cemmap
centre for microdata methods and practice

Instrumental variables estimation of a generalized correlated random coefficients model

Matthew Masten
Alexander Torgovitsky

The Institute for Fiscal Studies
Department of Economics, UCL

cemmap working paper CWP02/14

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

Instrumental Variables Estimation of a Generalized Correlated Random Coefficients Model*

Matthew A. Masten† Alexander Torgovitsky‡

January 1, 2014

Abstract

We study identification and estimation of the average treatment effect in a correlated random coefficients model that allows for first stage heterogeneity and binary instruments. The model also allows for multiple endogenous variables and interactions between endogenous variables and covariates. Our identification approach is based on averaging the coefficients obtained from a collection of ordinary linear regressions that condition on different realizations of a control function. This identification strategy suggests a transparent and computationally straightforward estimator of a trimmed average treatment effect constructed as the average of kernel-weighted linear regressions. We develop this estimator and establish its $\sqrt{n}$–consistency and asymptotic normality. Monte Carlo simulations show excellent finite-sample performance that is comparable in precision to the standard two-stage least squares estimator. We apply our results to analyze the effect of air pollution on house prices, and find substantial heterogeneity in first stage instrument effects as well as heterogeneity in treatment effects that is consistent with household sorting.

JEL classification: C14; C26; C51

Keywords: correlated random coefficients, instrumental variables, unobserved heterogeneity, semiparametrics, hedonic models, residential sorting, valuation of clean air

*This paper was presented at the UW–Milwaukee Mini-Conference on Microeconometrics, Duke, Northwestern, Chicago, the Harvard/MIT joint workshop, Ohio State, Boston College, the Triangle econometrics conference, University College London, and the First Conference in Econometric Theory at Universidad de San Andrés. We thank audiences at those seminars as well as Federico Bugni, Ivan Canay, Matias Cattaneo, Andrew Chesher, Bryan Graham, Jim Heckman, Stefan Hoderlein, Joel Horowitz, Max Kasy, Shakeeb Khan, Jia Li, Arnaud Maurel, Whitney Newey, Maya Rossin-Slater, Elie Tamer, Duncan Thomas, Chris Timmins, and Ed Vytlacil for helpful conversations and comments. We thank Fu Ouyang for research assistance.

†Department of Economics, Duke University, matt.masten@duke.edu

‡Department of Economics, Northwestern University, a-torgovitsky@northwestern.edu

1 Introduction

This paper is about the linear correlated random coefficients (CRC) model. In its simplest form, the model can be written as

$$Y = B_0 + B_1 X, \tag{1}$$

where Y is an outcome variable, X is an explanatory variable and $B \equiv (B_0, B_1)$ are unobservable variables. The explanatory variable X is endogenous in the sense that it may be statistically dependent with B_0 and B_1. Concerns about endogeneity are widespread in economic applications and are often addressed by using the variation of an instrumental variable, Z, that is plausibly independent (or uncorrelated) with (B_0, B_1), but correlated with X. The most common tool for doing this is the two-stage least squares (TSLS) estimator. However, unless the partial effect of X on Y, i.e. B_1, is a degenerate random variable (a constant), the estimand of the TSLS estimator is not necessarily an easily interpretable feature of the distribution of B_1. Assuming that B_1 is constant is tantamount to assuming that the treatment effect of X on Y is homogenous. As many authors have discussed, the theoretical and empirical evidence does not support the assumption of homogenous treatment effects. See Heckman (2001) and Imbens (2007) for thorough expositions of this point.

To address this problem, several authors have explored auxiliary assumptions under which the TSLS estimand becomes a parameter of interest. The most influential finding is that of Imbens and Angrist (1994), who show that if both X and Z are binary and if Z affects X monotonically, then the TSLS estimator is consistent for the local average treatment effect (LATE). The LATE parameter has generated significant debate over whether it is actually a quantity that economists should be interested in; see, for example, the *Journal of Economic Perspectives* (2010) and the *Journal of Economic Literature* (2010) symposia. However, as the support of X grows from binary to multi-valued discrete to continuous, the TSLS estimand becomes an increasingly complicated weighted average of LATEs between different X realizations (Angrist and Imbens 1995). Even if one finds a solitary LATE to be an interesting parameter, the interpretation, economic significance, and policy relevance of such weighted averages of LATEs is more tenuous. A less controversial parameter of interest is the average treatment effect (ATE), which due to the linearity in (1) is determined by the average partial effect (APE), $\mathbb{E}(B_1)$. In a series of papers, Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2008) showed that if the effect of Z on X is homogenous, then TSLS will be consistent for $\mathbb{E}(B_1)$. This type of homogeneity assumption is somewhat

unsatisfying, since accounting for heterogeneity is the main motivation for considering this problem to begin with. (See also Li and Tobias (2011), who consider Bayesian inference in those models.)

An alternative is to consider different instrumental variables estimators besides TSLS. Florens, Heckman, Meghir, and Vytlacil (2008) take this approach in considering a polynomial version of (1) plus an additive nonparametric function of X common to all units. They show that the ATE is identified if X is continuously distributed and there exists a function h that is strictly increasing in a scalar unobservable V such that $X = h(Z, V)$. This type of first stage restriction allows for heterogeneity in the effect of Z on X, albeit in a limited form, and so directly addresses the concerns about previous work by Heckman, Vytlacil and Wooldridge. The utility of the first stage restriction is in creating a random variable R which is a control function in the sense that $X \perp\!\!\!\perp B|R$. A central contribution of our paper is to exploit this control function property to provide an alternate identification approach to the one considered by Florens et al. (2008). Our approach has three main benefits relative to that of Florens et al. (2008). First, while Florens et al. (2008) require a continuous instrument (see the discussion on page 8), we can achieve identification with binary and discrete instruments in many cases. Second, our approach enables us to include multiple endogenous variables, non-polynomial terms and interactions between endogenous variables and covariates in more general linear-in-coefficients specifications of (1). Third, it suggests a computationally straightforward estimator that appears to have good finite sample properties. The main drawbacks of our approach relative to that of Florens et al. (2008) is that their model allows for a common additive nonparametric function of X, and, when Z is continuous, their "measurable separability" assumption may hold in some cases that our corresponding relevance condition does not.

Our results build on recent research on nonparametric identification in nonseparable models. A recurring finding in this work is a trade-off between the dimension of heterogeneity and the required variation in the instrument Z. At one extreme lie the papers by Imbens and Newey (2009) and Kasy (2013), who show that unrestricted forms of heterogeneity can be allowed in the outcome and/or first stage equations while still attaining point identification of the ATE, as long as Z satisfies a large support assumption (they also provide sharp partial identification results when the large support assumption does not hold). Despite their ubiquity across the econometric theory literature, such large support assumptions are unlikely to ever be even approximately satisfied in practice. In particular, they rule out the binary and discrete instruments that are commonly found in applied work, such as

policy shifts, institutional changes, and natural experiments. On the other hand, work by Chernozhukov and Hansen (2005), Torgovitsky (2012) and D'Haultfœuille and Février (2012) has shown that binary and discrete instruments of this sort can secure identification, as long as the dimension of heterogeneity is sufficiently restricted. These restrictions on heterogeneity rule out simple, parsimonious specifications like (1) which contain more than one unobservable. Between these two extremes lies the paper by Florens et al. (2008) and also those of Chesher (2003) and Masten (2012), both of which require a continuous instrument with small support but also allow for additional heterogeneity. Our paper contributes to this middle ground and, among other things, provides an example where a broadly interesting parameter can be identified in a model with high-dimensional heterogeneity and discrete instruments.

The recent work of Graham and Powell (2012) (who build on work by Chamberlain 1992) on CRC models with panel data is related in motivation to this paper. Both papers seek to identify the APE—at least among some subpopulation—but the analysis is fundamentally different due to differences between using panels and instruments as sources of identification. Partially related to their paper as well as ours is the literature on random uncorrelated coefficient models; for example, Beran and Hall (1992) and Hoderlein, Klemelä, and Mammen (2010). That literature assumes X and (B_0, B_1) are independent and centers on estimating the distribution of (B_0, B_1). In contrast, we limit our focus to identifying averages, but have to contend with the difficulty of dependence between X and (B_0, B_1).

An advantage of our identification approach and the linear structure in (1) is that it facilitates estimators that are precise, easy to implement, and which do not suffer from the curse of dimensionality. A main contribution of our paper is to develop such an estimator of $\mathbb{E}(B)$ and establish its $\sqrt{n}$–consistency and asymptotic normality. (Due to uniformity issues, we actually develop asymptotic theory for an estimator of a trimmed version of $\mathbb{E}(B)$; see section 4.) Our estimator is essentially an average of ordinary linear regressions run conditional on a realization of a control function and so shares similarities with the control function approaches of, for example, Blundell and Powell (2004), Imbens and Newey (2009), Rothe (2009), Hoderlein and Sherman (2013) and Torgovitsky (2013). The control function is estimated with a first stage quantile or distribution regression and the conditioning is approximated with kernel weights. Hence, our estimator reduces to a straightforward average of weighted linear regressions, where the weights are determined by a first stage quantile or distribution regression of X on Z. Incorporating covariates is a simple matter of including them in these linear mean and quantile regressions. Monte Carlo experiments show that

our estimator can perform as well or better than the TSLS estimator under conditions when both would be consistent, while remaining consistent in situations where TSLS would be inconsistent.

We apply our results to study the effect of air pollution on house prices. We follow the empirical approach of Chay and Greenstone (2005), who argue that instrumenting is necessary to deal with unobserved economic shocks and sorting of households based on unobserved preferences for clean air. They also argue that this sorting leads to correlated random coefficients. They define a binary instrument based on regulation implemented by the 1970 Clean Air Act Amendments. We demonstrate substantial first stage heterogeneity in the effect of this instrument, which strongly suggests that the simpler estimators discussed by Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2008) would be inconsistent for the APE. Likewise, the binary instrument precludes approaches which rely on continuous variation, such as Florens et al. (2008). For two subsets of counties where the instrument has a statistically significant effect on pollution levels, we estimate unweighted APEs of changes in pollution on changes in house prices. These estimates demonstrate patterns that are consistent with household sorting. Taken together, these estimates along with TSLS suggest there is substantial heterogeneity in households' valuation of clean air.

The structure of the paper is as follows. In the next section we formally discuss the model, assumptions and our identification results. In Section 3, we describe our estimator and discuss its implementation. In Section 4, we analyze the asymptotic properties of our estimator. In Section 5, we report the results of Monte Carlo studies that demonstrate the performance of our estimator. Finally, in Section 6, we present our application to air pollution and house prices. Section 7 concludes.

2 Model and Identification

A general version of model (1) is

$$Y = B_0 + \sum_{j=1}^{d_x} B_j X_j + \sum_{j=1}^{d_1} B_{d_x+j} Z_{1j} \equiv W'B, \tag{2}$$

where $X \in \mathbb{R}^{d_x}$ is a vector of potentially endogenous variables, $Z_1 \in \mathbb{R}^{d_1}$ is a vector of included exogenous variables with j^{th} component Z_{1j}, $W \equiv [1, X', Z_1']' \in \mathbb{R}^{d_w}$ with $d_w \equiv 1 + d_x + d_1$, and $B \in \mathbb{R}^{d_w}$ is a vector of unobservable variables. In addition to Z_1, there is a vector of excluded exogenous variables (instruments) $Z_2 \in \mathbb{R}^{d_2}$ that do not directly affect Y

in (2). We write the exogenous variables together as $Z \equiv [Z_1', Z_2']' \in \mathbb{R}^{d_z}$ with $d_z \equiv d_1 + d_2$.

We divide the vector of endogenous variables into subvectors of lengths $d_b \geq 1$ and $d_x - d_b \geq 0$. We refer to the first d_b components of X as the *basic* endogenous variables and the last $d_x - d_b$ components of X as the *derived* endogenous variables. We assume that the basic endogenous variables satisfy a particular first stage structure that is specified in the assumptions ahead. In contrast, the derived endogenous variables are assumed to be functions of the basic endogenous variables and the included exogenous variables Z_1. For example, we could have $d_b = 1$ and derived endogenous variables $X_k = X^k$ for $k > d_b$, as in the model of Florens et al. (2008). Or, we could have $X_k = X_1 Z_1$ for some $k > d_b$ be an interaction variable, which would allow the distribution of partial effects to differ arbitrarily across values of Z_1. This allows, for example, men and women to have different distributions of treatment effects, allowing for heterogeneity on observables to be dealt with in the usual way.

Throughout our analysis we frequently use the random vector

$$R \equiv [F_{X_1|Z}(X_1|Z), \ldots, F_{X_{d_b}|Z}(X_{d_b}|Z)]',$$

which we refer to as the *conditional rank* of X. We are only concerned with the conditional ranks of the basic endogenous variables, since under our assumptions the conditional ranks of the derived endogenous variables $F_{X_k|Z}(X_k|Z)$ for $k = d_b + 1, \ldots, d_x$ will contain less information. Below, we will restrict X_k to be continuously distributed for $k = 1, \ldots, d_b$ so that R_k is distributed uniformly on $[0, 1]$ for these k. Note, however, that if $d_b > 1$ then the support of R may be a proper subset of $[0, 1]^{d_b}$. Consider the following assumptions.

Assumption I.

I1. (Existence of moments) $\mathbb{E}(B) < \infty$ *and* $\mathbb{E}(\|W\|^2) < \infty$.

I2. (First stage equation) *For each basic endogenous variable $k = 1, \ldots, d_b$, there exists a scalar random variable V_k and a possibly unknown function h_k that is strictly increasing in its second argument, for which $X_k = h_k(Z, V_k)$. The vector $V \equiv (V_1, \ldots, V_{d_b})$ is continuously distributed.*

I3. (Derived endogenous variables) *For each $k = d_b + 1, \ldots, d_x$, there exists a known function g_k such that $X_k = g_k(X_1 \ldots, X_{d_b}, Z_1)$.*

I4. (Instrument exogeneity) $(B, V) \perp\!\!\!\perp Z$.

I5. (Instrument relevance) $\mathbb{E}[WW'|R=r]$ *is invertible for almost every* r *in a known Lebesgue measurable set* $\mathcal{R} \subseteq \text{supp}(R)$.

Theorem 1. *Define* $\beta(r) \equiv \mathbb{E}[B|R=r]$. *Under Assumptions I,*

$$\beta(r) = \mathbb{E}[WW'|R=r]^{-1}\,\mathbb{E}[WY|R=r]$$

for any $r \equiv (r_1, \ldots, r_{d_b}) \in \mathcal{R}$. *Hence both* $\beta(r)$ *and* $\beta_{\mathcal{R}} \equiv \mathbb{E}[B|R \in \mathcal{R}]$ *are point identified.*

The proof of Theorem 1 uses the following implication of I2 and I4, which has been used and analyzed in various forms by Imbens (2007), Florens et al. (2008), Imbens and Newey (2009), Kasy (2011) and Torgovitsky (2012). Since our version is a slight extension, we provide a short proof in the appendix.

Proposition 1. *I2 and I4 imply that* $(R, B) \perp\!\!\!\perp Z$. *If I3 also holds, then* $W \perp\!\!\!\perp B|R$.

Proof of Theorem 1. I1 ensures that all conditional moments of interest exist. Premultiplying both sides of (2) by W and taking expectations conditional on $R=r$ for any $r \in \mathcal{R}$, we have

$$\mathbb{E}[WY|R=r] = \mathbb{E}[WW'B|R=r] = \mathbb{E}[WW'|R=r]\beta(r),$$

by Proposition 1. Given I5, we can premultiply both sides by the inverse of $\mathbb{E}[WW'|R=r]$ to obtain the claimed expression for $\beta(r)$. Since $\mathbb{E}[WW'|R=r]^{-1}$ and $\mathbb{E}[WY|R=r]$ are features of the observable data, this shows that $\beta(r)$ and $\beta_{\mathcal{R}} \equiv \mathbb{E}[B|R \in \mathcal{R}]$ are both identified for any known $\mathcal{R} \subseteq \text{supp}(R)$ satisfying I5. *Q.E.D.*

The intuition behind Theorem 1 is as follows. After conditioning on $R=r$, all of the variation in the basic endogenous variables is due to variation in Z, by the definition of R. Since the derived endogenous variables are functions of the basic endogenous variables and Z_1, all of the variation in W conditional on $R=r$ is also due to variation in Z. Variation in Z, however, is independent of B conditional on $R=r$ by instrument exogeneity (I4) via Proposition 1. As a result, a linear regression of Y on X conditional on $R=r$ identifies $\beta(r) \equiv \mathbb{E}[B|R=r]$. Averaging $\mathbb{E}[B|R=r]$ over $r \in \mathcal{R}$ then yields $\beta_{\mathcal{R}} \equiv \mathbb{E}[B|R \in \mathcal{R}]$. If instrument relevance (I5) holds for some measure one subset of $\text{supp}(R)$, then $\beta_{\mathcal{R}} = \mathbb{E}[B]$ is identified. This intuition is illustrated in Figure 1.

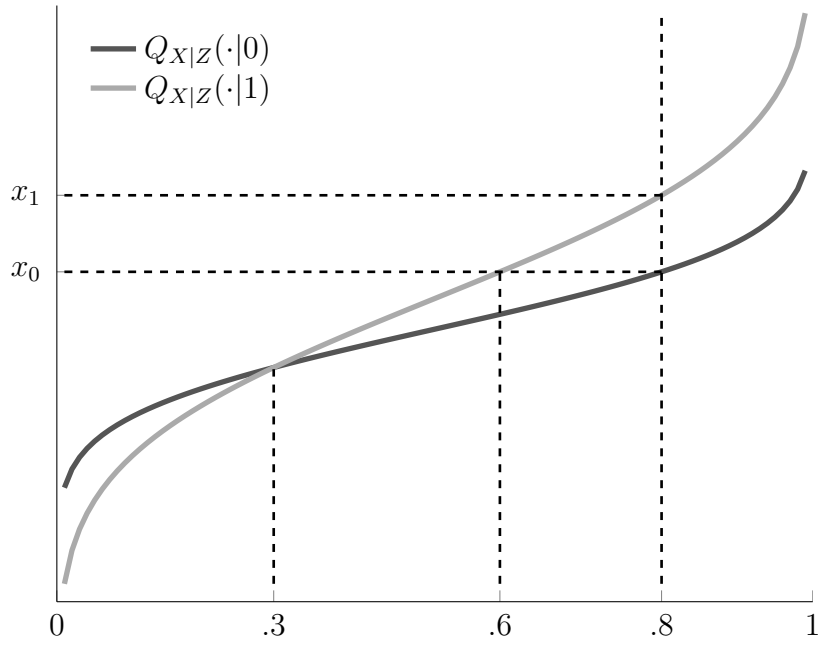

Figure 1: Consider the simple CRC model (1), $Y = B_0 + B_1 X$, and suppose Z is binary. Conditional on $R = 0.8$, X assumes two values (x_0 and x_1) depending on the realization of Z. Since $Z \perp\!\!\!\perp B \mid \{R = 0.8\}$, a mean regression of Y on X conditional on $R = 0.8$ identifies the means of the intercept and slope coefficients, $\mathbb{E}(B_0 \mid R = 0.8)$ and $\mathbb{E}(B_1 \mid R = 0.8)$. For the plotted quantile functions, the relevance condition I5 holds for almost every $r \in (0, 1)$, since the curves intersect at only one point. Hence the previous argument yields $\mathbb{E}(B \mid R = r)$ for all $r \in (0, 1)$ and averaging then gives $\mathbb{E}(B)$. Note also that the instrument's effect is nonmonotonic—it is positive for units with large R (above $R = 0.3$) and negative for units with small R.

Theorem 1 is complementary to a result by Florens et al. (2008). Those authors consider a model with a single basic endogenous variable X and the outcome equation

$$Y = \varphi(X) + B_0 + B_1 X + B_2 X^2 + \cdots + B_K X^K,$$

for some pre-specified K, where φ is an unknown function, and $(B_0, \ldots, B_K)$ are random coefficients that are potentially correlated with X. Except for φ, this outcome equation can be obtained from (2) with basic endogenous variable X, and derived endogenous variables $(X^2, \ldots, X^K)$. The price of including the φ function is that Florens et al. (2008) require a continuous small support instrument (see their identification proof on page 1203, the step from equation 10 to the next line). We do not include the φ function, but are generally able to achieve identification of the average coefficients in the polynomial outcome equation model so long as the distribution of Z has at least $K + 1$ support points. Florens et al. (2008) also

maintain I2 and I4, but in place of I5 they impose a "measurable separability" condition that is somewhat high-level. As those authors discuss, their measurable separability condition may fail if the first stage equation is not continuous in V. In contrast, our relevance condition I5 does not require such continuity. This allows for the support of X conditional on Z to be disjoint.

I5 is directly analogous to the standard no-multicollinearity condition in ordinary least squares and consequently requires the analyst to avoid standard causes of failure, such as the dummy variable trap. When $d_x = d_b = 1$, so that there is a single basic endogenous variable and no derived endogenous variables, I5 requires that $\text{Var}[Q_{X|Z}(r \mid Z)] > 0$ for all $r \in \mathcal{R}$. If $Z \in \{0, 1\}$ is binary, then $\text{Var}[Q_{X|Z}(r \mid Z)] > 0$ happens if and only if $Q_{X|Z}(r \mid 0) \neq Q_{X|Z}(r \mid 1)$; that is, the two curves in Figure 1 are separated at r. Since $Q_{X|Z}(r \mid Z) = h[Z, Q_V(r)]$ by strict monotonicity (I2) and independence (I4), we must have that for each $r \in \mathcal{R}$ there are distinct $z, z' \in \text{supp}(Z)$ with $h[z, Q_V(r)] \neq h[z', Q_V(r)]$. Hence, for all units with first stage unobservables $v = Q_V(r)$ for which we want to learn $\mathbb{E}(B)$, the instrument must affect those units' endogenous variable. Generally, whether I5 holds is an empirical matter in the sense that the condition only depends on the distribution of observables and so, at least in principle, can be checked in the data.

When I5 only holds for some proper subset $\mathcal{R}$ of $\text{supp}(R)$ then Theorem 1 identifies $\beta_{\mathcal{R}} \equiv \mathbb{E}[B|R \in \mathcal{R}]$, which generally will not equal $\mathbb{E}[B]$. Nevertheless, $\beta_{\mathcal{R}}$ has an interpretation similar to the unweighted LATE of Imbens and Angrist (1994). That is, $\beta_{\mathcal{R}}$ is the unweighted average of B for those agents for whom the instrument has an effect. Note that we do not require this effect to be monotonic. If Z is assumed to have a monotonic effect on X (as in Imbens and Angrist 1994), then $\beta_{\mathcal{R}}$ is the unweighted average of B for those agents who are induced to increase their treatment intensity X due to a change in Z. This type of parameter may be of comparable (or even greater) interest than $\mathbb{E}[B]$ for a policy maker considering a policy change that affects the determination of X through an incentive Z.

While I5 may fail for some subset of $\text{supp}(R)$, it is an intuitively appealing requirement for an instrument. Agents characterized by an r at which $\mathbb{E}[WW'|R = r]$ is singular do not experience independent variation in W due to variation in Z, and so it is natural that $\mathbb{E}[B|R = r]$ should not be identifiable for those agents. Assuming that the effect of Z on X is homogenous, as in Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2008), ignores this distinction and explicitly includes agents in the average for whom the instrument might be completely ineffectual—in effect, extrapolating from Z-sensitive agents to Z-insensitive agents. Similarly, the measurable separability condition of Florens et al. (2008) could ap-

parently hold even if there is a non-negligible subset of agents for whom the instrument is irrelevant.

As in standard linear regression analysis, identification of $\mathbb{E}(B)$ (or some conditional version of it) via Theorem 1 provides identification of the ATE and APE when the outcome equation includes nonlinear functions of X or interactions with covariates Z_1. This is an elementary point, but we mention it for clarity. Suppose that $Y = B_0 + B_1 X + B_2 X Z_1$. Then the APE is given by $\mathbb{E}[B_1] + \mathbb{E}[B_2]\,\mathbb{E}[Z_1]$ while the ATE for an exogenous change from x to $\overline{x}$ is given by $(\mathbb{E}[B_1] + \mathbb{E}[B_2]\,\mathbb{E}[Z_1])(\overline{x} - x)$. Both of these quantities can be obtained from estimates of $\mathbb{E}[B_1]$, $\mathbb{E}[B_2]$ and $\mathbb{E}[Z_1]$. Alternatively, an analyst may be interested in the APE for some predetermined value of z_1, which would be given by $\mathbb{E}[B_1] + \mathbb{E}[B_2] z_1$. When (2) contains nonlinear terms, e.g. $Y = B_0 + B_1 X + B_2 X^2$, then an analyst may be more interested in reporting $\mathbb{E}[B_1] + 2\,\mathbb{E}[B_2] x$ as the APE when X is exogenously set to x. All of these quantities can be obtained after applying our identification results.

Among the maintained assumptions for Theorem 1, I2 is generally the most controversial. While it is more flexible than the homogenous effect specifications of Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2008), it does restrict the basic endogenous variables to be continuous and also limits the heterogeneity in their first stage equations to have dimension one. One-dimensional heterogeneity of the sort in I2 can be interpreted as "rank invariance" in the effect of Z on each basic component of X. (The concept of rank invariance was first introduced by Doksum 1974.) Rank invariance means that the ordinal ranking of any two agents in terms of any component of X_k ($k \leq d_b$) would be the same if both agents received the same realization of Z, for any realization of Z. See Heckman, Smith, and Clements (1997), Chernozhukov and Hansen (2005) and Torgovitsky (2012) for further discussions of rank invariance. While one-dimensional heterogeneity is restrictive, there are few alternatives in the literature that allow for high-dimensional heterogeneity in both the outcome and first stage equations while attaining point identification of a broadly interpretable parameter. An important exception to this is the work of Kasy (2013), who obtains such a result but under the assumption that Z affects a scalar X monotonically and also has large support.

Assumptions I2 and I4 together generally imply that correlation between X and the random coefficients B must occur through V. For example, specifying B as a direct function of X, such as setting $B_0 = X$, implies that, conditional on R, some variation remaining in B is due to Z, and hence I4 will typically not hold. Thus, I2 and I4 should be viewed as also placing restrictions on the manner in which X and B may be dependent. This

point is not unique to our model—even in a simple textbook model (1) with a constant B_1, data generating processes like $B_0 = X$ will violate the usual uncorrelatedness assumption $\mathbb{E}(ZB_0) = 0$. Consequently, if we wish to express model (1) in terms of potential outcomes, it is helpful to view (B_0, B_1, V) as unobserved heterogeneity parameters which are intrinsic to each unit. After the instrument is assigned, the value of X is determined via the first stage equation of I2 and then the value of Y is determined through (1). Thus the average partial effect $\mathbb{E}(B_1)$ tells us the average effect of exogenously increasing X by one for all units.

In addition to the overall average of $\mathbb{E}(B)$ identified in Theorem 1, the following result shows that averages for groups determined by their treatment intensity are also identified. This parameter is analogous to the "effect of the treatment on the treated" parameter defined in Florens et al. (2008).

Theorem 2. *Under Assumptions I, the "average effect of treatment on the treated" parameter* $\mathbb{E}(B \mid X_k = x_k, k \leq d_b)$ *is point identified for any* $x = (x_1, \ldots, x_{d_b}) \in \text{supp}(X_1, \ldots, X_{d_b})$ *such that*

$$\left\{ \left(F_{X_1|Z}(x_1|z), \ldots, F_{X_{d_b}|Z}(x_{d_b}|z) \right) : z \in \text{supp}(Z|(X_1, \ldots, X_{d_b}) = x) \right\} \subseteq \mathcal{R}.$$

Proof of Theorem 2. From the proof of Theorem 1, $\beta(r) \equiv \mathbb{E}(B \mid R = r)$ is identified for all $r \in \mathcal{R}$. For notational convenience, let $\widetilde{X} \equiv (X_1, \ldots, X_{d_b})$. By iterated expectations, the definition of R, and Proposition 1, we have

$$\begin{aligned}
\mathbb{E}(B \mid \widetilde{X} = x) &= \mathbb{E}_{R|\widetilde{X}}[\mathbb{E}(B \mid \widetilde{X} = x, R) \mid \widetilde{X} = x] \\
&= \mathbb{E}_{Z|\widetilde{X}}[\mathbb{E}(B \mid \widetilde{X} = x, R = (F_{X_1|Z}(x_1|Z), \ldots, F_{X_{d_b}|Z}(x_{d_b}|Z))) \mid \widetilde{X} = x] \\
&= \mathbb{E}_{Z|\widetilde{X}}[\mathbb{E}(B \mid R = (F_{X_1|Z}(x_1|Z), \ldots, F_{X_{d_b}|Z}(x_{d_b}|Z))) \mid \widetilde{X} = x] \\
&= \mathbb{E}_{Z|\widetilde{X}}[\beta((F_{X_1|Z}(x_1|Z), \ldots, F_{X_{d_b}|Z}(x_{d_b}|Z))) \mid \widetilde{X} = x],
\end{aligned}$$

which is identified since $(F_{X_1|Z}(x_1|z), \ldots, F_{X_{d_b}|Z}(x_{d_b}|z)) \in \mathcal{R}$ for all $z \in \text{supp}(Z|\widetilde{X} = x)$. *Q.E.D.*

The support condition in Theorem 2 holds trivially if $\mathcal{R} = (0, 1)$. To interpret the support condition when $\mathcal{R}$ is a strict subset of $(0, 1)$, suppose for simplicity there is a single basic endogenous variable. Then the condition states that for every $z \in \text{supp}(Z)$ such that there is an r with $x = h[z, Q_V(r)]$, or equivalently $x = Q_{X|Z}(r|z)$, that r must be such that we can identify $\beta(r)$. That is, the value x must be obtainable via some r for which we can identify

$\beta(r)$. For example, in the simple model $Y = B_0 + B_1 X$, $\text{Var}[F_{X|Z}(x \mid Z)] > 0$ is sufficient for the support condition. This variance condition says there are at least two different instrument values z and z' which could have yielded x, and which correspond to different conditional ranks r and r'. That is, $x = h[z, Q_V(r)] = h[z', Q_V(r')]$. By strict monotonicity in the first stage equation, $h[z, Q_V(r)] \neq h[z, Q_V(r')]$ and $h[z', Q_V(r)] \neq h[z', Q_V(r')]$. Thus $h[z, Q_V(r)] \neq h[z', Q_V(r)]$ and $h[z, Q_V(r')] \neq h[z', Q_V(r')]$, and hence the relevance condition I5 holds so that $r, r' \in \mathcal{R}$. For example, see Figure 1 in which x_0 can be obtained via either $(z, r) = (1, 0.6)$ or $(z', r') = (0, 0.8)$. Both $r = 0.6$ and $r' = 0.8$ are points at which $\beta(r)$ is identified. The figure also shows why this condition is not necessary: consider an x value much larger than x_1. Such a value may be obtained only through $z = 1$, and yet the corresponding conditional rank may be a point at which we can identify $\beta(r)$.

The treatment on the treated parameter $\mathbb{E}(B \mid (X_1, \ldots, X_{d_b}) = x)$ provides one way of exploring heterogeneity in treatment effects. A truly constant treatment effect would yield a function $\mathbb{E}(B \mid (X_1, \ldots, X_{d_b}) = x)$ which is constant over x. An increasing function would show positive correlation between received treatment and the coefficients, while a decreasing function would show negative correlation between received treatment and the coefficients. Indeed, if $\mathcal{R} = (0, 1)^{d_b}$ then $\mathbb{E}(B \mid (X_1, \ldots, X_{d_b}) = x)$ is identified for all $x \in \text{supp}(X_1, \ldots, X_{d_b})$ and hence the correlations $\mathbb{E}[B_j X_l] = \mathbb{E}(\mathbb{E}[B_j \mid X_k = x_k, k \leq d_b] X_l)$ are identified for any j and any $l \leq d_b$.

3 Estimation

We construct estimators of $\beta(r)$ and $\beta_{\mathcal{R}}$ from an i.i.d. sample $\{(Y_i, X_i, Z_i)\}_{i=1}^n$ using the sample analog of the expressions in Theorem 1. We limit our focus to the case where there is one basic endogenous variable ($d_b = 1$), although there may be any number of known derived endogenous variables and exogenous variables Z. We discuss generalizations to $d_b > 1$ at the end of the section. To simplify notation, we let X denote the one basic endogenous variable in both this section and the next. As a first step towards approximating the event that $R = r$, we construct estimates $\widehat{R}_i$ of $R_i \equiv F_{X|Z}(X_i|Z_i)$ for $i = 1, \ldots, n$ as

$$\widehat{R}_i \equiv \widehat{F}_{X|Z}(X_i|Z_i), \tag{3}$$

where $\widehat{F}_{X|Z}(x|z)$ is an estimator of $F_{X|Z}(x|z)$. This step of our estimation procedure is similar to those of Imbens and Newey (2009) and Jun (2009), among others.

The asymptotic theory we develop in the next section is general enough to allow for

many different $\sqrt{n}$–consistent estimators $\widehat{F}_{X|Z}$. One could use a direct estimator such as the empirical conditional distribution function in the case that all Z variables are discrete. Alternatively, as pointed out by Chernozhukov, Fernández-Val, and Galichon (2010), one can estimate $Q_{X|Z}(s|z)$ at several quantiles s and then use the "pre-rearrangement" operator to construct an indirect estimator

$$\widehat{F}_{X|Z}(x|z) = \int_0^1 \mathbb{1}[\widehat{Q}_{X|Z}(s|z) \leq x]\, ds. \tag{4}$$

Chernozhukov, Fernández-Val, and Melly (2009, 2012) discuss several different parametric direct and indirect estimators.

For our purposes, we prefer nonparametric direct estimators (such as the empirical conditional distribution function) when the dimension of Z is small and discrete, and parametric indirect estimators when there are more than a few covariates. The latter are easier than direct estimators to link to primitives under I2, since, by strict monotonicity and independence, $Q_{X|Z}(r|z) = h(z, Q_V(r))$. For example, the linear quantile regression model of Koenker and Bassett (1978) implies that $h(z, Q_V(r)) = Q_{X|Z}(r|z) = z'\pi(r)$ for a function π that is strictly increasing in r. Substituting $F_V(v)$ for r, we have $h(z, v) = z'\pi(F_V(v))$, so that the linear quantile regression model imposes that h is linear with respect to z, while I2 links together the components of π to depend on a single underlying random variable V. For practical implementation when Z has more than just a few components, we advocate using linear quantile regression together with (4) to construct $\widehat{F}_{X|Z}$ and $\widehat{R}_i$. Besides being easy to interpret under I2, the linear quantile regression estimator has the additional benefits of being straightforward to compute, amenable to high-dimensional Z, and widely available in statistical packages. The integral in (4) can be evaluated using a uniform grid $\{s_j\}_{j=1}^J \subset (0, 1)$.

Having constructed $\widehat{R}_i$, we estimate $\beta(r)$ for a given r as

$$\widehat{\beta}(r) \equiv \left(\frac{1}{n}\sum_{i=1}^n \widehat{k}_i^h(r) W_i W_i'\right)^+ \left(\frac{1}{n}\sum_{i=1}^n \widehat{k}_i^h(r) W_i Y_i\right), \tag{5}$$

where $(\cdot)^+$ is the Moore-Penrose inverse and $\widehat{k}_i^h(r) \equiv h^{-1}K((\widehat{R}_i - r)/h)$ are weights constructed through a kernel function K with bandwidth parameter h that tends to 0 asymptotically. The Moore-Penrose inverse is useful here because the matrix in question may not be invertible for all values of r and h in small samples, although our assumptions in the next section will ensure invertibility asymptotically. Since R is always distributed uniformly with

support $[0,1]$ when $d_b = 1$, we can use our estimates of $\beta(r)$ to estimate $\beta_{\mathcal{R}}$ by

$$\widehat{\beta}_{\mathcal{R}} \equiv \lambda(\mathcal{R})^{-1} \int_{\mathcal{R}} \widehat{\beta}(r)\, dr, \tag{6}$$

where $\mathcal{R}$ is a measurable subset of $[0,1]$ that is specified by the analyst and λ is the Lebesgue measure.[1] As we show in Section 4, this estimator is $\sqrt{n}$–consistent and asymptotically normal for $\beta_{\mathcal{R}}$ under relatively weak regularity conditions. In studying a related problem for a different model, Hoderlein and Sherman (2013) described the strategy of an estimator like $\widehat{\beta}_{\mathcal{R}}$ as "localize-then-average." We find this terminology appealing as it captures the idea that for any given r, $\widehat{\beta}(r)$ only depends on the portion of the data local to the event $R = r$, while $\widehat{\beta}_{\mathcal{R}}$ forms an average of these various local estimators.

Overall, the computational complexity of (6) is very light for modern computing systems. A typical implementation would first estimate $\widehat{R}_i$, e.g. by using (4) with a moderate sized grid. Next, one would numerically integrate to compute (6). A simple and effective way to do this is to use variance-reducing pseudo-random draws, such as Halton sequences (see e.g. Section 9.3.3 of Train 2009) or a uniform grid. Typically, a few hundred draws should be more than sufficient. Moreover, unlike Monte Carlo integration, deterministic sequences can yield the same numerical results for all researchers. At each draw, one would estimate $\widehat{\beta}(r)$ using (5), which is essentially just a weighted linear regression. Finally, the draws are averaged together to obtain $\widehat{\beta}_{\mathcal{R}}$.

As we discuss in the next section, $\widehat{\beta}_{\mathcal{R}}$ is $\sqrt{n}$–consistent and asymptotically normal, but the asymptotic variance turns out to be complicated due to the effect of estimating R_i. Consequently, we use the nonparametric bootstrap to obtain standard errors. The typical procedure draws S sets of n observations with replacement from $\{(Y_i, X_i, Z_i)\}_{i=1}^n$, say $\{(Y_{si}, X_{si}, Z_{si})\}_{i=1}^n$ for $s = 1, \ldots, S$. These observations are used to compute $\widehat{\beta}_{\mathcal{R}}^s$ for $s = 1, \ldots, S$. Then

$$\widehat{\Sigma} \equiv \frac{1}{S-1} \sum_{s=1}^{S} (\widehat{\beta}_{\mathcal{R}}^s - \overline{\beta}_{\mathcal{R}})(\widehat{\beta}_{\mathcal{R}}^s - \overline{\beta}_{\mathcal{R}})'$$

with $\overline{\beta}_{\mathcal{R}} \equiv S^{-1} \sum_{s=1}^{S} \widehat{\beta}_{\mathcal{R}}^s$ forms a bootstrapped estimate of the variance of $\widehat{\beta}_{\mathcal{R}}$. This estimator can be used to construct confidence intervals or conduct hypothesis tests in the usual fashion. For example, a two-sided confidence interval of level α for the first component of $\beta_{\mathcal{R}}$ would

[1] Here and throughout the paper, the integration of vectors as in (6) should be understood as component-wise.

be given by

$$\left[\widehat{\beta}_{\mathcal{R},1} - \widehat{\Sigma}_{11}^{1/2}\Phi^{-1}(1-\alpha/2), \widehat{\beta}_{\mathcal{R},1} + \widehat{\Sigma}_{11}^{1/2}\Phi^{-1}(1-\alpha/2)\right],$$

where $\widehat{\beta}_{\mathcal{R},1}$ is the first component of $\widehat{\beta}_{\mathcal{R}}$, $\widehat{\Sigma}_{11}$ is the $(1,1)$ component of $\widehat{\Sigma}$, and Φ is the cumulative distribution function for the standard normal distribution.

Extending our estimator to the case where there are multiple basic endogenous variables ($d_b > 1$) requires a few modifications. First, we need to estimate $R_{ki} \equiv F_{X_k|Z}(X_{ki}|Z_i)$ for each $k = 1, \ldots, d_b$. This can be done the same way as in the $d_b = 1$ case. Second, $\widehat{\beta}(r)$ in (5) needs to be modified so that the kernel weights are multivariate. The curse of dimensionality would accompany this sort of multivariate smoothing, and while $\widehat{\beta}_{\mathcal{R}}$ could still be expected to be formally $\sqrt{n}$–convergent under certain conditions on the kernel function, K, its small sample behavior will likely be quite poor with realistic sample sizes if d_b is greater than 3 or 4. Third, when $d_b > 1$, the density of R is no longer known *a priori*, so that $\widehat{\beta}_{\mathcal{R}}$ could no longer be constructed by integrating as in (6). A natural solution to the latter problem is to use the empirical measure to approximate the integral by taking

$$\widehat{\beta}_{\mathcal{R}} = \frac{\sum_{i=1}^n \mathbb{1}[\widehat{R}_i \in \mathcal{R}]\widehat{\beta}(\widehat{R}_i)}{\sum_{i=1}^n \mathbb{1}[\widehat{R}_i \in \mathcal{R}]}.$$

The asymptotic analysis of this estimator involves third-order U-statistics and is much more complicated than that for (6). Given this complication and since the case $d_b = 1$ is by far the most commonly encountered in applications, we focus our formal analysis in the next section on $\widehat{\beta}_{\mathcal{R}}$ defined by (6).

4 Asymptotic Theory

In this section we discuss an asymptotic normality result for $\widehat{\beta}_{\mathcal{R}}$. The proof is in Appendix A. In the following, we let $P(r) \equiv \mathbb{E}[WW'|R = r]$ and use $\rightsquigarrow$ to denote convergence in distribution.

Theorem 3. *Under Assumptions I and E,*

$$\sqrt{n}(\widehat{\beta}_{\mathcal{R}} - \beta_{\mathcal{R}}) \rightsquigarrow N\left(0, \lambda(\mathcal{R})^{-2}\,\mathbb{E}[(\zeta_{1i} + \zeta_{2i})(\zeta_{1i} + \zeta_{2i})']\right),$$

where

$$\zeta_{1i} \equiv \mathbb{1}[R_i \in \mathcal{R}]P(R_i)^{-1}W_iW_i'(B_i - \beta(R_i))$$
$$\zeta_{2i} \equiv -\mathbb{E}[\mathbb{1}[R_j \in \mathcal{R}]\xi_i(X_j|Z_j)P(R_j)^{-1}W_jW_j'\dot{\beta}(R_j)|i] \qquad (j \neq i),$$

with all additional notation being defined below in Assumptions E.

Assumptions E.

E1. (Random sample) (Y_i, X_i, Z_i) *is an i.i.d. sample.*

E2. (Integration set) $\mathcal{R}$ *is a closed, measurable subset of* $[\delta, 1-\delta]$ *for some* $\delta > 0$.

E3. (Kernel) K *has support* $[-1, 1]$ *and is twice continuously differentiable and symmetric around* 0 *with* $\int_{-1}^{1} K(\eta)d\eta = 1$ *and* $\int_{-1}^{1} \eta^2 K(\eta)d\eta < \infty$.

E4. (Bandwidth) *As* $n \to \infty$, $\sqrt{n}h^2 \to 0$ *and* $\sqrt{n}h/\log(n) \to \infty$.

E5. (Smoothness) *Every component of* $P(r)$ *and* $\beta(r)$ *is twice continuously differentiable over* $r \in \mathcal{R}$ *with first and second component-wise derivatives* $\dot{P}(r), \ddot{P}(r), \dot{\beta}(r), \ddot{\beta}(r)$.

E6. (Existence of moments) $\mathbb{E}(\|WW'\|^4|R \in \mathcal{R})$ *and* $\mathbb{E}(\|B\|^4|R \in \mathcal{R})$ *are both finite.*

E7. (Rank estimation) $\widehat{R}_i$ *is constructed from* (3) *and for all* $(x, z) \in \mathcal{XZ}(\mathcal{R}) \equiv \{(x, z) : F_{X|Z}(x|z) \in \mathcal{R}\}$,

$$\sqrt{n}(\widehat{F}_{X|Z}(x|z) - F_{X|Z}(x|z)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \xi_i(x|z) + \rho_n(x|z) \tag{7}$$

with $\mathbb{E}[\xi_i(x|z)] = 0$, $\mathbb{E}[\mathbb{1}[R_i \in \mathcal{R}]\xi_j(X_i|Z_i)^4] < \infty$ *for both* $j = i$ *and* $j \neq i$, *and* $\sup_{(x,z)\in\mathcal{XZ}(\mathcal{R})} |\rho_n(x|z)| = o_{\mathbb{P}}(1)$. *Also, with probability approaching* 1, $\widehat{F}_{X|Z}$ *belongs to a class of functions* $\mathcal{F}$ *such that* $\log N(\epsilon, \mathcal{F}, \|\cdot\|_\infty) < C\epsilon^{-1/2}$ *for some* $C > 0$.[2]

The i.i.d. assumption E1 is standard for microeconometric applications and could in principle be extended to cover non-identical and/or dependent data frameworks. The assumption that $\mathcal{R}$ is closed and measurable in E2 is mild and of no practical significance. The additional restriction that $\mathcal{R}$ is a subset of $[\delta, 1-\delta]$ for some small $\delta > 0$ is made to avoid boundary issues. While these issues could potentially be addressed by using local linear weights, we

[2] The notation $N(\epsilon, \mathcal{F}, \|\cdot\|_\infty)$ stands for the ϵ–covering number of $\mathcal{F}$ under the sup–norm; that is, the minimal number of $\|\cdot\|_\infty$–balls of radius ϵ that are required to cover $\mathcal{F}$. Intuitively, the covering number is a measure of the complexity of the class of functions $\mathcal{F}$. See van der Vaart and Wellner (1996) for more details.

have found that these work poorly in practice. This is perhaps not surprising in our framework since the density of R is uniform, which is a particularly unfavorable case for local linear regression (see Remark 4 of Ruppert and Wand 1994). Additionally, as we discuss further below, E7 will in practice also require that $\mathcal{R}$ does not contain extremal ranks. We therefore see E2 as a natural restriction given our identification strategy, although it does imply that we can only estimate a trimmed version of $\mathbb{E}(B)$. It is likely possible to adjust our estimator to allow for $\delta \to 0$ asymptotically, or to use a different smoothing approach that is less sensitive to boundary effects, such as sieves. We leave these modifications for future research.

The restrictions on the kernel in E3 are relatively mild and allow for a broad range of commonly used kernels, such as the uniform or biweight kernel. We rule out kernels with unbounded support such as the Gaussian kernel in order to apply results in the literature on kernel regression with generated regressors (specifically, those in Mammen, Rothe, and Schienle 2012). Our bandwidth conditions in E4 prescribe a choice of h that undersmooths (goes to zero faster) relative to the usual optimal bandwidth choice for nonparametric kernel regression. This is standard given the semiparametric nature of our estimator $\widehat{\beta}_{\mathcal{R}}$ and appears in similar contexts like the average derivative estimator of Powell, Stock, and Stoker (1989). Intuitively, while $\widehat{\beta}(r)$ only uses a portion of the data, $\widehat{\beta}_{\mathcal{R}}$ uses all the data and thus has a much smaller variance. Consequently, the bandwidth h can be sent to 0 more quickly in order to remove the bias of $\widehat{\beta}_{\mathcal{R}}$ at a $\sqrt{n}$–rate and achieve the overall $\sqrt{n}$–rate of convergence asserted in Theorem 3.

Assumption E5 places some standard smoothness conditions on the population objects $P(r)$ and $\beta(r)$. In combination with I5 and E2, these imply that $P(r)$ is invertible uniformly over $\mathcal{R}$, and so serves to strengthen I5 in a way that is theoretically important for the asymptotics. The practical implication is that $\mathcal{R}$ should not include neighborhoods of isolated points where I5 fails, such as where the curves cross in Figure 1 ($r = .3$). Assumption E6 is a standard type of assumption regarding the number of existing moments for W and B. Since WW' contains squared terms, E6 essentially requires each component of W to have a finite eighth moment.

The conditions in E7 require the estimator of $F_{X|Z}$ used in constructing $\widehat{R}_i$ to be asymptotically linear and $\sqrt{n}$–convergent. This assumption is not very restrictive for parametric models. Chernozhukov et al. (2009, 2012) provide several examples of direct conditional distribution function estimators that satisfy this condition. In addition, Chernozhukov et al. (2010) show that (4), viewed as a functional mapping from conditional quantile to condi-

tional distribution functions, is Hadamard differentiable. As a result, asymptotically linear representations for conditional quantile estimators give rise to asymptotically linear representations for conditional distribution estimators defined by (4) after applying the functional delta method. The results from the vast literature on quantile regression can therefore be transferred fairly easily to conditional distribution estimators defined by (4).

The more restrictive part of E7 is the requirement that the estimation error ρ_n be convergent uniformly over x and z such that $F_{X|Z}(x|z) \in \mathcal{R}$. For some estimators of $F_{X|Z}$, such as the conditional empirical distribution function, this condition does not present a problem. For our preferred estimator that uses (4) and a linear quantile regression estimator of $Q_{X|Z}$, it is well-known that this condition will generally not hold for subsets $\mathcal{R}$ that include extremal points in $[0, 1]$ unless strong restrictions are placed on the tail behavior of X. However, since we rule out extremal ranks in E2, this does not represent a substantive additional restriction in our setting. The following result, which is Theorem 3 in Chernozhukov, Fernández-Val, and Kowalski (2011), provides sufficient conditions for E7 for our preferred estimator of $\widehat{R}_i$.

Proposition 2. *Suppose that $\widehat{F}_{X|Z}$ is estimated using (4) with $\widehat{Q}_{X|Z}$ taken as the linear quantile regression estimator of Koenker and Bassett (1978). Then E7 holds under Assumptions QR with*

$$\begin{aligned}\xi_i(x|z) = f_{X|Z}(x|z)z' \, \mathbb{E}\left[f_{X|Z}(Z'\pi_0(F_{X|Z}(x|z))|Z)ZZ'\right]^{-1} \\ \times \left(F_{X|Z}(x|z) - \mathbb{1}\left[X_i \leq \pi_0(F_{X|Z}(x|z))\right]\right) Z_i.\end{aligned}$$

Assumptions QR.

QR1. (Well-specified) *$Q_{X|Z}(r|z) = z'\pi_0(r)$ for all $r \in \mathcal{R}$ and $z \in \text{supp}(Z)$.*

QR2. (Smooth quantile function) *$Q_{X|Z}(r|z)$ is three times continuously differentiable in r over $\mathcal{R}$ with a uniformly bounded third derivative.*

QR3. (Well-behaved density) *$f_{X|Z}(x|z)$ is uniformly continuous, uniformly bounded and uniformly bounded away from 0 over $(x, z) \in \mathcal{XZ}(\mathcal{R})$.*

QR4. (Existence of moments) *$\mathbb{E}(\|Z\|^8) < \infty$.*

QR5. (No multicollinearity) *$\mathbb{E}(ZZ')$ is invertible.*

The asymptotic variance of $\widehat{\beta}_{\mathcal{R}}$ given in Theorem 3 depends on two components. If R_i were known and did not need to be estimated by $\widehat{R}_i$, the asymptotic variance would only

depend on ζ_{1i} and would be given by $\lambda(\mathcal{R})^{-2}\,\mathbb{E}[\zeta_{1i}\zeta_{1i}']$. To interpret this quantity, rewrite $Y_i = W_i'B_i$ as $Y_i = W_i'\beta(r) + U_i(r)$, where $U_i(r) \equiv W_i'(B_i - \beta(r))$ satisfies $\mathbb{E}[U_i(r)|R_i = r] = 0$ by Proposition 1. Suppose that we were to regress Y on X in a large sample drawn from the subpopulation $R = r$. (Of course, even if we knew R_i *a priori*, this wouldn't be feasible since the event $R = r$ has measure zero.) Then the asymptotic variance of the coefficient vector would be given by the usual sandwich form, $P(r)^{-1}\,\mathbb{E}[U_i(r)^2 W_i W_i'|R_i = r]P(r)^{-1}$, where all of the typical components have been conditioned on $R = r$. This sandwich form is exactly what appears in

$$\lambda(\mathcal{R})^{-2}\,\mathbb{E}[\zeta_{1i}\zeta_{1i}'] = \mathbb{E}\left[P(R_i)^{-1}\,\mathbb{E}[U_i(R_i)^2 W_i W_i'|R_i]P(R_i)^{-1}|R_i \in \mathcal{R}\right]\lambda(\mathcal{R})^{-1},$$

except that it is now being integrated over all $r \in \mathcal{R}$ under consideration and scaled to account for the size of $\mathcal{R}$.

The second component of the asymptotic variance expression, ζ_{2i}, accounts for the effect of estimating $\widehat{R}_i$. This term involves the influence function from the first stage, ξ_i, and so will depend on the estimator of $F_{X|Z}$ that is used. It appears to generally have a complicated form, and at least for our preferred rank estimator discussed in Proposition 2, we have not found that the form of ξ_i provides any useful simplification in the expression for ζ_{2i}. Note also that ζ_{2i} depends multiplicatively on the first derivative of $\beta(r)$. Hence, in the case of no treatment effect heterogeneity, ζ_{2i} is identically zero and the asymptotic variance of $\widehat{\beta}_{\mathcal{R}}$ is determined exclusively by ζ_{1i}.

Constructing a direct estimator of the asymptotic variance of $\widehat{\beta}_{\mathcal{R}}$ would be tedious and difficult, likely requiring an additional estimator of $\dot{\beta}(r)$ as a function of r. Instead, we propose bootstrapping to approximate the limiting distribution of $\widehat{\beta}_{\mathcal{R}}$. The procedure for constructing bootstrapped standard errors and confidence intervals was outlined in Section 3. This type of bootstrap procedure is generally consistent, and our framework does not possess any of the usual causes of inconsistency that have been studied in the literature. We therefore anticipate that the bootstrap is consistent, although we have not attempted a formal proof.

5 Monte Carlo Simulations

This section contains the results of Monte Carlo simulations on the finite-sample behavior of $\widehat{\beta}_{\mathcal{R}}$. We consider a data generating process with an outcome equation specified as

$$Y = B_0 + B_1 X,$$

and with a first stage equation given by

$$X = \pi Z + \gamma ZV + V.$$

We draw V independently from a normal distribution with mean 0.1 and standard deviation 0.4. The random coefficients in the outcome equation are then generated as $B_j = \rho_j V + \epsilon_j$ for $j = 0, 1$ with ϵ_j distributed $N(\mu_j, \sigma_j^2)$ independently of all other variables. In particular, we take $\rho_0 = .3$, $\mu_0 = .2$, $\sigma_0 = .2$ and $\rho_1 = .7, \mu_1 = .45$, $\sigma_1 = 1$, which implies that $\mathbb{E}(B_0) = .23$ and $\mathbb{E}(B_1) = .52$. Since both $\rho_j \neq 0$, there is a strong endogeneity problem in this data generating process in the sense that B and X are highly correlated through their mutual dependence on V. As a consequence, the ordinary least squares (OLS) estimator will be inconsistent for $\mathbb{E}(B)$.

In the first stage equation we set $\pi = .2$ and consider the cases where $\gamma = 0$ and $\gamma = .4$. In the first case, the effect of Z on X is homogenous, so the results of Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2008) imply that the TSLS estimator will be consistent for $\mathbb{E}(B_1)$, although it is still generally inconsistent for $\mathbb{E}(B_0)$. In the second case, the effect of Z on X varies with V, so that TSLS will generally be inconsistent for both $\mathbb{E}(B_0)$ and $\mathbb{E}(B_1)$. In contrast, $\widehat{\beta}_{\mathcal{R}}$ will be consistent for both components of $\mathbb{E}(B)$ for either value of γ. The instrument Z is a binary random variable that takes values $\{0, 1\}$ with equal probability and is drawn independently from $(V, \epsilon_0, \epsilon_1)$. We used a conditional empirical distribution function to estimate $\widehat{R}_i$, a biweight kernel for K and specified $\mathcal{R} = [0, 1]$. Although this choice of $\mathcal{R}$ does not satisfy E2, we have found the results of these simulations to be insensitive to different values of δ. The number of replications in all simulations presented is 1000 and the integrals in the definition of $\widehat{\beta}_{\mathcal{R}}$ were evaluated using 300 Halton draws.

Table 1 reports the performance of the first and second components of $\widehat{\beta}_{\mathcal{R}}$ as estimators of $\mathbb{E}(B_0)$ and $\mathbb{E}(B_1)$ relative to both the OLS and TSLS estimators in the case without first stage heterogeneity, $\gamma = 0$. As expected, the OLS estimator is inconsistent for both parameters. The results support the prediction of Heckman and Vytlacil (1998) and Wooldridge

(1997, 2003, 2008) that TSLS is consistent for $\mathbb{E}(B_1)$ but inconsistent for $\mathbb{E}(B_0)$. The performance of $\widehat{\beta}_{\mathcal{R}}$ is reported for a variety of bandwidth choices h. Our prediction of the consistency of both components of $\widehat{\beta}_{\mathcal{R}}$ at the $\sqrt{n}$ rate is supported by the decrease in mean squared error (mse) that occurs when increasing N from 500 to 1000. Most remarkable is the performance of the second component of $\widehat{\beta}_{\mathcal{R}}$ as an estimator of $\mathbb{E}(B_1)$ relative to TSLS. Our results suggest a mean-squared error that is actually slightly lower than TSLS across a broad range of bandwidth values. For smaller bandwidth values, both the bias and standard deviation (std) are comparable to TSLS, perhaps even being a bit smaller.

Table 2 reports the same type of results as Table 1 for the case with first stage heterogeneity, $\gamma = .4$. Here, we see that the heterogeneity in the effect of Z on X leads to severe inconsistency for the TSLS estimator. On the other hand, $\widehat{\beta}_{\mathcal{R}}$ remains consistent and performs similarly to the case where $\gamma = 0$. We interpret these results as promising evidence in support of the practical applicability of our estimator to situations where heterogeneity in the first stage cannot be ruled out.

6 Air Pollution and House Prices

In this section, we apply our results to analyze the relationship between air pollution and house prices. A large literature on hedonic methods uses relationships like this to infer the value of non-market amenities, such as clean air (e.g. Rosen 1974, Smith and Huang 1995, Ekeland, Heckman, and Nesheim 2004, Palmquist 2005, Heckman, Matzkin, and Nesheim 2010). Reliable measurements of these valuations are important for quantifying the economic benefit of air quality regulation. We follow the empirical approach of Chay and Greenstone (2005). They argue that previous analysis based on cross-sectional OLS or first-differences yields small, zero, or perverse-signed effects due to omitted variables, such as unobserved economic shocks, or sorting of households based on unobserved preferences for clean air. To remedy this, they use regulation introduced by the 1970 Clean Air Act Amendments to define a binary instrument for change in total suspended particulates (TSP) from 1970 to 1980 and then use TSLS to estimate the effect of TSP changes on county-level house price changes.

Our analysis builds on Chay and Greenstone in several directions. As they note (page 393), a correlated random coefficients model is appropriate due to sorting of households based on unobserved preferences for clean air (we also discuss this below). We demonstrate substantial first stage heterogeneity in the effect of the instrument, which strongly suggests

that the simpler estimators discussed by Heckman and Vytlacil (1998) and Wooldridge (1997, 2003, 2008) would be inconsistent for the APE. Likewise, the binary instrument precludes approaches which rely on continuous variation, such as Florens et al. (2008). For two subsets of counties where the instrument has a statistically significant effect on pollution levels, we estimate unweighted average partial effects of changes in pollution on changes in house prices. These estimates demonstrate patterns that are consistent with household sorting. Taken together, these estimates along with TSLS suggest there is substantial heterogeneity in households' value of clean air.

6.1 The dataset and institutional background

The 1970 Clean Air Act Amendments set national ambient air quality standards (NAAQS) for TSPs with the goal that all counties would eventually meet these standards. The law requires the U.S. Environmental Protection Agency (EPA) to annually designate each county as either attainment, if the county meets the standard, or as nonattaintment if it does not. Firms in nonattainment counties are subject to much stricter pollution regulations than firms in attainment counties. Consequently, nonattainment status should affect counties' pollution levels. Chay and Greenstone argue (pages 395–406) that a county's nonattainment status in 1975 and 1976 is plausibly independent of unobserved variables which change between 1970 and 1980 and affect housing prices, such as unobserved economic shocks, as well as unobserved changes in clean air preferences from 1970 to 1980 due to sorting. In addition, there is no reason to expect that households care about nonattainment status above and beyond its effect on pollution; i.e., the exclusion restriction holds. For these reasons, Chay and Greenstone conclude that mid-decade nonattaintment status is a valid instrument for identifying causal effects of changes in TSP from 1970 to 1980 on changes in house prices from 1970 to 1980. We take the instrument definition and validity arguments as given and investigate the implications of allowing for first stage heterogeneity via our CRC estimator.

Our dataset is essentially identical to that of Chay and Greenstone (2005), as described in their data appendix (pages 419–421). We obtain house price data as well as covariates from the 1972 and 1983 County and City Data Books (obtained via ICPSR). This price and covariate data is only available at the county level and hence the units of analysis are counties. TSP pollution data may be downloaded from the EPA. One minor difference between our dataset and Chay and Greenstone's is that we do not have TSP data from 1969; this data is not available for download and the EPA has not responded to our requests. Chay and Greenstone define TSP levels for 1970 as the average of TSP levels for 1969–

1972. Since we are missing 1969, we average over just 1970–1972. To define the 1980 TSP level, we average over 1977–1980 levels, as in Chay and Greenstone. The annual TSP levels are derived by aggregating observations throughout the year at different pollution monitors located across the country, as in Chay and Greenstone (page 384). Also as in their paper (page 391), we use the TSP data from 1974 and 1975 to define the instrument as the binary indicator variable for mid-decade nonattaintment status, since data on the actual EPA designated nonattaintment status does not exist. A second difference between their paper and our analysis arises here. Of our 989 observations, our definition of the instrument yielded 300 nonattaintment counties, whereas Chay and Greenstone have only 280 nonattaintment counties out of 988 observations.[3] This difference may explain why our TSLS results in table 4 differ somewhat from theirs.

Table 3 shows summary statistics along with a list of all covariates included in the analysis (see pages 420–421 of Chay and Greenstone for further explanation of the covariates). This table is comparable to Chay and Greenstone's table 1. All prices are adjusted to 1982–1984 dollars. Mean house prices increased from around \$40,000 to around \$53,000 while TSP levels fell by about 9 μg/m^3. The goal of the instrumental variable analysis is to determine to what extent this correlation between the rise in house prices and fall in pollution levels reflects causal effects.

6.2 Empirical results

Let X denote the change in TSP between 1970 and 1980, Z_2 denote mid-decade nonattainment status, and Z_1 denote the vector of 22 covariates. As discussed in Section 3, we begin by estimating a linear quantile regression of the treatment variable X on the instrument and the covariates for several different quantiles (similar results obtain when the covariates are omitted). Figure 2 plots the coefficient on the instrument Z_2 against the quantile used. Recall that the first stage assumption I2 implies that countries with small conditional ranks generally have smaller values of X—that is, larger drops in pollution—than counties with

[3] This difference may arise due to an ambiguity in determining whether a county violates the "bad day rule", which says that the second largest daily TSP value within a year must not exceed 260 μg/m^3 and would place a county in nonattainment. For counties with multiple monitors, there are at least two approaches: (1) compute the second highest daily TSP value for each monitor, and say a county violates the rule if any monitor within the county violates that rule, and (2) compute a county-level daily reading by averaging all monitors for a given day, and then compute the second highest daily TSP from that averaging. Our reading of the EPA regulations suggest that (1) is the approach EPA used and hence is what we use as well. Approach (2) leads to far fewer counties being designated as nonattaintment—222 out of 989. Hence this approach cannot be what Chay and Greenstone used either. We are unsure what they used, and neither author has responded to our requests for clarification.

larger conditional ranks. A rth-quantile regression tells us the effect of the instrument for counties with conditional rank $R = r$. For example, the median quantile regression tells us the effect of the instrument for counties generally at the middle of the distribution of changes in TSP. For these counties, the coefficient is around -0.1, which suggests that being in a nonattaintment county caused pollution to drop by -10 μg/m^3 relative to attainment counties, all else equal. Recall from table 3 that the average TSP level in 1970 was 65.5 μg/m^3, and it fell by around 9 μg/m^3. So a -10 μg/m^3 effect is quite large. The effect of -20 μg/m^3 for counties with the smallest conditional ranks is even larger. Note that the coefficient we find at the median, about -10 μg/m^3, is essentially equal to the coefficient obtained from a linear mean regression, as in Chay and Greenstone's table 4 panel A column 2.

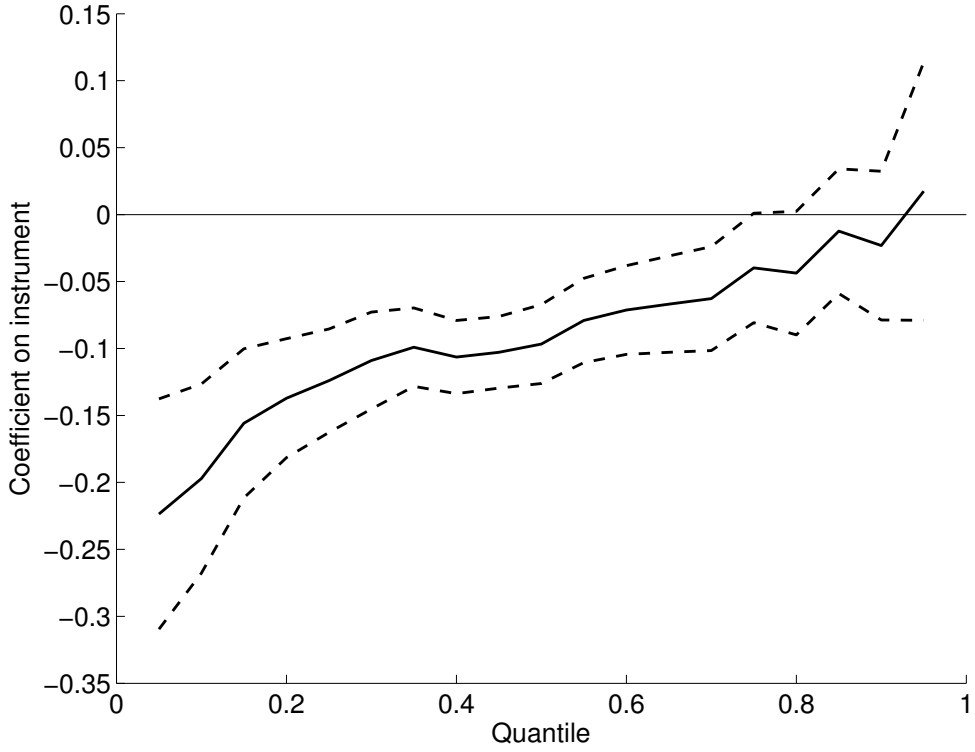

Figure 2: Plot of estimation results from several linear quantile regressions of treatment X (change in TSP, where TSP here is in units of 1×10^{-4} grams/m^3 rather than 1×10^{-6} grams/m^3 = μg/m^3) on the instrument Z_2 (nonattainment status) and controls. The solid line plots the estimated coefficient on the instrument on the vertical axis against the corresponding quantile on the horizontal axis, from .05 to .95 in .05 increments. The dotted lines plot simultaneous confidence intervals for each of these quantiles.

The plot shows a heterogeneous effect of the instrument on treatment—the instrument

has a strong negative effect at low quantiles, but this effect decreases towards zero for higher quantiles. For quantiles from about 0.75 to 1 the instrument does not have a statistically significant effect on treatment. Even though the standard F-statistic for the instrument suggests that there is no weak instrument problem ($F = 25$), figure 2 shows that the instrument is not uniformly strong for all counties, and indeed there are many counties (about 25% of them) where the instrument appears to have no effect at all.

Counties with smaller conditional ranks have the smallest values of X—that is, the largest drops in pollution over the decade. Large drops in pollution are strongly negatively correlated with having a high baseline level of pollution in 1970 ($\rho = -0.78$). Consequently, the heterogeneous effect of the instrument is to be expected. Counties in nonattainment are more heavily regulated than counties in attainment. But being regulated only matters if a county has a pollution problem to begin with. Hence the counties with the highest baseline pollution are also the ones where the instrument has a strongest effect. Conversely, the counties with the lowest baseline pollution (and hence lowest potential drops in pollution) are the ones where the instrument has essentially a zero effect.

Next we implement our generalized CRC estimator. We choose two different sets of conditional ranks to consider: $\mathcal{R} = [0.1, 0.4]$ and $\mathcal{R} = [0.4, 0.7]$. First, we omit estimating average partial effects near the tails as discussed in Section 4. Second, we omit estimating average partial effects near the region where the instrument is weak or irrelevant, as discussed in Sections 2 and 4. We split up the region $[0.1, 0.7]$ for which we can estimate average partial effects into two pieces in order to examine potential heterogeneity in the effect of pollution on house prices. For each choice of $\mathcal{R}$ we use the following tuning parameters: 1999 points for the first stage grid (equation 4), which corresponds to 1999 linear quantile regressions (an equally spaced grid with step size 0.0005). 2000 Halton draws for integration of $\widehat{\beta}(r)$ over $r \in \mathcal{R}$ (equation 6). We use 500 bootstrap draws to compute 95% confidence intervals. Finally, we present a range of bandwidths from $h = 0.04$ to $h = 0.085$. In our Monte Carlo simulations, the bandwidth $h = 0.07$ minimized MSE for the sample size $N = 1000$ and when there was first stage heterogeneity (Table 2).

Table 4 shows the main results. There are four columns. Columns (1) and (2) show estimation results without any control variables while columns (3) and (4) show estimation results with the control variables. Chay and Greenstone present additional specifications which use a "flexible functional form", but they do not specify what precisely they mean, and they have not responded to our requests for clarification. Hence we present only their first two specifications. Columns (1) and (3) show the generalized CRC estimates for the

choice $\mathcal{R} = [0.1, 0.4]$ while columns (2) and (4) show estimates for $\mathcal{R} = [0.4, 0.7]$.

First, notice that OLS—which is a first-differences regression here—has the perverse sign (implying prices go up when pollution goes up), and is even statistically significant without covariates, as also found by Chay and Greenstone (their Table 3 panel C columns 1 and 2). The TSLS results also mirror the main findings of Chay and Greenstone (their Table 5 panel A columns 1 and 2): Without covariates, we find a large negative effect of pollution on log house prices: a 1 μg/m^3 reduction in mean TSP causes a 0.42 percent increase in property values. With covariates, this effect size is cut in half and becomes marginally statistically insignificant. Chay and Greenstone's corresponding TSLS point estimate is -0.213 with a 95 percent confidence interval of $[-0.4, -0.025]$, which is not too different from the confidence interval we obtain (the difference is likely due to the data discrepancies previously mentioned).

Second, consider the bootstrap confidence intervals for the generalized CRC estimates. None of them are statistically significant. We have already seen that the instrument is strongest for smaller conditional ranks. Consequently, the confidence intervals are nearly twice as wide for the 'weaker' instrument region $\mathcal{R} = [0.4, 0.7]$ than for the 'stronger' instrument region $\mathcal{R} = [0.1, 0.4]$. For the region $\mathcal{R} = [0.1, 0.4]$, the confidence intervals are roughly the same length as for TSLS for column (3), and they're actually smaller for column (1). But the point estimates for both of these columns are close to zero, as expected under the sorting story discussed below. Consequently, since the TSLS confidence interval was already quite close to zero, the generalized CRC estimate confidence intervals here have just been shifted over to be centered near zero.

Next consider the generalized CRC estimates. Begin by considering the point estimates. The point estimates are fairly insensitive to changes in bandwidths, especially after controlling for covariates. When controlling for covariates, the estimates for $\mathcal{R} = [0.4, 0.7]$ are roughly twice as large as those for $\mathcal{R} = [0.1, 0.4]$. A similar finding is true when comparing column (2) to column (1), although the difference in magnitudes is much larger without the controls. Moreover, all the CRC estimates are smaller than the TSLS estimates. These findings are consistent with the possibility that households sort according to their unobserved preference for clean air during the baseline year of 1970. In that case, households with the strongest taste for clean air will move to counties with low baseline pollution, while households who do not care about clean air will move to counties with high baseline pollution. As we saw earlier, the baseline pollution and conditional rank are strongly correlated. Hence counties where the instrument has the strongest effect are also counties where most

households do not care about clean air, and so we find close to zero effects for $\mathcal{R} = [0.1, 0.4]$. For counties with households that moderately care about clean air, $\mathcal{R} = [0.4, 0.7]$, we find moderate sized effects. For counties with households that care strongly about clean air, $\mathcal{R} = [0.7, 0.1]$, we are unable to identify their preferences, since those are precisely the counties where the instrument has little or no effect—because those are the counties with little pollution to begin with. The finding that TSLS is larger than all the CRC estimates suggests that the effects for counties with $\mathcal{R} = [0.7, 1]$ are larger than -0.15, the effect we found for $\mathcal{R} = [0.4, 0.7]$. This is because TSLS is a weighted average effect, where the weights depend on the strength of the instrument. Although counties in $\mathcal{R} = [0.7, 1]$ will receive close to zero weight in forming the TSLS estimand, if their actual effect size is large enough it can still pull up the overall estimate.

In this application, we showed that the instrument has a naturally interpreted heterogeneous effect due to differences in baseline pollution levels. We showed that the data is consistent with sorting, and that comparing our CRC estimates with TSLS allows us to draw conclusions about effects for counties where the instrument is weak. Overall, our findings suggest that there is substantial heterogeneity in the size of the effect of pollution on housing prices.

7 Conclusion

In this paper we have studied a linear correlated random coefficients model. We provided conditions under which we can point identify the average partial and treatment effects of an endogenous treatment variable by using variation in an instrumental variable. In contrast to previous research, these conditions allow for heterogeneous effects of the instrument in the first stage equation, as well as binary or discrete instruments in many cases. Our identification argument led directly to a simple estimator of a trimmed average of the outcome coefficients. This estimator is just an average of weighted least squares regressions, where the weights depend on a first stage estimator. We established $\sqrt{n}$-consistency and asymptotic normality of this estimator, and showed that it performs well in finite sample simulations. We have illustrated how allowing for and analyzing heterogeneity in the first stage and in the outcome equation can be easily and fruitfully incorporated into a typical applied instrumental variables analysis.

Several issues remain for future research. First, it may be theoretically interesting to modify our estimator to better account for boundary effects both due to kernel smoothing

and first stage rank estimation. Second, we have not provided a method for choosing the bandwidth h, which is an important question in practice. Third, we assumed the set $\mathcal{R}$ for which the relevance condition I5 holds is known *a priori*. In some applications, this is a reasonable assumption, such as in our empirical application. In other applications, it may not be reasonable. In principle, this set can be estimated in a preliminary step, and then the previous analysis can be repeated by using this estimated set $\widehat{\mathcal{R}}$ in place of $\mathcal{R}$. This extension is both nontrivial and of independent interest, and we leave it to future work. Fourth, it may be helpful to explore modifications of our proposed estimator to achieve efficiency gains. Finally, we are coding a Stata module that will enable practitioners to apply the estimator in this paper with minimal investment.

A Proofs

Proof of Proposition 1. For any $r \equiv (r_1, \ldots, r_{d_b}) \in \operatorname{int}\operatorname{supp}(R)$, $z \in \operatorname{supp}(Z)$ and $b \in \mathbb{R}^{d_w}$ we have

$$\begin{aligned}
\mathbb{P}\left[R_k \leq r_k \, \forall k, B \leq b | Z = z\right] &= \mathbb{P}\left[X_k \leq Q_{X_k|Z}(r_k|z) \, k = 1, \ldots, d_b, B \leq b | Z = z\right] \\
&= \mathbb{P}\left[h_k(z, V_k) \leq h_k(z, Q_{V_k|Z}(r_k|z)) \, \forall k, B \leq b | Z = z\right] \\
&= \mathbb{P}\left[V_k \leq Q_{V_k|Z}(r_k|z) \, \forall k, B \leq b | Z = z\right] \\
&= \mathbb{P}\left[V_k \leq Q_{V_k}(r_k) \, \forall k, B \leq b\right],
\end{aligned}$$

where the first equality follows because for $k = 1, \ldots, d_b$, $X_k | Z = z$ is continuous by I2 and $r_k \in (0, 1)$, the second follows by I2 and the equivariance to monotone transformations property of quantiles, the third follows because $h_k(z, \cdot)$ is strictly increasing by I2, and the fourth uses I4. Since the right-hand side does not depend on z, we conclude that $(R, B) \perp\!\!\!\perp Z$.

The second statement follows because $R = r$ if and only if $X_k = Q_{X_k|Z}(r_k|Z)$ for $k = 1, \ldots, d_b$. Hence, conditional on $R = r$, X_k is a stochastic function of Z for $k = 1, \ldots, d_b$. Since, by I3, the derived endogenous variables, X_k, $k > d_b$, are functions of $X_1, \ldots, X_{d_b}$ and Z_1, they are also stochastic functions of Z alone after conditioning on $R = r$. Thus, conditional on $R = r$, $W \equiv [1, X', Z_1']'$ is only stochastic through Z. Since $Z \perp\!\!\!\perp B | R$ (as implied by $(R, B) \perp\!\!\!\perp Z$), this shows that $W \perp\!\!\!\perp B | R$ as well. *Q.E.D.*

Proof of Theorem 3. We begin by rewriting the model as $Y_i \equiv W_i' \beta(r) + U_i(r)$, where $U_i(r) \equiv W_i'(B_i - \beta(r))$. Substituting into (5), we have $\widehat{\beta}(r) = \widehat{P}(r)^+ \widehat{P}(r) \beta(r) + \widehat{P}(r)^+ \widehat{A}(r)$

with $\widehat{P}(r) \equiv n^{-1}\sum_{i=1}^n \widehat{k}_i^h(r) W_i W_i'$ and $\widehat{A}(r) \equiv n^{-1}\sum_{i=1}^n \widehat{k}_i^h(r) W_i U_i(r)$. For any square matrix M, let $\sigma(M)$ denote the absolute value of the smallest eigenvalue of M. Then I5 and E5 imply that $\inf_{r\in\mathcal{R}} \sigma(P(r)) = C$ for some $C > 0$. Defining $\widehat{1} \equiv \mathbb{1}[\inf_{r\in\mathcal{R}} \sigma(\widehat{P}(r)) > C/2]$, we have $\widehat{1}(\widehat{\beta}(r) - \beta(r)) = \widehat{1}\widehat{P}(r)^+\widehat{A}(r)$, since $\widehat{1}(\widehat{P}(r)^+\widehat{P}(r) - I) = 0$ for all $r \in \mathcal{R}$.[4] Centering, integrating over $\mathcal{R}$ and scaling by $\sqrt{n}$, we obtain by Lemma A.1 that

$$\widehat{1}\sqrt{n}\left(\int_{\mathcal{R}} \widehat{\beta}(r)\,dr - \int_{\mathcal{R}} \beta(r)\,dr\right) = \widehat{1}\sqrt{n}\int_{\mathcal{R}} P(r)^{-1}\widehat{A}(r)\,dr + o_{\mathbb{P}}(1).$$

Next, define $T_i(r) \equiv P(r)^{-1}W_iU_i(r)$ and note that our assumptions only ensure that $T_i(r)$ is defined for $r \in \mathcal{R}$. For $r \notin \mathcal{R}$ we use the convention that $\mathbb{1}[r \in \mathcal{R}]T_i(r) = 0 \cdot T_i(r) = 0$, which will help to ease notation.[5] Then

$$\begin{aligned}
\int_{\mathcal{R}} P(r)^{-1}\sqrt{n}\widehat{A}(r)\,dr &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int \frac{1}{h}K\left(\frac{r-\widehat{R}_i}{h}\right)\mathbb{1}[r\in\mathcal{R}]T_i(r)\,dr \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^n \int K(\eta)\mathbb{1}[\widehat{R}_i + h\eta \in \mathcal{R}]T_i(\widehat{R}_i + h\eta)\,d\eta \\
&\overset{\text{a.s.}}{=} \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{1}[R_i \notin \mathrm{bd}(\mathcal{R})]\int K(\eta)\mathbb{1}[\widehat{R}_i + h\eta \in \mathcal{R}]T_i(\widehat{R}_i + h\eta)\,d\eta, \quad (8)
\end{aligned}$$

where the second equality follows by changing the variable of integration from r to $\eta \equiv (r - \widehat{R}_i)/h$ and the third equality follows with $\mathrm{bd}(\mathcal{R})$ denoting the boundary of $\mathcal{R}$ and $\overset{\text{a.s.}}{=}$ denoting almost-sure equality because $\mathbb{P}[R_i \in \mathrm{bd}(\mathcal{R}),\ \text{any } i] = 0$.[6] Note that every component of the vector $\mathbb{1}[r \in \mathcal{R}]T_i(r)$ is twice continuously differentiable at all $r \notin \mathrm{bd}(\mathcal{R})$, since I5 and E5 imply that $T_i(r)$ is twice continuously differentiable at all $r \in \mathcal{R}^\circ$ (the interior of $\mathcal{R}$) with first and second component-wise derivatives $\dot{T}_i(r)$ and $\ddot{T}_i(r)$, while $\mathbb{1}[r \in \mathcal{R}]T_i(r) = 0$ for $r \notin \mathcal{R}$.[7] For $R_i \in \mathcal{R}^\circ$, a second order element-by-element application of Young's form of

[4] That $\widehat{1}$ is indeed a random variable (in the sense of being a measurable function on the underlying sample space) follows from Theorem 18.19 of Aliprantis and Border (2006) combined with standard results.

[5] The alternative would be to define $T_i(r)$ and related functions (its derivatives, etc.) case by case, which seems unnecessarily tedious.

[6] Since almost-sure equality is sufficient for determining limiting distributions, we will drop the "a.s." qualifier in the following.

[7] The differentiability of $T_i(r)$ for $r \in \mathcal{R}^\circ$ can be determined using the calculus rules for matrices of functions derived in Section 6.5 of Horn and Johnson (1991).

Taylor's Theorem yields

$$\begin{aligned}
&\mathbb{1}[\widehat{R}_i + h\eta \in \mathcal{R}]T_i(\widehat{R}_i + h\eta) \\
&\quad = \mathbb{1}[R_i \in \mathcal{R}]\Big[T_i(R_i) + (\widehat{R}_i - R_i)\dot{T}_i(R_i) + h\eta\dot{T}_i(R_i) + (\widehat{R}_i - R_i)^2(\ddot{T}_i(R_i) + o(1)) \\
&\qquad\qquad + 2(\widehat{R}_i - R_i)h\eta(\ddot{T}_i(R_i) + o(1)) + (h\eta)^2(\ddot{T}_i(R_i) + o(1))\Big]. \qquad (9)
\end{aligned}$$

Let $\alpha_i^h \equiv \mathbb{1}[R_i \notin \mathrm{bd}(\mathcal{R})] \int K(\eta)\mathbb{1}[\widehat{R}_i + h\eta \in \mathcal{R}]T_i(\widehat{R}_i + h\eta)\, d\eta$ so that from (8) we have $\int_{\mathcal{R}} P(r)^{-1}\sqrt{n}\widehat{A}(r)\, dr = n^{-1/2}\sum_{i=1}^n \alpha_i^h$. Substituting the Taylor expansion in (9) into the definition of α_i^h and using the symmetry of K, $\int K(\eta)\, d\eta = 1$ and $\mathbb{1}[R_i \in \mathcal{R}]\mathbb{1}[R_i \notin \mathrm{bd}(\mathcal{R})] = \mathbb{1}[R_i \in \mathcal{R}^\circ]$, we obtain

$$\begin{aligned}
\frac{1}{\sqrt{n}}\sum_{i=1}^n \alpha_i^h &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{1}[R_i \in \mathcal{R}^\circ]\left[T_i(R_i) + (\widehat{R}_i - R_i)\dot{T}_i(R_i)\right] \qquad (10) \\
&\quad + \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{1}[R_i \in \mathcal{R}^\circ]\left[(\widehat{R}_i - R_i)^2 + h^2\int \eta^2 K(\eta)\, d\eta\right](\ddot{T}_i(R_i) + o(1)).
\end{aligned}$$

It can be shown through some tedious algebra that I5, E5 and E6 imply that $\mathbb{E}[\mathbb{1}[R_i \in \mathcal{R}^\circ]\|\ddot{T}_i(R_i)\|]$ is finite. As a result, the second term in (10) is $o_{\mathbb{P}}(1)$, since by E7

$$\begin{aligned}
&\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{1}[R_i \in \mathcal{R}^\circ](\widehat{R}_i - R_i)^2(\ddot{T}_i(R_i) + o(1))\right\| \\
&\quad \le \max_{i:R_i\in\mathcal{R}^\circ} \sqrt{n}(\widehat{R}_i - R_i)^2\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}[R_i \in \mathcal{R}^\circ]\left(\|\ddot{T}_i(R_i)\| + o(1)\right)\right] = O_{\mathbb{P}}(n^{-1/2}),
\end{aligned}$$

and by $\int \eta^2 K(\eta)\, d\eta$ finite and E4,

$$\begin{aligned}
&\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \mathbb{1}[R_i \in \mathcal{R}^\circ]\left[h^2\int \eta^2 K(\eta)\, d\eta\right](\ddot{T}_i(R_i) + o(1))\right\| \\
&\quad \le \sqrt{n}O(h^2)\left[\frac{1}{n}\sum_{i=1}^n \mathbb{1}[R_i \in \mathcal{R}^\circ]\left(\|\ddot{T}_i(R_i)\| + o(1)\right)\right] = o_{\mathbb{P}}(1).
\end{aligned}$$

Substituting the asymptotically linear form for $\widehat{R}_i - R_i$ given in E7 into the first term in

(10), we obtain

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\alpha_i^h = \frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}[R_i\in\mathcal{R}^\circ]\left(T_i(R_i)+\xi_j(X_i|Z_i)\dot{T}_i(R_i)\right) + \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[R_i\in\mathcal{R}^\circ]\rho_n(X_i|Z_i)\dot{T}_i(R_i)+o_{\mathbb{P}}(1). \tag{11}$$

Some tedious algebra shows that I5, E5 and E6 imply that $\mathbb{E}[\mathbb{1}[R_i\in\mathcal{R}^\circ]\|\dot{T}_i(R_i)\|^4]$ is finite. Consequently, the second term in (11) is asymptotically negligible under E7 since

$$\left\|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[R_i\in\mathcal{R}^\circ]\rho_n(X_i|Z_i)\dot{T}_i(R_i)\right\| \leq \left[\sup_{(x,z)\in\mathcal{XZ}(\mathcal{R})}|\rho_n(x|z)|\right]\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}[R_i\in\mathcal{R}^\circ]\|\dot{T}_i(R_i)\| = o_{\mathbb{P}}(1).$$

The first term in (11) can be written as a second-order V-statistic with a symmetric kernel by defining $M_{ij}\equiv\mathbb{1}[R_i\in\mathcal{R}^\circ]\left(T_i(R_i)+\xi_j(X_i|Z_i)\dot{T}_i(R_i)\right)$ and $\zeta_{ij}\equiv\frac{1}{2}(M_{ij}+M_{ji})$, so that

$$\frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbb{1}[R_i\in\mathcal{R}^\circ]\left(T_i(R_i)+\xi_j(X_i|Z_i)\dot{T}_i(R_i)\right) = \frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\zeta_{ij}.$$

As noted, it can be shown that $\mathbb{E}[\mathbb{1}[R_i\in\mathcal{R}^\circ]\|\dot{T}_i(R_i)\|^4]$ is finite under our moment and smoothness assumptions. It can similarly be shown that I5, E5 and E6 imply that $\mathbb{E}[\mathbb{1}[R_i\in\mathcal{R}^\circ]\|T_i(R_i)\|^2]$ is also finite. These observations and E7 imply that

$$\mathbb{E}(\|M_{ij}\|^2)\leq\mathbb{E}\left(\mathbb{1}[R_i\in\mathcal{R}^\circ]\|T_i(R_i)\|^2\right) + \mathbb{E}\left(\mathbb{1}[R_i\in\mathcal{R}^\circ]\|\dot{T}_i(R_i)\|^4\right)^{1/2}\mathbb{E}\left(\xi_j(X_i|Z_i)^4\right)^{1/2}<\infty,$$

for both $j\neq i$ and $j=i$. Since $\mathbb{E}(\|\zeta_{ij}\|^2)\leq\frac{1}{2}\mathbb{E}(\|M_{ij}\|^2)$, this implies that $\mathbb{E}(\|\zeta_{ij}\|^2)<\infty$ for all i and j, by which we can conclude that

$$\frac{\sqrt{n}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\zeta_{ij} = \sqrt{n}\binom{n}{2}^{-1}\sum_{i<j}\zeta_{ij}+o_{\mathbb{P}}(1).$$

That is, the second-order V-statistic is asymptotically equivalent to a second-order U-

statistic, see e.g. pg. 206 of Serfling (1980).[8] Recapping, we have now shown that

$$\widehat{1}\sqrt{n}\left(\int_{\mathcal{R}}\widehat{\beta}(r)\,dr-\int_{\mathcal{R}}\beta(r)\,dr\right)=\widehat{1}\sqrt{n}\binom{n}{2}^{-1}\sum_{i<j}\zeta_{ij}+o_{\mathbb{P}}(1). \tag{12}$$

We determine the projection of the U-statistic in (12) by computing $\mathbb{E}(\zeta_{ij}|i)$ for $j\neq i$, where the notation $\mathbb{E}(\cdot|i)$ is shorthand for $\mathbb{E}(\cdot|Y_i,X_i,Z_i)$. First, since $\mathbb{E}[\xi_j(X_i|Z_i)|i]=0$ for $R_i\in\mathcal{R}^\circ$ by E1 and E7,

$$\begin{aligned}\mathbb{E}(M_{ij}|i)&=\mathbb{E}\left[\mathbb{1}[R_i\in\mathcal{R}^\circ]\left(T_i(R_i)+\xi_j(X_i|Z_i)\dot{T}_i(R_i)\right)|i\right]\\&=\mathbb{1}[R_i\in\mathcal{R}^\circ]\left(T_i(R_i)+\mathbb{E}(\xi_j(X_i|Z_i)|i)\dot{T}_i(R_i)\right)\overset{\text{a.s.}}{=}\mathbb{1}[R_i\in\mathcal{R}]T_i(R_i)\equiv\zeta_{1i},\end{aligned}$$

which has mean zero, because $\mathbb{E}(\zeta_{1i}|R_i=r)=\mathbb{1}[r\in\mathcal{R}]P(r)\,\mathbb{E}(W_iU_i(r)|R_i=r)=0$ and $\mathbb{E}(W_iU_i(r)|R_i=r)=\mathbb{E}(W_iW_i'(B_i-\beta(r))|R_i=r)=0$ by Proposition 1. Second,

$$\begin{aligned}\mathbb{E}(M_{ji}|i)&=\mathbb{E}\left[\mathbb{1}[R_j\in\mathcal{R}^\circ]\left(T_j(R_j)+\xi_i(X_j|Z_j)\dot{T}_j(R_j)\right)|i\right]\\&=\mathbb{E}[\mathbb{1}[R_j\in\mathcal{R}^\circ]T_j(R_j)]+\mathbb{E}\left[\mathbb{1}[R_j\in\mathcal{R}^\circ]\xi_i(X_j|Z_j)\dot{T}_j(R_j)|i\right]\\&=\mathbb{E}(\zeta_{1j})+\mathbb{E}\left[\mathbb{1}[R_j\in\mathcal{R}^\circ]\xi_i(X_j|Z_j)\,\mathbb{E}(\dot{T}_j(R_j)|R_j,Z_j)|i\right]\\&=-\mathbb{E}[\mathbb{1}[R_j\in\mathcal{R}]\xi_i(X_j|Z_j)P(R_j)^{-1}W_jW_j'\dot{\beta}(R_j)|i]\equiv\zeta_{2i},\end{aligned}$$

where the third equality uses the law of iterated expectations (noting that X_j is deterministic conditional on R_j,Z_j) and the fourth equality uses $\dot{T}_j(r)=-P(r)^{-1}[\dot{P}(r)P(r)^{-1}W_jU_j(r)+W_jW_j'\dot{\beta}(r)]$ and $\mathbb{E}[U_j(R_j)|R_j,Z_j]=0$.[9] Applying the law of iterated expectations shows that $\mathbb{E}(\zeta_{2i})=0$, since $\mathbb{E}(\xi_i(x|z))=0$ for $(x,z)\in\mathcal{XZ}(\mathcal{R})$ by E7. We have now shown that $\mathbb{E}(\zeta_{ij}|i)=\frac{1}{2}(\zeta_{1i}+\zeta_{2i})$, $\mathbb{E}(\zeta_{ij})=\mathbb{E}[\mathbb{E}(\zeta_{ij}|i)]=0$ and $\mathbb{E}(\|\zeta_{ij}\|^2)<\infty$, so that by the central limit theorem for U-statistics (e.g. Theorem A on page 192 of Serfling 1980)[10]

$$\sqrt{n}\binom{n}{2}^{-1}\sum_{i<j}\zeta_{ij}\rightsquigarrow N\left(0,\mathbb{E}[(\zeta_{1i}+\zeta_{2i})(\zeta_{1i}+\zeta_{2i})']\right). \tag{13}$$

[8] Serfling's discussion is limited to scalar-valued variables, but the modification for vector-valued variables is immediate.

[9] This expression for the component-wise derivative of $T_j(r)$ follows immediately from the rules in Section 6.5 of Horn and Johnson (1991).

[10] Serfling's discussion is limited to scalar-valued variables. The original work by Hoeffding (1948) contains an explicit statement of the vector case; see also Chapter 5 of Kowalski and Tu (2008) for a modern treatment.

Combining equations (12) and (13) and applying Slutsky's theorem with $\widehat{1} \to_{\mathbb{P}} 1$ from Lemma A.1, we have

$$\sqrt{n}\left(\int_{\mathcal{R}} \widehat{\beta}(r)\,dr - \int_{\mathcal{R}} \beta(r)\,dr\right) \rightsquigarrow N\left(0, \mathbb{E}[(\zeta_{1i}+\zeta_{2i})(\zeta_{1i}+\zeta_{2i})']\right).$$

The result now follows after scaling both sides by $\lambda(\mathcal{R})^{-1}$. *Q.E.D.*

Lemma A.1. *Under the assumptions of Theorem 3,* $\widehat{1} \to_{\mathbb{P}} 1$ *and*

$$\widehat{1}\int_{\mathcal{R}} [\widehat{P}(r)^{+} - P(r)^{-1}]\widehat{A}(r)\,dr = o_{\mathbb{P}}(n^{-1/2}).$$

Proof of Lemma A.1. Let $J(r) \equiv \widehat{P}(r)^{+} - P(r)^{-1}$, let $\|\cdot\|_1$ denote the l_1 norm and let $\|\cdot\|_{1,op}$ denote the matrix operator norm induced by the l_1 norm. Then

$$\begin{aligned}\left\|\widehat{1}\int_{\mathcal{R}} J(r)\widehat{A}(r)\,dr\right\|_1 &\leq \widehat{1}\int_{\mathcal{R}} \|J(r)\widehat{A}(r)\|_1\,dr \\ &\leq \widehat{1}\int_{\mathcal{R}} \|J(r)\|_{op,1}\|\widehat{A}(r)\|_1\,dr \leq \widehat{1}\sup_{r\in\mathcal{R}}\|J(r)\|_{op,1}\sup_{r\in\mathcal{R}}\|\widehat{A}(r)\|, \qquad (14)\end{aligned}$$

where the first inequality follows because for vector-valued function $x : \mathbb{R} \to \mathbb{R}^K$ and component-wise integration we have $\sum_{k=1}^K |\int x(r)\,dr| \leq \sum_{k=1}^K \int |x(r)|\,dr = \int \|x(r)\|_1\,dr$, and the second inequality uses the sub-multiplicative property of the matrix operator norm.

First, we consider the behavior of

$$\sup_{r\in\mathcal{R}}\|\widehat{P}(r) - P(r)\| \leq \sup_{r\in\mathcal{R}}\|\widehat{P}(r) - \widetilde{P}(r)\| + \sup_{r\in\mathcal{R}}\|\widetilde{P}(r) - P(r)\|,$$

where $\widetilde{P}(r) \equiv n^{-1}\sum_{i=1}^n k_i^h(r) W_i W_i'$ with $k_i^h \equiv h^{-1}K((R_i - r)/h)$. Notice that both $\widehat{P}(r)$ and $\widetilde{P}(r)$ are Nadaraya-Watson kernel regression estimators of the matrix $P(r) \equiv \mathbb{E}[WW'|R = r]f_R(r)$ with $f_R(r) = 1$, but that the weights in $\widehat{P}(r)$ use the generated regressor $\widehat{R}_i$, while the weights in $\widetilde{P}(r)$ use R_i. Recent work on nonparametric regression with generated regressors has established that $\sup_{r\in\mathcal{R}}\|\widehat{P}(r) - \widetilde{P}(r)\| = O_{\mathbb{P}}(\log(n)n^{-1/2})$ under our assumptions.[11] Using standard results in the literature, our assumptions also ensure that

[11] In particular, we appeal to Lemma 1 of Mammen et al. (2012), but see also Sperlich (2009), Mammen, Rothe, and Schienle (2013), Hahn and Ridder (2013), Lee (2013) and Escanciano, Jacho-Chavez, and Lewbel (2014) for related results. We verify the conditions for Lemma 1 of Mammen et al. (2012). In their Assumption 1, (i) is E1, (ii) is satisfied with $R \sim \text{Unif}[0,1]$, (iii) is E5, (iv) is not used in the proof of their Lemma 1, (v) is E3 and (vi) is met under E4 and E7. Their Assumptions 2 and 3 are

$\sup_{r\in\mathcal{R}} \|\widetilde{P}(r)-P(r)\| = O_{\mathbb{P}}((\log(n)/nh)^{1/2}+h^2)$, with the dominant rate being $(\log(n)/nh)^{1/2}$ given E4.[12] Since this rate also dominates $\log(n)n^{-1/2}$, it follows that $\sup_{r\in\mathcal{R}} |\widehat{P}(r)-P(r)| = O_{\mathbb{P}}((\log(n)/nh)^{1/2})$. Note that this also implies that $\widehat{1} \to_{\mathbb{P}} 1$. Using these observations, the definition of $\widehat{1}$ and I5 with E5, we have that

$$\begin{aligned}\widehat{1}\sup_{r\in\mathcal{R}} \|J(r)\| &= \widehat{1}\sup_{r\in\mathcal{R}} \|\widehat{P}(r)^{-1}(P(r)-\widehat{P}(r))P(r)^{-1}\| \\ &\leq \widehat{1}\sup_{r\in\mathcal{R}} \|\widehat{P}(r)^{-1}\| \sup_{r\in\mathcal{R}} \|\widehat{P}(r)-P(r)\| \sup_{r\in\mathcal{R}} \|P(r)^{-1}\| = O_{\mathbb{P}}((\log(n)/nh)^{1/2}).\end{aligned}$$

The same rate applies to $\widehat{1}\sup_{r\in\mathcal{R}} \|J(r)\|_{op,1}$, because finite-dimensional norms are equivalent.

A rate of convergence for $\sup_{r\in\mathcal{R}} \|\widehat{A}(r)\|$ follows similarly after using the definition of $U_i(r) \equiv W_i'(B_i-\beta(r))$ to write

$$\widehat{A}(r) = \frac{1}{n}\sum_{i=1}^n \widehat{k}_i^h(r) W_i W_i' B_i - \widehat{P}(r)\beta(r)$$

$$\text{and} \quad \widetilde{A}(r) = \frac{1}{n}\sum_{i=1}^n k_i^h(r) W_i W_i' B_i - \widetilde{P}(r)\beta(r).$$

The difference of the first two terms in these expressions is uniformly $O_{\mathbb{P}}(\log(n)n^{-1/2})$ again by Lemma 1 of Mammen et al. (2012).[13] Since $\beta(r)$ is bounded uniformly over $\mathcal{R}$ by E2 and E5, the difference of the second two terms is also $O_{\mathbb{P}}(\log(n)n^{-1/2})$ using the already established rate for $\sup_{r\in\mathcal{R}} \|\widehat{P}(r)-\widetilde{P}(r)\|$, and hence $\sup_{r\in\mathcal{R}} \|\widehat{A}(r)-\widetilde{A}(r)\| = O_{\mathbb{P}}(\log(n)n^{-1/2})$. Also, since $\mathbb{E}[WW'B|R=r]-\mathbb{E}[WW'|R=r]\beta(r)=0$ by Proposition 1, standard results again imply that $\sup_{r\in\mathcal{R}} \|\widetilde{A}(r)\| = O_{\mathbb{P}}((\log(n)/nh)^{1/2})$ under our assumptions. We conclude that $\sup_{r\in\mathcal{R}} \|\widehat{A}(r)\| = O_{\mathbb{P}}((\log(n)/nh)^{1/2})$. This establishes the claim via (14), since under E4, $\sqrt{n}\log(n)(nh)^{-1} = \log(n)/(\sqrt{n}h) \to 0$. *Q.E.D.*

satisfied by our E7, while their Assumption 4 is not used in the proof of their Lemma 1. The rate of $O_{\mathbb{P}}(\log(n)n^{-1/2})$ is determined by computing κ_1 on pg. 1141 and observing their notational convention of leaving out $\log(n)$ terms.

[12] For example, see Lemma B3 of Newey (1994).

[13] The verification for most of their conditions is as in footnote 11. Their Assumption 1 (iii) is satisfied since each component of $\mathbb{E}[WW'B|R=r] = P(r)\beta(r)$ is twice continuously differentiable by E5.

References

Aliprantis, C. D. and K. C. Border (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Springer, 3 ed. 29

Angrist, J. D. and G. W. Imbens (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *The Journal of the American Statistical Association*, 90, 431–442. 2

Angrist, J. D., M. P. Keane, E. E. Leamer, A. Nevo, J.-S. Pischke, C. A. Sims, J. H. Stock, and M. D. Whinston (2010): "Symposia: Con out of Economics," *The Journal of Economic Perspectives*, 24, 3–94.

Beran, R. and P. Hall (1992): "Estimating Coefficient Distributions in Random Coefficient Regressions," *The Annals of Statistics*, 1970–1984. 4

Blundell, R. W. and J. L. Powell (2004): "Endogeneity in Semiparametric Binary Response Models," *The Review of Economic Studies*, 71, 655–679. 4

Chamberlain, G. (1992): "Efficiency Bounds for Semiparametric Regression," *Econometrica*, 60, 567–596. 4

Chay, K. Y. and M. Greenstone (2005): "Does Air Quality Matter? Evidence from the Housing Market," *The Journal of Political Economy*, 113, 376–424. 5, 21, 22

Chernozhukov, V., I. Fernández-Val, and A. Galichon (2010): "Quantile and Probability Curves Without Crossing," *Econometrica*, 78, 1093–1125. 13, 17

Chernozhukov, V., I. Fernández-Val, and A. Kowalski (2011): "Quantile Regression with Censoring and Endogeneity," *Working paper*. 18

Chernozhukov, V., I. Fernández-Val, and B. Melly (2009): "Inference on Counterfactual Distributions," *Working paper*. 13, 17

——— (2012): "Inference on Counterfactual Distributions," *Econometrica (forthcoming)*. 13, 17

Chernozhukov, V. and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73, 245–261. 4, 10

Chesher, A. (2003): "Identification in Nonseparable Models," *Econometrica*, 71, 1405–1441. 4

Deaton, A., J. J. Heckman, and G. W. Imbens (2010): "Forum on the Estimation of Treatment Effects," *The Journal of Economic Literature*, 48, 356–455.

D'Haultfœuille, X. and P. Février (2012): "Identification of Nonseparable Models with Endogeneity and Discrete Instruments," *Working paper*. 4

Doksum, K. (1974): "Empirical Probability Plots and Statistical Inference for Nonlinear Models in the Two-Sample Case," *The Annals of Statistics*, 2, 267–277. 10

Ekeland, I., J. J. Heckman, and L. Nesheim (2004): "Identification and Estimation of Hedonic Models," *The Journal of Political Economy*, 112, 60. 21

Escanciano, J. C., D. T. Jacho-Chavez, and A. Lewbel (2014): "Uniform Convergence of Weighted Sums of Non and Semiparametric Residuals for Estimation and Testing," *Journal of Econometrics*, 178, 426–443. 33

Florens, J. P., J. J. Heckman, C. Meghir, and E. Vytlacil (2008): "Identification of Treatment Effects Using Control Functions in Models With Continuous, Endogenous Treatment and Heterogeneous Effects," *Econometrica*, 76, 1191–1206. 3, 4, 5, 6, 7, 8, 9, 11, 22

Graham, B. S. and J. L. Powell (2012): "Identification and Estimation of Average Partial Effects in Irregular Correlated Random Coefficient Panel Data Models," *Econometrica*, 80, 2105–2152. 4

Hahn, J. and G. Ridder (2013): "Asymptotic Variance of Semiparametric Estimators With Generated Regressors," *Econometrica*, 81, 315–340. 33

Heckman, J. and E. Vytlacil (1998): "Instrumental Variables Methods for the Correlated Random Coefficient Model: Estimating the Average Rate of Return to Schooling When the Return is Correlated with Schooling," *The Journal of Human Resources*, 33, 974–987. 2, 5, 9, 10, 20, 22

Heckman, J. J. (2001): "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *The Journal of Political Economy*, 109, 673–748. 2

Heckman, J. J., R. L. Matzkin, and L. Nesheim (2010): "Nonparametric Identification and Estimation of Nonadditive Hedonic Models," *Econometrica*, 78, 1569–1591. 21

Heckman, J. J., J. Smith, and N. Clements (1997): "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *The Review of Economic Studies*, 64, 487–535. 10

Hoderlein, S., J. Klemelä, and E. Mammen (2010): "Analyzing the Random Coefficient Model Nonparametrically," *Econometric Theory*, 26, 804–837. 4

Hoderlein, S. and R. Sherman (2013): "Identification and Estimation in a Correlated Random Coefficients Binary Response Model," *Working paper*. 4, 14

Hoeffding, W. (1948): "A Class of Statistics with Asymptotically Normal Distribution," *The Annals of Mathematical Statistics*, 19, 293–325. 32

Horn, R. A. and C. R. Johnson (1991): *Topics in Matrix Analysis*, Cambridge: Cambridge University Press. 29, 32

Imbens, G. (2007): "Nonadditive Models with Endogenous Regressors," in *Advances in Economics and Econometrics: Theory and Applications, Ninth World Congress*, ed. by R. Blundell, W. Newey, and T. Persson, Cambridge University Press, New York, vol. 3. 2, 7

Imbens, G. W. and J. D. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467–475. 2, 9

Imbens, G. W. and W. K. Newey (2009): "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity," *Econometrica*, 77, 1481–1512. 3, 4, 7, 12

Jun, S. J. (2009): "Local Structural Quantile Effects in a Model with a Nonseparable Control Variable," *Journal of Econometrics*, 151, 82–97. 12

Kasy, M. (2011): "Identification In Triangular Systems Using Control Functions," *Econometric Theory*, 27, 663–671. 7

——— (2013): "Identification in General Triangular Systems," *Working paper*. 3, 10

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50. 13, 18

Kowalski, J. and X. M. Tu (2008): *Modern Applied U-statistics*, Wiley. 32

Lee, Y.-Y. (2013): "Partial Mean Processes with Generated Regressors: Continuous Treatment Effects and Nonseparable Models," *Working paper*. 33

Li, M. and J. L. Tobias (2011): "Bayesian Inference in a Correlated Random Coefficients Model: Modeling Causal Effect Heterogeneity with an Application to Heterogeneous Returns to Schooling," *Journal of Econometrics*, 162, 345–361. 3

Mammen, E., C. Rothe, and M. Schienle (2012): "Nonparametric Regression With Nonparametrically Generated Covariates," *The Annals of Statistics*, 40, 1132–1170. 17, 33, 34

——— (2013): "Semiparametric Estimation with Generated Covariates," *Working paper*. 33

Masten, M. A. (2012): "Random Coefficients on Endogenous Variables in Simultaneous Equations Models," *Working paper*. 4

Newey, W. K. (1994): "Kernel Estimation of Partial Means and a General Variance Estimator," *Econometric Theory*, 10, 233–253. 34

Palmquist, R. B. (2005): "Property Value Models," *Handbook of Environmental Economics*, 2, 763–819. 21

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica*, 57, 1403–1430. 17

Rosen, S. (1974): "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *The Journal of Political Economy*, 82, 34–55. 21

Rothe, C. (2009): "Semiparametric Estimation of Binary Response Models with Endogenous Regressors," *Journal of Econometrics*, 153, 51–64. 4

Ruppert, D. and M. P. Wand (1994): "Multivariate Locally Weighted Least Squares Regression," *The Annals of Statistics*, 22, 1346–1370. 17

Serfling, R. J. (1980): *Approximation Theorems of Mathematical Statistics*, New York: John Wiley & Sons. 32

Smith, V. K. and J.-C. Huang (1995): "Can Markets Value Air Quality? A Meta-analysis of Hedonic Property Value Models," *The Journal of Political Economy*, 209–227. 21

Sperlich, S. (2009): "A Note on Non-parametric Estimation with Predicted Variables," *Econometrics Journal*, 12, 382–395. 33

Torgovitsky, A. (2012): "Identification of Nonseparable Models with General Instruments," *Working paper*. 4, 7, 10

——— (2013): "Minimum Distance from Independence Estimation of Nonseparable Instrumental Variables Models," *Working paper*. 4

Train, K. (2009): *Discrete Choice Methods with Simulation*, Cambridge University Press, 2 ed. 14

U.S. Department of Commerce. Bureau of the Census (2008): *County and City Data Book [United States], 1983. ICPSR version*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICPSR) [distributor], http://doi.org/10.3886/ICPSR08256.v1.

——— (2012): *County and City Data Book [United States] Consolidated File: County Data, 1947-1977, ICPSR07736-v2*, Ann Arbor, MI: Inter-university Consortium for Political and Social Research (ICSPR) [distributor], http://doi.org/10.3886/ICPSR07736.v2.

U.S. Environmental Protection Agency (2001): "Historical TSP data, 1970–1999," http://www.epa.gov/ttn/airs/airsaqs/archived%20data, accessed: 10-8-2013.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer. 16

Wooldridge, J. M. (1997): "On Two Stage Least Squares Estimation of the Average Treatment Effect in a Random Coefficient Model," *Economics Letters*, 56, 129–133. 2, 5, 9, 10, 20, 22

——— (2003): "Further Results on Instrumental Variables Estimation of Average Treatment Effects in the Correlated Random Coefficient Model," *Economics Letters*, 79, 185–191. 2, 5, 9, 10, 20, 21, 22

——— (2008): "Instrumental Variables Estimation of the Average Treatment Effect in the Correlated Random Coefficient Model," in *Modelling and Evaluating Treatment Effects in Econometrics (Advances in Econometrics, Volume 21)*, ed. by D. Millimet, J. Smith, and E. Vytlacil, Emerald Group Publishing Limited, 93–116. 2, 5, 9, 10, 20, 21, 22

estimators of	$N = 500$			$N = 1000$		
$\mathbb{E}(B_0) = .23$	bias	(std)	mse	bias	(std)	mse
OLS	0.0291	(0.0228)	0.0014	0.0294	(0.0172)	0.0012
TSLS	0.1124	(0.0581)	0.0160	0.1105	(0.0393)	0.0137
h = 0.01	-0.0136	(0.1784)	0.0320	-0.0164	(0.1085)	0.0120
h = 0.03	-0.0394	(0.1322)	0.0190	-0.0391	(0.0830)	0.0084
h = 0.05	-0.0590	(0.1118)	0.0160	-0.0571	(0.0703)	0.0082
h = 0.07	-0.0750	(0.0988)	0.0154	-0.0724	(0.0620)	0.0091
h = 0.09	-0.0878	(0.0896)	0.0157	-0.0852	(0.0558)	0.0104
h = 0.11	-0.0980	(0.0822)	0.0164	-0.0957	(0.0510)	0.0118
h = 0.13	-0.1066	(0.0762)	0.0172	-0.1045	(0.0472)	0.0132
h = 0.15	-0.1137	(0.0716)	0.0181	-0.1120	(0.0444)	0.0145
estimators of	$N = 500$			$N = 1000$		
$\mathbb{E}(B_1) = .52$	bias	(std)	mse	bias	(std)	mse
OLS	0.4136	(0.0889)	0.1790	0.4142	(0.0648)	0.1757
TSLS	-0.0024	(0.2757)	0.0760	0.0104	(0.1877)	0.0353
h = 0.01	-0.0125	(0.2668)	0.0713	0.0094	(0.1700)	0.0290
h = 0.03	0.0207	(0.2256)	0.0513	0.0338	(0.1491)	0.0234
h = 0.05	0.0532	(0.2012)	0.0433	0.0621	(0.1333)	0.0216
h = 0.07	0.0853	(0.1826)	0.0406	0.0908	(0.1211)	0.0229
h = 0.09	0.1158	(0.1674)	0.0414	0.1192	(0.1111)	0.0265
h = 0.11	0.1442	(0.1549)	0.0448	0.1463	(0.1028)	0.0320
h = 0.13	0.1706	(0.1447)	0.0501	0.1719	(0.0963)	0.0388
h = 0.15	0.1948	(0.1368)	0.0566	0.1955	(0.0914)	0.0466

Table 1: Performance of $\widehat{\beta}_{\mathcal{R}}$ as an estimator of $\mathbb{E}(B)$ relative to ordinary least squares (OLS) and two stage least squares (TSLS) in the dgp without first stage heterogeneity ($\gamma = 0$).

estimators of	$N = 500$			$N = 1000$		
$\mathbb{E}(B_0) = .23$	bias	(std)	mse	bias	(std)	mse
OLS	0.0532	(0.0254)	0.0035	0.0537	(0.0190)	0.0032
TSLS	0.0927	(0.0542)	0.0115	0.0912	(0.0360)	0.0096
h = 0.01	-0.0067	(0.1562)	0.0244	-0.0157	(0.1146)	0.0134
h = 0.03	-0.0226	(0.0993)	0.0104	-0.0228	(0.0684)	0.0052
h = 0.05	-0.0349	(0.0887)	0.0091	-0.0331	(0.0595)	0.0046
h = 0.07	-0.0451	(0.0836)	0.0090	-0.0432	(0.0549)	0.0049
h = 0.09	-0.0543	(0.0797)	0.0093	-0.0528	(0.0516)	0.0055
h = 0.11	-0.0626	(0.0761)	0.0097	-0.0615	(0.0486)	0.0061
h = 0.13	-0.0704	(0.0729)	0.0103	-0.0696	(0.0461)	0.0070
h = 0.15	-0.0776	(0.0702)	0.0109	-0.0771	(0.0439)	0.0079
estimators of	$N = 500$			$N = 1000$		
$\mathbb{E}(B_1) = .52$	bias	(std)	mse	bias	(std)	mse
OLS	0.3678	(0.0926)	0.1439	0.3690	(0.0665)	0.1406
TSLS	0.1888	(0.2578)	0.1021	0.2002	(0.1742)	0.0704
h = 0.01	-0.0194	(0.3205)	0.1031	-0.0105	(0.2315)	0.0537
h = 0.03	-0.0077	(0.1970)	0.0389	0.0053	(0.1310)	0.0172
h = 0.05	0.0101	(0.1655)	0.0275	0.0213	(0.1100)	0.0126
h = 0.07	0.0304	(0.1512)	0.0238	0.0389	(0.1009)	0.0117
h = 0.09	0.0515	(0.1423)	0.0229	0.0579	(0.0949)	0.0124
h = 0.11	0.0724	(0.1354)	0.0236	0.0769	(0.0901)	0.0140
h = 0.13	0.0923	(0.1298)	0.0254	0.0956	(0.0864)	0.0166
h = 0.15	0.1113	(0.1253)	0.0281	0.1138	(0.0836)	0.0199

Table 2: Performance of $\widehat{\beta}_{\mathcal{R}}$ as an estimator of $\mathbb{E}(B)$ relative to ordinary least squares (OLS) and two stage least squares (TSLS) in the dgp with first stage heterogeneity ($\gamma = 0.4$).

Table 3: Summary Statistics, 1970 and 1980

	1970	1980
Mean housing value	40,268	53,046
Mean TSPs	65.5	56.3
Income per capita	7,530	9,279
Total population	163,880,811	175,516,811
Unemployment rate	.0455	.068
% employment in manufacturing	.249	.226
Population density	613	476
% $\geq$ high school graduate	.504	.646
% $\geq$ college graduate	.0971	.146
% urban	.576	.593
% poverty	.124	.0976
% white	.902	.877
% female	.51	.511
% senior citizens	.0997	.113
% overall vacancy rate	.0336	.0782
% owner-occupied	.676	.638
% of houses without plumbing	.0822	.0253
Per capita government revenue	748	1,138
Per capita property taxes	314	366
Per capita general expenditures	769	1,111
% spending on education	.549	.509
% spending on highways	.0909	.0698
% spending on welfare	.0462	.0371
% spending on health	.0486	.0669
Observations	989	989

Statistics are based on the 989 counties with data on TSP in 1970, 1980, and 1974 or 1975, as well as nonmissing price data in both 1970 and 1980. Mean TSP for 1970 is the average of 1970 to 1972 annual TSP. Mean TSP for 1980 is the average of 1977 to 1980 annual TSP. Annual TSP for a county is the weighted average of the geometric mean of each monitor's TSP readings in the county, using the number of observations per monitor as weights. All dollar quantities are adjusted to 1982-1984 dollars (housing values use the housing only part of the CPI, series CUUR0000SAH; all other values use overall CPI, series CUUR0000SA0).

Table 4: Estimates of the effect of 1970–1980 changes in TSP pollution on changes in log housing values

Estimator				
OLS	0.0861		0.0288	
	[0.0079, 0.1643]		[-0.0200, 0.0776]	
TSLS	-0.4149		-0.2073	
	[-0.7616, -0.0682]		[-0.4258, 0.0113]	
	(1)	(2)	(3)	(4)
Generalized CRC estimator	$\mathcal{R} = [0.1, 0.4]$	$\mathcal{R} = [0.4, 0.7]$	$\mathcal{R} = [0.1, 0.4]$	$\mathcal{R} = [0.4, 0.7]$
$h = 0.040$	-0.0241	-0.2067	-0.0664	-0.1517
	[-0.3011, 0.2314]	[-0.7879, 0.4544]	[-0.4395, 0.1585]	[-0.7414, 0.4363]
$h = 0.0475$	-0.0252	-0.1766	-0.0640	-0.1535
	[-0.2994, 0.2239]	[-0.7635, 0.4540]	[-0.3981, 0.1407]	[-0.6894, 0.3974]
$h = 0.055$	-0.0261	-0.1545	-0.0670	-0.1580
	[-0.2974, 0.2210]	[-0.7419, 0.4498]	[-0.3864, 0.1215]	[-0.6441, 0.3634]
$h = 0.0625$	-0.0266	-0.1364	-0.0703	-0.1592
	[-0.2940, 0.2187]	[-0.6967, 0.4410]	[-0.3719, 0.1179]	[-0.6258, 0.3310]
$h = 0.0775$	-0.0258	-0.1073	-0.0755	-0.1544
	[-0.2856, 0.2047]	[-0.6551, 0.4255]	[-0.3422, 0.1057]	[-0.5988, 0.2751]
$h = 0.085$	-0.0248	-0.0954	-0.0784	-0.1524
	[-0.2807, 0.2009]	[-0.6329, 0.4430]	[-0.3321, 0.1031]	[-0.5927, 0.2636]
Observations	983		983	
County data book controls?	No		Yes	

Entries show estimates of coefficients on change in TSP over 1970-1980, and corresponding 95-percent confidence intervals for several different estimators: ordinary least squares, two-stage least squares, and a variety of bandwidths for our generalized correlated random coefficient model estimator. TSP here is in units of 1×10^{-4} grams/m^3 rather than 1×10^{-6} grams/m^3 $= \mu$g/m^3. Columns (1) and (3) show the generalized CRC estimates for conditional ranks over $\mathcal{R} = [0.1, 0.4]$ while columns (2) and (4) use $\mathcal{R} = [0.4, 0.7]$. All regressions are first differenced from 1970 to 1980. The outcome variable is 1980 log-housing value minus 1970 log-housing value. The treatment variable of interest is the 1980 TSP value minus the 1970 TSP value. All controls are also first differenced. The instrument is mid-decade nonattainment status (see body text for further details). OLS and TSLS confidence intervals are computed via asymptotic plug-in estimators of the heteroskedasticity robust standard errors. The generalized CRC model confidence intervals are computed using 500 bootstrap draws.