Sarnetzki, Florian; Dzemski, Andreas

Conference Paper

Overidentification test in a nonparametric treatment model with unobserved heterogeneity

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Econometric Theory, No. C20-V1

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Sarnetzki, Florian; Dzemski, Andreas (2014) : Overidentification test in a nonparametric treatment model with unobserved heterogeneity, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2014: Evidenzbasierte Wirtschaftspolitik - Session: Econometric Theory, No. C20-V1, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:
https://hdl.handle.net/10419/100620

Overidentification test in a nonparametric treatment model with unobserved heterogeneity

Andreas Dzemski and Florian Sarnetzki*

First Draft: January 14, 2014
This Version: February 15, 2014

We provide an instrument test for a treatment model in which individuals select into treatment based on unobserved gains (Imbens and Angrist 1994). We augment a standard model by assuming that both a binary and a continuous instrument are available. Under treatment monotonicity a parameter that is closely related to the Marginal Treatment Effect (cf. Heckman and Vytlacil 2005) is overidentified. We suggest a test statistic and characterize its asymptotic distribution and behavior under local alternatives. In simulations, we investigate the validity and finite sample performance of a wild bootstrap procedure. Finally, we illustrate the applicability of our method by studying two instruments from the literature on teenage pregnancies.

JEL codes: C21, C14

Keywords: treatment effects, unobserved heterogeneity, overidentification test, instrumental variables, generated regressors, wild bootstrap, teenage pregnancies

1. Introduction

The canonical treatment effect evaluation problem in Economics can be phrased as the problem of recovering the coefficient β from the outcome equation

$$Y = \alpha + \beta D, \tag{1}$$

*Center for Doctoral Studies in Economics, Universität Mannheim. We are indebted to our advisors Enno Mammen and Markus Frölich for constant support. We are grateful for comments from Steffen Reinhold, Martin Huber, Allie Carnegie, Peter Aronow, Anne Leucht and Carsten Jentsch.

where D is a binary indicator of treatment status, and α and β are random coefficients. If the treatment effect β is known to be constant then β can be identified by classical instrumental variables methods. In this framework it is straightforward to test the validity of the instruments by classical GMM overidentification tests (Hansen 1982, Sargan 1958). In many applications the more natural assumption is to assume that the treatment effect β is non-constant and correlated with D. Economically this means that individuals differ in their gains from participating in the treatment and that when deciding whether to participate or not individuals take into account possible gains from participation. This setting is often referred to as one of *essential heterogeneity* (Heckman, Urzua, and Vytlacil 2006). It was first considered in the seminal papers by Imbens and Angrist 1994 and Angrist, Imbens, and Rubin 1996. These authors give assumptions under which a binary instrument identifies the average treatment effect for the subpopulation of compliers which they dub the Local Average Treatment Effect (LATE). The compliers are the individuals that respond to a change in the realizations of the binary instrument by changing their participation decision. Different instruments may induce different subpopulations to change their treatment status and therefore estimate different LATEs. Hence, if a GMM overidentification test rejects, this no longer constitutes compelling evidence that one instrument is invalid. Rather, it might as well be interpreted as evidence for a non-constant treatment effect (Heckman, Schmierer, and Urzua 2010).

In this paper we present an instrument test that is valid under essential heterogeneity. A key assumption of Imbens and Angrist 1994, which we maintain as well, is treatment monotonicity. Intuitively, this assumption says that individuals can be ordered by their willingness to participate in the treatment. As we show below, an immediate consequence of the monotonicity assumption is that the propensity score, i.e., the proportion of individuals who participate in the treatment, subsumes all information about observed outcomes that is included in a vector of instruments. This type of index sufficiency is a testable restriction since both observed outcomes and propensity scores are identified from the data. More concretely, we assume that a binary and a continuous instrument are available. The purpose of the binary instrument is to split the population into two subpopulations with distinct propensity scores in a way that is independent of unobserved characteristics. We then test whether observed outcomes conditional on the propensity score are identical in the two subpopulations. The reason why we assume continuity of the second instrument is that this offers a plausible way to argue that the supports of the propensity scores in the two subpopulations overlap.

Our test is related to the test of the validity of the matching approach suggested in Heckman et al. 1996 and Heckman et al. 1998. Their test also exploits index sufficiency under the null hypothesis. Moreover, the role that random assignment to a control group serves in their testing approach is similar to the part that the binary instrument plays in our overidentification result. The testing theory that we develop in this paper translates with slight modifications to the testing problem of Heckman et al. 1996 and Heckman et al. 1998. We hope that it will prove useful in other settings where the null hypothesis imposes some kind of index sufficiency as well.

Our testable restriction in terms of a conditional mean function is closely related to

a similar restriction in terms of the Marginal Treatment Effect (MTE, see Heckman and Vytlacil 2005 for a discussion of the MTE). The characterization of the restriction in terms of the MTE, while certainly the less practical one for testing, has a lot of theoretical appeal as it illustrates that our test is based on the overidentification of a structural parameter of the model.

We are not the first to consider the problem of testing instruments in a model with essential heterogeneity. Following previous work by Balke and Pearl 1997, Kitagawa 2008 and Huber and Mellace 2011 consider testing the validity of a discrete instrument in a LATE model. They test inequalities for the densities and the mean of the outcomes for always takers and never takers, i.e. two subpopulations for which treatment status is not affected by the instrument. In stark contrast, our test focuses on the subpopulation which responds to the instrument. Angrist and Fernandez-Val 2010 develop a LATE overidentification test under the additional assumption that the heterogeneity is captured by observed covariates. We do not require such an assumption. Our test lends itself naturally to testing continuous instruments, whereas previous tests can handle continuous instruments only via a discretization.

Our method works if both a binary and a continuous instrument are available. This is the case in many relevant applications. In this paper we apply our method to test for the validity of instruments that have been used to investigate the effect of teenage child bearing on high school completion. For another example of an evaluation problem where our method would come to bear consider Carneiro, Heckman, and Vytlacil 2011. They estimate returns to schooling using as instruments a binary indicator of distance to college, tuition fees, as well as continuous measures of local labor market conditions.

Our test reduces to the problem of testing the equality of two nonparametric regression curves. This is a problem with a rich history in the statistical literature (cf., e.g., Hall and Hart 1990; King, Hart, and Wehrly 1991; Delgado 1993; Dette and Neumeyer 2001; Neumeyer and Dette 2003). Our testing problem, however, does not fit directly into any of the frameworks analyzed in the previous literature as it comes with the added complication of generated regressors. We propose a test statistic and quantify the effect of the first stage estimation error on the asymptotic distribution of the test statistic. We find that in order to have good power against local alternatives we have to reduce the nonparametric bias from the first stage estimation. With our particular choice of second stage estimator no further bias reduction is necessary.

We propose a bootstrap procedure to compute critical values. In the context of a treatment model with nonparametrically generated regressors Y. Lee 2013 establishes the validity of a multiplier bootstrap that is based on the first order terms in an asymptotic expansion of the underlying process. We suggest a wild bootstrap procedure that does not rely on first order asymptotics and that is easy to implement in standard software. In exploratory simulations our procedure is faithful to its nominal size in small and medium sized samples.

The paper is structured as follows. Section 2 defines our heterogeneous treatment model. In Section 3 we state our central overidentification result, discuss nonparametric parameter estimation, and define the test statistic. The asymptotic behavior of our test statistic is discussed in Section 4. Our simulations are presented in Section 5. In

Section 6 we apply our approach to real data and study the validity of instruments in the context of teenage child bearing and high school graduation. Section 7 concludes.

2. Model definition

Our version of a treatment model with unobserved heterogeneity in the spirit of Imbens and Angrist 1994 is owed in large part to Vytlacil 2002. As in Abadie 2003 and Frölich 2007 we assume that our assumptions hold conditional on a set of covariates. We restrict ourselves to covariates that take values in a finite set. Our main overidentification result carries over to more general covariate spaces in a straightforward manner. The purpose of the restriction is exclusively to facilitate estimation by keeping the estimation of infinite dimensional nuisance parameters free of the curse of dimensionality. Without loss of generality assume that we can enumerate all possible covariate configurations by $\{1, \ldots, J^{\max}\}$ and let J denote the covariate configuration of an individual. Treatment status is binary and is denoted by D. The latent outcomes are denoted by Y^0 and Y^1 and $Y = (1-D)Y^0 + DY^1$ denotes the observed outcome. Note that by setting $\alpha = Y^0$ and $\beta = Y^1 - Y^0$ we recover the correlated random effects model from equation (1). Let S denote a continuous random variable and let Z denote a binary random variable. Below, S and Z are required to fulfill certain conditional independence assumptions that render them valid instruments in a heterogeneous treatment model. We observe a sample $(Y_i, D_i, S_i, Z_i, J_i)_{i \leq n}$ from (Y, D, S, Z, J). Treatment status is determined by the threshold crossing decision rule

$$D = 1_{\{r_{Z,J}(S) \geq V\}},$$

with $r_{z,j}$ a function that is bounded between zero and one and V satisfying

$$V \sim U[0,1] \quad \text{and} \quad V \perp\!\!\!\perp (S, Z) \mid J. \tag{I-V}$$

Under this assumption the function $r_{z,j}$ is a propensity score and V can be interpreted as an individual's type reflecting her natural proclivity to select into the treatment group. As pointed out in Vytlacil 2002 the threshold crossing model imposes treatment monotonicity.[1] The assumption that V is uniformly distributed is merely a convenient normalization that allows us to identify $r_{z,j}$. The crucial part of this assumption is that the instruments are jointly independent of the heterogeneity parameter V. This allows us to use the instruments as a source of variation in treatment participation that is independent of the unobserved types. Furthermore, we assume that for given V, Z and J the latent outcomes are independent of S,

$$Y^d \perp\!\!\!\perp S \mid V, Z, J \quad d = 0, 1. \tag{CI-S}$$

Also, for given V and J the latent outcomes are independent of Z,

$$Y^d \perp\!\!\!\perp Z \mid V, J \quad d = 0, 1. \tag{CI-Z}$$

[1] Consider two types $v_1 \leq v_2$. Under the threshold model v_1 participates if v_2 participates. This is independent of the shape of the propensity score function. In particular, monotonicity of the propensity score function in its parameters is not required.

Intuitively, these assumptions state that once the unobserved type is controlled for, the instruments are uninformative about latent outcomes. Note that we do not place any restrictions on the joint distribution of potential outcomes and V. Economically this means that unobserved characteristics such as personal taste that enter into the decision to participate in the treatment are allowed to be correlated with the latent outcomes. The more commonly assumed instrument condition is

$$(Y^0, Y^1, V) \perp\!\!\!\perp (S, Z) \mid J$$

which implies the conditional independence assumptions stated above. To argue the validity of an instrument it is helpful to split up the instrument condition in a way that allows us to disentangle participation and outcome effects. In our application, for example, assumptions CI-S and CI-Z seem quite plausible. The problematic assumption is to assume that the variation in treatment participation induced by the instrument is independent of the variation that is driven by the unobserved types.

Throughout, we let E_z and $\mathrm{E}_{z,j}$ denote the expectation operator conditional on $Z = z$, and $(J, Z) = (j, z)$, respectively.

3. Testing approach

Overidentification result

For $z = 0, 1$ and $j = 1, \ldots, J^{\max}$ define $m_{z,j}(x) = \mathrm{E}_{z,j}[Y \mid r_{z,j}(S) = x]$. The propensity score is identified from

$$r_{z,j}(s) = \mathrm{E}_{z,j}[D \mid S = s].$$

and therefore $m_{z,j}$ is identified on the interior of the support of $r_{z,j}(S) \mid Z = z$. Our test is based on the following overidentification result.

Proposition 1 (Overidentification): *Fix $j \in \{1, \ldots, J^{max}\}$ and suppose that conditional on $J = j$ x lies in the interior of the support of both $r_{0,j}(S) \mid Z = 0$ and $r_{1,j}(S) \mid Z = 1$. Then $m_{z,j}$ does not depend on z, i.e., $m_{0,j}(x) = m_{1,j}(x)$. Let $m_j(x)$ denote the common value for all j and x that satisfy the assumption.*

Proof

$$\begin{aligned}
m_{z,j}(x) =& \mathrm{E}[Y \mid r_{z,j}(S) = x, Z = z, J = j] \\
=& (1 - x)\,\mathrm{E}[Y^0 \mid r_{z,j}(S) = x, V > x, Z = z, J = j] \\
&+ x\,\mathrm{E}[Y^1 \mid r_{z,j}(S) = x, V \leq x, Z = z, J = j] \\
=& (1 - x)\,\mathrm{E}[Y^0 \mid V > x, Z = z, J = j] + x\,\mathrm{E}[Y^1 \mid V \leq x, Z = z, J = j] \\
=& (1 - x)\,\mathrm{E}[Y^0 \mid V > x, J = j] + x\,\mathrm{E}[Y^1 \mid V \leq x, J = j]
\end{aligned}$$

Now note that the right hand side does not depend on z. □

The result says that under the null hypothesis that the model is correctly specified the parameter m_j can be identified from two different subpopulations. Under alternatives the instruments have a direct effect on outcomes that is not mediated through the propensity score. The overidentification restriction has some power to detect such alternatives because in the two subpopulations distinct values of the instrument vector are used to identify the same parameter.

Suppose that for $j = 1, \ldots, J^{\max}$ there are $\underline{x}_j$ and $\bar{x}_j$, $\underline{x}_j \leq \bar{x}_j$, and open sets $\mathcal{G}_j$ such that

$$\operatorname{supp} r_{0,j}(S) \mid Z = 0, J = j \cap \operatorname{supp} r_{1,j}(S) \mid Z = 1, J = j \supseteq \mathcal{G}_j \supseteq [\underline{x}_j \bar{x}_j].$$

Proposition 1 implies that on $[\underline{x}_j \bar{x}_j]$ we have

$$m_{0,j}(x) - m_{1,j}(x) = 0. \tag{2}$$

We are testing this equality. For the test to have some bite we need $[\underline{x}_j \bar{x}_j]$ to be non-empty. Intuitively, what is required is that for fixed Z the continuous instrument is strong enough to induce as many individuals to change their treatment status as would be swayed to change their participation decision by a change in Z while keeping S fixed. An important case where this is not possible is if Z is a deterministic function of S.

The basic idea of the overidentification result does not rely on the continuity of S. However, continuity of S is crucial as it offers a way to ensure that the common support of the propensity scores in the two subpopulations with $Z = 0$ and $Z = 1$ can plausibly have positive probability. For a given j we refer to an interval $[\underline{x}_j, \bar{x}_j]$ that satisfies the above condition as a testable subpopulation. It consists of a set of unobserved types that can be induced to select in and out of treatment by marginal changes in the continuous instrument regardless of the value of the binary instrument. Therefore the types in this interval are part of the complier population as defined in Angrist, Imbens, and Rubin 1996.

Proposition 1 is implied by the stronger result

$$\mathrm{E}_j[Y \mid S, Z] = \mathrm{E}_j[Y \mid r_{Z,j}(S)] \quad a.s.. \tag{3}$$

This says that conditional on covariates, the propensity score aggregates all information that the instruments provide about observed outcomes. In that sense, our approach can be interpreted as a test of index sufficiency that is similar in spirit to the test of the validity of the matching approach suggested in Heckman et al. 1996; Heckman et al. 1998. The equivalence (3) remains true if Y is replaced by a measurable function of Y. By considering different functions of Y a whole host of testable restrictions can be generated. One implication, for example, is that a conditional distribution function is overidentified. In this paper we only consider overidentified conditional mean outcomes and leave the obvious extensions to future research. Our testable restriction (2) is closely related to the marginal treatment effect (MTE)

$$\beta_j(x) = \mathrm{E}_j[Y^1 - Y^0 \mid V = x]$$

which has been proposed as a natural way to parameterize a heterogeneous treatment model (Heckman and Vytlacil 2005). In fact, $\beta_j(x) = \partial_x m_j(x)$. Since we are testing for

overidentification of a function, we are also testing for overidentification of its derivative. If we were to base our test directly on the MTE instead of mean outcomes we would not be able to detect alternatives where instruments are uncorrelated with the treatment effect β but have a direct effect on the base outcome α. Another advantage of our mean outcome approach over a test based on the MTE is that we avoid having to estimate a derivative. In our nonparametric setting derivatives are much harder to estimate than conditional means. However, if the econometrician is not interested in a direct effect on the base outcome and if a large sample is available it might be beneficial to look at β_j rather than at m_j. The reason is that as m_j is a smoothed version of β_j it might not provide good evidence for perturbations of β_j that oscillate around zero. Another maybe more compelling reason to consider overidentification of β_j is that it allows us to investigate the source of a rejection of the null hypothesis. If a test based on m_j rejects while at the same time a test based on β_j does not reject it seems likely that instruments have a direct effect on the base outcome but not on the treatment effect. In this paper we focus on the test based on conditional outcomes and leave a test considering the MTE to future research.

It is helpful to think of alternatives as violations of the index sufficiency condition (3). Economically this means that instruments have a direct effect on outcomes, i.e., instruments have an effect on observed outcomes that can not be squared with their role as providers of independent variation in the participation stage. To formalize how our test detects such alternatives ignore covariates for the moment and define the prediction error from regressing on the propensity score instead of on the instruments

$$\varphi(S,Z) = \mathrm{E}[Y \mid S,Z] - \mathrm{E}[Y \mid r_Z(S)].$$

Now suppose that the model is correctly specified up to possibly a violation of the index sufficiency condition. The restricted null hypothesis is

$$H_0 : \varphi(S,Z) = 0 \quad a.s..$$

Using this notation we can rewrite the testable restriction (2) as

$$\mathrm{E}[\varphi(S,Z) \mid r_0(S) = x, Z = 0] - \mathrm{E}[\varphi(S,Z) \mid r_1(S) = x, Z = 1] = 0$$

for all $x \in [\underline{x}, \bar{x}]$. This is a necessary condition for

$$\mathrm{E}[\varphi(S,Z) \mid r_z(S) = x, Z = z] = 0 \quad \text{for } z = 0,1 \text{ and } x \in \operatorname{supp} r_z(S) \mid Z = z$$

which in turn is necessary for the restricted null. Since we are only testing a necessary condition not all alternatives can be detected. As an extreme case consider the case of identical propensity scores, i.e., $r_0 = r_1$. In this particular case our testable restriction does not have the power to detect a direct effect of S on outcomes.

Parameter estimation and test statistic

Let $\hat{m}_{z,j}$ denote an estimator of $m_{z,j}$ and let $\underline{x} = (\underline{x}_1, \ldots, \underline{x}_{J^{\max}})$ and $\bar{x} = (\bar{x}_1, \ldots, \bar{x}_{J^{\max}})$. Suppose that under the null hypothesis m_j is overidentified on $[\underline{x}_j, \bar{x}_j]$ for $j = 1, \ldots, J^{\max}$

and define the test statistic

$$T_n = T_n(\underline{x}, \bar{x}) = \sum_{j=1}^{J} \int_{\underline{x}_j}^{\bar{x}_j} (\hat{m}_{0,j}(x) - \hat{m}_{1,j}(x))^2 \pi_j(x)\, dx. \tag{4}$$

Here π_j is a weight function that can be used to fine-tune power against certain alternatives. What constitutes a sensible choice for π_j will depend on the specifics of the application. For simplicity we assume that π_j is unity from here on. In the following we will refer to the subsample with $J_i = j$ and $Z_i = z$ as the (j, z)-cell. We estimate $\hat{m}_{z,j}$ by a two step procedure. In the first step we estimate the function $r_{z,j}$ by local polynomial regression of S on D within the (j, z)-cell. We will refer to this step as the participation regression. The first step estimator is denoted by $\hat{r}_{z,j}$. In the second step we estimate $m_{z,j}$ by local linear regression of Y on the predicted regressors $\hat{r}_{z,j}(S_i)$ within the (j, z)-cell. This step will be referred to as outcome regression. We let L and K denote the kernel functions for the participation and outcome regression, respectively. Also let g and h denote the respective bandwidth sequences. To reduce notational clutter, we assume that the bandwidths do not depend on j and z. It is straightforward to extend the model to allow cell dependent bandwidths. Let q denote the degree of the local polynomial in the participation regression. It is necessary to choose $q \geq 2$ to remove troublesome bias terms. If these bias terms are not removed the test will behave asymptotically like a linear test, i.e., it will favor the rejection of alternatives that point into a certain direction. A formal definition of the estimators is provided in Appendix A.

In many applications the bounds $\underline{x}$ and $\bar{x}$ are not a priori known and have to be estimated. Below we show that replacing the bounds by a consistent estimator does not affect the asymptotic distribution of the test statistic under weak assumptions. Since we assume $r_{z,j}$ to be continuous, the set on which m_j is overidentified will always be an interval $(x_{L,j}, x_{U,j})$. To avoid boundary problems we fix some positive c_δ and estimate the smaller interval $[\underline{x}_j, \bar{x}_j] = [x_{L,j} + c_\delta, x_{U,j} - c_\delta]$ by its sample equivalent.

Inference and bootstrap method

In Proposition 2 below we characterize the asymptotic distribution of the test statistic under the null. However, as we explain below, we do not recommend to use this distributional result as a basis for approximating critical values. In a related problem with nonparametrically generated regressors Y. Lee 2013 establishes the validity of a multiplier bootstrap procedure. We conjecture that, building on the asymptotic influence function from Lemma 3 in the appendix, a similar approach can be taken in our setting. However, simulating the distribution by multiplier methods has some disadvantages. First, as the approach is based on asymptotic influence functions no improvements beyond first order asymptotics can be expected. Secondly, the method requires significant coding effort which makes it unattractive in applied work. This is why we propose a wild bootstrap procedure that is straightforward to implement instead. We provide simulation evidence that illustrates that the procedure can have good properties in small and medium sized samples. A theoretical proof of the validity of the method is beyond the scope of the present paper and left to future research.

First, estimate the bounds $\underline{x}$ and $\bar{x}$. In the bootstrap samples these bounds can be taken as given. For all j and all z estimate $\hat{r}_{z,j}$ from the (j,z)-cell and predict $R_i^0 = \hat{r}_{Z_i,J_i}(S_i)$ and $\hat{\zeta}_i^0 = D_i - R_i^0$. Next, pool all observations with $J = j$ and estimate m_j by local linear regression of Y_i on R_i^0 with kernel K and bandwidth h. Predict $M_i^0 = \hat{m}_{J_i}(R_i^0)$ and $\hat{\epsilon}_i^0 = Y_i - M_i^0$. Now generate B bootstrap samples in the following way. Draw a sample of n independent Rademacher random variables $(W_i)_{i\leq n}$, let

$$\begin{pmatrix} D_i^* \\ Y_i^* \end{pmatrix} = \begin{pmatrix} R_i^0 \\ M_i^0 \end{pmatrix} + W_i \begin{pmatrix} \hat{\zeta}_i^0 \\ \hat{\epsilon}_i^0 \end{pmatrix},$$

and define the bootstrap sample $(Y_i^*, D_i^*, S_i, Z_i, J_i)_{i\leq n}$.

While we use Rademacher variables as an auxiliary distribution, other choices such such as the two-point distribution from Mammen 1993 or a standard normal distribution are also possible.

4. Asymptotic analysis

In this section we derive the asymptotic distribution of our test statistic. This analysis gives rise to a number of interesting insights. First, it allows us to consider local alternatives. A lesson implicit in the existing literature on L^2-type test statistics is that a naive construction of such a statistic often leads to a test with the undesirable property of treating different local alternatives disparately. Loosely speaking, such a tests behaves like a linear test in that it only looks for alternatives that point to the same direction as a certain bias term (cf. Härdle and Mammen 1993). We find that in order to avoid such behavior it suffices to employ bias-reducing methods when estimating the propensity scores. We recommend to fit a local polynomial of at least quadratic degree. The outcome estimation does not contribute to the problematic bias term. Secondly, our analysis allows us to consider the case when the bounds of integration $\underline{x}$ and $\bar{x}$ are unknown and have to be estimated. We show that, provided that the estimators satisfy a very weak assumption, the asymptotic distribution is unaffected by the estimation. Thirdly, our results allow us to make recommendations about the choice of the smoothing parameters. Our main asymptotic result implies that our test has good power against a large class of local alternatives if the outcome stage estimator oversmoothes compared to the participation stage estimator but not by too much. For convenience of notation, in the following we focus on the case $J^{\max} = 1$ and omit the j subscript. Proofs for the results in this section can be found in the appendix.

Assumptions

Define the sampling errors $\epsilon = Y - \mathrm{E}[Y \mid r_Z(S)]$ and $\zeta = D - E[D \mid S, Z]$. Under the null hypothesis the conditional variances $\sigma_\epsilon^2(x) = \mathrm{E}[\epsilon^2 \mid r_Z(S) = x]$, $\sigma_\zeta^2(x) = \mathrm{E}[\zeta^2 \mid r_Z(S) = x]$ and $\sigma_{\epsilon\zeta}(x) = \mathrm{E}[\epsilon\zeta \mid r_Z(S) = x]$ remain unchanged if the unconditional expectation operator is replaced by the conditional expectation operator E_z, $z = 0, 1$. Also note that $\sigma_\zeta^2(x) = x(1-x)$. For our local estimation approach to work we have to impose

some smoothness on the functions m_z and r_z. We now give conditions in terms of the primitives of the model to ensure that the functions that we are estimating are sufficiently smooth.

Assumption 1: *Assume that m is overidentified on an open interval (x_L, x_U) and*

(i) there is a positive ρ such that

$$\mathrm{E}[\exp(\rho\left|Y^d\right|)] < \infty, \quad d = 0, 1.$$

(ii) Conditional on $Z = z$, $z = 0, 1$, S is continuously distributed with density $f_{S|Z=z}$ and $r_z(S)$ is continuously distributed with density $f_{R|Z=z}$. Moreover, $f_{S|Z=z}$ is bounded away from zero and has one bounded derivatives and $f_{R|Z=z}$ is bounded away from zero and is twice continuously differentiable.

(iii) $\mathrm{E}[Y^0 \mid V > x]$ and $\mathrm{E}[Y^1 \mid V \leq x]$ are twice continuously differentiable on (x_L, x_U).

(iv) The functions $\mathrm{E}[(Y^0)^2 \mid V > x]$ and $\mathrm{E}[(Y^1)^2 \mid V \leq x]$ are continuous on (x_L, x_U).

(v) r_z, $z = 0, 1$, is $(q+1)$-times continuously differentiable on (x_L, x_U).

The assumption implies standard regularity conditions for m, σ_ϵ^2 and $\sigma_{\epsilon\zeta}$ that are summarized in Assumption 3 in the appendix. These conditions include that m is twice continuously differentiable and that σ_ϵ^2 and $\sigma_{\epsilon\zeta}$ are continuous. A consequence of Assumption 1(ii) is that x_L and x_U are identified by

$$\begin{aligned} x_L &= \max\left\{\inf_s r_0(s), \inf_s r_1(s)\right\} \quad \text{and} \\ x_U &= \min\left\{\sup_s r_0(s), \sup_s r_1(s)\right\}. \end{aligned} \tag{5}$$

Fix a small constant $c_\delta > 0$. We can choose $\underline{x} = x_L + c_\delta$ and $\bar{x} = x_U - c_\delta$. We also need some assumptions about the kernel functions.

Assumption 2: *K and L are symmetric probability density functions with bounded support. K has two bounded and continuous derivatives. The bandwidth sequences are parametrized by $g \sim n^{-\eta^*}$ and $h \sim n^{-\eta}$.*

Implicit in this assumption is that the bandwidths are not allowed to depend on z. In particular, the bandwiths are tied to the overall sample size rather than the size of the two subsamples corresponding to $Z = z$, $z = 0, 1$. This is for expositional convenience only.

Local alternatives

To investigate the behavior of the test under local alternatives we now consider a sequence of models that converges to a model in the null hypothesis.

Definition 1 (Local alternative): *A sequence of local alternatives is a sequence of models*

$$\mathcal{M}^n = (Y^{0,n}, Y^{1,n}, V^n, S, Z, r_0, r_1)$$

in the alternative that converges to a model

$$\mathcal{M}^{null} = (Y^{0,null}, Y^{1,null}, V^{null}, S, Z, r_0, r_1)$$

in the null hypothesis in the following sense:

$$\sup_x \mathrm{E}\left[\left(1_{\{V^n \le x\}} - 1_{\{V^{null} \le x\}}\right)^2 \mid S, Z\right] = O_{a.s}\left(c_n^2\right) \tag{6a}$$

$$\mathrm{E}\left[\left(Y^{d,n} - Y^{0,null}\right)^2 \mid S, Z\right] = O_{a.s}\left(c_n^2\right) \quad d = 0, 1 \tag{6b}$$

for a vanishing sequence c_n. For n large enough there are positive constants ρ and C such that

$$\mathrm{E}[\exp(\rho\left|Y^{d,n} - \mathrm{E}[Y^{d,n} \mid S, Z]\right|) \mid S, Z] \le C \quad d = 0, 1.$$

We let Y^n and Y^{null} denote the observed outcome under the model $\mathcal{M}^n$ and $\mathcal{M}^{null}$, respectively.

Write φ_n for the index prediction error under the sequence of models $\mathcal{M}^n$ and note that

$$\begin{aligned}\varphi_n(S, Z) &= \mathrm{E}[Y^n \mid S, Z] - \mathrm{E}[Y^n \mid r_Z(S)] \\ &= \mathrm{E}[Y^n - Y^{\text{null}} \mid S, Z] - \mathrm{E}[Y^n - Y^{\text{null}} \mid r_Z(S)] = O_{a.s}(c_n)\end{aligned}$$

so that index sufficiency holds approximately in large samples. Formally, we are testing the sequence of local alternatives

$$H_{0,n} : \Delta_n(x) = 0 \quad \text{for } x \in [\underline{x}, \bar{x}]$$

with

$$\Delta_n(x) = \mathrm{E}[\varphi_n(S, Z) \mid r_Z(S) = x, Z = 0] - \mathrm{E}[\varphi_n(S, Z) \mid r_Z(S) = x, Z = 1].$$

To analyze the behavior of our test under local alternatives we suppose that we are observing a sequence of samples where the n-th sample is drawn from $\mathcal{M}^n$. For vanishing c_n we interpret $\mathcal{M}^{\text{null}}$ as a hypothetical data generating process that satisfies the restriction of the null and that is very close to the observed model $\mathcal{M}^n$. Our objective is to show that our test can distinguish $\mathcal{M}^n$ from $\mathcal{M}^{\text{null}}$. The fastest rate at which local alternatives can be detected is $c_n = n^{-1/2}h^{-1/4}$. This is the standard rate for this type of problem (cf. Härdle and Mammen 1993). At this rate the smoothed and scaled version of the local alternative

$$\Delta_{K,h}(x) = c_n^{-1} \int \Delta_n(x + ht) K(t)\, dt$$

enters the asymptotic distribution of the test statistic.

Asymptotic behavior of the test statistic

Our main asymptotic result states that, provided that c_n vanishes fast enough, the asymptotic distribution of the test statistic can be described by the asymptotic distribution of the statistic under the hypothetical model $\mathcal{M}^{\text{null}}$ shifted by a deterministic sequence that measures the distance of the observed model from $\mathcal{M}^{\text{null}}$. The behavior of the test statistic under the null is obtained as a special case by choosing a trivial sequence of local alternatives.

Proposition 2: *Let $c_n = n^{-1/2}h^{-1/4}$ and consider a model $\mathcal{M}^{null}$ satisfying Assumption 1 for $x_L < \underline{x} < \bar{x} < x_U$ and corresponding local alternatives $\mathcal{M}^n$ satisfying Definition 1. Assume that the functions $\mathrm{E}[Y^n \mid r_Z(S) = x]$ and $\mathrm{E}[Y^n \mid r_Z(S) = x, Z = z]$, $z = 0, 1$, are Riemann integrable on (x_L, x_U). The bandwidth parameters η and η^* satisfy*

$$3\eta + 2\eta^* < 1 \quad (7a) \qquad \eta > 1/6 \quad (7d)$$
$$2\eta > \eta^* \quad (7b) \qquad (q+1)\eta^* > 1/2 \quad (7e)$$
$$\eta^* + \eta < 1/2 \quad (7c) \qquad \eta^* > \eta\,. \quad (7f)$$

Then

$$n\sqrt{h}T_n - \frac{1}{\sqrt{h}}\gamma_n - \int_{\underline{x}}^{\bar{x}} \Delta_{K,h}^2(x)\,dx \xrightarrow{d} N(0, V)\,,$$

where

$$V = 2K^{(4)}(0)\int_{\underline{x}}^{\bar{x}} \left[x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x) + \sigma_\epsilon^2(x)\right]^2 \left(\sum_{z\in\{0,1\}} \frac{1}{p_z f_{R,z}(x)}\right)^2 dx$$

and γ_n is a deterministic sequence such that $\gamma_n \to \gamma$ for

$$\gamma = K^{(2)}(0)\int_{\underline{x}}^{\bar{x}} \left[x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x) + \sigma_\epsilon^2(x)\right] \sum_{z\in\{0,1\}} \frac{1}{p_z f_{R,z}(x)}\,dx.$$

Here $m(x) = E[Y^{null} \mid r_Z(S) = x]$ and all conditional moments are computed under $\mathcal{M}^{null}$. $K^{(v)}$ denotes the v-fold convolution product of K. For $q \geq 2$ the set of admissible bandwidths is non-empty.

The result implies that the test can detect local alternatives that converge to a model in the null hypothesis at the rate $c_n = n^{-1/2}h^{-1/4}$ and that satisfy

$$\liminf_n \int_{\underline{x}}^{\bar{x}} \Delta_{K,h}^2(x)\,dx > 0.$$

Both the first and the second stage estimation contribute to the asymptotic variance. The term $x(1-x)m'(x)^2 - 2\sigma_{\epsilon\zeta}(x)m'(x)$ in the expression for the asymptotic variance is due to the first stage estimation. Under our assumptions this term can not be signed, so that the first stage estimation might increase or decrease the asymptotic variance. However, while it is possible to construct models under which this term is negative, these models have some rather unintuitive features and we do not consider them to be

typical. If the estimated regression function is rather flat, the influence of the first stage regression on the asymptotic variance is small. To gain an intuition as to why this is so, note that if $m'(x)$ is small then a large interval of index values around x is informative about $m(x)$. This helps to reduce the first stage estimation error, because on average the index is estimated more reliably over large intervals than over smaller intervals.

An essential ingredient in the proof of Proposition 2 is a result from Mammen, Rothe, and Schienle 2012. They provide a stochastic expansion of a local linear smoother that regresses on generated regressors around the oracle estimator. The oracle estimator is the infeasible estimator that regresses on the true instead of the estimated regressors. This expansion allows us to additively separate the respective contributions of the participation and the outcome regression to the overall bias of our estimator of $m_0 - m_1$. Under the null the oracle estimator is free of bias. This is intuitive. Under the null $m = m_0 = m_1$ so that $\hat{m}_0$ and $\hat{m}_1$ estimate the same function in two subpopulations with non-identical designs. A well-known property of the local linear estimator is that its bias is design independent (Ruppert and Wand 1994) which makes it attractive for testing problems that compare nonparametric fits (Gørgens 2002). Hence, only the bias of the participation regression has to be reduced.

We do not recommend using the distributional result in Proposition 2 to compute critical values. The exact shape of the distribution is very sensitive to bandwidth choice. As explained below, one does not know in practice if bandwidths satisfy the conditions in the theorem. Even if bandwidths are chosen incorrectly, in many cases the statistic still converges to a normal and most of the lessons we draw from the asymptotic analysis still hold up. However, the expressions for the asymptotic bias and variance would look different. Furthermore, to estimate the asymptotic bias and variance we have to estimate derivatives and conditional variances. These are quantities that are notoriously difficult to estimate. Instead, our inference is based on the wild bootstrap procedure introduced in Section 3. We investigate the validity of our bootstrap procedure in simulations in Section 5 below.

Proposition 2 requires that the bandwidth parameters satisfy a system of inequalities. The restrictions are satisfied for example if $q = 2$, $\eta^* = 1/5$ and $1/6 < \eta < 1/5$. The inequalities (7a)-(7c) ensure that our estimators satisfy the assumptions of Theorem 1 in Mammen, Rothe, and Schienle 2012. Condition (7d) ensures that up to parametric order the bias of the oracle estimator is design independent. When the inequality (7f) is satisfied, the error terms from both the participation and outcome regression contribute to the asymptotic distribution. Finally, inequality (7e) says that the bias from the participation regression must vanish at a faster than parametric rate. This is precisely the condition needed to get rid of the troublesome bias terms discussed above. While the proposition offers conditions on the rates at which the bandwidths should vanish it offers little guidance on how to choose the bandwidths in finite samples. There are no bandwidth selection procedures that produce deliberately under- or oversmoothing bandwidths. This problem is by no means specific to our model but on the contrary quite ubiquitous in the kernel smoothing literature (cf. Hall and Horowitz 2012). In our application we circumvent the problem of bandwidth selection by reporting results for a large range of bandwidth choices.

In practice, the bounds of integration $\underline{x}$ and $\bar{x}$ are additional parameters that have to be chosen. In most applications this means that they have to be estimated from the data. The following result states that a rather slow rate of convergence of these estimated bounds suffices to ensure that bound estimation does not affect the asymptotic distribution.

Proposition 3: *Suppose that the assumptions of Proposition 2 hold. Assume also that* $\underline{x}_n$ *and* $\bar{x}_n$ *are sequences of random variables such that*

$$(\underline{x}_n, \bar{x}_n) - (\underline{x}, \bar{x}) = o_p\left(h^\ell\right)$$

for a constant $\ell > 1/2$. *Then*

$$T_n(\underline{x}_n, \bar{x}_n) - T_n(\underline{x}, \bar{x}) = o_p\left(\frac{1}{n\sqrt{h}}\right).$$

Let $\hat{x}_L$ and $\hat{x}_U$ denote the sample equivalents of the right hand side of the equation (5) that identifies x_L and x_U, respectively. Under the bandwidth restrictions of Proposition 2 the assumptions in Proposition 3 are satisfied if we set $\underline{x}_n = \hat{x}_L + c_\delta$ and $\bar{x}_n = \hat{x}_U - c_\delta$.

5. Simulations

We simulate various versions of the random coefficient model from equation (1) and compute empirical rejection probabilities for our bootstrap test for two sample sizes and a large number of bandwidth choices. As in the previous section we assume $J^{\max} = 1$ and drop the j subscript.

Our basic setup is a model in the null hypothesis. Simulating our test for this model allows us to compare the nominal and empirical size of our test. We then generate several models in the alternative by perturbing outcomes in the basic model for the $Z = 1$ subpopulation. For the basic model we define linear propensity scores $r_0(s) = 0.1 + 0.5s$ and $r_1(s) = 0.5s$. The binary instrument Z is a Bernoulli random variable with $P(Z = 0) = P(Z = 1) = 0.5$ and the continuous instrument S is distributed uniformly on the unit interval. The base outcome α follows a mean-zero normal distribution with variance 0.5. The treatment effect is a deterministic function of V, $\beta = -2V$. As alternatives we consider perturbations of the base outcome α as well as perturbations of the treatment effect β. These perturbations are obtained by adding Δ_α to α and Δ_β to β in the $Z = 1$ subpopulation. The specifications for the alternatives are summarized in Table 1. The first three alternatives consider perturbations of the base outcome, whereas alternatives 4-6 are derived from perturbations of the treatment effect. Alternatives 1 and 4 consider the case that base outcome and treatment effect, respectively, are shifted independently of the unobserved heterogeneity V. The perturbations generating alternatives 2 and 5 are linear functions of V. Finally, alternatives 3 and 6 are generated by perturbing by functions of V that change sign. These alternatives are expected to be particularly hard to detect because our test is based on the m_z function which smoothes over the unobserved heterogeneity as is apparent in the proof of Proposition 1. As bandwidths

alternative	perturbation
1	$\Delta_\alpha = 0.2$
2	$\Delta_\alpha = -\frac{1}{2}V$
3	$\Delta_\alpha = 40(V-0.3)\exp\left(-80(V-0.3)^2\right)$
4	$\Delta_\beta = 0.2$
5	$\Delta_\beta = -V$
6	$\Delta_\beta = 40(V-0.3)\exp\left(-80(V-0.3)^2\right)$

Table 1: Specification of simulated alternatives.

we choose $g = C_g n^{-\frac{1}{5}}$ and $h = C_h n^{-\frac{1}{6}}$. We report results for a number of choices for the constants C_g and C_h. We set $q = 2$ and choose an Epanechnikov kernel for both K and L. The sample size is set to $n = 200, 400$. These should be considered rather small numbers considering the complexity of the problem. We consider the nominal levels $\theta = 0.1, 0.05$ as these are the most commonly used ones in econometric applications. As bound estimation has only a higher order effect we take $\underline{x} = 0.15$ and $\bar{x} = 0.45$ as given. To simulate the bootstrap distribution we are using $B = 999$ bootstrap iterations. For each model we conduct 999 simulations. Empirical rejection probabilities are reported in Table 2 for $n = 400$ and in Table 3 in the appendix for $n = 200$.

We discuss only the results for $n = 400$ in detail. Under the null hypothesis the empirical rejection probabilities are very close to the nominal levels. While this is not conclusive evidence that our bootstrap approach will always work, it is suggestive of the validity of the procedure.

Alternative 1 and Alternative 2 are detected with high probability. These alternatives are particularly easy to detect for two reasons. First, the perturbation affects a large subpopulation so that the alternative is easy to detect due to abundance of relevant data. Secondly, the smoothing inherent in the quantities that our test considers does not smear out the perturbations in a way that makes the alternatives hard to detect. To understand the first effect contrast Alternative 1 and Alternative 2 with Alternative 4 and Alternative 5. Both pairs of alternatives arise from similar perturbations. However, the whole subsample with $Z = 1$ can be used to detect the first pair. In contrast, only treated individuals in the $Z = 1$ subsample provide data that helps to detect the second pair. A back-of-the-envelope calculation reveals that on average only about $400 \times 1/2 \times 1/4 = 25$ observations fall into the subsample with $Z = 1$ and $D = 1$. As cell sizes are observed in applications, a lack of relevant data is a problem that can readily be accounted for when interpreting test results. To shed light on the second effect recall that m_z is derived from smoothing outcomes over $V \leq x$ and $V > x$. Therefore, if a perturbation changes sign, positive and negative deviations from the null will cancel each other out. This effect is precisely what makes it so hard to detect perturbations such as those underlying Alternative 3 and Alternative 6. Luckily, these kinds of alternatives are not what should be expected in many applications. The problem that applied researchers have in mind most of the time is that instruments might have a direct effect on outcomes that can

	$\theta = 0.10$						$\theta = 0.05$					
C_h	0.50	0.75	1.00	1.25	1.50	1.75	0.50	0.75	1.00	1.25	1.50	1.75
null												
$C_g = 0.50$	9.3	8.9	8.4	7.7	8.6	9.6	4.2	3.4	4.7	4.2	4.1	4.1
$C_g = 0.75$	10.1	9.9	9.4	8.2	7.7	9.3	4.8	4.5	4.0	3.3	3.6	4.0
$C_g = 1.00$	8.9	8.7	7.4	9.0	8.9	8.1	4.2	4.1	3.2	4.0	3.6	3.3
alternative 1												
$C_g = 0.50$	94.3	93.8	93.7	93.6	92.8	94.7	88.5	87.1	87.2	86.7	87.7	88.4
$C_g = 0.75$	94.8	91.9	93.0	92.6	94.0	93.8	88.6	86.9	87.2	85.8	87.3	87.2
$C_g = 1.00$	94.0	93.4	94.8	93.5	93.8	93.3	86.7	88.4	89.6	87.2	87.2	89.3
alternative 2												
$C_g = 0.50$	96.9	97.5	97.5	98.1	98.6	98.0	93.3	94.4	95.4	96.0	96.4	95.4
$C_g = 0.75$	96.9	97.9	97.2	97.8	97.1	97.5	93.0	95.6	94.6	94.7	94.3	95.3
$C_g = 1.00$	97.7	97.2	97.4	97.8	97.4	97.8	94.5	95.3	94.1	94.1	95.3	94.4
alternative 3												
$C_g = 0.50$	8.3	8.7	7.2	9.3	8.7	8.9	3.4	3.6	3.5	4.6	4.0	4.0
$C_g = 0.75$	6.9	9.1	8.9	8.6	8.9	9.3	3.5	4.4	3.6	3.6	4.0	3.9
$C_g = 1.00$	8.3	8.2	7.9	8.8	8.9	8.7	4.0	3.7	3.7	3.7	4.6	3.9
alternative 4												
$C_g = 0.50$	25.5	23.8	22.9	24.2	22.6	22.7	15.1	13.9	13.5	13.8	12.3	13.3
$C_g = 0.75$	25.1	26.3	26.1	22.3	23.3	24.7	15.0	14.6	15.0	13.1	13.3	14.5
$C_g = 1.00$	25.4	23.5	24.5	23.7	23.7	23.6	15.2	13.0	15.6	13.8	14.1	13.8
alternative 5												
$C_g = 0.50$	24.3	22.5	21.5	22.7	22.8	21.8	14.9	12.9	11.9	13.8	12.0	12.4
$C_g = 0.75$	21.1	22.0	21.3	20.9	22.7	22.3	10.4	10.8	12.1	11.5	12.7	12.5
$C_g = 1.00$	21.8	21.5	21.4	23.7	21.9	22.5	13.1	12.0	11.3	12.7	12.6	12.2
alternative 6												
$C_g = 0.50$	45.1	44.3	42.3	45.2	46.6	47.7	30.9	30.7	29.3	31.2	35.0	31.2
$C_g = 0.75$	45.3	43.5	44.9	44.2	45.8	44.2	32.3	31.4	32.0	30.7	33.0	30.3
$C_g = 1.00$	44.2	45.5	47.6	44.3	44.6	46.6	32.6	33.7	34.0	30.6	30.9	33.9

Table 2: Empirical rejection probabilities in percentage points under nominal level θ. Sample size is $n = 400$.

readily be signed by considering the economic context. In that respect, Alternative 1 and Alternative 2 are more typical of issues that applied economists worry about than Alternative 3.

It might seem puzzling that Alternative 6 is detected much more frequently than Alternative 3. The reason is that in Alternative 3 negative deviations in the $V \leq x$ population are offset by positive deviations in the $V > x$ population. This does not happen in Alternative 6 as only the treated population is affected by the perturbation.

Accounting for the complexity of the problem the sample size $n = 200$, for which we report results in the appendix, should be considered very small. Therefore, it is not surprising that the deviations from the nominal size are slightly more pronounced than in the larger sample. The deviations err on the conservative side, but that might be a particularity of our setup. The pattern in the way alternatives are detected is similar to the $n = 400$ sample with an overall lower detection rate.

Our simulations show that our approach has good empirical properties in finite samples. For the simulated model the test holds its size which indicates that the bootstrap procedure works well. Very particular alternatives that perturb outcomes by a function of the unobserved types that oscillates around zero are difficult to detect by our procedure. Alternatives that we consider to be rather typical are reliably detected provided that the subsample affected by the alternative is large enough.

6. Application

To illustrate the applicability of our method we now consider the effect of teenage childbearing on the mother's probability of graduating from high-school. This topic has been discussed extensively in the literature. An early survey can be found in Hoffman 1998. To deal with the obvious endogeneity of motherhood, many authors (Ribar 1994; Hotz, McElroy, and Sanders 2005; Klepinger, Lundberg, and Plotnick 1995) have used instrumental variables methods. It has been suggested that treatment effect heterogeneity is a reason why estimated effects depend strongly on the choice of instrument (Reinhold 2007). In fact, it is very natural to assume that the effect of motherhood on graduation is heterogeneous. For a simple economic model that generates treatment effect heterogeneity suppose that the time cost of child care is the same for students of different abilities whereas the time cost of studying to improve the odds of graduating is decreasing in ability. To translate the problem into our heterogeneous treatment model let D denote a binary indicator of teenage motherhood and let Y denote a binary indicator of whether the woman has obtained a high school diploma[2]. We consider two instruments from the literature. The first one, henceforth labelled S, is age at first menstrual period which has been used in the studies by Ribar 1994 and Klepinger, Lundberg, and Plotnick 1995. This instrument acts as a random shifter of female fecundity and is continuous in nature. Its validity is discussed briefly in Klepinger, Lundberg, and Plotnick 1995 and Levine and Painter 2003. The second instrument, denoted by Z, is an indicator of whether the

[2]We do not include equivalency degrees (GED's). There is a discussion in the literature as to what the appropriate measure is (cf. Hotz, McElroy, and Sanders 2005).

individual experienced a miscarriage as a teenager. Miscarriage has been used as an unexpected fertility shock in the analysis of adult fertility choices (Miller 2011) and also to study teenage child bearing in Hotz, Mullin, and Sanders 1997; Hotz, McElroy, and Sanders 2005. The population studied in Hotz, McElroy, and Sanders 2005 consists of all women who become pregnant in their teens, whereas we focus on the larger group of all women who are sexually active in their teens. This turns out to be a crucial difference. It stands to investigate the plausibility of the assumptions I-V, CI-S and CI-Z. Arguably, age at first menstrual period is drawn independently of V and fulfills the instrument specific conditional independence assumption CI-S if one controls for race. Possible threats to a linear version of CI-Z are discussed in Hotz, Mullin, and Sanders 1997. Hotz, McElroy, and Sanders 2005 conclude that the linear version of CI-Z holds in good approximation in the population that they are considering. The most problematic assumption to maintain is that Z is orthogonal to V. In a simplified behavioral model teenagers choose to become pregnant based on their unobserved type and then a random draw from nature determines how that pregnancy is resolved. This implies a sort of maximal dependence between Z and V, i.e., teenagers select into treatment and into $Z = 1$ in exactly the same way. Our test substantiates this heuristic argument by rejecting the null hypothesis that the assumptions I-V, CI-S and CI-Z hold simultaneously. Furthermore, it gives instructive insights into the role that heterogeneity plays in the failure of the assumptions.

We use data from the National Longitudinal Survey of Youth 1997[3] (henceforth NLSY97) from round 1 through round 15. We only include respondents who were at least 21 of age at the last interview they participated in. This is to ensure that we capture our outcome variable. A miscarriage is defined as a teenage miscarriage if the woman experiencing the miscarriage was not older than 18 at the time the pregnancy ended. Similarly, a young woman is defined as a teenage mother if she was not older than 18 when the child was born. We control for race for two reasons. First, this is required to make the menarche instrument plausible. Secondly, this takes care of the oversampling of minorities in the NLSY97 so that we are justified in using unweighed estimates. We remove respondents who report "mixed race" as race/ethnicity because the cell size is too small to conduct inference. Table 4 in the appendix gives some summary statistics for our sample. An unfortunate side effect of using the low probability event of a teenage miscarriage as an instrument is that cell sizes can become rather small. This makes it impossible to control for additional covariates while preserving reasonable power. In Section 7 we briefly discuss a model that permits a much larger number of covariates. The estimated propensity scores $\hat{r}_{z,j}$ are plotted in Figure 1. For each j the functions $\hat{r}_{0,j}$ and $\hat{r}_{1,j}$ are not identical almost everywhere and their ranges exhibit

[3]Most of the previous studies relied on data from the National Longitudinal Survey of Youth 1979 (NLSY79). In that study the date of the first menstrual period was asked for for the first time in 1984 when the oldest respondents were 27 years old. As is to be expected, a lot of respondents had trouble recalling the date such a long time after the fact. The NLSY97 contained the relevant question starting from the very first survey when the oldest respondents were still in their teens. Since our method relies on a good measurement of the continuous variable the NLSY97 data is a better choice than the NLSY79 data.

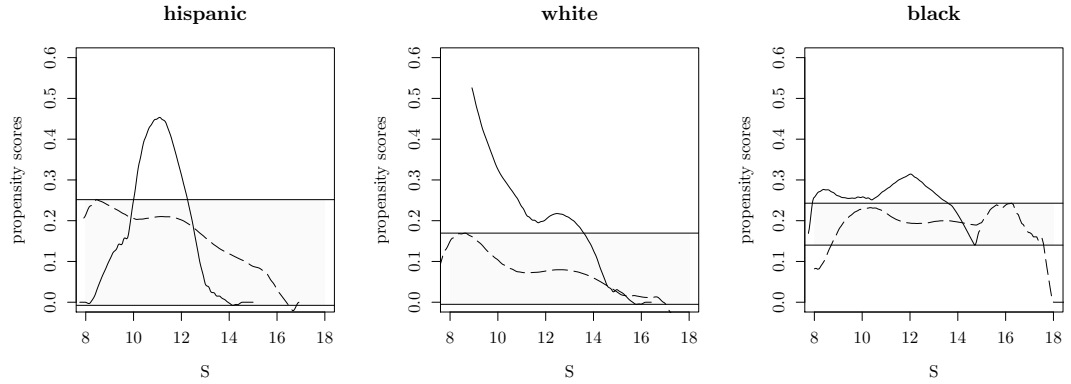

Figure 1: Probability of entering treatment conditional on age of first menstrual period (S) plotted separately for the subpopulations with $Z = 0$ (no miscarriage as a teenager, *dashed line*) and $Z = 1$ (miscarriage as a teenager, *solid line*). Plotted with $q = 1$ and bandwidth $g = 2.00$.

considerable overlap. We require the same properties from their population counterparts to have good power. It should be noted at this point that the shape of the estimated propensity scores is already indicative of the way that miscarriage fails as an instrument. In a naive telling of the story, the propensity score for women who had a teenage miscarriage is shifted upward, contrary to what we observe in Figure 1. Our test rejects if,

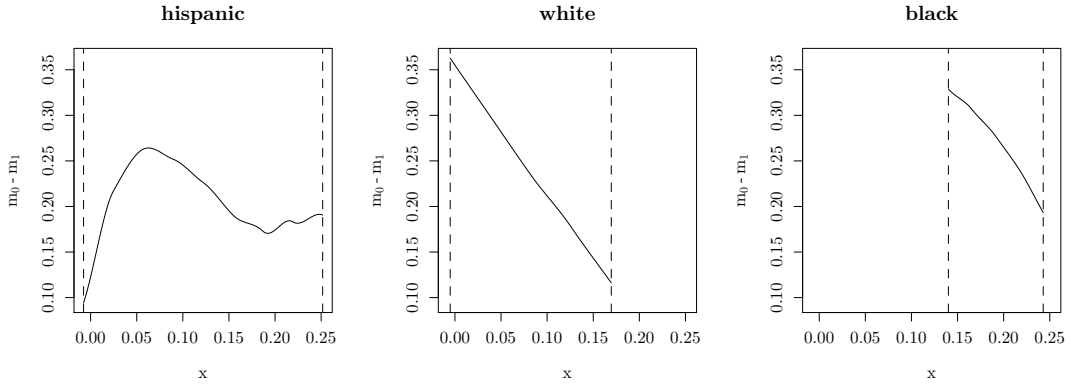

Figure 2: Difference in expected outcomes conditional on probability of treatment between the subpopulations with $Z = 0$ and $Z = 1$. Plotted with $q = 1$ and bandwidths $h = 0.25$ and $g = 2.00$.

keeping the probability of treatment fixed, the difference between the outcomes of the subpopulation with $Z = 0$ and the subpopulation with $Z = 1$ is large. Figure 2 plots $\hat{m}_{0,j}(x) - \hat{m}_{1,j}(x)$ for all values of j. The dashed lines indicate our estimates of $x_{L,j}$ and $x_{U,j}$. We observe that the estimated outcome difference is positive and decreasing in the probability of treatment x. This means that for a low treatment probability x women who have a miscarriage do much worse in terms of high school graduation than

do women who do not have a miscarriage. For larger x, however, this difference in outcomes becomes smaller. This feature is in line with our story-based criticism of the instrument. Suppose that the underlying heterogeneity selects women into pregnancy rather than into motherhood. For concreteness think of the heterogeneity as the amount of unprotected sex that a woman has and suppose that this variable is highly correlated with outcomes. In a Bayesian sense a woman who has a miscarriage reveals herself to be of the type that is prone to have unprotected sex. In that sense she is very similar to women with a high probability of becoming pregnant and carrying the child to term and very different from women who become pregnant only with small probability. To turn this eye-balling of the plots in Figure 2 into a rigorous argument we now take into account sampling error by applying our formal testing procedure. For both the first and the second stage regression we choose an Epanechnikov kernel. To have good power against local alternatives we choose $q = 2$. To keep the problem tractable and to reduce the number of parameters we have to choose, we set $g_j = g$ and $h_j = h$ for all j. We then run the test for a large number of bandwidth choices letting h vary between 0.1 and 0.5 and letting g vary between 1 and 3. To determine the bounds of integration $\underline{x}_j$ and $\bar{x}_j$ we use the naive sample equivalence approach suggested in Section 4 with different values for c_δ. Table 5 in the appendix reports results for $c_\delta = 0.05$ and Table 6 reports results for $c_\delta = 0.075$. For these two choices of c_δ the test rejects at moderate to high significance levels for a large range of smoothing parameter choices.

Our approach can also be used to investigate other instruments that have been suggested in the literature on teen pregnancies. For example, Z or S could be based on local variation in abortion rates or in availability of fertility related health services (cf. Ribar 1994; Klepinger, Lundberg, and Plotnick 1995).

7. Conclusion

So far, inference about heterogeneous treatment effect models mostly relies on theoretical considerations about the relationship between instruments and unobserved individual characteristics that are not investigated empirically. This paper shows that under the assumption that a binary and a continuous instrument are available, a parameter is overidentified. This provides a way to test whether the model is correctly specified. The overidentification result is not merely a theoretical curiosity, it has bite when applied to real data. We illustrate this by applying our method to a dataset on teenage child bearing and high school graduation.

Apart from suggesting a new test, we also contribute to the statistical literature by developing testing theory that with slight modifications can be applied to other settings where index sufficiency holds under the null hypothesis. We accommodate an index that is not observed and enters the test statistic as a nonparametrically generated regressor. This setting is encountered, e.g., when testing the validity of the matching approach along the lines suggested in Heckman et al. 1996 and Heckman et al. 1998. Heckman et al. 1998 employ a parametric first-stage estimator. As a result, their second-stage estimator is, to first order, identical to the oracle estimator. Our analysis suggests that

replacing the parametric first-stage estimator by a non- or semiparametric estimator is not innocuous. In particular, it can affect the second-stage bandwidth choice and the behavior of the test under local alternatives.

A theoretical analysis of our wild bootstrap procedure is beyond the scope of this paper. Developing resampling methods for models with nonparametrically generated regressors is an interesting direction for future research. We hope to corroborate the findings in our exploratory simulations by theoretical results in the future.

To apply our method to a particular data set, additional considerations are necessary. In many applications the validity of an instrument is only plausible provided that a large set of observed variables is controlled for. It is hard to accommodate a rich covariate space in a completely nonparametric model. This is partly due to a curse of dimensionality. Another complicating factor is that our testing approach has good power only if, for fixed covariate values, the instruments provide considerable variation in participation. This is what allows us to test the model for a wide range of unobserved types. Typically, however, instruments become rather weak once the model is endowed with a rich covariate space. These issues can be dealt with by imposing a semiparametric model. As an example, consider the following simple variant of a model suggested in Carneiro and S. Lee 2009. We let X denote a vector of covariates with possibly continuous components and assume that the unobserved type V is independent of X. Treatment status is determined by $D = 1_{\{R \geq V\}}$ with $R = r_1(X) + r_2(S, Z)$. The unobserved type affects the treatment effect and not the base outcome. The observed outcome is

$$Y = \mu_\alpha(X) + D[\mu_\beta(X) + \lambda(V)].$$

The functions r_1, μ_α and μ_β are known up to a finite dimensional parameter. A semiparametric version of our test would compare $\mathrm{E}[D\lambda(V) \mid R = x, Z]$ in the $Z = 0$ and $Z = 1$ subpopulations. The fact that X is uninformative about V and the additive structure allow for an overidentification result that uses variation in X to extend the interval on which a function is overidentified. This contrasts sharply with Proposition 1 which relies on variation in S keeping the value of covariates fixed. In terms of asymptotic rates this semiparametric model with a large covariate space is not harder to estimate than our fully nonparametric model with a small covariate space and there is no curse of dimensionality.

Appendix

A. Definition of estimators

Let $L_g(\cdot) = g^{-1}L(\cdot/g)$ and $K_h(\cdot) = h^{-1}K(\cdot/h)$. For the first-stage estimator set $\hat{r}_{z,j}(s) = a_0$, where a_0 satisfies

$$(a_0, \ldots, a_q) \in \arg\min_{(a_0,\ldots,a_1)\in\mathbb{R}^{q+1}} \sum_{i:Z_i=z,J_i=j} \big(Y_i - a_0 - a_1(S_i - s) - \cdots - a_q(S_i - s)^q\big)^2 L_g(S_i - s).$$

For the second stage estimator set $\hat{m}_{z,j} = b_0$, where b_0 satisfies

$$(b_0, b_1) \in \arg\min_{(b_0,b_1)\in\mathbb{R}^2} \sum_{i:Z_i=z,J_i=j} \big(Y_i - b_0 - b_1(\hat{r}_{z,j}(S_i) - x)\big)^2 K_h(\hat{r}_{z,j}(S_i) - x).$$

B. Proofs

Proof of Proposition 2

The proposition follows from a sequence of lemmas. We first prove that the second-stage regression function and the error terms from the first- and second-stage regressions behave nicely under our assumptions about the primitives of the model.

Assumption 3: *For each* $z \in \{0, 1\}$

(i) m_z *is twice continuously differentiable on* (x_L, x_U).

(ii) there is a positive ρ *such that* $\mathrm{E}_z[\exp(\rho|\zeta|) \mid S]$ *and* $\mathrm{E}_z[\exp(\rho|\epsilon|) \mid S]$ *are bounded,*

(iii) $\sigma^2_{\zeta,z}(x) = \mathrm{E}_z[\zeta^2 \mid r_z(S) = x]$, $\sigma^2_{\epsilon,z}(x) = \mathrm{E}_z[\epsilon^2 \mid r_z(S) = x]$, *and* $\sigma_{\epsilon\zeta,z}(x) = \mathrm{E}_z[\epsilon\zeta \mid r_z(S) = x]$ *are continuous on* (x_L, x_U).

Lemma 1
Assumption 1 is sufficient for Assumption 3.

Proof The lemma follows from plugging in the structural treatment model into the observed quantities and arguing similarly to the proof of Proposition 1. □

In the next lemma we give a complete description of the relevant properties of our first-stage estimator. We provide an explicit expression of a smoothed version of the first-stage estimator that completely characterizes the impact of estimating the regressors on the asymptotic behavior of the test statistic.

Lemma 2 (First stage estimator)
The first stage local polynomial estimator can be written as

$$\hat{r}_z(s) = \rho_n(s) + R_n,$$

where

$$\sup_s |R_n| = O_p\left(g^{q+1}\sqrt{\frac{\log n}{ng}} + \frac{\log n}{ng}\right)$$

and ρ_n *is given explicitly in equation* (8). *Wpa1* ρ_n *is contained in a function class* $\mathcal{R}$ *that for some constant* K, *any* $\xi > \frac{5}{4}\eta^* - \frac{1}{4}$ *and all* $\epsilon > 0$ *can be covered by* $K\exp(n^\xi\epsilon^{-1/2})$ ϵ*-balls with respect to the sup norm. The true propensity score is contained in* $\mathcal{R}$. *Furthermore,*

$$-m'(x)\int K_h(r_z(s) - x)(\hat{r}_z(s) - r_z(s))f_{S|Z=z}(s)\,ds = \frac{1}{n}\sum_{i:Z_i=z}\psi^{(2)}_{n,z,i}(x) + o_p\big(n^{-1/2}\big),$$

with $\psi^{(2)}_{n,z,i}$ *as defined in Lemma 3. Moreover,*

$$\sup_s |\hat{r}_z(s) - r_z(s)| = O_p\left(n^{-\frac{1}{2}(1-\eta^*)}\right).$$

PROOF Throughout, condition on the subsample with $Z = z$. Let $e_1 = (1, 0, \ldots, 0)^\top$ and $\mu(t) = (1, t, \ldots, t^q)^\top$. Furthermore, define

$$\bar{M}_n(s) = \mathrm{E}\,\mu\left(\frac{S_i - s}{g}\right)\mu^\top\left(\frac{S_i - s}{g}\right)L_g(S_i - s).$$

Since we defined g in terms of the total sample size it behaves like a random variable when we work conditionally on the subsample $Z = z$. We have $g = a_n n_z^{-\eta^*} + O_p\left(n^{-\frac{1}{2}-\eta^*}\right)$ for a bounded deterministic sequence a_n. From a straightforward extension of standard arguments for the case of a deterministic bandwidth (c.f. Masry 1996) it can be shown that $\hat{r}_z$ can be written as

$$\hat{r}_z(s) = \rho_n(s) + R_n,$$

where

$$\rho_n(s) = r_z(s) + g^{q+1}b_n(s) + e_1^\top \bar{M}_n^{-1}(s)\frac{1}{n}\sum_i \mu\left(\frac{S_i - s}{g}\right)L_g(S_i - s)\zeta_i, \tag{8}$$

b_n is a bounded function and R_n has the desired order. To show that the desired entropy condition holds, note that $\bar{M}_n$ is a deterministic sequence that is bounded away from zero so that it suffices to derive an entropy bound for the functions

$$\frac{1}{n}\sum_i \mu\left(\frac{S_i - s}{g}\right)L_g(S_i - s)\zeta_i.$$

Wpa1 these functions have a second derivative that is bounded by $\sqrt{n^{-1}g^5 \log n}$. The desired bound on the covering number then follows from a straightforward corollary to Theorem 2.7.1 in Van der Vaart and Wellner 1996. To prove the statement about the smoothed first stage estimator note that under our assumptions we only have to consider the smoothed error term

$$\frac{1}{n}\sum_{i:Z_i=z} \psi_n^*(x, S_i)\zeta_i,$$

where

$$\begin{aligned}\psi_n^*(x, s) &= -m'(x)\int K_h(r_z(u) - x)e_1'\bar{M}_n^{-1}(u)\mu\left(\frac{s-u}{g}\right)L_g(s-u)f_{S|Z=z}(u)\,du \\ &= -m'(x)\int K_h(r_z(s - gu) - x)e_1'\bar{M}_n^{-1}(s - gu)\mu(u)L(u)f_{S|Z=z}(s - gu)\,du.\end{aligned}$$

Since $f_{S|Z=z}$ is bounded and has a bounded derivative there is a function $D_n(s, u)$ bounded uniformly in s, u and x satisfying

$$\bar{M}_n^{-1}(s - ug)f(s - ug) - M^{-1} = gD_n(s, u).$$

By standard kernel smoothing arguments

$$\frac{1}{n_z}\sum_{i:Z_i=z}\left\{\int K_h(r_z(S_i-ug)-x)D_n(S_i,u)\mu(u)L(u)\,du\right\}\zeta_i = O_p\left(\sqrt{\frac{\log n}{nh}}\right).$$

Noting that $L^*(u) = e_1^\top M^{-1}\mu(u)L(u)$ we have

$$\frac{1}{n}\sum_{i:Z_i=z}\psi_n^*(x,S_i)\zeta_i = \frac{1}{n}\sum_{i:Z_i=z}\psi_{n,z,i}^{(2)}(x) + o_p\big(n^{-1/2}\big).$$

□

Next, we give an asymptotic expansion of the integrand in (4) up to parametric order. The result states that the integrand can be characterized by a deterministic function that summarizes the deviation from index sufficiency under the alternative and an asymptotic influence function calculated under the hypothetical model $\mathcal{M}^{\text{null}}$.

Lemma 3 (Expansion)
Uniformly in x

$$\hat{m}_0(x) - \hat{m}_1(x) = \Delta_{K,h}(x) + \frac{1}{n}\sum_i \psi_{n,i}(x) + o_p\big(n^{-1/2}\big)$$

where $\psi_{n,i} = \psi_{n,i}^{(1)} + \psi_{n,i}^{(2)}$ and $\psi_{n,i}^{(j)} = \sum_{z=0,1}\psi_{n,z,i}^{(j)}$, $j = 1, 2$,

$$\psi_{n,z,i}^{(1)}(x) = \frac{1_{\{Z_i=z\}}(-1)^z}{p_z f_{R,z}(x)}K_h(r_z(S_i)-x)\epsilon_i,$$
$$\psi_{n,z,i}^{(2)}(x) = -m'(x)\frac{1_{\{Z_i=z\}}(-1)^z}{p_z f_{R,z}(x)}\int K_h(r_z(S_i-gu)-x)L^*(u)\,du\,\zeta_i.$$

Here $\epsilon_i = Y^{null} - \mathrm{E}[Y^{null} \mid r_Z(S)]$, i.e, ϵ_i is the residual under the hypothetical model $\mathcal{M}^{null}$, and L^* denotes the equivalent kernel of the first step local polynomial regression.

Proof The statement follows from an expansion of $\hat{m}_z$. Work conditionally on the subsample with $Z = z$ and let n_z denote the number of observations in the subsample. To avoid confusion, we write h_n for the second-stage bandwidth, as h will sometimes denote a generic element of a set of bandwidths. Let $h^z = n_z^{-\eta}$. Note that for C large enough h_n is contained in the set

$$\mathcal{H}_{n_z} = \left\{h' : \left|h' - h^z\right| \le C n_z^{-1/2-\eta}\right\}$$

wpa1. Let $e_1 = (1,0)^\top$, $\mu(t) = (1,t)^\top$ and

$$M_h^r(x) = \frac{1}{n}\sum_{i:Z_i=z}\mu\big((r(S_i)-x)/h\big)\mu^\top\big((r(S_i)-x)/h\big)K_h(r(S_i)-x).$$

For arbitrary $\mathbb{R}^n$-valued random variables W define the local linear smoothing operator

$$\mathcal{K}_{h,x,z}^r W = e_1^\top\left(M_h^r(x)\right)^{-1}\frac{1}{n_z}\sum_{i:Z_i=z}W_i\mu\left(\frac{r(S_i)-x}{h}\right)K_h(r(S_i)-x).$$

Decompose the estimator as

$$\begin{aligned}\hat{m}_z(x) =& \mathcal{K}^{\hat{r}}_{h_n,x,z} Y^n + \mathcal{K}^{\hat{r}}_{h_n,x,z} \left\{ (Y^n - Y^{\text{null}}) - \mathrm{E}[Y^n - Y^{\text{null}} \mid S, Z] \right\} \\ &+ \mathcal{K}^{\hat{r}}_{h_n,x,z} \mathrm{E}[Y^n - Y^{\text{null}} \mid S, Z] \\ =& J_1 + J_2 + J_3.\end{aligned}$$

We now proceed to show that

$$\begin{aligned}J_1 &= m(x) + b_{1,n}(x) + \frac{1}{n} \sum_i \left\{ \psi^{(1)}_{n,z,i}(x) + \psi^{(2)}_{n,z,i}(x) \right\} + o_p\big(n^{-1/2}\big), \\ J_2 &= o_p\big(n^{-1/2}\big), \\ J_3 &= b_{2,n}(x) + \int \mathrm{E}[\varphi_n(S, Z) \mid r_Z(S) = x + hr, Z = z] K(r)\, dr + o_p\big(n^{-1/2}\big),\end{aligned}$$

where $b_{j,n}$, $j = 1, 2$, are independent of z and all order symbols hold uniformly in x. For the J_1 term we apply the approach from Mammen, Rothe, and Schienle 2012 (MRS) and expand J_1 around the oracle estimator. Write

$$J_1 = \mathcal{K}^{\hat{r}}_{h_n,x,z} \epsilon_i + \mathcal{K}^{\hat{r}}_{h_n,x,z} m(r_z(S_i)) = J_{1,a} + J_{1,b}.$$

For the $J_{1,a}$ term note that $e_1^\top \left(M^r_h(x) \right)^{-1}$ is stochastically bounded by a uniform over $\mathcal{H}_{n_z}$ version of Lemma 2 in MRS. For ρ_n as defined in Lemma 2 write

$$\begin{aligned}&\frac{1}{n_z} \sum_{i:Z_i=z} K_{h_n}(\hat{r}_z(S_i) - x)\epsilon_i - \frac{1}{n_z} \sum_{i:Z_i=z} K_{h_n}(r_z(S_i) - x)\epsilon_i \\ =& \frac{1}{n_z} \sum_{i:Z_i=z} \left(K_{h_n}(\hat{r}(S_i) - x) - K_{h_n}(\rho_n(S_i) - x) \right) \epsilon_i \\ &+ \frac{1}{n_z} \sum_{i:Z_i=z} \left(K_{h_n}(\rho_n(S_i) - x) - K_{h_n}(r_z(S_i) - x) \right) \epsilon_i = I_1 + I_2.\end{aligned}$$

By the mean-value theorem $I_1 = o_p(n^{-1/2})$. For I_2 note that $\mathrm{E}_z[\epsilon \mid S] = 0$ so that following the arguments in the proof of Lemma 2 in MRS

$$\sup_{x; h \in \mathcal{H}_{n_z}} P\left(\sup_{r_1, r_2 \in \mathcal{R}} \left| \frac{1}{n_z} \sum_{i:Z_i=z} \left(K_h(r_1(S_i) - x) - K_h(r_2(S_i) - x) \right) \epsilon_i \right| > C^* n^{-\kappa_1} \right) \leq \exp(-cn^c),$$

where κ_1 is defined in MRS and C^* is a large constant. To check that $\kappa_1 > 1/2$ note that Theorem 1 in MRS allows bandwidth exponents in an open set so that it suffices to check the conditions for h^z. It is now straightforward to show that a polynomial number of points in $[\underline{x}, \bar{x}] \times \mathcal{H}_{n_z}$ provide a good enough approximation to ensure that

$$\sup_{x, h \in \mathcal{H}_{n_z}, \rho \in \mathcal{R}} \left| \frac{1}{n_z} \sum_{i:Z_i=z} \left(K_h(\rho(S_i) - x) - K_h(r_z(S_i) - x) \right) \epsilon_i \right| = O_p(n^{-\kappa_1})$$

and hence $I_2 = o_p(n^{-1/2})$. Similar arguments apply to

$$\frac{1}{n_z}\sum_{i:Z_i=z}\frac{\hat{r}_z(S_i)-x}{h_n}K_{h_n}(\hat{r}_z(S_i)-x)\epsilon_i.$$

Therefore, $J_{1,a}$ can be replaced by its oracle counterpart at the expense of a remainder term that vanishes at the parametric rate:

$$J_{1,a} = \frac{1}{n}\sum_i \psi^{(1)}_{n,z,i}(x) + o_p(n^{-1/2}).$$

Note that in the last step we also replaced n_z by $p_z n$. Decompose $J_{1,b}$ as in the proof of Theorem 1 in MRS. It is straightforward to extend their results to hold uniformly over bandwidths in $\mathcal{H}_{n_z}$. Deduce that

$$J_{1,b} = m(x) + b_{1,n}(x) - m'(x)\int K_{h_n}(r_z(s)-x)(\hat{r}_z(s)-r_z(s))f_{S|Z=z}(s)\,ds + o_p\big(n^{-1/2}\big).$$

for a sequence of functions $b_{1,n}$ that does not depend on the design. The previous results use standard results about the Bahadur representation of the oracle estimator (cf. Masry 1996; Kong, Linton, and Xia 2010). The desired representation for J_1 follows from Lemma 2. For the J_2 term apply Lemma 2 in MRS in a similar way as described above to argue that

$$J_2 - \mathcal{K}^{r_z}_{h_n,x,z}\left\{(Y^n - Y^{\text{null}}) - \mathrm{E}[Y^n - Y^{\text{null}} \mid S, Z]\right\} = J_2 - J_2^* = o_p\big(n^{-1/2}\big).$$

By standard kernel smoothing arguments $J_2^* = o_p\big(n^{-1/2}\big)$. For the J_3 term let $A_i = \mathrm{E}[Y_i^n - Y_i^{\text{null}} \mid S_i, Z_i]$ and consider the behavior of the terms

$$\frac{1}{n_z}\sum_{i:Z_i=z} A_i\left(\frac{\hat{r}_z(S_i)-x}{h_n}\right)^a K_{h_n}(\hat{r}_z(S_i)-x), \quad a = 0, 1.$$

We focus on $a = 0$. The argument for the other case is similar. Let $K'_h(\cdot) = h^{-1}K'(\cdot/h)$. For any $\tilde{r}$ (pointwise) between $\hat{r}_z$ and r_z

$$\sup_x\left|\frac{1}{n_z}\sum_{i:Z_i=z}K'_{h_n}(\tilde{r}(S_i)-x)\right| \le C\sup_x\frac{1}{n_z h^z}\sum_{i:Z_i=z}1_{\{|r_z(S_i)-x|\le Ch^z\}} = O_p(1)$$

for a positive constant C. Noting that $\max_{i\le n}|A_i| = O_p(c_n)$ it is now easy to see that

$$\begin{aligned}
&\frac{1}{n_z}\sum_{i:Z_i=z}A_i K_{h_n}(\hat{r}_z(S_i)-x)\\
=&\frac{1}{n_z}\sum_{i:Z_i=z}A_i\left[K_{h_n}(r_z(S_i)-x) + K'_{h_n}(\tilde{r}(S_i)-x)\frac{\hat{r}_z(S_i)-r_z(S_i)}{h_n}\right]\\
=&\frac{1}{n_z}\sum_{i:Z_i=z}A_i K_{h_n}(r_z(S_i)-x) + o_p\big(n^{-1/2}\big)
\end{aligned}$$

uniformly in x. Let $M = \int \mu(t)\mu^\top(t)K(t)\,dt$, $M_n = M_{h_n}^{r_z}$ and $\bar{M}_n = \mathrm{E}\, M_n$. By Lemma 2 in Mammen, Rothe, and Schienle 2012 and standard arguments we have

$$\begin{aligned} M_{h_n}^{\hat{r}_z}(x) - f_{R|Z=z} M =& M_{h_n}^{\hat{r}_z}(x) - M_n(x) + M_n(x) - \bar{M}_n(x) + \bar{M}_n(x) - f_{R|Z=z}(x) M \\ =& O_p\left(n^{-\frac{1}{2}(1-3\eta)} + \sqrt{\frac{\log n}{n h_n}} + h_n \right) \end{aligned}$$

uniformly in x. Therefore,

$$J_3 - f_{R|Z=z}^{-1}(x) \frac{1}{n_z} \sum_{i:Z_i=z} A_i K_{h_n}(r_z(S_i) - x) = J_3 + J_3^* = o_p\big(n^{-1/2}\big).$$

It is straightforward to show that under our assumptions J_3^* can be replaced by its expectation at the expense of an uniform $o_p(n^{-1/2})$ term. Since

$$\mathrm{E}[Y^n - Y^{\mathrm{null}} \mid S, Z] = \mathrm{E}[Y^n - Y^{\mathrm{null}} \mid r_Z(S)] + \varphi_n(S, Z),$$

and since $f_{R|Z=z}$ has a bounded derivative

$$\begin{aligned} \mathrm{E}_z\, J_3^* = & \int \mathrm{E}[Y^n - Y^{\mathrm{null}} \mid r_Z(S) = x + h_n r] K(r)\, dr \\ & + \int \mathrm{E}[\varphi_n(S, Z) \mid r_Z(S) = x + h_n r, Z = z] K(r)\, dr + o\big(n^{-1/2}\big). \end{aligned}$$

Here we keep implicit that we are treating h_n as a constant in the above expectations, i.e., we are integrating with respect to the marginal measure of (Z, S). The conclusion follows by noting that the first term on the right-hand side is independent of z. □

Plugging in from Lemma 3 gives an asymptotic expansion of the test statistic.

Lemma 4

$$T_n = T_{n,a} + T_{n,b} + \int \Delta_{K,h}^2(x)\, dx + o_p(n\sqrt{h}),$$

where

$$T_{n,a} = \frac{2}{n^2} \sum_{i<j} \int \psi_{n,i}(x) \psi_{n,j}(x)\, dx \quad \textit{and} \quad T_{n,b} = \frac{1}{n^2} \int \sum_i \psi_{n,i}^2(x)\, dx.$$

Proof Plug in from Lemma 3, expand the square and inspect each term separately. □

Lemma 5 (Variance)
For $T_{n,a}$ as defined in Lemma 4

$$\begin{aligned} & \mathrm{var}(T_{n,a}) = n^{-2} h^{-1} V + o\left(n^{-2} h^{-1}\right) \quad \textit{and} \\ & n\sqrt{h} T_{n,a} \xrightarrow{d} \mathcal{N}(0, V). \end{aligned}$$

PROOF For the first part of the lemma, note that

$$\int K_h(r_z(s-gu)-x)L^*(u)\,du = \int \{K_h(r_z(s)-x) + \underbrace{K'(\chi_1/h)\partial_s r_z(\chi_2)u}_{\equiv a(s,u,x)}\frac{g}{h^2}\}L^*(u)\,du,$$

where χ_1 is an intermediate value between $r_z(s-hu)-x$ and $r_z(s)-x$, and χ_2 is an intermediate value between $s-hu$ and s. As K and r_z have bounded derivatives

$$\tilde{a}(r,x) = \mathrm{E}\Big[\int a(S,u,x)L^*(u)\,du \mid r_z(S)=r\Big]$$

is a bounded function. By standard U-statistic arguments

$$\begin{aligned}&\operatorname{var}\left(2\sum_{i<j}\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx\right) = 4\sum_{i<j}\mathrm{E}\Big[\int \psi_{n,i}(x)\psi_{n,j}(x)\,dx\Big]^2\\ &= 4\binom{n}{2}\int h\left\{\mathrm{E}[\psi_{n,1}(x)\psi_{n,1}(x+hx')]\right\}^2\,dx'\,dx.\end{aligned}$$

Note that

$$\mathrm{E}[\psi_{n,1}(x)\psi_{n,1}(x+hx')] = \sum_{z\in\{0,1\}}\mathrm{E}[\psi_{n,z,1}(x)\psi_{n,z,1}(x+hx')].$$

We consider here only one of the terms composing $\mathrm{E}[\psi_{n,z,1}(x)\psi_{n,z,1}(x+hx')]$. For the other terms similar arguments apply. Let

$$q(x) = -\frac{m'(x)1_{\{Z=z\}}}{p_z f_{R|Z=z}}\int K_h(r_z(S-gu)-x)L^*(u)\,du.$$

Using $\mathrm{E}_z[\zeta^2 \mid r_Z(S)=x] = x(1-x)$ we have

$$\begin{aligned}&h[\mathrm{E}\,q(x)q(x+hx')\zeta_1^2]\\ =&h\,\mathrm{E}\left[\frac{1_{\{Z=z\}}m_z'(x)m_z'(x+hx')}{p_z^2 f_{R|Z=z}(x)f_{R|Z=z}(x+hx')}(K_h(r_z(S)-x)+\frac{g}{h^2}\tilde{a}(r_z(S),x))\cdots\right.\\ &\left.\cdots(K_h(r_z(S)-x-hx')+\frac{g}{h^2}\tilde{a}(r_z(S),x-hx')\zeta^2\right]\\ =&\frac{x(1-x)[m_z'(x)]^2}{p_z f_{R|Z=z}(x)}\int K(y)K(x'-y)\,dy + o(1) = \frac{x(1-x)[m_z'(x)]^2}{p_z f_{R|Z=z}(x)}K^{(2)}(x') + o(1).\end{aligned}$$

For the second part of the lemma it suffices to check the two conditions of Theorem 2.1 in de Jong 1987. Let $W_{ij} = 2n^{-1}\sqrt{h}\int \psi_i(x)\psi_j(x)$ and show that

$$\begin{aligned}\operatorname{var}^{-1}\Big(\sum_{i<j}W_{ij}\Big)\max_{1\le i\le n}\sum_{1\le j\le n}\operatorname{var}(W_{ij}) &\to 0\\ \operatorname{var}^{-2}\Big(\sum_{i<j}W_{ij}\Big)\mathrm{E}\Big\{\sum_{i<j}W_{ij}\Big\}^4 &\to 3.\end{aligned}$$

The first condition holds trivially. To show that the second condition is satisfied note that $\operatorname{var}(\sum_{i<j} W_{ij})$ converges to a constant. It is easy to see that asymptotically only terms of the form $\mathrm{E}\, W_{ij}^2 W_{kl}^2$ with $\{i,j\} \cap \{k,l\} = \varnothing$ will contribute to $\mathrm{E}\left[\sum_{i<j} W_{ij}\right]^4$. There are

$$\binom{4}{2} \frac{\binom{n}{2}\left[\binom{n}{2}-1\right]}{2!} \approx \frac{3}{4} n^4$$

such terms when expanding $\mathrm{E}\left[\sum_{i<j} W_{ij}\right]^4$. The condition then follows by noting that

$$\operatorname{var}\Big(\sum_{i<j} W_{ij}\Big) = \sum_{i<j} \mathrm{E}\, W_{ij}^2$$

and that $\mathrm{E}\, W_{ij}^2 W_{kl}^2$ factors as $\mathrm{E}\, W_{ij}^2\, \mathrm{E}\, W_{kl}^2$. □

We now apply standard U-statistic theory. As the next two lemmas show, $T_{n,b}$ contributes to the asymptotic bias and $T_{n,b}$ contributes to the asymptotic variance.

Lemma 6 (Bias)
For $T_{n,b}$ as defined in Lemma 4

$$n\sqrt{h}\, T_{n,b} = \frac{1}{\sqrt{h}} \gamma_n + o_p(1),$$

where γ_n is a deterministic sequence converging to γ.

Proof Write

$$n\sqrt{h}\, T_{n,b} = \frac{\sqrt{h}}{n} \sum_i \int \psi_{n,i}^2(x)\, dx = \mathrm{E}\left\{ \frac{\sqrt{h}}{n} \sum_i \int \psi_{n,i}^2(x)\, dx \right\} + o_p(1) \equiv \gamma_n + o_p(1).$$

Define the function a as in the proof for Lemma 5. To compute γ_n write

$$\begin{aligned}
\psi_{n,z,i}^2(x) =& \frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)} \Big\{ K_h^2(r_z(S)-x)\epsilon^2 + [m'(x)]^2 K_h^2(r_z(S)-x)\zeta^2 \\
& - 2m'(x) K_h(r_z(S)-x)\epsilon\zeta \Big\} \\
& + \frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)} \left(\frac{g}{h^2} \int g(S,u,x) L^*(u)\, du \right)^2 \zeta^2 + \\
& \frac{1_{\{Z=z\}}}{p_z^2 f_{R,z}^2(x)} \frac{g}{h^2} \left(\int g(S,u,x) L^*(u)\, du \right) K_h(r_z(S)-x)\epsilon\zeta \\
=& \Gamma_1(S,x) + \Gamma_2(S,x) + \Gamma_3(S,x).
\end{aligned}$$

Note that

$$h \sum_{z=0,1} \mathrm{E} \int \Gamma_1(S,x)\, dx \to \gamma,$$

where we kept the dependence of Γ_1 on z implicit. Now show that the other terms entering γ_n vanish. To show that $h\sum_{z=0,1}\mathrm{E}\int\Gamma_3(S,x)\,dx\to 0$ it suffices to show that

$$\mathrm{E}_z\left[\left(\int g(S,u,x)L^*(u)\,du\right)\epsilon\zeta\mid r_z(S)\right]$$

is bounded. This follows immediately from the fact that $\int g(S,u,x)L^*(u)\,du$ is bounded and hence

$$\mathrm{E}_z\left[\int g(S,u,x)L^*(u)\,du\,\epsilon\zeta\mid r_z(S)\right]\lesssim \mathrm{E}_z[|\epsilon\zeta|\mid r_z(S)]\le\sqrt{\sigma^2_\epsilon(r_z(S))}\le C$$

for some constant C. For $h\sum_{z=0,1}\mathrm{E}\int\Gamma_2(S,x)\,dx$ argue similarly. □

Proof of Proposition 3

Proof Using the expansion from Lemma 3 and applying standard smoothing arguments to the stochastic term we get that for a small enough open set $\mathcal{G}_x\supset[\underline{x},\bar{x}]$

$$\sup_{x\in\mathcal{G}_x}|\hat{m}_0(x)-\hat{m}_1(x)|^2=O\left(\frac{1}{n\sqrt{h}}+g^{2(q+1)}\right)+O_p\left(\frac{\log n}{nh}\right)+o_p\left(\frac{1}{n}\right).$$

Write

$$T_n(\underline{x}_n,\bar{x}_n)-T_n(\underline{x},\bar{x})=T_n(\underline{x}_n,\underline{x})-T_n(\bar{x},\bar{x}_n).$$

We can bound $T_n(\underline{x}_n,\underline{x})$ by

$$|\underline{x}_n-\underline{x}|\sup_{x\in\mathcal{G}_x}|\hat{m}_0(x)-\hat{m}_1(x)|=o_p(n\sqrt{h}).$$

Similarly, we can find a bound for $T_n(\bar{x},\bar{x}_n)$. □

C. Tables

	$\theta = 0.10$						$\theta = 0.05$					
C_h	0.50	0.75	1.00	1.25	1.50	1.75	0.50	0.75	1.00	1.25	1.50	1.75
null												
$C_g = 0.50$	6.7	5.7	5.8	8.9	7.0	7.9	2.7	2.6	1.9	4.6	3.3	2.8
$C_g = 0.75$	9.2	6.4	8.2	6.5	6.4	7.0	4.6	2.0	3.2	2.8	3.2	2.8
$C_g = 1.00$	6.4	6.7	8.1	6.8	8.8	7.1	2.2	2.9	2.9	3.1	3.2	2.8
alternative 1												
$C_g = 0.50$	65.8	65.8	67.7	63.8	65.3	65.7	50.5	49.9	53.2	47.5	50.6	50.8
$C_g = 0.75$	65.1	65.8	64.8	65.3	65.8	65.9	49.7	47.7	49.9	49.5	50.1	52.3
$C_g = 1.00$	66.3	65.0	66.4	67.9	64.8	66.5	50.4	51.2	50.3	51.1	50.9	49.2
alternative 2												
$C_g = 0.50$	82.4	79.9	80.2	80.5	81.6	78.0	67.9	66.8	68.4	67.3	68.6	65.5
$C_g = 0.75$	79.2	81.0	79.9	80.6	80.4	79.8	66.1	68.3	68.0	68.2	66.5	65.8
$C_g = 1.00$	80.9	81.4	80.1	80.3	80.3	78.2	68.4	67.5	66.1	66.9	64.0	64.7
alternative 3												
$C_g = 0.50$	6.9	8.1	8.8	7.7	5.0	6.7	2.3	3.9	3.3	4.2	1.8	3.2
$C_g = 0.75$	7.2	8.1	6.8	7.4	6.7	6.9	2.9	2.6	3.7	3.9	2.1	3.0
$C_g = 1.00$	7.7	8.0	6.2	6.7	7.8	7.1	2.6	3.3	3.1	2.3	3.5	2.6
alternative 4												
$C_g = 0.50$	15.0	10.5	15.1	14.0	13.1	12.2	7.0	4.8	6.5	5.7	7.0	6.6
$C_g = 0.75$	12.5	13.9	13.8	12.9	13.6	13.3	5.2	6.2	7.0	7.0	6.3	5.9
$C_g = 1.00$	10.0	15.7	14.1	15.7	11.5	14.2	4.2	6.9	7.4	9.5	4.7	6.7
alternative 5												
$C_g = 0.50$	12.0	12.4	15.5	13.5	14.2	13.2	5.7	4.6	7.4	4.9	5.8	6.0
$C_g = 0.75$	13.4	14.5	12.5	12.1	12.0	11.1	6.0	6.9	5.7	4.2	5.3	5.7
$C_g = 1.00$	12.2	14.3	13.1	12.6	12.9	12.2	5.3	5.8	5.7	5.9	6.4	6.2
alternative 6												
$C_g = 0.50$	22.5	23.0	22.9	24.0	21.6	23.1	12.3	12.4	11.2	14.3	10.7	12.8
$C_g = 0.75$	23.3	20.6	25.3	23.5	23.3	20.0	12.5	11.4	13.2	12.0	13.5	12.1
$C_g = 1.00$	22.0	22.0	20.9	25.7	24.0	20.8	11.7	11.6	9.9	13.4	12.7	9.9

Table 3: Simulation. Empirical rejection probabilities in percentage points under nominal level θ. Sample size is $n = 200$.

			D		Y	
Race	Z	n	mean	sd	mean	sd
black	0	787	0.19949	0.3999	0.8183	0.3858
	1	67	0.26866	0.4466	0.6269	0.4873
hispanic	0	549	0.18033	0.3848	0.7687	0.4221
	1	36	0.27778	0.4543	0.5278	0.5063
white	0	1394	0.07389	0.2617	0.8479	0.3592
	1	77	0.20779	0.4084	0.6234	0.4877

Table 4: Teenage child bearing (D) and high-school graduation (Y).

	g	h	T_n	$\underline{x}_1$	$\bar{x}_1$	$\underline{x}_2$	$\bar{x}_2$	$\underline{x}_3$	$\bar{x}_3$	$P(>T_n)$	test result
1	1.00	0.15	0.086	0.03	0.32	0.01	0.21	0.05	0.46	0.225	no rejection
2	1.50	0.15	0.053	0.03	0.27	0.03	0.18	0.05	0.41	0.224	no rejection
3	2.00	0.15	0.084	0.03	0.21	0.02	0.15	0.05	0.36	0.059	*
4	2.50	0.15	0.054	0.04	0.19	0.02	0.13	0.11	0.27	0.012	**
5	3.00	0.15	0.022	0.02	0.18	0.05	0.11	0.13	0.19	0.092	*
6	1.00	0.20	0.064	0.03	0.32	0.01	0.21	0.05	0.46	0.060	*
7	1.50	0.20	0.042	0.03	0.27	0.03	0.18	0.05	0.41	0.084	*
8	2.00	0.20	0.067	0.03	0.21	0.02	0.15	0.05	0.36	0.010	**
9	2.50	0.20	0.043	0.04	0.19	0.02	0.13	0.11	0.27	0.010	**
10	3.00	0.20	0.019	0.02	0.18	0.05	0.11	0.13	0.19	0.083	*
11	1.00	0.25	0.045	0.03	0.32	0.01	0.21	0.05	0.46	0.012	**
12	1.50	0.25	0.037	0.03	0.27	0.03	0.18	0.05	0.41	0.036	**
13	2.00	0.25	0.051	0.03	0.21	0.02	0.15	0.05	0.36	0.008	***
14	2.50	0.25	0.035	0.04	0.19	0.02	0.13	0.11	0.27	0.025	**
15	3.00	0.25	0.017	0.02	0.18	0.05	0.11	0.13	0.19	0.090	*
16	1.00	0.30	0.040	0.03	0.32	0.01	0.21	0.05	0.46	0.010	**
17	1.50	0.30	0.035	0.03	0.27	0.03	0.18	0.05	0.41	0.036	**
18	2.00	0.30	0.044	0.03	0.21	0.02	0.15	0.05	0.36	0.021	**
19	2.50	0.30	0.030	0.04	0.19	0.02	0.13	0.11	0.27	0.022	**
20	3.00	0.30	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.080	*
21	1.00	0.35	0.039	0.03	0.32	0.01	0.21	0.05	0.46	0.005	***
22	1.50	0.35	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.024	**
23	2.00	0.35	0.041	0.03	0.21	0.02	0.15	0.05	0.36	0.014	**
24	2.50	0.35	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.018	**
25	3.00	0.35	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.064	*
26	1.00	0.40	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.003	***
27	1.50	0.40	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.021	**
28	2.00	0.40	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.007	***
29	2.50	0.40	0.028	0.04	0.19	0.02	0.13	0.11	0.27	0.011	**
30	3.00	0.40	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.064	*
31	1.00	0.50	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.003	***
32	1.50	0.50	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.012	**
33	2.00	0.50	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.012	**
34	2.50	0.50	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.005	***
35	3.00	0.50	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.065	*

Table 5: Test results for varying bandwidths and c_δ = 0.050. (*) reject at 0.10 level, (**) reject at 0.05 level, (***) reject at 0.01 level.

	g	h	T_n	$\underline{x}_1$	$\bar{x}_1$	$\underline{x}_2$	$\bar{x}_2$	$\underline{x}_3$	$\bar{x}_3$	$P(>T_n)$	test result
1	1.00	0.15	0.057	0.06	0.29	0.03	0.18	0.07	0.44	0.170	no rejection
2	1.50	0.15	0.033	0.06	0.24	0.05	0.15	0.08	0.39	0.208	no rejection
3	2.00	0.15	0.066	0.06	0.19	0.04	0.13	0.07	0.33	0.042	**
4	2.50	0.15	0.037	0.06	0.16	0.04	0.10	0.13	0.25	0.011	**
5	3.00	0.15	0.009	0.04	0.16	0.07	0.09	0.16	0.16	0.114	no rejection
6	1.00	0.20	0.041	0.06	0.29	0.03	0.18	0.07	0.44	0.038	**
7	1.50	0.20	0.028	0.06	0.24	0.05	0.15	0.08	0.39	0.108	no rejection
8	2.00	0.20	0.048	0.06	0.19	0.04	0.13	0.07	0.33	0.009	***
9	2.50	0.20	0.029	0.06	0.16	0.04	0.10	0.13	0.25	0.014	**
10	3.00	0.20	0.009	0.04	0.16	0.07	0.09	0.16	0.16	0.101	no rejection
11	1.00	0.25	0.033	0.06	0.29	0.03	0.18	0.07	0.44	0.011	**
12	1.50	0.25	0.025	0.06	0.24	0.05	0.15	0.08	0.39	0.044	**
13	2.00	0.25	0.034	0.06	0.19	0.04	0.13	0.07	0.33	0.012	**
14	2.50	0.25	0.023	0.06	0.16	0.04	0.10	0.13	0.25	0.020	**
15	3.00	0.25	0.008	0.04	0.16	0.07	0.09	0.16	0.16	0.113	no rejection
16	1.00	0.30	0.031	0.06	0.29	0.03	0.18	0.07	0.44	0.010	**
17	1.50	0.30	0.024	0.06	0.24	0.05	0.15	0.08	0.39	0.036	**
18	2.00	0.30	0.031	0.06	0.19	0.04	0.13	0.07	0.33	0.015	**
19	2.50	0.30	0.020	0.06	0.16	0.04	0.10	0.13	0.25	0.023	**
20	3.00	0.30	0.007	0.04	0.16	0.07	0.09	0.16	0.16	0.138	no rejection
21	1.00	0.35	0.030	0.06	0.29	0.03	0.18	0.07	0.44	0.005	***
22	1.50	0.35	0.024	0.06	0.24	0.05	0.15	0.08	0.39	0.024	**
23	2.00	0.35	0.029	0.06	0.19	0.04	0.13	0.07	0.33	0.017	**
24	2.50	0.35	0.018	0.06	0.16	0.04	0.10	0.13	0.25	0.013	**
25	3.00	0.35	0.007	0.04	0.16	0.07	0.09	0.16	0.16	0.124	no rejection
26	1.00	0.40	0.030	0.06	0.29	0.03	0.18	0.07	0.44	0.008	***
27	1.50	0.40	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.020	**
28	2.00	0.40	0.029	0.06	0.19	0.04	0.13	0.07	0.33	0.016	**
29	2.50	0.40	0.028	0.04	0.19	0.02	0.13	0.11	0.27	0.005	***
30	3.00	0.40	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.076	*
31	1.00	0.50	0.038	0.03	0.32	0.01	0.21	0.05	0.46	0.001	***
32	1.50	0.50	0.033	0.03	0.27	0.03	0.18	0.05	0.41	0.012	**
33	2.00	0.50	0.040	0.03	0.21	0.02	0.15	0.05	0.36	0.012	**
34	2.50	0.50	0.029	0.04	0.19	0.02	0.13	0.11	0.27	0.010	**
35	3.00	0.50	0.015	0.02	0.18	0.05	0.11	0.13	0.19	0.062	*

Table 6: Test results for varying bandwidths and c_δ = 0.075. (*) reject at 0.10 level, (**) reject at 0.05 level, (***) reject at 0.01 level.

References

Abadie, Alberto (2003). "Semiparametric instrumental variable estimation of treatment response models". In: *Journal of Econometrics* 113.2, pp. 231–263.

Angrist, Joshua D. and Ivan Fernandez-Val (2010). "ExtrapoLATEing: External validity and overidentification in the LATE framework". Working Paper.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin (1996). "Identification of Causal Effects Using Instrumental Variables". In: *Journal of the American Statistical Association* 91.434, pp. 444–455.

Balke, Alexander and Judea Pearl (1997). "Bounds on treatment effects from studies with imperfect compliance". In: *Journal of the American Statistical Association* 92.439, pp. 1171–1176.

Carneiro, Pedro, James Heckman, and Edward Vytlacil (2011). "Estimating Marginal Returns to Education". In: *American Economic Review* 101.6, pp. 2754–2781.

Carneiro, Pedro and Sokbae Lee (2009). "Estimating distributions of potential outcomes using local instrumental variables with an application to changes in college enrollment and wage inequality". In: *Journal of Econometrics* 149.2, pp. 191–208.

de Jong, Peter (1987). "A central limit theorem for generalized quadratic forms". In: *Probability Theory and Related Fields* 75.2, pp. 261–277.

Delgado, Miguel A (1993). "Testing the equality of nonparametric regression curves". In: *Statistics & probability letters* 17.3, pp. 199–204.

Dette, Holger and Natalie Neumeyer (2001). "Nonparametric analysis of covariance". In: *the Annals of Statistics* 29.5, pp. 1361–1400.

Frölich, Markus (2007). "Nonparametric IV estimation of local average treatment effects with covariates". In: *Journal of Econometrics* 139.1, pp. 35–75.

Gørgens, Tue (2002). "Nonparametric comparison of regression curves by local linear fitting". In: *Statistics & probability letters* 60.1, pp. 81–89.

Hall, Peter and Jeffrey D Hart (1990). "Bootstrap test for difference between means in nonparametric regression". In: *Journal of the American Statistical Association* 85.412, pp. 1039–1049.

Hall, Peter and Joel Horowitz (2012). *A simple bootstrap method for constructing nonparametric confidence bands for functions*. Tech. rep. working paper.

Hansen, Lars Peter (1982). "Large sample properties of generalized method of moments estimators". In: *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

Härdle, Wolfgang and Enno Mammen (1993). "Comparing nonparametric versus parametric regression fits". In: *The Annals of Statistics* 21.4, pp. 1926–1947.

Heckman, James, Daniel Schmierer, and Sergio Urzua (2010). "Testing the correlated random coefficient model". In: *Journal of Econometrics* 158.2, pp. 177–203.

Heckman, James, Sergio Urzua, and Edward Vytlacil (2006). "Understanding instrumental variables in models with essential heterogeneity". In: *The Review of Economics and Statistics* 88.3, pp. 389–432.

Heckman, James and Edward Vytlacil (2005). "Structural Equations, Treatment Effects, and Econometric Policy Evaluation". In: *Econometrica*, pp. 669–738.

Heckman, James et al. (1996). "Sources of selection bias in evaluating social programs: An interpretation of conventional measures and evidence on the effectiveness of matching as a program evaluation method". In: *Proceedings of the National Academy of Sciences* 93.23, pp. 13416–13420.

— (1998). "Characterizing selection bias using experimental data". In: *Econometrica: Journal of the Econometric Society* 66.5, pp. 1017–1098.

Hoffman, Saul D (1998). "Teenage childbearing is not so bad after all... or is it? A review of the new literature". In: *Family Planning Perspectives* 30.5, pp. 236–243.

Hotz, V Joseph, Susan Williams McElroy, and Seth G Sanders (2005). "Teenage Childbearing and Its Life Cycle Consequences Exploiting a Natural Experiment". In: *Journal of Human Resources* 40.3, pp. 683–715.

Hotz, V Joseph, Charles H Mullin, and Seth G Sanders (1997). "Bounding causal effects using data from a contaminated natural experiment: analysing the effects of teenage childbearing". In: *The Review of Economic Studies* 64.4, pp. 575–603.

Huber, Martin and Giovanni Mellace (2011). "Testing instrument validity for LATE identification based on inequality moment constraints". Working Paper.

Imbens, Guido W. and Joshua D. Angrist (1994). "Identification and estimation of local average treatment effects". In: *Econometrica*, pp. 467–475.

King, Eileen, Jeffrey D Hart, and Thomas E Wehrly (1991). "Testing the equality of two regression curves using linear smoothers". In: *Statistics & Probability Letters* 12.3, pp. 239–247.

Kitagawa, Toru (2008). "A Bootstrap Test for Instrument Validity in the Heterogeneous Treatment Effect Model". Working Paper.

Klepinger, Daniel H, Shelly Lundberg, and Robert D Plotnick (1995). "Adolescent fertility and the educational attainment of young women". In: *Family planning perspectives*, pp. 23–28.

Kong, Efang, Oliver Linton, and Yingcun Xia (2010). "Uniform bahadur representation for local polynomial estimates of M-regression and its application to the additive model". In: *Econometric Theory* 26.05, pp. 1529–1564.

Lee, Ying-Ying (2013). "Partial mean processes with generated regressors: Continuous Treatment Effects and Nonseparable models." Working Paper.

Levine, David I and Gary Painter (2003). "The schooling costs of teenage out-of-wedlock childbearing: analysis with a within-school propensity-score-matching estimator". In: *Review of Economics and Statistics* 85.4, pp. 884–900.

Mammen, Enno (1993). "Bootstrap and wild bootstrap for high dimensional linear models". In: *The Annals of Statistics*, pp. 255–285.

Mammen, Enno, Christoph Rothe, and Melanie Schienle (2012). "Nonparametric regression with nonparametrically generated covariates". In: *The Annals of Statistics* 40.2, pp. 1132–1170.

Masry, Elias (1996). "Multivariate local polynomial regression for time series: uniform strong consistency and rates". In: *Journal of Time Series Analysis* 17.6, pp. 571–599.

Miller, Amalia R (2011). "The effects of motherhood timing on career path". In: *Journal of Population Economics* 24.3, pp. 1071–1100.

Neumeyer, Natalie and Holger Dette (2003). "Nonparametric comparison of regression curves: an empirical process approach". In: *The Annals of Statistics* 31.3, pp. 880–920.

Reinhold, Steffen (2007). "Essays in demographic Economics". PhD thesis. John Hopkins University.

Ribar, David C (1994). "Teenage fertility and high school completion". In: *The Review of Economics and Statistics*, pp. 413–424.

Ruppert, David and Matthew P Wand (1994). "Multivariate locally weighted least squares regression". In: *The Annals of Statistics*, pp. 1346–1370.

Sargan, John D (1958). "The estimation of economic relationships using instrumental variables". In: *Econometrica: Journal of the Econometric Society*, pp. 393–415.

Van der Vaart, Aad W and Jon A Wellner (1996). *Weak Convergence and Empirical Processes*. Springer.

Vytlacil, Edward (2002). "Independence, monotonicity, and latent index models: An equivalence result". In: *Econometrica* 70.1, pp. 331–341.