Journal of Econometrics 226 (2022) 423–450

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

A test of the selection on observables assumption using a discontinuously distributed covariate

Umair Khalil [a], Neşe Yıldız [b,*]

[a] Centre for Development Economics and Sustainability, Monash University, Building H, Caulfield Campus, Melbourne, VIC, Australia
[b] Department of Economics, University of Rochester, 222 Harkness Hall, Rochester, NY 14627, United States of America

ARTICLE INFO

Article history:
Received 23 June 2016
Received in revised form 24 March 2021
Accepted 30 September 2021
Available online 1 December 2021

JEL classification:
C12
C14
C21

Keywords:
Selection on observables
Testing
Partial effects
Essential heterogeneity

ABSTRACT

We present a test of the selection on observables assumption that neither requires instruments nor excluded covariates from the structural function. Instead, we rely on the presence of a discontinuously distributed variable among the set of controls. We develop formal testing procedures for a non-parametric additively separable model for binary and finite treatment variables. We also outline a nonparametric nonseparable extension. Our test is easy to implement and should be useful in many empirical settings. Specifically, we employ it to study selection concerns in the estimation of the impact of a nutritional aid program for pregnant women on birth weight.

1. Introduction

Applied economists are often interested in estimating the partial effect of a given variable, X, on an outcome of interest, Y. The majority of this research is done in a linear or partially linear setting, where the unobservables in the model are represented by an additive unobservable variable, η. Furthermore, a bulk of studies in the empirical literature rely on the crucial assumption of the exogeneity of X (the so-called selection on observables assumption) to identify and estimate the relevant partial effect. Specifically, they assume that η is stochastically unrelated to X conditional on other covariates (controls). This assumption, however, is notoriously difficult to test without additional structure. Caetano (2015) provides a fully non-parametric test of exogeneity, but her test only applies to a situation in which the treatment variable X has a discontinuous distribution.

In this paper we provide the first test of exogeneity of X, where X could be a variable of any type (discrete, continuous, neither discrete, nor continuous). Our null hypothesis comprises of two joint hypotheses: (i) the structural equation is additively separable in X and the unobservable variable η, and (ii) η is conditionally mean independent of X given W. When the null hypothesis is true η equals the structural unobservable, which we call U. On the other hand, when the null is rejected, X is endogenous in one of two ways. First, if the true structural equation is not additively separable but the researcher models it as such, then $\eta \neq U$ making X endogenous due to misspecification. Second, the true structural equation is additively separable, but η, which equals U in this case, is not mean independent of X conditional on W.

* Corresponding author.
E-mail addresses: umair.khalil@monash.edu (U. Khalil), nese.yildiz@rochester.edu (N. Yıldız).

https://doi.org/10.1016/j.jeconom.2021.09.018
0304-4076/

Note that either source of endogeneity will lead to incorrect estimation of the average partial effect of X on Y in a setup assuming additive separability of X and the unobservables.

Our testing procedure exploits the presence of a covariate, W, whose distribution has a discontinuity at a known point w_0, but is continuous near that point. We are assuming the researcher is interested in identifying the average partial effect of X on Y, not the partial effect of W on Y. However, W is a variable the researcher believes should be controlled for in their analysis. The main intuition for how our test works is as follows: suppose that the structural equation is actually additively separable in U. Then if W is endogenous, that is, if U stochastically depends on W conditional on X, then one might suspect that the conditional distribution of U given X and W will also be discontinuous in W. This follows since W itself has a discontinuous distribution. If X drops from this conditional distribution, then the discontinuity in $\mathbb{E}(U|X = x, W = w)$ will not depend on X. On the other hand, if X does not drop from this conditional distribution, the discontinuity in $\mathbb{E}(U|X = x, W = w)$ will generally vary with the value of X, and our baseline test will have power.

For implementing the above testing idea, we propose a test statistic that is based on a direct sample analog of the function whose continuity we wish to verify. This computation only involves standard nonparametric regression techniques and can be implemented using existing Stata routines. We also derive the asymptotic behavior of this test statistic. Through a Monte Carlo study we show that in finite samples our test has good size and power properties.

Finally, we also implement our test with an empirical application. The treatment variable we study is participation in a federal aid program called the Supplemental Program for Women Infants and Children (WIC). WIC provides nutritional aid to low-income pregnant mothers with the aim of improving birth outcomes. While the goals of WIC are clear, the empirical literature has found it difficult to assess the true effect of WIC receipt, since participation into the program is not random.[1]

The outcome of interest is an infant's birth weight (Y) modeled as a function of WIC participation (X), average number of cigarettes smoked daily during pregnancy (W), other controls like mother's demographics and pregnancy's characteristics, plus some unobservable, U. The latter measures the conscientiousness of the mother among other unobserved factors. We maintain the assumption that the above function is continuous in the smoking variable.

Around 80% of mothers in our sample do not report smoking anytime during pregnancy. In addition, since it is possible to smoke part of a cigarette, smoking is a continuous variable for positive amounts smoked.[2] Thus, average cigarettes smoked daily is discontinuously distributed with a discontinuity at 0, and is continuously distributed on the positive side near 0. Moreover, expected birth weight conditional on the amount smoked, and WIC participation status is discontinuous in the average daily cigarettes smoked as shown in Fig. 9.2(a) in Section 8. Together these two observations give us the structure required on the discontinuously distributed covariate W. Intuitively, one can think of an unobservable measuring a mother's conscientiousness that positively determines birth weight but reduces smoking behavior. Since highly conscientious mothers are more likely to have babies with higher birth weight, and they cannot smoke negative cigarettes, they will bunch at zero cigarettes creating the discontinuity in the distribution of birth weight.

Given that the function relating smoking and WIC participation (and possibly other controls) to birth weight is assumed to be continuous in W, this means that the conditional expectation of the unobserved variable must be *discontinuous* in smoking due to the above outlined bunching behavior of conscientious mothers. However, if the selection on observables assumption holds, that is, if the unobservables in the outcome equation are mean independent of WIC participation conditional on smoking and other controls, then this discontinuity must be the same for WIC participants and non-participants. For instance, this would fail if more conscientious mothers are more likely to enroll in WIC as well, implying differential positive selection of mothers into WIC, creating a divergence in the discontinuities for participants and non-participants.[3] One could design a test of selection on observables assumption by checking if the discontinuities are equal or not. These arguments would work even if treatment was not a binary variable.

The paper is organized as follows. Section 2 discusses how our paper fits into the existing literature. Section 3 discusses empirical scenarios where our methods can be useful. In Section 4 we introduce the baseline model and its main testable implications. Section 5 discusses the implementation of our testing idea. Section 6 presents some extensions of our baseline model including for nonseparable models. Section 7 illustrates finite sample properties of our test statistic. Section 8 discusses our empirical application in detail. Section 9 concludes. The Appendix provides the proofs.

2. Literature review

The most closely related paper to ours is Caetano (2015), given the common assumption on W of a discontinuity at a known value, but being otherwise continuously distributed. Both papers further assume that the function relating W to the outcome of interest is continuous. However, in Caetano (2015), W itself is the treatment variable, and the focus is on estimating the average partial effect of W on Y. In addition, both papers present a test of the exogeneity of the treatment

[1] In Section 8, we lay out the problems in estimation of WIC treatment effects in much more detail.

[2] In Section 7, we provide simulation evidence regarding the sensitivity of our procedure when researchers do not have access to a 'truly' continuous W near the discontinuity point.

[3] In other words, under the assumption that the birth weight equation is additively separable in WIC participation and the unobservables, the discontinuity in expected birth weight conditional on smoking, other controls and WIC status should be the same for both participants and non-participants.

variable in models with nonparametric structural functions, but our paper imposes an additional weak monotonicity assumption. Finally, in her paper W, the treatment variable, is neither discrete nor continuous but has the particular structure of bunching at a known point and being continuously distributed otherwise. Our treatment variable X, on the other hand, can be of any type.

Chapter 21 of Imbens and Rubin (2015) discusses some of the existing approaches to assessing the unconfoundedness assumption. For instance, consider the subset unconfoundedness assumption: treatment is independent of potential outcomes both conditional on a set of observable controls ($\tilde{Z}$) and on a strict subset of these controls ($\tilde{Z}^r$). Equality of the two different average treatment effects (ATE) expressions can provide insights on the unconfoundedness assumption. A second approach relies on the presence of a proxy for one of the potential outcomes, which is observed regardless of treatment status. One can then test whether treatment is independent of this proxy variable, $\tilde{Z}^p$, conditional on $\tilde{Z}^r$.[4] The third approach is based on the availability of a two-component control group. Let G_i denote the group indicator with $G_i \in \{c_1, c_2\}$ implying treatment status of observation i is 0 (untreated), and t (treated). Moreover, it is assumed that G is independent of potential outcomes conditional on controls, $\tilde{Z}$. The group unconfoundedness assumption then implies that the group indicator is independent of the observed outcome conditional on $\tilde{Z}$ and the event that $G \in \{c_1, c_2\}$. Then, by testing whether $\mathbb{E}(\mathbb{E}[Y|\tilde{Z}, G = c_1] - \mathbb{E}[Y|\tilde{Z}, G = c_2])$ is equal to 0 or not, one can assess the group unconfoundedness assumption. In contrast, the testing procedures we propose exploit the special structure of the distribution of W, which is one of the controls. The approaches described by Imbens and Rubin (2015) do not require such a control, but they test assumptions more restrictive than unconfoundedness.

Similarly, one could also test the selection on observables assumption if one has access to an instrument. See, for example, Huber (2013), Donald et al. (2014) and de Luna and Johansson (2014). Our approach does not require availability of an instrument.

Crump et al. (2008) present nonparametric tests for treatment heterogeneity for binary treatments under the selection on observables assumption. They test whether the treatment effect conditional on $\{W = w, Z = z\}$ is constant in (w, z). The test we present for our baseline, additively separable model can thus also be interpreted as a test of a form of leftover unobserved heterogeneity. In other words we are testing the so-called essential heterogeneity of Heckman et al. (2006), which is present if the effect of the treatment is different across individuals with the same value of (W, Z).

This paper is also related to the literature on testing the validity of identifying assumptions in nonlinear models. Caetano et al. (2016) use a discontinuity in the distribution of X (the treatment variable) to test for the validity of a control function in nonseparable triangular models. Kitagawa (2015) presents a joint test for validity of an instrument and the LATE monotonicity assumption in the context of a binary treatment variable X. Lu and White (2014) present a test for additive separability of the structural function in X and U under the assumption that U and X are independent conditional on Z. Here, Z is a variable that is excluded from the structural function. Under the same conditional independence assumption, Hoderlein et al. (2014) present a procedure for testing whether the structural function is strictly monotone in a scalar, continuously distributed unobservable variable, U. Similarly, Lewbel et al. (2015) present a procedure for testing $Y = G(H_1(X, W) + U)$, with strictly increasing G under the assumption that U is independent of either (X, W) or U is independent of (X, W) conditional on Z, where Z is again excluded from the structural function. With $G(y) = y$ their test becomes a test of additive separability. We do not require W to be independent of U, and we do not assume that it is excluded from the structural function.

3. Potential applications

As mentioned above, the testable implications we provide exploit a covariate W that is continuously distributed, except for having a discontinuity in its distribution at a known point. As discussed in Sections 4.2 and 7, our methods work best when the bunching variable W is related in some way to both unobservables in the Y equation and in the X, or participation equation. We believe that this is indeed likely to be the case in many empirical settings.

- **College Major and Labor Market Outcomes:** In studying the effect of choosing a STEM (science, technology, engineering, and mathematics) major on future labor market outcomes, one could use SAT scores as the W variable. Quantitative SAT scores might have a bunching point at the top of the distribution due to a clustering of high ability students. In addition, more able students are likely to choose STEM majors in the first place, implying a differential discontinuity in both post-college wages and unobservables.
- **Export Promotion Programs and Export Sales:** Trade economists have been interested in studying the impact of firm-level export promotion programs (X) on export sales (Y).[5] For these firms, number of markets served, or products produced, are often bunched at 1 (Martincus and Carballo, 2010), which can provide the W in our setup.[6]

[4] Lagged (pre-treatment) values of the outcome variable are particularly plausible proxies for potential outcome $Y(0)$.

[5] For instance, Broocks and Biesebroeck (2017) study this problem for Belgium under the selection on observables assumption.

[6] Furthermore, export oriented firms might have better managerial staff or more entrepreneurial owners who can be more aware of both the existence of promotion programs as well as of the benefits of specializing in one product/market. This can lead to a differential discontinuity at 1 for participants and non-participants.

- **Unemployment Duration and Labor Market Outcomes:** Using time-use diaries, Krueger and Mueller (2010) present evidence of bunching at zero for job search intensity among unemployed workers (W).[7] This can be used to study the effect of unemployment duration, a continuous X, on future labor market outcomes like wages (Y).
- **Mother's Labor Force Participation and Children's Outcomes:** A mother's labor force participation (LFP) decision (X) can have profound impacts on children's academic outcomes (Y). The number of hours a mother allows her children to watch television on school days can provide the bunching variable (W).[8] Arguments for the relationship between W and the unobservables in the outcome equation and LFP equation are similar to our WIC example.
- **Dynamic Spillovers in Crime:** The effect of lagged crime rates (X) on current crime (Y) (Jacob et al., 2007) is complicated by endogeneity concerns due to unobserved location specific criminogenic factors influencing crime rates across time. We can use bunching at zero in police response to calls for service as our discontinuously distributed variable (W).[9] Neighborhoods where police are quick to respond might have discontinuously zero crime due to a deterrent effect, and police efficiency may vary negatively with increased values of lagged crime (X).

Other examples where our methods can be used are: exploiting bunching associated with minimum wage thresholds to study the impact of job training on future labor market outcomes; using discontinuities at zero in firm level investments to investigate the impact of management policies on worker productivity; and estimation of the effect of unionization on wages, using firm size as the W variable, as firm size is likely to have a bunching point due to regulations (For example, see Gourio and Roys (2014)).

4. The model and its testable implications

Before we introduce the model and the testing procedure, we introduce some notation. In particular, for random variable T, X, W let $\mu_T(x, w) := \mathbb{E}(T|X = x, W = w)\, a.s.$ and $\mu_T(x, 0^+) = \lim_{w \downarrow 0} \mathbb{E}(T|X = x, W = w)\, a.s.$, whenever it is defined. Also for expositional purposes other controls, which we will call Z, are suppressed until the next section when we discuss the implementation of our methods. In this section all the assumptions should be taken to hold conditional on each value z of the vector of controls Z.

Throughout the paper the general notation we will use for the structural equation determining the outcome of interest is as follows:

$$Y = m(X, W, U), \tag{4.1}$$

where Y is the outcome variable (denoting birth weight in our application), X is the treatment variable (participation in WIC), W is a scalar control (smoking during pregnancy), and U represents an unobservable variable.

4.1. Baseline model

Our baseline model starts with additively separable unobservables.[10] The outcome equation is given by

$$Y = g(X, W) + U, \tag{4.2}$$

so that in our baseline model $m(X, W, U) = g(X, W) + U$. We assume that $\mu_U(x, w)$ exists for each x, w in the support of (X, W) and that $\mu_U(x, 0^+)$ exists for each x in the support of X.

When the outcome equation is of this form, relevant treatment effects are identified under the following conditional independence assumption:

$$\mathbb{E}(U|X = x, W = w) = \mathbb{E}(U|W = w). \tag{4.3}$$

To see this, suppose that X is binary taking values 0 and 1 with positive probability. Then the potential outcomes are given by

$$Y_1 = g(1, W) + U,$$
$$Y_0 = g(0, W) + U.$$

Therefore, $Y_1 - Y_0 = g(1, W) - g(0, W)$ and the average treatment effect (ATE) equals $\mathbb{E}[g(1, W) - g(0, W)]$ where the expectation is taken with respect to the distribution of W. If the conditional mean independence condition (4.3) holds,

[7] It might be easier to envisage the applicability of our setup in naturally occurring thresholds like time inputs which are naturally bounded to the left by zero.

[8] Caetano et al. (2019) provide extensive evidence of such bunching in their study of the effect of children's time allocation on cognitive skill development.

[9] In fact, Caetano and Maheshri (2018) actually have access to such a variable in their high frequency crime data for the city of Dallas, and they document exactly such a bunching in police response at zero.

[10] In Section 6, we discuss how a modification of the ideas presented here can lead to testable implications in nonseparable models as well.

we have

$$\begin{aligned}\mathbb{E}(Y|X=1,W)-\mathbb{E}(Y|X=0,W) &= g(1,W)-g(0,W)+\mathbb{E}(U|W)-\mathbb{E}(U|W),\\ &= g(1,W)-g(0,W),\end{aligned}$$

and

$$ATE = \mathbb{E}[\mathbb{E}(Y|X=1,W)-\mathbb{E}(Y|X=0,W)] = \int [g(1,w)-g(0,w)]\mathrm{d}F_W(w).$$

Similarly, under the conditional mean independence condition (4.3), the average treatment effect on the treated is also identified and is given by

$$\mathbb{E}[\mathbb{E}(Y|X=1,W)-\mathbb{E}(Y|X=0,W)|X=1] = \int \{g(1,w)-g(0,w)\}\,\mathrm{d}F_{W|X}(w|1).$$

The conditional independence condition is crucial for identifying these (and possibly other) treatment effects when the structural equation relating X, W and U (the structural unobservable) to Y is as in Eq. (4.2). If, in fact, $m(X,W,U)$ is not additively separable in X and U, then the conditional mean independence condition (4.3) is not sufficient to identify the average treatment effect (or the average treatment effect on the treated). Moreover, if we incorrectly assume that the structural equation is additively separable in treatment and the unobservables, and write

$$Y = g(X,W)+\eta,$$

then η will not be the same as the structural unobservable U, but instead be equal to $m(X,W,U)-g(X,W)$. Thus, η will not necessarily satisfy the conditional mean independence assumption even when the true structural unobservable U is independent of X conditional on W. Therefore, we formulate our key identifying assumption as follows:

Assumption 1. (i) The structural equation is as in Eq. (4.2). (ii) the conditional mean independence condition (4.3) holds.

It is not possible to test for this assumption without imposing additional restrictions. Our aim is to demonstrate that we can devise a testable implication of this assumption in certain empirically relevant settings. In particular, we propose a testing procedure that relies on the existence of a control, W, that has a discontinuous distribution. In particular, we model W as being a censored variable:[11]

$$W = \max\{0, W^*\}. \tag{4.4}$$

Moreover, we assume that W^* has a Lebesgue density conditional on each value of X and that:

$$0 < \mathbb{P}(W^* \le 0|X=x) < 1, \text{ for each } x. \tag{4.5}$$

This assumption simply means that W has bunching at 0, and is a continuous random variable when it is greater than 0. It is not essential that we model W as the maximum of 0 and partially latent variable W^*. All we really need is that the CDF of W has a jump of size less than 1 at 0 and is continuous in a neighborhood of 0. Modeling W this way simplifies exposition. Nevertheless, in our empirical application, W can be thought of as average daily cigarette consumption. Then as discussed in Caetano et al. (2016) women with a large negative realization of W^* can be thought of as those who would require a substantial compensation in utility-equivalent units in order to smoke even a single cigarette per day, whereas women with W^* close to zero are the ones who are almost indifferent between smoking and not smoking.

Finally, we assume for each (x,u) in the support of (X,U)

$$\lim_{w\downarrow 0} m(x,w,u) - m(x,0,u) = 0. \tag{4.6}$$

This assumption requires the structural function to be continuous in w at $w = 0$ for each value of X. Continuity of the structural function is a reasonable assumption in many economic settings. In addition, smoothness of structural functions is commonly assumed when defining average treatment effects, or partial effects for continuous treatments. Under Assumption 1(i), a sufficient condition for this assumption is that $\lim_{w\downarrow 0} g(x,w) = g(x,0)$ for each x in the support of X.

Now we are ready to state the hypotheses that we would like to test:

$$\mathbb{H}_0\text{: Assumption 1 holds} \quad \textit{vs.} \quad \mathbb{H}_1\text{: Assumption 1 is violated;} \tag{4.7}$$

We would like to perform this test under the maintained assumptions that conditions (4.5)–(4.6) hold, and that $\mathbb{E}(Y|X=x,W=0)$ and $\lim_{w\downarrow 0}\mathbb{E}(Y|X=x,W=w)$ exist for each x.

[11] For the testing procedure to work the mass, or bunching point of W could be at the boundary or in the interior of the support of W. In our empirical application, the mass point is at the lower bound of the support of W. For this reason, the assumptions and all the limits are written to fit this setting. Everything we do could also be done, at the expense of added notation, if W has multiple mass points either in the interior or the upper bound of the support of W.

To motivate our testing procedure, note that

$$\mathbb{E}(Y|X=x,W=w)=\begin{cases}\mathbb{E}(Y|X=x,W^*=w), & \text{if } w>0,\\ \mathbb{E}(Y|X=x,W^*\leq w), & \text{if } w=0.\end{cases}$$

Since the conditioning sets on the right-hand side of the previous equation differ, one would generally expect the function $\mathbb{E}(Y|X=x,W=w)$ to be discontinuous at $w=0$. If, however, Assumption 1 holds, under the continuity condition (4.6), the size of this discontinuity should not depend on x. As a result, we must have for $x\neq x'$

$$\begin{aligned}\lambda(x,x') &:= \lim_{w\downarrow 0}\mathbb{E}(Y|X=x,W=w)-\mathbb{E}(Y|X=x,W=0)\\ &-\left(\lim_{w\downarrow 0}\mathbb{E}(Y|X=x',W=w)-\mathbb{E}(Y|X=x',W=0)\right)=0,\end{aligned} \tag{4.8}$$

with probability 1. In contrast, when Assumption 1 does not hold, one would generally expect that the discontinuity depends on x, which would imply that the above entity is different from 0 with positive probability. The result below summarizes this discussion.

Theorem 1. *Under our maintained assumptions, a necessary condition for $\mathbb{H}_0$ to hold is that $\lambda(x,x')=0$ a.s.*

Remark 1. If one was interested in testing $(X,W)\perp\!\!\!\perp U$, then the result of Theorem 1 would still apply even when the structural equation is of the form $Y=m(X,W,U)$, under the assumption that $m(1,w,u)-m(0,w,u)$ is continuous in w at $w=0$ for each u. See Section 6 for more on this.[12]

4.2. Power properties of the baseline test

In a general setting such as ours, deriving testable implications of an identification condition like Assumption 1 is notoriously difficult. Thus, our test will lack power against some alternatives. To give empirical researchers better guidance we analyze the behavior of $\lambda(x,x')$ under different data generating processes that violate Assumption 1. We also provide conditions that only depend on observable quantities, which can be used to check whether our testing procedure is going to have power in a given empirical setting.

First, consider alternatives under which the first part of Assumption 1 is violated. That is, suppose that the true structural function is not additively separable, but we mistakenly write

$$Y=g(X,W)+\eta,$$

with g continuous in w for each x, and $\eta=m(X,W,U)-g(X,W)$. Then,

$$\begin{aligned}\lambda(x,x') = &\lim_{w\downarrow 0}\int[m(x,w,u)-g(x,w)]\left[\mathrm{d}F_{U|X,W}(u|x,w)-\mathrm{d}F_{U|X,W}(u|x,0)\right]\\ &-\lim_{w\downarrow 0}\int[m(x',w,u)-g(x',w)]\left[\mathrm{d}F_{U|X,W}(u|x',w)-\mathrm{d}F_{U|X,W}(u|x',0)\right].\end{aligned}$$

Therefore, under this type of alternatives there is no reason to expect $\lambda(x,x')$ to be zero. This is not surprising, because if the researcher uses a misspecified model for inference, the treatment variable is generally going to be endogenous. Table 2 in Section 7 provides evidence that this is indeed the case.

Next, consider alternatives under which the first part of Assumption 1 holds, but the second part is violated. The following two conditions are necessary for our testing procedure to have power under such alternatives: (I) $\mathbb{E}(U|X=x,W=w)$ is discontinuous in w at $w=0$ for each $x\in A\subseteq \mathrm{Supp}(X)$ with $\mathbb{P}(A)>0$: (II) this discontinuity is different for $(x,x')\in A\times A$.

The first of these necessary conditions is violated if $\mathbb{E}(U|X=x,W=w)=\mathbb{E}(U|X=x)$. This situation is depicted in rows 2–5 of column (1) in each panel in Table 1 7. Note that whether violation of Condition (I) is a concern for a potential empirical application can be verified by using (Caetano, 2015) to check for each given value of x of X, if

$$\lim_{w\downarrow 0}\mathbb{E}(Y|X=x,W=w)-\mathbb{E}(Y|X=x,W=0)$$

equals 0 or not. In our empirical application $\lim_{w\downarrow 0}\mathbb{E}(Y|X=x,W=w)-\mathbb{E}(Y|X=x,W=0)$ is different from 0 for both treated and untreated observations.

Condition (II) above will be violated if for some functions ψ_1 and ψ_2, which are measurable w.r.t. X and W, respectively, we have

$$\mathbb{E}(U|X,W)=\psi_1(X)+\psi_2(W).$$

[12] We are grateful to an anonymous referee for pointing this out.

Table 1
Power analysis – U additively separable case – $3 * Bandwidth$.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $N = 2000$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.067	0.066	0.072	0.068	0.137	0.223	0.069	0.285	0.534
0.2	0.064	0.072	0.073	0.067	0.162	0.251	0.066	0.316	0.524
0.4	0.067	0.081	0.079	0.068	0.197	0.299	0.074	0.374	0.583
0.6	0.067	0.091	0.102	0.066	0.266	0.384	0.075	0.475	0.682
0.8	0.064	0.118	0.165	0.066	0.370	0.532	0.079	0.632	0.815
Panel B: $N = 5000$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.062	0.065	0.066	0.064	0.246	0.426	0.067	0.571	0.883
0.2	0.065	0.076	0.076	0.067	0.309	0.493	0.067	0.628	0.883
0.4	0.065	0.088	0.096	0.066	0.393	0.594	0.068	0.716	0.919
0.6	0.066	0.122	0.143	0.065	0.513	0.717	0.074	0.826	0.960
0.8	0.067	0.196	0.262	0.071	0.693	0.868	0.082	0.940	0.991

Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section 7.1.

Table 2
Power analysis — Non-separable case, κXU wuth $\kappa = 0.25\beta$.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $N = 2000$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.064	0.063	0.068	0.153	0.363	0.491	0.303	0.773	0.931
0.2	0.062	0.067	0.066	0.156	0.361	0.457	0.324	0.756	0.873
0.4	0.064	0.071	0.067	0.178	0.394	0.453	0.378	0.780	0.852
0.6	0.065	0.066	0.071	0.207	0.442	0.495	0.458	0.826	0.880
0.8	0.061	0.066	0.078	0.256	0.547	0.580	0.665	0.911	0.934
Panel B: $N = 5000$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.061	0.062	0.064	0.275	0.696	0.846	0.612	0.987	0.999
0.2	0.063	0.070	0.064	0.282	0.709	0.822	0.642	0.984	0.999
0.4	0.065	0.062	0.063	0.329	0.739	0.816	0.708	0.988	0.997
0.6	0.065	0.068	0.075	0.391	0.799	0.851	0.817	0.994	0.999
0.8	0.068	0.073	0.080	0.516	0.886	0.912	0.951	0.999	0.999

Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section 7.1. Missing entries correspond to combinations of the parameter space where the variance covariance matrix is not positive semi-definite.

Then $\lambda(x, x')$ will be 0 for $x \neq x'$, even when ψ_1 is not equal to 0, since the discontinuity in $\mathbb{E}(U|X, W)$ does not depend on x in this case. While this seems troubling, it is less problematic than it seems at first, because of the structure of W. To explain this, let us assume that the equation generating X is given by

$$X = h(W, V),$$

where V is a vector of unobservables. If $\mathbb{E}(U|V, W^*)$ is not additively separable in V and W^*, then because of the bunching in W, $\mathbb{E}(U|X, W)$ is not additively separable in X and W. To study the worst case scenario for our testing procedure, let us further assume that $\mathbb{E}(U|V, W^*) = \zeta_1(V) + \zeta_2(W^*)$. Note that his situation arises when (U, W^*, V) are jointly normal, for example. Then

$$\mathbb{E}(U|X = x, W = w) = \mathbb{E}[\zeta_1(V)|h(w, V) = x, W = w] + \zeta_2(w), \text{ if } w > 0,$$
$$\mathbb{E}(U|X = x, W = 0) = \mathbb{E}[\zeta_1(V)|h(0, V) = x, W = 0] + \mathbb{E}[\zeta_2(W^*)|h(0, V) = x, W^* \leq 0].$$

If W^* is independent of V, and $\lim_{w\downarrow 0} \mathbb{E}[\zeta_1(V)|h(w, V) = x] = \mathbb{E}[\zeta_1(V)|h(0, V) = x]$, we can easily show that $\lim_{w\downarrow 0} \mathbb{E}(U|X = x, W = w) - \mathbb{E}(U|X = x, W = 0)$ does not vary with x, and our testing procedure does not have power. On the other hand, if $h(0, V)$ is not independent of W^*, even if $\lim_{w\downarrow 0} \mathbb{E}[\zeta_1(V)|h(w, V) = x, W = w] = \mathbb{E}[\zeta_1(V)|h(0, V) = x, W^* = 0]$, we have

$$\begin{aligned}
&\lim_{w\downarrow 0} \mathbb{E}(U|X = x, W = w) - \mathbb{E}(U|X = x, W = 0) \\
&\quad = \mathbb{E}[\zeta_1(V)|h(0, V) = x, W^* = 0] - \mathbb{E}[\zeta_1(V)|h(0, V) = x, W = 0] \\
&\qquad + \lim_{w\downarrow 0} \zeta_2(w) - \frac{\mathbb{E}[\zeta_2(W^*)1\{W^* \leq 0\}|h(0, V) = x]}{\mathbb{P}(W^* \leq 0|h(0, V) = x)},
\end{aligned}$$

which varies with x. These arguments illustrate that whether W^* and $h(0, V)$ are independent or not plays a crucial role in whether our testing procedure has power or not. Furthermore, under the weak assumption that

$$\lim_{w\downarrow 0}\mathbb{P}(h(w,V)\leq x|W=w)=\lim_{w\downarrow 0}\mathbb{P}(h(0,V)\leq x|W=w), \tag{4.9}$$

whether W^* and $h(0, V)$ are independent can be verified by checking

$$\begin{aligned}&\lim_{w\downarrow 0}\mathbb{P}(X\leq x|W=w)-\mathbb{P}(X\leq x|W=0)\\&\quad=\lim_{w\downarrow 0}\mathbb{P}(h(0,V)\leq x|W^*=w)-\mathbb{P}(h(0,V)\leq x|W^*\leq 0)\end{aligned} \tag{4.10}$$

equals zero or not. If $h(w, v)$ is continuous in w (at 0), then condition (4.9) will hold, since probabilities are naturally bounded. Even when $h(w, v)$ is not continuous, however, condition (4.9) can still hold. In particular, suppose X is discrete, with $1\{X \leq x\} = 1\{V \leq b_x(w)\}$ with $b_x(w)$ continuous. Then $\lim_{w\downarrow 0}\mathbb{P}(X \leq x|W = w) = \lim_{w\downarrow 0}\mathbb{P}(V \leq b_x(0)|W = w)$, and condition (4.9) holds. For instance, in our empirical application we see such a pattern holds for WIC participation in Fig. 9.2(b) in Section 8.

We end this section by pointing out that the data generating process in our Monte Carlo study is such that $X = \{\alpha + \beta W \geq V\}$, with (U, V, W^*) jointly normal. Then $V \perp\!\!\!\perp W^* \iff \sigma_{vw^*} = 0$, and our test has power only when $\sigma_{vw^*} \neq 0$. In rows 2–5 of columns (4) and (7) of each panel of Table 1, $\sigma_{vw^*} = 0$, and Condition (II) fails.

5. Test statistics and their large sample properties

In this section, we devise our test statistic. Recall that the null hypothesis in Section 4 consists of two parts: (i) that the data generating process is such that $Y = g(X, W, Z)+U$, and (ii) $\mathbb{P}_{XWZ}\big(\mathbb{E}(U|X, W, Z) = \mathbb{E}(U|W, Z)\big) = 1$, where Z now represents a vector of other observed covariates. Theorem 1 states that under our maintained assumptions, a necessary condition for this null hypothesis to hold is that

$$\begin{aligned}\lambda(x,x',z)&:=\lim_{w\downarrow 0}\mathbb{E}(Y|X=x,W=w,Z=z)-\mathbb{E}(Y|X=x,W=0,Z=z)\\&-\left(\lim_{w\downarrow 0}\mathbb{E}(Y|X=x',W=w,Z=z)-\mathbb{E}(Y|X=x',W=0,Z=z)\right)=0.\end{aligned}$$

The test statistic we propose replaces the population conditional expectations with their corresponding sample versions. We expect the dimension of Z to be quite large in many empirical applications.[13] This means that fully nonparametric estimation of $\lambda(x, x', z)$ may not be feasible for many empirical settings because of data constraints.[14] For this reason, we implement our testing procedure for models in which

$$g(X,Z,W)=\tilde{g}(X,W)+Z^{\top}\gamma, \tag{5.1}$$

so that

$$Y=\tilde{g}(X,W)+Z^{\top}\gamma+U. \tag{5.2}$$

We also assume that Z is exogenous, in the following sense:

Assumption 2. $\mathbb{E}\big([U-\mathbb{E}(U|X,W)][Z-\mathbb{E}(Z|X,W)]\big)=0.$

Note that

$$Y-\mathbb{E}(Y|X,W)=[Z-\mathbb{E}(Z|X,W)]\top\gamma+U-\mathbb{E}(U|X,W),$$

and by Robinson (1988), $\sqrt{n}(\widehat{\gamma}-\gamma)\xrightarrow{d}N(0,V_\gamma)$ under Assumption 2.

Once γ is known we can write

$$\tilde{Y}:=Y-Z^{\top}\gamma=\tilde{g}(X,W)+U.$$

Then under the null hypothesis, we have

$$\tilde{\lambda}(1)-\tilde{\lambda}(0)=0,$$

where $\tilde{\lambda}(x) := \lim_{w\downarrow 0}\mathbb{E}(\tilde{Y}|X = x, W = w) - \mathbb{E}(\tilde{Y}|X = x, W = 0)$.

The advantage of this additional structure is that at each step, one has to essentially perform one-dimensional, non-parametric local regressions. If the dimension of Z is large, however, estimation of γ would still require a large

[13] In our empirical application, even in our baseline specification, the dimension of Z is around 40.

[14] For instance, cell sizes around W greater than zero can diminish drastically for distinct Z and particularly so for interactions of various covariates, which we employ in our empirical application as well.

number of non-parametric regressions. If, in addition, $\mathbb{E}(Z_k|X,W)$ is linear in W for each $k \in \{1,2,\ldots,d_z\}$, that is $\mathbb{E}(Z_k|X,W) = X(\alpha_{1k} + W\beta_{1k}) + (1-X)(\alpha_{0k} + W\beta_{0k})$, then γ can be consistently estimated by a single OLS regression of Y on X and Z.[15] In the discussion below, we will simply assume that we have a $\sqrt{n}$-normal estimator $\widehat{\gamma}$ of γ, so that $\tilde{Y}$ is identified.

In our empirical application, we adopt a slightly more flexible partial linear specification. In particular, for the empirical application we assume that

$$Y = g^*(X,W) + XZ^\top\gamma_1 + (1-X)Z^\top\gamma_0 + U. \tag{5.3}$$

Assuming that $\mathbb{E}(U|X,W,Z) = \mathbb{E}(U|X,W) =: \rho(X,W)$ and letting $V = U - \rho(X,W)$, we get

$$Y = g^*(X,W) + XZ^\top\gamma_1 + (1-X)Z^\top\gamma_0 + \rho(X,W) + V,$$

and

$$Y - \mathbb{E}(Y|W,X) = X[Z - \mathbb{E}(W|X)]^\top\gamma_1 + (1-X)[Z - \mathbb{E}(W|X)]^\top\gamma_0 + V.$$

In this case, as before, γ_0, γ_1 can be estimated at a $\sqrt{n}$-rate.

We estimate $\mathbb{E}(\tilde{Y}|X=x,W=0)$ by the sample average of $\widehat{\tilde{Y}} := Y - Z^\top\widehat{\gamma}$, where the average is taken over observations for which $X=x$, $W=0$. For $x=0$, 1, we estimate $\lim_{w\downarrow 0}\mathbb{E}(\tilde{Y}|X=x,W=w)$ by using local linear regression of $\widehat{\tilde{Y}}$ on W at $W=0$ using only observations for which $W>0$ and $X=x$.

To derive the asymptotic distribution of our test statistic under the null hypothesis, we impose the following additional restrictions:

Assumption 3. (i) $\{Y_i, X_i, W_i, Z_i^\top\}_{i=1}^n$ is a random sample. For some $\alpha > 0$, $\mathbb{E}\left(|Y|^{2+\alpha}\right) < \infty$ and $\mathbb{E}\left(\|Z\|^{2+\alpha}\right) < \infty$.

(ii) The density $f_{W|X,\delta\geq W>0}(w,x)$ is bounded and bounded away from 0 for $x=0$, 1. It is also continuously differentiable on $(0,\delta)$ for $x=0$, 1.

(iii) For each $w \in (0,\delta)$ and $x=0$, 1, $\mathbb{E}[\tilde{Y}|W=w,X=x]$ is twice continuously differentiable in w.

(iv) For each $w \in (0,\delta)$, $x=0$, 1, and $j=1,2,\ldots,d_z$, $\mathbb{E}(Z_{ji}|W_i=w,X_i=x)$ is continuous in w.

(v) We have a first stage estimator $\widehat{\gamma}$ such that $\sqrt{n}(\widehat{\gamma}-\gamma) = O_P(1)$.

(vi) $Var(\varepsilon_i) < \infty$, and $\mathbb{E}(\varepsilon_i^2|W_i=w)$ is a continuous function of w for $w \in (0,\delta]$, $\lim_{w\downarrow 0}\mathbb{E}(\varepsilon_i^2|W_i=w,X=x)$ exists for $x=0$, 1.

(vii) The kernel function K has compact support and is twice continuously differentiable in the interior of its support. In addition, it satisfies the following conditions: $\int K(u)\mathrm{d}u = 1$ and $\int uK(u)\mathrm{d}u = 0$.

(viii) The bandwidth satisfies the following conditions as $n\to\infty$: $nh^5 \to 0$ and $\frac{\sqrt{nh}}{\log n} \to \infty$.

Before stating the main asymptotic result, we have to introduce some notation:

$$\begin{array}{rclcrcl} f_{W|X}(0^+|x) & := & \lim_{w\downarrow 0} f_{W|X}(w|x), & & \sigma^2_{\varepsilon|X,W}(x,0^+) & := & \lim_{w\downarrow 0}\mathbb{E}(\varepsilon_i^2|X_i=x,W_i=w), \\ \kappa_j & := & \int_0^\infty u^j K(u)\mathrm{d}u, & & \nu_j & := & \int_0^\infty u^j K^2(u)\mathrm{d}u, \end{array}$$

for $j=0,1,2$ and $x=0,1$.

Theorem 2. *Suppose the maintained assumptions discussed in Section 4.1 hold. In addition, suppose Assumption 3 holds. Then*

$$\sqrt{nh}\left(\widehat{\tilde{\lambda}}(1) - \widehat{\tilde{\lambda}}(0) - (\tilde{\lambda}(1) - \tilde{\lambda}(0))\right) \xrightarrow{d} N(0,V), \tag{5.4}$$

where $V = V_1 + V_0$, *where* $V_x = \frac{\kappa_2^2\nu_0 - 2\kappa_2\kappa_1\nu_1 + \kappa_1^2\nu_2}{(\kappa_0\kappa_2 - \kappa_1^2)^2}\frac{\sigma^2_{\varepsilon|X}(x,0^+)}{\mathbb{P}(X=x)f_{W|X}(0^+|x)}$.

The variance V is the sum of two variances. The terms in the sum are variances of the local linear estimator of the limit of the conditional expectation of $\tilde{Y}$ given $W=w$ as w goes to 0 for treated and untreated people. Since γ and $\mathbb{E}(\tilde{Y}|W=0,X=x)$ can be estimated at a $\sqrt{n}$ rate, first stage estimation of them does not influence the asymptotic distribution, and hence, the asymptotic variance of the test statistic. The covariance term disappears since $1\{X_i=1\}1\{X_i=0\}=0$ for each i. Each of the two variances whose sum equals V can be estimated in a straightforward fashion. In particular, for i such that $W_i>0$, we could estimate $\widehat{\varepsilon}_i$ as

$$\widehat{\varepsilon}_i := \widehat{\tilde{Y}}_i - \widehat{\mu}_{\widehat{\tilde{Y}}|X,W}(X_i,W_i),$$

and $\sigma^2_{\varepsilon|X}(x,0^+)$ can be estimated by using local linear regression of $\widehat{\varepsilon}$ on W at $W=0$ using only observations for which $W>0$ and $X=x$. As in the estimation of $\lim_{w\downarrow 0}\mathbb{E}(Y|W=w,X=x)$, the fact that $\widehat{\tilde{Y}}$, as opposed to $\tilde{Y}$ is generating $\widehat{\varepsilon}$

[15] This is an application of Frisch–Waugh–Lowell Theorem.

will not have a first order effect on the asymptotic behavior of $\widehat{\sigma}^2_{\varepsilon|X,W}(x,0^+)$ because of the faster convergence of the first stage estimator. Moreover, $\widehat{\sigma}^2_{\varepsilon|X,W}(x,0^+)$ will be consistent for $\sigma^2_{\varepsilon|X,W}(x,0^+)$. $f_{W|X}(0^+|x)$ can be consistently estimated as

$$\widehat{f}_{W|X}(0^+|x) := \frac{2}{\sum_{i=1}^n 1\{X_i = x\}} \sum_{i=1}^n \frac{1}{h_f} K_f\left(\frac{W_i}{h_f}\right) 1\{W_i > 0, X_i = x\},$$

where h_f is a bandwidth that goes to 0 as $n \to \infty$, and $\int_0^\infty K_f(u) = 0.5$. $\mathbb{P}(X = x)$ can be consistently estimated by the fraction of observations with $X = x$. Finally, the terms κ_j and ν_j can be calculated for each specific choice of the kernel. Thus, for $x = 0, 1$, we can consistently estimate V_x by $\widehat{V}_x = \frac{\kappa_2^2\nu_0 - 2\kappa_2\kappa_1\nu_1 + \kappa_1^2\nu_2}{(\kappa_0\kappa_2 - \kappa_1^2)^2} \frac{\widehat{\sigma}^2_{\varepsilon|X,W}(x,0^+)}{\widehat{\mathbb{P}}(X=x)\widehat{f}_{W|X}(0^+|x)}$. In our empirical application we use the estimators given in Stata.

In light of this theorem, we can define

$$\tilde{t}_n = \sqrt{nh}\frac{\widehat{\tilde{\lambda}}(1,0)}{\sqrt{\widehat{V}}}, \tag{5.5}$$

with $\widehat{\tilde{\lambda}}(1,0) = \widehat{\tilde{\lambda}}(1) - \widehat{\tilde{\lambda}}(0)$ and $\widehat{V}$ is some consistent estimator for V. Then we can reject the hypothesis that $\tilde{\lambda} = 0$ when $\tilde{t}_n \in \mathcal{R}$, $\mathcal{R} = (-\infty, c_{\alpha/2}] \cup [c_{1-\alpha/2}, \infty)$, and $c_{\alpha/2}$ and $c_{1-\alpha/2}$ are, respectively, $\alpha/2$ and $1-\alpha/2$ quantiles of the standard normal, respectively. By these arguments, we have the following corollary of Theorem 2:

Corollary 1. *Suppose the conditions of Theorem 2 hold and $\widehat{V}$ is some consistent estimator for V. Then if $\tilde{\lambda}(1,0) = 0$, then $\mathbb{P}(\tilde{t}_n \in \mathcal{R}) \to \alpha$ as $n \to \infty$.*

Since our null comprises a joint hypothesis, we can analyze the asymptotic properties of our baseline test statistic under different kind of alternative hypotheses. First, we consider the data generating process where the outcome equation is given by

$$Y = g(X,W) + Z^\top\gamma + U,$$

but the conditional mean independence assumption (4.3) is violated. Second, we study outcome equations of the type.

$$Y = m(X,W,Z,U).$$

Thus in this case, our null hypothesis fails regardless of whether $\mathbb{E}(U|X,W)$ equals $\mathbb{E}(U|W)$ with probability 1 or not. Under both types of alternative hypotheses $\tilde{\lambda}(1,0) = \tilde{\lambda}(1) - \tilde{\lambda}(0)$ is not necessarily 0. Since the conclusion of Theorem 2 holds even when $\tilde{\lambda}(1,0) \neq 0$, we get the following result on asymptotic consistency and power of our test statistic, which is a corollary of Theorem 2:

Corollary 2. *Suppose the conditions of 2 hold and $\widehat{V}$ is some consistent estimator for V.*

(i) For any fixed alternative that implies $\tilde{\lambda}(1,0) \neq 0$, $Pr(\tilde{t}_n \in \mathcal{R}) \to 1$ as $n \to \infty$.

(ii) Under any local alternative that implies $\tilde{\lambda}(1,0) = \frac{\delta}{\sqrt{nh}}$ with $\delta \neq 0$, $Pr(\tilde{t}_n \in \mathcal{R}) \to 1 - \Phi\left(c_{1-\alpha/2} - \frac{\delta}{\sqrt{V}}\right) + \Phi\left(c_{\alpha/2} - \frac{\delta}{\sqrt{V}}\right)$, as $n \to \infty$, where $\Phi(\cdot)$ denotes the standard normal distribution.

The proof of this Corollary can be found in the Appendix.

5.1. Test statistic for finite X

We now discuss how to implement our testing procedure when X takes finitely many different values $x_0, x_1, \ldots, x_K$ with positive probability. We start by noting that $\tilde{\lambda}(x_j)$ can be estimated using local linear regression in the same way as $\tilde{\lambda}(x)$ in the previous subsection. Then using standard arguments, which are presented in the Appendix we have

$$\sqrt{nh}\left(\widehat{\tilde{\lambda}}(x_j) - \tilde{\lambda}(x_j)\right) = \sqrt{nh}S^*_{nj} + o_P(1), \tag{5.6}$$

where

$$S^*_{nj} := e_1^\top \frac{A^{-1}}{\mathbb{P}(X = x_j)f_{W|X}(0^+|x_j)} \sum_{i=1}^n \begin{pmatrix} 1 \\ W_i/h \end{pmatrix} 1\{W_i > 0, X_i = x_k\} K_h(W_i)\varepsilon_i. \tag{5.7}$$

and A is a matrix of constants defined in the Appendix. This implies that

$$\sqrt{nh}\widehat{\Gamma}_n \xrightarrow{d} N(0,V), \tag{5.8}$$

where $\widehat{\Gamma}_n$ is a $\frac{K(K+1)}{2} \times 1$ vector whose components are $\widehat{\tilde{\lambda}}(x_l) - \widehat{\tilde{\lambda}}(x_k) - \left(\tilde{\lambda}(x_l) - \tilde{\lambda}(x_k)\right)$ for $k = 1, 2, \ldots, K-1$ and $l = k+1, \ldots, K$. The entries of the asymptotic variance matrix V are asymptotic covariances between $\sqrt{nh}(S^*_{nl} - S^*_{ns})$

and $\sqrt{nh}(S^*_{nt} - S^*_{nr})$ for all (l, s) and (t, r) pairs. The specific formulas for these entries are provided in the Appendix. Given the discussion in the previous subsection, we know that V can be estimated consistently. Moreover, under the null, $\tilde{\lambda}(x_k) - \tilde{\lambda}(x_l) = 0$ for each (k, l). Based on these arguments, we propose the following test statistic for the non-binary, finite X case

$$T_n = nh(L^*_n)^\top \widehat{V}^{-1}(L^*_n),$$

where L^*_n is a $\frac{K(K+1)}{2} \times 1$ vector whose components are $\widehat{\tilde{\lambda}}(x_l) - \widehat{\tilde{\lambda}}(x_k)$ for $k = 1, 2, \ldots, K-1$ and $l = k+1, \ldots, K$. The testing decision is thus to reject the null if $T_n > \chi^2_{\frac{K(K+1)}{2}}(1-\alpha)$, where $\alpha \in (0, 1)$ denotes the Type I error, and $\chi^2_{\frac{K(K+1)}{2}}(1-\alpha)$ denotes the $(1-\alpha)$-quantile of the χ^2-distribution with $\frac{K(K+1)}{2}$ degrees of freedom.

6. Extensions

We now discuss how our testing idea can be extended to a model in which X and the unobservable variable are not separable.

6.1. Monotone transformation of baseline model

We first illustrate our ideas in the context of a transformation of the baseline model in Section 4. This will help us build intuition for testing the fully nonseparable, nonparametric case. Consider

$$Y = m(X, W, U) = \tilde{m}(g(X, W) + U), \tag{6.1}$$

where Y, X, W, U are as before, g is continuous in w at $w = 0$ for each x, and $\tilde{m}$ is a strictly monotone, normalized increasing function. Finally, assume that conditional on $X = x, W = w$ for each x and w, the CDF of U is strictly increasing and continuous on $\mathbb{R}$; these assumptions ensure that the support of Y is $\mathbb{R}$.

In this case the potential outcomes and the ATE are given by

$$\begin{aligned} Y_1 &= \tilde{m}(g(1, W) + U), \\ Y_0 &= \tilde{m}(g(0, W) + U), \\ \mathbb{E}(Y_1 - Y_0) &= \mathbb{E}_{UW}[\tilde{m}(g(1, W) + U) - \tilde{m}(g(0, W) + U)]. \end{aligned}$$

The sufficient condition for identification of the ATE in this case is the conditional independence condition

$$U \perp\!\!\!\perp X | W. \tag{6.2}$$

We assume that the data generating process for W is as in the previous section.

Let $p(a|x, w) := \mathbb{P}(Y \leq a | X = x, W = w)$ *a.s.* Then if $g(x, w)$ is continuous in w for each x, under condition (6.2) we have

$$p(a|x, w) = \mathbb{P}\big(U \leq \tilde{m}^{-1}(a) - g(x, w) | W = w\big), \text{ and}$$

$$\lim_{w \downarrow 0} p(a|x, w) - p(a|x, 0) = \int_{-\infty}^{\tilde{m}^{-1}(a) - g(x,0)} [\lim_{w \downarrow 0} \mathrm{d}F_{U|W}(u|w) - \mathrm{d}F_{U|W}(u|0)].$$

If W is endogenous, the last expression will not be 0, and because of the upper limit of the integral, it will depend on x. As a result, even if condition (6.2) holds, we do not necessarily have

$$\lim_{w \downarrow 0} p(a|x, w) - p(a|x, 0) - \left(\lim_{w \downarrow 0} p(a|x', w) - p(a|x', 0)\right) = 0,$$

for $x \neq x'$. To get a testable implication of condition (6.2) for this model, we need to modify the difference-in-differences approach we used in the previous section. This modification is inspired by Vytlacil and Yıldız (2007), and it involves looking at different quantiles of Y conditional on $X = x, W = w$, as well as considering different x values. Define,

$$\theta(x, a) := \lim_{w \downarrow 0} p(a|x, w) - p(a|x, 0). \tag{6.3}$$

Suppose without loss of generality that $m^{-1}(a) - g(x, 0) \geq m^{-1}(b) - g(\tilde{x}, 0)$. Then when condition (6.2) holds, we have

$$\theta(x, a) - \theta(\tilde{x}, b) = \int_{\tilde{m}^{-1}(b) - g(\tilde{x}, 0)}^{\tilde{m}^{-1}(a) - g(x,0)} [\lim_{w \downarrow 0} \mathrm{d}F_{U|W}(u|w) - \mathrm{d}F_{U|W}(u|0)]. \tag{6.4}$$

This difference will be 0 when $\tilde{m}^{-1}(a) - g(x, 0) = \tilde{m}^{-1}(b) - g(\tilde{x}, 0)$, even when $\mathrm{d}F_{U|W}(u|w)$ is discontinuous in w at 0. Moreover, if the conditional distribution of $U|W = 0$ is strictly increasing on all of $\mathbb{R}$, then we have[16]

$$p(a|x, 0) = p(b|\tilde{x}, 0) \iff \tilde{m}^{-1}(a) - g(x, 0) = \tilde{m}^{-1}(b) - g(\tilde{x}, 0). \tag{6.5}$$

[16] Note that when $g(x, w) = \alpha + x\beta + w\delta$, $m^{-1}(a) - g(x, 0) = \tilde{m}^{-1}(b) - g(\tilde{x}, 0) \iff \tilde{m}^{-1}(a) - \tilde{m}^{-1}(b) = (x - \tilde{x})\beta$.

These arguments lead us to the following result:

Theorem 3. *Suppose the structural equation is as given in (6.1), and*

(i) $g(x, w)$ *is continuous in* w *for each* x*;*
(ii) $\tilde{m}$ *is strictly increasing;*
(iii) the conditional distribution of U *given* $W = 0$ *is strictly increasing on all of* $\mathbb{R}$*;*
(iv) $\mathbb{P}(\mathcal{S}) > 0$*, where*

$$\mathcal{S} := \{(x, \tilde{x}) : (x, 0), (\tilde{x}, 0) \in \text{Supp}(X, W) \text{ and } \exists a, b \in \mathbb{R} \text{ with } \mathbb{P}(Y \leq a | X = x, W = 0) = \mathbb{P}(Y \leq b | X = \tilde{x}, W = 0)\}. \tag{6.6}$$

Then the following statement is a necessary condition for conditional independence $U \perp\!\!\!\perp X|W$*: For each* $(x, \tilde{x}) \in \mathcal{S}$ *with their respective* a *and* b *values,* $\theta(x, a) - \theta(\tilde{x}, b) = 0$*.*

This theorem gives a testable implication of condition (6.2). In the above, (iv) gives a support condition. Since the conditional CDF of Y given $X = x, W = 0$ is identified from the data, we can check whether the condition holds or not. Nevertheless, we note that it will be satisfied under the null if either $g(x, 0)$ does not depend on x; or if there exists a subset, A, of the support of $X|W = 0$ with $\mathbb{P}(A) > 0$ such that for each $x \in A$, the support of $g(x, 0) + U$ conditional on $W = 0$ equals $\mathbb{R}$. Note that this latter assumption will hold if the conditional support of $U|W = 0$ equals $\mathbb{R}$, since the support of $g(x, 0) + U$ given $W = 0$ is just a translation of the support of $U|W = 0$ for each x. On the other hand, if a is such that $\mathbb{P}(Y \leq a|X = x, W = 0) \in (0, 1)$, then a is in the interior of the support of $\tilde{m}(g(x, 0) + U)$ given $W = 0$, and $\tilde{m}^{-1}(a)$ is in the support of $g(x, 0) + U$ given $W = 0$ under the null. Suppose $g(x, 0) \neq g(\tilde{x}, 0)$. Without loss of generality, assume that $g(x, 0) > g(\tilde{x}, 0)$. Then $\tilde{m}^{-1}(a) - g(x, 0) < \tilde{m}^{-1}(a) - g(\tilde{x}, 0)$. Since $\tilde{m}$ is strictly increasing, so is $\tilde{m}^{-1}$. Since $\mathbb{R}$ is connected we can find b such that $\tilde{m}^{-1}(a) - g(x, 0) = \tilde{m}^{-1}(b) - g(\tilde{x}, 0)$. Since support of $g(\tilde{x}, 0) + U$ given $W = 0$ is $\mathbb{R}$, $\tilde{m}^{-1}(b)$ belongs to the support of $g(\tilde{x}, 0) + U$ given $W = 0$.

To study what happens to $\theta(x, a) - \theta(\tilde{x}, b)$ when $U \not\perp\!\!\!\perp X|W$, note that because $((x, a), (\tilde{x}, b))$ are chosen to satisfy $p(a|x, 0) = p(b|\tilde{x}, 0)$ in the above theorem,

$$\theta(x, a) - \theta(\tilde{x}, b) = \lim_{w \downarrow 0} \big(\mathbb{P}(Y \leq a | X = x, W = w) - \mathbb{P}(Y \leq b | X = \tilde{x}, W = w)\big),$$

both when U is independent of X given W, and when it is not. When $U \perp\!\!\!\perp X|W$, this difference equals 0. When $U \not\perp\!\!\!\perp X|W$, it equals

$$\int_{-\infty}^{\tilde{m}^{-1}(a) - g(x,0)} \lim_{w \downarrow 0} \mathrm{d}F_{U|X,W}(u|x, w) - \int_{-\infty}^{\tilde{m}^{-1}(b) - g(\tilde{x},0)} \lim_{w \downarrow 0} \mathrm{d}F_{U|X,W}(u|\tilde{x}, w). \tag{6.7}$$

This testing procedure will have power when the conditional distribution of U given $\{X = x, W = 0\}$ is different from the limit of the conditional distribution of U given $\{X = x, W = w\}$ as w goes to 0 for a set of x values that has strictly positive probability. Since the distribution of W has a discontinuity at $W = 0$ in our set up, the two conditional distributions of U will be different from each other when W is endogenous. Intuitively, this is why we expect this testing procedure to have power against many alternatives. Thus, when W is endogenous we can expect (6.7) to be different from 0 even if

$$\int_{-\infty}^{\tilde{m}^{-1}(a) - g(x,0)} \mathrm{d}F_{U|X,W}(u|x, 0) - \int_{-\infty}^{\tilde{m}^{-1}(b) - g(\tilde{x},0)} \mathrm{d}F_{U|X,W}(u|\tilde{x}, 0) = 0.$$

To further analyze against which alternatives this procedure will have power, note that (6.7) equals

$$\int_{-\infty}^{\tilde{m}^{-1}(a) - g(x,0)} \Big[\lim_{w \downarrow 0} \mathrm{d}F_{U|X,W}(u|x, w) - \lim_{w \downarrow 0} \mathrm{d}F_{U|X,W}(u|\tilde{x}, w)\Big] \tag{6.8}$$

$$+ \int_{\min\{\tilde{m}^{-1}(a) - g(x,0),\, \tilde{m}^{-1}(b) - g(\tilde{x},0)\}}^{\max\{\tilde{m}^{-1}(a) - g(x,0),\, \tilde{m}^{-1}(b) - g(\tilde{x},0)\}} \lim_{w \downarrow 0} \mathrm{d}F_{U|X,W}(u|\tilde{x}, w). \tag{6.9}$$

When U is not independent of X conditional on W, $p(a|x, 0) = p(b|\tilde{x}, 0)$ does not necessarily imply that $\tilde{m}^{-1}(a) - g(x, 0) = \tilde{m}^{-1}(b) - g(\tilde{x}, 0)$. Thus, (6.9) is not necessarily 0. On the other hand, when X is endogenous conditional on W, we would expect that the conditional density of U given $X = x$ and $W = w$ as $w \downarrow 0$ to vary with the value of X and expression (6.8) to be different from 0. When X is endogenous, for some pair of $(x, \tilde{x})$ values we might find ourselves in a pathological case, in the sense that the sum of (6.8) and (6.9) equals 0, but it is unlikely that this will happen for all pairs of $(x, \tilde{x})$. Therefore, having more pairs satisfying $p(a|x, 0) = p(b|\tilde{x}, 0)$ helps with power.

Finally, we outline the implementation of this procedure for the binary X case. Assuming we have a random sample of size n of (Y, X, W) values, for $d = 0, 1$ we could estimate $\mathbb{P}(Y \leq y|X = d, W = 0)$ by

$$\widehat{p}(d, y) := \frac{\frac{1}{n}\sum_{i=1}^{n} 1\{Y_i \leq y, X_i = d, W_i = 0\}}{\frac{1}{n}\sum_{i=1}^{n} 1\{X_i = d, W_i = 0\}}.$$

Next, we could estimate $\lim_{w\to 0}\mathbb{P}(Y\le y|X=d,W=w)$ using local linear/polynomial regression:

$$\widehat{\theta}(d,y)=\frac{1}{\sum_{j=1}^{n}1\{X_j=d\}}e_1^{\top}\underset{(a_1,a_2^{\top})}{\operatorname{argmin}}\sum_{i=1}^{n}(1\{Y_i\le y\}-a_1-a_2W_i)^2K_g(W_i)1\{W_i>0,X_j=d\},$$

where $K_g(u)=\frac{1}{g}k\left(\frac{u}{g}\right)$, g is the bandwidth, and e_1 is the two dimensional first unit vector. Then one possible test statistic is

$$\sum_j\sum_{k\neq j}K\left(\frac{\widehat{p}(1,Y_j)-\widehat{p}(0,Y_k)}{h_n}\right)\left[\widehat{\theta}(1,Y_j)-\widehat{\theta}(0,Y_k)\right]^2,$$

where K is a symmetric kernel maximized at 0, with $K(0)>0$ and h_n is a bandwidth that goes to 0 as $n\to\infty$.

6.2. Fully nonparametric model

The ideas presented above can be used for testing the exogeneity of X in a more general nonseparable model. Assume that the structural equation is now given by

$$Y=m(X,W,U),\tag{6.10}$$

instead of (6.1). Here Y,X,W are as before, but now U is a vector of unobservables with $U=(U_1^{\top},U_2)^{\top}$, where U_1 is $J\times 1$ and U_2 is scalar. We assume that m is continuous in w at $w=0$ for each value of x, u_1 and u_2, and strictly increasing in u_2 for each x, w and u_1. Finally, we assume that the conditional distribution of (U_1,U_2) given $X=x,W=w$ has a Lebesgue density that is positive on all of $\mathbb{R}^{J+1}$, for each x and each $w\in[0,\delta]$ for some $\delta>0$.

Similar to the previous subsection, condition (6.2) is a sufficient condition for identification of the ATE for this model too. To get a testable implication of this condition, we are going to look at a different conditional quantile of Y given $X=x,W=0$ as we change the value of X from x to $\tilde{x}$ as in the previous subsection. In this more general model,

$$\theta(x,a)=\lim_{w\downarrow 0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}1\{m(x,w,u_1,u_2)\le a\}\left[\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|x,w)-\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|x,0)\right],$$

which is further equal to

$$\theta(x,a)=\lim_{w\downarrow 0}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}1\{m(x,w,u_1,u_2)\le a\}\left[\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|w)-\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|0)\right],$$

when condition (6.2) holds. On the other hand, since $1\{m(x,w,u_1,u_2)\le a\}$ is clearly bounded for each (x,w,u_1,u_2), if $\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|w)\le\nu(u_1,u_2)$ for some (U_1,U_2)-measurable function ν for each $w\in(0,\delta)$[17]

$$\theta(x,a)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\lim_{w\downarrow 0}1\{m(x,w,u_1,u_2)\le a\}\left[\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|w)-\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|0)\right],$$

when condition (6.2) holds. Moreover, since m is strictly increasing in u_2 for each (x,w,u_1), continuous in w at 0 for each (x,u_1,u_2), and the conditional distribution of (U_1,U_2) given $W=w$ has a Lebesgue density that is positive on all of $\mathbb{R}^{J+1}$, for each $w\in[0,\delta]$, $\mathbb{P}(m(x,w,U_1,U_2)=a|W=w)=0$, so that when (6.2) holds

$$\theta(x,a)=\int_{-\infty}^{\infty}\int_{-\infty}^{m^{-1}(x,0,u_1,a)}\left[\lim_{w\downarrow 0}\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|w)-\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|0)\right].$$

where $m^{-1}(x,w,u_1,y)$ satisfies $m(x,w,u_1,m^{-1}(x,w,u_1,y))=y$ and $m^{-1}(x,w,u_1,m(x,w,u_1,u_2))=u_2$. Finally, note when (6.2) holds

$$p(a|x,0)=\int_{-\infty}^{\infty}\int_{-\infty}^{m^{-1}(x,0,u_1,a)}\mathrm{d}F_{U_1,U_2|W}(u_1,u_2|0).$$

Therefore, for $(x,\tilde{x})\in\mathcal{S}$, which means that $p(x,a)=p(\tilde{x},b)$ for corresponding a and b, $\theta(x,a)-\theta(\tilde{x},b)=0$ if condition (6.2) holds. These arguments are summarized in the following result:

Theorem 4. *Suppose the structural equation is as given in (6.10), with m continuous in w at $w=0$ for each value of x, u_1 and u_2, strictly increasing in u_2 for each x, u_1 and w. Suppose also that the conditional density of (U_1,U_2) given $W=w$ is continuous in w for $w\in[0,\delta]$ for some $\delta>0$, the density of W^* is strictly positive for $w^*\in[0,\delta]$ and that $\mathbb{P}(\mathcal{S})>0$. Then the following statement is a necessary condition for the conditional independence condition $U\perp\!\!\!\perp X|W$: For each $(x,\tilde{x})\in\mathcal{S}$ with their respective a and b values, $\theta(x,a)-\theta(\tilde{x},b)=0$.*

[17] Note we make this assumption to ensure that the Dominated Convergence Theorem holds. In particular, if conditional density of (U_1,U_2) given $W=w$ is continuous in w for $w\in[0,\delta]$, and if the density of W^* is strictly positive for $w^*\in[0,\delta]$, then this condition will hold.

Remark 2. The implication of this theorem is almost immediately applicable to the case in which

$$Y = \max\{0, m(X, W, U_1, U_2)\}, \tag{6.11}$$

if $\mathcal{S}$ in the theorem is replaced by its slightly modified version

$$\begin{aligned}\mathcal{S}_+ := \{&(x, \tilde{x}) : (x, 0), (\tilde{x}, 0) \in \text{Supp}(X, W) \text{ and } \exists a, b \in \mathbb{R}_+ \text{ with} \\ &\mathbb{P}(Y \leq a | X = x, W = 0) = \mathbb{P}(Y \leq b | X = \tilde{x}, W = 0)\}\end{aligned} \tag{6.12}$$

Remark 3. For models given in Eqs. (6.10) and (6.11), if U is scalar, the support condition will be satisfied under the null if either $m(x, 0, U)$ does not depend on x, or if the support of U given $W = 0$ is $\mathbb{R}$. This is because under the null, if $\mathbb{P}(Y \leq a | X = x, W = 0) = \mathbb{P}(m(x, 0, U) \leq a | W = 0) \in (0, 1)$ then a must be in the support of $m(x, 0, U)$ given $W = 0$, and $m^{-1}(x, 0, a)$ must be in the support of U given $W = 0$. If, on the other hand, dimension of U_1 is not 0, then the support condition will hold if the support of $U_2 | U_1 = u_1, W = 0 = \mathbb{R}$ for each u_1.

The discussion of why a testing procedure based on $\theta(x, a) - \theta(\tilde{x}, b)$ with $\big((x, a), (\tilde{x}, b)\big)$ satisfying $p(a|x, 0) = p(b|\tilde{x}, 0)$ will have power is almost identical to the discussion in the previous subsection. Now we have multiple unobservables instead of a single one (although U_1 plays the same role as U in the previous model), and $\tilde{m}^{-1}(a) - g(x, 0)$ and $\tilde{m}^{-1}(b) - g(\tilde{x}, 0)$ in Eqs. (6.7)–(6.9) have to be replaced with $m^{-1}(x, 0, u_1, a)$, and $m^{-1}(\tilde{x}, 0, u_1, b)$ in the integration with respect to u_1 and we have to integrate each quantity in those equations over all possible values of U_2 as well.

6.2.1. Including controls

In this section we outline how our proposed testing procedure can be performed when controls other than W are added to the fully non-parametric model. First, it is straightforward to see that in principle controls could be included in the model fully nonparametrically, by interpreting all the conditional expectations in the above analysis as being also conditioned on a fixed value of the vector of controls Z. However, such an approach would likely suffer from the curse of dimensionality and will not have good finite sample properties. A popular approach in applied research to reduce dimensionality is allowing other controls to enter the model linearly:

$$Y = m(X, W, U) + Z^\top \gamma, \tag{6.13}$$

where Y, X, W and U are as before, and Z are other controls. If $Z \perp\!\!\!\perp U | X, W$, then $\mathbb{E}(Y|X, W, Z)$ has a linear index structure in Z, since $\mathbb{E}(Y|X, W, Z) = \mathbb{E}(m(X, W, U)|X, W) + Z^\top \gamma$, which has a linear index structure in Z. Therefore, a $\sqrt{n}$-consistent estimator $\widehat{\gamma}$ for γ can be obtained under standard regularity conditions by, for example, using the method of Ichimura and Lee (1991). Once γ is identified the above analysis can be performed on $\tilde{Y} := Y - Z^\top \gamma = m(X, W, U)$ as before.

6.3. Discrete Y

The extensions proposed in the previous two subsections are not applicable when Y is discrete. This is because the structural function m was strictly increasing in one of its arguments over part of its domain. We now demonstrate how the endogeneity of X can be tested when Y is discrete and X is continuous on at least part of its support. In particular, suppose

$$Y = \sum_{j=1}^{J} 1\{U > m_j(X, W)\}, \tag{6.14}$$

with $J > 1$ and $m_j(x, w) < m_{j+1}(x, d)$ for each j. We also assume that for each j, $m_j(x, w)$ is continuous in w at $w = 0$, for each x and each j. If $U \perp\!\!\!\perp X | W$,

$$\begin{aligned}&\mathbb{P}(Y \leq j | X = x, W = 0) = \mathbb{P}(U \leq m_j(x, 0) | W = 0), \\ &\mathbb{P}(Y \leq \tilde{j} | X = x', W = 0) = \mathbb{P}(U \leq m_{\tilde{j}}(\tilde{x}, 0) | W = 0).\end{aligned}$$

Therefore under this conditional independence condition, $\mathbb{P}(Y \leq j | X = x, W = 0) = \mathbb{P}(Y \leq \tilde{j} | X = \tilde{x}, W = 0)$ for some (x, j) and $(\tilde{x}, \tilde{j})$ is equivalent to $m_j(x, 0) = m_{\tilde{j}}(\tilde{x}, 0)$, and as a result we have

$$\begin{aligned}&\lim_{w \downarrow 0} \mathbb{P}(Y \leq j | X = x, W = w) - \mathbb{P}(Y \leq j | X = x, W = 0) \\ &\quad - \left\{ \lim_{w \downarrow 0} \mathbb{P}(Y \leq \tilde{j} | X = \tilde{x}, W = w) - \mathbb{P}(Y \leq \tilde{j} | X = \tilde{x}, W = 0) \right\} = 0.\end{aligned}$$

The theorem below summarizes our arguments:

Theorem 5. *Suppose the structural equation is as given in (6.11), with m_j continuous in w at $w = 0$ for each value of x and j. Suppose also that the conditional distribution of U given $X = x, W = w$ has a Lebesgue density that is positive on all of $\mathbb{R}$, for each x and each $w \in [0, \delta]$ for some $\delta > 0$, and that $\mathbb{P}(\mathcal{S}) > 0$. Then the following statement is a necessary condition for the conditional independence condition $U \perp\!\!\!\perp X | W$: For each $(x, \tilde{x}) \in \mathcal{S}$ with their respective j and $\tilde{j}$ values, $\theta(x, j) - \theta(\tilde{x}, \tilde{j}) = 0$.*

If X is discrete, in general we would not expect to have $\mathbb{P}(\mathcal{S}) > 0$. So this procedure is more applicable when X is at least partially continuous.

7. Monte Carlo analysis

7.1. Setup

In this section we present results from a Monte Carlo exercise for the binary X case. The data generating process (DGP) is as follows,

$$Y = \alpha + \beta X + \theta W + \phi XW + \kappa XU + U \tag{7.1}$$

$$X = 1\{\gamma + \delta W \geq V\} \tag{7.2}$$

$$W = \max\{0, W^*\} \tag{7.3}$$

$$\begin{pmatrix} U \\ V \\ W^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_{w^*} \end{pmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{uw^*} \\ \sigma_{uv} & \sigma_v^2 & \sigma_{vw^*} \\ \sigma_{uw^*} & \sigma_{vw^*} & \sigma_{w^*}^2 \end{bmatrix} \right) \tag{7.4}$$

We use the following parameter values to characterize the DGP: $\alpha = 1; \beta = 1.5; \theta = 1; \phi = 0.5; \gamma = 0.75; \delta = -0.75; \mu_{w^*} = -0.80; \sigma_u^2 = \sigma_v^2 = \sigma_{w^*}^2 = 1$. We draw samples of sizes 2000 and 5000 and repeat the experiment 10,000 times for each. Before proceeding with analyzing the simulation results, we point out that in our Monte Carlo setup, $\mathbb{E}(U|V, W^*) = aV + b(W^* - \mu_{w^*})$, as argued in Section 4.2. Thus, our Monte Carlo study is designed to analyze the finite sample properties of our test in a case where our test is least likely to have power.

7.2. Additively separable case: $\kappa = 0$

In Fig. 9.1 we provide a graphical representation of statistical power and endogeneity as a function of the three covariances of interest σ_{uv}, σ_{uw^*}, and σ_{vw^*} for the $\kappa = 0$ case. Under this setting, the null hypothesis of no endogeneity of X is represented by the case when either all three covariances are zero, the origin in Fig. 9.1, or when $\sigma_{uv} = 0$, and exactly one of σ_{uw^*} and σ_{vw^*} is not, as represented by the y-axis and the z-axis, respectively. All other regions in Fig. 9.1 represent the DGP under the alternative hypothesis of endogeneity of X.

As the blue dotted and dashed lines in the Figure indicate, our test has power on the planes $\sigma_{uv} = 0, \sigma_{uw^*} \neq 0, \sigma_{vw^*} \neq 0$, and $\sigma_{uw^*} = 0, \sigma_{uv} \neq 0, \sigma_{vw^*} \neq 0$. Our test, however, has no power on the plane $\sigma_{vw^*} = 0, \sigma_{uv} \neq 0, \sigma_{uw^*} \neq 0$. This plane is depicted by green dashed lines in the Figure. Finally, when all three covariances are non-zero, then X is endogenous and our test has statistical power. In fact, as mentioned in the Introduction, this is likely to be the most empirically relevant setting as well. In our empirical setting of welfare program participation and birth weight, unobservable mothering ability, or conscientiousness of a mother, is likely to be part of U as it would affect birth weight (Y). At the same time, it is likely to determine participation in WIC, and hence also be part of V, implying $\sigma_{uv} \neq 0$. As less conscientious mothers are likely to smoke higher amounts of cigarettes as well, this will induce a correlation between both V and W^* ($\sigma_{vw^*} \neq 0$) and U and $W^*(\sigma_{uw^*} \neq 0)$. Intuitively, the interaction of V with the bunching variable W provides us with the required statistical power, which we establish in the remainder of this section.

Table 1 presents simulation results for the above DGP with $\kappa = 0$. Under the null hypothesis of no endogeneity, columns (1) to (4) and (7) when $\sigma_{uv} = 0$, we report rejections of between 6 to 7% for both sample sizes, which signify a slight size distortion for our test. It is crucial to note that the effective number of observations within the bandwidths is around 240 for treated and 135 for the control group in Panel A, and around 600 for treated and 340 for the control group in Panel B.

The remaining cells in Table 1 present the DGP under the alternative hypothesis. When both σ_{vw^*} and σ_{uw^*} are non-zero, columns (5), (6) and (8), (9), the rejection rates rise with the degree of endogeneity in the structural equation as captured by increasing values of σ_{uv}. In Panel B, with a sample size of 5000, we achieve rejection rates of over 90% for these empirically relevant cases.[18]

Our test performs less than ideally in columns (1) to (3) when $\sigma_{uw^*} = 0$, and X is endogenous, i.e. $\sigma_{uv} > 0$ where even with high correlation between V and W^*, rejection rates are under 30%. This is despite Fig. 9.1 depicting that analytically we should have power as shown by the dashed blue lines. However, we believe it is difficult to encounter empirical settings where such configuration of the covariances is likely to hold.

Finally, when $\sigma_{vw^*} = 0$, columns (1), (4), and (7), even with massive endogeneity our rejection rates remain below 10%. This remains true in Panel B as well with the higher sample size. This is the scenario that is depicted by the green

[18] In Tables B.1–B.3, we present findings for the additively separable case for different chosen bandwidths as well. The results remain consistent across DGPs. This is true even though the effective number of observations varies substantially across different bandwidths.

Table 3
Power analysis – U additively separable case – W rounded off 2 decimal places.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $N = 2000$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.067	0.067	0.070	0.079	0.135	0.219	0.072	0.282	0.527
0.2	0.064	0.072	0.073	0.066	0.156	0.245	0.066	0.309	0.513
0.4	0.067	0.082	0.079	0.069	0.198	0.288	0.074	0.371	0.572
0.6	0.068	0.090	0.100	0.064	0.257	0.373	0.074	0.467	0.666
0.8	0.064	0.117	0.160	0.062	0.368	0.519	0.079	0.624	0.800
Panel B: $N = 5000$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.063	0.064	0.067	0.064	0.239	0.415	0.066	0.558	0.876
0.2	0.066	0.075	0.073	0.066	0.301	0.484	0.065	0.621	0.877
0.4	0.067	0.084	0.094	0.065	0.383	0.587	0.066	0.707	0.911
0.6	0.064	0.122	0.140	0.067	0.504	0.708	0.072	0.819	0.955
0.8	0.068	0.191	0.260	0.071	0.685	0.857	0.079	0.935	0.990

Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section 7.1. In Panel A the number of distinct values of W fall by around 66% on average, and in Panel B by 82% relative to the baseline case in Table 1.

dashed lines in Fig. 9.1. Intuitively, this stems from the fact that the bunching variable W^* carries no information about the unobservables in equation (2), V, whose correlation with the unobservables, U, in equation (1) causes the endogeneity of X.

Since the configuration of the alternative hypothesis is slightly complex, Fig. B.1 presents only a subset of the results in Table 1. In both panels we restrict to the DGPs where the origin always represents the null hypothesis. In panel (a), as σ_{vw^*} increases along the x-axis, the rejection rates rise as well. However, as shown in Table 1, they approach 1 only when σ_{uw^*} is sufficiently high as well. Fig. B.1(b) presents the results for when we fix $\sigma_{uw*} = 0$, and in this scenario our test performs less than ideally. However, as mentioned earlier, in most empirical settings we expect all three covariances involved to be non-zero. In such a scenario our test performs really well as argued above.

7.3. Additively non-separable case: $\kappa \neq 0$

The test that we develop in this paper is a joint test of both the selection on observables assumption and of misspecification of functional form in terms of the relationship between X and U. We next present the performance of our test under the setting where κ is equal to one-fourth of the partial effect of X on Y.[19] In Table 2, columns (5), (6), (8), and (9), when all three covariances are non-zero, we see substantial improvements in the rejection rates. In addition, in columns (4) and (7), where $\sigma_{vw^*} = 0$, and where we had no power in the additively separable case, we now see rejection rates above 50%. For $N = 5000$ in Panel B, rejection rates approach 90%. This is because in these DGPs σ_{uw^*} is non-zero, and hence W^* contains information regarding misspecification which is operating through the multiplicative XU term in equation (1). This is also depicted by the fact that in columns (1) to (3) when $\sigma_{uw^*} = 0$ we do not have power to detect misspecification.

7.4. Extensions

In this section we present some extensions to the basic Monte Carlo setup discussed above. Ideally, in our empirical setting we would want a bunching variable which is continuous throughout its domain and has a known mass point. However, at times it is difficult to have 'truly' continuous variables as is the case with smoking during pregnancy, which is measured in discrete units of cigarettes, or for years of education or experience. Thus, we now explore whether our test is likely to suffer under such settings. In Table 3, we keep the above Monte Carlo setup but round off the bunching variable W to 2 decimal places. This reduces the number of distinct values of W by 66% and 82% in Panel A and B, respectively. The performance of our tests is very similar to the baseline case in Table 1. In Table B.4, we repeat this exercise where we round off to 1 decimal place, reducing the number of distinct values by over 95%. Our rejection rates fall slightly for the smaller sample size of 2000 but remain robust for the larger sample of 5000 observations.

Finally, we also consider the case by explicitly including covariates in the DGP, and repeating the simulation exercise. We add a covariate Z to the above DGP, equation (1), with a partial effect depicted by π. We construct it as follows, $Z = \tau_0 + \tau_1 X + \tau_2 W + \varepsilon$, where ε is a standard normal variable uncorrelated with the other unobservables in the system. For brevity, Table 4 only presents the results for the small sample size of 2000. In Panel A, we set $\pi = 1$, which is slightly less than the partial effect of X on Y. This is the standard way of modeling omitted variables. And as Panel A shows, our rejection rates fall drastically across DGPs, though are still substantially high. This can be reflective of the case where the

[19] We repeat the same exercise for when $\kappa = 0.5\beta$ and the results are qualitatively similar to those presented here.

Table 4
Power analysis — U additively separable case with covariate $Z = \tau_0 + \tau_1 X + \tau_2 W + \varepsilon$.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $\pi = 1$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.062	0.066	0.068	0.069	0.098	0.134	0.068	0.152	0.249
0.2	0.065	0.067	0.073	0.064	0.107	0.152	0.067	0.169	0.262
0.4	0.070	0.071	0.080	0.065	0.127	0.173	0.069	0.187	0.297
0.6	0.067	0.076	0.081	0.067	0.148	0.208	0.070	0.211	0.335
0.8	0.065	0.088	0.094	0.065	0.172	0.253	0.067	0.271	0.396
Panel B: $\pi = 4.5$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.061	0.070	0.071	0.065	0.066	0.075	0.066	0.077	0.088
0.2	0.065	0.070	0.070	0.061	0.070	0.073	0.066	0.076	0.082
0.4	0.066	0.066	0.071	0.071	0.072	0.076	0.067	0.079	0.092
0.6	0.067	0.068	0.069	0.070	0.075	0.077	0.065	0.079	0.087
0.8	0.067	0.073	0.070	0.065	0.075	0.084	0.065	0.081	0.094

Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section 7.1. Results reported for only $N = 2000$.

omitted variable has limited explanatory power. In Panel B, we increase the values of π to three times that of β or the effect of X on Y. Under this scenario all rejection rates fall well below 10%, even in the case when there is substantial presence of endogeneity depicted by high values of σ_{uv}, and we have power to detect them as in columns (5), (6), and (8), (9). Overall, this shows that the test performs precisely as expected when we add a covariate to the data generating process.

8. Empirical application

8.1. Background

A healthy intrauterine environment is considered to be of critical importance for positive birth outcomes. Low birth weight and other complications at birth are linked to significant health costs especially during infancy and early childhood.[20] Given these concerns, the U.S. government operates an extensive welfare program: the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC). The program targets low-income pregnant mothers, and provides food supplements, nutrition education, and access to health services with the objective of improving birth outcomes.

Nutritional risk is determined by an income threshold, but due to a lack of data on actual income levels for participants, concerns about selection into treatment are hard to deal with.[21] Moreover, given a lack of potential exclusion restrictions, most of the literature has resorted to using a selection on observables approach, and finds treatment effects on average birth weight ranging from no effect to gains upwards of 60 g (Bitler et al., 2005; Figlio et al., 2009).[22] The framework developed in our paper is thus ideally suited to studying the above problem. We have a continuous outcome variable in terms of birth weight, a binary treatment variable in terms of WIC participation, and we use smoking during pregnancy as our discontinuously distributed and potentially endogenous variable with bunching at zero.

8.2. Data

We use the Vital Statistics Data that compiles information from birth certificates of all infants born in the United States in a given year. Along with birth outcomes and WIC participation, we observe detailed information on the demographics of the parents, current and past pregnancies, prenatal care, mother's smoking behavior, etc. We pool together cross-sectional data from 2010–2012 covering more than 80% of all births in the U.S. in the given time period. In this pooled sample, 47% of mothers were on WIC during their current pregnancy, signifying the immense scope of the federal aids program.

[20] A review of the literature by Almond and Currie (2011) even establishes important links between poor birth outcomes and health and human capital accumulation well into adulthood.

[21] In our data set, we find that WIC participants are more likely to be teenagers (6pp), more likely to be unmarried (7pp). 32.7% of mothers on WIC are high school dropouts compared to 23.6% in the controls, and 8.7% went to college compared to 17% for the nonparticipants. Very similar patterns hold for fathers as well. Thus, non-random selection into the program is a valid concern.

[22] Figlio et al. (2009) is one of the few papers which exploits an exclusion restriction to identify the effect of WIC participation on birth outcomes.

8.3. Estimation

We first flexibly control for a wide variety of observables which can explain participation into WIC. Specifically, the set of other covariates, Z, includes parental age, race, education, marital status, various interactions between the demographic variables of the mother, measures of prenatal care, risk factors during the current pregnancy, and whether she had a poor outcome for a previous pregnancy. We also control for a cubic polynomial in prepregnancy BMI, non-parametric controls for gestation, and flexible controls for mother's smoking behavior across trimesters during pregnancy.[23]

We use mother's smoking behavior in the third trimester of pregnancy as the bunching variable W with bunching at zero, which manifests itself in a discontinuity in birth weight as well. Fig. 9.2(a) plots conditional means of birth weight for each level of cigarette smoked for both WIC and non-WIC participants. We see a large discontinuity at zero cigarettes smoked for both groups of observations.

After removing the direct effect of our extensive set of covariates for both treatment and control groups from birth weight Y, we separately implement a local linear estimator on the 'cleaned' variable $Y - Z^\top \gamma$ with a bandwidth of 4 and the standard Epanechnikov kernel. We first present the test statistic using only a basic set of controls in Z. These mainly include information on the demographics of the parents and other controls which are readily available in most data sets that record birth outcomes.[24]

The test statistic from this basic specification in Fig. 9.3(a), using third trimester smoking as W, is statistically significantly different from zero at -20.17 g, implying the existence of substantial amount of selection even after a fairly detailed set of covariates. The test statistic is even larger for a sparser set of covariates. For instance, if we control only for mother's race, it is upwards of -40 g. Fig. 9.3(b) next presents results from the full specification detailed above. Most importantly, it includes controls for previous and current pregnancy characteristics, smoking behavior across pregnancy, and various interactions involving demographic variables.[25] The value of the test statistic falls down to -1.64 and is statistically indistinguishable from zero, indicating a substantial decrease in potential selection concerns.

As an alternative specification, we try the above analysis with W as prepregnancy smoking behavior instead of third trimester smoking. To the extent that WIC participation affects smoking cessation differently for participants and non-participants, even after controlling for a rich set of covariates, one might consider the prepregnancy behavior as more suited to our test. Fig. 9.3(c) and (d) thus present our test statistic with prepregnancy smoking as our bunching variable W. Because of the detailed set of controls that we use in Z, we fail to reject null hypothesis of the failure of the selection on observables assumption as seen in Fig. 9.3(d), with a test statistic of 2.11.

We next extend our results in Table 5 for various bandwidths and degrees of the polynomial for both third trimester and prepregnancy smoking. We use both the basic and detailed specification. Results show more robust results for the local linear estimator, as expected, given the spread of data shown in 9.3. However, for either prepregnancy or third trimester smoking as W, the estimated discontinuities are relatively small in all cases, especially under the local linear estimator, and we fail to reject the null hypothesis. This implies that the selection on observables assumption is likely to hold when we use our most detailed specification.

8.4. Test statistic using average smoking

Given the requirement that W needs to be continuous, one might be concerned about using the trimester specific measure. We construct an average smoking during pregnancy measure that calculates the mean of smoking across all three trimesters for mothers in our sample. This variable is necessarily 'more' continuous than smoking during a given trimester, which is recorded in increments of one cigarette, whereas the average measure is in increments of 1/3rd of a cigarette.

Fig. 9.4 shows a much noisier scatter compared to the single trimester scatters in Fig. 9.3. However, this is not surprising given that we are slicing the data into finer bins.[26] Fig. 9.4 (a) presents a higher value of the test statistic for the basic specification than the one in (b), which uses the full specification. However, in this case the former is statistically insignificant as well. Similarly, Table B.5 then presents results for various combinations of bandwidths for both the local linear and cubic case. Results for the local linear case tell a similar story as the single trimester analysis and we fail to reject the null for our most preferred specification in all instances except one. However, one concern is that the estimated test statistic is noisier compared to the results in Table 5. But this again is not surprising, given the flux in the scatter around zero for the average smoking case.

9. Conclusion

In this paper, we developed a joint test of additive separability of the outcome equation in treatment and unobservables as well as the selection on observables assumption. The treatment variable of interest could be of any type. We develop

[23] We employ a similar specification as the one used first by Almond et al. (2005).

[24] Like birth order, information on prenatal care, gestation and linear controls for smoking three months prior to the pregnancy as well as in the first two trimesters. However, we still remain extremely flexible in specifying how these covariates affect birth weight.

[25] The basic specification only includes linear controls for smoking in other trimester, however, in our full specification we flexibly control for various 'types' of mothers as depicted by their changing smoking behavior across trimesters.

[26] We top-code the plot at 15 average cigarettes, since there are very few mothers that smoke more than this.

Table 5
Test statistic with varying bandwidths.

	Bandwidth			
	3	4	5	6
Panel A: 3rd trimester smoking				
Degree = 1				
Basic specification	−17.10	−20.17∗	−22.37∗∗	−19.18∗∗
	(12.64)	(8.468)	(7.028)	(5.384)
Full specification	−2.615	−1.642	−4.927	−3.308
	(12.14)	(8.223)	(6.825)	(5.229)
Degree = 3				
Basic specification	−17.18∗∗	−16.63	−13.05	−1.018
	(6.539)	(10.16)	(41.54)	(26.51)
Full specification	−2.393	−2.753	−16.63	12.00
	(6.343)	(9.850)	(40.29)	(25.71)
Panel A: Prepregnancy smoking				
Degree = 1				
Basic specification	23.12	17.27∗	18.37∗∗	16.62∗∗
	(12.58)	(8.482)	(6.981)	(5.452)
Full specification	9.141	2.111	1.098	2.975
	(12.21)	(8.233)	(6.777)	(5.294)
Degree = 3				
Basic specification	17.79∗∗	21.35∗	50.61	22.20
	(6.594)	(10.29)	(42.53)	(26.57)
Full specification	4.508	8.198	37.16	23.36
	(6.372)	(9.948)	(41.10)	(25.68)

∗∗, ∗ represents significance at the 1% and 5% level, respectively. Treated group has 4,488,328 while the control group has 4,950,406.

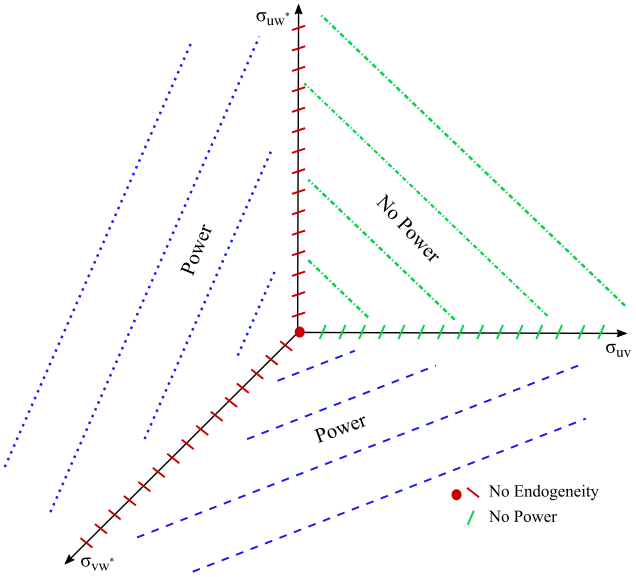
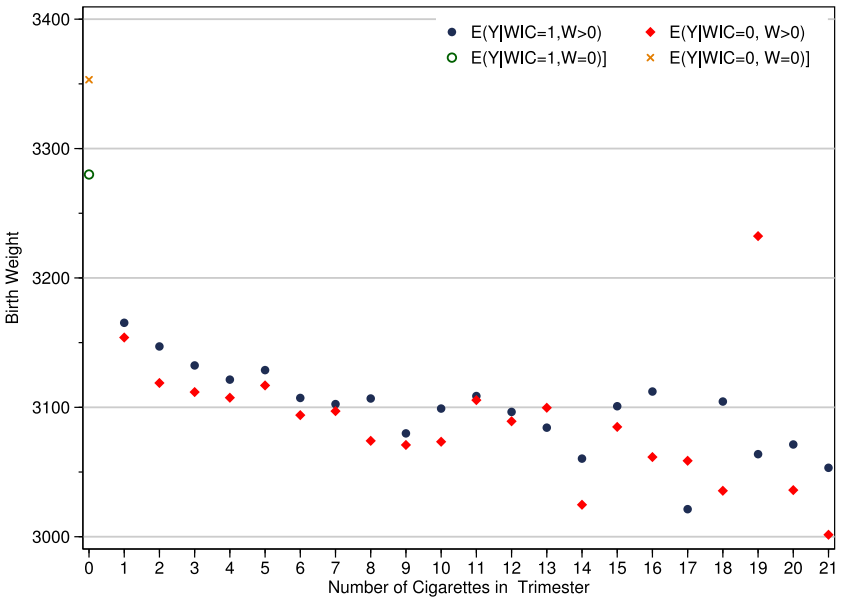
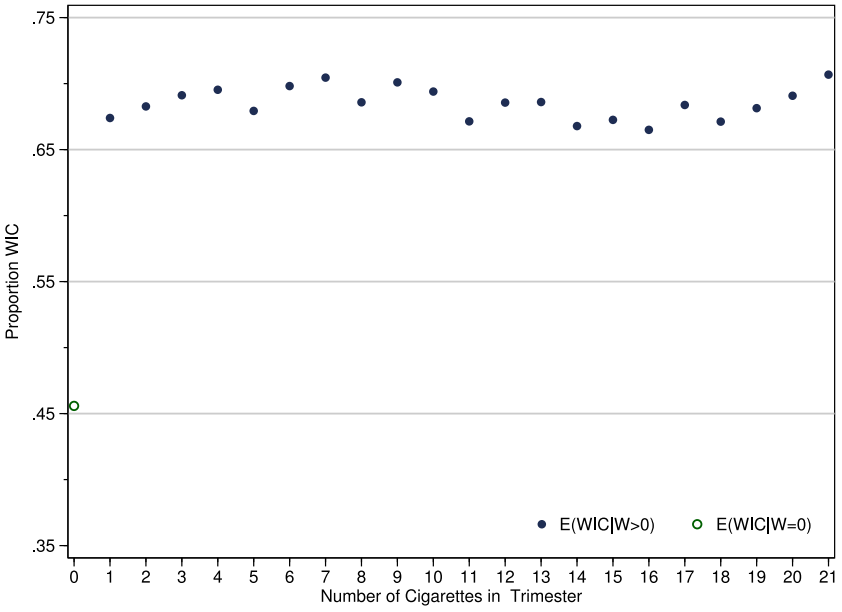


Fig. 9.1. Illustration of statistical power.

formal testing procedures for binary and finite X hinging crucially on two conditions. First, there has to be a variable, W, among the set of controls that has a positive probability taking a known value but is otherwise continuously distributed. Second, the structural function relating W to the outcome of interest, Y, must be continuous in W. For our testing procedure to have power, the expected outcome, Y, conditional on W, treatment X and other possible controls (Z), has to be discontinuous in W at the bunching point under the alternative for at least some values of the treatment variable. In other words, we need W to be endogenous under our alternative hypothesis. Moreover, the endogeneity of W has to interact with that of X when X is endogenous (this last part is a testable condition). Since W is not the treatment variable, it could also be endogenous under the null. The test then checks whether the discontinuity in the expected outcome conditional on W, treatment, and other possible controls is the same for treated and untreated individuals. We also outline extensions of our testing idea to nonparametric nonseparable models under a weak monotonicity assumption.



(a) Mean Birth Weight



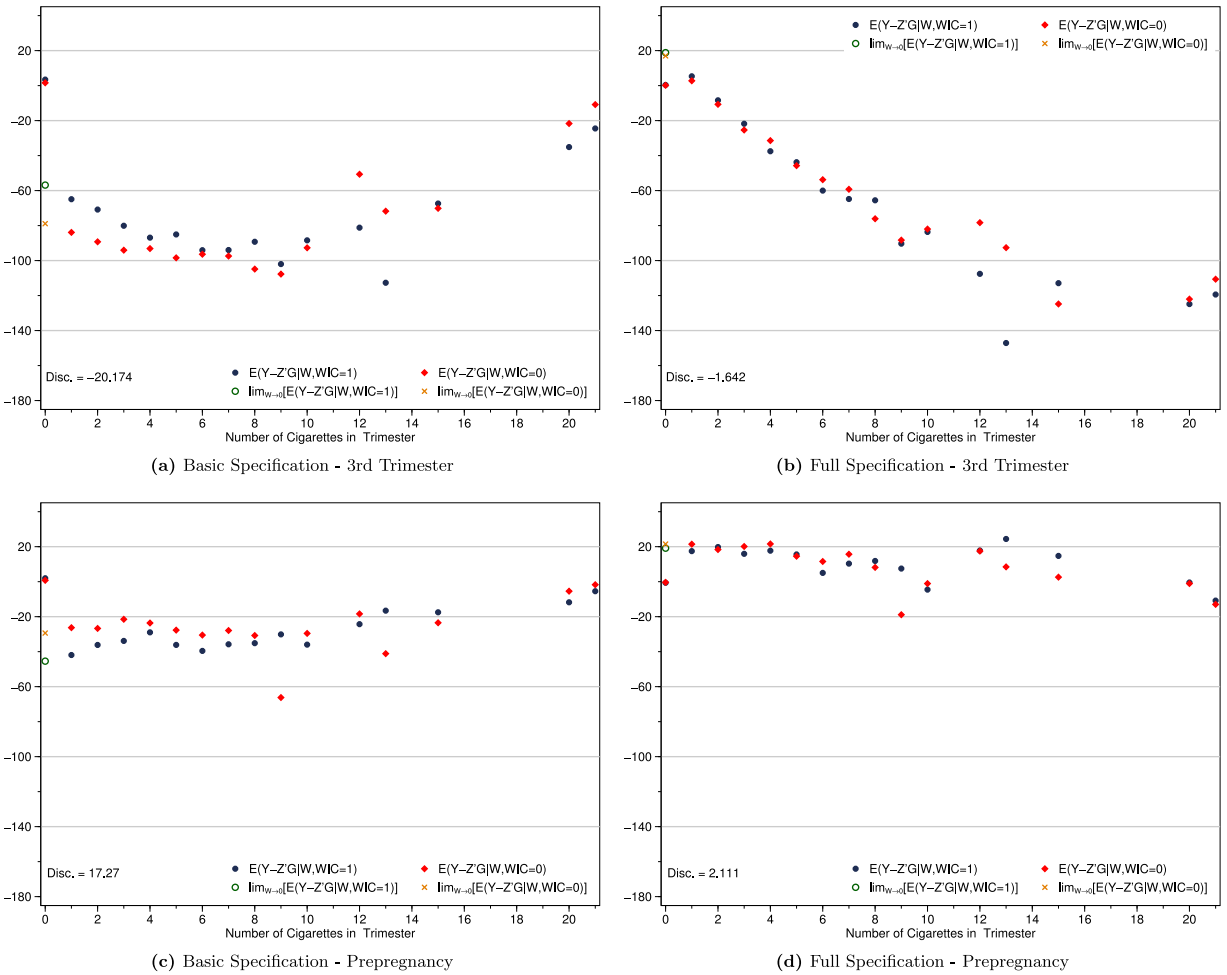
(b) Proportion WIC

Fig. 9.2. Mean birth weight and WIC participation by 3rd trimester smoking.

The testing procedures we suggest are easy to implement, and the assumptions under which our testing procedures work are likely to hold in many empirical situations. As such we expect that our paper will be appealing to empirical economists.

Acknowledgments

We would like to thank the editor and the referees for their valuable feedback. We are also extremely grateful to our anonymous associate editor for their thorough and detailed reading of our work. Many of the results of this paper were presented as part of a larger project at 2015 Winter Meetings of the Econometric Society, 2015 IAAE Conference, 2015 Econometric Society World Congress, University of Toronto, UT at Austin, University of Western Ontario, Ohio



(a) Basic Specification - 3rd Trimester

(b) Full Specification - 3rd Trimester

(c) Basic Specification - Prepregnancy

(d) Full Specification - Prepregnancy

Fig. 9.3. Test statistic for WIC participation.

State University, North Carolina State University, Iowa State University, California Institute of Technology and University of Missouri. We would like to thank participants of those seminars for valuable comments and questions. Neşe Yıldız gratefully acknowledges financial support through The National Science Foundation, United States of America grant SES-1918985.

Appendix A

A.1. Proof of Theorem 2

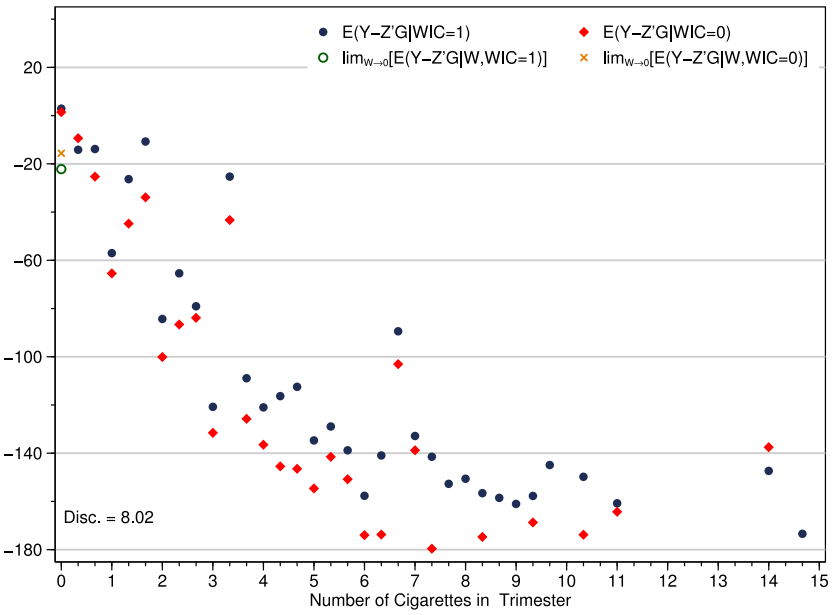
Define the infeasible estimators that use $\tilde{Y}_i$ instead of $\widehat{\tilde{Y}}_i$ as

$$\widehat{\mu}_{\tilde{Y}|X,W}(x,0) := \frac{1}{n_{x0}} \sum_{i=1}^{n} \tilde{Y}_i 1\{X_i = x, W_i = 0\},$$

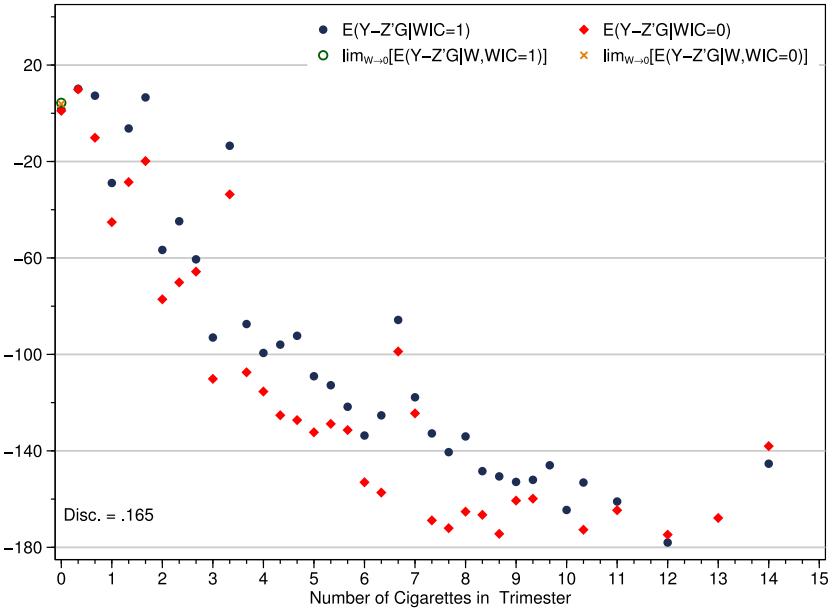
$$\widehat{\mu}_{\tilde{Y}|X,W}(x,0^+) := \frac{1}{n_x} e_1^\top \operatorname*{argmin}_{a_0,a_1} \sum_{i=1}^{n} (\tilde{Y}_i - a_0 - a_1 W_i/h)^2 K_h(W_i) 1\{W_i > 0, X_i = x\}.$$

We first start by analyzing the large sample behavior of the infeasible test statistic and note that

$$\sqrt{nh}\left[\widehat{\mu}_{\tilde{Y}|X,W}(1,0^+) - \widehat{\mu}_{\tilde{Y}|X,W}(1,0) - \left(\widehat{\mu}_{\tilde{Y}|X,W}(0,0^+) - \widehat{\mu}_{\tilde{Y}|X,W}(0,0)\right)\right]$$



(a) Basic Specification - Avg. Trim. 1,2,3



(b) Full Specification - Avg. Trim. 1,2,3

Fig. 9.4. Test statistic for WIC participation — Average smoking.

$$=\sqrt{nh}\left[\widehat{\mu}_{\tilde{Y}|X,W}(1,0^{+})-\mu_{\tilde{Y}|X,W}(1,0^{+})-\left(\widehat{\mu}_{\tilde{Y}|X,W}(0,0^{+})-\mu_{\tilde{Y}|X,W}(0,0^{+})\right)\right] \tag{A.1}$$

$$-\sqrt{nh}\left[\widehat{\mu}_{\tilde{Y}|X,W}(1,0)-\mu_{\tilde{Y}|X,W}(1,0)-\left(\widehat{\mu}_{\tilde{Y}|X,W}(0,0)-\mu_{\tilde{Y}|X,W}(0,0)\right)\right] \tag{A.2}$$

$$+\sqrt{nh}\left[\mu_{\tilde{Y}|X,W}(1,0^{+})-\mu_{\tilde{Y}|X,W}(1,0)-\left(\mu_{\tilde{Y}|X,W}(0,0^{+})-\mu_{\tilde{Y}|X,W}(0,0)\right)\right]. \tag{A.3}$$

Since,

$$\mu_{\tilde{Y}|X,W}(1,0^+) - \mu_{\tilde{Y}|X,W}(1,0) = \mu_{U|W}(0^+) - \mu_{U|W}(0),$$
$$\mu_{\tilde{Y}|X,W}(0,0^+) - \mu_{\tilde{Y}|X,W}(0,0) = \mu_{U|W}(0^+) - \mu_{U|W}(0),$$

with $\mu_{U|W}(0^+) := \lim_{w\downarrow 0} \mathbb{E}(U|W=w)$, (A.3) equals 0. Moreover, by the Law of Large Numbers $\widehat{\mu}_{\tilde{Y}|X,W}(1,0) - \mu_{\tilde{Y}|X,W}(1,0) - \left(\widehat{\mu}_{\tilde{Y}|X,W}(0,0) - \mu_{\tilde{Y}|X,W}(0,0)\right) = O_P\left(n^{-1/2}\right)$, and therefore, (A.2) is $o_P(1)$.

To analyze the asymptotic behavior of (A.1), we recall that

$$\widehat{\mu}_{\tilde{Y}|X,W}(x,0^+) = \frac{1}{n_x}\sum_{i=1}^{n} e_1^\top M_{nx}^{-1} L_{ix} K_h(W_i)\tilde{Y}_i = e_1^\top M_{nx}^{-1}\frac{1}{n_x}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i.$$

where

$$L_{ix} := (1, W_i/h)^\top 1\{W_i > 0, X_i = x\},$$
$$M_{nx} := \frac{1}{n_x}\sum_{i=1}^{n} L_{ix}L_{ix}^\top K_h(W_i),$$

Lemma 1. *Suppose the conditions of Theorem 2 hold. Then*

$$\sqrt{nh}\left(\widehat{\mu}_{\tilde{Y}|X,W}(1,0^+) - \mu_{\tilde{Y}|X,W}(1,0^+) - (\widehat{\mu}_{\tilde{Y}|X,W}(0,0^+) - \mu_{\tilde{Y}|X,W}(0,0^+))\right) \xrightarrow{d} N(0,V). \tag{A.4}$$

Proof. First, we note that

$$\begin{aligned}\sqrt{nh}e_1^T M_{nx}^{-1}\frac{1}{n_x}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i &= \sqrt{n}\left(\frac{1}{\frac{n_x}{n}} - \frac{1}{\mathbb{P}(X=x)}\right)\sqrt{h}e_1^T M_{nx}^{-1}\frac{1}{n}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i \\ &\quad + \sqrt{nh}e_1^T M_{nx}^{-1}\frac{1}{n\mathbb{P}(X=x)}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i \\ &= \sqrt{nh}e_1^T M_{nx}^{-1}\frac{1}{n\mathbb{P}(X=x)}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i + o_P(1).\end{aligned} \tag{A.5}$$

In addition,

$$\begin{aligned}M_{nx} &= \frac{1}{n_x/n}\frac{1}{n}\sum_{i=1}^{n} L_{ix}L_{ix}^\top K_h(W_i) \xrightarrow{P} \frac{1}{\mathbb{P}(X=x)}\mathbb{E}(L_{ix}L_{ix}^\top K_h(W_i)) \\ &= \mathbb{E}(L_{ix}L_{ix}^\top K_h(W_i)|X_i = x) =: N_{nx}.\end{aligned} \tag{A.6}$$

Define

$$S_{nx} = \frac{1}{\mathbb{P}(X_i=x)}\frac{1}{n}\sum_{i=1}^{n} e_1^\top N_{nx}^{-1} L_{ix}K_h(W_i)\varepsilon_i.$$

Note that

$$\mathbb{E}\left[L_{ix}K_h(W_i)\frac{\varepsilon_i}{\mathbb{P}(X_i=x)}\right] = 0.$$

Then using standard results as in Masry (1996), for example, we have

$$e_1^\top M_{nx}^{-1}\frac{1}{n\mathbb{P}(X=x)}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i = \mu_{\tilde{Y}|W,X}(0^+,x) + S_{nx} + O(h^2) + O_P\left(\frac{\log(n)}{nh}\right), \tag{A.7}$$

where

$$\mu_{\tilde{Y}|X,W}(x,0^+) := \lim_{w\downarrow 0}\int_{-\infty}^{\infty} y\frac{f_{\tilde{Y},W|X}(y,w|x)}{f_{W|X}(w|x)}dy. \tag{A.8}$$

These arguments show that the asymptotic distribution of our test statistic will be determined by the limiting distribution of $\sqrt{nh}(S_{n1} - S_{n0})$. Letting $p = \mathbb{P}(X=1)$ we can write

$$\sqrt{nh}(S_{n1} - S_{n0}) = e_1^\top\frac{1}{p}N_{n1}^{-1}\sum_{i=1}^{n}\sqrt{\frac{h}{n}}L_{i1}K_h(W_i)\varepsilon_i - e_1^\top\frac{1}{1-p}N_{n0}^{-1}\sum_{i=1}^{n}\sqrt{\frac{h}{n}}L_{i0}K_h(W_i)\varepsilon_i.$$

Below we will argue that

$$\sum_{i=1}^{n}\sqrt{\frac{h}{n}}L_{ix}K_h(W_i)\varepsilon_i = O_P(1).$$

for $x = 0, 1$. As a result,

$$\begin{aligned}\sqrt{nh}(S_{n1} - S_{n0}) &= e_1^\top A^{-1}\frac{1}{pf_{W|X}(0^+|1)}\sum_{i=1}^{n}\sqrt{\frac{h}{n}}L_{i1}K_h(W_i)\varepsilon_i \\ &\quad - e_1^\top A^{-1}\frac{1}{(1-p)f_{W|X}(0^+|0)}\sum_{i=1}^{n}\sqrt{\frac{h}{n}}L_{i0}K_h(W_i)\varepsilon_i + o_P(1), \\ &=: \sum_{i=1}^{n} T_{ni} + o_P(1),\end{aligned}$$

with

$$A = \begin{bmatrix}\kappa_0 & \kappa_1 \\ \kappa_1 & \kappa_2\end{bmatrix},$$

where κ_j for $j = 0, 1, 2$ is as in Assumption 3(vii). We will apply the Lindeberg–Feller Theorem to $\sum_{i=1}^{n} T_{ni}$. Note that when we compute $E(T_{ni}^2)$ the cross terms will be 0 since $1\{X_i = 1\}1\{X_i = 0\} = 0$. Also note that

$$e_1^\top A^{-1}\begin{pmatrix}1 \\ \frac{W_i}{h}\end{pmatrix} = \frac{\kappa_2 - \kappa_1\frac{W_i}{h}}{\kappa_0\kappa_2 - \kappa_1^2}.$$

$$\begin{aligned}Var(\sum_{i=1}^{n} T_{ni}) &= \mathbb{E}\left[\frac{(\kappa_2 - \kappa_1\frac{W_i}{h})^2}{(\kappa_0\kappa_2 - \kappa_1^2)^2 p[f_{W|X}(0^+|1)]^2}\frac{1}{h}K^2\left(\frac{W_i}{h}\right)1\{W_i > 0\}\sigma^2_{\varepsilon|W,X}(W_i, 1)|X_i = 1\right] \\ &\quad + \mathbb{E}\left[\frac{(\kappa_2 - \kappa_1\frac{W_i}{h})^2}{(\kappa_0\kappa_2 - \kappa_1^2)^2 (1-p)[f_{W|X}(0^+|0)]^2}\frac{1}{h}K^2\left(\frac{W_i}{h}\right)1\{W_i > 0\}\sigma^2_{\varepsilon|W,X}(W_i, 0)|X_i = 0\right].\end{aligned}$$

Using the standard change of variables argument with ν_0, ν_1, ν_2 as in Assumption 3(vii) we get

$$Var\left(\sum_{i=1}^{n} T_{ni}\right) \to \frac{\kappa_2^2\nu_0 - 2\kappa_2\kappa_1\nu_1 + \kappa_1^2\nu_2}{(\kappa_0\kappa_2 - \kappa_1^2)^2}\left[\frac{\sigma^2_{\varepsilon|W,X}(0^+, 1)}{pf_{W|X}(0^+|1)} + \frac{\sigma^2_{\varepsilon|W,X}(0^+, 0)}{(1-p)f_{W|X}(0^+|0)}\right].$$

To apply the Lindeberg–Feller Theorem we also need to verify that

$$\sum_{i=1}^{n}\mathbb{E}\left(T_{ni}^2 1\{|T_{ni}| > \epsilon\}\right) \to 0,$$

for each $\epsilon > 0$.This is bounded by
$(P(|T_{ni}| > \epsilon))^{(1+\alpha)/(2+\alpha)}\sum_{i=1}^{n}\left(\mathbb{E}\left(|T_{ni}|^{2+\alpha}\right)\right)^{1/(2+\alpha)} \to 0$ by Hölder's inequality. □

Lemma 2.

$$\sqrt{nh}\left(\widehat{\mu}_{\widehat{\tilde{Y}}|X,W}(x, 0^+) - \widehat{\mu}_{\tilde{Y}|X,W}(x, 0^+)\right) = o_P(1), \text{ and}$$
$$\sqrt{nh}\left(\widehat{\mu}_{\widehat{\tilde{Y}}|X,W}(x, 0) - \widehat{\mu}_{\tilde{Y}|X,W}(x, 0)\right) = o_P(1).$$

Proof.

$$\begin{aligned}&\sqrt{nh}\left(\widehat{\mu}_{\widehat{\tilde{Y}}|X,W}(x, 0^+) - \widehat{\mu}_{\tilde{Y}|X,W}(x, 0^+)\right) \\ &= -\sqrt{h}e_1^\top M_{nx}^{-1}\frac{1}{n_x}\sum_{i=1}^{n} L_{ix}K_h(W_i)Z_i^\top\sqrt{n}(\widehat{\gamma} - \gamma) = o_P(1)O_P(1) = o_P(1).\end{aligned}$$

Finally, recall $n_{x0} := \sum_{i=1}^{n} 1\{W_i = 0, X_i = x\}$, and note that

$$\sqrt{nh}\frac{1}{n_{x0}}\sum_{i=1}^{n}\left(\widehat{\tilde{Y}}_i - \tilde{Y}_i\right)1\{W_i = 0, X_i = x\}$$

Table B.1
Power analysis – U additively separable case – $2 * Bandwidth$.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $N = 2000$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.068	0.072	0.073	0.071	0.127	0.200	0.074	0.241	0.471
0.2	0.066	0.071	0.073	0.069	0.136	0.210	0.069	0.257	0.442
0.4	0.067	0.079	0.075	0.071	0.166	0.238	0.073	0.294	0.471
0.6	0.071	0.078	0.079	0.070	0.201	0.280	0.072	0.359	0.539
0.8	0.063	0.092	0.101	0.064	0.266	0.379	0.077	0.488	0.659
Panel B: $N = 5000$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.065	0.066	0.071	0.067	0.212	0.373	0.070	0.484	0.821
0.2	0.067	0.075	0.071	0.068	0.246	0.396	0.068	0.516	0.800
0.4	0.071	0.074	0.073	0.066	0.299	0.465	0.071	0.583	0.825
0.6	0.067	0.093	0.090	0.066	0.381	0.547	0.070	0.687	0.883
0.8	0.070	0.124	0.126	0.066	0.521	0.697	0.072	0.831	0.946

Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section 7.1. Missing entries correspond to combinations of the parameter space where the variance covariance matrix is not positive semi-definite.

$$= -\sqrt{nh}\frac{1}{n_{x0}}\sum_{i=1}^{n} Z_i^\top(\widehat{\gamma} - \gamma) = o_P(1). \quad \square$$

The conclusion of Theorem 2 follows from combining the conclusions of Lemmas 1 and 2.

A.2. Proof of Corollary 2

Proof. Since $\widehat{V} \xrightarrow{P} V$, by Theorem 2, we have

$$\sqrt{nh}\frac{\widehat{\tilde{\lambda}}(1,0)}{\sqrt{\widehat{V}}} = \sqrt{nh}\frac{\left[\widehat{\tilde{\lambda}}(1,0) - \tilde{\lambda}(1,0)\right]}{\sqrt{V}} + \sqrt{nh}\frac{\tilde{\lambda}(1,0)}{\sqrt{\widehat{V}}} + o_P(1).$$

To prove the first statement, note that if $\tilde{\lambda}(1,0)$ is a fixed number different from 0, since $\sqrt{nh} \to \infty$, depending on its sign, $\sqrt{nh}\tilde{\lambda}(1,0)$ will go to either positive or negative infinity as $\sqrt{nh}\frac{\left[\widehat{\tilde{\lambda}}(1,0) - \tilde{\lambda}(1,0)\right]}{\sqrt{V}} = O_P(1)$ by Theorem 2, this implies the first statement.

To prove the second statement, consider a local alternative so that $\tilde{\lambda}(1,0) = \frac{\delta}{\sqrt{nh}}$ with $\delta \neq 0$. Then

$$\sqrt{nh}\frac{\widehat{\tilde{\lambda}}(1,0)}{\sqrt{\widehat{V}}} = \sqrt{nh}\frac{\left[\widehat{\tilde{\lambda}}(1,0) - \tilde{\lambda}(1,0)\right]}{\sqrt{V}} + \frac{\delta}{\sqrt{V}} + o_P(1).$$

Therefore, under such a local alternative,

$$\mathbb{P}(\tilde{t}_n \leq c_{\alpha/2}) \to \Phi\left(c_{\alpha/2} - \frac{\delta}{\sqrt{V}}\right), \text{ and}$$

$$\mathbb{P}(\tilde{t}_n > c_{1-\alpha/2}) = 1 - \mathbb{P}(\tilde{t}_n \leq c_{1-\alpha/2}) \to 1 - \Phi\left(c_{1-\alpha/2} - \frac{\delta}{\sqrt{V}}\right). \quad \square$$

A.3. Description of the asymptotic variance of $\widehat{\Gamma}_n$

The diagonal elements of V are asymptotic variances of $\sqrt{nh}(S^*_{nl} - S^*_{ns})$,[27] which equals $V_l + V_s$, with $V_k = \frac{\kappa_2^2\nu_0 - 2\kappa_2\kappa_1\nu_1 + \kappa_1^2\nu_2}{(\kappa_0\kappa_2 - \kappa_1^2)^2}\frac{\sigma^2_{\varepsilon|W,X}(0^+, x_k)}{\mathbb{P}(X = x_k) f_{W|X}(0^+|x_k)}$ for $k \in \{0, 1, \ldots, K\}$. Each off diagonal element of V corresponds to the asymptotic covariance between $\sqrt{nh}(S^*_{nl} - S^*_{ns})$ and $\sqrt{nh}(S^*_{nt} - S^*_{nr})$, for $(l,s) \neq (t,r)$ and $(l,s) \neq (r,t)$. Note that this asymptotic covariance will be 0, when t and r are each different from both l and s. It will be

(a) V_l, when $l = t$, $l \neq r$, $s \neq t$, $s \neq r$;
(b) $-V_l$, when $l \neq t$, $l = r$, $s \neq t$, $s \neq r$;
(c) $-V_s$, when $l \neq t$, $l \neq r$, $s = t$, $s \neq r$;
(d) V_s, when $l \neq t$, $l \neq r$, $s \neq t$, $s = r$.

[27] Thus, for $j \in \{1, 2, \ldots, \frac{K(K+1)}{2}\}$, $V_{j,j}$ element of this matrix is equal to $V_l + V_s$, with l satisfying $\frac{l(l+1)}{2} \leq j < \frac{(l+1)(l+2)}{2}$ and $s \in \{0, 1, 2, \ldots, l-1\}$.

Table B.2
Power analysis – U additively separable case – $4 * Bandwidth$.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $N = 2000$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.067	0.067	0.072	0.069	0.145	0.232	0.069	0.311	0.556
0.2	0.064	0.076	0.080	0.069	0.183	0.278	0.066	0.353	0.564
0.4	0.064	0.087	0.093	0.066	0.234	0.353	0.075	0.436	0.646
0.6	0.067	0.110	0.137	0.065	0.323	0.472	0.074	0.566	0.765
0.8	0.069	0.160	0.257	0.070	0.466	0.659	0.087	0.739	0.898
Panel B: $N = 5000$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.064	0.064	0.071	0.066	0.263	0.437	0.066	0.610	0.895
0.2	0.069	0.080	0.081	0.069	0.351	0.550	0.068	0.693	0.915
0.4	0.067	0.103	0.131	0.067	0.468	0.679	0.068	0.796	0.952
0.6	0.065	0.163	0.224	0.069	0.618	0.824	0.078	0.902	0.987
0.8	0.073	0.289	0.457	0.079	0.808	0.948	0.094	0.978	0.999

Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section 7.1. Missing entries correspond to combinations of the parameter space where the variance covariance matrix is not positive semi-definite.

Table B.3
Power analysis – U additively separable case – $5 * Bandwidth$.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $N = 2000$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.066	0.067	0.074	0.067	0.151	0.229	0.069	0.319	0.554
0.2	0.064	0.078	0.086	0.069	0.202	0.295	0.068	0.373	0.586
0.4	0.065	0.096	0.111	0.069	0.264	0.395	0.075	0.482	0.685
0.6	0.072	0.128	0.180	0.068	0.366	0.540	0.078	0.622	0.815
0.8	0.073	0.210	0.357	0.078	0.536	0.742	0.098	0.805	0.940
Panel B: $N = 5000$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.066	0.068	0.074	0.067	0.272	0.434	0.068	0.626	0.890
0.2	0.068	0.086	0.089	0.068	0.378	0.582	0.072	0.727	0.924
0.4	0.068	0.119	0.162	0.068	0.518	0.733	0.070	0.838	0.967
0.6	0.069	0.207	0.310	0.074	0.689	0.885	0.085	0.935	0.995
0.8	0.083	0.384	0.633	0.086	0.879	0.979	0.113	0.992	0.999

Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section 7.1. Missing entries correspond to combinations of the parameter space where the variance covariance matrix is not positive semi-definite.

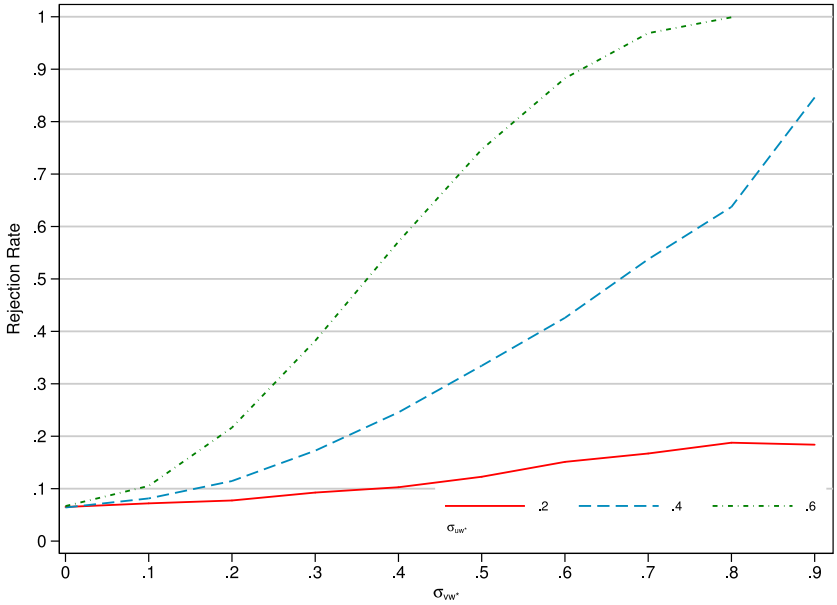
Table B.4
Power analysis – U additively separable case – W rounded off 1 decimal places.

	$\sigma_{uw^*} = 0.0$			$\sigma_{uw^*} = 0.4$			$\sigma_{uw^*} = 0.6$		
Panel A: $N = 2000$	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
σ_{vw^*}	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.069	0.063	0.067	0.068	0.121	0.167	0.072	0.227	0.405
0.2	0.065	0.068	0.072	0.070	0.138	0.194	0.066	0.251	0.403
0.4	0.065	0.075	0.079	0.070	0.167	0.235	0.069	0.300	0.443
0.6	0.067	0.085	0.094	0.063	0.209	0.295	0.069	0.391	0.538
0.8	0.061	0.109	0.143	0.062	0.305	0.414	0.074	0.531	0.681
Panel B: $N = 5000$	0.0	0.4	0.6	0.0	0.4	0.6	0.0	0.4	0.6
$\sigma_{uv} = 0$	0.065	0.065	0.066	0.066	0.197	0.317	0.064	0.461	0.760
0.2	0.064	0.069	0.066	0.064	0.244	0.370	0.065	0.521	0.758
0.4	0.062	0.080	0.084	0.064	0.319	0.472	0.064	0.605	0.809
0.6	0.064	0.111	0.124	0.067	0.426	0.584	0.068	0.725	0.886
0.8	0.069	0.173	0.221	0.072	0.592	0.753	0.080	0.873	0.958

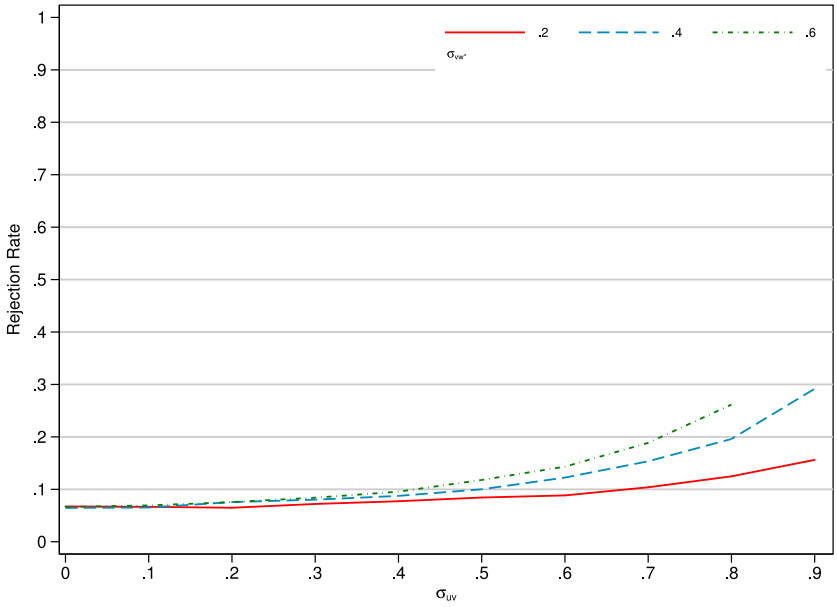
Table entries present empirical rejection rates. Monte Carlo setup and parameter details are given in Section XX. Missing entries correspond to combinations of the parameter space where the variance covariance matrix is not positive semi-definite. In Panel A the number of distinct values of W fall by around 94% on average, and in Panel B by over 97% relative to the baseline case in Table 1.

Appendix B. Monte Carlo extensions

See Tables B.1–B.5 and Fig. B.1.



(a) $\sigma_{uv} = 0; \sigma_{vw^*}, \sigma_{uw^*}\ vary$



(b) $\sigma_{uw^*} = 0; \sigma_{uv}, \sigma_{vw^*}\ vary$

Fig. B.1. Power analysis.

Table B.5
Estimation of test statistic — average smoking across trimesters 1, 2 and 3.

Bandwidth	1.5	2	2.5	3	3.5
Degree = 1					
Basic Specification	8.320	8.020	2.821	0.380	−4.962
	(11.82)	(8.927)	(7.332)	(6.765)	(5.478)
Full Specification	2.832	0.165	−5.039	−7.801	−12.51∗
	(11.49)	(8.673)	(7.123)	(6.572)	(5.322)
Degree = 3					
Basic Specification	74.17	35.54	15.28	15.35	20.84
	(64.65)	(43.28)	(26.16)	(22.24)	(18.08)
Full Specification	61.67	34.97	16.05	12.19	16.34
	(62.71)	(41.98)	(25.37)	(21.58)	(17.54)

** and * represents significance at the 1% and 5% level, respectively. The treated group has 4,488,328 while the control group has 4,950,406.

References

Almond, D., Chay, K.Y., Lee, D.S., 2005. The costs of low birth weight. Q. J. Econ. 120 (3), 1031–1083.
Almond, D., Currie, J., 2011. Killing me softly: The fetal origins hypothesis. J. Econ. Perspect. 25 (3), 153–172.
Bitler, M.P., Hotz, J., Imbens, G.W., Mitnik, O.A., 2005. Does WIC work? The effects of WIC on pregnancy and birth outcomes. Rev. Econ. Stat. 90 (3), 389–405.
Caetano, C., 2015. A test of exogeneity without instrumental variables in models with bunching. Econometrica 83 (4), 1581–1600.
Caetano, G., Kinsler, J., Teng, H., 2019. Towards causal estimates of children's time allocation on skill development. J. Appl. Econometrics 34 (4), 588–605.
Caetano, G., Maheshri, V., 2018. Identifying dynamic spillovers of crime with a causal approach to model selection. Quant. Econ. 9 (1), 343–394.
Caetano, C., Rothe, C., Yıldız, N., 2016. A discontinuity test for identification in triangular nonseparable models. J. Econometrics 193 (1), 113–122.
Crump, R.K., Hotz, V.J., Imbens, G.W., Mitnik, O.A., 2008. Nonparametric tests for treatment effect heterogeneity. Rev. Econ. Stat. 90 (3), 389–405.
Donald, S.G., Hsu, Y.-C., Lieli, R.P., 2014. Testing the unconfoundedness assumption via inverse probability weighted estimators of (L)ATT. J. Bus. Econom. Statist. 32 (3), 395–415.
Figlio, D., Hamersma, S., Roth, J., 2009. Does prenatal WIC participation improve birth outcomes? New evidence from Florida. J. Public Econ. 93 (1–2), 235–245.
Gourio, F., Roys, N., 2014. Size-dependent regulations, firm size distribution, and reallocation. Quant. Econ. 5 (2), 377–416.
Heckman, J.J., Urzua, S., Vytlacil, E., 2006. Understanding instrumental variables in models with essential heterogeneity. Rev. Econ. Stat. 88 (3), 389–432.
Hoderlein, S., Su, L., White, H.L., Yang, T., 2014. Testing for monotonicity in unobservables under unconfoundedness. Available at SSRN 2448681.
Huber, M., 2013. A simple test for the ignorability of non-compliance in experiments. Econom. Lett. 120 (3), 389–391.
Ichimura, H., Lee, L.-F., 1991. Semiparametric least squares estimation of multiple index models: single equation estimation. In: Nonparametric and Semiparametric Methods in Econometrics and Statistics. pp. 3–49.
Imbens, G.W., Rubin, D.B., 2015. Causal Inference in Statistics, Social, and Biomedical Sciences. Cambridge University Press.
Jacob, B., Lefgren, L., Moretti, E., 2007. The dynamics of criminal behavior evidence from weather shocks. J. Hum. Resour. 42 (3), 489–527.
Kitagawa, T., 2015. A test for instrument validity. Econometrica 83 (5), 2043–2063.
Krueger, A.B., Mueller, A., 2010. Job search and unemployment insurance: New evidence from time use data. J. Public Econ. 94 (3–4), 298–307.
Lewbel, A., Lu, X., Su, L., 2015. Specification testing for transformation models with an application to generalized accelerated failure-time models. J. Econometrics 184 (1), 81–96.
Lu, X., White, H., 2014. Testing for separability in structural equations. J. Econometrics 182 (1), 14–26.
de Luna, X., Johansson, P., 2014. Testing for the unconfoundedness assumption using an instrumental assumption. J. Causal Inference 2 (2), 187–199.
Martincus, C.V., Carballo, J., 2010. Beyond the average effects: The distributional impacts of export promotion programs in developing countries. J. Dev. Econ. 92 (2), 201–214.
Robinson, P., 1988. Root-N-consistent semiparametric regression. Econometrica 56 (4), 931–954.
Vytlacil, E., Yıldız, N., 2007. Dummy endogenous variables in weakly separable models. Econometrica 75 (3), 757–779.