Informational Content of Special Regressors in Heteroskedastic Binary Response Models[1]

Songnian Chen, Shakeeb Khan and Xun Tang[2]

This Version: December 2, 2015

We quantify the informational content of special regressors in heteroskedastic binary response models with median-independent or conditionally symmetric errors. Based on (Lewbel 1998), a special regressor is one that is additively separable from all other components in the latent payoff and is independent from the error term conditional on all other regressors. We measure the informational content using two criteria: the set of regressor values that help to identify the coefficients in the latent payoff as in (Manski 1988); and the Fisher information of coefficients as in (Chamberlain 1986). We find that under median-independent errors the presence of a special regressor that is conditionally independent from the error term does not increase the identifying power or lead to positive information for coefficients, even though it helps to recover the average structural function. For conditionally symmetric errors, the presence of a special regressor improves the identifying power by the criterion of (Manski 1988), and the Fisher information for coefficients is positive under mild conditions. We also propose two new estimators for coefficients in a binary response model under conditionally symmetric errors and a special regressor.

Key words: Binary response, heteroskedasticity, identification, information, median independence, conditional symmetry

JEL codes: C14, C21, C25

[1]We are grateful to Brendan Kline, Arthur Lewbel, Jim Powell, Frank Schorfheide and Haiqing Xu for useful feedback. We also thank conference and seminar participants at Cambridge, CEMFI, China ES Meetings, Cornell, CREST, Northwestern, Pompeu Fabra, SETA (Taiwan), Texas Econometrics Camp, Tilburg, UCL, UT Austin and Yale for comments. We thank Bingzhi Zhao, Yichong Zhang and Michelle Tyler for capable research assistance. The usual disclaimer applies.

[2]CHEN: Department of Economics, Hong Kong University of Science and Technology, email: snchen@hst.hk; KHAN: Department of Economics, Duke University, email: shakeebk@duke.edu; TANG: Department of Economics, Rice University, email: xun.tang@rice.edu.

1 Introduction

In this paper we explore the informational content of a special regressor in binary choice models. A special regressor is one that is additively separable from all other components in the latent payoff and that satisfies an exclusion restriction (i.e., it is independent from the error term conditional on all other regressors). In this paper, our definition of a special regressor per se does not require it to satisfy any "large support" requirement. We examine how a special regressor contributes to the identification and the Fisher information of coefficients in semiparametric binary response models with heteroskedastic errors. We focus on the role of special regressors in two models where errors are median-independent or conditionally symmetric, respectively.

Special regressors arise in various social-economic contexts. (Lewbel 2000) used a special regressor to recover coefficients in semiparametric binary response models where heteroskedastic errors are mean-independent from regressors. He showed coefficients for all regressors and the error distribution are identified up to scale, provided that the support of special regressor is large enough. (Lewbel 2000) then proposed a two-step inverse-density-weighted estimator. Since then, arguments based on special regressors have been used to identify structural micro-econometric models in a variety of contexts. These include multinomial-choice demand models with heterogeneous consumers (Berry and Haile 2010); static games of incomplete information with player-specific regressors excluded from interaction effects (Lewbel and Tang 2015); and matching games with unobserved heterogeneity (Fox and Yang 2012).

Using a special regressor to identify coefficients in binary response models with heteroskedastic errors typically requires additional conditions on the support of the special regressor. For instance, in the case with mean-independent errors, identification of linear coefficients requires the support of special regressors to be at least as large as that of errors. (Khan and Tamer 2010) argued that point identification of coefficients under mean-independent errors is lost whenever the support of special regressor is bounded.[3] They also showed that when the support of a special regressor is unbounded, the Fisher information for coefficients is zero when the second moment of regressors is finite.

The econometrics literature on semiparametric binary response models has largely been silent about how to use special regressors in combination with alternative stochastic restrictions on errors that require less stringent conditions on the support of special regressors. (Magnac and Maurin 2007) introduced a new restriction on the tail behavior of the latent utility distribution outside the support of special regressors. They established the identification of coefficients under such restrictions. The tail condition they used is not directly linked to more conventional stochastic restrictions on heteroskedastic errors, such as median independence or conditional symmetry. (We show in Appendix B that

[3] They showed in a stylized model that there is no informative partial identification result for the intercept in this case.

the tail condition in (Magnac and Maurin 2007) and the conditional symmetry considered in our paper are non-nested.) We show the information for coefficients is positive in our model under conditional symmetry with a special regressor.

We contribute to the literature on binary choice models with several findings. First, we quantify the change in the identifying power of the model due to the presence of special regressors under median-independent or conditionally symmetric errors. This is done following the approach used in (Manski 1988), which involves comparing the set of states where the conditional choice probabilities can be used for distinguishing true coefficients from other elements in the parameter space. For the model with median-independent errors, we find that further restricting one of the regressors to be special does not improve the identifying power for coefficients. For the model with conditionally symmetric errors, we find that using a special regressor does add to the identifying power for coefficients in the sense that it leads to an additional set of (paired states) that can be used for recovering the true coefficients. This is a surprising insight, because (Manski 1988) showed that, in the absence of a special regressor, the stronger restriction of conditional symmetry adds no identifying power relative to the weaker restriction of median independence.

Second, we show how the presence of a special regressor contributes to the information for coefficients in these two semiparametric binary response models with heteroskedastic errors. For models with median-independent errors, we find the Fisher information for coefficients is zero even when one of the regressors is special. In comparison, for models with conditionally symmetric errors, we find the presence of a special regressor does yield positive information for coefficients. We also propose two consistent estimators for linear coefficients when errors are conditionally symmetric. These two results seem to suggest that there exists a link between the two distinct ways of quantifying informational content in such a semiparametric model: the set of states that help identify the true coefficients in (Manski 1988), and the Fisher information for coefficients in semiparametric binary response models in (Chamberlain 1986).

Our third set of results (Section 3.3) provide a more positive perspective on the role of special regressors in structural analyses. We argue that, even though a special regressor does not add to the identifying power or information for coefficients when heteroskedastic errors are median-independent, it is instrumental for recovering the average structural function, as long as the support of the special regressor is large enough.

This paper contributes to a broad econometrics literature on the identification, estimation and information of semiparametric limited response models with heteroskedastic errors. A partial list of other papers that discussed related topics include (Chamberlain 1986), (Chen and Khan 2003), (Cosslett 1987), (Horowitz 1992), (Khan 2013), (Magnac and Maurin 2007), (Manski 1988) and (Zheng 1995) (which studied semiparametric binary response models with various specifications of heteroskedastic errors); as well as (Andrews 1994), (Newey and McFadden 1994), (Powell 1994) and (Ichimura and Lee 2010) (which discussed asymptotic properties of semiparametric M-estimators).

2 Preliminaries

Consider a binary response model:

$$Y = 1\{X\beta - V \geq \epsilon\} \tag{1}$$

where $X \in \mathbb{R}^k, V \in \mathbb{R}$ and $\epsilon \in \mathbb{R}$, and the first coordinate in X is a constant. We use upper-case letters for random variables and lower-case letters for their realized values. Let F_R, f_R and Ω_R denote the distribution, the density and the support of a random vector R respectively; let $F_{R_1|R_2}$, $f_{R_1|R_2}$ and $\Omega_{R_1|R_2}$ denote the conditional distribution, density and support in the data-generating process (DGP); let $F_{R_1|r_2}$ be shorthand for $F_{R_1|R_2=r_2}$ and likewise for $f_{R_1|r_2}$ and $\Omega_{R_1|r_2}$. Assume the marginal effect of V is known to be negative, and set it to -1 as a scale normalization. We maintain the following condition throughout the paper.

Assumption 2.1 *(Special Regressor) V is independent from ϵ given any $x \in \Omega_X$.*

For the rest of the paper, we sometimes refer to this condition as an "exclusion restriction", and use the terms "special regressors" and "excluded regressors" interchangeably. Let Θ be the parameter space for $F_{\epsilon|X}$ (i.e. Θ is a collection of all conditional distributions of errors that satisfy the model restrictions imposed on $F_{\epsilon|X}$). Let $Z \equiv (X, V)$ denote the vector of regressors reported in the data. The distribution $F_{V|X}$ and the conditional choice probabilities $\Pr(Y = 1 \mid Z)$ are both directly identifiable from the data and considered known in the discussion about identification. Let $p(z)$ denote $\Pr(Y = 1 \mid Z = z)$. Let (Z, Z') be a pair of independent draws from the same marginal distribution F_Z. Assume the distribution of Z has positive density with respect to a σ-finite measure, which consists of the counting measure for discrete coordinates and the Lebesgue measure for continuous coordinates.

To quantify the informational content, we first follow the approach in (Manski 1988). For a generic pair of coefficients and a nuisance distribution $(b, G_{\epsilon|X}) \in \mathbb{R}^k \otimes \Theta$, define $\xi\left(b, G_{\epsilon|X}\right) \equiv \{z : p(z) \neq \int 1\left(\epsilon \leq xb - v\right) dG_{\epsilon|x}\}$ and $\tilde{\xi}\left(b, G_{\epsilon|X}\right) \equiv$

$$\left\{(z, z') : (p(z), p(z')) \neq \left(\int 1\left(\epsilon \leq xb - v\right) dG_{\epsilon|x}, \int 1\left(\epsilon \leq x'b - v'\right) dG_{\epsilon|x'}\right)\right\}. \tag{2}$$

In words, the set $\xi(b, G_{\epsilon|X})$ consists of states for which the conditional choice probabilities implied by $(b, G_{\epsilon|X})$ differ from those in the actual data-generating process (DGP) $(\beta, F_{\epsilon|X})$. In comparison, the set $\tilde{\xi}$ in (2) consists of paired states where implied conditional choice probabilities differ from those in the actual DGP. We say β is *identified relative to* $b \neq \beta$ if

$$\text{either } \int 1\{z \in \xi(b, G_{\epsilon|X})\} dF_Z > 0 \text{ or } \int 1\{(z, z') \in \tilde{\xi}(b, G_{\epsilon|X})\} dF_{(Z,Z')} > 0$$

for all $G_{\epsilon|X} \in \Theta$.

As is clear from this definition, the identification of β depends on the restrictions that define Θ. In Sections 3 and 4, we discuss identification of β when Assumption 2.1 is combined with *one* of the two stochastic restrictions on errors below:

Assumption 2.2 *(Median Independence) For all x, ϵ is continuously distributed with positive density in a neighborhood around 0, and $Med(\epsilon|x) = 0$.*

Assumption 2.3 *(Conditional Symmetry) For all x, ϵ is continuously distributed with positive density over the support $\Omega_{\epsilon|x}$ and $F_{\epsilon|x}(t) = 1 - F_{\epsilon|x}(-t)$ for all $t \in \Omega_{\epsilon|x}$.*

We discuss the Fisher information for β under these assumptions together with Assumption 2.1 in Section 3.2 and 4.2. This amounts to finding smooth parametric submodels which are nested in the semiparametric models and which have the least Fisher information for β. Following (Chamberlain 1986), we define the semiparametric efficiency bound as follows. Let μ denote a measure on $\{0,1\} \otimes \Omega_Z$ such that $\mu(\{0\} \otimes \Omega_0) = \mu(\{1\} \otimes \Omega_0) = F_Z(\Omega_0)$, where Ω_0 is a Borel subset of Ω_Z. A *path* that goes through $F_{\epsilon|X}$ is a function $\lambda(\varepsilon, x; \delta)$ such that $\lambda(\varepsilon, x; \delta_0) = F_{\epsilon|x}(\varepsilon)$ for some $\delta_0 \in \mathbb{R}$, and $\lambda(.,.;\delta) \in \Theta$ for all δ in an open neighborhood around δ_0. Let $f_\lambda(y \mid z; b, \delta)$ denote the probability mass function of Y conditional on z and given coefficients b and a nuisance parameter $\lambda(.,.;\delta)$. A *smooth* parametric submodel is characterized by a path λ such that there exists $\{(\psi_k)_{k \leq K}, \psi_\lambda\}$ such that

$$f_\lambda^{1/2}(y \mid z; b, \delta) - f_\lambda^{1/2}(y \mid z; \beta, \delta_0) = \textstyle\sum_k \psi_k(y,z)(b - \beta) + \psi_\lambda(y,z)(\delta - \delta_0) + r(y, z; b, \delta)$$

with

$$(\|b - \beta\| + \|\delta - \delta_0\|)^{-2} \int r^2(y, z; b, \delta) d\mu \to 0 \text{ as } b \to \beta \text{ and } \delta \to \delta_0.$$

The *path-wise partial information* for the k-th coordinate in β is

$$I_{\lambda,k} \equiv \inf_{\{(\alpha_j)_{j \neq k}, \alpha_\lambda\}} 4 \int \left(\psi_k - \textstyle\sum_{j \neq k} \alpha_j \psi_j - \alpha_\lambda \psi_\lambda\right)^2 d\mu. \tag{3}$$

The *information* for β_k is the infimum of $I_{\lambda,k}$ over all smooth parametric submodels λ.

3 Exclusion Restriction and Median Independence

This section discusses the identification of and the information for β in heteroskedastic binary response models under Assumption 2.1 and 2.2. The model differs from that in (Manski 1988), (Horowitz 1992) and (Khan 2013) in that one of the regressors V is independent from the error term conditional on the other regressors X. It also differs from that considered in (Lewbel 2000) and (Khan and Tamer 2010), for the error term is median-independent rather than mean-independent from X.

3.1 Identification

Our first finding is that the exclusion restriction on a special regressor (Assumption 2.1) does not help with identifying β under median-independent errors (Assumption 2.2). We show the set of $z \equiv (x, v)$ that helps to identify β relative to any $b \neq \beta$ remains the same with or without a special regressor.

Proposition 1 *Suppose Assumption 2.1 and 2.2 hold in (1). Then β is identified relative to b if and only if* $\Pr(Z \in Q_b) > 0$, *where* $Q_b \equiv \{z : x\beta \leq v < xb \text{ or } xb \leq v < x\beta\}$.

For a model with median independent errors but no special regressor ($F_{\epsilon|Z}(0) = 1/2$ where the distribution of ϵ depends on V and X), (Manski 1988) showed that Q_b is the set of states that help to identify β relative to $b \neq \beta$. Therefore, a main conclusion from Proposition 1 is that under median-independent errors, the presence of a special regressor does not help with identifying β in the sense that the set of states that help to identify β relative to any $b \neq \beta$ remains the same.

Proposition 1 is based on two facts. First, if the states in Q_b help to identify β relative to b under median independent errors without a special regressor as in (Manski 1988), then they must also do so when a special regressor is added. Second, if $\Pr(Z \in Q_b) = 0$, then there exists some $G_{\epsilon|X} \neq F_{\epsilon|X}$ that satisfies Assumption 2.1 and Assumption 2.2 and, when combined with $b \neq \beta$, generates the same conditional choice probabilities for all states as the actual DGP. Unlike the result in (Manski 1988), the proof of Proposition 1 requires the construction of a nuisance parameter $G_{\epsilon|X}$ that also satisfies Assumption 2.1.

By the same argument as in (Manski 1988), the following support conditions are sufficient for point-identifying β under Assumption 2.1 and 2.2.

Assumption 3.1 *(Sufficient Variation) For all x, V is continuously distributed with positive density in a neighborhood around $x\beta$.*

Assumption 3.2 *(Full Rank)* $\Pr(X\gamma \neq 0) > 0$ *for all non-zero vectors* $\gamma \in \mathbb{R}^k$.

We discuss these conditions and how they help to point identify β under Assumption 2.1 and 2.2 in Appendix B.

3.2 Zero Fisher Information

We now show the information for β under Assumption 2.1 and 2.2 is zero when Z has finite second moments. In addition to Section 3.1, our finding in this subsection provides

an alternative way to compare the information for β under median-independent errors with or without the exclusion restriction in Assumption 2.1.

Assumption 3.3 *(Regularity) For each* $\left(b, G_{\epsilon|X}\right)$ *in the interior of the parameter space, there exists a measurable function* $q : \{0,1\} \otimes \Omega_Z \to \mathbb{R}$ *such that*

$$\left| \left[\frac{\partial}{\partial B} f^{1/2}\left(y, z; B, G_{\epsilon|X}\right) \right]_{B=\tilde{b}} \right| \leq q(y,z)$$

for all $\tilde{b}$ *in an neighborhood around* b*; and* $\int q^2(y,z)d\mu < \infty$.

Assumption 3.3 is needed to establish mean-square differentiability of the square-root likelihood of (y,x) with respect to the linear coefficient for each $G_{\epsilon|X}$. The parameter space Θ for the distribution of ϵ given X satisfies Assumption 2.1, 2.2 and 3.3. We show that a set of paths similar to those used in Theorem 5 of (Chamberlain 1986) (which considered binary response models with median-independent errors but no special regressor) leads to zero information for β under Assumption 2.1 and 2.2. Specifically, define a set of paths with the following form:

$$\lambda(\varepsilon, x; \delta) \equiv F_{\epsilon|x}(\varepsilon)\left[1 + (\delta - \delta_0)\, h(\varepsilon, x)\right], \tag{4}$$

where $F_{\epsilon|X}$ is the actual conditional distribution in the DGP; and $h : \mathbb{R}^{K+1} \to \mathbb{R}$ is continuously differentiable, is zero outside some compact set and satisfies $h(0,x) = 0$ for all $x \in \mathbb{R}^K$.

Due to an argument similar to (Chamberlain 1986), $\lambda(.,.;\delta)$ in (4) is in Θ for δ close enough to δ_0. Besides, $f_\lambda^{1/2}(.;b,\delta)$ is mean-square differentiable at $(b,\delta) = (\beta, \delta_0)$ with:

$$\begin{aligned}
\psi_k(y,z) &\equiv \frac{1}{2}\left\{ yF_{\epsilon|x}(w)^{-1/2} - (1-y)\left[1 - F_{\epsilon|x}(w)\right]^{-1/2} \right\} f_{\epsilon|x}(w) x_k; \text{ and} \\
\psi_\lambda(y,z) &\equiv \frac{1}{2}\left\{ yF_{\epsilon|x}(w)^{-1/2} - (1-y)\left[1 - F_{\epsilon|x}(w)\right]^{-1/2} \right\} F_{\epsilon|x}(w) h(w,x)
\end{aligned}$$

where w is shorthand for $x\beta - v$. Note the excluded regressor v is dropped from $F_{\epsilon|x}$ and $f_{\epsilon|x}$ due to Assumption 2.1.

Proposition 2 *Suppose Assumptions 2.1, 2.2, 3.1, 3.2 and 3.3 hold in (1),* Z *has finite second moments and* $\Pr(X\beta = V) = 0$. *Then the information for* β_k *is zero for all* $k = 1, 2, ..., K$.

The existence of a special regressor V restricts admissible parametric paths in the model, but such restrictions are not sufficient for raising the minimum path-wise information above zero. We omit the formal proof of Proposition 2 because of its similarity to

that of Theorem 5 in (Chamberlain 1986). Instead, we provide a heuristic argument to illustrate the intuition and main difference.

In (Chamberlain 1986), the information for β_k, when $F_{\epsilon|Z}$ is median-independent but depends on V as well as X, takes the form of:

$$\inf_{\lambda\in\Lambda} 4\int \phi(z)\left[f_{\epsilon|z}(w)\left(x_k - \textstyle\sum_{j\neq k}\alpha_j^* x_j\right) - \alpha_\lambda^*\lambda_\delta(w,x;\delta_0)\right]^2 dF_Z$$

where $\phi(z) \equiv \left[F_{\epsilon|z}(w)(1-F_{\epsilon|z}(w))\right]^{-1} \geq 0$ with $w \equiv x\beta - v$; $\{(\alpha_j^*)_{j\neq k},\ \alpha_\lambda^*\}$ solve the minimization problem (3) that defines the path-wise information for λ; and Λ is a set of admissible smooth parametric paths. To show zero information for β_k, (Chamberlain 1986) considered a set of paths $\lambda(\varepsilon,z;\delta) = F_{\epsilon|z}(\varepsilon)\left[1+(\delta-\delta_0)\tilde{h}(\varepsilon,z)\right]$, where $\tilde{h}:\mathbb{R}^{K+2}\to\mathbb{R}$ is continuously differentiable, equals zero outside some compact set and $\tilde{h}(0,z)=0$ for all z. First off, (Chamberlain 1986) noted that by construction the pathwise information in such a λ is bounded above by

$$4\int \phi(z)F_{\epsilon|z}(w)^2\left[b(w,z)-\tilde{h}(w,z)\right]^2 dF_Z \tag{5}$$

where $b(w,z) \equiv f_{\epsilon|z}(w)x_k/F_{\epsilon|z}(w)$. Then (Chamberlain 1986) showed: (i) there exists a continuously differentiable function $a(z)$ arbitrarily close to $b(w,z)$ (recall w is a function of z itself); and (ii) one can construct $\tilde{h}(w,z)$ such that it is arbitrarily close to $a(z)$ (e.g. $\tilde{h}(w,z) = c(w)a(z)$ with $c(w)$ being 1 except in a small neighbor around 0). Together these two arguments imply that one can pick a path indexed by $\tilde{h}$ that pushes the upper bound of this path-wise information (5) to be arbitrarily close to zero.

In comparison, after we impose the additional Assumption 2.1 condition, $F_{\epsilon|z}$ must be independent from v and $b(w,z)$ is restricted to take the form $b(w,x) \equiv f_{\epsilon|x}(w)x_k/F_{\epsilon|x}(w)$. However, the arguments in (i) and (ii) above remain valid despite this specialization. First, $b(w,x)$ can still be arbitrarily approximated by some continuously differentiable $a(z)$ because w is a function of z. Next, for a fixed β, we can write $a(z)$ as $a^*(w,x)$ because there is a one-to-one mapping between w and v once conditional on x. Therefore we can use the same $c(.)$ function as in (Chamberlain 1986) to construct a continuously differentiable function $h(w,x) \equiv c(w)a^*(w,x)$ so that it is zero outside some compact set, and $h(0,x)=0$ for all x. This implies we can always construct a path in the form of (4) that pushes the upper bound on the path-wise information in (5) to be arbitrarily close to zero.

We conclude this subsection with several remarks. First, the zero information for β_k under Assumption 2.1 and Assumption 2.2 is closely related to two facts: there is zero information for β_k under Assumption 2.2 alone; and there is no incremental identifying power for β when Assumption 2.1 and 2.2 both hold. Second, root-n estimator for β is possible when the second moments for regressors are infinite. In such a case, (Khan and Tamer 2010) showed that the parametric rate can be achieved in the estimation of β

under Assumption 2.1 and mean-independent errors. We expect a similar result to hold in the model under Assumption 2.1 and median-independent errors. Third, if there are multiple excluded regressors (i.e., V in Assumption 2.1 is a vector), then after a scale normalization the information for the coefficients of the other components in V is positive, and root-n estimation of these coefficients is possible (say, using the average-derivative approach).

3.3 Counterfactual Prediction and Average Structural Function

The previous subsections show that a special regressor does not improve the identification of or the information for coefficients for the non-special regressors under median-independent errors. We now provide a positive perspective on the role of special regressors by explaining how they help with predicting counterfactual choice probabilities and estimating average structural functions.[4]

(Lewbel 2000) pointed out the special regressor is useful for identifying the heteroskedastic error distribution under mean-independent errors. The same result holds under Assumption 2.1 and 2.2: With β identified, $F_{\epsilon|x}(t)$ can be recovered for all t over the support of $X\beta - V$ given $X = x$ as $\mathbb{E}(Y|X = x, V = x\beta - t)$. This can then be used to predict counterfactual choice probabilities. To see how, let's consider a stylized model of retirement decisions. Let $Y = 1$ if an individual decides to retire and $Y = 0$ otherwise. An individual's decision is determined by:

$$Y = 1\{X_1\beta_1 + X_2\beta_2 - V \geq \epsilon\}$$

where $X \equiv (X_1, X_2)$ are *log age* and *health status* respectively, and V denotes the total market value of the individual's assets. Suppose asset values are uncorrelated with idiosyncratic elements (unobserved family factors such as money or other resources spent on offspring) conditional on age and health. Suppose we want to predict retirement patterns among another population of senior workers not observed in data, who share the same β_1 and $F_{\epsilon|X}$ but differ in the coefficient for health status $\tilde{\beta}_2$ (where $\tilde{\beta}_2 > \beta_2$). Then knowledge of $F_{\epsilon|X}$ helps to at least bound the counterfactual retirement probabilities conditional on (X_1, X_2, V). If the magnitude of the difference between $\tilde{\beta}_2$ and β_2 is also known, then such a counterfactual conditional retirement probability is point-identified for z, provided that the support $\Omega_{V|x}$ is large enough. (That is, the index $x_1\beta_1 + x_2\tilde{\beta}_2 - v$ is within the support of $X_1\beta_1 + X_2\beta_2 - V$ given $X = x$.)

Second, the special regressor helps to identify the average structural function defined in (Blundell and Powell 2003) under the large support condition of V. To see this, recall

[4]We are grateful to Arthur Lewbel for pointing out that the variation in the special regressor helps to recover the average structural function in binary regressions.

that the average structural function is defined as $G(x,v) \equiv \int 1\{\varepsilon \leq x\beta - v\}dF_\epsilon(\varepsilon) = \Pr(\epsilon \leq x\beta - v)$. If $\Omega_{V|x} = \mathbb{R}$ for all $x \in \Omega_X$, then

$$G(x,v) = \int \varphi(s,x,v)dF_X(s)$$

where

$$\varphi(s,x,v) \equiv \mathbb{E}[Y|X = s, V = v + (s-x)\beta] = F_{\epsilon|X=s}(x\beta - v).$$

With β identified, $\varphi(s,x,v)$ can be constructed as long as the support of V spans the real line for all x. If this large support condition fails, then the point identification of $G(x,v)$ is lost at any (x,v) such that there exists $s \in \Omega_X$ where $v + (s-x)\beta$ falls outside of the support $\Omega_{V|X=s}$. This large support condition is not necessary for point identification of coefficients, because the conditions (Assumption 3.1 and 3.2) in Section 3.1, which are sufficient for point identifying β, hold even when regressors have bounded support. Nevertheless, for identifying the average structural function when errors are median-independent, we need the large support condition on the special regressor as defined in (Lewbel 2000).

We propose the following estimator for the average structural function:

$$\hat{G}(x,v) \equiv \textstyle\sum_{i=1}^n \hat{\varphi}(x_i,x,v),$$

where

$$\hat{\varphi}(x_i,x,v) \equiv \frac{\sum_{j\neq i} y_j \mathcal{H}_\sigma\left(x_j - x_i, v_j - (v + (x_i - x)\tilde{\beta})\right)}{\sum_{j\neq i} \mathcal{H}_\sigma\left(x_j - x_i, v_j - (v + (x_i - x)\tilde{\beta})\right)}$$

with $\mathcal{H}_\sigma(.) \equiv \sigma^{-(k+1)}\mathcal{H}(./\sigma^{k+1})$ where $\mathcal{H}$ is a product kernel, and $\tilde{\beta}$ being a first-stage preliminary estimator such as the one defined in Appendix B1, or the maximum score estimator proposed in (Manski 1985).

4 Exclusion Restriction and Conditional Symmetry

This section discusses the identification of and the information for β under Assumption 2.1 and 2.3. That is, we now replace the location restriction of median-independent errors with a stronger location and shape restriction of conditionally symmetric errors in addition to the special regressor. To motivate a model satisfying Assumption 2.1 and 2.3, consider a binary response model where the individual's latent choice-specific payoff is $y_j^* = h_j^*(x) + a_j v + e_j$ for $j \in \{0,1\}$, with $a_1 \neq a_0$ and the joint distribution of (e_1, e_0) being exchangeable and independent from v conditional on x. Then a rational individual chooses $y = 1$ if and only if $h(x) + v\beta + e_1 - e_0 \geq 0$, where $h \equiv h_1^* - h_0^*$, $\beta \equiv a_1 - a_0$ and the error term $e_1 - e_0$ is symmetric around zero. This fits in the model we consider in this section, with the marginal effect of v on the differential payoff $a_1 - a_0$ normalized to -1

when $a_1 < a_0$. The sign of $a_1 - a_0$, if unknown a priori, can be identified from the sign of the partial derivative of the propensity score with respect to the special regressor.[5]

For example, an investment decision model studied in (Lewbel 2007) would fit in this framework under mild conditions. In this case, e_j are unobserved idiosyncratic factors in a firm's profits with or without investment ($j = 1$ or 0); and v is the firm's plant size reported in the data. The plant size is measured as the log of employment in the previous year, and has a different impact on profits depending on the investment decision. Suppose the idiosyncratic noises in payoffs are independently drawn from some distribution that depends on the other observed profit factors in x but not the firm size v. If in addition the latent payoff is additively separable in the firm size (which is implied by any linear index specification), then the model considered in this section is empirically relevant in such a context.

4.1 Identification

First, we show that further requiring the median-independent errors to be conditionally symmetric when there is a special regressor *does* help with the identification of β. For models with no special regressors, (Manski 1988) showed that strengthening median independence into conditional symmetry *does not* add to the identifying power for β. He showed that the sets of states that help to distinguish β from $b \neq \beta$ under both cases are the same. In comparison, our result in the next proposition shows that such an equivalence does not hold in the presence of a special regressor: Replacing Assumption 2.2 with Assumption 2.3 leads to an additional set of *paired* states that help to identify β relative to any $b \neq \beta$. In this sense, having a special regressor does add to the informational content of the model with conditionally symmetric errors.

Let $X \equiv (X_c, X_d)$, with X_c and X_d denoting continuous and discrete coordinates respectively. Let Θ_{CS} denote the parameter space for the distribution of ϵ given X under Assumption 2.1 and 2.3. We need restrictions on Θ_{CS} due to continuous coordinates in X_c.

Assumption 4.1 *(Equicontinuity) For any $\eta > 0$ and (x, ε), there exists $\delta_\eta(x, \varepsilon) > 0$ such that for all $G_{\epsilon|X} \in \Theta_{CS}$,*

$$|G_{\epsilon|\tilde{x}}(\tilde{\varepsilon}) - G_{\epsilon|x}(\varepsilon)| \leq \eta \text{ whenever } \|\tilde{x} - x\|^2 + \|\tilde{\varepsilon} - \varepsilon\|^2 \leq \delta_\eta(x, \varepsilon).$$

[5] In this model, the deterministic parts of latent payoffs $h_j^*(z)$ differ across $j \in \{0, 1\}$, while the marginal distributions of the error term e_j given x are the same. This kind of specification is general enough to subsume popular parametric models (such as multinomial logit or probit). Nonetheless, we acknowledge that such a treatment of the deterministic and unobserved parts in latent payoffs could be less plausible in certain contexts than others.

This condition requires pointwise continuity in (x, ε) to hold with equal variation all over the parameter space Θ_{CS} , in the sense that the same $\delta_\eta(x, \varepsilon)$ is used to satisfy the "δ-η-neighborhood" definition of pointwise continuity at (x, ε) for all elements in Θ_{CS}.[6] Such an equicontinuity condition is a technicality introduced only because of the need to modify the definition of identification in (Manski 1988) when X contains continuous coordinates. A sufficient condition for Assumption 4.1 is that all $G_{\epsilon|X}$ in Θ_{CS} are Lipschitz-continuous in (x, ε) with their modulus uniformly bounded by a finite constant.

To quantify the incremental identifying power due to Assumption 2.3, define:

$$R_b(x) \equiv \{(v_i, v_j) : x\beta < (v_i + v_j)/2 < xb \text{ or } x\beta > (v_i + v_j)/2 > xb\}$$

for any x. Let $F_{V_i,V_j|X}$ denote the joint distribution of V_i and V_j drawn independently from the same marginal distribution $F_{V|X}$. In addition, we also need the joint distribution of V and X_c given X_d to be continuous.

Assumption 4.2 *(Continuity)* (V, X_c) *is continuously distributed conditional on* X_d.

Under Assumption 4.2, if $\Pr(V_i \in \mathcal{A}|(x_c, x_d)) > 0$ for any set $\mathcal{A}$, then $\Pr(V_i \in \mathcal{A}|(\tilde{x}_c, x_d)) > 0$ for $\tilde{x}_c$ close enough to x_c. This does not impose any large support restriction on $F_{V_i,X_c|x_d}$.

Proposition 3 *Under 2.1, 2.3, 4.1 and 4.2,* β *is identified relative to* b *if and only if either (i)* $\Pr(Z \in Q_b) > 0$ *or (ii)* $\Pr(X \in \Omega_{X,b}) > 0$, *where* $\Omega_{X,b} \equiv \{x : \int 1\{(v_i, v_j) \in R_b(x)\} dF_{V_i,V_j|x} > 0\}$.

By Proposition 3, an additional set of states that help to identify β relative to any $b \neq \beta$ under Assumption 2.1 and 2.3 is $\{z : \exists z' \text{ s.t. } x = x' \text{ and either "}(x+x')\beta \leq v+v' < (x+x')b\text{" or "}(x+x')b \leq v+v' < (x+x')\beta\text{"}\}$. As $b \to \beta$, this set converges to $\{z : \exists z' \text{ s.t. } x = x' \text{ and } (x + x')\beta = v + v'\}$, which has a positive measure under the marginal distribution of Z and some mild support conditions implied by Assumption 3.1. This is in contrast with the case under Assumption 2.1 and 2.2, where $\Pr\{Z \in Q_b\} \to 0$ as $b \to \beta$.

It is probably more transparent to illustrate this incremental identifying power by conditioning on non-special regressors. Consider a fixed vector x and $b \neq \beta$ such that $x\beta < xb$. Under Assumption 2.1 and 2.2, the set of v that help us to identify β relative to $b \neq \beta$ is $\{v : x\beta \leq v < xb\}$, which is reduced to a singleton as $b \to \beta$. In comparison, the set of v that contributes to the identification of β relative to $b \neq \beta$ under Assumption 2.1 and 2.3 is $\{v : \exists v' \text{ s.t. } x\beta \leq (v + v')/2 < xb\}$, which does not collapse to a singleton

[6] An alternative way to formulate Assumption 4.1 is that for any $\eta > 0$ and (x, ε), the infimum of $\delta_\eta(x, \varepsilon; G_{\epsilon|X})$ (i.e. the radius of neighborhood around x in the definition of pointwise continuity) over $G_{\epsilon|X} \in \Theta_{CS}$ is bounded below by a positive constant.

as $b \to \beta$. In other words, with conditionally symmetric errors and a special regressor, the subset of states with $v < x\beta$ also contribute to the identification of β, provided there exist $v' \geq xb$ with $(v + v')/2$ is between $x\beta$ and xb.

To see the intuition for Proposition 3, consider a simple model where all components in X are discrete. For a fixed $b \neq \beta$, consider a pair $(z_i, z_j) \in \tilde{Q}_{b,S}$ where

$$\tilde{Q}_{b,S} \equiv \{(z_i, z_j) : x_i = x_j \text{ and } (v_i, v_j) \in R_b(x_i)\}.$$

Then for any $(z_i, z_j) \in \tilde{Q}_{b,S}$, either

$$\begin{aligned} \text{either } & \text{“}x_i\beta - v_i < -(x_j\beta - v_j) \text{ and } x_ib - v_i > -(x_jb - v_j)\text{”} \\ \text{or } & \text{“}x_i\beta - v_i > -(x_j\beta - v_j) \text{ and } x_ib - v_i < -(x_jb - v_j)\text{”}. \end{aligned} \tag{6}$$

In the first case, the actual conditional choice probabilities from the DGP satisfy $p(z_i) + p(z_j) < 1$ while those implied by $b \neq \beta$ and *any* $G_{\epsilon|X} \in \Theta_{CS}$ at z_i and z_j must add up to a number that is greater than 1. This suggests that any pair (z_i, z_j) from $\tilde{Q}_{b,S}$ should help us to identify β from $b \neq \beta$, because the sign of $p(z_i) + p(z_j) - 1$ differs from that of $(x_ib - v_i) + (x_jb - v_j)$. Thus, if condition (ii) in Proposition 3 holds for b and if all coordinates in X are discrete, then $\Pr((Z_i, Z_j) \in \tilde{\xi}(b, G_{\epsilon|X})) > 0$ for all $G_{\epsilon|X} \in \Theta_{CS}$. On the other hand, if both (i) and (ii) fail, then β is not identified to b because some $G_{\epsilon|X} \neq F_{\epsilon|X}$ can be constructed so that $(b, G_{\epsilon|X})$ is observationally equivalent to the true parameters $(\beta, F_{\epsilon|X})$. That is, $(b, G_{\epsilon|X})$ imply conditional choice probabilities identical to the true conditional choice probabilities in the DGP almost everywhere.

Assumption 4.1 and 4.2 are technical conditions on the parameter space for $F_{\epsilon|X}$, which are introduced in order to extend the intuition above to the case with continuous components in X. With continuous X, $\Pr((Z, Z') \in \tilde{Q}_{b,S}) = 0$ for all $b \neq \beta$. However, under Assumption 4.1 and 4.2, (6) holds for paired states in some small “δ-expansion" of $\tilde{Q}_{b,S}$ defined as:

$$\tilde{Q}^{\delta}_{b,S} \equiv \{(z, \tilde{z}) : x_d = \tilde{x}_d \text{ and } \|\tilde{x}_c - x_c\| \leq \delta \text{ and } (v, \tilde{v}) \in R_b(x)\},$$

provided $\delta > 0$ is small enough. To identify β relative from b, it then suffices to require $\tilde{Q}^{\delta}_{b,S}$ to have positive probability for such small δ, which is possible with continuous coordinates in X.

The point identify β under Assumption 2.1 and 2.3, it is sufficient to use Assumption 3.2 and a weaker version of the support condition in Assumption 3.1.

Assumption 4.3 *(Symmetric Positive Density) For any given x, V is continuously distributed with positive density in a neighborhood around $x\beta - c_0(x)$ and a neighborhood around $x\beta + c_0(x)$ for some constant $c_0(x)$.*

Assumption 3.1 is a special case of Assumption 4.3 with $c_0(x) = 0$ for all x. Under Assumption 3.2, for all $b \neq \beta$, there exists a set $\Omega'_{X,b} \subseteq \Omega_X$ with a positive measure such that $x\beta \neq xb$ for all $x \in \Omega'_{X,b}$. Without loss of generality, suppose $x\beta < xb$ and let $\hat{\varepsilon} < x(b-\beta)/2$. Assumption 4.3 implies that the density of V given $X = x$ is positive over two open subintervals of $(x\beta - c_0(x), x\beta - c_0(x) + \frac{\hat{\varepsilon}}{2})$ and $(x\beta + c_0(x), x\beta + c_0(x) + \frac{\hat{\varepsilon}}{2})$ respectively. It then follows that $\int 1\{x\beta < (v_i + v_j)/2 < xb\} dF_{V_i,V_j|x} > 0$. Symmetric arguments show that, for all $x \in \Omega'_{X,b}$ with $x\beta > xb$, Assumption 4.3 implies $\int 1\{xb < (v_i + v_j)/2 < x\beta\} dF_{V_i,V_j|x} > 0$. Thus point-identification of β follows from Proposition 3.

4.2 Positive Fisher Information

Whether a model contains positive Fisher information for a parameter has an impact on its inference procedure. For example, (Andrews and Schafgans 1998) and (Khan and Tamer 2010) proposed rate-adaptive, "studentized" inference procedures in models where there is zero Fisher information for parameters of interests. (Khan and Tamer 2010) showed (in their Theorem 3.1) that in a binary response model with zero Fisher information for coefficients, regular root-n estimation of coefficients requires regressors to have infinite second moments (such as in Cauchy distribution). Their result suggests that a rate-adaptive inference procedure should be used if in the data-generating process the regressors are in fact drawn from a distribution with finite second moments.

We now show that the information for coefficients is positive in a binary response model with a special regressor and conditional symmetric errors. (Zheng 1995) showed that without any special regressor, the information for β is zero in a binary response model with a conditionally symmetric error distribution. In contrast, we show in this subsection that with a special regressor, the conditional symmetry of the error distribution implies positive information for β under mild regularity conditions. In Section 4.3 and 4.4 we build on this result to propose two new estimators for β.

Assumption 4.4 *(Additional Regularity) (i) There exists a constant $c_f > 0$ such that for any x, the conditional density $f_{\epsilon|x}$ is bounded below by c_f over two open neighborhoods around $-c_0(x)$ and $c_0(x)$ respectively (where $c_0(.)$ is defined in Assumption 4.3). (ii) For any Ω_0 such that $\Pr(X \in \Omega_0) > 0$ there exists no non-zero $\alpha \in \mathbb{R}^K$ such that $\Pr(X\alpha = 0 | X \in \Omega_0) = 1$.*

Let Λ consist of paths $\lambda : \Omega_{\epsilon,X} \otimes \mathbb{R} \to [0,1]$ such that: (a) for some $\delta_0 \in \mathbb{R}$, $\lambda(\varepsilon, x; \delta_0) = F_{\epsilon|x}(\varepsilon)$ for all ε, x; (b) for δ in an neighborhood around δ_0, $\lambda(\varepsilon, x; \delta)$ is a conditional distribution of ϵ given X that satisfies:

$$\lambda(\varepsilon, x; \delta) = 1 - \lambda(-\varepsilon, x; \delta) \text{ for all } \varepsilon, x \in \Omega_{\epsilon,X}; \tag{7}$$

and (c) the square-root density $f_\lambda^{1/2}(y,z;b,\delta)$ is mean-square differentiable at $(b,\delta) = (\beta,\delta_0)$, with the pathwise derivative with respect to δ being:

$$\psi_\lambda(y,z) \equiv \frac{1}{2}\left\{ yF_{\epsilon|x}(w)^{-1/2} - (1-y)\left[1-F_{\epsilon|x}(w)\right]^{-1/2}\right\}\lambda_\delta(w,x;\delta_0)$$

where $w \equiv x\beta - v$ and $\lambda_\delta(\varepsilon,x;\delta_0) \equiv \partial\lambda(\varepsilon,x;\delta)/\partial\delta\mid_{\delta=\delta_0}$.

Proposition 4 *Under Assumption 2.1, 2.3, 3.2, 3.3, 4.1-4.4, the information for β_k is positive for all k.*

We now sketch the intuition behind this result. By the properties of μ (the measure on $\{0,1\}\otimes\Omega_Z$ defined in Section 2), we can show the Fisher information for β_k takes the form of $\inf_{\lambda\in\Lambda} 4\int \phi(z)\left[f_{\epsilon|x}(w)\left(x_k - \sum_{j\neq k}\alpha_j^* x_j\right) - \alpha_\lambda^*\lambda_\delta(w,x;\delta_0)\right]^2 dF_Z$ where $\phi(z) \equiv \left[F_{\epsilon|x}(w)(1-F_{\epsilon|x}(w))\right]^{-1} \geq 0$; and $(\alpha_j^*)_{j\neq k}$ and α_λ^* constitute a solution to the minimization problem in (3) that defines path-wise information $I_{\lambda,k}$. To begin with, note that if $I_{\lambda,k}$ were to be zero for any $\lambda\in\Lambda$, it must be the case that $\alpha_\lambda^* \neq 0$. (If $\alpha_\lambda^* = 0$, the pathwise information $I_{\lambda,k}$ under λ would equal that of a parametric model where the actual $F_{\epsilon|X}$ in the DGP is known, and would therefore be positive. This would contradict the claim that $I_{\lambda,k} = 0$.) Since each path λ in Λ needs to satisfy conditional symmetry for δ close to δ_0, $\lambda_\delta(w,x;\delta_0)$ (and consequently its product with the non-zero α_λ^*) must be odd functions in w once x is fixed. At the same time, $f_{\epsilon|x}(w)$ is by construction an even function of w (symmetric in w around 0) given x. Then the pathwise information for β_k under λ amounts to a weighted integral of squared distance between an odd and an even function. Provided the true index $W = X\beta - V$ spans both sides of zero with positive probability, the information for β_k must be positive because an even function can not approximate an odd function well enough to reduce $I_{\lambda,k}$ arbitrarily close to zero.

Showing positive information for a parameter under a given set of semiparametric assumptions is interesting in its own right. For example, (Cosslett 1987), (Chamberlain 1986) and (Magnac and Maurin 2007) addressed such a question for binary response models under different assumptions. To the best of our knowledge, Section 4.2 marks the first effort to study the information of linear coefficients in binary response models with a special regressor and conditionally symmetric errors. Proposition 4 provides a foundation for existence of regular root-n estimators under these conditions.

(Magnac and Maurin 2007) considered binary response models under Assumption 2.1, mean-independent errors and some tail conditions that restrict the truncated expectation of $F_{\epsilon|X}$ outside the support of V given X.[7] They showed that the information for β_k

[7] See equation (5) in Proposition 5 of (Magnac and Maurin 2007) for the tail restriction. This restriction is sufficient for extending the proof of idenification of β in (Lewbel 2000), a model with a special regressor and mean-independent errors, when the support of the excluded regressor V is bounded between $v_L > -\infty$ and $v_H < \infty$.

is positive in such a model. The tail condition in (Magnac and Maurin 2007) is a joint restriction on the location of the support of V and the tail behavior of $F_{\epsilon|X}$ outside the support of V. In comparison, the conditional symmetry condition in Assumption 2.3 is a transparent restriction on the shape of $F_{\epsilon|X}$ over its full support. The conditional symmetry we consider here and the tail conditions in (Magnac and Maurin 2007) are non-nested. (See Appendix B for details.)

4.3 Weighted Least Absolute Deviation (WLAD) Estimator

This section proposes a consistent weighted least absolute deviation (WLAD) estimator when the error satisfies Assumption 2.1 and 2.3. The least absolute deviation approach is useful for estimating binary response models with median-independent errors. For example, the maximum-score estimator in (Manski 1985) can be expressed in the form of a least absolute deviation estimator that is numerically equivalent. For simplicity in exposition, we assume all components in $X \in \mathbb{R}^k$ are continuous. Extension to cases with mixed covariates is straightforward and omitted for brevity.

Let $[.]_- \equiv -\min\{.,0\}$ and $[.]_+ \equiv \max\{.,0\}$. The WLAD estimator is

$$\hat{\beta} \equiv \arg\min_{b \in \mathcal{B}} \hat{H}_n(b),$$

where $\mathcal{B}$ is a compact parameter space for coefficients and

$$\hat{H}_n(b) \equiv \tfrac{1}{n(n-1)} \sum_{j \neq i} K_h\left(x_i - x_j\right) \left\{ \kappa\left(\hat{w}_{i,j} - 1\right) [\varphi_{i,j}(b)]_- + \kappa(1 - \hat{w}_{i,j}) [\varphi_{i,j}(b)]_+ \right\}, \quad (8)$$

$$\varphi_{i,j}(b) \equiv \tfrac{x_i + x_j}{2} b - \tfrac{v_i + v_j}{2},\ \hat{w}_{i,j} \equiv \hat{p}_i + \hat{p}_j \text{ and}$$

$$\hat{p}_l \equiv \hat{p}(z_l) \equiv \frac{\sum_{s \neq l} y_s \mathcal{K}_\sigma(z_s - z_l)}{\sum_{s \neq l} \mathcal{K}_\sigma(z_s - z_l)} \text{ for } l = i, j,$$

where $K_h(.) \equiv h_n^{-k} K(./h_n)$ and $\mathcal{K}_\sigma(.) \equiv \sigma_n^{-(k+1)} \mathcal{K}(./\sigma_n)$ with $K, \mathcal{K}$ being kernel functions and h_n, σ_n bandwidths. (If X contains a discrete component X_d, then simply modify (8) by replacing the matching kernel for this component with an indicator function $1\{x_{i,d} = x_{j,d}\}$.)

Assumption 5.1 *(Weight Function)* $\kappa : \mathbb{R} \to [0,1]$ *with* $\kappa(t) = 0$ *for all* $t \leq 0$ *and* $\kappa(t) > 0$ *for all* $t > 0$. *In addition,* κ *is increasing and bounded over* $[0, +\infty)$, *and is twice continuously differentiable with bounded derivatives in an open neighborhood around* 0.

The weight function, when evaluated at $\hat{w}_{i,j} - 1$, serves as a smooth replacement for the indicator function $1\{\hat{w}_{i,j} \geq 1\}$. To establish consistency of $\hat{\beta}$, we show that $\hat{H}_n \xrightarrow{p} H_0$

uniformly over the parameter space, where

$$H_0(b) = \mathbb{E}\left\{ f(X)\mathbb{E}\left[\kappa(W_{i,j}-1)[\varphi_{i,j}(b)]_- + \kappa(1-W_{i,j})[\varphi_{i,j}(b)]_+ | X_j = X_i = X\right]\right\}, \quad (9)$$

with f being the density of X in the DGP, and $w_{i,j}$ the sum of actual conditional choice probability $p(z_i)$ and $p(z_j)$. The second expectation in (9) is with respect to V_i, V_j given $X_j = X_i = X$ while the first with respect to X. The next proposition shows β is the unique minimizer of H_0 in $\mathcal{B}$.

Proposition 5 *Under Assumption 2.1, 2.3, 3.1, 3.2 and 5.1, $H_0(b) > 0$ for all $b \neq \beta$ and $H_0(\beta) = 0$.*

To see why $H_0(\beta) = 0$, note $\kappa(p_i + p_j - 1)[\varphi_{i,j}(\beta)]_- = 0$ whenever z_i, z_j are such that $x_i = x_j$. This is because under Assumption 5.1 $\kappa(w_{i,j} - 1) > 0$ if and only if $1\{p(z_i) + p(z_j) > 1\}$, which under Assumption 2.1 and 2.3 is equivalent to $2x_i\beta > v_i + v_j$ given $x_i = x_j$. However, $[\varphi_{i,j}(\beta)]_- > 0$ if and only if $(x_i + x_j)\beta = 2x_i\beta < v_i + v_j$ when $x_i = x_j$. By a symmetric argument, $\kappa(1 - p_i - p_j)[\varphi_{i,j}(\beta)]_+ = 0$ if $x_i = x_j$. Hence $H_0(\beta) = 0$. On the other hand, for any $b \neq \beta$, Assumption 3.2 implies $\Pr(X\beta \neq Xb) > 0$. Without loss of generality, suppose $\Pr(X\beta > Xb) > 0$. Assumption 3.1 ensures for all x with $x\beta > xb$ there exists a set of pairs (v_i, v_j) which satisfies $xb - v_i + xb - v_j < 0$ and $x\beta - v_i + x\beta - v_j > 0$ and has positive measure under $F_{V_i,V_j|X}$. Therefore under conditions of the proposition, the expectation of the product $\kappa(W_{i,j} - 1)[\varphi_{i,j}(b)]_-$ conditional on "$V_i + V_j \leq 2X_i\beta$ and $X_j = X_i = x$" is positive. It then follows that $H_0(b) > 0$ for all $b \neq \beta$. To show the consistency of $\hat{\beta}$, we need the following conditions.

Assumption 5.2 *(Parameter Space) β is in the interior of a compact parameter space $\mathcal{B}$.*

Assumption 5.3 *(Smoothness) (i) The density of Z is continuous and bounded away from zero uniformly over its full support.(ii) The propensity score $p(Z)$* and *the density of Z are twice continuously differentiable with Lipschitz continuous derivatives. (iii) $\mathbb{E}\{[Y - p(z)]^2 | z\}$ is continuous in z. (iv) $H_0(b)$ is continuous in b in an open neighborhood around β. (v) For all x_i, $\mathbb{E}\left[\tilde{\varphi}_{i,j}(b)\middle| X_i = x_i, X_j = x_j\right] f(x_j)$ is twice continuously differentiable in x_j around $x_j = x_i$, where*

$$\tilde{\varphi}_{i,j}(b) \equiv \kappa\left(w_{i,j} - 1\right)[\varphi_{i,j}(b)]_- + \kappa(1 - w_{i,j})[\varphi_{i,j}(b)]_+.$$

Assumption 5.4 *(Kernel for Estimating Propensity Scores) (i) $\mathcal{K}$ is the product of $k+1$ univariate kernel functions $\tilde{\mathcal{K}}$, each of which is symmetric around zero, bounded and has compact support, integrates to 1. (ii) $\|t\|^l \mathcal{K}(t)$ is Lipschitz continuous for $0 \leq l \leq 3$.*

Assumption 5.5 *(Bandwidth for Estimating Propensity Scores) $\sigma_n \to 0$ and $n\sigma_n^{k+1} \to \infty$ as $n \to \infty$.*

Assumption 5.6 *(Finite Moments)* $\mathbb{E}\left[\left(\mathcal{C}_{i,j} - \frac{V_i+V_j}{2}\right)^2\right]$ *and* $\mathbb{E}\left[\left(\mathcal{D}_{i,j} - \frac{V_i+V_j}{2}\right)^2\right]$ *are finite, where* $\mathcal{C}_{i,j} \equiv \inf_{b\in\mathcal{B}}(X_i + X_j)b/2$ *and* $\mathcal{D}_{i,j} \equiv \sup_{b\in\mathcal{B}}(X_i + X_j)b/2$.

Assumption 5.7 *(Kernel for Matching)* $K(.)$ *is the product of* k *univariate kernel functions* $\tilde{K}(.)$*, where* $\tilde{K}(.)$ *is bounded over a compact support, symmetric around* 0 *and integrates to one, and the order of* $\tilde{K}$ *is two.*

Assumption 5.8 *(Bandwidth for Matching)* $h_n \propto n^{-\rho}$*, where* $\rho \in (0, 2/k)$.

Proposition 6 *Under Assumption 2.1, 2.3, 3.1, 3.2 and 5.1-5.8,* $\hat{\beta} \xrightarrow{p} \beta$.

The objective function $\hat{H}_n$ in 8 can also be used to construct a contour set estimator for the identified set of coefficients (denoted $\mathcal{B}_I$), when Assumption 2.1 and 2.3 hold but some of the conditions for point identification (e.g., Assumption 3.1 or 3.2) fail. This is because $\hat{H}_n \xrightarrow{p} H_0$ uniformly over the parameter space under appropriate conditions, and H_0 satisfies $H_0(b) = 0$ for all $b \in \mathcal{B}_I$ and $H_0(b) > 0$ for all $b \notin \mathcal{B}_I$ under the conditions in Proposition 3. Specifically, if Condition C.1 in (Chernozhukov, Hong, and Tamer 2007) holds, one can define a contour set estimator for $\mathcal{B}_I$ by $\widehat{\mathcal{B}_I} \equiv \{b \in \mathcal{B} : a_n \hat{H}_n(b) \leq \hat{c}\}$, where $a_n \to \infty$ is such that $\inf_{b\in\mathcal{B}_I} \hat{H}_n(b) = O_p(1/a_n)$ and $\hat{c}$ is proportional to $\ln n$. By Theorem 3.1 in (Chernozhukov, Hong, and Tamer 2007), such a set estimator $\widehat{\mathcal{B}_I}$ is consistent for $\mathcal{B}_I$ in the Hausdorff metric. On the other hand, if in addition a degeneracy condition (Condition C.3 (a) in (Chernozhukov, Hong, and Tamer 2007)) holds, then consistency for $\mathcal{B}_I$ can be achieved by setting $\hat{c} = O_p(1)$ with $\hat{c} \geq \inf_{b\in\mathcal{B}} \hat{H}_n(b)$ with probability one.

To define a root-n asymptotically normal estimator for β based on the idea in Section 4.3, we may replace the weight function in (8) with a data-dependent version that assigns increasing weights to a shrinking neighborhood around zero as $n \to \infty$ (e.g., by dividing the argument of κ with a sequence of bandwidths). This requires introducing an additional smoothing parameter such as that in (Horowitz 1992). Such an estimator essentially minimizes kernel-weighted least absolute deviation (KWLAD), and is therefore a least absolute deviation analog of the kernel-weighted least squares (KWLS) estimator introduced in Section 4.4 below. The KWLAD estimator is computationally more expensive than the KWLS estimator, for KWLAD involves three bandwidths and the same matching kernels as KWLS but does not have a closed form.

4.4 Kernel Weighted Least Squares (KWLS) Estimator

This section introduces a kernel-weighted least squares (KWLS) estimator that has a closed form and converges at the parametric rate under Assumption 2.1 and 2.3. Let $\tau(.)$ be a bounded, non-negative trimming function that is continuous and attains positive

values over a bounded subset of the support of $Z = (X, V)$ where the propensity score is bounded away from 0 and 1. Similar trimming functions are used in (Ahn and Powell 1993) and (Chen and Khan 2003). Define $\tau_i \equiv \tau(z_i)$.

To see how the KWLS estimator works, suppose we have a pair of observations where the x's are the same but the v's differ. Denote such a pair by $(x, v_i), (x, v_j)$. If for such a pair the propensity scores add up to one, then $v_i + v_j = (x_i + x_j)\beta = x\beta$ by Assumption 2.1 and 2.3. The KWLS estimator is

$$\begin{aligned}\hat{\beta}_{KWLS} \equiv & \left[\sum_{i\neq j} \tau_i\tau_j K_1\left(\tfrac{x_i - x_j}{h_{1,n}}\right) K_2\left(\tfrac{\hat{p}_i+\hat{p}_j-1}{h_{2,n}}\right)(x_i+x_j)(x_i+x_j)'\right]^{-1} \times \\ & \left[\sum_{i\neq j} \tau_i\tau_j K_1\left(\tfrac{x_i - x_j}{h_{1,n}}\right) K_2\left(\tfrac{\hat{p}_i+\hat{p}_j-1}{h_{2,n}}\right)(x_i+x_j)(v_i+v_j)\right], \end{aligned} \tag{10}$$

where K_1 is a product kernel; K_2 is a univariate kernel; $h_{1,n}, h_{2,n}$ are sequences of bandwidths that converge to 0 as $n \to \infty$; and $\hat{p}_i, \hat{p}_j$ are kernel estimates of the propensity scores at z_i, z_j. That is, $\hat{p}_i \equiv \hat{p}(z_i) \equiv \hat{h}(z_i)/\hat{f}(z_i)$, where

$$\hat{h}(z_i) \equiv \frac{1}{n}\sum_{l=1}^n \mathcal{K}_\sigma(z_l - z_i)\, y_l \text{ and } \hat{f}(z_i) \equiv \frac{1}{n}\sum_{l=1}^n \mathcal{K}_\sigma(z_l - z_i)$$

with $\mathcal{K}_\sigma$ defined as in Section 4.3.The estimator in (10) implements the intuition above. It uses kernel smoothing to collect the matched pairs of z_i, z_j, and then estimates the coefficient through the best linear fit of $v_i + v_j$ on $x_i + x_j$.

Let $\mathcal{G}(p, x) \equiv \inf\{t : \Pr(\epsilon_i \leq t \mid X_i = x) \geq p\}$. By the conditional symmetry of the distribution of ϵ_i,

$$x_i\beta - v_i = \mathcal{G}(p_i, x_i) \text{ and } v_i - x_i\beta = \mathcal{G}(1 - p_i, x_i).$$

For $s = 1, 2, ..., k+1$, let $\mathcal{G}_s(p, x)$ denote the partial derivative with respect to the s-th argument. Let $\tilde{\mu}_{\tau,i}$ be shorthand for $\mu_\tau(x_i, 1 - p_i)$, where

$$\mu_\tau(x_i, 1-p_i) \equiv \mathbb{E}[\tau(Z_j) \mid X_j = x_i, P_j = 1-p_i] = \tau(x_i, x_i\beta - \mathcal{G}(1-p_i, x_i)) = \tau(x_i, 2x_i\beta - v_i).$$

Define $\Sigma \equiv 4\mathbb{E}[\tau_i\tilde{\mu}_{\tau,i}X_iX_i'f_{X,P}(X_i, 1 - P_i)]$.

Assumption 6.1 *(Non-singularity) The matrix Σ is non-singular.*

Assumption 6.2 *(Kernels for matching) (i) K_1 and K_2 have compact supports, are symmetric around 0, integrate to 1, are twice continuously differentiable and are eighth-order. (ii) $\sup_{t\in\mathbb{R}} h_1^{-1}|K_1(t/h_1)|$, $\sup_{t\in\mathbb{R}} h_2^{-1}|K_2'(t/h_2)|$, $\sup_{t\in\mathbb{R}} h_2^{-1}|K_2''(t/h_2)|$ are $O(1)$ as $h_1, h_2 \to 0$.*

Assumption 6.3 *(Bandwidths for matching) $h_{1,n} \propto n^{-\delta_1}, h_{2,n} \propto n^{-\delta_2}$, where $\delta_2 \in (\frac{1}{12}, \frac{1}{9})$ and $k\delta_1 < (\frac{2}{3} - \delta_2)$.*

Assumption 6.4 *(Smoothness) The functions $\tau, \mathcal{G}$ and the joint density of (X, P) are all $M = 6$ times continuously differentiable with bounded derivatives.*

Assumption 6.5 *(Kernel for estimating propensity scores) (i) $\mathcal{K}$ has compact support, is symmetric around zero, integrates to one, and is twice continuously differentiable. (ii) $\mathcal{K}$ has an m-th order with $m > 4(1+k)$ where $k \equiv \dim(X)$.*

Assumption 6.6 *(Smoothness of population moments) The propensity score $p(z)$ and the density of Z are continuously differentiable of order m with bounded derivatives, where $m > 4(1+k)$.*

Assumption 6.7 *(Bandwidth for estimating propensity scores) $\sigma_n \propto n^{-\gamma/(1+k)}$, where*

$$\tfrac{1+k}{m}\left(\tfrac{1}{3} + \delta_2\right) < \gamma < \tfrac{1}{3} - 2\delta_2.$$

Assumption 6.8 *(Finite moments) Define $v^*(z_i) \equiv 2\tau_i x_i \tilde{\mu}_{\tau,i} \mathcal{G}_1(1-p_i, x_i) f_{X,P}(x_i, 1-p_i)[1, -p_i]$. The function v^* is continuous almost surely; $\int \|v^*(z)\|\, dz < \infty$; and $\exists c > 0$ s.t. $\mathbb{E}[\sup_{\|\eta\| \le c} \|v^*(Z+\eta)\|^4] < \infty$.*

Proposition 7 *Under Assumptions 2.1, 2.3 and 6.1-6.8,*

$$\sqrt{n}\left(\hat{\beta}_{KWLS} - \beta\right) \xrightarrow{d} N\left(0, \Sigma^{-1}\mathbb{E}[\chi_i \chi_i'] \left(\Sigma^{-1}\right)'\right)$$

where

$$\chi_i \equiv 4\tau_i x_i \tilde{\mu}_{\tau,i} \mathcal{G}_1(1-p_i, x_i) f_{X,P}(x_i, 1-p_i)(y_i - p_i).$$

For inference, we propose the use of plug-in estimators for the components in the variance of the limiting distribution Σ and $\mathbb{E}(\chi_i \chi_i')$. It is known that plug-in estimators for these variance components are typically consistent under mild conditions. See, for example, Theorem 8.13 in (Newey and McFadden 1994) and Theorem 4 in (Chen and Khan 2003).

The KWLS estimator does not require minimizing any non-linear objective function, but it uses three bandwidths: two for the matching kernels K_1, K_2 and one for the kernel $\mathcal{K}$ used in estimating $\hat{p}_i, \hat{p}_j$. In comparison, the WLAD estimator requires two bandwidths in K and $\mathcal{K}$. A comparison between the KWLS and the WLAD estimators under Assumption 2.1 and 2.3 is reminiscent of that between Ichimura's two-step local quantile regression estimator in (Ichimura 1993) and the maximum score estimator in (Manski 1985). Both of those estimators are for β in binary response models with median-independent errors but no special regressors. Ichimura's estimator has a closed form and involves an additional choice of bandwidth in the estimation of conditional choice probabilities, while Manski's maximum score estimator has no closed form but does not require any choice of bandwidth.

A third estimator that is valid under Assumption 2.1 and 2.3 is the inverse-density-weighted estimator (IDW) proposed by (Lewbel 2000) for binary response models with a

special regressor and a mean-independent error. This estimator has an advantage of being consistent under a weaker stochastic restriction on the error term (mean independence implied by conditional symmetry around zero). It has a closed form, and only requires the choice of a single smoothing parameter (i.e., bandwidth in kernel estimates for the density of special regressor). The estimator converges at the parametric rate to a normal distribution when the special regressor has a thick tail with infinite second moments. More generally the rate of convergence of this estimator depends on the tail behavior in the distribution of special regressors relative to that of the errors.

In comparison, the KWLS estimator also has a closed form using three bandwidths. It takes a pair-wise form, and is root-n asymptotically normal regardless of relative thickness of the tail in the distribution of regressors. Our simulation results below suggest it yields the smallest mean squared errors in various designs with heteroskedastic as well as homoskedastic errors.

The WLAD estimator does not have a closed form. It requires minimizing an objective function estimated with the choices of two bandwidths. Nonetheless, based on the idea from (Chernozhukov, Hong, and Tamer 2007), the WLAD estimator may be used for consistent estimation of the identified set of coefficients under Assumption 2.1 and 2.3 when the conditions for point identification fail. This advantage of the WLAD estimator over the KWLS and IDW estimators is again reminiscent of the contrast between the maximum score estimator in (Manski 1985) and the local quantile regression estimator in (Ichimura 1993): The latter does not provide a consistent estimator for the identified set for coefficients when the support conditions necessary for point-identification under median-independent errors fail.

5 Simulation

This section presents simulation evidence for the performance of the WLAD estimator, the KWLS estimator and the IDW estimator in (Lewbel 2000). The IDW estimator was introduced in (Lewbel 2000) under Assumption 2.1 and mean independence, which is implied by Assumption 2.1 and Assumption 2.3.

We report performance of these estimators under various designs of data-generating processes. In all designs, $Y = 1\{\alpha + X\beta + V + \epsilon \geq 0\}$ where X, V, ϵ are all scalar variables. We let V be drawn from the standard normal or the standard Laplace distribution; and let (X, ϵ) be both drawn from the standard normal or the standard Laplace. We also include designs where the errors are heteroskedastic. (For example, we call ϵ heter. normal if $\epsilon = (1 + |X|)U$ where U and X are both drawn from a standard normal.) For designs 1-8 we set $(\alpha, \beta) = (0.2, 0.5)$ and summarize the specifications as follows:

V ~ normal			V ~ Laplace		
Design #	X	ϵ	Design #	X	ϵ
1	normal	normal	5	normal	normal
2	Laplace	Laplace	6	Laplace	Laplace
3	normal	heter. normal	7	normal	heter. normal
4	Laplace	heter. Laplace	8	Laplace	heter. Laplace

Designs 9-16 use the same specifications as designs 1-8 respectively but with $(\alpha, \beta) = (0, 1)$. Except for designs with heteroskedastic errors, X, V and ϵ are mutually independent. In the designs with heteroskedastic errors, the variance of ϵ varies with X. It is common in practice to adopt a normal specification for the error terms in simulation designs when there is unknown heteroskedasticity (e.g., (Chen and Khan 2003)). It is known from the literature (e.g., (Horowitz 1993)) that parametric/semiparametric estimators in binary response models are usually less sensitive to the form of error distribution than to the form of heteroskedasticity. As is customary in the literature, we use these simple simulation designs as benchmarks to study estimator performance. In particular, we choose these designs to illustrate the estimator performance when the error term has different tail thickness (the tail of the standard normal distribution is thinner than that of the standard Laplace).

For each choice of sample size $n = 100, 200, 400, 800$ and 1600, we simulate 3000 samples and report the bias, the standard deviation, the root mean-square error and the median absolute deviation of WLAD, KWLS and IDW estimates out of these 3000 replications. For simplicity in implementation, we use second-order Gaussian kernels for estimating the conditional choice probabilities in WLAD and KWLS and the conditional density $f_{V|X}$ in IDW, and for matching the non-excluded regressors X_i, X_j in WLAD and KWLS and the estimated conditional choice probabilities $\hat{p}_i, \hat{p}_j$ in KWLS. Our simulation results show that using higher-order kernels in these places does not change qualitatively the performance of the estimators. Following the Silverman's Rule of Thumb (SRT), we use $\hat{\sigma}_X n^{-1/6}$ and $\hat{\sigma}_V n^{-1/6}$ for estimating the conditional choice probabilities and the conditional density $f_{V|X}$, where $\hat{\sigma}_X, \hat{\sigma}_V$ are sample standard deviations for X, V respectively. (That the constant in SRT equals one is due to the Gaussian kernels we use, and the dimension of explanatory variables is 2.) As a simple benchmark in implementation, we replace $\kappa(.)$ in WLAD with a simple indicator function $1\{. \geq 0\}$, and simply let the trimming function τ in KWLS be the identity function. In WLAD, we use $\sqrt{2}\hat{\sigma}_X n^{-1/5}$ for matching the non-excluded regressors. This is because the use of the matching kernel here is analogous to that used for estimating the univariate density of the difference $X_i - X_j$ around 0, thus we follow SRT with the sample standard deviation of $X_i - X_j$ being $\sqrt{2}\hat{\sigma}_X$. To pick the smoothing parameters for the KWLS estimator, we let $\delta_1 \equiv 1/3$, $\delta_2 \equiv 1/11$ and $\gamma \equiv 1/7$, which jointly respects the inequalities restricting bandwidths in Assumption 6.3 and 6.7 with $m > 8$. We then adjust the Silverman's Rule of Thumb for the matching bandwidth in the form $c_0\hat{\sigma}_p n^{-\delta_2}$ and $c_0\hat{\sigma}_X n^{-\delta_1}$, where $\hat{\sigma}_p$ is the sample standard deviation of $\hat{p}_i$. As is typically called for in multi-step estimators that use first-stage kernel

estimates, we implement undersmoothing by setting $c_0 = 0.7$.[8] We use the fminsearch function in Matlab to solve the minimization problem in the WLAD estimator, with initial values set at $(0.1, 0.1)$.

The speed at which the root mean square error (RMSE) of the estimators diminishes varies across the designs. For example, with $(\alpha, \beta) = (0.2, 0.5)$ and (X, ϵ) both being standard normal, all three estimators are more or less converging at the parametric rate as n increase to 1600 regardless of the distribution of the excluded regressor V (see designs 1 and 5). On the other hand, when $(\alpha, \beta) = (0.2, 0.5)$, X is drawn from Laplace and ϵ has a heteroskedastic Laplace specification, the rates of convergence for the slope estimators appear to be slower than $\sqrt{n}$ (see designs 4 and 8). This suggests that these designs may require larger samples in order to manifest the asymptotic properties of KWLS and IDW estimators.

None of these estimators seem to consistently outperform the others in terms of having the smallest RMSE in large samples: The IDW estimator has the smallest RMSE when $n = 1600$ in some of the designs with homogenous errors (see designs 1 and 5); on the other hand, either the WLAD or the KWLS estimator has the smallest RMSE when the errors are heteroskedastic (see designs 3, 4, 7, 8). In most of the designs considered, the bias for slope coefficients diminishes at a somewhat slower pace in the IDW estimator than in the other two estimators. The KWLS estimator outperforms the other two estimators in terms of RMSE in some of the heteroskedastic designs (7 and 8) and homoskedastic design (6). A notable advantage of the WLAD estimator is that it produces the smallest RMSE for the slope parameter under various homogenous and heteroskedastic designs, *especially* where the tail in the error distribution is thick relative to that of the excluded regressor V (designs 2 and 4).

In general, WLAD and KWLS tend to converge faster than IDW as n increases from 100 to 1600, with the only exceptions being designs 1 and 5 where all estimators demonstrate more or less the same rates of convergence. That the IDW estimator seems to converge faster when X and ϵ are normally distributed (as in designs 1 and 5) conforms with the following observation in (Khan and Tamer 2010): The inverse-density-weighted estimator tends to perform better when the tail of the distribution of the excluded regressor is thick relative to that of the error distribution.

For each of the three estimators, the RMSE for the intercept parameter remains steady regardless of the true values for (α, β). On the other hand, for all three estimators, the RMSEs for the slope estimators are higher when $(a, \beta) = (0, 1)$. This is more notable when the error distribution is homogenous or heterogenous Laplace. Among the three

[8] In our simulations, we experimented with smaller choices of bandwidths for estimating the conditional density in order to reduce the bias (such as $\hat{\sigma}_V n^{-\alpha}$ for $\alpha \in (\frac{1}{5}, \frac{1}{4})$ or $c\hat{\sigma}_X n^{-1/6}$ and $c\hat{\sigma}_V n^{-1/6}$ for $.5 < c < .9$). In addition, we also experimented with undersmoothing in the bandwidth for matching propensity scores by varying c_0 between 0.5 and 0.9. The results are insensitive and comparable to those reported in the paper.

alternatives, the WLAD estimator reports relatively smaller increase in RMSE for the slope parameter in most of the cases. This pattern may be associated with the fact that WLAD is based on a weighted least absolute deviation, and is therefore more robust against outliers when the error distribution has a thick tail.

6 Concluding Remarks

In semiparametric binary response models with heteroskedastic errors, we study how a special regressor, which is additively separable in the latent payoff and independent from errors given the other regressors, contributes to the identifying power and the Fisher information for coefficients. We consider binary choice models with a special regressor and either median-independent errors or conditionally symmetric errors.

We find that with median-independent errors, using a special regressor does not directly add to the identifying power or information for coefficients. Nonetheless it does help with counterfactual prediction and the identification of the average structural function. In contrast, with conditional symmetry in the error distribution, using a special regressor improves the identifying power by the criterion in (Manski 1988), and the information for coefficients becomes positive under mild conditions. In other words, the joint restrictions of conditional symmetry (Assumption 2.3) and exclusion restriction (Assumption 2.1) *together* add the informational content for coefficients, whereas neither of them does so *individually*. Therefore, an interesting alternative interpretation of our results is about the informational content of conditional symmetry with and without excluded regressors. We propose two root-n estimators for binary response models with heteroskedastic but conditionally symmetric errors, and report their decent performances in finite samples.

In this paper we do not calculate the semiparametric efficiency bound on linear coefficients under Assumption 2.1 and 2.3. As discussed in (Newey 1990), the first step for such calculation is to characterize a tangent set, or the mean-square closure of linear combination of scores with respect to the nuisance parameters. The semiparametric efficiency bound is then the inverse of the variance of residuals from projecting the score with respect to linear coefficients on such a tangent set. The first step requires some ad hoc "guess-and-verify" approach that exploits the parametric component of the model; the second step involves original argument based on the specific nature of the semiparametric restriction imposed. All in all, such calculation requires substantial additional work, and is left as a topic for future investigation.

Other directions of future research could include similar exercises for other limited dependent variable models such as censored or truncated regressions, and further exploration of the link between the notion of informational content from the support-based approach in (Manski 1988) and the semiparametric efficiency perspective in (Chamberlain 1986).

Appendix A. Proofs

Proof of Proposition 1. (Sufficiency) Under Assumption 2.1 and Assumption 2.2, $p(x,v) \leq 1/2$ if and only if $x\beta \leq v$. Consider $b \neq \beta$ with $\Pr(Z \in Q_b) > 0$. Without loss of generality, consider some $(x,v) \in Q_b$ with $x\beta \leq v < xb$. Then for any $G_{\epsilon|X} \in \Theta$ (where Θ here in Section 3.1 is the set of conditional distributions that satisfy Assumption 2.1 and Assumption 2.2), we have $\int 1(\epsilon \leq xb - v)\, dG_{\epsilon|x} > 1/2$, which implies $(x,v) \in \xi(b, G_{\epsilon|X})$. Therefore, $\Pr(Z \in \xi(b, G_{\epsilon|X})) > 0$ for such a b and all $G_{\epsilon|X} \in \Theta$. Since $(Z, \tilde{Z})$ is a pair of states drawn independently from the same marginal, this also implies $\Pr((Z,\tilde{Z}) \in \tilde{\xi}(b, G_{\epsilon|X})) > 0$ for such a b and all $G_{\epsilon|X} \in \Theta$.Thus β is identified relative to b.

(Necessity) Consider $b \neq \beta$. Suppose $\Pr(Z \in Q_b) = 0$ so that $sign(V - X\beta) = sign(V - Xb)$ with probability one. Construct a $\tilde{G}_{\epsilon|x}$ so that $\tilde{G}_{\epsilon|x}(t;b) = \mathbb{E}(Y|x, V = xb - t)$ for all t on the support of $Xb - V$ given $X = x$. For t outside the support of $Xb - V$ given $X = x$, define $\tilde{G}_{\epsilon|x}(t;b)$ arbitrarily subject to the requirement that $\tilde{G}_{\epsilon|x}(t;b)$ is monotone in t over the support $\Omega_{\epsilon|x}$. By construction, $\tilde{G}_{\epsilon|x}(xb - v; b) = \mathbb{E}(Y|x, V = v) \equiv p(z)$ for all $z \equiv (x,v)$. If $xb \in \Omega_{V|x}$, then $\tilde{G}_{\epsilon|x}(0;b) = 1/2$ by construction. Otherwise (i.e. zero is outside the support of $V - xb$ given x), construct $\tilde{G}_{\epsilon|x}(.;b)$ outside the support of $Xb - V$ given $X = x$ subject to the requirement that $\tilde{G}_{\epsilon|x}(0;b) = 1/2$. This can be done, because $\Pr(Z \in Q_b) = 0$ implies that $p(x,v) \geq 1/2$ for all (x,v) if and only if $v - xb \geq 0$ for all (x,v). Hence as long as $\Pr(Z \in Q_b) = 0$, there exists $\tilde{G}_{\epsilon|X} \in \Theta$ satisfying Assumption 2.1 and Assumption 2.2 such that $\Pr(Z \in \xi(b, \tilde{G}_{\epsilon|X})) = 0$. Furthermore, with any pair of Z and $\tilde{Z}$ that is drawn independently from the same marginal, that $\Pr(Z \in Q_b) = 0$ implies "$sign(X\beta - V) = sign(Xb - V)$ *and* $sign(\tilde{X}\beta - \tilde{V}) = sign(\tilde{X}b - \tilde{V})$" with probability 1. Thus the distribution $\tilde{G}_{\epsilon|X}$ constructed as above is in Θ and also satisfies $\Pr((Z,\tilde{Z}) \in \tilde{\xi}(b, \tilde{G}_{\epsilon|X})) = 0$. Thus β is not identified relative to b. □

Proof of Proposition 3. (Sufficiency) Proposition 1 shows that β is identified relative to b under Assumption 2.1 and 2.2 whenever (i) holds. It follows immediately that (i) also implies identification of β relative to b under the stronger assumptions of Assumption 2.1 and 2.3. To see how (ii) is sufficient for identification of β relative to b, define $\tilde{Q}_{b,S} \equiv \{(z,\tilde{z}) : \tilde{x} = x \text{ and } (v,\tilde{v}) \in R_b(x)\}$. By construction, for any $(z,\tilde{z}) \in \tilde{Q}_{b,S}$, either "$x\beta - v < \tilde{v} - x\beta$ and $xb - v > \tilde{v} - xb$" or "$x\beta - v > \tilde{v} - x\beta$ and $xb - v < \tilde{v} - xb$". Under Assumption 2.1 and 2.3, this implies that for any $G_{\epsilon|X} \in \Theta_{CS}$ and any $(z,\tilde{z}) \in \tilde{Q}_{b,S}$, either

$$\text{"}F_{\epsilon|x}(x\beta - v) + F_{\epsilon|\tilde{x}}(\tilde{x}\beta - \tilde{v}) < 1 \text{ and } G_{\epsilon|x}(xb - v) + G_{\epsilon|\tilde{x}}(\tilde{x}b - \tilde{v}) > 1\text{"} \tag{11}$$

or

$$\text{"}F_{\epsilon|x}(x\beta - v) + F_{\epsilon|\tilde{x}}(\tilde{x}\beta - \tilde{v}) > 1 \text{ and } G_{\epsilon|x}(xb - v) + G_{\epsilon|\tilde{x}}(\tilde{x}b - \tilde{v}) < 1\text{"}.$$

Thus $\tilde{Q}_{b,S} \subseteq \tilde{\xi}(b, G_{\epsilon|X})$ for any $G_{\epsilon|X} \in \Theta_{CS}$. Next, for any $\delta > 0$, define a "δ-expansion" of $\tilde{Q}_{b,S}$ as:

$$\tilde{Q}^{\delta}_{b,S} \equiv \{(z,\tilde{z}) : \tilde{x}_d = x_d \text{ and } (v,\tilde{v}) \in R_b(x) \text{ and } \|\tilde{x}_c - x_c\| \leq \delta\}.$$

Without loss of generality, suppose all $(z, \tilde{z}) \in \tilde{Q}_{b,S}$ satisfy (11) for all $G_{\epsilon|X} \in \Theta_{CS}$. Then Assumption 4.1 implies that when $\delta > 0$ is small enough, $\|\tilde{x}_c - x_c\|^2$ and $\|(\tilde{x} - x)\,\beta\|^2$ and $\|(\tilde{x} - x)\,b\|^2$ are also small enough so that (11) holds for all $(z, \tilde{z})$ in $\tilde{Q}^{\delta}_{b,S}$ and all $G_{\epsilon|X} \in \Theta_{CS}$. Thus with such a small δ, we have $\tilde{Q}^{\delta}_{b,S} \subseteq \tilde{\xi}(b, G_{\epsilon|X})$ for all $G_{\epsilon|X} \in \Theta_{CS}$. Finally, suppose condition (ii) in Proposition 3 holds for $b \neq \beta$ and some set $\Omega_{X,b}$ with positive measure. Then Assumption 4.2 implies

$$\int 1\{(v_i, v_j) \in R_b(x)\} dF_{V_i|\tilde{x}}(v_j) dF_{V_i|x}(v_i) > 0 \tag{12}$$

for all $(x, \tilde{x})$ with $x \equiv (x_c, x_d) \in \Omega_{X,b}$, $\tilde{x}_d = x_d$ and $\|\tilde{x}_c - x_c\| \leq \tilde{\delta}$, where $\tilde{\delta} > 0$ is small enough. Apply the law of total probability to integrate out $\left(\tilde{X}, X\right)$ on the left-hand side of (12), which then implies that $\Pr((Z, \tilde{Z}) \in \tilde{Q}^{\tilde{\delta}}_{b,S}) > 0$ for such a small $\tilde{\delta}$. Hence for such a $b \neq \beta$, $\Pr((Z_i, Z_j) \in \tilde{\xi}(b, G_{\epsilon|X})) > 0$ for all $G_{\epsilon|X} \in \Theta_{Assumption 2.3}$, and β is identified relative to b.

(Necessity) Consider some $b \neq \beta$ such that (i) $\Pr(Z \in Q_b) = 0$ and (ii) $\Pr(X \in \Omega_{X,b}) = 0$, where $\Omega_{X,b}$ is as defined in the proposition. We now show how to construct a nuisance parameter $G^*_{\epsilon|X}$ which satisfies Assumption 2.1, Assumption 2.3 and Assumption 4.2, and, together with b, are observationally equivalent to the true parameters β and $F_{\epsilon|X}$. Specifically, for each fixed x, construct $G^*_{\epsilon|x}$ through the steps (i)-(iii) below: (i) For any $t > 0$, find v such that $t = xb - v$ and define $G^*_{\epsilon|x}(t) \equiv G^*_{\epsilon|x}(xb - v) \equiv F_{\epsilon|x}(x\beta - v)$. If no such v exists on the support of V given x, then find $v\prime$ such that $t = v' - xb$ and define $G^*_{\epsilon|x}(t) \equiv G^*_{\epsilon|x}(v' - xb) \equiv F_{\epsilon|x}(v' - x\beta)$. (ii) For all $t < 0$, find v' such that $t = v' - xb$ and define $G^*_{\epsilon|x}(t) \equiv G^*_{\epsilon|x}(v' - xb) \equiv F_{\epsilon|x}(v' - x\beta)$. If no such v' exists, then find v such that $t = xb - v$ and define $G^*_{\epsilon|x}(t) \equiv G^*_{\epsilon|x}(xb - v) \equiv F_{\epsilon|x}(x\beta - v)$. (Because $\Pr(X \in Q_b) = 0$ and $\Pr(X \in \Omega_{X,b}) = 0$, $G^*_{\epsilon|x}$ constructed from (i) and (ii) is increasing over its domain.[9] This function $G^*_{\epsilon|x}$ also satisfies the conditional symmetry around zero. To see this, consider $t_1 = -t_2 > 0$ where there exists $xb - v_1 = t_1$. This implies that there exists $v_2 = v_1$ such that $v_2 - xb = t_2$. Then by construction, $G^*_{\epsilon|x}(t_1) + G^*_{\epsilon|x}(t_2) = F_{\epsilon|x}(x\beta - v_1) + F_{\epsilon|x}(v_1 - x\beta) = 1$. On the other hand, if $t_1 = -t_2 > 0$ is such that there exits no v_1 with $xb - v_1 = t_1$ but there exists v_1 with $v_1 - xb = t_1$. This implies there exists no v_2 such that $v_2 - xb = t_2$ but there exists $v_2 = v_1$ with $xb - v_2 = t_2$. Hence $G^*_{\epsilon|x}(t_1) + G^*_{\epsilon|x}(t_2) = F_{\epsilon|x}(v_1 - x\beta) + F_{\epsilon|x}(x\beta - v_1) = 1$.) (iii) Finally, as for the range of t such that there exists no v with $xb - v = t$ or v' with $v' - xb = v$, construct $G^*_{\epsilon|x}$ over this range by extrapolating from the sections constructed in (i) and (ii) while respecting the increasingness and conditional symmetry around zero. It then follows that, for all (x, v),

[9] To see why the increasingness holds, first consider the case with $t_2 > t_1 > 0$ where there exists v_s with $t_s = xb - v_s$ for $s = 1, 2$. This implies $v_1 > v_2$ and $G^*_{\epsilon|x}(t_2) \equiv F_{\epsilon|x}(x\beta - v_2) > F_{\epsilon|x}(x\beta - v_1) \equiv G^*_{\epsilon|x}(t_1)$. Next, consider the case with $t_2 \geq 0 > t_1$ where there exists v_1, v_2 such that $xb - v_2 = t_2$ and $v_1 - xb = t_1$. It then follows from $\Pr\{Z \in Q_b\} = 0$ and $\Pr\{X \in \Omega_{X,b}\} = 0$ that $x\beta - v_2 \geq 0 > v_1 - x\beta$. Hence $G^*_{\epsilon|x}(t_2) \geq \frac{1}{2} > G^*_{\epsilon|}(t_1)$. The ranking between $G^*_{\epsilon|x}(t_s)$ in the other cases with $t_2 > t_1$ follow from the same set of arguments.

either $G^*_{\epsilon|x}(xb-v) = F_{\epsilon|x}(x\beta - v)$, or $G^*_{\epsilon|x}(v-xb) = F_{\epsilon|x}(v-x\beta)$ (which implies the first equality given the conditional symmetry of $F_{\epsilon|X}$ and $G^*_{\epsilon|X}$). This implies that β is not identified relative to b under the conditions of the proposition. □

Proof of Proposition 4. Under Assumption 2.1, 2.3, 4.1, 4.2, 4.3 and 3.2, β is identified relative to all $b \neq \beta$. With μ consisting of a counting measure for $y \in \{0,1\}$ and the probability measure for Z, we can show that path-wise information for β_k under a path $\lambda \in \Lambda$ (denoted by $I_{\lambda,k}$) takes the form

$$4\int \left(\psi_k - \alpha^*_\lambda \psi_\lambda - \textstyle\sum_{j\neq k} \alpha^*_j \psi_j\right)^2 d\mu = 4\int_{\Omega_Z} \frac{\left[f_{\epsilon|x}(w)\left(x_k - \sum_{j\neq k}\alpha^*_j x_j\right) - \alpha^*_\lambda \lambda_\delta(w,x;\delta_0)\right]^2}{F_{\epsilon|x}(w)\left[1-F_{\epsilon|x}(w)\right]} dF_Z \quad (13)$$

where $(\alpha^*_j)_{j\neq k}$ and α^*_λ are constants that solve the minimization problem in the definition of $I_{\lambda,k}$ in (3).

We prove the proposition through contradiction. Suppose $I_{\lambda,k} = 0$ for some $\lambda \in \Lambda$. First off, note α^*_λ must be nonzero for such a λ, because otherwise the path-wise information $I_{\lambda,k}$ would equal the Fisher information for β in a parametric model where the true error distribution $F_{\epsilon|X}$ is known, which is positive. This would lead to a contradiction.

Suppose $I_{\lambda,k} = 0$ for some $\lambda \in \Lambda$ with $\alpha^*_\lambda \neq 0$. For any given x, Assumption 4.3 states the conditional density $f_{V|x}$ is positive around $x\beta - c_0(x)$ and $x\beta + c_0(x)$ for some constant $c_0(x)$. Thus for $\varepsilon^* > 0$ small enough, $W \equiv X\beta - V$ is continuously distributed with positive density over two open intervals $(c_0(x)-\varepsilon^*, c_0(x)+\varepsilon^*)$ and $(-c_0(x)-\varepsilon^*, -c_0(x)+\varepsilon^*)$ given any x. Note these two intervals are symmetric around 0. To simplify exposition, suppose $c_0(x) = 0$ for all x for now, so that these two intervals collapse into one symmetric interval around 0, i.e. $(-\varepsilon^*, \varepsilon^*)$.

Note the integrand in (13) is non-negative by construction. Thus the right-hand side of (13) is bounded below by

$$4\int_{\Omega_X}\int_{-\varepsilon^*}^{\varepsilon^*} \frac{\left[f_{\epsilon|x}(w)\left(x_k - \sum_{j\neq k}\alpha^*_j x_j\right) - \alpha^*_\lambda \lambda_\delta(w,x;\delta_0)\right]^2}{F_{\epsilon|x}(w)\left[1-F_{\epsilon|x}(w)\right]} dF_{W,X}.$$

Differentiating both sides of (7) with respect to δ at δ_0 shows $\lambda_\delta(-\varepsilon, x; \delta_0) = -\lambda_\delta(\varepsilon, x; \delta_0)$ for all x and ε. This implies that $\alpha^*_\lambda \lambda_\delta(w,x;\delta_0)$ is an odd function in w given any x. On the other hand, conditional symmetry of errors implies that $f_{\epsilon|x}(w)\left(x_k - \sum_{j\neq k}\alpha^*_j x_j\right)$ is even in w (i.e. symmetric in w around 0) given any x. Due to Assumption 4.4-(i), $F_{\epsilon|x}(t)^{-1}\left[1-F_{\epsilon|x}(t)\right]^{-1}$ is uniformly bounded between positive constants for all $t \in (-\varepsilon^*, \varepsilon^*)$ and $x \in \Omega_X$. It follows that for any constant $\varphi > 0$,

$$\int_{\Omega_X}\int_{-\varepsilon^*}^{\varepsilon^*} \left[f_{\epsilon|x}(w)\left(x_k - \textstyle\sum_{j\neq k}\alpha^*_j x_j\right) - \alpha^*_\lambda \lambda_\delta(w,x;\delta_0)\right]^2 dF_{W|x}(w) dF_X(x) < \varphi$$

for ε^* sufficiently small. Then for any $\varphi > 0$, there must exist $\mathcal{I} \subset [0, \varepsilon^*) \otimes \Omega_X$ or $\mathcal{I} \subset (-\varepsilon^*, 0] \otimes \Omega_X$ with $\Pr((W, X) \in \mathcal{I}) > 0$ and

$$\left| f_{\epsilon|x}(t) \left(x_k - \textstyle\sum_{j \neq k} \alpha_j^* x_j \right) - \alpha_\lambda^* \lambda_\delta(t, x; \delta_0) \right| < \varphi \tag{14}$$

for all $(t, x) \in \mathcal{I}$. Without loss of generality, suppose $\mathcal{I} \subset [0, \varepsilon^*) \otimes \Omega_X$, and define $\bar{\omega} \equiv \{x : \exists t \text{ with } (t, x) \in \mathcal{I}\}$.

Assumption 4.4-(ii) implies that $\Pr(X_k - \sum_{j \neq k} \alpha_j^* X_j > 0 \mid X \in \bar{\omega}\} > 0$. Consider $\bar{x} \in \bar{\omega}$ with $a(\bar{x}) \equiv \bar{x}_k - \sum_{j \neq k} \alpha_j^* \bar{x}_j > 0$. Thus $f_{\epsilon|\bar{x}}(t) \left(\bar{x}_k - \sum_{j \neq k} \alpha_j^* \bar{x}_j \right)$ is positive and bounded below by $a(\bar{x})c > 0$ for all t such that $(t, \bar{x}) \in \mathcal{I}$. Pick $\varphi \leq \frac{a(\bar{x})c}{2}$. Then (14) implies $\alpha_\lambda^* \lambda_\delta(t, \bar{x}; \delta_0) \geq \frac{a(\bar{x})c}{2} > 0$ for all t with $(t, \bar{x}) \in \mathcal{I}$. By symmetry of $f_{\epsilon|x}$ and oddness of $\lambda_\delta(t, x; \delta_0)$ in t given any x, $\left| f_{\epsilon|\bar{x}}(-t) \left(\bar{x}_k - \sum_{j \neq k} \alpha_j^* \bar{x}_j \right) - \alpha_\lambda^* \lambda_\delta(-t, \bar{x}; \delta_0) \right| \geq \frac{3}{2} a(\bar{x})c > 0$ for all t with $(t, \bar{x}) \in \mathcal{I}$. A symmetric argument applies to show that such a distance is also bounded below by positive constants for any $\bar{x} \in \bar{\omega}$ with $a(\bar{x}) < 0$ and any t such that $(t, x) \in \mathcal{I}$. Due to Assumption 4.3, $\Pr((W, X) \in \mathcal{I}^-) > 0$ where $\mathcal{I}^- \equiv \{(t, x) : (-t, x) \in \mathcal{I}\}$. Thus $\left| f_{\epsilon|x}(t) \left(x_k - \sum_{j \neq k} \alpha_j^* x_j \right) - \alpha_\lambda^* \lambda_\delta(t, x; \delta_0) \right|$ is bounded away from zero by some positive constant over $\mathcal{I}^-$. It then follows that

$$\int_{\mathcal{I}^-} \left[f_{\epsilon|x}(w) \left(x_k - \textstyle\sum_{j \neq k} \alpha_j^* x_j \right) - \alpha_\lambda^* \lambda_\delta(w, x; \delta_0) \right]^2 dF_{W,X}$$

is bounded away from zero by some positive constant. This contradicts the claim that $I_{\lambda,k} = 0$ for $\lambda \in \Lambda$ where $\alpha_\lambda^* \neq 0$.

The proof for the general case where $c_0(x) \neq 0$ follows from similar arguments based on the discrepancy between even and odd functions, only with the interval $(-\varepsilon^*, \varepsilon^*)$ above replaced by the union of $(c_0(x) - \varepsilon^*, c_0(x) + \varepsilon^*)$ and $(-c_0(x) - \varepsilon^*, -c_0(x) + \varepsilon^*)$ and with the definition of $\mathcal{I}, \mathcal{I}^-$ adjusted accordingly. $\square$

Appendix B. Further Discussion

B1. Identification under Assumption 2.1 and 2.2

Assumption 3.1 in the text holds when the support of V given each x is bounded, provided that the parameter space for β is bounded. It differs from the large support condition needed to point identify β when errors are mean-independent or median-independent without exclusion restrictions. In the latter case, the support of V needs to include the support of $-X\beta + \epsilon$ conditional on X. Assumption 3.2 is a typical full-rank condition analogous to that in (Manski 1988). With the first coordinate in X being a constant intercept, Assumption 3.2 implies that there exists no $\tilde{\gamma} \neq 0$ in $\mathbb{R}^{k-1}$ and $c \in \mathbb{R}$ with $\Pr(X_{-1}\tilde{\gamma} = c) = 1$. It also means $\Pr(X(\beta - b) \neq 0) > 0$ for any $b \neq \beta$.

To see why Assumption 3.1 and 3.2 in Section 3.1 imply the point identification of β under Assumption 2.1 and 2.2, we use essentially the same argument as in (Manski 1988): Suppose without loss of generality that $\Pr(X\beta < Xb) > 0$. Under Assumption 3.1, for any x with $x\beta < xb$, there exists an interval of v with $x\beta \leq v < xb$. This implies that $\Pr(Z \in Q_b) > 0$ and thus β is identified relative to all $b \neq \beta$. For estimation, we propose a new extremum estimator for β that differs qualitatively from the Maximum Score estimator in (Manski 1985), based on the following corollary.

Corollary 1 *(Proposition 1) Suppose Assumption 2.1, 2.2, 3.1 and 3.2 hold in (1), and* $\Pr(X\beta = V) = 0$. *Then*

$$\beta = \arg\min_b \mathbb{E}_Z[1\{p(Z) \geq \tfrac{1}{2}\}(Xb - V)_- + 1\{p(Z) < \tfrac{1}{2}\}(Xb - V)_+] \tag{15}$$

where $(.)_+ \equiv \max\{., 0\}$ *and* $(.)_- \equiv -\min\{., 0\}$.

Proof of Corollary 1. The objective function in (15) is non-negative by construction. We show it is positive for all $b \neq \beta$, and 0 for $b = \beta$. Consider $b \neq \beta$. Then $\Pr(Xb \neq X\beta) = \Pr(Xb > X\beta \text{ or } Xb < X\beta) > 0$ under Assumption 3.2. W.L.O.G. suppose $\Pr(Xb > X\beta) > 0$. Assumption 3.1 implies for any x with $xb > x\beta$, there exists an interval of v with $xb > v \geq x\beta$. Hence $\Pr(Xb - V > 0 \geq X\beta - V) = \Pr(p(Z) \leq 1/2 \text{ and } Xb - V > 0) > 0$. With $\Pr(X\beta = V) = 0$ (and thus $\Pr(p(Z) = 1/2) = 0$), this implies $1\{p(Z) \leq 1/2\}(Xb - V)_+ > 0$ with positive probability. Thus the objective function in (15) is positive for $b \neq \beta$. On the other hand, Assumption 2.1 and 2.2 imply $p(Z) \geq 1/2$ if and only if $X\beta - V \geq 0$, and the objective function in (15) is 0 for $b = \beta$. □

Let n denote the sample size and let $\hat{p}_i$ denote kernel estimator for $\mathbb{E}(Y \mid Z = z_i)$. An alternative estimator is

$$\tilde{\beta} \equiv \arg\min \textstyle\sum_i \kappa\left(\hat{p}_i - \frac{1}{2}\right)(x_i b - v_i)_- + \kappa\left(\frac{1}{2} - \hat{p}_i\right)(x_i b - v_i)_+ \tag{16}$$

where the weight function $\kappa : \mathbb{R} \to [0,1]$ satisfies: $\kappa(t) = 0$ for all $t \leq 0$; $\kappa(t) > 0$ for all $t > 0$; and κ is increasing over $[0, +\infty)$.[10]

B2. Tail condition in (Magnac and Maurin 2007)

We now discuss how Assumption 2.3 is related to the tail condition in (Magnac and Maurin 2007). We give an example of some $F_{\epsilon|X}$ that satisfy Assumption 2.3 but fail to meet the tail condition in (Magnac and Maurin 2007). Suppose the distribution of a continuous random variable R is such that $\lim_{r\to-\infty} rF_R(r) = 0$. Then for any c,

$$\mathbb{E}\left[(R-c)\,1(R<c)\right] = \int_{-\infty}^{c} (r-c)\,dF_R(r) = 0 - 0 - \int_{-\infty}^{c} F_R(r)dr$$

and

$$\begin{aligned}\mathbb{E}\left[(R-c)\,1(R>c)\right] &= \mathbb{E}(R-c) - \mathbb{E}\left[(R-c)\,1(R<c)\right] \\ &= \mathbb{E}(R) - c + \int_{-\infty}^{c} F_R(r)dr.\end{aligned}$$

Let $Y_H \equiv -(X\beta + \epsilon + v_H)$ and $Y_L \equiv X\beta + \epsilon + v_L$. Therefore, for any given x,

$$\mathbb{E}\left[Y_H 1(Y_H > 0)|x\right] = \int_{-\infty}^{-v_H} F_{X\beta+\epsilon|X=x}(s)ds \tag{17}$$

$$\mathbb{E}\left[Y_L 1(Y_L > 0)|x\right] = x\beta + v_L + \int_{-\infty}^{-v_L} F_{X\beta+\epsilon|X=x}(s)ds \tag{18}$$

so that the difference of (18) minus (17) is given by

$$x\beta + v_L + \int_{-v_H}^{-v_L} F_{X\beta+\epsilon|X=x}(s)ds. \tag{19}$$

Suppose $F_{\epsilon|X}$ satisfies Assumption 2.3, then $F_{X\beta+\epsilon|x}$ is symmetric around $x\beta$ for all x. If $x\beta = \frac{-v_L-v_H}{2}$, then (19) equals

$$v_L - \frac{1}{2}(v_H + v_L) + \frac{1}{2}(v_H - v_L) = 0.$$

If $x\beta < \frac{-v_L-v_H}{2}$, then (19) is strictly less than 0 by the symmetry of $F_{X\beta+\epsilon|x}$ around $x\beta$ given x. Likewise if $x\beta > \frac{-v_L-v_H}{2}$, then (19) is strictly greater than 0 by the symmetry of $F_{X\beta+\epsilon|x}$ around $x\beta$ given x. Now suppose $x\beta < \frac{-v_L-v_H}{2}$ for all x on the support $\Omega_X \subseteq \mathbb{R}^K_{++}$. Then $\mathbb{E}[X'Y_H 1(Y_H > 0)] \neq \mathbb{E}[X'Y_L 1(Y_L > 0)]$, and the tail condition in Proposition 5 of (Magnac and Maurin 2007) does not hold.

[10] In implementation, one may choose κ to be twice continuously differentiable with bounded derivatives in an open neighborhood around 0 for technical convenience in deriving asymptotic properties of $\tilde{\beta}$.

Appendix C. Consistency of WLAD Estimator

Proof of Proposition 5. Consider any $b \neq \beta$. Under Assumption 3.2, $\Pr(X\beta - Xb \neq 0) > 0$. Without loss of generality, suppose that $\Pr(X\beta - Xb > 0) > 0$ and let $\omega \equiv \{x : x\beta > xb\}$. Then under Assumption 3.1,

$$\int 1\{2x\beta > v_i + v_j > 2xb\} dF_{V_i,V_j|x}(v_i, v_j) > 0$$

for all $x \in \omega$. By construction, whenever $x_i = x_j$, $p(x_i, v_i) + p(x_j, v_j) > 1$ if and only if $v_i + v_j < 2x_i\beta = 2x_j\beta$. Thus for all $x \in \omega$, properties of κ in Assumption 5.1 imply that:

$$\begin{aligned} & \mathbb{E}\left\{ \kappa\left(W_{i,j}-1\right)\left[\varphi_{i,j}(b)\right]_- + \kappa(1-W_{i,j})[\varphi_{i,j}(b)]_+ \middle| X_j = X_i = x \right\} \\ \geq & \left(\begin{array}{l} \mathbb{E}\left\{ \kappa(W_{i,j}-1)\left[\varphi_{i,j}(b)\right]_- \middle| V_i + V_j \leq 2x\beta, X_j = X_i = x \right\} \\ \times \Pr(V_i + V_j \leq 2x\beta | X_j = X_i = x) \end{array} \right) > 0. \end{aligned} \tag{20}$$

By construction, the conditional expectation on the left-hand side can never be negative for any x. Multiply both sides of (20) by $f(x)$ and then integrate out x over its full support (including ω) with respect to the distribution of non-special regressors. Thus we get $H_0(b) > 0$ for all $b \neq \beta$. Likewise, if $b \neq \beta$ and $\Pr(X\beta < Xb) > 0$, then for any x with $x\beta < xb$, Assumption 3.1 implies

$$\begin{aligned} & \mathbb{E}\left\{ \kappa\left(W_{i,j}-1\right)\left[\varphi_{i,j}(b)\right]_- + \kappa(1-W_{i,j})\left[\varphi_{i,j}(b)\right]_+ \middle| X_j = X_i = x \right\} \\ \geq & \left(\begin{array}{l} \mathbb{E}\left\{ \kappa(1-W_{i,j})\left[\varphi_{i,j}(b)\right]_+ \middle| V_i + V_j > 2x\beta, X_j = X_i = x \right\} \\ \times \Pr(V_i + V_j > 2x_i\beta | X_j = X_i = x) \end{array} \right) > 0. \end{aligned}$$

Then $H_0(b) > 0$ for all $b \neq \beta$ by the same argument as above. Next, consider $b = \beta$. For any x,

$$\begin{aligned} & H_0(\beta) \\ = & \mathbb{E}\left\{ f(X)\mathbb{E}\left[\kappa\left(W_{i,j}-1\right)\left[\varphi_{i,j}(\beta)\right]_- + \kappa(1-W_{i,j})\left[\varphi_{i,j}(\beta)\right]_+ \middle| X_j = X_i = X \right] \right\} \\ = & \mathbb{E}\left\{ f(X)\mathbb{E}\left[\kappa(W_{i,j}-1)\left[\varphi_{i,j}(\beta)\right]_- 1\{W_{i,j} \geq 1\} \middle| X_j = X_i = X \right] \right\} \\ + & \mathbb{E}\left\{ f(X)\mathbb{E}\left[\kappa(1-W_{i,j})\left[\varphi_{i,j}(\beta)\right]_+ 1\{W_{i,j} < 1\} \middle| X_i = X_i = X \right] \right\}. \end{aligned} \tag{21}$$

The first conditional expectation on the right-hand side of (21) is 0, because when $x_i = x_j$, $w_{i,j} \geq 1$ if and only if $v_i + v_j \leq 2x_i\beta$. Likewise the second conditional expectation is also 0. Thus $H_0(\beta) = 0$. □

Define an "infeasible" version of the objective function as follows:

$$H_n(b) = \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} K_h\left(x_i - x_j\right)\left\{ \kappa\left(w_{i,j}-1\right)\left[\varphi_{i,j}(b)\right]_- + \kappa(1-w_{i,j})[\varphi_{i,j}(b)]_+ \right\}$$

where $w_{i,j}$ is the sum of the true propensity scores (i.e. $w_{i,j} \equiv p_i + p_j$ with $p_l \equiv p(z_l)$).

Proof of Proposition 6. The first step of the proof is to establish that

$$\sup_{b\in\mathcal{B}} |\hat{H}_n(b) - H_n(b)| = o_p(1). \tag{22}$$

Let $\varphi^-_{i,j}(b)$, $\varphi^+_{i,j}(b)$ be shorthand for $[\varphi_{i,j}(b)]_-$, $[\varphi_{i,j}(b)]_+$. Applying the Taylor's expansion around $w_{i,j}$ and using the boundedness conditions in Assumption 5.6 and 5.7, we have:

$$\begin{aligned}
& \sup_{b\in\mathcal{B}} \left| \tfrac{1}{n(n-1)} \sum_{j\neq i} \begin{pmatrix} K_h(x_i - x_j)\,\varphi^-_{i,j}(b)\times \\ [\kappa(\hat{w}_{i,j}-1) - \kappa(w_{i,j}-1) - \kappa'(w_{i,j}-1)(\hat{w}_{i,j}-w_{i,j})] \end{pmatrix} \right| \\
= \; & \sup_{b\in\mathcal{B}} \left| \tfrac{1}{n(n-1)} \textstyle\sum_{j\neq i} K_h(x_i - x_j)\,\varphi^-_{i,j}(b)\kappa''(\tilde{w}_{i,j}-1)\,\|\hat{w}_{i,j} - w_{i,j}\|^2 \right| \\
\leq \; & a \sup_z \|\hat{p}(z) - p(z)\|^2 \sup_{b\in\mathcal{B}} \left\{ \tfrac{1}{n(n-1)} \textstyle\sum_{j\neq i} \left| K_h(x_i - x_j)\,\varphi^-_{i,j}(b)\kappa''(\tilde{w}_{i,j}-1) \right| \right\}
\end{aligned} \tag{23}$$

where κ', κ'' are first- and second-order derivatives of κ; $\tilde{w}_{i,j}$ is a random variable between $\hat{w}_{i,j}$ and $w_{i,j}$; and $a > 0$ is some finite constant. Under Assumption 5.1, 5.6, 5.7, the supreme of the term in the braces on the right-hand side of the inequality is $O_p(1)$.

Under Assumption 5.3 (i)-(iii), 5.4 and 5.5 $\sup_z |\hat{p}(z) - p(z)| = o_p(1)$ almost surely by Theorem 2.6 of (Li and Racine 2007). Hence the remainder term on the right-hand side of the inequality (23) is $o_p(1)$. Next, note:

$$\begin{aligned}
& \sup_{b\in\mathcal{B}} \left| \tfrac{1}{n(n-1)} \textstyle\sum_{j\neq i} K_h(x_i - x_j)\,\kappa'(w_{i,j}-1)(\hat{w}_{i,j} - w_{i,j})\varphi^-_{i,j}(b) \right| \\
\leq \; & 2 \sup_z \|\hat{p}(z) - p(z)\| \sup_{b\in\mathcal{B}} \left\{ \tfrac{1}{n(n-1)} \textstyle\sum_{j\neq i} \left| K_h(x_i - x_j)\,\kappa'(w_{i,j}-1)\varphi^-_{i,j}(b) \right| \right\}.
\end{aligned}$$

By similar arguments, the second term is $O_p(1)$, and the first term is $o_p(1)$. Thus (22) holds by a symmetric argument.

Next, decompose $H_n(b)$ as

$$H_n(b) = \mathbb{E}[g_n(Z_i, Z_j; b)] + \tfrac{2}{n} \textstyle\sum_{i\leq n} g_{n,1}(z_i; b) + \tfrac{2}{n(n-1)} \sum_{j\neq i} g_{n,2}(z_i, z_j; b) \tag{24}$$

where

$$\begin{aligned}
g_n(z_i, z_j; b) &\equiv K_h(x_i - x_j) \left[\kappa(w_{i,j}-1)\,\varphi^-_{i,j}(b) + \kappa(1 - w_{i,j})\varphi^+_{i,j}(b) \right]; \\
g_{n,1}(z_i; b) &\equiv \mathbb{E}[g_n(Z, Z'; b)|Z = z_i] + \mathbb{E}[g_n(Z, Z'; b)|Z' = z_i] - 2\mathbb{E}[g_n(Z, Z'; b)] \\
g_{n,2}(z_i, z_j; b) &\equiv g_n(z_i, z_j; b) - \mathbb{E}[g_n(Z, Z'; b)|Z = z_i] - \mathbb{E}[g_n(Z, Z'; b)|Z' = z_j] \\
&\quad + \mathbb{E}[g_n(Z, Z'; b)].
\end{aligned}$$

By construction, $\mathbb{E}[g_{n,1}(Z_i; b)] = 0$ and $\mathbb{E}[g_{n,2}(Z_i, Z_j; b)|Z_i = z_i] = \mathbb{E}[g_{n,2}(Z_i, Z_j; b)|Z_j = z_j] = 0$ for all z_i, z_j.

We now show that the second and the third term in (24) are $o_p(1)$. Under Assumption 5.2 and 5.7, we get

$$\begin{aligned} & \sup_{n,b\in\mathcal{B}} |h_n^k g_n(z_i,z_j;b)| \\ \leq \;& \mathcal{F}(z_i,z_j) \equiv a' \left\{ \kappa\left(w_{i,j}-1\right)\left[\mathcal{C}_{i,j}-\tfrac{v_i+v_j}{2}\right]_- + \kappa(1-w_{i,j})\left[\mathcal{D}_{i,j}-\tfrac{v_i+v_j}{2}\right]_+ \right\} \end{aligned}$$

for all (z_i, z_j), where $\mathcal{C}_{i,j}$ and $\mathcal{D}_{i,j}$ are defined in Assumption 5.6 and $a' > 0$ is some finite constant. By arguments as in (Pakes and Pollard 1989), the class of functions:

$$\{h_n^k g_n(z_i,z_j;b) : b \in \mathcal{B}\}$$

is Euclidean with a constant envelop $\mathcal{F}$, which satisfies $\mathbb{E}\left[\mathcal{F}(Z_i,Z_j)^2\right] < \infty$ under Assumption 5.6 and 5.7. Besides, $\mathbb{E}[\sup_{b\in\mathcal{B}} h_n^{2k} g_n\left(Z_i,Z_j;b\right)^2] = O(h_n^k)$ under Assumption 5.6 and 5.7 by an argument based on changing variables and dominated convergence theorem. It then follows from Theorem 3 in (Sherman 1994) that the second and the third terms in the decomposition in (24) are $o_p(1)$ and $O_p(n^{-1}h_n^{-k/2})$ uniformly over $b \in \mathcal{B}$ respectively. Under our choice of bandwidth in Assumption 5.8, $nh_n^{k/2} \to \infty$ as $n \to \infty$ and hence these two terms are both $o_p(1)$.

Next, we deal with the first term in the H-decomposition (24). Let

$$\tilde{\varphi}(z_i,z_j;b) \equiv \kappa\left(w_{i,j}-1\right)\left[\varphi_{i,j}(b)\right]_- + \kappa(1-w_{i,j})\left[\varphi_{i,j}\left(b\right)\right]_+ .$$

By definition,

$$\begin{aligned} & \mathbb{E}[g_n(Z_i,Z_j;b)] \\ = \;& \int K_h\left(x_i-x_j\right)\tilde{\varphi}(z_i,z_j;b)dF(z_i,z_j) \\ = \;& \int K_h\left(x_i-x_j\right)\mathbb{E}\left[\tilde{\varphi}(Z_i,Z_j;b)\middle|\, x_i,x_j\right]dF(x_i,x_j) \\ = \;& \int\int K(u)\mathbb{E}\left[\tilde{\varphi}(Z_i,Z_j;b)\middle|\, X_i=x_i, X_j=x_i+h_nu\right] f(x_i+h_nu)dudF(x_i), \end{aligned}$$

where the last equality follows from changing variables between x_j and $u \equiv (x_j - x_i)/h_n$. Under Assumption 5.3-(v), we can apply a Taylor expansion around x_i and apply the dominated convergence theorem to get $\mathbb{E}[g_n(Z_i,Z_j;b)] = H_0(b) + O(kh_n^2) = H_0(b) + o(1)$ for all $b \in \mathcal{B}$. Thus the sum of the three terms on the right-hand side of (24) is $o_p(1)$ uniformly over $b \in \mathcal{B}$.

Combine this result with (22), we get:

$$\sup_{b\in\mathcal{B}} |\hat{H}_n(b) - H_0(b)| = o_p(1). \tag{25}$$

The limiting function $H_0(b)$ is continuous under Assumption 5.3-(iv) in an open neighborhood around β. Besides, Proposition 5 has established that $H_0(b)$ is uniquely minimized at β. It then follows from Theorem 2.1 in (Newey and McFadden 1994) that $\hat{\beta} \xrightarrow{p} \beta$. $\square$

Appendix D. Limit Distribution for the KWLS Estimator

To simplify exposition of the limiting distribution of $\hat{\beta}_{KWLS}$, we let the non-special regressor X be a scalar.[11] Define

$$\omega_{i,j} \equiv \tau_i \tau_j \frac{1}{h_{1,n} h_{2,n}} K_1\left(\frac{x_i - x_j}{h_{1,n}}\right) K_2\left(\frac{p_i + p_j - 1}{h_{2,n}}\right). \tag{26}$$

Let $\hat{\omega}_{i,j}$ denote the sample analog of $\omega_{i,j}$ where the actual conditional choice probabilities p_i, p_j in (26) are replaced by kernel estimates $\hat{p}_i, \hat{p}_j$. The closed-form estimator is:

$$\hat{\beta}_{KWLS} \equiv \left(\sum_{i \neq j} \hat{\omega}_{i,j}(x_i + x_j)\,(x_i + x_j)'\right)^{-1} \left(\sum_{i \neq j} \hat{\omega}_{i,j}(x_i + x_j)\,(v_i + v_j)\right).$$

Thus we can write:

$$\begin{aligned} \hat{\beta}_{KWLS} - \beta \;=\; & \left(\frac{1}{n(n-1)} \sum_{i \neq j} \hat{\omega}_{i,j}(x_i + x_j)(x_i + x_j)'\right)^{-1} \times \\ & \left(\frac{1}{n(n-1)} \sum_{i \neq j} \hat{\omega}_{i,j}(x_i + x_j)[(v_i + v_j) - (x_i + x_j)\beta]\right). \end{aligned} \tag{27}$$

Lemma D1. (Lemma 3.1 in (Powell, Stock and Stocker 1989)) For an i.i.d. sequence of random variables $\{\xi_i : i = 1, 2, ..., n\}$, define a J-th order U-statistic of the form

$$U_n = \frac{1}{n(n-1)(n-J+1)} \sum \varphi_n(\xi_{i_1}, ..., \xi_{i_J})$$

where the sum is over all permutations of m distinct elements $\{i_1, ..., i_J\}$ from the set $\{1, ..., n\}$. Let

$$\hat{U}_n \equiv \theta_n + \frac{1}{n} \sum_{j=1}^{J} \sum_{i=1}^{n} \left[r_n^{(j)}(\xi_i) - \theta_n\right]$$

with $r_n^{(j)}(\xi_i) \equiv E[\varphi_n(\xi_{i_1}, ...\xi_{i_{j-1}}, \xi_i, \xi_{i_{j+1}}, ..., \xi_{i_J})|\xi_i]$, $\theta_n \equiv E[r_n^{(j)}(\xi_i)] = E[\varphi_n(\xi_1, ..., \xi_J)]$. If $E[\|\varphi_n(\xi_1, ..., \xi_J)\|^2] = O(n)$, then (i) $U_n = \theta_n + o_p(1)$; and (ii) $U_n = \hat{U}_n + o_p(n^{-1/2})$.

Lemma D2. Under Assumption 6.3, 6.5, 6.6 and 6.7,

$$\sup_{z \in \mathcal{Z}_\tau^+} |\hat{p}(z) - p(z)| = O_p(n^{-1/3 - \delta_2}).$$

[11] When X is J-dimensional, the choice of bandwidth may be component-specific $(h_{s,n}^1, h_{s,n}^2, ..., h_{s,n}^J)$ for $s = 1, 2$, and the kernels would be scaled by $(\Pi_{j=1}^J h_{s,n}^j)^{-1}$.

where $\mathcal{Z}_\tau^+ \equiv \{z \in \Omega_Z : \tau(z) > 0\}$.

Lemma D2 follows from the same argument as in (Newey 1994), (Ahn and Powell 1993) and (Chen and Khan 2003). It is therefore omitted here.

Lemma D3. Under Assumption 2.1, 2.3, 6.2, 6.3, 6.4, 6.5, 6.6 and 6.7,

$$\frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \hat{\omega}_{i,j}(x_i + x_j)(x_i + x_j)' \xrightarrow{p} \Sigma. \tag{28}$$

Proof of Lemma D3. An application of the Mean Value Theorem (using the smoothness of K_2 in Assumption 6.2) implies that:

$$\frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \hat{\omega}_{i,j}(x_i + x_j)(x_i + x_j)' = \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \omega_{i,j}(x_i + x_j)(x_i + x_j)' + R_n \tag{29}$$

where

$$\begin{aligned} R_n &\equiv \frac{1}{n(n-1)} \sum\nolimits_{i \neq j} \omega_{i,j}^{(1)}(\tilde{p}_i, \tilde{p}_j)(x_i + x_j)(x_i + x_j)'(\hat{p}_i + \hat{p}_j - p_i - p_j) \text{ and} \\ \omega_{i,j}^{(1)}(p_i, p_j) &\equiv \tau_i \tau_j \frac{1}{h_{1,n}} K_1\left(\frac{x_i - x_j}{h_{1,n}}\right) \frac{1}{h_{2,n}^2} K_2'\left(\frac{p_i + p_j - 1}{h_{2,n}}\right) \end{aligned}$$

for some $\tilde{p}_i$ between p_i and $\hat{p}_i$ and some $\tilde{p}_j$ between p_j and $\hat{p}_j$. Assumption 6.2 implies as $n \to \infty$, the absolute value of R_n is eventually bounded above by the product of some finite positive constant and

$$h_{2,n}^{-1} \sup\nolimits_{z \in \mathcal{Z}_\tau^+} |\hat{p}(z) - p(z)| \tag{30}$$

where $\mathcal{Z}_\tau^+ \equiv \{z : \tau(z) > 0\}$. It then follows from Lemma D2 that $|R_n|$ is $O_p(n^{-1/3})$.

Next, we derive the probability limit of the first term on the right-hand side of (29). Let $\xi_i \equiv (z_i, p_i)$ and define $f_n(\xi_i, \xi_j) \equiv \omega_{i,j}(x_i + x_j)(x_i + x_j)'$. (Note that $p_i \equiv p(z_i)$ is a function of z_i, so conditioning on ξ_i is equivalent to conditioning on z_i.) Then by an argument similar to Lemma 8 in (Chen and Khan 2003), $E[\left\|f_n(\xi_i, \xi_j)\right\|^2] = O(h_{1,n}^{-k} h_{2,n}^{-1}) = o(n)$, where the second equality is due to the properties of $h_{1,n}, h_{2,n}$ in Assumption 6.3. Thus we can apply Lemma D1. Decompose f_n into $f_{1,n} + f_{2,n} + f_{3,n} + f_{4,n}$ where $f_{1,n}(\xi_i, \zeta_j) \equiv \omega_{i,j} x_i x_i'$; $f_{2,n}(\xi_i, \zeta_j) \equiv \omega_{i,j} x_i x_j'$; $f_{3,n}(\xi_i, \zeta_j) \equiv \omega_{i,j} x_j x_i'$; $f_{4,n}(\xi_i, \zeta_j) \equiv \omega_{i,j} x_j x_j'$. By definition,

$$\begin{aligned} & E[f_{1,n}(\xi_i, \xi_j) \mid \xi_i] \\ = \; & \tau_i x_i x_i' \int \int \tau_j \tfrac{1}{h_{1,n}} K_1\left(\tfrac{x_j - x_i}{h_{1,n}}\right) \tfrac{1}{h_{2,n}} K_2\left(\tfrac{p_i + p_j - 1}{h_{2,n}}\right) f_{X,P}(x_j, p_j) dx_j dp_j. \end{aligned}$$

Note by Assumption 2.1, $\tau_j \equiv \tau(x_j, v_j) = \tau(x_j, x_j\beta - \mathcal{G}(p_j, x_j))$. By changing variables between $(u, \eta) \equiv ((x_j - x_i)/h_{1,n}, [p_j - (1 - p_i)]/h_{2,n})$ and (x_j, p_j) while fixing ξ_i, the right-hand side in the display above is

$$\begin{aligned} & \tau_i x_i x_i' \int\int \tau_j \tfrac{1}{h_{1,n}} K_1\left(\tfrac{x_j - x_i}{h_{1,n}}\right) K_2(\eta) f_{X,P}(x_j, 1 - p_i + \eta h_{2,n}) dx_j d\eta \\ = \; & \tau_i x_i x_i' \int\int \tau_{i,n} K_1(u)\, K_2(\eta) f_{X,P}(x_i + u h_{1,n}, 1 - p_i + \eta h_{2,n}) du d\eta, \end{aligned}$$

where $\tau_{i,n}$ is a shorthand for $\tau(x_i + u h_{1,n}, (x_i + u h_{1,n})\beta - \mathcal{G}(1 - p_i + \eta h_{2,n}, x_i + u h_{1,n}))$. Because $h_{1,n}, h_{2,n} \to 0$, it follows from the continuity of τ, $\mathcal{G}$ that $\tau_{i,n}$ converges to $\tau(x_i, x_i\beta - \mathcal{G}(1 - p_i, x_i)) = \tau(x_i, 2x_i\beta - v_i) \equiv \tilde{\mu}_{\tau,i}$, where the equality follows from Assumption 2.3. Furthermore, by the continuity of $f_{X,P}$ in Assumption 6.4 and the Dominated Convergence Theorem, the right-hand side converges to

$$\begin{aligned} & \tau_i x_i x_i' \int\int \tau(x_i, 2x_i\beta - v_i) K_1(u)\, K_2(\eta) f_{X,P}(x_i, 1 - p_i) du d\eta \\ = \; & \tau_i \tilde{\mu}_{\tau,i} x_i x_i' f_{X,P}(x_i, 1 - p_i). \end{aligned}$$

By another application of the Dominated Convergence Theorem, $E[f_{1,n}(\xi_i, \xi_j)]$ converges to

$$E[\tau_i \tilde{\mu}_{\tau,i} X_i X_i' f_{X,P}(X_i, 1 - P_i)]. \tag{31}$$

Using a similar argument based on changing variables and the Dominated Convergence Theorem, we can show that $E[f_{2,n}(\xi_i, \xi_j) \mid \xi_i]$ also converges to (31). A symmetric argument that swaps the indices of i, j above shows $E[f_{3,n}(\xi_i, \xi_j)]$ and $E[f_{4,n}(\xi_i, \xi_j)]$ also converge to (31). Thus $E[f_n(\xi_i, \xi_j)] = 4E[\tau_i \tilde{\mu}_{\tau,i} X_i X_i' f_{X,P}(X_i, 1 - P_i)]$. It then follows from Lemma D1 that (28) holds. □

The next step is to derive a linear representation of the numerator (the second term) on the right-hand side of (27). Henceforth we simply write $\omega_{i,j}^{(1)}(p_i, p_j)$ as $\omega_{i,j}^{(1)}$ when the function $\omega_{i,j}^{(1)}$ is evaluated at the true propensity scores. Also, let $\psi_{i,j} \equiv \psi(z_i, z_j) f \equiv (x_i + x_j)[(v_i + v_j) - (x_i + x_j)'\beta]$.

First off, apply a second-order Taylor expansion (allowed by the smoothness condition on K_2) to write the numerator as

$$\frac{1}{n(n-1)} \textstyle\sum_{i \neq j} \omega_{i,j} \psi_{i,j} + \frac{1}{n(n-1)} \sum_{i \neq j} \omega_{i,j}^{(1)} \psi_{i,j} (\hat{p}_i + \hat{p}_j - p_i - p_j) + \mathcal{R}_n \tag{32}$$

where the remainder term is

$$\mathcal{R}_n \equiv \frac{1}{n(n-1)} \textstyle\sum_{i \neq j} \tau_i \tau_j \frac{1}{h_{1,n} h_{2,n}^3} K_1\left(\frac{x_i - x_j}{h_{1,n}}\right) K_2''\left(\frac{p_i' + p_j' - 1}{h_{2,n}}\right) \psi_{i,j} (\hat{p}_i + \hat{p}_j - p_i - p_j)^2$$

for some p_i', p_j' between (p_i, p_j) and $(\hat{p}_i, \hat{p}_j)$. Assumption 6.2 ensures the absolute value of $\mathcal{R}_n$ is bounded above by the product of a finite constant and $h_{2,n}^{-2} \sup_{z \in \mathcal{Z}_\tau^+} |\hat{p}(z) - p(z)|^2$,

which is $O_p(n^{-2/3})$ under the conditions in Lemma D2. This means $\mathcal{R}_n$ is asymptotically negligible in the linear representation of the numerator on the right-hand side of (27).

Lemma D4. Under Assumption 2.1, 2.3, 6.2, 6.3 and 6.4,

$$\frac{1}{n(n-1)}\sum_{i\neq j}\omega_{i,j}\psi_{i,j} = o_p(n^{-1/2}).$$

Proof of Lemma D4. As before, let $\xi_i \equiv (z_i, p_i)$ where $z_i = (x_i, v_i)$ and $p_i \equiv p(z_i)$. Let $g_n(\xi_i, \xi_j) \equiv \omega_{i,j}\psi_{i,j}$. By an argument similar to Lemma 8 in (Chen and Khan 2003), $E[\left\|g_n(\xi_i,\xi_j)\right\|^2] = O(h_{1,n}^{-k}h_{2,n}^{-1}) = o(n)$, where the second equality is due to the properties of $h_{1,n}, h_{2,n}$ in Assumption 6.3. This allows us to apply results from Lemma D1. By definition, we can write $E[g_n(\xi_i,\xi_j) \mid \xi_i]$ as:

$$\begin{aligned}
&\tau_i \int\int \left(\begin{array}{c} \frac{\tau_j}{h_{1,n}h_{2,n}} K_1\left(\frac{x_j - x_i}{h_{1,n}}\right) K_2\left(\frac{p_i+p_j-1}{h_{2,n}}\right) \times (x_i + x_j) \\ \times\{v_i + [x_j\beta - \mathcal{G}(p_j, x_j)] - (x_i + x_j)\beta\} \end{array} \right) f_{X_j,P_j}(x_j, p_j) dp_j dx_j \\
= \;& \tau_i \int\int \left(\begin{array}{c} \tau_{i,n} \times K_1(u) \times K_2(\eta) \times (2x_i + h_{1,n}u) \\ \times[v_i + (x_i + h_{1,n}u)\beta - \mathcal{G}_{i,n} - (2x_i + h_{1,n}u)'\beta] \end{array} \right) f_{X_j,P_j}(x_i + h_{1,n}u, 1 - p_i + h_{2,n}\eta) d\eta du
\end{aligned}$$

where $\tau_{i,n}$ is defined in the proof of Lemma D3 and $\mathcal{G}_{i,n}$ is a shorthand for $\mathcal{G}(1 - p_i + h_{2,n}\eta, x_i + h_{1,n}u)$. The equality is due to the change of variables between $(u, \eta) \equiv (\frac{x_j - x_i}{h_{1,n}}, \frac{p_i+p_j-1}{h_{2,n}})$ and (x_j, p_j) while fixing ξ_i. Under Assumption 6.4, an M-th order Taylor expansion around $(x_i, 1 - p_i)$ in

$$\tau_{i,n}(2x_i + h_{1,n}u)(v_i - x_i\beta - \mathcal{G}_{i,n}) f_{X_j,P_j}(x_i + h_{1,n}u, 1 - p_i + h_{2,n}\eta),$$

together with the properties of the kernels in Assumption 6.2 and the bandwidths in Assumption 6.3, imply for any ξ_i, the conditional expectation $E[g_n(\xi_i,\xi_j) \mid \xi_i]$ is:

$$2\tilde{\mu}_{\tau,i} x_i [v_i - x_i\beta - \mathcal{G}(1 - p_i, x_i)] f_{X_j,P_j}(x_i, 1 - p_i) + o(n^{-1/2}). \tag{33}$$

Recall $v_i - x_i\beta = \mathcal{G}(1 - p_i, x_i)$ due to Assumption 2.1 and 2.3. Because the magnitude of the remainder term is $o(n^{-1/2})$ uniformly in ξ_i, this implies $\frac{1}{n}\sum_{i=1}^n E[g_n(\xi_i,\xi_j)|\xi_i]$ is $o_p(n^{-1/2})$. By a symmetric argument, $\frac{1}{n}\sum_{i=1}^n E[g_n(\xi_i,\xi_j) \mid \xi_j]$ is also $o_p(n^{-1/2})$. Lemma D4 then follows from the second conclusion in Lemma D1. □

It remains to establish the linear representation of the second term in (32). For convenience, let $Q \equiv [Y, 1]'$ and $q_l \equiv [y_l, 1]'$. With a slight abuse of notation, we now denote $\xi \equiv (z, p(z), y)$ henceforth. Let $\mathcal{K}_\sigma(.)$ be a shorthand for $\sigma_n^{-(1+k)}\mathcal{K}\left(\frac{\cdot}{\sigma_n}\right)$ as before.

Lemma D5. Under Assumption 6.2, 6.3, 6.5, 6.6 and 6.7,

$$\frac{1}{n(n-1)}\sum_{i\neq j}\omega_{i,j}^{(1)}\psi_{i,j}(\hat{p}_i - p_i) = \frac{1}{n(n-1)(n-2)}\sum_{i\neq j\neq l}\varphi_n(\xi_i,\xi_j,\xi_l) + o_p(n^{-1/2}) \quad (34)$$

where $\varphi_n(\xi_i,\xi_j,\xi_l) \equiv \omega_{i,j}^{(1)}\left(\psi_{i,j}/f_i\right)\mathcal{K}_\sigma\left(z_l - z_i\right)\left(y_l - p_i\right)$.

Proof of Lemma D5. Let $h_i, \hat{h}_i, f_i, \hat{f}_i$ denote $h(z_i), \hat{h}(z_i), f(z_i), \hat{f}(z_i)$ respectively so that $\hat{p}_i = \hat{h}_i/\hat{f}_i$ and $p_i = h_i/f_i$; let $\gamma_i \equiv [h_i, f_i]$ and $\hat{\gamma}_i \equiv [\hat{h}_i, \hat{f}_i]$. A second-order Taylor expansion implies the left-hand side of (34) is

$$\frac{1}{n(n-1)}\sum_{i\neq j}\omega_{i,j}^{(1)}\psi_{i,j}\frac{1}{f_i}\left[\hat{h}_i - h_i - (\hat{f}_i - f_i)p_i\right] + \tilde{R}_n \quad (35)$$

where

$$\tilde{R}_n \equiv \frac{1}{2n(n-1)}\sum_{i\neq j}\omega_{i,j}^{(1)}\psi_{i,j}[\hat{h}_i - h_i, \hat{f}_i - f_i]\begin{bmatrix} 0, & -\tilde{f}_i^{-2} \\ -\tilde{f}_i^{-2}, & 2\tilde{h}_i\tilde{f}_i^{-3}\end{bmatrix}[\hat{h}_i - h_i, \hat{f}_i - f_i]'$$

for some $\tilde{h}_i$ between $\hat{h}_i$ and h_i, and $\tilde{f}_i$ between $\hat{f}_i$ and f_i. By condition (ii) in Assumption 6.2 and an argument analogous to (8.9) in Newey and McFadden (1994), $|\tilde{R}_n|$ is bounded above by the product of a finite constant and $h_{2,n}^{-1}$ and $\sup_{z\in\mathcal{Z}_\tau^+}|\hat{h}(z)-h(z)|^2+|\hat{f}(z)-f(z)|^2$, which is $o_p(n^{-1/2})$ under the conditions in Lemma D2. Furthermore, the lead term in (35) is

$$\frac{1}{n^2(n-1)}\sum_{i=1}^n\sum_{j=1,j\neq i}^n\sum_{l=1}^n\varphi_n(\xi_i,\xi_j,\xi_l).$$

By standard argument, the difference between this and the third-order U-statistic on the right-hand side of (34) is $o_p(n^{-1/2})$. □

Lemma D6. Under Assumptions 6.2-6.8,

$$\frac{1}{n(n-1)}\sum_{i\neq j}\omega_{i,j}^{(1)}\psi_{i,j}(\hat{p}_i - p_i) = \frac{1}{n}\sum_{l=1}^n\phi_0(z_l)(y_l - p_l) + o_p(n^{-1/2})$$

where $\phi_0(z_l) \equiv 2\tau_l x_l\tilde{\mu}_{\tau,l}\mathcal{G}_1(1-p_l, x_l)f_{X,P}(x_l, 1-p_l)$.

Proof of Lemma D6. By an argument using change of variables, $E[\left\|\varphi_n(\xi_i,\xi_j,\xi_l)\right\|^2] = O(h_{1,n}^{-k}h_{2,n}^{-3}\sigma_n^{-1-k})$, which is $o(n)$ under Assumption 6.2„ 6.3 and 6.7. Hence Lemma D1 implies the third-order U-statistic on the right-hand side of (34) can be written as:

$$\theta_n + \frac{1}{n}\sum_{i=1}^n\sum_{s=1}^3\left[r_n^{(s)}(z_i,y_i) - \theta_n\right] + o_p(n^{-1/2}). \quad (36)$$

where $r_n^{(s)}(\xi) \equiv E[\varphi_n(\xi_1,\xi_2,\xi_3)|\xi_s = \xi]$ for $s = 1,2,3$; and $\theta_n \equiv E[\varphi_n(\xi_i,\xi_j,\xi_l)]$. By construction,

$$\begin{aligned} r_n^{(1)}(\xi_i) &= E\left\{ E\left[\omega_{i,j}^{(1)}\left(\psi_{i,j}/f_i\right)\mathcal{K}_\sigma\left(z_l - z_i\right)\left(y_l - p_i\right)\middle|\, \xi_i, z_l\right]\middle|\, \xi_i\right\} \\ &= E\left[\omega_{i,j}^{(1)}\left(\psi_{i,j}/f_i\right)\middle|\, \xi_i\right] E\left[\mathcal{K}_\sigma\left(z_l - z_i\right)\left(p_l - p_i\right)\middle|\, \xi_i\right] \end{aligned}$$

where for any ξ_i

$$\begin{aligned} & E\left[\omega_{i,j}^{(1)}\left(\psi_{i,j}/f_i\right)\middle|\, \xi_i\right] = \int\int \omega_{i,j}^{(1)}(\psi_{i,j}/f_i) f_{X,P}(x_j,p_j)dp_j dx_j \\ = \ & \frac{\tau_i}{f_i}\int h_{1,n}^{-1}K_1\left(\frac{x_i - x_j}{h_{1,n}}\right)\left[\int \tau_j\tilde{\psi}_i(x_j,p_j) f_{X,P}(x_j,p_j) h_{2,n}^{-2}K_2'\left(\frac{p_i + p_j - 1}{h_{2,n}}\right)dp_j\right]dx_j \end{aligned} \tag{37}$$

with $\tilde{\psi}_i(x_j,p_j) \equiv (x_i + x_j)\{v_i + [x_j\beta - \mathcal{G}(p_j,x_j)] - (x_i + x_j)\beta\}$. Now let:

$$\tilde{\lambda}_i(x_j,p_j) \equiv \tau(x_j, x_j\beta - \mathcal{G}(p_j,x_j))\tilde{\psi}_i(x_j,p_j) f_{X,P}(x_j,p_j)$$

and $\tilde{\lambda}_{i,1}$ denote the partial derivative of $\tilde{\lambda}_i$ w.r.t. the second argument p_j. For any (ξ_i, x_j), we can use integration by parts and conditions Assumption 6.2 to write the term in the square brackets in the integrand on the right-hand side of (37) as:

$$\int \tilde{\lambda}_i(x_j,p_j)d\left[h_{2,n}^{-1}K_2\left(\frac{p_i + p_j - 1}{h_{2,n}}\right)\right] = \int h_{2,n}^{-1}K_2\left(\frac{p_i + p_j - 1}{h_{2,n}}\right)\tilde{\lambda}_{i,1}(x_j,p_j)dp_j.$$

Substitute this into the right-hand side of (37) and change variables between $(u,\eta) \equiv (\frac{x_j - x_i}{h_{1,n}}, \frac{p_j - (1-p_i)}{h_{2,n}})$ and (x_j,p_j). This leads to:

$$\begin{aligned} & \frac{\tau_i}{f_i}\int h_{1,n}^{-1}K_1\left(\frac{x_j - x_i}{h_{1,n}}\right)\left[\int h_{2,n}^{-1}K_2\left(\frac{p_j - (1-p_i)}{h_{2,n}}\right)\tilde{\lambda}_{i,1}(x_j,p_j)dp_j\right]dx_j \\ = \ & \frac{\tau_i}{f_i}\int K_1(u)\left(\int K_2(\eta)\,\tilde{\lambda}_{i,1}(x_i + h_{1,n}u, 1 - p_i + h_{2,n}\eta)d\eta\right)du. \end{aligned}$$

An M-th order Taylor expansion around $(x_i, 1-p_i)$, together with the higher order of K_1, K_2, imply:

$$\sup_{\xi_i}\left\|E\left[\omega_{i,j}^{(1)}\psi_{i,j}/f_i\middle|\, \xi_i\right] - \tau_i\tilde{\lambda}_{i,1}(x_i, 1-p_i)/f_i\right\| = O(h_{1,n}^M + h_{2,n}^M) = o(n^{-1/2}).$$

(Note that we could have relaxed the condition and only require $O(h_{1,n}^{M_1}) = o(n^{-1/2})$ and $O(h_{2,n}^{M_2}) = o(n^{-1/2})$ where $\max\{M_1, M_2\} = M$ here.) Furthermore, under Assumption 6.5 and by an argument using change of variables,

$$\sup_{\xi_i} E\left[\mathcal{K}_\sigma\left(z_l - z_i\right)\left(p_l - p_i\right)\middle|\, \xi_i\right] = O(\sigma_n^m).$$

which is $o(1)$ as long as $\sigma_n \to 0$. It then follows that $E[r_n^{(1)}(\xi_i)] = \theta_n = o(n^{-1/2})$. Furthermore, an application of the Dominated Convergence Theorem shows the unconditional

variance of $r_n^{(1)}(\xi_i)$ is $o(1)$ under condition D. It then follows from an application of the Chebyshev's Inequality that $n^{-1/2}\sum_{i=1}^n \left[r_n^{(1)}(z_i, y_i) - \theta_n\right] = o_p(1)$. Likewise:

$$r_n^{(2)}(\xi_j) = E\left[\omega_{i,j}^{(1)}\left(\psi_{i,j}/f_i\right)\middle|\,\xi_j\right] E\left[\mathcal{K}_\sigma\left(z_l - z_i\right)\left(p_l - p_i\right)\right],$$

where the equality follows from the mutual independence between ξ_i, ξ_j, ξ_l. Hence by similar argument, $n^{-1/2}\sum_{i=1}^n \left[r_n^{(2)}(z_i, y_i) - \theta_n\right]$ is $o_p(1)$.

Next, for any given z_l, define:

$$\phi_n(z_l) \equiv E\left[\left(\omega_{i,j}^{(1)}\psi_{i,j}/f_i\right)\mathcal{K}_\sigma(z_l - z_i)\middle|\, z_l\right].$$

Then

$$\frac{1}{n}\sum_{l=1}^n r_n^{(3)}(\xi_l) = \frac{1}{n}\sum_{l=1}^n \phi_0(z_l)(y_l - p_l) + \tilde{\rho}_{1,n} + \tilde{\rho}_{2,n} \tag{38}$$

where $\tilde{\rho}_n = \tilde{\rho}_{1,n} + \tilde{\rho}_{2,n}$ with

$$\begin{aligned}
\tilde{\rho}_{1,n} &\equiv \frac{1}{n}\sum_{l=1}^n E\left\{\left(\omega_{i,j}^{(1)}\psi_{i,j}/f_i\right)\mathcal{K}_\sigma(z_l - z_i)(p_l - p_i)\middle|\,\xi_l\right\}; \\
\tilde{\rho}_{2,n} &\equiv \frac{1}{n}\sum_{l=1}^n [\phi_n(z_l) - \phi_0(z_l)](y_l - p_l).
\end{aligned}$$

By an argument based on the Chebyshev's Inequality and several applications of Dominated Convergence Theorem (similar to that in the previous paragraph), $\tilde{\rho}_n = o_p(n^{-1/2})$ under the smoothness conditions in Assumption 6.4 and the finite-moment conditions in Assumption 6.8. Lemma D6 then follows from (36) and the two results shown above: "$\frac{1}{n}\sum_{i=1}^n \left[r_n^{(s)}(z_i, y_i) - \theta_n\right]$ is $o_p(n^{-1/2})$ for $s = 1, 2$; and $\theta_n = o(n^{-1/2})$". $\square$

The limit distribution in Proposition 7 then follows from Lemma D1-D6 and the non-singularity of Σ in Assumption 6.1.

Design 1: $X \sim Normal$, $V \sim Normal$, $\epsilon \sim Normal$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0401	0.0377	0.0299	0.0193	0.0160
		β	-0.0579	-0.0302	-0.0121	-0.0186	-0.0061
	KWLS	α	-0.0336	-0.0059	-0.0022	0.0003	0.0033
		β	-0.1519	-0.0899	-0.0679	-0.0524	-0.0400
	IDW	α	-0.0244	-0.0138	-0.0135	-0.0084	-0.0061
		β	-0.0954	-0.0716	-0.0568	-0.0527	-0.0392
STD	WLAD	α	0.2989	0.2028	0.1428	0.0982	0.0706
		β	0.3319	0.2361	0.1761	0.1252	0.0913
	KWLS	α	0.1607	0.1124	0.0785	0.0557	0.0406
		β	0.1613	0.1171	0.0908	0.0658	0.0480
	IDW	α	0.1694	0.1145	0.0768	0.0534	0.0392
		β	0.1627	0.1155	0.0838	0.0629	0.0465
RMSE	WLAD	α	0.3015	0.2063	0.1458	0.1001	0.0723
		β	0.3368	0.2380	0.1765	0.1265	0.0915
	KWLS	α	0.1641	0.1126	0.0785	0.0557	0.0407
		β	0.2216	0.1476	0.1134	0.0841	0.0625
	IDW	α	0.1711	0.1153	0.0779	0.0541	0.0397
		β	0.1886	0.1359	0.1012	0.0820	0.0608
MAD	WLAD	α	0.2110	0.1365	0.0978	0.0659	0.0490
		β	0.2246	0.1625	0.1202	0.0782	0.0612
	KWLS	α	0.1137	0.0778	0.0546	0.0376	0.0278
		β	0.1632	0.1074	0.0766	0.0599	0.0428
	IDW	α	0.1257	0.0757	0.0506	0.0356	0.0276
		β	0.1274	0.0947	0.0768	0.0607	0.0451

Design 2: $X \sim Laplace$, $V \sim Normal$, $\epsilon \sim Laplace$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0393	0.0262	0.0212	0.0201	0.0167
		β	-0.0829	-0.0765	-0.0752	-0.0690	-0.0614
	KWLS	α	-0.0486	-0.0183	-0.0101	0.0024	0.0004
		β	-0.2372	-0.1790	-0.1446	-0.1133	-0.0978
	IDW	α	-0.0545	-0.0495	-0.0430	-0.0405	-0.0423
		β	-0.2392	-0.2253	-0.2088	-0.1943	-0.1771
STD	WLAD	α	0.3991	0.2712	0.1860	0.1288	0.0879
		β	0.3158	0.1889	0.1296	0.0980	0.0652
	KWLS	α	0.2359	0.1754	0.1207	0.0891	0.0621
		β	0.1383	0.1123	0.0828	0.0600	0.0422
	IDW	α	0.2206	0.1682	0.1407	0.1073	0.0886
		β	0.0842	0.0665	0.0543	0.0417	0.0335
RMSE	WLAD	α	0.4010	0.2724	0.1872	0.1304	0.0894
		β	0.3265	0.2038	0.1498	0.1198	0.0896
	KWLS	α	0.2408	0.1763	0.1211	0.0891	0.0621
		β	0.2745	0.2113	0.1666	0.1282	0.1065
	IDW	α	0.2272	0.1753	0.1471	0.1147	0.0981
		β	0.2536	0.2349	0.2157	0.1987	0.1803
MAD	WLAD	α	0.2536	0.1925	0.1165	0.0934	0.0572
		β	0.2038	0.1587	0.1071	0.0900	0.0585
	KWLS	α	0.1551	0.1021	0.0798	0.0609	0.0428
		β	0.2556	0.1834	0.1571	0.1158	0.0966
	IDW	α	0.1589	0.1154	0.1015	0.0779	0.0719
		β	0.2491	0.2284	0.2112	0.1946	0.1776

Design 3: $X \sim Normal$, $V \sim Normal$, $\epsilon \sim Het.Normal$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0289	0.0275	0.0200	0.0149	0.0111
		β	-0.1655	-0.1102	-0.0713	-0.0706	-0.0455
	KWLS	α	-0.0606	-0.0204	-0.0097	-0.0058	-0.0005
		β	-0.2975	-0.2131	-0.1535	-0.1174	-0.0818
	IDW	α	-0.0435	-0.0417	-0.0346	-0.0305	-0.0273
		β	-0.2445	-0.2180	-0.1825	-0.1672	-0.1343
STD	WLAD	α	0.3398	0.2277	0.1656	0.1159	0.0834
		β	0.3567	0.2700	0.2046	0.1482	0.1075
	KWLS	α	0.1824	0.1364	0.1023	0.0775	0.0544
		β	0.2049	0.1623	0.1307	0.1014	0.0732
	IDW	α	0.2092	0.1460	0.1153	0.0899	0.0753
		β	0.2219	0.1694	0.1305	0.1064	0.0822
RMSE	WLAD	α	0.3410	0.2294	0.1668	0.1169	0.0841
		β	0.3931	0.2916	0.2167	0.1642	0.1168
	KWLS	α	0.1922	0.1379	0.1028	0.0777	0.0544
		β	0.3612	0.2679	0.2016	0.1551	0.1097
	IDW	α	0.2137	0.1518	0.1204	0.0949	0.0801
		β	0.3301	0.2761	0.2244	0.1982	0.1575
MAD	WLAD	α	0.2345	0.1526	0.1181	0.0748	0.0541
		β	0.2528	0.2003	0.1554	0.1157	0.0764
	KWLS	α	0.1196	0.0931	0.0655	0.0541	0.0318
		β	0.2956	0.2100	0.1605	0.1162	0.0850
	IDW	α	0.1488	0.0993	0.0837	0.0698	0.0584
		β	0.2630	0.2165	0.2082	0.1990	0.1719

Design 4: $X \sim Laplace$, $V \sim Normal$, $\epsilon \sim Het.Laplace$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0269	0.0216	0.0161	0.0134	0.0055
		β	-0.2495	-0.2228	-0.1990	-0.1724	-0.1474
	KWLS	α	-0.0753	-0.0383	-0.0191	-0.0009	-0.0034
		β	-0.3874	-0.3518	-0.3043	-0.2593	-0.2154
	IDW	α	-0.0928	-0.0901	-0.0846	-0.0812	-0.0768
		β	-0.3975	-0.3598	-0.3344	-0.3046	-0.2727
STD	WLAD	α	0.4774	0.3324	0.2194	0.1533	0.1035
		β	0.2801	0.2139	0.1523	0.1187	0.0806
	KWLS	α	0.2435	0.1986	0.1487	0.1107	0.0796
		β	0.1357	0.1271	0.0866	0.0780	0.0585
	IDW	α	0.2741	0.2158	0.1905	0.1469	0.1181
		β	0.1038	0.0813	0.0708	0.0554	0.0488
RMSE	WLAD	α	0.4782	0.3331	0.2200	0.1539	0.1036
		β	0.3751	0.3089	0.2505	0.2093	0.1680
	KWLS	α	0.2549	0.2022	0.1499	0.1107	0.0797
		β	0.4104	0.3740	0.3163	0.2707	0.2232
	IDW	α	0.2894	0.2339	0.2084	0.1678	0.1409
		β	0.4108	0.3689	0.3418	0.3096	0.2770
MAD	WLAD	α	0.3030	0.2383	0.1567	0.1036	0.0603
		β	0.3017	0.2501	0.2045	0.1801	0.1482
	KWLS	α	0.1603	0.1251	0.1003	0.0769	0.0514
		β	0.3930	0.3453	0.3066	0.2642	0.2153
	IDW	α	0.1882	0.1437	0.1288	0.1100	0.1077
		β	0.3841	0.3398	0.3077	0.2873	0.2516

Design 5: $X \sim Normal$, $V \sim Laplace$, $\epsilon \sim Normal$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.1330	0.1032	0.0772	0.0604	0.0498
		β	0.0601	0.0267	0.0416	0.0452	0.0484
	KWLS	α	0.0216	0.0249	0.0303	0.0316	0.0264
		β	-0.0207	-0.0023	0.0058	0.0176	0.0220
	IDW	α	0.0027	-0.0006	0.0005	0.0010	0.0046
		β	-0.0118	-0.0201	-0.0164	-0.0086	-0.0055
STD	WLAD	α	0.5351	0.3349	0.2354	0.1634	0.1069
		β	0.5669	0.3578	0.2684	0.1886	0.1411
	KWLS	α	0.2380	0.1677	0.1150	0.0795	0.0541
		β	0.2444	0.1752	0.1240	0.0904	0.0626
	IDW	α	0.2232	0.1474	0.0992	0.0669	0.0474
		β	0.1984	0.1366	0.0994	0.0730	0.0493
RMSE	WLAD	α	0.5513	0.3504	0.2477	0.1742	0.1180
		β	0.5700	0.3588	0.2716	0.1939	0.1492
	KWLS	α	0.2389	0.1696	0.1189	0.0855	0.0602
		β	0.2452	0.1752	0.1241	0.0921	0.0664
	IDW	α	0.2232	0.1474	0.0992	0.0669	0.0477
		β	0.1988	0.1380	0.1008	0.0735	0.0496
MAD	WLAD	α	0.3360	0.2301	0.1641	0.1229	0.0786
		β	0.3332	0.2280	0.1726	0.1292	0.0970
	KWLS	α	0.1628	0.1119	0.0811	0.0641	0.0354
		β	0.1507	0.1072	0.0802	0.0628	0.0448
	IDW	α	0.1604	0.1009	0.0712	0.0453	0.0291
		β	0.1241	0.0938	0.0692	0.0513	0.0381

Design 6: $X \sim Laplace$, $V \sim Laplace$, $\epsilon \sim Laplace$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.1109	0.0817	0.0762	0.0630	0.0446
		β	0.0540	0.0160	0.0157	-0.0108	-0.0059
	KWLS	α	0.0055	0.0286	0.0199	0.0269	0.0263
		β	-0.1172	-0.0861	-0.0582	-0.0476	-0.0401
	IDW	α	-0.0582	-0.0238	-0.0286	-0.0203	-0.0189
		β	-0.1331	-0.1318	-0.1206	-0.1080	-0.0961
STD	WLAD	α	0.6484	0.4175	0.2847	0.1803	0.1186
		β	0.4398	0.2880	0.2076	0.1365	0.1051
	KWLS	α	0.2963	0.2263	0.1662	0.1180	0.0792
		β	0.1978	0.1339	0.1018	0.0756	0.0565
	IDW	α	0.2929	0.2192	0.1780	0.1283	0.0952
		β	0.1423	0.0992	0.0740	0.0587	0.0429
RMSE	WLAD	α	0.6577	0.4254	0.2947	0.1910	0.1267
		β	0.4430	0.2884	0.2081	0.1369	0.1052
	KWLS	α	0.2963	0.2281	0.1674	0.1210	0.0835
		β	0.2298	0.1592	0.1173	0.0893	0.0693
	IDW	α	0.2986	0.2204	0.1803	0.1298	0.0970
		β	0.1949	0.1649	0.1415	0.1229	0.1052
MAD	WLAD	α	0.4271	0.3143	0.1820	0.1064	0.0902
		β	0.2494	0.1753	0.1353	0.0962	0.0739
	KWLS	α	0.1798	0.1439	0.1145	0.0760	0.0537
		β	0.1634	0.1161	0.0785	0.0632	0.0506
	IDW	α	0.1842	0.1501	0.1130	0.0827	0.0698
		β	0.1487	0.1409	0.1228	0.1108	0.1020

Design 7: $X \sim Normal$, $V \sim Laplace$, $\epsilon \sim Het.Normal$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.1529	0.1074	0.0841	0.0640	0.0488
		β	-0.0440	-0.0496	-0.0182	0.0047	0.0193
	KWLS	α	0.0121	0.0067	0.0232	0.0214	0.0207
		β	-0.1740	-0.1303	-0.0918	-0.0542	-0.0241
	IDW	α	-0.0058	-0.0112	0.0021	-0.0045	-0.0063
		β	-0.1377	-0.1353	-0.1240	-0.0897	-0.0752
STD	WLAD	α	0.5328	0.4059	0.2687	0.1855	0.1141
		β	0.5444	0.4224	0.3076	0.2083	0.1584
	KWLS	α	0.2860	0.2120	0.1456	0.0991	0.0650
		β	0.2839	0.2176	0.1611	0.1205	0.0899
	IDW	α	0.2967	0.2160	0.1418	0.1023	0.0809
		β	0.3013	0.2222	0.1636	0.1308	0.1037
RMSE	WLAD	α	0.5543	0.4198	0.2816	0.1962	0.1241
		β	0.5461	0.4253	0.3081	0.2083	0.1596
	KWLS	α	0.2862	0.2121	0.1474	0.1014	0.0683
		β	0.3330	0.2536	0.1854	0.1321	0.0931
	IDW	α	0.2967	0.2163	0.1418	0.1024	0.0811
		β	0.3313	0.2601	0.2053	0.1586	0.1281
MAD	WLAD	α	0.3631	0.2451	0.1834	0.1286	0.0892
		β	0.4039	0.2810	0.1922	0.1403	0.1115
	KWLS	α	0.1855	0.1520	0.1014	0.0676	0.0431
		β	0.2346	0.1835	0.1332	0.0861	0.0606
	IDW	α	0.2133	0.1322	0.0908	0.0662	0.0509
		β	0.2039	0.1620	0.1551	0.1192	0.0919

Design 8: $X \sim Laplace$, $V \sim Laplace$, $\epsilon \sim Het.Laplace$; $(\alpha, \beta) = (0.2, 0.5)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.1200	0.0794	0.0669	0.0579	0.0491
		β	-0.0930	-0.1176	-0.0978	-0.0919	-0.0718
	KWLS	α	-0.0525	-0.0007	0.0016	0.0290	0.0281
		β	-0.3136	-0.2617	-0.2174	-0.1700	-0.1305
	IDW	α	-0.0814	-0.0529	-0.0509	-0.0441	-0.0342
		β	-0.3091	-0.2703	-0.2419	-0.2126	-0.1925
STD	WLAD	α	0.6049	0.4933	0.3451	0.2082	0.1450
		β	0.4990	0.3107	0.2281	0.1702	0.1240
	KWLS	α	0.3348	0.2876	0.2214	0.1440	0.0981
		β	0.2117	0.1657	0.1357	0.1119	0.0855
	IDW	α	0.4691	0.3751	0.3042	0.2333	0.1995
		β	0.2198	0.1685	0.1427	0.1144	0.0986
RMSE	WLAD	α	0.6167	0.4996	0.3514	0.2082	0.1530
		β	0.5075	0.3321	0.2481	0.1934	0.1433
	KWLS	α	0.3388	0.2875	0.2214	0.1469	0.1020
		β	0.3784	0.3097	0.2563	0.2035	0.1560
	IDW	α	0.4760	0.3787	0.3084	0.2374	0.2024
		β	0.3793	0.3185	0.2808	0.2414	0.2163
MAD	WLAD	α	0.4960	0.3352	0.2163	0.1391	0.1076
		β	0.3096	0.2185	0.1759	0.1334	0.1059
	KWLS	α	0.2040	0.1970	0.1440	0.0955	0.0646
		β	0.3102	0.2636	0.2199	0.1718	0.1265
	IDW	α	0.3072	0.2495	0.1993	0.1389	0.1404
		β	0.3160	0.2873	0.2628	0.2355	0.2176

Design 9: $X \sim Normal$, $V \sim Normal$, $\epsilon \sim Normal$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0104	0.0053	-0.0014	0.0017	-0.0002
		β	-0.0525	-0.0344	-0.0325	-0.0371	-0.0169
	KWLS	α	0.0034	0.0077	0.0051	0.0043	-0.0001
		β	-0.2907	-0.1896	-0.1466	-0.1221	-0.0920
	IDW	α	0.0047	0.0043	0.0062	0.0040	0.0015
		β	-0.2428	-0.1980	-0.1758	-0.1554	-0.1275
STD	WLAD	α	0.3468	0.2253	0.1489	0.1042	0.0743
		β	0.4158	0.2865	0.2191	0.1540	0.1214
	KWLS	α	0.1672	0.1212	0.0856	0.0642	0.0429
		β	0.2112	0.1576	0.1212	0.0898	0.0685
	IDW	α	0.1627	0.1158	0.0836	0.0606	0.0483
		β	0.1654	0.1203	0.0949	0.0718	0.0533
RMSE	WLAD	α	0.3469	0.2254	0.1489	0.1042	0.0743
		β	0.4191	0.2885	0.2215	0.1583	0.1225
	KWLS	α	0.1672	0.1214	0.0857	0.0643	0.0429
		β	0.3593	0.2466	0.1902	0.1516	0.1147
	IDW	α	0.1627	0.1159	0.0838	0.0607	0.0483
		β	0.2938	0.2316	0.1998	0.1712	0.1382
MAD	WLAD	α	0.2297	0.1440	0.0946	0.0706	0.0526
		β	0.3086	0.1950	0.1604	0.1131	0.0774
	KWLS	α	0.0982	0.0782	0.0608	0.0416	0.0288
		β	0.3218	0.2069	0.1541	0.1260	0.0977
	IDW	α	0.1020	0.0782	0.0562	0.0385	0.0332
		β	0.2464	0.1984	0.1758	0.1616	0.1294

Design 10: $X \sim Laplace$, $V \sim Normal$, $\epsilon \sim Laplace$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0075	-0.0204	-0.0121	0.0051	0.0025
		β	-0.1884	-0.1909	-0.1824	-0.1748	-0.1600
	KWLS	α	-0.0065	-0.0093	-0.0040	0.0052	0.0008
		β	-0.4814	-0.3891	-0.3219	-0.2729	-0.2460
	IDW	α	-0.0015	-0.0078	-0.0044	-0.0019	-0.0057
		β	-0.5819	-0.5112	-0.4741	-0.4200	-0.3408
STD	WLAD	α	0.3956	0.2945	0.2038	0.1385	0.0934
		β	0.4173	0.2533	0.1873	0.1410	0.0993
	KWLS	α	0.2265	0.1797	0.1327	0.0970	0.0635
		β	0.2104	0.1636	0.1224	0.0917	0.0711
	IDW	α	0.2232	0.1748	0.1541	0.1205	0.0959
		β	0.1050	0.0799	0.0658	0.0528	0.0416
RMSE	WLAD	α	0.3956	0.2951	0.2041	0.1386	0.0934
		β	0.4577	0.3171	0.2614	0.2246	0.1883
	KWLS	α	0.2265	0.1800	0.1328	0.0971	0.0635
		β	0.5254	0.4221	0.3444	0.2879	0.2560
	IDW	α	0.2231	0.1749	0.1542	0.1205	0.0961
		β	0.5913	0.5174	0.4786	0.4233	0.3433
MAD	WLAD	α	0.2563	0.2012	0.1395	0.1003	0.0661
		β	0.3255	0.2452	0.2139	0.1786	0.1643
	KWLS	α	0.1408	0.1099	0.0840	0.0584	0.0413
		β	0.5037	0.4007	0.3294	0.2788	0.2511
	IDW	α	0.1532	0.1187	0.1045	0.0824	0.0589
		β	0.5827	0.5587	0.5330	0.5108	0.4846

Design 11: $X \sim Normal$, $V \sim Normal$, $\epsilon \sim Het.Normal$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0092	0.0095	0.0020	-0.0044	-0.0043
		β	-0.2516	-0.1929	-0.1517	-0.1350	-0.0940
	KWLS	α	-0.0069	0.0069	0.0034	0.0018	-0.0006
		β	-0.5431	-0.4079	-0.3015	-0.2291	-0.1747
	IDW	α	-0.0052	0.0007	-0.0001	0.0028	0.0002
		β	-0.4737	-0.4198	-0.3278	-0.2570	-0.2004
STD	WLAD	α	0.3571	0.2455	0.1680	0.1243	0.0847
		β	0.4182	0.3027	0.2349	0.1743	0.1360
	KWLS	α	0.1894	0.1434	0.1033	0.0797	0.0528
		β	0.2412	0.1895	0.1463	0.1168	0.0880
	IDW	α	0.2157	0.1612	0.1178	0.0927	0.0746
		β	0.2253	0.1686	0.1327	0.1101	0.0838
RMSE	WLAD	α	0.3571	0.2456	0.1680	0.1244	0.0848
		β	0.4880	0.3589	0.2796	0.2204	0.1653
	KWLS	α	0.1895	0.1435	0.1034	0.0797	0.0528
		β	0.5943	0.4497	0.3351	0.2571	0.1956
	IDW	α	0.2157	0.1612	0.1178	0.0927	0.0746
		β	0.5245	0.4524	0.3536	0.2796	0.2172
MAD	WLAD	α	0.2486	0.1596	0.1112	0.0850	0.0506
		β	0.3715	0.2824	0.2172	0.1772	0.1191
	KWLS	α	0.1148	0.0935	0.0692	0.0544	0.0366
		β	0.5496	0.4187	0.3060	0.2353	0.1755
	IDW	α	0.1404	0.1156	0.0738	0.0578	0.0466
		β	0.4757	0.4036	0.3242	0.2511	0.1973

Design 12: $X \sim Laplace$, $V \sim Normal$, $\epsilon \sim Het.Laplace$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0171	-0.0198	-0.0058	0.0125	0.0066
		β	-0.4754	-0.4103	-0.3829	-0.3483	-0.3144
	KWLS	α	0.0005	-0.0018	-0.0014	0.0060	0.0026
		β	-0.6863	-0.5963	-0.4964	-0.3988	-0.3202
	IDW	α	0.0179	-0.0130	-0.0033	-0.0026	-0.0031
		β	-0.6801	-0.6269	-0.5541	-0.4508	-0.3704
STD	WLAD	α	0.4490	0.3247	0.2348	0.1632	0.1046
		β	0.3507	0.2824	0.1932	0.1481	0.1060
	KWLS	α	0.2428	0.1885	0.1521	0.1174	0.0789
		β	0.1620	0.1553	0.1131	0.1007	0.0758
	IDW	α	0.2751	0.2155	0.1873	0.1457	0.1158
		β	0.1091	0.0906	0.0749	0.0591	0.0481
RMSE	WLAD	α	0.4493	0.3252	0.2349	0.1636	0.1048
		β	0.5908	0.4980	0.4289	0.3784	0.3318
	KWLS	α	0.2427	0.1885	0.1521	0.1175	0.0789
		β	0.7052	0.6162	0.5091	0.4113	0.3291
	IDW	α	0.2757	0.2159	0.1873	0.1457	0.1158
		β	0.6888	0.6334	0.5591	0.4546	0.3735
MAD	WLAD	α	0.2905	0.2169	0.1567	0.1043	0.0700
		β	0.5323	0.4461	0.3928	0.3608	0.3156
	KWLS	α	0.1488	0.1127	0.0944	0.0739	0.0527
		β	0.7938	0.7064	0.5979	0.4992	0.4279
	IDW	α	0.1776	0.1454	0.1339	0.1017	0.0823
		β	0.7799	0.7121	0.6054	0.5126	0.4510

Design 13: $X \sim Normal$, $V \sim Laplace$, $\epsilon \sim Normal$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0803	0.0476	0.0209	0.0151	0.0113
		β	0.1368	0.0749	0.0773	0.0702	0.0666
	KWLS	α	0.0197	0.0128	0.0075	0.0095	0.0021
		β	-0.0709	-0.0388	-0.0135	0.0058	0.0161
	IDW	α	0.0078	-0.0032	-0.0039	0.0028	0.0014
		β	-0.0862	-0.0794	-0.0692	-0.0501	-0.0393
STD	WLAD	α	0.5303	0.3523	0.2527	0.1828	0.1158
		β	0.6908	0.4222	0.3172	0.2131	0.1610
	KWLS	α	0.2481	0.1675	0.1138	0.0806	0.0568
		β	0.2680	0.1962	0.1434	0.1006	0.0680
	IDW	α	0.2222	0.1517	0.1001	0.0681	0.0495
		β	0.2174	0.1523	0.1071	0.0810	0.0561
RMSE	WLAD	α	0.5363	0.3555	0.2535	0.1834	0.1163
		β	0.7041	0.4287	0.3264	0.2243	0.1742
	KWLS	α	0.2489	0.1679	0.1140	0.0811	0.0569
		β	0.2772	0.2000	0.1440	0.1007	0.0699
	IDW	α	0.2223	0.1517	0.1001	0.0681	0.0495
		β	0.2339	0.1717	0.1275	0.0953	0.0684
MAD	WLAD	α	0.3190	0.2203	0.1811	0.1222	0.0768
		β	0.3770	0.2498	0.1813	0.1490	0.1088
	KWLS	α	0.1549	0.1098	0.0765	0.0563	0.0387
		β	0.1794	0.1311	0.1009	0.0715	0.0496
	IDW	α	0.1354	0.1040	0.0672	0.0444	0.0338
		β	0.1650	0.1148	0.0871	0.0683	0.0487

Design 14: $X \sim Laplace$, $V \sim Laplace$, $\epsilon \sim Laplace$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	-0.0194	-0.0175	0.0121	0.0114	0.0047
		β	-0.0878	-0.0742	-0.0693	-0.0466	-0.0212
	KWLS	α	-0.0325	-0.0039	-0.0112	0.0010	0.0029
		β	-0.2655	-0.2276	-0.1789	-0.1529	-0.1341
	IDW	α	-0.0554	-0.0210	-0.0201	-0.0062	-0.0028
		β	-0.3979	-0.3530	-0.3167	-0.2817	-0.2404
STD	WLAD	α	0.6482	0.4387	0.2966	0.1865	0.1275
		β	0.5364	0.3575	0.2710	0.1865	0.1338
	KWLS	α	0.3124	0.2334	0.1706	0.1188	0.0827
		β	0.2537	0.1654	0.1271	0.0951	0.0660
	IDW	α	0.3102	0.2461	0.1847	0.1311	0.1046
		β	0.1636	0.1240	0.0973	0.0738	0.0508
RMSE	WLAD	α	0.6484	0.4390	0.2968	0.1868	0.1276
		β	0.5435	0.3651	0.2797	0.1922	0.1355
	KWLS	α	0.3140	0.2334	0.1710	0.1188	0.0828
		β	0.3672	0.2813	0.2194	0.1800	0.1494
	IDW	α	0.3150	0.2469	0.1858	0.1312	0.1046
		β	0.4302	0.3741	0.3313	0.2912	0.2457
MAD	WLAD	α	0.4389	0.2935	0.1798	0.1289	0.0887
		β	0.3458	0.2407	0.1964	0.1497	0.1100
	KWLS	α	0.2106	0.1530	0.1046	0.0834	0.0594
		β	0.2896	0.2346	0.1888	0.1517	0.1334
	IDW	α	0.2233	0.1497	0.1211	0.0855	0.0689
		β	0.4099	0.3781	0.3437	0.3038	0.2740

Design 15: $X \sim Normal$, $V \sim Laplace$, $\epsilon \sim Het.Normal$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	0.0557	0.0333	0.0220	0.0186	0.0134
		β	-0.0233	-0.0324	-0.0027	0.0028	0.0119
	KWLS	α	0.0042	-0.0071	-0.0013	0.0014	-0.0034
		β	-0.3098	-0.2186	-0.1344	-0.0796	-0.0461
	IDW	α	-0.0134	-0.0208	-0.0073	0.0018	0.0022
		β	-0.2705	-0.2444	-0.2127	-0.1743	-0.1469
STD	WLAD	α	0.6060	0.4050	0.2806	0.1906	0.1260
		β	0.6896	0.4743	0.3483	0.2305	0.1784
	KWLS	α	0.2824	0.2038	0.1417	0.0989	0.0723
		β	0.3262	0.2504	0.1962	0.1398	0.1035
	IDW	α	0.2998	0.2195	0.1514	0.1122	0.0890
		β	0.3078	0.2455	0.1824	0.1454	0.1173
RMSE	WLAD	α	0.6085	0.4063	0.2814	0.1915	0.1267
		β	0.6898	0.4754	0.3483	0.2305	0.1788
	KWLS	α	0.2824	0.2039	0.1416	0.0989	0.0724
		β	0.4498	0.3324	0.2378	0.1608	0.1133
	IDW	α	0.3000	0.2204	0.1516	0.1122	0.0891
		β	0.4097	0.3464	0.2802	0.2270	0.1880
MAD	WLAD	α	0.3350	0.2280	0.1852	0.1202	0.0857
		β	0.4110	0.3251	0.2130	0.1608	0.1217
	KWLS	α	0.1846	0.1358	0.1021	0.0645	0.0466
		β	0.3519	0.2625	0.1728	0.1039	0.0823
	IDW	α	0.2014	0.1269	0.1003	0.0687	0.0597
		β	0.3182	0.2588	0.2229	0.1841	0.1509

Design 16: $X \sim Laplace$, $V \sim Laplace$, $\epsilon \sim Het.Laplace$; $(\alpha, \beta) = (0, 1)$

			N= 100	N= 200	N= 400	N= 800	N= 1600
Bias	WLAD	α	-0.0109	0.0081	0.0097	0.0150	0.0082
		β	-0.2468	-0.2702	-0.2175	-0.2147	-0.1851
	KWLS	α	-0.0324	0.0014	-0.0125	0.0029	0.0037
		β	-0.5498	-0.4473	-0.3231	-0.2348	-0.1692
	IDW	α	-0.0342	-0.0052	-0.0076	-0.0095	0.0013
		β	-0.5543	-0.4922	-0.4155	-0.3512	-0.2798
STD	WLAD	α	0.7358	0.5076	0.3547	0.2155	0.1519
		β	0.5581	0.3477	0.2760	0.1894	0.1426
	KWLS	α	0.3080	0.2792	0.2185	0.1465	0.1057
		β	0.2376	0.1921	0.1559	0.1358	0.0981
	IDW	α	0.4461	0.3641	0.3122	0.2344	0.2021
		β	0.2131	0.1657	0.1411	0.1174	0.0977
RMSE	WLAD	α	0.7358	0.5076	0.3548	0.2160	0.1521
		β	0.6101	0.4402	0.3514	0.2863	0.2337
	KWLS	α	0.3097	0.2792	0.2188	0.1465	0.1057
		β	0.5989	0.4868	0.3587	0.2712	0.1956
	IDW	α	0.4473	0.3640	0.3122	0.2346	0.2021
		β	0.5939	0.5193	0.4388	0.3703	0.2964
MAD	WLAD	α	0.4971	0.3330	0.2260	0.1425	0.1089
		β	0.4441	0.3252	0.2817	0.2317	0.1763
	KWLS	α	0.2056	0.1926	0.1462	0.0983	0.0690
		β	0.6540	0.5694	0.4240	0.3293	0.2754
	IDW	α	0.2855	0.2440	0.1962	0.1533	0.1429
		β	0.6429	0.5204	0.4375	0.3766	0.3084

References

Andrews, D. (1994): "Asymptotics for Semiparametric Econometric Models via Stochastic Equicontinuity," *Econometrica*, 62, 43–72.

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

Berry, S., and P. Haile (2010): "Nonparametric Identification of Multinomial Choice Demand Models with Heterogeneous Consumers," Discussion paper no. 1718, Yale University.

Blundell, R., and J. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," in *Advances in Economics and Econonometrics: Theory and Applications, Eighth World Congress*, ed. by M. Dewatripont, L. Hansen, and S. Turnovsky, vol. II. Cambridge University Press.

Chamberlain, G. (1986): "Asymptotic Efficiency in Semiparametric Models with Censoring," *Journal of Econometrics*, 32, 189–218.

Chen, S., and S. Khan (2003): "Rates of Convergence for Estimating Regression Coefficients in Heteroskedastic Discrete Response Models," *Journal of Econometrics*, 117, 245–278.

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and Conƀdence Regions for Parameter Sets in Econometric Models," *Econometrica*, 75, 1243–1284.

Cosslett, S. (1987): "Efficiency Bounds for Distribution Free Estimators of the Binary Choice Model," *Econometrica*, 51, 765–782.

Fox, J., and C. Yang (2012): "Unobserved Heterogeneity in Matching Games," Working paper, University of Michigan.

Horowitz, J. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica*, 60, 505–531.

——— (1993): "Semiparametric and nonparametric estimation of quantal response models," in *Handbook of Statistics*, ed. by C. G.S.Maddala, and H.D.Vinod, vol. 11, pp. 45–72. Elsevier.

Ichimura, H. (1993): "Local quantile regression estimation of binary response models with conditional heteroskedasticity," Working paper, University of Minnesota.

Ichimura, H., and S. Lee (2010): "Characterizing Asymptotic Distributions of Semiparametric M-Estimators," *Journal of Econometrics*, 159, 252–266.

Khan, S. (2013): "Distribution free estimation of heteroskedastic binary response models using Probit/Logit criterion functions," *Journal of Econometrics*, 172, 168–182.

KHAN, S., AND E. TAMER (2010): "Irregular Identification, Support Conditions, and Inverse Weight Estimation," *Econometrica*, 78, 2021–2042.

LEWBEL, A. (1998): "Semiparametric Latent Variable Model Estimation with Endogenous or Mismeasured Regressors," *Econometrica*, 66(1), 105–121.

——— (2000): "Semiparametric qualitative response model estimation with unknown heteroskedasticity or instrumental variables," *Journal of Econometrics*, 97, 145–177.

——— (2007): "Endogenous Selection or Treatment Model Estimation," *Journal of Econometrics*, 141, 777–806.

LEWBEL, A., AND X. TANG (2015): "Identification and Estimation of Games with Incomplete Information Using Excluded Regressors," *Journal of Econometrics*, 189, 229–244.

LI, Q., AND J. RACINE (2007): *Nonparametric Econometrics*. Princeton University Press.

MAGNAC, T., AND E. MAURIN (2007): "Identification and information in monotone binary models," *Journal of Econometrics*, 139, 76–104.

MANSKI, C. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics*, 27, 313–333.

——— (1988): "Identification of binary response models," *Journal of the American Statistical Association*, 83, 729–738.

NEWEY, W. (1990): "Semiparametric Efficiency Bounds," *Journal of Applied Econometrics*, 5(2), 99–135.

NEWEY, W., AND D. MCFADDEN (1994): "Large sample estimation and hypothesis testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4 (36), pp. 2111–2245. Elsevier.

PAKES, A., AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica*, 57(5), 1027–57.

POWELL, J. (1994): "Estimation of Semiparametric Models," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. 4 (41), pp. 2443–2521. Elsevier.

SHERMAN, R. (1994): "U-processes in the analysis of a generalized semiparametric regression estimator," *Econometric Theory*, 10, 372–395.

ZHENG, X. (1995): "Semiparametric efficiency bounds for the binary choice and sample selection models under conditional symmetry," *Economics Letters*, 47, 249–253.