$
\begin{aligned}
\bar{Y}_1 - \bar{Y}_0 &= \left[(\theta_1 - \theta_0) + \bar{\boldsymbol{X}}_0^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\right] + (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_0)^\top \boldsymbol{\beta}_1 \\
&\quad + (E[U_1 | D_1 = 1] - E[U_0 | D_0 = 1]) && (25)\\
&= \left[(\theta_1 - \theta_0) + \bar{\boldsymbol{X}}_1^\top (\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)\right] + (\bar{\boldsymbol{X}}_1 - \bar{\boldsymbol{X}}_0)^\top \boldsymbol{\beta}_0 \\
&\quad + (E[U_1 | D_1 = 1] - E[U_0 | D_0 = 1]) && (26)\\
&\equiv A + B + C, && (27)
\end{aligned}
$$

where A is that part of the gap due to differences in wage structures between genders; B is due to observable differences between men and women in wage-determining characteristics and C is the contribution of differential self-selection into the labour force. Following Schafgans (2000) the quantity $\bar{Y}_1 - \bar{Y}_0 - C = A + B$ is referred to as the *selection-corrected log-wage gap*.

Wage discrimination in favor of members of gender $j = 1$ is empirically plausible if the overall log-wage gap $\bar{Y}_1 - \bar{Y}_0$ cannot be entirely explained by differential self-selection into paid work, differences in observed endowments or by differing returns to those endowments. Moreover, given the definitions of the quantities A and B appearing above in (27), the extent of plausible wage discrimination favoring gender 1 may be equated with the difference in intercepts $\theta_1 - \theta_0$.

The analysis that follows considers a subset of the sample taken from the MFLS2 of 1988–89 that was analyzed by Schafgans (2000). This particular dataset is publicly available from the *Journal of Applied Econometrics* Data Archive at `http://qed.econ.queensu.ca/jae/1998-v13.5/schafgans/`. Each observation in this sample corresponds to a member of the labour force. I specifically consider ethnic Malays residing in non-urban settings who were observed to have some level of unearned household income in terms of dividends, interests or rents, and who were also observed to have passed the highest level of schooling (i.e., primary on the one hand, or secondary or above) corresponding to the number of years of schooling observed. This subset of the MFLS2 consisted of 965 women and 878 men.

I also use the same variable specifications used by Schafgans (2000). In particular, the outcome variable Y_j^* is LWAGE, the log hourly real wage in the local currency deflated using the 1985 consumer price index. The selection variable D_j is the indicator PAIDWORK for whether the individual in question is in fact a wage worker. The exogenous variables appearing in the selection equations for each gender include UNEARN, a measure of household unearned income in terms of dividends, interest and rents; HOUSEH, the value of household real estate owned, computed as the product of an indicator variable for house ownership and the cost of the house owned; and AMTLAND, the extent of household landholding in hundreds of acres. In addition, selection into wage work is also assumed to be determined by AGE, in years; AGESQ, the square of AGE divided by 100; YPRIM, years of primary schooling and YSEC, years of schooling at the secondary level or above. The variables appearing on the right-hand side of the outcome equations for each gender or ethnic group are AGE, AGESQ, YPRIM and YSEC. Schafgans (2000, Section 4) contains further details regarding variable definitions.

For each gender $j \in \{\text{female, male}\}$, the proposed estimator and that of Heckman (1990) and Andrews and Schafgans (1998) rely on the preliminary procedure described in Schafgans (1998, p. 484–487) to estimate the nuisance parameters $\boldsymbol{\beta}_j$ and $\boldsymbol{\gamma}_j$ appearing in (22) and (23), respectively. This involves estimating the selection equation for each group via the method of Klein and Spady (1993) and estimating the slope parameters in each outcome equation using the method of Robinson (1988). This is followed by estimation of the intercept parameter in each outcome equation via the proposed estimator. Standard errors are calculated by bootstrapping with replacement with $B = 200$ replications.

Estimates of the outcome-equation parameters obtained via the proposed estimator are given in which the proposed estimator is implemented using the same kernel and estimated MSE-optimal bandwidth $\hat{h}_n^*$ used in the simulations reported in Section 4. In common with the results given earlier in Section 4, I also considered implementations of the proposed estimator in which the bandwidth was set to $h_n = (2/3)\hat{h}_n^*$ and $h_n = (3/2)\hat{h}_n^*$.

The decomposition of the observed gender log-wage gaps for ethnic Malay workers is presented in Table 9. In keeping with the theory developed above, the focus is on the extent of plausible gender wage discrimination, which is identified with the difference between the estimated intercepts. A striking result is the evidence provided by the proposed estimator of positive wage discrimination in favor of women. In particular, Table 9 indicates that all three implementations of the proposed estimator imply a large, positive and significant difference in intercepts, while the OLS and 2-step procedures generated estimated intercept differences that were both insignificant. The implementation of the H90 procedure with b_n set to the .90-quantile of the estimated selection index generated a similarly insignificant estimate of the difference in intercepts. The other implementation of the H90 procedure, along with the AS98 procedure, proved to be numerically unstable. Table 9 indicates that there exists a clear difference in inferences regarding the extent of gender wage discrimination amongst ethnic Malay workers between estimates generated by the proposed estimator and those generated by established procedures.

Table 9: Female–male log-wage decomposition, Malays. Standard errors in parentheses

Wage gap (overall)	-0.2882 (0.0425)							
Female (endowment)	-0.0638 (0.0337)							
Male (endowment)	-0.0369 (0.0377)							
	$\hat{\theta}_n$			OLS	2-step	H90		AS98
	$(h_n = \hat{h}_n^*)$	$(h_n = (2/3)\hat{h}_n^*)$	$(h_n = (3/2)\hat{h}_n^*)$			$(b_n = \hat{F}^{-1}_{\mathbf{Z}^\top\hat{\boldsymbol{\gamma}}_n}(.90))$	$(b_n = \hat{F}^{-1}_{\mathbf{Z}^\top\hat{\boldsymbol{\gamma}}_n}(.95))$	$(b_n = \hat{F}^{-1}_{\mathbf{Z}^\top\hat{\boldsymbol{\gamma}}_n}(.95))$
Wage gap (selection-corrected)	0.9283 (0.7516)	0.9328 (0.7515)	0.9263 (0.7517)	-1.337 (0.0812)	-10.1628 (1.0385)	-0.2709 (1.0617)	-0.0516 (NaN)	-0.1015 (NaN)
Female (coefficients)	0.1098 (0.6483)			-1.2553 (0.6977)	-8.0955 (2.2356)	0.1098 (0.6483)		0.1098 (0.6483)
Male (coefficients)	0.0829 (0.6457)			-1.2951 (0.6989)	-7.988 (2.2296)	0.0829 (0.6457)		0.0829 (0.6457)
Difference in intercepts	0.8823 (0.3787)	0.8867 (0.3784)	0.8803 (0.3789)	-0.0364 (0.7526)	-2.1082 (3.1712)	-0.3169 (0.8401)	-0.0976 (NaN)	-0.1475 (NaN)

6 Conclusion

This paper has developed a new estimator of the intercept of a sample-selection model in which the joint distribution of the unobservables and the selection index is unspecified. It has been shown that the new estimator can be made under mild conditions to converge in probability at an $n^{-p/(2p+1)}$-rate, where $p \geq 2$ is an integer that indexes the strength of certain smoothness assumptions as given above in Assumption 2.4. This rate of convergence is shown to be the optimal rate of convergence for estimation of the intercept parameter in terms of a minimax criterion. The new estimator is under mild conditions consistent and asymptotically normal with a rate of convergence that is the same regardless of the joint distribution of the unobservables and the selection index. This differs from other proposals in the literature and is convenient in practice, as the extent to which selection is endogenous is typically unknown in applications. In addition, the rate of convergence of the new estimator, unlike those of better known estimators, does not depend on assumptions regarding the relative tail behaviours of the determinants of selection beyond those necessary for the identification of the estimand. This similarly facilitates statistical inference regarding the intercept. Simulations presented above show the potential accuracy of the proposed estimator relative to that of established procedures across different model specifications. An empirical example using individual labour-market data from Malaysia shows the potential of the proposed estimator to generate inferences regarding the extent of plausible gender wage discrimination that differ from those available from better known estimators.

A Appendix

A.1 Further discussion of the finiteness of $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1}$ for any $u \in \mathbb{R}$

This appendix supplies details regarding the assertion made earlier that the conditional density $r_{U|Q}(u|q)$ given in (12) satisfies $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1} < \infty$ for any $u \in \mathbb{R}$. Recall in this connection that the finiteness of $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1}$ implies in turn the previously stated differentiability condition regarding the conditional mean function $m_{F_0}(q)$ given in (11). In particular, the differentiability of $m_{F_0}(q)$ on $(0,1)$ to pth order, where $p \geq 2$ is a constant specified in Assumption 2.4, along with the left-continuity of the pth derivative of $m_{F_0}(q)$ at $q = 1$, is sufficient to control the asymptotic bias of the proposed estimator $\hat{\theta}_n$; see Appendix A.2 below for details.

The finiteness of $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1}$ is a consequence firstly of the fact, developed in Lemma 1 in what follows, that identification of $\boldsymbol{\gamma}_0$ subject to Assumption 1.2 implies that $r_{U|Q}(u|q)$ is the conditional density of U given $F_0(V) = q$ for any $q \in [0, 1]$:

Lemma 1. *Identification of $\boldsymbol{\gamma}_0$ subject to the conditions of Assumption 1.2 implies that the random variable $F_0(V)$ satisfies the following:*

1. *The distribution of $F_0(V)$ has support equal to $[0, 1]$;*
2. *the conditional distribution of U given $F_0(V) = q$ for any $q \in [0, 1]$ is absolutely continuous with density given by $r_{U|Q}(u|q)$ in (12).*

Proof. 1. Begin by observing that $\boldsymbol{\gamma}_0$ is identified up to the location and scale normalization specified in Assumption 1.2 iff the mapping $\tilde{\boldsymbol{\gamma}} \to P\left[V \le z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}} \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right]$ is 1–1 on $\mathbb{R}^{l-1}$ for any $\boldsymbol{x}$ in the support of $\boldsymbol{X}$ and any $\boldsymbol{z} = \begin{bmatrix} z_1 & \tilde{\boldsymbol{z}}^\top \end{bmatrix}^\top$ in the support $Supp[\boldsymbol{Z}]$ of $\boldsymbol{Z}$. We have

$$P\left[V \le z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}} \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right] = P\left[F_0(V) \le F_0\left(z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}}\right) \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right]$$

since F_0 is a distribution function, so identification of $\boldsymbol{\gamma}_0$ subject to the conditions of Assumption 1.2 holds iff the mapping $\tilde{\boldsymbol{\gamma}} \to P\left[F_0(V) \le F_0\left(z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}}\right) \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right]$ is 1–1 on $\mathbb{R}^{l-1}$ for any $\begin{bmatrix} \boldsymbol{x}^\top & \boldsymbol{z}^\top \end{bmatrix}$ in the support of $\boldsymbol{X}$ and $\boldsymbol{Z}$.

Let the support of the conditional distribution of V given $\begin{bmatrix} \boldsymbol{X}^\top & \boldsymbol{Z}^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}^\top & \boldsymbol{z}^\top \end{bmatrix}$ be given by the interval $[v_1, v_2]$ for constants $-\infty \le v_1 < v_2 \le \infty$. Suppose that $F_0(v_2) < 1$. Then writing $\boldsymbol{z} = \begin{bmatrix} z_1 & \tilde{\boldsymbol{z}}^\top \end{bmatrix}^\top$, there exists a $\tilde{\boldsymbol{\gamma}}' \in \mathbb{R}^{l-1}$ with $\tilde{\boldsymbol{\gamma}}' \ne \boldsymbol{0}$ such that $F_0\left(z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}}'\right) \ge F_0(v_2)$, which implies that

$$\begin{aligned} & P\left[F_0(V) \le F\left(z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}}'\right) \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right] \\ = \; & P\left[F_0(V) \le F_0\left(z_1 + \tilde{\boldsymbol{z}}^\top (2\tilde{\boldsymbol{\gamma}}')\right) \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right] \\ = \; & 1, \end{aligned}$$

from which it follows that $\boldsymbol{\gamma}_0$ is not identified. A failure of identification accordingly ensues when $F_0(v_2) < 1$.

Similarly, if $F_0(v_1) > 0$ we have that $F_0\left(z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}}''\right) < F_0(v_1)$ for some $\tilde{\boldsymbol{\gamma}}'' \in \mathbb{R}^{l-1}$ with $\tilde{\boldsymbol{\gamma}}'' \ne \boldsymbol{0}$, so that

$$\begin{aligned} & P\left[F_0(V) \le F_0\left(z_1 + \tilde{\boldsymbol{z}}^\top \tilde{\boldsymbol{\gamma}}''\right) \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right] \\ = \; & P\left[F_0(V) \le F_0\left(z_1 + \tilde{\boldsymbol{z}}^\top (.5\tilde{\boldsymbol{\gamma}}'')\right) \middle| \boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Z} = \boldsymbol{z} \right] \\ = \; & 0. \end{aligned}$$

This implies a similar failure of identification when $F_0(v_1) > 0$.

It follows that identification of $\boldsymbol{\gamma}_0$ subject to Assumption 1.2 implies that the support of the conditional distribution of $F_0(V)$ given $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ] = [\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$, for any $[\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ in the support of $\boldsymbol{X}$ and $\boldsymbol{Z}$, is $[0, 1]$. The support of the conditional distribution given $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ] = [\ \boldsymbol{x}^\top \quad \boldsymbol{z}^\top\ ]$ coincides with that of the marginal distribution.

2. Part 1 of this lemma shows that the distribution of $F_0(V)$ has support equal to $[0, 1]$. Assumption 1.2e implies that the mapping $v \longrightarrow F_0(v)$ is 1–1 and strictly monotone on the support $Supp[V]$ of V. A standard argument accordingly shows that the density of $Q \equiv F_0(V)$ is given by

$$r_Q(q) \equiv \frac{g_V\left(F_0^{-1}(q)\right)}{f_0\left(F_0^{-1}(q)\right)}, \tag{28}$$

where $g_V(\cdot)$ denotes the marginal density of V. Similarly, the joint density of $[\ U \quad Q\ ]$ is given by

$$r_{UQ}(u, q) \equiv \frac{g_{UV}\left(u, F_0^{-1}(q)\right)}{f_0\left(F_0^{-1}(q)\right)}, \tag{29}$$

where $g_{UV}(\cdot, \cdot)$ denotes the joint density of U and V.

That the conditional density of U given $F_0(V) = q$ for any $q \in [0, 1]$ has the desired form is immediate.

□

It should be noted that the marginal distribution of the disturbance term V in the selection equation is restricted to have a right tail that is related to that of the selection index $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ via the finiteness for all $q \in [0, 1]$ of the marginal density $r_Q(q)$ in (28). For example, in the case where V and $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ are both normally distributed with the scale normalization $Var[V] = 1$, the condition $r_Q(1) < \infty$ implies that the variance of V is no greater than that of $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$, i.e., $r_Q(1) < \infty$ in this case implies that $Var\left[\boldsymbol{Z}^\top\boldsymbol{\gamma}_0\right] \geq 1$. More generally, the condition $r_Q(1) < \infty$ rules out situations where the marginal distribution of V has an upper tail that is strictly heavier than that of $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$.

This restriction on the relative upper-tail behaviours of the distributions of V and $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$ is weaker than the restrictions on the joint distribution of $[\ V \quad \boldsymbol{Z}^\top\boldsymbol{\gamma}_0\ ]$ that feature in e.g., Andrews and Schafgans (1998) or Lewbel (2007). In particular, Assumption 1 does not imply restrictions on the relative upper tail thicknesses of the distributions of V and the selection index $\mathbf{Z}^\top \boldsymbol{\gamma}_0$ that are beyond those necessary for the identification of $\boldsymbol{\gamma}_0$.

Lemma 1 and the differentiability conditions of Assumption 2.4 imply the desired finiteness of $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1}$:

Lemma 2. *The conditional density $r_{U|Q}(u|q)$ in (12) satisfies $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=0} < \infty$ and $(\partial^p/\partial q^p)\, r_{U|Q}(u|q)\big|_{q=1} < \infty$ under the conditions of Assumptions 1 and 2.4 for any $u \in \mathbb{R}$.*

Proof. Recall from the proof of Lemma 1 that $r_{U|Q}(u|q) = r_{UQ}(u,q)/r_Q(q)$, where $r_Q(q)$ and $r_{UQ}(u,q)$ are as given above in (28) and (29), respectively. We have for any $u \in \mathbb{R}$ that $r_{UQ}(u,\cdot)$ is right- and left-continuous at $q = 0$ and $q = 1$, respectively, given the conclusion of Lemma 1 that the distribution of $F_0(V)$ is absolutely continuous with support $[0,1]$. Similarly, the marginal density $r_Q(q)$ is bounded away from zero for all $q \in [0,1]$ by virtue of the conclusion of Lemma 1 that Q has support $[0,1]$. It follows that for any $u \in \mathbb{R}$, the conditional density $r_{U|Q}(u|q)$ is right- and left-continuous as a function of q at $q = 0$ and $q = 1$, respectively.

Next, observe that for any $u \in \mathbb{R}$, the conditional density $r_{U|Q}(u|q)$ is $(p+1)$-times differentiable in q on $(0,1)$ by virtue of Assumption 2.4. The desired conclusion is a special case of the following argument. Let $\phi(\cdot)$ denote a differentiable function on $(0,1)$ so that $\sup_{x\in(0,1)} |\phi'(x)| < \infty$. Assume that $\phi(\cdot)$ is right- and left-continuous at 0 and 1, respectively. Then

$$
\begin{aligned}
& \phi(1) - \phi(0) && (30)\\
= \; & \phi(1-) - \phi(0+) && (31)\\
= \; & \int_0^1 \phi'(x)dx \\
\leq \; & \sup_{x\in(0,1)} |\phi'(x)| \\
< \; & \infty,
\end{aligned}
$$

where (30)–(31) follows by the right- and left-continuity of $\phi(\cdot)$ at 0 and 1, respectively. □

A.2 Proof of Theorem 1

Begin by recalling the definition of $\hat{\eta}_n(\cdot)$ given above in (4). Define in addition

$$\hat{\eta}_0(z) \equiv \frac{1}{n}\sum_{i=1}^{n} 1\left\{(\boldsymbol{Z}_i - z)^\top \boldsymbol{\gamma}_0 \leq 0\right\}. \tag{32}$$

Next, let $F_n(\cdot)$ denote the cdf of $\boldsymbol{Z}^\top\hat{\boldsymbol{\gamma}}_n$, and define

$$\eta_n(z) \equiv F_n\left(z^\top \hat{\boldsymbol{\gamma}}_n\right) \tag{33}$$

and

$$\eta_0(z) \equiv F_0\left(z^\top \boldsymbol{\gamma}_0\right). \tag{34}$$

Consider the following preliminary result that will be used repeatedly in the sequel:

Lemma 3. *Under the conditions of Assumptions 1 and 2,*

$$\sup_{z\in\mathbb{R}^l}\left|\sqrt{n}\left[(\hat{\eta}_n(z)-\eta_n(z))-(\hat{\eta}_0(z)-\eta_0(z))\right]\right| = o_p(1).$$

Proof. Lemma 3 involves an application of van der Vaart and Wellner (2007, Theorem 2.1). In particular, let $\delta \in \mathbb{R}$, $\tilde{\boldsymbol{\gamma}} \in \mathbb{R}^{l-1}$ and $\tilde{z} \in \mathbb{R}^{l-1}$ be fixed, and define the function $g_{\delta,\tilde{\gamma},\tilde{z}} : \mathbb{R}^{l-1} \to \mathbb{R}$ as $g_{\delta,\tilde{\gamma},\tilde{z}}(\boldsymbol{w}) = \delta + (\boldsymbol{w} - \tilde{z})^\top \tilde{\boldsymbol{\gamma}}$. Consider the corresponding function class $\mathcal{G} \equiv \left\{g_{\delta,\tilde{\gamma},\tilde{z}} : \delta \in \mathbb{R}, \tilde{\boldsymbol{\gamma}} \in \mathbb{R}^{l-1}, \tilde{z} \in \mathbb{R}^{l-1}\right\}$. Observe that $\mathcal{G}$ is contained in a finite-dimensional vector space. To see this, note that for an arbitrary non-zero constant $\lambda \in \mathbb{R}$, $\lambda g_{\delta,\tilde{\gamma},\tilde{z}}(\boldsymbol{w}) = g_{\lambda\delta,\lambda\tilde{\gamma},\tilde{z}}(\boldsymbol{w})$, while for fixed $[\ \delta_1 \quad \tilde{\boldsymbol{\gamma}}_1^\top \quad \tilde{z}_1^\top\ ]$, $[\ \delta_2 \quad \tilde{\boldsymbol{\gamma}}_2^\top \quad \tilde{z}_2^\top\ ] \in \mathbb{R}^{2l-1}$,

$$\begin{aligned} & g_{\delta_1,\tilde{\gamma}_1,\tilde{z}_1}(\boldsymbol{w}) + g_{\delta_2,\tilde{\gamma}_2,\tilde{z}_2}(\boldsymbol{w}) \\ = & \left(\delta_1 + \delta_2 + \tilde{z}_1^\top \tilde{\boldsymbol{\gamma}}_2 + \tilde{z}_2^\top \tilde{\boldsymbol{\gamma}}_1\right) + \left[\boldsymbol{w} - (\tilde{z}_1 + \tilde{z}_2)\right]^\top (\tilde{\boldsymbol{\gamma}}_1 + \tilde{\boldsymbol{\gamma}}_2) \\ = & g_{\delta_1+\delta_2+\tilde{z}_1^\top\tilde{\gamma}_2+\tilde{z}_2^\top\tilde{\gamma}_1,\tilde{\gamma}_1+\tilde{\gamma}_2,\tilde{z}_1+\tilde{z}_2}(\boldsymbol{w}). \end{aligned}$$

It follows that $\mathcal{G}$ is a VC-class, which implies that its negativity sets also constitute a VC-class (e.g., van der Vaart and Wellner, 1996, Lemma 2.6.18).

As such, it follows that the class of indicator functions of $\left\{\boldsymbol{w} \in \mathbb{R}^l : g_{\delta,\tilde{\gamma},\tilde{z}}(\boldsymbol{w}) \leq 0\right\}$, indexed by $[\ \delta \quad \tilde{\boldsymbol{\gamma}}^\top \quad \tilde{z}^\top\ ] \in \mathbb{R}^{2l-1}$, is a Donsker class. Lemma 3 follows immediately from van der Vaart and Wellner (2007, Theorem 2.1). $\square$

Recall the definition of $\hat{W}_i$ given above in (5) and define $W_i \equiv D_i \left(Y_i - \boldsymbol{X}_i^\top \boldsymbol{\beta}_0\right)$ for each $i \in \{1, \ldots, n\}$. Recall in addition the definitions of $\boldsymbol{S}_i$ and K_i given above in (7) and (8), respectively. The proposed estimator given above in (6) may be rewritten as

$$
\begin{aligned}
&\hat{\theta}_n \\
= \; & \boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i W_i \\
& - \boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{X}_i^\top \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) \\
\equiv \; & \hat{m}_{n1}(1) + \hat{m}_{n2}(1).
\end{aligned} \tag{35}
$$

Consider $\hat{m}_{n1}(1)$. For each $i \in \{1, \ldots, n\}$ we can write

$$
\begin{aligned}
W_i &= m_{F_0}\left(\hat{\eta}_n(\mathbf{Z}_i)\right) + \left(W_i - m_{F_0}\left(\hat{\eta}_n(\mathbf{Z}_i)\right)\right) \\
&\equiv m_{F_0}\left(\hat{\eta}_n(\mathbf{Z}_i)\right) + \zeta_{ni} \\
&= m_{F_0}(1) + \sum_{j=1}^{p-1} \frac{1}{j!} \left(\hat{\eta}_n(\mathbf{Z}_i) - 1\right)^j m_{F_0}^{(j)}(1) \\
&\quad + \frac{1}{p!} \left(\hat{\eta}_n(\mathbf{Z}_i) - 1\right)^p \cdot m_{F_0}^{(p)}\left(\hat{\eta}_{ni}^*\right) + \zeta_{ni} \\
&\equiv \boldsymbol{S}_i^\top \begin{bmatrix} \theta_0 \\ \theta_0^{(1)} \end{bmatrix} + \sum_{j=2}^{p-1} \frac{1}{j!} \left(\hat{\eta}_n(\mathbf{Z}_i) - 1\right)^j m_{F_0}^{(j)}(1) \\
&\quad + \frac{1}{p!} \left(\hat{\eta}_n(\mathbf{Z}_i) - 1\right)^p \bar{\theta}_{ni}^{(p)} + \zeta_{ni},
\end{aligned} \tag{36}
$$

where $m_{F_0}^{(j)}(1)$ for each $j \in \{1, \ldots, p\}$ denotes the left-hand limit of $m_{F_0}^{(j)}(q)$ as $q \uparrow 1$, i.e., $m_{F_0}^{(j)}(1) = \lim_{q \uparrow 1} \left(d^j / dq^j\right) m_{F_0}(q')\big|_{q'=q}$. In addition, for each $i \in \{1, \ldots, n\}$, $\hat{\eta}_{ni}^*$ denotes a point between $\hat{\eta}_n(\mathbf{Z}_i)$ and one. It follows that

$$
\begin{aligned}
\hat{m}_{n1}(1) &= \theta_0 + \boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \sum_{j=2}^{p-1} \frac{1}{j!} \left(\hat{\eta}_n(\mathbf{Z}_i) - 1\right)^j m_{F_0}^{(j)}(1) \\
&\quad + \frac{1}{p!} \boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \left(\hat{\eta}_n(\mathbf{Z}_i) - 1\right)^p \bar{\theta}_{ni}^{(p)}
\end{aligned}
$$

$$+\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_{ni}. \tag{37}$$

Observe that for each $i \in \{1, \ldots, n\}$,

$$\begin{aligned} & \hat{\eta}_n(\boldsymbol{Z}_i) - 1 \\ = \; & \eta_0(\boldsymbol{Z}_i) - 1 + [(\hat{\eta}_n(\boldsymbol{Z}_i) - \eta_n(\boldsymbol{Z}_i)) - (\hat{\eta}_0(\boldsymbol{Z}_i) - \eta_0(Z_i))] \\ & + (\eta_n(\boldsymbol{Z}_i) - \eta_0(\boldsymbol{Z}_i)) + (\hat{\eta}_0(\boldsymbol{Z}_i) - \eta_0(\boldsymbol{Z}_i)) \\ \equiv \; & \eta_0(\boldsymbol{Z}_i) - 1 + R_{ni1} + R_{ni2} + R_{ni3}. \end{aligned} \tag{38}$$

Notice that

$$\max_{1 \le i \le n} |R_{ni1}| = o_p\left(n^{-\frac{1}{2}}\right) \tag{39}$$

by Lemma 3 and that

$$\max_{1 \le i \le n} |R_{ni2}| = O_p\left(n^{-\frac{1}{2}}\right) \tag{40}$$

given the $\sqrt{n}$-consistency of $\hat{\boldsymbol{\gamma}}_n$ and the assumption that $F_0(\cdot)$ has a bounded derivative on the support of $\boldsymbol{Z}^\top\boldsymbol{\gamma}_0$. Finally, we have

$$\max_{1 \le i \le n} |R_{ni3}| = O_p\left(n^{-\frac{1}{2}}\right) \tag{41}$$

by Donsker's theorem.

Now consider $(nh_n)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top$. We have

$$\begin{aligned} & \frac{1}{nh_n} \sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \\ = \; & \frac{1}{nh_n} \sum_{i=1}^n \\ & \begin{bmatrix} K\left(\frac{1}{h_n}(\hat{\eta}_n(\boldsymbol{Z}_i) - 1)\right) & (\hat{\eta}_n(\boldsymbol{Z}_i) - 1) K\left(\frac{1}{h_n}(\hat{\eta}_n(\boldsymbol{Z}_i) - 1)\right) \\ (\hat{\eta}_n(\boldsymbol{Z}_i) - 1) K\left(\frac{1}{h_n}(\hat{\eta}_n(\boldsymbol{Z}_i) - 1)\right) & (\hat{\eta}_n(\boldsymbol{Z}_i) - 1)^2 K\left(\frac{1}{h_n}(\hat{\eta}_n(\boldsymbol{Z}_i) - 1)\right) \end{bmatrix}. \end{aligned} \tag{42}$$

We have for each $i \in \{1, \ldots, n\}$ that

$$K_i$$

$$
\begin{aligned}
= \quad & K\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right) \\
& +\frac{1}{h_n}\left(R_{ni1}+R_{ni2}+R_{ni3}\right)K^{(1)}\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right) \\
& +\frac{1}{h_n^2}\left(R_{ni1}+R_{ni2}+R_{ni3}\right)^2 K^{(2)}\left(\Delta_{ni}\right),
\end{aligned} \tag{43}
$$

where R_{ni1}, R_{n2} and R_{ni3} are as given above in (38), and where Δ_{ni} is a point between $h_n^{-1}\left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)$ and $h_n^{-1}\left(\eta_0(\mathbf{Z}_i)-1\right)$. It follows that

$$
\begin{aligned}
& \frac{1}{nh_n}\sum_{i=1}^{n} K_i \\
= \quad & \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right) \\
& +\frac{1}{nh_n^2}\sum_{i=1}^{n}\left(R_{ni1}+R_{ni2}+R_{ni3}\right)K^{(1)}\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right) \\
& +\frac{1}{2nh_n^3}\sum_{i=1}^{n}\left(R_{ni1}+R_{ni2}+R_{ni3}\right)^2 K^{(2)}\left(\Delta_{ni}\right) \\
= \quad & \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right)+O_p\left(\frac{1}{\sqrt{n}}\right)+O_p\left(\frac{1}{nh_n^3}\right) \\
= \quad & \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right)+o_p(1)
\end{aligned} \tag{44}
$$

given the results (39)–(41) above and the assumptions that $nh_n^3 \to \infty$ and that $K^{(2)}(\cdot)$ is bounded.

From (42) and (44) one can use standard calculations (e.g., Ruppert and Wand, 1994) to deduce that

$$
\frac{1}{nh_n}\sum_{i=1}^{n} \boldsymbol{S}_i K_i \boldsymbol{S}_i^{\top} = \begin{bmatrix} 1 & 0 \\ 0 & h_n^p \int u^p K(u)du \end{bmatrix} + o_p(1),
$$

which implies that

$$
\boldsymbol{e}_1^{\top}\left(\frac{1}{nh_n}\sum_{i=1}^{n} \boldsymbol{S}_i K_i \boldsymbol{S}_i^{\top}\right)^{-1} = \boldsymbol{e}_1^{\top} + o_p(1). \tag{45}
$$

Similar calculations yield

$$
\begin{aligned}
& \frac{1}{nh_n}\sum_{i=1}^{n} \boldsymbol{S}_i K_i \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^j m_{F_0}^{(j)}(1) \\
= \quad & \frac{1}{nh_n}\sum_{i=1}^{n} \begin{bmatrix} \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^j K_i \cdot m_{F_0}^{(j)}(1) \\ \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^{j+1} K_i \cdot m_{F_0}^{(j)}(1) \end{bmatrix} \\
= \quad & \begin{bmatrix} 0 \\ h_n^{j+1}\int u^{j+1}K(u)du \cdot m_{F_0}^{(j)}(1) \end{bmatrix} + o_p(1)
\end{aligned} \tag{46}
$$

for each $j \in \{2,\ldots,p-1\}$, and

$$
\begin{aligned}
& \frac{1}{nh_n}\sum_{i=1}^{n} \boldsymbol{S}_i K_i \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^p m_{F_0}^{(p)}(1) \\
= \quad & \frac{1}{nh_n}\sum_{i=1}^{n} \begin{bmatrix} \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^p K_i \bar{\theta}_{ni}^{(p)} \\ \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^{p+1} K_i \bar{\theta}_{ni}^{(p)} \end{bmatrix} \\
= \quad & \begin{bmatrix} h_n^{p}\int u^{p}K(u)du \cdot m_{F_0}^{(p)}(1) \\ h_n^{p+1}\int u^{p+1}K(u)du \cdot m_{F_0}^{(p+1)}(1) \end{bmatrix} + o_p(1).
\end{aligned} \tag{47}
$$

Combining (45) with (46) and (47) we get

$$
\begin{aligned}
& \boldsymbol{e}_1^\top \left(\sum_{i=1}^{n} \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^{n} \boldsymbol{S}_i K_i \sum_{j=2}^{p-1} \frac{1}{j!} \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^j m_{F_0}^{(j)}(1) \\
& +\frac{1}{p!}\boldsymbol{e}_1^\top \left(\sum_{i=1}^{n} \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^{n} \boldsymbol{S}_i K_i \left(\hat{\eta}_n(\mathbf{Z}_i)-1\right)^{p+1} \bar{\theta}_{ni}^{(p)} \\
= \quad & \frac{h_n^p}{p!}\int u^p K(u)du \cdot m_{F_0}^{(p)}(1) + o_p(1).
\end{aligned} \tag{48}
$$

Now consider $\zeta_{ni} = W_i - m_{F_0}\left(\hat{\eta}_n(\mathbf{Z}_i)\right)$ for each $i \in \{1,\ldots,n\}$. We have

$$
\begin{aligned}
& m_{F_0}\left(\hat{\eta}_n(\mathbf{Z}_i)\right) \\
= \quad & m_{F_0}\left(\eta_0(\mathbf{Z}_i)\right) \\
& + \{[(\eta_n(\mathbf{Z}_i)-\eta_n(\mathbf{Z}_i)) + (\hat{\eta}_0(\mathbf{Z}_i)-\eta_0(\mathbf{Z}_i))] + (\eta_n(\mathbf{Z}_i)-\eta_0(\mathbf{Z}_i)) \\
& + (\hat{\eta}_0(\mathbf{Z}_i)-\eta_0(\mathbf{Z}_i))\} \cdot m_{F_0}^{(1)}\left(\hat{\eta}_{ni}^{**}\right)
\end{aligned}
$$

$$= \quad m_{F_0}(\eta_0(\mathbf{Z}_i)) + (R_{ni1} + R_{ni2} + R_{ni3}) \cdot m_{F_0}^{(1)}\left(\hat{\eta}_{ni}^{**}\right),$$

where $\hat{\eta}_{ni}^{**}$ is an intermediate value. It follows that

$$\begin{aligned} \zeta_{ni} &= W_i - m_{F_0}(\eta_0(\mathbf{Z}_i)) + (R_{ni1} + R_{ni2} + R_{ni3}) \cdot m_{F_0}^{(1)}\left(\hat{\eta}_{ni}^{**}\right) \\ &\equiv \zeta_i + (R_{ni1} + R_{n2} + R_{ni3})\,\bar{\theta}_{ni}^{(1)}, \end{aligned} \tag{49}$$

and so

$$\begin{aligned} &\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\boldsymbol{S}_iK_i\zeta_{ni} \\ = \;& \frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\boldsymbol{S}_iK_i\zeta_i + \frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\begin{bmatrix} K_i(R_{ni1} + R_{n2} + R_{ni3})\,\bar{\theta}_{ni}^{(1)} \\ K_i(\hat{\eta}_n(\mathbf{Z}_i) - 1)(R_{ni1} + R_{n2} + R_{ni3})\,\bar{\theta}_{ni}^{(1)} \end{bmatrix}. \end{aligned} \tag{50}$$

Recall that identification of $\boldsymbol{\gamma}_0$ subject to the conditions of Assumption 1 and the smoothness conditions in Assumptions 2.4a–2.4(b)i jointly imply that $m_{F_0}^{(1)}(q)$ is bounded for all $q \in (0,1)$. It follows that there exists a constant $C_1 \in (0,\infty)$ such that

$$\begin{aligned} \left|\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}K_i(R_{ni1} + R_{ni2} + R_{ni3})\,\bar{\theta}_{ni}^{(1)}\right| &\leq \frac{1}{\sqrt{nh_n}} \cdot C_1 n^{-\frac{1}{2}} \cdot (nh_n) \cdot \frac{1}{nh_n}\sum_{i=1}^{n}|K_i| \\ &= O_p\left(\sqrt{h_n}\right) \\ &= o_p(1). \end{aligned}$$

Similar calculations show that the second component of the second term in (50) is $o_p(1)$. It follows that

$$\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\boldsymbol{S}_iK_i\zeta_{ni} = \frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\boldsymbol{S}_iK_i\zeta_i + o_p(1). \tag{51}$$

Combining (45), (48) and (51) yields

$$\hat{m}_{n1}(1) = \theta_0 + \frac{h_n^p}{p!}\int u^pK(u)du \cdot m_{F_0}^{(p)}(1) + \boldsymbol{e}_1^\top \cdot \frac{1}{nh_n}\sum_{i=1}^{n}\boldsymbol{S}_iK_i\zeta_i + o_p(1). \tag{52}$$

Exploiting the decomposition in (43) produces the result

$$\boldsymbol{e}_1^\top \cdot \frac{1}{nh_n}\sum_{i=1}^{n}\boldsymbol{S}_iK_i\zeta_i$$

$$
\begin{aligned}
&= \frac{1}{nh_n}\sum_{i=1}^{n}\zeta_i \\
&= \frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{1}{h_n}\left(\eta_0(\boldsymbol{Z}_i)-1\right)\right)\zeta_i + O_p\left(n^{-\frac{1}{2}}\right) + O_p\left(\frac{1}{nh_n^3}\right) \\
&= \frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{1}{h_n}\left(\eta_0(\boldsymbol{Z}_i)-1\right)\right)\zeta_i + o_p(1) \qquad (53)
\end{aligned}
$$

under the condition that $nh_n^3 \to \infty$. It follows from (52) and (53) that

$$
\begin{aligned}
&\hat{m}_{n1}(1) \\
&= \theta_0 + \frac{h_n^p}{p!}\int u^p K(u)du \cdot m_{F_0}^{(p)}(1) \\
&\quad + \frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{1}{h_n}\left(\eta_0(\boldsymbol{Z}_i)-1\right)\right)\zeta_i + o_p(1). \qquad (54)
\end{aligned}
$$

Next, consider the term $\hat{m}_{n2}(1) = -\boldsymbol{e}_1^\top\left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1}\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{X}_i^\top\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right)$. We have

$$
\boldsymbol{e}_1^\top \cdot \frac{1}{nh_n}\sum_{i=1}^{n}\boldsymbol{S}_i K_i \boldsymbol{X}_i^\top\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right) = \frac{1}{nh_n}\sum_{i=1}^{n}K_i \boldsymbol{X}_i^\top\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right),
$$

where

$$
\left|\frac{1}{nh_n}\sum_{i=1}^{n}K_i \boldsymbol{X}_i^\top\left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right)\right| \le \frac{1}{nh_n}\sum_{i=1}^{n}|K_i|\,\|\boldsymbol{X}_i\| \cdot \left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right\| = O_p\left(n^{-\frac{1}{2}}\right), \qquad (55)
$$

and where the decomposition appearing above in (43) has been applied, along with the assumptions that $E\left[\|\boldsymbol{X}_1\|\right] < \infty$ and $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right\| = O_p\left(n^{-1/2}\right)$. Combining (45) with (55) yields the result

$$
\hat{m}_{n2}(1) = O_p\left(n^{-\frac{1}{2}}\right), \qquad (56)
$$

while combining (56) with (54) produces

$$
\begin{aligned}
&\hat{\theta}_n \\
&= \theta_0 + \frac{h_n^p}{p!}\int u^p K(u)du \cdot m_{F_0}^{(p)}(1)
\end{aligned}
$$

$$+\frac{1}{nh_n}\sum_{i=1}^{n}K\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right)\zeta_i+o_p(1).$$

It follows that

$$\begin{aligned}&\sqrt{nh_n}\left(\hat{\theta}_n-\theta_0-\frac{h_n^p}{p!}\int u^pK(u)du\cdot m_{F_0}^{(p)}(1)\right)\\ =\ &\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}K\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_i)-1\right)\right)\zeta_i+o_p(1),\end{aligned}\tag{57}$$

where the leading term is asymptotically normal mean-zero with variance

$$\frac{1}{h_n}E\left[K^2\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_1)-1\right)\right)\zeta_1^2\right]\tag{58}$$

$$=\ \frac{1}{h_n}E\left[K^2\left(\frac{1}{h_n}\left(\eta_0(\mathbf{Z}_1)-1\right)\right)E\left[\zeta_1^2\middle|\eta_0(\mathbf{Z}_1)\right]\right]\tag{59}$$

$$\to\ E\left[U_1^2\middle|F_0\left(\mathbf{Z}^\top\boldsymbol{\gamma}_0\right)=1\right]\int K^2(u)du\tag{60}$$

$$=\ \sigma^2_{U|F_0(\mathbf{Z}^\top\boldsymbol{\gamma}_0)}(1)\int K^2(u)du\tag{61}$$

The conclusion of Theorem 1 is immediate.

A.3 Proof of Theorem 2

For any $s>0$ and $t\in\mathbb{R}$, define $L_s(t)\equiv 1\{|t|>s\}$. Let $\boldsymbol{\psi}_{10}\equiv(\theta_0,\boldsymbol{\beta}_0^\top,\boldsymbol{\gamma}_0^\top)$ denote a point in $\mathbb{R}^{1+k+l}$, and let $\boldsymbol{\psi}_{1n}$ denote a generic vector in the corresponding set $\Psi_{1n}^*\equiv\Theta_n\times B_n\times\Gamma_n$, where

$$\Theta_n\ =\ \left\{\theta\in\mathbb{R}:\ n^{\frac{p}{2p+1}}|\theta-\theta_0|\le\kappa_1\right\},\tag{62}$$

$$B_n\ =\ \left\{\boldsymbol{\beta}\in\mathbb{R}^k:\ \sqrt{n}\,\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\le\kappa_2\right\},\tag{63}$$

$$\Gamma_n\ =\ \left\{\boldsymbol{\gamma}\in\mathbb{R}^l:\ \sqrt{n}\,\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|\le\kappa_3\right\}\tag{64}$$

for some positive constants κ_1, κ_2 and κ_3. Let $g_{\psi_{2n}}$ denote a joint conditional density for (U,V) given $\boldsymbol{X}$ and $\boldsymbol{Z}$ lying on some curve in a shrinking neighbourhood Ψ_{2n}^* of a bivariate density g_0 satisfying all the relevant conditions of Assumptions 1 and 2 for a conditional density of (U,V) given $\boldsymbol{X}=\boldsymbol{x}$ and $\boldsymbol{Z}=\boldsymbol{z}$, and such that $g_{\psi_{2n0}}=g_0$ for some $\psi_{2n0}\in\Psi_{2n}^*$.

Let $E_{\boldsymbol{\psi}_{1n},g_{\psi_{2n}}}[\cdot]$ denote expectation under the corresponding point $(\boldsymbol{\psi}_{1n}, g_{\psi_{2n}}) \in \Psi_n$. Begin by noting that (17) and (18) may be rewritten as

$$\liminf_{n\to\infty} \sup_{\boldsymbol{\psi}_{1n}\in\Psi^*_{1n},\, \psi_{2n}\in\Psi^*_{2n}} E_{\boldsymbol{\psi}_{1n},g_{\psi_{2n}}} \left[L_s \left(n^{\frac{p}{2p+1}} (\theta_n - \theta) \right) \right] > 0 \tag{65}$$

and

$$\lim_{s\to 0} \liminf_{n\to\infty} \sup_{\boldsymbol{\psi}_{1n}\in\Psi^*_{1n},\, \psi_{2n}\in\Psi^*_{2n}} E_{\boldsymbol{\psi}_{1n},g_{\psi_{2n}}} \left[L_s \left(n^{\frac{p}{2p+1}} (\theta_n - \theta) \right) \right] = 1, \tag{66}$$

respectively.

Consider the generalization of the Hájek–Le Cam asymptotic minimax theorem (e.g., Ibragimov and Has'minskii, 1981, Theorem 12.1) given in Ibragimov and Has'minskii (1981, inequality (II.12.18)). One can deduce from Ibragimov and Has'minskii (1981, inequality (II.12.18)) that (65)–(66) hold, thus implying (17)–(18) and (14)–(15), if for some $\boldsymbol{\psi}_{10} \in \mathbb{R}^{1+k+l}$ and $\psi_{20} \in \mathbb{R}$, there exists a parametrization $\psi_{2n} \longrightarrow g_{\psi_{2n}}$ on Ψ^*_{2n} with $g_{\psi_{20}} = g_0$ such that the conditional joint distribution of (D, Y) given $\boldsymbol{X}$ and $\boldsymbol{Z}$ is locally asymptotically normal (LAN) at $[\ \boldsymbol{\psi}_{10}^\top \quad \psi_{20}\ ]$ in the sense of Condition 1 given below.

Let $\psi_{2n} \longrightarrow g_{\psi_{2n}}$ be such a parametrization of the conditional joint density of $[\ U \quad V\ ]$ given $\boldsymbol{X}$ and $\boldsymbol{Z}$, and let $l\,(\boldsymbol{\psi}_{1n}, \psi_{2n}; D, Y|\, \boldsymbol{X}, \boldsymbol{Z})$ denote the conditional log-likelihood of (D, Y) given $\boldsymbol{X}$ and $\boldsymbol{Z}$ evaluated at the point $(\boldsymbol{\psi}_{1n}, g_{\psi_{2n}}) \in \Psi^*_{1n} \times \Psi^*_{2n}$. Let

$$\left\{ (D_i, Y_i, \boldsymbol{X}_i^\top, \boldsymbol{Z}_i^\top) : \ i = 1, \ldots, n \right\}$$

denote a sample of ordered $(2+k+l)$-tuples generated by (1)–(3). The LAN condition is specified as follows:

Condition 1 (LAN)**.** *The point $(\boldsymbol{\psi}_{10}, g_0)$, where $\boldsymbol{\psi}_{10} = (\theta_0, \boldsymbol{\beta}_0^\top, \boldsymbol{\gamma}_0^\top)^\top \in \mathbb{R}^{1+k+l}$, identifies the conditional joint distribution of each (D_i, Y_i) given $\boldsymbol{X}_i$ and $\boldsymbol{Z}_i$. In addition, the point $\boldsymbol{\psi}_{1n} \equiv (\theta_n, \boldsymbol{\beta}_n^\top, \boldsymbol{\gamma}_n^\top)^\top$ is such that $n^{p/(2p+1)}(\theta_n - \theta_0) \to \omega_1^{p^*}$ for some $\omega_1 \neq 0$, where $p^* \geq 3$ is the odd integer specified above in Assumption 2.4b. In addition, $\sqrt{n}[(\boldsymbol{\beta}_n - \boldsymbol{\beta}_0)^\top, (\boldsymbol{\gamma}_n - \boldsymbol{\gamma}_0)^\top]^\top \longrightarrow \boldsymbol{\omega}_2$ for some $\boldsymbol{\omega}_2 \neq \boldsymbol{0}$, while ψ_{2n} is such that $\sqrt{n}\psi_{2n} \longrightarrow \omega_3$ for some constant $\omega_3 \neq 0$.*

There exists a random $(2+k+l)$-vector $\boldsymbol{S}_{n0}$ and a $[(2+k+l)\times(2+k+l)]$-matrix $\boldsymbol{I}_0$ of full rank such that the conditional distribution $\boldsymbol{S}_{n0}|\,(\boldsymbol{X}^\top, \boldsymbol{Z}^\top) \xrightarrow{d} N_{2+k+l}(\boldsymbol{0}, \boldsymbol{I}_0)$ and

$$\begin{aligned} &\sum_{i=1}^{n} \left(l\,(\boldsymbol{\psi}_{1n}, \psi_{2n}; D_i, Y_i|\, \boldsymbol{X}_i, \boldsymbol{Z}_i) - l\,(\boldsymbol{\psi}_{10}, 0; Y_i, D_i|\, \boldsymbol{X}_i, \boldsymbol{Z}_i) \right) \\ &= [\ \omega_1^{p^*} \quad \boldsymbol{\omega}_2 \quad \omega_3\ ]\boldsymbol{S}_{n0} - \frac{1}{2}[\ \omega_1^{p^*} \quad \boldsymbol{\omega}_2 \quad \omega_3\ ]\boldsymbol{I}_0 \begin{bmatrix} \omega_1^{p^*} \\ \boldsymbol{\omega}_2 \\ \omega_3 \end{bmatrix} + o_p(1). \end{aligned}$$

It follows that Theorem 2 is proved if for a given $\boldsymbol{\psi}_{10} \in \mathbb{R}^{1+k+l}$, one can exhibit a parametrization $\psi_{2n} \longrightarrow g_{\psi_{2n}}$ on a shrinking neighbourhood Ψ^*_{2n} of g_0 such that the corresponding conditional log likelihood of (D, Y) given $[\ \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$ satisfies Condition 1 at the point $[\ \boldsymbol{\psi}_{10}^\top \quad 0\ ]$.

In this connection consider arbitrary points $\theta, \theta_0 \in \mathbb{R}$ and $\boldsymbol{\beta}, \boldsymbol{\beta}_0 \in \mathbb{R}^k$. Let $\delta_1 \equiv \theta - \theta_0$ and $\boldsymbol{\delta}_2 \equiv \boldsymbol{\beta} - \boldsymbol{\beta}_0$, and let Δ_{1n} and Δ_{2n} denote neighbourhoods of the origin given by

$$\Delta_{1n} = \left\{\delta_1 : n^{\frac{p}{2p+1}} |\delta_1| \leq \kappa_1\right\}, \tag{67}$$

$$\Delta_{2n} = \left\{\boldsymbol{\delta}_2 : n^{\frac{1}{2}} \|\boldsymbol{\delta}_2\| \leq \kappa_2\right\} \tag{68}$$

for positive constants κ_1 and κ_2. Next, let $g_{\mathbf{0}U|V,X,Z}(\cdot|\cdot)$ denote a conditional joint density of U given $[\ V \quad \boldsymbol{X}^\top \quad \boldsymbol{Z}^\top\ ]$ that satisfies all relevant conditions of Assumptions 1 and 2. Let $\eta_1(u|\boldsymbol{x}, z)$ be a non-constant measurable function such that

$$E\left[\eta_1\left(U|\,\boldsymbol{X}, \boldsymbol{Z}\right)|\, D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] = 0$$

and

$$E\left[\eta_1^2\left(U|\,\boldsymbol{X}, \boldsymbol{Z}\right)\middle|\, D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] < \infty$$

with probability one. Let Δ_{3n} be a neighbourhood of the origin on $\mathbb{R}$ given by

$$\Delta_{3n} = \left\{\delta_3 : n^{\frac{1}{2}} |\delta_3| \leq \kappa_3\right\} \tag{69}$$

for some positive constant κ_3. Define the following curve on $\Delta_{1n} \times \Delta_{2n} \times \Delta_{3n}$ parameterized by $(\delta_1, \boldsymbol{\delta}_2^\top, \delta_3)$ and passing through $g_{\mathbf{0}U|V,X,Z}(y - \theta_0 - \boldsymbol{x}^\top\boldsymbol{\beta}_0|v, \boldsymbol{x}, z)$:

$$\begin{aligned} & g_{\boldsymbol{\delta}U|V}\left(y - \theta_0 - \boldsymbol{x}^\top\boldsymbol{\beta}_0\middle|\, v, \boldsymbol{x}, z)\right) \\ = \;& \left(1 + \delta_3\eta_1(y - \theta_0 - \boldsymbol{x}^\top\boldsymbol{\beta}_0|\boldsymbol{x}, z)\right) \\ & \cdot g_{\mathbf{0}U|V}\left(y - \theta_0 - \frac{1}{p^*!}\eta_2\left(y - \theta_0 - \boldsymbol{x}^\top\beta_0\middle|\, \boldsymbol{x}, z\right)\delta_1^{p^*}\right. \\ & \left.-\boldsymbol{x}^\top(\boldsymbol{\beta}_0 + \boldsymbol{\delta}_2)\middle|\, v - \frac{1}{p^*!}\eta_2(y - \theta_0 - \boldsymbol{x}^\top\beta_0|\boldsymbol{x}, z)\delta_1^{p^*}, \boldsymbol{x}, z\right), \end{aligned} \tag{70}$$

where $\eta_2(u|\boldsymbol{x}, z)$ is a non-constant function such that $E\left[\eta_2^2\left(U|\,\boldsymbol{X}, \boldsymbol{Z}\right)\middle|\, D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] < \infty$ and where

$$E\left[\eta_2(U|\boldsymbol{X}, \boldsymbol{Z}) \cdot \left.\frac{\partial^{p^*}}{\partial\delta_1^{p^*}} l(\boldsymbol{\delta})\right|_{\boldsymbol{\delta}=\mathbf{0}} \middle|\, D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] = 0$$

for $l(\boldsymbol{\delta})$ denoting the sub-model conditional likelihood function given below in (72).

Now let $(d, y, \boldsymbol{x}^\top, \boldsymbol{z}^\top) \in \{0,1\} \times \mathbb{R}^{2+k+l}$ be a point in the support of $(D, Y, \boldsymbol{X}^\top, \boldsymbol{Z}^\top)$. Let $\boldsymbol{\gamma}, \boldsymbol{\gamma}_0 \in \mathbb{R}^l$ be arbitrary points, and define $\boldsymbol{\delta}_4 \equiv \boldsymbol{\gamma} - \boldsymbol{\gamma}_0$. Let Δ_{4n} be a neighbourhood of the origin given by

$$\Delta_{4n} = \left\{\boldsymbol{\delta}_4 : n^{\frac{1}{2}} \|\boldsymbol{\delta}_4\| \leq \kappa_4\right\} \tag{71}$$

for some positive constant κ_4, and let $\Delta_n \equiv \Delta_{1n} \times \Delta_{2n} \times \Delta_{3n} \times \Delta_{4n}$ and $\boldsymbol{\delta} \equiv [\ \delta_1 \quad \boldsymbol{\delta}_2^\top \quad \delta_3 \quad \boldsymbol{\delta}_4^\top\ ]^\top$.

Let $g_{\mathbf{0}V}(\cdot|\cdot)$ denote a conditional density of V given $\boldsymbol{X}$ and $\boldsymbol{Z}$ satisfying all relevant conditions of Assumptions 1 and 2. For a given n the conditional log-likelihood of (d, y) given $\boldsymbol{x}$ and $\boldsymbol{z}$ of the submodel indexed by $\left(\boldsymbol{\delta}, g_{\boldsymbol{\delta}U|V,\boldsymbol{X},\boldsymbol{Z}}\right)$, where $g_{\boldsymbol{\delta}U|V,\boldsymbol{X},\boldsymbol{Z}}(y - \theta_0 - \boldsymbol{x}^\top\boldsymbol{\beta}_0|v, \boldsymbol{x}, \boldsymbol{z})$ is as given above in (70), is

$$\begin{aligned} & l(\boldsymbol{\delta}) \\ \equiv\ & l\left(\boldsymbol{\delta}; d, y|\,\boldsymbol{x}, \boldsymbol{z}\right) \\ \equiv\ & d \log \int_{-\infty}^{\boldsymbol{z}^\top(\boldsymbol{\gamma}_0+\boldsymbol{\delta}_4)} g_{\boldsymbol{\delta}U|V}\left(y - \theta_0 - \boldsymbol{x}^\top\boldsymbol{\beta}_0\middle|\, v, \boldsymbol{x}, \boldsymbol{z}\right) g_{\mathbf{0}V}(v|\boldsymbol{x}, \boldsymbol{z}) dv \\ & +(1-d) \log \int_{\boldsymbol{z}^\top(\boldsymbol{\gamma}_0+\boldsymbol{\delta}_4)}^{\infty} g_{\mathbf{0}V}(v|\boldsymbol{x}, \boldsymbol{z}) dv. \end{aligned} \tag{72}$$

Observe from (72) that $\left(\partial^m/\partial\delta_1^m\right) l(\boldsymbol{\delta})\big|_{\boldsymbol{\delta}=\mathbf{0}} \equiv 0$ for each $m \in \{1, \ldots, p^* - 1\}$ and all $(d, y, \boldsymbol{x}^\top, \boldsymbol{z}^\top)$, while $s_{\delta_1}^{(p^*)}(\mathbf{0}) \equiv \left(\partial^{p^*}/\partial\delta_1^{p^*}\right) l(\boldsymbol{\delta})\Big|_{\boldsymbol{\delta}=\mathbf{0}}$ is both nonzero with positive probability and linearly independent, with probability one, of the submodel scores corresponding to $\boldsymbol{\delta}_2$, δ_3 and $\boldsymbol{\delta}_4$. In particular, for $s_{\delta_3}(\mathbf{0}) \equiv (\partial/\partial\delta_3)\, l(\boldsymbol{\delta})|_{\boldsymbol{\delta}=\mathbf{0}}$ we have

$$E\left[s_{\delta_1}^{(p^*)}(\mathbf{0})\middle| D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] = E\left[s_{\delta_3}(\mathbf{0})\middle| D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] = 0$$

and

$$E\left[s_{\delta_1}^{(p^*)}(\mathbf{0}) s_{\delta_3}(\mathbf{0})\middle| D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] = 0.$$

Similarly, for $\boldsymbol{s}_{\boldsymbol{\delta}_2}(\mathbf{0}) \equiv (\partial/\partial\boldsymbol{\delta}_2)\, l(\boldsymbol{\delta})|_{\boldsymbol{\delta}=\mathbf{0}}$ and $\boldsymbol{s}_{\boldsymbol{\delta}_4}(\mathbf{0}) \equiv (\partial/\partial\boldsymbol{\delta}_4)\, l(\boldsymbol{\delta})|_{\boldsymbol{\delta}=\mathbf{0}}$, one can show that $E\left[s_{\delta_1}^{(p^*)}(\mathbf{0})\boldsymbol{s}_{\boldsymbol{\delta}_2}(\mathbf{0})\middle| D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] = \mathbf{0}_{k\times 1}$ and $E\left[s_{\delta_1}^{(p^*)}(\mathbf{0})\boldsymbol{s}_{\boldsymbol{\delta}_4}(\mathbf{0})\middle| D = 1, \boldsymbol{X}, \boldsymbol{Z}\right] = \mathbf{0}_{l\times 1}$, which indicates that $s_{\delta_1}^{(p^*)}(\mathbf{0})$ is almost surely conditionally uncorrelated given $\boldsymbol{X}$ and $\boldsymbol{Z}$ with the submodel scores corresponding to $\boldsymbol{\delta}_2$, δ_3 and $\boldsymbol{\delta}_4$.

In what follows, Condition 1 is shown to apply to a condensed version of the submodel with conditional log-likelihood given in (72). This simplification involves assuming that the finite-dimensional nuisance parameters $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ are known, in which case the argument $\boldsymbol{\delta}$ appearing in (72) reduces to the ordered pair $\boldsymbol{\delta} = [\ \delta_1 \quad \delta_3\ ]^\top = [\ \theta - \theta_0 \quad \delta_3\ ]^\top$.

In addition, the set Δ_n is understood to have the form $\Delta_n = \Delta_{1n} \times \Delta_{3n}$, where Δ_{1n} and Δ_{3n} are as given above in (67) and (69), respectively. It is shown that the the family of conditional joint distributions of (d, y) given $(\boldsymbol{x}^\top, \boldsymbol{z}^\top)$ and indexed by $(\boldsymbol{\delta}, g_{\delta U|V})$ for $\boldsymbol{\delta} \in \Delta_n$ is LAN at the point $\boldsymbol{\delta} = \mathbf{0}_{2\times 1}$. The argument for the general case in which Condition 1 is shown to apply to the conditional log-likelihood appearing in (72) in which $\boldsymbol{\beta}_0$ and $\boldsymbol{\gamma}_0$ are both unknown is similar, although rather more notationally complex.

For $\boldsymbol{\delta} = (\delta_1, \delta_3)$ as discussed above and (j_1, j_2) denoting an ordered pair of non-negative integers, define the derivatives $l^{(j_1,j_2)}(\boldsymbol{\delta}) \equiv \left(\partial^{j_1+j_2}/\partial\delta_1^{j_1}\partial\delta_3^{j_2}\right) l(\boldsymbol{\delta})$ and also $l_0^{(j_1,j_2)} \equiv l^{(j_1,j_2)}(\mathbf{0})$, where $l(\boldsymbol{\delta})$ is now taken to be the analogue of the conditional log-likelihood given in (72) corresponding to the submodel in which $\boldsymbol{\delta}_2$ and $\boldsymbol{\delta}_4$ are both fixed. Suppose δ_{1n} and δ_{3n} are such that $\left[\ (n^{p/(2p+1)}\delta_{1n})^{1/p^*} \quad \sqrt{n}\delta_{3n}\ \right] \to \left[\ \omega_1 \quad \omega_3\ \right]$ for some $\left[\ \omega_1 \quad \omega_3\ \right] \neq \mathbf{0}$. Let $\boldsymbol{\delta}_n \equiv \left[\ \delta_{1n} \quad \delta_{3n}\ \right]^\top$. For $j_1, j_2 \geq 0$ with $j_1 + j_2 = 2p^* + 1$, define $R_0^{(j_1+j_2)} \equiv l^{(j_1,j_2)}(\bar{\boldsymbol{\delta}}) - l_0^{(j_1,j_2)}$, for some point $\bar{\boldsymbol{\delta}} \in \Delta_n$ such that $\left\|\bar{\boldsymbol{\delta}}\right\| < \|\boldsymbol{\delta}_n\|$.

Observe from previous discussion that $l_0^{(1,0)} = \cdots = l_0^{(p^*-1,0)} = 0$. It follows that

$$
\begin{aligned}
& l(\boldsymbol{\delta}_n) - l(\mathbf{0}) \\
= \ & \omega_1^{p^*}\Bigg[n^{-\frac{1}{2}} \cdot \frac{l_0^{(p^*,0)}}{p^*!} + n^{-\frac{1}{2p^*}}\left(n^{-\frac{1}{2}} \cdot \frac{l_0^{(p^*+1,0)}}{(p^*+1)!}\omega_1\right) + n^{-\frac{1}{2p^*}}\left(\sum_{j_1=2}^{p^*-1} n^{-\frac{1}{2}}\frac{l_0^{(p^*+j_1,0)}}{(p^*+j_1)!}n^{\frac{1-j_1}{2p^*}}\omega_1^{j_1}\right) \\
& +n^{-1}\frac{l_0^{(2p^*,0)}}{(2p^*)!}\omega_1^{p^*} + n^{-\frac{1}{2p^*}}\left(n^{-1}\frac{l_0^{(2p^*+1,0)}}{(2p^*+1)!}\omega_1^{p^*+1} + n^{-1} \cdot \frac{R_0^{(2p^*+1,0)}}{(2p^*+1)!}\omega_1^{p^*+1}\right)\Bigg] \\
& +\omega_3\Bigg\{n^{-\frac{1}{2}} \cdot l_0^{(0,1)} + n^{-\frac{1}{2p^*}} \cdot \left[n^{-\frac{1}{2}}l_0^{(1,1)}\omega_1 + \left(\sum_{j_1=2}^{p^*-1} n^{-\frac{1}{2}}\frac{l_0^{(j_1,1)}}{j_1!}n^{\frac{1-j_1}{2p^*}}\omega_1^{j_1}\right)\right] \\
& +n^{-1}\frac{l_0^{(p^*,1)}}{p^*!}\omega_1^{p^*} + n^{-\frac{1}{2p^*}}\Bigg[n^{-1}\frac{l_0^{(p^*+1,1)}}{(p^*+1)!}\omega_1^{p^*+1} + \left(\sum_{j_1=p^*+2}^{2p^*} n^{-1}\frac{l_0^{(j_1,1)}}{j_1!}n^{\frac{1-j_1}{2p^*}}\omega_1^{j_1}\right) \\
& +n^{-1}\frac{R_0^{(2p^*,1)}}{(2p^*)!}n^{\frac{1-p^*}{2p^*}}\omega_1^{2p^*}\Bigg] \\
& +n^{-1}\frac{l_0^{(0,2)}}{2}\omega_3 + n^{-\frac{1}{2p^*}}\Bigg[n^{-1}\frac{l_0^{(1,2)}}{2}\omega_1\omega_3 + \left(\sum_{j_1=2}^{2p^*-1} n^{-1}\frac{l_0^{(j_1,2)}}{(2+j_1)!}n^{\frac{1-j_1}{2p^*}}\omega_1^{j_1}\omega_3\binom{2+j_1}{j_1}\right) \\
& +n^{-1}\frac{R_0^{(2p^*-1,2)}}{(2p^*+1)!}n^{\frac{2-2p^*}{2p^*}}\omega_1^{2p^*-1}\omega_3\binom{2p^*+1}{2p^*-1}
\end{aligned}
$$

$$
\begin{aligned}
&+\left(\sum_{m=3}^{2p^*+1}\sum_{j_1+j_2=m:\, j_2\geq 3, j_1\geq 0} n^{-1}\frac{l_0^{(j_1,j_2)}}{m!}n^{\frac{(1-j_1)+(2-j_2)p^*}{2p^*}}\omega_1^{j_1}\omega_3^{j_2-1}\binom{m}{j_1}\right)\\
&+\left(\sum_{j_1+j_2=2p^*+1:\, j_2\geq 3, j_1\geq 0} n^{-1}\frac{R_0^{(j_1,j_2)}}{(2p^*+1)!}n^{\frac{1-j_1+(2-j_2)p^*}{2p^*}}\omega_1^{j_1}\omega_3^{j_2-1}\binom{2p^*+1}{j_1}\right)\Bigg]\Bigg\}\\
\equiv\ & \omega_1^{p^*}[A_{1n}+n^{-\frac{1}{2p^*}}A_{2n}+n^{-\frac{1}{2p^*}}A_{3n}+A_{4n}+n^{-\frac{1}{2p^*}}(A_{5n}+A_{6n})]\\
&+\omega_3\{A_{7n}+n^{-\frac{1}{2p^*}}[A_{8n}+A_{9n}]+A_{10n}+n^{-\frac{1}{2p^*}}[A_{11n}+A_{12n}+A_{13n}]+A_{14n}\\
&+n^{-\frac{1}{2p^*}}[A_{15n}+A_{16n}+A_{17n}+A_{18n}+A_{19n}]\}.
\end{aligned}
$$

Let $f_{\boldsymbol{\delta}}(y, d|\, \boldsymbol{x}, z) \equiv \exp(l(\boldsymbol{\delta}))$, where $\boldsymbol{\delta} = [\ \delta_1 \quad \delta_3\ ]^\top$, denote the joint conditional density of (Y, D) given $(\boldsymbol{X}^\top, \boldsymbol{Z}^\top) = (\boldsymbol{x}^\top, z^\top)$ corresponding to the condensed version of the conditional log-likelihood in (72) where $\boldsymbol{\delta}_2$ and $\boldsymbol{\delta}_4$ are fixed. Define

$$
\begin{aligned}
I_{011} &\equiv E\left[\left(\frac{\partial^{p^*}}{\partial\delta_1^{p^*}}\log f_{\boldsymbol{\delta}}(Y, D|\, \boldsymbol{x}, z)\bigg|_{\boldsymbol{\delta}=\mathbf{0}}\right)^2\right];\\
I_{033} &\equiv E\left[\left(\frac{\partial}{\partial\delta_3}\log f_{\boldsymbol{\delta}}(Y, D|\, \boldsymbol{x}, z)\bigg|_{\boldsymbol{\delta}=\mathbf{0}}\right)^2\right]
\end{aligned}
$$

and

$$
I_{031} \equiv I_{013} \equiv E\left[\frac{\partial^{p^*}}{\partial\delta_1^{p^*}}\log f_{\boldsymbol{\delta}}(Y, D|\, \boldsymbol{x}, z)\bigg|_{\boldsymbol{\delta}=\mathbf{0}}\cdot\frac{\partial}{\partial\delta_3}\log f_{\boldsymbol{\delta}}(Y, D|\, \boldsymbol{x}, z)\bigg|_{\boldsymbol{\delta}=\mathbf{0}}\right],
$$

where each expectation is taken at $\boldsymbol{\delta} = \mathbf{0}$, and let

$$
\boldsymbol{I}_0 \equiv \begin{bmatrix} I_{011} & I_{013} \\ I_{031} & I_{033} \end{bmatrix}.
$$

For any ordered pair of nonnegative integers (r_1, r_2) with $3 \leq r_1 + r_2 \leq 2p^* + 1$, one can exploit the form of the parametrization of the conditional density $g_{U|V,\boldsymbol{X},\boldsymbol{Z}}(y - \theta_0 - x\beta_0|v, \boldsymbol{x}, z)$ given above in (70) to deduce that

$$
E\left[\left(\frac{\partial^{r_1+r_2}}{\partial\delta_1^{r_1}\partial\delta_3^{r_2}}f_{\boldsymbol{\delta}}(Y, D|\, \boldsymbol{x}, z)\bigg|_{\boldsymbol{\delta}=\mathbf{0}}\right)^2\right] < \infty,
$$

where the expectation is also taken at $\boldsymbol{\delta} = \mathbf{0}$. The parametrization of the function $g_{U|V,\boldsymbol{X},\boldsymbol{Z}}(\cdot|\cdot)$ given in (70) allows one to apply Rotnitzky et al. (2000, Corollary 1, p. 268) to deduce the following:

- $A_{3n} = O_p\left(n^{-1/(2p^*)}\right) = o_p(1)$.
- $A_{9n} = O_p\left(n^{-1/(2p^*)}\right) = o_p(1)$.
- $A_{12n} = O_p\left(n^{-1/(2p^*)}\right) = o_p(1)$.
- $A_{16n} = O_p\left(n^{-1/(2p^*)}\right) = o_p(1)$.
- $A_{18n} = O_p\left(n^{-1/p^*}\right) = o_p(1)$.
- $A_{6n} = O_p\left(n^{-1/(2p^*)}\right) = o_p(1)$.
- $A_{13n} = o_p\left(n^{-1/p^*}\right) = o_p(1)$.
- $A_{17n} = o_p\left(n^{-2/p^*}\right) = o_p(1)$.
- $A_{19n} = o_p\left(n^{-1/p^*}\right) = o_p(1)$.

Now define

$$\begin{aligned}
C_{011} &\equiv E\left[\left.\frac{\partial^{p^*}}{\partial\delta_1^{p^*}}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\cdot\left.\frac{\partial^{p^*+1}}{\partial\delta_1^{p^*+1}}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\right];\\
C_{013} &\equiv E\left[\left.\frac{\partial^{p^*}}{\partial\delta_1^{p^*}}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\cdot\left.\frac{\partial^{p^*-1}}{\partial\delta_1\partial\delta_3}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\right];\\
C_{031} &\equiv E\left[\left.\frac{\partial}{\partial\delta_3}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\cdot\left.\frac{\partial^{p^*+1}}{\partial\delta_1^{p^*+1}}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\right];
\end{aligned}$$

and

$$C_{033} \equiv E\left[\left.\frac{\partial}{\partial\delta_3}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\cdot\left.\frac{\partial^{p^*-1}}{\partial\delta_1\partial\delta_3}\log f_{\boldsymbol{\delta}}\left(Y,D|\,\boldsymbol{x},\boldsymbol{z}\right)\right|_{\boldsymbol{\delta}=\mathbf{0}}\right],$$

where each expectation is taken at $\boldsymbol{\delta} = \mathbf{0}$, and let

$$\boldsymbol{C}_0 \equiv \begin{bmatrix} C_{011} & C_{013} \\ C_{031} & C_{033} \end{bmatrix}.$$

A further application of Rotnitzky et al. (2000, Corollary 1) yields the following:

- $A_{4n} = \omega_1^{p^*} \left[-(1/2) \cdot I_{011} + o_p \left(n^{-1/(2p^*)} \right) \right]$.
- $A_{5n} = \omega_1^{p^*+1} \left(-C_{011} + o_p \left(n^{-1/(2p^*)} \right) \right)$.
- $A_{10n} = \omega_1^{p^*} \left(-I_{013} + o_p \left(n^{-1/(2p^*)} \right) \right)$.
- $A_{11n} = \omega_1^{p^*+1} \left[-\left(C_{013} + C_{031} \right) + o_p \left(n^{-1/(2p^*)} \right) \right]$.
- $A_{14n} = \omega_3 \left[-(1/2) \cdot I_{013} + o_p \left(n^{-1/(2p^*)} \right) \right]$.
- $A_{15n} = \omega_1 \omega_3 \left(-C_{033} + o_p \left(n^{-1/(2p^*)} \right) \right)$.

Collecting terms, one gets $l(\boldsymbol{\delta}_n) - l(\mathbf{0}) = G_{n0} \left(\omega_1^{p^*}, \omega_3 \right) + H_{n0} \left(\omega_1, \omega_3 \right)$, where

$$\begin{aligned} G_{n0} \left(\omega_1^{p^*}, \omega_3 \right) &= \omega_1^{p^*} \cdot n^{-\frac{1}{2}} \cdot \frac{l_0^{(p^*,0)}}{(p^*)!} + \omega_3 \cdot n^{-\frac{1}{2}} \cdot l_0^{(0,1)} - \frac{1}{2} [\begin{matrix} \omega_1^{p^*} & \omega_3 \end{matrix}] \boldsymbol{I}_0 \begin{bmatrix} \omega_1^{p^*} \\ \omega_3 \end{bmatrix} \\ &= [\begin{matrix} \omega_1^{p^*} & \omega_3 \end{matrix}] \cdot n^{-\frac{1}{2}} \begin{bmatrix} \frac{1}{(p^*)!} l_0^{(p^*,0)} \\ l_0^{(0,1)} \end{bmatrix} - \frac{1}{2} [\begin{matrix} \omega_1^{p^*} & \omega_3 \end{matrix}] \boldsymbol{I}_0 \begin{bmatrix} \omega_1^{p^*} \\ \omega_3 \end{bmatrix}, \end{aligned}$$

and where

$$H_{n0} \left(\omega_1, \omega_3 \right) = n^{-\frac{1}{2p^*}} \omega_1 \left(T_{n0} \left(\omega_1^{p^*}, \omega_3 \right) + o_p(1) \right) + O_p \left(n^{-\frac{1}{2}} \right),$$

where

$$T_{n0} \left(\omega_1^{p^*}, \omega_3 \right) = \omega_1^{p^*} \cdot n^{-\frac{1}{2}} \cdot \frac{l_0^{(p^*+1,0)}}{(p^* + 1)!} + \omega_3 \cdot n^{-\frac{1}{2}} \cdot l_0^{(1,1)} - [\begin{matrix} \omega_1^{p^*} & \omega_3 \end{matrix}] \boldsymbol{C}_0 \begin{bmatrix} \omega_1^{p^*} \\ \omega_3 \end{bmatrix}.$$

An application of Rotnitzky et al. (2000, Corollary 1) and the Cramér–Wold device yields

$$n^{-\frac{1}{2}} \begin{bmatrix} \frac{1}{p^*!} l_0^{(p^*,0)} \\ l_0^{(0,1)} \end{bmatrix} \xrightarrow{d} N_2 \left(\mathbf{0}, \boldsymbol{I}_0 \right).$$

In addition, Rotnitzky et al. (2000, Corollary 1) also implies that

$$n^{-\frac{1}{2}} \cdot \frac{l_0^{(p^*+1,0)}}{(p^*+1)!} = O_p(1)$$

and that

$$n^{-\frac{1}{2}} l_0^{(1,1)} = O_p(1).$$

It follows that $T_{n0}\left(\omega_1^{p^*}, \omega_3\right) = O_p(1)$ and that $H_{n0}(\omega_1, \omega_3) = O_p\left(n^{-1/(2p^*)}\right) = o_p(1)$.

In summary, we have the first-order representation

$$l(\boldsymbol{\delta}_n) - l(\mathbf{0}) = [\ \omega_1^{p^*} \quad \omega_3\ ] \cdot n^{-\frac{1}{2}} \begin{bmatrix} \frac{1}{p^*!} l_0^{(p^*,0)} \\ l_0^{(0,1)} \end{bmatrix} - \frac{1}{2} [\ \omega_1^{p^*} \quad \omega_3\ ] \boldsymbol{I}_0 \begin{bmatrix} \omega_1^{p^*} \\ \omega_3 \end{bmatrix} + o_p(1),$$

where for $(\boldsymbol{X}^\top, \boldsymbol{Z}^\top) = (\boldsymbol{x}^\top, \boldsymbol{z}^\top)$,

$$n^{-\frac{1}{2}} \begin{bmatrix} \frac{1}{p^*!} l_0^{(p^*,0)} \\ l_0^{(0,1)} \end{bmatrix} \xrightarrow{d} N_2\left(\mathbf{0}, \boldsymbol{I}_0\right).$$

It follows that the condensed version of the submodel with conditional log-likelihood given above in (72) and where $\boldsymbol{\delta}_2$ and $\boldsymbol{\delta}_4$ are fixed is LAN at the point $[\ \delta_1 \quad \delta_3\ ]^\top = \mathbf{0}$, and as such, satisfies Condition 1. The general case in which $\boldsymbol{\delta}_2$ and $\boldsymbol{\delta}_4$ are unknown follows *mutatis mutandis*.

A.4 Proof of Theorem 3

The approach taken involves showing that $\sqrt{nh_n^*}(\hat{\theta}_n^* - \theta) = O_p(1)$ as $n \to \infty$ uniformly across sequences $\{\psi_n\} \equiv \{(\boldsymbol{\psi}_{1n}, g)\} \subset \Psi_n$, where the set Ψ_n is as given above in (13). In particular, for each n, $\boldsymbol{\psi}_{1n} = (\theta, \boldsymbol{\beta}^\top, \boldsymbol{\gamma}^\top)^\top \in \mathbb{R}^{1+k+l}$, while g denotes the joint conditional density of (U, V) given $\boldsymbol{X}$ and $\boldsymbol{Z}$.

It suffices to show that for any $\epsilon > 0$ there exists a constant $\delta(\epsilon) \in (0, \infty)$ such that

$$\lim_{n\to\infty} \sup_{\psi_n \in \Psi_n} P_{\psi_n}\left[\sqrt{nh_n^*}|\hat{\theta}_n^* - \theta| > \delta(\epsilon)\right] < \epsilon, \tag{73}$$

where $P_{\psi_n}[\cdot]$ denotes probability measure under ψ_n. One can show using Chebyshev's inequality that the following conditions suffice for (73) to hold with $\delta(\epsilon) = 4/\epsilon$, in particular:

$$\lim_{n\to\infty} \sqrt{nh_n^*} \sup_{\psi_n \in \Psi_n} \left|E_{\psi_n}\left[\theta_n^*\right] - \theta\right| \ < \ \infty; \tag{74}$$

$$\lim_{n\to\infty} nh_n^* \cdot \sup_{\psi_n \in \Psi_n} Var_{\psi_n}\left[\theta_n^*\right] \quad < \quad \infty, \tag{75}$$

where $E_{\psi_n}[\cdot]$ and $Var_{\psi_n}[\cdot]$ respectively denote expectation and variance under a given $\psi_n \in \Psi_n$.

To see that (74) and (75) jointly imply (73), note that $\left|\hat{\theta}_n^* - \theta\right| \leq \left|\hat{\theta}_n^* - E_{\psi_n}\left[\hat{\theta}_n^*\right]\right| + \left|E_{\psi_n}\left[\hat{\theta}_n^*\right] - \theta\right|$, so

$$\begin{aligned}
& P_{\psi_n}\left[\sqrt{nh_n^*}\left|\hat{\theta}_n^* - \theta\right| > \delta(\epsilon)\right] \\
\leq\ & P_{\psi_n}\left[\sqrt{nh_n^*}\left|\hat{\theta}_n^* - E_{\psi_n}\left[\hat{\theta}_n^*\right]\right| + \sqrt{nh_n^*}\left|E_{\psi_n}\left[\hat{\theta}_n^*\right] - \theta\right| > \delta(\epsilon)\right] \\
\leq\ & P_{\psi_n}\left[\sqrt{nh_n^*}\left|\hat{\theta}_n^* - E_{\psi_n}\left[\hat{\theta}_n^*\right]\right| > \frac{\delta(\epsilon)}{2}\right] + P_{\psi_n}\left[\sqrt{nh_n^*}\left|E_{\psi_n}\left[\hat{\theta}_n^*\right] - \theta\right| > \frac{\delta(\epsilon)}{2}\right] \\
\leq\ & \frac{4}{\delta^2(\epsilon)} Var_{\psi_n}\left[\sqrt{nh_n^*}\hat{\theta}_n^*\right] + \frac{2\sqrt{nh_n^*}\left|E_{\psi_n}\left[\hat{\theta}_n^*\right] - \theta\right|}{\delta(\epsilon)},
\end{aligned}$$

from which it follows that

$$\begin{aligned}
& \sup_{\psi_n \in \bar{\Psi}_n} P_{\psi_n}\left[\sqrt{nh_n^*}\left|\hat{\theta}_n^* - \theta\right| > \delta(\epsilon)\right] \\
\leq\ & \frac{4nh_n^*}{\delta^2(\epsilon)} \sup_{\psi \in \Psi_n} Var_{\psi}\left[\hat{\theta}_n^*\right] + \frac{2\sqrt{nh_n^*}}{\delta(\epsilon)} \sup_{\psi_n \in \bar{\Psi}_n}\left|E_{\psi_n}\left[\hat{\theta}_n^*\right] - \theta\right|.
\end{aligned}$$

In what follows, (74) and (75) are proved in sequence.

A.4.1 Proof of (74)

Recall the expression for $\hat{m}_{n1}(1)$ given above in (37). In particular, consider the first bias term appearing in (37). The assumption of a pth-order kernel, the uniform boundedness of $\hat{\eta}_n(\cdot)$ over $n \geq 1$, the boundedness of $K(\cdot)$ and of $m_{F_0}^{(j)}(\cdot)$ for each $j \in \{0, 1, \ldots, p\}$ imply via the bounded convergence theorem that for each sequence $\{\psi_n : \psi_n \in \Psi_n\}$,

$$\begin{aligned}
& \sqrt{nh_n^*} \\
& \cdot E_{\psi_n}\left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \left[\sum_{j=2}^{p-1} \frac{1}{j!}\left(\hat{\eta}_n(\tilde{\boldsymbol{Z}}_i) - 1\right)^j m_{F_0}^{(j)}(1)\right.\right. \\
& \left.\left. + \frac{1}{p!}\left(\hat{\eta}_n(\tilde{\boldsymbol{Z}}_i) - 1\right)^p \bar{\theta}_{ni}^{(p)}\right]\right]
\end{aligned}$$

$$= \quad O(1) \tag{76}$$

in view of the assumption that $\sqrt{nh_n^*} \cdot (h_n^*)^p = \sqrt{n(h_n^*)^{2p+1}} = \sqrt{c} < \infty$.

Notice that the expectation in (76) does not depend on g, while K_i is nonzero only for those observations i such that $1 - \hat{\eta}_n(\boldsymbol{Z}_i) \leq h_n$. It follows that the bound in (76) is uniform in Ψ_n, i.e.,

$$\begin{aligned}
&\lim_{n\to\infty} \sqrt{nh_n^*} \\
&\cdot \sup_{\psi_n \in \Psi_n} \left| E_{\psi_n} \left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \left[\sum_{j=2}^{p-1} \frac{1}{j!} (\hat{\eta}_n(\boldsymbol{Z}_i) - 1)^j \, m_{F_0}^{(j)}(1) \right.\right.\right. \\
&\left.\left.\left. + \frac{1}{p!} (\hat{\eta}_n(\boldsymbol{Z}_i) - 1)^p \, \bar{\theta}_{ni}^{(p)} \right]\right]\right| \\
< \quad & \infty.
\end{aligned} \tag{77}$$

Next, consider that $\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right\| = O_p\left(n^{-1/2}\right)$ by Assumption 2, so there exists a constant $C_1 \in (0, \infty)$ such that $\sqrt{n}\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right\| \leq C_1$ with probability approaching one as $n \to \infty$. Let A_{n1} denote the event in which $\sqrt{n}\left\|\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\right\| \leq C_1$. Let $\mathcal{A}_{n1}$ be the σ-algebra generated by A_{n1}. We have for any sequence $\{\psi_n : \psi_n \in \Psi_n\}$ that

$$\begin{aligned}
& \sqrt{nh_n^*} \left| E_{\psi_n} \left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{X}_i^\top \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right) \middle| \mathcal{A}_{n1} \right] \right| \\
\leq \quad & \sqrt{h_n^*} E_{\psi_n} \left[\left\| \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \right)^{-1} \right\| \sum_{i=1}^n \|\boldsymbol{S}_i\| \, |K_i| \, \|\boldsymbol{X}_i\| \middle| \mathcal{A}_{n1} \right] \cdot C_1 \\
= \quad & O_p\left(\sqrt{h_n^*}\right) \\
= \quad & o_p(1),
\end{aligned} \tag{78}$$

where use has been made of the assumption that $E\left[\|\boldsymbol{X}_1\|\right] < \infty$, as well as of the uniform boundedness of $\hat{\eta}_n(\cdot)$ over all n, the boundedness of $K(\cdot)$ and the bounded convergence theorem. From (78) it follows that there exists a random variable $M_{n1} = O_p\left(\sqrt{h_n^*}\right)$ such that

$$\sqrt{nh_n^*} \sup_{\psi_n \in \Psi_n} \left| E_{\psi_n} \left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{X}_i^\top \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right) \middle| \mathcal{A}_{n1} \right] \right|$$

$$\leq \quad M_{n1}.$$

But $P_{\psi_n}[A_{n1}] \to 1$ as $n \to \infty$, so

$$\begin{aligned}
&\sqrt{nh_n^*} \sup_{\psi_n \in \Psi_n} \left| E_{\psi_n} \left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{X}_i^\top \left(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 \right) \right] \right| \\
= \quad & O\left(\sqrt{h_n^*}\right) \\
= \quad & o(1).
\end{aligned}$$

Next, note that the uniform boundedness of $\hat{\eta}_n(\cdot)$ over n and the boundedness of $K(\cdot)$ and $m_{F_0}^{(1)}(\cdot)$ imply via the bounded convergence theorem that

$$\sqrt{nh_n^*} E_{\psi_n} \left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \left(\zeta_{ni} - \zeta_i \right) \right] = o(1) \tag{79}$$

for all sequences $\{\psi_n : \psi_n \in \Psi_n\}$, where ζ_{ni} and ζ_i are as given above in (36) and (49), respectively.

The expectation in (79) does not depend upon the conditional joint density g of (U, V) given $\boldsymbol{X}$ and $\boldsymbol{Z}$, while K_i is nonzero only for those observations i where $1 - \hat{\eta}_n(\boldsymbol{Z}_i) \leq h_n$. It follows that

$$\sqrt{nh_n^*} \sup_{\psi_n \in \Psi_n} \left| E_{\psi_n} \left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top \right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \left(\zeta_{ni} - \zeta_i \right) \right] \right| = o(1). \tag{80}$$

Next, consider that the uniform boundedness of $\hat{\eta}_n(\cdot)$ over all n and the boundedness of $K(\cdot)$ imply that there exists constants $M_{n2}, M_{n3} < (0, \infty)$ not depending on ψ_n such that for every $\psi_n \in \Psi_n$,

$$\begin{aligned}
&\left| E_{\psi_n} \left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i \right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i \right] - \frac{1}{nh_n^*} E_{\psi_n} \left[\boldsymbol{e}_1^\top \sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i \right] \right| \\
\leq \quad & M_{n2} \cdot \frac{1}{h_n^*} \left| E_{\psi_n} \left[K \left(\frac{1}{h_n^*} \left(\eta_0(\boldsymbol{Z}_1) - 1 \right) \right) \zeta_1 \right] \right| \\
\leq \quad & M_{n3} (h_n^*)^p,
\end{aligned} \tag{81}$$

where use has been made of the assumptions that $E\left[\|\boldsymbol{X}_1\|\right] < \infty$, $E\left[U_1\right] = 0$, that U_1 and $\eta_0(\boldsymbol{Z}_1)$ are independent and that $K(\cdot)$ is a kernel of p-th order. Since (81) holds for every $\psi_n \in \Psi_n$, we find that

$$\begin{aligned} & \sup_{\psi_n \in \Psi_n} \left| E_{\psi_n}\left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i\right] - \frac{1}{nh_n^*} E_{\psi_n}\left[\boldsymbol{e}_1^\top \sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i\right]\right| \\ \leq \quad & M_{n3}\left(h_n^*\right)^p. \end{aligned} \tag{82}$$

A similar calculation shows that there exists a constant $M_{n4} \in (0, \infty)$ such that

$$\sup_{\psi_n \in \Psi_n} \left| \frac{1}{nh_n^*} E_{\psi_n}\left[\boldsymbol{e}_1^\top \sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i\right]\right| \leq M_{n4}\left(h_n^*\right)^p. \tag{83}$$

Combine (80), (82) and (83) with the assumption that $n(h_n^*)^{2p+1} = c < \infty$ to deduce that

$$\sqrt{nh_n^*} \sup_{\psi_n \in \Psi_n} \left| E_{\psi_n}\left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_{ni}\right]\right| = 0. \tag{84}$$

The desired bound, i.e., (74), follows from (77) and (84).

A.4.2 Proof of (75)

Arguments similar to those used in the proof of (74) based on the uniform boundedness of $\hat{\eta}_n(\cdot)$ over all n and on the boundedness of $K(\cdot)$ and of $m_{F_0}^{(j)}(\cdot)$ for each $j \in \{0, 1, \ldots, p\}$ show that

$$\begin{aligned} & nh_n^* \sup_{\psi_n \in \Psi_n} \Bigg| Var_{\psi_n}\left[\hat{\theta}_n^*\right] \\ & -E_{\psi_n}\left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i\right)\left(\sum_{i=1}^n \boldsymbol{S}_i^\top K_i \zeta_i\right)\right. \\ & \left. \cdot \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \boldsymbol{e}_1\right]\Bigg| \\ = \quad & o(1). \end{aligned} \tag{85}$$

Next, consider that the uniform boundedness of $\hat{\eta}_n(\cdot)$ over all n and the boundedness of $K(\cdot)$, and of $m_{F_0}^{(1)}(\cdot)$ imply that there exists constants $M_{n5}, M_{n6} \in (0, \infty)$ not depending on ψ_n such that for every sequence $\{\psi_n : \psi_n \in \Psi_n\}$:

$$
\begin{aligned}
& E_{\psi_n}\left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i\right)\left(\sum_{i=1}^n \boldsymbol{S}_i^\top K_i \zeta_i\right)\right. \\
& \quad \left.\cdot \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \boldsymbol{e}_1\right] \\
\leq & \ M_{n5} \cdot \frac{1}{n(h_n^*)^2} E_{\psi_n}\left[K^2\left(\frac{1}{h_n^*}(\eta_0(\boldsymbol{Z}_1) - 1)\right)\zeta_1^2\right] \\
\leq & \ M_{n6} \cdot \frac{1}{nh_n^*},
\end{aligned} \tag{86}
$$

where use has been made of the assumptions that $E_{\psi_n}\left[U_1^2\right] < \infty$ for all $\{\psi_n\}$, and that $\int K^2(u)du < \infty$.

It follows from (86) that

$$
\begin{aligned}
& nh_n^* \sup_{\psi_n \in \Psi_n} E_{\psi_n}\left[\boldsymbol{e}_1^\top \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \zeta_i\right)\left(\sum_{i=1}^n \boldsymbol{S}_i^\top K_i \zeta_i\right)\right. \\
& \quad \left.\cdot \left(\sum_{i=1}^n \boldsymbol{S}_i K_i \boldsymbol{S}_i^\top\right)^{-1} \boldsymbol{e}_1\right] \\
< & \ \infty.
\end{aligned} \tag{87}
$$

Combining (85) and (87) enables one to deduce that $nh_n^* \sup_{\psi_n \in \Psi_n} Var_{\psi_n}\left[\hat{\theta}_n^*\right] < \infty$. This completes the proof of (75).

References

Amemiya, T. (1984) 'Tobit models: A survey.' *Journal of Econometrics* 24, 3–61

Andrews, D. W. K. (1991) 'Asymptotic normality of series estimators for nonparametric and semiparametric regression models.' *Econometrica* 59, 307–345

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

Arabmazar, A., and P. Schmidt (1982) 'An investigation of the robustness of the Tobit estimator to non-normality.' *Econometrica* 50, 1055–1063

Bushway, S., B. D. Johnson, and L. Slocum (2007) 'Is the magic still there? The use of the Heckman two-step correction for selection bias in criminology.' *Journal of Quantitative Criminology* 23, 151–178

Collier, D., and J. Mahoney (1996) 'Insights and pitfalls: Selection bias in qualitative research.' *World Politics* 49, 56–91

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

_ (1996) *Local Polynomial Modelling and Its Applications* (Boca Raton, Florida: Chapman & Hall/CRC)

Gallant, A. R., and D. Nychka (1987) 'Semi-nonparametric maximum likelihood estimation.' *Econometrica* 55, 363–390

Goldberger, A. S. (1983) 'Abnormal selection bias.' In *Studies in Econometrics, Time Series and Multivariate Statistics: In commemoration of T. W. Anderson's 65th birthday,* ed. S. Karlin, T. Amemiya, and L. A. Goodman (San Diego, Calif.: Academic Press) pp. 67–84

Gronau, R. (1974) 'Wage comparisons: A selectivity bias.' *Journal of Political Economy* 82, 1119–1143

Han, A. (1987) 'Non-parametric analysis of a generalized regression model.' *Journal of Econometrics* 35, 303–316

Heckman, J. J. (1974) 'Shadow prices, market wages and labor supply.' *Econometrica* 42, 679–694

_ (1976) 'The common structure of statistical models of truncation, sample selection and limited dependent variables, and a simple estimator for such models.' *Annals of Economic and Social Measurement* 5, 475–492

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

_ (1990) 'Varieties of selection bias.' *American Economic Review* 80, 313–318

Helpman, E., M. Melitz, and Y. Rubinstein (2008) ‘Estimating trade flows: Trading partners and trading volumes.’ *Quarterly Journal of Economics* 123, 411–487

Horowitz, J. L. (1993) ‘Optimal rates of convergence of parameter estimators in the binary response model with weak distributional assumptions.’ *Econometric Theory* 9, 1–18

Ibragimov, I. A., and R. Z. Has’minskii (1981) *Statistical Estimation: Asymptotic Theory* (New York: Springer-Verlag New York). Translated by Samuel Kotz

Ichimura, H. (1993) ‘Semiparametric least squares (SLS) and weighted SLS estimation of single-index models.’ *Journal of Econometrics* 58, 71–120

Ichimura, H., and L.-F. Lee (1991) ‘Semiparametric least squares estimation of multiple index models: Single equation estimation.’ In *Nonparametric and semiparametric methods in econometrics and statistics: Proceedings of the Fifth International Symposium in Economic Theory and Econometrics,* ed. W. A. Barnett, J. Powell, and G. E. Tauchen (Cambridge, U.K.: Cambridge University Press) chapter 1, pp. 3–49

Jiménez, G., S. Ongena, J. L. Peydró, and J. Saurina (2014) ‘Hazardous times for monetary policy: What do twenty-three million bank loans say about the effects of monetary policy on credit risk-taking?’ *Econometrica* 82, 463–505

Khan, S., and D. Nekipelov (2013) ‘On uniform inference in nonlinear models with endogeneity.’ Unpublished paper. Department of Economics, Duke University

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

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

Lee, L.-F. (1983) ‘Generalized econometric models with selectivity.’ *Econometrica* 51, 507–512

Lennox, C. S., J. R. Francis, and Z. Wang (2012) ‘Selection models in accounting research.’ *Accounting Review* 87, 589–616

Lewbel, A. (1997) ‘Semiparametric estimation of location and other discrete choice moments.’ *Econometric Theory* 13, 32–51

_ (2007) ‘Endogenous selection or treatment model estimation.’ *Journal of Econometrics* 141, 777–806

Lewis, H. G. (1974) 'Comments on selectivity biases in wage comparisons.' *Journal of Political Economy* 82, 1145–1155

Little, R. J. A. (1982) 'Models for nonresponse in sample surveys.' *Journal of the American Statistical Association* 77, 237–250

Mulligan, C., and Y. Rubinstein (2008) 'Selection, investment, and women's relative wages over time.' *Quarterly Journal of Economics* 123, 1061–1110

Newey, W. K. (2009) 'Two-step series estimation of sample selection models.' *Econometrics Journal* 12, S217–S229

Oaxaca, R. (1973) 'Male-female wage differentials in urban labor markets.' *International Economic Review* 14, 693–709

Olsen, R. J. (1982) 'Distributional tests for selectivity bias and a more robust likelihood estimator.' *International Economic Review* 23, 223–240

Paarsch, H. J. (1984) 'A Monte Carlo comparison of estimators for censored regression models.' *Journal of Econometrics* 24, 197–213

Powell, J. L. (2001) 'Semiparametric estimation of censored selection models.' In *Nonlinear statistical modeling: Proceedings of the Thirteenth International Symposium in Economic Theory and Econometrics: Essays in Honor of Takeshi Amemiya,* ed. C. Hsiao, K. Morimune, and J. L. Powell (Cambridge, U.K.: Cambridge University Press) chapter 6, pp. 165–196

Powell, J. L., J. H. Stock, and T. M. Stoker (1989) 'Semiparametric estimation of index coefficients.' *Econometrica* 57, 1403–1430

Robinson, P. M. (1988) 'Root-*n*-consistent semiparametric regression.' *Econometrica* 56, 931–954

Rotnitzky, A., D. R. Cox, M. Bottai, and J. Robins (2000) 'Likelihood-based inference with singular information matrix.' *Bernoulli* 6, 243–284

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

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

_ (2000) 'Gender wage differences in Malaysia: Parametric and semiparametric estimation.' *Journal of Development Economics* 63, 351–378

_ (2004) 'Finite sample properties for the semiparametric estimation of the intercept of a censored regression model.' *Statistica Neerlandica* 58, 35–56

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

Stone, C. J. (1980) 'Optimal rates of convergence for non-parametric estimators.' *Annals of Statistics* 8, 1348–1360

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

Stute, W., and L.-X. Zhu (2005) 'Nonparametric checks for single-index models.' *Annals of Statistics* 33, 1048–1083

van der Vaart, A. W., and J. A. Wellner (1996) *Weak Convergence and Empirical Processes: With Applications to Statistics* (New York: Springer-Verlag New York)

_ (2007) 'Empirical processes indexed by estimated functions.' In *Asymptotics: Particles, Processes and Inverse Problems: Festschrift for Piet Groeneboom,* ed. E. A. Cator, G. Jongbloed, C. Kraaikamp, H. P. Lopuhaä, and J. A. Wellner (Beachwood, Ohio: Institute of Mathematical Statistics) pp. 234–252

Vella, F. (1998) 'Estimating models with sample selection bias: A survey.' *Journal of Human Resources* 33, 127–169

Winship, C., and R. D. Mare (1992) 'Models for sample selection bias.' *Annual Review of Sociology* 18, 327–350

Yang, S. S. (1981) 'Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function.' *Journal of the American Statistical Association* 76, 658–662

Zuehlke, T. W., and A. R. Zeman (1991) 'A comparison of two-stage estimators of censored regression models.' *Review of Economics and Statistics* 73, 185–188