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.\footnote{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).} 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.\footnote{They showed in a stylized model that there is no informative partial identification result for the intercept in this case.} 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-
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 re- 
mains 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 dis- 
tinct 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), (Mancac 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:

\[ Y = 1 \{ X \beta - V \geq \epsilon \} \]

(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 \( F_R, f_R, \Omega_R \) 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 \( \epsilon \) 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 \( \Theta \) be the parameter space for \( F_{\epsilon|X} \) (i.e. \( \Theta \) 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 \( \sigma \)-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 \( \xi(b, G_{\epsilon|X}) \equiv \{ z : p(z) \neq \int 1 (\epsilon \leq xb - v) \ dG_{\epsilon|X} \} \) and \( \tilde{\xi}(b, G_{\epsilon|X}) \equiv \{(z, z') : (p(z), p(z')) \neq \left( \int 1 (\epsilon \leq xb - v) \ dG_{\epsilon|X}, \int 1 (\epsilon \leq x'b - v') \ dG_{\epsilon|X'} \right) \} \).

(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 $\xi$ in (2) consists of pairs of states where implied propensity scores differ from those in the true DGP. We say $\beta$ 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 $\beta$ hinges on model restrictions defining $\Theta$, i.e. parameter space for error distributions given $X$. In Sections 3 and 4, we discuss identification of $\beta$ when CI is paired with one of the following two stochastic restrictions on errors respectively.

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

\textbf{CS} \textit{(Conditional Symmetry)} For all $x$, $\epsilon$ 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 $\beta$ 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 $\beta$. The semiparametric efficiency bound is formally defined as follows. Let $\mu$ denote some measure on $\{0,1\} \otimes \Omega_Z$ such that $\mu(\{0\} \otimes \Omega) = \mu(\{1\} \otimes \Omega) = F_Z(\omega)$, where $\omega$ is a Borel subset of $\Omega_Z$. A path that goes through $F_{\epsilon|x}$ is a function $\lambda(\epsilon,x;\delta)$ such that $\lambda(\epsilon,x;\delta_0) = F_{\epsilon|x}(\epsilon)$ for some $\delta_0 \in \mathbb{R}$, and $\lambda(.,.;\delta) \in \Theta$ for all $\delta$ in an open neighborhood around $\delta_0$. 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 $\lambda$ such that there exists $\{(\psi_k)_{k \leq K}, \psi_\lambda\}$ such that
\[
f^{1/2}_\lambda(y|z;b,\delta) = f^{1/2}_\lambda(y|z;b,\delta_0) = \sum_k \psi_k(y,z)(b - \beta) + \psi_\lambda(y,z)(\delta - \delta_0) + r(y,z;b,\delta) \quad (3)
\]
with
\[
(\|b - \beta\| + \|\delta - \delta_0\|)^{-2} \int r^2(y,z;b,\delta)d\mu \rightarrow 0 \text{ as } b \rightarrow \beta \text{ and } \delta \rightarrow \delta_0. \quad (4)
\]
The path-wise partial information for the $k$-th coordinate in $\beta$ is
\[
I_{\lambda,k} \equiv \inf_{\{(\alpha_j)_{j \neq k}, \alpha_\lambda\}} 4 \int \left(\psi_k - \sum_{j \neq k} \alpha_j \psi_j - \alpha_\lambda \psi_\lambda\right)^2 d\mu. \quad (5)
\]
The information for $\beta_k$ is the infimum of $I_{\lambda,k}$ over all smooth parametric submodels $\lambda$. 
3 Exclusion plus Median Independence

This section discusses the identification and information for $\beta$ 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 $\beta$. 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 $\beta$ 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 $\beta$ is identified relative to $b$ if and only if $\Pr\{z \in Q_b\} > 0$, where $Q_b \equiv \{z : x\beta \leq v < xb \text{ or } xb < v < x\beta\}$.

For a model satisfying MI but not CI (i.e. $F_{\epsilon|X,V}(0) = 1/2$ and $\epsilon$ 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 $\beta$, 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 $\beta$, as the set of states that help identify $\beta$ relative to $b$ remains unchanged.

The intuition for such an equivalence builds on two observations. First, if states in $Q_b$ help identify $\beta$ 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 $\beta$, 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 $\beta$. That is, $Q_b$ becomes a “thin set" as $b$ approaches $\beta$.

It also follows from Proposition 1 that conditions that yield point identification of $\beta$ in (Manski 1988) are also sufficient for point identification of $\beta$ 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 $\beta$ is bounded. It differs from the large support conditions needed to point identify $\beta$ 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 $\gamma$ in $\mathbb{R}^{K-1}$ and $c \in \mathbb{R}$ with $\Pr \{X \gamma = c\} = 1$. 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 $\beta$ is identified relative to all $b \neq \beta$.

For estimation, we propose a new extremum estimator for $\beta$ 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

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

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

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

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 $\kappa$ 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 $\beta$ 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. We conjecture the estimator converges at the cubic rate under conditions in (Chernozhukov, Hong, and Tamer 2007).

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 \( \beta \) 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 \( \beta \).

**RG (Regularity)** For each \((b,G_{\epsilon|X})\) in the parameter space, there exists a measurable function \( q : \{0,1\} \otimes \Omega_Z \to \mathbb{R} \) such that \( \left| \frac{\partial f^{1/2}(y,z;\eta,G_{\epsilon|X})}{\partial b} \right| \leq q(y,z) \) for all \( \eta \) 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 \( \Theta \) denote the parameter space for the distribution of \( \epsilon \) 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 \( \beta \) under CI, MI and RG. Let \( \Lambda \) consist of paths

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

where \( F_{\epsilon|x} \) is the true conditional distribution in DGP from \( \Theta \); and \( h : \mathbb{R}^{K+1} \rightarrow \mathbb{R} \) 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 \( \beta \) in a model under MI alone (without the exclusion restrictions in CI). In that latter

---

\(^7\)The identified set for \( \beta \) 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 \( \kappa \) 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
\[ \lambda(\varepsilon, x; \delta) \equiv F_{\varepsilon|x}(\varepsilon) \left[ 1 + (\delta - \delta_0) \tilde{h}(\varepsilon, z) \right] \]
with \( \tilde{h} : \mathbb{R}^{K+2} \rightarrow \mathbb{R} \) 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 \( \lambda(\cdot, \cdot; \delta) \) in (8) is in \( \Theta \) for \( \delta \) close enough to \( \delta_0 \). Besides, \( f^{1/2}_\lambda(\cdot; b, \delta) \) is mean-square differentiable at \( (b, \delta) = (\beta, \delta_0) \) with:

\[
\psi_k(y, z) = \frac{1}{2} \left\{ yF_{\varepsilon|x}(w)^{-1/2} - (1 - y) \left[ 1 - F_{\varepsilon|x}(w) \right]^{-1/2} \right\} f_{\varepsilon|x}(w)x_k
\]

\[
\psi_\lambda(y, z) = \frac{1}{2} \left\{ yF_{\varepsilon|x}(w)^{-1/2} - (1 - y) \left[ 1 - F_{\varepsilon|x}(w) \right]^{-1/2} \right\} F_{\varepsilon|x}(w)h(w, x)
\]

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

**Proposition 2** Suppose CI, MI, SV, FR and RG hold in (1); \( Z \) has finite second moments; and \( \text{Pr}(X\beta = V) = 0 \). Then the information for \( \beta_k \) 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 \( \beta \) 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 \( \beta_k \) under CI and MI is closely related to two facts: there is no incremental identifying power for \( \beta \) from CI given MI; and there is zero information for \( \beta_k \) under MI alone. Second, root-n estimator for \( \beta \) 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 \( \beta \) 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 $F_{e|X}$ under MI and CL, note $\mathbb{E}(Y|x,v) = F_{e|x}(x\beta - v)$. With $\beta$ identified, $F_{e|x}(t)$ 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 $\beta_1$ and $F_{e|X_1, X_2}$ but different weights for health status $\beta_2$ (where $\beta_2 > \beta_1$). Then knowledge of $F_{e|X}$ as well as $\beta_1, \beta_2$ helps at least bound the counterfactual retirement probabilities conditional on $Z \equiv (X_1, X_2, V)$. If the magnitude of the difference between $\beta_2$ and $\beta_1$ 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 $x_1\beta_1 + x_2\beta_2 - v$ 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 $G(x,v) \equiv \int 1\{\epsilon \leq x\beta - v\}dF_{e}(\epsilon) = \Pr(\epsilon \leq x\beta - v)$. 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

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

With $\beta$ 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:

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

where

$$\hat{\varphi}(x_i,x,v) \equiv \frac{\sum_{j \neq i} y_j K_\sigma \left((x_j - x_i, v_j - (v + (x_i - x)\beta))\right)}{\sum_{j \neq i} K_\sigma \left((x_j - x_i, v_j - (v + (x_i - x)\beta))\right)}$$
with $\mathcal{K}_\sigma(\cdot) \equiv \sigma^{-(k+1)} \mathcal{K}(\cdot/\sigma^{k+1})$ 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 $\beta$ 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 $Y = 1\{h_1(Z) + \varepsilon_1 \geq h_0(Z) + \varepsilon_0\} = 1\{h^*(Z) \geq \varepsilon^*\}$, where $h^* \equiv h_1 - h_0$ and $\varepsilon^* \equiv \varepsilon_0 - \varepsilon_1$. As long as $\varepsilon_1$ and $\varepsilon_0$ are i.i.d. draws from the same marginal, the normalized error $\varepsilon^*$ must be symmetrically distributed around 0 given $Z$.

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

4.1 Identification

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

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

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

This condition requires the pointwise continuity in $(x, \varepsilon)$ to hold with equal variation all over the parameter space $\Theta_{CS}$, in the sense that the same $\delta_\eta(x, \varepsilon)$ is used to satisfy the “$\delta$-$\eta$-neighborhood” definition of pointwise continuity at $(x, \varepsilon)$ for all elements in $\Theta_{CS}$.

An alternative way to formulate EC is that for any $\eta > 0$ and $(x, \varepsilon)$, the infimum of $\delta_\eta(x, \varepsilon; G_{\varepsilon|x})$ (i.e. the radius of neighborhood around $x$ in the definition of pointwise continuity) over $G_{\varepsilon|x} \in \Theta_{CS}$ is bounded away from zero by a positive constant.
Such an equicontinuity condition is needed because the identification of $\beta$ relative to $b \neq \beta$ states that $b$ cannot be paired with any $G_{i|X} \neq F_{i|X}$ in $\Theta_{CS}$ 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 $G_{i|X}$ in $\Theta_{CS}$ are Lipschitz-continuous with their modulus uniformly bounded by a finite constant.

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

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

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 $\Pr\{V_i \in A|(x_c, x_d)\} > 0$ for any set $A$, then $\Pr\{V_i \in A|\tilde{x}_c, x_d\} > 0$ for $\tilde{x}_c$ close enough to $x_c$.

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

$$\int 1\{(v_i, v_j) \in R_b(x)\}dF_{V_i,V_j|x} > 0.$$  \hspace{1cm} (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 $(z_i, z_j) \in Q_{b,S}$ 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

$$“x_i \beta - v_i < -(x_j \beta - v_j) \text{ and } x_i b - v_i > -(x_j b - v_j)”$$

or

$$“x_i \beta - v_i > -(x_j \beta - v_j) \text{ and } x_i b - v_i < -(x_j b - v_j)”$$

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 $G_{i|X} \in \Theta_{CS}$ 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 $\beta$ 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 $\Pr\{(Z_i, Z_j) \in \tilde{\xi}(b, G_{i|x})\} > 0$ for all $G_{i|x} \in \Theta_{CS}$. On the other hand, if both (i) and (ii) fail, then $\beta$ is not identified to $b$ because some $G_{i|x} \neq F_{i|x}$ can be constructed so that $(b, G_{i|x})$ is observationally equivalent to the true parameters $(\beta, F_{i|x})$. That is, $(b, G_{i|x})$ 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 $F_{i|x}$. 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 “$\delta$-expansion" of $\tilde{Q}_{b,S}$ defined as:

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

provided $\delta > 0$ is small enough. To identify $\beta$ relative from $b$, it then suffices to require $\tilde{Q}_{b,S}^\delta$ to have positive probability for such small $\delta$, 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 $\beta$. He showed the sets of states that help distinguish $\beta$ 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 $\beta$ relative to $b$. Thus in that sense a regressor being special does add to the informational content of the model.

Finally, note by construction any condition that identifies $\beta$ under CI and MI also identifies $\beta$ 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 $\int 1\{x\beta < \frac{v_i + v_j}{2} < xb\}dF_{i,v_i,j|x} > 0$ or $\int 1\{xb < \frac{v_i + v_j}{2} < x\beta\}dF_{i,v_i,j|x} > 0$ for all $x \in \omega$. This is because under SV, $V_i$ and $V_j$ are independent draws from $F_{i,v_i,j|x}$ and both fall in an open neighborhood around $x\beta$ with positive probability. Identification of $\beta$ follows from Proposition 3.

### 4.2 Positive Fisher Information

We now show how CS together with CI leads to positive information for $\beta$ in heteroskedastic binary regressions. (Zheng 1995) showed without any excluded regressors the information for $\beta$ 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 $\beta$ 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 $\beta$ in the next subsection.

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

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

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

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

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

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

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

**Proposition 4** Under CI, CS’, EC, CT, FR, SV and RG’, the information for $\beta_k$ 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 $\mu$ (the measure on $\{0, 1\} \otimes \Omega_Z$ defined in Section 2), we can show the Fisher information for $\beta_k$ takes the form of

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

where $\phi(z) \equiv \left[ F_{\varepsilon|x}(w)(1 - F_{\varepsilon|x}(w)) \right]^{-1} \geq 0$; and $(\alpha_j^*)_{j \neq k}$ and $\alpha_k^*$ 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 $\alpha_k^* \neq 0$. (Otherwise the pathwise information $I_{\lambda,k}$ under $\lambda$ would equal that of a parametric model where true error distribution $F_{\varepsilon|x}$ is known, and be positive. This would contradict the claim that $I_{\lambda,k} = 0$.) Since each path $\lambda$ in $\Lambda$ needs to satisfy conditional symmetry for $\delta$ close to $\delta_0$, $\lambda_\delta(w, x; \delta_0)$ (and consequently its product with the nonzero $\alpha_k^*$) must be odd functions in $w$ once $x$ is fixed. At the same time, $f_{\varepsilon|x}(w)$ is an even function of $w$ (i.e. symmetric in $w$ around 0) given $x$. Then the pathwise information for $\beta_k$ under $\lambda$ 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 $\beta_k$ 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 $\epsilon$ is symmetric around 0 given $Z = (X, V)$ with unrestricted dependence between $\epsilon$ and all coordinates in $Z$. The information for $\beta_k$ is zero in that case because the scores $\psi_k$ and $\psi_\lambda$ 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 $\psi_\lambda$ that are arbitrarily good approximation to $\psi_k$ in $L^2(\mu)$-norm, provided for the path $\lambda$ is appropriately a chosen. To see this, note $I_{\lambda,k} =$

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

(17)

Indeed the same path used for showing zero information for $\beta_k$ under MI with no excluded regressors (see Theorem 5 in (Chamberlain 1986)) also drives the information for $\beta_k$ to zero in (Zheng 1995). Specifically, the path is $\lambda(\epsilon, z; \delta) = F_{\epsilon|z}(\epsilon) [1 + (\delta - \delta_0) h(\epsilon, z)]$, where $h$ is continuously differentiable, equals zero outside of some compact set, and $h(0, z) = 0$ for all $z$ so that $\lambda_\delta(\epsilon, z; \delta_0) = h(\epsilon, z)$. Since there is no restriction on how the vector $z$ enters $\lambda_\delta$, 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 $\lambda_\delta(\epsilon, x; \delta_0)$ through the index $w = x^\beta - v$. This additional form restriction is what delivers the positive information for $\beta_k$.

(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 $X$.\textsuperscript{10} They showed the information for $\beta_k$ 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 $\beta$ can be constructed. We consider the case where all coordinates in $Z$ are continuously distributed. Extensions

\textsuperscript{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 $\beta$ 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 \((.)_-=\min\{.,0\}\) and \((.)_+=\max\{.,0\}\). Our estimator is

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

where:

\[
\hat{H}_n(b) \equiv \frac{1}{n(n-1)} \sum_{j \neq i} K_n(x_i - x_j) \left[ \kappa(\hat{w}_{i,j} - 1) \varphi^-(Z_i, Z_j; b) + \kappa(1 - \hat{w}_{i,j}) \varphi^+(Z_i, Z_j; b) \right];
\]

\[
\varphi^-(z_i, z_j; b) \equiv \left( \frac{(x_i + x_j)^t}{2} b - \frac{v_i + v_j}{2} \right)_- \quad \text{and} \quad \varphi^+(z_i, z_j; b) \equiv \left( \frac{(x_i + x_j)^t}{2} b - \frac{v_i + v_j}{2} \right)_+;
\]

\[
\hat{w}_{i,j} \equiv \hat{p}_i + \hat{p}_j; \quad \text{and} \quad \hat{p}_i \equiv \hat{p}(z_i) \equiv \frac{\sum_{s \neq i} y_n \mathcal{K}_\sigma(z_s - z_i)}{\sum_{s \neq l} \mathcal{K}_\sigma(z_s - z_l)} \quad \text{for} \ i, j.
\]

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 \(h_n, \sigma_n\) being kernel functions and bandwidths whose properties are to be specified below. The weighting function \(\kappa\) 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 \(t > 0\); \(\kappa\) 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

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

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 \(\beta\) is identified as the unique minimizer of \(H_0\) in \(B\).

**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)** \(\beta\) lies in the interior of a compact parameter space \(B\).
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 $m_K$-times continuously differentiable (where $m_K \geq k+2$); 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 $\beta$. (v) For all $x_i, \mathbb{E}[\tilde{\psi}(Z_i, Z_j; b)] X_i = x_i, X_j = x_j f(x_j)$ is twice continuously differentiable in $x_j$ around $x_j = x_i$, where

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

KF1 (Kernel Function for Estimating Propensity Scores) (i) $K$ 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 $\tilde{K}$ is $m_K$. (iii) $\|t\|\hat{\tilde{K}}(t)$ is Lipschitz continuous for $0 \leq l \leq m_K$.

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

FM1 (Finiteness) $\mathbb{E}\{[C(X_i, X_j) - (V_i + V_j)/2]^2\} \text{ and } \mathbb{E}\{[D(X_i, X_j) - (V_i + V_j)/2]^2\}$ are finite, where $C(X_i, X_j) \equiv \text{inf}_{b \in B}(X_i + X_j)^b/2$ and $D(X_i, X_j) \equiv \text{sup}_{b \in B}(X_i + X_j)^b/2$.

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

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

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 $H_n$ 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 $p(z)$ 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 $b \in B$, 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 $B$ 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 $\rho_\sigma$ 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. $1/\sqrt{n} \sigma_{\hat{p}}^{k-1} > \sigma_{\sigma}^{m\kappa}$); (b) the resulted rates of uniform converge of $\hat{p}$ is faster than $n^{-1/2}$ (i.e. $1/\sqrt{n} \sigma_{\hat{p}}^{k-1} < n^{-1/4}$); and (c) the order of $\sigma_{\sigma}^{m\kappa}$ 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 $\beta$ 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 $\sigma_n$, 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 $\beta$. (ii) For all $x$ and $x'$, $\varphi^-(x, x'; b)$ and $\varphi^+(x, x'; b)$ are twice continuously differentiable in $b$ in an open neighborhood around $\beta$ where for $\hat{\diamond} \in \{+,-\}$,

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

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

$$
\varphi^{\hat{\diamond}}(x, x'; b) \equiv \mathbb{E} \left[ |\Delta \varphi^{\hat{\diamond}}_{i,j}(b) | X_i = x, X_j = x' \right].
$$
For all \( x' \), \( \nabla_b \omega^-(x, x'; \beta) f(x) \) and \( \nabla_b \omega^+(x, x'; \beta) f(x) \) 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 \( \beta, \tilde{\mu}^-(z, x; b) f(x') \) and \( \tilde{\mu}^+(z, x; b) f(x') \) are \( m_\varphi \)-times continuously differentiable with respect to \( X' \) around \( X' = x \), where for \( \diamond \in \{+, -\} \)

\[
\tilde{\mu}^\diamond(z, x; b) \equiv E[\kappa'_0(W_{i,j}) \Delta \varphi_{i,j}^\diamond(b) | Z_i = z, X_j = x'].
\]

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

\[
m^\diamond_\varphi(z; b) \equiv \nabla w(z) f(x) \tilde{\mu}^\diamond(z, x; b), \quad \text{with} \quad \nabla w(z) \equiv [1/f(z), -p(z)f(z)/f(z)^2].
\]

Besides, \( \nabla_b m^*_\varphi(z; \beta) f(z) \) and \( \nabla_b m^*_\varphi(z; \beta) f(z) \) are both \( m_\kappa \)-times continuously differentiable with bounded derivatives in \( z \) over its full support.

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

\[
\int \sup_{b \in \mathcal{N}(\beta)} \| \nabla_b \varphi^\diamond(x, x; b) \| f(x) dF(x) < \infty \quad \text{and} \quad \int \sup_{b \in \mathcal{N}(\beta)} \| \nabla_b \omega^\diamond(x, x; b) \| f(x) dF(x) < \infty,
\]

for \( \diamond \in \{+, -\} \). (ii) For \( \diamond \in \{+, -\} \), \( \int \| \nabla_b m^\diamond_\varphi(z; \beta) f(z) \| < \infty \) and there exists \( \tilde{\varepsilon} > 0 \) such that

\[
E[\sup_{\| \varphi \| \geq \tilde{\varepsilon}} \| \nabla_b m^\diamond_\varphi(Z + \tilde{\varepsilon}; \beta) f(Z + \tilde{\varepsilon}) \|^4] < \infty.
\]

(iii) For \( \diamond \in \{+, -\} \), \( \int \| \nabla_b m^\diamond_\varphi(z; \beta) f(z) \| \, dz < \infty. \)

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

\[
\delta^\diamond_\varphi(y, z) \equiv q \nabla_b m^\diamond_\varphi(z; \beta) f(z) - E[Q \nabla_b m^\diamond_\varphi(Z; \beta) f(Z)].
\]

**Proposition 7** Suppose conditions for Proposition 6 hold. Under additional conditions \( \text{SM2} \) and \( \text{FM2} \),

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

where

\[
\Sigma \equiv \nabla_b H_0(\beta) \quad \text{and} \quad \Omega \equiv 4E[\delta^* Y, Z) \delta^* (Y, Z)].
\]

The proof follows steps similar to (Khan 2001). The continuity of \( H_0 \) under \( \text{SM1-(v)} \) is strengthened to \( \text{SM2-(i)} \), which helps showing that the limiting function \( H_0 \) to have quadratic approximation that is sufficiently precise over an open neighborhood around \( \beta \).
4.4 Root-N Estimation: Close-Form Estimator

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

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

(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 $K$ 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 $K$. 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 $\beta$ 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. This is in part due to the uniform convergence of $\hat{H}_n$ in (18) to $H_0$ over $B$ (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 $\beta$ 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.

\[\text{In the context of set estimators, consistency can be defined using the Hausdorff set metric as in (Manski and Tamer 2002).}\]
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 $4K_1(0) \sum x_i x_i$ and $4K_1(0) \sum x_i v_i$ respectively. That is, if we were to use unpaired observations only (as opposed to pairs) in the summard 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 $\beta$ 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 $\epsilon$ are scalar variables. In the first three designs, the triplet $(X, V, \epsilon)$ are mutually independent, and we experiment with three sets of parametric specifications of marginal distributions for $(X, \epsilon)$, 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 $\alpha$ and $\beta$ 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 $K$ 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 statistics 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): $X \sim \text{Normal}(0,1)$, $\epsilon \sim \text{Normal}(0,1)$

<table>
<thead>
<tr>
<th></th>
<th>$N = 50$</th>
<th>$N = 100$</th>
<th>$N = 200$</th>
<th>$N = 400$</th>
<th>$N = 800$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.0366</td>
<td>0.0376</td>
<td>0.0352</td>
<td>0.0254</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.0160</td>
<td>0.0079</td>
<td>-0.0048</td>
<td>-0.0047</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>-0.0448</td>
<td>-0.0356</td>
<td>-0.0423</td>
<td>-0.0456</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>-0.1490</td>
<td>-0.1417</td>
<td>-0.1408</td>
<td>-0.1375</td>
</tr>
<tr>
<td><strong>Std</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.4532</td>
<td>0.3045</td>
<td>0.2210</td>
<td>0.1511</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.5268</td>
<td>0.3620</td>
<td>0.2479</td>
<td>0.1730</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>0.1912</td>
<td>0.1361</td>
<td>0.0967</td>
<td>0.0681</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.1832</td>
<td>0.1287</td>
<td>0.0922</td>
<td>0.0671</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.4544</td>
<td>0.3066</td>
<td>0.2238</td>
<td>0.1532</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.5268</td>
<td>0.3619</td>
<td>0.2479</td>
<td>0.1730</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>0.1962</td>
<td>0.1406</td>
<td>0.1058</td>
<td>0.0819</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.2360</td>
<td>0.1914</td>
<td>0.1683</td>
<td>0.1530</td>
</tr>
<tr>
<td><strong>MAD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.3095</td>
<td>0.2092</td>
<td>0.1542</td>
<td>0.1034</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.3239</td>
<td>0.2349</td>
<td>0.1635</td>
<td>0.1238</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>0.1301</td>
<td>0.0953</td>
<td>0.0724</td>
<td>0.0583</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.1728</td>
<td>0.1497</td>
<td>0.1395</td>
<td>0.1401</td>
</tr>
</tbody>
</table>
Table 1(b): $X^\sim \text{Laplace}(0, 1)$, $\epsilon^\sim \text{Laplace}(0, 1)$

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

Table 1(c): $X^\sim \text{Logistic}(0, 1)$, $\epsilon^\sim \text{Logistic}(0, 1)$

<table>
<thead>
<tr>
<th></th>
<th>$N = 50$</th>
<th>$N = 100$</th>
<th>$N = 200$</th>
<th>$N = 400$</th>
<th>$N = 800$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>-0.0150</td>
<td>-0.0102</td>
<td>0.0027</td>
<td>0.0055</td>
<td>0.0076</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.0024</td>
<td>-0.0063</td>
<td>-0.0062</td>
<td>0.0044</td>
<td>0.0022</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>-0.0848</td>
<td>-0.0864</td>
<td>-0.0817</td>
<td>-0.0815</td>
<td>-0.0804</td>
</tr>
<tr>
<td>$\beta$</td>
<td>-0.2887</td>
<td>-0.2742</td>
<td>-0.2669</td>
<td>-0.2582</td>
<td>-0.2469</td>
</tr>
<tr>
<td><strong>Std</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.6120</td>
<td>0.4008</td>
<td>0.2722</td>
<td>0.1921</td>
<td>0.1383</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.4989</td>
<td>0.3489</td>
<td>0.2370</td>
<td>0.1641</td>
<td>0.1201</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.2312</td>
<td>0.1660</td>
<td>0.1279</td>
<td>0.0942</td>
<td>0.0746</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.1079</td>
<td>0.0776</td>
<td>0.0577</td>
<td>0.0432</td>
<td>0.0332</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.6118</td>
<td>0.4008</td>
<td>0.2721</td>
<td>0.1920</td>
<td>0.1385</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.4987</td>
<td>0.3488</td>
<td>0.2370</td>
<td>0.1640</td>
<td>0.1200</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.2462</td>
<td>0.1871</td>
<td>0.1517</td>
<td>0.1246</td>
<td>0.1097</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.3082</td>
<td>0.2849</td>
<td>0.2731</td>
<td>0.2618</td>
<td>0.2491</td>
</tr>
<tr>
<td><strong>MAD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.3364</td>
<td>0.2607</td>
<td>0.1737</td>
<td>0.1279</td>
<td>0.0928</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.2726</td>
<td>0.2123</td>
<td>0.1567</td>
<td>0.1052</td>
<td>0.0802</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.1638</td>
<td>0.1298</td>
<td>0.1023</td>
<td>0.0920</td>
<td>0.0829</td>
</tr>
<tr>
<td>$\beta$</td>
<td>0.2909</td>
<td>0.2763</td>
<td>0.2688</td>
<td>0.2596</td>
<td>0.2482</td>
</tr>
</tbody>
</table>
Table 1(d): Heteroskedastic Design with $X \sim \text{Normal}(0, 1)$, $\epsilon \sim \text{Normal}(0, 1)$

<table>
<thead>
<tr>
<th></th>
<th>$N = 50$</th>
<th>$N = 100$</th>
<th>$N = 200$</th>
<th>$N = 400$</th>
<th>$N = 800$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Bias</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.0224</td>
<td>0.0186</td>
<td>0.0193</td>
<td>0.0167</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>-0.1169</td>
<td>-0.0867</td>
<td>-0.0748</td>
<td>-0.0646</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>-0.0638</td>
<td>-0.0674</td>
<td>-0.0642</td>
<td>-0.0675</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>-0.2842</td>
<td>-0.2679</td>
<td>-0.2591</td>
<td>-0.2564</td>
</tr>
<tr>
<td><strong>Std</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.5415</td>
<td>0.3483</td>
<td>0.2381</td>
<td>0.1638</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.5596</td>
<td>0.3707</td>
<td>0.2815</td>
<td>0.2087</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>0.2379</td>
<td>0.1758</td>
<td>0.1324</td>
<td>0.0979</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.2579</td>
<td>0.1803</td>
<td>0.1395</td>
<td>0.1024</td>
</tr>
<tr>
<td><strong>RMSE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.5417</td>
<td>0.3486</td>
<td>0.2388</td>
<td>0.1646</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.5714</td>
<td>0.3805</td>
<td>0.2911</td>
<td>0.2184</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>0.2462</td>
<td>0.1882</td>
<td>0.1471</td>
<td>0.1188</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.3836</td>
<td>0.3228</td>
<td>0.2942</td>
<td>0.2761</td>
</tr>
<tr>
<td><strong>MAD</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pairwise</td>
<td>$\alpha$</td>
<td>0.3224</td>
<td>0.2258</td>
<td>0.1607</td>
<td>0.1084</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.3633</td>
<td>0.2482</td>
<td>0.1981</td>
<td>0.1413</td>
</tr>
<tr>
<td>Inverse-DW</td>
<td>$\alpha$</td>
<td>0.1597</td>
<td>0.1270</td>
<td>0.0990</td>
<td>0.0838</td>
</tr>
<tr>
<td></td>
<td>$\beta$</td>
<td>0.2889</td>
<td>0.2696</td>
<td>0.2615</td>
<td>0.2568</td>
</tr>
</tbody>
</table>

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 $\alpha_1, \beta_1$) and the inverse-density-weighted estimators (labeled $\alpha_2, \beta_2$) 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 \( G_{t|X} \in \Theta \) (where \( \Theta \) here in Section 3.1 is the set of conditional distributions that satisfy CI and MI), we have \( \int 1 (\epsilon \leq xb - v) dG_{t|X} > 1/2 \), which implies \( (x, v) \in \xi(b, G_{t|X}) \). Therefore, \( \Pr \{ Z \in \xi(b, G_{t|X}) \} > 0 \) for such a \( b \) and all \( G_{t|X} \in \Theta \). Since \((Z, \tilde{Z})\) is a pair of states drawn independently from the same marginal, this also implies \( \Pr \{ (Z, \tilde{Z}) \in \xi(b, G_{t|X}) \} > 0 \) for such a \( b \) and all \( G_{t|X} \in \Theta \). Thus \( \beta \) is identified relative to \( b \).

(Necessity) Consider some \( b \neq \beta \). Suppose \( \Pr \{ Z \in Q_b \} = 0 \) so that \( \text{sign}(V - X \beta) = \text{sign}(V - Xb) \) with probability one. Construct a \( \hat{G}_{t|X} \) so that \( \hat{G}_{t|X}(t; b) = \mathbb{E}(Y|x, V = Xb - t) \) for all \( t \) on the support of \( V - Xb \) given \( x \). For \( t \) outside the support of \( V - Xb \) given \( x \), define \( \hat{G}_{t|X}(t; b) \) arbitrarily subject to the requirement that \( \hat{G}_{t|X}(t; b) \) is monotone in \( t \) over the support \( \Omega_{t|X} \). By construction, \( \hat{G}_{t|X}(xb - v; b) = \mathbb{E}(Y|x, V = v) = p(z) \) for all \( z \equiv (x, v) \). If \( xb \in \Omega_{t|X} \), then \( \hat{G}_{t|X}(0; b) = 1/2 \) by construction. Otherwise (i.e. zero is outside the support of \( V - Xb \) given \( x \)), construct \( \hat{G}_{t|X}(.; b) \) outside the support of \( V - Xb \) given \( x \) subject to the requirement that \( \hat{G}_{t|X}(0; b) = 1/2 \). 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 \( G_{t|X} \in \Theta \) satisfying CI and MI such that \( \Pr \{ Z \in \xi(b, G_{t|X}) \} = 0 \). 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 “\( \text{sign}(X \beta - V) = \text{sign}(Xb - V) \) and \( \text{sign}(X' \beta - \tilde{V}) = \text{sign}(Xb' - \tilde{V}) \)” with probability one. Thus the distribution \( \hat{G}_{t|X} \) constructed as above is in \( \Theta \) and also satisfies \( \Pr \{ (Z, \tilde{Z}) \in \xi(b, \hat{G}_{t|X}) \} = 0 \). Thus \( \beta \) is not identified relative to \( b \). \( \text{Q.E.D.} \)

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 \( xb > v > x\beta \). 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 \( b = \beta \). \( \text{Q.E.D.} \)

Proof of Proposition 3. Proposition 1 shows \( \beta \) is identified relative to \( b \) under CI and MI whenever (i) holds. It follows immediately that (i) also implies identification of \( \beta \) relative to \( b \) under the stronger assumptions of CI and CS. To see how (ii) is also sufficient
for identification of $\beta$ relative to $b$, define $\tilde{Q}_{b,S} \equiv \{(z, \tilde{z}) : \tilde{x} = x$ 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

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

(21)

or

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

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 "$\delta$-expansion" of $\tilde{Q}_{b,S}$ as:

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

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 $\tilde{Q}_{b,S}^\delta$ and all $G_{\epsilon|x} \in \Theta_{CS}$. Thus with such a small $\delta$, we have $\tilde{Q}_{b,S}^\delta \subseteq \tilde{\xi}(b, G_{\epsilon|x})$ for all $G_{\epsilon|x} \in \Theta_{CS}$. Finally, suppose condition (ii) in Proposition 3 holds for some $b \neq \beta$ and a set $\omega$ open in $\Omega_X$. Then CT implies

$$\int 1\{(v_i, v_j) \in R_b(x)\}dF_{\epsilon|x}(v_j)dF_{\epsilon|x}(v_i) > 0$$

(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 $(\tilde{X}, X)$ on the left-hand side of (22) then implies $\Pr\{(Z, \tilde{Z}) \in \tilde{Q}_{b,S}^\delta\} > 0$ 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 $\beta$ is identified relative to $b$. The necessity of these two conditions for identifying $\beta$ 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, $\beta$ is identified relative to all $b \neq \beta$. With $\mu$ consisting of counting measure for $y \in \{0, 1\}$ and probability measure for $Z$, we can show path-wise information for $\beta_k$ under a path $\lambda \in \Lambda$ (denoted by $I_{\lambda,k}$) takes the form

$$4 \int \left(\psi_k - \alpha_k^* \lambda^* \psi_\lambda - \sum_{j \neq k} \alpha_j^* \psi_j\right)^2 d\mu = 4 \int_{\Omega_Z} \frac{F_{\epsilon|x}(w)\left(x_k - \sum_{j \neq k} \alpha_j^* x_j\right) - \alpha_k^* \lambda_d(w, x, \delta_0)^2}{F_{\epsilon|x}(w)\left[1 - F_{\epsilon|x}(w)\right]}dF_Z$$

(23)

where $(\alpha_j^*)_{j \neq k}$ and $\alpha_k^*$ 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 $\alpha_k^*$ must be nonzero for such a $\lambda$, because otherwise the path-wise
information $I_{\lambda,k}$ would equal the Fisher information for $\beta$ in a parametric model where the true error distribution $F_{e|x}$ 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 $x_0$ in its interior for all $x$. Thus there exists an open interval $(-\varepsilon^*, \varepsilon^*)$ such that $W \equiv X - V$ 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

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

Differentiating both sides of (14) with respect to $\delta$ at $\delta_0$ suggests $\lambda_\delta(-\varepsilon, x; \delta_0) = -\lambda_\delta(\varepsilon, x; \delta_0)$ for all $x$ and $\varepsilon$. 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 $f_{e|x}(w) \left( x_k - \sum_{j \neq k} \alpha^*_j x_j \right)$ is even in $w$ (i.e. symmetric in $w$ around 0) given any $x$. Due to CS', $F_{e|x}(t)^{-1} \left[ 1 - F_{e|x}(t) \right]^{-1}$ is uniformly bounded below by $\varepsilon$ for all $t \in (-\varepsilon^*, \varepsilon^*)$ and $x \in \Omega_X$. It follows that for any constant $\varphi > 0$,

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

Thus for any $\varphi > 0$, there exist $T \subset [0, \varepsilon^*) \otimes \Omega_X$ or $T \subset (-\varepsilon^*, 0] \otimes \Omega_X$ with $\Pr\{(W, X) \in T\} > 0$ and

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

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

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 $f_{e|\bar{x}}(t) \left( \bar{x}_k - \sum_{j \neq k} \alpha^*_j \bar{x}_j \right)$ is positive and bounded below by $a(\bar{x})c > 0$ for all $t$ such that $(t, \bar{x}) \in T$. 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 $(t, \bar{x}) \in T$. By symmetry of $f_{e|\bar{x}}$ and oddness of $\lambda_\delta(t, x; \delta_0)$ in $t$ given any $x$, $f_{e|\bar{x}}(-t) \left( \bar{x}_k - \sum_{j \neq k} \alpha^*_j \bar{x}_j \right) - \alpha^*_\lambda \lambda_\delta(-t, \bar{x}; \delta_0) \geq \frac{3}{2} a(\bar{x})c > 0$ for all $t$ with $(t, \bar{x}) \in T$. 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 $(t, x) \in T$. Due to SV, $\Pr\{(W, X) \in \Omega^-\} > 0$ where $\Omega^- \equiv \{(t, x) : (-t, x) \in T\}$. Thus $f_{e|x}(t) \left( x_k - \sum_{j \neq k} \alpha^*_j x_j \right) - \alpha^*_\lambda \lambda_\delta(t, x; \delta_0)$ is bounded away from zero by some positive constant over $\Omega^-$. It then follows that

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

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 \( F_{j|X} \) 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 \( \lim_{t \to \infty} t F_W(t) = 0 \). Then for any \( c \),

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

and \( \mathbb{E}[(W - c) 1(W > c)] = \mathbb{E}(W - c) - \mathbb{E}[(W - c) 1(W < c)] = \mu_W - c + \int_{-\infty}^{c} F_W(w) \, dw \).

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

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

(25)

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

(26)

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

\[
x\beta + v_L + \int_{-v_H}^{-v_L} F_{X\beta+\epsilon|X=x}(s) \, ds.
\]

(27)

Suppose \( F_{j|X} \) 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:

\[
H_n(z_i, z_j; b) = \frac{1}{n(n-1)} \sum_{j \neq i} K_h(x_i - x_j) \left[ \kappa(w_{i,j} - 1) \varphi^-(z_i, z_j; b) + \kappa(1-w_{i,j}) \varphi^+(z_i, z_j; b) \right]
\]
where \( w_{i,j} \) is the sum of the true propensity scores (i.e. \( w_{i,j} \equiv p_i + p_j \) with \( p_i \equiv p(z_i) \)).

**Proof of Proposition 5.** Consider any \( b \neq \beta \). Under FR, \( \text{Pr}(X\beta - Xb \neq 0) > 0 \).

Without loss of generality, suppose \( \text{Pr}(X\beta - Xb > 0) > 0 \) and let \( \omega \equiv \{x : x\beta > xb\} \). Then under SV,

\[
\int 1\{2x\beta > v_i + v_j > 2xb\}dF_{v_i,v_j}(v_i, v_j) > 0
\]

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 \( \kappa \) in WF imply that:

\[
\begin{align*}
\mathbb{E} \left[ \kappa(W_{i,j} - 1)\varphi^-(Z_i, Z_j; b) + \kappa(1 - W_{i,j})\varphi^+(Z_i, Z_j; b) \mid X_j = X_i = x \right] \\
\geq \mathbb{E} \left[ \kappa(W_{i,j} - 1)\varphi^-(Z_i, Z_j; b) \mid V_i + V_j \leq 2x\beta, X_j = X_i = x \right] \text{Pr}(V_i + V_j \leq 2x\beta \mid X_j = X_i = x) > (28)
\end{align*}
\]

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 \( \omega \)) 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 \( \text{Pr}(X\beta < Xb) > 0 \), then for any \( x \) with \( x\beta < xb \), SV implies

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

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

Next, consider \( b = \beta \). For any \( x \),

\[
H_0(\beta) = \mathbb{E} \left\{ f(x)\mathbb{E} \left[ \kappa(W_{i,j} - 1)\varphi^-(Z_i, Z_j; \beta) + \kappa(1 - W_{i,j})\varphi^+(Z_i, Z_j; \beta) \mid X_j = X_i = x \right] \right\}
\]

\[
= \mathbb{E} \left\{ f(x)\mathbb{E} \left[ \kappa(W_{i,j} - 1)\varphi^-(Z_i, Z_j; \beta) \mid W_{i,j} \geq 1, X_j = X_i = X \right] \right\} \text{Pr}(W_{i,j} \geq 0 \mid X_j = X_i = X) \quad (29)
\]

\[
+ \mathbb{E} \left\{ f(x)\mathbb{E} \left[ \kappa(1 - W_{i,j})\varphi^+(Z_i, Z_j; \beta) \mid W_{i,j} < 1, X_j = X_i = X \right] \right\} \text{Pr}(W_{i,j} < 0 \mid X_j = X_i = X).
\]

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

\[
\sup_{b \in B} |\hat{H}_n(b) - H_n(b)| = o_p(1).
\]

(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 KP2, we have:

\[
\sup_{b \in B} \left| \frac{1}{n(n-1)} \sum_{j \neq i} K_h(x_i - x_j) \varphi^-_{i,j}(b) [\kappa(\hat{w}_{i,j} - 1) - \kappa(w_{i,j} - 1) - k'(w_{i,j} - \hat{w}_{i,j} - w_{i,j})] \right|
\]

\[
= \sup_{b \in B} \left| \frac{1}{n(n-1)} \sum_{j \neq i} K_h(x_i - x_j) \varphi^-_{i,j}(b) \kappa''(\hat{w}_{i,j} - 1) \|\hat{w}_{i,j} - w_{i,j}\|^2 \right|
\]

\[
\leq \alpha \sup_z \|\bar{p}(z) - p(z)\|^2 \sup_{b \in B} \left\{ \frac{1}{n(n-1)} \sum_{j \neq i} K_h(x_i - x_j) \varphi^-_{i,j}(b) \kappa''(\hat{w}_{i,j}) \right\}
\]

(31)
where κ', κ'' are first- and second-order derivatives of κ; \( \tilde{w}_{i,j} \) is a random variable between \( w_{i,j} \) and \( \bar{w}_{i,j} \); and \( a > 0 \) is some finite constant. Under KF2-(iii), FM1-(i) and W, 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, \( \sup_z |\hat{p}(z) - p(z)| = O_p\left( \frac{\log n}{\sqrt{n \sigma^k + (k + 1) \sigma_m}} \right) \) 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:

\[
\sup_{b \in B} \left| \frac{1}{n(n-1)} \sum_{j \neq i} K_h(x_i - x_j) \varphi_{i,j}^-(b) \kappa'(w_{i,j} - 1)(\tilde{w}_{i,j} - w_{i,j}) \right| \\
\leq 2 \sup_z \|\hat{p}(z) - p(z)\| \sup_{b \in B} \left\{ \frac{1}{n(n-1)} \sum_{j \neq i} \left| K_h(x_i - x_j) \varphi_{i,j}^-(b) \kappa'(w_{i,j} - 1) \right| \right\}.
\]

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

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

where

\[
\begin{align*}
g_n(z_i, z_j; b) &= K_h(x_i - x_j) \left[ \kappa(w_{i,j} - 1) \varphi_{i,j}^- + \kappa'(w_{i,j} - 1)(\tilde{w}_{i,j} - w_{i,j}) \right] ; \\
g_n,1(z_i; b) &= \mathbb{E}[g_n(Z_i, Z'_i; b) | Z = z_i] + \mathbb{E}[g_n(Z_i, Z'_i; b) | Z' = z_i] - 2\mathbb{E}[g_n(Z_i, Z'_i; b)]; \\
g_n,2(z_i, z_j; b) &= g_n(z_i, z_j; b) - \mathbb{E}[g_n(Z_i, Z'_i; b) | Z = z_i] - \mathbb{E}[g_n(Z_i, Z'_i; b) | Z' = z_i] + \mathbb{E}[g_n(Z_i, Z'_i; b)].
\end{align*}
\]

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

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

for all \((z_i, z_j)\), where \( C(.) \) and \( D(.) \) are defined in FM1 and \( \alpha' > 0 \) is some finite constant. By arguments as in (Pakes and Pollard 1989), the class of functions:

\[
\{ h^k_{n} g_n(z_i, z_j; b) : b \in B \}
\]

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, \( \mathbb{E}[\sup_{b \in B} h^2_{n} g_n(Z_i, Z_j; b)^2] = O(1) \) 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 \( O_p(n^{-1/2} h^{-k}_{n}) \) and \( O_p(n^{-1} h^{-k}_{n}) \) uniformly over \( b \in B \) 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

\[
\varphi(z_i, z_j; b) \equiv \kappa(w_{i,j} - 1) \varphi_{i,j}^- + \kappa(1 - w_{i,j}) \varphi_{i,j}^+ (b)
\]
to facilitate derivations. By definition,

\[ \mathbb{E}[g_n(Z_i, Z_j; b)] = \int K_h(x_i - x_j) \tilde{\varphi}(z_i, z_j; b) dF(z_i, z_j) \]

\[ = \int K_h(x_i - x_j) \mathbb{E}[\tilde{\varphi}(Z_i, Z_j; b) | x_i, x_j] dF(x_i, x_j) \]

\[ = \int K(u) \mathbb{E}[\tilde{\varphi}(Z_i, Z_j; b) | X_i = x_i, X_j = x_i + h_n^k u] f(x_i + h_n^k u) dudF(x_i) \]

Changing variables between \( x_j \) and \( u = (x_j - x_i)/h_n^k \) 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 \( b \in B \). Thus the sum of the three terms on the right-hand side of (32) is \( o_p(1) \) uniformly over \( b \in B \).

Combine this result with (30), we get:

\[ \sup_{b \in B} |\hat{H}_n(b) - H_0(b)| = o_p(1). \quad (33) \]

The limiting function \( H_0(b) \) is continuous under SM1 in an open neighborhood around \( \beta \). Besides, Proposition 5 has established that \( H_0(b) \) is uniquely minimized at \( \beta \). 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:

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

By construction, the optimizers of \( \hat{\mathcal{H}}_n \) 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 \( \hat{\mathcal{H}}_n(.) \) locally in a neighborhood of \( \beta \) 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 \( \beta \) 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 $\beta$ 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 $\beta$; and

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

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

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 $\beta$ 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 $\beta$.

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

$$H_0(b) = \frac{1}{2}(b - \beta)' \nabla b H_0(\tilde{b})(b - \beta)$$

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

To simplify notations in what follows, we let

$$\hat{H}_{1,n}(b) \equiv \frac{1}{n(n-1)} \sum_{j \neq i} K_h(x_i - x_j) \kappa(\hat{w}_{i,j} - 1)[\varphi_{i,j}^{-}(b) - \varphi_{i,j}^{-}(\beta)]$$

denote the first half of the “location-normalized” objective function $\hat{H}_n$ (which only involve $(.)_-$); and likewise let $H_{1,n}(b)$ and $H_{1,0}(b)$ denote the first halves of $H_n$ and $H_0$ respectively. Similarly, define $\hat{H}_{2,n}$, $H_{2,n}$ 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,

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

uniformly over an $o_p(1)$ neighborhood around $\beta$ in $B$; 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 $\beta$.

**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:

$$\{E[\tilde{g}_n(Z_i, Z_j; b)] - H_{1,0}(b)\} + \frac{1}{n} \sum_{i=1}^n \tilde{g}_{n,1}(z_i; b) + \frac{1}{n(n-1)} \sum_{j \neq i} \tilde{g}_{n,2}(z_i, z_j; b) \quad (35)$$
where
\[
\hat{g}_n(z_i, z_j; b) = K_h(x_i - x_j) \kappa^-(z_i, z_j) \left[ \varphi^-(z_i, z_j; b) - \varphi^-(z_i, z_j; \beta) \right];
\]
\[
g_{n,1}(z_i; b) = \mathbb{E}\left[\hat{g}_n(Z, Z'; b) | Z = z_i \right] + \mathbb{E}\left[\hat{g}_n(Z, Z'; b) | Z' = z_i \right] - 2\mathbb{E}\left[\hat{g}_n(Z, Z'; b) \right];
\]
\[
g_{n,2}(z_i, z_j; b) = \hat{g}_n(z_i, z_j; b) - \mathbb{E}\left[\hat{g}_n(Z, Z'; b) | Z = z_i \right] - \mathbb{E}\left[\hat{g}_n(Z, Z'; b) | Z' = z_j \right] + \mathbb{E}\left[\hat{g}_n(Z, Z'; b) \right].
\]

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 \( b \):
\[
\mathbb{E}\left[\hat{g}_n(Z_i, Z_j; b) \right] = \int K_h(x_i - x_j) \varphi(x_i, x_j; b) dF(x_i, x_j)
\]
where
\[
\varphi(x, x'; b) \equiv \mathbb{E}\left\{ \kappa^-(Z_i, Z_j) \Delta \varphi^-(Z_i, Z_j; b) | X_i = x, X_j = x' \right\}.
\]

By construction, \( \varphi(x_i, x_j; \beta) = 0 \) and under SM2-(ii),
\[
\varphi(x_i, x_j; b) = \nabla_b \varphi(x_i, x_j; \beta)(b - \beta) + \frac{1}{2}(b - \beta)^t \nabla_{bb} \varphi(x_i, x_j; \beta)(b - \beta) + o(\|b - \beta\|^2)
\]

for all \( b \) in an \( o(1) \) neighborhood around \( \beta \); where \( \nabla_b \varphi \) and \( \nabla_{bb} \varphi \) 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 \( \mathbb{E}[\hat{g}_n(Z_i, Z_j; b)] \) as
\[
(b - \beta)^t \left[ \int K_h(x_i - x_j) \frac{1}{2} \nabla_{bb} \varphi(x_i, x_j; \beta) dF(x_i, x_j) \right] (b - \beta)
\]
\[
+ \left\{ \int K_h(x_i - x_j) \nabla_b \varphi(x_i, x_j; \beta) dF(x_i, x_j) \right\} (b - \beta) + o(\|b - \beta\|^2).
\]

for all \( b \) in an \( o(1) \) neighborhood around \( \beta \). Changing variables between \( x_i \) and \( u \equiv (x_i - x_j)/h^k_n \), we can write the square bracket term in (36) as
\[
\int \left[ \int K(u) \frac{1}{2} \nabla_{bb} \varphi(x_j + h^k u, x_j; \beta) f(x_j + h^k u) du \right] dF(x_j)
\]
\[
= \frac{1}{2} \int \left[ \nabla_b \varphi(x, x; \beta) f(x) + O(h^2_n) \right] dF(x) = \frac{1}{2} \int \left[ \nabla_b \varphi(x, x; \beta) f(x) \right] dF(x) + o(1) = \frac{1}{2} \nabla_b H_{1,0}(\beta) + o(1).
\]

The first equality above follows from a Taylor expansion of \( \nabla_{bb} \varphi(x, x; \beta) f(x) \) 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 \( O(h^2_n) \) 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
\[
\int \left[ \int K(u) \nabla_b \varphi(x_j + h^k u, x_j; \beta) f(x_j + h^k u) du \right] dF(x_j)
\]
\[
= \int \left[ \nabla_b \varphi(x, x; \beta) f(x) + O(h^m_n) \right] dF(x) = \int \left[ \nabla_b \varphi(x, x; \beta) f(x) \right] dF(x) + o(n^{-1/2})
\]
where \( \int [\nabla \hat{\varphi}(x, x; \beta) f(x)] dF(x) = \nabla H_{1,0}(\beta) = 0 \) 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

\[
\frac{1}{2} (b - \beta)' \nabla H_{1,0}(\beta)(b - \beta) + o_p(\|b - \beta\| / \sqrt{n}) + o_p(\|b - \beta\|^2).
\]

By standard Taylor expansion using SM-(i), \( H_{1,0}(b) = \frac{1}{2} (b - \beta)' \nabla H_{1,0}(\beta)(b - \beta) + o_p(\|b - \beta\|^2) \) over an \( o(1) \) neighborhood of \( \beta \), 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 \( \beta \) to the second term in (35) to get

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

where \( \tilde{b} \) is on the line segment between \( b \) and \( \beta \) (and possibly depends on \( z_i \)). By construction, \( \hat{g}_{n,1}(z_i; \beta) = 0 \) 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 \( b \):

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

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,

\[
\nabla_b \mathbb{E}[\tilde{g}_{n}(Z, Z'; b)|Z = z] = \int K_1(p - x') \nabla_b \hat{\varphi}(z, x'; b) f(x') dx' = \int K(u) \nabla_b \hat{\varphi}(z, x - h^k u; b) f(x - h^k u) du
\]

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 \( \nabla_b \mathbb{E}[\tilde{g}_{n}(Z, b)|Z' = z] \).

It then follows from boundedness of \( K \) in \( \text{KF} \) and the finite moment condition in \( \text{FM2} \) that \( \mathbb{E}[\nabla_{b} \tilde{g}_{n,1}(Z; b) \nabla_{b} \tilde{g}_{n,1}(Z; b)'] = O(1) \) for all \( b \) in an open neighborhood around \( \beta \). Thus for any fixed \( b \) in an open neighborhood around \( \beta \), the Lapunov 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 \( \text{FM1,2} \). With \( \tilde{b} \) between \( b \) and \( \beta \), and with \( b \xrightarrow{p} \beta \), an application of Lemma 2.17 in (Pakes and Pollard 1989) shows \( \frac{1}{n} \sum_{i=1}^{n} \nabla_{b} \tilde{g}_{n,1}(z_i; \tilde{b}) = o_p(n^{-1/2}) \) uniformly over an \( o_p(1) \) neighborhood around \( \beta \). Thus the second term in the decomposition in (35) is \( o_p(\|b - \beta\| / \sqrt{n}) \) uniformly over an \( o_p(1) \) neighborhood of \( \beta \).

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 \( \beta \). Furthermore, this term is reduced to \( O_p(n^{-3/2} h^{-3k/2}) \) uniformly over an \( O_p(1/\sqrt{n h^k}) \) neighborhood around \( \beta \), which is \( o_p(n^{-1}) \) due to our choice of bandwidth in \( \text{BW2} \). Q.E.D.

Next, we show the difference between \( \tilde{\mathcal{H}}_{1,n}(b) \) 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 \( \beta \) in \( \mathcal{B} \).
Lemma C3. Suppose conditions for Proposition 6 hold. Under additional conditions in SM2 and FM2,

\[ |\hat{\mathcal{H}}_{1,n}(b) - \mathcal{H}_{1,n}(b)| = \frac{2}{n} \sum_{i=1}^{n} \delta_i(y_i, z_i)(b - \beta) + o_p(\|b - \beta\|/\sqrt{n}) + o_p\left(\|b - \beta\|^2\right) + O_p(n^{-1}h^{-k}) \]  

(37)

uniformly over an \(o_p(1)\) neighborhood of \(\beta\) in \(B\); where

\[ \delta_i(y, z) = q \nabla_{b} m^*(z; \beta) f(z) - \mathbb{E}[Q \nabla_{b} m^*(Z; \beta) f(Z)] \]  

with \(q \equiv (y, 1)\); Besides, the \(O_p(n^{-1}h^{-k})\) term in (37) is further reduced to \(o_p(n^{-1})\) uniformly over an \(O_p\left(1/\sqrt{n}h^{-k}\right)\) neighborhood around \(\beta\).

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

\[ \Delta_{1,n} = \frac{1}{n(n-1)} \sum_{j \neq i} K_h (x_i - x_j) \Delta \varphi_{i,j}^{-}(b) \kappa'(w_{i,j} - 1)(\hat{w}_{i,j} - w_{i,j}) \]  

and

\[ R_{1,n} = \frac{1}{n(n-1)} \sum_{j \neq i} K_h (x_i - x_j) \Delta \varphi_{i,j}^{-}(b) \kappa''(\hat{w}_{i,j} - 1)(\hat{w}_{i,j} - w_{i,j})^2 \]  

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

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

(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,

\[ \sup_{z} |\hat{p}(z) - p(z)| = o_p(1) \]  

(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:

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

(40)

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

\[ \varpi_{n,1}(z_i; b) \equiv \mathbb{E}[\varpi_n(Z, Z'; b)|Z = z_i] + \mathbb{E}[\varpi_n(Z, Z'; b)|Z' = z_i] - 2\mathbb{E}[\varpi_n(Z, Z'; b)]; \]  

and

\[ \varpi_{n,2}(z_i, z_j; b) \equiv \varpi_n(z_i, z_j; b) - \mathbb{E}[\varpi_n(Z, Z'; b)|Z = z_i] - \mathbb{E}[\varpi_n(Z, Z'; b)|Z' = z_j] + \mathbb{E}[\varpi_n(Z, Z'; b)]. \]

By standard arguments in (Pakes and Pollard 1989), the class of functions \(\{h^k \varpi_n(z_i, z_j; b) : b \in B\} \) is Euclidean with a constant envelop that has finite second moments. Our conditions in FM1, the boundedness of \(\hat{K}\) 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 \(\delta_n\) and \(\gamma_n\) therein being \(o(1)\) and \(O(1)\) respectively. Hence

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

(41)
uniformly over an $o_p(1)$ neighborhood around $\beta$.

As for the unconditional expectation in (40), by definition, it equals
\[
\mathbb{E}[\varpi_n(Z_i, Z_j; b)] = \int |K_h(x_i - x_j)| \varpi(x_i, x_j; b) dF(x_i, x_j)
\]
where $\varpi(x, x'; b) \equiv \mathbb{E}\{|\Delta \varphi_{i,j}^{-1}(b)| | X_i = x, X_j = x' \}$. By construction, $\varpi(x_i, x_j; \beta) = 0$ and under SM2,
\[
\varpi(x_i, x_j; b) = \nabla_b \varpi(x_i, x_j; \beta)(b - \beta) + o(\|b - \beta\|)
\]
(43) for all $b$ in an $o(1)$ neighborhood around $\beta$; 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
\[
\int |K_h(x_i - x_j)| \nabla_b \varpi(x_i, x_j; \beta) dF(x_i, x_j)
\]
\[= \int \left[ \int |K(u)| \nabla_b \varpi(x_j + h^k u, x_j; \beta) f(x_j + h^k u) du \right] dF(x_j)
\]
\[= \kappa_1 \nabla_b \mathbb{E}[f(X) \varpi(X, X; \beta)] + o(1)
\]
(44)
where $\kappa_1 \equiv \int |K(u)| du$ 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
\[
\nabla_b \mathbb{E}[f(X) \varpi(X, X; \beta)] = 0
\]
(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:
\[
\mathbb{E}[\varpi_n(Z_i, Z_j; b)] = o(\|b - \beta\|).
\]
(46)
Combining results from (39), (41) and (46), we know the order of $|R_{1,n}|$ is bounded above by
\[
o_p(\|b - \beta\|/\sqrt{n}) + o_p(n^{-1}h^{-k}) + o_p(n^{-3/2}h^{-k})
\]
(47) uniformly over an $o_p(1)$ neighborhood of $\beta$. 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 $\beta$, 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 $\gamma_0 \equiv (\gamma_{0,1}, \gamma_{0,2})'$ 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:

\[
\frac{1}{n(n-1)} \sum_{j \neq i} \kappa'(z_i, z_j) K_h(x_i - x_j) \Delta \phi_{i,j}(b)[\hat{p}(z_i) - p(z_i)] = \frac{1}{n(n-1)} \sum_{j \neq i} \kappa'(z_i, z_j) K_h(x_i - x_j) \Delta \phi_{i,j}(b)[\hat{\gamma}_1(z_i)/\hat{\gamma}_2(z_i) - \gamma_{0,1}(z_i)/\gamma_{0,2}(z_i)]
\]

By arguments similar to those apply to the rst term on the right-hand side of (50), we have:

\[
\exp(b R(h)) \leq \sum_{j \neq i} \kappa'(z_i, z_j) K_h(x_i - x_j) \Delta \phi_{i,j}(b) \exp(b \sup_{z} |\hat{\gamma}(z_i) - \gamma_0(z_i)|).
\]

By triangular inequality, we have:

\[
R_{1,n} \leq \left\{ \frac{1}{n(n-1)} \sum_{j \neq i} \left| \kappa'(z_i, z_j) K_h(x_i - x_j) \Delta \phi_{i,j}(b) \exp(b \sup_{z} |\hat{\gamma}(z_i) - \gamma_0(z_i)|) \right| \right\} \sup_{z} |\hat{\gamma}(z_i) - \gamma_0(z_i)|.
\]

By arguments similar to those apply to the rst term in \( R_{1,n} \), the rst 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 \( O(\sigma_n^{m_k}) \) due to Lemma 8.9 in (Newey and McFadden 1994), which is \( o_p(n^{-1/2}) \) by our choice of \( \sigma_n \) in BW1 and smoothness condition in SM1. Thus the order of \( R_{1,n} \) 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 \( \beta \). Again, similar to the case with \( R_{1,n} \), the \( o_p(n^{-1}h^{-k}) \) term in \( R_{1,n} \) above is further reduced to \( o_p(n^{-1}) \) over an \( O_p(n^{-1/2}h^{-k/2}) \) neighborhood around \( \beta \) by a repeated application of Theorem 3 in (Sherman 1994b).

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

\[
\frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq s} \phi_n(\tilde{z}_i, \tilde{z}_j, \tilde{z}_s; b)
\]

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

\[
\phi_n(\tilde{z}_i, \tilde{z}_j, \tilde{z}_s; b) = \kappa'(z_i, z_j) K_h(x_i - x_j) \Delta \phi_{i,j}(b) [\exp(b) \{q \kappa_{\sigma}(z_i - z_s) - \exp(b) \{q \kappa_{\sigma}(z_i - Z_s) \} - \exp(b) \{q \kappa_{\sigma}(z_i - Z_s) \})
\]

with \( \kappa_{\sigma} \) being a shorthand for \( \sigma^{-(k+1)} \kappa(\cdot/\sigma^{k+1}) \); and the expectation is taken w.r.t. \( \tilde{Z}_s \) while \( z_i \) is some realized value of \( Z_i \). Note \( \phi_n \) 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

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

(52)

where

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

(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 "\(\gamma_n\) therein being \(O(1)\) for any sequence of \(\delta_n\) converging to 0. Hence uniformly over an \(o_p(1)\) neighborhood of \(\beta\) 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 \(O_p(h^{-k}n^{-1})\) over an \(o_p(1)\) neighborhood of \(\beta\). 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 \(\beta\), the second-order term is \(O_p(\delta_n^{-1}) = O_p(h^{-k/2}n^{-3/2})\), 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,

\[
\mathbb{E}[\phi_n \left( \tilde{Z}_{i,j,s}; b \right)| \tilde{Z}_i = (y, z)] = \mathbb{E}\left\{ \kappa'(z, Z_j) K_h(\nabla w(Z_i) | Q, K_\sigma(z - Z_s)) - \mathbb{E}[Q, K_\sigma(z - Z_s)] | Z_i = z \right\}
\]

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

\[
\mathbb{E}[\phi_n \left( \tilde{Z}_{i,j,s}; b \right)| \tilde{Z}_j = (y, z)] = \mathbb{E}_{Z_i}\left\{ \kappa'(Z_i, Z_j) \nabla w(Z_i) K_h(X_i - x) \Delta \varphi_{ij}(b) \right\} E_{Z_i} \left\{ Q, K_\sigma(Z_i - Z_s) - \mathbb{E}[Q, K_\sigma(Z_i - Z_s)] | Z_i = z \right\} \]

(54)
where \( q_l \equiv (y_l, 1)' \); and

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

To clarify notations, note on the second line of (54), \( \mathbb{E}[Q'K_\sigma (Z_i - Z')] \) 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 \( o_p(\|b - \beta\|/\sqrt{n}) + o_p(\|b - \beta\|^2) + o_p(n^{-1}) \) uniformly over \( o_p(1) \) neighborhood around \( \beta \).

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

\[
\tilde{m}_n(z_i; b) = \nabla w(z_i) \int \kappa'(z_i, z_j) K_h(x_i - x_j) \Delta \varphi_{i,j}^-(z_i, z_j; b) dF(z_j)
= \nabla w(z_i) \int K_h(x_i - x_j) \tilde{\mu}^- (z_i, x_j; b) dF(x_j)
= \nabla w(z_i) \int K(u) \tilde{\mu}^- (z_i, x_i - h^k u; b) f(x_i - h^k u) du
= \nabla w(z_i) \tilde{\mu}^- (z_i, x_i; b) f(x_i) + O(h_n^m) \]

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

\[
\int \tilde{m}_n(z; b) \left( \frac{1}{n} \sum_{l=1}^n q_l \kappa_{\sigma} (z - z_l) - \int \mathbb{E}[Q'|x']\kappa_{\sigma} (z - z') f(z') dz' \right) dF(z)
= \frac{1}{n} \sum_{l=1}^n \int q_l \left[ m^*_n(z; b) + O(h_n^m) \right] \kappa_{\sigma} (z - z_l) f(z) dz
- \int f(z') \mathbb{E}[Q'|x'] \left( \int \left[ m^*_n(z; b) + O(h_n^m) \right] \kappa_{\sigma} (z - z') f(z) dz \right) dz'.
\]

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:

\[
\frac{1}{n} \sum_{l=1}^n \int q_l m^*_n(z; b) \kappa_{\sigma} (z - z_l) f(z) dz - \int f(z') \mathbb{E}[Q'|x'] \left( \int m^*_n(z; b) \kappa_{\sigma} (z - z') f(z) dz \right) dz' \quad (55)
\]

and

\[
O(h_n^m) \left\{ \frac{1}{n} \sum_{l=1}^n \int q_l \kappa_{\sigma} (z - z_l) f(z) dz - \int f(z') \mathbb{E}[Q'|x'] \left( \int \kappa_{\sigma} (z - z') f(z) dz \right) dz' \right\}. \quad (56)
\]
We first examine the term in (56). Note \( \int \mathcal{K}_\sigma (z - z') f(z) dz = f(z') + O(\sigma_n^{m_\kappa}) \) for all \( z' \) under our conditions, and the term in the braces above can be written as
\[
\frac{1}{n} \sum_{i=1}^{n} \int q_i \mathcal{K}_\sigma (z - z_i) f(z) dz - \int f(z') \mathbb{E} (Q' | x') [f(z')] + O(\sigma_n^{m_\kappa}) dz'
= \frac{1}{n} \sum_{i=1}^{n} \int f(z) q_i \mathcal{K}_\sigma (z - z_i) dz - \mathbb{E} [f(Z) Q] + O(\sigma_n^{m_\kappa})
= \frac{1}{n} \sum_{i=1}^{n} \{ f(z_i) q_i - \mathbb{E} [f(Z) Q] \} + o_p(n^{-1/2}) + O(\sigma_n^{m_\kappa})
\]
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 \( O(\delta_1^{m_\kappa}) \) and \( O(\sigma_n^{m_\kappa}) \) are \( o(n^{-1/2}) \). Note \( \frac{1}{n} \sum_{i=1}^{n} \{ f(z_i) q_i - \mathbb{E} [f(Z) Q] \} \) 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 \( \beta \).

Next, to deal with (55), for any \( z \), we can apply a Taylor expansion of \( m^*_\kappa \) around \( b = \beta \) to get:
\[
m^*_\kappa (z; b) = 0 + \nabla b m^*_\kappa (z; \beta)(b - \beta) + o (\|b - \beta\|) .
\]
Substituting this into (55) above, we decompose it into the sum of
\[
(b - \beta) \left\{ \frac{1}{n} \sum_{i=1}^{n} \int q_i \nabla b m^*_\kappa (z; \beta) \mathcal{K}_\sigma (z - z_i) f(z) dz - \int f(z') \mathbb{E} (Q' | x') \left( \int \nabla b m^*_\kappa (z; \beta) \mathcal{K}_\sigma (z - z') f(z) dz \right) dz' \right\}
\]
and
\[
o (\|b - \beta\|) \left\{ \frac{1}{n} \sum_{i=1}^{n} \int q_i \mathcal{K}_\sigma (z - z_i) f(z) dz - \int f(z') \mathbb{E} (Q' | x') \left( \int \mathcal{K}_\sigma (z - z') f(z) dz \right) dz' \right\}
\]
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:
\[
\int \nabla b m^*_\kappa (z; \beta) f(z) \mathcal{K}_\sigma (z - z') dz = \nabla b m^*_\kappa (z'; \beta) f(z') + O(\sigma_n^{m_\kappa}).
\]
Thus the term in the braces of (57) is
\[
\frac{1}{n} \sum_{i=1}^{n} \int q_i \nabla b m^*_\kappa (z; \beta) f(z) \mathcal{K}_\sigma (z - z_i) dz - \int \mathbb{E} (Q' | x') \nabla b m^*_\kappa (z'; \beta) f(z')^2 dz' + O(\sigma_n^{m_\kappa})
= \left\{ \frac{1}{n} \sum_{i=1}^{n} \int q_i \nabla b m^*_\kappa (z; \beta) f(z) \mathcal{K}_\sigma (z - z_i) dz - \mathbb{E} Q f(z) \nabla b m^*_\kappa (Z; \beta) \right\} + O(\sigma_n^{m_\kappa})
\]
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
\[
\frac{1}{n} \sum_{i=1}^{n} \delta^*_\kappa (y_i, z_i) + o_p(n^{-1/2}) \text{ where } \delta^*_\kappa (y, z) \equiv q \nabla b m^*_\kappa (z; \beta) f(z) - \mathbb{E} [Q \nabla b m^*_\kappa (Z; \beta) f(Z)];
\]
while $O(\sigma_n^2) = o(n^{-1/2})$ under our choice of bandwidth. To sum up, we have shown

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\phi_n(\tilde{Z}_{i,j,k}; b) | \tilde{Z}_n = (y_i, z_i)] = \frac{1}{n} \sum_{i=1}^{n} \delta_+^*(y_i, z_i)(b - \beta) + o_p(\|b - \beta\|/\sqrt{n}) + o_p(n^{-1})$$

uniformly over an $o_p(1)$ neighborhood around $\beta$ in $\mathcal{B}$, where $\delta_+^*$ 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 $|\hat{H}_{2,n}(b) - \mathcal{H}_{2,n}(b)|$, except that $\delta_-^*$ needs to be replaced by

$$\delta_-^*(y, z) \equiv q \nabla w_m^+ (z; \beta) f(z) - \mathbb{E}[Q f(Z) \nabla w_m^+(Z; \beta)]$$

where

$$m^+_\beta(z) \equiv \nabla w(z) f(x) \hat{p}^+(z, x; b); \quad \hat{p}^+(z_i, x_j; b) \equiv \mathbb{E}[\kappa'(Z_i, Z_j) \Delta \varphi^+(Z_i, Z_j; b) | Z_i = z_i, X_j = x_j].$$

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,

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

uniformly over an $o_p(1)$ neighborhood around $\beta$ 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 $\beta$ at a rate of $1/\sqrt{n}h^k$. 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 $\|\hat{\beta} - \beta\| = O_p(n^{-1/2})$.

Recall that by a second-order Taylor expansion, $H_0(b) = \frac{1}{2}(b - \beta)\nabla \varphi_0 H_0(\beta)(b - \beta) + o_p(\|b - \beta\|^2)$ over an $o(1)$ neighborhood of $\beta$, for $\nabla \varphi_0 H_0(\beta) = 0$ by construction. This, together with Lemma C2 and Lemma C3 and the root-n convergence shown in the previous paragraph, suggests that

$$\hat{H}_n(b) = \frac{1}{2}(b - \beta)' \nabla \varphi_0 H_0(\beta)(b - \beta) + \frac{1}{n} \left( \delta_+^*(\hat{Z}_i) + \delta_-^*(\hat{Z}_i) \right)(b - \beta) + o_p(n^{-1})$$

uniformly over an $O_p(n^{-1/2})$ neighborhood around $\beta$. The limiting distribution then follow from Theorem 2 in (Sherman 1994b) and that $\mathbb{E}[\delta_+^*(\hat{Z})\delta_-^*(\hat{Z})'] < \infty$ under FM2. Q.E.D.
Figure 1. \((X, \epsilon)\) (Logistic,Logistic). First two rows: Pairwise extremum estimator. Last two rows: Inverse-density weighted estimator.
Figure 2. \((X, \epsilon)^-(\text{Laplace, Laplace})\)
Figure 3. $(X, \epsilon)^{(\text{Normal,Normal})}$. 
Figure 4. Heteroskedastic Error. \((X, \epsilon)\sim \text{(Normal,Normal)}\).
References


Abstract

We study the informational content of widely used identifying assumptions in discrete panel data models. To do so, we begin by exploring the information in the static binary choice panel data introduced in Manski (1987), who showed that point identification was achievable under a stationarity assumption. Our first result is that while these conditions suffice for point identification, they do not to achieve regular identification, thereby ruling out the ability to conduct standard inference. This motivates exploring the informational content of stronger identifying conditions, such as exchangeability, e.g. Honoré (1992), and exclusion, e.g. Honoré and Lewbel (2002). Here we find that while exchangeability adds no informational content, exclusion does in the sense that the identification is now regular. Based on this identification result we propose an estimation procedure that has conventional asymptotic properties, such as root-n consistency and asymptotic normality. We then extend our analysis to dynamic discrete choice panel data models, e.g. Honore and Kyriaizidou (2000), Arellano and Carrasco (2003), and censored panel transformation/duration models, e.g. Khan and Tamer (2007), finding analogous results to the base model, in that exclusion restrictions add much informational content.

Keywords: Panel Data, Dynamic Discrete Choice, Duration Models.
1 Introduction

In this paper we consider the identification and information for regression coefficients in discrete panel data models. Models with discrete data are widespread in econometrics, as many variables of interest in empirical economic models are binary, such as consumer choice or employment status. Paralleling the increase in popularity of estimating binary regression models is the estimation of panel data models. The increased availability of longitudinal panel data sets has presented new opportunities for econometricians to control for individual unobserved heterogeneity across agents. In linear panel data models, unobserved additive individual specific heterogeneity, if assumed constant over time (i.e. “fixed effects”), can be controlled for when estimating the slope parameters by first differencing the observations.

Discrete panel data models have received a great deal of interest in both the econometrics and statistics literature, beginning with the seminal paper of Anderson (1970). For a review of the early work on this model see Chamberlain (1984), and for a survey of more recent contributions see Arellano and Honoré (2001). More generally speaking there is a vibrant and growing literature on both partial and point identification in nonlinear panel data models. There are a set of recent papers that deal with various nonlinearities in models with (short $T$) panels. See for example the work of Arellano and Bonhomme (2009), Bester and Hansen (2009), Bonhomme (2012), Chernozhukov, Fernandez-Val, Hahn, and Newey (2010), Evdokimov (2010), Graham and Powell (2012) and Hoderlein and White (2012). See also the survey in Arellano and Honoré (2001).

Our approach will be to begin with a set of weak assumptions which will result in what we will refer to as the base model. This base model will effectively be the same on considered in Manski (1987). This base model can be expressed as

$$y_i = I[\alpha_i + x_i'\beta_0 + \epsilon_{it} > 0] \quad i = 1, 2, \ldots n \quad t = 1, 2$$

(1.1)

where $y_{it}$ is the observed binary dependent variable, $x_{it}$ are the observed $K$ dimensional regressors, and $\beta_0$ is unknown $K$ dimensional parameter of interest. The scalar variables $\alpha_i$ and $\epsilon_{it}$ are unobserved, but will be assumed to satisfy varying conditions as we consider variations of the base model. The panel structure of the data set implies we have a few observations for each individual in a large cross section, so we are effectively assuming the number of individuals, $n$, is large whereas the number of time periods, $t$ is small, and set it 2, w.l.o.g.
The main objective of this paper will be to evaluate the identification power for the parameter of interest, $\beta_0$ of the varying restrictions on the unobservables of the model. Consequently, we structure the paper as follows. In the next section, we first explore the informational content for $\beta_0$ in the base model, formally defined by stating the conditions on both the observables and unobservables. We then explore the additional informal content of strengthening the assumptions on the unobservables considering those assumed in the recent panel data literature. In those cases that there is increased informational content of the stronger assumptions we propose new estimation procedures and derive their properties.

Section Three then considers a dynamic variation of the base model, where a lagged dependent variable is an explanatory variable, as considered in Arellano and Carrasco (2003), Honore and Kyriazidou (2000), Honoré and Lewbel (2002), among others. In these models of particular interest is the coefficient on the lagged dependent variable. As with the static model, we consider varying identifying assumptions so we can explore the informational content of the different conditions. In cases where we do find additional informational content we propose new estimation procedure. Section Four then considers panel data duration models modeled as censored monotone transformation models—see Ridder and Tunali (1999), Khan and Tamer (2007), by following the pattern of the previous sections by demonstrating which conditions add informational content, and providing new estimators for the set of assumptions that do. Section 5 concludes and the appendix collects the proofs of the main theorems.

2 Identification and Information in Binary Panel Data

In this section we consider the identification of our base models, which is a binary choice model with fixed effects. Anderson (1970) considered the problem of inference on fixed effects linear models from binary response panel data. He showed that inference is possible if the disturbances for each panel member are known to be white noise with the logistic distribution and if the observed explanatory variables vary over time. Nothing need be known about the distribution of the fixed effects and he proved that a conditional maximum likelihood estimator consistently estimates the model parameters up to scale.

See ? for a comprehensive survey on duration analysis which includes work on multiple durations.
Manski (1987) showed that inference remains possible if the disturbances for each panel member are known only to be time-stationary with unbounded support and if the explanatory variables vary enough over time.

Specifically, he considered the model:

\[ y_{it} = I[\alpha_i + x_{it}'\beta_0 + \epsilon_{it} > 0] \]  \hspace{1cm} (2.2)

where \( i = 1, 2, \ldots n, t = 1, 2 \). The binary variable \( y_{it} \) and the \( K \)-dimensional regressor vector \( x_{it} \) are each observed and the parameter of interest is the \( K \) dimensional vector \( \beta_0 \). The unobservables are \( \alpha_i \) and \( \epsilon_{it} \), the former not varying with \( t \) and often referred to as the “fixed effect” or the individual specific effect. Manski (1987) imposes no restrictions on the conditional distribution of \( \alpha_i \) conditional on \( x_i \equiv x_{i1}, x_{i2} \). Manski (1987) imposed the following conditions:

1. \( F_{\epsilon_{i1}|x_i,\alpha_i}(\cdot|x_i,\alpha_i) = F_{\epsilon_{i2}|x_i,\alpha_i}(\cdot|x_i,\alpha_i) \) for all \( \alpha_i, x_i \).

2. The support of \( \epsilon_{it}|x_i,\alpha_i \) is \( R \).

3. Let \( w_i = x_{i2} - x_{i1} \). Then the support of \( w_i \) does not lie in proper liner subspace of \( R^K \).

4. There exists at least one \( k \in \{1, 2, \ldots K\} \) such that \( \beta_k \neq 0 \) and \( w_k \) is distributed continuously with support on the real line conditional on the other components of \( w_i \).

5. The sample \( (y_i \equiv (y_{i1}, y_{i2}), x_i) \) \( i = 1, 2, \ldots N \) is i.i.d.

The above conditions imply that the parameter of interest \( \beta_0 \) is identified up to scale, as shown in Manski (1985), Manski (1987).

**Theorem 2.1** Under the 4 conditions above, \( \beta_0 \) is identified up to scale.

While in one sense this is a positive results, it turns out it is quite limited from an inference point of view. That is because the identification is on an arbitrarily small set of the observables, making the identification "irregular". One way to see this is to consider a sequence of "imposter" values of \( \beta_0 \), denoted by \( \beta_n \) such that \( \beta_n \to \beta_0 \). This was done in Chen, Khan, and Tang (2013) to show that the models in Manski (1985), Manski (1988),
where identified on a set of measure 0, thereby ruling out the possibility of conducting standard inference. We show that the identification for the panel data model considered here is of the same type:

**Theorem 2.2** Under assumptions 1-4, $\beta_0$ is identified, but on a set of measure 0.

The proof is actually a corollary from our next theorem and an impossibility result in Chamberlain (1984). One of the implications of this result is that the conditional maximum score estimator proposed in Manski (1987), while converging at the sub parametric rate as the original maximum score, will also be rate optimal under the stated conditions.

Given these conclusions, that inference will be nonstandard and difficult to conduct for this model, the next step to be able to conduct inference might be to add stronger assumptions which will hopefully add informational content for $\beta_0$. We note that Assumption 1 above imposes no restrictions on the form of the serial correlation between $\epsilon_{i1}, \epsilon_{i2}$. This was desirable in the sense it made results less sensitive to possible misspecification. But given our previous result one may want to impose more structure on the correlation in errors if it has informational content for $\beta_0$. Unfortunately, our next result proves it does not. Specifically, we consider imposing a much stronger assumption that Assumption 1 above. Specifically we now impose:

1': Conditional on $\alpha_i, x_i, \epsilon_{i1}, \epsilon_{i2}$ are independently and identically distributed.

Assumption 1' is quite stronger than Assumption 1, now imposing independence as well as identical conditional distributions.

Unfortunately, the stronger assumption has no informational content for $\beta_0$:

**Theorem 2.3** The model under Assumptions 1',2,3,4 is *informationally equivalent* to the model under Assumptions 1,2,3,4.

The notion of informational equivalence was introduced in Manski (1988), and recently extended in Chen, Khan, and Tang (2013). What this means in the context of the conclusions of our two previous theorems is that the researcher may want to impose alternative assumptions than Assumption 1 or 1'.

Consequently, we will now look at an alternative assumption to see if the informational content for $\beta_0$ can be increased. The alternative assumption we adopt will be referred to as
an exclusion restriction, analogous to those used in cross sectional models in Lewbel (1998) and Chen, Khan, and Tang (2013), and for panel data models in Honoré and Lewbel (2002). For the model at hand, by exclusion restriction, the based way to denote what we mean, is following the notation of Honoré and Lewbel (2002), express

\[ x_{it}^ \beta_0 = \tilde{x}_{it}^ \theta_0 + v_{it} \]

where \( \tilde{x}_{it} \) denotes the first \( K - 1 \) components of \( x_{it} \) and \( v_{it} \) denotes the \( K_{th} \) component, whose coefficient we normalized to 1. \( \theta_0 \) denotes the first \( K - 1 \) components of \( \beta_0 \).

With knew notation, our new conditions are:

1. Let \( e_{it} = \alpha_i + \epsilon_{it} \). The the conditional distribution of \( e_{it} \) is independent of \( v_{it} \), conditioning on all values of \( x_i \)

2. \( F_{\epsilon_{i1}|x_i,\alpha_i}(\cdot|\cdot) = F_{\epsilon_{i2}|x_i,\alpha_i}(\cdot|\cdot) \) for all \( \alpha_i, x_i \).

3. The support of \( \epsilon_{it}|x_i,\alpha_i \) is \( R \).

4. Let \( w_i = ((x_{i2} - x_{i1})', v_{i2} - v_{i1})' \). Then the support of \( w_i \) does not lie in proper linear subspace of \( R^K \).

5. Let \( w_K \equiv v_{i2} - v_{i1} \). Then \( w_K \) is distributed continuously with support on the real line conditional on the other components of \( w_i \).

6. The sample \( (y_i, x_i, v_i) \quad i = 1, 2, \ldots N \) is i.i.d., where \( v_i \equiv (v_{i1}, v_{i2}), y_i \equiv (y_{i1}, y_{i2}) \)

We see that our knew conditions are unequivocally stronger than in our first model, as we have maintained stationarity but added conditional independence. But our new conditions are neither stronger nor weaker than in the second model, as we do not require independence in \( \epsilon_{it} \) but we do impose the exclusion restriction. Our next theorem establishes that the new conditions provide the most informational content for \( \beta_0 \).

**Theorem 2.4** Under Assumptions 1", 2", 3", 4", \( \theta_0 \) is identified. Furthermore, unlike in the first two sets of conditions, the identification here is regular, implying the semi parametric variance bound is finite.
The conclusion of the theorem is that in the binary panel data model, the exclusion restriction has identifying power. This result is quite surprising when compared to the cross sectional setting. In that case, when analyzing the informational content in the model in Manski (1988), it was found in Chen, Khan, and Tang (2013) that adding the exclusion assumption added no informational content whatsoever. The two models, with and without the exclusion restriction were informationally equivalent.

The conclusion of the theorem immediately motivates the construction of a new estimator procedure which exploits this added informational content, and consequently converges at the parametric rate.

We consider and estimation procedure bee on the following probabilities:

Let 

\[ p_1(x_i, v_i, \alpha_i) = P(y_{i1} = 1|x_i, v_i, \alpha_i) = P(\epsilon_{i1} \leq x_{i1}' \beta_0 + v_{i1} - \alpha_i|x, \alpha_i) \]

Let 

\[ p_1^*(x_i, v_i) = \int p_1(x_i, v_i, \alpha_i)dF(\alpha_i|x_i) \]

Note \( p_1^*(x_i, v_i) \) is identified from our assumptions. Now consider the pair \((x_i, v_i), (x_i, v_j)\), by above and our assumptions, we have

\[ p_1^*(x_i, v_j) = p_1^*(x_i, v_i) \iff x_{i2}' \beta_0 + v_{i2} = x_{i1}' \beta_0 + v_{j1} \] (2.3)

This means if we could find such matches we could identify \( \beta_0 \) by regressing \( v_{j1} - v_{i2} \) on \( x_{i2} - x_{i1} \).

Of course, given our assumptions such matches will never occur, and we do not know \( p_1^*(x_i, v_i) \), nor will there be such exact matches.

But given suitable regularity conditions, listed below, \( p_1^*(x_i, v_i) \) can be estimated non parametrically, and although exact matches cannot be found, kernel weights can be assigned to each pair, with the weight larger for pairs that are closer to a match. Letting \( k_{h_n}(\cdot) \equiv \frac{1}{h_n} k\left(\frac{\cdot}{h_n}\right) P \) denote this kernel function, we could then define:

\[ \omega_{ij} = k_{h_n}(\hat{p}_2^*(x_i, v_j) - \hat{p}_1^*(x_i, v_i)) \] (2.4)
where

\[ \hat{p}_1^*(x_i, v_i) = \frac{\sum_{l \neq i} y_{il} K_1 h(v_l - v_i) K_2 h(x_l - x_j)}{\sum_{l \neq i} y_{il} K_1 h(v_l - v_i) K_2 h(x_l - x_j)} \]  

(2.5)

With these weights we could then define our estimator of \( \beta_0 \) as

\[ \hat{\beta} = \left( \sum_{i \neq j} \omega_{ij} (x_{i2} - x_{i1}) (x_{i2} - x_{i1})^t \right)^{-1} \sum_{i \neq j} \omega_{ij} (x_{i2} - x_{i1}) (v_{j1} - v_{i2}) \]  

(2.6)

We conclude this section by summarizing our results found for the static binary choice panel data model introduced in Manski (1987). Under condition in Manski (1987), the regression coefficients are identified, but on a set of measure 0, implying the difficulty to conduct standard inference on them. This result was found under a stationarity condition, but we found no additional informational content from strengthening the stationarity assumption to serial independence and stationarity. In contrast, the exclusion restriction considered in Honoré and Lewbel (2002) did have informational content in the sense that the regression coefficients were regularly identified and so could be estimated at the parametric rate. In the next section we explore how these results change when we consider dynamic panel data binary models.

3 Identification and Information in Dynamic Binary Panel Data

In this section we extend the base model of the previous section by introducing dynamics into the model. The dynamics we consider will be of the form of including lagged binary dependent variables as one of the explanatory variables. This model is well motivate in empirical settings.

In many situations, such as in the study of labor force and union participation, it is observed that an individual who has experienced an event in the past is more likely to experience the event in the future than an individual who has not experienced the event. This section expands results from the previous one by presenting identification and estimation
methods for discrete choice models with structural state dependence that allow for the presence of unobservable individual heterogeneity in panels with a large number of individuals observed over a small number of time periods.

Specifically we consider a discrete choice panel data model of the form:

$$y_{it} = I[\alpha_i + x_{it}'\beta_0 + \gamma_0y_{it-1} + \alpha_i + \epsilon_{it} > 0]$$

(3.7)

where $x_{it}$ is a vector of exogenous variables, $\epsilon_{it}$ is an unobservable error term, $\alpha_i$ is an unobservable individual-specific effect, and $\beta_0$ and $\gamma_0$ are the parameters of interest to be estimated.

This model or variations thereof have been considered in Chamberlain (1984), Honore and Kyriazidou (2000), Arellano and Carrasco (2003), Honoré and Lewbel (2002).

Our Assumptions, detailed below, imply that

$$P(y_{i0} = 1|x_i, \alpha_i) = p_0(x_i, \alpha_i)$$

(3.8)

$$P(y_{it} = 1|x_i, \alpha_i, y_{i0}, ... y_{it-1}) = F(x_{it}'\beta_0 + \gamma_0y_{it-1} + \alpha_i) \quad (t = 1, 2, 3)$$

(3.9)

Crucial to our results below is our “initial conditions” assumptions, which we keep along the lines of, for example, Chamberlain (1984), Honore and Kyriazidou (2000). For important work based on alternative initial conditions assumptions, see Honoré and Tamer (2006), Wooldridge (2002).

Note the above equality follows from the crucial assumption the the time varying errors are independently and identically distributed over time, as well as cross sectional units. With this assumption and the assumption of having data on 4 time periods, including the "initial" period 0, Honore and Kyriazidou (2000) proposed estimating the parameters by maximizing the objective function:

$$\frac{1}{n} \sum_{i=1}^{n} K\left(\frac{x_{i2} - x_{i3}}{h_n}\right) (y_{i2} - y_{i1})sgn((x_{i2} - x_{i1})'\beta + \gamma(y_{i3} - y_{i0}))$$

(3.10)

with respect to $\beta, \gamma$. Analogous to some of the estimation procedures discussed before $K(\cdot)$ denotes a kernel function assigning greater weight to observations where $x_{i2}, x_{i3}$ are close.
Related this this point, this estimator irregularly identifies the parameter $\beta_0, \gamma_0$ and so the estimator will not converge at the parametric rate, though it will be consistent.

The sub parametric rate of convergence of the estimator reflects the difficulty in attaining identification in this dynamic model. Hahn (2001) shows that there does not exist any estimator for $\beta_0, \gamma_0$ converging at the parametric rate under the condition in Honore and Kyriazidou (2000). This is even with the strong assumptions of iid errors across time and the cross section.

To see if it is possible to add informational content to this model, we again consider the exclusion restriction from the previous section.

For the dynamic model the exclusion restriction we assume is of the form:

$$ (\alpha_i, \epsilon_i) \perp v_i | x_i $$  \hspace{1cm} (3.11)

where as before variables without time subscripts indicate a lll time periods- e.g. $x_i = (x_{i1}, x_{i2})$.

Also, like in Honore and Kyriazidou (2000), we assume the error terms are iid across time, meaning that conditional on $\alpha_i, x_i, \epsilon_{i1}, \epsilon_{i2}$ are iid.

Under these assumptions we take the following approach to identify $\beta_0, \gamma_0$: Let $x_{it}^*$ denote $(x_{it}, y_{it-1})$, and let $x_i^*$ denote $(x_{i1}^*, x_{i2}^*)$.

$$ p_2(x_i^*, v_i, \alpha_i) \equiv P(y_{i2} = 1 | y_{i1}, x_i, v_i, \alpha_i) $$

$$ = P(\epsilon_{i2} \leq x_{i2}' \beta_0 + \gamma_0 y_{i1} + v_{i2} - \alpha_i | y_{i1}, x_i, \alpha_i) $$

$$ = P(\epsilon_{i2} \leq x_{i2}' \beta_0 + \gamma_0 y_{i1} + v_{i2} - \alpha_i | x_i, \alpha_i) $$

and

$$ p_1(x_i^*, v_i, \alpha_i) \equiv P(y_{i1} = 1 | y_{i0}, x_{i1}^*, v_i, \alpha_i) $$

$$ = P(\epsilon_{i1} \leq x_{i1}' \beta_0 + \gamma_0 y_{i0} + v_{i1} - \alpha_i | y_{i0}, x_i, \alpha_i) $$

$$ = p(\epsilon_{i1} \leq x_{i1}' \beta_0 + \gamma_0 y_{i0} + v_{i1} - \alpha_i | x_i, \alpha_i) $$
And as in the static case, we will define:

\[ p_t^*(x^*_i, v_i) \equiv \int p_t(x^*_i, v_i, \alpha_i) dF(\alpha_i | x_i) \]

and analogous to the static case:

\[ p^*_1(x^*_i, v_j) = p^*_2(x^*_i, v_i) \iff x'_{i2} \beta_0 + \gamma_0 y_{i1} v_{i2} = x'_{i1} \beta_0 + \gamma_0 y_{i0} + v_{j1} \quad (3.12) \]

and

\[ p^*_1(x^*_i, v_i) = p^*_2(x^*_i, v_j) \iff x'_{i1} \beta_0 + \gamma_0 y_{i0} v_{i2} = x'_{i2} \beta_0 + \gamma_0 y_{i1} + v_{j2} \quad (3.13) \]

This would suggest the kernel weighted least squares estimator:

\[ (\hat{\beta}, \hat{\gamma}) = \left( \sum_{i \neq j} \omega_{ij} (x^*_{i2} - x^*_i)(x^*_{i2} - x^*_i)' \right)^{-1} \sum_{i \neq j} \omega_{ij} (x^*_{i2} - x^*_i)(v_{j1} - v_{i2}) \quad (3.14) \]

Under standard conditions, this estimator will converge at the parametric rate with an asymptotic normal distribution. This fully demonstrates the identification power and informational content of the exclusion restriction. As demonstrated in both Honore and Kyriazidou (2000) and Hahn (2001), even with the stationarity and serial independence assumptions regular identification cannot be achieved without the exclusion restriction. Honoré and Lewbel (2002) impose the exclusion restriction but they relax the stationarity and serial independence assumption that we impose here. They demonstrate identification and propose an estimation procedure based on “inverse weighting” of the density function of the excluded variable \( v_i \). As shown in Khan and Tamer (2010), such an approach will generally not yield estimators converging at the parametric rate can even converge at rates slower than maximum score type procedures.
4 Identification and Information in Censored Duration Panel Data

In this section we consider estimation of a right censored duration model with fixed/group effects. As in the previous sections, we allow for general forms of censoring. Of particular interest in the panel data setting is to permit the distribution of the censoring variable to be spell-specific and individual/group specific. The vast existing literature does not address this type of problem. Honoré, Khan, and Powell (2002) allow for random censoring, but requires a linear transformation, and the censoring variables to be distributed independently of the covariates with the same distribution across spells. Extensions of the linear specification can be found in Abrevaya (2000),Abrevaya (1999), which allow for a generalized transformation function, but rule out fixed and/or general random censoring. See also ? Other work in the panel duration literature parametrically specifies the distribution of the error terms. Examples include Chamberlain (1984), Honore (1993), Ridder and Tunali (1999), Lancaster (1979), Horowitz and Lee (2004) and Lee (2008). Some of these also rule out censoring distributions that vary across spells and/or are independent of covariates. In the context of multiple spell data, we wish to allow for distribution of the censoring variable to vary across spells, for one of two reasons: for one, the censoring distribution may depend on time-varying covariates. Also, even if the censoring distribution does not depend on the covariates, and is purely a result of the observation plan, the observation plans may vary across spells.

To be precise, we will focus on the following model:

$$ T(\nu_{it}) = \min(\alpha_i + x'_{it}\beta_0 + \epsilon_{it}, c_{it}) $$
$$ d_{it} = I[\alpha_i + x'_{it}\beta_0 + \epsilon_{it} \leq c_{it}] \quad i = 1, 2, ... n, \quad t = 1, 2 $$

In the duration framework the subscript \( t \) denotes the spell, as opposed to the time period. \( T(\cdot) \) is unknown strictly monotonic function, \( \nu_{it} \) is an observed scalar dependent variable, \( x_{it} \) if a vector of observed covariates as before, and \( c_{it} \) is a censoring variables which is permitted to be random, and only observed for the censored observations. The observed binary variable \( d_{it} \) is an indicator if the observation is censored. As before the disturbance term \( \epsilon_{it} \) is unobserved, as is the individual specific effect \( \alpha_i \). The vector \( \beta_0 \) is again the parameter of interest we wish to identify.
As alluded to above identification becomes very complicated when we allow for the general censoring structure. Khan and Tamer (2007) establish point identification of $\beta_0$ and proposed the following maximum score type estimator:

$$\hat{\beta} = \arg \max_{\beta} \frac{1}{n} \sum_{i=1}^{n} d_{i1} I[v_{i1} < v_{i2} I[x_{i2}^t < x_{i1}^t \beta] + d_{i2} I[v_{i2} < v_{i1}] I[x_{i2}^t < x_{i1}^t \beta] \quad (4.15)$$

As with the other maximum score type estimators referred to in the previous sections, the estimator is consistent as point identification was proven. But the identification is irregular and the estimator converges at a sub parametric rate with non normal limiting distribution. So like in the previous sections we explore what additional restrictions may be used to add informational content. Given that we already are attaining irregular identification despite the assumption of stationarity and serial independence of $\epsilon_{it}$, like before we turn to an exclusion restriction.

To introduce an excluded variable into the above model we adopt the following notation:

$$T(v_{it}) = \min(\alpha_i + x_{it}^t \beta_0 + v_{it} + \epsilon_{it}, c_{it})$$

$$d_{it} = I[\alpha_i + x_{it}^t \beta_0 + v_{it} + \epsilon_{it} \leq c_{it}] \quad i = 1, 2, ..., n, \quad t = 1, 2$$

and we impose the following assumptions:

1. $v_i \perp (\alpha_i, \epsilon_i, c_i)|x_i$

2. $(\epsilon_i, c_i)$ are stationary and serially independent conditional on $x_i$.

To identify $\beta_0$ in this case, following steps analogous to before, let

$$p_i(x_i, v_i) = P(d_{it} = 0|x_i, v_i) = P(\alpha_i + \epsilon_{it} + c_{it} \leq x_{it}^t \beta_0 + v_{it}|x_i)$$

So given our assumptions, we have

$$p_1(x_i, v_i) = p_2(x_i, v_j) \iff x_{i1}^t \beta_0 + v_{i1} = x_{i2}^t \beta_0 + v_{j2} \quad (4.16)$$
Hence, as before we can search for pairs where the probabilities match, and regress \((v_{j2} - v_{i1})\) on \(x_{i1} - x_{i2}\). As before, while such exact matches will never occur, kernel weights can be used. Under our stated conditions the estimator will converge at the parametric rate, fully demonstrating the identification power of the exclusion restriction.

To our knowledge, this will be the only estimation procedure for the censored transformation panel data model converging at the parametric rate with this generality in the censoring structure.

5 Simulation Study

In the this section we explore the relative finite sample performances of the new proposed estimation procedures in both static and dynamic designs. Beginning with simulating data from a static model, we randomly generate data from the following design:

\[
y_{it} = I[\alpha_i + x_{it}\beta_0 + v_{it} + \epsilon_{it} > 0] \quad t = 1, 2
\]

(5.17)

where \(x_{it}\) was distributed bivariate normal, mean 0, variances 1, and correlation 0.5. The error terms \(\epsilon_{it}\) were distributed independently of \(x_{it}, \alpha_i\) and serially independent standard normal. The special regressor was distributed standard normal, independently of \(x_{it}, \alpha_i, \epsilon_{it}\). For the fixed effect \(\alpha_i\) we considered two designs, on where \(\alpha_i\) was binary with probability 0.5 taking the value 1, and probability 0.5 taking the value 0. In the other design we set \(\alpha_i = (x_{i1} + x_{i2})/2 + u_i\), where \(u_i\) was distributed standard uniform, independently of the all the other variables. Table 1 reports the mean bias, median bias and mean squared error from 401 replications, for two estimators of \(\beta_0\), the panel maximum score estimator in Manski (1987) and the estimator proposed in this paper. The estimator proposed here required the selection of tuning parameters for both the conditional probability estimation as well as the matching of both propensity scores and the nonspecial regressors \((x_i)\). To select these tuning parameters we followed the procedure outlined in Chen and Khan (2008), which emphasized that the bandwidth in the propensity score estimation has to be under smoothed when compared to the bandwidth for matching.

As the results indicate, the finite sample performances coincide somewhat with the asymptotic theory. Both estimators clearly demonstrate consistency, with biases and MSE’s declining with the sample size. Also both estimators indicate to a certain extent finite sam-
ple symmetry with both small mean and median biases that are generally the same sign. The new estimator demonstrates smaller MSE than the maximum score estimator, which is expected by the asymptotic theory, as the new estimator is designed to exploit the stronger exclusion restriction that the maximum score estimator is not.

Table two reports results from simulating data for a dynamic design where we generated data from the following model:

\[ y_{it} = I[\alpha_i + x_{it}\beta_0 + \gamma_0 y_{i,t-1} + v_{it} + \epsilon_{it} > 0] \quad t = 0, 1, 2, 3 \]  \hspace{1cm} (5.18)

Here we used 4 time periods, as is required for consistency of the estimator in Honore and Kyriazidou (2000). As in the static design we assumed \( \epsilon_{it}, v_{it} \) here are each serially independent, and distributed standard normal, independent of each other and independent of \( \alpha_i, x_i \). We again consider two designs for \( \alpha_i \), distributed analogously to how they were in the static design, and \( x_i \) was now distributed multivariate normal with pairwise correlations equal to 0.5. The initial variable \( y_{0i} \) was distributed standard normal, independent of all the other variables. Results are reported for the new estimator for and the kernel weighted maximum core estimator proposed in Honore and Kyriazidou (2000). To implement the latter, we used a normal kernel function and a bandwidth selected as if estimating the density function of the random variable \( x_{i3} - x_{i2} \) evaluated at 0. As the new estimator can be implemented with three time periods, we estimated \( \beta_0, \gamma_0 \) twice, once for periods 0,1,2 and again for periods 1,2,3, and averaged the results. For both the new estimator and the estimator in Honore and Kyriazidou (2000), the same statistics as in table 1 are reported for 50,100,200 and 400 observations, with 401 replications, for each of the parameters \( \beta_0, \gamma_0 \).

The results for the dynamic design are somewhat analogous to those of the static design, the new estimator has noticeable smaller MSE than the estimator in Honore and Kyriazidou (2000), which is to be expected as the latter converges at a slower rate than the former since the latter is not based on the special regressor assumption. One difference between the results in table 2 and table 1 is in the dynamic design the new estimator exhibits a larger bias than in the static design. This may well be due to the fact that first stage estimator now involves more regressors, but a more formal and rigorous analysis is warranted for an explanation.
<table>
<thead>
<tr>
<th></th>
<th>Design 1</th>
<th>Design 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>50 obs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Bias</td>
<td>-0.6491</td>
<td>-0.4005</td>
</tr>
<tr>
<td>Median Bias</td>
<td>-0.7841</td>
<td>-0.3905</td>
</tr>
<tr>
<td>MSE</td>
<td>0.5955</td>
<td>0.2435</td>
</tr>
<tr>
<td>100 obs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Bias</td>
<td>-0.4742</td>
<td>-0.3005</td>
</tr>
<tr>
<td>Median Bias</td>
<td>-0.5241</td>
<td>-0.3205</td>
</tr>
<tr>
<td>MSE</td>
<td>0.4039</td>
<td>0.1321</td>
</tr>
<tr>
<td>200 obs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Bias</td>
<td>-0.3041</td>
<td>-0.2205</td>
</tr>
<tr>
<td>Median Bias</td>
<td>-0.3641</td>
<td>-0.2305</td>
</tr>
<tr>
<td>MSE</td>
<td>0.2557</td>
<td>0.0821</td>
</tr>
<tr>
<td>400 obs.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean Bias</td>
<td>-0.1871</td>
<td>-0.1676</td>
</tr>
<tr>
<td>Median Bias</td>
<td>-0.2241</td>
<td>-0.1705</td>
</tr>
<tr>
<td>MSE</td>
<td>0.1549</td>
<td>0.0468</td>
</tr>
</tbody>
</table>
### Table 2
Simulation Results for Dynamic Panel Binary Regression Estimators

<table>
<thead>
<tr>
<th></th>
<th>Design 1</th>
<th></th>
<th></th>
<th>Design 2</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\beta_0$</td>
<td>$\gamma_0$</td>
<td>$\beta_0$</td>
<td>$\gamma_0$</td>
<td>$\beta_0$</td>
<td>$\gamma_0$</td>
</tr>
<tr>
<td>50 obs.</td>
<td>Mean Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>Median Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0239</td>
<td>0.0321</td>
<td>0.0524</td>
<td>0.0619</td>
<td>0.0141</td>
</tr>
<tr>
<td>100 obs.</td>
<td>Mean Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>Median Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0239</td>
<td>0.0321</td>
<td>0.0524</td>
<td>0.0619</td>
<td>0.0141</td>
</tr>
<tr>
<td>200 obs.</td>
<td>Mean Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>Median Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0239</td>
<td>0.0321</td>
<td>0.0524</td>
<td>0.0619</td>
<td>0.0141</td>
</tr>
<tr>
<td>400 obs.</td>
<td>Mean Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>Median Bias</td>
<td>0.0141</td>
<td>0.0505</td>
<td>-0.0306</td>
<td>0.0292</td>
<td>0.0141</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.0239</td>
<td>0.0321</td>
<td>0.0524</td>
<td>0.0619</td>
<td>0.0141</td>
</tr>
</tbody>
</table>

### 6 Conclusions

In this paper we explored the identifying power of varying restrictions in discrete response panel data models. Our approach is to treat the conditions in Manski (1987) as the base model and to continue adding restrictions to explore their informational content. We first strengthen the level of serial dependence by assuming the unobserved time varying heterogeneity is independently and identically distributed, and find that this adds no informational content. The notion of information equivalence we use for this finding is based on Manski (1988), which is a more refined notion than the Fisher information.

We then consider an exclusion restriction, in the form of a special regressor, and find that it indeed adds informational content, which is in contrast to the cross sectional discrete response model, as found in Chen, Khan, and Tang (2013). Specifically, we find that the Fisher information for the regression coefficients was positive with the exclusion restriction, whereas Chamberlain (1984) found there to be zero information without an exclusion re-
striction. We then extend our analysis to consider the informational content in dynamic models, such as those considered in Chamberlain (1984), Honore and Kyriazidou (2000), Arellano and Carrasco (2003), Honoré and Lewbel (2002). Here again we find informational content with a special regressor, this time in the form of attaining regular identification for the coefficient of the lagged dependent variable. Furthermore, we propose a new semi parametric estimator for the coefficient of the lagged dependent variable that converges at the paramedic rate with a limiting normal distribution. Analogous new results are also found for the regression coefficients in a panel data (i.e. multiple spell) duration model, where the exclusion restriction is shown to add strong informational content.

The results in this paper do suggest areas for future research. For example, in the models where we do attain positive Fisher information, it would be useful to derive the semi parametric efficiency bound as well as propose estimators that are asymptotically efficient. Furthermore, most of the results in this paper are for 2 or 3 time periods, and it would be useful to formally analyze the additional informational content in models with wider panels.
References


A Informational Equivalence Result

In this section we prove the informational equivalence between the stationarity assumption and the i.i.d. assumption on the disturbance terms. Recall from the paper we began with the stationarity assumption and then we imposed:

1’: Conditional on $\alpha_i, x_i, \epsilon_{i1}, \epsilon_{i2}$ are independently and identically distributed.

Assumption 1’ is quite stronger than Assumption 1 (stationarity), now imposing independence as well as identical conditional distributions.

Recall the statement of the theorem.
Theorem A.1 The model under Assumptions 1',2,3,4 is informationally equivalent to the model under Assumptions 1,2,3,4.

Proof: The proof involves a series of steps, and we first recall the definitions of the two models we wish to show are informationally equivalent.

Model A: \( y_{it} = 1\{x_{it}\beta + \alpha_i \geq \epsilon_{it}\} \) for \( t = 1, 2 \) where \( F_{\epsilon_{i1}|x_i,\alpha_i} = F_{\epsilon_{i2}|x_i,\alpha_i} \) for all \( x_i, \alpha_i \) (conditional stationarity)

We drop subscripts \( i \) to simplify notation below. Let \( x \equiv (x_1, x_2) \) and \( \epsilon \equiv (\epsilon_1, \epsilon_2) \). Use lower cases to denote realizations. Let \( X \) be the support of \( x \); and let \( F \equiv \{F_{\epsilon,\alpha|x} : x \in X\} \) be the parameter space for (or, equivalently, the restrictions on) \( F \equiv \{F_{\epsilon,\alpha|x} : x \in X\} \) that satisfy stationarity. Let \( (\beta, F) \) be the true parameters in the DGP. Let \( p(x; \beta, F) \equiv (p_1(x; \beta, F), p_2(x; \beta, F), p_{12}(x; \beta, F)) \) where \( p_t(x; \beta, F) \equiv E[Y_t | X = x; \beta, F] \) and \( p_{12}(x; \beta, F) \equiv E[Y_1Y_2 | X = x; \beta, F] \).

Definition: We say \((b, G)\) is observationally equivalent to \((\beta, F)\) (denoted by \((b, G) \mathcal{\equiv}(\beta, F)\)) when \( p(x; \beta, F) \equiv p(x; b, G) \) for almost all \( x \). We say \( \beta \) is observationally equivalent to \( b \) under \( F \) (denoted by \( \beta \mathcal{\equiv}_F b \)) if there exists \( G \in F \) such that \((b, G) \mathcal{\equiv}(\beta, F)\). We say \( \beta \) is identified relative to \( b \) under \( F \) (denoted by \( \beta \mathcal{\equiv}_{i.d.} b \)) if \( \beta \) is not observational equivalent to \( b \) under \( F \) (or equivalently, if for all \( G \in F \) there exists a subset \( Q_b \subseteq X \) with positive measure such that \( p(x; \beta, F) \neq p(x; b, G) \) for all \( x \in Q_b \).

We now apply the criterion in Manski (JASA 1988) to analyze the identification of this model. For any \( b \neq \beta \), define

\[
Q(b) \equiv \{x \in X : \text{sgn}[(x_1 - x_2)b] \neq \text{sgn}[(x_1 - x_2)\beta]\}
\]

As is shown in the corollary to Lemma 1 in Manski (1987), we can write \( Q(b) \) equivalently as

\[
Q(b) = \{x \in X : \text{sgn}[(x_1 - x_2)b] \neq \text{Med}[Y_1 - Y_2 | Y_1 \neq Y_2, X = x]\}.
\]

We now characterize the set of states that help identify \( \beta \) relative to a generic “imposter” \( b \neq \beta \).
**Lemma A.1** If $\Pr\{X \in Q(b)\} > 0$, then $\beta \overset{i.d.}{\rightarrow} x b$.

**Proof:** It suffices to show that for any $x \in Q(b)$, there exists no $G \in \mathcal{F}$ such that $p(x; b, G) = p(x; \beta, F)$. Without loss of generality, consider a generic $\bar{x} \in Q(b)$ such that $\bar{x}_1 \beta \geq \bar{x}_2 \beta$ and $\bar{x}_1 b < \bar{x}_2 b$. By construction,

$$p_t(\bar{x}; b, G) = \int G_{\epsilon_1|X=\bar{x},\alpha}(\bar{x}_t \beta + \alpha) dG_{\alpha|X=\bar{x}} \text{ for } t = 1, 2$$

where $G_{\epsilon_1|X,\alpha}$, $G_{\alpha|X}$ denote the conditional C.D.F. according to $G$. By construction,

$$G_{\epsilon_1|X=\bar{x},\alpha}(\bar{x}_1 b + \alpha) < G_{\epsilon_2|X=\bar{x},\alpha}(\bar{x}_2 b + \alpha)$$

for all $G \in \mathcal{F}$ (because, $G_{\epsilon_1|X,\alpha} = G_{\epsilon_2|X,\alpha}$) and all $\alpha$. Integrating out $\alpha$ with respect to the same conditional distribution $G_{\alpha|X=\bar{x}}$ on both sides of the inequality implies $p_1(\bar{x}; b, G) < p_2(\bar{x}; b, G)$. On the other hand, by similar derivation, we can show $p_1(\bar{x}; \beta, F) \geq p_2(\bar{x}; \beta, F)$ because $F \in \mathcal{F}$. Thus there exists no $G \in \mathcal{F}$ such that $p(\bar{x}; b, G) = p(\bar{x}; \beta, F)$. A symmetric argument proves the same result for $\bar{x} \in Q(b)$ where $\bar{x}_1 \beta < \bar{x}_2 \beta$ and $\bar{x}_1 b \geq \bar{x}_2 b$. Q.E.D.

**Lemma A.2** If $\Pr\{X \in Q(b)\} = 0$, then $\beta \overset{o.e.}{\rightarrow} x b$.

**Proof:** It suffices to show that there exists $G \in \mathcal{F}$ (which possibly depends on $b$ and $\beta$) such that $p(x; b, G) = p(x; \beta, F)$ for all $x \not\in Q(b)$. Without loss of generality, consider a generic $\bar{x} \not\in Q(b)$ such that $\bar{x}_1 \beta \geq \bar{x}_2 \beta$ and $\bar{x}_1 b \geq \bar{x}_2 b$. We want to construct a joint distribution $G_{(\epsilon_1, \epsilon_2); \alpha|X=\bar{x}}$ so that $p(\bar{x}; b, G) = p(\bar{x}; \beta, F)$.

We start by constructing the conditional joint distribution $G_{(\epsilon_1, \epsilon_2)|X=\bar{x}, \alpha}$. For any given $\alpha$, construct a marginal distribution $\bar{G}(.; \bar{x}, \alpha)$ (whose form depends on $\bar{x}, \alpha$) so that:

$$\bar{G}(\bar{x}_t b + \alpha; \bar{x}, \alpha) = F_{\epsilon_t|X=\bar{x},\alpha}(\bar{x}_t \beta + \alpha) \text{ for } t = 1, 2.$$ 

Recall that $F \in \mathcal{F}$ implies $F_{\epsilon_1|X,\alpha} = F_{\epsilon_2|X,\alpha}$. Hence the construction of such a marginal distribution $\bar{G}(.; \bar{x}, \alpha)$ requires $sgn((\bar{x}_1 - \bar{x}_2)\beta) = sgn((\bar{x}_1 - \bar{x}_2)b)$, an equality guaranteed to hold at $\bar{x}$ because $\bar{x} \in Q_b$ by construction. Let $C_\alpha : [0, 1]^2 \rightarrow [0, 1]$ denote the copula function such that

$$C_\alpha(F_{\epsilon_1|X=\bar{x},\alpha}(\varepsilon), F_{\epsilon_2|X=\bar{x},\alpha}(\varepsilon)) = F_{(\epsilon_1, \epsilon_2)|X=\bar{x},\alpha}(\varepsilon, \varepsilon') \quad (A.19)$$

23
for all \( \varepsilon, \varepsilon' \). The Sklar’s Theorem states such a copula exists. Construct a joint distribution \( G_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha} \) as follows: First, define two (identical) marginal distributions \( G_{\epsilon_1|X=\bar{x}, \alpha}(\cdot) \equiv \tilde{G}(\cdot; \bar{x}, \alpha) \) for \( t = 1, 2 \). Then construct a joint distribution

\[
G_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha}(\varepsilon, \varepsilon') \equiv C_\alpha(G_{\epsilon_1|X=\bar{x}, \alpha}(\varepsilon), G_{\epsilon_2|X=\bar{x}, \alpha}(\varepsilon')).
\]

By the Sklar’s Theorem, \( G_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha} \) is a well-defined joint C.D.F. For any fixed \( \alpha \), the joint distribution constructed in this way satisfies the conditional stationarity; and

\[
G_{\epsilon_t|X=\bar{x}, \alpha}(\bar{x}_t b + \alpha) = F_{\epsilon_t|X=\bar{x}, \alpha}(\bar{x}_t \beta + \alpha) \quad \text{for} \quad t = 1, 2.
\]

(A.20)

Furthermore, it then follows that

\[
G_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha}(\bar{x}_1 b + \alpha, \bar{x}_2 b + \alpha) = C_\alpha(G_{\epsilon_1|X=\bar{x}, \alpha}(\bar{x}_1 b + \alpha), G_{\epsilon_2|X=\bar{x}, \alpha}(\bar{x}_2 b + \alpha)) = C_\alpha(F_{\epsilon_1|X=\bar{x}, \alpha}(\bar{x}_1 \beta + \alpha), F_{\epsilon_2|X=\bar{x}, \alpha}(\bar{x}_2 \beta + \alpha)) = F_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha}(\bar{x}_1 \beta + \alpha, \bar{x}_2 \beta + \alpha)
\]

(A.21)

where the first equality is due to the construction of \( G_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha} \) using the copula \( C_\alpha \); the second equality follows from (A.20); and the third equality follows from the definition of \( C_\alpha \).

Next, let \( G_{\alpha|X=\bar{x}} = F_{\alpha|X=\bar{x}} \). It then follows from (A.20) and (A.21) that

\[
p_t(\bar{x}; b, G) = \int G_{\epsilon_t|X=\bar{x}, \alpha}(\bar{x}_t b + \alpha)dG_{\alpha|X=\bar{x}} = \int F_{\epsilon_t|X=\bar{x}, \alpha}(\bar{x}_t \beta + \alpha)dF_{\alpha|X=\bar{x}} = p_t(x; \beta, F) \quad \text{for} \quad t = 1, 2
\]

and

\[
p_{12}(\bar{x}; b, G) = \int G_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha}(\bar{x}_1 b + \alpha, \bar{x}_2 b + \alpha)dG_{\alpha|X=\bar{x}} = \int F_{\epsilon_1, \epsilon_2|X=\bar{x}, \alpha}(\bar{x}_1 \beta + \alpha, \bar{x}_2 \beta + \alpha)dF_{\alpha|X=\bar{x}} = p_{12}(x; \beta, F).
\]

To sum up, \( p(\bar{x}; b, G) = p(\bar{x}; \beta, F) \). This proves the claim. Q.E.D.

Next, we consider a more restrictive version of Model A, where the conditional stationarity assumption “\( F_{\epsilon_1|x, \alpha} = F_{\epsilon_2|x, \alpha} \) for all \( x, \alpha \)” is strengthened into a conditional i.i.d. assumption (i.e. “\( \epsilon_1 \) is independent from \( \epsilon_2 \) conditional on \( (x, \alpha) \) and \( F_{\epsilon_1|x, \alpha} = F_{\epsilon_2|x, \alpha} \)”). Refer to this
model with a stronger assumption as Model B, and denote the set of nuisance parameter satisfying conditional i.i.d. assumption as $\mathcal{F}_0$, which by construction is a strict subset of $\mathcal{F}$. The next claim shows the set of states leading to the identification of $\beta$ relative to a generic “imposter” $b \neq \beta$ is the same as that in Model A.

\textbf{Lemma A.3} \ $\beta \overset{i.d.}{\sim} \mathcal{F}_0 b$ iff $\Pr\{X \in Q(b)\} > 0$.

\textbf{Proof:} Sufficiency follows from the same argument as in the proof of Lemma A.1 (because $\mathcal{F}_0 \subset \mathcal{F}$). Proof of necessity also follows from exactly the same argument as the proof of Lemma A.2. The notable difference in the proof is that the copula defined in (A.19) under $\mathcal{F}_0$ is necessarily the independence copula for all $\alpha$. That is $C_\alpha(s, s') = s \ast s'$ for $(s, s') \in [0, 1]^2$. Q.E.D.

To sum up, we have shown that in a static binary panel model, strengthening the conditional stationarity assumption in Manski (1987) into a conditional i.i.d. assumption considered in ? does not lead to any informational advantage by the support-based criterion in Manski (1988). This is consistent with the finding in ? that the Fisher information for $\beta$ in the latter model is zero (which implies the Fisher information for $\beta$ must also be zero in the model under weaker conditional stationarity assumption).