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

Songnian Chen[2], Shakeeb Khan[3] and Xun Tang[4]

This Version: April 23, 2013

We quantify the informational content of special regressors in heteroskedastic binary regressions with median-independent or conditionally symmetric errors. We measure informational content by two criteria: the set of regressor values that help point identify coefficients in latent payoffs as in (Manski 1988); and the Fisher information of coefficients as in (Chamberlain 1986). We find for median-independent errors, requiring one of the regressors to be "special" in a sense similar to (Lewbel 2000) does not add the identifying power or the information for coefficients. Nonetheless it does help identify the error distribution and the average structural function. For conditionally symmetric errors (which were shown to add no informational content by (Manski 1988) and (Zheng 1995) without special regressors), the presence of a special regressor improves the identifying power by the criterion of (Manski 1988), and the Fisher information for coefficients is strictly positive under mild conditions. We propose a new estimator for coefficients that converges at the parametric rate under symmetric errors and a special regressor, and report its decent performance in small samples through simulations.

Key words: Binary regression, 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, Haiqing Xu and seminar participants at UT Austin and Yale for comments. We thank Bingzhi Zhao for capable research assistance. The usual disclaimer applies.

[2]Economics Department, Hong Kong University of Science and Technology. Email: snchen@ust.hk

[3]Economics Department, Duke University. Email: shakeebk@duke.edu.

[4]Economics Department, University of Pennsylvania. Email: xuntang@sas.upenn.edu.

1 Introduction

In this paper we explore the informational content of a special regressors in binary choice models. In a binary choice model, a special regressor is one that is additively separable from all other components in the latent payoffs and that satisfies an exclusion restriction (i.e. being independent from the error conditional on all other regressors). Note in this paper, our definition of a special regressor per se does not require it to satisfy any "large support" requirement.[5] We examine how a special regressor contributes to the identification and the Fisher information of coefficients in semiparametric binary regressions with heteroskedastic errors. We focus on the role of special regressors in two models where errors are median independent or conditionally symmetric respectively. These models are of particular interest, because identification of coefficients in them does not require the "large support" condition (i.e. the support of special regressors includes that of the error), a condition typically used in identification-at-infinity arguments.

Special regressor arise in various social-economic contexts. (Lewbel 2000) used a special regressor to recover coefficients in semiparametric binary regressions where heteroskedastic errors are mean-independent from regressors. He showed coefficients for all regressors along with the error distribution are identified up to scale, provided 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 2012); and matching games with unobserved heterogeneity (Fox and Yang 2012).

Using a special regressor to identify coefficients in binary regressions 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.[6] They also showed that when support of special regressor is unbounded, Fisher information for coefficients becomes zero when the second moment of regressors is finite.

The econometrics literature on semiparametric binary regressions has largely been silent about how to use special regressors in combination of alternative stochastic restric-

[5] This differs from the definition of a special regressor in (Lewbel 2000), which requires a special regressor to be conditionally independent from the error, and have large support compared to the support of all other components in the latent payoff (including the errors).

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

tions 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 latent utility distribution outside the support of special regressors. They established identification for coefficients under such restrictions. Nonetheless the tail condition they use 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 the tail conditions in (Magnac and Maurin 2007) and the conditional symmetry considered in our paper are non-nested. We also provide a formal proof for positive information for coefficients in our model.

We contribute to the large literature on binary choice models by deriving several new results. First, we quantify the change in 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 amounts to comparing the size of the set of states where the propensity scores 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 a special one 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 regressions with heteroskedastic errors. For models with median-independent errors, we find the information for coefficients remains zero even after one of the regressors is required to be special. In comparison, for models with conditionally symmetric errors, the presence of a special regressor does yield positive information for coefficients. We provide some intuition for such positive information in this case, and propose a new two-step extremum estimator. Asymptotic properties of the estimator are derived and some monte carlo evidence for its performance is reported. These two results seem to suggest 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 regressions in (Chamberlain 1986).

Our third set of results (Section 3.3) provides 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 identifying power or information for coefficients when heteroskedastic errors are only required to be median independent, it is instrumental for recovering the distribution of the heteroskedastic error. This in turn can be used to predict counterfactual choice probabilities; and helps to recover the average structural function as defined in

(Blundell and Powell 2003) as long as the support of the special regressor is large enough.

This paper contributes to a broad econometrics literature on identification, inference 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 regressions 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 regression:

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

where $X \in \mathbb{R}^K, V \in \mathbb{R}$ and $\epsilon \in \mathbb{R}^1$ and the first coordinate in X is a constant. We use upper cases for random variables and lower cases for their realizations. Let F_R, f_R, Ω_R denote the distribution, the density and the support of a random vector R respectively, and let $F_{R_1|R_2}$, $f_{R_1|R_2}$ and $\Omega_{R_1|R_2}$ denote conditional distributions, densities and supports in the data-generating process (DGP). Assume the marginal effect of V is known to be negative, and is set to -1 as a scale normalization. We maintain the following exclusion restriction throughout the paper.

CI *(Conditional Independence) V is independent from ϵ given any $x \in \Omega_X$.*

For the rest of the paper, we also 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 $F_{\epsilon|X}$). The distribution $F_{V|X}$ and the propensity scores $\Pr(Y = 1|Z)$ are both directly identifiable from data and considered known in the identification exercise. Let $Z \equiv (X, V)$, and let $p(z)$ denote $\Pr(Y = 1|z)$ (which is directly identifiable from data). 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 Lesbegue measure for continuous coordinates.

To quantify informational content, we first follow the approach taken in (Manski 1988). For a generic pair of coefficients and the 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 propensity scores implied by $(b, G_{\epsilon|X})$ differ from those in the true data-generating process (DGP) characterized by $(\beta, F_{\epsilon|X})$. In comparison, the set $\tilde{\xi}$ in (2) consists of pairs of states where implied propensity scores differ from those in the true DGP. We say β is *identified relative to* $b \neq \beta$ if

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

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

As is clear from this definition, the identification of β hinges on model restrictions defining Θ, i.e. parameter space for error distributions given X. In Sections 3 and 4, we discuss identification of β when CI is paired with *one* of the following two stochastic restrictions on errors respectively.

MI *(Median Independence) For all x, ϵ is continuously distributed with $Med(\epsilon|x) = 0$ and with strictly positive densities in an open neighborhood around 0.*

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

We also discuss the information for β under these two assumptions and CI in Sections 3.2 and 4.2. We do so to explore any relation between the two distinct notions of informational content in (Manski 1988) and (Chamberlain 1986). This amounts to finding smooth parametric submodels which are nested in the semiparametric models and which have the least Fisher information for β. The semiparametric efficiency bound is formally defined as follows. Let μ denote some measure on $\{0, 1\} \otimes \Omega_Z$ such that $\mu(\{0\} \otimes \omega) = \mu(\{1\} \otimes \omega) = F_Z(\omega)$, where ω 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 $f_\lambda(y|z; b, \delta)$ denote the probability mass function of Y conditional on z and given coefficients b as well as a nuisance parameter $\lambda(., .; \delta)$. A *smooth* parametric submodel is characterized by a path λ such that there exists $\left\{(\psi_k)_{k \leq K}, \psi_\lambda\right\}$ such that

$$f_\lambda^{1/2}(y|z; b, \delta) - f_\lambda^{1/2}(y|z; \beta, \delta_0) = \textstyle\sum_k \psi_k(y, z)(b - \beta) + \psi_\lambda(y, z)(\delta - \delta_0) + r(y, z; b, \delta) \quad (3)$$

with

$$(\|b - \beta\| + \|\delta - \delta_0\|)^{-2} \int r^2(y, z; b, \delta) d\mu \to 0 \text{ as } b \to \beta \text{ and } \delta \to \delta_0. \quad (4)$$

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

$$I_{\lambda,k} \equiv \inf_{(\{\alpha_j\}_{j \neq k},\ \alpha_\lambda)} 4 \int \left(\psi_k - \textstyle\sum_{j \neq k} \alpha_j \psi_j - \alpha_\lambda \psi_\lambda\right)^2 d\mu. \quad (5)$$

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

3 Exclusion plus Median Independence

This section discusses the identification and information for β in heteroskedastic binary regressions under CI and MI. The model differs from that in (Manski 1988), (Horowitz 1992), and (Khan 2013) in that one of the regressors (V) is required to be independent from the error conditional on all other regressors (X). It also differs from that considered in (Lewbel 2000) and (Khan and Tamer 2010), for its error is median-independent, rather than mean-independent, from X. We are not aware of any previous work that discusses both identification and Fisher information of coefficients in such a model.

3.1 Identification

Our first finding is, with median-independent errors (MI), the exclusion restriction (CI) does not add any identifying power for recovering β. We formalize this result in Proposition 1 by noting that under MI the set of states z that help detect a given $b \neq \beta$ from β remains unchanged, regardless of whether an exclusion restriction is added to one of the regressors.

Proposition 1 *Suppose CI and MI 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 satisfying MI but not CI (i.e. $F_{\epsilon|X,V}(0) = 1/2$ and ϵ depends on both V and X) (Manski 1988) showed Q_b is the set of states that can be used to detect $b \neq \beta$ from β, based on observed propensity scores. Thus Proposition 1 suggests adding the exclusion restriction (CI) to a model with median-independent errors does not improve the identifying power for recovery of β, as the set of states that help identify β relative to b remains unchanged.

The intuition for such an equivalence builds on two observations. First, if states in Q_b help identify β relative to b under the weaker assumption of MI alone in (Manski 1988), they also do so under a stronger set of assumptions MI and CI. Second, if $\Pr\{Z \in Q_b\} = 0$, then certain distribution of structural errors $G_{\epsilon|X} \neq F_{\epsilon|X}$ can be constructed to satisfy CI and MI and, together with $b \neq \beta$, can generate the same propensity scores as those from the DGP. Proposition 1 differs qualitatively from that of Manski's result in that the construction of such a distribution $G_{\epsilon|X}$ needs to respect the additional assumption exclusion restriction in CI.

Although not helping with identifying β, CI *does* help recover the error distribution $F_{\epsilon|X}$, which in turn is useful for counterfactual predictions of propensity scores, and for estimating an average structural function of the model. We discuss this in greater details in Section 3.3.

Proposition 1 is also related to (Khan and Tamer 2010). To see this, suppose X consists of continuous coordinates only. Then $\Pr\{Z \in Q_b\} \to 0$ as b converges to β. That is, Q_b becomes a "thin set" as b approaches β.

It also follows from Proposition 1 that conditions that yield point identification of β in (Manski 1988) are also sufficient for point identification of β in the current model with CI and MI.

SV *(Sufficient Variation) For all x, V is continuously distributed with positive densities over $\Omega_{V|x}$, which includes $x\beta$ in the interior.*

FR *(Full Rank)* $\Pr\{X\gamma \neq 0\} > 0$ *for all nonzero vector* $\gamma \in \mathbb{R}^K$.

Note SV can be satisfied when the support of V given each x is bounded, provided the parameter space for β is bounded. It differs from the large support conditions needed to point identify β when errors are mean-independent, where the support of V needs to include the support of $-X\beta+\epsilon$ conditional on X. FR is a typical full-rank condition analogous to that in (Manski 1988). With the first coordinate in X being a constant intercept, FR implies that there exists no nonzero $\tilde{\gamma}$ in $\mathbb{R}^{K-1}$ and $c \in \mathbb{R}$ with $\Pr\{X_{-1}\tilde{\gamma} = c\} = 1$. FR implies $\Pr\{X(\beta - b) \neq 0\} > 0$ for any $b \neq \beta$. To see how these are sufficient for identification, suppose, without loss of generality, $\Pr\{X\beta < Xb\} > 0$. Under SV, for any x with $x\beta < xb$, there exists an interval of v with $x\beta \leq v < xb$. This implies $\Pr\{Z \in Q_b\} > 0$ 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 CI, MI, SV and FR hold in (1), and* $\Pr(X\beta = V) = 0$. *Then*

$$\beta = \arg\min_b \mathbb{E}_Z[1\{p(Z) \geq \tfrac{1}{2}\}(Xb - V)_- + 1\{p(Z) < \tfrac{1}{2}\}(Xb - V)_+] \tag{6}$$

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

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

$$\tilde{\beta} \equiv \arg\min \textstyle\sum_i \kappa\left(\hat{p}_i - \frac{1}{2}\right)(x_i b - v_i)_- + \kappa\left(\frac{1}{2} - \hat{p}_i\right)(x_i b - v_i)_+ \tag{7}$$

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 $[0, +\infty)$.

A few remarks about the asymptotic properties of the estimator and its comparison with the maximum score estimator are in order. If either SV or FR fails, then β is only

set-identified and the objective function in (6) have multiple minimizers. The estimator in (7) is a random set that is consistent for the identified set (under the Hausdorff set metric), under conditions that ensure the uniform convergence of the objective function in (7) to its population counterpart over the parameter space.[7] We conjecture the estimator converges at the cubic rate under conditions in (Chernozhukov, Hong, and Tamer 2007).[8]

Compared with the maximum score estimator, $\tilde{\beta}$ in (7) appears to have computational advantages once the propensity scores are estimated. The argument b enters the estimand continuously through $(.)_-$ and $(.)_+$, as opposed to in the indicator function in maximum score estimators. The flip side of our estimator is that it does require the choice of an additional smoothing parameters in $\hat{p}_i$.

3.2 Zero Fisher Information

We now show the information for β under CI and MI is zero, provided $Z = (X, V)$ has finite second moments and certain regularity condition on the coefficient and the error distribution holds. In addition to Section 3.1, our finding in this subsection provides an alternative way to formalize equivalence between the two models (i.e. binary regressions with "MI alone" versus "MI and CI") when it comes to estimating β.

RG *(Regularity) For each* $\left(b, G_{\epsilon|X}\right)$ *in the parameter space, there exists a measurable function* $q : \{0,1\} \otimes \Omega_Z \to \mathbb{R}$ *such that* $\left|\partial f^{1/2}\left(y, z; \eta, G_{\epsilon|X}\right)/\partial b\right| \leq q(y,z)$ *for all* η *in an neighborhood around* b*; and* $\int q^2(y,z)d\mu < \infty$.

RG is needed to establish mean-square differentiability of the square-root likelihood of (y, x) with respect to b for each $G_{\epsilon|X}$. Let Θ denote the parameter space for the distribution of ϵ given X, which needs to satisfy CI, MI and RG now. We show that a set of paths similar to those considered in Theorem 5 of (Chamberlain 1986) yields zero information for β under CI, MI and RG. Let Λ consist of paths

$$\lambda(\varepsilon, x; \delta) \equiv F_{\epsilon|x}(\varepsilon)\left[1 + (\delta - \delta_0)\, h(\varepsilon, x)\right], \tag{8}$$

where $F_{\epsilon|X}$ is the true conditional distribution in DGP from Θ; and $h : \mathbb{R}^{K+1} \to \mathbb{R}$ is continuously differentiable, is zero outside of some compact set; and satisfies $h(0, x) = 0$ for all $x \in \mathbb{R}^K$. Such a set of paths differs from those leading to zero information of β in a model under MI alone (without the exclusion restrictions in CI). In that latter

[7] The identified set for β is defined as the set of all coefficients that could generate propensity scores identical to that in the DGP for *all* z when paired with some nuisance parameter $F_{\epsilon|X}$ that satisfies CI and MI.

[8] 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 $\tilde{\beta}$.

case, the paths that lead to zero information is $\lambda(\varepsilon, x; \delta) \equiv F_{\epsilon|z}(\varepsilon)\left[1 + (\delta - \delta_0)\,\tilde{h}(\varepsilon, z)\right]$ with $\tilde{h} : \mathbb{R}^{K+2} \to \mathbb{R}$ continuously differentiable; is zero outside of some compact set; and satisfying $\tilde{h}(0, x, v) = 0$. (See (Chamberlain 1986) for details.)

Using arguments similar to (Chamberlain 1986), we can show $\lambda(.,.;\delta)$ in (8) 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:

$$\psi_k(y, z) \equiv \frac{1}{2}\left\{yF_{\epsilon|x}(w)^{-1/2} - (1-y)\left[1 - F_{\epsilon|x}(w)\right]^{-1/2}\right\} f_{\epsilon|x}(w)x_k \tag{9}$$

$$\psi_\lambda(y, z) \equiv \frac{1}{2}\left\{yF_{\epsilon|x}(w)^{-1/2} - (1-y)\left[1 - F_{\epsilon|x}(w)\right]^{-1/2}\right\} F_{\epsilon|x}(w)h(w, x) \tag{10}$$

where w is a shorthand for $x\beta - v$. Again note the excluded regressor v is dropped from $F_{\epsilon|x}$ and $f_{\epsilon|x}$ due to CI.

Proposition 2 *Suppose CI, MI, SV, FR and RG hold in (1); Z has finite second moments; and* $\Pr(X\beta = V) = 0$. *Then the information for β_k is zero for all $k \leq K$.*

Proof of Proposition 2 is similar to that of Theorem 5 in (Chamberlain 1986) for binary regressions under MI alone, and is omitted for brevity. It suffices to note that the main difference between Proposition 2 and Theorem 5 in (Chamberlain 1986) is that the path leading to zero information for β under the additional assumption of CI has to respect the exclusion restriction imposed on V.

A few remarks related to this zero information result are in order. First, the zero information for β_k under CI and MI is closely related to two facts: there is no incremental identifying power for β from CI given MI; and there is zero information for β_k under MI alone. 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 parametric rate can be attained in estimation of β under CI and mean-independent errors. A similar result holds in the current model under CI and median independent errors as well. Third, if there are multiple excluded regressors satisfying CI (i.e. V is a vector rather than scalar), then, after a scale normalization (e.g. setting one of coefficients for V to have absolute value 1), the information for the other coefficients is positive and root-n estimation of coefficients for V exist (e.g. using the average-derivative approach).

3.3 Error Distribution and Average Structural Function

The previous subsections show the presence of a special regressor does not help improve the identification or information of coefficients for the non-special regressors with median-independent errors. In contrast, this subsection provides a positive perspective on the

role of excluded regressors by explain how they help to predict counterfactual choice probabilities and estimate average structural functions.

First, exclusion restriction *does* help recover the heteroskedastic error distributions, which in turn is useful for counterfactual predictions. To see how to recover $F_{\epsilon|X}$ under MI and CI, note $\mathbb{E}(Y|x,v) = F_{\epsilon|x}(x\beta - v)$. With β identified, $F_{\epsilon|x}(t)$ can be recovered for all t over the support of $X\beta - V$ give $X = x$ as $\mathbb{E}(Y|X = x, V = x\beta - t)$. To get counterfactual predictions, consider a stylized model of retirement decisions. Let $Y = 1$ if the individual decides to retire and $Y = 0$ otherwise. The decision is given 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 individual's assets. Suppose conditional on age and health, asset values are uncorrelated with idiosyncratic elements (e.g. unobserved family factors such as money or energy spent on offsprings). Suppose we want to predict retirement patterns among another population of senior workers not observed in data, which has the same β_1 and $F_{\epsilon|X_1,X_2}$ but different weights for health status $\tilde{\beta}_2$ (where $\tilde{\beta}_2 > \beta_2$). Then knowledge of $F_{\epsilon|X}$ as well as β_1, β_2 helps at least bound the counterfactual retirement probabilities conditional on $Z \equiv (X_1, X_2, V)$. If the magnitude of the difference between $\tilde{\beta}_2$ and β_2 is also known, then point-identification of such a counterfactual conditional retirement probability is also attained for z, provided 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.)

Second, exclusion restriction helps identify the average structural function defined in (Blundell and Powell 2003) under the large support condition of V. To see this, note 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 $\Omega_{V|x} = \mathbb{R}^1$ for all $x \in \Omega_X$, then

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

where

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

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 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 $\Omega_{V|s}$.

Based on the analog principle, we propose the following estimator for the average structural function as follows:

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

where

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

with $\mathcal{K}_\sigma(.) \equiv \sigma^{-(k+1)}\mathcal{K}(./\sigma^{k+1})$ where $\mathcal{K}$ is a product kernel; and $\tilde{\beta}$ being some first-stage preliminary estimator such as the one defined in (7), or the maximum score estimator as proposed in (Manski 1985).

4 Exclusion plus Conditional Symmetry

This section discusses identification and information of β under CI while the location restriction of median independence (MI) is replaced by the stronger location and shape restriction of conditional symmetry (CS). To motivate the CS assumption in binary regressions, suppose the latent utility associated with binary actions are $h_j(z)+\varepsilon_j$ for $j \in \{0,1\}$; and the action is governed by $Y = 1\{h_1(Z) + \epsilon_1 \geq h_0(Z) + \epsilon_0\} = 1\{h^*(Z) \geq \epsilon^*\}$, where $h^* \equiv h_1 - h_0$ and $\epsilon^* \equiv \epsilon_0 - \epsilon_1$. As long as ϵ_1 and ϵ_0 are i.i.d. draws from the same marginal, the normalized error ϵ^* must be symmetrically distributed around 0 given Z.

In Section 4.1, we characterize a set of paired states (z, z') that help distinguish β from some $b \neq \beta$ based on observed propensity scores. Building on this result, we then specify sufficient conditions for the point identification of β. In Section 4.2 we show the Fisher information for β is zero under mild regularity conditions. We then conclude this section with the introduction of a root-N estimator for β.

4.1 Identification

Our first finding is that replacing MI with CS while maintaining CI *does* help with the identification of β. Let $X \equiv (X_c, X_d)$, with X_c and X_d denoting continuous and discrete coordinates respectively. Let Θ_{CS} denote parameter space for the distribution of ϵ given X under the restrictions of CI and CS. We need further restrictions on Θ_{CS} due to continuous coordinates in X_c.

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

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

This condition requires the 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}.[9]

[9] An alternative way to formulate EC 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 away from zero by a positive constant.

Such an equicontinuity condition is needed because the identification of β relative to $b \neq \beta$ states that b cannot be paired with *any* $G_{\epsilon|X} \neq F_{\epsilon|X}$ in Θ_{CS} to generate propensity scores identical to those from the true DGP at all pairs (z, z'). It is a technicality introduced only due to the need to modify the definition of identification in (Manski 1988) when X contains continuous coordinates. A sufficient condition for EC is that all $G_{\epsilon|X}$ in Θ_{CS} are Lipschitz-continuous with their modulus uniformly bounded by a finite constant.

To formally quantify the incremental identifying power due to CS, define:

$$R_b(x) \equiv \left\{(v_i, v_j) : x\beta < \frac{v_i + v_j}{2} < xb \text{ or } x\beta > \frac{v_i + v_j}{2} > xb\right\} \tag{11}$$

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 $F_{V_i|X}$. In addition we also need the joint distribution of V and X_c given X_d to be continuous.

CT *(Continuity) For any x_d, the distribution $F_{V,X_c|x_d}$ is continuous with positive densities almost everywhere with respect to the Lesbegue measure.*

Under CT, 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.

Proposition 3 *Under CI, CS, EC and CT, β is identified relative to b if and only if either (i) $\Pr\{Z \in Q_b\} > 0$; or (ii) there exists a set ω open in Ω_X such that for all $x \in \omega$,*

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

Proof of Proposition 3 is included in Appendix A. To see intuition for this result, consider a simple model where X only consists of discrete regressors. 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 either

$$\begin{array}{rcl} \text{“}x_i\beta - v_i & < & -(x_j\beta - v_j) \text{ and } x_ib - v_i > -(x_jb - v_j)\text{”} \\ & & \text{or} \\ \text{“}x_i\beta - v_i & > & -(x_j\beta - v_j) \text{ and } x_ib - v_i < -(x_jb - v_j)\text{”} \end{array} \tag{13}$$

for $(z_i, z_j) \in \tilde{Q}_{b,S}$. In the former case, the true propensity scores 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 necessarily add up to be greater than 1. This suggests any pair (z_i, z_j) from $\tilde{Q}_{b,S}$ should

help distinguish β from $b \neq \beta$, as the sign of $p(z_i) + p(z_j) - 1$ differs from that of $(x_i b - v_i) + (x_j b - v_j)$. 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, $(b, G_{\epsilon|X})$ yields propensity scores identical to the true propensity scores in the DGP almost everywhere.

To extend this intuition when there are continuous coordinates in X, we invoke EC and CT as additional restrictions on the parameter space for $F_{\epsilon|X}$. With continuous coordinates in X, $\Pr\{(Z, Z') \in \tilde{Q}_{b,S}\} = 0$ for all $b \neq \beta$. However, under EC and CT, the inequalities in (13) also hold for paired states in some small "δ-expansion" of $\tilde{Q}_{b,S}$ defined as:

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

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.

(Manski 1988) showed in a model without excluded regressors that strengthening median independence into conditional symmetry does not add to the identifying power for β. He showed the sets of states that help distinguish β from $b \neq \beta$ under both cases are the same. Our finding in Proposition 3 shows this equivalence fails when the vector of states contain an excluded regressor: Strengthening MI into CS leads to an additional set of paired states R_b that help identify β relative to b.Thus in that sense a regressors being special does add to the informational content of the model.

Finally, note by construction any condition that identifies β under CI and MI also identifies β under CI and CS. Under FR, for all $b \neq \beta$, there exists an open set $\omega \subseteq \Omega_X$ with $x\beta \neq xb$ for all $x \in \omega$. SV then implies either $\int 1\{x\beta < \frac{v_i + v_j}{2} < xb\} dF_{V_i, V_j|x} > 0$ or $\int 1\{xb < \frac{v_i + v_j}{2} < x\beta\} dF_{V_i, V_j|x} > 0$ for all $x \in \omega$. This is because under SV, V_i and V_j are independent draws from $F_{V|x}$ and both fall in an open neighborhood around $x\beta$ with positive probability. Identification of β follows from Proposition 3.

4.2 Positive Fisher Information

We now show how CS together with CI leads to positive information for β in heteroskedastic binary regressions. (Zheng 1995) showed without any excluded regressors the information for β is zero in binary regressions with a conditionally symmetric error distribution. In contrast, we show in this subsection that with excluded regressors, the conditional symmetry of error distribution does lead to positive information for β under mild regularity conditions. This demonstrates a further link between the two distinct notions of information we have discussed in this paper. We then build on this result to propose a

new root-N consistent estimators for β in the next subsection.

CS' *CS holds; and there exists an open interval $\mathcal{I}^*$ around 0 and a constant $c > 0$ such that for all $x \in \Omega_X$ $f_{\epsilon|x}(\varepsilon) \geq c$ for all $\varepsilon \in \mathcal{I}^*$.*

RG' *RG holds and for any $\bar{\omega}$ such that $\Pr(X \in \bar{\omega}) > 0$, there exists no nonzero $\alpha \in \mathbb{R}^K$ such that $\Pr\{X\alpha = 0 | X \in \bar{\omega}\} = 1$.*

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

$$\lambda(\varepsilon, x; \delta) = 1 - \lambda(-\varepsilon, x; \delta) \text{ for all } \varepsilon, x \in \Omega_{\epsilon,X}; \tag{14}$$

and (iii) 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:

$$\psi_\lambda(y, z) \equiv \frac{1}{2}\left\{ y F_{\epsilon|x}(w)^{-1/2} - (1-y)\left[1 - F_{\epsilon|x}(w)\right]^{-1/2} \right\} \lambda_\delta(w, x; \delta_0) \tag{15}$$

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

Proposition 4 *Under CI, CS', EC, CT, FR, SV and RG', the information for β_k is positive for all k.*

Proof of Proposition 4 is presented in the appendix. We sketch the heuristics of the idea here. Exploiting 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

$$\inf_{\lambda \in \Lambda} 4 \int \phi(z) \left[f_{\epsilon|x}(w) \left(x_k - \textstyle\sum_{j \neq k} \alpha_j^* x_j \right) - \alpha_\lambda^* \lambda_\delta(w, x; \delta_0) \right]^2 dF_Z \tag{16}$$

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 (5) 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 $\alpha_\lambda^* \neq 0$. (Otherwise the pathwise information $I_{\lambda,k}$ under λ would equal that of a parametric model where true error distribution $F_{\epsilon|X}$ is known, and 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 nonzero α_λ^*) must be odd functions in w once x is fixed. At the same time, $f_{\epsilon|x}(w)$ is an even function of w (i.e. 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 $W = X\beta - V$ falls to both sides of zero with positive probabilities, the information

for β_k must be positive because an even function can never approximate an odd function well enough to reduce $I_{\lambda,k}$ arbitrarily close to zero.

Some discussions relating Proposition 4 to the existing literature are in order. Recall the model in (Zheng 1995) where ϵ is symmetric around 0 given $Z = (X, V)$ with unrestricted dependence between ϵ and *all* coordinates in Z. The information for β_k is zero in that case because the scores ψ_k and ψ_λ are both flexible in the sense of depending on V as well as X. Hence one can construct linear combinations of $(\psi_j)_{j\neq k}$ and ψ_λ that are arbitrarily good approximation to ψ_k in $L^2(\mu)$-norm, provided for the path λ is appropriately a chosen. To see this, note $I_{\lambda,k} \equiv$

$$\inf_{\alpha_\lambda,(\alpha_j)_{j\neq k}} \int \left(\psi_k - \alpha_\lambda \psi_\lambda - \textstyle\sum_{j\neq k} \alpha_j \psi_j\right)^2 d\mu \leq 4 \int \phi(z) \left[f_{\epsilon|z}(w) x_k - \lambda_\delta(w, z; \delta_0)\right]^2 dF_Z. \tag{17}$$

Indeed the same path used for showing zero information for β_k under MI with no excluded regressors (see Theorem 5 in (Chamberlain 1986)) also drives the information for β_k to zero in (Zheng 1995). Specifically, the path is $\lambda(\varepsilon, z; \delta) = F_{\epsilon|z}(\varepsilon)\left[1 + (\delta - \delta_0) h(\varepsilon, z)\right]$, where h is continuously differentiable, equals zero outside of some compact set, and $h(0, z) = 0$ for all z so that $\lambda_\delta(\varepsilon, z; \delta_0) = h(\varepsilon, z)$. Since there is no restriction on how the vector z enters λ_δ, one can exploit such flexibility to make the approximation on the right-hand side of (17) arbitrarily good and establish zero information in (Zheng 1995)'s case. In contrast, in our model under CI as well as CS, the excluded regressor v can only enter $\lambda_\delta(w, x; \delta_0)$ through the index $w = x\beta - v$. This additional form restriction is what delivers the positive information for β_k.

(Magnac and Maurin 2007) considered binary regressions under CI, mean-independent errors ($E(\epsilon|X) = 0$), and some tail conditions that restrict the truncated expectation of $F_{\epsilon|X}$ outside of the support of V given X.[10] They showed the information for β_k 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 behaviors outside the support of V. In comparison, the conditional symmetry condition (CS) considered here in Section 4 is a transparent restriction on the shape of $F_{\epsilon|X}$ over its full support. The conditions in Section 4 and those in (Magnac and Maurin 2007) are non-nested. (See Appendix B for detailed discussions.)

4.3 Root-N Estimation: Extremum Estimator

Our findings in Proposition 4 suggest root-N regular estimators for β can be constructed. We consider the case where all coordinates in Z are continuously distributed. Extensions

[10] See equation (5) in Proposition 5 of (Magnac and Maurin 2007) for the tail restriction. Essentially, this is sufficient and necessary for to extending the proof of idenification of β in (Lewbel 2000), a model with CI and mean independence, when the support of the excluded regressor V is bounded between $v_L > -\infty$ and $v_H < \infty$.

to cases with mixed co-variates are straightforward and omitted for brevity.

Let $(.)_- \equiv -\min\{.,0\}$ and $(.)_+ \equiv \max\{.,0\}$. Our estimator is

$$\hat{\beta} \equiv \arg\min_{b\in\mathcal{B}} \hat{H}_n(b), \tag{18}$$

where:

$$\hat{H}_n(b) \equiv \tfrac{1}{n(n-1)} \textstyle\sum_{j\neq i} K_h\left(x_i - x_j\right)\left[\kappa\left(\hat{w}_{i,j}-1\right)\varphi^-(Z_i,Z_j;b) + \kappa(1-\hat{w}_{i,j})\varphi^+(Z_i,Z_j;b)\right];$$
$$\varphi^-(z_i,z_j;b) \equiv \left(\tfrac{(x_i+x_j)'}{2}b - \tfrac{v_i+v_j}{2}\right)_- \text{ and } \varphi^+(z_i,z_j;b) \equiv \left(\tfrac{(x_i+x_j)'}{2}b - \tfrac{v_i+v_j}{2}\right)_+;$$
$$\hat{w}_{i,j} \equiv \hat{p}_i + \hat{p}_j; \text{ and } \hat{p}_l \equiv \hat{p}(z_l) \equiv \frac{\sum_{s\neq l} y_s \mathcal{K}_\sigma\left(z_s - z_l\right)}{\sum_{s\neq l} \mathcal{K}_\sigma\left(z_s - z_l\right)} \text{ for } l = i,j.$$

where $K_h(.) \equiv h^{-k}K(./h^k)$ and $\mathcal{K}_\sigma(.) \equiv \sigma^{-(k+1)}\mathcal{K}(./\sigma^{k+1})$, with $K, \mathcal{K}$ and h_n, σ_n being kernel functions and bandwidths whose properties are to be specified below. The weighting function κ satisfies the following properties.

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

The weight function, evaluated at $\hat{w}_{i,j} - 1$, could be intuitively interpreted as a smooth replacement for the indicator function $1\{\hat{w}_{i,j} \geq 1\}$. To derive asymptotic properties of $\hat{\beta}$, we first show $\hat{H}_n$ converges in probability to a limiting function H_0 uniformly over the parameter space, where

$$H_0(b) = \mathbb{E}\left\{f(X)\mathbb{E}\left[\kappa(W_{i,j}-1)\varphi^-(Z_i,Z_j;b) + \kappa(1-W_{i,j})\varphi^+(Z_i,Z_j;b)\middle| X_j = X,\ X_i = X\right]\right\} \tag{19}$$

where f is the true density for non-special regressors X in the data-generating process; and $w_{i,j}$ is the sum of true propensity scores $p(z_i)$ and $p(z_j)$. The inner expectation of (19) is taken with respect to V_i, V_j given $X_j = X_i = X$ while the outer expectation is taken w.r.t. X (distributed according to f). The next proposition shows β is identified as the unique minimizer of H_0 in $\mathcal{B}$.

Proposition 5 *Suppose CI, CS, EC, CT, SV, FR and WF hold. Then* $H_0(b) > 0$ *for all* $b \neq \beta$ *and* $H_0(\beta) = 0$*.*

Proof of this proposition follows from arguments similar to that of Proposition 3, and is included in Appendix C. We now list the conditions for our estimator to be consistent.

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

SM1 *(Smoothness) (i) The density of $Z = (X, V)$ is bounded away from zero by some positive constant over its compact support. (ii) The density of Z and the propensity score $p(Z)$ is $m_{\mathcal{K}}$-times continuously differentiable (where $m_{\mathcal{K}} \geq k + 2$); and the derivatives are all Lipschitz continuous. (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, $\mathbb{E}[\tilde{\varphi}(Z_i, Z_j; b)| X_i = x_i, X_j = x_j] f(x_j)$ is twice continuously differentiable in x_j around $x_j = x_i$, where*

$$\tilde{\varphi}(z_i, z_j; b) \equiv \kappa(w_{i,j} - 1)\varphi^-(z_i, z_j; b) + \kappa(1 - w_{i,j})\varphi^+(z_i, z_j; b).$$

KF1 *(Kernel Function for Estimating Propensity Scores) (i) $\mathcal{K}$ is the product of $k + 1$ univariate kernel functions (denoted $\tilde{K}$), each of which is symmetric around 0, bounded over a compact support, and integrates to 1. (ii) The order of $\tilde{\mathcal{K}}$is $m_{\mathcal{K}}$. (iii) $\|t\|^l \tilde{\mathcal{K}}(t)$ is Lipschitz continuous for $0 \leq l \leq m_{\mathcal{K}}$.*

BW1 *(Bandwidth for Estimating Propensity Scores) σ_n is proportional to $n^{-\rho_\sigma}$, where $\rho_\sigma \in \left(\frac{1}{2m_{\mathcal{K}}}, \frac{1}{2(k+1)}\right)$.*

FM1 *(Finiteness) $\mathbb{E}\{[\mathcal{C}(X_i, X_j) - (V_i + V_j)/2]^2\}$ and $\mathbb{E}\{[\mathcal{D}(X_i, X_j) - (V_i + V_j)/2]^2\}$ are finite, where $\mathcal{C}(X_i, X_j) \equiv \inf_{b\in\mathcal{B}}(X_i + X_j)'b/2$ and $\mathcal{D}(X_i, X_j) \equiv \sup_{b\in\mathcal{B}}(X_i + X_j)'b/2$.*

KF2 *(Kernel Functions for Matching) $K(.)$ is the product of k univariate kernel functions (each denoted $\tilde{K}(.)$) such that (i) $\tilde{K}(.)$ is bounded over a compact support, symmetric around 0 and integrates to one. (ii) The order of $\tilde{K}(.)$ is m_φ, where $m_\varphi > 2k$.*

BW2 *(Bandwidths for Matching) h_n is proportional to $n^{-\rho_h}$ with $\rho_h \in \left(\frac{1}{4k}, \frac{1}{3k}\right)$.*

Proposition 6 *Suppose conditions for Proposition 5 hold; and in addition, PS, SM1, FM1, KF1,2 and BW1,2 also hold. Then $\hat{\beta} \xrightarrow{p} \beta$.*

Proof of Proposition 6 amounts to checking conditions for basic consistency theorems for extreme estimators, such as Theorem 4.1 in (Amemiya 1985) and Theorem 2.1 in (Newey and McFadden 1994). A key step of the proof is to show that our objective function $\hat{H}_n$ converges the limiting function H_0 uniformly over the parameter space. Our approach is to first show the difference between the objective function to an infeasible version, where estimates for propensity scores $\hat{p}(z)$ are replaced by the truth $p(z)$, is negligible in a uniform sense. Since the infeasible objective function takes the form of a second-order U-process indexed by $b \in \mathcal{B}$, it can be decomposed by the H-decomposition into the sum of an unconditional expectation involving the matching kernel; and two degenerate U-processes with orders one and two respectively. We then use known results from (Sherman 1994b) to show the two U-processes converge to 0 uniformly over $\mathcal{B}$ given

our choices of kernels and bandwidths; and show the unconditional expectation is $H_0(b) + o(1)$ for all b by a standard approach of changing variables.

The kernel and bandwidth conditions in KF1 and BW1, together with smoothness conditions (i)-(iii) in SM1, ensure the preliminary estimates of propensity scores converge uniformly to the true propensity score from data-generating process. This is useful for showing that replacing $\hat{p}$ with the true propensity scores only results in negligible differences. The choice of ρ_σ in BW1 ensures that: (a) the order of the part of mean-square error due to bias is dominated by that ascribed to variance (i.e. $1/\sqrt{n\sigma_n^{k+1}} > \sigma_n^{m\kappa}$); (b) the resulted rates of uniform converge of $\hat{p}$ is faster than $n^{-1/4}$ (i.e. $1/\sqrt{n\sigma_n^{k+1}} < n^{-1/4}$); and (c) the order of $\sigma_n^{m\kappa}$ is smaller than $o(n^{-1/2})$. The requirements (b) and (c) are sufficient but not necessary for consistency. As is explained later, (b) and (c) help to show a quadratic approximation of the objective function is accurate enough in a uniform sense over certain shrinking neighborhood around the true β to lead to the parametric rate.

BW2 is also sufficient but not necessary for consistency. This is because the uniform convergence of U-processes over $\mathcal{B}$ in the H-decomposition only require $n^{-1/2}h_n^{-k}$ (and therefore $n^{-1}h_n^{-k}$) to be $o(1)$; and the convergence of the unconditional expectation only requires $h_n \to 0$. Nonetheless, just as with σ_n, the specific range of magnitude for h_n is needed for showing the quadratic approximation H_0 uniformly over $\mathcal{B}$ is fast enough to induce the parametric rate of our estimator. Continuity of H_0 in SM1-(iv) is a necessary condition for applying the consistency theorem for extremum estimators. The other conditions in SM1 are also useful for showing uniform convergence of $\hat{H}_n$ to H_0. The finiteness condition in FM1 is instrumental as it is need for applying the results on uniform convergence of degenerate U-processes in (Sherman 1994b).

To establish that $\hat{\beta}$ attains the parametric rate with normal limiting distribution, we need the following additional restriction on smoothness and finiteness of some population moments. To simplify notations, let $\Delta\varphi_{i,j}^-(b) \equiv \varphi^-(Z_i, Z_j; b) - \varphi^-(Z_i, Z_j; \beta)$ and likewise define $\Delta\varphi_{i,j}^+$. Let $\kappa_-(W_{i,j}) \equiv \kappa(W_{i,j} - 1)$ and $\kappa_+(W_{i,j}) \equiv \kappa(1 - W_{i,j})$; and let $\kappa'_-(W_{i,j}) \equiv \kappa'(W_{i,j} - 1)$ and $\kappa'_+(W_{i,j}) \equiv \kappa'(1 - W_{i,j})$.

SM2 *(Smoothness of Population Moments) (i) $H_0(b)$ is twice continuously differentiable in an open neighborhood around β. (ii) For all x and x', $\bar{\varphi}^-(x, x'; b)$ and $\bar{\varphi}^+(x, x'; b)$ are twice continuously differentiable in b in an open neighborhood around β where for $\Diamond \in \{+, -\}$,*

$$\bar{\varphi}^\Diamond(x, x'; b) \equiv \mathbb{E}\left[\kappa_\Diamond(W_{i,j})\Delta\varphi_{i,j}^\Diamond(b) | X_i = x, X_j = x'\right].$$

For all x', $\nabla_b\bar{\varphi}^-(x, x'; \beta)f(x)$ and $\nabla_b\bar{\varphi}^+(x, x'; \beta)f(x)$ are m_φ-times continuously differentiable in x at $x = x'$ with bounded derivatives; and $\nabla_{bb}\bar{\varphi}^-(x, x'; \beta)f(x)$ and $\nabla_{bb}\bar{\varphi}^+(x, x'; \beta)f(x)$ are both continuously differentiable in x at $x = x'$ with bounded derivatives. (iii) For all x, x', $\varpi^-(x, x'; b)$ and $\varpi^+(x, x'; b)$ are continuously differentiable in an open neighborhood around β, where for $\Diamond \in \{+, -\}$,

$$\varpi^\Diamond(x, x'; b) \equiv \mathbb{E}\left[|\Delta\varphi_{i,j}^\Diamond(b)| \; | X_i = x, X_j = x'\right].$$

For all x', $\nabla_b \varpi^-(x,x';\beta)f(x)$ and $\nabla_b \varpi^+(x,x';\beta)f(x)$ are continuously differentiable in x around $x = x'$with bounded derivatives. (iv) For all $z \equiv (x,v)$ and all b in an open neighborhood around β, $\tilde{\mu}^-(z,x';b)f(x')$ and $\tilde{\mu}^+(z,x';b)f(x')$ are m_φ-times continuously differentiable with respect to X' around $X' = x$, where for $\Diamond \in \{+,-\}$

$$\tilde{\mu}^{\Diamond}(z,x';b) \equiv E[\kappa'_{\Diamond}(W_{i,j})\Delta\varphi^{\Diamond}_{i,j}(b)|Z_i = z, X_j = x'].$$

The derivatives are all bounded over support of z. (v) For all z, $m^*_-(z;b)$ and $m^*_+(z;b)$ are continuously differentiable in b around β with bounded derivatives, where for $\Diamond \in \{+,-\}$

$$m^*_{\Diamond}(z;b) \equiv \nabla w(z) f(x) \tilde{\mu}^{\Diamond}(z,x;b), \text{ with } \nabla w(z) \equiv \left[1/f(z), -p(z)f(z)/f(z)^2\right].$$

Besides, $\nabla_b m^*_-(z;\beta)f(z)$ and $\nabla_b m^*_+(z;\beta)f(z)$ are both $m_{\mathcal{K}}$-times continuously differentiable with bounded derivatives in z over its full support.

FM2 *(Finiteness of Population Moments) (i) There exists an open neighborhood around β in $\mathcal{B}$, denoted $\mathcal{N}(\beta)$, such that*

$$\int \sup_{b\in\mathcal{N}(\beta)} \left\|\nabla_{bb}\bar{\varphi}^{\Diamond}(x,x;b)\right\| f(x)dF(x) < \infty \text{ and } \int \sup_{b\in\mathcal{N}(\beta)} \left\|\nabla_b \varpi^{\Diamond}(x,x;b)\right\| f(x)dF(x) < \infty,$$

for $\Diamond \in \{+,-\}$. (ii) For $\Diamond \in \{+,-\}$, $\int \left\|\nabla_b m^*_{\Diamond}(z;\beta)f(z)\right\| < \infty$ and there exists $\tilde{\varepsilon} > 0$ such that

$$E[\sup_{\|\tilde{\varepsilon}\|\geq 0} \left\|\nabla_b m^*_{\Diamond}(Z+\tilde{\varepsilon};\beta)f(Z+\tilde{\varepsilon})\right\|^4] < \infty.$$

(iii) For $\Diamond \in \{+,-\}$, $\int \left\|\nabla_b m^*_{\Diamond}(z;\beta)f(z)\right\| dz < \infty$.

Let $Q \equiv (Y,1)$. Define $\delta^* = \delta^*_- + \delta^*_+$, where for subscripts $\Diamond \in \{+,-\}$,

$$\delta^*_{\Diamond}(y,z) \equiv q\nabla_b m^*_{\Diamond}(z;\beta)f(z) - \mathbb{E}[Q\nabla_b m^*_{\Diamond}(Z;\beta)f(Z)]$$

Proposition 7 *Suppose conditions for Proposition 6 hold. Under additional conditions SM2 and FM2,*

$$\sqrt{n}\left(\hat{\beta} - \beta\right) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1}\Omega\left(\Sigma^{-1}\right)')$$

where

$$\Sigma \equiv \nabla_{bb} H_0(\beta) \text{ and } \Omega \equiv 4\mathbb{E}[\delta^*(Y,Z)\delta^*(Y,Z)'].$$

The proof follows steps similar to (Khan 2001). The continuity of H_0 under SM1-(v) is strengthened to SM2-(i), which helps showing that the limiting function H_0 to have quadratic approximation that is sufficiently precise over an open neighborhood around β.

4.4 Root-N Estimation: Close-Form Estimator

This subsection introduces an alternative estimator under CI and CS that has a close form:

$$\hat{\beta}_{CF} \equiv \left[\sum_{i,j} K_1\left(\frac{x_i - x_j}{h_{1,n}}\right) K_2\left(\frac{\hat{p}_i + \hat{p}_j - 1}{h_{2,n}}\right)(x_i + x_j)'(x_i + x_j)\right]^{-1} \times \left[\sum_{i,j} K_1\left(\frac{x_i - x_j}{h_{1,n}}\right) K_2\left(\frac{\hat{p}_i + \hat{p}_j - 1}{h_{2,n}}\right)(x_i + x_j)'(v_i + v_j)\right] \quad (20)$$

where K_1 is a product kernel; K_2 is a univariate kernel; $\hat{p}_i, \hat{p}_j$ are kernel estimates of propensity scores as before; and $h_{1,n}, h_{2,n}$ are sequences of bandwidths. The intuition for this estimator is as follows: Suppose one can collect pairs of observations z_i, z_j with $i \neq j$ such that $x_i = x_j$ and $p_i + p_j = 1$. Then CI and CS imply for any such pair, $v_i + v_j = (x_i + x_j)'\beta$. The estimator in (20) implements this intuition by using kernel smoothing to collect such matched pairs of z_i, z_j and then estimates the coefficient by finding a vector that provides the best linear fit of $v_i + v_j$ as a function of $x_i + x_j$.

A couple of remarks regarding close-form and extremum estimators are in order. The close-form estimator has a computational advantage in that it does not require minimizing a non-linear objective function. On the other hand, it does require choosing three bandwidths: two in the matching kernels K_1, K_2 and one in the kernel $\mathcal{K}$ for estimating $\hat{p}_i, \hat{p}_j$. In comparison, the extremum estimator in the previous subsection requires two choices of bandwidths: one in K and one in $\mathcal{K}$. We do not expect the additional choice of bandwidth in K_2 in the close-form estimator to pose much computational problem due to its low dimension. We also conjecture the close-form estimator has a smaller asymptotic variance than the extremum estimator.

The extremum estimator, on the other hand, has an advantage of being robust to the loss of point identification in the following sense: In case the conditions for point identifying β under CI and CS (e.g. SV and FR) fail, the estimator in (18) is consistent for the set $\{b : H_0(b) = 0\}$, or the identified set of coefficients under CI and CS due to Proposition 3.[11] This is in part due to the uniform convergence of $\hat{H}_n$ in (18) to H_0 over $\mathcal{B}$ (as shown in Appendix C).

A comparison between the close-form and extremum estimators under CI and CS is reminiscent of that between Ichimura's two-step estimator in (Ichimura 1993) and the maximum score estimator in (Manski 1985). In the latter comparison, both are estimators for β in binary regressions under MI alone. Ichimura's estimator has a close-form and involves an additional choice of bandwidth in the preliminary estimates of propensity scores, while Manski's maximum score estimator has no close form but does not require any choice of bandwidth.

[11] In the context of set estimators, consistency can be defined using the Hausdorff set metric as in (Manski and Tamer 2002).

We conclude this subsection with a technical note on the number of observations used in both estimators. The close-form estimator uses n^2 pairs of observations in total, including n pairs with $i = j$. For such pairs with $i = j$, the sums in the square brackets in (20) are reduced to $4K_1(0)\sum_i K_2\left(\frac{\hat{p}_i - 1/2}{h_{2,n}}\right) x_i' x_i$ and $4K_1(0)\sum_i K_2\left(\frac{\hat{p}_i - 1/2}{h_{2,n}}\right) x_i' v_i$ respectively. That is, if we were to use unpaired observations only (as opposed to pairs) in the summand of (20), then the estimator would be reduced to the two-step close-form estimator proposed by (Ichimura 1993) for the case of heteroskedastic binary regressions under MI alone (known to converge at a rate slower than $\sqrt{n}$).

By the same token, the extremum estimator in the preceding subsection could also be modified to include n pairs with $i = j$. Indeed, if H_n in (18) were to be defined only using pairs with $i = j$, then it would lead to an estimator numerically equivalent to the one proposed under CI and MI in (7), which is known to converge at a rate slower than root-N due to zero information for β under CI and MI (Proposition 2).

While the inclusion of "$i = j$" pairs in the definition of extremum and close-form estimators has asymptotically negligible impact on these estimators, we expect them to improve the finite sample performance of both estimators.

4.5 Monte Carlo

We now present some simulation evidence for performance of our extremum estimator and the two-step, inverse-density-weighted estimator in (Lewbel 2000). The estimator in (Lewbel 2000) was introduced under CI and mean independence, which is a weaker set of assumptions than CI and CS.

We report performance of both estimators under four designs of data-generating processes (DGP). In all four, $Y = 1\{\alpha + X\beta + V + \epsilon \geq 0\}$ where V is a scalar variable following the standard normal distribution. Both X and ϵ are scalar variables. In the first three designs, the triplet (X, V, ϵ) are mutually independent, and we experiment with three sets of parametric specifications of marginal distributions for (X, ϵ), where both of them are either (a) standard normal; (b) standard logistic; or (c) standard Laplace. We also include a fourth design to allow for heteroskedastic errors by letting $\epsilon = (1 + |X|)U$, where X, V, U are mutually independent and all standard normal. We choose these distributional designs in order to understand better the performance of these estimators when the errors have different thickness of tails. Among the three parametric classes of normal, logistic and Laplace, the normal distribution has the thinnest tail while the logistic distribution has the thickest.

The true values for α and β in DGP are set to 0.2 and 0.5 respectively. For each choice of sample sizes (N = 50, 100, 200, 400 and 800), we simulate 1000 data sets and apply our extremum estimator (labeled as "Pairwise") and the inverse-density weighted

estimator in (Lewbel 2000).

Following the rule-of-thumb for bandwidths in kernel density and regression estimates (Section 1.7 and Section 2.2 in (Li and Racine 2007)), we use $\sigma_n = n^{-1/6}$ for the product kernel $\mathcal{K}$ while estimating the propensity scores at z in the pairwise estimator. For the same reason, we use $n^{-1/5}$ in the univariate kernel while estimating $f(v)$ in the inverse-density weighted estimator. The matching kernel $K(.)$ in the pairwise estimator can be viewed as a kernel for estimating the univariate density of $X_i - X_j$ at 0. Thus following the same rule-of-thumb, we choose $h_n = \sqrt{2}n^{-1/5}$ in $K(.)$.

We report descriptive statics from sampling distributions of these estimators out of 1000 simulations. These include bias, standard deviation, square-root of mean-square errors, and median absolute deviations).

Table 1(a): $X \tilde{} Normal(0,1)$, $\epsilon \tilde{} Normal(0,1)$

			$N = 50$	$N = 100$	$N = 200$	$N = 400$	$N = 800$
$Bias$	Pairwise	α	0.0366	0.0376	0.0352	0.0254	0.0251
		β	0.0160	0.0079	-0.0048	-0.0047	-0.0081
	Inverse-DW	α	-0.0448	-0.0356	-0.0423	-0.0456	-0.0443
		β	-0.1490	-0.1417	-0.1408	-0.1375	-0.1369
Std	Pairwise	α	0.4532	0.3045	0.2210	0.1511	0.1041
		β	0.5268	0.3620	0.2479	0.1730	0.1235
	Inverse-DW	α	0.1912	0.1361	0.0967	0.0681	0.0505
		β	0.1832	0.1287	0.0922	0.0671	0.0493
$RMSE$	Pairwise	α	0.4544	0.3066	0.2238	0.1532	0.1070
		β	0.5268	0.3619	0.2479	0.1730	0.1237
	Inverse-DW	α	0.1962	0.1406	0.1058	0.0819	0.0672
		β	0.2360	0.1914	0.1683	0.1530	0.1455
MAD	Pairwise	α	0.3095	0.2092	0.1542	0.1034	0.0726
		β	0.3239	0.2349	0.1635	0.1238	0.0835
	Inverse-DW	α	0.1301	0.0953	0.0724	0.0583	0.0498
		β	0.1728	0.1497	0.1395	0.1401	0.1384

Table 1(b): $X \tilde{} Laplace(0,1)$, $\epsilon \tilde{} Laplace(0,1)$

			$N=50$	$N=100$	$N=200$	$N=400$	$N=800$
$Bias$	Pairwise	α	0.0022	-0.0044	0.0124	0.0148	0.0113
		β	0.0052	-0.0084	-0.0177	-0.0078	-0.0035
	Inverse-DW	α	-0.0909	-0.0906	-0.0866	-0.0853	-0.0807
		β	-0.3184	-0.3097	-0.3017	-0.2925	-0.2821
Std	Pairwise	α	0.5745	0.3953	0.2810	0.2030	0.1376
		β	0.5342	0.3526	0.2405	0.1784	0.1229
	Inverse-DW	α	0.2258	0.1680	0.1282	0.0969	0.0760
		β	0.0912	0.0667	0.0495	0.0355	0.0270
$RMSE$	Pairwise	α	0.5742	0.3951	0.2811	0.2035	0.1380
		β	0.5339	0.3525	0.2411	0.1784	0.1228
	Inverse-DW	α	0.2434	0.1908	0.1547	0.1291	0.1108
		β	0.3312	0.3168	0.3057	0.2946	0.2834
MAD	Pairwise	α	0.3595	0.2583	0.1773	0.1443	0.0918
		β	0.2723	0.2237	0.1517	0.1163	0.0790
	Inverse-DW	α	0.1580	0.1256	0.1042	0.0942	0.0833
		β	0.3212	0.3135	0.3047	0.2939	0.2825

Table 1(c): $X \tilde{} Logistic(0,1)$, $\epsilon \tilde{} Logistic(0,1)$

			$N=50$	$N=100$	$N=200$	$N=400$	$N=800$
$Bias$	Pairwise	α	-0.0150	-0.0102	0.0027	0.0055	0.0076
		β	-0.0024	-0.0063	-0.0062	0.0044	0.0022
	Inverse-DW	α	-0.0848	-0.0864	-0.0817	-0.0815	-0.0804
		β	-0.2887	-0.2742	-0.2669	-0.2582	-0.2469
Std	Pairwise	α	0.6120	0.4008	0.2722	0.1921	0.1383
		β	0.4989	0.3489	0.2370	0.1641	0.1201
	Inverse-DW	α	0.2312	0.1660	0.1279	0.0942	0.0746
		β	0.1079	0.0776	0.0577	0.0432	0.0332
$RMSE$	Pairwise	α	0.6118	0.4008	0.2721	0.1920	0.1385
		β	0.4987	0.3488	0.2370	0.1640	0.1200
	Inverse-DW	α	0.2462	0.1871	0.1517	0.1246	0.1097
		β	0.3082	0.2849	0.2731	0.2618	0.2491
MAD	Pairwise	α	0.3364	0.2607	0.1737	0.1279	0.0928
		β	0.2726	0.2123	0.1567	0.1052	0.0802
	Inverse-DW	α	0.1638	0.1298	0.1023	0.0920	0.0829
		β	0.2909	0.2763	0.2688	0.2596	0.2482

Table 1(d): Heteroskedastic Design with $X \tilde{} Normal(0,1)$, $\epsilon \tilde{} Normal(0,1)$

			$N=50$	$N=100$	$N=200$	$N=400$	$N=800$
Bias	Pairwise	α	0.0224	0.0186	0.0193	0.0167	0.0090
		β	-0.1169	-0.0867	-0.0748	-0.0646	-0.0562
	Inverse-DW	α	-0.0638	-0.0674	-0.0642	-0.0675	-0.0724
		β	-0.2842	-0.2679	-0.2591	-0.2564	-0.2513
Std	Pairwise	α	0.5415	0.3483	0.2381	0.1638	0.1221
		β	0.5596	0.3707	0.2815	0.2087	0.1561
	Inverse-DW	α	0.2379	0.1758	0.1324	0.0979	0.0722
		β	0.2579	0.1803	0.1395	0.1024	0.0836
RMSE	Pairwise	α	0.5417	0.3486	0.2388	0.1646	0.1223
		β	0.5714	0.3805	0.2911	0.2184	0.1658
	Inverse-DW	α	0.2462	0.1882	0.1471	0.1188	0.1023
		β	0.3836	0.3228	0.2942	0.2761	0.2649
MAD	Pairwise	α	0.3224	0.2258	0.1607	0.1084	0.0835
		β	0.3633	0.2482	0.1981	0.1413	0.1131
	Inverse-DW	α	0.1597	0.1270	0.0990	0.0838	0.0763
		β	0.2889	0.2696	0.2615	0.2568	0.2509

In all three designs, both the pairwise extremum estimator and the inverse-density-weighted estimator are shown to converge to the true parameter values as sample sizes increase. The pairwise estimator converges at approximately the root-n rate regardless of parametrization of error distributions. The inverse-density-weighted estimator appears to converge faster under the normal errors than under logistic and Laplace errors. This conforms with earlier observations in (Khan and Tamer 2010) that the performance of the inverse-density-weighted estimator could be sensitive to the thickness of the tails of error distributions relative to that of the special regressor.

Besides, when sample sizes are as small as $N = 50$, the inverse-density-weighted estimator seems to outperform the pairwise estimator in terms of RMSE under all designs. Nevertheless, it is shown to converge more slowly than the pairwise estimator. The inverse-density-weighted estimator demonstrates smaller variances than the pairwise estimator uniformly across all designs and sample sizes. On the other hand, the pairwise estimator shows lower bias than the inverse-density-weighted estimator in almost all designs and sample sizes. The figures in the appendix show both our estimator (labeled α_1, β_1) and the inverse-density-weighted estimators (labeled α_2, β_2) appear to be approximately normally distributed in the simulated samples.

5 Concluding Remarks

In semiparametric binary regressions with heteroskedastic errors, we study how some special regressors, which are additively separable in the latent payoff and independent

from errors given all other regressors, contribute to the identifying power of the model and the Fisher information for coefficients. We consider two classes of models where identification of coefficients do not depend on "large support" of the special regressors: one with median independent errors; and one with 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 recover error distributions and average structural functions. 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 strictly positive under mild conditions. In other words, the joint restrictions of conditional symmetry (CS) and exclusion restriction (CI) *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 root-n estimators for a binary regressions with heteroskedastic but conditionally symmetric errors, and report its decent performance in finite samples.

Directions of future investigations 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 CI and MI, $p(x,v) \leq 1/2$ 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 CI and MI), we have $\int 1(\epsilon \leq xb - v)\, dG_{\epsilon|x} > 1/2$, which implies $(x,v) \in \xi(b, G_{\epsilon|X})$. Therefore, $\Pr\left\{Z \in \xi(b, G_{\epsilon|X})\right\} > 0$ 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 some $b \neq \beta$. 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 $V - xb$ given x. For t outside the support of $V - xb$ given 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 $V - xb$ given 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 CI and MI such that $\Pr\left\{Z \in \xi(b, \tilde{G}_{\epsilon|X})\right\} = 0$. Furthermore, with any pair of Z and $\tilde{Z}$ that are 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 $\Pr\left\{(Z,\tilde{Z}) \in \tilde{\xi}(b, \tilde{G}_{\epsilon|X})\right\} = 0$. Thus β is not identified relative to b. *Q.E.D.*

Proof of Corollary 1. The objective function in (6) 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 FR. W.L.O.G. suppose $\Pr(Xb > X\beta) > 0$. SV 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 hence $\Pr(p(Z) = 1/2) = 0$), this implies $1\{p(Z) \leq 1/2\}(Xb - V)_+ > 0$ with positive probability. Thus the objective function in (6) is positive for $b \neq \beta$. On the other hand, CI and MI implies $p(Z) \geq 1/2$ if and only if $X\beta - V \geq 0$, and the objective function in (6) is 0 for $b = \beta$. *Q.E.D.*

Proof of Proposition 3. Proposition 1 shows β is identified relative to b under CI and MI whenever (i) holds. It follows immediately that (i) also implies identification of β relative to b under the stronger assumptions of CI and CS. To see how (ii) is also 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 $(z,\tilde{z}) \in \tilde{Q}_{b,S} \subseteq \Omega_Z \otimes \Omega_Z$, 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 CI and CS, this implies for any $G_{\epsilon|X} \in \Theta_{CS}$ and any $(z,\tilde{z}) \in \tilde{Q}_{b,S}$, either

$$\text{"}F_{\epsilon|x}(x\beta - v) + F_{\epsilon|\tilde{x}}(\tilde{x}\beta - \tilde{v}) < 1 \text{ and } G_{\epsilon|x}(xb - v) + G_{\epsilon|\tilde{x}}(\tilde{x}b - \tilde{v}) > 1\text{"} \tag{21}$$

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 $(z,\tilde{z}) \in \tilde{Q}_{b,S}$ satisfies (21) for all $G_{\epsilon|X} \in \Theta_{CS}$. Then EC implies 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 (21) 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 some $b \neq \beta$ and a set ω open in Ω_X. Then CT implies

$$\int 1\{(v_i, v_j) \in R_b(x)\} dF_{V_i|\tilde{x}}(v_j) dF_{V_i|x}(v_i) > 0 \tag{22}$$

for all $(x,\tilde{x})$ with $x \equiv (x_c, x_d) \in \omega$, $\tilde{x}_d = x_d$ 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 (22) then implies $\Pr\{(Z,\tilde{Z}) \in \tilde{Q}^{\tilde{\delta}}_{b,S}\} > 0$ 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 $G_{\epsilon|X} \in \Theta_{CS}$, and β is identified relative to b. The necessity of these two conditions for identifying β relative to b follows from constructive arguments similar to that in the proof of (1), and is hence omitted for brevity. *Q.E.D.*

Proof of Proposition 4. Under CI, CS', EC, CT, SV and FR, β is identified relative to all $b \neq \beta$. With μ consisting of counting measure for $y \in \{0,1\}$ and probability measure for Z, we can show path-wise information for β_k under a path $\lambda \in \Lambda$ (denoted by $I_{\lambda,k}$) takes the form

$$4\int \left(\psi_k - \alpha^*_\lambda \psi_\lambda - \textstyle\sum_{j \neq k} \alpha^*_j \psi_j\right)^2 d\mu = 4\int_{\Omega_Z} \frac{\left[f_{\epsilon|x}(w)\left(x_k - \sum_{j\neq k} \alpha^*_j x_j\right) - \alpha^*_\lambda \lambda_\delta(w,x;\delta_0)\right]^2}{F_{\epsilon|x}(w)\left[1 - F_{\epsilon|x}(w)\right]} dF_Z \tag{23}$$

where $(\alpha^*_j)_{j\neq k}$ and α^*_λ are constants that solve the minimization problem in the definition of $I_{\lambda,k}$ in (5).

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 $\alpha^*_\lambda \neq 0$. SV states the support $\Omega_{V|x}$ includes $x\beta$ in its interior for all x. Thus there exists an open interval $(-\varepsilon^*, \varepsilon^*)$ such that $W \equiv X\beta - V$ is continuously distributed with positive densities over $(-\varepsilon^*, \varepsilon^*)$ given any x. Note the integrand in (23) is non-negative by construction. Thus the right-hand side of (23) 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 (14) with respect to δ at δ_0 suggests $\lambda_\delta(-\varepsilon, x; \delta_0) = -\lambda_\delta(\varepsilon, x; \delta_0)$ for all x and ε. This implies $\alpha^*_\lambda \lambda_\delta(w, x; \delta_0)$ 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 CS', $F_{\epsilon|x}(t)^{-1}\left[1 - F_{\epsilon|x}(t)\right]^{-1}$ is uniformly bounded between positive constants for all $t \in (-\varepsilon^*, \varepsilon^*)$ and $x \in \Omega_X$. It follows that for any constant $\varphi > 0$,

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

Thus for any $\varphi > 0$, there 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

$$\left|f_{\epsilon|x}(t)\left(x_k - \sum_{j\neq k}\alpha^*_j x_j\right) - \alpha^*_\lambda \lambda_\delta(t,x;\delta_0)\right| < \varphi \tag{24}$$

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

The new condition RG' implies $\Pr\{X_k - \sum_{j\neq k}\alpha^*_j X_j > 0 \neq 0 | X \in \bar{\omega}\} > 0$. 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 (24) 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 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 SV, $\Pr\{(W, X) \in \mathcal{I}^-\} > 0$ 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

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

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

Appendix B: CS and Tail Conditions in (Magnac and Maurin 2007)

We now give an example of some $F_{\epsilon|X}$ that satisfies CS but fail to meet tail requirements in (Magnac and Maurin 2007). Suppose the distribution of a continuous random variable W is such that $\lim_{t\to-\infty} tF_W(t) = 0$. Then for any c,

$$\mathbb{E}\left[(W-c)\,1(W<c)\right] = \int_{-\infty}^{c}(s-c)\,dF_W(s) = 0 - 0 - \int_{-\infty}^{c} F_W(s)ds$$

and $\mathbb{E}\left[(W-c)\,1(W>c)\right] = \mathbb{E}(W-c) - \mathbb{E}\left[(W-c)\,1(W<c)\right] = \mu_W - c + \int_{-\infty}^{c} F_W(w)dw$. Let $Y_H \equiv -(X\beta+\epsilon+v_H)$ and $Y_L \equiv X\beta+\epsilon+v_L$. Therefore, for any given x,

$$\mathbb{E}\left[Y_H 1(Y_H>0)|x\right] = \int_{-\infty}^{-v_H} F_{X\beta+\epsilon|X=x}(s)ds \tag{25}$$

$$\mathbb{E}\left[Y_L 1(Y_L>0)|x\right] = x\beta + v_L + \int_{-\infty}^{-v_L} F_{X\beta+\epsilon|X=x}(s)ds \tag{26}$$

so that the difference of (26) minus (25) is given by

$$x\beta + v_L + \int_{-v_H}^{-v_L} F_{X\beta+\epsilon|X=x}(s)ds. \tag{27}$$

Suppose $F_{\epsilon|X}$ satisfies CS, then $F_{X\beta+\epsilon|x}$ is symmetric around $x\beta$ for all x. If $x\beta = \frac{-v_L-v_H}{2}$, then (27) 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 (27) is strictly less than 0. Likewise if $x\beta > \frac{-v_L-v_H}{2}$, then (27) is strictly greater than 0. Now suppose $x\beta < \frac{-v_L-v_H}{2}$ for all x on the support $\Omega_X \subseteq \mathbb{R}^K_{++}$. Then $\mathbb{E}[X'Y_H 1(Y_H>0)] < \mathbb{E}[X'Y_L 1(Y_L>0)]$, and the tail condition in Proposition 5 of (Magnac and Maurin 2007) does not hold.

Appendix C: Asymptotic Properties of $\hat{\beta}$

Our proof follows steps similar to those in (Sherman 1994b), (Khan 2001), (Khan and Tamer 2010) and (Abrevaya, Hausman, and Khan 2010).

C1. Consistency

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

$$H_n(z_i,z_j;b) = \frac{1}{n(n-1)}\sum_{j\neq i} K_h\left(x_i-x_j\right)\left[\kappa\left(w_{i,j}-1\right)\varphi^-(z_i,z_j;b) + \kappa(1-w_{i,j})\varphi^+(z_i,z_j;b)\right]$$

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

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

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 WF imply that:

$$\begin{aligned} & \mathbb{E}\left[\kappa\left(W_{i,j}-1\right)\varphi^{-}(Z_i,Z_j;b)+\kappa(1-W_{i,j})\varphi^{+}(Z_i,Z_j;b)\middle|\, X_j = X_i = x\right] \\ \geq\ & \mathbb{E}\left[\kappa(W_{i,j}-1)\varphi^{-}(Z_i,Z_j;b)\middle|\, V_i+V_j \leq 2x\beta, X_j = X_i = x\right]\Pr(V_i+V_j \leq 2x\beta | X_j = X_i = x) > 0 \end{aligned} \quad (28)$$

By construction, the conditional expectation on the left-hand side can never be negative for any x. Multiply both sides of (28) 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 $x\beta < xb$, SV implies

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

Then $H_0(b) > 0$ for all $b \neq \beta$ by the same argument as above.

Next, consider $b = \beta$. For any x,

$$\begin{aligned} & H_0(\beta) = \mathbb{E}\left\{f(X)\mathbb{E}\left[\kappa\left(W_{i,j}-1\right)\varphi^{-}(Z_i,Z_j;\beta)+\kappa(1-W_{i,j})\varphi^{+}(Z_i,Z_j;\beta)\middle|\, X_j = X_i = X\right]\right\} \\ =\ & \mathbb{E}\left\{f(X)\mathbb{E}\left[\kappa(W_{i,j}-1)\,\varphi^{-}(Z_i,Z_j;\beta)\middle|\, W_{i,j} \geq 1, X_j = X_i = X\right]\Pr(W_{i,j} \geq 0 | X_j = X_i = X)\right\} \quad (29) \\ +\ & \mathbb{E}\left\{f(X)\mathbb{E}\left[\kappa(1-W_{i,j})\,\varphi^{+}(Z_i,Z_j;\beta)\middle|\, W_{i,j} < 1, X_i = X_i = X\right]\Pr(W_{i,j} < 0 | X_j = X_i = X)\right\}. \end{aligned}$$

The first conditional expectation on the right-hand side of (29) is 0, because whenever $x_i = x_j$, we have $w_{i,j} \geq 1$ if and only if $v_i + v_j \leq 2x_i\beta$. Likewise the second conditional expectation is also 0. Thus $H_0(\beta) = 0$. *Q.E.D*

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

$$\sup_{b\in\mathcal{B}} |\hat{H}_n(b) - H_n(b)| = o_p(1). \quad (30)$$

Let $\varphi^{-}_{i,j}(b)$ be a shorthand for $\varphi^{-}(z_i, z_j; b)$ and likewise for $\varphi^{+}_{i,j}(b)$. Applying the Taylor's expansion around $w_{i,j}$ and using the boundedness conditions in FM1 and KF2, we have:

$$\begin{aligned} & \sup_{b\in\mathcal{B}}\left|\tfrac{1}{n(n-1)}\textstyle\sum_{j\neq i} K_h\left(x_i - x_j\right)\varphi^{-}_{i,j}(b)\left[\kappa(\hat{w}_{i,j}-1) - \kappa(w_{i,j}-1) - \kappa'(w_{i,j}-1)(\hat{w}_{i,j} - w_{i,j})\right]\right| \\ =\ & \sup_{b\in\mathcal{B}}\left|\tfrac{1}{n(n-1)}\textstyle\sum_{j\neq i} K_h\left(x_i - x_j\right)\varphi^{-}_{i,j}(b)\kappa''_1(\tilde{w}_{i,j}-1)\left\|\hat{w}_{i,j} - w_{i,j}\right\|^2\right| \\ \leq\ & a\sup_z\left\|\hat{p}(z) - p(z)\right\|^2 \sup_{b\in\mathcal{B}}\left\{\tfrac{1}{n(n-1)}\textstyle\sum_{j\neq i}\left|K_h\left(x_i - x_j\right)\varphi^{-}_{i,j}(b)\kappa''_1(\tilde{w}_{i,j})\right|\right\} \end{aligned} \quad (31)$$

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 KF2-(iii), FM1-(i) and WF, the second term on the right-hand side (i.e. the supreme of the term in the braces) is $O_p(1)$. Under SM1 and KF1, $\sup_z |\hat{p}(z) - p(z)| = O_p\left((\log n)/\sqrt{n\sigma_n^{k+1}} + (k+1)\sigma_n^{m\kappa}\right)$ almost surely by Theorem 2.6 of (Li and Racine 2007). Our choice of bandwidth in BW1 implies this term is $o_p(n^{-1/4})$. Hence the remainder term of the approximation (l.h.s. of (31))is $o_p(1)$. Next, note:

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

By similar arguments, the second term is bounded in probability, and the first term is $o_p(n^{-1/4})$. Thus (30) holds.

Next, decompose $H_n(z_i, z_j; b)$ as

$$H_n(z_i,z_j;b) = \mathbb{E}[g_n(Z_i,Z_j;b)] + \tfrac{2}{n}\textstyle\sum_{i\leq n} g_{n,1}(z_i;b) + \tfrac{2}{n(n-1)}\sum_{j\neq i} g_{n,2}(z_i,z_j;b) \tag{32}$$

where

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

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 the second and third term in (32) are $o_p(1)$ under our conditions. Under KF2 and PS, we get

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

for all (z_i, z_j), where $\mathcal{C}(.)$ and $\mathcal{D}(.)$ are defined in FM1 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 KF2 and FM1. Besides, $\mathbb{E}[\sup_{b\in\mathcal{B}} h_n^{2k} g_n\left(Z_i,Z_j;b\right)^2] = O(1)$ under KF2 and FM1. It then follows from Theorem 3 in (Sherman 1994b) that the second and the third terms in the decomposition in (32) are $O_p(n^{-1/2}h_n^{-k})$ and $O_p(n^{-1}h_n^{-k})$ uniformly over $b \in \mathcal{B}$ respectively. Under our choice of bandwidth in BW2, these two terms are both $o_p(1)$.

Next, we deal with the first term in the H-decomposition above. Let $\kappa^-(z_i,z_j) \equiv \kappa(w_{i,j}-1)$ and $\kappa^+(z_i,z_j) \equiv \kappa(1-w_{i,j})$ and

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

to facilitate derivations. By definition,

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

Changing variables between x_j and $u \equiv (x_j - x_i)/h_n^k$ and applying the dominated convergence theorem, we can show that $\mathbb{E}[g_n(Z_i,Z_j;b)] = H_0(b) + O(kh_n^2) = H_0(b) + o(1)$ for all $b \in \mathcal{B}$. Thus the sum of the three terms on the right-hand side of (32) is $o_p(1)$ uniformly over $b \in \mathcal{B}$.

Combine this result with (30), we get:

$$\sup_{b\in\mathcal{B}}|\hat{H}_n(b) - H_0(b)| = o_p(1). \tag{33}$$

The limiting function $H_0(b)$ is continuous under SM1 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 $\hat{\beta} \xrightarrow{p} \beta$. *Q.E.D.*

C2. Root-N and Asymptotic Normality

For convenience of proof in this section, define:

$$\widehat{\mathcal{H}}_n(b) = \hat{H}_n(b) - \hat{H}_n(\beta) \text{ and } \mathcal{H}_n(b) = H_n(b) - H_n(\beta).$$

By construction, the optimizers of $\widehat{\mathcal{H}}_n$ and $\mathcal{H}_n$ are the same as those for $\hat{H}_n$ and H_n.

Having shown consistency, our strategy for deriving the limiting distribution of $\hat{\beta}$ is to approximate $\widehat{\mathcal{H}}_n(.)$ locally in a neighborhood of β by some function that is quadratic in b. The approximation needs to accommodate the fact that the objective function is not smooth in b. Quadratic approximation of such objective functions have been provided in, for example, (Pakes and Pollard 1989), and (Sherman 1994a), (Sherman 1994b) among others. A preliminary step is to show $||\hat{\beta} - \beta||$ converges at a rate no slower than $\sqrt{n}$. Once established, this result allows us to focus on such a shrinking neighborhood around β where quadratic approximation mentioned above becomes more precise so that root-n consistency and asymptotic normality can be established in one step. A useful theorem that will be invoked for showing these results is Theorem 1 in (Sherman 1994b), which require the following conditions:

1. $\hat{\beta} - \beta = O_p(\delta_n)$;

2. There exists a neighborhood of β and a constant $\tilde{a} > 0$ such that $H_0(b) - H_0(\beta) \geq \tilde{a}\|b - \beta\|^2$ for all b in this neighborhood of β; and

3. Uniformly over an $O_p(\delta_n)$ neighborhood of β:

$$\widehat{\mathcal{H}}_n(b) = H_0(b) + O_p(\|b - \beta\|/\sqrt{n}) + o_p(\|b - \beta\|^2) + O_p(\epsilon_n). \tag{34}$$

Under these three conditions, Theorem 1 in (Sherman 1994b) states $\hat{\beta} - \beta_0 = O_p(\max\{\sqrt{\epsilon_n}, 1/\sqrt{n}\})$.

Lemma C1. *Under SM2-(i), there exists an open neighborhood of β and some constant $\tilde{a} > 0$ such that $H_0(b) - H_0(\beta) \geq \tilde{a}\|b - \beta\|^2$ for all b in this neighborhood of β.*

Proof of Lemma C1. Under SM-(i), we can apply the Taylor's expansion to write:

$$H_0(b) = \tfrac{1}{2}(b - \beta)' \bigtriangledown_{bb} H_0(\tilde{b})(b - \beta)$$

where $\tilde{b}$ is on the line segment linking b and β. Note we have used $H_0(\beta) = 0$ and $\bigtriangledown_b H_0(\beta) = 0$ due to the identification result in Proposition 5. The claim in the this lemma then follows from the positive definiteness of $\bigtriangledown_{bb} H_0(\beta)$ and its continuity at β. *Q.E.D.*

To simplify notations in what follows, we let

$$\widehat{\mathcal{H}}_{1,n}(b) \equiv \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} K_h(x_i - x_j)\, \kappa(\hat{w}_{i,j} - 1) \left[\varphi^-_{i,j}(b) - \varphi^-_{i,j}(\beta)\right]$$

denote the first half of the "location-normalized" objective function $\widehat{\mathcal{H}}_n$ (which only involve $(.)_-$); and likewise let $\mathcal{H}_{1,n}(b)$ and $H_{1,0}(b)$ denote the first halves of $\mathcal{H}_n$ and H_0 respectively. Similarly, define $\widehat{\mathcal{H}}_{2,n}$, $\mathcal{H}_{2,n}$ and $H_{2,0}$ as the second halves involving $(.)_+$. Recall that $H_{1,0}(\beta) = H_{2,0}(\beta) = 0$ by construction.

Lemma C2. *Suppose conditions for Proposition 6 hold. Under additional conditions SM2 and FM2,*

$$\mathcal{H}_n(b) - H_0(b) = o_p\left(\|b - \beta\|^2\right) + o_p\left(\|b - \beta\|/\sqrt{n}\right) + O_p(n^{-1}h^{-k})$$

uniformly over an $o_p(1)$ neighborhood around β in $\mathcal{B}$; and the $O_p(n^{-1}h^{-k})$ term is further reduced to $o_p(n^{-1})$ uniformly over an $O_p(1/\sqrt{nh^k})$ neighborhood of β.

Proof of Lemma C2. We analyze the order of magnitude of $\mathcal{H}_{1,n} - H_{1,0}$ in this proof. The case with $\mathcal{H}_{2,n} - H_{2,0}$ follows from the same arguments and is omitted for brevity. For any b, decomposed $\mathcal{H}_{1,n}(b) - H_{1,0}(b)$ as:

$$\{\mathbb{E}[\tilde{g}_n(Z_i, Z_j; b)] - H_{1,0}(b)\} + \frac{1}{n}\textstyle\sum_{i=1}^n \tilde{g}_{n,1}(z_i; b) + \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} \tilde{g}_{n,2}(z_i, z_j; b) \tag{35}$$

where

$$\begin{aligned}\tilde{g}_n(z_i,z_j;b) &\equiv K_h\left(x_i-x_j\right)\kappa^-(z_i,z_j)\left[\varphi^-(z_i,z_j;b)-\varphi^-(z_i,z_j;\beta)\right];\\ \tilde{g}_{n,1}(z_i;b) &\equiv \mathbb{E}[\tilde{g}_n(Z,Z';b)|Z=z_i]+\mathbb{E}[\tilde{g}_n(Z,Z';b)|Z'=z_i]-2\mathbb{E}[\tilde{g}_n(Z,Z';b)];\text{ and}\\ \tilde{g}_{n,2}(z_i,z_j;b) &\equiv \tilde{g}_n(z_i,z_j;b)-\mathbb{E}[\tilde{g}_n(Z,Z';b)|Z=z_i]-\mathbb{E}[\tilde{g}_n(Z,Z';b)|Z'=z_j]+\mathbb{E}[\tilde{g}_n(Z,Z';b)].\end{aligned}$$

where $\kappa^-(z_i,z_j)$ is a shorthand for $\kappa(w_{i,j}-1)$.

We first deal with the first term in (35). With a slight abuse of notation, let F denote distributions and f denote densities. Let $\Delta\varphi^-(z_i,z_j;b)\equiv\varphi^-(z_i,z_j;b)-\varphi^-(z_i,z_j;\beta)$. Note by the Law of Iterated Expectation, we can write for all b:

$$\mathbb{E}[\tilde{g}_n(Z_i,Z_j;b)]=\int K_h\left(x_i-x_j\right)\bar{\varphi}(x_i,x_j;b)dF(x_i,x_j)$$

where

$$\bar{\varphi}(x,x';b)\equiv\mathbb{E}\left\{\kappa^-(Z_i,Z_j)\Delta\varphi^-(Z_i,Z_j;b)|X_i=x,X_j=x'\right\}.$$

By construction, $\bar{\varphi}(x_i,x_j;\beta)=0$ and under SM2-(ii),

$$\bar{\varphi}(x_i,x_j;b)=\nabla_b\bar{\varphi}(x_i,x_j;\beta)(b-\beta)+\tfrac{1}{2}(b-\beta)'\nabla_{bb}\bar{\varphi}(x_i,x_j;\beta)(b-\beta)+o(\|b-\beta\|^2)$$

for all b in an $o(1)$ neighborhood around β; where $\nabla_b\bar{\varphi}$ and $\nabla_{bb}\bar{\varphi}$ are gradient and Hessian w.r.t. b respectively. Since the magnitude of the remainder is invariant in x_i,x_j, we can decompose $\mathbb{E}[\tilde{g}_n(Z_i,Z_j;b)]$ as

$$\begin{aligned}&(b-\beta)'\left[\int K_h\left(x_i-x_j\right)\tfrac{1}{2}\nabla_{bb}\bar{\varphi}(x_i,x_j;\beta)dF(x_i,x_j)\right](b-\beta)\\ &+\left\{\int K_h\left(x_i-x_j\right)\nabla_b\bar{\varphi}(x_i,x_j;\beta)dF(x_i,x_j)\right\}(b-\beta)+o(||b-\beta||^2).\end{aligned}\tag{36}$$

for all b in an $o(1)$ neighborhood around β. Changing variables between x_i and $u\equiv(x_i-x_j)/h_n^k$, we can write the square bracket term in (36) as

$$\begin{aligned}&\int\left[\int K(u)\tfrac{1}{2}\nabla_{bb}\bar{\varphi}(x_j+h^ku,x_j;\beta)f(x_j+h_n^ku)du\right]dF(x_j)\\ =\;&\tfrac{1}{2}\int\left[\nabla_{bb}\bar{\varphi}(x,x;\beta)f(x)+O(h_n^2)\right]dF(x)=\tfrac{1}{2}\int\left[\nabla_{bb}\bar{\varphi}(x,x;\beta)f(x)\right]dF(x)+o(1)=\tfrac{1}{2}\nabla_{bb}H_{1,0}(\beta)+o(1).\end{aligned}$$

The first equality above follows from a Taylor expansion of $\nabla_{bb}\bar{\varphi}(x,x_j;\beta)f(x)$ around $x=x_j$ under SM2; the order K in KF2; and the fact that $\int K(u)du=1$. The second equality is due to the facts that the expansion applies for all x_j; that the order of remainder is invariant in x_j; and that $O(h_n^2)$ is $o(1)$ under BW2. The third equality follows from the fact that the order of differentiation and integration can be exchanged under SM2-(ii) and FM2-(i). Similarly we can show the term in the braces in (36) is

$$\begin{aligned}&\int\left[\int K(u)\nabla_b\bar{\varphi}(x_j+h^ku,x_j;\beta)f(x_j+h_n^ku)du\right]dF(x_j)\\ =\;&\int\left[\nabla_b\bar{\varphi}(x,x;\beta)f(x)+O(h_n^{m_\varphi})\right]dF(x)=\int\left[\nabla_b\bar{\varphi}(x,x;\beta)f(x)\right]dF(x)+o(n^{-1/2})\end{aligned}$$

where $\int [\nabla_b \bar{\varphi}(x,x;\beta) f(x)]\, dF(x) = \nabla_b H_{1,0}(\beta) = 0$ because the order of integration and differentiation can be exchanged and $H_{1,0}(b)$ is uniquely minimized at $b = \beta$. To sum up, (36) is

$$\tfrac{1}{2}(b-\beta)'\nabla_{bb} H_{1,0}(\beta)(b-\beta) + o_p(\|b-\beta\|/\sqrt{n}) + o_p(\|b-\beta\|^2).$$

By standard Taylor expansion using SM-(i), $H_{1,0}(b) = \frac{1}{2}(b-\beta)'\nabla_{bb} H_{1,0}(\beta)(b-\beta) + o_p(\|b-\beta\|^2)$ over an $o(1)$ neighborhood of β, it then follows that the first term in (35) is $o_p(\|b-\beta\|/\sqrt{n}) + o_p(\|b-\beta\|^2)$.

Next, we turn to the second term in (35). By SM2-(ii), we can apply the Taylor expansion around β to the second term in (35) to get

$$\frac{1}{n}\textstyle\sum_{i=1}^n \tilde{g}_{n,1}(z_i;b) = \frac{1}{n}\sum_{i=1}^n \tilde{g}_{n,1}(z_i;\beta) + \left(\frac{1}{n}\sum_{i=1}^n \nabla_b \tilde{g}_{n,1}(z_i;\tilde{b})\right)(b-\beta)$$

where $\tilde{b}$ is on the line segment between b and β (and possibly depends on z_i). By construction, $\tilde{g}_{n,1}(z_i;\beta) = 0$ for all n and z_i. Besides, for any given n and b, $\nabla_b \tilde{g}_{n,1}(z_i;b)$ has mean zero for all z_i. To see this, note for any fixed n and b:

$$\mathbb{E}[\nabla_b \tilde{g}_{n,1}(Z_i;b)] = \mathbb{E}\left\{\nabla_b \mathbb{E}[\tilde{g}_n(Z,Z';b)|Z=Z_i]\right\} + \mathbb{E}\left\{\nabla_b \mathbb{E}[\tilde{g}_n(Z,Z';b)|Z'=Z_i]\right\} - 2\nabla_b \mathbb{E}[\tilde{g}_n(Z,Z';b)] = 0$$

because under FM1,2 and SM1,2 the order of integration and differentiation in the first two terms on the right-hand side can be exchanged. Also note that by definition,

$$\nabla_b \mathbb{E}[\tilde{g}_n(Z,Z';b)|Z=z] = \int K_h(x-x')\nabla_b \hat{\varphi}(z,x';b) f(x')dx' = \int K(u)\nabla_b \hat{\varphi}(z,x-h^k u;b) f(x-h^k u)du$$

where $\hat{\varphi}^-(z,x';b) \equiv \mathbb{E}[\kappa^-(Z_i,Z_j)\Delta\varphi^-_{i,j}(b)|Z_i = z, X_j = x']$. The first equality follows from an interchange of integration and differentiation; and the second equality follows from a change of variables between x' and $u \equiv (x-x')/h^k$. Likewise we can derive a similar expression for $\nabla_b \mathbb{E}[\tilde{g}_n(Z,Z';b)|Z'=z]$. It then follows from boundedness of K in KF and the finite moment condition in FM2 that $\mathbb{E}[\nabla_b \tilde{g}_{n,1}(Z;b)\,\nabla_b \tilde{g}_{n,1}(Z_i;b)'] = O(1)$ for all b in an open neighborhood around β. Thus for any fixed b in an open neighborhood around β, the Liapunov Central Limit Theorem applies and $\frac{1}{n}\sum_{i=1}^n \nabla_b \tilde{g}_{n,1}(z_i;b) = O_p(n^{-1/2})$ under FM1,2. With $\tilde{b}$ between b and β, and with $b \xrightarrow{p} \beta$, an application of Lemma 2.17 in (Pakes and Pollard 1989) shows $\frac{1}{n}\sum_{i=1}^n \nabla_b \tilde{g}_{n,1}(z_i;\tilde{b})$ is $o_p(n^{-1/2})$ uniformly over an $o_p(1)$ neighborhood around β. Thus the second term in the decomposition in (35) is $o_p(\|b-\beta\|/\sqrt{n})$ uniformly over an $o_p(1)$ neighborhood of β.

Next, arguments similar to Proposition 6 suggest conditions for Theorem 3 in (Sherman 1994b) hold for the third term in the decomposition in (35) multiplied with h^k, which is a second-order degenerate U-process. Hence the third term is $O_p(n^{-1}h^{-k})$ uniformly over $o_p(1)$ neighborhood around β. Furthermore, this term is reduced to $O_p(n^{-3/2}h^{-3k/2})$ uniformly over an $O_p(1/\sqrt{nh^k})$ neighborhood around β, which is $o_p(n^{-1})$ due to our choice of bandwidth in BW2. *Q.E.D.*

Next, we show the difference between $\widehat{\mathcal{H}}_{1,n}(b)$ and $\mathcal{H}_{1,n}(b)$ can be expressed in terms of a simple sample average plus some negligible approximation errors over a shrinking neighborhood of β in $\mathcal{B}$.

Lemma C3. Suppose conditions for Proposition 6 hold. Under additional conditions in SM2 and FM2,

$$|\widehat{\mathcal{H}}_{1,n}(b) - \mathcal{H}_{1,n}(b)| = \frac{2}{n}\sum_{i=1}^n \delta_1^*(y_i, z_i)(b-\beta) + o_p(\|b-\beta\|/\sqrt{n}) + o_p\left(\|b-\beta\|^2\right) + O_p(n^{-1}h^{-k}) \quad (37)$$

uniformly over an $o_p(1)$ neighborhood of β in $\mathcal{B}$; where

$$\delta_1^*(y,z) \equiv q\nabla_b m_-^*(z;\beta)f(z) - \mathbb{E}[Q\nabla_b m_-^*(Z;\beta)f(Z)] \text{ with } q \equiv (y,1)';$$

Besides, the $O_p(n^{-1}h^{-k})$ term in (37) is further reduced to $o_p\left(n^{-1}\right)$ uniformly over an $O_p\left(1/\sqrt{nh^k}\right)$ neighborhood around β.

Proof of Lemma C3. Let $\Delta\varphi_{i,j}^-(b) \equiv \Delta\varphi^-(z_i, z_j; b) \equiv \varphi_{i,j}^-(b) - \varphi_{i,j}^-(\beta)$. By smoothness of κ in WF, we can use the Taylor's expansion to decompose $\widehat{\mathcal{H}}_{1,n}(b) - \mathcal{H}_{1,n}(b)$ into:

$$\Delta_{1,n} \equiv \tfrac{1}{n(n-1)}\sum_{j\neq i} K_h\left(x_i - x_j\right)\Delta\varphi_{i,j}^-(b)\kappa'(w_{i,j}-1)(\hat{w}_{i,j} - w_{i,j})$$

and

$$R_{1,n} \equiv \tfrac{1}{n(n-1)}\sum_{j\neq i} K_h\left(x_i - x_j\right)\Delta\varphi_{i,j}^-(b)\kappa''(\tilde{w}_{i,j}-1)(\hat{w}_{i,j} - w_{i,j})^2$$

where $\tilde{w}_{i,j}$ is between $\hat{w}_{i,j}$ and $w_{i,j}$. Using the triangular inequality and by the fact that the second-order derivative κ'' is bounded, we have:

$$|R_{1,n}| \leq \hat{a}\left\{\tfrac{1}{n(n-1)}\sum_{j\neq i}\left|K_h\left(x_i - x_j\right)\Delta\varphi_{i,j}^-(b)\right|\right\}\sup_{z,z'}\|\hat{p}(z) + \hat{p}(z') - p(z) - p(z')\|^2) \quad (38)$$

for some finite constant $\hat{a} > 0$. The second term on the right-hand side of (38) is $o_p(n^{-1/2})$ since under our conditions of SM1, KF1 and BW1,

$$\sup_z |\hat{p}(z) - p(z)| = o_p(n^{-1/4}). \quad (39)$$

As for the first term in the braces of (38), we use the H-decomposition to break it down into the sum of an unconditional expectation and two degenerate U-processes:

$$\mathbb{E}[\varpi_n(Z_i, Z_j; b)] + \frac{1}{n}\sum_{i=1}^n \varpi_{n,1}(z_i; b) + \tfrac{1}{n(n-1)}\sum_{j\neq i}\varpi_{n,2}(z_i, z_j; b) \quad (40)$$

where $\varpi_n(z_i, z_j; b) \equiv \left|K_h\left(x_i - x_j\right)\Delta\varphi_{i,j}^-(b)\right|$; and

$$\begin{aligned}
\varpi_{n,1}(z_i;b) &\equiv \mathbb{E}[\varpi_n(Z,Z';b)|Z=z_i] + \mathbb{E}[\varpi_n(Z,Z';b)|Z'=z_i] - 2\mathbb{E}[\varpi_n(Z,Z';b)]; \text{ and} \\
\varpi_{n,2}(z_i,z_j;b) &\equiv \varpi_n(z_i,z_j;b) - \mathbb{E}[\varpi_n(Z,Z';b)|Z=z_i] - \mathbb{E}[\varpi_n(Z,Z';b)|Z'=z_j] + \mathbb{E}[\varpi_n(Z,Z';b)].
\end{aligned}$$

By standard arguments in (Pakes and Pollard 1989), the class of functions $\{h^k\varpi_n(z_i, z_j; b) : b \in \mathcal{B}\}$ is Euclidean with a constant envelop that has finite second moments. Our conditions in FM1, the boundedness of K over its compact support in KF2, and boundedness of derivatives in WF all imply that both conditions for Theorem 3 of (Sherman 1994b) hold with δ_n and γ_n therein being $o(1)$ and $O(1)$ respectively. Hence

$$\frac{1}{n}\sum_{i=1}^n \varpi_{n,1}(z_i;b) = O_p\left(n^{-1/2}h^{-k}\right); \text{ and } \tfrac{1}{n(n-1)}\sum_{j\neq i}\varpi_{n,2}(z_i,z_j;b) = O_p(n^{-1}h^{-k}) \quad (41)$$

uniformly over an $o_p(1)$ neighborhood around β.

As for the unconditional expectation in (40), by definition, it equals

$$\mathbb{E}[\varpi_n(Z_i, Z_j; b)] = \int |K_h(x_i - x_j)| \varpi(x_i, x_j; b) dF(x_i, x_j) \tag{42}$$

where $\varpi(x, x'; b) \equiv \mathbb{E}\left\{|\Delta\varphi^-_{i,j}(b)| \; | X_i = x, X_j = x'\right\}$. By construction, $\varpi(x_i, x_j; \beta) = 0$ and under SM2,

$$\varpi(x_i, x_j; b) = \nabla_b \varpi(x_i, x_j; \beta)(b - \beta) + o(\|b - \beta\|) \tag{43}$$

for all b in an $o(1)$ neighborhood around β; where $\nabla_b \varpi$ is a gradient w.r.t. b. Change variables between x_i and $u \equiv (x_i - x_j)/h^k$ given any x_j on the right-hand side of (42), and we get

$$\begin{aligned} & \int |K_h(x_i - x_j)| \nabla_b \varpi(x_i, x_j; \beta) dF(x_i, x_j) \\ = & \int \left[\int |K(u)| \nabla_b \varpi(x_j + h^k u, x_j; \beta) f(x_j + h_n^k u) du \right] dF(x_j) \\ = & \tilde{\kappa}_1 \nabla_b \mathbb{E}[f(X) \varpi(X, X; \beta)] + o(1) \end{aligned} \tag{44}$$

where $\tilde{\kappa}_1 \equiv \int |K(u)| du$ is finite under KF2. The second equality follows from an application of a first-order Taylor expansion of $\nabla_b \varpi(x_i, x_j; \beta) f(x_i)$ around $x_i = x_j$; and from changing the order of integration and differentiation allowed under SM2 and FM2. Note that

$$\nabla_b \mathbb{E}[f(X) \varpi(X, X; \beta)] = 0 \tag{45}$$

because $\mathbb{E}[f(X) \varpi(X, X; \beta)]$ is minimized to 0 at $b = \beta$. Hence combining results from (42), (43), (44) and (45), we have:

$$\mathbb{E}[\varpi_n(Z_i, Z_j; b)] = o(\|b - \beta\|). \tag{46}$$

Combining results from (39), (41) and (46), we know the order of $|R_{1,n}|$ is bounded above by

$$o_p(\|b - \beta\| / \sqrt{n}) + o_p(n^{-1} h^{-k}) + o_p(n^{-3/2} h^{-k}) \tag{47}$$

uniformly over an $o_p(1)$ neighborhood of β. The third term in (47) is $o_p(n^{-1})$ due to choice of bandwidth in BW2. The second term is due to the product of $\sup_z |\hat{p}(z) - p(z)|$ and a degenerate empirical process $\frac{1}{n} \sum_{i=1}^n \varpi_{n,1}(z_i; b)$. Next, let $\delta_n = O(n^{-1/2} h^{-k/2})$. Following the same arguments in (Khan 2001), another application of Theorem 3 in (Sherman 1994b) implies that over an $O_p(\delta_n)$ neighborhood of β, the magnitude of this product would be $h^{-k} O_p(\delta_n n^{-1/2}) o_p(n^{-1/2}) = o_p(h^{-3k/2} n^{-3/2})$, which is $o_p(n^{-1})$ given our choice of bandwidth in BW2.

We now deal with $\Delta_{1,n}$. We first derive the correction term due to estimation errors in $\hat{p}(z_i)$. Let $\gamma_0 \equiv \left(\gamma_{0,1}, \gamma_{0,2}\right)'$ denote $\mathbb{E}[Y_i | z_i] f(z_i)$ and density $f(z_i)$ in the population and let $\hat{\gamma} \equiv (\hat{\gamma}_1, \hat{\gamma}_2)'$ denote their kernel estimates respectively so that $\hat{\gamma}_1 / \hat{\gamma}_2 = \hat{p}$. With

a slight abuse of notation, let $\kappa'(z_i, z_j)$ be a shorthand for $\kappa'(w_{i,j} - 1)$, and write the first half of $\Delta_{1,n}$ as:

$$\begin{aligned} & \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} \kappa'(z_i, z_j) K_h\left(x_i - x_j\right) \Delta\varphi_{i,j}^{-}(b)[\hat{p}(z_i) - p(z_i)] \\ = \; & \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} \kappa'(z_i, z_j) K_h\left(x_i - x_j\right) \Delta\varphi_{i,j}^{-}(b)[\hat{\gamma}_1(z_i)/\hat{\gamma}_2(z_i) - \gamma_{0,1}(z_i)/\gamma_{0,2}(z_i)] \\ = \; & \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} \kappa'(z_i, z_j) K_h\left(x_i - x_j\right) \Delta\varphi_{i,j}^{-}(b) \nabla w(z_i)[\hat{\gamma}(z_i) - \gamma_0(z_i)]' + \tilde{R}_{1,n} \end{aligned} \tag{48}$$

where $\nabla w(z_i) \equiv [1/\gamma_{0,2}(z_i),\ -\gamma_{0,1}(z_i)/\gamma_{0,2}^2(z_i)]$; and $\tilde{R}_{1,n}$ is of order $o_p(\|b - \beta\|/\sqrt{n}) + O_p(n^{-1}h^{-k}) + o_p(n^{-1})$ uniformly over an $o_p(1)$ neighborhood around β due to Taylor-expansion-based arguments similar to those applied to $R_{1,n}$. Also similar to the case with $R_{1,n}$, the second term in $\tilde{R}_{1,n}$, which is of the order $O_p(n^{-1}h^{-k})$ uniformly over $o_p(1)$ neighborhood of β, is further reduced to $o_p(n^{-1})$ over an $O_p\left(n^{-1/2}h^{-k/2}\right)$ neighborhood around β, due to a repeated application of Theorem 3 in (Sherman 1994b) and our choice of bandwidth in BW2.

Next, let $q \equiv (y, 1)'$. Write the first term on the last line in (48) as

$$\begin{aligned} & \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} \kappa'(z_i, z_j) K_h\left(x_i - x_j\right) \Delta\varphi_{i,j}^{-}(b) \nabla w(z_i)[\hat{\gamma}(z_i) - \mathbb{E}\left(\hat{\gamma}(z_i)\right)] + \hat{R}_{1,n}; \\ & \text{where } \hat{R}_{1,n} \equiv \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} \kappa'(z_i, z_j) K_h\left(x_i - x_j\right) \Delta\varphi_{i,j}^{-}(b)[\mathbb{E}\left(\hat{\gamma}(z_i)\right) - \gamma_0(z_i)]. \end{aligned} \tag{49}$$

By triangular inequality, we have:

$$\hat{R}_{1,n} \leq \left\{ \tfrac{1}{n(n-1)} \textstyle\sum_{j \neq i} \left|\kappa'(z_i, z_j) K_h\left(x_i - x_j\right) \Delta\varphi_{i,j}^{-}(b) \nabla w(z_i)\right| \right\} \sup_z [\mathbb{E}\left(\hat{\gamma}(z_i)\right) - \gamma_0(z_i)]. \tag{50}$$

By arguments similar to those apply to the first term in $R_{1,n}$, the first term in the product on the right-hand side of (50) is

$$o_p(\|b - \beta\|) + o_p(n^{-1/2}h^{-k}) + o_p(n^{-1}h^{-k})$$

Furthermore, the second term on the right-hand side of (50) is $O(\sigma_n^{m_\kappa})$ due to Lemma 8.9 in (Newey and McFadden 1994), which is $o_p(n^{-1/2})$ by our choice of σ_n in BW1 and smoothness condition in SM1. Thus the order of $\hat{R}_{1,n}$ is no greater than $o_p(\|b - \beta\|/\sqrt{n}) + o_p(n^{-1}h^{-k}) + o_p(n^{-3/2}h^{-k})$ uniformly over an $o_p(1)$ neighborhood around β. Again, similar to the case with $R_{1,n}$, the $o_p(n^{-1}h^{-k})$ term in $\hat{R}_{1,n}$ above is further reduced to $o_p(n^{-1})$ over an $O_p\left(n^{-1/2}h^{-k/2}\right)$ neighborhood around β by a repeated application of Theorem 3 in (Sherman 1994b).

We now write the first term in (49) as a third-order U-statistic:

$$\tfrac{1}{n(n-1)(n-2)} \textstyle\sum_{i \neq j \neq s} \phi_n(\tilde{z}_i, \tilde{z}_j, \tilde{z}_s; b) \tag{51}$$

where $\tilde{z} \equiv (y, z) \equiv (y, x, v)$ and

$$\phi_n(\tilde{z}_i, \tilde{z}_j, \tilde{z}_s; b) \equiv \kappa'(z_i, z_j) K_h\left(x_i - x_j\right) \Delta\varphi_{i,j}^{-}(b) \nabla w(z_i) \left\{ q_s \mathcal{K}_\sigma\left(z_i - z_s\right) - \mathbb{E}\left[Q_s \mathcal{K}_\sigma\left(z_i - Z_s\right)\right] \right\}$$

with $\mathcal{K}_\sigma$ being a shorthand for $\sigma^{-(k+1)}\mathcal{K}\left(./\sigma^{k+1}\right)$; and the expectation is taken w.r.t. $\tilde{Z}_s$ while z_i is some realized value of Z_i. Note ϕ_n is not symmetric in the three arguments,

for it depends on y_s but not y_i and y_j. Let $\tilde{Z}_{i,j,s}$ be a shorthand for $(\tilde{Z}_i, \tilde{Z}_j, \tilde{Z}_s)$. Then apply the H-decomposition to write this third-order U-statistic in (51) as

$$\mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)] + \tfrac{1}{n}\textstyle\sum_{i=1}^n \phi_n^{(1)}(\tilde{z}_i; b) + U^2\phi_n^{(2)}(b) + U^3\phi_n^{(3)}(b) \tag{52}$$

where

$$\phi_n^{(1)}(\tilde{z}_i) \equiv \mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)|\tilde{Z}_i = \tilde{z}_i] + \mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)|\tilde{Z}_j = \tilde{z}_i] + \mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)|\tilde{Z}_s = \tilde{z}_i] - 3\mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)] \tag{53}$$

and $U^2\phi_n^{(2)}(b)$ and $U^3\phi_n^{(3)}(b)$ are second- and third-order degenerate U-statistics as defined in (Sherman 1994b).

To deal with the second- and third-order processes $U^2\phi_n^{(2)}(b)$ and $U^3\phi_n^{(3)}(b)$, we use the same arguments as in (Khan 2001). It follows from our conditions on BW2 and KF2, FM1,2 that the two classes $\{h^k\phi_n^{(2)}(b) : b \in \mathcal{B}\}$ and $\{h^k\phi_n^{(3)}(b) : b \in \mathcal{B}\}$ are both Euclidean. Besides, these conditions ensure condition (ii) of Theorem 3 in (Sherman 1994b) holds with the "γ_n" therein being $O(1)$ for any sequence of δ_n converging to 0. Hence uniformly over an $o_p(1)$ neighborhood of β in $\mathcal{B}$, the third-order term $U^3\phi_n^{(3)}(b)$ is $h^{-k}O_p(n^{-3/2})$, which is $o_p(n^{-1})$ under our choice of bandwidth in BW2. The second-order term is $O_p\left(h^{-k}n^{-1}\right)$ over an $o_p(1)$ neighborhood of β. Let $\delta_n = O(h^{-k/2}n^{-1/2})$. Furthermore, following the same arguments in (Khan 2001), another application of Theorem 3 implies that over an $O_p(\delta_n)$ neighborhood of β, the second-order term is $O_p(\delta_n n^{-1}) = O_p(h^{-k/2}n^{-3/2})$, which is $o_p(n^{-1})$ given our choice of bandwidth in BW2.

Next, we deal with the first-order term $\phi_n^{(1)}$. By definition,

$$\begin{aligned}
& \mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)|\tilde{Z}_i = (y,z)] \\
\equiv\ & \mathbb{E}\left\{\kappa'(z, Z_j)K_h(x - X_j)\Delta\varphi_{i,j}^-(b)\nabla w(z)\left[Q_s\mathcal{K}_\sigma(z - Z_s) - \mathbb{E}\left[Q_s\mathcal{K}_\sigma(z - Z_s)\right]\right]\middle| Z_i = z\right\} \\
=\ & \mathbb{E}\left\{\kappa'(z, Z_j)\nabla w(z)K_h(x - X_j)\Delta\varphi_{i,j}^-(b)\middle| Z_i = z\right\}\left\{\mathbb{E}\left[Q_s\mathcal{K}_\sigma(z - Z_s) - \mathbb{E}\left[Q_s\mathcal{K}_\sigma(z - Z_s)\right]| Z_i = z\right]\right\}
\end{aligned}$$

where the second term on the right-hand side is 0 by construction. Besides,

$$\begin{aligned}
& \mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)|\tilde{Z}_j = (y,z)] \\
=\ & \mathbb{E}_{Z_i}\left\{\kappa'(Z_i, z)\nabla w(Z_i)K_h(X_i - x)\Delta\varphi_{i,j}^-(b)\mathbb{E}_{Z_s}\left\{Q_s\mathcal{K}_\sigma(Z_i - Z_s) - \mathbb{E}\left[Q_s\mathcal{K}_\sigma(Z_i - Z_s)\right]| Z_i, Z_j = z\right\}|Z_j = z\right\} \\
=\ & \mathbb{E}_{Z_i}\left\{\kappa'(Z_i, z)\nabla w(Z_i)K_h(X_i - x)\Delta\varphi_{i,j}^-(b)\mathbb{E}_{Z_s}\left\{Q_s\mathcal{K}_\sigma(Z_i - Z_s) - \mathbb{E}\left[Q_s\mathcal{K}_\sigma(Z_i - Z_s)\right]| Z_i\right\}|Z_j = z\right\}
\end{aligned}$$

where $\mathbb{E}_{Z_s}\left[Q_s\mathcal{K}_\sigma(Z_i - Z_s) - \mathbb{E}\left[Q_sK_\sigma(Z_i - Z_s)\right]| Z_i\right] = 0$ conditional on all Z_i. Hence this term is also degenerate at 0 for all Z_i and b. It then follows that the unconditional expectation $\mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)] = 0$ for all b. Hence the first two terms in the H-decomposition in (52) are reduced to:

$$\begin{aligned}
& \tfrac{1}{n}\textstyle\sum_{l=1}^n \mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s}; b\right)|\tilde{Z}_s = (y_l, z_l)] \\
=\ & \tfrac{1}{n}\textstyle\sum_{l=1}^n \mathbb{E}_{Z_i}\left\{\tilde{m}_n(Z_i; b)\left[Q_s\mathcal{K}_\sigma(Z_i - Z_s) - \mathbb{E}_{Q',Z'}\left(Q'\mathcal{K}_\sigma(Z_i - Z')\right)\right]|\tilde{Z}_s = (y_l, z_l)\right\}
\end{aligned} \tag{54}$$

where $q_l \equiv (y_l, 1)'$; and

$$\tilde{m}_n(z;b) \equiv \mathbb{E}_{Z_j}[\nabla w(Z_i)\kappa'(Z_i,Z_j)K_h\left(X_i - X_j\right)\Delta\varphi^-_{i,j}(b)|Z_i = z].$$

To clarify notations, note on the second line of (54), $\mathbb{E}\left[Q'\mathcal{K}_\sigma\left(Z_i - Z'\right)\right]$ is a function of Z_i and the expectation is taken w.r.t. Q', Z'.

It remains to show that we can write the right-hand side of (54) as a sample average of some function of (z_l, y_l) plus a term that is smaller than $o_p(\|b-\beta\|/\sqrt{n}) + o_p\left(\|b-\beta\|^2\right) + o_p(n^{-1})$ uniformly over $o_p(1)$ neighborhood around β.

By changing variables between x_j and $u \equiv h^{-k}\left(x_i - x_j\right)$ while fixing z_i, we have

$$\begin{aligned}
\tilde{m}_n(z_i;b) &= \nabla w(z_i)\int \kappa'(z_i,z_j)K_h\left(x_i - x_j\right)\Delta\varphi^-(z_i,z_j;b)dF(z_j) \\
&= \nabla w(z_i)\int K_h(x_i - x_j)\tilde{\mu}^-(z_i,x_j;b)dF(x_j) \\
&= \nabla w(z_i)\int K\left(u\right)\tilde{\mu}^-(z_i,x_i - h^k u;b)f(x_i - h^k u)du \\
&= \nabla w(z_i)\tilde{\mu}^-(z_i,x_i;b)f(x_i) + O(h_n^{m_\varphi})
\end{aligned}$$

where $\tilde{\mu}^-(z_i,x_j;b) \equiv \mathbb{E}[\kappa'(Z_i,Z_j)\Delta\varphi^-_{i,j}(b)|Z_i = z_i, X_j = x_j]$. The first equality follows from independence between Z_i and Z_j; the second from the law of iterated expectation; the third from changing variables between u and x_j; and the last from applying a Taylor expansion of x_j around x_i, and using the boundedness of the derivatives under SM2 and WF and the order of K in KF2.

Let $m^*_-(z;b) \equiv \nabla w(z)f(x)\tilde{\mu}^-(z,x;b)$. Then (54) can be written as:

$$\begin{aligned}
&\int \tilde{m}_n(z;b)\left(\tfrac{1}{n}\textstyle\sum_{l=1}^n q_l\mathcal{K}_\sigma\left(z - z_l\right) - \int \mathbb{E}\left(Q'|x'\right)\mathcal{K}_\sigma\left(z - z'\right)f(z')dz'\right)dF(z) \\
= \quad &\tfrac{1}{n}\textstyle\sum_{l=1}^n \int q_l\left[m^*_-(z;b) + O(h_n^{m_\varphi})\right]\mathcal{K}_\sigma\left(z - z_l\right)f(z)dz \\
- \quad &\int f(z')\mathbb{E}\left(Q'|x'\right)\left(\int \left[m^*_-(z;b) + O(h_n^{m_\varphi})\right]\mathcal{K}_\sigma\left(z - z'\right)f(z)dz\right)dz'.
\end{aligned}$$

Thus this suggests $\frac{1}{n}\sum_{l=1}^n \mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s};b\right)|\tilde{Z}_s = (y_l,z_l)]$ can be decomposed into the sum of the following two terms:

$$\tfrac{1}{n}\textstyle\sum_{l=1}^n \int q_l m^*_-(z;b)\mathcal{K}_\sigma\left(z - z_l\right)f(z)dz - \int f(z')\mathbb{E}\left(Q'|x'\right)\left(\int m^*_-(z;b)\mathcal{K}_\sigma\left(z - z'\right)f(z)dz\right)dz' \quad (55)$$

and

$$O(h_n^{m_\varphi})\left\{\tfrac{1}{n}\textstyle\sum_{l=1}^n \int q_l\mathcal{K}_\sigma\left(z - z_l\right)f(z)dz - \int f(z')\mathbb{E}\left(Q'|x'\right)\left(\int \mathcal{K}_\sigma\left(z - z'\right)f(z)dz\right)dz'\right\}. \quad (56)$$

We first examine the term in (56). Note $\int \mathcal{K}_\sigma(z-z')f(z)dz = f(z') + O(\sigma_n^{m_\mathcal{K}})$ for all z' under our conditions, and the term in the braces above can be written as

$$\begin{aligned} & \tfrac{1}{n}\textstyle\sum_{l=1}^n \int q_l \mathcal{K}_\sigma(z-z_l) f(z)dz - \int f(z')\mathbb{E}(Q'|x')[f(z') + O(\sigma_n^{m_\mathcal{K}})]dz' \\ = & \tfrac{1}{n}\textstyle\sum_{l=1}^n \int f(z) q_l \mathcal{K}_\sigma(z-z_l)\, dz - \mathbb{E}[f(Z)Q] + O(\sigma_n^{m_\mathcal{K}}) \\ = & \tfrac{1}{n}\textstyle\sum_{l=1}^n \{f(z_l)q_l - \mathbb{E}[f(Z)Q]\} + o_p(n^{-1/2}) + O(\sigma_n^{m_\mathcal{K}}) \end{aligned}$$

where the last equality follows from arguments identical to Theorem 8.11 in (Newey and McFadden 1994) and our choice of bandwidth in BW1. Also by our choice of bandwidth in BW1,2, both $O(h_n^{m_\varphi})$ and $O(\sigma_n^{m_\mathcal{K}})$ are $o(n^{-1/2})$. Note $\frac{1}{n}\sum_{l=1}^n \{f(z_l)q_l - \mathbb{E}[f(Z)Q]\}$ is $O_p(n^{-1/2})$ by the Central Limit Theorem. Thus the term in (56) is $o_p(n^{-1})$ uniformly over an $o_p(1)$ neighborhood around β.

Next, to deal with (55), for any z, we can apply a Taylor expansion of m_-^* around $b = \beta$ to get:

$$m_-^*(z;b) = 0 + \nabla_b m_-^*(z;\beta)(b-\beta) + o(\|b-\beta\|).$$

Substituting this into (55) above, we decompose it into the sum of

$$(b-\beta)\left\{\tfrac{1}{n}\sum_{l=1}^n \int q_l \nabla_b m_-^*(z;\beta)\mathcal{K}_\sigma(z-z_l)f(z)dz - \int f(z')\mathbb{E}(Q'|x')\left(\int \nabla_b m_-^*(z;\beta)\mathcal{K}_\sigma(z-z')f(z)dz\right)dz'\right\} \tag{57}$$

and

$$o(\|b-\beta\|)\left\{\tfrac{1}{n}\textstyle\sum_{l=1}^n \int q_l \mathcal{K}_\sigma(z-z_l)f(z)dz - \int f(z')\mathbb{E}(Q'|x')\left(\int \mathcal{K}_\sigma(z-z')f(z)dz\right)dz'\right\}, \tag{58}$$

where the latter term is of order smaller than $o_p(\|b-\beta\|/\sqrt{n})$ as the term in the braces in (58) is $O_p(n^{-1/2})$ by the same arguments above.

As for the term in the braces in (57), first note by standard arguments using change of variables, we have:

$$\int \nabla_b m_-^*(z;\beta) f(z) \mathcal{K}_\sigma(z-z')\, dz = \nabla_b m_-^*(z';\beta)f(z') + O(\sigma_n^{m_\mathcal{K}}).$$

Thus the term in the braces of (57) is

$$\begin{aligned} & \tfrac{1}{n}\textstyle\sum_{l=1}^n \int q_l \nabla_b m_-^*(z;\beta) f(z) \mathcal{K}_\sigma(z-z_l)\, dz - \int \mathbb{E}(Q'|x')\nabla_b m_-^*(z';\beta) f(z')^2 dz' + O(\sigma_n^{m_\mathcal{K}}) \\ = & \left\{\tfrac{1}{n}\textstyle\sum_{l=1}^n \int q_l \nabla_b m_-^*(z;\beta) f(z) \mathcal{K}_\sigma(z-z_l)\, dz - \mathbb{E}[Qf(Z)\nabla_b m_-^*(Z;\beta)]\right\} + O(\sigma_n^{m_\mathcal{K}}) \end{aligned} \tag{59}$$

Again by arguments similar to above and citing same arguments from Theorem 8.11 in (Newey and McFadden 1994) under SM2 and FM2, the term in the braces of (59) is

$$\tfrac{1}{n}\textstyle\sum_{l=1}^n \delta_-^*(y_l, z_l) + o_p(n^{-1/2}) \text{ where } \delta_-^*(y,z) \equiv q\nabla_b m_-^*(z;\beta)f(z) - \mathbb{E}[Q\nabla_b m_-^*(Z;\beta)f(Z)];$$

while $O(\sigma_n^{m\kappa}) = o(n^{-1/2})$ under our choice of bandwidth. To sum up, we have shown

$$\frac{1}{n}\sum_{l=1}^{n}\mathbb{E}[\phi_n\left(\tilde{Z}_{i,j,s};b\right)|\tilde{Z}_s=(y_l,z_l)] = \frac{1}{n}\sum_{l=1}^{n}\delta_-^*(y_l,z_l)(b-\beta)+o_p(\|b-\beta\|/\sqrt{n})+o_p(n^{-1})$$

uniformly over an $o_p(1)$ neighborhood around β in $\mathcal{B}$, where δ_-^* is the correction term due to $\hat{p}(z_i)$ in $\hat{H}_n$.

Because $\hat{p}_i$ and $\hat{p}_j$ enter the objective function in the same way, and p_i and p_j are additively separable in the first-order expansion, we can apply identical arguments above to derive another identical correction term due to the use of $\hat{p}(z_j)$ in $\hat{H}_n$. This proves the claim of the lemma. Q.E.D.

Replicating the arguments in the preceding lemma we can prove a result similar to Lemma C3 holds for the other half of the difference between "feasible" and "infeasible" objective function $|\widehat{\mathcal{H}}_{2,n}(b)-\mathcal{H}_{2,n}(b)|$, except that δ_-^* needs to be replaced by

$$\delta_+^*(y,z) \equiv q\nabla_b m_+^*(z;\beta)f(z) - \mathbb{E}[Qf(Z)\nabla_b m_+^*(Z;\beta)$$

where

$$\begin{aligned} m_+^*(z) &\equiv \nabla w(z)f(x)\tilde{\mu}^+(z,x;b); \text{ with} \\ \tilde{\mu}^+(z_i,x_j;b) &\equiv \mathbb{E}[\kappa'(Z_i,Z_j)\Delta\varphi^+(Z_i,Z_j;b)|Z_i=z_i,X_j=x_j]. \end{aligned}$$

Building on the preceding Lemmas, we are now ready to prove the final result about the limiting distribution of $\hat{\beta}$.

Proof of Proposition 7. By Lemma C2 and Lemma C3,

$$\hat{H}_n(b) = H_0(b) + O_p(\|b-\beta\|/\sqrt{n}) + o_p(\|b-\beta\|^2) + O_p(n^{-1}h^{-k}) \tag{60}$$

uniformly over an $o_p(1)$ neighborhood around β in $\mathcal{B}$. Recall $H_0(b)$ is minimized at $b=\beta$ due to Proposition 5. Hence it follows from (60), Lemma C1 above and Theorem 1 in (Sherman 1994b) that $\hat{\beta}$, as the minimizer of $\hat{H}_n(b)$ over $b\in\mathcal{B}$, converges to β at a rate of $1/\sqrt{nh^k}$. As stated in Lemma C2 and Lemma C3, the $O_p(n^{-1}h^{-k})$ term in (60) is further reduced to $o_p(n^{-1})$ under conditions of the proposition. Hence another application of Theorem 1 in (Sherman 1994b) suggests $\left\|\hat{\beta}-\beta\right\| = O_p(n^{-1/2})$.

Recall that by a second-order Taylor expansion, $H_0(b) = \frac{1}{2}(b-\beta)'\nabla_{bb}H_0(\beta)(b-\beta) + o_p(\|b-\beta\|^2)$ over an $o(1)$ neighborhood of β, for $\nabla_b H_0(\beta)=0$ by construction. This, together with Lemma C2 and Lemma C3 and the root-n convergence shown in the previous paragraph, suggests that

$$\hat{H}_n(b) = \tfrac{1}{2}(b-\beta)'\nabla_{bb}H_0(\beta)(b-\beta) + \frac{1}{n}\sum_{i=1}^{n}2[\delta_-^*(\tilde{z}_i)+\delta_+^*(\tilde{z}_i)](b-\beta)+o_p(n^{-1})$$

uniformly over an $O_p(n^{-1/2})$ neighborhood around β. The limiting distribution then follow from Theorem 2 in (Sherman 1994b) and that $\mathbb{E}[\delta^*(\tilde{Z})\delta^*(\tilde{Z})'] < \infty$ under FM2. *Q.E.D.*

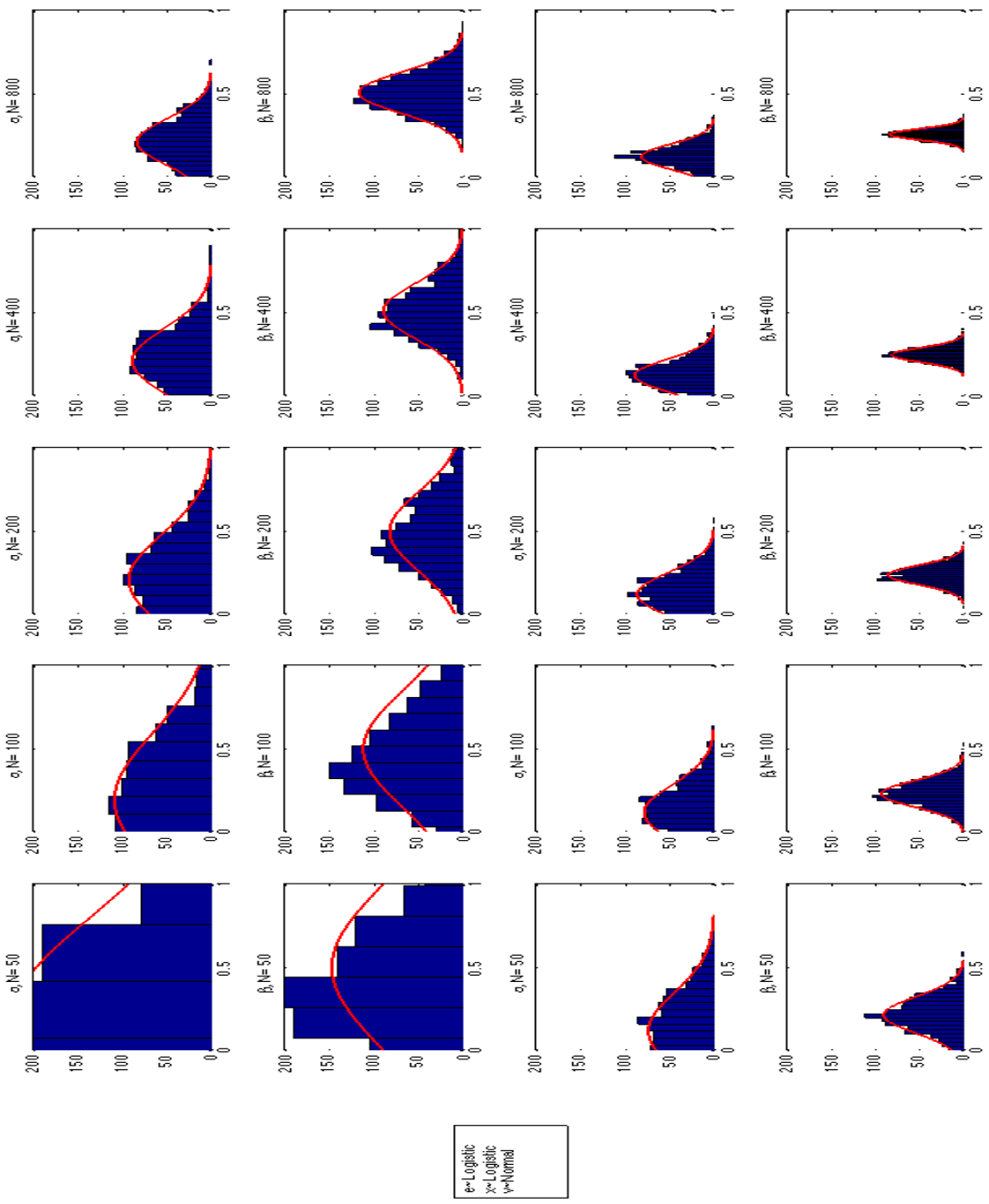

Figure 1. $(X, \epsilon)\tilde{}$(Logistic,Logistic). First two rows: Pairwise extremum estimator. Last two rows: Inverse-density weighted estimator.

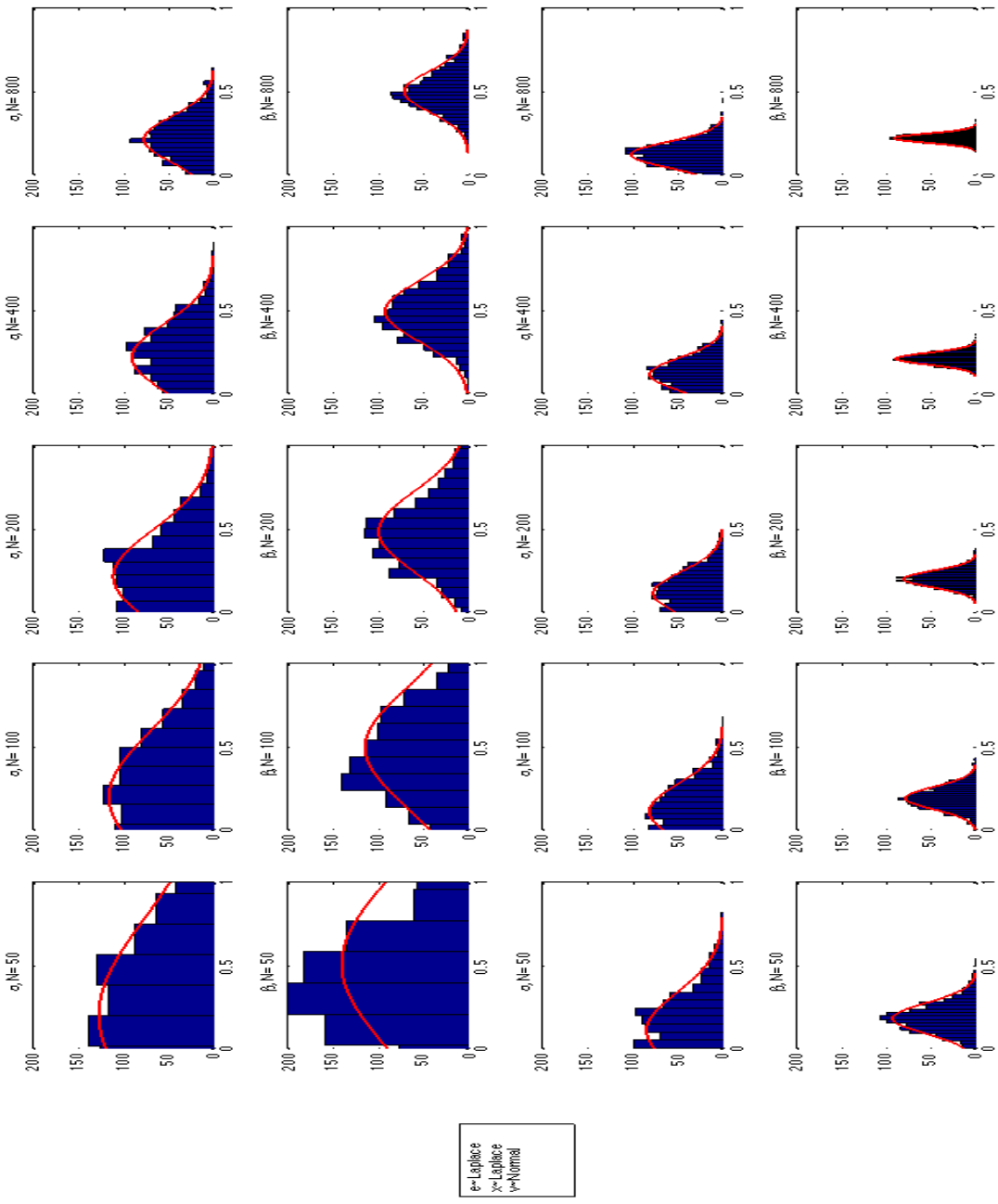

Figure 2. $(X, \epsilon)\tilde{}$(Laplace,Laplace)

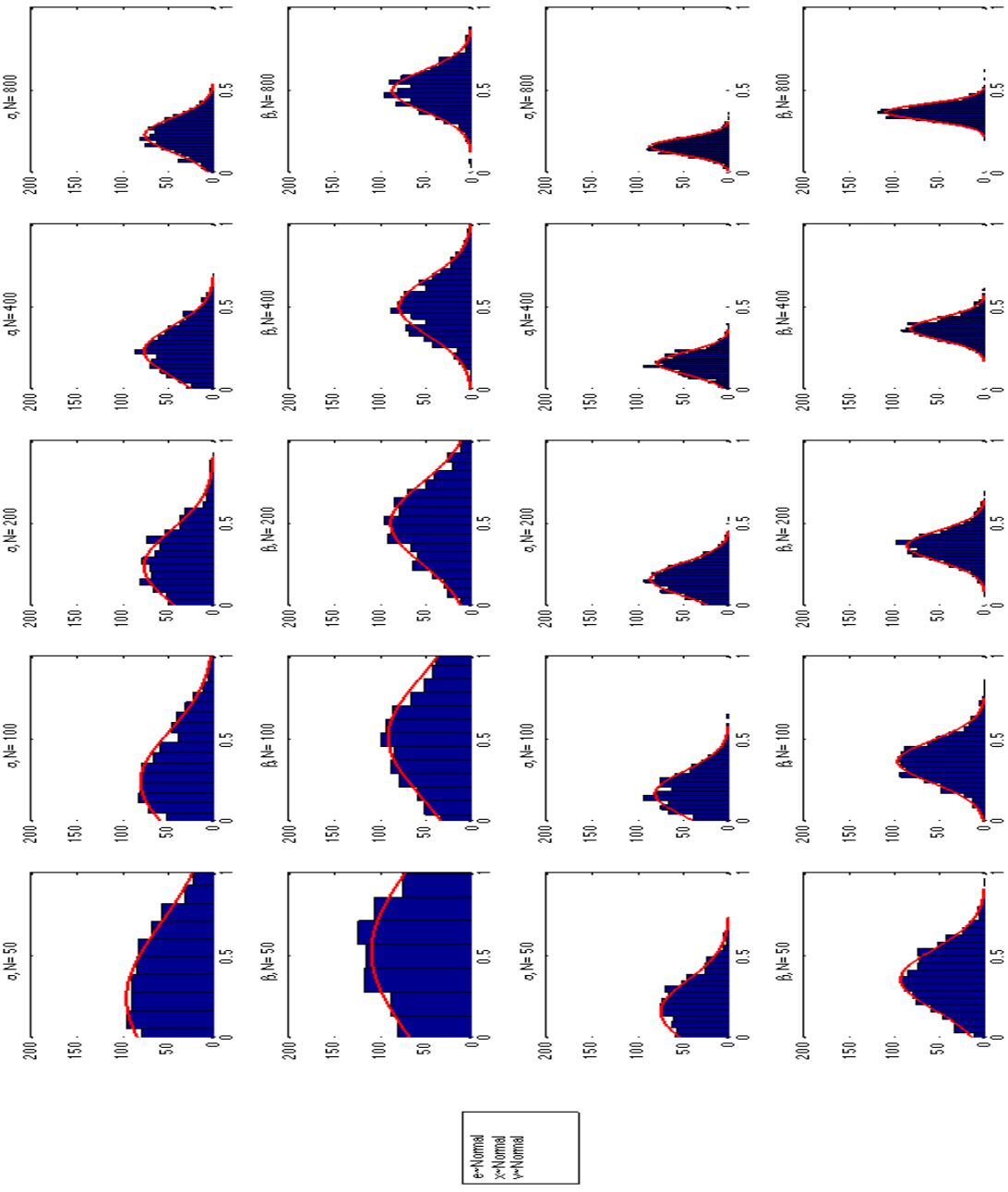

Figure 3. $(X, \epsilon)\tilde{}$(Normal,Normal).

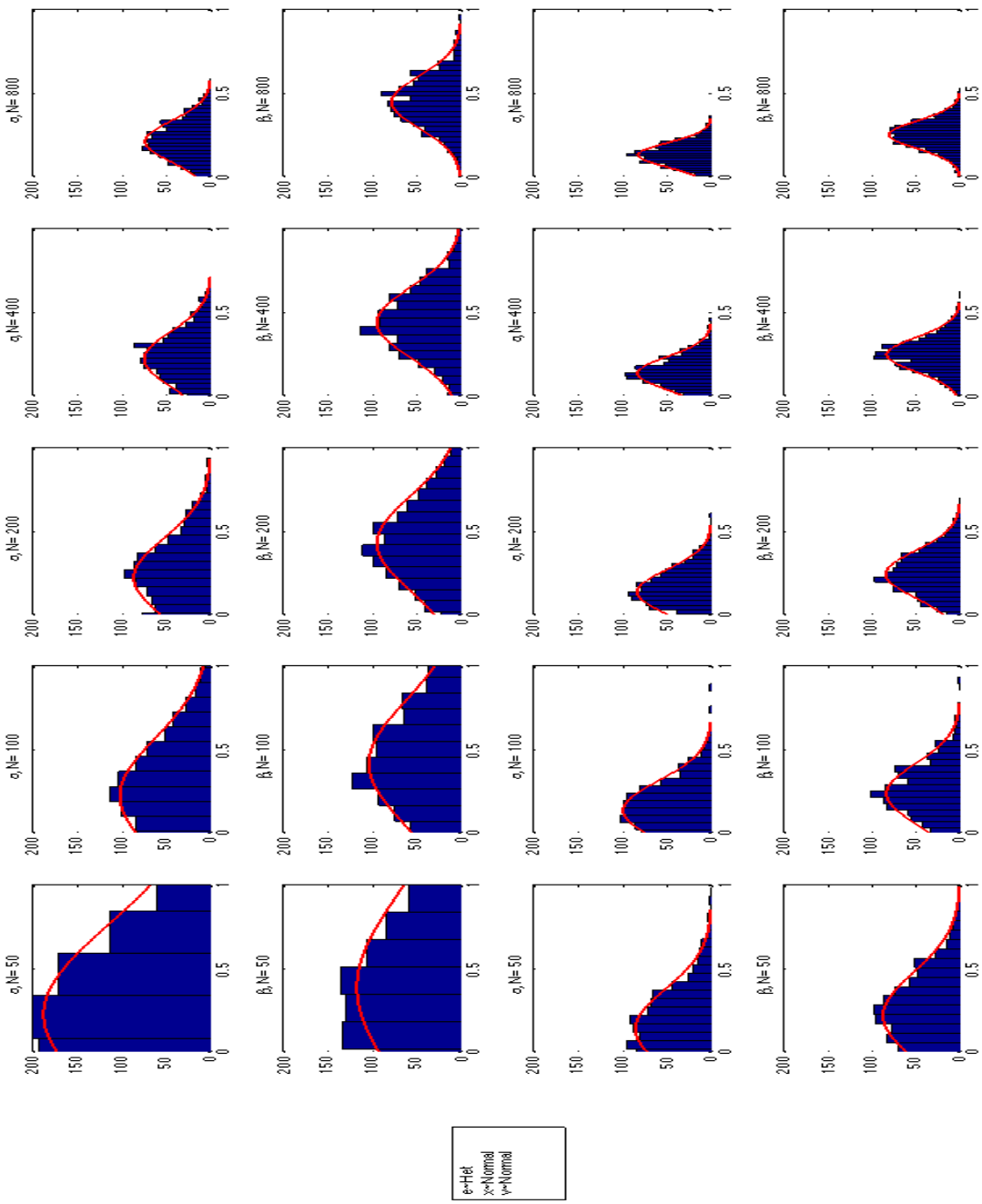

Figure 4. Heteroskedastic Error. (X, ϵ)˜(Normal,Normal).

References

Abrevaya, J., J. Hausman, and S. Khan (2010): "Testing for Causal Effects in a Generalized Regression Model with Endogenous Regressors," *Econometrica*, pp. 2043–2061.

Amemiya, T. (1985): *Advanced Econometrics*. Harvard University Press.

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

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 Conþdence 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.

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. (2001): "Two Stage Rank Estimation of Quantile Index Models," *Journal of Econometrics*, 100, 319–355.

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

LEWBEL, A., AND X. TANG (2012): "Identification and Estimation of Games with Incomplete Information Using Excluded Regressors," Working paper, Boston College and U Penn.

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.

MANSKI, C., AND E. TAMER (2002): "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70(2), 519–546.

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. (1994a): "Maximal inequalities for degenerate U-processes with applications to optimization estimators," *Annal of Statistics*, 22, 439–459.

——— (1994b): "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.