A Test Of Selection On Observables Assumption Using A Discontinuously Distributed Covariate*

Umair Khalil† Neşe Yıldız‡

Abstract

We present a test of the selection on observables assumption that does not require the presence of any excluded instruments or any covariates excluded from the structural function. Our test relies on the presence of a variable among the set of controls that has a discontinuous distribution. We develop formal testing procedures for binary and continuous treatment variables. The testing procedures we suggest are easy to implement, and should be useful in many empirical situations. Specifically, we demonstrate how our test handles selection concerns which may potentially complicate the study of the impact of participation in the Supplemental Program for Women, Infants, and Children (WIC) on birth weight.

JEL Numbers: C12, C14, C21.

KEYWORDS: Selection on observables, testing, partial effects, essential heterogeneity.

1 Introduction

Applied economists are often interested in estimating the partial effect of a given variable on an outcome of interest. 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 U. A bulk of studies in the empirical literature work under the crucial assumption of selection on observables, to identify and estimate the partial effect of a variable X on an outcome Y. Specifically, they assume that U is stochastically unrelated to X conditional on the other covariates (controls). This assumption, however, is notoriously difficult to test

*We would like to thank Carol Caetano, Greg Caetano, Joshua Kinsler and Ronni Pavan for useful discussions. 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 State University and North Carolina State University. We would like to thank participants of those seminars for valuable comments and questions.

†Department of Economics, West Virginia University, Morgantown, WV, 26505; Email: umair.khalil@mail.wvu.edu.

‡Corresponding author: Department of Economics, University of Rochester, 231 Harkness Hall, Rochester, NY 14627; Email: nese.yildiz@rochester.edu; Phone: 585-275-5782; Fax: 585-256-2309.

without additional structure. This paper provides a testing procedure for the identifying assumption of the partial effect of X in such a setup.

Our testing procedure hinges on the presence of a covariate, W, whose distribution has a discontinuity at a known point w_0. We are not interested in the average partial effect of W on Y, but on identifying the partial effect of X on Y. Our main variable of interest, X could be any type of random variable, although in our empirical application we focus on the binary X case.

To fix ideas, consider the federal aid program called the Supplemental Program for Women Infants and Children (WIC) which provides nutritional aid and health awareness to pregnant mothers with the aim of improving birth outcomes. While the goals of WIC are clear, the empirical literature has found assessing the true effect of WIC participation difficult, since participation into the program is not random.[1] Thus, having a procedure to judge whether the controls in a given study do a good job eliminating the effects of this possible non-random selection is crucial.

Our outcome variable of interest is an infant's birth weight. Both health policy and the medical literature has largely focused on this variable given that low birth weight can lead to various complications at birth and in turn is linked to significant health costs especially during infancy and early childhood (See, for example, Almond and Currie (2011)). Now suppose that birth weight is a function of smoking during pregnancy (average number of cigarettes smoked daily), WIC participation, other controls like mother's demographics and pregnancy's characteristics plus some unobservable, U. We maintain the assumption that this function is continuous in the smoking variable, W. Around 80% of mothers in our sample do not report smoking anytime during pregnancy.[2] In addition, since it is possible to smoke part of a cigarette, for positive amounts smoked, smoking is a continuous variable.[3] Thus, average cigarettes smoked daily is discontinuously distributed with a discontinuity at 0 and is continuously distributed on the positive side near 0. Moreover, as can be seen in Figures 1 and 2, expected birth weight conditional on the amount smoked, other controls and WIC participation status is discontinuous in the average daily cigarettes smoked. Together these two observations give us the required structure on our required discontinuously distributed covariate, W.

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 expectation of the unobserved variable conditional on smoking, other controls and WIC participation must be *discontinuous* in smoking. If, however, the selection-on-observables assumption holds, that is if the unobservables in the outcome equation are independent of WIC participation condition on smoking and other controls, then this discontinuity must be the same for WIC participants and non-participants. conditional on smoking and other controls. In other

[1] In section 5, we use the same setup for our empirical application and layout the problem in estimation of WIC treatment effects in much more detail.

[2] The exact percentage of non-smokers, however, varies by trimester. For example, before pregnancy (0^{th}-trimester) 82.4% of women do not smoke. By the third trimester, this number increases to 88%.

[3] We treat average amount of smoking per day as a continuously distributed variable to the right of 0 as continuously distributed in the same way education is treated as continuously distributed by many empirical researchers.

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. Thus, one could design a test of selection-on-observables assumption by checking if this is indeed the case. These arguments would work even if treatment was not a discrete variable.

As mentioned above, the testable implications we provide hinge crucially on having a covariate/control W that is continuously distributed except for having an atom at a known point. Such a scenario can be found in widely diverse empirical settings than one would expect at first. Caetano (2015) discusses a number of examples of such variables and we use one of these, namely smoking during pregnancy as our W, in the empirical application to evaluate whether participation in WIC is random conditional on maternal smoking and other controls. However, our testing procedure can be applied in other situations. For example, our methods could be useful in evaluating the effect of choosing a STEM (science, technology, engineering, and mathematics) major in college on future labor market outcomes, using SAT scores as the W variable. We expect that quantitative SAT scores might have bunching point at 800, which is the maximum attainable score. On the other hand, it might be easier to envisage the applicability of our setup in naturally occurring thresholds, for instance, variables that are bounded to the left by zero. One example is time inputs, like job search intensity for unemployed workers or parental time investments in their children. In the latter, our setup can be used to investigate the effect of a mother's labor force participation decision X, on children's academic outcomes, Y. As our bunching variable, we can use the number of hours a mother allows her children to watch television on school days. This variable can have a bunching point at zero and hence can be used for our testing procedure.[4]

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; 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 (See Gourio and Roys (2014), for example.).

Our test can also be used for a continuous treatment variable. Consider the problem of estimating dynamic spillovers in crime where researchers are interested in estimating the effect of lagged crime rates on current crime (Jacob, Lefgren and Moretti (2007); Caetano and Maheshri (2015)). There are clear endogeneity concerns here with unobserved location specific criminogenic factors influencing crime rates in both time periods in a given city. We can potentially use bunching at zero in police response to calls for service as our discontinuously distributed variable, W, and estimate the effects for our treatment variable of interest, lagged crime rate, which is continuos in this setup.[5]

[4]Allowing children to watch television is just one example of a potentially discontinuously distributed covariate in this set up. A number of time use surveys provide detailed information on household members' daily activities. Hence, we can find similar covariates that satisfy our requirement for W, largely associated with children's entertainment activities.

[5]In fact, Caetano and Maheshri (2015) 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.

The paper is organized as follows. Section 2 discusses how our paper fits into the existing literature. In Section 3 we introduce the formal model and the main Theorem. Section 4 discusses the implementation of the basic testing idea. In particular, we provide our formal test statistics and derive their asymptotic distribution in this section. Section 5 discusses our empirical application in detail. Section 6 concludes. The Appendix provides the proofs.

2 Literature Review:

The most closely related paper to ours is Caetano (2015), given the common assumption that there is a variable W that has a bunching point at a known value, but is otherwise continuously distributed. Both papers further assume that the function relating W to outcome of interest is continuous. In Caetano (2015), however, W itself is the variable of interest/treatment variable, and the focus is on estimating the average partial effect of W on Y. If W is independent of the unoservables in the outcome equation, the expected value of the outcome given W and the controls should be continuous in w at the bunching value of W. However, if this does not hold, then we can conclude that W must be stochastically related to U conditional on controls, at least around the bunching point. In Caetano (2015) the structural function is nonparametric (except for the requirement of continuity and existence of some conditional moments no restrictions are placed on the structural function). In her case W, the treatment, however, is neither discrete nor continuous. It has this particular structure of having bunching at a known point and being continuously distributed except at the bunching point.

In our paper we provide a formal test for the selection-on-observables assumption for models in which the outcome equation is additively separable in treatment, X, and unobservables, U, using the discontinuously distributed variable, W, in our set of covariates. We do this by checking if the discontinuities in $E(Y|X = x, W = w)$ and $E(Y|X = x', W = w)$ with respect to w are equal for different values x and x' of X. In particular, the variable U in our set up does not have to represent the structural unobservables; it might be resulting from misspecification. Our treatment variable, however, could be any type of variable. Thus, the approach in this paper and that in Caetano (2015) are not nested. On the other hand, for the testing procedure we present and the one proposed by Caetano (2015) to have power, conditional on W the distribution of U must be discontinuous in W at $W = w_0$. In our paper, since the main variable of interest is X, not W, conditional distribution of U given $W = w$ could be discontinuous under the null hypothesis as well.

Superficially the approach in this paper and in Caetano (2015) seems to be related to the regression discontinuity approach. In actuality, neither this paper nor Caetano (2015) is a regression discontinuity design (RDD) paper. The only similarity between the RDD and our approach is the local nature of the conclusion obtained. In the RDD the identified effect is a local one, since the assumed exclusion restriction is a local one. In this paper, it is the conclusion about whether the variable of interest stochastically depends on the unobservables conditional on controls that is local.[6]

[6]Note that local failure of the exclusion restriction being tested implies its global failure.

Similarly, Crump, Hotz, Imbens and Mitnik (2008) presents nonparametric tests for treatment heterogeneity for discrete treatments. Specifically, focusing on the binary treatment case, they make the selection-on-observables assumption and test two null hypotheses under this assumption. The first is whether $\tau(w,z) = 0$ for each (w,z), where $\tau(w,z)$ denotes the effect of treatment for individuals with covariate values equal to (w,z). The second hypothesis they test is that $\tau(w,z) = \tau$ for all (w,z), that is the treatment effect conditional on (W,Z) is constant, except for distributional effects that average out to zero. The test we present can thus also be interpreted as a test of a kind of test of leftover unobserved heterogeneity: we are testing whether the effect of X is the same for individuals with the same value of (W,Z). In other words, what we are testing is if there is any remaining heterogeneity in the effect of treatment after we control for W and Z. Heckman calls this essential heterogeneity (see Heckman, Urzua and Vytlcail (2006)). In addition, while we provide formal test statistics and derive their asymptotic behavior for binary and continuous X cases, our testing procedure could be used for any other type of treatment variable.

This paper is also related to the literature on testing the validity of identifying assumptions in nonlinear models. Catano, Rothe and Yıldız (2016) use discontinuity in the distribution of X (the main variable of interest) 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 LATE monotonicity assumption in the context of a binary treatment variable X. Lu and White (2014) present a test for additive separability in X and U, of the structural function relating X to Y, under the assumption that U and X are independent conditional on Z, where Z is a variable that is excluded from the structural function (i.e. the structural function does not depend on Z). Under the same conditional independence assumption Hoderlein, Su, White and Yang (2014) presents a testing procedure for testing whether the structural function relating X to outcome Y is strictly monotone in a scalar continuously distributed unobservable variable, U. We do not require any covariate to be excluded from the structural function.Lewbel, Lu and Su (2015) present a testing 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 The Model and Its Testable Implications:

In this section, we introduce the model and derive associated testable implications given a discontinuously distributed covariate W. For expositional purposes we omit other covariates in this section. In the next section, we discuss how the testing ideas we introduce in this section could be implemented; we also discuss how we introduce other covariates into the model in that section. To facilitate exposition for any random variable T, let $\mu_T(x,w)$ denote a version of $E(T|X = x, W = w)$. For any function $g(x,w)$, let $g(x,0^+)$ denote $\lim_{w \downarrow 0} g(x,w)$, whenever that limit exists. Let $\mathcal{X}$ denote the support of x. The model we

study is given by

$$Y = m(X, W) + U, \tag{1}$$

where Y is the outcome variable (denoting birth weight in our application), X is the variable whose partial effect we are interested in (in our example it denotes participation in WIC) and W is a scalar control (maternal smoking during pregnancy in our application), and U represents an unobservable variable. A vector of other covariates represented by Z is introduced later in this section. Note that the model given in equation 1 allows some heterogeneity in the effect of X on Y. In particular, the effect of treatment on people with different values of (W, Z) are allowed to be different.

Since in our empirical application the variable of interest is binary, we introduce and develop the testing idea and the test statistic for the binary X case first. In the next subsection, we develop a second test statistic that can be used for continuous X. When X is binary the potential outcomes, Y_1 and Y_0, are

$$Y_1 = m(1, W) + U, \tag{2}$$

$$Y_0 = m(0, W) + U. \tag{3}$$

The full selection on observables assumption is that $Y_1, Y_0 \perp\!\!\!\perp X|W$. In the current context, the selection on observables assumption can be written as

$$U \perp\!\!\!\perp X|W.$$

This assumption states that conditional on W, the effect of the treatment X on outcome is homogeneous, that is, the effect of treatment on people with same W value must be the same. Arguably, this is a strong assumption. Nevertheless, a significant part of the reduced form empirical analysis is conducted under this assumption. To estimate average treatment effect, the assumption needed is in fact the conditional mean independence assumption given by

$$\mathrm{H}_0 : \quad E(U|X = x, W = w) = E(U|W = w) \text{ for } a.e.\,(x, w). \tag{4}$$

Here we provide a formal test to assess the validity of this assumption. Later in this section we outline how our testing procedure could be extended to test $U \perp\!\!\!\perp X|W$. We will assume that m is continuous in w for each z and x, and the random variable W has an atom/ positive mass at a known point, and is continuously distributed in a neighborhood of this point. The point at which W has positive mass is normalized to be 0. This mass point 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 the situation in which W has a single mass point and that mass point is at the lower bound of the support of W. Everything we do could also be done if W has multiple mass points and they might be in the interior or the upper bound of the support of W, but at the expense of added notation. Now we will state the crucial assumption we will make to get the main testable implication of the selection on observables assumption.

Assumption 3.1. (i) m is continuous in w for each x; (ii) There exists $A \in \sigma(X)$ with $P(A) > 0$ such that $\mu_U(x, 0^+)$ exists for each $x \in A$; (iii) For some $\delta > 0$ and the set

A in (ii), W has a conditional density $f_{W|X}(w|x)$ that is strictly greater than 0, whenever $w \in (0, \delta)$ and $x \in A$.

The first part of this assumption requires that the function linking treatment X to the outcome is continuous in the variable W. In the context of our empirical application, this assumption will hold if the structural effect of maternal smoking on infant's birth weight is continuous. The second part says that the limit of $\mu_U(x, w)$ exists as $w \downarrow 0$ for a set of x values that has positive probability. The third part of the assumption ensures that we can meaningfully talk about this limit.

Theorem 3.1. *Suppose the model given in equation 1 and Assumption 3.1 hold. Then*

$$E(U|X = x, W = w) = E(U|W = w) \text{ for } a.e.\ (x, w) \implies \lambda(x, x') := \lambda(x) - \lambda(x') = 0, \quad (5)$$

where

$$\lambda(x) := \mu_Y(x, 0^+) - \mu_Y(x, 0). \quad (6)$$

Proof. The results in the theorem follow once we notice that under Assumption 3.1

$$\lambda(x) = \mu_U(x, 0^+) - \mu_U(x, 0).$$

□

At this point, it is not obvious that a testing procedure based on $\lambda(x, x')$ would have have any power at all. Before we explain why in certain situations a testing procedure based on $\lambda(x, x')$ will have power let us emphasize that testing $E(U|X = x, W = w) = E(U|W = w)$ for $a.e.\ (x, w)$ is challenging. As far as we know, there is no other testing procedure that has power under scenarios which we consider empirically relevant and in which the procedure we propose has power. In particular, given that the basic model we study has an additively separable unobservable term, it might be tempting to define $T := Y - E(Y|X, W)$, and try to test whether T and X are independent conditional on W.[7] Note that

$$E[XT|W] = E[XY|W] - E[XE(Y|X, W)|W] = E[XY|W] - E[E(XY|X, W)|W] = 0,$$

so this alternative test would have no power when H_0. When the model is really as given in equation 1, then the assumption that identifies the average treatment effect is, in fact, H_0. Such an alternative testing procedure can, however, have power in testing whether higher order moments of X and T are independent from each other conditional on W.

Our testing procedure will be based on $\lambda(x, x')$ and will exploit the discontinuity in the distribution of W. We therefore, will be exploiting a particular type of W. Specifically, our W not only has to have a discontinuity in its distribution, but around its mass point W has to be continuously distributed. If W were simply discrete, we would have $\mu_Y(x, w) = m(x, w) + \mu_U(x, w)$ and comparisons of $\mu_Y(x, w)$ at different (x, w) values are likely to tell us nothing about how μ_U is related to x. And if W were continuously distributed at a particular

[7] This was suggested by a previous anonymous referee. We thank this referee for this suggestion.

value, say 0, it would be hard to believe that $\mu_U(x,w)$ would necessarily have a discontinuity in w at $w=0$. Fortunately, variables with the above defined structure are readily available in empirically relevant settings, as cited in Caetano (2015) and in the introduction to this paper. To see how this type of discontinuity in the distribution of W may help us get some power in testing H_0 for binary X, consider

$$
\begin{aligned}
& E[Y|W=w,X=1]-E[Y|W=w,X=0] && (7)\\
= \; & m(w,1)-m(w,0)\\
+ \; & E[U|W=w,X=1]-E[U|W=w,X=0]\\
= \; & m(w,1)-m(w,0) && (8)\\
+ \; & \frac{\int_{-\infty}^{\infty}\int_{\mathcal{V}(w)} u\mathrm{d}F_{UV|W}(u,v|w)}{P(w)}-\frac{\int_{-\infty}^{\infty}\int_{\mathcal{V}^C(w)} u\mathrm{d}F_{UV|W}(u,v|w)}{1-P(w)}\\
= \; & m(w,1)-m(w,0) && (9)\\
+ \; & \frac{\int_{-\infty}^{\infty}\int_{\mathcal{V}(w)} u\mathrm{d}F_{U|V,W}(u|v,w)\mathrm{d}F_{V|W}(v|w)}{P(w)}\\
- \; & \frac{\int_{-\infty}^{\infty}\int_{\mathcal{V}^C(w)} u\mathrm{d}F_{U|V,W}(u|v,w)\mathrm{d}F_{V|W}(v|w)}{1-P(w)},
\end{aligned}
$$

where V denotes the unobservables in the treatment equation, $P(w)=E(X|W=w)$, and $\mathcal{V}(w)$ denotes the set of V values associated with treatment value equal to 1 conditional on $W=w$. Note that in this case testing $E(U|X,W)=E(U|W)\,a.s.$ is closely linked to testing $U \perp\!\!\!\perp V|W$. If $(U,V) \perp\!\!\!\perp W$, then $F_{U,V|W}(u,v|w)=F_{U,V}(u,v)$, and therefore, $F_{U,V|W}(u,v|w)$ is continuous in w. If, in addition, the set $\mathcal{V}(w)$ varies continuously in w, as in the case $\mathcal{V}(w)=(-\infty,h(w)]$, with h continuous in w, then the last expression above will be continuous in x even if $U \not\perp\!\!\!\perp V|W$. In that case, both $\lambda(1)$ and $\lambda(0)$, and hence, $\lambda(1,0)$ will be 0, and any test statistic based on λ will have no power for testing $U \not\perp\!\!\!\perp V|W$.

While tests based on λ will have no power as long as $F_{U,V|W,Z}(u,v|w,z)$ is continuous in w at $w=0$, we have some way of checking whether this is the the reason $\lambda(z)=0$. In particular, using Caetano (2015) we can check if $P(w,z)=E(X|W=w,Z=z)$ is continuous in w or not at $w=0$. If this is discontinuous in w then we can conclude that $F_{V|W,Z}(v|w,z)$ is not continuous in w at $w=0$.

On the other hand, if V is independent of U conditional on (W,Z), even if the joint distribution of U and V conditional on (W,Z) varies discontinuously with w, this discontinuity is canceled. In contrast, if $U \not\perp\!\!\!\perp V|W,Z$ the discontinuity of the joint distribution of U and V conditional on W,Z will be different for treated and untreated people, and the difference between $E(Y|X=1,W=w,Z=z)$ and $E(Y|X=0,W=w,Z=z)$ will be discontinuous in w at $w=0$, except in non-generic/knife-edge cases. In particular, in a supplementary appendix we show $\lambda(1,0)$ will be different from 0, and the test we propose will have power when $\sigma_{UW^*}=0,\sigma_{UV}\neq 0,\sigma_{VW^*}\neq 0$ and when $\sigma_{VW^*}=0,\sigma_{UV}\neq 0,\sigma_{UW^*}\neq 0$ in the special case in which

$$
\begin{aligned}
X &= 1\{\alpha+W\beta\geq V\},\\
W &= \max\{0,W^*\},
\end{aligned}
$$

with

$$\begin{pmatrix} U \\ V \\ W^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_{w^*} \end{pmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{u\eta} & \sigma_{uw^*} \\ \sigma_{u\eta} & 1 & \sigma_{vw^*} \\ \sigma_{uw^*} & \sigma_{vw*} & \sigma_{w^*}^2 \end{bmatrix} \right). \tag{10}$$

In the supplementary appendix we also study the case

$$\begin{aligned} Y &= \alpha + X\beta + \theta W + U, & (11) \\ W &= \max\{0, \gamma + \delta X + \eta\} & (12) \end{aligned}$$

and

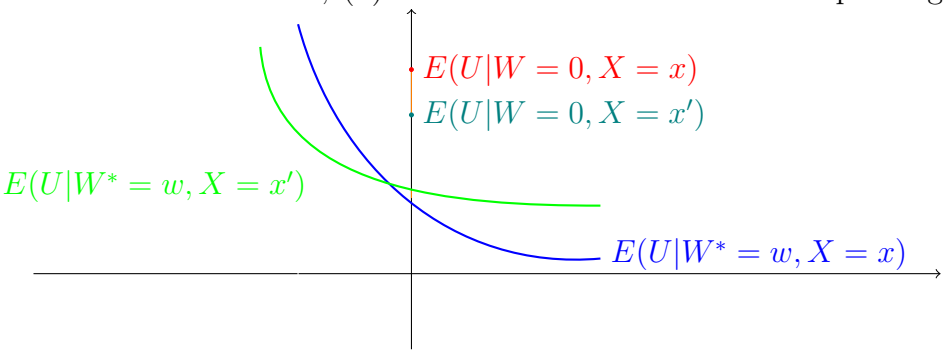
$$\begin{pmatrix} U \\ \eta \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_x \end{pmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{u\eta} & \sigma_{ux} \\ \sigma_{u\eta} & \sigma_\eta^2 & \sigma_{\eta x} \\ \sigma_{ux} & \sigma_{\eta x} & \sigma_x^2 \end{bmatrix} \right). \tag{13}$$

In this case

$$\begin{aligned} E(U|W,X) &= \pi_x(X-\mu_x) + \pi_\eta(W-\gamma-\delta X)1\{W>0\} \\ &+ \left[\pi_\eta \frac{\sigma_{x\eta}}{\sigma_x^2}(X-\mu_x) - \pi_\eta \sqrt{\sigma_\eta^2(1-\rho_{x\eta}^2)} \frac{\phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho_{x\eta}^2}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho_{x\eta}^2}} X \right)}{\Phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho_{x\eta}^2}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho_{x\eta}^2}} X \right)} \right] 1\{W=0\}, \end{aligned} \tag{14}$$

where $\pi_x = \frac{\sigma_{ux}\sigma_\eta^2 - \sigma_{u\eta}\sigma_{\eta x}}{\sigma_x^2\sigma_\eta^2 - \sigma_{\eta x}^2}$ and $\pi_\eta = \frac{\sigma_{u\eta}\sigma_x^2 - \sigma_{ux}\sigma_{\eta x}}{\sigma_x^2\sigma_\eta^2 - \sigma_{\eta x}^2}$, and $\rho_{x\eta}$ is the correlation coefficient between X and η. In this example, if $\sigma_{\eta x} \neq 0$ and $\sigma_{ux} \neq 0$, then π_x and π_η are both different from 0, and our testing procedure will have power. Alternatively, if $\sigma_{\eta x} = 0$, but $\delta \neq 0$, and $\sigma_{u\eta} \neq 0, \sigma_{ux} \neq 0$, again our testing procedure will have power.

To see how the particular structure of W is related to $\lambda(x, x')$ in general, the graph below may be helpful. In the graph, we also assume that $W = \max\{0, W^*\}$. In this graph we assume that $E(U|X = x, W^* = w^*)$ is a continuous function of w^* for both x and x'. In this graph, $\lambda(x, x')$ is the total length of two intervals: (i) the interval between red and teal dots, (ii) the interval between the intercepts of green and blue curves.



As the discussion as well as the graph above indicate our testing idea is based on exploiting the discontinuity in the distribution of a covariate, W. For our testing procedure to have power this covariate has to be endogenous, and this endogeneity has to interact with the

variable of interest X. In particular, our testing procedure will not have any power if for some functions ψ_1 and ψ_2, which are measurable w.r.t. X and W, respectively, we have

$$E(U|X,W) = \psi_1(X) + \psi_2(W) \; a.s.,$$

then $\lambda(x,x')$ will be 0 for each $x \neq x'$, even when ψ_1 is not equal to 0. While this seems troubling, it is less problematic than it seems at first, since W has this particular structure of having a mass point and being continuously distributed in a neighborhood of its mass point. In particular, if in fact, W^* is continuously distributed conditional on X, and $W = W^*$ when W^* is outside some bunching set B, and $W = 0$ when $W^* \in B$. Moreover, suppose that $P(W^* \in B|X = x) > 0$ whenever $x \in A$, where A is as in Assumption 3.1. Then if we let B^C denote the complement of B, we have

$$\mu_U(x,w) = \mu_U(x,w^*)1\{w^* \in B^C\} + \frac{\int 1\{w^* \in B\}\mu_U(x,w^*)\mathrm{d}F_{W^*|X}(w^*|x)}{P(W^* \in B|X = x)}.$$

This shows that for $\mu_U(x,w)$ to be additively separable, it must be that $\mu_U(x,w^*)$ must be additively separable in x and w^*. That is, there must be functions $\tilde{\psi}_1$ and $\tilde{\psi}_2$, measurable with respect to X and W^*, respectively, such that $\mu_U(x,w^*) = \tilde{\psi}_1(x) + \tilde{\psi}_2(w^*)$. In that case, we have

$$\begin{aligned}\mu_U(x,w) =& [\tilde{\psi}_1(x) + \tilde{\psi}_2(w^*)]1\{w^* \in B^C\} + \frac{\int 1\{w^* \in B\}[\tilde{\psi}_1(x) + \tilde{\psi}_2(w^*)]\mathrm{d}F_{W^*|X}(w^*|x)}{P(W^* \in B|X = x)} \\ =& \tilde{\psi}_1(x) + \tilde{\psi}_2(w)1\{w \in B^C\} + \frac{\int 1\{w^* \in B\}\tilde{\psi}_2(w^*)\mathrm{d}F_{W^*|X}(w^*|x)}{P(W^* \in B|X = x)}, \end{aligned} \tag{15}$$

which means that even if $\mu_U(x,w^*)$ is additively separable in x and w^*, $\mu_U(x,w)$ will be additively separable in x and w only if W^* and X are independent. Furthermore, if W^* and X are independent then W and X must be independent, and independence of W and X is a testable. Since maternal smoking and WIC participation are likely to be correlated as similar types of unobserved heterogeneity may make women not smoke during pregnancy and participate in WIC, we would expect W and X not to be independent in our application. Alternatively, participation in WIC and the educational programs it involves may convince a woman to stop smoking during pregnancy.

This discussion suggests that the covariate W should be deemed relevant for the outcome the researcher is interested in. In particular, we suggest using a W that the researcher strongly believes has a structural effect on the outcome, and as a result, not controlling for W would very likely lead to biased estimate of the partial effect of X.

3.1 Further discussion:

It is crucial to note that our testing procedure is a joint test of additive separability of the outcome equation in treatment and unobservables as well as the selection on observables assumption. In particular, if the true structural relationship is

$$Y = h(W,X,\tilde{U}), \tag{16}$$

and the researcher mistakenly assumes that it is as in equation 1, then $U = h(W, X\tilde{U}) - m(W, X)$. In this case, then the discontinuities in $E(Y|W = w, X = x')$ and in $E(Y|W = w, X = x)$ will, typically, not be equal and the population value of our test statistic will be nonzero. Therefore, the endogeneity we are testing includes the one arising from misspecification.[8,9]

Before proceeding, we note that if along with selection on observables assumption, one uses OLS to estimate the effect of treatment, then finding that $E[Y|X = 1, W = w] - E[Y|X = 0, W = w]$ is discontinuous in w at $w = w_0$ is evidence that the estimator for the effect of X on Y is not valid. If one makes the selection on observables assumption and uses propensity score matching, however, finding that $E[Y|X = 1, W = w] - E[Y|X = 0, W = w]$ is discontinuous in w at $w = 0$ does not necessarily mean that the estimator of the treatment effect is invalid. For this reason, it is important to have a procedure which we could use to judge the validity of the selection on observables assumption.

A diagnostic procedure for judging the validity of the selection on observables assumption if the structural function is not additively separable in X and the unobservables is the following: In the first step, test whether $P(w)$ is continuous in w at $w = 0$. We will do the second step only if we find that $P(w)$ is discontinuous in w. Note that finding that $P(w)$ is discontinuous in w suggests that W is not independent of the unobservables in the treatment equation, say V, if $\mathcal{V}(w)$ varies continuously in w. In the second step, test whether $E(Y|X = 1, W = w)$ and $E(Y|X = 0, W = w)$ are continuous in w at x_0. Finding that they both are continuous in w conditional on having found that $P(w)$ is discontinuous suggests that $U \perp\!\!\!\perp (X, W)$, since given the discontinuity in $P(x)$ this is an unlikely scenario if selection on observables assumption is violated.

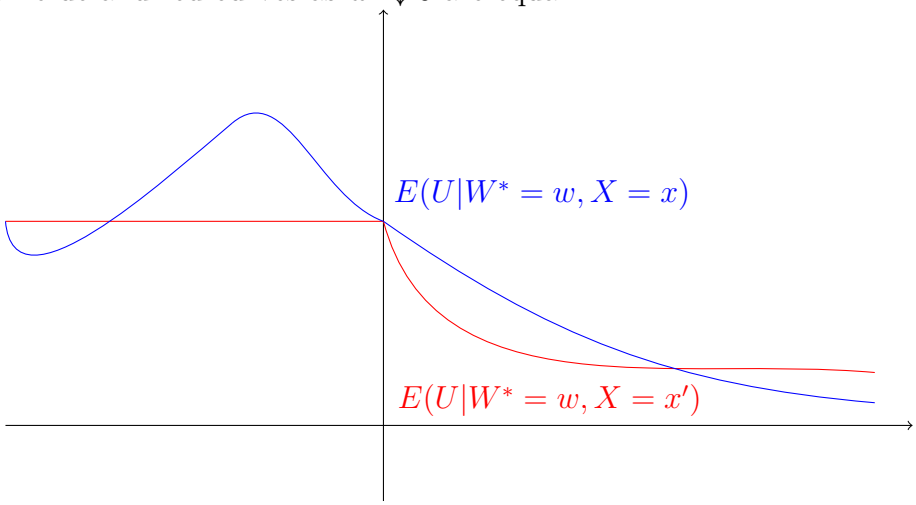
From an empirical standpoint, failing to reject the null hypothesis after conditioning for a rich enough set of covariates can be taken as evidence that aspects of unboserved heterogeneity that could potentially contaminate estimation of partial effect of X on Y have been controlled for. Nevertheless, one should exercise caution for multiple reasons. First, our testing procedure will have power only when W is endogenous itself. Moreover, the endogeneity of this covariate has to interact with X somehow. This means that the covariate W cannot be any arbitrary covariate. In particular, one should only condition on W if one believes not controlling for W would result in omitted variables bias. If a researcher controls for a W that is not really expected to be related to the outcome one is interested in, then controlling for such a W may introduce endogeneity when there is none. This statement,

[8]Endogeneity resulting from misspecification can also be thought of as arising from omitted variables bias. To see this suppose that h has a series expansion. Then $m(W, X)$ represents the parts of this expansion that depend on W and X only. All the other terms are part of U. In our empirical application, when we use only basic controls, we cannot reject that the leftover term is stochastically related to (not independent of) treatment conditional on these controls. However, when we add a detailed set of covariates based on various interactions between parental socio-economics variables and mother's pregenancy charactersitics, along with flexible controls for smoking across trimesters, we see that the leftover terms no longer seem to be related to treatment conditional on these new set of controls.

[9]If U is not he structural unobservable term, then m is not the structural function. If the true structural function is $h(W, X, \tilde{U})$ where h has a series expansion and m represents the terms in this expansion that depend on (W, X) only, if h is continuous in W at $W = w_0$, then so will m be. Thus, Assumption 3.1 can be argued by arguing that the true structural relationship is continuous at $W = w_0$.

however, applies to all empirical research, not just to our testing procedure.

On the other hand, one should, for example, keep in mind that like the testing procedure in Caetano (2015) the testing procedures we propose cannot detect dependence of $E(U|X,W)$ on X or W away from the bunching point of W. In addition, we could have pathological situations in which $\mu_U(x,w^*)$ is not additively separable in x and w^*, but $\lambda(x,x')=0$. The graph below illustrates such a situation. Note that in the graph, the blue curve is sometimes above and sometimes below the red line when $w^*<0$. If the weighted average (where the weight used is $\mathrm{d}F_{W^*|X}(w^*|x)/P(W^*\leq 0|X=x)$) of the parts of the blue curve above and below the red line exactly cancel each other out then $\lambda(x,x')$ will be 0, since the limits of both blue and red curves as $w^*\downarrow 0$ are equal.



Any test based on $\lambda(x,x')$ will only be a test of whether $E(U|X,W)=E(U|X,W)\,a.s.$; such a test will not give us any information about the dependence of distribution U on X conditional in W beyond the first conditional moment of U. As mentioned above, for higher moment dependence one can check whether higher order moments of $Y-E(Y|X,W)$ are conditionally independent of higher order moments of X conditional on W. An alternative test of can be constructed by considering the characteristic function of Y, instead. Specifically, suppose, as before, the model given by equation 1 holds. Note that

$$E\left[e^{itY}|X=1,W=w\right]=e^{itm(w,1)}E[e^{itU}|X=1,W=w].$$

Then

$$\kappa(t) := \frac{\lim_{w\downarrow 0} E\left[e^{itY}|X=1,W=w\right]}{E\left[e^{itY}|X=1,W=0\right]} - \frac{\lim_{w\downarrow 0} E\left[e^{itY}|X=0,W=w\right]}{E\left[e^{itY}|X=0,W=0\right]} \tag{17}$$

$$= \frac{\lim_{w\downarrow 0} E\left[e^{itU}|X=1,W=w\right]}{E\left[e^{itU}|X=1,W=0\right]} - \frac{\lim_{w\downarrow 0} E\left[e^{itU}|X=0,W=w\right]}{E\left[e^{itU}|X=0,W=0\right]}$$

$$= \frac{E\left[e^{itU}|X=0,W=0\right]}{E\left[e^{itU}|X=1,W=0\right]E\left[e^{itU}|X=0,W=0\right]}$$

$$\times \left(\lim_{w\downarrow 0} E\left[e^{itU}|X=1,W=w\right] - \lim_{w\downarrow 0} E\left[e^{itU}|X=0,W=w\right]\right) \tag{18}$$

$$- \frac{\lim_{w\downarrow 0} E\left[e^{itU}|X=0,W=w\right]}{E\left[e^{itU}|X=1,W=0\right]E\left[e^{itU}|X=0,W=0\right]}$$

$$\times \left(E\left[e^{itU}|X=1,W=0\right] - E\left[e^{itU}|X=0,W=0\right]\right). \tag{19}$$

Under the null hypothesis that $U \perp\!\!\!\perp X|W$, both 18 and 19, and hence $\kappa(t)$ must be 0 for each t.

4 Implementation of the Test

In this section we discuss how based on Theorem 3.1 and the discussion following it, we can devise a test statistic for testing

$$H_0 : P_{XWZ}\big(E(U|X,W,Z) = E(U|W,Z)\big) = 1, \tag{20}$$

against

$$H_1 : P_{XWZ}\big(E(U|X,W,Z) = E(U|W,Z)\big) < 1. \tag{21}$$

Here Z represents a vector of other observed covariates. We first propose a test statistic for binary X case. In a subsection we propose another test statistic for continuous X case.

If X takes values 1 and 0 only with positive probability, for each z value we will only have $\lambda(1,0,z)$, but we still need to aggregate this population parameter over different z values. In particular, we could use

$$T \;=\; \int_{z\in\mathcal{A}} \psi(\lambda(1,0,z))\omega(z)\mathrm{d}z, \tag{22}$$

where $\psi : \mathbb{R} \mapsto \mathbb{R}_+$ such that $\psi(s) = 0 \iff s = 0$, and ω is a non-negative weight function. We could come up with an estimator $\widehat{T}$ for T, and reject the null hypothesis when $\widehat{T}$ is "too large". To estimate T, we would first have to estimate $\lambda(x,z)$ for $x = 0, 1$. Specifically, $E[Y|X = x, W = 0, Z = z]$ can be estimated with a local linear regression of Y onto Z at z using only observations such that $W = 0$ and $X = x$, and $\lim_{w\downarrow 0} E[Y|W = w, Z = z, X = x]$ can be estimated with a local linear regression of Y onto W and Z at $W = 0$ and $Z = z$ using only observations such that $W > 0$ and $X = x$. Under standard regularity conditions, in estimation of $\lambda(x,z)$, $\lim_{w\downarrow 0} E[Y|W = w, Z = z, X = x]$ term will dominate

the asymptotic variance because it has one more estimation dimension and thus converges slower. Using standard results from the regression discontinuity literature we expect that $\sqrt{nh^{d_z+1}}(\widehat{\lambda}(x,z) - \lambda(x,z)) \xrightarrow{d} N(0, V_x)$, where d_z denotes the dimension of Z.

In our application, the dimension of Z is quite large,[10] which means that fully nonparametric estimation of $\lambda_x(z)$ is not feasible in practice. For this reason, for estimation, we assume the structural function is partially linear. One possibility is to assume that

$$m(W, X, Z) = g(W, X) + Z^\intercal \gamma, \tag{23}$$

$$Y = g(W, X) + Z^\intercal \gamma + U. \tag{24}$$

We also assume that Z is exogenous:

Assumption 4.1. $E(U|X, Z, W) = E(U|X, W)$.

Under this assumption, we have

$$E(U|X, Z, W) = E(U|X, W) =: \rho(W, X),$$

which means that

$$Y = g(W, X) + Z^\intercal \gamma + \rho(W, X) + \varepsilon,$$

where $\varepsilon := U - E(U|X, W, Z) = U - E(U|X, W)$. Therefore,

$$Y - E(Y|X, W) = [Z - E(Z|X, W)]^\intercal \gamma + \varepsilon,$$

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

Once γ is known we can write

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

Then under under our null hypothesis, we have

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

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

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 requires a large number of non-parametric regressions. If, in addition, $E(Z_k|X, W)$ is linear in W for each $k \in \{1, 2, ..., d_z\}$, that is $E(Z_k|X, W) = X(\alpha_{1k} + W\beta_{1k}) + (1 - X)(\alpha_{0k} + W\beta_{0k})$, then γ can be consistently estimated by a single OLS regression of Y on X and Z.[11]

We estimate $E(\tilde{Y}|W = 0, X = x)$ by

$$\frac{1}{n_{x0}} \sum_{i=1}^{n} \hat{\tilde{Y}}_i 1\{X_i = x, W_i = 0\},$$

[10] Even in our baseline specification the dimension of Z is around 40.

[11] See the supplementary Appendix.

where $n_{x0} = \sum_{i=1}^{n} 1\{X_i = x, W_i = 0\}$. For $x = 0, 1$, we estimate

$$\mu_{\tilde{Y}|X,W}(x, 0^+) := \lim_{w \downarrow 0} E(\tilde{Y}|W = w, X = x)$$

as

$$\hat{\mu}_{\hat{\tilde{Y}}|X,W}(x, 0^+) := e_1^\intercal \text{argmin}_{a_0,a_1} \sum_{i=1}^{n} (\hat{\tilde{Y}}_i - a_0 - a_1 W_i/h)^2 K_h(W_i) 1\{W_i > 0, X_i = x\},$$

where $K_h(w) = \frac{1}{h} K\left(\frac{w}{h}\right)$, h is a bandwidth that goes to 0 as $n \to \infty$, and $e_1 = (1, 0)^\intercal$.

Assumption 4.2. (i) $\{Y_i, X_i, W_i, Z_i^\intercal\}_{i=1}^n$ is a random sample. For some $\alpha > 0$, $E\left(|Y|^{2+\alpha}\right) < \infty$ and $E\left(\|Z\|^{2+\alpha}\right) < \infty$.

(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 $x = 0, 1$, $E[\tilde{Y}|W = w, X = x]$ is twice continuously differentiable in w.

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

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

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

$$\begin{aligned}
p &: P(X_i = 1), \\
f_{W|X}(0^+|x) &:= \lim_{w \downarrow 0} f_{W|X}(w|x), \\
\sigma^2_{\varepsilon|W,X}(0^+, x) &:= \lim_{w \downarrow 0} E(\varepsilon_i^2|W_i = w, X_i = x), \\
\kappa_j &:= \int_0^\infty u^j K(u)\mathrm{d}u, \\
\nu_j &:= \int_0^\infty u^j K^2(u)\mathrm{d}u,
\end{aligned}$$

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

Theorem 4.1. *Suppose the model in 1 holds. In addition, suppose Assumptions 3.1, 4.1 and 4.2 hold. Then*

$$\sqrt{nh}\left(\hat{\tilde{\lambda}}(1)-\hat{\tilde{\lambda}}(0)-(\tilde{\lambda}(1)-\tilde{\lambda}(0))\right)\xrightarrow{d}N(0,V), \tag{25}$$

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

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. Note that since γ can be estimated at $\sqrt{n}$ rate, first stage estimation of γ does not influence the asymptotic distribution of the test statistic. Moreover, since $0<P(X=1,W=0)<P(W=0)<1$, $E(\tilde{Y}|W=0,X=x)$ can be estimated at $\sqrt{n}$ rate for both $x=0,1$. This means that the estimation of these quantities does not influence asymptotic variance of the test statistic either. 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 $\hat{\varepsilon}_i$ as

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

and estimate

$$\hat{\sigma}^2_{\varepsilon|W,X}(0^+,x):=\frac{1}{n_x}e_1^{\intercal}\text{argmin}_{a_0,a_1}\sum_{i=1}^{n}(\hat{\varepsilon}_i^2-a_0-a_1W_i/h_\sigma)^2K_{h_\sigma}(W_i)1\{W_i>0,X_i=x\},$$

where n_x denotes the number of observations with $X=x$, and h_σ is a bandwidth that goes to 0 as $n\to\infty$. As in the estimation of $\lim_{w\downarrow 0}E(Y|W=w,X=x)$, the fact that $\hat{\tilde{Y}}$, as opposed to $\tilde{Y}$ is generating $\hat{\varepsilon}$ will not have a first order effect on the asymptotic behavior of $\hat{\sigma}^2_{\varepsilon|W,X}(0^+,x)$ because of the faster convergence of the first stage estimator. Moreover, $\hat{\sigma}^2_{\varepsilon|W,X}(0^+,x)$ will be consistent for $\sigma^2_{\varepsilon|W,X}(0^+,x)$.

$f_{W|X}(0^+|x)$ can be consistently estimated as

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

where h_f is a bandwidth that goes to 0 as $n\to\infty$, and $K_{h_f}(u)=\frac{1}{h_f}K_f(u)$, with $\int_0^\infty K_f(u)=0.5$. $P(X=x)$ 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 $\hat{V}_x=\frac{\kappa_2^2\nu_0-2\kappa_2\kappa_1\nu_1+\kappa_1^2\nu_2}{(\kappa_0\kappa_2-\kappa_1^2)^2}\frac{\hat{\sigma}^2_{\varepsilon|W,X}(0^+,x)}{\hat{P}(X=x)\hat{f}_{W|X}(0^+|x)}$. In our empirical application we use the estimators given in Stata.

In light of this theorem, we can define

$$\tilde{t}_n=\sqrt{nh}\frac{\hat{\tilde{\lambda}}}{\sqrt{\hat{V}}}, \tag{26}$$

$\hat{\tilde{\lambda}} = \hat{\tilde{\lambda}}(1) - \hat{\tilde{\lambda}}(0)$ and $\hat{V}$ is some consistent estimator for V. Then we can reject the hypothesis that $\tilde{\lambda} = 0$ when $\tilde{t}_n \in \mathcal{R}$, where $\mathcal{R} = (-\infty, c_{\alpha/2}] \cup [c_{1-\alpha/2}, \infty)$, where $c_{\alpha/2}$ and $c_{1-\alpha/2}$ $\alpha/2$ and $1-\alpha/2$ quantiles of the standard normal, respectively.

Theorem 4.2. *Suppose the conditions of 4.1 hold and $\hat{V}$ is some consistent estimator for V. Then*

(i) If $\tilde{\lambda} = 0$, then $Pr(\tilde{t}_n \in \mathcal{R}) \to \alpha$ as $n \to \infty$.

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

(iii) Under any local alternative that implies $\tilde{\lambda} = \frac{\delta}{\sqrt{nh}}$ 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.

Proof. The proof of this theorem follows from Theorem 4.1 using straightforward arguments. □

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

$$m(W, X, Z) = g(W, X) + XZ\gamma_1 + (1 - X)Z\gamma_0, \tag{27}$$

so that

$$Y = g(W, X) + XZ\gamma_1 + (1 - X)Z\gamma_0 + U. \tag{28}$$

Assuming that $E(U|W, X, Z) = E(U|W, X) =: \rho(W, X)$ we have

$$Y = XY_1 + (1 - X)Y_0,$$

where for $x = 0, 1$

$$Y_x = g(W, X) + \rho(W, X) + Z\gamma_x + \varepsilon,$$

and $\varepsilon = U - \rho(W, X)$ as before. In this case, we have

$$Y - E(Y|W, X) = X[Z - E(W|X)]\gamma_1 + (1 - X)[Z - E(W|X)]\gamma_0 + \varepsilon.$$

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

4.1 Implementation for continuous X:

As in the previous section we are going to assume equations 23 and 24 as well as Assumption 4.1. Then $\tilde{Y}$ and $\hat{\tilde{Y}}$ are as defined before. To describe our test statistic, for a random variable S, let

$$\begin{aligned}
\hat{\mu}_{S|X,W}(x, 0^+) &:= \frac{1}{n}\sum_j e_1^\intercal [\hat{M}_n(x, 0^+)]^{-1} \begin{pmatrix} 1 \\ \frac{X_j - x}{h} \\ \frac{W_j}{h} \end{pmatrix} 1\{W_j > 0\} \frac{1}{h^2} K\left(\frac{X_j - x}{h}\right) K\left(\frac{W_j}{h}\right) S_j, \\
\hat{\mu}_{S|X,W}(x) &:= \frac{1}{n}\sum_j e_1^\intercal [\hat{M}_n(x)]^{-1} \begin{pmatrix} 1 \\ \frac{X_j - x}{h} \end{pmatrix} \frac{1}{h} K\left(\frac{X_j - x}{h}\right) S_j,
\end{aligned}$$

where

$$\begin{aligned}\hat{M}_n(x,0^+) &=:= \frac{1}{n}\sum_l \begin{pmatrix} 1 \\ \frac{X_l-x}{h} \\ \frac{W_l}{h} \end{pmatrix} \begin{pmatrix} 1, \frac{X_l-x}{h}, \frac{W_l}{h} \end{pmatrix} 1\{W_l > 0\}, \\ \hat{M}_n(x) &:= \frac{1}{n}\sum_l \begin{pmatrix} 1 \\ \frac{X_l-x}{h} \end{pmatrix} \begin{pmatrix} 1, \frac{X_l-x}{h} \end{pmatrix}.\end{aligned}$$

In addition,

$$\hat{\mu}_{S|X,W}(x,0) := \frac{\hat{\mu}_{1\{W=0\}S}(x)}{\hat{\mu}_{1\{W=0\}}(x)}.$$

Our basic test statistic will be

$$\hat{t}_n = \frac{1}{n^2}\sum_i \sum_j \left[\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0^+) - \hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0) - \left(\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_j,0^+) - \mu_{\hat{\tilde{Y}}|X,W}(X_j,0)\right)\right]^2. \tag{29}$$

Note that $\hat{t}_n$ is an estimator for

$$t = E[\tilde{\lambda}^2(X_i,X_j)], \tag{30}$$

which equals 0 under the null hypothesis and under the semiparametric structure we have imposed. This statistic is special case of t^C with $w(x) = f_X(x)$ and $\psi(a) = a^2$, where $\tilde{\lambda}$ is used in place of λ to take advantage of the semiparametric structure we have imposed to reduce the dimensions of the nonparametric estimators we have to estimate.

To analyze the asymptotic behavior of $\hat{t}_n$ we make the following assumption:

Assumption 4.3.

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

2. Suppose the support of X, denoted $\mathcal{X}$ is compact, and $P_{XW}(\mathcal{X} \times [0,\delta]) > 0$ for some $\delta > 0$.

3. The joint density f_{XW} of X, W, is continuous on $\mathcal{X}\times[0,\delta]$ and $\inf_{(x,w)\in\mathcal{X}\times[0,\delta]} f_{XW}(x,w) > 0$. It is also continuously differentiable on the interior of $\mathcal{X} \times [0,\delta]$

4. $\mu_Y(x,w) = E(Y|X=x,W=w)$ $a.s.$, and for or each $j = 1,...,d_z$, $\mu_{Z_l}(x,w) = E(Z_l|X=x,W=w)$ $a.s.$. $\mu_Y(x,w)$ and $\mu_{Z_l}(x,w)$ are continuously differentiable on the interior of $\mathcal{X} \times [0,\delta]$ for each $j = 1,...,d_z$.

5. The kernel function K has compact support and is continuously differentiable in the interior of its support. In addition, it satisfies the following conditions: $\int_{-\infty}^{\infty} K(u) = 1$, $\int_{-\infty}^{\infty} uK(u) = 0$. The matrix A is invertible, where $A := \begin{bmatrix} \kappa_0 & 0 & \kappa_1 \\ 0 & \kappa_0\kappa_2^* & 0 \\ \kappa_1 & 0 & \kappa_2 \end{bmatrix}$, with $\kappa_j = \int_0^{\infty} u^j K(u)\mathrm{d}u$ for $j = 0,1,2$ and $\kappa_2^* = \int_{-\infty}^{\infty} u^2 K(u)\mathrm{d}u$.

6. The bandwidth h satisfies the following conditions: (i) $\frac{nh^2}{\log n} \to \infty$.

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

8. $\sigma_\epsilon^2(x,w)=E(\varepsilon_i^2|X_i=x,W_i=w)$ *a.s.* is continuous on $\mathcal{X}\times(0,\delta]$ and $\lim_{w\to 0}\sigma_\epsilon^2(x,w)$ exists for each x.

This assumption lists standard conditions in non-parametric local polynomial estimation. As shown in the appendix, the asymptotic distribution of $\hat{t}_n$ is the same as the asymptotic distribution of $\hat{t}_n^{Inf}$, where

$$\hat{t}_n^{Inf}=\frac{1}{n^2}\sum_i\sum_j\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\left(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+)-\mu_{\tilde{Y}|X,W}(X_j,0)\right)\right]^2. \tag{31}$$

This is because it turns out that the infeasible test statistic is $O_P\left((nh)^{-\frac{1}{2}}\right)$, whereas the difference between the feasible and infeasible test statistic is $O_P\left(n^{-\frac{1}{2}}\right)$. At first glance that the infeasible test statistic is $\sqrt{nh}$-normal as opposed to $\sqrt{n}$-normal seems counter intuitive, since the infeasible test statistic seems to aggregate information across different observations. While it is true that the test statistic does aggregate information across observations, the convergence rate of the test statistic is essentially determined by the convergence rate of $\frac{1}{n}\sum_i[\hat{\mu}_Y(X_i,0^+)-\mu_Y(X_i,0^+)]$, and this quantity only aggregates information only on X dimension, not on both X and W dimensions. The theorem below summarizes the asymptotic behavior of t_n^{Inf}. The proof of this theorem is in the appendix.

Theorem 4.3. *Suppose the model in 1 holds. In addition, suppose Assumptions* **??**, *4.1 and 4.3 hold. Then*

$$\sqrt{nh}(\hat{t}_n-t)\xrightarrow{d}N(0,V), \tag{32}$$

where $V=4C\int_{-\infty}^{\infty}\sigma_\varepsilon^2(X_i,0^+)\frac{\mu^2_{\tilde{Y}|X,W}(X_i,0^+)+\mu^2_{\tilde{Y}|X,W}(X_i,0)}{f_{XW}(X_i,0^+)}f_X^2(X_i)\mathrm{d}X_i$, *with*
$C=\left(\int_{-\infty}^{\infty}\int_0^{\infty}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\u_x\\u_w\end{pmatrix}\right]^2K^2(u_w)K^2(u_w)\mathrm{d}u_w\mathrm{d}u_x\right)$, $f_X(x)$ *denoting the density of* X, $\sigma_\varepsilon^2(x,0^+)=\lim_{w\to 0}\sigma_\varepsilon^2(x,w)$, *and* $f_{XW}(x,0^+)=\lim_{w\to 0}f_{XW}(x,w)$.

The constant C in the asymptotic variance depends on the kernel and can be calculated once the kernel is chosen. The other term in V can be consistently estimated by

$$\frac{1}{n}\sum_i\hat{\sigma}_\varepsilon^2(X_i,0^+)\frac{\hat{\mu}^2_{\hat{\tilde{Y}}|X,W}(X_i,0^+)+\hat{\mu}^2_{\hat{\tilde{Y}}|X,W}(X_i,0)}{\hat{f}_{XW}(X_i,0^+)}\hat{f}_X(X_i),$$

where $\hat{\mu}^2_{\hat{\tilde{Y}}}(X_i,0^+)$ and $\hat{\mu}^2_{\hat{\tilde{Y}}}(X_i,0)$ are as previously defined, $\hat{f}_X(x)$ is the usual kernel density estimator. In addition, for i such that $W_i>0$, we could estimate $\hat{\varepsilon}_i$ as

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

and estimate

$$\hat{\sigma}^2_{\varepsilon|X,W}(x,0^+) := e_1^\intercal \text{argmin}_{a_0,a_1,a_2} \sum_{i=1}^n (\hat{\varepsilon}_i^2 - a_0 - a_1(X_i - x)/h_\sigma - a_2 W_i/h_\sigma)^2 K_{h_\sigma}(W_i) 1\{W_i > 0\}.$$

Finally, $\hat{f}_{XW}(x,0^+)$ is a boundary corrected kernel density estimator, for example,

$$\hat{f}_{XW}(x,0^+) = \frac{1}{n}\sum_{i=1}^n K\left(\frac{X_i - x}{h_f}\right)\left[K\left(\frac{W_i}{h_f}\right) + K\left(\frac{-W_i}{h_f}\right)\right] 1\{W_i > 0\}.$$

In light of this theorem, we can define

$$\tilde{\hat{t}}_n = \sqrt{nh}\frac{\hat{t}_n}{\sqrt{\hat{V}}}, \tag{33}$$

where $\hat{V}$ is some consistent estimator for V. Then we can reject the hypothesis that $\tilde{\lambda}(x,x') = 0$ for $a.e.(x,x')$ when $\tilde{\hat{t}}_n \in \mathcal{R}$, where $\mathcal{R} = (-\infty, c_{\alpha/2}] \cup [c_{1-\alpha/2}, \infty)$, where $c_{\alpha/2}$ and $c_{1-\alpha/2}$ $\alpha/2$ and $1-\alpha/2$ quantiles of the standard normal, respectively.

Theorem 4.4. *Suppose the conditions of 4.3 hold and $\hat{V}$ is some consistent estimator for V. Then*

(i) If $\tilde{\lambda}(x,x') = 0$ for a.e.(x,x'), then $Pr(\tilde{\hat{t}}_n \in \mathcal{R}) \to \alpha$ as $n \to \infty$.

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

(iii) Under any local alternative that implies $\tilde{\lambda}(x,x') = \frac{\delta(x,x')}{(nh)^{1/4}}$ for a.e. (x,x') with $Pr(\delta(X_i,X_j) \neq 0) > 0$, $Pr(\tilde{\hat{t}}_n \in \mathcal{R}) \to 1 - \Phi\left(c_{1-\alpha/2} - \frac{\Delta}{\sqrt{V}}\right) + \Phi\left(c_{\alpha/2} - \frac{\Delta}{\sqrt{V}}\right)$, as $n \to \infty$, where $\Phi(\cdot)$ denotes the standard normal distribution and $\Delta := E[\delta^2(X_i,X_j)]$.

Proof. The proof of this theorem follows from Theorem 4.3 using straightforward arguments. □

5 Empirical Application

5.1 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 in turn linked to significant health costs especially during infancy and early childhood.[12] Given these concerns the U.S. government operates a widely applicable $6.2 billion welfare program, the Special

[12] 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.

Supplemental Nutrition Program for Women, Infants, and Children (WIC) targeting at-risk low income pregnant mothers. The program 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.[13] Previous literature thus lacks conclusive evidence on the actual treatment effect of WIC in improving birth outcomes for participants. Moreover, given a lack of other potential exclusion restriction most of the literature has resorted to using a selection on observables approach and concludes treatment effects on average birth weight ranging from no effect to gains upwards of 60g (Bitler and Currie (2005); Figlio (2009); Khalil (2015).[14] 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 discontinuosly distributed and potentially endogenous variable with an atom at zero.[15]

5.2 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. After 2003 the birth certificate underwent major changes in its format and included a set of new variables which are especially useful for our purposes. Most importantly, it asked the mother about her WIC status during the current pregnancy.[16] In addition it provides immensely detailed information on the demographics of the parents, socio-economic variables, rich information on 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.

[13] 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. Almost exactly similar patterns hold for the fathers. Thus, non-random selection into the program may be a valid concern.

[14] Figlio (2009) is one of the few papers which has managed to exploit an exclusion restriction to identify the effect of WIC participation on birth outcomes and deal with non-random selection beyond a selection on observables approach.

[15] Our main set of results uses smoking during the third trimester of pregnancy, however we also present results using a predetermined measure of smoking as given by maternal smoking behavior during the three months before pregnancy.

[16] Beginning in 2003 different states set different time lines to move to the new birth certificate protocol with relatively few states following it in the first few years. By 2012, 38 states had implemented the revision including, California, Colorado, Delaware, Florida, Georgia, Idaho, Illinois, Indiana, Iowa, Kansas, Kentucky, Louisiana, Maryland, Massachusetts, Michigan, Minnesota, Missouri, Montana, Nebraska, Nevada, New Hampshire, New Mexico, New York, North Carolina, North Dakota, Ohio, Oklahoma, Oregon, Pennsylvania, South Carolina, South Dakota, Tennessee, Texas, Utah, Vermont, Washington, Wisconsin, and Wyoming. These 38 states along with the District of Columbia cover 86.3 of all birth to U.S. residents in 2012.

5.3 Estimation of the Test Statistic

Following previous literature we can partially reduce selection concerns by restricting our sample to only mothers whose pregnancy was paid for by Medicaid. Given that we do not observe actual income of the respondents and that all individuals on Medicaid are automatically eligible for WIC, this restriction can give us comparable low income mothers from both treatment and control groups. However, we present results for both the full sample and the restricted Medicaid sample. The former sample gives us an ideal opportunity to implement our test in a case where massive selection is prevalent across treatment groups. Next we 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, and marital status, various interactions between the demographic variables of the mother, total number of prenatal visits, initiation of prenatal visits, whether the mother was suffering from hypertension or diabetes during or before 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.[17]

We use mother's smoking behavior in the third trimester of pregnancy as the bunching variable W, in our framework with an atom at zero. As Caetano (2015) shows there is prevalence of significant selection across smokers and non-smokers which is especially evident right at the threshold. We, however, will use these selection concerns combined with the idea that there should be no differential selection patterns between smokers and non-smokers, after controlling for Z, across our treatment and control groups to test for the presence of non-random selection into treatment.

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 $Y - Z^{\intercal}\gamma$ with a bandwidth of 4 and the standard Epanechnikov kernel. Figure 1 first presents the test statistic for 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 datasets that record birth outcomes 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.[18]

The test statistic from this specification in Figure 1, using third trimester smoking as W, is statistically significantly different from zero at -20.17 grams implying the existence of substantial amount of selection even after a fairly detailed set of covariates.[19] The test statistic is even larger for 'worse' set of covariates, for instance, if we control only for mother's race it is upwards of 40 grams. Figure 2 next presents results from the full specification detailed above. Most importantly it includes controls for previous and current pregnancy character-

[17] We employ a similar specification as the one used first by Almond, Chay and Lee (2005). For a complete list of covariates refer to Khalil (2015) which implements this specification to calculate actual treatment effects under the selection-on-observables assumption, for participation in WIC for a range of birth outcomes.

[18] However, we still remain extremely flexible in specifying how these covariates affect birth weight.

[19] These figures use the full sample to calculate the test statistic, however, Table 1 and Table 2 provide results using both the Medicaid and the full sample for various bandwidth and degree of polynomial combinations.

istics, any complications during current pregnancy, flexible controls for smoking behavior across pregnancy, and various interactions involving demographic variables./footnoteThe 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. The estimated test statistic falls down to -1.64 and is statistically indistinguishable from zero, indicating a substantial decrease in potential selection concerns.

Figure 1: Basic Specification - Third Trimester Smoking

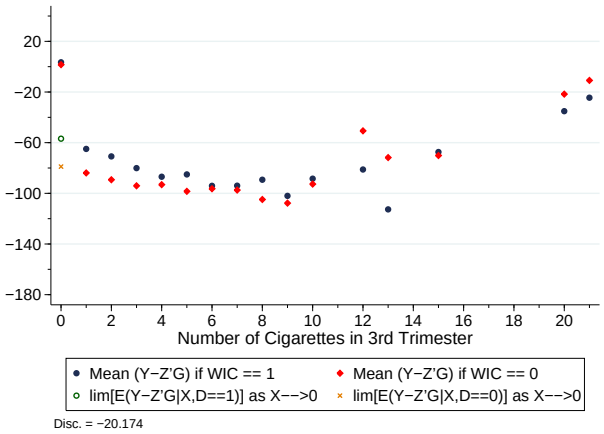


Figure 2: Full Specification - Third Trimester

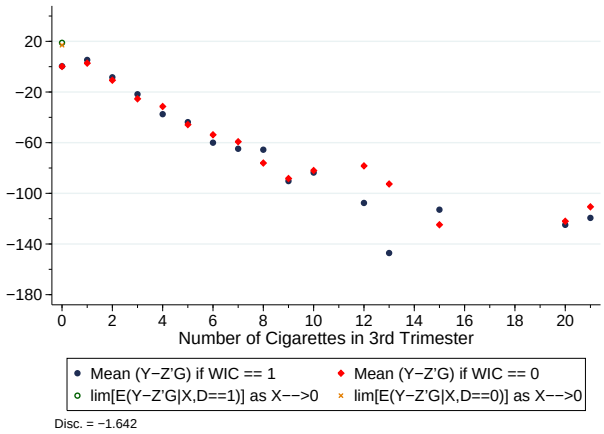


Figure 3: Basic Specification - Prepregnancy Smoking

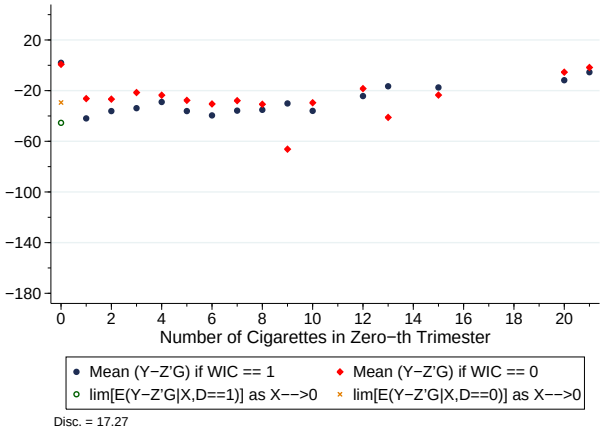
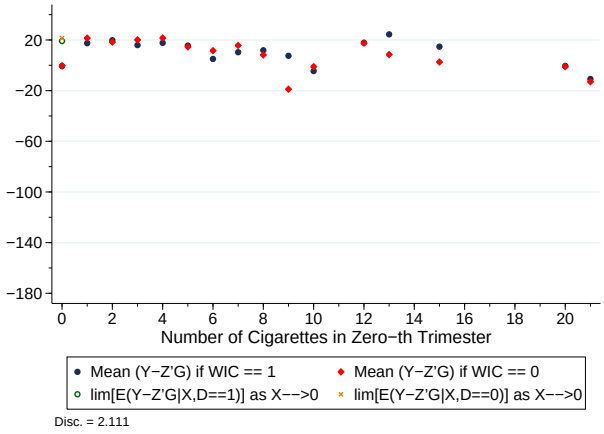


Figure 4: Full Specification - Prepregnancy Smoking



As an alternative specification, we try the above analysis with W as prepregnancy smok-

Table 1: Estimation of Test Statistic - 3rd Trimester Smoking

Bandwidth	3	4	5	6
Panel A: Full Sample				
Degree = 1	−2.615	−1.642	−4.927	−3.308
	(12.14)	(8.223)	(6.825)	(5.229)
Degree = 3	−2.393	−2.753	−16.63	12.00
	(6.343)	(9.850)	(40.29)	(25.71)
Panel B: Medicaid Sample				
Degree = 1	−11.34	−3.349	−2.724	0.107
	(16.21)	(10.94)	(9.039)	(6.927)
Degree = 3	−4.119	−8.983	−44.42	−17.67
	(8.525)	(13.19)	(53.13)	(34.03)

** represents significance at the 1% level. Treated groups in Medicaid sample has 3,278,311 individuals while the Full sample has 4,488,328. Control group in the Medicaid sample has 852,820 observations and the full sample has 4,950,406. Specification controls for the full set of covariates.

ing behavior instead of third trimester smoking.[20] To the extent that WIC participation affects smoking cessation differently for participants and non-participants, even after controlling for a rich set of covariates, we might consider the prepregnancy behavior as more suited to our test. Figure 3 and Figure 4, 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 Figure 4, with a test statistic of 2.11.

We next extend our results in table 1 and 2 with the estimation of the test statistic for various bandwidths and degree of the polynomial for both third trimester smoking and prepregnancy smoking. We use our most detailed specification and implement it for both the full sample and for the restricted Medicaid sample. Results show more robust results for the local linear estimator, as expected, given the spread of data shown in Figure 1 and 2. Moreover, the standard errors for the Medicaid sample are slightly larger owing to the smaller sample size of this restricted sample. However, for both the full and the restricted sample, across 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. There are a myriad of methods available, from hereon, for the estimation of the treatment effect under the selection on observables assumption. However, our framework provides the first method in the literature to test this crucial assumption in a binary treatment setting.

[20]We use the term 'zero-th' trimester of smoking for prepregnancy smoking to align with smoking during pregnancy across trimesters.

Table 2: Estimation of Test Statistic - Prepregnancy Smoking

Bandwidth	3	4	5	6
Panel A: Full Sample				
Degree = 1	9.141	2.111	1.098	2.975
	(12.21)	(8.233)	(6.777)	(5.294)
Degree = 3	4.508	8.198	37.156	23.363
	(6.372)	(9.948)	(41.103)	(25.681)
Panel B: Medicaid Sample				
Degree = 1	9.700	3.573	−1.957	-5.532
	(18.70)	(12.54)	(10.26)	(8.014)
Degree = 3	2.155	6.584	21.16	30.40
	(9.771)	(15.19)	(61.69)	(38.61)

** represents significance at the 1% level. Treated groups in Medicaid sample has 3,278,311 individuals while the Full sample has 4,488,328. Control group in the Medicaid sample has 852,820 observations and the full sample has 4,950,406. Specification controls for the full set of covariates.

6 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 main variable of interest could be any type of variable. We develop formal testing procedures for binary and continuous X. Our testing procedure hinges crucially on two conditions. First, there has to be a variable, W, among the set of controls that has a positive probability at a known point, but is otherwise continuous. Second, the structural function relating this variable 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 variable of interest, or the treatment variable, it could also be endogenous under the null. The test then checks whether the discontinuity in the expected outcome conditional on the variable with the bunching point, treatment and other possible controls is the same for treated and untreated individuals. 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.

A Appendix

A.1 Proof of Theorem 4.1:

We first analyze the asymptotic behavior of the infeasible estimator

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

where

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

Lemma A.1. *Suppose the conditions of Theorem 4.1 hold. Then*

$$\sqrt{nh}\left(\hat{\mu}_{\tilde{Y}|X,W}(1,0^+) - \hat{\mu}_{\tilde{Y}|X,W}(0,0^+) - (\mu_{\tilde{Y}|X,W}(1,0^+) - \mu_{\tilde{Y}|X,W}(0,0^+))\right) \xrightarrow{d} N(0,V). \quad (34)$$

Proof. First, we note that

$$\begin{aligned}\sqrt{nh}e_1^T M_{nx}^{-1}\frac{1}{n_x}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i &= \sqrt{n}\left(\frac{1}{\frac{n_x}{n}} - \frac{1}{P(X=x)}\right)\sqrt{h}e_1^T M_{nx}^{-1}\frac{1}{n}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i \\ &\quad + \sqrt{nh}e_1^T M_{nx}^{-1}\frac{1}{nP(X=x)}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i \\ &= \sqrt{nh}e_1^T M_{nx}^{-1}\frac{1}{nP(X=x)}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i + o_P(1). \quad (35)\end{aligned}$$

Define

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

where

$$N_{nx} := E(L_{ix}L_{ix}^\intercal K_h(W_i)|X_i = x).$$

Note that

$$E\left[L_{ix}K_h(W_i)\frac{\varepsilon_i}{P(X_i=x)}\right] = \frac{1}{P(X_i=x)}E\left[E\left(L_{ix}K_h(W_i)\varepsilon_i|W_i, X_i=x\right)\right] = 0.$$

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

$$e_1^T M_{nx}^{-1}\frac{1}{nP(X=x)}\sum_{i=1}^{n} L_{ix}K_h(W_i)\tilde{Y}_i = \mu_{\tilde{Y}|W,X}(0^+,x) + S_{nx} + O(h^2) + O_P\left(\frac{\log(n)}{nh}\right), \quad (36)$$

where

$$\mu_{\tilde{Y}|X,W}(x,0^+) := \lim_{w\downarrow 0}\int_{-\infty}^{\infty} y\frac{f_{\tilde{Y},W|X}(y,w|x)}{f_{W|X}(w|x)}\mathrm{d}y. \tag{37}$$

Note that $\mu_{\tilde{Y}|X,W}(x,0^+) = g(0,x) + \lim_{w\downarrow 0}\rho(w,x)$, since $E(\varepsilon|W,X)=0$ by definition, and g is continuous in w at $w=0$.

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 $p = P(X=1)$ we can write

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

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,

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

with

$$A = \left[\begin{array}{cc}\kappa_0 & \kappa_1\\ \kappa_1 & \kappa_2\end{array}\right],$$

where κ_j for $j=0,1,2$ is as in Assumption 4.2(vii). We will apply 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

$$e_1^TA^{-1}\left(\begin{array}{c}1\\ \frac{W_i}{h}\end{array}\right) = \frac{\kappa_2-\kappa_1\frac{W_i}{h}}{\kappa_0\kappa_2-\kappa_1^2}.$$

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

Using the standard change of variables argument with ν_0,ν_1,ν_2 as in Assumption 4.2(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 Lindeberg-Feller Theorem we also need to verify that

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

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

Lemma A.2. $\sqrt{nh}\left(\hat{\mu}_{\hat{\tilde{Y}}|X,W}(x,0^+) - \hat{\mu}_{\tilde{Y}|X,W}(x,0^+)\right) = o_P(1)$.

Proof.

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

□

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

$$\begin{aligned}
&\sqrt{nh} \frac{1}{n_{x0}} \sum_{i=1}^{n} \left(\hat{\tilde{Y}}_i - \tilde{Y}_i\right) 1\{W_i = 0, X_i = x\} \\
&= -\sqrt{nh} \frac{1}{n_{x0}} \sum_{i=1}^{n} Z_i^\intercal (\hat{\gamma} - \gamma) = o_P(1).
\end{aligned}$$

Similarly, $\sqrt{nh}\left(\frac{1}{n_{x0}} \sum_{i=1}^{n} \tilde{Y}_i 1\{W_i = 0, X_i = x\} - E[\tilde{Y}_i | W_i = 0, X_i = x]\right) = o_P(1)$. Thus, the conclusion of Theorem 4.1 follows from combining these last two statements with the conclusions of Lemmas A.1 and A.2.

A.2 Proof of Theorem 4.3:

First, let $\hat{P}_W(X_i), P_W(X_i)$ denote $\hat{E}(1\{W = 0\}|X_i)$ and $E(1\{W = 0\}|X_i)$, respectively. Note that using Theorem 37 of Pollard (1984) and the fact that $\inf_{x \in \mathcal{X}} P_W(x) = \alpha > 0$, which is

implied by Assumption 4.3, we can argue that

$$\begin{aligned}
\sup_{x\in\mathcal{X}}\left|\hat{M}_n(x,0^+) - Af_{XW}(x,0^+)\right| &\xrightarrow{P} 0,\\
\sup_{x\in\mathcal{X}}\left|\hat{\mu}_{\tilde{Y}|X,W}(x,0^+) - \mu_{\tilde{Y}|X,W}(x,0^+)\right| &\xrightarrow{P} 0,\\
\sup_{x\in\mathcal{X}}\left|\frac{1}{\hat{P}_W(x)} - \frac{1}{P_W(x)}\right| &\xrightarrow{P} 0,\\
\sup_{x\in\mathcal{X}}\left|\frac{\mu_{1\{W=0\}\tilde{Y}}(X_i)|X}{P_W(x)}\left[\frac{1}{\hat{P}_W(x)} - \frac{1}{P_W(x)}\right]\right| &\xrightarrow{P} 0,\\
\sup_{x\in\mathcal{X}}\left|\hat{\mu}_{\tilde{Y}|X,W}(x,0) - \mu_{\tilde{Y}|X,W}(x,0)\right| &\xrightarrow{P} 0,\\
\frac{1}{\sqrt{n}}\sum_i\left[\hat{\mu}_{1\{W=0\}\tilde{Y}|X}(X_i) - \mu_{1\{W=0\}\tilde{Y}|X}(X_i)\right] &= O_P(1),\\
\frac{1}{\sqrt{n}}\sum_i\left[\hat{P}_W(X_i) - P_W(X_i)\right] &= O_P(1).
\end{aligned}$$

To prove Theorem 4.3 we first analyze the asymptotic behavior of

$$\frac{\sqrt{nh}}{n^2}\sum_i\sum_j\left\{[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+) - \hat{\mu}_{\tilde{Y}|X,W}(X_i,0) - (\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+) - \mu_{\hat{\tilde{Y}}|X,W}(X_j,0))]^2 - t_n\right\}, \tag{38}$$

where

$$t_n = \frac{1}{n^2}\sum_i\sum_j\tilde{\lambda}^2(X_i,X_j).$$

Now (38) equals

$$\frac{2\sqrt{h}}{\sqrt{n}}\sum_i\left[(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+) - \hat{\mu}_{\tilde{Y}|X,W}(X_i,0))^2 - (\mu_{\tilde{Y}|X,W}(X_i,0^+) - \mu_{\tilde{Y}|X,W}(X_i,0))^2\right] \tag{39}$$

$$-2\frac{\sqrt{nh}}{n^2}\sum_i\sum_j\Big[(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+) - \hat{\mu}_{\tilde{Y}|X,W}(X_i,0))(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+) - \hat{\mu}_{\tilde{Y}|X,W}(X_j,0))$$

$$- (\mu_{\tilde{Y}|X,W}(X_i,0^+) - \mu_{\tilde{Y}}(X_i,0))(\mu_{\tilde{Y}|X,W}(X_j,0^+) - \mu_{\tilde{Y}}(X_j,0))\Big] \tag{40}$$

Next, note that 39 equals

$$\frac{2\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0^+) - \mu^2_{\tilde{Y}|X,W}(X_i,0^+)\right] \tag{41}$$

$$+\frac{2\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0) - \mu^2_{\tilde{Y}|X,W}(X_i,0)\right] \tag{42}$$

$$-\frac{4\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0) - \mu_{\tilde{Y}|X,W}(X_i,0^+)\mu_{\tilde{Y}|X,W}(X_i,0)\right]. \tag{43}$$

Lemma A.3. *Suppose the conditions of Theorem 4.3 hold. Then*

$$\frac{2\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0^+)-\mu^2_{\tilde{Y}|X,W}(X_i,0^+)\right]\xrightarrow{d}N(0,4V_1),$$

where $V_1=C\int_{-\infty}^{\infty}\sigma^2_\varepsilon(x,0^+)\frac{\mu^2_{\tilde{Y}|X,W}(x,0^+)}{f_{XW}(x,0^+)}f^2_X(x)\mathrm{d}x$ *with*
$C=\int_{-\infty}^{\infty}\int_0^{\infty}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\u_x\\u_w\end{pmatrix}\right]^2K^2(u_w)K^2(u_w)\mathrm{d}u_w\mathrm{d}u_x.$

Proof. Now 41 equals

$$\frac{2\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)+\mu_{\tilde{Y}|X,W}(X_i,0^+)\right]\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\mu_{\tilde{Y}|X,W}(X_i,0^+)\right]\tag{44}$$

We know that $\hat{\mu}_{\tilde{Y}|X,W}(x,0^+)$ converges uniformly and almost surely to $\mu_{\tilde{Y}|X,W}(x,0^+)$. In addition,

$$\hat{\mu}_{\tilde{Y}|X,W}(x,0^+)=\mu_{\tilde{Y}|X,W}(x,0^+)+\tilde{S}_n(x)+O(h^2)+O_P\left(\frac{\log(n)}{nh^2}\right)\tag{45}$$

uniformly, where

$$\tilde{S}_n(x)=e_1^\intercal A^{-1}f^{-1}_{XW}(x,0^+)\frac{1}{n}\sum_{l=1}^n\begin{pmatrix}1\\\frac{X_l-x}{h}\\\frac{W_l}{h}\end{pmatrix}1\{W_l>0\}\frac{1}{h^2}K\left(\frac{X_l-x}{h}\right)K\left(\frac{W_l}{h}\right)\varepsilon_l,\tag{46}$$

with $\varepsilon_l=Y_l-E(Y_l|X_l,W_l)$ and A as defined above. Combining these facts we get that 41 equals

$$\frac{4\sqrt{h}}{n^{3/2}}\sum_i\sum_l\frac{\mu_{\tilde{Y}|X,W}(X_i,0^+)}{f_{XW}(X_i,0^+)}e_1^\intercal A^{-1}\begin{pmatrix}1\\\frac{X_l-X_i}{h}\\\frac{W_l}{h}\end{pmatrix}\frac{1\{W_l>0\}}{h^2}K\left(\frac{X_l-X_i}{h}\right)K\left(\frac{W_l}{h}\right)\varepsilon_l+o_P(1).\tag{47}$$

Let $\xi_i=(X_i,W_i,\varepsilon_i)$ and

$$\begin{aligned}\kappa_n(\xi_i,\xi_l):=&\frac{\mu_{\tilde{Y}|X,W}(X_i,0^+)}{f_{XW}(X_i,0^+)}e_1^\intercal A^{-1}\begin{pmatrix}1\\\frac{X_l-X_i}{h}\\\frac{W_l}{h}\end{pmatrix}\frac{1\{W_l>0\}}{h^{3/2}}K\left(\frac{X_l-X_i}{h}\right)K\left(\frac{W_l}{h}\right)\varepsilon_l\\&+\frac{\mu_{\tilde{Y}|X,W}(X_l,0^+)}{f_{XW}(X_l,0^+)}e_1^\intercal A^{-1}\begin{pmatrix}1\\\frac{X_i-X_l}{h}\\\frac{W_i}{h}\end{pmatrix}\frac{1\{W_i>0\}}{h^{3/2}}K\left(\frac{X_i-X_l}{h}\right)K\left(\frac{W_i}{h}\right)\varepsilon_i\end{aligned}\tag{48}$$

The term 41 has the same asymptotic distribution as

$$\frac{2}{\sqrt{n}(n-1)}\sum_i\sum_{l\neq i}\kappa_n(\xi_i,\xi_l).$$

$$\begin{aligned}
\kappa_n^2(\xi_i,\xi_l) =&\frac{\mu^2_{\tilde{Y}|X,W}(X_i,0^+)}{f^2_{XW}(X_i,0^+)}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ \frac{X_l-X_i}{h}\\ \frac{W_l}{h}\end{pmatrix}\right]^2\frac{1\{W_l>0\}}{h^3}K^2\left(\frac{X_l-X_i}{h}\right)K^2\left(\frac{W_l}{h}\right)\varepsilon_l^2\\
&+\frac{\mu^2_{\tilde{Y}|X,W}(X_l,0^+)}{f^2_{XW}(X_l,0^+)}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ \frac{X_i-X_l}{h}\\ \frac{W_i}{h}\end{pmatrix}\right]^2\frac{1\{W_i>0\}}{h^3}K^2\left(\frac{X_i-X_l}{h}\right)K^2\left(\frac{W_i}{h}\right)\varepsilon_i^2\\
&+2\frac{\mu_{\tilde{Y}|X,W}(X_i,0^+)\mu_{\tilde{Y}|X,W}(X_l,0^+)}{f_{XW}(X_i,0^+)f_{XW}(X_l,0^+)}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ \frac{X_l-X_i}{h}\\ \frac{W_l}{h}\end{pmatrix}\right]\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ \frac{X_i-X_l}{h}\\ \frac{W_i}{h}\end{pmatrix}\right]\\
&\times\frac{1\{W_l>0,W_i>0\}}{h^3}K\left(\frac{X_l-X_i}{h}\right)K\left(\frac{X_i-X_l}{h}\right)K\left(\frac{W_l}{h}\right)K\left(\frac{W_i}{h}\right)\varepsilon_l\varepsilon_i.
\end{aligned}$$

By the law of iterated expectations the expectation of the last term equals 0.

$$\begin{aligned}
\frac{1}{n}E[\kappa_n^2(\xi_i,\xi_l)] =&\frac{2}{n}E\left[\frac{\mu^2_{\tilde{Y}|X,W}(X_i,0^+)}{f^2_{XW}(X_i,0^+)}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ \frac{X_l-X_i}{h}\\ \frac{W_l}{h}\end{pmatrix}\right]^2\frac{1\{W_l>0\}}{h^3}K^2\left(\frac{X_l-X_i}{h}\right)K^2\left(\frac{W_l}{h}\right)\varepsilon_l^2\right]\\
=&\frac{2}{nh^2}E\left\{\int_{-\infty}^{\infty}\frac{\mu^2_{\tilde{Y}|X,W}(X_l-u_xh,0^+)}{f^2_{XW}(X_l-u_xh,0^+)}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ u_x\\ \frac{W_l}{h}\end{pmatrix}\right]^2K^2(u_x)\right.\\
&\left.\times1\{W_l>0\}K^2\left(\frac{W_l}{h}\right)\sigma^2_\varepsilon(X_l,W_l)f_X(X_l-u_xh)\mathrm{d}u_x\right\}\\
=&\frac{2}{nh}\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty}\frac{\mu^2_{\tilde{Y}|X,W}(X_l-u_xh,0^+)}{f^2_{XW}(X_l-u_xh,0^+)}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ u_x\\ u_w\end{pmatrix}\right]^2K^2(u_x)\\
&\times K^2(u_w)\,\sigma^2_\varepsilon(X_l,W_l)f_X(X_l-u_xh)f_{XW}(X_l,u_wh)\mathrm{d}u_x\mathrm{d}u_w\mathrm{d}X_l
\end{aligned}$$

Since $nh^2\to\infty$, $\sigma^2_\varepsilon(x,w)$ is continuous in $(x,w)\in\mathrm{Supp}(X)\times[0,\delta]$ and $\int\frac{\mu^2_{\tilde{Y}|X,W}(x,0^+)}{f_{XW}(x,0^+)}\sigma^2_\varepsilon(x,0^+)f_X(x)\mathrm{d}x<\infty$,then $E[\kappa_n^2(\xi_i,\xi_j)]=o(n)$. This means that by the H-P-S-S Lemma of Powell, Stock and Stoker (1989) we have

$$\frac{2}{\sqrt{n}(n-1)}\sum_i\sum_{l\neq i}\kappa_n(\xi_i,\xi_l)=\frac{2}{\sqrt{n}}\sum_{i=1}^n E[\kappa_n(\xi_i,\xi_l)|Z_i]+o_P(1).$$

Next, note that

$$\begin{aligned}
E[\kappa_n(\xi_i,\xi_l)|\xi_i] =&E\left[\frac{\mu_{\tilde{Y}|X,W}(X_l,0^+)}{f_{XW}(X_l,0^+)}e_1^\intercal A^{-1}\begin{pmatrix}1\\ \frac{X_i-X_l}{h}\\ \frac{W_i}{h}\end{pmatrix}\frac{1\{W_i>0\}}{h^{3/2}}K\left(\frac{X_i-X_l}{h}\right)K\left(\frac{W_i}{h}\right)\varepsilon_i|\varepsilon_i,X_i\right]\\
=&\frac{\varepsilon_i1\{W_i>0\}}{\sqrt{h}}K\left(\frac{W_i}{h}\right)e_1^\intercal A^{-1}\int_{-\infty}^{\infty}\begin{pmatrix}1\\ u_x\\ \frac{W_i}{h}\end{pmatrix}\frac{\mu_{\tilde{Y}}(X_i-u_xh,0^+)K(u_x)}{f_{XW}(X_i-u_xh,0^+)}f_X(X_i-u_xh)\mathrm{d}u_x.
\end{aligned}$$

Since

$$\frac{\varepsilon_i1\{W_i>0\}}{\sqrt{h}}K\left(\frac{W_i}{h}\right)e_1^\intercal A^{-1}\int_{-\infty}^{\infty}\begin{pmatrix}1\\ u_x\\ \frac{W_i}{h}\end{pmatrix}K(u_x)\mathrm{d}u_x=O_P(1),$$

and

$$\sup_{(x,u_x)\in\mathcal{X}\times[a,b]}\left|\frac{\mu_{\tilde{Y}|X,W}(x-u_xh,0^+)f_X(x-u_xh))}{f_{XW}(x-u_xh,0^+)}-\frac{\mu_{\tilde{Y}|X,W}(x,0^+)f_X(x))}{f_{XW}(x,0^+)}\right|\to 0,$$

we have

$$\frac{2}{\sqrt{n}}\sum_{i=1}^{n}E[\kappa_n(\xi_i,\xi_l)|\xi_i]=2\sum_{i=1}^{n}r_n(\xi_i)+o_P(1),$$

where $r_n(\xi_i):=\frac{\varepsilon_i 1\{W_i>0\}}{\sqrt{nh}}K\left(\frac{W_i}{h}\right)\frac{\mu_{\tilde{Y}|X,W}(X_i,0^+)f_X(X_i)}{f_{XW}(X_i,0^+)}e_1^\intercal A^{-1}\int_{-\infty}^{\infty}\begin{pmatrix}1\\ u_x\\ \frac{W_i}{h}\end{pmatrix}K(u_x)\mathrm{d}u_x.$

$$\begin{aligned}\sum_i E\left(r_n(\xi_i)\right)^2=&\int_{-\infty}^{\infty}\int_0^{\infty}\int_{-\infty}^{\infty}\sigma_\varepsilon^2(X_i,u_wh)K^2(u_w)\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ u_x\\ u_w\end{pmatrix}\right]^2\\ &\times K^2(u_x)\frac{\mu^2_{\tilde{Y}|X,W}(X_i,0^+)}{f^2_{XW}(X_i,0^+)}f_X^2(X_i)f_{XW}(X_i,u_wh)\mathrm{d}u_x\mathrm{d}u_w\mathrm{d}X_i\\ \to&\int_{-\infty}^{\infty}\int_0^{\infty}\left[e_1^\intercal A^{-1}\begin{pmatrix}1\\ u_x\\ u_w\end{pmatrix}\right]^2K^2(u_w)K^2(u_w)\mathrm{d}u_w\mathrm{d}u_x\\ &\times\int_{-\infty}^{\infty}\sigma_\varepsilon^2(X_i,0^+)\frac{\mu^2_{\tilde{Y}|X,W}(X_i,0^+)}{f_{XW}(X_i,0^+)}f_X^2(X_i)\mathrm{d}X_i=:V_1.\end{aligned}$$

Next, let $\delta>0$ and consider $\sum_i E[\|r_n(\xi_i)\|^2 1\{\|r_n(\xi_i)\|>\delta\}]$. By Holder's inequality this is less than or equal to

$$\left(P(\|r_n(\xi)\|>\delta)\right)^{(1+\alpha)/(2+\alpha)}\sum_i\left(E\|r_n(\xi)\|^{2+\alpha}\right)^{1/(2+\alpha)}\to 0.$$

Therefore, by the Lindeberg-Feller Central Limit Theorem we have

$$2\sum_i r_n(Z_i)\xrightarrow{d}N(0,4V_1).$$

Given the analysis so far, this shows that the limiting distribution of 41 is $N(0,4V_1)$. □

Lemma A.4. *Suppose the conditions of Theorem 4.3 hold. Then*

$$\frac{2\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0)-\mu^2_{\tilde{Y}|X,W}(X_i,0)\right]=o_P(1).$$

Proof.

$$\begin{aligned}\frac{1}{\sqrt{n}}\sum_i\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\mu_{\tilde{Y}|X,W}(X_i,0)\right]=&\frac{1}{\sqrt{n}}\sum_i\left[\frac{\hat{\mu}_{1\{W=0\}\tilde{Y}|X}(X_i)}{\hat{P}_W(X_i)}-\frac{\mu_{1\{W=0\}\tilde{Y}|X}(X_i)}{P_W(X_i)}\right]\\ =&\frac{1}{\sqrt{n}}\sum_i\frac{\hat{\mu}_{1\{W=0\}\tilde{Y}|X}(X_i)-\mu_{1\{W=0\}\tilde{Y}|X}(X_i)}{\hat{P}_W(X_i)}\\ &-\frac{1}{\sqrt{n}}\sum_i\frac{\hat{P}_W(X_i)-P_W(X_i)}{\hat{P}_W(X_i)P_W(X_i)}\mu_{1\{W=0\}\tilde{Y}|X}(X_i).\end{aligned}$$

By the arguments at the beginning of this subsection, we have

$$\frac{2}{\sqrt{n}}\sum_i\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\mu_{\tilde{Y}|X,W}(X_i,0)\right]=O_P(1),$$

and

$$\begin{aligned}\tfrac{2}{\sqrt{n}}\textstyle\sum_i\left[\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0)-\mu^2_{\tilde{Y}|X,W}(X_i,0)\right]=&\tfrac{2}{\sqrt{n}}\textstyle\sum_i\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)+\mu_{\tilde{Y}|X,W}(X_i,0)\right]\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\mu_{\tilde{Y}|X,W}(X_i,0)\right]\\=&\tfrac{4}{\sqrt{n}}\textstyle\sum_i\mu_{\tilde{Y}|X,W}(X_i,0)\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\mu_{\tilde{Y}|X,W}(X_i,0)\right]+o_P(1)\\=&O_P(1).\end{aligned}$$

Thus, $\frac{2\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0)-\mu^2_{\tilde{Y}|X,W}(X_i,0)\right]=o_P(1)$. □

Lemma A.5. *Suppose the conditions of Theorem 4.3 hold. Then*

$$\frac{4\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\mu_{\tilde{Y}|X,W}(X_i,0^+)\mu_{\tilde{Y}|X,W}(X_i,0)\right]\xrightarrow{d}N(0,4V_2),\quad(49)$$

where $V_2=C\int_{-\infty}^{\infty}\sigma^2_\varepsilon(x,0^+)\frac{\mu^2_{\tilde{Y}|X,W}(x,0)}{f_{XW}(x,0^+)}f^2_X(x)\mathrm{d}x$.

Proof. Given the arguments in the proof of the previous lemma, we have

$$\begin{aligned}&\frac{4\sqrt{h}}{\sqrt{n}}\sum_i\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\mu_{\tilde{Y}|X,W}(X_i,0^+)\mu_{\tilde{Y}|X,W}(X_i,0)\right]\\=&\frac{4\sqrt{h}}{\sqrt{n}}\sum_i\mu_{\tilde{Y}|X,W}(X_i,0)\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\mu_{\tilde{Y}|X,W}(X_i,0^+)\right]+o_P(1).\end{aligned}$$

Then using arguments analogous to those given in the analysis of asymptotic behavior of 41 we can show that that

$$\frac{4\sqrt{h}}{\sqrt{n}}\sum_i\mu_{\tilde{Y}|X,W}(X_i,0)\left[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\mu_{\tilde{Y}|X,W}(X_i,0^+)\right]=\frac{2}{\sqrt{n}(n-1)}\sum_i\sum_{l\neq i}\lambda_n(\xi_i,\xi_l)+o_P(1),$$

where

$$\begin{aligned}\lambda_n(\xi_i,\xi_l):=&\frac{\mu_{\tilde{Y}|X,W}(X_i,0)}{f_{XW}(X_i,0^+)}e_1^\intercal A^{-1}\begin{pmatrix}1\\\frac{X_l-X_i}{h}\\\frac{W_l}{h}\end{pmatrix}\frac{1\{W_l>0\}}{h^{3/2}}K\left(\frac{X_l-X_i}{h}\right)K\left(\frac{W_l}{h}\right)\varepsilon_l\\&+\frac{\mu_{\tilde{Y}|X,W}(X_l,0)}{f_{XW}(X_l,0^+)}e_1^\intercal A^{-1}\begin{pmatrix}1\\\frac{X_i-X_l}{h}\\\frac{W_i}{h}\end{pmatrix}\frac{1\{W_i>0\}}{h^{3/2}}K\left(\frac{X_i-X_l}{h}\right)K\left(\frac{W_i}{h}\right)\varepsilon_i.\quad(50)\end{aligned}$$

Moreover, by the H-P-S-S lemma we also have

$$\begin{aligned}\frac{2}{\sqrt{n}(n-1)}\sum_i\sum_{l\neq i}\lambda_n(\xi_i,\xi_l)=&\frac{2}{\sqrt{n}}\sum_i E[\lambda_n(\xi_i,\xi_l)|\xi_i]+o_P(1)\\\xrightarrow{P}&N(0,4V_2).\end{aligned}$$

□

Lemma A.6. *Suppose the conditions of Theorem 4.3 hold. Then*

$$\frac{\sqrt{nh}}{n^2}\sum_i\sum_j\Big[(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0))(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_j,0))$$
$$-(\mu_{\tilde{Y}|X,W}(X_i,0^+)-\mu_{\tilde{Y}}(X_i,0))(\mu_{\tilde{Y}|X,W}(X_j,0^+)-\mu_{\tilde{Y}}(X_j,0))\Big]=o_P(1).$$

Proof. Note that the expression in the statement of the lemma equals

$$2\left[\frac{\sqrt{h}}{\sqrt{n}}\sum_i\left(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\mu_{\tilde{Y}|X,W}(X_i,0^+)\right)\right]\left[\frac{1}{n}\sum_j\left(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+)-\mu_{\tilde{Y}|X,W}(X_j,0^+)\right)\right]$$
$$-4\left[\frac{\sqrt{h}}{\sqrt{n}}\sum_i\left(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\mu_{\tilde{Y}|X,W}(X_i,0^+)\right)\right]\left[\frac{1}{n}\sum_j\left(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0)-\mu_{\tilde{Y}|X,W}(X_j,0)\right)\right]$$
$$+2\sqrt{h}\left[\frac{1}{\sqrt{n}}\sum_i\left(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-\mu_{\tilde{Y}|X,W}(X_i,0)\right)\right]\left[\frac{1}{n}\sum_j\left(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0)-\mu_{\tilde{Y}|X,W}(X_j,0)\right)\right]$$
$$=2O_P(1)o_P(1)-4O_P(1)o_P(1)+2\sqrt{h}O_P(1)o_P(1)=o_P(1).$$

□

Combining Lemmas A.3-A.6 we get the following result:

Lemma A.7. *Suppose the conditions of Theorem 4.3 hold. Then*

$$\tfrac{\sqrt{nh}}{n^2}\textstyle\sum_i\sum_j\left[[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+)-\mu_{\hat{\tilde{Y}}|X,W}(X_j,0))]^2-t_n\right]\xrightarrow{d}N(0,4(V_1+V_2)).$$

The following lemma is obvious:

Lemma A.8. *Suppose the conditions of Theorem 4.3 hold. Then* $\sqrt{nh}(t_n-t)=o_P(1)$.

Next, consider

$$\frac{1}{n^2}\sum_i\sum_j\Big\{[\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0^+)-\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0)-(\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_j,0^+)-\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_j,0))]^2$$
$$-[\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)-(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_j,0))]^2\Big\}$$
$$=\frac{2}{n}\sum_i\left[(\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0^+)-\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0))^2-(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0))^2\right] \quad (51)$$
$$-2\frac{1}{n^2}\sum_i\sum_j\Big[(\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0^+)-\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0))(\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_j,0^+)-\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_j,0))$$
$$-(\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0))(\hat{\mu}_{\tilde{Y}|X,W}(X_j,0^+)-\hat{\mu}_{\tilde{Y}|X,W}(X_j,0))\Big]. \quad (52)$$

Lemma A.9. *Suppose the conditions of Theorem 4.3 hold. Then term (52) is* $O_P\left(n^{-1/2}\right)$.

Proof. $\sqrt{n}$ times (52) can be written as

$$\begin{aligned}
&\left\{\frac{2}{n}\sum_{j=1}^{n}\left[\hat{\mu}_{Y|X,W}(X_j,0^+)-\hat{\mu}_{Y|X,W}(X_j,0)\right]-\frac{1}{n}\sum_{j=1}^{n}\left[\hat{\mu}_{Z}^{\intercal}(X_j,0^+)-\hat{\mu}_{Z|X,W}^{\intercal}(X_j,0)\right](\hat{\gamma}+\gamma)\right\}\\
&\qquad\times\frac{2}{n}\sum_{j=1}^{n}\left[\hat{\mu}_{Z|X,W}^{\intercal}(X_j,0^+)-\hat{\mu}_{Z|X,W}^{\intercal}(X_j,0)\right]\sqrt{n}(\hat{\gamma}-\gamma)\\
=&\left\{\frac{2}{n}\sum_{j=1}^{n}\left[\mu_{Y|X,W}(X_j,0^+)-\mu_{Y|X,W}(X_j,0)\right]-\frac{2}{n}\sum_{j=1}^{n}\left[\mu_{Z|X,W}^{\intercal}(X_j,0^+)-\mu_{Z|X,W}^{\intercal}(X_j,0)\right]\gamma)\right\}\\
&\qquad\times\frac{2}{n}\sum_{j=1}^{n}\left[\mu_{Z|X,W}^{\intercal}(X_j,0^+)-\mu_{Z|X,W}^{\intercal}(X_j,0)\right]\sqrt{n}(\hat{\gamma}-\gamma)+o_P(1).
\end{aligned}$$

Since all the sums in the above expression are $O_P(1)$ and $\sqrt{n}(\hat{\gamma}-\gamma)=O_P(1)$, this proves the result. □

$\sqrt{n}$ times (52) is $O_P(1)$, which means that $\sqrt{nh}$ times (52) is $o_P(1)$. Next we turn to the analysis of 51, which equals

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n}\left[\hat{\mu}^2_{\hat{\hat{Y}}|X,W}(X_i,0^+)-\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0^+)\right]+\frac{2}{n}\sum_{i=1}^{n}\left[\hat{\mu}^2_{\hat{\hat{Y}}|X,W}(X_i,0)-\hat{\mu}^2_{\tilde{Y}}(X_i,0)\right]\\
&\qquad-\frac{4}{n}\sum_{i=1}^{n}\left[\hat{\mu}_{\hat{\hat{Y}}|X,W}(X_i,0^+)\hat{\mu}_{\hat{\hat{Y}}|X,W}(X_i,0)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)\right]
\end{aligned}$$

Lemma A.10. *Suppose the conditions of Theorem 4.3 hold. Then*

$$\begin{aligned}
\frac{2\sqrt{nh}}{n}\sum_{i=1}^{n}\left[\hat{\mu}^2_{\hat{\hat{Y}}|X,W}(X_i,0^+)-\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0^+)\right] &= o_P(1),\\
\frac{2\sqrt{nh}}{n}\sum_{i=1}^{n}\left[\hat{\mu}^2_{\hat{\hat{Y}}|X,W}(X_i,0)-\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0)\right] &= o_P(1).
\end{aligned}$$

Proof.

$$\begin{aligned}
\frac{2}{n}\sum_{i=1}^{n}\left[\hat{\mu}^2_{\hat{\hat{Y}}|X,W}(X_i,0^+)-\hat{\mu}^2_{\tilde{Y}|X,W}(X_i,0^+)\right]=&-\frac{4}{n}\sum_{i=1}^{n}\hat{\mu}_{Y|X,W}(X_i,0^+)\hat{\mu}_{Z|X,W}^{\intercal}(X_i,0^+)(\hat{\gamma}-\gamma)\\
&+2(\hat{\gamma}+\gamma)^{\intercal}\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{Z|X,W}(X_i,0^+)\hat{\mu}_{Z|X,W}^{\intercal}(X_i,0^+)(\hat{\gamma}-\gamma).
\end{aligned}$$

Both of the sums on the right side of the above equation are $O_P(1)$ and $(\hat{\gamma}-\gamma)=O_P\left(n^{-1/2}\right)$. This proves the first result. The second result can be proved in a similar way. □

Lemma A.11. *Under the conditions of Theorem 4.3, we have*

$$-4\frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n}[\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0^+)\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)]=o_P(1).$$

Proof. Finally,

$$\begin{aligned}
&-4\frac{1}{n}\sum_{i=1}^{n}[\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0^+)\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)]\\
&\qquad\qquad=-4(\hat{\gamma}-\gamma)^{\intercal}\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{Z|X,W}(X_i,0)\hat{\mu}^{\intercal}_{Z|X,W}(X_i,0^+)(\hat{\gamma}-\gamma)\\
&+4(\hat{\gamma}-\gamma)^{\intercal}\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)\hat{\mu}_{Z|X,W}(X_i,0^+)+4(\hat{\gamma}-\gamma)^{\intercal}\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{Z|X,W}(X_i,0)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+).
\end{aligned}$$

Since, $\hat{\gamma}-\gamma=o_P(1)$ and $\frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{Z|X,W}(X_i,0)\hat{\mu}^{\intercal}_{Z|X,W}(X_i,0^+)=O_P(1)$,

$$\begin{aligned}
&-4\frac{\sqrt{h}}{\sqrt{n}}\sum_{i=1}^{n}[\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0^+)\hat{\mu}_{\hat{\tilde{Y}}|X,W}(X_i,0)-\hat{\mu}_{\tilde{Y}|X,W}(X_i,0^+)\hat{\mu}_{\tilde{Y}|X,W}(X_i,0)]\\
&=4\sqrt{nh}(\hat{\gamma}-\gamma)^{\intercal}E\big[\mu_{\tilde{Y}|X,W}(X_i,0)\mu_{Z|X,W}(X_i,0^+)+\mu_{\tilde{Y}|X,W}(X_i,0^+)\mu_{Z|X,W}(X_i,0)\big]+o_P(1)\\
&\qquad\qquad=o_P(1).
\end{aligned}$$

□

Combining Lemmas A.9-A.11, we get

$$\sqrt{nh}\left(\hat{t}_n-\hat{t}_n^{Inf}\right)=o_P(1).$$

This completes the proof of of Theorem 4.3.

References

[1] Almond, D., K. Chay, and D. Lee (2005): The Costs of Low Birth Weight, *The Quarterly Journal of Economics*, **120**, 1031-1083.

[2] Almond, D. and J. Currie (2011): Killing me softly: The fetal origins hypothesis, *The Journal of Economic Perspectives*, **25**, 153-172.

[3] Bitler, M. P. and J. Currie (2005): Does WIC work? The effects of WIC on pregnancy and birth outcomes, *Journal of Policy Analysis and Management*, **24**, 73-91.

[4] Caetano, C. (2015): A Test of Exogeneity Without Instrumental Variables in Models With Bunching, *Econometrica*, **83**, 1581-1600.

[5] Caetano, C., C. Rothe and N. Yıldız (2016): (2015): A Discontinuity Test for Identification in Triangular Nonseparable Models, forthcoming in *Journal of Econometrics*.

[6] Caetano, G. and V. Maheshri (2015): Identifying Dynamic Spillovers of Crime: An Empirical Approach to Model Selection, *Unpublished Manuscript*, University of Rochester.

[7] Crump, R. K., V. J. Hotz, G. W. Imbens and O. A. Mitnik (2008): Nonparametric Tests for Treatment Effect Heterogeneity, *Review of Economics and Statistics*, **90**, 389-405.

[8] Currie, J. and E. Moretti (2003): Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings, *Quarterly Journal of Economics*, **118**, 1495-1532.

[9] Figlio, D. (2009): Does prenatal WIC participation improve birth outcomes? New evidence from Florida, *Journal of Public Economics*, **93**, 235-245.

[10] Goldberger, A. S., 1991, *A Course in Econometrics* (Harvard University Press: Cambridge, MA).

[11] Gourio, F. and N. Roys (2014): Size-Dependent Regulations, Firm Size Distribution And Allocation *Quantitative Journal of Economics*, **5**, 377-416.

[12] Heckman, J. J., S. Urzua, , and E. J. Vytlacil (2006): Understanding instrumental variables in models with essential heterogeneity, *Review of Economics and Statistics*, 88, 389432.

[13] Hoderlein, S., L. Su, H. White and T.T. Yang (2014): Testing for Monotonicity in Unobservables under Unconfoundedness, *Working Paper*, Department of Economics, Boston College.

[14] Jacob, B., L. Lefgren, and E. Moretti (2007): The dynamics of criminal behavior evidence from weather shocks, *Journal of Human Resources*, **42**, 489-527

[15] Khalil, U. (2015): *Heterogenous Effects of WIC Participation on Birth Outcomes*, unpublished manuscript, University of Rochester.

[16] Kitagawa, T. (2015): A Test for Instrument Validity, *Econometrica*, **83**, 2043-2063.

[17] Pollard, D., 1984: *Convergence of Stochastic Processes*, (New York: Springer-Verlag).

[18] Powell, J.L, J.H. Stock, and I.M. Stoker (1989): Semiparametric estimation of index coefficients, *Econometrica*, **57**, 14031430.

[19] Robinson, P. M. (1988): Root-N-Consistent Semiparametric Regression, *Econometrica*, **56**, 931-954.

B Supplementary Appendix:

B.1 Illustrative Examples:

B.1.1 Discrete X:

Suppose $X = 1\{\alpha + W\beta \geq V\}$, and

$$\begin{pmatrix} U \\ V \\ W^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_{w^*} \end{pmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{uw^*} \\ \sigma_{uv} & 1 & \sigma_{vw^*} \\ \sigma_{uw^*} & \sigma_{vw*} & \sigma_{w^*}^2 \end{bmatrix}\right) \tag{53}$$

Then

$$\begin{aligned}
E[U|V=v,W^*=w] &= E(U) + \begin{bmatrix} \sigma_{uv} & \sigma_{uw^*} \end{bmatrix} \begin{bmatrix} 1 & \sigma_{vw^*} \\ \sigma_{vw^*} & \sigma_{w^*}^2 \end{bmatrix}^{-1} \begin{pmatrix} v \\ (w-\mu_{w^*}) \end{pmatrix} \\
&= \frac{1}{\sigma_{w^*}^2 - \sigma_{vw^*}^2} \begin{bmatrix} \sigma_{uv} & \sigma_{uw^*} \end{bmatrix} \begin{bmatrix} \sigma_{w^*}^2 & -\sigma_{vw^*} \\ -\sigma_{vw^*} & 1 \end{bmatrix} \begin{pmatrix} v \\ (w-\mu_{w^*}) \end{pmatrix} \\
&= \frac{1}{\sigma_{w^*}^2 - \sigma_{vw^*}^2} \begin{bmatrix} \sigma_{uv} & \sigma_{uw^*} \end{bmatrix} \begin{bmatrix} \sigma_{w^*}^2 v - \sigma_{vw^*}(w-\mu_{w^*}) \\ -\sigma_{vw^*} v + (w-\mu_{w^*}) \end{bmatrix} \\
&= \frac{1}{\sigma_{w^*}^2 - \sigma_{vw^*}^2} \left[\sigma_{uv}\sigma_{w^*}^2 v - \sigma_{uv}\sigma_{vw^*}(w-\mu_{w^*}) - \sigma_{uw^*}\sigma_{vw^*} v + \sigma_{uw^*}(w-\mu_{w^*})\right] \\
&= \frac{1}{\sigma_{w^*}^2 - \sigma_{vw^*}^2} \left[(\sigma_{uv}\sigma_{w^*}^2 - \sigma_{uw^*}\sigma_{vw^*})v + (\sigma_{uw^*} - \sigma_{uv}\sigma_{vw^*})(w-\mu_{w^*})\right] \\
&= av + b(w-\mu_{w^*}),
\end{aligned}$$

where $a = \frac{\sigma_{uv}\sigma_{w^*}^2 - \sigma_{uw^*}\sigma_{vw^*}}{\sigma_{w^*}^2 - \sigma_{vw^*}^2}$, and $b = \frac{\sigma_{uw^*} - \sigma_{uv}\sigma_{vw^*}}{\sigma_{w^*}^2 - \sigma_{vw^*}^2}$.

$$\begin{aligned}
Var[U|V=v,W^*=w] &= \sigma_u^2 - \begin{bmatrix} \sigma_{uv} & \sigma_{uw^*} \end{bmatrix} \begin{bmatrix} 1 & \sigma_{vw^*} \\ \sigma_{vw^*} & \sigma_{w^*}^2 \end{bmatrix}^{-1} \begin{pmatrix} \sigma_{uv} \\ \sigma_{uw^*} \end{pmatrix} \\
&= \sigma_u^2 - \frac{1}{\sigma_{w^*}^2 - \sigma_{vw^*}^2} \begin{bmatrix} \sigma_{uv} & \sigma_{uw^*} \end{bmatrix} \begin{bmatrix} \sigma_{w^*}^2\sigma_{uv} - \sigma_{vw^*}\sigma_{uw^*} \\ -\sigma_{vw^*}\sigma_{uv} + \sigma_{uw^*} \end{bmatrix}
\end{aligned}$$

Then for $x > 0$

$$\begin{aligned}
E[U|X=1,W^*=w] &= \frac{\int_{-\infty}^{\alpha+\beta w}\int_{-\infty}^{\infty} u f_{UV|W^*}(u,v|w)\mathrm{d}u\mathrm{d}v}{P(X=1|W^*=w)} \\
&= \frac{\int_{-\infty}^{\alpha+\beta w}\int_{-\infty}^{\infty} u f_{U|V,W^*}(u|v,w)\mathrm{d}u f_{V|W^*}(v|w)\mathrm{d}v}{P(X=1|W^*=w)} \\
&= \frac{\int_{-\infty}^{\alpha+\beta w} E(U|V=v,W^*=w) f_{V|W^*}(v|w)\mathrm{d}v}{P(X=1|W^*=w)} \\
&= \frac{\int_{-\infty}^{\alpha+\beta w} (av + b(w-\mu_{w^*})) f_{V|W^*}(v|w)\mathrm{d}v}{P(X=1|W^*=w)} \\
&= a\frac{\int_{-\infty}^{\alpha+\beta w} v f_{V|W^*}(v|w)\mathrm{d}v}{P(X=1|W^*=w)} + b(w-\mu_{w^*}).
\end{aligned}$$

Now,

$$V|W^* \sim N\left(\mu_{V|W^*}(x), \sigma^2_{v|w^*}\right),$$

where $\mu_{V|W^*}(w) = \frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(w - \mu_{w^*})$ and $\sigma^2_{v|w^*} = 1 - \frac{\sigma^2_{vw^*}}{\sigma^2_{w^*}}$. Then

$$\begin{aligned}
a\frac{\int_{-\infty}^{\alpha+\beta w} v f_{V|W^*}(v|w)\mathrm{d}v}{P(X=1|W^*=w)} &= a\frac{\int_{-\infty}^{\alpha+\beta w}(v-\mu_{V|W^*}(w))f_{V|W^*}(v|w)\mathrm{d}v}{P(X=1|W^*=w)} + a\mu_{V|W^*}(w) \\
&= a\sigma_{v|w^*}\frac{\int_{-\infty}^{\alpha+\beta w}\frac{v-\mu_{V|W^*}(w)}{\sigma_{v|w^*}}f_{V|W^*}(v|w)\mathrm{d}v}{P(X=1|W^*=w)} + a\mu_{V|W^*}(w) \\
&= -a\sigma_{v|w^*}\frac{\phi\left(\frac{\alpha+\beta w-\mu_{V|W^*}(w)}{\sigma_{v|w^*}}\right)}{\Phi\left(\frac{\alpha+\beta w-\mu_{V|W^*}(w)}{\sigma_{v|w^*}}\right)} + a\mu_{V|W^*}(w),
\end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote the standard normal density and distribution functions, respectively. Thus,

$$E[U|X=1,W^*=w] = -a\sigma_{v|w^*}\frac{\phi\left(\frac{\alpha+\beta w-\mu_{V|W^*}(w)}{\sigma_{v|w^*}}\right)}{\Phi\left(\frac{\alpha+\beta w-\mu_{V|W^*}(w)}{\sigma_{v|w^*}}\right)} + a\mu_{V|W^*}(w) + b(w-\mu_{w^*}).$$

Similarly,

$$E[U|X=0,W^*=w] = a\sigma_{v|w^*}\frac{\phi\left(\frac{\alpha+\beta w-\mu_{V|W^*}(w)}{\sigma_{v|w^*}}\right)}{1-\Phi\left(\frac{\alpha+\beta w-\mu_{V|W^*}(w)}{\sigma_{v|w^*}}\right)} + a\mu_{V|W^*}(w) + b(w-\mu_{w^*}).$$

Note that if $\sigma_{uw^*} = \sigma_{uv} = 0$, then we have $a = b = 0$.

Next, consider

$$\begin{aligned}
\lim_{x\downarrow 0} E[U|X=1,W=w] &= \lim_{w\downarrow 0} E[U|X=1,X^*=w] \\
&= -a\sigma_{v|w^*}\frac{\phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma^2_{w^*}}\mu_{w^*}}{\sqrt{1-\frac{\sigma^2_{vw^*}}{\sigma^2_{w^*}}}}\right)}{\Phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma^2_{w^*}}\mu_{w^*}}{\sqrt{1-\frac{\sigma^2_{vw^*}}{\sigma^2_{w^*}}}}\right)} - a\frac{\sigma_{vw^*}}{\sigma^2_{w^*}}\mu_{w^*} - \frac{\sigma_{uw^*}-\sigma_{uv}\sigma_{vw^*}}{\sigma^2_{w^*}-\sigma^2_{vw^*}}\mu_{w^*}.
\end{aligned}$$

Now

$$\begin{aligned}
a\frac{\sigma_{vw^*}}{\sigma^2_{w^*}} + b = a\frac{\sigma_{vw^*}}{\sigma^2_{w^*}} + \frac{\sigma_{uw^*}-\sigma_{uv}\sigma_{vw^*}}{\sigma^2_{w^*}-\sigma^2_{vw^*}} &= \frac{\sigma_{uv}\sigma^2_{w^*}-\sigma_{uw^*}\sigma_{vw^*}}{\sigma^2_{w^*}-\sigma^2_{vw^*}}\frac{\sigma_{vw^*}}{\sigma^2_{w^*}} + \frac{\sigma_{uw^*}-\sigma_{uv}\sigma_{vw^*}}{\sigma^2_{w^*}-\sigma^2_{vw^*}} \\
&= \frac{1}{\sigma^2_{w^*}(\sigma^2_{w^*}-\sigma^2_{vw^*})}[\sigma_{uv}\sigma^2_{w^*}\sigma_{vw^*} - \sigma_{uw^*}\sigma^2_{vw^*} + \sigma_{uw^*}\sigma^2_{w^*} - \sigma_{uv}\sigma^2_{w^*}\sigma_{vw^*}] \\
&= \frac{\sigma_{uw^*}}{\sigma^2_{w^*}}.
\end{aligned}$$

So we have

$$\lim_{w\downarrow 0} E[U|X=1,W=w] = -a\sqrt{1-\frac{\sigma_{vw^*}^2}{\sigma_{w^*}^2}}\frac{\phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma_{w^*}^2}\mu_{w^*}}{\sqrt{1-\frac{\sigma_{vw^*}^2}{\sigma_{w^*}^2}}}\right)}{\Phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma_{w^*}^2}\mu_{w^*}}{\sqrt{1-\frac{\sigma_{vw^*}^2}{\sigma_{w^*}^2}}}\right)} - \frac{\sigma_{uw^*}}{\sigma_{w^*}^2}\mu_{w^*}.$$

Similarly, we have

$$\lim_{x\downarrow 0} E[U|X=0,W=w] = a\sqrt{1-\frac{\sigma_{vw^*}^2}{\sigma_{w^*}^2}}\frac{\phi\left(\frac{\alpha+\frac{\sigma_{vw^*}}{\sigma_{w^*}^2}\mu_{w^*}}{\sqrt{1-\frac{\sigma_{vw^*}^2}{\sigma_{w^*}^2}}}\right)}{1-\Phi\left(\frac{\alpha+\frac{\sigma_{vw^*}}{\sigma_{w^*}^2}\mu_{w^*}}{\sqrt{1-\frac{\sigma_{vw^*}^2}{\sigma_{w^*}^2}}}\right)} - \frac{\sigma_{uw^*}}{\sigma_{w^*}^2}\mu_{w^*}.$$

Next, we study $E[U|X=1,W=0]$ and $E[U|X=0,W=0]$.

$$\begin{aligned}
E[U|X=1,W=0] &= E[U|V\le\alpha,W^*\le 0] \\
&= \frac{\int_{-\infty}^{0}\int_{-\infty}^{\alpha}\int_{-\infty}^{\infty} u f_{UVW^*}(u,v,t)\mathrm{d}u\mathrm{d}v\mathrm{d}t}{P(V\le\alpha,W^*\le 0)} \\
&= \frac{\int_{-\infty}^{0}\int_{-\infty}^{\alpha}[av+b(t-\mu_{W^*})]f_{VW^*}(v,t)\mathrm{d}v\mathrm{d}t}{P(V\le\alpha,W^*\le 0)} \\
&= \frac{\int_{-\infty}^{0}\int_{-\infty}^{\alpha} a(v-\mu_{V|W^*}(t))f_{V|W^*}(v|t)\mathrm{d}v f_{W^*}(t)\mathrm{d}t}{P(V\le\alpha,W^*\le 0)} \\
&+ \frac{\int_{-\infty}^{0}\int_{-\infty}^{\alpha}[a\mu_{V|W^*}(t)+b(t-\mu_{W^*})]f_{VW^*}(v,t)\mathrm{d}v\mathrm{d}t}{P(V\le\alpha,W^*\le 0)}.
\end{aligned}$$

Since $\mu_{V|W^*}(t) = \frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t - \mu_{w^*})$, this becomes

$$
\begin{aligned}
&= -a\sigma_{V|W^*} \frac{\int_{-\infty}^{0} \phi\left(\frac{\alpha - \mu_{V|W^*}(t)}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t}{\int_{-\infty}^{0} \Phi\left(\frac{\alpha - \mu_{V|W^*}(t)}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t} \\
&+ \frac{\int_{-\infty}^{0}\int_{-\infty}^{\alpha} [a\frac{\sigma_{VW^*}}{\sigma^2_{W^*}} + b](t - \mu_{W^*}) f_{VW^*}(v,t)\mathrm{d}v\mathrm{d}t}{\int_{-\infty}^{0} \Phi\left(\frac{\alpha - \mu_{V|W^*}(t)}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t} \\
&= -a\sigma_{V|W^*} \frac{\int_{-\infty}^{0} \phi\left(\frac{\alpha - \mu_{V|W^*}(t)}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t}{\int_{-\infty}^{0} \Phi\left(\frac{\alpha - \mu_{V|W^*}(t)}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t} \\
&+ \frac{\sigma_{UW^*}}{\sigma^2_{W^*}} \frac{\int_{-\infty}^{0} (t - \mu_{W^*}) \Phi\left(\frac{\alpha - \mu_{V|W^*}(t)}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}v\mathrm{d}t}{\int_{-\infty}^{0} \Phi\left(\frac{\alpha - \mu_{V|W^*}(t)}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t},
\end{aligned}
$$

where we use

$$a\frac{\sigma_{VW^*}}{\sigma^2_{W^*}} + b = \frac{\sigma_{UW^*}}{\sigma^2_{W^*}}.$$

Plugging in for $\mu_{v|w^*}(t) = \frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t - \mu_{w^*})$ yields

$$
\begin{aligned}
E[U|X = 1, W = 0] &= E[U|V \leq \alpha, W^* \leq 0] \\
&= -a\sigma_{V|W^*} \frac{\int_{-\infty}^{0} \phi\left(\frac{\alpha - \frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t - \mu_{w^*})}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t}{\int_{-\infty}^{0} \Phi\left(\frac{\alpha - \frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t - \mu_{w^*})}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t} \\
&+ \frac{\sigma_{UW^*}}{\sigma^2_{W^*}} \frac{\int_{-\infty}^{0} (t - \mu_{W^*}) \Phi\left(\frac{\alpha - \frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t - \mu_{w^*})}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}v\mathrm{d}t}{\int_{-\infty}^{0} \Phi\left(\frac{\alpha - \frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t - \mu_{w^*})}{\sigma_{V|W^*}}\right) f_{W^*}(t)\mathrm{d}t},
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
E[U|X=0,W=0] &= E[U|V>\alpha, W^*\le 0] \\
&= a\sigma_{V|W^*}\frac{\int_{-\infty}^{0}\phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t-\mu_{w^*})}{\sigma_{V|W^*}}\right)f_{W^*}(t)\mathrm{d}t}{\int_{-\infty}^{0}\left[1-\Phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t-\mu_{w^*})}{\sigma_{V|W^*}}\right)\right]f_{W^*}(t)\mathrm{d}t} \\
&+ \frac{\sigma_{UW^*}}{\sigma^2_{W^*}}\frac{\int_{-\infty}^{0}(t-\mu_{W^*})\left[1-\Phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t-\mu_{w^*})}{\sigma_{V|W^*}}\right)\right]f_{W^*}(t)\mathrm{d}v\mathrm{d}t}{\int_{-\infty}^{0}\left[1-\Phi\left(\frac{\alpha-\frac{\sigma_{vw^*}}{\sigma^2_{w^*}}(t-\mu_{w^*})}{\sigma_{V|W^*}}\right)\right]f_{W^*}(t)\mathrm{d}t},
\end{aligned}
$$

1. As mentioned above, when $\sigma_{UW^*}=\sigma_{UV}=0$, then $a=b=0$, so that

$$
\begin{aligned}
\lim_{w\downarrow 0}E[U|X=1,W=w] = E[U|X=1,W=0] &= 0, \\
\lim_{w\downarrow 0}E[U|X=0,W=w] = E[U|X=0,W=0] &= 0.
\end{aligned}
$$

2. If $\sigma_{UW^*}=\sigma_{VW^*}=0$, then $b=0$, but $a=\sigma_{UV}$. In addition, $\mu_{V|W^*}(t)=\mu_V=0$ for each t and $\sigma_{V|W^*}=\sigma_V=1$. As a result, we have

$$
\begin{aligned}
\lim_{w\downarrow 0}E[U|X=1,W=w] = E[U|X=1,W=0] &= -\sigma_{UV}\frac{\phi(\alpha)}{\Phi(\alpha)}, \\
\lim_{w\downarrow 0}E[U|X=0,W=w] = E[U|X=0,W=0] &= \sigma_{UV}\frac{\phi(\alpha)}{1-\Phi(\alpha)}.
\end{aligned}
$$

   Thus, we have no power in this case.

3. If $\sigma_{UW^*}=0$, but $\sigma_{UV}\neq 0$ and $\sigma_{VW^*}\neq 0$, we have

$$
\begin{aligned}
\lim_{w\downarrow 0}E[U|X=1,W=w] &= -a\sigma_{V|W^*}\frac{\phi\left(\frac{\alpha+\mu_{W^*}\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)}{\Phi\left(\frac{\alpha+\mu_{W^*}\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)}, \\
E[U|X=1,W=0] &= -a\sigma_{V|W^*}\frac{\int_{-\infty}^{0}\phi\left(\frac{\alpha+(t-\mu_{W^*})\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)f_{W^*}(t)\mathrm{d}t}{\int_{-\infty}^{0}\Phi\left(\frac{\alpha+(t-\mu_{W^*})\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)f_{W^*}(t)\mathrm{d}t}, \\
\lim_{w\downarrow 0}E[U|X=0,W=w] &= a\sigma_{V|W^*}\frac{\phi\left(\frac{\alpha+\mu_{W^*}\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)}{1-\Phi\left(\frac{\alpha+\mu_{W^*}\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)}, \\
E[U|X=0,W=0] &= a\sigma_{V|W^*}\frac{\int_{-\infty}^{0}\phi\left(\frac{\alpha+(t-\mu_{W^*})\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)f_{W^*}(t)\mathrm{d}t}{\int_{-\infty}^{0}\left[1-\Phi\left(\frac{\alpha+(t-\mu_{W^*})\sigma_{VW^*}/\sigma^2_{W^*}}{\sigma_{V|W^*}}\right)\right]f_{W^*}(t)\mathrm{d}t}.
\end{aligned}
$$

   Moreover, in this case $a=\frac{\sigma_{UV}\sigma^2_{W^*}}{\sigma^2_{W^*}-\sigma^2_{VW^*}}\neq 0$. It seems that in this case we have power.

4. Suppose $\sigma_{VW^*} = 0$, but $\sigma_{UV} \neq 0$ and $\sigma_{UW^*} \neq 0$. In this case, we have

$$\begin{aligned}
\lim_{w\downarrow 0} E[U|X=1,W=w] &= -\sigma_{UV}\tfrac{\phi(\alpha)}{\Phi(\alpha)} - \tfrac{\sigma_{UW^*}}{\sigma^2_{W^*}}\mu_{W^*},\\
E[U|X=1,W=0] &= -\sigma_{UV}\tfrac{\phi(\alpha)}{\Phi(\alpha)} - \tfrac{\sigma_{UW^*}}{\sigma_{W^*}}\tfrac{\phi(-\mu_{W^*}/\sigma_{W^*})}{\Phi(-\mu_{W^*}/\sigma_{W^*})},\\
\lim_{w\downarrow 0} E[U|X=0,W=w] &= \sigma_{UV}\tfrac{\phi(\alpha)}{1-\Phi(\alpha)} - \tfrac{\sigma_{UW^*}}{\sigma^2_{W^*}}\mu_{W^*},\\
E[U|X=0,W=0] &= \sigma_{UV}\tfrac{\phi(\alpha)}{1-\Phi(\alpha)} + \tfrac{\sigma_{UW^*}}{\sigma_{W^*}}\tfrac{\phi(-\mu_{W^*}/\sigma_{W^*})}{1-\Phi(-\mu_{W^*}/\sigma_{W^*})}.
\end{aligned}$$

Again, we seem to have power in this case.

B.1.2 Continuous X:

Suppose the data generating process is given by

$$\begin{aligned}
Y &= \alpha + X\beta + \theta W + U, &(54)\\
W &= \max\{0, \gamma + \delta X + \eta\} &(55)
\end{aligned}$$

and

$$\begin{pmatrix} U \\ \eta \\ X \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_x \end{pmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{u\eta} & \sigma_{ux} \\ \sigma_{u\eta} & \sigma_\eta^2 & \sigma_{\eta x} \\ \sigma_{ux} & \sigma_{\eta x} & \sigma_x^2 \end{bmatrix}\right). \tag{56}$$

Then

$$E(U|X=x,\eta=e) = \pi_x(X-\mu_x) + \pi_\eta e,$$

with $\pi_x = \frac{\sigma_{ux}\sigma_\eta^2 - \sigma_{u\eta}\sigma_{\eta x}}{\sigma_x^2\sigma_\eta^2 - \sigma_{\eta x}^2}$ and $\pi_\eta = \frac{\sigma_{u\eta}\sigma_x^2 - \sigma_{ux}\sigma_{\eta x}}{\sigma_x^2\sigma_\eta^2 - \sigma_{\eta x}^2}$. Therefore,

$$\lim_{w\downarrow 0} E(U|X=x,W=w) = \lim_{w\downarrow 0} E(U|X=x,\eta=w-\gamma-\delta x) = \pi_x(x-\mu_x) + \pi_\eta(-\gamma-\delta x)$$

and

$$\begin{aligned}
E(U|X=x,W=0) = E(U|X=x,\eta\leq -\gamma-\delta x) &= \int_{-\infty}^{-\gamma-\delta x} \frac{\pi_x(x-\mu_x)+\pi_\eta e}{F_{\eta|X}(-\gamma-\delta x|x)} \mathrm{d}F_{\eta|X}(e|x)\\
&= \pi_x(x-\mu_x) + \pi_\eta \frac{1}{F_{\eta|X}(-\gamma-\delta x|x)} \int_{-\infty}^{-\gamma-\delta x} e\mathrm{d}F_{\eta|X}(e|x)\\
&= \pi_x(x-\mu_x)\\
&+ \pi_\eta \frac{1}{\Phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho_{x\eta}^2}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho_{x\eta}^2}}x\right)} \int_{-\infty}^{-\gamma-\delta x} \frac{(e-\mu_{\eta|x})}{\sqrt{2\pi\sigma_{\eta|x}^2}} e^{-\frac{(e-\mu_{\eta|x})^2}{2\sigma_{\eta|x}^2}} \mathrm{d}e\\
&+ \pi_\eta \mu_{\eta|x}.
\end{aligned}$$

Using the facts that $\mu_{\eta|x} = \frac{\sigma_{x\eta}}{\sigma_x^2}$ and $\sigma^2_{\eta|x} = \sigma^2_\eta(1-\rho^2_{x\eta})$ with $\rho_{x\eta} = \frac{\sigma_{x\eta}}{\sigma_x\sigma_\eta}$. we can write the last term above as

$$\begin{aligned}
E(U|X=x,W=0) =& \pi_x(x-\mu_x) \\
&-\pi_\eta\sqrt{\sigma^2_\eta(1-\rho^2_{x\eta})}\frac{\phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}}x\right)}{\Phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}}x\right)} \\
&+\pi_\eta\frac{\sigma_{x\eta}}{\sigma_x^2}(x-\mu_x).
\end{aligned}$$

Then

$$\begin{aligned}
E(U|W,X) =& \pi_x(X-\mu_x) + \pi_\eta(W-\gamma-\delta X)1\{W>0\} \\
&+\left[\pi_\eta\frac{\sigma_{x\eta}}{\sigma_x^2}(X-\mu_x) - \pi_\eta\sqrt{\sigma^2_\eta(1-\rho^2_{x\eta})}\frac{\phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}}X\right)}{\Phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}}X\right)}\right]1\{W=0\},
\end{aligned} \tag{57}$$

and

$$\lim_{w\downarrow 0} E(U|X=x,W=w) - E(U|X=x,W=0) \tag{58}$$

$$\begin{aligned}
=& -\pi_\eta\frac{\sigma_{\eta x}}{\sigma_x^2}(x-\mu_x) + \pi_\eta(-\gamma-\delta x) + \pi_\eta\sqrt{\sigma^2_\eta(1-\rho^2_{x\eta})}\frac{\phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}}x\right)}{\Phi\left(\frac{-\gamma+\frac{\sigma_{\eta x}}{\sigma_x^2}\mu_x}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}} - \frac{\frac{\sigma_{\eta x}}{\sigma_x^2}+\delta}{\sigma_\eta\sqrt{1-\rho^2_{x\eta}}}x\right)}
\end{aligned} \tag{59}$$

Things this example illustrates:

1. If $\theta = 0$ and $\sigma_{ux} = 0$, then running an OLS on X only would consistently estimate β. Moreover, conditioning on W would introduce endogeneity. In particular, our test statistic would be different from 0 in the following cases:

   (i) $\sigma_{u\eta} \neq 0$ (which implies $\pi_\eta \neq 0$) and $\delta \neq 0$. In this case, regardless of whether $\sigma_{\eta x} = 0$ or not the second and third terms in 59 will depend on x, our test statistic will be different from 0.

   (ii) $\sigma_{u\eta} \neq 0$ and $\sigma_{\eta x} \neq 0$. Again, π_v will be different from 0 and the third term in 59 will depend on x, and our test statistic will be different from 0. This is true even if $\delta = 0$.

   On the other hand, if θ in equation (2) is really different from 0, then running an OLS regression of Y on X, without controlling for W would not yield a consistent of β because of omitted variable bias.

2. If $\theta \neq 0$, but $\pi_\eta = 0$ and $\sigma_{u\eta} \neq 0$, $\sigma_{\eta x} \neq 0$. Note $[\pi_\eta = 0, \sigma_{u\eta} \neq 0, \sigma_{\eta x} \neq 0] \implies \sigma_{ux} \neq 0$. In this case, π_x is not necessarily 0. If $\pi_x \neq 0$, but $\pi_\eta = 0$, then $E(U|W = w, X = x) = \pi_x(x - \mu_x)$. Even though X is still endogenous after controlling for W, our test will not detect this, and as such will have no power. This case, however, seems to be a non-generic case.

3. If $\theta \neq 0$, but $\pi_\eta = 0$, and $\sigma_{u\eta} = 0$, $\sigma_{ux} \neq 0$. This means $\sigma_{\eta x} = 0$, and $\pi_x = \frac{\sigma_{ux}}{\sigma_x^2}$. As in the previous case, X is still endogenous after controlling for W, but our test will not detect this, and as such will have no power. Moreover, this is not a knife edge case. In this and the case W actually drops out from $E(U|W = w, X = x)$. In general, however, one might imagine situations in which $E(U|W = w, X = x)$ is of the form $\psi_1(X) + \psi_2(W)$. In those cases, our test statistic will be 0, even though X is still endogenous after controlling for W.

4. If $\delta = \sigma_{\eta x} = 0$, then

$$E(U|W, X) = \pi_x(X - \mu_x) + \pi_\eta(W - \gamma)1\{W > 0\} - \pi_\eta \sqrt{\sigma_\eta^2(1 - \rho_{x\eta}^2)} \frac{\phi\left(\frac{-\gamma}{\sqrt{\sigma_\eta^2(1-\rho_{x\eta}^2)}}\right)}{\Phi\left(\frac{-\gamma}{\sqrt{\sigma_\eta^2(1-\rho_{x\eta}^2)}}\right)} 1\{W = 0\}.$$

   $E(U|W, X)$ is additively separable in X and W, and our test statistic will be 0, even though X is still endogenous after controlling for W.

5. One might think that π_x is the parameter that determines whether X is endogenous after controlling for W, but this would be wrong. In particular, an anonymous referee claimed that "We reject the test when $\delta \neq 0$ even if X is exogenous ($\pi_x = 0$)." As equation 57 indicates $E(U|W, X)$ depends on X through multiple terms, not just through $\pi_x(x - \mu_x)$. Even when both $\pi_x = 0$ and $\delta = 0$, $E(U|W, X)$ depends on X as long as $\pi_\eta \neq 0$ and $\sigma_{x\eta} \neq 0$, and our test correctly detects endogeneity of X after controlling for W.

6. The same referee also claimed that "Conversely, we accept the null hypothesis when $\delta = 0$ even though X is endogenous ($\pi_x \neq 0$)." This is also false. Even if $\delta = 0$ our test statistic will be different from 0 as long as π_η and $\sigma_{x\eta}$ are both different from 0.

7. The referee also says "More generally, we can expect the test to have low power when the correlation between X and W becomes small." "... by introducing covariates Z, the partial correlation between X and W decreases, and the power of the test decreases. But endogeneity of X could still be strong." This claim is vague. To evaluate this claim suppose that Y and W are still as in equations 2 and 3, and

$$\begin{pmatrix} U \\ \eta \\ X \end{pmatrix} | Z \sim N\left(\begin{pmatrix} \mu_{u|z} \\ \mu_{\eta|z} \\ \mu_{x|z} \end{pmatrix}, \begin{bmatrix} \sigma_{u|z}^2 & \sigma_{u\eta|z} & \sigma_{ux|z} \\ \sigma_{u\eta|z} & \sigma_{\eta|z}^2 & \sigma_{\eta x|z} \\ \sigma_{ux|z} & \sigma_{\eta x|z} & \sigma_{x|z}^2 \end{bmatrix} \right). \tag{60}$$

In this case,

$$E(U|X,W,Z) = \pi_{x|z}(X-\mu_{x|z}) + \pi_{\eta|z}(W-\mu_{\eta|z}-\gamma-\delta X)1\{W>0\} + 1\{W=0\}$$
$$\times \left[\pi_{\eta|z}\frac{\sigma_{x\eta|z}}{\sigma^2_{x|z}}(X-\mu_{x|z}) - \pi_{\eta|z}\sqrt{\sigma^2_{\eta|z}(1-\rho^2_{x\eta|z})}\frac{\phi\left(\frac{-\mu_{\eta|z}-\gamma+\frac{\sigma_{\eta x|z}}{\sigma^2_{x|z}}\mu_{x|z}}{\sigma_{\eta|z}\sqrt{1-\rho^2_{x\eta|z}}} - \frac{\frac{\sigma_{\eta x|z}}{\sigma^2_{x|z}}+\delta}{\sigma_{\eta|z}\sqrt{1-\rho^2_{x\eta|z}}}X\right)}{\phi\left(\frac{-\mu_{\eta|z}-\gamma+\frac{\sigma_{\eta x|z}}{\sigma^2_{x|z}}\mu_{x|z}}{\sigma_{\eta|z}\sqrt{1-\rho^2_{x\eta|z}}} - \frac{\frac{\sigma_{\eta x|z}}{\sigma^2_{x|z}}+\delta}{\sigma_{\eta|z}\sqrt{1-\rho^2_{x\eta|z}}}X\right)} \right], \tag{61}$$

where $\pi_{x|z} = \frac{\sigma_{ux|z}\sigma^2_{\eta|z} - \sigma_{u\eta|z}\sigma_{\eta x|z}}{\sigma^2_{x|z}\sigma^2_{\eta|z} - \sigma^2_{\eta x|z}}$ and $\pi_{\eta|z} = \frac{\sigma_{u\eta|z}\sigma^2_{x|z} - \sigma_{ux|z}\sigma_{\eta x|z}}{\sigma^2_{x|z}\sigma^2_{\eta|z} - \sigma^2_{\eta x|z}}$.

In this case, for our test to lack power both $\pi_{v|z}$ and $\sigma_{\eta x|z}$ must be 0, but $\pi_{x|z} \neq 0$. $\pi_{\eta|z}$ and $\sigma_{\eta x|z}$ will both be 0 and well defined only if both $\sigma_{\eta x|z} = 0$ and $\sigma_{u\eta|z} = 0$. For $\pi_{x|z} \neq 0$ we must, at the same time, have that $\sigma_{ux|z} \neq 0$.