cemmap
centre for microdata methods and practice

Identification and estimation in a correlated random coefficients binary response model

Stefan Hoderlein
Robert Sherman

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP42/12

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL
An ESRC Research Centre

IDENTIFICATION AND ESTIMATION IN A CORRELATED RANDOM COEFFICIENTS BINARY RESPONSE MODEL

By Stefan Hoderlein and Robert Sherman[1]

October, 2011.

Abstract

We study identification and estimation in a binary response model with random coefficients B allowed to be correlated with regressors X. Our objective is to identify the mean of the distribution of B and estimate a trimmed mean of this distribution. Like Imbens and Newey (2009), we use instruments Z and a control vector V to make X independent of B given V. A consequent conditional median restriction identifies the mean of B given V. Averaging over V identifies the mean of B. This leads to an analogous localize-then-average approach to estimation. We estimate conditional means with localized smooth maximum score estimators and average to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator of a trimmed mean of the distribution of B. The method can be adapted to models with nonrandom coefficients to produce $\sqrt{n}$-consistent and asymptotically normal estimators under the conditional median restrictions. We explore small sample performance through simulations, and present an application.

Keywords: Heterogeneity, Correlated Random Coefficients, Endogeneity, Binary Response Model, Instrumental Variables, Control Variables, Conditional Median Restrictions.

1. Introduction

Individuals with the same observed characteristics often respond in systematically different ways to the same stimulus, especially when that stimulus is chosen by the individual. For example, people of the same race and gender, with the same level of education from the same school and in the same

[1]We thank the seminar participants at Rochester as well as Kim Border and Lan Nguyen for helpful discussions.

area of study, may realize different rates of return to education due to unobserved variables like ability and motivation. This is an example of the phenomenon of heterogeneous marginal effects. In linear regression, for example, this type of heterogeneity can be modeled with coefficients that are random, and so vary across the population. Of interest are various characteristics of the distribution of random coefficients, such as the mean. Each component of the mean is the average marginal effect of the corresponding regressor on the structural conditional expectation.

Unobserved factors may cause random coefficients and observed regressors to be correlated. For example, love of learning may motivate a person to pursue an academic career, even though that person could succeed in a more lucrative profession. Such a person may acquire more education but earn less money than a counterpart who is otherwise identical, but loves learning less. In other words, ceteris paribus, the more one loves learning, the more education one may seek at the expense of the rate of return to education. But rate of return to education is the coefficient of education in a standard linear model of wages. Thus, this coefficient and education may be negatively correlated. Heckman and Vytlacil (1998) note the plausibility of this negative correlation and show that nonzero correlation makes education endogenous, resulting in inconsistency of the least squares estimator of the mean of the distribution of random coefficients. This leads these authors to posit a correlated random coefficients model and to develop corresponding estimation procedures to consistently estimate the mean.

Similar considerations motivate the development of binary response models that allow for correlated random coefficients. For example, the decision of a married woman to work or not may depend on education level, and the coefficient of education may be positively correlated with education. Unobserved ability and motivation may drive this correlation. It may be that women of higher ability and motivation not only seek more education but give greater weight to education in

making their work decisions. If so, then the coefficient of education is not the same positive constant for all women, but tends to be higher for more educated women. As with the linear model, this type of heterogeneity can be modeled with a binary response model with random coefficients allowed to be correlated with regressors. We call such a model a correlated random coefficients binary response model, or a CRCBR model, for short. The objective of this paper is to identify the mean of the distribution of random coefficients in a CRCBR model, and then develop an estimator of a trimmed mean of this distribution.

More formally, let a latent scalar random variable $Y^* = XB^*$, where $X = (X_1, X_2, \ldots, X_k)$ is a $1 \times k$ vector of random explanatory variables and $B^* = (B_1^*, B_2^*, \ldots, B_k^*)'$ is a $k \times 1$ vector of random coefficients allowed to be correlated with X. Take $X_2 \equiv 1$, the intercept term. Then $XB^* = X_1B_1^* + B_2^* + X_3B_3^* + \cdots + X_kB_k^*$. Define $Y = \{Y^* > 0\} = \{XB^* > 0\}$. Note that $Y = \{XB > 0\}$ where $B = \lambda B^*$ for any $\lambda > 0$. That is, Y is invariant to positive scale normalizations of B^*, and so the distribution of B^* can only be identified up to scale. We assume that $B_1^* > 0$ and take $\lambda = 1/B_1^*$, so that the vector of random coefficients that can be identified is $B = B^*/B_1^* = (1, B_2, \ldots, B_k)$. For simplicity, from now on, we take $Y^* = XB = X_1 + B_2 + X_3B_3 + \cdots + X_kB_k$.

As indicated in the last paragraph, we normalize on the coefficient of X_1. For ease of exposition, throughout this paper we assume that X_1 has support $\mathbb{R}$. However, we stress that this support assumption is not necessary. For example, as we discuss later, X_1 could have bounded support. Apart from special cases, what is necessary for purposes of identification is that X_1 be continuously distributed. The full support assumption is invoked simply to make subsequent identification and trimming arguments more transparent.

For ease of exposition, we assume that the mean of B exists and write $\beta = I\!E B$.[2] The model is

$$Y = \{XB > 0\} \quad (1)$$

$$= \{\epsilon < X\beta\} \quad (2)$$

where $\epsilon = -X(B - \beta)$. This is the general CRCBR model.

The general CRCBR model nests a number of special cases that are interesting in their own right. For example, it reduces to the standard nonrandom coefficients binary response model with X exogenous when B_2, the coefficient of the intercept term, is the only random coefficient. In this model, B_2 plays the role of the error term (plus nonrandom intercept) and is uncorrelated with X. Another special case is the standard nonrandom coefficients binary choice model with endogenous regressors. Again, the only random coefficient is B_2, but this coefficient is allowed to be correlated with components of X. In addition, the CRCBR model reduces to the independent random coefficients binary response model when B is independent of X.

While the CRCBR model nests a number of interesting special cases, these special cases allow either no correlation between B and X, or correlation only between B_2 and X. The general model allows a richer correlation structure between B and X, and so can capture richer forms of heterogeneity, as suggested previously.

Refer to (2) and recall that $\epsilon = -X(B - \beta)$. We see that if any component of X is correlated with any component of B, then $I\!E X'\epsilon \neq 0$. In this sense, correlation between regressors and random coefficients implies endogeneity. Thus, we say that a component of X is endogenous if it is correlated with at least one component of B. To handle this endogeneity, we require instruments

[2] Our methods do not require that the mean of B exists, and even if the mean exists, β need not be the mean, but can denote any reasonable measure of center of the distribution of B.

and control variables.

Let $\mathcal{X}_1$ denote an endogenous component of X. We assume that there exists a vector of instruments, Z, for $\mathcal{X}_1$. This means that at least one component of Z is correlated with $\mathcal{X}_1$, and Z is unrelated to B in a sense to be defined shortly. In addition, we assume that $\mathcal{X}_1 = \phi_1(Z, V_1)$ where V_1 is a random variable independent of Z, and ϕ_1 is a real-valued function invertible in V_1 for each possible value of Z. If Z is independent of B given V_1, then $\mathcal{X}_1$ is independent of B given V_1, and so V_1 is a control variable for $\mathcal{X}_1$ with respect to B, as defined in Imbens and Newey (2009). Under these restrictions, $\mathcal{X}_1$ can be either a separable or nonseparable function of Z and V_1. For example, we allow the standard separable case $\mathcal{X}_1 = \phi_1(Z, V_1) = M(Z) + V_1$ where $M(Z) = I\!E(\mathcal{X}_1 \mid Z)$ and $V_1 = \mathcal{X}_1 - M(Z)$.

Let e denote the number of endogenous components of X and let $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_e)$ denote the vector of endogenous components of X. For simplicity, assume that Z is a vector of instruments for each component of $\mathcal{X}$. Let $V = (V_1, \ldots, V_e)$ denote the vector of control variables for $\mathcal{X}$. That is, as above, for each j, $\mathcal{X}_j = \phi_j(Z, V_j)$ where V_j is independent of Z and ϕ_j is a invertible in V_j conditional on Z. If Z is independent of B given V, then $\mathcal{X}$ is independent of B given V. If, in addition, the exogenous components of X are independent of B, then X is independent of B given V. In this setting, conditioning on V, the source of endogeneity, produces conditional independence. This suggests a localize-then-average approach to both identification and estimation.

If X is independent of B given V, and (B, V) satisfies a type of joint symmetry condition, then certain conditional median restrictions hold, which generalize the median restrictions of Manski (1975,1985). These conditional median restrictions are sufficient to identify $I\!E(B \mid V)$ in a distribution-free way that allows arbitrary forms of heteroscedasticity in a conditional error term. Averaging over V then identifies β.

We estimate $I\!E(B \mid V)$ with a localized version of the smoothed maximum score estimator of Horowitz (1992), and then average the trimmed estimated conditional expectations to obtain an estimator of a trimmed mean of the distribution of B. This trimmed mean can be made arbitrarily close to β by choosing a large enough trimming constant. The conditional expectation estimators suffer from a curse of dimensionality, but the estimator of the trimmed mean does not. The averaging overcomes the curse and yields a $\sqrt{n}$-consistent and asymptotically normal estimator of the trimmed mean. An interesting aspect of the estimation procedure is the localization step. We do not localize directly on V, which would require estimating V, a difficult task, in general. Rather, we localize on an invertible transformation of V, which is easily estimated with kernel regression methods. This simplified localization step is made possible by a simple generalization of a result in Matzkin (2003).

If the conditional median restrictions hold, then the localize-then-average method can be adapted to models with nonrandom slope coefficients, with either exogenous or endogenous regressors, to produce $\sqrt{n}$-consistent and asymptotically normal estimators of the nonrandom slope coefficients. For example, under the conditional median restrictions, it is possible to construct $\sqrt{n}$-consistent and asymptotically normal estimators of the nonrandom slope coefficients in the models of Manski (1975,1985) and Horowitz (1992). It is well known (Horowitz, 1993) that it is not possible to achieve $\sqrt{n}$-consistency under the maintained assumptions of these authors. However, the extra information provided by instruments, along with the stronger conditional median restrictions and the estimation procedure that exploits this information, makes this possible.

Our localize-then-average identification and estimation strategy is very similar in spirit to that of Imbens and Newey (2009). These authors use instruments and condition on unobserved control variables to break the dependence of endogenous regressors on multiple sources of unobserved

heterogeneity, and then average over the control variables to identify various structural effects of interest. They average over estimated control variables to estimate these effects. They do this in the context of nonseparable models with continuous outcomes. By contrast, we treat the case of binary outcomes and estimate a measure of center of the distribution of random coefficients, leading to a very different type of estimator and analysis. Altonji and Matzkin (2005) use a similar localize-then-average identification and estimation strategy in their earlier work, but do not consider random coefficients. Blundell and Powell (2004) and Petrin and Train (2006) use related control function strategies. Our approach is also related to the work of Matzkin (1992), who was the first to consider nonparametric identification in the exogenous binary choice model.

Lewbel (2000) can estimate a measure of center of the distribution of random coefficients in a CRCBR model provided there exists a special regressor with heavy tails that, conditional on instruments, is independent of the structural error and the other regressors in the model. Bajari, Fox, Kim, and Ryan (2008) and Gautier and Kitamura (2009) identify and estimate the entire distribution of random coefficients in a binary response model, but focus on the case of exogenous regressors. Fox and Ghandi (2010) focus on identification of the distribution of random coefficients in a more general nonlinear set-up. Hoderlein (2010) identifies and estimates the mean of the distribution of random coefficients in a binary response model, allowing correlation between a random intercept coefficient and regressors. However, he does not allow correlation between random slope coefficients and regressors. He also does not allow endogenous regressors to be nonseparable functions of control variables and instruments. As already mentioned, in the linear model, Heckman and Vytlacil (1998) identify and estimate the mean of the distribution of correlated random coefficients. However, their approach does not extend to the binary response model.

A number of authors study identification and estimation of nonrandom coefficients in the binary

response model, allowing correlation between the structural error term and the regressors, the usual notion of endogeneity. Rivers and Voung (1988) treat the parametric case. Lewbel (2000), mentioned previously, treats the semiparametric case. Vytlacil and Yildiz (2007) treat nonparametric identification and estimation of the effect of an endogenous binary regressor.

There is a huge semiparametric literature treating identification and estimation of nonrandom coefficients in the binary response model with exogenous regressors. A seminal reference for a direct, or nonoptimization, approach, is the work of Powell, Stock and Stoker (1989) on average derivative estimation. There are many optimization estimators, including the semiparametric least squares estimator of Ichimura (1993) and the semiparametric maximum likelihood estimator of Klein and Spady (1993). Most of these of estimators do not handle general forms of heteroscedasticity in X, even when X is exogenous. A notable exception is the maximum score (MS) estimator of Manski (1975,1985). However, the MS estimator converges at rate $n^{-1/3}$ and has a nonstandard limiting distribution, making asymptotic inference problematic. In response to this difficulty, Horowitz (1992) develops the smoothed maximum score (SMS) estimator. The SMS estimator also handles general forms of heteroscedasticity, but can converge at a rate arbitrarily close to the parametric rate of $n^{-1/2}$, depending on the smoothness of the distribution of model primitives, and has a limiting normal distribution, making standard asymptotic inference possible. Because of these attractive properties, we use a localized version of the SMS estimator in our localization step in estimation.

The rest of the paper is organized as follows. The next section develops the localize-then-average approach to identifying $\beta = I\!E B$. Section 3 develops the localize-then-average approach to estimating a trimmed mean of the distribution of B. We sketch an outline of the argument and state and discuss the assumptions used to prove that this estimator is $\sqrt{n}$-consistent and asymptotically

normally distributed. The proof is quite complicated due to the fact that the estimator is an average of optimization estimators evaluated at generated variables. A technical innovation involves establishing a uniform strong maximization condition, used to prove a consistency result uniform over the localizing variable. A formal statement of the asymptotic result is given for the special case of one endogenous regressor. We also discuss application of the asymptotic result to a number of interesting special cases. Section 4 presents simulation results and Section 5 presents an application. We conclude and give directions for future work in Section 6. A complete proof of the asymptotic result for the special case of one endogenous regressor is given in Appendix A. Issues of trimming and local polynomial estimation are addressed in Appendix B.

2. Identification

Recall the CRCBR model as defined in (1) and (2). Recall the definition of the vector of instruments Z and the vector of control variables V. We observe (Y, X, Z). Our objective is to identify $\beta = I\!EB$. We do so in two stages. In the first stage, we develop conditions under which $I\!E(B \mid V)$ is identified. We then average over V to identify $I\!EB$. Moreover, we identify $I\!E(B \mid V)$ without assuming knowledge of the functional form of distribution of B, while at the same time allowing arbitrary forms of heteroscedasticity, in X and V, in a conditional error term. Because of this, conditional versions of the sort of median restrictions imposed by Manski (1975,1985) and Horowitz (1992) play a role.

Let S_V denote the support of V. Define the conditional expectation $\beta(v) \equiv I\!E(B \mid V = v)$. For each $x \in S_X$ and each $v \in S_V$, we make the following conditional median independence and conditional median zero assumptions:

CMI. $med(x(B - \beta(v)) \mid X = x, V = v) = med(x(B - \beta(v)) \mid V = v)$.

CMZ. $med(x(B - \beta(v)) \mid V = v) = 0$.

Under CMI and CMZ in the general CRCBR model, we get that

$$\begin{aligned} med(Y^* \mid X = x, V = v) &= med(X\beta(V) + X(B - \beta(V)) \mid X = x, V = v) \\ &= x\beta(v) + med(x(B - \beta(v)) \mid X = x, V = v) \\ &= x\beta(v) + med(x(B - \beta(v)) \mid V = v) \quad (CMI) \\ &= x\beta(v) \quad (CMZ)\,. \end{aligned}$$

It follows that

$$\begin{aligned} Y &= \{XB > 0\} \\ &= \{\epsilon(V) < X\beta(V)\} \end{aligned}$$

where $\epsilon(V) = -X(B - \beta(V))$. Thus, if CMI and CMZ hold, then for each $x \in S_X$ and each $v \in S_V$,

$$med(\epsilon(v) \mid X = x, V = v) = 0\,. \tag{3}$$

This is the key result used to identify $\beta(v) = I\!E(B \mid V = v)$ for each $v \in S_V$. It also follows from this result that arbitrary forms of heteroscedasticity in x and v are allowed in $\epsilon(v)$. By averaging out over v, we identify $\beta = I\!EB$.

Assumption CMI defines the precise sense in which Z is assumed to be unrelated to B. Stronger, more intuitive conditions imply CMI. For example, if Z is independent of B given V and the exogenous components of X are independent of B, then X is independent of B given V, which implies CMI. When $\beta(v) = I\!E(B \mid V = v)$, CMZ is implied by a type of joint symmetry condition.

For example, CMZ holds if (B, V) is distributed multivariate normal. Both CMI and CMZ are satisfied if $B = \beta(V) + \nu$, where ν is a k-vector, independent of Z and V, whose components are independent and symmetric about zero. These conditions can be relaxed. For example, the symmetry conditions can be relaxed if we make a different choice of $\beta(v)$.[3]

In the next section, we show how to estimate a trimmed mean of the distribution of B with an analogous two-stage, localize-then-average estimation procedure. One possibility is to localize directly on the unknown vector V. This requires estimating V, which can be difficult without further information about the ϕ_j functions. Other possibilities arise from the fact that $\beta(V) = \beta(M(V))$ for any e-dimensional invertible function M. It follows that $\beta = I\!E\beta(V) = I\!E\beta(M(V))$ where the expectation can either be over V or $M(V)$. If CMI and CMZ (and therefore condition (3)) hold for V, then they must hold with $M(V)$ in place of V. Thus, by choosing M judiciously, we may replace localization on V with localization on $M(V)$ and thereby simplify the localization step. There are many possible choices for M. We choose one that leads to the simplest estimation procedure.

Recall that $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_e)$ denotes the vector of endogenous components of X. Write $S_{\mathcal{X}}$ for the support of $\mathcal{X}$. Also, write S_Z for the support of Z. Fix $v \in S_V$ and note that $v = (\phi_1^{-1}(x_1, z), \ldots, \phi_e^{-1}(x_e, z))$ for some $x = (x_1, \ldots, x_e) \in S_{\mathcal{X}}$ and $z \in S_Z$. It is easy to show (see Lemma 0 in Appendix A) that there exists an invertible function M such that

$$M(v) = (I\!P\{\mathcal{X}_1 \leq x_1 \mid Z = z\}, \ldots, I\!P\{\mathcal{X}_e \leq x_e \mid Z = z\}).$$

[3]The symmetry conditions reflect the choice $\beta(v) = I\!E(B \mid V = v)$. We make this choice for ease of exposition, so that β is the familiar object $I\!EB$. However, other measures of center, corresponding to other localizing functions, are possible. For example, take $\tilde{\beta}(v)$ to be the parameter of interest in a localized version of the binary median regression models of Manski (1975,1985) and Horowitz (1992). For fixed v, CMI and CMZ combined constitute a localized version of their median independence assumption, which does not require symmetry. Define $\tilde{\beta} = I\!E\tilde{\beta}(V)$. In general, $\tilde{\beta} \neq I\!EB$, but $\tilde{\beta}$ may still be a reasonable measure of center of the distribution of B.

For example, when $e = 1$, $v = v_1 = \phi_1^{-1}(x_1, z)$ and

$$M(v_1) = F_{V_1}(v_1) = F_{V_1}(\phi_1^{-1}(x_1, z)) = I\!P\{\mathcal{X}_1 \leq x_1 \mid Z = z\}$$

where F_{V_1} is the invertible cdf of V_1. Therefore, direct localization on v can be replaced by localization on $M(v)$, where each component of $M(v)$ can be easily estimated with kernel regression methods. We adopt this approach in the next section.

3. Estimation and Asymptotics

In this section, we develop a two-stage, localize-then-average estimator of a trimmed mean of the distribution of B. We prove that this estimator is $\sqrt{n}$-consistent and asymptotically normally distributed. Since the asymptotic analysis is complicated, we first discuss the overall strategy used to prove the result. Then, after stating and discussing assumptions used to obtain the general result, we state the main theorem for the special case of one endogenous regressor. Finally, we discuss application of the result to other models, including binary response models where X is exogenous or endogenous, and slope coefficients are nonrandom. A formal proof of the main theorem for the case of one endogenous regressor is given in Appendix A.

We begin by describing some of the model primitives in more detail. Let the instrument vector Z be a $1 \times m$ vector, and let c denote the number of components of Z that are continuous. We allow the case $c = 0$, and since Z contains the intercept term, we have that $0 \leq c \leq m - 1$. Let Z^C denote the $1 \times c$ vector of continuous components of Z.

Recall that e denotes the number of endogenous components of X and $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_e)$ denotes the $1 \times e$ vector of endogenous components of X. For now, we assume that $e \geq 1$, and since X_2 is

unity, we have that $1 \leq e \leq k-1$.[4] Recall that $V = (V_1, \ldots, V_e)$ denotes the $1 \times e$ vector of control variables corresponding to $\mathcal{X}$. Let $d \leq e$ denote the number of continuous components of $\mathcal{X}$ and let $\mathcal{X}^C$ denote the $1 \times d$ vector of continuous components of $\mathcal{X}$.

We now develop the first-stage estimator. As discussed at the end of the last section, instead of localizing directly on $V = v$ we localize on

$$M(v) = (I\!P\{\mathcal{X}_1 \leq x_1 \mid Z = z\}, \ldots, I\!P\{\mathcal{X}_e \leq x_e \mid Z = z\})$$

where $v = (\phi_1^{-1}(x_1, z), \ldots, \phi_e^{-1}(x_e, z))$, with $x \in S_{\mathcal{X}}$ and $z \in S_Z$.

Let (Y_i, X_i, Z_i), $i = 1, 2, \ldots, n$, denote iid observations from the CRCBR model described previously. Then $V_i = (V_{i1}, \ldots, V_{ie})$ where, for $j = 1, \ldots, e$, $V_{ij} = \phi_j^{-1}(\mathcal{X}_{ij}, Z_i)$. Define $U_i = (U_{i1}, \ldots, U_{ie})$ where, for $j = 1, \ldots, e$, $U_{ij} = I\!P\{\mathcal{X}_j \leq \mathcal{X}_{ij} \mid Z = Z_i\}$. Note that $U_i = M(V_i)$. We estimate U_{ij} with the standard kernel regression estimator:

$$\hat{U}_{ij} = \hat{I\!P}\{\mathcal{X}_j \leq \mathcal{X}_{ij} \mid Z = Z_i\} = \sum_{a=1}^{n} \{\mathcal{X}_{aj} \leq \mathcal{X}_{ij}\} K_n(Z_a - Z_i) / \sum_{a=1}^{n} K_n(Z_a - Z_i) \tag{4}$$

where $K_n(\boldsymbol{t}) = K(\boldsymbol{t}/\alpha_n)$ with $\boldsymbol{t} \in I\!R^m$ and α_n a positive bandwidth, and K an m-dimensional kernel function. Define $\hat{U}_i = (\hat{U}_{i1}, \ldots, \hat{U}_{ie})$. For $\boldsymbol{t} = (t_1, \ldots, t_m)$, we take

$$K(\boldsymbol{t}) = \prod_{b=1}^{m} K^{(b)}(t_b)\,. \tag{5}$$

That is, the m-dimensional multivariate kernel is a product of m univariate kernels. If the bth component of Z is discrete, then $K^{(b)}(t_b) \equiv \{t_b = 0\}$. If the bth component of Z is continuous,

[4] The case $e = 0$ is the exogenous case. We stress that our methods cover this important special case, and we say more about this case at the end of this section. However, different notation is required and so for now, we take $e \geq 1$.

then $K^{(b)}(t_b)$ is a smooth univariate kernel function, defined in A14 below.[5]

Write U for $M(V)$ and S_U for the support of U. We assume $S_U = [0,1] \otimes \cdots \otimes [0,1]$ (e factors).[6] Note that $\beta(V) = \beta(M^{-1}(U))$. By a slight abuse of notation, for each $u \in S_U$, we write $\beta(u)$ for $I\!E(B \mid U = u)$. Fix $u \in S_U$. Write B_u for a compact set known to contain $\beta(u)$. Following Horowitz (1992), we estimate $\beta(u)$ with $\hat{\beta}(u) = \underset{b \in B_u}{\operatorname{argmax}}\, \hat{S}_n(b \mid u)$ where

$$\hat{S}_n(b \mid u) = \frac{1}{n\tau_n^e} \sum_{j=1}^{n} (2Y_j - 1) K_n^*(X_j b) K_n(\hat{U}_j - u) \tau_\kappa(Z_j^C) \tag{6}$$

where $K_n^*(t) = K^*(t/\sigma_n)$ with $t \in I\!R$ and σ_n a positive bandwidth, and $K_n(\boldsymbol{t}) = K(\boldsymbol{t}/\tau_n)$ with $\boldsymbol{t} \in I\!R^e$, τ_n a positive bandwidth, and K an e-dimensional kernel function. For $\kappa > 0$, the trimming function $\tau_\kappa(z) = \{|z| \leq \kappa\}$.

In (6), the function $K_n^*(t)$ smoothly approximates the indicator function $\{t > 0\}$. The function $K^*(t)$ is the integral of a smooth univariate kernel function, defined in A14 below. We define $K(\boldsymbol{t})$ as in (5), with e replacing m. Since all the components of U are continuous, each $K^{(b)}(t_b)$ is a smooth univariate kernel function, also defined in A14 below.

In (6), note that we only trim on Z^C, the continuous components of Z. For ease of exposition, we assume that Z^C has support $I\!R^c$. Then τ_κ trims the jth summand in (6) when the (rescaled) denominator of $\hat{U}_j$ gets too close to zero. This prevents so-called ratio bias, as explained in Appendix B. If $c = 0$, then Z is discrete and so we do not trim. Equivalently, $\tau_\kappa(z) \equiv 1$.

[5] We need not condition on all the components of the instrument vector Z_i. If the jth component of X is endogenous, we may define $U_{ij} = I\!P\{\mathcal{X}_j \leq \mathcal{X}_{ij} \mid Z = Z_i^j\}$ where Z_i^j is comprised of any nonnull subset of the components of Z_i that are instruments for X_{ij}. Then $\hat{U}_{ij}$ is defined as in (4) above, except the regression is on a smaller set of instruments, a computational advantage. We have chosen not to present the estimator in this generality to avoid the extra notational burden this would entail.

[6] This support assumption is made for convenience. It is not necessary, as we later discuss prior to the proof of Lemma 0 in Appendix A. However, this assumption does imply that U is continuously distributed. Since $V = M^{-1}(U)$, if M^{-1} is continuous, then V is continuously distributed. Note, however, that this need not imply that $\mathcal{X}$ has all continuous components. That is, this support assumption does not preclude discrete endogenous regressors.

Recall that $\beta = I\!E B = I\!E\beta(U)$. For $\kappa > 0$, define the trimmed mean

$$\beta_\kappa = I\!E\beta(U)\tau_\kappa(\mathcal{X}^C, Z^C)\,. \tag{7}$$

where $\tau_\kappa(x,z) = \{|x| \le \kappa\}\{|z| \le \kappa\}$. We view β_κ as the parameter of interest. If $\beta < \infty$, then $\beta_\kappa \to \beta$ as $\kappa \to \infty$. That is, β_κ can be made arbitrarily close to β by choosing κ large enough.[7] The trimming in (7) prevents boundary bias and ratio bias in the estimator of β_κ. If $c = d = 0$, then $\mathcal{X}$ and Z, and therefore, U, are discrete and so we do not trim. Equivalently, we take $\tau_\kappa(x,z) \equiv 1$ and the estimand is β. If $\beta = \infty$, then β_κ can be viewed as a robust measure of the center of the distribution of B.

We estimate β_κ with

$$\hat{\beta}_\kappa = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}(\hat{U}_i)\tau_\kappa(\mathcal{X}_i^C, Z_i^C)\,. \tag{8}$$

As explained in Appendix B, trimming on both $\mathcal{X}_i^C$ and Z_i^C prevents boundary bias by trimming $\hat{\beta}(\hat{U}_i)$ when $\hat{U}_i$ gets too close to the boundary of S_U, while trimming on Z_i^C prevents ratio bias by trimming $\hat{\beta}(\hat{U}_i)$ when the denominator of $\hat{U}_i$ gets too close to zero.[8]

We get that

$$\hat{\beta}_\kappa - \beta_\kappa = \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}(\hat{U}_i) - \beta(U_i)\right]\tau_\kappa(\mathcal{X}_i^C, Z_i^C) + \frac{1}{n}\sum_{i=1}^{n}\left[\beta(U_i)\tau_\kappa(\mathcal{X}_i^C, Z_i^C) - \beta_\kappa\right]\,. \tag{9}$$

[7] It is not necessary to use the same trimming constant for $\mathcal{X}^C$ and Z^C. Also, if we divide the RHS of (7) by $p_\kappa = I\!E\tau_\kappa(\mathcal{X}^C, Z^C)$, then in some special cases, $\beta_\kappa = \beta$ for all $\kappa > 0$. This is true, for example, in the separable model $X = M(Z) + V$ where $M(Z) = I\!E(X \mid Z)$ and $V = X - M(Z)$, when (B, X, Z) is multivariate normal and X and Z have zero means. Nonzero means can be accommodated by subtracting them in the trimming functions. If the RHS of (7) is divided by p_κ, then the RHS of (8) must be divided by the sample analogue of p_κ. Such an adjustment makes a first order contribution to the asymptotic distribution of $\hat{\beta}_\kappa$. This contribution is small when κ is large.

[8] It may be possible to eliminate boundary bias by using a higher-order local polynomial estimator of U_i. While this would eliminate the need to trim on $\mathcal{X}_i^C$ in (7) and (8), it would still be necessary to trim on Z_i^C to prevent ratio bias. As explained in Appendix B, higher-order local polynomial estimation may prevent boundary bias, but it does not prevent ratio bias. It also leads to a much more complicated analysis.

The second term on the RHS of (9) converges to zero as $n \to \infty$ by the LIE and a LLN, and when rescaled by $\sqrt{n}$, has an asymptotic normal distribution by a standard CLT. The first term on the RHS of (9) equals

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}(\hat{U}_i) - \hat{\beta}(U_i)\right]\tau_\kappa(\mathcal{X}_i^C, Z_i^C) + \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}(U_i) - \beta(U_i)\right]\tau_\kappa(\mathcal{X}_i^C, Z_i^C)\,. \tag{10}$$

Write $\hat{\delta}(u)$ for the $k \times e$ matrix $\frac{\partial}{\partial u}\hat{\beta}(u)$. By a Taylor expansion of each $\hat{\beta}(\hat{U}_i)$ about the corresponding U_i, we get that the first term in (10) equals

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\delta}(\hat{U}_i^*)(\hat{U}_i - U_i)'\tau_\kappa(\mathcal{X}_i^C, Z_i^C) \tag{11}$$

where $\hat{U}_i^*$ is between $\hat{U}_i$ and U_i.

Three of the above terms make first order contributions to the asymptotic distribution of $\hat{\beta}_\kappa$: the second term in (9), the second term in (10), and the term in (11). The second term in (9) quantifies the first order contribution to the asymptotic distribution of $\hat{\beta}_\kappa$ when the U_i's are observed. The second term in (10) and the term in (11) quantify the penalties paid for having to estimate each U_i. We show that the second term in (10) can be decomposed into two terms, each of which makes a first order asymptotic contribution.

Notice that $\hat{\beta}_\kappa$ is not a standard optimization estimator, but rather an average of optimization estimators. Moreover, each component optimization estimator is based on generated variables. As such, $\hat{\beta}_\kappa$ requires a different method of analysis than a more standard estimator. We now sketch the basic strategy used in establishing the limiting distribution of $\hat{\beta}_\kappa$. We begin by stating and discussing the significance of several results that will be useful in this regard.

The first is a uniform consistency result. Recall that $S_U = [0,1] \otimes \cdots \otimes [0,1]$ (e factors). Let

U_κ denote a compact subset of S_U defined below. We show that as $n \to \infty$,

$$\sup_{u \in U_\kappa} |\hat{\beta}(u) - \beta(u)| = o_p(1)\,. \tag{12}$$

We use (12) to help establish a rate of uniform consistency, stated and discussed next. A technical novelty in proving (12) involves establishing a uniform version of an identification condition called a strong maximization condition. That is, we show that for any $\delta > 0$,

$$\inf_{u \in U_\kappa} \left[S(\beta(u) \mid u) - \sup_{|b - \beta(u)| \geq \delta} S(b \mid u) \right] > 0 \tag{13}$$

where $S(b \mid u)$ is the population analogue of $\hat{S}_n(b \mid u)$.

Standard results can be used to establish $\sqrt{n}$-consistency and asymptotic normality of the second term in (9). However, more is needed to handle the second term in (10) and the term in (11). These terms are averages of differences (or derivatives) of optimization estimators involving generated variables. The strategy used to analyze these terms is to replace each summand involving generated variables with a summand that involves only original observations plus a remainder term that has order $o_p(1/\sqrt{n})$ uniformly over S_U. One can then neglect the average of remainder terms and analyze the friendlier average of approximating terms using more standard methods. The principal means to this end are two results involving rates of uniform convergence.

Recall $\mathcal{X} = (\mathcal{X}_1, \ldots, \mathcal{X}_e)$ denotes the vector of endogenous components of X, and $S_\mathcal{X}$ the support of $\mathcal{X}$. For $x = (x_1, \ldots, x_e) \in S_\mathcal{X}$ and $z \in S_Z$, define $U(x, z) = (U_1(x_1, z), \ldots, U_e(x_e, z))$ where $U_j(x_j, z) = I\!P\{\mathcal{X}_j \leq x_j \mid Z = z)$ and $\hat{U}(x, z) = (\hat{U}_1(x_1, z), \ldots, \hat{U}_e(x_e, z))$ where $\hat{U}_j(x_j, z) = \hat{I\!P}\{\mathcal{X}_j \leq x_j \mid Z = z)$. For each $x \in S_\mathcal{X}$, define x^C to be the subvector of x corresponding to the continuous components of $\mathcal{X}$. For each $z \in S_Z$ define z^C to be the subvector of z corresponding to

the continuous components of Z. Define $\mathcal{X}_\kappa = \{x \in S_\mathcal{X} : |x^C| \leq \kappa\}$ and $Z_\kappa = \{z \in S_Z : |z^C| \leq \kappa\}$. Define $U_\kappa = \{U(x,z) \in S_U : x \in \mathcal{X}_\kappa, z \in Z_\kappa\}$, a compact subset of S_U. We show that

$$\sup_{x \in \mathcal{X}_\kappa, z \in Z_\kappa} |\hat{U}(x,z) - U(x,z)| = O_p(1/\sqrt{n}\alpha_n^c) \tag{14}$$

$$\sup_{u \in U_\kappa} |\hat{\beta}(u) - \beta(u)| = O_p(1/\sqrt{n}\alpha_n^c \sigma_n \tau_n^{e+1}) \,. \tag{15}$$

An analysis of the second term in (10) leads to an influence function representation for $\hat{\beta}(u) - \beta(u)$. A remainder term associated with this representation is a product of two factors, one associated with the gradient of $\hat{S}_n(b \mid u)$, the other associated with the hessian of $\hat{S}_n(b \mid u)$. (Of all the remainder terms associated with $\hat{\beta}_\kappa$, this one requires the most delicate analysis.) Using (14) and (15), we show that the factor associated with the gradient has order $O_p(1/\sqrt{n}\alpha_n^c \sigma_n \tau_n^{e+1})$ and the factor associated with the hessian has order $O_p(1/\sqrt{n}\alpha_n^c \sigma_n^2 \tau_n^{e+1})$. The product of these two factors has order $O_p(1/n\alpha_n^{2c}\sigma_n^3\tau_n^{2e+2})$. Provided $\alpha_n^{2c}\sigma_n^3\tau_n^{2e+2} \gg n^{-1/2}$, this product has order $o_p(1/\sqrt{n})$ and so can be ignored.

It is instructive to identify the sources of the term $\alpha_n^{2c}\sigma_n^3\tau_n^{2e+2}$ in the rates just discussed. The gradient contributes a factor of $\alpha_n^c \sigma_n \tau_n^{e+1}$: a factor of τ_n^e is automatically inherited from $\hat{S}_n(b \mid u)$, while the σ_n factor comes from differentiating $\hat{S}_n(b \mid u)$ with respect to b to form the gradient. The additional factor $\alpha_n^c \tau_n$ comes from a one term Taylor expansion of the jth summand of the gradient about U_j: the linear term, $\hat{U}_j - U_j$, accounts for the factor of α_n^c, while the coefficient of this linear term accounts for the factor of τ_n, since differentiating a kernel function results in a rescaling by the reciprocal of the bandwidth. Next, consider the factor of $\alpha_n^c \sigma_n^2 \tau_n^{e+1}$ contributed by the hessian of $\hat{S}_n(b \mid u)$. The sources of bandwidth factors are the same as for the gradient, except there is an extra factor of σ_n from differentiating $\hat{S}_n(b \mid u)$ twice with respect to b to form the hessian.

Note that, especially when both e and c are positive, there is a curse of dimensionality when estimating $\beta(u)$ for each $u \in U_\kappa$. However, there is no curse of dimensionality when estimating β_κ, no matter how large the values of e and c. This is due to averaging the optimization estimators to construct $\hat{\beta}_\kappa$. Consider once again the second term in (10). When the influence function representation of $\hat{\beta}(U_i) - \beta(U_i)$ is expanded and combined with the sum over i, what remains are essentially various zero-mean U-statistics of orders two, three and four. We apply the Hoeffding decomposition to represent each as a sum of degenerate U-statistics. The higher order degenerate statistics are negligible, while the degenerate U-statistics of order one drive the asymptotics. But these statistics are averages of summands that are themselves population averages over all but one of their arguments. Through a sequence of conditional expectations, each of which involves a change of variable resulting in a rescaling by a bandwidth factor, we flush out all bandwidth factors from denominators. This results in a $\sqrt{n}$ rate of convergence for the second term in (10) for arbitrary e and c. Something similar happens for other terms, leading to a $\sqrt{n}$ rate of convergence for $\hat{\beta}_\kappa$. It is also important to note that to flush out all the bandwidth factors it is critical that the bias reducing kernels that are used have derivatives that are odd functions over symmetric intervals. Bias reducing kernels that are sums of even degree polynomials defined on symmetric intervals (cf. Müller (1984)) possess this property, for example.

We now formally state the assumptions used to derive the limiting distribution of $\hat{\beta}_\kappa$. Recall the definitions of the regressor vectors X, $\mathcal{X}$, and $\mathcal{X}^C$, along with their respective supports S_X, $S_\mathcal{X}$, and $\mathbb{R}^d$, with respective dimensions k, e, and d satisfying $k \geq e \geq d$. Recall the definitions of the instrument vectors Z and Z^C, with respective supports S_Z and $\mathbb{R}^c$, with respective dimensions m and c satisfying $m \geq c$. Recall the definitions of the latent random vectors V and U, with respective supports S_V and S_U. Finally, recall the definitions of $\mathcal{X}_\kappa$, Z_κ, and U_κ, the respective

trimmed supports of $\mathcal{X}^C$, Z^C, and U.

A1. (Y_i, X_i, Z_i), $i = 1, \ldots, n$ are iid observations from the CRCBR model. (Y, X, Z) is a generic observation from this model: $Y = \{XB > 0\}$ where $X = (X_1, X_2, \ldots, X_k) \equiv (X_1, \tilde{X})$. X_1 has support $\mathbb{R}$. $X_2 \equiv 1$, the intercept term. $B = (1, B_2, \ldots, B_k)$, where $B_i = B_i^*/B_1^*$, with $B_1^* > 0$. $\mathcal{X} = (\mathcal{X}_1 \ldots, \mathcal{X}_e)$ is the vector of endogenous components of X. Each $\mathcal{X}_j = \phi_j(Z, V_j)$ where V_j is independent of Z and ϕ_j is differentiable in V_j and invertible in V_j given Z. $V = (V_1, \ldots, V_e)$.

A2. $U \equiv M(V)$, where M is a differentiable, invertible function from S_V onto S_U such that for each $u \in S_U$ with corresponding $v \in S_V$, there exist $x = (x_1, \ldots, x_e) \in S_{\mathcal{X}}$ and $z \in S_Z$ such that $u = M(v) = (\mathbb{P}\{\mathcal{X}_1 \leq x_1 \mid Z = z\}, \ldots, \mathbb{P}\{\mathcal{X}_e \leq x_e \mid Z = z\})$.

A3. At least one component of Z is correlated with each endogenous component of X.

A4. The support of X given $U = u$ is not contained in any proper linear subspace of $\mathbb{R}^k$.

A5. The distribution of X_1 conditional on $\tilde{X} = \tilde{x}$ and $U = u$ has an everywhere positive density.

A6. $f(x_1 \mid \tilde{x}, u)$ denotes the pdf of X_1 given $\tilde{X} = \tilde{x}$ and $U = u$, and is continuous in x_1 and u.

A7. $f(\tilde{x} \mid u)$ denotes the pdf/pmf of $\tilde{X}$ given $U = u$, and is continuous in u.

A8. $\beta(u)$ is a twice continuously differentiable function of u on S_U.

A9. For each $u \in U_\kappa$, there exists a known compact set B_u containing $\beta(u)$.

A10. The correspondence from $u \in U_\kappa$ to B_u is upper hemi-continuous.

A11. $med(x(B - \beta(u)) \mid X = x, U = u) = med(x(B - \beta(u)) \mid U = u)$.

A12. $med(x(B - \beta(u)) \mid U = u) = 0$.

A13. When $c = 0$ and $e > 0$, take $\alpha_n \equiv 1$, $\sigma_n \propto n^{-\sigma}$, and $\tau_n \propto n^{-\tau}$, with $0 < \sigma < 1/6$, $0 < \tau < 1/(4e+4)$, $\sigma < \tau$, and $3\sigma + \tau(2e+2) < 1/2$. When $c > 0$ and $e > 0$, take $\alpha_n \propto n^{-\alpha}$, $\sigma_n \propto n^{-\sigma}$, and $\tau_n \propto n^{-\tau}$, with $0 < \alpha < 1/4c$, $0 < \sigma < 1/6$, $0 < \tau < 1/(4e+4)$, $\sigma < \tau$, and $2\alpha c + 3\sigma + \tau(2e+2) < 1/2$.

A14. When estimating U_{ij} in (4), $K(\boldsymbol{t}) = \prod_{b=1}^{m} K^{(b)}(t_b)$, with $K^{(b)}(t) = \{t = 0\}$ when t_b is discrete, and $K^{(b)}(t) = \mathcal{K}_\alpha(t)$ when t_b is continuous. $\mathcal{K}_\alpha(t)$ has support $T = [-1, 1]$, integrates to unity, and satisfies $\int_T t^p \mathcal{K}_\alpha(t)dt = 0$, $p = 1, \ldots, p_\alpha$, where p_α is the smallest even integer greater that $1/2\alpha$. $K^*(t)$ satisfies $\frac{d}{dt}K^*(t) = \mathcal{K}_\sigma(t)$ where $\mathcal{K}_\sigma(t)$ has support T, integrates to unity, and satisfies $\int_T t^p \mathcal{K}_\sigma(t)dt = 0$, $p = 1, \ldots, p_\sigma$, where p_σ is the smallest even integer greater that $1/2\sigma$. When localizing on u in (6), $K(\boldsymbol{t}) = \prod_{b=1}^{e} K^{(b)}(t_b)$, with $K^{(b)}(t) = \mathcal{K}_\tau(t)$ where $\mathcal{K}_\tau(t)$ has support T, integrates to unity, and satisfies $\int_T t^p \mathcal{K}_\tau(t)dt = 0$, $p = 1, \ldots, p_\tau$, where p_τ is the smallest even integer greater that $1/2\tau$. Moreover, $\mathcal{K}'_\tau(t)$ exists and satisfies $\int_T \mathcal{K}'_\tau(t)dt = 0$.

A15. $Z^C \equiv (Z_1, \ldots, Z_c)$. Let D_z be the vector of discrete components of Z. Let $f(Z^C \mid D_z)$ denote the density of Z^C given D_z. Then $f(Z^C \mid D_z) = f(Z_1 \mid Z_2, \ldots, Z_c, D_z) \times \cdots \times f(Z_c \mid D_z)$. Each univariate conditional density in this product is bounded away from zero on Z_κ, and has p_a bounded derivatives. On $\mathcal{X}_\kappa \otimes Z_\kappa$, $U(x, z)$ has p_α bounded partial derivatives with respect to each component of Z^C.

A16. Let $f(U \mid Z, X\beta(u))$ denote the density of U given Z and $X\beta(u)$. Then $f(U \mid Z, X\beta(u)) = f(U_1 \mid U_2, \ldots, U_e, Z, X\beta(u)) \times \cdots \times f(U_e \mid Z, X\beta(u))$. Each univariate density in this product has p_τ bounded derivatives. Define $g(U, Z) \equiv 2\mathbb{P}(\epsilon(U) < X\beta(U) \mid U, Z) - 1$, where $\epsilon(U) = -X(B - \beta(U))$. $g(U, Z)$ has p_τ bounded partial derivatives with respect to each component of U. The marginal density of U is continuously differentiable.

A17. The pdf/pmf of Z given $X\beta(u) = s$ and the marginal pdf of $X\beta(u)$ at s have p_σ bounded derivatives with respect to s.

A18. Let $S(b \mid u) = I\!Eg(u, Z)f(u \mid Z)\{Xb > 0\}\tau_\kappa(Z)$. Let $H(b \mid u)$ denote the hessian matrix $\frac{\partial^2}{\partial b \partial b'} S(b \mid u)$. Each component of $H(b \mid u)$ is continuously differentiable with respect to b.

A19. For each $u \in U_\kappa$, $H(\beta(u) \mid u)$ is positive definite.

Assumption A1 describes the sampling assumptions as well as properties of X and B. Assuming that X_1 has support $I\!R$ is done for ease of exposition. X_1 can have bounded support, in which case a straightforward adaptation of Corollary 3.1.1 in Horowitz (1998) suffices for identification. Trimming is also more complicated in the bounded support case. It is convenient, but not necessary to include an intercept term. Assumption A2 concerns localization. Lemma 0 in Appendix A states conditions under which A2 holds. A3 describes the relationship between Z and X. A4 and A5, along with A11 and A12, serve to identify $\beta(u)$ for each $u \in U_\kappa$. A11 is CMI and A12 is CMZ, with U in place of V. These last two conditions imply the key identification condition

$$med(\epsilon(u) \mid X = x, U = u) = 0 \tag{16}$$

for all $x \in S_X$ and $u \in S_U$, where $\epsilon(U) = -X(B - \beta(U))$. Note that (16) is just (3) with U in place of V. Assumptions A6 and A7 are used to show the negligibility of a bias term in the proof of (12). Assumptions A8 through A10 relate to the conditional expectation $\beta(u)$. A8 imposes smoothness assumptions on $\beta(u)$ needed to carry out various Taylor expansions. A9 is a standard compactness assumption. A10 is a mild regularity condition used in establishing (13). A13 gives sufficient conditions on bandwidths α_n, σ_n, and τ_n for allowable values for e and c.

Note that $\sigma < \tau$. This implies that $n^{-\sigma} \gg n^{-\tau}$. That is, σ_n, the bandwidth used to smoothly approximate $\{t > 0\}$, is asymptotically wider than τ_n, the bandwidth used to localize on U. This is used to handle a bias term in a consistency proof. Note that the bandwidths must be wide to overcome the estimation uncertainty introduced by the generated regressor vector $\hat{U}$. For example, if $e = c = 1$ and $\alpha = \sigma = \tau - \epsilon$, for ϵ very small positive, then A13 states that the bandwidths must be proportional to $n^{-\alpha}$ where $\alpha < 1/18$. We conjecture that the values of α, σ, and τ in A13 can all be multiplied by two using more refined uniformity methods such as those in Bickel and Rosenblatt (1973), but we do not pursue this here. A14 gives conditions on the kernel functions. Note that we use the same univariate kernel function, namely $\mathcal{K}_\alpha$, for each continuous component of Z. Likewise, we use the same univariate kernel function, $\mathcal{K}_\sigma$, for each component of U. This is done for notational convenience. Bias reducing kernels satisfying the stated conditions are found in Müller (1984). A15 gives conditions on the joint density of the continuous components of Z given the discrete components of Z, as well as conditions on $U(x, z)$. These conditions are used along with A14 to show that the bias term $I\!E\hat{U}(x, z) - U(x, z)$ is negligible. A16 gives smoothness conditions on the components of the joint density of U given Z and $X\beta(u)$. It also gives conditions on a certain conditional expectation. Note that $A16$ implies that $d \geq 1$. That is, at least one of the endogenous regressors must be continuously distributed. A17 gives smoothness conditions on other densities. A16 and A17 are used along with A14 to show that a bias term consisting of a certain expected gradient is negligible. A18 and A19 give regularity conditions on the Hessian of the population analogue of the sample criterion function $S_n(b \mid u)$.

Under assumptions A1 through A19, we can show that $\hat{\beta}_\kappa - \beta_\kappa = \frac{1}{n}\sum_{i=1}^n f_n(W_i) + o_p(1/\sqrt{n})$ where $W_i = (Y_i, X_i, Z_i, U_i)$ and $f_n(W_i)$ has mean zero and finite variance. Moreover, we can show that $f_n(W_i) = f^{(1)}(W_i) + f_n^{(2)}(W_i) + f_n^{(3)}(W_i) + f_n^{(4)}(W_i)$ where each term has mean zero and finite

variance. The term $f^{(1)}(W_i)$ comes from the second term in (9), both $f_n^{(2)}(W_i)$ and $f_n^{(3)}(W_i)$ come from the second term in (10), and $f_n^{(4)}(W_i)$ comes from the term in (11). It follows from a standard CLT and Slutsky's theorem that

$$\sqrt{n}(\hat{\beta}_\kappa - \beta_\kappa) \rightsquigarrow N(0, \Sigma)$$

where $\Sigma = I\!E f_n(W) f_n(W)'$. This result holds for $e > 0$ and $c \geq 0$.[9] The cases $e > 0$ and $c = 0$ require the choice of bandwidths σ_n and τ_n for smoothly approximating the indicator $\{t > 0\}$ and localizing on U, respectively. The cases $e > 0$ and $c > 0$ are more complex to analyze since they require the choice of bandwidths α_n, σ_n, and τ_n for estimating U, approximating $\{t > 0\}$, and localizing on U, respectively. Proving the asymptotic results in complete generality is quite tedious and adds nothing to understanding. Because of this, we prove the result for the case $e = c = 1$, while letting k, the number of components of X, and m, the number of components of Z, be arbitrary. The arguments used to handle this special case readily extend to cover general $e > 0$ and $c > 0$.

Consider the special case $e = c = 1$ with k and m arbitrary. Assumption A16 requires that $d \geq 1$, and so when $e = 1$ we require that $d = 1$. In this case, $\mathcal{X} = \mathcal{X}^C$, $U = M(V)$ denotes the scalar random error associated with $\mathcal{X}$, and from A15, Z^C denotes the single continuous instrument for $\mathcal{X}$. For ease of notation, we write $\mathcal{Z}$ for Z^C. Recall from A14 the definitions of $\mathcal{K}_\alpha$, $\mathcal{K}_\sigma$, and $\mathcal{K}_\tau$. For ease of notation, we suppress the subscripts α, σ, and τ, because each kernel function is easily identified by its argument. Thus, we write $\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_i)$ for $\mathcal{K}_\alpha((\mathcal{Z}_k - \mathcal{Z}_i)/\alpha_n)$, $\mathcal{K}_n((X_j\beta(U_i))$ for $\mathcal{K}_\sigma(X_j\beta(U_i))/\sigma_n)$, and $\mathcal{K}_n(U_j - U_i)$ for $\mathcal{K}_\tau((U_j - U_i)/\tau_n)$.

[9] The result also holds for the case $e = 0$ of no endogenous regressors, but under slightly different conditions. We briefly discuss these conditions at the end of this section.

We begin by noting that the term $f^{(1)}(W_i)$ has the same form for all values of e and c. We get

$$f^{(1)}(W_i) = \beta(U_i)\tau_\kappa(\mathcal{X}_i)\tau_\kappa(\mathcal{Z}_i) - \beta_\kappa\,.$$

We now give the forms for $f_n^{(2)}(W_i)$, $f_n^{(3)}(W_i)$, and $f_n^{(4)}(W_i)$ for the special case $e = c = 1$, with k and m arbitrary.

We start with the most complicated term, $f_n^{(2)}(W_i)$. For $i \neq j \neq k \neq i$, define the function

$$\begin{aligned} f_n(W_i, W_j, W_k) &= \frac{\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)\tau_\kappa(\mathcal{Z}_j)}{\alpha_n \sigma_n \tau_n^2} H(\beta(U_i) \mid U_i)]^{-1} \\ &\times (2Y_j - 1)\mathcal{K}_n(X_j\beta(U_i))\tilde{X}_j'\mathcal{K}_n'(U_j - U_i) \\ &\times \frac{\{\mathcal{X}_k \leq \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\} - I\!E_j\{\mathcal{X}_k \leq \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\}}{f(\mathcal{Z}_j, D_j)} \end{aligned} \tag{17}$$

where the expectation $I\!E_j$ in the last line of the display is an expectation over W_k given W_j.

Next, for (i, j, k, l) with no equal components, define the function

$$\begin{aligned} f_n(W_i, W_j, W_k, W_l) &= \frac{\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)\tau_\kappa(\mathcal{Z}_j)}{\alpha_n^2 \sigma_n \tau_n^2} [H(\beta(U_i) \mid U_i)]^{-1} \\ &\times (2Y_j - 1)\mathcal{K}_n(X_j\beta(U_i))\tilde{X}_j'\mathcal{K}_n'(U_j - U_i) \\ &\times \frac{\{\mathcal{X}_k \leq \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\}}{f(\mathcal{Z}_j, D_j)} \\ &\times \frac{I\!E_j\mathcal{K}_n(\mathcal{Z}_l - \mathcal{Z}_j)\{D_l = D_j\} - \mathcal{K}_n(\mathcal{Z}_l - \mathcal{Z}_j)\{D_l = D_j\}}{f(\mathcal{Z}_j, D_j)} \end{aligned} \tag{18}$$

where the expectation $I\!E_j$ in the last line of the display is an expectation over W_l given W_j.

Finally, define $f_n^{(2)}(W_i) = f_n(P, P, W_i) + f_n(P, P, P, W_i)$, where, for example, $f_n(P, P, W_i)$ is the expectation of $f_n(W_i, W_j, W_k)$ over W_i and W_j given W_k, evaluated at $W_k = W_i$.

Next, we define $f_n^{(3)}(W_i)$. For $i \neq j$, define the function

$$f_n(W_i, W_j) = \frac{\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)\tau_\kappa(\mathcal{Z}_j)}{\sigma_n \tau_n}[H(\beta(U_i) \mid U_i)]^{-1}(2Y_j - 1)\mathcal{K}_n(X_j\beta(U_i))\tilde{X}_j'\mathcal{K}_n(U_j - U_i)\,. \tag{19}$$

Define $f_n^{(3)}(W_i) = f_n(P, W_i) - f_n(P, P)$.

Finally, we define $f_n^{(4)}(W_i)$. Let $\delta(u) = \frac{\partial}{\partial u}\beta(u)$. For $i \neq j$ and $i \neq j \neq k \neq i$, define the functions

$$\begin{aligned} f_n(W_i, W_j) &= \frac{\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)}{\alpha_n}\delta(U_i) \qquad (20)\\ &\times \frac{\{\mathcal{X}_j \leq \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\} - I\!E_i\{\mathcal{X}_j \leq \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\}}{f(\mathcal{Z}_i, D_i)} \\ f_n(W_i, W_j, W_k) &= \frac{\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)}{\alpha_n^2}\delta(U_i) \qquad (21)\\ &\times \frac{\{\mathcal{X}_j \leq \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\}}{f(\mathcal{Z}_i, D_i)} \\ &\times \frac{I\!E_i\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_i)\{D_k = D_i\} - \mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_i)\{D_k = D_i\}}{f(\mathcal{Z}_i, D_i)} \end{aligned}$$

where $I\!E_i$ in $f_n(W_i, W_j)$ denotes expectation over W_j given W_i, and $I\!E_i$ in $f_n(W_i, W_j, W_k)$ denotes expectation over W_k given W_i. Define $f_n^{(4)}(W_i) = f_n(P, W_i) + f_n(P, P, W_i)$.

We are now in a position to state the formal result.

THEOREM 1. *Let $e = c = 1$ with k and m arbitrary. If assumptions A1 through A19 hold, then*

$$\sqrt{n}(\hat{\beta}_\kappa - \beta_\kappa) \rightsquigarrow N(0, \Sigma)$$

where $\Sigma = I\!E f_n(W_i)f_n(W_i)'$ with $f_n(W_i) = f^{(1)}(W_i) + f_n^{(2)}(W_i) + f_n^{(3)}(W_i) + f_n^{(4)}(W_i)$.

The proof of this result is given in Appendix A. A consistent estimator of Σ is obtained by

replacing, in the expressions above, Ps with P_ns, U_is with $\hat{U}_i$s and $I\!E$s with the corresponding sample averages.

REMARKS. Various special cases of the $\sqrt{n}$-consistency and asymptotic normality result for $\hat{\beta}_\kappa$ are of interest.

Consider the binary response model analyzed by Manski (1975, 1985) and Horowitz (1992). In this model, $Y^* = X\beta + \epsilon$, where X is exogenous and $\beta = (1, \beta_2, \beta_3, \ldots, \beta_k)'$ is a vector of nonrandom coefficients defined implicitly by the median independence restriction $med(\epsilon \mid X = x) = 0$ for each $x \in S_X$. In the terminology of the random coefficients model, the only random coefficient is $B_2 = \beta_2 + \epsilon$. That is, $B = \beta + (0, \epsilon, 0, \ldots, 0)$ and the object of interest is the nonrandom coefficient vector β. It is well known (see Horowitz, 1993) that no estimator of β can achieve $\sqrt{n}$-consistency under the assumptions maintained by these authors.

However, the localize-then-average procedure can produce a $\sqrt{n}$-consistent and asymptotically normal estimator of β if instruments exist and assumptions A1 through A19 hold. Suppose there is a component of X, say X_1, for which there exists an instrument Z that is not a component of X. Formally, X_1 plays the role of an endogenous regressor, even though X_1 is exogenous. Further, assume that A1 through A19 hold where, for convenience, we take $\beta(U)$ to be the componentwise median of B. That is, $\beta(U) = (1, \beta_2 + med(\epsilon \mid U), \beta_3 \ldots \beta_k)$. Note that for this choice of $\beta(U)$, the conditional median zero assumption A12 automatically holds. This implies, in particular, that no symmetry assumption on the distribution of ϵ given U is required. We get

$$\beta_\kappa \equiv I\!E\beta(U)\tau_\kappa(\mathcal{X}^C, Z^C) = \beta I\!E\tau_\kappa(\mathcal{X}^C, Z^C) + (0, I\!Emed(\epsilon \mid U)\tau_\kappa(\mathcal{X}^C, Z^C), 0, \ldots, 0)\,.$$

It follows from Theorem 1 that for all $\kappa > 0$, $\hat{\beta}_\kappa$ is a $\sqrt{n}$-consistent and asymptotically nor-

mal estimator of β_κ. Deduce that $\hat{\beta}_\kappa / \hat{I\!E}\tau_\kappa(\mathcal{X}^C, Z^C)$, where $\hat{I\!E}\tau_\kappa(\mathcal{X}^C, Z^C) = \frac{1}{n}\sum_{i=1}^n \tau_\kappa(\mathcal{X}_i^C, Z_i^C)$, provides a $\sqrt{n}$-consistent and asymptotically normal estimator of the vector of nonrandom slope coefficients $(\beta_3, \ldots, \beta_k)$. The additional information provided by the instruments and assumptions, together with the localize-then-average estimation procedure which exploits this extra information, makes this possible. Suppose that $med(\epsilon \mid U) \equiv I\!E(\epsilon \mid U)$, and we make the additional palatable assumption that $I\!E(\epsilon \mid X, Z) \equiv 0$. Since U is a deterministic function of X and Z, it follows that $I\!E(\epsilon \mid U) \equiv 0$. In this case, β_κ reduces to $\beta I\!E\tau_\kappa(\mathcal{X}^C, Z^C)$, and so $\hat{\beta}_\kappa / \hat{I\!E}\tau_\kappa(\mathcal{X}^C, Z^C)$ is a $\sqrt{n}$-consistent and asymptotically normal estimator of the entire vector β.

Next, consider the same set-up as in the previous paragraph, except that X contains endogenous components. This model includes the very interesting subcase where one of the endogenous regressors is a binary treatment. Under A1 through A19, for all $\kappa > 0$, $\hat{\beta}_\kappa / \hat{I\!E}\tau_\kappa(\mathcal{X}^C, Z^C)$ provides a $\sqrt{n}$-consistent and asymptotically normal estimator of the nonrandom slope coefficient vector $(\beta_3, \ldots, \beta_k)$.

Finally, consider the independent random coefficients model, where X is independent of B. As in the case of the correlated random coefficients model, we exploit instruments to localize on U and then average to achieve a $\sqrt{n}$-consistent and asymptotically normal estimator of β_κ, a robust measure of center of the distribution of B.

4. Simulations

In this section, we report results of several simulations exploring aspects of the finite sample behavior of the estimator $\hat{\beta}_\kappa$. We study the effects of varying bandwidths, sample sizes, and the number of estimated coefficients. We also compare $\hat{\beta}_\kappa$ to two estimators of nonrandom coefficients in a binary response model. While these estimators are not designed to handle correlated random coefficients, they can account for endogeneity and so may be considered competitors of $\hat{\beta}_\kappa$ in

practice. Specifically, we consider the parametric estimator of Rivers and Voung (1988) and the semiparametric 2SLS estimator assuming the linear probability model.

We start by considering the following DGP: For $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \{X_{i1} + B_{i2} + B_{i3}X_{i3} > 0\} \\ X_{i1} &= Z_{i1} + V_i \,. \end{aligned}$$

We consider the case of one endogenous regressor. We take X_{i1} to be the endogenous regressor and the vector of instruments $Z_i = (Z_{i1}, X_{i3})$. We specify the random coefficients to be functions of V_i to generate the correlation with X_{i1}. Specifically, for $i = 1, \ldots, n$ and $j = 2, 3$, we take $B_{ij} = 1 + V_i + \nu_{ij}$ where the ν_{ij} are exogenous components. To summarize the dependence structure, we assume that the drivers of the data (Z_i, V_i, ν_i) are iid draws from the following distribution: $V_i \sim U[-.5, .5]$, $\nu_i \sim N(\mathbf{0}_2, \Sigma_\nu)$, and $Z_i \sim N(\mathbf{0}_2, \Sigma_Z)$ where V_i, ν_i, and Z_i are mutually independent. Here, $\mathbf{0}_k$ is a vector of zeros of length k, $\Sigma_\nu = 0.5\Sigma_Z$, and

$$\Sigma_Z = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}.$$

Note that in this simulation, the endogenous regressor is an additively separable function of instruments and control variables. In the notation of the paper, $X_1 = \phi(Z, V_1) = I\!E(X_1 \mid Z) + V_1$ where $I\!E(X_1 \mid Z) = Z_1$ and $V_1 = X_1 - I\!E(X_1 \mid Z)$. In this standard set-up, it is straightforward to directly estimate V_1 with $\hat{V}_1 = X_1 - \hat{I\!E}(X_1 \mid Z)$ where $\hat{I\!E}(X_1 \mid Z)$ is the usual kernel regression estimator of the conditional mean $I\!E(X_1 \mid Z)$. In other words, it is just as easy to localize directly on $\hat{V}_1$ as it is to localize indirectly on $\hat{U}_1 = \hat{I\!P}\{X_1 \leq x \mid Z = z\}$, as in the main text. Moreover,

either localization is valid. All definitions and the theoretical results go through with $\hat{V}_1$ in place of $\hat{U}_1$ and V_1 in place of U_1. We illustrate the use of both localizations in our simulations.

Both $\hat{V}_1$ and $\hat{U}_1$ are estimated using a product kernel with higher order "Epanechnikov type" kernel in every dimension. The individual kernel function is

$$\begin{aligned} K(t) &= c_0(6864 - 240240\,t^2 + 2450448\,t^4 - 11085360\,t^6 \\ &+ 25865840\,t^8 - 32449872\,t^{10} + 20801200\,t^{12} - 5348880\,t^{14})\{-1 \le t \le 1\} \end{aligned}$$

where $c_0 = 0.0006712228$. We use this same kernel to localize. In addition, we take K^* in (6) to be the integral of this kernel. This kernel satisfies the conditions of A13 and A14 when $c = 2$, $e = 1$, and $\alpha = \sigma = \tau - \epsilon$, for ϵ very small positive. The optimization is performed via the R routine *nlminb*, a program for unconstrained and constrained optimization using PORT routines.[10] We need to select the grid over which the search is performed, and the absolute or relative numerical tolerance. We have chosen the first to be the rectangle $[-10, 10] \times [-10, 10]$, while we specified the absolute tolerance to be e^{-15}. We take $\tau_\kappa(Z_i) = \{|Z_{i1}| \le \kappa\}\{|Z_{i2}| \le \kappa\}$ and $\tau_\kappa(X_{i1}) = \{|X_{i1}| \le \kappa\}$ where $\kappa = 3.25$. This corresponds to trimming 5.2% of the data.

The results in the following table are for localization on $\hat{U}_1$, with a sample size of $n = 2000$ observations and 100 Monte Carlo repetitions. Recall that α_n denotes the bandwidth for estimating $\hat{U}_1$, σ_n the bandwidth for the integrated kernel function K^*, and τ_n the bandwidth for localizing. In all the reported simulations we take $\alpha_n = 1.5$ and vary the values of σ_n and τ_n as indicated in the tables. In other simulations which we do not report, we varied α_n and found that results were robust to moderate changes in this bandwidth.

[10] A detailed description of the PORT-routines involved can be found in "Usage Summary for Selected Optimization Routines", at http://netlib.bell-labs.com/cm/cs/cstr/153.pdf.

Table 1: RMSE results for $\hat{\beta}_\kappa$ at different bandwidths for $n = 2000$ (localizing on $\hat{U}_1$)

(τ_n, σ_n)	RMSE on Coefficient 1	RMSE on Coefficient 2
$(0.12, 2.0)$	0.1049993	0.1176348
$(0.16, 2.0)$	0.1134255	0.1114663
$(0.08, 2.0)$	0.1517702	0.1212661
$(0.24, 2.0)$	0.1137844	0.1168878
$(0.16, 1.4)$	0.1520292	0.1204596
$(0.24, 0.8)$	0.1913334	0.1534957
$(0.08, 4.0)$	0.1812628	0.2651299

Several things are noteworthy: First, there is a large area of plausible bandwidths where the MSE is of the same order of magnitude, meaning that our method is rather robust to changes in bandwidth. Second, the two coefficients are differently affected by the two bandwidths. Given that only the first regressor is endogenous, it is not surprising that the coefficients on the two regressors exhibit different behavior. Third, compared to an estimator which does not use trimming, the RMSE is slightly smaller, and in particular, the bias reduces significantly.[11]

Qualitatively similar results were obtained with $n = 500$ observations. To focus on essentials, we do not report the table, but note that the RMSE is again rather insensitive to bandwidth choice in a large area. The RMSE becomes uniformly larger, compared to $n = 2000$. At the optimal bandwidth $(0.16, 2.5)$, the RMSE of the first components of $\hat{\beta}_\kappa$ is 0.225062 while the RMSE on the second component is 0.190486. The steady improvement with sample size is corroborated for

[11] Results are available from the authors upon request. The estimator also becomes less sensitive to bandwidth choice, which reflects the fact that the model becomes more robust as tail behavior plays less of a role.

$n = 5000$. At the optimal bandwidth $(0.08, 2.0)$ the RMSE on the first coefficient is 0.06960242, while the RMSE on the second is 0.07380443. As expected, the RMSE decreases steadily as the sample size increases.

It is interesting to compare the estimator with alternative estimators. The first comparison is with a Linear Probability Model (LProb). Among the advantages conventional wisdom attributes to the LProb is that it lets one tackle the issue of endogeneity by using standard IV methods. The logic is that, despite its apparent misspecification in terms of the discrete nature of the dependent variable, the LProb captures endogeneity well. That is, provided the misspecification in the error distribution is not too severe (e.g., we are in the center of the cdf of an unimodal symmetric cdf), it provides a way to capture the first order effect due to endogeneity. Note that this is the case with our simulation design, and so a priori one might expect somewhat satisfactory performance from the LProb.

However, our results strikingly belie these expectations. Running a standard 2SLS model and normalizing all coefficients such that the coefficient on X_1 is unity, we obtain the following results:

Table 2: RMSE results for the LProb procedure at different sample sizes

Sample Size	RMSE on Coefficient 1	RMSE on Coefficient 2
500	2.177558	0.3459466
2000	2.151667	0.3226654
5000	2.159087	0.3058008

Observe that the first coefficient in particular is estimated much more imprecisely than the second one. In particular, the means of the intercept (whose true value is 1) stay close to 3 (3.091092, 3.084162, and 3.091984, respectively) and so are severely biased. Also, compared to $\hat{\beta}_\kappa$,

the second coefficient has roughly two to five times the RMSE, while the first has more than ten times the RMSE. This reflects the feature that endogeneity, and apparently also misspecification, seems to have a much stronger effect on the intercept. Finally, note that both RMSEs decrease slowly with sample size, indicating that the main problem is the fairly persistent bias, which in the case of the second coefficient diminishes very slowly. In particular, the mean is 0.6235943, 0.6225125, and 0.6428279, respectively.

The second commonly used estimator with which we compare our results is the estimator of Rivers and Voung (1988), henceforth, RV. This estimator features a probit, but adds an estimated control function residual as additional regressor, and probably corresponds to standard practise in good applied work. Note that, compared to our approach, the RV estimator is twice misspecified. First, it does not account for the heteroscedasticity that inevitably results from the random coefficients structure; second, it imposes a parametric structure on the link function (i.e., probit).

With this procedure, we obtain the following results:

Table 3: RMSE results for the Rivers-Voung procedure at different sample sizes

Sample Size	RMSE on Coefficient 1	RMSE on Coefficient 2
500	1.066009	0.3608533
2000	1.091336	0.3460289
5000	1.125813	0.3322708

It is noteworthy that the bias of the estimator for the intercept increases, with the mean of the intercept moving further away from the true mean of 1 (-0.1003702, -0.146636 and -0.17692040, respectively) as n increases, while the mean of the second coefficient stays roughly constant (0.6028553, 0.6095537, and 0.6159592, respectively). This is also reflected in the RMSE which is two to three

times as large as the corresponding RMSE for $\hat{\beta}_\kappa$. Once again, most of the RMSE is due to the bias component.

To represent our results graphically, we display estimated densities in Figures 1 through 6. In these graphs, the estimator $\hat{\beta}_\kappa$ is referred to as the Hoderlein-Sherman estimator. The graphs differ according to sample size. Figures 1 through 3 show the behavior of the RV estimator for the first coefficient compared to ours (the density of the LProb is in fact so far outside of the graph that we do not display it here) at $n = 500, 2000, 5000$. The true trimmed mean, β_κ, in our setup is 0.948. The estimated means for $\hat{\beta}_\kappa$ are 0.8622154, 0.904023 and 0.9158743, respectively, for the first coefficient (the intercept), while the means of the RV estimator remain slightly below zero regardless of sample size. In fact, the bias seems to increase slightly in absolute value. Figures 4 through 6 show the same results for the second coefficient, but now including the LProb, which in fact is slightly less biased than the RV estimator, though both have a downward bias of around 0.4, while $\hat{\beta}_\kappa$ has a slight and rapidly diminishing upward bias, with values 1.029436, 0.961271 0.9585179, respectively, while the true trimmed mean $\beta_\kappa = 0.948$. We see that both misspecified estimators exhibit a significant downward bias, though slightly less variance due to the parametric nature of the estimators. The fact that the variance vanishes for all estimators can be clearly seen by the fact that the area of positive density is diminishing. As expected the parametric estimators are less dispersed, though not by too much. In summary, the simulation results confirm the consistency of $\hat{\beta}_\kappa$ and the inconsistency of two likely competitors, both of which exhibit a large and persistent bias.

Next, we perform an exercise where we make explicit use of the additive form in the IV equation. Moreover, we slightly stack the deck against $\hat{\beta}_\kappa$ by ignoring the fact that $\hat{\beta}_\kappa$ estimates the trimmed mean β_κ, rather than the mean β, and we report all RMSE results relative to the β. Thus the

RMSE results for $\hat{\beta}_\kappa$ are slightly inflated. The DGP is exactly as before, but we change the uniform distribution of the residual to $V_i \sim N(0,1)$. The estimation procedure is exactly as before, except now we localize directly on the $\hat{V}_i$, the estimated mean regression residuals. We consider this setup particularly relevant for applications.

For $\hat{\beta}_\kappa$, we obtain the following results at the optimal bandwidths:

Table 4: RMSE results for $\hat{\beta}_\kappa$ for different sample sizes
(localizing on $\hat{V}_1$)

Sample Size and Bandwidth	RMSE on Coefficient 1	RMSE on Coefficient 2
$n = 500$, $(\tau_n, \sigma_n) = (1.0, 4)$	0.3198675	0.2658185
$n = 2000$, $(\tau_n, \sigma_n) = (0.8, 3)$	0.1999790	0.1607219
$n = 5000$, $(\tau_n, \sigma_n) = (0.7, 2.3)$	0.1765224	0.1069096

Comparing these RMSE results with those for the LProb and the RV estimators again clearly demonstrates the superiority of $\hat{\beta}_\kappa$. And this is in spite of the fact that the RMSE results for $\hat{\beta}_\kappa$ are inflated. The results for the LProb procedure are displayed in the following table.

Table 5: RMSE results for Lprob Procedure for different sample sizes

Sample Size	RMSE on Coefficient 1	RMSE on Coefficient 2
500	4.60288	0.48481
2000	4.41378	0.37517
5000	4.33741	0.34142

Not surprisingly, the results are very similar to those previously obtained. In fact, the results are even less flattering to the LProb procedure. The same is true for the RV estimator, as the following table illustrates.

Table 6: RMSE results for RV Procedure for different sample sizes

Sample Size	RMSE on Coefficient 1	RMSE on Coefficient 2
500	0.77488	0.51530
2000	0.73682	0.43207
5000	0.72775	0.40373

Observe that even compared to the untrimmed population mean of unity, the RMSE of $\hat{\beta}_\kappa$ is at most half as large, and in some cases one tenth of the RMSE of the LProb and the RV estimators. In particular, the bias of the latter two estimators is much larger, even when we ignore the fact that $\hat{\beta}_\kappa$ estimates the trimmed mean $\beta_\kappa = .948$.

Finally, we examine the effect of increasing the number of regressors in the previous set-up. Considering the following DGP: For $i = 1 \ldots, n$, we take

$$\begin{aligned} Y_i &= \{X_{i1} + B_{i2} + B_{i3}X_{i3} + B_{i4}X_{i4} + B_{i5}X_{i5} + B_{i6}X_{i6} > 0\} \\ X_{i1} &= Z_{i1} + X_{i4} + V_i \,. \end{aligned}$$

As before, we take X_{i1} to be the single endogenous regressor, and we take the vector of instruments to be $Z_i = (Z_{i1}, X_{i2}, ..., X_{i5})$. We specify the random coefficients as functions of V_i to generate the correlation with X_{i1}. Specifically, for $i = 1, \ldots, n$ and $j = 1, \ldots, 5$ we take $B_{ij} = 1 + V_i + \nu_{ij}$ where the ν_{ij} are exogenous components. To summarize the dependence structure, we assume that the drivers of the data (Z_i, V_i, ν_i), $i = 1, .., n$, are iid draws from the following distribution: $V_i \sim N(0, 1)$, $\nu_i \sim N(\mathbf{0}_5, \Sigma_\nu)$, and $Z_i \sim N(\mathbf{0}_5, \Sigma_Z)$ where V_i, ν_i, and Z_i are mutually independent.

Here $\mathbf{0}_5$ is a vector of zeros of length 5, $\Sigma_\nu = 0.5\Sigma_Z$, and

$$\Sigma_Z = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 \\ 1 & 2 & 0 & 0 & 0 \\ 0 & 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 2 & 1 \\ 0 & 0 & 1 & 1 & 2 \end{bmatrix}.$$

Finally, when implementing the estimator, we follow exactly the same strategy as above, and the kernels are again product kernels of the higher order Epanechnikov type kernel variety in every dimension, where we use the same higher order Epanechnikov integral kernel as above.[12]

Table 7: RMSE results for $\hat{\beta}_\kappa$ for different sample sizes
(localizing on $\hat{V}_1$)

Sample Size	1	2	3	4	5
$n = 500$	0.56025	0.42609	0.43926	0.38557	0.39332
$n = 2000$	0.36793	0.31890	0.27856	0.26856	0.25894
$n = 5000$	0.32034	0.23281	0.20177	0.19483	0.19227

The reduction in RMSE with increasing sample size is obvious. Note also that due to the largely symmetric setup, all four slope coefficients are equally affected, even those whose regressors less correlated with the endogenous one, X_{i1}. Observe, however, that the intercept exhibits a larger RMSE, as was also the case in the two-dimensional model.

[12] Technically, A13 and A14 must be satisfied for $c = 5$ and $e = 1$, but, in terms of bias reduction, there is no practical difference between using the previous kernel and using a technically correct one.

5. Application: The Effect of Smoking on Abnormal Birth Conditions

Among the most harmful effects of smoking are the effects of a mother's smoking during pregnancy on the health of her newborn child. Therefore, these effects have been extensively studied in the health economics literature (e.g., Rosenzweig and Schultz (1983), Evans and Ringel (1999), Lien and Evans (2005)). In this paper, we study whether children are more likely to be born with abnormal birth conditions, like congenital abnormalities or low birth weight, when their mothers increase the number of cigarettes smoked during pregnancy. The goal of our analysis is to provide a detailed assessment of the effect of smoking. In particular, we want to allow for heterogeneous marginal effects of an extra cigarette smoked, and for endogenous choice of the number of cigarettes smoked, reflecting the fact that smoking more is a deliberate decision. This is important because women who smoke more may be more likely to exhibit other types of behavior that increase the risk of health defects in their newborns. That is, we suspect that the coefficient of the regressor "number of cigarettes smoked daily" is positively correlated with this regressor, and so our methods apply. Throughout this section, we condition on mothers who already smoke, since we want to isolate the effect of smoking an additional cigarette each day, not the effect of smoking as opposed to not smoking.

To account for the endogenous nature of the regressor "number of cigarettes smoked daily", we use an older idea of Evans and Ringel (1999) who use cigarette excise tax rate as a source of exogenous variation to mitigate confounding factors in identifying the effects of smoking. To see why this is a sensible instrument, observe that this tax is set on the state level by the state government. Since setting the cigarette tax is only one of many issues, and certainly not an issue that decides elections or causes people to move on a large scale, it can be seen as exogenous to the individuals' choice set. We follow this idea, letting the tax rate (denoted Z_1) be the principal instrument for

number of cigarettes smoked per day (X_3), the principal endogenous regressor. Other variables included in the model are an indicator of alcohol consumption during pregnancy (X_1), the number of live births experienced (X_4), and mother's age (X_5). The latter three variables are associated with congenital abnormalities or abnormal conditions at birth. We let Y denote an indicator of the above mentioned abnormalities at a child's birth.

The causal model is thus given by

$$\begin{aligned} Y &= \{B_1^* X_1 + B_2^* + B_3^* X_3 + B_4^* X_4 + B_5^* X_5 > 0\} \\ X_1 &= I\!E(X_1 \mid Z) + V \end{aligned}$$

where $V = X_1 - I\!E(X_1 \mid Z)$ and the instrument vector $Z = (Z_1, X_1, X_4, X_5)$. The components of $B^* = (B_1^*, \ldots, B_5^*)'$ are unobserved factors related to the lifestyle of the mother that impact the child's health at birth. Subsequently, we compare our estimator of a trimmed mean of B, where $B = B^*/B_1^*$, to analogous estimators based on Rivers and Vuong (1988) and 2SLS. These comparison estimators directly estimate five unknown parameters, corresponding to each of the regressors in the model.

5.1 Description of Data and Variables

We use a cross section of the natality data from the Natality Vital Statistics System of the National Center for Health Statistics. From this data set we extract a random sample of size 100,000 from the time period between 1989 to 1999.

We focus on the subset of this sample who smoke, since there are serious discontinuities in terms of unobservables near zero.[13] If interest centered on the total effect of smoking, then the

[13]The difference between the population that does not smoke and the one that smokes one cigarette a day is

decision to smoke would have to be modeled as well. However, we argue that understanding the effect of smoking an additional cigarette for the subpopulation of smokers is of interest, since any government tax measure aimed at reducing adverse health consequences may primarily affect the amount smoked rather than the decision to smoke or not.

Given our simulation results, we think that a subsample of smokers of size $n = 10000$ is sufficient to obtain a precise estimate, and we therefore draw such a subsample randomly without replacement from the data. The descriptive statistics for this sample can be found in Table 12.

5.2 Empirical Results

Recall that our aim is to determine the effect of smoking more cigarettes on abnormal birth conditions. To show the performance of our estimator in this regard, it is instructive to start with the standard practise of estimating a linear probability model. This procedure suffers from misspecification that yields highly implausible results. We condition on the subsample of males. The result is as follows:

Table 8: OLS estimates of Linear Probability Model

Variable	Estimate	Std. Error	t value	p value
Intercept	0.126150	0.0210541	5.992	0.000
Number of cigs	−0.000532	0.0005853	−0.909	0.3633
Age of mother	−0.000750	0.0008328	−0.901	0.3677
Number of births	0.002097	0.0040905	0.513	0.6081
Alcohol	0.056212	0.0235104	2.391	0.0168

qualitatively different from the difference between the population that smokes one cigarette a day and the one that smokes two cigarettes a day.

Observe that the only significant coefficients are those on the alcohol indicator and the intercept. Indeed, the former shows the expected positive sign: if women drink alcohol during pregnancy, the children are more likely to develop abnormal birth conditions. No other variables show a significant effect - including number of cigarettes. Indeed, even very heavy smokers have the same probability of giving birth to healthy children after a full 40 weeks of pregnancy. According to these results, they may even be more likely to have healthy children, since the point estimate is negative.

An obvious flaw in this approach is that it does not account for endogeneity in the choice of the number of cigarettes smoked. We now try to correct for endogeneity using 2SLS. The results are displayed in Table 9:

Table 9: Linear Probability Model - 2SLS

Variable	Estimate	Std. Error	t value	p value
Intercept	0.123846	0.021232	5.833	0.000
Number of cigs	−0.001376	0.001162	−1.184	0.2364
Age of mother	−0.000683	0.000836	−0.816	0.4143
Number of births	0.002970	0.004220	0.704	0.4815
Alcohol	0.056432	0.023512	2.400	0.0164

The results are qualitatively the same as those for the OLS estimator. Note that the estimate of the coefficient of number of cigarettes is actually more negative and slightly more significant, suggesting that more smoking, if it has any systematic effect, is actually beneficial to the health of newborns. Clearly a nonintuitive result. Since we want to compare the effects with our theoretically superior estimator, and we know that β_κ is only identified up to scale, for reference we divide all coefficients by the coefficient on alcohol. We obtain values of 2.194613, -0.02438979, -0.0121049, and 0.05264387, in the order of appearance.

We conclude that the linear probability model shows strong signs of misspecification. But as we have seen from the simulation evidence, we would also expect the linear probability model to be the most biased. Hence we implement the less biased model of Rivers and Voung (1988), using a probit and nonparametric control function residuals estimated in the same fashion as in the simulation part and enter additively. The results are displayed in the following table.

Table 10: Rivers Voung Estimator

Variable	Estimate	Std. Error	t value	p value
Intercept	−1.14901	0.11721	−9.803	0.000
Number of cigs	−0.00786	0.00656	−1.199	0.2307
Age of mother	−0.00370	0.00463	−0.798	0.4251
Number of births	0.01596	0.02313	0.690	0.4903
Alcohol	0.27001	0.11668	2.314	0.0207
Control Function	0.13911	0.16507	0.843	0.3994

The main results very much correspond to the linear model. Alcohol is again the only significant explanatory variable. The odd effect of cigarettes and its rather low p-value is also in line with the above results. The misspecification seems to have a comparable effect. Finally, again for comparison, we have that the coefficient relative to alcohol are, again in order of appearance -4.255303, -0.02912374, -0.01370293, and 0.05911417. Note the similarity with the 2SLS estimates of the effect of number of cigarettes.

Finally, we implement $\hat{\beta}_\kappa$. The individual elements are exactly as in the simulation section. Since the coefficient is only identified up to scale and sign, we choose to normalize on alcohol. All scientific evidence as well as the rather shaky evidence obtained in this paper point to alcohol

having an adverse effect on the pregnancy. We hence assume that the sign is positive and the coefficient is unity. Moreover, we use the bootstrap to compute the variance σ^2_{boot}, and construct the confidence intervals using the normal approximation with $\pm 1.96\sigma_{boot}$.

Relative to the alcohol coefficient, we obtain the following results.

Table 11: $\hat{\beta}_\kappa$ with bootstrap standard errors

	Estimate	Bootstrap Std. Err	95% CI Lower Bound	95% CI Upper Bound
Intercept	−0.73961406	0.08695773	−0.9100512	−0.5691769
Nr cigarettes	0.25516531	0.03810400	0.1804815	0.3298491
Age mother	−0.05818467	0.03069635	−0.1183495	0.0019801
Nr. births	−0.04479768	0.04838861	−0.1396394	0.0500440

First, note that neither the estimated coefficient on age nor that on the number of births is significant at the 95% confidence level, though the former is significant at the 90% level. However, what is highly significant is the number of cigarettes. More important, the coefficient on number cigarettes has the expected sign, as well as a very plausible magnitude. One may object that the magnitude seems somewhat large. Note, however, that the alcohol indicator covers cases where individuals consume from small to large amounts, and it is not really clear whether a low level of consumption has a strong effect. Also, there may be a problem with underreporting so that there may be measurement error, whereas we expect the number of cigarettes to be more accurately reported, since the mother has already admitted to smoking. For both reasons, the coefficient on alcohol may in fact be larger. In either case, it is conceivable that smoking four additional cigarettes (say, six instead of two per day, i.e., switching from being an occasional smoker to a moderate one) has a negative effect which on average is as bad as consuming alcohol. The bottom line is that only

$\hat{\beta}_\kappa$ identifies a higher number of cigarettes to be a serious risk in the health development of a child, while the other two estimators do not identify smoking more as a health risk. The policy advice is clear: Pregnant women should reduce smoking as much as possible if they do not find the strength to quit outright.

Table 12: Descriptive Statistics

Variable	Description(10000 obs)	Mean	SD	min	max
fipsst	State of Residence			1	56
stoccfip	State of Occurrence			1	56
monpre	Month of Pregnancy Prenatal Care Began	2.72	1.55	0	9
nprevist	Total number of Prenatal Visits	11.1	4.14	0	45
weight		3162	581	225	5410
dplural	1=single 2=twin 3=triplet 4=quadruplet 5=Quintuplet or highrt	1.02	.162	1	4
apgar1	One minute Apgar Score (not observed after 1994)	8.01	1.32	0	10
apgar5	Five minute Apgar Score	8.95	.746	0	10
gestat	Gestation	39.0	2.81	18	47
momage	Age in years	25.7	5.73	13	47
momedu	Education in years	11.7	1.75	0	17
cigar	average number of cigarettes per day	12.3	7.72	1	75
alcohol	Alcohol use during pregnancy	.038	.191	0	1
drink	average number of drinks per day	.129	1.23	0	44
dadedu	(not observed after 1994)	11.8	1.82	0	17
male	newborn is male	.514	.499	0	1
lbw	1 if birth weight below 2500	.104	.305	0	1
ormoth	Hispanic Origin of Mother	.051	.413	0	5
mrace	1=white 2=black 3=american indian 4=chinese 5=japanese 6=hawaiian 7=filipino 8=other asian 9=all other race	1.28	3.03	1	78
biryr	year of birth	1993	3.10	1989	1999
lmpyr	year last normal menses began	1993	3.13	1988	1999
nlbnl	number of live birth, now living	1.17	1.22	0	11
married	1 if mother married	.617	.485	0	1
vaginal	method of delivery is vaginal	.788	.408	0	1
vbac	method of delivery is vaginal birth after previous C-section	.027	.162	0	1
primac	method of delivery is Primary C-section	.127	.333	0	1
repeac	method of delivery is Repeat C-section	.084	.277	0	1

Variable	Description(10000 obs)	Mean	SD	min	max
forcep	method of delivery is Forceps	.041	.199	0	1
vacuum	method of delivery is vacuum	.056	.230	0	1
weekday	day of week child born			1	7
datayear				1989	1999
mhisp	1 if race of mother is hispanic	.022	.147	0	1
mblack	1 if race of mother is black	.104	.305	0	1
masian	1 if race of mother is asian	.003	.058	0	1
multbirth	1 if multiple birth	.024	.155	0	1
abnormalc	1 if abnormal conditions of newborn reported	.083	.276	0	1
abnmcong	1 if congenital abnormalities reported	.017	.130	0	1
abnm	1 if abnormalc+abnmcong > 0	.094	.293	0	1
gestun37	1 if gestation < 37	.115	.320	0	1
state					
year		1993	3.10	1989	1999
sttax	state tax in cents per pack	26.3	17.7	2	100
pcpacks	pe capita packs	102	21.3	32.5	186
p	nominal price per pack	169	31.2	103	327
taxbyp	% of retail price that is tax	27.0	6.27	13.6	44.6
fedtax	federal tax per pack	21.8	3.09	16	24
tottax	total tax per pack	48.2	18.7	18	124

For more details, see User's Guide in http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm

6. Summary and Directions for Future Work

This paper identifies and estimates a measure of center of the distribution of random coefficients in a binary response model where the random coefficients are allowed to be correlated with regressors. Like Imbens and Newey (2009), we use instruments and control variables to break the dependence of regressors on coefficients. Independence of regressors and coefficients given controls implies conditional median restrictions which drive identification of the mean of the coefficients given controls. Averaging over controls identifies the mean of the coefficients. This suggests an

analogous two-stage, localize-then-average approach to estimation. We first estimate unknown controls, and then estimate conditional means with localized smooth maximum score estimators. We average the estimated conditional means to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator of a trimmed mean of the distribution of coefficients. The asymptotic analysis of the estimator, though conceptually fairly straightforward, is very complicated due to the fact that the estimator is an average of optimization estimators evaluated at generated variables. Technical novelties include localizing indirectly on a monotone transformation of the control variable, and establishing a uniform strong maximization condition for a uniform consistency proof. Simulations illustrate that the estimator performs well relative to likely competitors. We apply our estimator to data on the effect of mothers' smoking on the health of their newborns. We find that the new estimator gives more sensible estimates of the effect of smoking than those of likely competitors.

The results of this paper have interesting implications for more standard models. For example, with the addition of instruments, we show how to use the new estimator to obtain a $\sqrt{n}$-consistent and asymptotically normal estimator of the vector of slope parameters in the model proposed by Manski (1975,1985) and Horowitz (1992).

In future work, we plan to estimate other characteristics of the distribution of correlated random coefficients in the binary response model. In particular, we plan to adapt the localize-then-average procedure to estimate the distribution of each random coefficient, conditional on centered values of the other coefficients. We propose to do this by first estimating a center of the distribution as done in this paper. The center need not be the mean of the distribution and so need not require strong symmetry assumptions. Then, using the estimator of Kordas (2006) to estimate the quantiles of the distribution of a given random coefficient conditional on the centered values of the other coefficients, we recover the entire conditional distribution of that random coefficient.

Aliprantis, C., and Border, K. (1994): *Infinite Dimensional Analysis*, New York, Springer-Verlag.

Altonji, J., and Matzkin, R. (2005): "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, **73**, 1053–1102.

Bajari, P., Fox, J., Kim, K., and Ryan, S. (2011): "Discrete choice models with nonparametric distribution of random coefficients," Working paper, University of Minnesota.

Berge, P. (1997): *Topological Spaces*, New York, Dover.

Bickel, P., and Rosenblatt, M. (1973): "On some global measures of the deviations of density function estimates," *Annals of Statistics*, **1**, 1071–1096.

Blundell, R. and Powell, J. (2004): "Endogeneity in semiparametric binary response models," *Review of Economic Studies*, **71**, 655–679.

Evans, W. and Ringel, J. (1999): "Can higher cigarette taxes improve birth outcomes?" *Journal of Public Economics*, **72**, 135–154.

Fan, J., and Gijbels, I. (1996): *Local polynomial modeling and its applications*, Boca Raton, Chapman & Hall.

Fox, J., and Ghandi, A. (2010): "Nonparametric identification and estimation of random coefficients in nonlinear economic models," Working paper, University of Michigan.

Gautier, E. and Kitamura, Y. (2009): "Nonparametric estimation of random coefficients binary choice models," Cowles Foundation working paper, Yale University.

Heckman, J. and Vytlacil, E. (1998): "Instrumental variable methods for the correlated random coefficient model," *Journal of Human Resources*, **33**(4), 974–987.

Horowitz, J. (1992): "A smoothed maximum score estimator for the binary response model,"

Econometrica, 60(3), 505–531.

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

Horowitz, J. (1998): *Semiparametric Methods in Econometrics*, New York, Springer.

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

Imbens, G. and Newey, W. (2009): "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, **77**(5), 1481–1512.

Klein, R. and Spady, R. (1993): "An efficient semiparametric estimator for binary response models," *Econometrica*, **61**, 387–421.

Kordas, G. (2006): "Smoothed Binary Regression Quantiles," *Journal of Applied Econometrics*, **21**, 387–407.

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

Lien, D. and Evans, W. (2005): "Estimating the impact of large cigarette tax hikes: the case of maternal smoking and infant birth weight," *Journal of Human resources*, **40**, 373–392.

Manski, C. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics*, 3, 205–228.

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

Matzkin, R. (1992): "Nonparametric and distribution-free estimation of the binary choice model and the threshold crossing models," *Econometrica*, **60**(2), 239–270.

Matzkin, R. (2003): "Nonparametric estimation of nonadditive random functions," *Econometrica*,

71(5), 1339–1375.

Powell, J., Stock, J., and Stoker, T. (1989): "Semiparametric estimation of index coefficients," *Econometrica*, **57**, 1403–1430.

Petrin, A. and Train, K. (2010): "A control function approach to endogeneity in consumer choice models," *Journal of Marketing Research*, **47**(1), 3–13.

Rivers, D. and Vuong, Q. (1988): "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, **39**, 347–366.

Rosenzweig, M., and Shultz, T. (1983): "Estimating a household production function: heterogeneity, the demand for health inputs, and their effect on birth weight," *Journal of Political Economy*, **91**, 723–746.

Serfling, R. (1980): *Approximation Theorems of Mathematical Statistics*, New York, Wiley.

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

Vytlacil, E., and Yildiz, N. (2007): "Dummy endogenous variables in weakly separable models," *Econometrica*, **75**, 757–779.

In this appendix, we prove Theorem 1, which treats the case $e = c = 1$, with k and m arbitrary. We start by proving some preliminary lemmas.

The first lemma states conditions under which assumption A2 holds. We begin by developing a key localization condition. For $j = 1, \ldots, e$, write $F_{V_j}(\cdot)$ for the cdf of V_j.

L1. For each $j = 1, \ldots, e$, $F_{V_j}(\cdot)$ is an invertible map from the support of V_j onto $[0, 1]$.

In L1, it is not necessary that the maps be onto $[0, 1]$. More generally, we allow maps onto compact subsets of $[0, 1]$, but then more complicated notation is needed. We avoid this by making the stronger assumption. We do note that isolated jumps in the cdfs can be accommodated. For example, it is possible to accommodate discrete marginal distributions. However, nonisolated jumps and flat spots in the cdfs are not allowed. Informally, nonisolated jumps imply that close v_j values (v_j values on either side of a point v_0 at which a jump occurs) do not correspond to close u_j values (cdf values). Flat spots imply the reverse: close u_j values (u_j values on either side of a point u_0 equal to the cdf value on the flat spot) do not correspond to close v_j values. When either nonisolated jumps or flat spots occur, localization on $u \in S_U$ does not correspond to localization on $v \in S_V$.

Lemma 0 (Localization). *If L1 holds, then*

$$M(v) = (F_{V_1}(v_1), \ldots, F_{V_e}(v_e))$$

is an invertible map from S_V onto S_U. If, in addition, A1 holds, then for each $u \in S_U$ with

corresponding $v \in S_V$, there exist $x = (x_1, \ldots, x_e) \in S_{\mathcal{X}}$ and $z \in S_Z$ such that

$$u = M(v) = (I\!P\{\mathcal{X}_1 \leq x_1 \mid Z = z\}, \ldots, I\!P\{\mathcal{X}_e \leq x_e \mid Z = z\}) .$$

PROOF FOR $e = 2$. Let $(u_1, u_2) = (F_{V_1}(v_1), F_{V_2}(v_2))$ and $(\tilde{u}_1, \tilde{u}_2) = (F_{V_1}(\tilde{v}_1), F_{V_2}(\tilde{v}_2))$. Check that equal images imply equal preimages: $(u_1, u_2) = (\tilde{u}_1, \tilde{u}_2)$ and the invertibility of F_{V_1} and F_{V_2} imply that $(v_1, v_2) = (\tilde{v}_1, \tilde{v}_2)$. That the mapping is onto S_U follows from L1. This proves the first part of the Lemma.

Assume, for simplicity, that ϕ_1 and ϕ_2 are strictly increasing in their first arguments. Fix $v \in S_V$. By A1, there exist $x = (x_1, x_2) \in S_{\mathcal{X}}$ and $z \in S_Z$ such that $v = (v_1, v_2) = (\phi_1^{-1}(x_1, z), \phi_2^{-1}(x_2, z))$. It follows that

$$\begin{aligned}
(u_1, u_2) &\equiv (I\!P\{\mathcal{X}_1 \leq x_1 \mid Z = z\}, I\!P\{X_2 \leq x_2 \mid Z = z\}) \\
&= (I\!P\{\phi_1(z, V_1) \leq x_1 \mid Z = z\}, I\!P\{\phi_2(z, V_2) \leq x_2 \mid Z = z\}) \\
&= (I\!P\{V_1 \leq \phi_1^{-1}(x_1, z) \mid Z = z\}, I\!P\{V_2 \leq \phi_2^{-1}(x_2, z) \mid Z = z\}) \\
&= (I\!P\{V_1 \leq \phi_1^{-1}(x_1, z)\}, I\!P\{V_2 \leq \phi_2^{-1}(x_2, z)\}) \quad (V \text{ independent of } Z) \\
&= (F_{V_1}(\phi_1^{-1}(x_1, z)), F_{V_2}(\phi^{-1}(x_2, z))) \\
&= (F_{V_1}(v_1), F_{V_2}(v_2)) .
\end{aligned}$$

By the first part of the Lemma, each $u = (u_1, u_2) \in S_U$ can be so represented. This proves the second part of the Lemma. □

REMARK. Note that L1 does not preclude a discrete endogenous regressor. For example, take

$e = 1$ and $\mathcal{X}_1$ binary. By A1, $I\!P\{\mathcal{X}_1 = 1 \mid Z = z\} = F_{V_1}(\phi_1^{-1}(1, z))$. Deduce that L1 holds if $I\!P\{\mathcal{X}_1 = 1 \mid Z = z\}$ takes on all values in $[0, 1]$ as z ranges over S_Z.

LEMMA 1 (UNIFORM CONSISTENCY). *If assumptions A1 through A19 hold, then*

$$\sup_{u \in U_\kappa} |\hat{\beta}(u) - \beta(u)| = o_p(1)\,.$$

as $n \to \infty$.

PROOF. Fix $u \in U_\kappa$. Since $e = 1$, we estimate $\beta(u)$ with $\hat{\beta}(u) = \underset{b \in B_u}{\operatorname{argmax}}\, \hat{S}_n(b \mid u)$ where

$$\hat{S}_n(b \mid u) = \frac{1}{n\tau_n} \sum_{j=1}^{n} (2Y_j - 1) K_n^*(X_j b) \mathcal{K}_n(\hat{U}_j - u) \tau_\kappa(\mathcal{Z}_j)\,.$$

Define

$$\begin{aligned} S_n(b \mid u) &= \frac{1}{n\tau_n} \sum_{j=1}^{n} (2Y_j - 1) K_n^*(X_j b) \mathcal{K}_n(U_j - u) \tau_\kappa(\mathcal{Z}_j) \\ \bar{S}_n(b \mid u) &= I\!E S_n(b \mid u) = \frac{1}{\tau_n} I\!E (2Y - 1) K_n^*(Xb) \mathcal{K}_n(U - u) \tau_\kappa(\mathcal{Z})\,. \end{aligned}$$

Note that given U and Z, X is determined. By iterated expectations, with the inner expectation over Y given U and Z, we get that

$$\bar{S}_n(b \mid u) = \frac{1}{\tau_n} I\!E g(U, Z) K_n^*(Xb) \mathcal{K}_n(U - u) \tau_\kappa(\mathcal{Z})$$

where the expectation $I\!E$ in the last expression is over U and Z and

$$g(U, Z) = 2I\!P(\epsilon(U) < X\beta(U) \mid U, Z) - 1 \tag{22}$$

where $\epsilon(U) = -X(B - \beta(U))$. It follows from (16) that $g(u, Z) = 0$ whenever $X\beta(u) = 0$. This is a critical fact used in what follows. A1 and A2 imply that $\mathcal{X} = \phi(Z, M^{-1}(U))$, and we write $X \equiv X(U, Z)$ to acknowledge the functional dependence of X on U and Z. Apply a change of variable argument with $r = (U - u)/\tau_n$, to get

$$\bar{S}_n(b \mid u) = I\!E\left[\int g(u + r\tau_n, Z)K_n^*((X(u + r\tau_n, Z)b)f(u + r\tau_n)\mathcal{K}(r)dr\right]\tau_\kappa(\mathcal{Z})$$

where the expectation $I\!E$ is over Z and $f(\cdot)$ is the marginal density of U.

Next, define

$$\begin{aligned}\tilde{S}_n(b \mid u) &= I\!E\left[\int g(u + r\tau_n, Z)f(u + r\tau_n)\mathcal{K}(r)dr\right]\{X(u, Z)b > 0\}\tau_\kappa(\mathcal{Z}) \\ S(b \mid u) &= I\!Eg(u, Z)f(u)\{X(u, Z)b > 0\}\tau_\kappa(\mathcal{Z})\,.\end{aligned}$$

To show uniform consistency, we show that (i) a standard identification condition, called a strong maximization condition, holds uniformly over U_κ and (ii) as $n \to \infty$, $\hat{S}_n(b \mid u)$ converges in probability to $S(b \mid u)$ uniformly over the compact set $U_\kappa \otimes B_u$. (Pointwise strong maximization and a weak law of large numbers holding uniformly over B_u are sufficient to prove pointwise consistency while the stronger conditions (i) and (ii) are sufficient to prove uniform consistency.)

Start with the uniform strong maximization condition (i) above, which says that for any $\delta > 0$,

$$\inf_{u \in U_\kappa}\left[S(\beta(u) \mid u) - \sup_{|b - \beta(u)| \geq \delta} S(b \mid u)\right] > 0\,.$$

To establish this uniform condition, we first show the pointwise condition, namely, that the term

in brackets above is positive for each $u \in U_\kappa$.[14]

Fix $u \in U_\kappa$. To prove the pointwise result, we show that $S(b \mid u)$ is continuous in b on the compact set B_u, and is uniquely maximized at $\beta(u)$. To see this, note that $S(b \mid u)$ is continuous in b by a dominated convergence argument using the almost sure continuity and uniform boundedness of the integrand $g(u,Z)f(u)\{X(u,Z)b > 0\}\tau_\kappa(\mathcal{Z})$. By condition (16), for each $z \in S_Z$, the integrand attains its maximum value of $\max\{0, g(u,z)f(u)\tau_\kappa(z)\}$ at $b = \beta(u)$. It follows that $S(b \mid u)$ is maximized at $b = \beta(u)$. Unique maximization follows from (16), A4, A5, and arguments in Horowitz (1998, pp.59-60). This establishes the pointwise result.

We now establish the uniform result. By A16, $g(u,z)f(u)$ is a continuous function of u. This and a dominated convergence argument similar to the one used to prove that $S(b \mid u)$ is continuous in b, imply that $S(b \mid u)$ is also continuous in u. Since $S(b \mid u)$ is continuous in both b and u, the maximum theorem implies that $\beta(u)$ is an UHC correspondence. This and the fact that $\beta(u)$ uniquely maximizes $S(b \mid u)$ imply that $\beta(u)$ is a continuous function. Since B_u is compact, the correspondence from u to B_u is compact-valued. These last two facts and A10 imply that the constraint set $\{b \in B_u : |b - \beta(u)| \geq \delta\}$ is a compact-valued UHC correspondence. This and continuity of $S(b \mid u)$ in both b and u imply that the constrained value function $\sup_{|b-\beta(u)|\geq\delta} S(b \mid u)$ is an USC function of u. Since the unconstrained value function $S(\beta(u) \mid u)$ is continuous in u, it follows that $S(\beta(u) \mid u) - \sup_{|b-\beta(u)|\geq\delta} S(b \mid u)$ is a LSC function of u. By the Weierstrass Theorem, this function must attain its minimum value on the compact set U_κ. By the pointwise result, this function is positive for each $u \in U_\kappa$. It follows that its minimized value must also be positive, which establishes condition (i). See Aliprantis and Border (1994) and Berge (1997) for references.

[14] It is possible to replace U_κ with S_U in the uniform strong maximization argument. But this would require that S_V be compact.

To show (ii), we note that

$$\begin{aligned} |\hat{S}_n(b \mid u) - S(b \mid u)| &\leq |\hat{S}_n(b \mid u) - S_n(b \mid u)| \\ &+ |S_n(b \mid u) - \bar{S}_n(b \mid u)| \\ &+ |\bar{S}_n(b \mid u) - \tilde{S}_n(b \mid u)| \\ &+ |\tilde{S}_n(b \mid u) - S(b \mid u)| \,. \end{aligned}$$

Consider the first term on the RHS of the last expression. An argument based on a Taylor expansion of each $\hat{U}_j$ about U_j (as in the proof of (35) in Lemma 2 below) shows that the first term on the RHS is $o_p(1)$ uniformly over $U_\kappa \otimes B_u$. Standard empirical process results (see, for example, Lemma 3A in Sherman (1994)) show that as $n \to \infty$, the second term on the RHS is $o_p(1)$ uniformly over $U_\kappa \otimes B_u$. A Taylor expansion of $g(u + r\tau_n, Z)f(u + r\tau_n)$ about u implies that the fourth term is also $o_p(1)$ uniformly over $U_\kappa \otimes B_u$.

We now turn to the third term, which requires a bit more work. Recall the definition of $\bar{S}_n(b \mid u)$ and that $\tau_n \ll \sigma_n$. Write b_0 for the component of b corresponding to $\mathcal{X}$. By a Taylor expansion about u, we get that

$$\begin{aligned} K^*(X(u + r\tau_n, Z)b) &= K^*(X(u, Z)b) + (r\tau_n/\sigma_n)b_0\mathcal{K}_n(X^*b) \\ &= K^*(X(u, Z)b) + o(1) \end{aligned}$$

as $n \to \infty$ uniformly over $r \in [-1, 1]$ and $(u, b) \in U_\kappa \otimes B_u$. Assumption A16 implies that $|g(u + r\tau_n, Z)f(u + r\tau_n)|$ is bounded. Deduce that for some $c > 0$,

$$|\bar{S}_n(b \mid u) - \tilde{S}_n(b \mid u)| \leq c\mathbb{E}\,|K^*(X(u, Z)b/\sigma_n) - \{(X(u, Z)b > 0\}| + o(1)$$

where the last expectation is over $X(u, Z)$.

Recall that $X = (X_1, \tilde{X})$. Also, recall that the coefficient of X_1 is unity. Let $W = Xb = X_1 + \tilde{X}\tilde{b}$ and consider the transformation $(X_1, \tilde{X}) \mapsto (W, \tilde{X})$. This transformation is 1-1 and onto and the Jacobian of the transformation is unity. Let $f(x_1 \mid \tilde{x}, u)$ denote the density of X_1 given $\tilde{X} = \tilde{x}$ and $U = u$. By A6, this density is continuous in x_1. It follows that the last expectation is equal to

$$\int_{\tilde{x}} \left[\int_w |K^*(w/\sigma_n) - \{w > 0\}| f(w - \tilde{x}\tilde{b} \mid \tilde{x}, u)\, dw \right] f(\tilde{x} \mid u)\, d\tilde{x}\,.$$

Note that this transformation shifts the dependence on u and b from the integrand, which depends on n, to the argument of the conditional density, which does not depend on n. Since B_u and U_κ are compact, A6 implies that there exist $b^* \in B_u$ and $u^* \in U_\kappa$ such that $f(w - \tilde{x}\tilde{b}^* \mid \tilde{x}, u^*) = \sup_{b \in B_u, u \in U_\kappa} f(w - \tilde{x}\tilde{b} \mid \tilde{x}, u)$. It follows that $f(w - \tilde{x}\tilde{b}^* \mid \tilde{x}, u^*)$ is a density with respect to lebesgue measure on $\mathbb{R}$, and the integral in brackets is bounded by the integral with $f(w - \tilde{x}\tilde{b} \mid \tilde{x}, u)$ replaced by $f(w - \tilde{x}\tilde{b}^* \mid \tilde{x}, u^*)$. For each fixed w, $|K^*(w/\sigma_n) - \{w > 0\}|$ is bounded and converges to zero as $n \to \infty$. By the DCT, the integral in brackets converges to zero as $n \to \infty$. Since U_κ is compact, A7 implies that there exists $u^* \in U_\kappa$ such that $f(\tilde{x} \mid u^*) = \sup_{u \in U_\kappa} f(\tilde{x} \mid u)$. A similar DCT argument shows that the outer integral also converges to zero as $n \to \infty$. Moreover, this convergence is uniform over $U_\kappa \otimes B_u$. This establishes (ii), proving Lemma 1. □

LEMMA 2 (Rates of Uniform Convergence). *If assumptions A1 through A19 hold, then*

$$\sup_{x \in \mathcal{X}_\kappa, z \in Z_\kappa} |\hat{U}(x, z) - U(x, z)| = O_p(1/\sqrt{n}\alpha_n)$$

as $n \to \infty$.

PROOF. Recall that Z denotes the $1 \times m$ vector of instruments for the single endogenous regressor $\mathcal{X}$, and that $\mathcal{Z}$ denotes the single continuous instrument for $\mathcal{X}$. Also, recall that D denotes the $1 \times (m-1)$ vector of discrete instruments for $\mathcal{X}$. We write $z = (z_0, d)$ for a typical point in the support of Z, where z_0 is a point in the support of $\mathcal{Z}$, and d is a point in the support of D. Fix $x \in \mathcal{X}_\kappa$ and $z = (z_0, d) \in Z_\kappa$. We have that

$$\hat{U}(x,z) = \frac{1}{n\alpha_n} \sum_{k=1}^{n} \{\mathcal{X}_k \leq x\} \mathcal{K}_n(\mathcal{Z}_k - z_0) \{D_k = d\} / \hat{f}(z_0, d) \tag{23}$$

where

$$\hat{f}(z_0, d) = \frac{1}{n\alpha_n} \sum_{k=1}^{n} \mathcal{K}_n(\mathcal{Z}_k - z_0) \{D_k = d\} .$$

Note that $\hat{f}(z_0, d)$ estimates $f(z_0, d) = f(z_0 \mid d) I\!P\{D = d\}$ where $f(z_0 \mid d)$ denotes the conditional density of $\mathcal{Z}$ given d evaluated at z_0. Abbreviate $\hat{f}(z_0, d)$ to $\hat{f}$ and $f(z_0, d)$ to f. Note that

$$\begin{aligned}
\hat{U}(x,z) &= \frac{\frac{1}{n\alpha_n} \sum_{k=1}^{n} \{\mathcal{X}_k \leq x\} \mathcal{K}_n(\mathcal{Z}_k - z_0) \{D_k = d\}}{f(z_0, d)} \left[\frac{f}{\hat{f}} \right] \\
&= \frac{\frac{1}{n\alpha_n} \sum_{k=1}^{n} \{\mathcal{X}_k \leq x\} \mathcal{K}_n(\mathcal{Z}_k - z_0) \{D_k = d\}}{f(z_0, d)} \left[1 - \left(1 - \frac{\hat{f}}{f} \right) \right]^{-1} \\
&= \frac{\frac{1}{n\alpha_n} \sum_{k=1}^{n} \{\mathcal{X}_k \leq x\} \mathcal{K}_n(\mathcal{Z}_k - z_0) \{D_k = d\}}{f(z_0, d)} \left[1 + \left(1 - \frac{\hat{f}}{f} \right) + \left(1 - \frac{\hat{f}}{f} \right)^2 + \cdots \right] .
\end{aligned}$$

We now analyze the leading term in this last expansion. Nonleading terms can be handled similarly and have smaller stochastic order. By a slight abuse of notation, we take

$$\hat{U}(x,z) = \frac{\frac{1}{n\alpha_n} \sum_{k=1}^{n} \{\mathcal{X}_k \leq x\} \mathcal{K}_n(\mathcal{Z}_k - z_0) \{D_k = d\}}{f(z_0, d)} .$$

Write

$$\hat{U}(x,z) - U(x,z) = \hat{U}(x,z) - I\!E_z\hat{U}(x,z) + I\!E_z\hat{U}(x,z) - U(x,z)$$

where the expectation $I\!E_z$ is conditional on $Z = (z_0, d)$. By the first part of A15, $f(z_0 \mid d)$ is bounded above zero on Z_κ, precluding ratio bias. Since $x \in \mathcal{X}_\kappa$, $U(x,z)$ is eventually more than a bandwidth α_n from either boundary of S_U (0 or 1), precluding boundary bias. These facts, A14, the second part of A15, and a standard change of variable argument followed by a Taylor expansion to p_a terms implies that the bias term $I\!E_z\hat{U}(x,z) - U(x,z)$ has order $o_p(1/\sqrt{n})$ as $n \to \infty$, uniformly over $\mathcal{X}_\kappa \otimes Z_\kappa$. Note that

$$\hat{U}(x,z) - I\!E_z\hat{U}(x,z) = \frac{\frac{1}{n\alpha_n}\sum_{k=1}^{n}\{\mathcal{X}_k \le x\}\mathcal{K}_n(\mathcal{Z}_k - z_0)\{D_k = d\} - I\!E_z\{\mathcal{X}_k \le x\}\mathcal{K}_n(\mathcal{Z}_k - z_0)\{D_k = d\}}{f(z_0,d)}. \tag{24}$$

This is a zero-mean empirical process. A1, A14, and standard empirical process results (see, for example, the proof of Lemma 3A in Sherman (1994)) imply that this last term has order $O_p(1/\sqrt{n}\alpha_n)$ as $n \to \infty$, uniformly over $\mathcal{X}_\kappa \otimes Z_\kappa$. This proves Lemma 2. □

LEMMA 3 (Rates of Uniform Consistency). *If assumptions A1 through A19 hold, then*

$$\sup_{u \in U_\kappa} |\hat{\beta}(u) - \beta(u)| = O_p\left(1/\sqrt{n}\alpha_n\sigma_n\tau_n^2\right)$$

as $n \to \infty$.

Proof. Fix $u \in U_\kappa$. Recall $S_n(b \mid u) = \frac{1}{n\tau_n}\sum_{j=1}^{n}(2Y_j - 1)K_n^*(X_j b)\mathcal{K}_n(U_j - u)\tau_\kappa(\mathcal{Z}_j)$. Define $\bar{\beta}(u) = \underset{b \in B_u}{\operatorname{argmax}}\, S_n(b \mid u)$. Then

$$\hat{\beta}(u) - \beta(u) = [\hat{\beta}(u) - \bar{\beta}(u)] + [\bar{\beta}(u) - \beta(u)]. \tag{25}$$

Start with the second term on the RHS of (25). Define the gradient and hessian of $S_n(b \mid u)$:

$$\begin{aligned} G_n(b \mid u) &= \frac{1}{n\sigma_n\tau_n}\sum_{j=1}^{n}(2Y_j-1)\mathcal{K}_n(X_jb)\tilde{X}_j'\mathcal{K}_n(U_j-u)\tau_\kappa(\mathcal{Z}_j) \\ H_n(b \mid u) &= \frac{1}{n\sigma_n^2\tau_n}\sum_{j=1}^{n}(2Y_j-1)\mathcal{K}_n'(X_jb)\tilde{X}_j'\tilde{X}_j\mathcal{K}_n(U_j-u)\tau_\kappa(\mathcal{Z}_j)\,. \end{aligned}$$

The gradient and hessian of population criterion function $S(b \mid u)$ are denoted $G(b \mid u)$ and $H(b \mid u)$. By definition of $\bar{\beta}(u)$, $0 = G_n(\bar{\beta}(u) \mid u)$. A one term Taylor expansion of $G_n(\bar{\beta}(u) \mid u)$ about $\beta(u)$ implies that

$$\bar{\beta}(u) - \beta(u) = -[H_n(\bar{\beta}^*(u) \mid u)]^{-1}G_n(\beta(u) \mid u) \tag{26}$$

where $\bar{\beta}^*(u)$ is between $\bar{\beta}(u)$ and $\beta(u)$. Note that for each $u \in U_\kappa$,

$$\begin{aligned} H_n(\bar{\beta}^*(u) \mid u) &= H_n(\bar{\beta}^*(u) \mid u) - H(\bar{\beta}^*(u) \mid u) \\ &+ H(\bar{\beta}^*(u) \mid u) - H(\beta(u) \mid u)\,. \end{aligned}$$

The first term on the RHS of the last expression is bounded by

$$\sup_{(u,b)\in U_\kappa\otimes B_u} |H_n(b \mid u) - H(b \mid u)|\,.$$

The difference $H_n(b \mid u) - H(b \mid u)$ has mean zero for each $(u, b) \in U_\kappa \otimes B_u$. Standard empirical process arguments (once again, see Lemma 3A in Sherman (1994)) show that this last expression has order $O_p(1/\sqrt{n}\sigma_n^2\tau_n)$ as $n \to \infty$. Invoke A18. By a Taylor expansion of each of the k^2 components of $H(\bar{\beta}^*(u) \mid u)$ about $\beta(u)$, we get that the (i, j)th component of $H(\bar{\beta}^*(u) \mid u) - H(\beta(u) \mid u)$ equals $D_{ij}(\bar{\beta}^{**}(u) \mid u)(\bar{\beta}^*(u) - \beta(u))$, where $D_{ij}(b \mid u)$ is the partial derivative of the ijth component of

$H(b \mid u)$ with respect to b, and $\bar{\beta}^{**}(u)$ is between $\bar{\beta}(u)$ and $\beta(u)$. By A18, $D_{ij}(b \mid u)$ is a continuous function on the compact set $B_u \otimes U_\kappa$. Thus, this term has order $o_p(1)$ uniformly over $u \in U_\kappa$ provided $\sup_{u \in U_\kappa} |\bar{\beta}(u) - \beta(u)| = o_p(1)$ as $n \to \infty$. But this uniformity result holds by arguments similar to (and simpler than) those used to prove Lemma 1. Provided $\sigma_n^2 \tau_n \gg n^{-1/2}$, we get that uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$H_n(\bar{\beta}^*(u) \mid u) - H(\beta(u) \mid u) = O_p(1/\sqrt{n}\sigma_n^2\tau_n) + O_p(\sup_{u \in U_\kappa} |\bar{\beta}(u) - \beta(u)|) = o_p(1)$$

Now apply a Taylor expansion of $\left[H_n(\bar{\beta}^*(u) \mid u)\right]^{-1}$ about $H(\beta(u) \mid u)$. Provided $\sigma_n^2 \tau_n \gg n^{-1/2}$, we get that uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$\left[H_n(\bar{\beta}^*(u) \mid u)\right]^{-1} - \left[H(\beta(u) \mid u)\right]^{-1} = O_p(1/\sqrt{n}\sigma_n^2\tau_n) + O_p(\sup_{u \in U_\kappa} |\bar{\beta}(u) - \beta(u)|) = o_p(1)\,. \tag{27}$$

Further, note that A8, A18, A19, and continuity of the inverse function imply that uniformly over $u \in U_\kappa$,

$$\left[H(\beta(u) \mid u)\right]^{-1} = O(1)\,. \tag{28}$$

Deduce from (26), (27), and (28) that, uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$\bar{\beta}(u) - \beta(u) = -\left[\left[H(\beta(u) \mid u)\right]^{-1} + o_p(1)\right] G_n(\beta(u) \mid u) = O_p(1) G_n(\beta(u) \mid u)\,. \tag{29}$$

We now turn to an analysis of $G_n(\beta(u) \mid u)$. We have that

$$G_n(\beta(u) \mid u) = [G_n(\beta(u) \mid u) - I\!E G_n(\beta(u) \mid u)] + I\!E G_n(\beta(u) \mid u)\,. \tag{30}$$

Note that the term in brackets is a zero-mean empirical process. Standard empirical process arguments show that, uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$G_n(\beta(u) \mid u) - I\!E G_n(\beta(u) \mid u) = O_p(1/\sqrt{n}\sigma_n\tau_n)\,. \tag{31}$$

We now show that the bias term $I\!E G_n(\beta(u) \mid u)$ can be neglected. That is, we show that, uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$I\!E G_n(\beta(u) \mid u) = o_p(1/\sqrt{n})\,. \tag{32}$$

Note that

$$I\!E G_n(\beta(u) \mid u) = \frac{1}{\sigma_n\tau_n} I\!E(2Y_j - 1)\mathcal{K}_n(X_j\beta(u))\tilde{X}_j'\mathcal{K}_n(U_j - u)\tau_\kappa(\mathcal{Z}_j)\,. \tag{33}$$

Holding u fixed, we will evaluate this expectation in four steps: (i) average over Y_j given U_j and Z_j (ii) average over U_j given Z_j and $X_j\beta(u)$ (iii) average over Z_j given $X_j\beta(u)$ and (iv) average over $X_j\beta(u)$.

Recall the definition of $g(U, Z)$ given in (22), as well as the key identification result in (16) which follows from A11 and A12. After applying step (i), we get that the integrand in (33) equals

$$\frac{1}{\sigma_n\tau_n} g(U_j, Z_j)\mathcal{K}_n(X_j\beta(u))\tilde{X}_j'\mathcal{K}_n(U_j - u)\tau_\kappa(\mathcal{Z}_j)\,.$$

In applying step (ii), there are two cases to consider. The first is the case where $U_j = U_{j1}$. The second is the case where $U_j \neq U_{j1}$. We will analyze the former. The analysis of the latter is similar. Note that when $U_j = U_{j1}$, the random variable $\tilde{X}_j$ does not involve U_j. Apply step (ii), making

the change of variable $r = (U_j - u)/\tau_n$. After step (ii), the integrand in (33) equals

$$\frac{1}{\sigma_n}\left[\int g(u+\tau_n r, Z_j) f(u+\tau_n r \mid Z_j, X_j\beta(u))\mathcal{K}(r)dr\right]\mathcal{K}_n(X_j\beta(u))\tilde{X}_j'\tau_\kappa(\mathcal{Z}_j)$$

where $f(\cdot \mid Z, X\beta(u))$ denotes the density of U given Z and $X\beta(u)$.

In applying step (iii), write $\Gamma_n(X_j\beta(u))$ for the expectation over Z_j given $X_j\beta(u)$ of

$$\left[\int g(u+\tau_n r, Z_j) f(u+\tau_n r \mid Z_j, X_j\beta(u))\mathcal{K}(r)dr\right]\tilde{X}_j'\tau_\kappa(\mathcal{Z}_j)\,.$$

After applying step (iii), the integrand in (33) equals

$$\frac{1}{\sigma_n}\mathcal{K}_n(X_j\beta(u))\Gamma_n(X_j\beta(u))\,.$$

Finally, apply step (iv), making the change of variable $s = X_j\beta(u)/\sigma_n$ to get that the bias term in (33) equals

$$\int \Gamma_n(\sigma_n s) f(\sigma_n s)\mathcal{K}(s)ds$$

where $f(\cdot)$ denotes the density of $X_j\beta(u)$. Apply assumptions A14 and A16, and expand the product $g(u+\tau_n r, Z_j)f(u+\tau_n r \mid Z_j, X_j\beta(u))$ about $U_j = u$ to p_τ terms to replace it with $g(u, Z_j)f(u \mid Z_j, X_j\beta(u))$ plus a term that is $o_p(1/\sqrt{n})$ as $n \to \infty$. Then, apply A14 and A17 and expand $\Gamma_n(\sigma_n s)f(\sigma_n s)$ about $X_j\beta(u) = 0$ to p_σ terms to replace $\Gamma_n(\sigma_n s)f(\sigma_n s)$ with zero plus a term that is $o_p(1/\sqrt{n})$ as $n \to \infty$. The leading term in this expansion is zero because $g(u, Z_j) = 0$ when $X_j\beta(u) = 0$. The latter follows from (16). This proves (32).

It follows from (29), (30), (31), and (32) that, as $n \to \infty$,

$$\sup_{u \in U_\kappa} |\bar{\beta}(u) - \beta(u)| = O_p(1/\sqrt{n}\sigma_n\tau_n)\,. \tag{34}$$

Next we show that, as $n \to \infty$,

$$\sup_{u \in U_\kappa} |\hat{\beta}(u) - \bar{\beta}(u)| = O_p(1/\sqrt{n}\alpha_n\sigma_n\tau_n^2)\,. \tag{35}$$

Define the gradient and hessian of $\hat{S}_n(b \mid u)$:

$$\begin{aligned}
\hat{G}_n(b \mid u) &= \frac{1}{n\sigma_n\tau_n}\sum_{j=1}^{n}(2Y_j - 1)\mathcal{K}_n(X_j b)\tilde{X}_j'\mathcal{K}_n(\hat{U}_j - u)\tau_\kappa(\mathcal{Z}_j) \\
\hat{H}_n(b \mid u) &= \frac{1}{n\sigma_n^2\tau_n}\sum_{j=1}^{n}(2Y_j - 1)\mathcal{K}_n'(X_j b)\tilde{X}_j'\tilde{X}_j\mathcal{K}_n(\hat{U}_j - u)\tau_\kappa(\mathcal{Z}_j)\,.
\end{aligned}$$

By definition of $\hat{\beta}(u)$, $0 = \hat{G}_n(\hat{\beta}(u) \mid u)$. A one term Taylor expansion of $\hat{G}_n(\hat{\beta}(u) \mid u)$ about $\beta(u)$ implies that

$$\hat{\beta}(u) - \beta(u) = -[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1}\hat{G}_n(\beta(u) \mid u) \tag{36}$$

where $\hat{\beta}^*(u)$ is between $\hat{\beta}(u)$ and $\beta(u)$. Deduce from (26) and (36) that

$$\hat{\beta}(u) - \bar{\beta}(u) = -\left[\hat{H}_n(\hat{\beta}^*(u) \mid u)\right]^{-1}\hat{G}_n(\beta(u) \mid u) + \left[H_n(\bar{\beta}^*(u) \mid u)\right]^{-1}G_n(\beta(u) \mid u)\,. \tag{37}$$

Note that

$$\hat{G}_n(\beta(u) \mid u) = \frac{1}{n\sigma_n\tau_n}\sum_{j=1}^{n}(2Y_j - 1)\mathcal{K}_n(X_j\beta(u))\tilde{X}_j'\mathcal{K}_n(\hat{U}_j - u)\tau_\kappa(\mathcal{Z}_j)\,.$$

If we Taylor expand each summand about U_j, then the sum of the first terms in these expansions

equals $G_n(\beta(u) \mid u)$, a useful quantity to isolate in the subsequent analysis. By applying these expansions we get

$$\hat{G}_n(\beta(u) \mid u) = G_n(\beta(u) \mid u) + \frac{1}{n\sigma_n\tau_n} \sum_{j=1}^{n} \Lambda_n(\hat{U}_j^*, u)(\hat{U}_j - U_j)\tau_\kappa(\mathcal{Z}_j) \tag{38}$$

where $\hat{U}_j^*$ is between $\hat{U}_j$ and U_j, and

$$\begin{aligned} \Lambda_n(U,u) &= \frac{\partial}{\partial U}\left[(2Y-1)\mathcal{K}_n(X\beta(u))\tilde{X}'\mathcal{K}_n((U-u)\right] \\ &= (2Y-1)\mathcal{K}_n(X\beta(u))\tilde{X}'\mathcal{K}_n'((U-u)/\tau_n\,. \end{aligned}$$

Deduce from A14 that $\Lambda_n(U,u)\tau_\kappa(\mathcal{Z}) = O(1/\tau_n)$ as $n \to \infty$. Then apply Lemma 2 and (38) to get that, uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$\hat{G}_n(\beta(u) \mid u) = G_n(\beta(u) \mid u) + O_p(1/\sqrt{n}\alpha_n\sigma_n\tau_n^2)\,. \tag{39}$$

Note that (31) and (32) imply that, uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$G_n(\beta(u) \mid u) = O_p(1/\sqrt{n}\sigma_n\tau_n)\,. \tag{40}$$

Now, consider the term $\hat{H}_n(\hat{\beta}^*(u) \mid u)$ in (36). We have that

$$\begin{aligned} \hat{H}_n(\hat{\beta}^*(u) \mid u) - H(\beta(u) \mid u) &= \hat{H}_n(\hat{\beta}^*(u) \mid u) - H_n(\hat{\beta}^*(u) \mid u) \\ &+ H_n(\hat{\beta}^*(u) \mid u) - H(\hat{\beta}^*(u) \mid u) \\ &+ H(\hat{\beta}^*(u) \mid u) - H(\beta(u) \mid u)\,. \end{aligned}$$

By arguments very similar to those used to establish (39), we get that the first term in the decomposition, uniformly over $u \in U_\kappa$, has order $O_p(1/\sqrt{n}\alpha_n\sigma_n^2\tau_n^2)$ as $n \to \infty$. Arguments made previously show that the second term in the decomposition, uniformly over $u \in U_\kappa$, has order $O_p(1/\sqrt{n}\sigma_n^2\tau_n)$ as $n \to \infty$, while the third term, uniformly over $u \in U_\kappa$, has order $O_p(\sup_{u\in U_\kappa} |\hat{\beta}(u)-\beta(u)|) = o_p(1)$ as $n \to \infty$. Then the Taylor expansion arguments used to establish (27) can be used to show that uniformly over $u \in U_\kappa$, as $n \to \infty$,

$$\begin{aligned}\left[\hat{H}_n(\hat{\beta}^*(u) \mid u)\right]^{-1} - [H(\beta(u) \mid u)]^{-1} &= O_p(1/\sqrt{n}\alpha_n\sigma_n^2\tau_n^2) + O_p(1/\sqrt{n}\sigma_n^2\tau_n) + O_p(\sup_{u\in U_\kappa} |\hat{\beta}(u) - \beta(u)|) \\ &= o_p(1)\,. \end{aligned} \tag{41}$$

Recall (37). Deduce from (28), (41), and (39), and then (28), (27), and (40), that (35) holds. Lemma 3 now follows from (25), (35), and (34). □

We are now in a position to prove that $\hat{\beta}_\kappa$ is a $\sqrt{n}$-consistent and asymptotically normally distributed estimator of β_κ.

LEMMA 4 (THE SECOND TERM IN (10)). *If A1 through A19 hold, then*

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}(U_i) - \beta(U_i)\right]\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i) = \frac{1}{n}\sum_{i=1}^{n} f_n^{(2)}(W_i) + \frac{1}{n}\sum_{i=1}^{n} f_n^{(3)}(W_i) + o_p(1/\sqrt{n})$$

as $n \to \infty$, *where* $f_n^{(2)}(W_i) = f_n(P, P, W_i) + f_n(P, P, P, W_i)$ *and* $f_n^{(3)}(W_i) = f_n(P, W_i)$, *with* $f_n(W_i, W_j, W_k)$, $f_n(W_i, W_j, W_k, W_l)$, *and* $f_n(W_i, W_j)$ *defined in (17), (18), and (19), respectively.*

PROOF. Consider the second term in (10). This term equals

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}(U_i)-\beta(U_i)\right]\tau_\kappa(\mathcal{X}_i,\mathcal{Z}_i)\,. \tag{42}$$

For ease of notation, we suppress the trimming function $\tau_\kappa(\mathcal{X}_i,\mathcal{Z}_i)$. We get

$$\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\beta}(U_i)-\bar{\beta}(U_i)\right]+\frac{1}{n}\sum_{i=1}^{n}\left[\bar{\beta}(U_i)-\beta(U_i)\right]\,. \tag{43}$$

Start with the first term in (43). By (37), this term equals

$$-\frac{1}{n}\sum_{i=1}^{n}\left[\left[\hat{H}_n(\hat{\beta}^*(U_i)\mid U_i)\right]^{-1}\hat{G}_n(\beta(U_i)\mid U_i)-\left[H_n(\bar{\beta}^*(U_i)\mid U_i)\right]^{-1}G_n(\beta(U_i)\mid U_i)\right]\,. \tag{44}$$

By (41) and Lemma 3 we get that uniformly over $u\in U_\kappa$, as $n\to\infty$,

$$\left[\hat{H}_n(\hat{\beta}^*(u)\mid u)\right]^{-1}=\left[H(\beta(u)\mid u)\right]^{-1}+O_p(1/\sqrt{n}\alpha_n\sigma_n^2\tau_n^2)\,. \tag{45}$$

By (27) and (34) we get that uniformly over $u\in U_\kappa$, as $n\to\infty$,

$$\left[H_n(\bar{\beta}^*(u)\mid u)\right]^{-1}=\left[H(\beta(u)\mid u)\right]^{-1}+O_p(1/\sqrt{n}\sigma_n^2\tau_n)\,. \tag{46}$$

By (39) and (40), we get that uniformly over $u\in U_\kappa$, as $n\to\infty$,

$$\hat{G}_n(\beta(u)\mid u)=O_p(1/\sqrt{n}\alpha_n\sigma_n\tau_n^2)\,. \tag{47}$$

Equations (45) and (47), together with (46) and (40), imply that the expression in (44) equals

$$\frac{1}{n}\sum_{i=1}^{n}\left[[H(\beta(U_i)\mid U_i)]^{-1}\left[\hat{G}_n(\beta(U_i)\mid U_i)-G_n(\beta(U_i)\mid U_i)\right]\right]+O_p(1/n\alpha_n^2\sigma_n^3\tau_n^4)\,. \tag{48}$$

Note that the $O_p(1/n\alpha_n^2\sigma_n^3\tau_n^4)$ term has order $o_p(1/\sqrt{n})$ provided $\alpha_n^2\sigma_n^3\tau_n^4 \gg n^{-1/2}$. By (38),

$$\hat{G}_n(\beta(U_i)\mid U_i)-G_n(\beta(U_i)\mid U_i)=\frac{1}{n\sigma_n\tau_n}\sum_{j=1}^{n}\Lambda_n(\hat{U}_j^*,U_i)(\hat{U}_j-U_j)\tau_\kappa(\mathcal{Z}_j)\,.$$

Lemma 2 and a Taylor expansion of $\Lambda_n(\hat{U}_j^*,U_i)$ about U_j (see the expression following (38)) imply that, uniformly over i and j, as $n\to\infty$,

$$\hat{G}_n(\beta(U_i)\mid U_i)-G_n(\beta(U_i)\mid U_i)=\frac{1}{n\sigma_n\tau_n}\sum_{j=1}^{n}\Lambda_n(U_j,U_i)(\hat{U}_j-U_j)\tau_\kappa(\mathcal{Z}_j)+O_p(1/n\alpha_n^2\sigma_n\tau_n^3)\,. \tag{49}$$

Note that the $O_p(1/n\alpha_n^2\sigma_n\tau_n^3)$ term has order $o_p(1/\sqrt{n})$ provided $\alpha_n^2\sigma_n\tau_n^3 \gg n^{-1/2}$.

As in the proof of Lemma 2, we have that

$$\hat{U}_j=\frac{1}{n\alpha_n}\sum_{k=1}^{n}\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)\{D_k=D_j\}/\hat{f}(\mathcal{Z}_j,D_j) \tag{50}$$

where

$$\hat{f}(\mathcal{Z}_j,D_j)=\frac{1}{n\alpha_n}\sum_{k=1}^{n}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)\{D_k=D_j\}\,.$$

Note that $\hat{f}(\mathcal{Z}_j,D_j)$ estimates $f(\mathcal{Z}_j,D_j)=f(\mathcal{Z}_j\mid D_j)I\!P\{D=D_j\}$ where $f(\mathcal{Z}_j\mid D_j)$ denotes the conditional density of $\mathcal{Z}_j$ given D_j. Abbreviate $\hat{f}(\mathcal{Z}_j,D_j)$ to $\hat{f}$ and $f(\mathcal{Z}_j,D_j)$ to f. Then

$$\hat{U}_j \;=\; \frac{\frac{1}{n\alpha_n}\sum_{k=1}^{n}\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)\{D_k=D_j\}}{f(\mathcal{Z}_j,D_j)}\left[\frac{f}{\hat{f}}\right]$$

$$= \frac{\frac{1}{n\alpha_n}\sum_{k=1}^{n}\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\}}{f(\mathcal{Z}_j, D_j)}\left[1 - \left(1 - \frac{\hat{f}}{f}\right)\right]^{-1}$$

$$= \frac{\frac{1}{n\alpha_n}\sum_{k=1}^{n}\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\}}{f(\mathcal{Z}_j, D_j)}\left[1 + \left(1 - \frac{\hat{f}}{f}\right) + \left(1 - \frac{\hat{f}}{f}\right)^2 + \cdots\right]. \quad (51)$$

The first two terms in the last expansion, when combined with (49) and (48), make first order asymptotic contributions. The remaining terms lead to contributions of order $o_p(1/\sqrt{n})$ and so can be neglected.[15]

We now analyze the leading term in this last expansion. Analysis of the second term is very similar and so is omitted. By a slight abuse of notation, take

$$\hat{U}_j = \frac{\frac{1}{n\alpha_n}\sum_{k=1}^{n}\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\}}{f(\mathcal{Z}_j, D_j)}.$$

Write

$$\hat{U}_j - U_j = \hat{U}_j - I\!E_j\hat{U}_j + I\!E_j\hat{U}_j - U_j$$

where the expectation $I\!E_j$ is conditional on $(\mathcal{Z}_j, D_j)$. Invoke A13, A14, and A15 and apply a change of variable followed by a Taylor expansion to p_a terms to show that the bias term $I\!E_j\hat{U}_j - U_j$ has order $o_p(1/\sqrt{n})$. Therefore, it is enough to analyze

$$\hat{U}_j - I\!E_j\hat{U}_j = \frac{1}{n\alpha_n}\sum_{k=1}^{n}\frac{\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\} - I\!E_j\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\}}{f(\mathcal{Z}_j, D_j)}. \quad (52)$$

[15] The first two terms in the last expansion, apart from a $o_p(1/\sqrt{n})$ bias term, are zero-mean U-statistics of orders three and four, respectively. Each of these U-statistics has a nondegenerate projection (the first term in the Hoeffding decomposition), resulting in a first order asymptotic contribution. We demonstrate this fact with the first term in the expansion. However, this does not happen with the higher order terms in the expansion. Take the third term, for example. Apart from a $o_p(1/\sqrt{n})$ bias term, this term is a zero-mean U-statistic of order five. It is straightforward to show that the average of its kernel function over either of two arguments, conditional on the remaining four arguments, is zero. This implies a zero projection, resulting in no first order asymptotic contribution. Moreover, the tail process is easily shown to be $o_p(1/\sqrt{n})$.

Substitute (52) for $\hat{U}_j - U_j$ in (49), then combine with (48) and expand sums to get

$$\frac{1}{n^3}\sum_{i,j,k}\left[H(\beta(U_i)\mid U_i)\right]^{-1}\Lambda_n(U_j,U_i)\frac{\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)\{D_k=D_j\}-I\!E_j\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)\{D_k=D_j\}}{\alpha_n\sigma_n\tau_n f(\mathcal{Z}_j,D_j)} \tag{53}$$

where, to save space we suppress the trimming function $\tau_\kappa(\mathcal{Z}_j)$. Note that there are n^3 summands in (53). Define $n_{(3)} = n(n-1)(n-2)$ and $\boldsymbol{i_3} = (i,j,k)$ where $i \neq j \neq k \neq i$. Then the term in (53) equals

$$\frac{1}{n_{(3)}}\sum_{\boldsymbol{i_3}}\left[H(\beta(U_i)\mid U_i)\right]^{-1}\Lambda_n(U_j,U_i)\frac{\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)\{D_k=D_j\}-I\!E_j\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)\{D_k=D_j\}}{\alpha_n\sigma_n\tau_n f(\mathcal{Z}_j,D_j)} \tag{54}$$

plus a term that can be neglected asymptotically. The reason is that there are only $O(n^2)$ terms in the difference between the triple sums in (53) and (54). If $\alpha_n\sigma_n\tau_n \gg n^{-1/2}$, then the difference between (53) and (54) has order $o_p(1/\sqrt{n})$ as $n\to\infty$.

The term in (54) is a zero-mean U-statistic of order three. Define $W_i = (Y_i, X_i, Z_i, U_i)$ and $f_n(W_i,W_j,W_k)$ to be the (i,j,k)th summand in expression (54). Define $n_{(2)} = n(n-1)$ and $\boldsymbol{i_2} = (i,j)$ where $i \neq j$. Apply the Hoeffding decomposition (see Serfling, 1980,Chapter 5) to get that

$$\begin{aligned}\frac{1}{n_{(3)}}\sum_{\boldsymbol{i_3}} f_n(W_i,W_j,W_k) &= \frac{1}{n}\sum_{i=1}^{n}\left[f_n(W_i,P,P)+f_n(P,W_i,P)+f_n(P,P,W_i)\right] \\ &+ \frac{1}{n_{(2)}}\sum_{\boldsymbol{i_2}} g_n(W_i,W_j)+\frac{1}{n_{(3)}}\sum_{\boldsymbol{i_3}} h_n(W_i,W_j,W_k)\end{aligned} \tag{55}$$

where the second average in the decomposition is a degenerate U-statistic of order two, and the third average is a degenerate U-statistic of order three. It follows that as $n\to\infty$, the second and third averages have order $O_p(1/n\alpha_n\sigma_n\tau_n)$ and $O_p(1/n^{3/2}\alpha_n\sigma_n\tau_n)$, respectively. Thus, if $\alpha_n\sigma_n\tau_n \gg n^{-1/2}$,

then both of these terms have order $o_p(1/\sqrt{n})$ and so can be ignored.

We now show that the first term in (55) is $\sqrt{n}$-consistent and asymptotically normally distributed. First note that $f_n(W_i, P, P) = f_n(P, W_i, P) = 0$. To see this, fix W_i and W_j and note that $f_n(W_i, W_j, P) = 0$. So, it suffices to analyze the average of the $f_n(P, P, W_i)$'s in (55). For convenience, we will write this term as

$$\frac{1}{n}\sum_{k=1}^{n} f_n(P, P, W_k)\,. \tag{56}$$

The claim that this term is $\sqrt{n}$-consistent might initially be viewed with some suspicion. To see why, note that

$$\begin{aligned} f_n(W_i, W_j, W_k) &= \frac{1}{\alpha_n \sigma_n \tau_n^2} H(\beta(U_i) \mid U_i)]^{-1} \\ &\times (2Y_j - 1)\mathcal{K}_n(X_j\beta(U_i))\tilde{X}_j'\mathcal{K}_n'(U_j - U_i) \\ &\times \frac{\{\mathcal{X}_k \leq \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\} - I\!E_j\{\mathcal{X}_k \leq \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)\{D_k = D_j\}}{f(\mathcal{Z}_j, D_j)}\,. \end{aligned}$$

We see that $f_n(W_i, W_j, W_k)$ is a product of terms divided by $\alpha_n\sigma_n\tau_n^2$. However, this product involves only three kernel function factors: the kernel function used to estimate U corresponding to α_n, the kernel function used to smooth the indicator $\{t > 0\}$ corresponding to σ_n, and the derivative of the kernel function used to localize on U corresponding to τ_n. Integrating a kernel function or its derivative involves a change of variable, resulting in a rescaling by the corresponding bandwidth factor. Thus, one might expect that the one α_n factor, the one σ_n factor, and one of the τ_n factors can be accounted for, but not the remaining τ_n factor. If true, this would imply that the expression in (56) is at best $\sqrt{n\tau_n}$-consistent, but not $\sqrt{n}$-consistent. But, in fact, the expression in (56) is $\sqrt{n}$-consistent. The reason is that the derivative of the bias reducing kernel we use is an

odd function, which, when integrated, annihilates a leading constant term, thus accounting for the fourth bandwidth factor. We now show this.

To save space, we will consider the case $m = 1$ so that $Z_j = \mathcal{Z}_j$. The case of general m adds nothing to understanding and follows immediately from the argument given below by replacing marginal densities with joint densities (products of conditional and marginal densities) and adding summations over the discrete conditioning variables. From (54) and the expression following (38), we get that the term in question equals

$$\frac{1}{n}\sum_{k=1}^{n} f_n^{(1)}(P, P, W_k) \tag{57}$$

where $f_n^{(1)}(P, P, W_k)$ equals

$$\frac{1}{\alpha_n \sigma_n \tau_n^2} I\!E_k h(U_i)(2Y_j - 1)\mathcal{K}_n(X_j\beta(U_i))\tilde{X}_j'\mathcal{K}_n'(U_j - U_i)\frac{\{\mathcal{X}_k \leq \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j) - I\!E_j\{\mathcal{X}_k \leq \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)}{f(\mathcal{Z}_j)} \tag{58}$$

where the expectation $I\!E_k$ is over W_i and W_j given W_k, $h(U_i) = [H(\beta(U_i) \mid U_i)]^{-1}$, and the expectation $I\!E_j$ is the expectation over W_k given W_j.

In evaluating the expectation $I\!E_k$ in (58), we first fix W_i and average over W_j. Note that the integrand depends on W_i only through U_i. For ease of notation, when averaging out over W_j, we will replace each U_i with u. The averaging over W_j will be done in 4 steps: (i) average over Y_j given U_j and $\mathcal{Z}_j$ (ii) average over U_j given $\mathcal{Z}_j$ and $X_j\beta(u)$ (iii) average over $\mathcal{Z}_j$ given $X_j\beta(u)$ and (iv) average over $X_j\beta(u)$. After step (iv), we will average over u to get $f_n^{(1)}(P, P, W_k)$.

Recall the definition of $g(\cdot,\cdot)$ in (22). After applying step (i) the integrand in (58) equals

$$\frac{1}{\alpha_n\sigma_n\tau_n^2}h(u)g(U_j,Z_j)\mathcal{K}_n(X_j\beta(u))\tilde{X}_j'\mathcal{K}_n'(U_j-u)\frac{\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)-I\!E_j\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)}{f(\mathcal{Z}_j)}\,. \tag{59}$$

In applying step (ii), there are two cases to consider, namely, the case $U_j=U_{1j}$ and the case $U_j\neq U_{1j}$. We analyze the former case. Analysis of the latter case is similar. Note that when $U_j=U_{1j}$, then $\tilde{X}_j$ does not depend on U_j. Make the change of variable $r=(U_j-u)/\tau_n$. After applying step (ii), the integrand in (58) equals

$$\begin{aligned}\frac{1}{\alpha_n\sigma_n\tau_n}\mathcal{K}_n(X_j\beta(u)) &\times h(u)\tilde{X}_j'\frac{\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)-I\!E_j\{\mathcal{X}_k\le\mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k-\mathcal{Z}_j)}{f(\mathcal{Z}_j)}\\ &\times \int g(u+\tau_n r,Z_j)f(u+\tau_n r\mid\mathcal{Z}_j,X_j\beta(u))\mathcal{K}'(r)dr\,. \end{aligned}\tag{60}$$

where $f(\cdot\mid\mathcal{Z},X\beta(u))$ is the conditional density of U given $\mathcal{Z}$ and $X\beta(u)$. We now closely examine the integral in (60). By Taylor expansions about u we get that

$$\begin{aligned} g(u+\tau_n r,Z_j) &= g(u,Z_j)+\tau_n r g_1(u^*,Z_j)\\ f(u+\tau_n r\mid\mathcal{Z}_j,X_j\beta(u)) &= f(u\mid\mathcal{Z}_j,X_j\beta(u))+\tau_n r f_1(\bar{u}\mid\mathcal{Z}_j,X_j\beta(u))\end{aligned}$$

where g_1 is the partial derivative of g with respect to its first argument, and u^* is between u and $u+\tau_n r$, while $f_1(\cdot\mid\mathcal{Z}_j)$ is the partial derivative of $f(\cdot\mid\mathcal{Z}_j,X_j\beta(u))$ with respect to its first argument, and $\bar{u}$ is between u and $u+\tau_n r$. Since $\mathcal{K}'(\cdot)$ is an odd function integrated over a symmetric interval, the integral of the leading constant term is annihilated:

$$\int g(u,Z_j)f(u\mid\mathcal{Z}_j,X_j\beta(u))\mathcal{K}'(r)dr=g(u,Z_j)f(u\mid\mathcal{Z}_j,X_j\beta(u))\int\mathcal{K}'(r)dr=0\,. \tag{61}$$

Deduce that the integral in (60) equals

$$\tau_n \int r\left[g(u,Z_j)f_1(\bar{u} \mid \mathcal{Z}_j, X_j\beta(u)) + f(u \mid \mathcal{Z}_j, X_j\beta(u))g_1(u^*, Z_j) + r\tau_n f_1(\bar{u} \mid \mathcal{Z}_j, X_j\beta(u))g_1(u^*, Z_j)\right]\mathcal{K}'(r)dr\,. \tag{62}$$

Let $I_n(\mathcal{Z}_j, X_j\beta(u), u)$ denote the integral in (62). Thus, after applying step (ii), the integrand in (58) equals

$$\frac{1}{\alpha_n\sigma_n}\mathcal{K}_n(X_j\beta(u))h(u)\tilde{X}_j'\frac{[\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j) - I\!\!E_j\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)]}{f(\mathcal{Z}_j)}I_n(\mathcal{Z}_j, X_j\beta(u), u)\,. \tag{63}$$

We see that in applying step (ii), the two τ_n factors have been accounted for. Define $I_n(\mathcal{Z}_j) = I\!\!E_j\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}_n(\mathcal{Z}_k - \mathcal{Z}_j)$. Apply step (iii). Make the change of variable $s = (\mathcal{Z}_j - \mathcal{Z}_k)/\alpha_n$. After applying step (iii) the integrand in (58) equals

$$\frac{1}{\sigma_n}\mathcal{K}_n(X_j\beta(u))\int h(u)\tilde{X}_j'I_n(\mathcal{Z}_k{+}\tau_n s, X_j\beta(u), u)\frac{\{\mathcal{X}_k \le \mathcal{X}_j\}\mathcal{K}(s) - I_n(\mathcal{Z}_k + \tau_n s)}{f(\mathcal{Z}_k + \tau_n s)}f(\mathcal{Z}_k{+}\tau_n s \mid X_j\beta(u))ds\,. \tag{64}$$

where $f(\cdot \mid X_j\beta(u))$ is the conditional density of $\mathcal{Z}$ given $X\beta(u) = X_j\beta(u)$. Let $I_n(W_k, X_j\beta(u), u)$ denote the integral in (64). Then the integrand in (58) equals

$$\frac{1}{\sigma_n}\mathcal{K}_n(X_j\beta(u))I_n(W_k, X_j\beta(u), u)\,. \tag{65}$$

We now apply step (iv). Make the change of variable $t = X_j\beta(u)/\sigma_n$. After applying step (iv) the integrand in (58) equals

$$\int I_n(W_k, \sigma_n t, u)f(\sigma_n t \mid u)\mathcal{K}(t)dt \tag{66}$$

where $f(\cdot \mid u)$ is the density of $X_j\beta(u)$. Finally, we average out over u to get that the expression

in (58) equals

$$\int \left[\int I_n(W_k, \sigma_n t, u) f(\sigma_n t \mid u) \mathcal{K}(t) dt \right] f(u) du \tag{67}$$

where $f(\cdot)$ denotes the marginal density of U. That is, $f_n^{(1)}(P, P, W_k)$ in (57) is equal to this last expression. We see that the expression in (57) is an average of zero mean iid random vectors. Moreover, because all components of $f_n^{(1)}$ are bounded and the density of $\mathcal{Z}$ is bounded away from zero on $\{|\mathcal{Z}| \leq \kappa\}$, these variables have finite variance as well. Deduce from a standard CLT that the expression in (57) is $\sqrt{n}$-consistent and asymptotically normally distributed. This takes care of the first term in (43).

Now we analyze the second term in (43). This term is much easier to analyze than the first term in (43) because it does not involve estimated U_i's. By (26), we get that

$$\frac{1}{n} \sum_{i=1}^{n} \left[\bar{\beta}(U_i) - \beta(U_i) \right] = \frac{1}{n} \sum_{i=1}^{n} \left[H_n(\bar{\beta}^*(U_i) \mid U_i) \right]^{-1} G_n(\beta(U_i) \mid U_i) \,. \tag{68}$$

By (27), (34), and (40), we get that as $n \to \infty$,

$$\frac{1}{n} \sum_{i=1}^{n} \left[H_n(\bar{\beta}^*(U_i) \mid U_i) \right]^{-1} G_n(\beta(U_i) \mid U_i) = \frac{1}{n} \sum_{i=1}^{n} \left[H(\beta(U_i) \mid U_i) \right]^{-1} G_n(\beta(U_i) \mid U_i) + O_p(1/n\sigma_n^3 \tau_n^2) \,. \tag{69}$$

Provided $\sigma_n^3 \tau_n^2 \gg n^{-1/2}$, the $O_p(1/n\sigma_n^3\tau_n^2)$ term has order $o_p(1/\sqrt{n})$. Consider the main term on the RHS of equation (69). Recall that $W_i = (Y_i, X_i, Z_i, U_i)$ and $h(u) = [H(\beta(u) \mid u)]^{-1}$. Define

$$f_n(W_i, W_j) = \frac{1}{\sigma_n \tau_n} h(U_i)(2Y_j - 1) \mathcal{K}_n(X_j \beta(U_i)) \tilde{X}_j' \mathcal{K}_n(U_j - U_i) \,. \tag{70}$$

where, as before, to save space we have suppressed the trimming function $\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)\tau_\kappa(\mathcal{Z}_j)$. We get

that the main term on the RHS of (69) is equal to

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_n(W_i, W_j)\,. \tag{71}$$

There are only n terms in the double sum for which $i = j$. Provided $\sigma_n\tau_n \gg n^{-1/2}$, as $n \to \infty$,

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} f_n(W_i, W_j) = \frac{1}{n_{(2)}}\sum_{\boldsymbol{i_2}} f_n(W_i, W_j) + o_p(1/\sqrt{n})\,. \tag{72}$$

The first term on the RHS of this last equality is a U-statistic of order 2. By the Hoeffding decomposition, we get that

$$\begin{aligned}\frac{1}{n_{(2)}}\sum_{\boldsymbol{i_2}} f_n(W_i, W_j) &= f_n(P,P) + \frac{1}{n}\sum_{i=1}^{n}\left[f_n(W_i,P) + f_n(P,W_i) - 2f_n(P,P)\right] &(73)\\ &+ \frac{1}{n_{(2)}}\sum_{\boldsymbol{i_2}}\left[f_n(W_i,W_j) - f_n(W_i,P) - f_n(P,W_j) + f_n(P,P)\right]\,. &(74)\end{aligned}$$

The term in (74) is a degenerate U-statistic of order 2 having order $O_p(1/n\sigma_n\tau_n)$ as $n \to \infty$. Provided $\sigma_n\tau_n \gg n^{-1/2}$, this term has order $o_p(1/\sqrt{n})$ and so can be ignored. Consider (73). Note that $f_n(W_i, P) = h(U_i)\mathbb{E}G_n(\beta(U_i) \mid U_i)$. It follows from (32) that both $f_n(W_i, P)$ and $f_n(P, P)$ have order $o_p(1/\sqrt{n})$ as $n \to \infty$. Deduce that the only term in the last expression that makes a contribution to the first order asymptotic behavior of $\hat{\beta}_\kappa$ is

$$\frac{1}{n}\sum_{i=1}^{n}\left[f_n(P,W_i) - f_n(P,P)\right]\,. \tag{75}$$

For convenience, we will write this term as

$$\frac{1}{n}\sum_{j=1}^{n}\left[f_n(P,W_j)-f_n(P,P)\right]\,. \tag{76}$$

In order to evaluate $f_n(P,W_j)$, we will fix W_j in $f_n(W_i,W_j)$ and then average over W_i in 2 steps: (i) average over U_i given $X_j\beta(U_i)$ and (ii) average over $X_j\beta(U_i)$. Step (i) involves a change of variable argument and a rescaling by τ_n. Step (ii) involves a change of variable argument and a rescaling by σ_n. As before, we get that the term in (76) is an average of zero-mean iid random vectors with finite variance. A standard CLT shows this term to be $\sqrt{n}$-consistent and asymptotically normally distributed. This proves Lemma 4. □

LEMMA 5 (THE TERM IN (11)). *If assumptions A1 through A19 hold, then*

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\delta}(\hat{U}_i^*)(\hat{U}_i-U_i)\tau_\kappa(\mathcal{X}_i,\mathcal{Z}_i)=\frac{1}{n}\sum_{i=1}^{n}f_n^{(4)}(W_i)+o_p(1/\sqrt{n})$$

as $n\to\infty$ where $f_n^{(4)}(W_i)=f_n(P,W_i)+f_n(P,P,W_i)$ with $f_n(W_i,W_j)$ and $f_n(W_i,W_j,W_k)$ defined in (20) and (21), respectively.

PROOF. To save space, we suppress $\tau_\kappa(\mathcal{X}_i,\mathcal{Z}_i)$. We get that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\hat{\delta}(\hat{U}_i^*)(\hat{U}_i-U_i) &= \frac{1}{n}\sum_{i=1}^{n}\delta(U_i)(\hat{U}_i-U_i) &(77)\\ &+ \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\delta}(\hat{U}_i^*)-\delta(\hat{U}_i^*)\right](\hat{U}_i-U_i) &(78)\\ &+ \frac{1}{n}\sum_{i=1}^{n}\left[\delta(\hat{U}_i^*)-\delta(U_i)\right](\hat{U}_i-U_i)\,. &(79)\end{aligned}$$

We will show that the first term on the RHS is $\sqrt{n}$-consistent and asymptotically normally dis-

tributed, while the remaining two terms have order $o_p(1/\sqrt{n})$ and so can be neglected.

We start by analyzing the expression in (77). As in the proof of Lemma 4, averages associated with the first two terms in (51) lead to nondegenerate first order asymptotic contributions. Averages associated with the remaining terms make degenerate contributions and are ignored. We analyze the first of the two averages that make nondegenerate contributions. The analysis of the second such term is very similar and so is omitted.

We replace $\hat{U}_i - U_i$ in (77) with

$$\frac{1}{n\alpha_n}\sum_{j=1}^{n}\left[\{\mathcal{X}_j \le \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\} - I\!E_i\{\mathcal{X}_j \le \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\}\right]/f(\mathcal{Z}_i, D_i)\,. \tag{80}$$

Substitute (80) into (77), then combine sums to get that the expression in (77) equals

$$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\delta(U_i)\left[\{\mathcal{X}_j \le \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\} - I\!E_i\{\mathcal{X}_j \le \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\}\right]/\alpha_n f(\mathcal{Z}_i, D_i)\,. \tag{81}$$

As before, we may neglect the diagonal terms and take the term in (77) to equal the zero-mean second order U-statistic

$$\frac{1}{n_{(2)}}\sum_{\boldsymbol{i_2}}\delta(U_i)\left[\{\mathcal{X}_j \le \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\} - I\!E_i\{\mathcal{X}_j \le \mathcal{X}_i\}\mathcal{K}_n(\mathcal{Z}_j - \mathcal{Z}_i)\{D_j = D_i\}\right]/\alpha_n f(\mathcal{Z}_i, D_i)\,. \tag{82}$$

Define $f_n(W_i, W_j)$ to be equal to the (i,j)th summand in (82). Note that $f_n(W_i, P) = f_n(P, P) = 0$. By the Hoeffding composition,

$$\frac{1}{n_{(2)}}\sum_{\boldsymbol{i_2}} f_n(W_i, W_j) = \frac{1}{n}\sum_{i=1}^{n} f_n(P, W_i) + \frac{1}{n_{(2)}}\sum_{\boldsymbol{i_2}}\left[f_n(W_i, W_j) - f_n(P, W_j)\right]\,. \tag{83}$$

Standard U-statistic results show that the second term on the RHS of (83) has order $O_p(1/n\alpha_n)$ as $n \to \infty$. This term has order $o_p(1/\sqrt{n})$ provided $\alpha_n \gg n^{-1/2}$. The usual change of variable argument shows that the first term on the RHS of (83) is an average of zero-mean iid random vectors with finite variance. A standard CLT shows that this term is $\sqrt{n}$-consistent and asymptotically normally distributed.

Next, we analyze (79). Let $\gamma(u)$ denote the partial derivative of $\delta(u)$ with respect to U. By a Taylor expansion of each $\delta(\hat{U}_i^*)$ about U_i, we have that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left[\delta(\hat{U}_i^*) - \delta(U_i)\right](\hat{U}_i - U_i) &= \frac{1}{n}\sum_{i=1}^{n}\gamma(\hat{U}_i^{**})(\hat{U}_i^* - U_i)(\hat{U}_i - U_i) \quad (84) \\ &= \frac{1}{n}\sum_{i=1}^{n}\gamma(\hat{U}_i^{**})(U_i^* - \hat{U}_i)(U_i - \hat{U}_i)\,. \quad (85) \end{aligned}$$

where, for each i, $\hat{U}_i^{**}$ is between U_i and $\hat{U}_i^*$. Since $\gamma(\cdot)$ is a continuous function on the compact set U_κ, $\gamma(u)$ is uniformly bounded over U_κ. It follows from this and Lemma 2 that the expression in (79) has order $O_p(1/n\alpha_n^2)$. Thus, this term has order $o_p(1/\sqrt{n})$ as $n \to \infty$ provided $\alpha_n^2 \gg n^{-1/2}$.

Finally, we analyze the expression in (78). Recall the definition of $\gamma(u)$ above. Let $\hat{\gamma}(u)$ denote the partial derivative of $\hat{\delta}(u)$ with respect to U. By a Taylor expansion of $\hat{\delta}(\hat{U}_i^*) - \delta(\hat{U}_i^*)$ about U_i we get that

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\delta}(\hat{U}_i^*) - \delta(\hat{U}_i^*)\right](\hat{U}_i - U_i) &= \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\delta}(U_i) - \delta(U_i)\right](\hat{U}_i - U_i) \quad (86) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\left[\hat{\gamma}(\hat{U}_i^{**}) - \gamma(\hat{U}_i^{**})\right](U_i - \hat{U}_i^*)(U_i - \hat{U}_i) \quad (87) \end{aligned}$$

where $\hat{U}_i^{**}$ is between $\hat{U}_i^*$ and U_i. Start with (87). Just as integrating kernel functions with respect to u results in a rescaling by a factor of τ_n, differentiating kernel functions with respect to u

results in a rescaling by a factor of τ_n^{-1}. This principle can be applied together with (36), (45), and (47) to get that uniformly over i, as $n \to \infty$, $\hat{\gamma}(\hat{U}_i^{**}) - \gamma(\hat{U}_i^{**})$ has order $O_p(1/\sqrt{n}\alpha_n\sigma_n\tau_n^4)$. Combine this result with Lemma 2 to see that the term in (87), uniformly over i, as $n \to \infty$, has order $O_p(1/n^{3/2}\alpha_n^3\sigma_n\tau_n^4)$. Deduce that this term has uniform asymptotic order $o_p(1/\sqrt{n})$ provided $\alpha_n^3\sigma_n\tau_n^4 \gg n^{-1}$.

We now analyze the term in (86). Differentiate both sides of (36) with respect to u applying the product rule to get that

$$\begin{aligned}
\hat{\delta}(u) - \delta(u) &= \frac{\partial}{\partial u}\left[-[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1}\hat{G}_n(\beta(u) \mid u)\right] \\
&= -[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1}\frac{\partial}{\partial u}\left[\hat{G}_n(\beta(u) \mid u)\right] \qquad (88) \\
&+ \frac{\partial}{\partial u}\left[-[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1}\right]\hat{G}_n(\beta(u) \mid u)\,. \qquad (89)
\end{aligned}$$

Start with (89). Focus first on $-[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1}$. Recall (36). Since the LHS of (36) is continuously differentiable in u and $\hat{G}_n(\beta(u) \mid u)$ is continuously differentiable in u, it follows that $-[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1}$ is continuously differentiable in u. Thus, for each fixed n, $\frac{\partial}{\partial u}\left[-[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1}\right]$ is continuous in u on the compact set U_κ and so is bounded. Deduce from this together with (45) and the fact that $-[H(\beta(u) \mid u)]^{-1}$ does not depend on n and is bounded on U_κ, that $-[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1} = O_p(1)$ uniformly over u as $n \to \infty$. This, together with (47) imply that the term in (89) has order $O_p(1/\sqrt{n}\alpha_n\sigma_n\tau_n^2)$ uniformly over u as $n \to \infty$. Combine this with Lemma 2 and (86) to see that the contribution of (89) to (78) is $O_p(1/n\alpha_n^2\sigma_n\tau_n^2)$ as $n \to \infty$. Provided $\alpha_n^2\sigma_n\tau_n^2 \gg n^{-1/2}$, this contribution has order $o_p(1/\sqrt{n})$ as $n \to \infty$.

Finally, consider (88). Argue as in the previous paragraph to see that $-[\hat{H}_n(\hat{\beta}^*(u) \mid u)]^{-1} =$

$O_p(1)$ uniformly over u as $n \to \infty$. Note that by (38),

$$\frac{\partial}{\partial u}\left[\hat{G}_n(\beta(u) \mid u)\right] = \left[\frac{\partial}{\partial u} G_n(\beta(u) \mid u) - I\!E \frac{\partial}{\partial u} G_n(\beta(u) \mid u)\right] + \frac{\partial}{\partial u} I\!E G_n(\beta(u) \mid u) \quad (90)$$

$$+ \frac{\partial}{\partial u}\left[\frac{1}{n\sigma_n\tau_n}\sum_{j=1}^{n} \Lambda_n(\hat{U}_j^*, u)(\hat{U}_j - U_j)\tau_\kappa(\mathcal{Z}_j)\right] \quad (91)$$

where for the middle term in (90) we have used the fact that integration and differentiation can be interchanged. By (31) and the fact that differentiation results in a rescaling by τ_n^{-1}, we get that the term in brackets on the RHS of (90) has order $O_p(1/\sqrt{n}\sigma_n\tau_n^2)$ uniformly over u as $n \to \infty$. By (32) and the fact that differentiation results in a rescaling by τ_n^{-1}, we get that the second term on the RHS of (90) has order $o_p(1/\sqrt{n}\tau_n)$ uniformly over u as $n \to \infty$. These facts and Lemma 2 imply that the contribution of the term in (90) to (78) is $O_p(1/n\alpha_n\sigma_n\tau_n^2) + o_p(1/n\alpha_n\tau_n) = O_p(1/n\alpha_n\sigma_n\tau_n^2)$ as $n \to \infty$. Provided $\alpha_n\sigma_n\tau_n^2 \gg n^{-1/2}$, this contribution has order $o_p(1/\sqrt{n})$ as $n \to \infty$. Now consider the term in (91). By (38), (39), and Lemma 2, and the fact that differentiation results in a rescaling by τ_n^{-1}, we get that this term has order $O_p(1/\sqrt{n}\alpha_n\sigma_n\tau_n^3)$ as $n \to \infty$. This fact and another application of Lemma 2 imply that the contribution of the term in (91) to (78) is $O_p(1/n\alpha_n^2\sigma_n\tau_n^3)$ as $n \to \infty$. Provided $\alpha_n^2\sigma_n\tau_n^3 \gg n^{-1/2}$, this contribution has order $o_p(1/\sqrt{n})$ as $n \to \infty$. This proves Lemma 5. □

Recall the definitions of $f_n^{(j)}(W_i)$, $j = 1, 2, 3, 4$ given just prior to the statement of Theorem 1 in the main text.

THEOREM 1. ($\sqrt{n}$-Consistency and Asymptotic Normality) *Let $e = c = 1$ with k and m arbitrary. If A1 through A19 hold, then, as $n \to \infty$,*

$$\sqrt{n}(\hat{\beta}_\kappa - \beta_\kappa) \rightsquigarrow N(0, \Sigma)$$

where $\Sigma = I\!E f_n(W_i) f_n(W_i)'$ *with* $f_n(W_i) = f^{(1)}(W_i) + f_n^{(2)}(W_i) + f_n^{(3)}(W_i) + f_n^{(4)}(W_i)$.

PROOF. Put everything together. Apply a standard CLT for the second term in (9) together with Lemma 4 and Lemma 5 to get that

$$\hat{\beta}_\kappa - \beta_\kappa = \frac{1}{n} \sum_{i=1}^{n} f_n(W_i) + o_p(1/\sqrt{n})$$

where $f_n(W_i) = f^{(1)}(W_i) + f_n^{(2)}(W_i) + f_n^{(3)}(W_i) + f_n^{(4)}(W_i)$. This proves Theorem 1. □

Appendix B: Trimming and Local Polynomial Estimation

This appendix explains how our trimming scheme prevents boundary bias and ratio bias. It also explains why, in general, we choose a trimmed mean, rather than the mean of B, as the estimand of the localize-then-average estimation procedure. Finally, we explain why we do not estimate U_i with higher-order local polynomial estimators, despite their ability to automatically prevent boundary bias.

Recall that $\mathcal{X}^C$ denotes the vector of continuous endogenous components of X and Z^C denotes the vector of continuous components of Z. For simplicity, in the following discussion we assume that both X^C and Z^C are scalar random variables with joint support $\mathbb{R}^2$. We write $\mathcal{X}$ for $\mathcal{X}^C$ and $\mathcal{Z}$ for Z^C. We also assume for simplicity that $\mathcal{X}$ is the only endogenous component of X and $\mathcal{Z}$ is the only component of Z. The parameter of interest is the trimmed mean

$$\beta_\kappa = \mathbb{E}\beta(U)\tau_\kappa(\mathcal{X}, \mathcal{Z}) \tag{92}$$

which we estimate with

$$\hat{\beta}_\kappa = \frac{1}{n}\sum_{i=1}^{n} \hat{\beta}(\hat{U}_i)\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i) \tag{93}$$

where, for each $u \in S_U = [0,1]$, $\hat{\beta}(u) = \underset{b \in B_u}{\operatorname{argmax}}\, \hat{S}_n(b \mid u)$ and

$$\hat{S}_n(b \mid u) = \frac{1}{n\tau_n}\sum_{j=1}^{n}(2Y_j - 1)K_n^*(X_j b)K_n(\hat{U}_j - u)\tau_\kappa(\mathcal{Z}_j)\,. \tag{94}$$

Note that there are two trimming functions: $\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i) = \{|\mathcal{X}_i| \leq \kappa\}\{|\mathcal{Z}_i| \leq \kappa\}$ and $\tau_\kappa(\mathcal{Z}_j) = \{|\mathcal{Z}_j| \leq \kappa\}$. We discuss the role of each in preventing various types of bias.

Start with $\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)$. This trimming function prevents boundary bias and ratio bias. Start

with boundary bias. The standard kernel regression estimator $\hat{U}_i$ is an asymptotically biased estimator of U_i when U_i is within the bandwidth τ_n of the boundary of $S_U = [0,1]$. We call this boundary bias. It occurs for any fixed kernel used in standard kernel regression. Recall that $U_i = U(\mathcal{X}_i, \mathcal{Z}_i) = I\!P\{\mathcal{X} \leq \mathcal{X}_i \mid \mathcal{Z} = \mathcal{Z}_i\}$ and $\hat{U}_i = \hat{U}(\mathcal{X}_i, \mathcal{Z}_i) = \hat{I\!P}\{\mathcal{X} \leq \mathcal{X}_i \mid \mathcal{Z} = \mathcal{Z}_i\}$. Define

$$\begin{aligned} \mathcal{U} &= \sup_{|x|\leq\kappa, |z|\leq\kappa} U(x,z) \qquad \hat{\mathcal{U}} = \sup_{|x|\leq\kappa, |z|\leq\kappa} \hat{U}(x,z) \\ \mathcal{L} &= \inf_{|x|\leq\kappa, |z|\leq\kappa} U(x,z) \qquad \hat{\mathcal{L}} = \inf_{|x|\leq\kappa, |z|\leq\kappa} \hat{U}(x,z)\,. \end{aligned}$$

Since the support of $(\mathcal{X}, \mathcal{Z})$ is $I\!R^2$ and $\kappa < \infty$, $0 < \mathcal{L} < \mathcal{U} < 1$. By Lemma 2, $\hat{\mathcal{L}}$ converges in probability to $\mathcal{L}$ and $\hat{\mathcal{U}}$ converges in probability to $\mathcal{U}$. It follows that with probability tending to one as $n \to \infty$, $\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)$ trims $\hat{\beta}(\hat{U}_i)$ when $\hat{U}_i$ is within τ_n of 0 or $1 - \tau_n$ of 1. This guarantees that the only $\hat{\beta}(\hat{U}_i)$ values that play a role in the asymptotic analysis of $\hat{\beta}_\kappa$ are those whose $\hat{U}_i$ values are at least τ_n from the boundary of S_U where they are not subject to boundary bias.

The factor $\{|\mathcal{Z}_i| \leq \kappa\}$ in $\tau_\kappa(\mathcal{X}_i, \mathcal{Z}_i)$ also prevents so-called ratio bias. Consider the term in (11). This term involves the factors $\hat{U}_i - U_i$. In analyzing $\hat{U}_i - U_i$, terms of the form $[f(\mathcal{Z}_i) - \hat{f}(\mathcal{Z}_i)]/f(\mathcal{Z}_i)$ (and powers thereof) arise, where $\hat{f}(\mathcal{Z}_i)$ is a kernel density estimator of $f(\mathcal{Z}_i)$, the density of $\mathcal{Z}$ at $\mathcal{Z}_i$. (See, for example, the geometric expansion of $\hat{U}(x,z)$ in the proof of Lemma 2.) Note that

$$[f(\mathcal{Z}_i) - \hat{f}(\mathcal{Z}_i)]/f(\mathcal{Z}_i) = [f(\mathcal{Z}_i) - I\!E\hat{f}(\mathcal{Z}_i)]/f(\mathcal{Z}_i) + [I\!E\hat{f}(\mathcal{Z}_i) - \hat{f}(\mathcal{Z}_i)]/f(\mathcal{Z}_i)\,. \tag{95}$$

Conditional on $\mathcal{Z}_i$, the first term in this decomposition is a deterministic bias term and the second term is a zero-mean stochastic average. Both terms cause problems because of the presence of the density value $f(\mathcal{Z}_i)$ in their denominators. A bias reducing kernel of high enough order can make the numerator of the bias term arbitrarily small, but the ratio can still be large when $f(\mathcal{Z}_i)$ is

small. The stochastic term can cause even more serious problems. It's numerator cannot be made arbitrarily small, but rather has stochastic order no smaller than $O_p(1/\sqrt{n})$. Its ratio can be very large when $f(\mathcal{Z}_i)$ is small. However, since $\mathcal{Z}$ has support $\mathbb{R}$, these problems can only occur when $\mathcal{Z}_i$ is in one of the tails of the distribution of $\mathcal{Z}_i$. The trimming factor $\{|\mathcal{Z}_i| \leq \kappa\}$ prevents this from happening, by trimming the summand in (11) when $|\mathcal{Z}_i|$ gets too big. This prevents ratio bias.

Next, consider the trimming function $\tau_\kappa(\mathcal{Z}_j) = \{|\mathcal{Z}_j| \leq \kappa\}$ in (94). Asymptotic analysis of (10) involves Taylor expansions of the $\hat{U}_j$s about the corresponding U_js and so leads to analyses of the terms $\hat{U}_j - U_j$. By the same reasoning as given in the last paragraph, the function $\tau_\kappa(\mathcal{Z}_j)$ trims the jth summand in (10) when $|\mathcal{Z}_j|$ gets too big, thus preventing ratio bias in these terms.

We note that it is not necessary to do fixed trimming. Provided β exists, we can replace the fixed trimming constant κ with κ_n where $\kappa_n \to \infty$ as $n \to \infty$. The speed at which κ_n converges to infinity must be linked to assumptions about the tail behavior of $f(\mathcal{Z})$. However, such trimming implies that the estimand β_{κ_n} converges to β as $n \to \infty$. Practically speaking, the same effect is achieved by choosing a large fixed κ, and so for the sake of simplicity, we do fixed trimming.

While asymptotically negligible trimming of the sort just described is possible, it is not possible, in general, to take β itself as the estimand and still achieve $\sqrt{n}$-consistency. Establishing $\sqrt{n}$-consistency with β as the estimand would require showing that the difference $\beta_{\kappa_n} - \beta$ has order $O(1/\sqrt{n})$. A straightforward calculation shows that this would require that $\mathbb{P}\{|\mathcal{Z}| > \kappa_n\} = O(1/\sqrt{n})$. This, in turn, would require that the density of $\mathcal{Z}$ at $\pm\kappa_n$ be converging to zero very rapidly. However, this same density appears in the denominator of the terms in (95). To prevent ratio bias in these terms it is necessary that the density of $\mathcal{Z}$ at $\pm\kappa_n$ be converging to zero very slowly. It is easy to show that in general, these two conflicting demands cannot be met simultaneously. The real culprit is the stochastic term in decomposition (95). Even if the bias term is

identically zero, the stochastic term prevents these conflicting demands from being met. It follows that, apart from special cases when $\beta_\kappa = \beta$ for all $\kappa > 0$ or when Z is discrete so that instrument trimming is unnecessary, if we want to achieve a $\sqrt{n}$-consistent estimator, we must live with a trimmed mean of the distribution of B as an estimand. This is true no matter what estimator we use to estimate U_i. For example, this is true even if we were to replace the standard kernel regression estimators of U_i with general local polynomial (LP) estimators (Fan and Gijbels, 1996).

We estimate U_i with the standard kernel regression estimator, also known as the Nadaraya-Watson (NW) estimator. This is a local polynomial estimator where the local polynomial is a constant. While the NW estimator with bias reducing kernels of high enough order can achieve an arbitrary degree of bias reduction on the interior of the support of the localizing variable, it is an asymptotically biased estimator near the boundaries of the support. We trim on $\mathcal{X}_i$ as well as $\mathcal{Z}_i$ in (93) to prevent this bias, as explained above. However, a comparable higher-order local polynomial estimator can achieve the same degree of bias reduction on the interior as well as near or at the boundary of the support. There is no need to trim on $\mathcal{X}_i$ and $\mathcal{Z}_i$ to prevent boundary bias. So why not use the higher-order LP estimator instead of the NW estimator with bias reducing kernels?

We cite two reasons. First, it is not known (to the authors, at least) whether the known pointwise bounds on the bias of LP estimators at the boundaries of the support of the localizing variable are uniform in the localizing variable. This uniformity is needed to show that remainder terms in asymptotic arguments are small in the appropriate sense.

Secondly, even if the uniformity conditions hold, formally establishing $\sqrt{n}$-consistency and asymptotic normality of $\hat{\beta}_\kappa$ when U_i is estimated with a general local polynomial estimator would be extraordinarily complicated. To see why, assume once again for simplicity that U_i is scalar. A

local polynomial estimator of U_i of degree p can be written as the weighted average

$$\sum_{a=1}^{n}\{\mathcal{X}_a \leq \mathcal{X}_i\}w_a^{(p)} / \sum_{a=1}^{n} w_a^{(p)}$$

where, for $p > 0$, the weights $w_a^{(p)}$ depend on all the $\mathcal{Z}_i$. For any positive integer m, define $s_m = \sum_{a=1}^{n}(\mathcal{Z}_a - \mathcal{Z}_i)^m K_n(\mathcal{Z}_a - \mathcal{Z}_i)$. For simplicity, we suppress the dependence of s_m on n. Consider the cases $p = 0, 1, 2$, corresponding to the NW, local linear, and local quadratic estimators, respectively. It is straightforward (though tedious) to show that

$$\begin{aligned}
w_a^{(0)} &= K_n(\mathcal{Z}_a - \mathcal{Z}_i) \\
w_a^{(1)} &= K_n(\mathcal{Z}_a - \mathcal{Z}_i)[s_2 - (\mathcal{Z}_a - \mathcal{Z}_i)s_1] \\
w_a^{(2)} &= K_n(\mathcal{Z}_a - \mathcal{Z}_i)\left[[s_2 s_4 - s_3^2] - (\mathcal{Z}_a - \mathcal{Z}_i)[s_1 s_4 - s_2 s_3] + (\mathcal{Z}_a - \mathcal{Z}_i)^2[s_1 s_3 - s_2^2]\right].
\end{aligned}$$

Recall that the use of the NW estimator of U_i leads to a complicated analysis of U-statistics of orders 2, 3, and 4 in the proof on Lemma 4. Each of these U-statistics is painstakingly analyzed by means of the Hoeffding decomposition to extract its nonnegligible contribution to the asymptotic distribution of $\hat{\beta}_\kappa$. Now consider local linear estimation. The weight $w_a^{(1)}$ for the local linear estimator is itself a sum and would lead to an analysis of U-statistics of orders 3, 4, and 5 in the proof of Lemma 4. The weight $w_a^{(2)}$ for the local quadratic estimator is a double sum and would lead to an analysis of U-statistics of orders 4, 5, and 6. In general, the weight $w_a^{(p)}$ is a sum over p indices and would lead to the analysis of U-statistics of order $p+2$, $p+3$, and $p+4$. And this is only for scalar U_i. The analysis would be far more complicated for vector-valued U_i.

To avoid this added complexity, we use the NW estimator with higher-order bias reducing kernels. By doing so, we achieve the same order of bias reduction on the interior of the support of

the localizing variable as we would by using comparable higher-order local polynomial estimators. By trimming on $\mathcal{X}_i$ (as well as $\mathcal{Z}_i$) we eliminate the problems with bias near the boundary of the support of the localizing variable.

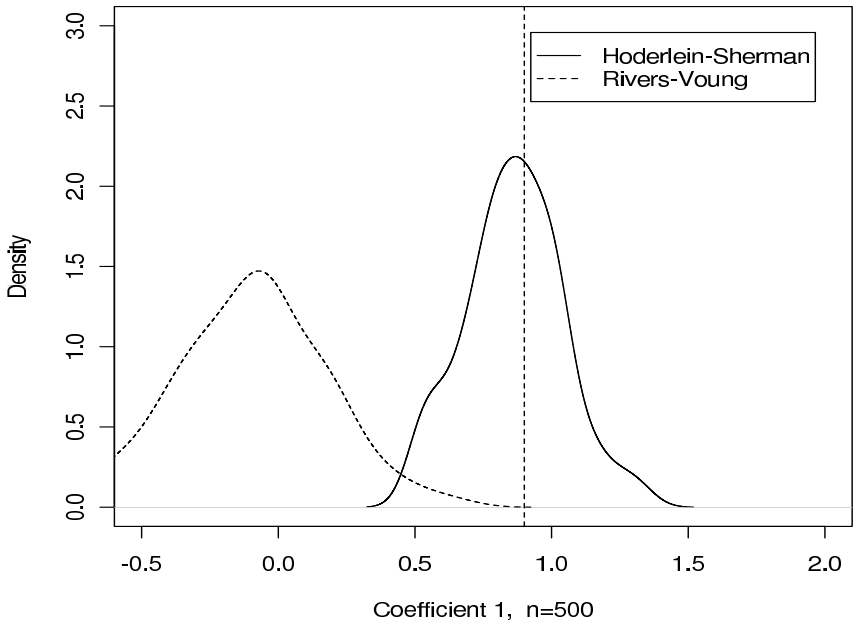
Hoderlein-Sherman
Rivers-Voung
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-0.5
0.0
0.5
1.0
1.5
2.0
Coefficient 1, n=500

Figure 1:

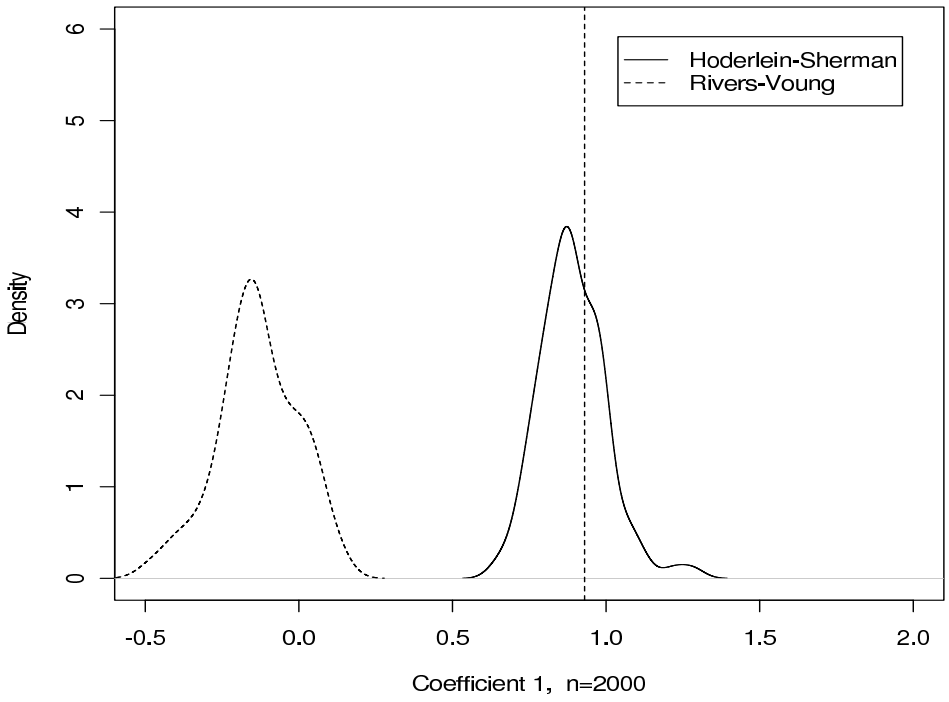
Hoderlein-Sherman
Rivers-Voung
Density
0
1
2
3
4
5
6
-0.5
0.0
0.5
1.0
1.5
2.0
Coefficient 1, n=2000

Figure 2:

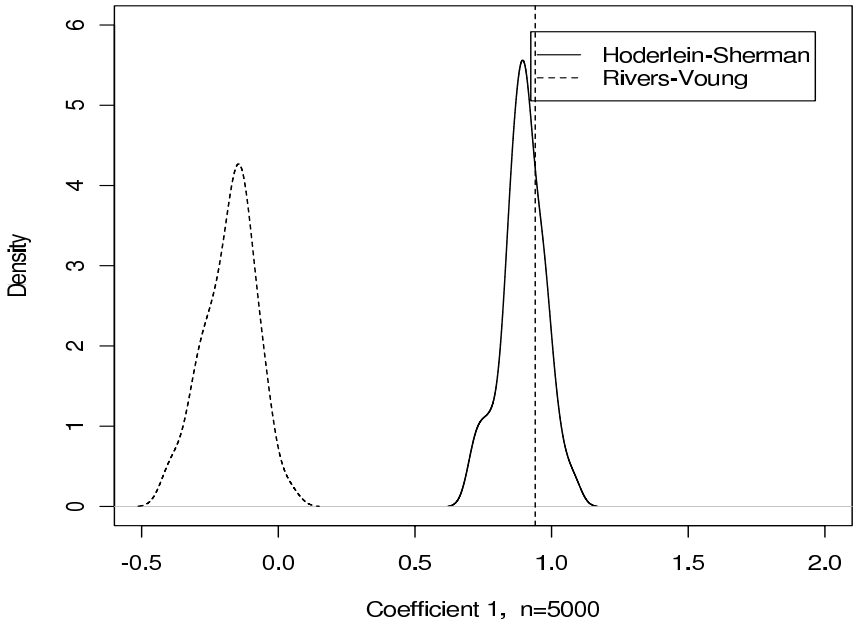
Hoderlein-Sherman
Rivers-Voung
Density
0
1
2
3
4
5
6
-0.5
0.0
0.5
1.0
1.5
2.0
Coefficient 1, n=5000

Figure 3:

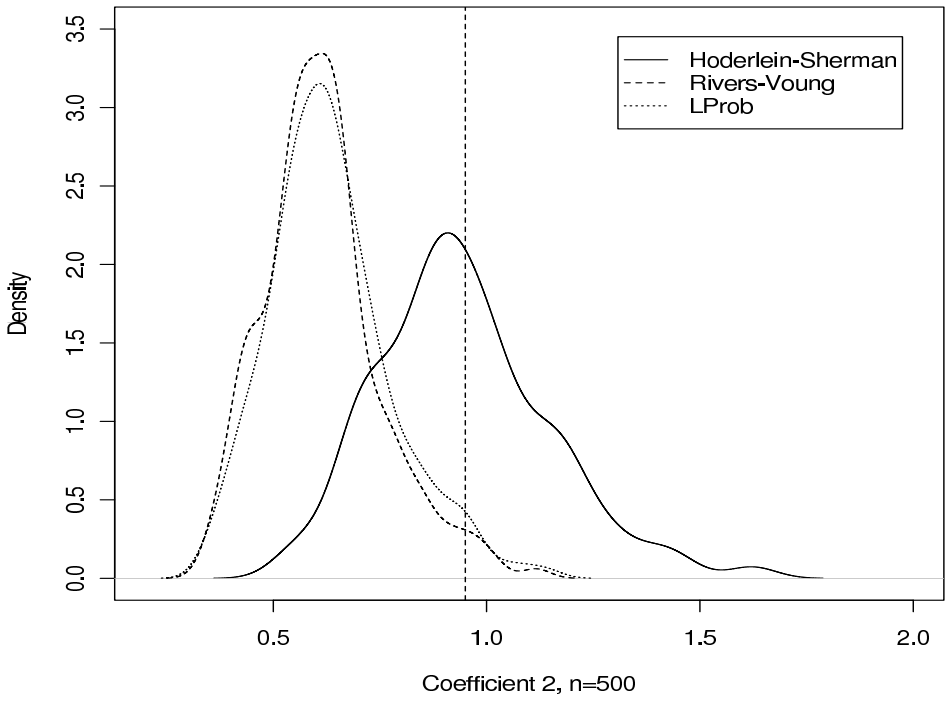
Hoderlein-Sherman
Rivers-Voung
LProb
Density
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
0.5
1.0
1.5
2.0
Coefficient 2, n=500

Figure 4:

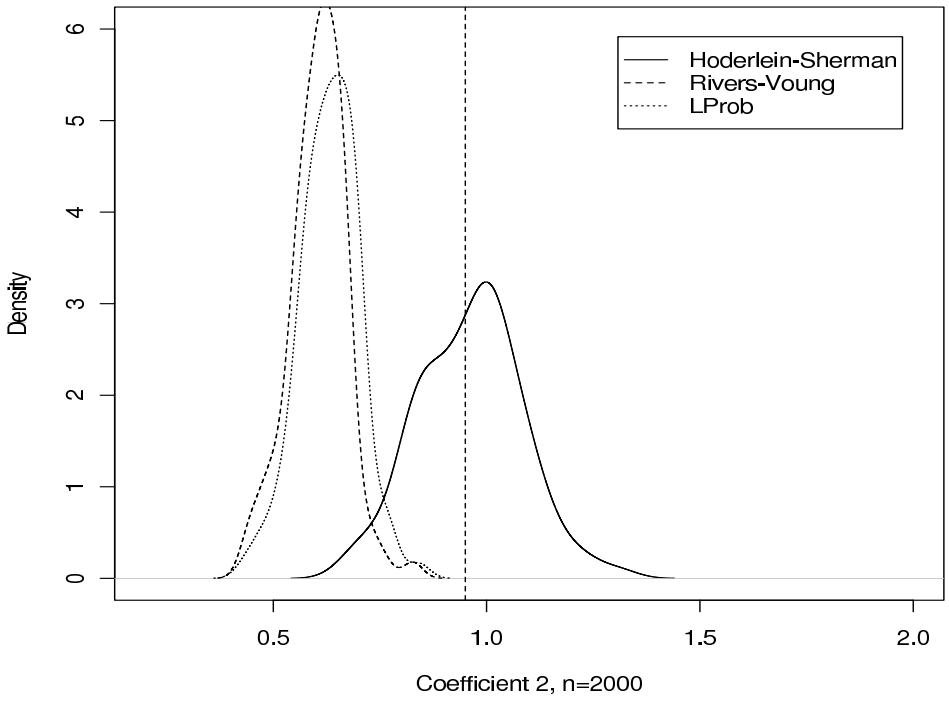
Hoderlein-Sherman
Rivers-Voung
LProb
Density
0
1
2
3
4
5
6
0.5
1.0
1.5
2.0
Coefficient 2, n=2000

Figure 5:

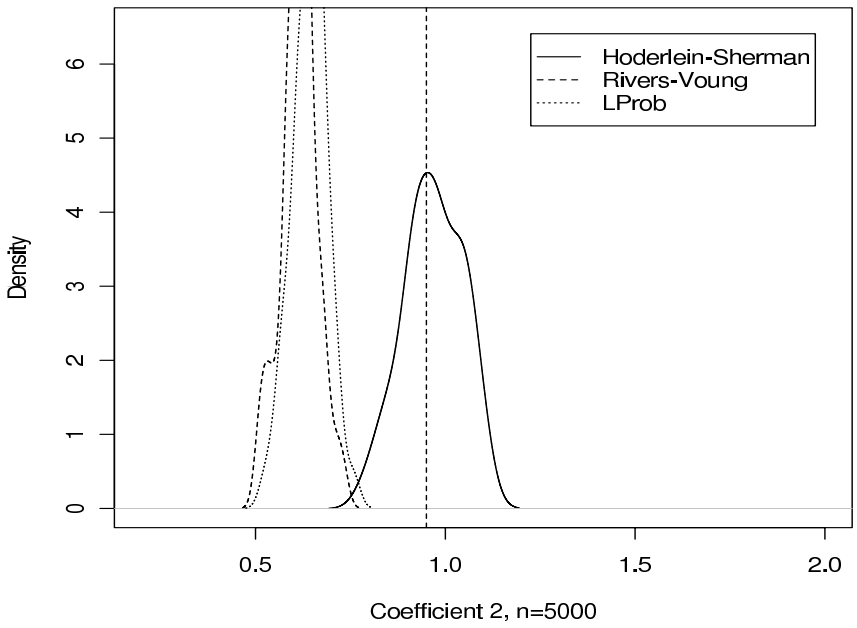
Hoderlein-Sherman
Rivers-Voung
LProb
Density
0
1
2
3
4
5
6
0.5
1.0
1.5
2.0
Coefficient 2, n=5000

Figure 6: