Diagnostic Tests for the Selection on Observables Assumption: The Case of Women Infants and Children Program
** PRELIMINARY AND INCOMPLETE **

Umair Khalil* Neşe Yıldız†

Abstract

In this paper we present diagnostic tests for selection on observables assumption. Our diagnostic tests are based on the presence of a variable among our controls that has an atom (bunching) at a known point, but is otherwise continuously distributed. This set up was first exploited by [4] to test exogeneity of smoking during pregnancy in considering the effect of smoking during pregnancy on baby's birth weight. Caetano's test, however, cannot be applied to test exogeneity of a discrete covariate, which is the focus of this paper. We demonstrate our diagnostic tests using birth certificate data, which covers more than 80% of all births in the U.S. from 2010 - 2012 and contains a rich, detailed set of parental and pregnancy covariates, to study the impact of the Supplemental Program for Women Infants and Children (WIC) on birth outcomes.

1 Introduction

Estimation of treatment effects is an important problem in economics. In the absence of experimental data or valid instruments, researchers rely on the selection on observables assumption for identifying treatment effects. Without additional structure, this assumption is difficult to test, however. In this paper we present diagnostic procedures for checking the validity of the selection on observables assumption. The procedures we present rely on the presence of a variable among our controls that has an atom (bunching) at a known point, but is otherwise continuously distributed. This set up was first exploited by [4]. The test introduced in [4] cannot be used to test the selection on observables assumption when treatment is discrete, which is the focus of our paper.

To explain how presence of a variable, say X, among the controls that has an atom (bunching) at a known point, but is otherwise continuously distributed, helps getting testable

*Department of Economics, University of Rochester, Harkness Hall, Rochester, NY, 14627; Email: khalil.umair@gmail.com.

†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.

implications of the selection on observables assumption, consider the problem of assessing the impact of WIC participation on baby's birth weight. While the goals of WIC are clear, assessing the true effect of WIC participation is difficult since participation into the program may not be random. 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 important. Our outcome variable of interest will be baby's birth weight. This is a natural variable to focus on since low birth weight and other complications at birth are in turn linked to significant health costs especially during infancy and early childhood. (See for example [2].)

Suppose that birth weight is a function of smoking during pregnancy (average number of daily cigarettes), WIC participation, other controls like mother's and pregnancy's characteristics plus some unobservable, U. Like [4], we maintain the assumption that this function (the structural function) is continuous in the smoking variable, X. Around 80% of mothers do not smoke in our sample. [1] In addition, since it is possible to smoke part of a cigarette, for positive amounts smoked, smoking is a continuous variable. Thus, average cigarettes smoked daily does have the structure we need. In addition, 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. Since the structural function is assumed to be continuous in this variable, this means that 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 conditional on smoking and other controls, then this discontinuity must be the same for WIC participants and non-participants. If the birth weight equation is additively separable in WIC participation and the unobservables, then this implies that the discontinuity in expected birth weight conditional on smoking, other controls and WIC participation 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 the case. These arguments work even if treatment was not a discrete variable. To the best of our knowledge, however, ours is the first paper to present a test of selection on observables assumption in the context of a binary treatment. The test presented in [4] is also a test of selection on observables assumption, but in her case, it is the treatment variable itself that has bunching at a known point, but is otherwise continuously distributed.

We provide a formal test for selection on observables assumption for models in which the outcome equation is additively separable in treatment and unobservables. In particular, we provide a test statistic whose population value must be 0 if the model is of this form and the selection on observables assumption holds. For this testing procedure to have power, we need a control variable, X with the structure described above that either depends on the unobservables in the outcome equation conditional on other covariates or depends on the unobservables determining treatment conditional on other covariates in neighborhood of the bunching point of the variable X when the selection on observables assumption fails. If expected outcome conditional on treatment, other covariates and X is continuous in x at the bunching point of X, then our test would have no power. On the other hand, this

[1] In our data set the exact percentage of non-smokers 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%.

testing procedure is a joint test of additive separability of outcome equation in treatment and unobservables as well as the selection on observables assumption. As discussed in Section 5, when the outcome equation is not additively separable in treatment and unobservables, our test statistic could be non-zero even if selection on observables assumption does not hold. If along with selection on observables assumption one uses OLS to estimate the treatment effects, then finding that the our statistic is significantly different from 0 means that these estimators are inconsistent and do not estimate the true treatment effect. If, however, one uses propensity score matching or inverse propensity score weighting, then finding a non-zero value for our test statistic does not necessarily invalidates the estimated treatment effects. For this reason we also discuss a diagnostic procedure for selection on observabels assumption for non-additively separable models.

As mentioned above, the testable implications of the selection on observables assumption we provide hinge crucially on having a covariate/control X that is continuously distributed except for having an atom at a known point. Examples of such variables were discussed in [4]. We use one of these, namely smoking during pregnancy, in our empirical application. Our testing procedure can also be used to estimate the effect of choosing a STEM (science, technology, engineering, and mathematics) major in college on future wages, using SAT scores as the X variable.[2]

The paper is organized as follows. In Section 2 we introduce the formal model and the main Theorem. Section 3 discusses the implementation of the basic testing idea. In particular, we provide our formal test statistic and derive its asymptotic distribution in this section. Section 4 discusses our empirical application in detail. Section 5 provides some discussion of our test and presents a partial diagnostic procedure to check the validity of the selection on observables assumption. Section 6 concludes. The Appendix provides the proofs.

2 The Model and the test statistic:

The model we study is given by

$$Y = g(X, D, Z) + U, \tag{1}$$

where Y is the outcome variable (it will denote baby's birthweight in our application), D is the binary treatment variable (in our example it denoted participation in WIC) and X is a scalar control, and Z is a vector of other observed variables, and U represents an unobservable variable. To keep the exposition simple, in this section we assume there are no other controls, Z. Later, we discuss how additional controls could be incorporated. Then the potential outcomes, Y_1 and Y_0, are

$$Y_1 = m(X, 1, Z) + U, \tag{2}$$
$$Y_0 = m(X, 0, Z) + U. \tag{3}$$

[2]We expect that quantitative SAT scores may have a bunching point at 800, which is the maximum attainable score.

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

$$U \perp\!\!\!\perp D|X, Z.$$

Our goal is to assess the validity of this assumption. We will assume that m is continuous in x for each z and d, and the random variable X has an atom/ positive mass at a known point, and is continuously distributed in a neighborhood of this point. The point at which X has positive mass is normalized to be 0.

Assumption 2.1. 1. m is continuous in x for each z and d.

2. There exists $\delta > 0$ such that (i) $0 < P(X = 0, D = d) < P(D = d) < 1$, and (ii) $0 < P(X \in (0, \delta], D = d) < P(D = d) < 1$ for $d = 0, 1$.

3. $\mathcal{A} := \cap_{d=0}^{1} \cap_{0 \le x \le \delta} \text{Supp}(Z|X = x, D = d) \neq \emptyset$.

4. For each $z \in \mathcal{A}$ and for each $d = 0, 1$, X has a conditional Lebesgue density, $f_{X|Z,D}(x|z, d)$ that is strictly positive for $x \in (0, \delta)$.

This assumption is crucial for our test. The first part of this assumption says that the structural function linking treatment to the outcome is continuous in the variable X. This is a fairly mild assumption which we would expect to hold in many situations. For example, it is likely that the effect of smoking on birth weight is continuous. Second part of the assumption ensures that X has a positive mass at 0 for both treatment and control groups. The third part ensures that the controls other than X have common support for both WIC participants and non-participants as well as for $X = 0$ and $x \in (0, \delta)$. Our test is based on checking whether the difference in two conditional expectations is continuous at $X = 0$. This requires that X be continuously distributed near 0. The neighborhood on which X is continuously distributed could be on either side or both sides of 0. In our application, this neighborhood is to the right of 0, since smoking during pregnancy cannot be negative. On the other hand, since it is possible to smoke fractions of cigarettes, it makes sense to treat smoking (or more precisely, average number of cigarettes smoked per day) as continuously distributed to the right of 0.

To implement our test we will look at $E(Y|D = 1, X = x, Z = z)$ and $E(Y|D = 0, X = x, Z = z)$. Note that under the selection on observables assumption, that is under the assumption that U is independent of D conditional on (X, Z),

$$\begin{aligned} E(Y|D = 1, X = x, Z = z) &= m(x, 1, z) + E(U|X = x, Z = z), \quad (4) \\ E(Y|D = 0, X = x, Z = z) &= m(x, 0, z) + E(U|X = x, Z = z), \quad (5) \end{aligned}$$

so that the difference equals

$$E(Y|D = 1, X = x, Z = z) - E[Y|D = 0, X = x, Z = z] = m(x, 1, z) - m(x, 0, z).$$

Since m is assumed to be continuous in x, this difference is continuous in x as $x \downarrow 0$. Note that this difference will be continuous in x as $x \downarrow 0$ even if X is correlated with the unobservables in the outcome equation. These arguments are summarized in the result below:

Theorem 2.1. *Suppose the model given in equation 1 and Assumption 2.1 hold. In addition, suppose and*
$E(U|D,X) = E(U|X)$*, then*

$$\lambda(z) := \lambda_1(z) - \lambda_0(z) = 0 \tag{6}$$

for almost every $z \in \mathcal{A}$*, where*

$$\lambda_d(z) := \lim_{x \downarrow 0} E[Y|D=d, X=x, Z=z] - E[Y|D=d, X=0, Z=z]. \tag{7}$$

If the selection on observables assumption does not hold, we instead have

$$\begin{aligned} &E[Y|X=x, D=1, Z=z] - E[Y|X=x, D=0, Z=z] = \\ &m(x,1,z) - m(x,0,z) + E[U|X=x, D=1, Z=z] - E[U|X=x, D=1, Z=z] = \\ &m(x,1,z) - m(x,0,z) + \frac{\int_{-\infty}^{\infty}\int_{\mathcal{V}(x,z)} u \mathrm{d}F_{UV|X,Z}(u,v|x,z)}{P(x,z)} - \frac{\int_{-\infty}^{\infty}\int_{\mathcal{V}^C(x,z)} u \mathrm{d}F_{UV|X,Z}(u,v|x,z)}{1-P(x,z)}, \end{aligned} \tag{8}$$

where V denotes the unobservables in the treatment equation, $P(x,z) = E[D|X=x, Z=z]$, and $\mathcal{V}(x,z)$ denotes the set of V values associated with treatment value equal to 1 conditional on $X=x, Z=z$. If $(U,V) \perp\!\!\!\perp X|Z$, then $F_{U,V|X,Z}(u,v|x,z) = F_{U,V|Z}(u,v|z)$, and therefore, $F_{U,V|X,Z}(u,v|x,z)$ is continuous in x. If in addition, the set $\mathcal{V}(x,z)$ varies continuously in x for almost every z, as in the case $\mathcal{V}(x,z) = (-\infty, h(x,z)]$, with h continuous in x for *a.e.* z, then the last expression above will be continuous in x even if $U \not\perp\!\!\!\perp V|(X,Z)$. In that case, both $\lambda_1(z)$ and $\lambda_0(z)$, and hence, $\lambda(z)$ will be 0 for each z, and any test statistic based on λ will have no power for testing $U \not\perp\!\!\!\perp V|(X,Z)$.

While tests based on λ will have no power as long as $F_{U,V|X,Z}(u,v|x,z)$ is continuous in x at $x=0$, we have some way of checking whether this is the the reason $\lambda(z)=0$. In particular, using Caetano (2012) we can check if $P(x,z) = E(D|X=x, Z=z)$ is continuous in x or not at $x=0$. If this is discontinuous in x then we can conclude that $F_{U,V|X,Z}(u,v|x,z)$ is not continuous in x at $x=0$.

On the other hand, if V is independent of U conditional on (X,Z), even if the joint distribution of U and V conditional on (X,Z) varies discontinuously with x, this discontinuity is canceled. In contrast, if $U \not\perp\!\!\!\perp V|X,Z$ the discontinuity of the joint distribution of U and V conditional on X,Z will likely be different for treated and untreated people, and the difference between $E[Y|D=1, X=x, Z=z]$ and $E[Y|D=0, X=x, Z=z]$ is likely to be discontinuous in x at $x=0$. In particular, in a supplementary appendix we show that this indeed occurs for the special case in which

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

with

$$\begin{pmatrix} U \\ V \\ X^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_{x^*} \end{pmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{ux^*} \\ \sigma_{uv} & 1 & \sigma_{vx^*} \\ \sigma_{ux^*} & \sigma_{vx*} & \sigma_{x^*}^2 \end{bmatrix} \right). \tag{9}$$

Based on Theorem 2.1 and the discussion following it, for testing

$$H_0 : U \perp\!\!\!\perp V|X, Z, \tag{10}$$

against

$$H_1 : U \not\perp\!\!\!\perp V|X, Z, \tag{11}$$

we could use either of the following test statistics

$$t_1 := \sup_{z\in\mathcal{A}} |\lambda(z)|, \tag{12}$$

$$t_2 = \int_{z\in\mathcal{A}} \psi(\lambda(z))\omega(z)\mathrm{d}z, \tag{13}$$

where $\psi : \mathbb{R} \mapsto \mathbb{R}_+$ such that $\psi(s) = 0 \iff s = 0$, and ω is a non-negative weight function.

Since $\lambda(z) = 0$ if $E(U|D = 1, X, Z) = E(U|D = 0, X, Z)$, it does not give us any information about the dependence of distribution U on D conditional in X and Z. We can alleviate this concern by considering the characteristic function of Y, instead. Specifically, suppose, as before, the model given by equation 1 holds. Note that

$$E\left[e^{itY}|D = 1, X = x, Z = z\right] = e^{itm(x,z,1)}E[e^{itU}|D = 1, X = x, Z = z].$$

Then

$$\begin{aligned}
\kappa(z) :=& \frac{\lim_{x\downarrow 0} E\left[e^{itY}|D = 1, X = x, Z = z\right]}{E\left[e^{itY}|D = 1, X = 0, Z = z\right]} - \frac{\lim_{x\downarrow 0} E\left[e^{itY}|D = 0, X = x, Z = z\right]}{E\left[e^{itY}|D = 0, X = 0, Z = z\right]} && (14)\\
=& \frac{\lim_{x\downarrow 0} E\left[e^{itU}|D = 1, X = x, Z = z\right]}{E\left[e^{itU}|D = 1, X = 0, Z = z\right]} - \frac{\lim_{x\downarrow 0} E\left[e^{itU}|D = 0, X = x, Z = z\right]}{E\left[e^{itU}|D = 0, X = 0, Z = z\right]}\\
=& \frac{E\left[e^{itU}|D = 0, X = 0, Z = z\right]}{E\left[e^{itU}|D = 1, X = 0, Z = z\right] E\left[e^{itU}|D = 0, X = 0, Z = z\right]}\\
&\times \left(\lim_{x\downarrow 0} E\left[e^{itU}|D = 1, X = x, Z = z\right] - \lim_{x\downarrow 0} E\left[e^{itU}|D = 0, X = x, Z = z\right]\right) && (15)\\
&- \frac{\lim_{x\downarrow 0} E\left[e^{itU}|D = 0, X = x, Z = z\right]}{E\left[e^{itU}|D = 1, X = 0, Z = z\right] E\left[e^{itU}|D = 0, X = 0, Z = z\right]}\\
&\times \left(E\left[e^{itU}|D = 1, X = 0, Z = z\right] - E\left[e^{itU}|D = 0, X = 0, Z = z\right]\right). && (16)
\end{aligned}$$

Under the null hypothesis that $U \perp\!\!\!\perp D|X$, both 15 and 16, and hence $\kappa_t(z)$ must be 0 for each t and *a.e.* z.

3 Implementation of the Test:

We could exploit the result from theorem 2.1 is to estimate t_1 or t_2 and reject the null hypothesis when the realization of estimated test statistic is "too large". To estimate either test statistic we would have to estimate $\lambda_d(z)$ for $d = 0, 1$. In particular, $E[Y|D = d, X = 0, Z = z]$ can be estimated with a local linear regression of Y onto Z at z using only

observations such that $X = 0$ and $D = d$, and $\lim_{x\downarrow 0} E[Y|X = x; Z = z, D = d]$ can be estimated with a local linear regression of Y onto X and Z at $X = 0$ and $Z = z$ using only observations such that $X > 0$ and $D = d$. Under standard regularity conditions, in estimation of $\lambda_d(z)$, $\lim_{x\downarrow 0} E[Y|X = x; Z = z, D = d]$ 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 $\sqrt{nh^{d_z+1}}(\widehat{\lambda}_d(z) - \lambda_d(z)) \xrightarrow{d} N(0, V_d)$, where d_z denotes the dimension of Z.

In our application, the dimension of Z is quite large,[3] which means that fully nonparametric estimation of $\lambda_d(z)$ is not feasible in practice. For this reason, for estimation, we assume the structural function is partially linear. In particular, we assume

$$m(X, D, Z) = g(X, D) + Z^T\gamma,$$

so that

$$Y = g(X, D) + Z^T\gamma + U.$$

We also assume that Z is exogenous:

Assumption 3.1. $E(U|X, Z, D) = E(U|X, D)$.

Under this assumption, we have

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

which means that

$$Y = g(X, D) + Z^T\gamma + \rho(X, D) + \varepsilon,$$

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

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

and by Robinson (1998?), $\sqrt{n}(\hat{\gamma} - \gamma) \xrightarrow{d} N(0, V_\gamma)$. Once γ is known we can write

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

Note that for $d = 0, 1$, and for each z,

$$\lambda_d(z) = \tilde{\lambda}_d,$$

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

The advantage of the additional structure we impose here is that at each step, we have to essentially perform one dimensional non-parametric local regressions. We estimate $E(\tilde{Y}|X = 0, D = d)$ by

$$\frac{1}{n_{d0}} \sum_{i=1}^{n} \hat{\tilde{Y}}_i 1\{D_i = d, X_i = 0\},$$

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

where $n_{d0} = \sum_{i=1}^{n} 1\{D_i = d, X_i = 0\}$. For $d = 0, 1$, we estimate

$$\mu^{+}_{\tilde{Y}|X,D}(0, d) := \lim_{x \downarrow 0} E(\tilde{Y}|X = x, D = d)$$

as

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

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

Assumption 3.2. (i) The density $f_{X|D,X>0}(x, d)$ is bounded and bounded away from 0 for $x \in [0, \delta]$ and for $d = 0, 1$. It is also continuously differentiable on $(0, \delta)$ for $d = 0, 1$.

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

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

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

(v) $Var(\varepsilon_i) < \infty$, and $E(\varepsilon_i^2|X_i = x)$ is a continuous function of x for $x \in (0, \delta]$.

(vi) 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 $\int uK(u)\mathrm{d}u = 0$.

(vii) 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:

$$\begin{aligned}
p &:= P(D = 1), \\
f^{+}_{X|D}(0|d) &:= \lim_{x \downarrow 0} f_{X|D}(x|d), \\
f^{+}_{X}(0) &:= \lim_{x \downarrow 0} f_X(x), \\
\sigma^2_{+}(0) &:= \lim_{x \downarrow 0} E(\varepsilon_i^2|X_i = x), \\
\kappa_j &:= \int_0^\infty u^j K(u)\mathrm{d}u, \\
\lambda_j &:= \int_0^\infty u^j K^2(u)\mathrm{d}u, \\
C &:= p^2(1-p)^2[f^{+}_{X|D}(0|1)]^2[f^{+}_{X|D}(0|0)]^2(\kappa_0\kappa_2 - \kappa_1^2)^2, \\
M &= \left[(1-p)f^{+}_{X|D}(0|0) - pf^{+}_{X|D}(0|1)\right]^2,
\end{aligned}$$

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

Theorem 3.1. *Suppose the model in 1 holds. In addition, suppose Assumptions 2.1 3.1 and 3.2 hold. Then*

$$\sqrt{nh}\left(\hat{\tilde{\lambda}}_1 - \hat{\tilde{\lambda}}_0 - (\tilde{\lambda}_1 - \tilde{\lambda}_0)\right) \xrightarrow{d} N(0, V), \tag{17}$$

where $V = \frac{M}{C}\left[\kappa_2^2\lambda_0 - 2\kappa_2\kappa_1\lambda_1 + \kappa_1^2\lambda_1\right]\sigma_+^2(0)f_X^+(0)$.

In light of this theorem, we can define

$$\tilde{t}_n = \sqrt{nh}\frac{\hat{\tilde{\lambda}}}{\sqrt{\hat{V}}}, \tag{18}$$

$\hat{\tilde{\lambda}} = \hat{\tilde{\lambda}}_1 - \hat{\tilde{\lambda}}_0$ 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 3.2. *Suppose the conditions of 3.1 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* $\delta \neq 0$*,* $Pr(\tilde{t}_n \in \mathcal{R}) \to 1 - \Phi\left(c_{1-\alpha/2} - \delta\right) + \Phi\left(c_{\alpha/2} - \delta\right)$*, as* $n \to \infty$*, where* $\Phi(\cdot)$ *denotes the standard normal distribution.*

Proof. The proof of this theorem follows from Theorem 3.1 using straightforward arguments. □

4 Empirical Application

4.1 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. [4] Given these concerns the U.S. government operates a widely applicable $6.2 billion welfare program, the Special 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. Previous literature is thus inconclusive on the actual treatment effect of WIC

[4] A review of the literature by [2] even establishes important links between poor birth outcomes and health and human capital accumulation well into adulthood.

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 ([3];[5];[6]). [5]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 continuous, potentially endogenous variable with an atom at zero.

4.2 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. [6] 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. This feature makes it particularly useful for the investigation of treatment effects under a selection-on-observable framework. 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.

4.3 Estimation of the Test Statistic

As a first step to deal with selection we restrict 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 gives us comparable low income mothers from both treatment and control groups. There is massive negative selection in the full sample which is substantially reduced in this low-income medicaid sample. Next we flexibly control for a wide variety of observables which can explain participation into WIC, specifically the set of covariates in 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

[5][5] 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.

[6]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.

in prepregnancy BMI, flexible controls for gestation, mother's smoking behavior 3 months before pregnancy and in first two trimester of pregnancy. [7]

We use mother's smoking behavior in the third trimester as the potentially endogenous variable, X, in our framework which has an atom at zero. As [4] shows there is prevalence of significant selection across smokers and non-smokers which is especially evident right at the threshold. We, in turn, will use these selection concerns combined with the idea that there should be no differential selection patterns between smokers and non-smokers across our treatment and control groups to test for the presence of non-random selection into treatment group.

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 $Y - Z'\gamma$ 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 some other simple controls which are readily available in most datasets that record birth outcomes. [8] The test statistic from this specification is estimated at -8.43 grams still implying the existence of decent amount of selection. 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 gestation, previous and current pregnancy characteristics and smoking behavior before the third trimester. The estimated test statistic thus falls down to -3.41 indicating a substantial decrease in potential selection concerns. There are a myriad of methods available for the estimation of the treatment effect under the selection-on-observables assumption. However, our framework thus provides the first method in the literature to test this crucial assumption in a binary treatment setting.

5 Discussion:

when the outcome equation is not additively separable in U, the testing ideas presented in the previous sections unfortunately break down. To see this, consider, for example, the model:

$$Y = g(X, D, U), \tag{19}$$
$$D = 1\{h(X) \geq V\}, \tag{20}$$

where the distribution of $V|X = x$ is assumed to be absolutely continuous with respect to Lebesgue measure for *a.e.* x. As before X is assumed to have an atom at 0, but is otherwise continuously distributed. We are going to maintain the following assumptions:

Assumption 5.1. The functions $h(x)$, $g(x, 1, u)$ and $g(x, 0, u)$ are assumed to be continuous in x for almost every u at $x = x_0$.

[7] We employ a similar specification as the one used first by [1]. For a complete list of covariates refer to [6]

[8] However, we still control fully flexibly for these covariates without imposing any functional form on how these variables affect birth weight.

Now

$$E[Y|D=1,X=x] = \frac{\int_{-\infty}^{h(x)} E[g(x,1,U)|V=v,X=x]f_{V|X}(v|x)\mathrm{d}v}{P(x)},$$
$$E[Y|D=0,X=x] = \frac{\int_{h(x)}^{\infty} E[g(x,0,U)|V=v,X=x]f_{V|X}(v|x)\mathrm{d}v}{1-P(x)}.$$

As before, selection on observables assumption is equivalent to assuming $U \perp\!\!\!\perp V|X$. Even when this assumption holds, however, in this case,

$$\begin{aligned} E[Y|D=1,X=x] - E[Y|D=0,X=x] &= E[g(x,1,U)|X=x] - E[g(x,0,U)|X=x] \\ &= \int [g(x,1,u) - g(x,0,u)]\mathrm{d}F_{U|X}(u|x), \end{aligned}$$

which may vary discontinuously if the distribution of $U|X$ varies discontinuously in X. Before proceeding, we note that if along with selection on observables assumption, one uses OLS to estimate the effect of treatment, then finding that $E[Y|D=1,X=x] - E[Y|D=0,X=x]$ is discontinuous in x at $x=0$ is evidence that the estimators are not valid. If one makes the selection on observables assumption and uses propensity score matching, however, finding that $E[Y|D=1,X=x] - E[Y|D=0,X=x]$ is discontinuous in x at $x=0$ 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 is the following: Suppose we test whether $P(x) := E(D|X=x)$ is continuous in x at $x=0$ and find that it is not. In addition, suppose we also test whether $E(Y|D=1,X=x)$ and $E(Y|D=0,X=x)$ are continuous in x at x_0, and find that they both are. Then this suggests that the assumption $U \perp\!\!\!\perp V|X$ is not violated. The reasoning is that under the assumption that h is continuous at $x=0$, discontinuity of $P(x)$ at $x=0$ suggests that $f_{V|X}(v|x)$ should be discontinuous at $x=0$. Therefore, if selection on observables assumption is violated then finding both $E[Y|D=1,X=x]$ and $E[Y|D=0,X=x]$ to be continuous at $x=0$ would be unlikely. Nevertheless, we could have pathological cases where the selection on observables assumption is violated, but both $E[Y|D=1,X=x]$ and $E[Y|D=0,X=x]$ are continuous in x at $x=0$. These cases can be detected by checking say continuity of $E[|Y||D=1,X=x]$ or other functions of Y. If $P(x)$ is discontinuous at $x=0$ and all these other functions are, then this would be a strong suggestion that treatment is independent of U conditional on X.

6 Conclusion:

In this paper, we presented diagnostic procedures for testing the selection on observables assumption. The procedures we presented hinge crucially on two conditions. First, there has to be a variable among the controls that has an atom at a known point, but is otherwise continuous. Second, the structural function relating this variable to the outcome of interest

must be continuous in this variable. The first testing procedure is a formal test of the joint hypothesis of additive separability of outcome equation in treatment and the unobservables. For this testing procedure to have power the expected outcome conditional on the variable with the bunching point, treatment and other possible controls has to have discontinuity in the variable with the bunching point at the bunching point. The tests consists of checking whether this discontinuity is the same for treated and untreated individuals. We also provide a partial diagnostic procedure for checking the selection on observables assumption for non-separable models

A Appendix

A.1 Proof of Theorem 3.1:

We first analyze the asymptotic behavior of the infeasible estimator

$$\widetilde{\mu}^{+}_{\tilde{Y}|X,D}(0,d) = \frac{1}{n_d}\sum_{i=1}^{n} e_1^T M_{nd}^{-1} L_{id} K_h(X_i)\tilde{Y}_i = e_1^T M_{nd}^{-1}\frac{1}{n_d}\sum_{i=1}^{n} L_{id} K_h(X_i)\tilde{Y}_i,$$

where

$$\begin{aligned} L_{id} &:= (1, X_i/h)^T 1\{X_i > 0, D_i = d\}, \\ M_{nd} &:= \frac{1}{n_d}\sum_{i=1}^{n} L_i L_i^T K_h(X_i), \end{aligned}$$

Lemma A.1. *Suppose the conditions of Theorem 3.1 hold. Then*

$$\sqrt{nh}\left(\widetilde{\mu}^{+}_{\tilde{Y}|X,D}(0,1) - \widetilde{\mu}^{+}_{\tilde{Y}|X,D}(0,0) - (\mu^{+}_{\tilde{Y}|X,D}(0,1) - \mu^{+}_{\tilde{Y}|X,D}(0,0))\right) \xrightarrow{d} N(0,V), \tag{21}$$

where

$$V = \frac{M}{C}\left[\kappa_2^2\lambda_0 - 2\kappa_2\kappa_1\lambda_1 + \kappa_1^2\lambda_1\right]\sigma_+^2(0) f_X^+(0). \tag{22}$$

Proof. First, we note that

$$\begin{aligned} \sqrt{nh}e_1^T M_{nd}^{-1}\frac{1}{n_d}\sum_{i=1}^{n} L_{id}K_h(X_i)\tilde{Y}_i &= \sqrt{n}\left(\frac{1}{\frac{n_d}{n}} - \frac{1}{P(D=d)}\right)\sqrt{h}e_1^T M_{nd}^{-1}\frac{1}{n}\sum_{i=1}^{n} L_{id}K_h(X_i)\tilde{Y}_i \\ &\quad + \sqrt{nh}e_1^T M_{nd}^{-1}\frac{1}{nP(D=d)}\sum_{i=1}^{n} L_{id}K_h(X_i)\tilde{Y}_i \\ &= \sqrt{nh}e_1^T M_{nd}^{-1}\frac{1}{nP(D=d)}\sum_{i=1}^{n} L_{id}K_h(X_i)\tilde{Y}_i + o_P(1). \end{aligned} \tag{23}$$

Define

$$S_{nd} = \frac{1}{P(D_i=d)}\frac{1}{n}\sum_{i=1}^{n} e_1^T N_{nd}^{-1}\widetilde{L}_i K_h(X_i)\varepsilon_i,$$

where

$$N_{nd} := E(L_{id}L_{id}^T K_h(X_i)|D_i = d),$$

and $\widetilde{L}_i := (1, X_i/h)^T 1\{X_i > 0\}$. Note that

$$\begin{aligned} E\left[\widetilde{L}_i K_h(X_i)\frac{\varepsilon_i}{P(D_i=d)}\right] &= E\left[\widetilde{L}_i K_h(X_i)\varepsilon_i|D_i=d\right] \\ &= E\left[E\left(\widetilde{L}_i K_h(X_i)\varepsilon_i|X,D_i=d\right)|D_i=d\right] = 0. \end{aligned}$$

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

$$e_1^T M_{nd}^{-1}\frac{1}{nP(D=d)}\sum_{i=1}^n L_{id}K_h(X_i)\tilde{Y}_i = \mu^+_{\tilde{Y}|X,D}(0,d) + S_{nd} + O(h^2) + O_P\left(\frac{\log(n)}{nh}\right), \quad (24)$$

where

$$\mu^+_{\tilde{Y}|X,D}(0,d) := \lim_{x\downarrow 0}\int_{-\infty}^{\infty} y\frac{f_{\tilde{Y},X|D}(y,x|d)}{f_{X|D}(x|d)}\mathrm{d}y. \quad (25)$$

Note that $\mu^+_{\tilde{Y}|X,D}(0,d) = g(0,d) + \lim_{x\downarrow 0}\rho(x,d)$, since $E(\varepsilon|X,D) = 0$ by definition.

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

$$\sqrt{nh}(S_{n1} - S_{n0}) = e_1^T\left[\frac{1}{p}N_{n1}^{-1} - \frac{1}{1-p}e_1^T N_{n0}^{-1}\right]\sum_{i=1}^n\sqrt{\frac{h}{n}}\widetilde{L}_i K_h(X_i)\varepsilon_i.$$

Below we will argue that

$$\sum_{i=1}^n\sqrt{\frac{h}{n}}\widetilde{L}_i K_h(X_i)\varepsilon_i = O_P(1).$$

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

$$\begin{aligned} \sqrt{nh}(S_{n1} - S_{n0}) &= e_1^T A^{-1}\sum_{i=1}^n\sqrt{\frac{h}{n}}\left[\frac{1}{pf^+_{X|D}(0|1)} - \frac{1}{(1-p)f^+_{X|D}(0|0)}\right]\widetilde{L}_i K_h(X_i)\varepsilon_i + o_P(1), \\ &=: \sum_{i=1}^n T_{ni} + o_P(1), \end{aligned}$$

with

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

$f^+_{X|D}(0|d) = \lim_{x\downarrow 0} f_{X|D}(x|d)$ and $\kappa_j = \int_0^\infty u^j K(u)\mathrm{d}u$ for $j = 0, 1, 2$.

We will apply Lindeberg-Feller Theorem to $\sum_{i=1}^n T_{ni}$. Note that

$$\begin{aligned} T_{ni} &= \sqrt{\frac{h}{n}}\frac{e_1^T A^{-1}\widetilde{L}_i K_h(X_i)}{p(1-p)f^+_{X|D}(0|1)f^+_{X|D}(0|0)}\left[(1-p)f^+_{X|D}(0|0) - pf^+_{X|D}(0|1)\right]\varepsilon_i \\ &= \sqrt{\frac{h}{n}}\frac{\left(\kappa_2 - \kappa_1\frac{X_i}{h}\right)1\{X_i>0\}K_h(X_i)}{p(1-p)f^+_{X|D}(0|1)f^+_{X|D}(0|0)(\kappa_0\kappa_2 - \kappa_1^2)}\left[(1-p)f^+_{X|D}(0|0) - pf^+_{X|D}(0|1)\right]\varepsilon_i \end{aligned}$$

Let

$$C := p^2(1-p)^2[f^+_{X|D}(0|1)]^2[f^+_{X|D}(0|0)]^2(\kappa_0\kappa_2 - \kappa_1^2)^2,$$
$$M = \left[(1-p)f^+_{X|D}(0|0) - pf^+_{X|D}(0|1)\right]^2$$
$$\xi_n(X_i) := \left(\kappa_2 - \kappa_1\frac{X_i}{h}\right)^2 1\{X_i > 0\}\frac{1}{h}K^2\left(\frac{X_i}{h}\right).$$

Then

$$\begin{aligned} Var\left(\sum_{i=1}^n T_{ni}\right) &= \frac{M}{C}E\left[\xi_n(X_i)\varepsilon_i^2\right] \\ &= \frac{M}{C}E\left[\xi_n(X_i)\sigma^2(X_i)\right] \\ &\to \frac{M}{C}\left[\kappa_2^2\lambda_0 - 2\kappa_2\kappa_1\lambda_1 + \kappa_1^2\lambda_1\right]\sigma^2_+(0)f^+_X(0), \end{aligned}$$

with $\lambda_j := \int_0^\infty u^j K^2(u)\mathrm{d}u$ for $j = 0, 1, 2$, $\sigma^2(X_i) = E(\varepsilon_i^2|X_i)$, $\sigma^2_+(0) = \lim_{x\downarrow 0}\sigma^2(X_i)$ and $f^+_X(0) = \lim_{x\downarrow 0} f(x)$. 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 $\sum_{i=1}^n E\left(T_{ni}^2\right) \to R < \infty$ and $1\{|T_{ni}| > \epsilon\} \xrightarrow{P} 0$. Thus, the condition holds by the dominated convergence theorem. □

Lemma A.2. $\sqrt{nh}\left(\widehat{\mu}_{\tilde{Y}|X,D}(0,d) - \widetilde{\mu}_{\tilde{Y}|X,D}(0,d)\right) = o_P(1)$.

Proof.

$$\begin{aligned} &\sqrt{nh}\left(\widehat{\mu}_{\tilde{Y}|X,D}(0,d) - \widetilde{\mu}_{\tilde{Y}|X,D}(0,d)\right) \\ =& -\sqrt{h}e_1^T M_{nd}^{-1}\frac{1}{n_d}\sum_{i=1}^n L_{id}K_h(X_i)Z_i^T\sqrt{n}(\hat{\gamma} - \gamma) = o_P(1)O_P(1) = o_P(1). \end{aligned}$$

□

Finally, let $n_{d0} := \sum_{i=1}^n 1\{X_i = 0, D_i = d\}$. and note that

$$\begin{aligned} &\sqrt{nh}\frac{1}{n_{d0}}\sum_{i=1}^n \left(\hat{\tilde{Y}}_i - \tilde{Y}_i\right)1\{X_i = 0, D_i = d\} \\ &= -\sqrt{nh}\frac{1}{n_{d0}}\sum_{i=1}^n Z_i^T(\hat{\gamma} - \gamma) = o_P(1). \end{aligned}$$

Similarly, $\sqrt{nh}\left(\frac{1}{n_{d0}}\sum_{i=1}^n \tilde{Y}_i 1\{X_i = 0, D_i = d\} - E[\tilde{Y}_i|X_i = 0, D_i = d]\right) = o_P(1)$. Thus, the conclusion of Theorem 3.1 follows from combining these last two statements with the conclusions of Lemmas A.1 and A.2.

B Supplementary Appendix

Suppose

$$Y = m(X, D) + U,$$

as before with m continuous everywhere in x for both $d = 0, 1$. In addition, suppose

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

with

$$\begin{pmatrix} U \\ V \\ X^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ \mu_{x^*} \end{pmatrix}, \begin{bmatrix} \sigma_u^2 & \sigma_{uv} & \sigma_{ux^*} \\ \sigma_{uv} & 1 & \sigma_{vx^*} \\ \sigma_{ux^*} & \sigma_{vx*} & \sigma_{x^*}^2 \end{bmatrix} \right) \tag{26}$$

Then

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

where $a = \frac{\sigma_{uv}\sigma_{x*}^2 - \sigma_{ux*}\sigma_{vx*}}{\sigma_{x*}^2 - \sigma_{vx*}^2}$, and $b = \frac{\sigma_{ux*} - \sigma_{uv}\sigma_{vx*}}{\sigma_{x*}^2 - \sigma_{vx*}^2}$.

$$\begin{aligned} Var[U|V = v, X^* = x] &= \sigma_u^2 - \begin{bmatrix} \sigma_{uv} & \sigma_{ux^*} \end{bmatrix} \begin{bmatrix} 1 & \sigma_{vx^*} \\ \sigma_{vx^*} & \sigma_{x^*}^2 \end{bmatrix}^{-1} \begin{pmatrix} \sigma_{uv} \\ \sigma_{ux^*} \end{pmatrix} \\ &= \sigma_u^2 - \frac{1}{\sigma_{x^*}^2 - \sigma_{vx^*}^2} \begin{bmatrix} \sigma_{uv} & \sigma_{ux^*} \end{bmatrix} \begin{bmatrix} \sigma_{x^*}^2\sigma_{uv} - \sigma_{vx^*}\sigma_{ux^*} \\ -\sigma_{vx^*}\sigma_{uv} + \sigma_{ux^*} \end{bmatrix} \end{aligned}$$

Then for $x > 0$

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

Now,

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

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

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

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

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

Similarly,

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

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

Next, consider

$$\lim_{x\downarrow 0} E[U|D=1,X=x] = \lim_{x\downarrow 0} E[U|D=1,X^*=x]$$
$$= -a\sigma_{v|x^*}\frac{\phi\left(\frac{\alpha-\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}\mu_{x^*}}{\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}}\right)}{\Phi\left(\frac{\alpha-\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}\mu_{x^*}}{\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}}\right)} - a\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}\mu_{x^*} - \frac{\sigma_{ux^*}-\sigma_{uv}\sigma_{vx^*}}{\sigma^2_{x^*}-\sigma^2_{vx^*}}\mu_{x^*}.$$

Now

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

So we have

$$\lim_{x\downarrow 0} E[U|D=1,X=x] = -a\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}\frac{\phi\left(\frac{\alpha-\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}\mu_{x^*}}{\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}}\right)}{\Phi\left(\frac{\alpha-\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}\mu_{x^*}}{\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}}\right)} - \frac{\sigma_{ux^*}}{\sigma^2_{x^*}}\mu_{x^*}.$$

Similarly, we have

$$\lim_{x\downarrow 0} E[U|D=0,X=x] = a\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}\frac{\phi\left(\frac{\alpha+\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}\mu_{x^*}}{\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}}\right)}{1-\Phi\left(\frac{\alpha+\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}\mu_{x^*}}{\sqrt{1-\frac{\sigma^2_{vx^*}}{\sigma^2_{x^*}}}}\right)} - \frac{\sigma_{ux^*}}{\sigma^2_{x^*}}\mu_{x^*}.$$

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

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

Since $\mu_{V|X^*}(t)=\frac{\sigma_{vx^*}}{\sigma^2_{x^*}}(t-\mu_{x^*})$, this becomes

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

where we use

$$a\frac{\sigma_{VX^*}}{\sigma^2_{X^*}}+b=\frac{\sigma_{UX^*}}{\sigma^2_{X^*}}.$$

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

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

Similarly,

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

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

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

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

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

   Thus, we have no power in this case.

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

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

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

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

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

Again, we seem to have power in this case.

References

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

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

[3] Bitler, Marianne 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. (2014): A Test of Endogeneity without Instrumental Variables, unpublished manuscript, University of Rochester.

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

[6] Khalil, U. (2014): *The Effect of WIC Participation on Birth Outcomes: New Evidence from over 9 million Births*, unpublished manuscript, University of Rochester.

Figure 1: Medicaid Sample - Basic Specification

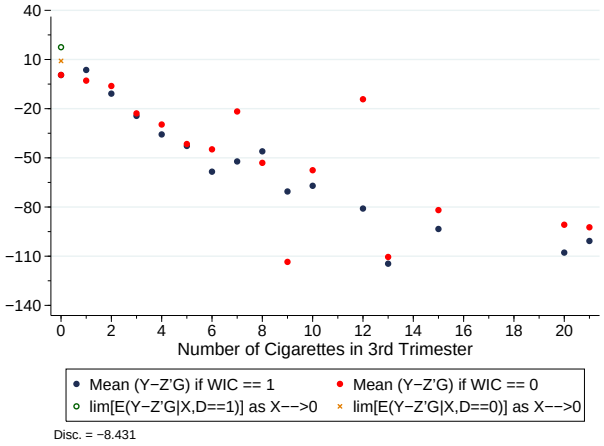

Figure 2: Medicaid Sample - Full Specification