Escanciano, J. C., & Zhu, L. (2014). A Simple Data-Driven Estimator for the Semiparametric Sample Selection Model. *Econometric Reviews*, 34 (6-10), pp. 734-762.

DOI: 10.1080/07474938.2014.956577

A Simple Data-Driven Estimator for the Semiparametric Sample Selection Model*

Juan Carlos Escanciano† *Indiana University*

Lin Zhu‡ *Tsinghua University*

May 22, 2013

Abstract

This paper proposes a simple fully data-driven version of Powell's (2001) two-step semiparametric estimator for the sample selection model. The main feature of the proposal is that the bandwidth used to estimate the infinite-dimensional nuisance parameter is chosen by minimizing the mean squared error of the fitted semiparametric model. We formally justify data-driven inference. We introduce the concept of asymptotic normality, uniform in the bandwidth, and show that the proposed estimator achieves this property for a wide range of bandwidths. The method of proof is different from that in Powell (2001) and permits straightforward extensions to other semiparametric or even fully nonparametric specifications of the selection equation. The results of a small Monte Carlo suggest that our estimator has excellent finite sample performance, comparing well with other competing estimators based on alternative choices of smoothing parameters.

Keywords: Two-step estimator; Semiparametric sample selection models; Empirical process theory.

JEL classification: C13, C14, C25

*We would like to thank the Editor, Essie Maasoumi, and two anonymous referees for various helpful suggestions and corrections that greatly improved the readability of the paper.

†Department of Economics, Indiana University, 105 Wylie Hall, 100 South Woodlawn Avenue, Bloomington, IN 47405–7104, USA. E-mail: jescanci@indiana.edu. Web Page: http://mypage.iu.edu/~jescanci/. Research funded by the Spanish Plan Nacional de I+D+I, reference number SEJ2007-62908.

‡School of Economics and Management, Tsinghua University, Beijing 100084, China. E-mail: zhulin@sem.tsinghua.edu.cn.

1 Introduction

Since the pioneering work by Heckman, sample selection models have achieved great success in many fields of social science. A large portion of subsequent researches have been focusing on developing robust alternative methods which relax the distributional restrictions or the functional form assumptions in the selection and outcome equations, leading to semiparametric estimators of parameters of interest; see Chapter 8 in Pagan and Ullah (1999) for an excellent review. Semiparametric methods require the choice of a smoothing parameter for estimating the nonparametric components, with optimal choices depending on the underlying data characteristics. This paper addresses the important, and yet unsolved, problem of fully data-driven inference in semiparametric estimation of sample selection models.

The estimator we consider is a fully data-driven version of the estimator proposed by Powell (2001). See also related estimators proposed by Robinson (1988) in the context of partially linear models. Our choice of estimator is driven by the ease of implementation and interpretation (the estimator is an ordinary least squares (OLS) estimator). This estimator is a two-step estimator. In a first step, we estimate the parameters of a single-index selection equation, and then we perform nonparametric projections to eliminate the nuisance selection correction term in the second step, before estimating the parameters of interest by a simple OLS regression of the projected quantities. Alternative semiparametric approaches in the literature approximate the selection correction function using series estimators. For example, Cosslett (1991) constructed a semiparametric analogue of Heckman's two-step estimator using piecewise-constant functions to approximate the selection correction function in the second step, whereas Newey (2009) proposed a two-step series estimation method via approximating the correction function with power series or spline series.

Semiparametric methods require a choice of smoothing parameters, namely, bandwidth in the nonparametric kernel estimation and the number of series in the series approximation. Powell (2001) and Newey (2009) suggest cross-validation methods to choose the optimal smoothing parameter, where the cross-validation is implemented to minimize the nonparametric regression error in the intermediate steps. Following Härdle, Hall and Ichimura (1993), we propose to choose the bandwidth by minimizing the mean squared error in the regression of projected quantities, and we provide a formal justification of this choice. Due to the simple form of our estimator, our data-driven choice of bandwidth does not incur much computation burden and is easy to implement.

The method of proof that we use to justify the data-driven bandwidth and the asymptotic distribution of the estimator is different from that in Powell (2001). The motivation for this alternative method of proof is that his method of proof requires the verification of an expan-

sion for the data-driven bandwidth choice that is difficult, if not impossible, to obtain, see e.g. Härdle, Hall and Ichimura (1993) and Hong (1999) for related expansions. In addition to being technically challenging, this approach has the drawback of requiring a different analysis for each specific model considered. In contrast, the new method of proof is applicable more generally to modifications and extensions of the semiparametric sample selection models, without the need to establish asymptotic expansions for data-driven bandwidth choices. It can be applied with simple modifications to related models with semiparametric or nonparametric generated regressors, including the models investigated in Ichimura and Lee (1991), Ahn and Powell (1993), Li and Wooldridge (2002) and Das, Newey and Vella (2003), among many others. We provide a uniform expansion for density-weighted empirical processes, adapting and extending to our setting recent empirical processes techniques developed in Escanciano, Jacho-Chávez and Lewbel (2012, henceforth EJL). The tools that we develop for proving our results are of independent interest, and have applications in other problems such as the development of nonparametric significance tests as studied in e.g. Delgado and González-Manteiga (2001).

The rest of the paper is organized as follows. Section 2 presents the semiparametric selection model and discusses identification and estimation. We derive the asymptotic distribution of our data-driven estimator in Section 3. Section 4 reports the results of a small Monte Carlo study to examine the finite sample performance of our estimator and Section 5 concludes and discusses further research. Mathematical proofs and technical results are relegated to an Appendix at the end of the paper.

2 A Simple Data-Driven Estimator for the Sample Selection Model

2.1 General Model and Identification

Let $Y^* = X'\theta_0 + \varepsilon$ be a latent outcome equation, where X is a d_X-dimensional vector of covariates, θ_0 is an unknown parameter in $\Theta \subset \mathbb{R}^{d_X}$, and ε is an unobservable error. Henceforth, A' denotes the transpose of A. We do not observe Y^* but instead observe (Y, D, X') where $Y = Y^*D = (X'\theta_0 + \varepsilon)D$, implying that Y^* is only observed when $D = 1$. Suppose $D = \mathbb{I}[s_0(Z) - u > 0]$ where $\mathbb{I}[A]$ denotes the indicator of the event A, i.e. =1 if A holds and =0 otherwise, the conditional distribution of u given ε is continuous and Z is a d_Z-dimensional observable vector that includes X and possibly other variables (hence $d_Z \geq d_X$). Nonrandom selection arises because the unobserved errors ε and u, though assumed independent of Z, are correlated with each other. We also assume that F_u, the cdf of u, is strictly

increasing. Henceforth, to simplify the notation, unless otherwise stated, all the conditional means are conditional on $D = 1$.

The semiparametric sample selection model is specified as

$$E(Y|Z) = X'\theta_0 + \lambda(g_0(Z)), \tag{1}$$

by noticing that $E(\varepsilon|Z, D = 1) = E(\varepsilon|Z, F_u(u) < F_u[s_0(Z)]) =: \lambda(g_0(Z))$, where $g_0(Z) := F_u[s_0(Z)]$ and F_u is marginal distribution function of u. In turn, by the law of iterated expectations, this implies that

$$E(Y|g_0(Z)) = E(X'|g_0(Z))\theta_0 + \lambda(g_0(Z)). \tag{2}$$

Taking the difference of (1) and (2), we obtain

$$E(Y_{g_0}|Z) = X'_{g_0}\theta_0, \tag{3}$$

where $Y_g := Y - E(Y|g(Z))$ and $X_g := X - E(X|g(Z))$. Then, these simple arguments show that θ_0 (and hence, λ) is identified under the following condition:

Assumption I: $V := E\left(DX_{g_0}X'_{g_0}\right)$ *is positive definite (pd).*

The singularity of V is equivalent to the existence of a measurable f such that $\alpha'X = f(g_0(Z))$, for some $\alpha \in \mathbb{R}^{d_X}$, $\alpha \neq 0$. In general we do not need exclusion restrictions to assure non-singularity of V, but for the commonly used single-index specification $g_0(Z) = F_u(Z'\gamma_0)$, the exclusion restriction is indeed necessary. To see this, suppose $Z = X$ and $s_0(Z) \equiv s_0(X) = X'\gamma_0$, then since F_u is assumed invertible, by choosing f to be the quantile of F_u and $\alpha = \gamma_0$, we obtain $\gamma'_0 X = f(g_0(X))$. Our identification assumption for the single-index selection model is the same as that of Powell (2001) and Newey (2009), among others. Under Assumption I, θ_0 is identified as

$$\theta_0 = \left(E\left[X_{g_0}X'_{g_0}\right]\right)^{-1} E\left[X_{g_0}Y_{g_0}\right].$$

This identification scheme also indicates the form of the estimator. For example, the conditional expectation terms can be replaced by corresponding kernel estimators and expectation terms can be replaced by sample expectations. However, the first-stage fully nonparametric estimation suffers from the curse of dimensionality as usual. To sidestep this problem, we focus our attention on the linear single-index selection equation which practitioners would commonly use for empirical applications of moderate sample size. For this purpose, we im-

pose the necessary exclusion restriction as discussed above, that is, we require $d_Z > d_X$. This necessary condition together with some other mild conditions would suffice to verify the identification assumption, see Newey (2009) for more discussion.

Henceforth, we assume $s_0(Z) = Z'\gamma_0$, so that $g_0(Z) = F_u(Z'\gamma_0)$ and by strict monotonicity of $F_u(\cdot)$, we have $E[X|g_0(Z)] = E[X|Z'\gamma_0]$ and $E[Y|g_0(Z)] = E[Y|Z'\gamma_0]$. Define $X_\gamma := X - E[X|Z'\gamma]$ and $X_0 := X - E[X|Z'\gamma_0]$, and similarly $Y_\gamma := Y - E[Y|Z'\gamma]$ and $Y_0 := Y - E[Y|Z'\gamma_0]$. With this notation, the parameter is identified as $\theta_0 = (E[X_0X_0'])^{-1} E[X_0Y_0]$. In the next section we investigate a data-driven OLS estimator for θ_0.

2.2 Data-Driven Semiparametric OLS Estimation

Given identification, we now describe in detail our data-driven estimator for the sample selection model. We assume that a random sample $\{(Y_i, D_i, Z_i)\}_{i=1}^n$ is observed from the joint distribution of $(Y, D, Z) \in \mathbb{R}^{2+d_Z}$. Hereafter, $\mathcal{X}_\xi$ denotes the support of a generic random vector ξ. Suppose that we obtained a $\sqrt{n}$-consistent estimator of $\widehat{\gamma}$ for γ_0 in the first stage, which can be done, for example, by Klein and Spady's (1993) quasi-MLE method, Powell, Stock and Stoker's (1989) density-weighted estimation method or Ichimura's (1993) Semiparametric Least Squares (SLS) estimator. Let $\Gamma \subset \mathbb{R}^{d_Z}$ be an arbitrary neighborhood of γ_0, and define the class of functions $\mathcal{W} := \{z \to z'\gamma : \gamma \in \Gamma\}$ and the set $\mathcal{X}_\mathcal{W} := \{z'\gamma : z \in \mathcal{X}_Z \text{ and } \gamma \in \Gamma\}$. We estimate X_0 and Y_0 by kernels. For γ in a neighborhood of γ_0, denote $f(\cdot|\gamma)$ as the density function of $Z'\gamma$ conditional on $D = 1$ and, for $w \in \mathcal{X}_\mathcal{W}$,

$$\begin{aligned} \mu_X(w|\gamma) &:= E[X|Z'\gamma = w, D = 1], \\ \mu_Y(w|\gamma) &:= E[Y|Z'\gamma = w, D = 1]. \end{aligned}$$

The regression function $\mu_X(w|\gamma)$ can be consistently estimated by the nonparametric Nadaraya-Watson kernel estimator

$$\begin{aligned} \widehat{\mu}_X(w|\gamma) &:= \widehat{T}_X(w|\gamma)/\widehat{f}(w|\gamma), \\ \widehat{T}_X(w|\gamma) &:= \frac{1}{\widehat{\pi}n}\sum_{j=1}^n D_j X_j K_{\widehat{h}_n}(Z_j'\gamma - w), \\ \widehat{f}(w|\gamma) &:= \frac{1}{\widehat{\pi}n}\sum_{j=1}^n D_j K_{\widehat{h}_n}(Z_j'\gamma - w), \end{aligned}$$

where $\widehat{\pi} := n^{-1}\sum_{i=1}^n D_i$, $K_h(t) := h^{-1}K(t/h)$ is a kernel function and $\widehat{h}_n$ denotes a possibly data-dependent bandwidth parameter. The same estimators are also constructed for

$\mu_Y(w|\gamma)$, denoted as $\widehat{\mu}_Y(w|\gamma)$.

To deal with the random denominator problem in the kernel estimation, we can either introduce a trimming sequence, as in e.g. Robinson (1988), or employ the density-weighted method, as in e.g. Li and Wooldridge (2002). In this paper, we consider for simplicity the density-weighted estimator, but results for randomly trimmed observations can be obtained by adapting the results of EJL. The population version of the density-weighted equation is

$$f_0(Z'\gamma_0) E[Y_0|Z] = f_0(Z'\gamma_0) X_0'\theta_0,$$

where $f_0(\cdot) = f(\cdot|\gamma_0)$ is the (Lebesgue) density function of $Z'\gamma_0$ conditional on $D = 1$. The parameter θ_0 is equivalently identified as

$$\theta_0 = E\left[f_0^2(Z'\gamma_0) X_0 X_0'\right]^{-1} E\left[f_0^2(Z'\gamma_0) X_0 Y_0\right].$$

The proposed density-weighted estimator takes the simple OLS form: for a given bandwidth $h \equiv \widehat{h}_n$ used in these kernel estimates,

$$\widehat{\theta}(h) = \left(\frac{1}{n}\sum_{i=1}^n D_i \widehat{f}_{\widehat{\gamma}i}^2(h) \widehat{X}_{\widehat{\gamma}i}(h) \widehat{X}_{\widehat{\gamma}i}'(h)\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n D_i \widehat{f}_{\widehat{\gamma}i}^2(h) \widehat{X}_{\widehat{\gamma}i}(h) \widehat{Y}_{\widehat{\gamma}i}(h)\right),$$

where $\widehat{f}_{\widehat{\gamma}i}(h) := \widehat{f}(Z_i'\widehat{\gamma}|\widehat{\gamma})$, $\widehat{\mu}_{Xi}(h) := \widehat{\mu}_X(Z_i'\widehat{\gamma}|\widehat{\gamma})$ and $\widehat{X}_{\widehat{\gamma}i}(h) := X_i - \widehat{\mu}_{Xi}(h)$ in which h indicates the bandwidth used in the kernel estimation. Similarly, $\widehat{\mu}_{Yi}(h) := \widehat{\mu}_Y(Z_i'\widehat{\gamma}|\widehat{\gamma})$ and $\widehat{Y}_{\widehat{\gamma}i}(h) := Y_i - \widehat{\mu}_{Yi}(h)$. Powell (2001) discusses asymptotic theory for $\widehat{\theta}(h)$, but did not address the challenging problem of data-driven choice of h, which is the main purpose of this paper.

Our theory is flexible and allows for plug-in bandwidths with deterministic rates or fully data-dependent bandwidths, for example, obtained by cross-validation, as long as Assumption 5 below is satisfied. However, the finite sample performance could be different by choosing different bandwidths. Here, we propose to choose the bandwidth that minimizes the mean squared error:

$$\widehat{h}_n^* = \arg\min_{a_n \le h \le b_n} \frac{1}{n}\sum_{i=1}^n \left(\widehat{Y}_{\widehat{\gamma}i}(h) - \widehat{X}_{\widehat{\gamma}i}(h)'\widehat{\theta}^{(i)}(h)\right)^2$$

where the bandwidth selection range $[a_n, b_n]$ satisfies Assumption 5 below and $\widehat{\theta}^{(i)}(h)$ is the leave-one-out version of $\widehat{\theta}(h)$.

We want to emphasize that our choice of bandwidth is directly targeted to minimize the mean squared error, as opposed to other data-dependent choices that are obtained by

optimizing the nonparametric estimation in the intermediate steps, see Härdle, Hall and Ichimura (1993) for more discussion and Hong (1999) for other proposals for choosing $\widehat{h}_n^*$ in the related partially linear model. Due to the closed-form OLS estimator, we do not undergo much computation burden, and the optimization problem can be carried out by a grid-search method. After obtaining the optimal bandwidth $\widehat{h}_n^*$, our data-driven estimator is given by $\widehat{\theta}(\widehat{h}_n^*)$. We establish the asymptotic properties of the data-driven estimator in the next section.

3 Asymptotic Properties of Data-Driven Estimates

There are several methods available for establishing the asymptotic properties of semiparametric estimators with random bandwidths. Recently, Li and Li (2010) have provided sufficient conditions for the first-order asymptotic properties of a large class of kernel-based semiparametric estimators to hold with data dependent bandwidths. Their method of proof requires one to use an estimated bandwidth first with a 'rule-of-thumb' asymptotic representation, i.e. a constant term times a known power of the sample size, and then establish the stochastic equicontinuity of the estimator with respect to this constant term. Similar treatments can be found in Boente and Fraiman (1995) and Martins-Filho and Saraiva (2012), among others. Powell (2001, p. 184) proposed an alternative method of proof, which exploits the index structure of the selection equation and it is based on the assumption that for the random bandwidth $\widehat{h}_n^*$, there exists a deterministic sequence h_n^* satisfying certain convergence conditions. Specifically, it must hold that

$$\sqrt{n}\left(\frac{\widehat{h}_n^*}{h_n^*} - 1\right) = O_P(1). \tag{4}$$

However, verifying (4) or obtaining a 'rule-of-thumb' asymptotic representation for $\widehat{h}_n^*$ is a challenging problem, and is closely related to the subject of a rather technical paper by Härdle, Hall and Ichimura (1993). Moreover, each different data-driven bandwidth choice would require a separate analysis. Here we consider a different approach for which there is no need to establish stochastic equicontinuity, 'rule-of-thumb' asymptotic representations, or representations such as (4). Instead, we obtain uniform representations for $\widehat{\theta}(h)$ over sets of admissible bandwidths, which include estimated bandwidths with 'rule-of-thumb' asymptotic representations as a special case. A useful by-product of our proposed method is that bandwidth choice procedures, including $\widehat{h}_n^*$, can be readily justified without further calculations under our assumptions. This method of proof has been suggested first by EJL,

but for a different class of models that does not nest the one we consider here.[1] Hence, the required analysis and proofs are different from those in EJL. Uniform in bandwidth consistency has been developed in the statistics literature for classical nonparametric kernel estimates. Einmahl and Mason (2005) studied uniform in bandwidth consistency of kernel estimators and Dony and Mason (2008) obtained similar results for a class of conditional U-statistics. We contribute to this literature in statistics by providing results for nonparametric kernel estimators with generated regressors of possibly unbounded support, and showing how these results can be applied to our semiparametric estimator.

Now, we list the assumptions needed for developing asymptotic distribution of our estimator as follows. We assume that Z includes a continuous random variable, say Z_1, which is associated with nonzero coefficient and normalized to 1 for identification purposes, i.e. $Z'\gamma = Z_1 - Z_2'\gamma_2$. For notational simplicity, we identify γ_2 with γ in what follows. Let $|\cdot|$ denote the Euclidean norm.

Assumption 1: *The sample observations* $\{Y_i, D_i, Z_i\}_{i=1}^n$ *are a sequence of independent and identically distributed (iid) variables, distributed as* $\{Y, D, Z\}$, *and* X_i *is a subvector of* Z_i *with* $E\left[|Z_i|^{s_Z}\right] < \infty$ *for some* $s_Z > 4$. *Furthermore,* $\sup_{z\in\mathcal{X}_Z} E\left[|Y|^{s_Y}\,|Z = z\right] \leq C$ *for some* $s_Y > 2$, and $\pi := \Pr(D = 1) > 0$.

Assumption 2: *The parameter space* Θ *is a compact subset of* $\mathbb{R}^{d_X}$ *and* θ_0 *is an element of its interior. The parameter space for the selection equation* $\Gamma \subset \mathbb{R}^{d_Z - 1}$ *is compact.*

Assumption 3: *For* $\xi = X$ *and* $\xi = Y$ *we assume that the conditional density* $f_{(\xi,Z_1)|(Z_2,D=1)}(\cdot)$ *of* (ξ, Z_1) *given* $Z_2 = z_2$ *and* $D = 1$ *is three times continuously differentiable with bounded derivatives for all* $z_2 \in \mathcal{X}_{Z_2}$.

Assumption 4: *The kernel function* $K(t) : \mathbb{R} \to \mathbb{R}$ *is bounded, symmetric, continuously differentiable, and satisfies the following conditions:* $\int K(t)\,dt = 1$, $\int t^2 K(t)\,dt = 0$ *and* $\int t^3 K(t)\,dt < \infty$; $\partial K(t)/\partial t$ *is bounded and for some* $\nu > 1$, $|\partial K(t)/\partial t| \leq C\,|t|^{-\nu}$ *for* $|t| > L$, $0 < L < \infty$.

Assumption 5: *The possibly data-dependent bandwidth* $\widehat{h}_n$ *satisfies* $\Pr\left(a_n \leq \widehat{h}_n \leq b_n\right) \to 1$ *as* $n \to \infty$, *where* $\{a_n, b_n\}$ *are deterministic sequences of positive numbers such that:* (i) $a_n \to 0$, $na_n^{5/2}/\ln n \to \infty$; (ii) $nb_n^6 \to 0$.

Assumption 6: *The first-stage estimator* $\widehat{\gamma}$ *has the asymptotic representation* $\sqrt{n}(\widehat{\gamma} - \gamma_0) = n^{-1/2}\sum_{i=1}^n \xi_i + o_P(1)$, *for a sequence* $\{\xi_i\}_{i=1}^n$ *of iid variables with finite variance.*

[1] EJL deal with weighted sample means of nonparametric residuals, with weights that are functions of covariates. In our paper, the estimator studied cannot be written in terms of these sample means.

The moment and compactness conditions in Assumptions 1 and 2 are standard in the literature. The uniform boundedness of conditional moments of Y given X can be relaxed. Assumption 3 provides primitive conditions such that $Z'\gamma$ is a continuous random variable for γ close to γ_0 and such that $f(w|\gamma)$, $\mu_X(w|\gamma)$, $\mu_Y(w|\gamma)$ and related functions are smooth in w and γ. Unlike the results in EJL, ours allow for regressors with unbounded support. Assumption 4 and 5 impose conditions on the kernel function and rate conditions for the bandwidth. We allow for stochastic bandwidths with optimal rates $n^{-1/5}$. The rate condition $na_n^{5/2}/\ln n \to \infty$ is used to prove that the infinite-dimensional nuisance parameters belong to a certain class of nonparametric functions with probability tending to one as the sample size increases; see Lemma P1 in the Appendix. An alternative is to require the latter as a high-level assumption, as typically done in the literature, and to require the weaker rate condition $na_n/\ln n \to \infty$, but we prefer to provide simple low-level primitive assumptions (although they might not be necessary). We use the same kernel and bandwidth for all estimators involved, but this could be relaxed without affecting our asymptotic results. Assumption 6 restricts the convergence rate of the first-stage estimator and assumes a linear asymptotic representation necessary to develop the asymptotic distribution of the proposed estimator. There are estimators for semiparametric binary choice model satisfying this condition, for example, Klein and Spady (1993)'s quasi-likelihood estimator and Ichimura (1993)'s semiparametric least squares estimator, among others. In a partial linear model setup with generated regressors, Li and Wooldridge (2002) imposes a similar condition for the first-stage estimator.

To simplify notations, let $f_{0i} := f(Z_i'\gamma_0|\gamma_0)$, $Y_{0i} := Y_i - \mu_Y(Z_i'\gamma_0|\gamma_0)$ and $X_{0i} := X_i - \mu_X(Z_i'\gamma_0|\gamma_0)$, also let $e_i := Y_{0i} - X_{0i}'\theta_0$, and define

$$\omega_i := e_i D_i f_{0i}^2 X_{0i} - B\xi_i,$$

where $B := E[D_i f_{\gamma_0 i}^2 \{\partial_w \mu_{Y_0 i} - (\partial_w \mu_{X_0 i}'\theta_0)\} X_{0i} Z_i']$, $\partial_w \mu_{Y_0 i} := \partial \mu_Y(w|\gamma_0)/\partial w|_{w=Z_i'\gamma_0}$ and $\partial_w \mu_{X_0 i} := \partial \mu_X(w|\gamma_0)/\partial w|_{w=Z_i'\gamma_0}$.

Assumption 7: $E\left[|\omega_i|^2\right] < \infty$.

We introduce the concept of asymptotic normality, uniformly in the bandwidth.

Definition: *Let* $Z_n(h_n) = \sqrt{n}\left(\widehat{\theta}(h_n) - \theta_0\right)$ *with cdf* $F_n(z;h_n)$, *and let* Z *be a normal random vector with zero mean and cdf* F. *Then, we say that* $\widehat{\theta}(h_n)$ *is asymptotically normal uniformly in the bandwidth at* $[a_n, b_n]$ *if*

$$\sup_{a_n \le h_n \le b_n} \sup_{z \in \mathbb{R}^{d_X}} |F_n(z;h_n) - F(z)| = o(1).$$

We are not aware of any formal justification of a semiparametric estimator satisfying the previously introduced uniform version of asymptotic normality. We establish this property for Powell's (2001) sample selection estimator in the next theorem.

Theorem: *Under Assumptions I, and 1-7,* $\widehat{\theta}(h_n)$ *is asymptotically normal uniformly in* $a_n \leq h_n \leq b_n$*, i.e.*

$$\sqrt{n}\left(\widehat{\theta}(h_n) - \theta_0\right) \longrightarrow_d N\left(0, \Sigma^{-1}\Omega\Sigma^{-1}\right)$$

where $\Sigma = E\left[D_i f_{0i}^2 X_0 X_{0i}'\right]$ *and* $\Omega = E\left[\omega_i \omega_i'\right]$.

Remark 1 *Estimation of the asymptotic covariance matrix is discussed in Powell (2001), and can be carried out by consistently estimating unknown parameters in the representation above. Namely,* Σ *can be estimated by*

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n} D_i \widehat{f}_{\widehat{\gamma}i}^2 \widehat{X}_{\widehat{\gamma}i} \widehat{X}_{\widehat{\gamma}i}',$$

and Ω *can also be estimated by the sample analog*

$$\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} \widehat{\omega}_i \widehat{\omega}_i',$$

where $\widehat{\omega}_i = \left(\widehat{Y}_{\widehat{\gamma}i} - \widehat{X}_{\widehat{\gamma}i}'\widehat{\theta}\right) D_i \widehat{f}_{\widehat{\gamma}i}^2 \widehat{X}_{\widehat{\gamma}i} - \widehat{B}\widehat{\xi}_i$ *with* $\widehat{B}$ *the sample analog of* B *and* $\widehat{\xi}_i$ *a consistent estimator for the asymptotic influence function of* $\widehat{\gamma}$*, see Newey (2009).*

Remark 2 *As an immediate consequence of the above theorem, we can conclude that*

$$\sqrt{n}\left(\widehat{\theta}\left(\widehat{h}_n^*\right) - \theta_0\right) \longrightarrow_d N\left(0, \Sigma^{-1}\Omega\Sigma^{-1}\right).$$

4 Monte Carlo Experiments

In this section, we examine the finite sample performance of our estimator via a Monte Carlo study and compare our estimator with Newey's estimator using series approximation. Newey (2009) proposed to use series approximation to deal with the sample selection bias term, and he gave rate conditions for the number of approximating terms. For comparison, we report the simulation results using different number of series $K \in \{3, 6, 9\}$. Newey's estimator is defined as

$$\widehat{\theta}_{NWY}(K) = \left[X_D'(I_n - \widehat{Q}_K)X_D\right]^{-1} X_D'\left(I_n - \widehat{Q}_K\right) Y_D,$$

where $X = (D_1X_1, \ldots, D_nX_n)'$, $Y = (D_1Y_1, \ldots, D_nY_n)'$, I_n is $n \times n$ identity matrix, $\widehat{Q}_K$ is the projection matrix $\widehat{Q}_K = \widehat{P}_K \left(\widehat{P}_K'\widehat{P}_K\right)^{-1} \widehat{P}_K'$ with $\widehat{P}_K = (D_1\widehat{p}_1, \ldots, D_n\widehat{p}_n)'$ and $\widehat{p}_i = (p_{1K}(Z_i'\widehat{\gamma}), \ldots, p_{KK}(Z_i'\widehat{\gamma}))'$, here $(p_{1K}(\cdot), \ldots, p_{KK}(\cdot))'$ is the vector of K series functions used to approximate the selection bias.

We draw random samples of size $n \in \{250, 500, 1000\}$ in 1000 replications. The data are generated from the model:

$$\begin{aligned} Y &= (X_1\theta_0 + \varepsilon) D \\ D &= \mathbb{I}(X_1\gamma_1 + X_2\gamma_2 - u \geq 0) \end{aligned}$$

where X_1, X_2 are independently sampled from standard normal distribution. The error terms (ε, u) are generated, independent of (X_1, X_2), from two sets of distributions: (1) bivariate normal distribution; (2) bivariate t-distribution with 3 degrees of freedom. The dependence between ε and u are measured by correlation coefficient ρ taking values in $\{0.1, 0.5, 0.9\}$. The true parameter values in the simulation are $\theta_0 = \gamma_1 = \gamma_2 = 1$.

Tables 1-3 report Bias, StdDev (standard deviation), RMSE (root mean squared error) and MAE (mean absolute deviation) for both estimators, $\widehat{\theta}_{NWY}(K)$ and $\widehat{\theta}\left(\widehat{h}_n^*\right)$, with $\rho =$ 0.1, 0.5 and 0.9, corresponding to low, moderate and high levels of selection, respectively. For implementing our estimator we choose $a_n = \hat{\sigma}_s n^{-2/5+\epsilon}$ and $b_n = \hat{\sigma}_s n^{-1/6-\epsilon}$, where $\epsilon =$ 0.01 and $\hat{\sigma}_s$ is the sample standard deviation of the index $\{Z_i'\widehat{\gamma}\}_{i=1}^n$. Strictly speaking, our assumption requires deterministic a_n and b_n, but it can be shown that our results are still valid with these choices, since $\hat{\sigma}_s$ is bounded in probability. For all cases, our estimator exhibits better performance in terms of root mean squared error, especially for the cases with small sample size and larger selection bias (measured by correlation between ε and u). For some cases, our estimator also shows smaller bias, for example, when $n = 250$ and $\rho = 0.1$. As sample size increases, our estimator shows similar performance as Newey's series estimator except that the convergence of bias is slower than that of Newey's series estimator.

In the same simulation setup as above, we also compare our data-driven estimator with a cross-validation version of Powell (2001)'s estimator. As we discussed in the introduction, cross-validated bandwidth is chosen to minimize nonparametric regression error in the intermediate steps, while our data-driven method is directly targeted at minimizing the mean squared regression error of the main equation. Hence, we would expect better performance of our estimator. The simulation results also confirm this point. In Tables 4-6, our estimator clearly dominates the cross-validation version, especially in terms of bias when the sample size is small. Even with large sample size, the difference is still notable.

5 Conclusion

We propose a simple fully data-driven two-stage semiparametric estimator for the sample selection model. Unlike other data-driven choices of bandwidth, such as cross-validation or plug-in bandwidth, the bandwidth we propose uses the same criteria as the semiparametric estimator, and hence, it is expected to deliver a better performance. It minimizes the mean squared error in the fitted semiparametric regression, after projecting out the selection equation from the outcome equation. We formally justify our data-driven choice of bandwidth by employing uniform-in-bandwidth arguments. Our method of proof can be readily extended to account for nonparametric selection equation or other semiparametric generated regressors with simple modifications. The asymptotic distribution of our estimator is also derived and a small Monte Carlo study shows its finite-sample performance.

The optimality properties of the proposed bandwidth have not been investigated, and will be the subject of future research. Optimal rates for bandwidths can be incorporated in the construction of the data-driven estimator, which may result in simpler implementations of the proposed procedure, since only a constant needs to be estimated. We will explore such simplifications in the future.

6 Appendix A

6.1 Notation

This section introduces some notations and preliminary results on empirical processes theory. For a measurable class of functions $\mathcal{G}$ from $\mathbb{R}^p$ to $\mathbb{R}$, let $\|\cdot\|$ be a generic pseudo-norm on $\mathcal{G}$, defined as a norm except for the property that $\|f\| = 0$ does not necessarily imply that $f \equiv 0$. Given two functions l, u, a bracket $[l, u]$ is the set of functions $f \in \mathcal{G}$ such that $l \leq f \leq u$. An ε-bracket with respect to $\|\cdot\|$ is a bracket $[l, u]$ with $\|l - u\| \leq \varepsilon$, $\|l\| < \infty$ and $\|u\| < \infty$ (note that u and l not need to be in $\mathcal{G}$). The *covering number with bracketing* $N_{[\cdot]}(\varepsilon, \mathcal{G}, \|\cdot\|)$ is the minimal number of ε-brackets with respect to $\|\cdot\|$ needed to cover $\mathcal{G}$. These definitions are extended to classes taking values in $\mathbb{R}^d$, with $d > 1$, by taking the maximum of the bracketing numbers of the coordinate classes. Let $N(\varepsilon, \mathcal{G}, \|\cdot\|)$ denote the *covering number with respect to* $\|\cdot\|$, i.e., the minimal number of ε-balls with respect to $\|\cdot\|$ needed to cover $\mathcal{G}$. For a Borel measure μ define $\|g\|_{L_2,\mu} := \left(\int_{\mathcal{X}} g^2 d\mu\right)^{1/2}$. When $\mu = P$, the underlying probability measure of the data, we simply write $\|g\|_{L_2} \equiv \|g\|_{L_2,P}$. Let $a \vee b := \max(a, b)$.

Define for any vector $a = (a_1, \ldots, a_q)$ of q non-negative integers the differential operator $\partial_\xi^a := \partial^{|a|_1} / \partial \xi_1^{a_1} \ldots \partial \xi_q^{a_q}$, where $|a|_1 := \sum_{i=1}^q a_i$. For any smooth function $g : \mathcal{X}_\xi \subseteq \mathbb{R}^q \to \mathbb{R}$ and some $\eta > 0$, let $\underline{\eta}$ be the largest integer smaller than η, and

$$\|g\|_{\infty,\eta} := \max_{|a|_1 \leq \underline{\eta}} \sup_{\xi \in \mathcal{X}_\xi} \left|\partial_\xi^a g(\xi)\right| + \max_{|a|_1 = \underline{\eta}} \sup_{\xi_1 \neq \xi_2} \frac{\left|\partial_\xi^a g(\xi_1) - \partial_\xi^a g(\xi_2)\right|}{|\xi_1 - \xi_2|^{\eta - \underline{\eta}}},$$

where $|\cdot|$ is the Euclidean norm. Further for $\beta \in \mathbb{R}$, let $C_M^\eta\left(\mathcal{X}_\xi, \langle\xi\rangle^\beta\right) := \left\{g : \left\|g \cdot \langle\xi\rangle^\beta\right\|_{\infty,\eta} \leq M\right\}$ be a bounded subset of the weighted Hölder space $C^\eta\left(\mathcal{X}_\xi, \langle\xi\rangle^\beta\right) := \left\{g : \left\|g \cdot \langle\xi\rangle^\beta\right\|_{\infty,\eta} < \infty\right\}$, where $\langle\xi\rangle^\beta = \left(1 + |\xi|^2\right)^{\beta/2}$. Notice that in this definition we allow the function spaces to have unbounded support, i.e. $\mathcal{X}_\xi = \mathbb{R}^q$ is possible. The entropy rates of this function class have been established by Nickl and Pötscher (2007). For the special case $\beta = 0$, with the corresponding function space denoted as $C_M^\eta(\mathcal{X}_\xi)$, Nickl and Pötscher (2007) showed that given the assumption $\|\langle\xi\rangle^s\|_{L_2,\mu} < \infty$ where $s > 0$ and $s \neq \eta$, the L_2-bracketing entropy satisfies

$$\log N_{[\cdot]}\left(\varepsilon, C_M^\eta(\mathcal{X}_\xi), \|\cdot\|_{L_2,\mu}\right) \leq C\varepsilon^{-q/\min(\eta,s)}. \tag{5}$$

Let $\Gamma \subset \mathbb{R}^{d_Z}$ be a compact subset that contains γ_0. We now introduce a class of functions which serves as the parameter space for the functions f, μ_X and μ_Y. For a given function $\overline{\phi}(z)$ with finite L_2-norm, i.e. $\left\|\overline{\phi}\right\|_{L_2} < \infty$, let $\mathcal{T}_M^\eta\left(\overline{\phi}\right)$ be a class of measurable functions defined

on $\mathcal{X}_Z$, $\{z \to q(z'\gamma|\gamma) : \gamma \in \Gamma\}$ such that for a universal constant C_L, and all $\gamma_1, \gamma_2 \in \Gamma$,

$$\|q(Z'\gamma_1|\gamma_1) - q(Z'\gamma_2|\gamma_2)\|_{L_2(\overline{\phi})} \leq C_L |\gamma_1 - \gamma_2|,$$

where the weighted L_2-pseudonorm $\|\cdot\|_{L_2(\overline{\phi})}$ is defined as $\|h\|_{L_2(\overline{\phi})} = \left(E\left[h^2(Z)\overline{\phi}^2(Z)\right]\right)^{1/2}$. Moreover, we assume that for each $\gamma \in \Gamma$, $q(\cdot|\gamma) \in C_M^\eta(\mathcal{X}_\mathcal{W})$. Let the Borel measure μ be given by $\mu(A) = E\left[1(Z \in A)\overline{\phi}^2(Z)\right]$, then $\|\cdot\|_{L_2,\mu} = \|\cdot\|_{L_2(\overline{\phi})}$. Notice that the function class $\mathcal{T}_M^\eta(\overline{\phi})$ is associated with the L_2-pseudonorm $\|\cdot\|_{L_2(\overline{\phi})}$ instead of the sup-norm $\|\cdot\|_\infty$ used in EJL.

Let $\overline{\phi}(z) = |x| + 1$, where x is a subvector of z as indicated in Assumption 1. Define $T_Y(\cdot|\gamma) = f(\cdot|\gamma)\mu_Y(\cdot|\gamma)$ and $T_X(\cdot|\gamma) = f(\cdot|\gamma)\mu_X(\cdot|\gamma)$.

Lemma P1: *let Assumptions I, and 1-7 hold. For $\overline{\phi}$ as specified above: (i) $f(\cdot|\gamma) \in \mathcal{T}_M^1(\overline{\phi})$, $T_Y(\cdot|\gamma) \in \mathcal{T}_M^1(\overline{\phi})$, and $T_X(\cdot|\gamma) \in \mathcal{T}_M^1(\overline{\phi})$; (ii) the kernel estimators satisfy: $P\left(\widehat{f} \in \mathcal{T}_M^1(\overline{\phi})\right) \to 1$, $P\left(\widehat{T}_Y \in \mathcal{T}_M^1(\overline{\phi})\right) \to 1$, and $P\left(\widehat{T}_X \in \mathcal{T}_M^1(\overline{\phi})\right) \to 1$.*

Proof of Lemma P1: Note that

$$f(w|\gamma) = E\left[f_{Z_1|Z_2,D=1}(w + Z_2'\gamma)\right]$$

and

$$\begin{aligned} T_Y(w|\gamma) &= f(w|\gamma)\int y\frac{E\left[f_{(Y,Z_1)|Z_2,D=1}(y, w + Z_2'\gamma)\right]}{f(w|\gamma)}dy \\ &= \int yE\left[f_{(Y,Z_1)|Z_2,D=1}(y, w + Z_2'\gamma)\right]dy. \end{aligned}$$

Hence, $f(\cdot|\gamma) \in \mathcal{T}_M^1(\overline{\phi})$ and $T_Y(\cdot|\gamma) \in \mathcal{T}_M^1(\overline{\phi})$ follows from Assumptions 1 and 3. Similarly, it holds that $T_X(\cdot|\gamma) \in \mathcal{T}_M^1(\overline{\phi})$.

We show that $P\left(\widehat{f} \in \mathcal{T}_M^1(\overline{\phi})\right) \to 1$. This will follow if we prove that

$$\sup_{w \in \mathcal{X}_\mathcal{W}} \frac{\left|\widehat{f}(w|\gamma_1) - \widehat{f}(w|\gamma_2)\right|}{|\gamma_1 - \gamma_2|} = O_\mathbb{P}(1) \tag{6}$$

and

$$P(\widehat{f}(\cdot|\gamma) \in C_M^1(\mathcal{X}_\mathcal{W})) \to 1 \text{ each } \gamma \in \Gamma. \tag{7}$$

Define $\psi := (\gamma_1, \gamma_2, w) \in \Psi := \Gamma^2 \times \mathcal{X}_{\mathcal{W}}$. To verify (6), we write

$$\begin{aligned} \widehat{m}_{h,1}(\psi) &:= \frac{\widehat{f}\,(w|\gamma_1) - \widehat{f}\,(w|\gamma_2)}{|\gamma_1 - \gamma_2|} \\ &:= \frac{1}{nh^{3/2}} \sum_{i=1}^{n} v_{h,1}(Z_i, \psi), \end{aligned}$$

where

$$v_{h,1}(Z_i, \psi) := \frac{\sqrt{h}}{|\gamma_1 - \gamma_2|} \left\{ K\left(\frac{w - Z_i'\gamma_1}{h}\right) - K\left(\frac{w - Z_i'\gamma_2}{h}\right) \right\}$$

From Lemma B.3 in the Appendix of EJL, for each $\psi \in \Psi$,

$$E[|v_{h,1}(Z_i, \psi)|^2] \leq Ch.$$

Then, arguing as in Corollary 3 in Einmahl and Mason (2005) one can show that

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} |\widehat{m}_{h,1}(\psi) - m_1(\psi)| = O_{\mathbb{P}}(d_{1n}),$$

where $m_1(\psi) = \{f\,(w|\gamma_1) - f\,(w|\gamma_2)\} / |\gamma_1 - \gamma_2|$ and

$$d_{1n} = \sqrt{\frac{\log a_n^{-1} \vee \log\log n}{n a_n^{5/2}}} + b_n^2.$$

By the arguments above and Assumption 5, (6) holds. Similarly, to prove (7) we need to show that

$$\sup_{w_1 \neq w_2 \in \mathcal{X}_{\mathcal{W}}} \frac{\left|\widehat{f}\,(w_1|\gamma) - \widehat{f}\,(w_2|\gamma)\right|}{|w_1 - w_2|} = O_{\mathbb{P}}(1)$$

To that end, define for $\psi := (\gamma, w_1, w_2) \in \Psi := \Gamma \times \mathcal{X}_{\mathcal{W}}^2$,

$$\begin{aligned} \widehat{m}_{h,2}(\psi) &:= \frac{1}{nh^{3/2}} \sum_{i=1}^{n} \frac{\sqrt{h}}{|w_1 - w_2|} \left\{ K\left(\frac{w_1 - Z_i'\gamma}{h}\right) - K\left(\frac{w_2 - Z_i'\gamma}{h}\right) \right\}, \\ &=: \frac{1}{nh^{3/2}} \sum_{i=1}^{n} v_{h,2}(Z_i, \psi). \end{aligned}$$

Then, again from Corollary 3 in Einmahl and Mason (2005), one can show that

$$\sup_{a_n \leq h \leq b_n} \sup_{\psi \in \Psi} |\widehat{m}_{h,2}(\psi) - m_2(\psi)| = O_{\mathbb{P}}(d_{1n}),$$

where $m_2(\psi) = \{f(w_1|\gamma) - f(w_2|\gamma)\}/|w_1 - w_2|$. This proves (7).

The proof for $P\left(\widehat{T}_Y \in \mathcal{T}_M^1\left(\overline{\phi}\right)\right) \to 1$ and $P\left(\widehat{T}_X \in \mathcal{T}_M^1\left(\overline{\phi}\right)\right) \to 1$ follows the same steps as that for $\widehat{f}(\cdot|\gamma)$ and hence, it is omitted. *Q.E.D.*

6.2 A General Result

We develop in this section an asymptotic representation for the density-weighted process:

$$\widehat{\Delta}_n((\phi, W)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi(X_i) D_i \widehat{f}(W_i|W)(Y_i - \widehat{\mu}_Y(W_i|W))$$

where $\widehat{f}(\cdot|W)$ is the kernel estimator of the density function $f(\cdot|W)$ of $W(X_i)$ conditioning on $D_i = 1$, $\widehat{\mu}_Y(W_i|W)$ is the regression kernel estimator of $E[Y_i|W(X_i), D_i = 1]$, W_i is abbreviated for $W(X_i)$ and ϕ belongs to a class of functions Φ satisfying some conditions below. More precisely,

$$\begin{aligned}
\widehat{\mu}_Y(w|W) &:= \widehat{T}_Y(w|W)/\widehat{f}(w|W), \\
\widehat{T}_Y(w|W) &:= \frac{1}{n}\sum_{i=1}^{n} Y_i K_{\widehat{h}_n}(w - W(X_i)), \\
\widehat{f}(w|W) &:= \frac{1}{n}\sum_{i=1}^{n} K_{\widehat{h}_n}(w - W(X_i)),
\end{aligned}$$

where $K_h(w) = \prod_{l=1}^{d_W} k_h(w_l)$, $k_h(w_l) = h^{-1}k(w_l/h)$, $k(\cdot)$ is a kernel function, $w = (w_1, \ldots, w_d)'$ and $\widehat{h}_n$ denotes a possibly data dependent bandwidth parameter satisfying regularity conditions described below.

The representation we provide for $\widehat{\Delta}_n$ is more general than needed for the main result of the paper and is of independent interest. The results derived below can be readily modified if the conditioning indicator D_i does not show up. For notational convenience, we suppress the conditional dependence on $D_i = 1$ in what follows. The assumptions used here are weaker than those in EJL which they need to deal with trimming functions. Let $\mathcal{X}_{\mathcal{W}} = \left\{W(x) \in \mathbb{R}^{d_W} : W \in \mathcal{W} \text{ and } x \in \mathcal{X}_X\right\}$.

Assumption A1: *The sample observations* $\{Y_i, X_i, D_i\}_{i=1}^n$ *are a sequence of independent and identically distributed (iid) variables, distributed as* $\{Y, X, D\}$, *satisfying* $E[|Y|^s|X = x, D = 1] \leq C$ *a.s. for some* $s > 2$, *and* $\Pr(D = 1) > 0$.

Assumption A2: *(i) The function class* Φ *is such that* $\log N\left(\varepsilon, \Phi, \|\cdot\|_{L_2}\right) \leq C\varepsilon^{-v_\Phi}$ *for some* $v_\Phi < 2$, *and it has an envelope function* $\widetilde{\phi}$, *redefined as* $\overline{\phi} = \widetilde{\phi} \vee 1$, *satisfying* $\left\|\overline{\phi}\right\|_{L_2} < \infty$;

(ii) assume for each $W \in \mathcal{W}$, $E\left[|1+W^2(X)|^{s_W/2}\overline{\phi}^2(X)\right] < \infty$ for some $s_W > d_W/2$; *and the class* $\mathcal{W}$ *satisfies* $\log N\left(\varepsilon, \mathcal{W}, \|\cdot\|_{L_2(\overline{\phi})}\right) \leq C\varepsilon^{-v_{\mathcal{W}}}$ for some $v_{\mathcal{W}} < 1/2$, *with respect to the weighted* L_2*-pseudonorm* $\|\cdot\|_{L_2(\overline{\phi})}$, *defined as* $\|W\|_{L_2(\overline{\phi})} = \left(E\left[|W|^2\left|\overline{\phi}\right|^2\right]\right)^{1/2}$.

Redefining the envelope function is for technical convenience and same conditions can be equivalently imposed on $\widetilde{\phi}$. The uniform boundedness assumption of conditional moments of Y given X can be relaxed by working with a weighted L_2-pseudonorm for Φ and a mild modification of the proofs. The weighted L_2-pseudonorm defined here is stronger than L_2-norm, but is much weaker than the sup-norm. We can also consider higher-order norms, for example L_4-norm, which guarantees the validity of weighted L_2-pseudonorm. For the index function class $\mathcal{W}$, the weighted L_2-pseudonorm is more appropriate for our application since the index functions are not required to be bounded, and here we do not impose boundedness assumption for the class Φ either which is also relevant for our purpose. Moreover, the moment condition on W is imposed to allow for the application of a result in Nickl and Pötscher (2007) as discussed below.

Assumption A3: *For all* $W \in \mathcal{W}$, $w \in \mathcal{X}_{\mathcal{W}}$, $f(w|W)$ *and* $\mu_Y(w|W)$ *are* r*-times continuously differentiable in* w *with uniformly (in* w, W*) bounded derivatives (including zero derivative), where* r *is as in Assumption A4.*

Assumption A4: *The kernel function* $k(t): \mathbb{R} \to \mathbb{R}$ *is bounded, symmetric, continuously differentiable, and satisfies the following conditions:* $\int k(t)\,dt = 1$, $\int t^l k(t)\,dt = 0$ *for* $0 < l < r$, *and* $\int |t^r k(t)|\,dt < \infty$, *for some* $r \geq 2$; $\partial k(t)/\partial t$ *is bounded and for some* $v > 1$, $|\partial k(t)/\partial t| \leq C|t|^{-v}$ *for* $|t| > L$, $0 < L < \infty$.

Assumption A5: *The possibly data-dependent bandwidth* $\widehat{h}_n$ *satisfies* $\Pr\left(a_n \leq \widehat{h}_n \leq b_n\right) \to 1$ *as* $n \to \infty$, *where* $\{a_n, b_n\}$ *are deterministic sequences of positive numbers such that:* (i) $a_n \to 0$, $na_n^{d_W}/\ln n \to \infty$; (ii) $nb_n^{2r} \to 0$.

These assumptions are standard in the literature. We first derive several parallel lemmas as in EJL with modifications for our setup. The following Lemma A1 modifies Lemma B.3 in the Appendix of EJL. Given a compact set $I \subset \mathcal{X}_{\mathcal{W}}$, define the class of functions

$$\mathcal{K}_0 = \left\{x \to K\left(\frac{w - W(x)}{h}\right) : w \in I, W \in \mathcal{W}, h \in (0,1]\right\}.$$

Lemma A1: *Under Assumption A4, for some positive constant C and C_1,*

$$N_{[\cdot]}\left(C_1\varepsilon, \mathcal{K}_0, \|\cdot\|_{L_2}\right) \leq C\varepsilon^{-v} N\left(\varepsilon^4, \mathcal{W}, \ \|\cdot\|_{L_2\left(\overline{\phi}\right)}\right), \text{ for some } v \geq 1. \tag{8}$$

Proof of Lemma A1: The proof follows the exactly same arguments as in the proof of Lemma B.3 in the Appendix of EJL by noticing that

$$\begin{aligned}
& E\left[\left|K\left(\frac{w-W_1(X)}{h}\right) - K\left(\frac{w-W_2(X)}{h}\right)\right|^2\right] \\
\leq\ & CE\left[\left|K\left(\frac{w-W_1(X)}{h}\right) - K\left(\frac{w-W_2(X)}{h}\right)\right|^{1/2}\right] \\
\leq\ & Ch^{-1/2}E\left[\left|W_1(X) - W_2(X)\right|^{1/2} K^{*1/2}\left(\frac{w-W_2(X)}{h}\right)\right] \\
\leq\ & Ch^{-1/2}\left(E\left[\left|W_1(X) - W_2(X)\right|\right]\right)^{1/2}\left(E\left[K^*\left(\frac{w-W_2(X)}{h}\right)\right]\right)^{1/2} \\
\leq\ & C\left\|W_1 - W_2\right\|_{L_2}^{1/2} \leq C\left\|W_1 - W_2\right\|_{L_2\left(\overline{\phi}\right)}^{1/2},
\end{aligned}$$

where the second inequality is due to the boundedness of the kernel function, $K^*(\cdot)$ is such that $|K(x) - K(y)| \leq |x-y| K^*(y)$ (see Lemma B.3 in the Appendix of EJL), and the last inequality comes from the definition of $\overline{\phi}$. *Q.E.D.*

Using Lemma A1 and under Assumption A1-A5, by redefining

$$\widehat{m}_h(\psi) = \frac{1}{nh^{d_W}} \sum_{i=1}^{n} \psi(X_i) D_i K\left(\frac{w - W(X_i)}{h}\right), \tag{9}$$

we can obtain the same results as displayed in Lemma B.4 in the Appendix of EJL, with which we can derive same results as in Lemma B.5 and B.6, and an analog result as in Lemma B.7. For the ease of reference, we list these results in our Lemma A2 below. Define

$$d_n := \sqrt{\frac{\log a_n^{-d_W} \vee \log\log n}{n a_n^{d_W}}} + b_n^r + n^{-1/2}.$$

Lemma A2: *Let Assumption A1-A5 hold, then we have:*

(i) the kernel density estimator (conditional on $D = 1$)

$$\sup_{a_n \leq h \leq b_n} \sup_{w \in \mathcal{X}_{\mathcal{W}}, W \in \mathcal{W}} \left|\widehat{f}(w|W) - f(w|W)\right| = O_{P^*}(d_n);$$

(ii) the kernel estimator for $T(w|W) = f(w|W)\,\mu_Y(w|W)$

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{X}_\mathcal{W}, W \in \mathcal{W}} \left| \widehat{T}(w|W) - T(w|W) \right| = O_{P^*}(d_n);$$

(iii) the regression kernel estimator

$$\sup_{a_n \le h \le b_n} \sup_{w \in \mathcal{X}_\mathcal{W}, W \in \mathcal{W}} \left| \widehat{f}(w|W) \left(\widehat{\mu}_Y(w|W) - \mu_Y(w|W) \right) \right| = O_{P^*}(d_n);$$

Proof of Lemma A2:

(i) Recall that

$$\widehat{f}(w|W) = \widehat{\pi}^{-1} \frac{1}{n} \sum_{i=1}^{n} D_i K_h \left(\frac{w - W(X_i)}{h} \right).$$

Then

$$\begin{aligned} \sup \left| \widehat{f}(w|W) - f(w|W) \right| &\le \widehat{\pi}^{-1} \sup \left| \frac{1}{n} \sum_{i=1}^{n} D_i K_h \left(\frac{w - W(X_i)}{h} \right) - E\left[D_i K_h \left(\frac{w - W(X_i)}{h} \right) \right] \right| \\ &\quad + \sup \left| \widehat{\pi}^{-1} E\left[D_i K_h \left(\frac{w - W(X_i)}{h} \right) \right] - f(w|W) \right| \\ &:= \widehat{\pi}^{-1} I_{1n} + I_{2n}, \end{aligned}$$

where the sup is taken over $a_n \le h \le b_n$, $w \in \mathcal{X}_W$ and $W \in \mathcal{W}$. Lemma B.4 in EJL implies that (with $\psi(\cdot) = 1$)

$$I_{1n} = O_{P^*} \left(\sqrt{\frac{\log a_n^{-d_W} \vee \log\log n}{n a_n^{d_W}}} \right).$$

For I_{2n}, notice that

$$E\left[D_i K_h \left(\frac{w - W(X_i)}{h} \right) \right] = \pi \int K_h \left(\frac{w - u}{h} \right) f(u|W)\, du,$$

where $\pi = \Pr(D = 1)$, then by classical change of variables, Taylor expansion and Assumption A3 and A4,

$$\begin{aligned} I_{2n} &\le \widehat{\pi}^{-1} \pi \sup \left| \int K_h \left(\frac{w - u}{h} \right) f(u|W)\, du - f(w|W) \right| \\ &\quad + \left| \widehat{\pi}^{-1} \pi - 1 \right| \sup_{w \in \mathcal{X}_W} |f(w|W)| \\ &\le O_{P^*}\left(b_n^r + n^{-1/2} \right). \end{aligned}$$

Additionally, Assumption A1 implies $\Pr(D = 1) > 0$, hence $\widehat{\pi}^{-1} < \infty$ with probability one.

Then the conclusion follows.

(ii) follows from same arguments and (iii) follows directly from (i) and (ii). *Q.E.D.*

Let $\mathcal{T}_M^\eta\left(\overline{\phi}\right)$ be the function class defined similarly as in Section 3 (with different $\overline{\phi}$ as introduced in Assumption A2), i.e. $q\left(w|W\right)$ is said to be in $\mathcal{T}_M^\eta\left(\overline{\phi}\right)$ if for all $W \in \mathcal{W}$, $q\left(\cdot|W\right) \in C_M^\eta\left(\mathcal{X}_\mathcal{W}\right)$ and for all W_1, $W_2 \in \mathcal{W}$, the uniform L_2-continuity holds, that is, for a universal constant C_L,

$$\left\|q\left(W_1\left(\cdot\right)|W_1\right) - q\left(W_2\left(\cdot\right)|W_2\right)\right\|_{L_2\left(\overline{\phi}\right)} \leq C_L \left\|W_1 - W_2\right\|_{L_2\left(\overline{\phi}\right)}.$$

For unbounded $\mathcal{X}_\mathcal{W}$, e.g. $\mathcal{X}_\mathcal{W} = \mathbb{R}^{d_W}$, Corollary 3.2 in Nickl and Pötscher (2007) showed the bracketing entropy rate of $C_M^\eta\left(\mathcal{X}_\mathcal{W}\right)$ (in which the weighting coefficient $\beta = 0$, $\gamma = s_W$, and the measure $\mu\left(A\right) = E\left[1\left(X \in A\right)\overline{\phi}^2\left(X\right)\right]$ according to their notations):

$$\log N_{[\cdot]}\left(\varepsilon, C_M^\eta\left(\mathcal{X}_\mathcal{W}\right), \left\|\cdot\right\|_{L_2,\mu}\right) \leq C\varepsilon^{-d_W/\min(\eta, s_W)}. \tag{10}$$

Next lemma provides bound for the entropy rate of the above defined function class.

Lemma A3: *Let $T_M^\eta\left(\overline{\phi}\right)$ be as defined above. Then for all $\varepsilon > 0$ and some positive constant C,*

$$N\left(2\varepsilon, \mathcal{T}_M^\eta\left(\overline{\phi}\right), \left\|\cdot\right\|_{L_2\left(\overline{\phi}\right)}\right) \leq N\left(\varepsilon, \mathcal{W}, \left\|\cdot\right\|_{L_2\left(\overline{\phi}\right)}\right) \times N_{[\cdot]}\left(\varepsilon, C_M^\eta\left(\mathcal{X}_\mathcal{W}\right), \left\|\cdot\right\|_{L_2\left(\overline{\phi}\right)}\right).$$

Proof of Lemma A3: For any given $\varepsilon > 0$, let $\left\{W_j : j = 1, ..., N_{1\varepsilon}\right\}$ be a sequence of centers of ε-balls covering $\mathcal{W}$ with respect to $\left\|\cdot\right\|_{L_2\left(\overline{\phi}\right)}$, and for each W_j, $\left\{q_i\left(\cdot|W_j\right), \delta_{W_j}\left(\cdot\right)\right\}_{i=1}^{N_{2\varepsilon}}$ be a sequence of bracket pairs covering the marginal class $\left\{w \to q\left(w|W_j\right)\right\}$ with respect to $\left\|\cdot\right\|_{L_2\left(\overline{\phi}\right)}$, i.e. for any $q\left(\cdot|W_j\right) \in C_M^\eta\left(\mathcal{X}_\mathcal{W}\right)$, there exists $\left\{q_i\left(\cdot|W_j\right), \delta_{W_j}\left(\cdot\right)\right\}$ such that $\left|q\left(w|W_j\right) - q_i\left(w|W_j\right)\right| \leq \delta_{W_j}\left(w\right)$ and $\left\|\delta_{W_j}\left(W_j\right)\right\|_{L_2\left(\overline{\phi}\right)} \leq \varepsilon$. Then for each $q\left(W\left(\cdot\right)|W\right) \in T_M^\eta\left(\overline{\phi}\right)$, there exist j and i such that

$$\begin{aligned}
& E\left[\left|q\left(W\left(X\right)|W\right) - q_i\left(W_j\left(X\right)|W_j\right)\right|^2 \overline{\phi}^2\left(X\right)\right] \\
\leq\ & 2E\left[\left|q\left(W\left(X\right)|W\right) - q\left(W_j\left(X\right)|W_j\right)\right|^2 \overline{\phi}^2\left(X\right)\right] \\
& +2E\left[\left|q\left(W_j\left(X\right)|W_j\right) - q_i\left(W_j\left(X\right)|W_j\right)\right|^2 \overline{\phi}^2\left(X\right)\right] \\
\leq\ & 4\varepsilon^2,
\end{aligned}$$

thus the result follows. *Q.E.D.*

By Assumption A2 that $v_\Phi < 2$ and $s_W > d_W/2$, then as long as $\eta > d_W/2$, we can

obtain that

$$\log N\left(2\varepsilon, \mathcal{T}_M^{\eta}\left(\overline{\phi}\right), \|\cdot\|_{L_2\left(\overline{\phi}\right)}\right) \leq C\varepsilon^{-v} \text{ with } v < 2$$

which indicates that the function class $\mathcal{T}_M^{\eta}\left(\overline{\phi}\right)$ is Donsker.

Now, we are ready to prove the asymptotic representation of $\widehat{\Delta}_n\left((\phi, W)\right)$. The proof is simpler than that of Theorem 2.1 in EJL since we do not need to deal with the indicator trimming function. Let

$$\begin{aligned}\Delta_n\left((\phi, W)\right) \quad &: \quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{D_i\phi_i\left(Y_i f_i - T_{Yi}\right) - E\left[D_i\phi_i\left(Y_i f_i - T_{Yi}\right)\right]\right\} \\ &\quad -\frac{1}{\sqrt{n}}\sum_{j=1}^{n} D_i f_i\left\{Y_i l_{\phi}\left(W_i\right) - l_{\phi Y}\left(W_i\right)\right\} \\ &\quad +\pi^{-1}E\left[D_i f_i\left\{Y_i l_{\phi}\left(W_i\right) - l_{\phi Y}\left(W_i\right)\right\}\right]\sqrt{n}\left(\widehat{\pi} - \pi\right),\end{aligned}$$

where $l_{\phi Y}\left(w\right) := E\left[\phi\left(X_i\right)Y_i | W_i = w, D_i = 1\right]$, $l_{\phi}\left(w\right) := E\left[\phi\left(X_i\right)|W_i = w, D_i = 1\right]$ and we use the shortened notations, $\phi_i = \phi\left(X_i\right)$, $f_i = f\left(W_i|W\right)$ and $T_{Yi} = f\left(W_i|W\right)\mu_Y\left(W_i|W\right)$. Notice that if there is no selection indicator D_i and all the conditional quantities are only conditioning on W_i, then the last term in the above representation will disappear.

For given $\mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right)$ and $\mathcal{T}_M^{\eta_Y}\left(\overline{\phi}\right)$, let $\mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right) = \left\{f \cdot \mu : \ f \in \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right), \text{ and } \mu \in \mathcal{T}_M^{\eta_Y}\left(\overline{\phi}\right)\right\}$ with $\eta_{TY} = \max\left(\eta_f, \eta_Y\right)$. Lemma A4 below shows that the covering number of $\mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right)$ is bounded by the product of the covering numbers of $\mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right)$ and $\mathcal{T}_M^{\eta_Y}\left(\overline{\phi}\right)$.

Lemma A4: *Let $\mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right)$ be defined as above, then we have for some constant C :*

$$N\left(\varepsilon, \mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right), \|\cdot\|_{L_2\left(\overline{\phi}\right)}\right) \leq CN\left(\varepsilon, \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right), \|\cdot\|_{L_2\left(\overline{\phi}\right)}\right) \times N\left(\varepsilon, \mathcal{T}_M^{\eta_Y}\left(\overline{\phi}\right), \|\cdot\|_{L_2\left(\overline{\phi}\right)}\right).$$

Proof of Lemma A4: Note that

$$\begin{aligned}&\left|fg - f_i g_j\right|^2 \\ \leq \ & 2\left|f - f_i\right|^2\left|g\right| + 2\left|f_i\right|^2\left|g - g_j\right|^2\end{aligned}$$

and that, g and f_i are bounded by the definition of function class $C_M^{\eta}\left(\mathcal{X}_{\mathcal{W}}\right)$. *Q.E.D.*

Assumption A6: *(i) $f\left(\cdot|W\right) \in \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right)$, $\mu_Y\left(\cdot|W\right) \in \mathcal{T}_M^{\eta_Y}\left(\overline{\phi}\right)$; (ii) the kernel estimators satisfies: $P\left(\widehat{f} \in \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right)\right) \to 1$ and $P\left(\widehat{T}_Y \in \mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right)\right) \to 1$* for some $\eta_f > d_W/2$, $\eta_Y > d_W/2$, and $M > 0$.

Assumption A6(i) and 2(ii) together imply that the function classes $\mathcal{T}_M^{\eta}\left(\overline{\phi}\right)$ for $\eta = \eta_f, \eta_Y, \eta_{TY}$ are all Donsker classes.

Theorem A1: *Under Assumption A1-A6, we have*

$$\sup_{a_n \leq \widehat{h}_n \leq b_n} \sup_{(\phi,W)\in\Phi\times\mathcal{W}} \left|\widehat{\Delta}_n\left((\phi,W)\right) - \Delta_n\left((\phi,W)\right)\right| = o_{P^*}(1).$$

Proof of Theorem A1: Define a class of functions $\mathcal{S} = \{(y,x,d) \to d\phi(x)(y\psi(x) - m(x)) : (\phi,\psi,m) \in \Phi \times \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right) \times \mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right)\}$. Notice that by triangle inequality

$$\begin{aligned}
& \left|\phi_2(X)(Y\psi_2(X) - m_2(X)) - \phi_1(X)(Y\psi_1(X) - m_1(X))\right|^2 \\
\leq\ & 2\left|\phi_2(X) - \phi_1(X)\right|^2 \left|Y\psi_1(X) - m_1(X)\right|^2 \\
& + 2\left|\phi_2(X)\right|^2 Y^2 \left|\psi_2(X) - \psi_1(X)\right|^2 + 2\left|\phi_2(X)\right|^2 \left|m_2(X) - m_1(X)\right|^2 \\
\leq\ & 2\left|\phi_2(X) - \phi_1(X)\right|^2 \left(2Y^2\left|\psi_1(X)\right|^2 + 2\left|m_1(X)\right|^2\right) \\
& + 2\left|\overline{\phi}(X)\right|^2 Y^2 \left|\psi_2(X) - \psi_1(X)\right|^2 + 2\left|\overline{\phi}(X)\right|^2 \left|m_2(X) - m_1(X)\right|^2.
\end{aligned}$$

Fix any $\delta > 0$, let the sup below be taken over $(\phi_2,\psi_2,m_2) \in \Phi \times \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right) \times \mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right)$ such that $\|\phi_2 - \phi_1\|_{L_2} < \delta$, $\|\psi_2 - \psi_1\|_{L_2(\overline{\phi})} < \delta, \|m_2 - m_1\|_{L_2(\overline{\phi})} < \delta$, then we have

$$\begin{aligned}
& E\left[\sup\left|\phi_2(X)(Y\psi_2(X) - m_2(X)) - \phi_1(X)(Y\psi_1(X) - m_1(X))\right|^2\right] \\
=\ & 2E\left[\left|\phi_2(X) - \phi_1(X)\right|^2 \left(E\left(2Y^2|X\right)\left|\psi_1(X)\right|^2 + 2\left|m_1(X)\right|^2\right)\right] + \\
& 2E\left[\left|\overline{\phi}(X)\right|^2 \left|\psi_2(X) - \psi_1(X)\right|^2 E\left(Y^2|X\right)\right] + 2E\left[\left|\overline{\phi}(X)\right|^2 \left|m_2(X) - m_1(X)\right|^2\right] \\
\leq\ & C\delta^2
\end{aligned}$$

where the last inequality is due to Assumption A1 and A2. Hence, by Lemma A.1 in EJL, we obtain that

$$N_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_{L_2}\right) \leq N\left(\frac{\varepsilon}{2C}, \Phi, \|\cdot\|_{L_2}\right) \times N\left(\frac{\varepsilon}{2C}, \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right), \|\cdot\|_{L_2(\overline{\phi})}\right) \times N\left(\frac{\varepsilon}{2C}, \mathcal{T}_M^{\eta_{TY}}\left(\overline{\phi}\right), \|\cdot\|_{L_2(\overline{\phi})}\right),$$

which implies that the function class $\mathcal{S}$ is Donsker class by Assumption A2 and A6, and Lemma A4.

Hence the centered process

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{D_i\phi(X_i)(Y_i\psi(X_i) - m(X_i)) - E\left[D_i\phi(X_i)(Y_i\psi(X_i) - m(X_i))\right]\right\}$$

is stochastically equicontinuous with respect to the pseudometric $\rho(\lambda_1,\lambda_2) = \|s(\cdot,\lambda_1) - s(\cdot,\lambda_2)\|_{L_2}$ with $\lambda_j = \left(\phi_j,\psi_j,m_j\right)$. Then by stochastic equicontinuity, Assumption A6 and Lemma A2,

we have uniformly in $(\phi, W) \in \Phi \times \mathcal{W}$,

$$\begin{aligned}\widehat{\Delta}_n((\phi, W)) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{D_i\phi_i\left(Y_i f_i - T_{Yi}\right) - E\left[D_i\phi_i\left(Y_i f_i - T_{Yi}\right)\right]\right\} \\ &\quad + \sqrt{n}E\left[D_i\phi\left(X_i\right)\widehat{f}\left(W_i|W\right)\left(Y_i - \widehat{\mu}_Y\left(W_i|W\right)\right)\right] + o_{P^*}(1).\end{aligned}$$

Now we shall find asymptotic representation for the second term. Recall the definitions $l_{\phi Y}(w) = E[\phi(X_i) Y_i | W_i = w, D_i = 1]$ and $l_\phi(w) = E[\phi(X_i) | W_i = w, D_i = 1]$. Then

$$\begin{aligned}E\left[D_i\phi\left(X_i\right)Y_i\widehat{f}\left(W_i|W\right)\right] &= \pi E\left[l_{\phi Y}\left(W_i\right)\widehat{f}\left(W_i|W\right)|D_i = 1\right] \\ &= \frac{\pi}{\widehat{\pi}}\frac{1}{n}\sum_{j=1}^{n}D_j\int l_{\phi Y}(w) K_h\left(\frac{W_j - w}{h}\right) f(w|W)\, dw \\ &= \frac{\pi}{\widehat{\pi}}\frac{1}{n}\sum_{j=1}^{n}D_j l_{\phi Y}\left(W_j\right) f\left(W_j|W\right) + O_{P^*}\left(h^r\right),\end{aligned}$$

$$\begin{aligned}E\left[D_i\phi\left(X_i\right)\widehat{T}_Y\left(W_i|W\right)\right] &= \pi E\left[l_\phi\left(W_i\right)\widehat{T}_Y\left(W_i|W\right)|D_i = 1\right] \\ &= \frac{\pi}{\widehat{\pi}}\frac{1}{nh^{dw}}\sum_{j=1}^{n}D_j Y_j\int l_\phi(w) K_h\left(\frac{W_j - w}{h}\right) f(w|W)\, dw \\ &= \frac{\pi}{\widehat{\pi}}\frac{1}{n}\sum_{j=1}^{n}D_j Y_j l_\phi\left(W_j\right) f\left(W_j|W\right) + O_{P^*}\left(h^r\right),\end{aligned}$$

hence, by Assumption A5 on the rates of bandwidths, the second term has the asymptotic representation as

$$\begin{aligned}&\sqrt{n}E\left[D_i\phi\left(X_i\right)\widehat{f}\left(W_i|W\right)\left(Y_i - \widehat{\mu}_Y\left(W_i|W\right)\right)\right] \\ = &\frac{\pi}{\widehat{\pi}}\frac{1}{\sqrt{n}}\sum_{j=1}^{n}D_j f\left(W_j|W\right)\left\{l_{\phi Y}\left(W_j\right) - Y_j l_\phi\left(W_j\right)\right\} + o_{P^*}(1) \\ = &\frac{1}{\sqrt{n}}\sum_{j=1}^{n}D_j f\left(W_j|W\right)\left\{l_{\phi Y}\left(W_j\right) - Y_j l_\phi\left(W_j\right)\right\} \\ &-\pi^{-1}E\left[D_j f\left(W_j|W\right)\left\{l_{\phi Y}\left(W_j\right) - Y_j l_\phi\left(W_j\right)\right\}\right]\sqrt{n}\left(\widehat{\pi} - \pi\right) + o_{P^*}(1).\end{aligned}$$

Q.E.D.

If $\phi(\cdot)$ is a function of W_i, then it is easy to see that $\Delta_n((\phi, W)) = 0$. Furthermore, due to uniformity of the above result, we obtain a direct corollary as follows.

Corollary A1: *Suppose* $\widehat{\phi}_n \to \phi_0 \in \Phi$ *and* $\widehat{W}_n \to W_0 \in W$ *in* L_2*-norm,* $\Pr\left(\widehat{\phi}_n \in \Phi\right) \to 1$

and $\Pr\left(\widehat{W}_n \in \mathcal{W}\right) \to 1$*, then under Assumption A1-A6,*

$$\left|\widehat{\Delta}_n\left(\left(\widehat{\phi}_n, \widehat{W}_n\right)\right) - \Delta_n\left(\left(\widehat{\phi}_n, \widehat{W}_n\right)\right)\right| = o_{P^*}(1).$$

6.3 Proof of Main Theorem

For brevity of notations, denote $\widehat{\mu}_{Xi} = \widehat{\mu}_{Xi}\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right)$, $\mu_{X_0 i} = E\left[X_i | Z_i'\gamma_0, D_i = 1\right]$ and similarly for $\widehat{\mu}_{Yi}$, $\mu_{Y_0 i}$, also let $\widehat{f}_{\widehat{\gamma}i} = \widehat{f}_i\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right)$ and $f_{0i} = f\left(Z_i'\gamma_0|\gamma_0\right)$.

Write

$$\begin{aligned}\sqrt{n}\left(\widehat{\theta}(\widehat{h}_n) - \theta_0\right) &= \left(\frac{1}{n}\sum_{i=1}^n D_i \widehat{f}^2_{\widehat{\gamma}i}\widehat{X}_{\widehat{\gamma}i}\widehat{X}'_{\widehat{\gamma}i}\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n D_i \widehat{f}^2_{\widehat{\gamma}i}\widehat{X}_{\widehat{\gamma}i}\left(\widehat{Y}_{\widehat{\gamma}i} - \widehat{X}'_{\widehat{\gamma}i}\theta_0\right)\right) \\ &= : \widehat{\Sigma}_n^{-1}\widehat{S}_n\end{aligned}$$

Define the index class of functions $\mathcal{W} = \{W(z) = z'\gamma : \gamma \in \Gamma\}$, and since $\widehat{f}_{\widehat{\gamma}i}$, and $\widehat{T}_{Xi}$ are all bounded in probability, we can let $\overline{\phi}(x) = |x| + 1$, and define the weighted L_2-pseudonorm as in Assumption A2(ii): $\|W_1 - W_2\|^2_{L_2(\overline{\phi})} \leq E\left[\overline{\phi}^2(X)|Z|^2\right]|\gamma_1 - \gamma_2|^2$. Hence, by Assumption 1 on the moment conditions of Z, the conditions on $\mathcal{W}$ are satisfied given that Γ is compact. In what follows, C is a generic constant and may change from equation to equation.

Step 1:

We first show that uniformly in $\widehat{h}_n \in [a_n, b_n]$, $\widehat{\Sigma}_n = \Sigma + o_{P^*}(1)$ where $\Sigma := E\left[D_i f^2_{0i} X_{0i} X'_{0i}\right]$. Notice that uniformly in $a_n \leq \widehat{h}_n \leq b_n$,

$$\begin{aligned}\left\|\widehat{f}_{\widehat{\gamma}i} - f_{0i}\right\|_{L_2} &\leq \left\|\widehat{f}_i\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right) - f\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right)\right\|_{L_2} + \left\|f\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right) - f\left(Z_i'\gamma_0|\gamma_0\right)\right\|_{L_2} \\ &\leq \sup_{w\in\mathcal{X}_{\mathcal{W}},\gamma\in\Gamma}\left|\widehat{f}(w|\gamma) - f(w|\gamma)\right| + \sup_{|\gamma-\gamma_0|\leq\delta_n}\left\|f\left(Z_i'\gamma|\gamma\right) - f\left(Z_i'\gamma_0|\gamma_0\right)\right\|_{L_2(\overline{\phi})} \\ &\quad + C\Pr\left(|\widehat{\gamma} - \gamma_0| > \delta_n\right) \\ &\leq O_{P^*}(d_n) + C\delta_n + o(1) \to 0\end{aligned}$$

where $d_n = \sqrt{\frac{\log a_n^{-1} \vee \log\log n}{n a_n}} + b_n^2 + n^{-1/2}$, $\delta_n \to 0$ and $n^{1/2}\delta_n \to \infty$, the third inequality follows from Lemma A2(i), Lemma P1 and Assumption 6, the last equality follows from our assumptions on the convergence rates of bandwidths and the first-stage estimator $\widehat{\gamma}$. Let $\widehat{T}_{Xi} = \widehat{T}_{Xi}\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right)$, $T_X(w|\gamma) = \mu_X(w|\gamma) f(w|\gamma)$ and $T_{X_0 i} = \mu_{X_0 i} f_{0i}$, then by Lemma A2(ii),

uniformly in $a_n \leq \widehat{h}_n \leq b_n$

$$\begin{aligned}\left\|\widehat{T}_{Xi} - T_{X_0 i}\right\|_{L_2} &\leq \left\|\widehat{T}_{Xi}\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right) - T_X\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right)\right\|_{L_2} + \left\|T_X\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right) - T_X\left(Z_i'\gamma_0|\gamma_0\right)\right\|_{L_2} \\ &\leq O\left(d_n\right) + \left\|T_X\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right) - T_X\left(Z_i'\gamma_0|\gamma_0\right)\right\|_{L_2}.\end{aligned}$$

By Triangle inequality and boundedness of $C_M^{\eta}\left(\mathcal{X}_{\mathcal{W}}\right)$, we have $\left\|T_X\left(Z_i'\widehat{\gamma}|\widehat{\gamma}\right) - T_X\left(Z_i'\gamma_0|\gamma_0\right)\right\|_{L_2} \leq C\left|\widehat{\gamma} - \gamma_0\right|$. Hence, $\left\|\widehat{T}_{Xi} - T_{X_0 i}\right\|_{L_2} = o\left(1\right)$. Then by Triangle inequalities, Cauchy-Schwarz inequality, Assumption 1, 5, 6, and Lemma P1, we obtain

$$\begin{aligned}& E\left|\widehat{\Sigma}_n - \frac{1}{n}\sum D_i f_{0i}^2 X_0 X_{0i}'\right| \\ \leq\ & E\left|D_i\{\left(X_i\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right)\left(X_i\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right)' - \left(X_i f_{0i} - T_{X_0 i}\right)\left(X_i f_{0i} - T_{X_0 i}\right)'\}\right| \\ \leq\ & C_1\left\|\widehat{f}_{\widehat{\gamma}i} - f_{0i}\right\|_{L_2} + C_2\left\|\widehat{T}_{Xi} - T_{X_0 i}\right\|_{L_2} = o\left(1\right).\end{aligned}$$

Hence, uniformly in $a_n \leq \widehat{h}_n \leq b_n$, $\widehat{\Sigma}_n = \frac{1}{n}\sum f_{0i}^2 X_{0i} X_{0i}' + o_{P^*}\left(1\right) = \Sigma + o_{P^*}\left(1\right)$.

Step 2:

We now turn to $\widehat{S}_n$. Write

$$\begin{aligned}\widehat{S}_n &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \widehat{f}_{\widehat{\gamma}i}^2 \widehat{X}_{\widehat{\gamma}i}\left(\widehat{Y}_{\widehat{\gamma}i} - \widehat{X}_{\widehat{\gamma}i}'\theta_0\right) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \widehat{f}_{\widehat{\gamma}i}^2 \widehat{X}_{\widehat{\gamma}i}\left(Y_{0i} - X_{0i}'\theta_0\right) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \widehat{f}_{\widehat{\gamma}i}^2 \widehat{X}_{\widehat{\gamma}i}\left(\mu_{Y_0 i} - \widehat{\mu}_{Yi}\right) \\ &\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \widehat{f}_{\widehat{\gamma}i}^2 \widehat{X}_{\widehat{\gamma}i}\left(\widehat{\mu}_{Xi} - \mu_{X_0 i}\right)'\theta_0 \\ &= : \widehat{S}_{n1} + \widehat{S}_{n2} + \widehat{S}_{n3}.\end{aligned}$$

(I) For $\widehat{S}_{n1}$, let $e_i = Y_{0i} - X_{0i}'\theta_0$, notice that $E\left[e_i|Z_i, D_i = 1\right] = 0$, we can apply Lemma A.3 in EJL. First define a class of functions, $s\left(d, z, \lambda\right) = dg\left(z\right)\left(xg\left(z\right) - k\left(z\right)\right)$, with $\lambda = \left(g\left(\cdot\right), k\left(\cdot\right)\right)$ and $\lambda \in \Lambda = \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right) \times \mathcal{T}_M^{\eta_{TX}}\left(\overline{\phi}\right)$ where $\mathcal{T}_M^{\eta_{TX}}\left(\overline{\phi}\right) := \left\{f \cdot \mu : f \in \mathcal{T}_M^{\eta_f} \text{ and } \mu \in \mathcal{T}_M^{\eta_X}\right\}$. Follow the same arguments as in the proof of Theorem A1, we can show that, letting $\mathcal{S} = \{s\left(\cdot, \lambda\right) : \lambda \in \Lambda\}$,

$$N_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_{L_2}\right) \leq N\left(\frac{\varepsilon}{C}, \mathcal{T}_M^{\eta_f}\left(\overline{\phi}\right), \|\cdot\|_{L_2(\overline{\phi})}^2\right) \times N\left(\frac{\varepsilon}{C}, \mathcal{T}_M^{\eta_{TX}}\left(\overline{\phi}\right), \|\cdot\|_{L_2(\overline{\phi})}^2\right).$$

According to Lemma A3 above, the bracketing number satisfies

$$\log N_{[\cdot]}\left(\varepsilon, \mathcal{S}, \|\cdot\|_{L_2}\right) \leq C\varepsilon^{-1}.$$

Furthermore, let $s_0(D,Z) = Df^2(Z'\gamma_0|\gamma_0)(X - \mu_X(Z'\gamma_0|\gamma_0))$ and its estimator $\widehat{s}(D,Z) = D\widehat{f}^2(Z'\widehat{\gamma}|\widehat{\gamma})(X - \widehat{\mu}_X(Z'\widehat{\gamma}|\widehat{\gamma}))$. It is easy to show $\|\widehat{s}(D_i,Z_i) - s_0(D_i,Z_i)\|_{L_2} \to 0$. Hence, we have, uniformly in $a_n \leq \widehat{h}_n \leq b_n$,

$$\begin{aligned}
\widehat{S}_{n1} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \widehat{f}^2_{\widehat{\gamma}i}\widehat{X}_{\widehat{\gamma}i}(Y_{0i} - X'_{0i}\theta_0) \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i D_i f^2_{\gamma_0 i} X_{0i} + o_{P^*}(1).
\end{aligned}$$

(II) For $\widehat{S}_{n2}$, denote $\mu_{\widehat{Y}i} = E[Y_i|Z_i'\widehat{\gamma}, D_i = 1]$ and $\mu_{\widehat{X}i} = E[X_i|Z_i'\widehat{\gamma}, D_i = 1]$, and write

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i \widehat{f}^2_{\widehat{\gamma}i}\widehat{X}_{\widehat{\gamma}i}\left(\mu_{Y_0 i} - \widehat{\mu}_{Yi}\right) \\
= &\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\left(X_i\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right)\left(\mu_{Y_0 i}\widehat{f}_{\widehat{\gamma}i} - \mu_{\widehat{Y}i}\widehat{f}_{\widehat{\gamma}i}\right) \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\left(\mu_{\widehat{Y}i}\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Yi}\right)\left(X_i\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right) \\
= &: \widehat{S}^{(1)}_{n2} + \widehat{S}^{(2)}_{n2}.
\end{aligned}$$

For the second term $\widehat{S}^{(2)}_{n2}$, let $\widehat{\phi}\left(\widehat{W}_i\right) := \mu_{\widehat{Y}i}\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Yi}$, then by Corollary A1,

$$\begin{aligned}
\widehat{S}^{(2)}_{n2} &= \Delta_n\left(\left(\widehat{\phi}, \widehat{W}_i\right)\right) + o_{P^*}(1) \\
&= o_{P^*}(1).
\end{aligned}$$

For the first term $\widehat{S}^{(1)}_{n2}$, write

$$\begin{aligned}
\widehat{S}^{(1)}_{n2} &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\widehat{f}_{\widehat{\gamma}i}\left(X_i\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right)\left(\mu_Y(Z_i'\gamma_0|\gamma_0) - \mu_Y(Z_i'\widehat{\gamma}|\gamma_0)\right) \\
&+\frac{1}{\sqrt{n}}\sum_{i=1}^{n} D_i\widehat{f}_{\widehat{\gamma}i}\left(X_i\widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right)\left(\mu_Y(Z_i'\widehat{\gamma}|\gamma_0) - \mu_Y(Z_i'\widehat{\gamma}|\widehat{\gamma})\right),
\end{aligned}$$

then by Taylor expansion, the first term equals

$$= -\left(\frac{1}{n}\sum_{i=1}^{n} D_i \widehat{f}_{\widehat{\gamma}i} \partial_w \mu_{Y_0 i} \left(X_i \widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right) Z_i'\right) \sqrt{n}\left(\widehat{\gamma} - \gamma_0\right) + o_{P^*}(1).$$

where $\partial_w \mu_{Y_0 i} = \frac{\partial}{\partial w}\mu_Y\left(w|\gamma_0\right)|_{w=Z_i'\gamma_0}$; while the second term is $o_{P^*}(1)$ by similar stochastic equicontinuity arguments as above. Then follow same arguments as in Step 1, we have

$$\begin{aligned} & \frac{1}{n}\sum_{i=1}^{n} D_i \widehat{f}_{\widehat{\gamma}i} \widehat{T}_{Di} \partial_w \mu_{Y_0 i} \left(X_i \widehat{f}_{\widehat{\gamma}i} - \widehat{T}_{Xi}\right) Z_i' \\ = \; & E\left[D_i f_{\gamma_0 i}^2 \partial_w \mu_{Y_0 i} X_{0i} Z_i'\right] + o_{P^*}(1). \end{aligned}$$

Hence, uniformly in $a_n \leq \widehat{h}_n \leq b_n$

$$\widehat{S}_{n2} = -E\left[D_i f_{\gamma_0 i}^2 \partial_w \mu_{Y_0 i} X_{0i} Z_i'\right] \sqrt{n}\left(\widehat{\gamma} - \gamma_0\right) + o_{P^*}(1)$$

(**III**) As for $\widehat{S}_{n3}$, follow the same arguments in (II), we obtain, uniformly in $a_n \leq \widehat{h}_n \leq b_n$

$$\widehat{S}_{n3} = E\left[D_i f_{0i}^2 X_{0i}\left(\partial_w \mu_{X_0 i}' \theta_0\right) Z_i'\right] \sqrt{n}\left(\widehat{\gamma} - \gamma_0\right) + o_{P^*}(1).$$

Combining (I), (II), (III), we obtain, uniformly in $a_n \leq \widehat{h}_n \leq b_n$,

$$\widehat{S}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i D_i f_{\gamma_0 i}^2 X_{0i} - B \cdot \sqrt{n}\left(\widehat{\gamma} - \gamma_0\right) + o_{P^*}(1)$$

where $B = E[D_i f_{\gamma_0 i}^2 \{\partial_w \mu_{Y_0 i} - \left(\partial_w \mu_{X_0 i}' \theta_0\right)\} X_{0i} Z_i']$ of full rank with dimension $d_X \times d_Z$. Using the asymptotic representation for $\sqrt{n}\left(\widehat{\gamma} - \gamma_0\right)$ in Assumption 6, we have

$$\widehat{S}_n = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(e_i D_i f_{\gamma_0 i}^2 X_{0i} - B\xi_i\right) + o_{P^*}(1)$$

which, by Assumption 7 and the standard Central Limit Theorem, converges in distribution to a normal random vector $N(0, \Omega)$ where $\Omega = E\left[\omega_i \omega_i'\right]$ with $\omega_i = e_i D_i f_{\gamma_0 i}^2 X_{0i} - B\xi_i$. *Q.E.D.*

7 Appendix B: Tables

Table 1: Comparison of data-driven estimator $(\widehat{\theta}_{DD})$
with power series estimators $(K = 3, 6, 9) : \rho = 0.1$

$\rho = 0.1$		Error type: $N(0,1)$				Error type: t_3			
Est.		Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE
	$\widehat{\theta}_{DD}$	-0.005	0.342	0.342	0.273	0.001	0.286	0.286	0.226
n=250	$\widehat{\theta}_{K=3}$	-0.031	0.424	0.425	0.307	-0.010	0.338	0.338	0.257
	$\widehat{\theta}_{K=6}$	-0.033	0.427	0.428	0.309	-0.012	0.343	0.343	0.261
	$\widehat{\theta}_{K=9}$	-0.032	0.425	0.426	0.309	-0.012	0.350	0.350	0.264
	$\widehat{\theta}_{DD}$	-0.003	0.242	0.242	0.191	0.017	0.216	0.217	0.171
n=500	$\widehat{\theta}_{K=3}$	-0.014	0.253	0.253	0.197	0.014	0.227	0.227	0.179
	$\widehat{\theta}_{K=6}$	-0.013	0.255	0.255	0.197	0.014	0.230	0.230	0.181
	$\widehat{\theta}_{K=9}$	-0.013	0.256	0.256	0.198	0.014	0.230	0.230	0.181
	$\widehat{\theta}_{DD}$	0.002	0.175	0.175	0.139	0.001	0.150	0.150	0.120
n=1000	$\widehat{\theta}_{K=3}$	-0.005	0.176	0.176	0.140	-0.004	0.151	0.151	0.122
	$\widehat{\theta}_{K=6}$	-0.006	0.177	0.177	0.140	-0.004	0.152	0.152	0.122
	$\widehat{\theta}_{K=9}$	-0.006	0.177	0.177	0.141	-0.004	0.152	0.152	0.122

Table 2: Comparison of data-driven estimator $(\widehat{\theta}_{DD})$
with power series estimators $(K = 3, 6, 9) : \rho = 0.5$

$\rho = 0.5$		Error type: $N(0,1)$				Error type: t_3			
Est.		Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE
	$\widehat{\theta}_{DD}$	0.024	0.332	0.334	0.265	0.030	0.280	0.281	0.222
n=250	$\widehat{\theta}_{K=3}$	-0.060	0.483	0.486	0.314	-0.022	0.332	0.333	0.253
	$\widehat{\theta}_{K=6}$	-0.061	0.482	0.486	0.314	-0.023	0.340	0.340	0.257
	$\widehat{\theta}_{K=9}$	-0.060	0.486	0.490	0.316	-0.023	0.341	0.341	0.260
	$\widehat{\theta}_{DD}$	0.019	0.234	0.235	0.187	0.035	0.215	0.218	0.171
n=500	$\widehat{\theta}_{K=3}$	-0.024	0.257	0.258	0.198	0.005	0.227	0.227	0.180
	$\widehat{\theta}_{K=6}$	-0.023	0.259	0.260	0.198	0.006	0.229	0.229	0.180
	$\widehat{\theta}_{K=9}$	-0.023	0.259	0.260	0.199	0.006	0.230	0.230	0.181
	$\widehat{\theta}_{DD}$	0.022	0.168	0.169	0.134	0.016	0.144	0.144	0.114
n=1000	$\widehat{\theta}_{K=3}$	-0.003	0.170	0.170	0.135	0.003	0.145	0.145	0.116
	$\widehat{\theta}_{K=6}$	-0.003	0.171	0.171	0.135	0.003	0.145	0.145	0.116
	$\widehat{\theta}_{K=9}$	-0.003	0.171	0.171	0.135	0.003	0.146	0.146	0.126

Table 3: Comparison of data-driven estimator $(\widehat{\theta}_{DD})$
with power series estimators $(K = 3, 6, 9)$: $\rho = 0.9$

$\rho = 0.9$		Error type: $N(0,1)$				Error type: t_3			
Est.		Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE
	$\widehat{\theta}_{DD}$	-0.053	0.298	0.303	0.243	0.055	0.259	0.264	0.210
n=250	$\widehat{\theta}_{K=3}$	-0.090	0.549	0.556	0.317	-0.042	0.335	0.337	0.253
	$\widehat{\theta}_{K=6}$	-0.089	0.540	0.547	0.317	-0.042	0.334	0.336	0.252
	$\widehat{\theta}_{K=9}$	-0.087	0.533	0.540	0.317	-0.044	0.340	0.342	0.256
	$\widehat{\theta}_{DD}$	0.038	0.217	0.221	0.176	-0.049	0.202	0.207	0.162
n=500	$\widehat{\theta}_{K=3}$	-0.037	0.263	0.265	0.199	-0.005	0.225	0.225	0.175
	$\widehat{\theta}_{K=6}$	-0.037	0.264	0.267	0.199	-0.006	0.225	0.225	0.175
	$\widehat{\theta}_{K=9}$	-0.037	0.265	0.267	0.199	-0.006	0.226	0.226	0.175
	$\widehat{\theta}_{DD}$	0.035	0.154	0.158	0.125	-0.035	0.133	0.138	0.109
n=1000	$\widehat{\theta}_{K=3}$	-0.009	0.165	0.165	0.128	-0.005	0.144	0.144	0.113
	$\widehat{\theta}_{K=6}$	-0.009	0.166	0.166	0.129	-0.004	0.143	0.143	0.113
	$\widehat{\theta}_{K=9}$	-0.009	0.166	0.166	0.129	-0.005	0.144	0.144	0.113

Table 4: Comparison of data-driven estimator $(\widehat{\theta}_{DD})$
with cross-validation $(\widehat{\theta}_{CV})$ estimators : $\rho = 0.1$

$\rho = 0.1$		Error type: $N(0,1)$				Error type: t_3			
Est.		Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE
n=250	$\widehat{\theta}_{DD}$	-0.005	0.341	0.342	0.273	0.001	0.286	0.286	0.226
	$\widehat{\theta}_{CV}$	-0.167	0.430	0.461	0.349	-0.144	0.342	0.371	0.289
n=500	$\widehat{\theta}_{DD}$	-0.003	0.242	0.242	0.191	0.017	0.216	0.217	0.172
	$\widehat{\theta}_{CV}$	-0.083	0.267	0.280	0.219	-0.063	0.239	0.247	0.193
n=1000	$\widehat{\theta}_{DD}$	0.002	0.175	0.175	0.139	0.001	0.150	0.150	0.120
	$\widehat{\theta}_{CV}$	-0.050	0.181	0.188	0.149	-0.052	0.158	0.166	0.132

Table 5: Comparison of data-driven estimator ($\widehat{\theta}_{DD}$)
with cross-validation ($\widehat{\theta}_{CV}$) estimators : $\rho = 0.5$

$\rho = 0.5$		Error type: $N(0,1)$				Error type: t_3			
Est.		Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE
n=250	$\widehat{\theta}_{DD}$	0.024	0.333	0.334	0.265	0.030	0.279	0.281	0.222
	$\widehat{\theta}_{CV}$	-0.118	0.447	0.462	0.319	-0.097	0.351	0.364	0.268
n=500	$\widehat{\theta}_{DD}$	0.019	0.235	0.235	0.188	0.035	0.215	0.218	0.171
	$\widehat{\theta}_{CV}$	-0.065	0.260	0.268	0.208	-0.040	0.238	0.241	0.189
n=1000	$\widehat{\theta}_{DD}$	0.022	0.168	0.169	0.134	0.018	0.143	0.144	0.114
	$\widehat{\theta}_{CV}$	-0.031	0.180	0.182	0.143	-0.032	0.155	0.158	0.125

Table 6: Comparison of data-driven estimator ($\widehat{\theta}_{DD}$)
with cross-validation ($\widehat{\theta}_{CV}$) estimators : $\rho = 0.9$

$\rho = 0.9$		Error type: $N(0,1)$				Error type: t_3			
Est.		Bias	Std. Dev.	RMSE	MAE	Bias	Std. Dev.	RMSE	MAE
n=250	$\widehat{\theta}_{DD}$	0.053	0.298	0.303	0.243	0.055	0.259	0.264	0.210
	$\widehat{\theta}_{CV}$	-0.113	0.503	0.515	0.323	-0.076	0.333	0.341	0.255
n=500	$\widehat{\theta}_{DD}$	0.038	0.218	0.221	0.177	-0.049	0.202	0.207	0.162
	$\widehat{\theta}_{CV}$	-0.057	0.264	0.270	0.203	-0.032	0.235	0.237	0.181
n=1000	$\widehat{\theta}_{DD}$	0.035	0.154	0.158	0.125	0.035	0.133	0.138	0.109
	$\widehat{\theta}_{CV}$	-0.024	0.175	0.177	0.138	-0.016	0.149	0.149	0.118

References

[1] Ahn, H. and J. L. Powell (1993), "Semiparametric estimation of censored selection models with a nonparametric selection mechanism",. *Journal of Econometrics* 58, 3-29.

[2] Boente, G. and R. Fraiman (1995), "Asymptotic distribution of data-driven smoothers in density and regression estimation under dependence", The Canadian Journal of Statistics, 23(4), 383-397.

[3] Cosslett, S. R. (1991), "Semiparametric estimation of a regression model with sample selectivity", In W. A. Barnett, J. L. Powell and G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, 175–97. Cambridge: Cambridge University Press.

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

[5] Delgado, M. A. and W. González Manteiga (2001), "Significance Testing in Nonparametric Regression Based on the Bootstrap," *Annals of Statistics*, 29(5), 1469–1507.

[6] Dony, J. and D. M. Mason (2008), "Uniform in Bandwidth Consistency of Conditional U-statistics", *Bernoulli*, 4, 1108-1133.

[7] Einmahl, J. H. J., and D. M. Mason (2005), "Uniform in Bandwidth Consistency of Kernel-Type Function Estimators," *Annals of Statistics*, 33(3), 1380–1403.

[8] Escanciano, J. C., D. T. Jacho-Chávez and A. Lewbel (2012), "Uniform Convergence of Weighted Sums of Non- and Semi-parametric Residuals for Estimation and Testing", Unpublished manuscript.

[9] Härdle, W., P. Hall and H. Ichimura, (1993), "Optimal semiparametric estimation in single index models", *Annals of Statistics* 21, 1, 157-178.

[10] Hong, S-Y (1999), "Automatic bandwidth choice in a semiparametric regression model", *Statistica Sinica* 9, 775-794.

[11] Ichimura, H. (1993), "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single Index Models," *Journal of Econometrics*, 58(1-2), 71–120.

[12] Ichimura, H. and L. Lee (1991), "Semiparametric least squares estimation of multiple index models: Single equation estimation", In W. A. Barnett, J. L. Powell and

G. Tauchen (Eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press.

[13] Klein, R. and R. Spady (1993): "An Efficient Semiparametric Estimator for Discrete Choice Models," *Econometrica*, 61, 387-421.

[14] Li, D. and Q. Li (2010): "Nonparametric/semiparametric Estimation and Testing of Econometric Models with Data Dependent Smoothing Parameters," *Journal of Econometrics*, 157(1), 179–190.

[15] Li, Q. and J.M. Wooldridge (2002), "Semiparametric estimation of partially linear model for dependent data with generated regressors", *Econometric Theory,* 18, 625-645.

[16] Martins-Filho, C. and P. Saraiva (2012), "On Asymptotic Normality of the Local Polynomial Regression Estimator with Stochastic Bandwidths", *Communications in Statistics - Theory and Methods*, 41, 1052-1068.

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

[18] Nickl R. and B. M. Pötscher (2007), "Bracketing Metric Entropy Rates and Empirical Central Limit Theorems for Function Classes of Besov- and Sobolev-Type", *Journal of Theoretical Probability*, 20, 177-199.

[19] Pagan, A. and A. Ullah (1999), "Nonparametric econometrics", Cambridge University Press: Cambridge, UK.

[20] Powell, J. L. (2001), "Semiparametric estimation of censored selection models", In C. Hsiao, K. Morimune and J. Powell (Eds.), *Nonlinear Statistical Modeling*, 165–96. Cambridge: Cambridge University Press.

[21] Powell, J. L., J. H. Stock and T. M. Stoker (1989), "Semiparametric Estimation of Index Coefficients", *Econometrica* 57, 1403-1430.

[22] Robinson, P.M. (1988), "Root-N consistent nonparametric regression", *Econometrica* 56, 931-954.