A Test of Exogeneity without Instrumental Variables in Models with Bunching

Carolina Caetano*

University of Rochester, Department of Economics

January 2015 (first version November 2012)

Abstract

This paper presents a test of the exogeneity of a single explanatory variable in a multivariate model. It does not require the exogeneity of the other regressors or the existence of instrumental variables. The fundamental maintained assumption is that the model must be continuous in the explanatory variable of interest. This test has power when unobservable confounders are discontinuous with respect to the explanatory variable of interest, and it is particularly suitable for applications in which that variable has bunching points. An application of the test to the problem of estimating the effects of maternal smoking in birth weight shows evidence of remaining endogeneity, even after controlling for the most complete covariate specification in the literature.

Keywords: Exogeneity Test, Discontinuity, Nonparametric Models, Bunching, Maternal Smoking, Birth Weight. JEL Code: C1, C12, C14, C21.

1 Introduction

Endogeneity is one of the most studied problems in econometrics. Failure to address it generally results in biased estimates and therefore wrong conclusions. There are a number of techniques designed to identify effects in models with endogeneity, such as instrumental variables, panel data with fixed effects, and proxy variables. Unfortunately, such techniques are often difficult to apply to several important problems because they rely on the existence of specific data or natural phenomena. Because of this, a test that can detect endogeneity can be useful, especially when it does not require that a solution to the problem be immediately available.

*I am especially indebted to Jim Powell, Gregorio Caetano, and David Brillinger for their time and valuable input. This paper also benefited from discussions with Bin Chen and Nese Yildiz, as well as several other colleagues and seminar participants, in fact too many to name. Umair Khalil and Liquan Huang provided excellent research assistance. Douglas Almond, Ken Chay, and David Lee graciously provided the data used in this paper. I am thankful to all of them. All errors are my own.
The available tests of exogeneity in the literature are either specification tests (e.g. Gourieroux et al. (1987), Bierens (1990), and Bierens and Ploberger (1997)), or require the existence of instrumental variables (e.g. Hausman (1978) and Blundell and Horowitz (2007)). This paper presents a test of the exogeneity of a single explanatory variable in a multivariate model. The test does not require the existence of instrumental variables, and the other variables in the model may be endogenous. Therefore, this test may be useful for two reasons. First, the test can be useful in the earlier stages of the analysis, before the researcher makes the effort to find and implement any new identification strategy (which could require, for example, a different data set or the existence of an instrumental variable) in order to deal with a possible endogeneity problem. Second, the test can be applied as an omitted variable test to provide guidance in choosing the appropriate model, even with an approach deliberately based on selection on observables.

The test is best explained in an applied example. Consider the problem of estimating the marginal effect of the amount a woman smokes during pregnancy on the baby’s birth weight. The variable “average number of cigarettes per day” is naturally prone to endogeneity, given that there are many pre-existing selection factors associated with both smoking and with birth weight. Examples of such factors include the mother’s education level, marital status, age, etc. The question is whether the amount smoked is exogenous after controlling for the observable factors available in the data. If this is the case, then it is possible to identify the marginal effects of maternal smoking on birth weight.

The fundamental maintained assumption of the exogeneity test is that the structural function must be continuous in the explanatory variable of interest. In the smoking example, it means that the mother’s smoking amount must have a continuous effect on the baby’s birth weight. Suppose that this is indeed true, and then consider Figure 1, which illustrates the expected birth weight for each amount smoked (see plot 19 on page 18 for the empirical version of this figure). If the expected birth weight conditional on the amount smoked is discontinuous, this discontinuity cannot be due to smoking, since smoking has a continuous effect on birth weight. However, this discontinuity may be due to selection on observables, since the observable mother’s characteristics may be discontinuous at zero cigarettes. This is indeed the case, as can be seen, for example, in figures 5 to 12 on page 12. Next, consider Figure 2, which depicts the expected birth weight for each amount smoked for a subgroup of mothers who share the same observable characteristics. If the birth weight per amount smoked is still discontinuous, this discontinuity cannot be caused by smoking (by assumption) nor by selection on observables (because the
observable mother’s characteristics are held fixed). The only explanation for this discontinuity is that there is at least one confounder that was not included in the structural equation, and thus smoking is endogenous.

![Figure 1](image1)

![Figure 2](image2)

The test consists of estimating the expected outcome variable conditional on all the observed variables, and assessing whether it is discontinuous in the variable whose exogeneity requires testing. If a discontinuity is found, then the variable is endogenous.

The smoking example evidences the source of the test’s power. The test has power when at least one unobservable confounder is discontinuously distributed with respect to the endogenous variable. This condition is discussed in more depth in Section 2.2, mentioning several examples in which such discontinuities can be found. Typically this phenomenon can be argued when the variable of interest has bunching point. This is the case in the maternal smoking example, where more than 80% of the observations do not smoke. It should be noted, however, that bunching points in the variable of interest are not a necessary requirement for the applicability of the test.

The paper is organized as follows. Section 2 presents the test idea, as well as discussions of the test’s maintained assumptions (Section 2.1), power (Section 2.2), and implementation (Section 2.3). An application of the test to the problem of the estimation of the effects of maternal smoking on birth weight can be seen in Section 3. That section discusses the applicability of the test (Section 3.1), introduces a test statistic that is particularly suited for this application and can be useful for practitioners interested in implementing the test (Section 3.2), and presents the test results (Section 3.3). The details about the test statistic, including the asymptotic results and small sample behavior in simulations, can be found in an online appendix. Finally, Section 4 concludes.
2 A Discontinuity Test of Exogeneity

This section presents the assumptions under which the test is built as well as the test idea. The following subsections discuss the assumptions themselves (Section 2.1), when such a test would have power (Section 2.2), and how it can be implemented in practice (Section 2.3).

Assumption 2.1. Suppose that the model satisfies

\[ Y = g(X, Z) + U, \]

where \( Y \) is scalar and observable; \( X \) is scalar, observable and \([0, \delta) \subset \text{supp}(X)\) for some \( \delta > 0 \); \( Z \) is a vector of observable variables, which may be continuously distributed or not; and \( U \) is scalar and unobservable.

If \( \mathbb{E}[U|X, Z] = 0 \), \( g \) is identifiable in the entire support of the joint distribution of \( X \) and \( Z \). This is the usual “exogeneity” condition commonly required in the literature (see for example Blundell and Horowitz (2007)). However, this condition is often too strong because it imposes the same level of mean independence between all the components of the vector \((X, Z)\) and the unobservable \( U \).

In order to identify partial effects, for example quantities such as \( \partial g(x, z)/\partial x \), \( \mathbb{E}[\partial g(X, Z)/\partial x] \), or \( \mathbb{E}[\partial g(X, Z)/\partial x|Z = z] \), it is not necessary that all the components of the vector \((X, Z)\) be exogenous. In fact, a weaker condition is sufficient:\( \mathbb{E}[U|X, Z] = \mathbb{E}[U|Z] \) with probability equal to one. (1)

In the remainder of this paper, if condition (1) is satisfied, then \( X \) is said to be exogenous. Otherwise \( X \) is said to be endogenous. Notice that condition (1) is agnostic about the exogeneity of \( Z \) to be endogenous, and thus \( g \) may not be identifiable even if this condition is satisfied.

This paper claims that there exists a functional of the distribution of \((Y, X, Z)\) that is equal to zero when \( X \) is exogenous and unequal to zero when \( X \) is endogenous in an important set of cases. Define

\[ \Delta(X, Z) = \mathbb{E}[Y|X = 0, Z] - \mathbb{E}[Y|X, Z], \]

---

1The notation \( \text{supp}(W) \) denotes the support of the distribution of \( W \).
2Exogeneity is always defined with respect to the model. Hence, since the model examined here is separable, exogeneity is defined as mean independence. If the model were non-separable, exogeneity would be defined as actual independence, i.e. \( U \perp X, Z \). See Blundell and Powell (2003).
3Along with regularity conditions that allow the order of the derivative and the expectation to be exchanged.
then $\tau(z) := \lim_{x \downarrow 0} \Delta(x, z)$ denotes the right discontinuity of $\mathbb{E}[Y|X = x, Z = z]$ as $x \downarrow 0$. In order to explore the relationship between $\tau(z)$ and the exogeneity definition, consider the following assumption.

**Assumption 2.2.** $(x, z) \mapsto g(x, z)$ is continuous at $x = 0$ uniformly for all $z$.

Observe that if the objective is the identification of the average partial effect of $X$, then the continuity requirement is redundant. For the partial effect to be well defined, $g$ must be differentiable in the first coordinate and thus continuous.

If Assumption 2.2 holds, then $\tau(z) = \mathbb{E}[U|X = 0, Z = z] - \lim_{x \downarrow 0} \mathbb{E}[U|X = x, Z = z]$. It is immediate to conclude that if $X$ is exogenous, then $\tau(z) = 0$. This quantity is attractive as a basis for a test of the exogeneity of $X$, provided it is identifiable. The following condition guarantees the identifiability of $\tau(z)$.

**Assumption 2.3.** $\text{supp}(Z) \subseteq \text{supp}(Z|X = 0) \cap \text{supp}(Z|0 < X < \epsilon)$ for all $\epsilon > 0$.

The pair of hypotheses of interest is thus

$$
\mathbb{H}_0 : X \text{ is exogenous} \quad \text{vs.} \quad \mathbb{H}_1 : X \text{ is endogenous}
$$

under the maintained Assumptions 2.1, 2.2 and 2.3. The previous arguments validate the following theorem.

**Theorem 1.** Suppose Assumptions 2.1-2.3 hold. Then $\tau(z)$ is identifiable for all $z \in \text{supp}(Z)$ and, under $\mathbb{H}_0$, $\tau(z) = 0$.

**Remark 2.1. Interpreting the Discontinuity:** One must exercise caution when giving an economic interpretation to the magnitude of $\tau(z)$. It is an incomplete measure of the bias of endogeneity at zero. To see this, denote $\mathbb{E}[U|X = x, Z = z] = m_1(x, z) + m_2(z)1(x = 0)$, where $m_1$ is continuous in $x$ for all $z$ (w.l.o.g.), then $\tau(z) = m_2(z)$, but the bias of endogeneity at $x = 0$ is given by $m_1(0, z) + m_2(z)$. Therefore, $\tau(z)$ does not account for the part of the bias that is caused by the continuous variation of the unobservables, $m_1(x, z)$, be it at zero or anywhere else.

In fact, because the endogeneity at zero has the two components $m_1$ and $m_2$, it may often be the case that the magnitude of the bias of endogeneity is larger at zero than at other points. One cannot claim this formally, since $m_1$ may vary freely, but it could often be the case. In the maternal smoking example, women that smoke

---

4The test presented here will be based on the discontinuity at zero to better align with the example of the effect of maternal smoking. One could write the test based on the discontinuity of $\mathbb{E}[Y|X = x, Z = z]$ at any given generic point $\bar{x}$, as long as $\bar{x}$ is known and the test assumptions hold for that point.
10 cigarettes may have different unobservables from those that smoke 9, but it is unlikely that they would be as different as the comparison between the women that smoked one cigarette and those that did not smoke at all. The group of non-smokers is much more heterogeneous, including women who are far from indifference.

Remark 2.2. Relation to the RDD: The technical similarities with the Regression Discontinuity Design methods may give rise to questions about the relation between the RDD and this test. The methods are not formally related, but this test can be seen as an inversion of sorts of the RDD. Assumption 2.2 is also a requirement in the RDD (it is an indirect consequence of Assumption A1 in Hahn et al. (2001)). However, Assumption 2.2 is the exact opposite (also as an indirect requirement of Assumption A1 in Hahn et al. (2001)).

2.1 Maintained Assumptions

Assumption 2.1 imposes two constraints: the separability of \( U \) and that \( \text{supp}(X) \) is connected in a neighborhood of zero. The first restriction is not vital for the test and is only made for expositional purposes. The ideas can be directly generalized to the fully non-separable case. In this case, the exogeneity condition tested is the conditional independence \( U \perp X | Z \) using \( \tau(z) \) exactly as defined, provided a limit theorem applies.

The second constraint may limit the applications mainly due to a lack of refined data. For example, since it is possible to smoke any fraction of a cigarette, the variable “average cigarettes per day” is likely continuous. However, surveys on maternal smoking behavior measure the average cigarettes per day in integers, which may be too coarse a measure. The concern is that if the expected birth weight varies sharply as the smoked amount approaches zero, the data point nearest to zero, “one cigarette per day,” is in fact too distant to be used to predict the right-limit.

The continuity requirement in Assumption 2.2 is common to most of the applied econometrics literature (often as an indirect consequence of linearity). However, in this paper it plays a crucial role, and it cannot be overlooked. For example, in the smoking example this test could not be implemented if the effect of smoking on birth weight was discontinuous at zero cigarettes. See more details in Section 3, especially Remark 3.1.

One example where Assumption 2.2 likely does not hold is for certain levels of schooling in a typical Mincer equation (e.g. Card (1999)), where \( X \) is education and \( Y \) is wages. Potential employers may use a high school degree as a proxy for ability and thus offer discontinuously higher wages to workers with a high school degree (this discontinuous treatment effect is often labeled as the “sheepskin effect”).
should be noted, however, that in principle the continuity condition can be argued for all other levels of schooling that are not susceptible to sheepskin effects.

Assumption 2.3 can be directly verified in the data. Put in terms of the smoking example, it requires that for each value of the mother’s characteristics (e.g. white, high school educated, unmarried, etc.) there exist both non-smoker mothers and mothers that smoke infinitesimal amounts. This could be restrictive, but the test can be easily modified to substantially relax this condition (see Remark ?? in Section 2.3). It should also be mentioned that Assumption 2.3 can be dropped if some semi-parametric restrictions are imposed on $E[Y|X, Z]$ (for example if $E[Y|X, Z] = \psi(X) + Z'\gamma$).

### 2.2 Detectable Alternatives

Under $H_1$, $E[U|X = x, Z]$ varies with $x$ with positive probability. However, a test based on $\tau(z)$ has non-trivial power only when the following condition holds.

**Assumption 2.4.** Under $H_1$, $E[U|X = x, Z = z]$ is right-discontinuous as $x \to 0$.

The set of cases where the discontinuity test can be applied is thus determined by Assumptions 2.1-2.4. It must be noted that the endogeneity is detected locally. If $X$ is locally exogenous in a neighborhood of zero ($E[U|X = x, Z] = E[U|Z]$, for $x < \delta$ with probability equal to one), then the test will have no power.

Assumption 2.4 is novel and grants further discussion. It can often be argued when the potentially endogenous variable has bunching points. Some of the most common causes of bunching are natural or law restrictions that generate corner solutions. Examples include wages in the problem of the estimation of Engel curves (bunching at the minimum wage), schooling in the problem of the estimation of Mincer equations (bunching at the minimum attendance laws’ thresholds), and weekly hours of work in the problem of the estimation of the effects of hours worked by the mother on the child’s test scores (bunching at zero hours of work). Number of cigarettes (as in the application in Section 3) is another example, as it cannot take non-negative values.

To illustrate how bunching due to a corner solution generates discontinuities in the unobservables, consider the maternal smoking example. It is helpful to think of $U$ as a specific variable, say an aggregate measure of mother’s quality. As the smoking level $X$ varies, so does the distribution of the mother’s quality. In general, it would be expected that incremental changes in smoking would generate incremental changes in the distribution of mother’s quality, and thus $E[U|X = x]$ would change continuously as $x \downarrow 0$. However, the group of mothers that do not smoke is different,
because it includes all the mothers that would have chosen \( X < 0 \) if it was possible. This group is likely to have a disproportional amount of higher quality mothers, and therefore the expected quality of the mothers will be discontinuously higher in comparison with that of the mothers that smoked marginal amounts.

This is indeed the case, as shown in Section 3. Every observable indicator of the mother’s quality in the application data (e.g. mother’s education and alcohol consumption) is discontinuous at zero cigarettes. The hope is that the vector \( Z \) can predict the entire part of the mother’s quality that is related to smoking. The remaining quality should be unrelated to smoking. If that is not the case, and thus \( X \) is still endogenous, then the test will have power if the remaining quality is actually discontinuous at zero. The following example proposes a model that explains how the discontinuities arise when bunching is generated by a lower boundary constraint, as in the case of maternal smoking.

**Example 1.** Suppose that smoking \( X \) is determined by the equation \( X = \max\{0, Z’\pi – Q\} \), where \( Q \) is a continuously distributed unobservable variable. One can understand \( Q \) as an index of the mother’s unobservable quality (the quality of the mother after accounting for the observables \( Z \)). Then

\[
\tau(z) = \mathbb{E}[U|Q \geq z’\pi, Z = z] - \lim_{x \downarrow 0} \mathbb{E}[U|Q = z’\pi - x, Z = z].
\]

Suppose, for example, that \((Q, U)|Z = z \sim \mathcal{N}(\langle 0, 0 \rangle, \langle 1, \rho(z); \rho(z), 1 \rangle)\). Then \( \tau(z) = \rho(z)(\lambda(-z’\pi) - z’\pi) \), where \( \lambda(\cdot) \) is the inverse Mills ratio. If \( \rho(z) \neq 0 \) (and thus \( X \) is endogenous), then \( \tau(z) \neq 0 \).

In general, the equality only holds if (i) \( \mathbb{P}(Q > z’\pi|Z = z) = 0 \). In this case no mother has quality above \( z’\pi \). Hence, the constraint that smoking cannot be negative is irrelevant, as there is effectively no censoring. (ii) \( \mathbb{E}[U|Q = q, Z = z] = \mathbb{E}[U|Z = z] \) for all \( q > z’\pi - \delta \) for some \( \delta > 0 \). In this case the mother’s quality is not a confounder for all mothers with quality above \( z’\pi - \delta \), and therefore \( X \) is locally exogenous in \([0, \delta]\). (iii) The shape of \( \mathbb{E}[U|Q = q, Z = z] \) for \( q > z’\pi \) is such that although it varies with \( q \), it incidentally averages exactly the same value as \( \lim_{x \downarrow 0} \mathbb{E}[U|Q = z’\pi - x, Z = z] \). In this case, if \( \mathbb{E}[U|Q = q, Z = z] \) is continuous in \( q \), it cannot be monotonic when \( q > z’\pi \). This means that for certain higher levels of the mother’s quality, more quality would have a decreasing effect on birth weight. If neither (i), (ii) or (iii) hold, then \( \tau(z) \neq 0 \) and the test has non-trivial power.

Other examples of bunching can be found in the empirical literature in several topics. For instance, Madrian and Shea (2001)’s study on 401(k) savings shows that people tend to bunch at default contract levels, and hence the test might be applied,
for example, to the problem of estimating the effects of savings for retirement on other savings. Similarly, Saez (2010) shows that taxpayers bunch at kink points of the U.S. income tax schedule, and hence the test might be applied, for example, to the problem of estimating the effect of taxable income on tax deductions or to the problem of estimating Engel curves. In both situations, to the extent that bunching occurs along any unobserved variable, that unobserved variable is likely to be discontinuous at the bunching point. If at least one such unobserved variable is not absorbed by the controls, then the test will have power.

Such discontinuities can also be found for reasons other than bunching. For example, figures 3 and 4 use the March Current Population Survey (CPS) data from 2000 to 2009 and show that workers who report working more than 40 hours per week are discontinuously more likely to be male and to work for the federal government compared with workers who report working strictly less than 40 hours.

**Weekly Hours of Work**

**Figure 3:** Percentage of Males  
**Figure 4:** Perc. of Federal Employees

Although it is true that there exists bunching of 25% of the labor force at 40 hours, the discontinuity is noticeable even without the bunching point. In fact, the group that reports working 40 hours or more per week is discontinuously more likely to contain individuals in positions with a fixed minimum workload, such as the typical “9 to 5” worker. Because of this, one may encounter discontinuities in the distribution of many variables related to the choice of profession, industry, or position at 40 hours.

This example is particularly instructive. Notice that the averages at the multiples of 5 (e.g. 20, 25, 30, 35, etc.) seem to follow a different process in comparison with the other points. This may reflect some unobservable characteristic of the pro-
fession or the tendency of individuals to round. Whatever the reason, these points can also be used to implement the test.

2.3 Implementing the Test

A natural way to exploit the result from Theorem 1 is to estimate $\tau(z)$ and to reject the null hypothesis when the realization of $\hat{\tau}(z)$ is “too large” in some suitable norm. For example, if $\mathbb{P}(X = 0) > 0$, $E[Y|X = 0, Z = z]$ can be estimated with a local linear regression of $Y$ onto $Z$ at $z$ using only observations such that $X = 0$, and $\lim_{x \downarrow 0} E[Y|X = x, Z = z]$ can be estimated with a local linear regression of $Y$ onto $X = 0$ and $Z = z$ using only observations such that $X > 0$. A good explanation of this approach for practitioners can be found on Imbens and Lemieux (2008), page 625.

From a technical point of view, this approach presents no difficulties. Given the standard assumptions for the convergence of the multivariate local linear estimator (see Masry (1996)), it is immediate to show that the $\lim_{x \downarrow 0} E[Y|X = x, Z = z]$ term will dominate the asymptotic variance because it has one more estimation dimension and thus converges slower. Therefore, under $H_0$, $\sqrt{n h^{d+1}} \hat{\tau}(z) \xrightarrow{d} \mathcal{N}(0, V)$, where $V$ is the asymptotic variance of the estimator of $\lim_{x \downarrow 0} E[Y|X = x, Z = z]$ as given in Fan and Gijbels (1992), page 2011.

In practice, it is worthwhile to aggregate the discontinuities $\tau(z)$ for several values of $z$. The aggregation can increase the rate of convergence of the test statistic considerably (often up to that of a univariate nonparametric regression) and thus significantly increase the power of the test. Possible aggregations include $E[\tau(Z)]$, $E[|\tau(Z)|]$ and $E[\tau(Z)^2]$. The last two have the desirable property that if the discontinuities have different signs for different values of $Z$, no power is lost because the positive discontinuities cancel out the negative ones. An alternative possibility is to build a test based on the quantity $\sup_{z \in \text{supp}(Z)} |\tau(z)|$.

Section 3 implements the test using an aggregation strategy that is different from the ones suggested above. Though it is not necessarily the most powerful alternative in general, it is well adjusted to the particular features of the data, and it is remarkably simple to implement on standard packaged software. Details of this approach can be found in Section 3 and in an online appendix.

Remark 2.3. Relaxing the support Assumption 2.3: If Assumption 2.3 is too restrictive in the particular application, the aggregation can be modified to include only the discontinuities over the set $A = \cap_{\epsilon > 0} \{ \text{supp}(Z|X = 0) \cap \text{supp}(Z|0 < X < \epsilon) \}$, for example $E[\tau(Z)^2|Z \in A]$, provided $\mathbb{P}(Z \in A) > 0$. 

10
3 An Application to the Effects of Maternal Smoking on Birth Weight

Investments before a person is born can substantially affect the person’s adult life, both with regards to health (e.g. Barker (1990)) and educational and labor market outcomes (e.g. Currie and Hyson (1999), Currie and Moretti (2007), Black et al. (2007), Oreopoulos et al. (2008) and Royer (2009)). For example, Black et al. (2007) find that a 10% increase in birth weight increases high school graduation by 1.2%, IQ (of men) by 1.2%, earnings by 0.9%, and height by 0.3% (see Almond and Currie (2011) for a recent survey on the topic.)

Smoking while pregnant is often viewed as the leading modifiable cause of low birth weight in the United States (Almond et al. (2005)). There is a very large literature of observational studies on the subject, of which Almond et al. (2005) seems to be the most exhaustive approach. The rest of this section applies the exogeneity test to the model given in Assumption 2.1, where $X$ represents the average cigarettes the mother smoked per day during pregnancy, $Y$ is the weight of the child at birth, and $Z$ is the entire vector of covariates used in the full specification in Almond et al. (2005) (the variables list can be seen in their footnote 36, p.1064) and using the same data set as in that paper.

3.1 Applicability of the Test

The first concern for the applicability of the exogeneity test to this problem is whether the causal effect of smoking on birth weight is continuous at zero. An observationally equivalent concern is that the observed smoking variable may not be continuous enough. The average number of cigarettes smoked per day is measured (at best) in integers. The fear is that smoking 1 cigarette may have a very steep effect on birth weight in comparison with subsequent cigarettes so that what is perceived as a discontinuity is in fact a very large derivative. Though this is indeed a possibility, there is currently no data set that can verify whether this is the case. In any event, the results are so pronounced that the qualitative results of the test are likely to hold anyway. See Remark 3.1 for more details.

The second concern is whether Assumption 2.4 is satisfied. Figures 5 to 12 provide heuristic evidence of discontinuities at $X = 0$ for several variables in the data set. The figures were cropped at $X = 40$, which represents 99.95% of the entire population. The dots are estimates of $E[Z|X]$, and the lines show the 95%
Mother’s Demographic Characteristics

**Figure 5:** Education (years)  
**Figure 6:** Age

**Figure 7:** Unmarried  
**Figure 8:** Race: black

Father’s Demographic Characteristics

**Figure 9:** Education (years)  
**Figure 10:** Age

**Figures 5 to 10:** Dots represent average values among pregnant mothers for each level of daily cigarette consumption. The vertical lines represent the 95% confidence interval of the mean.
Mother’s Behavior Variables

**Figure 11:** Consumed Alcohol

**Figure 12:** # of Prenatal Visits

**Figure 13:** Gender of Newborn

**Figure 14:** Order of Newborn

**Figures 11 to 14:** Dots represent average values among pregnant mothers for each level of daily cigarette consumption. The vertical lines represent the 95% confidence interval of the mean. Order of Newborn in figure 14 represents the order among live births.

Confidence interval of the mean for some values of $X$. The confidence intervals become much larger for $X > 20$, due to small samples. Similar discontinuities can be found for most of the confounders in the data, and the hope is that the same patterns would also be found with the unobservable confounders. There are a few exceptions, as seen in Figures 13 and 14, where no discontinuity is found. However, it is not clear that these variables are confounders at all.

The discontinuities are not only present in the first moment of the distribution. To illustrate this point, figures 15 to 18 show the (kernel density) distribution of a few variables for each level of cigarettes, from $X = 0$ to $X = 3$. These figures provide evidence that the distribution for $X = 0$ is not a simple mean shift of the distribution for $X$ positive. Rather, the variance and often the shape change as well.
3.2 Building the Test Statistic

As discussed in Section 2.3, because of concerns of severe loss of efficiency due to high dimensionality, it is recommendable to aggregate the discontinuities. This application will use the following aggregation:

\[
\theta = \lim_{x \downarrow 0} \mathbb{E}[\Delta(x, Z) | X = x] = \lim_{x \downarrow 0} \mathbb{E}[\mathbb{E}[Y | X = 0, Z] - Y | X = x].
\]

The advantages of this particular aggregation over other (perhaps more efficient) choices discussed in Section 2.3 are practical. With over 30 covariates (more than 300 covariates if one counts all the interactions used in the full specification of Almond et al. (2005)), the multivariate nonparametric boundary term \( \lim_{x \downarrow 0} \mathbb{E}[Y | X = x, Z] \)
cannot be estimated with a multivariate local linear estimator using reasonably small bandwidths. However, because of the law of iterated expectations, the aggregation in $\theta$ eliminates this term. The remaining multivariate nonparametric term $E[Y|X=0,Z]$ is not a boundary quantity and thus can be estimated with other methods such as series, or flexible OLS. Moreover, since 80% of the sample does not smoke, the estimation of $E[Y|X=0,Z]$ counts with over 400,000 observations. The only boundary quantity that has to be estimated is the limit of the outer expectation, which can be done with a univariate local linear regression without any practical difficulty.

The immediate concern with the aggregation in $\theta$ is that there could be a loss of power if the discontinuities $\tau(z)$ change signs for different values of $z$. In this application, this would become a problem if the mothers that do not smoke are often selected worse than the mothers that smoke marginal amounts. This seems unlikely. The practical advantages of this aggregation in this application far outweigh the risk of some loss of power due to this possibility.

Additionally, $\theta$ has a direct intuitive interpretation in relation to the model. If smoking is exogenous, then the observable mother characteristics $Z$ are enough to account for the selection to different levels of smoking treatment. In this case, the mothers that did not smoke can be used to calculate the counterfactual effect if the mothers that smoked marginal amounts stopped smoking altogether. If the effect of smoking is continuous, there should be no effect, since the treatment difference is only marginal across both groups. A difference can be found only if the non-smoking mothers cannot serve as counterfactual for the mothers that smoked marginal amounts, i.e. they have systematic unobservable differences that affect birth weight.

The test is thus based on $\theta$, which is estimated in two steps. The first step assumes that $E[Y|X=0,Z] = Z'\gamma$, where $Z$ is the full covariate specification in Almond et al. (2005), and estimates $\gamma$ with an OLS regression of $Y$ onto $Z$ using only observations such that $X = 0$. The second step is a local linear regression of $Z'\hat{\gamma} - Y$ onto $X$ at $X = 0$ using only observations such that $X > 0$. The results in the next section use the epanechnikov kernel.\(^5\) The test statistic is $T_n = \hat{\theta}/\sqrt{\hat{\Omega}/nh}$, where $\hat{\Omega}$ is the estimator of the asymptotic variance of $\sqrt{nh}(\hat{\theta} - \theta)$. $T_n$ should be compared to the percentiles of the standard normal distribution.

\(^5\)Rectangular and triangular kernels, as well as local polynomial regressions of degrees 2 and 3 in place of the local linear estimator were also used with very similar results. The Monte Carlo study in the online appendix suggests that the local linear estimator is more stable. This is likely a consequence of the fact that the estimation is done at the boundary, where Runge’s phenomenon causes large variability in the results for the higher order polynomials when larger bandwidths are used.
The online appendix provides more details about the estimators $\hat{\theta}$ and $\hat{\Omega}$. It also contains the results that establish the asymptotic behavior of $T_n$, as well as real-data Monte Carlo simulations that showcase its small sample behavior (although the sample in the application is anything but small). The test can be programmed directly in any packaged software as two regressions. A Stata code can be obtained in the author’s website.

### 3.3 Test Results

Table 1 shows the discontinuity test results for two outcome variables commonly studied in the medical literature. The first is the baby’s birth weight in grams. The second is the probability of low birth weight (LBW) expressed in percent likelihood, which is defined as birth weight below 2,500 grams. Specification I estimates the discontinuity in the outcome conditional on the amount smoked, without covariates, i.e. $\lim_{x \downarrow 0} \mathbb{E}[\mathbb{E}[Y|X = 0] - Y|X = x]$. For example, the values in the top row in table 1 can be interpreted as the expected difference in birth weight between a non-smoking mother and a mother that smokes a marginally positive amount. Of course, this difference can be explained by selection on observables. Specification II includes all the covariates. Hence, the numbers in the first row of Specification II in Table 1 are the values of $\hat{\theta}$.

The columns in Table 1 show the estimates for several bandwidths used in the second step of the estimator. For the smallest bandwidth, $h = 3$, the results for birth weight show that non-smoking women have babies that are 203 grams heavier than those that smoke a marginally positive amount. This difference is at least partially explained by selection on observables. Indeed, the results of specification II show that 66 grams are explained by the covariates. However, 137 grams remain unexplained and are thus evidence of endogeneity. The results show that the discontinuity is notably larger for $h = 3$ in comparison with the other bandwidths. However, $\hat{\theta}$ becomes even more significant as the bandwidth increases, and thus the qualitative point stands under any bandwidth.

The results for the probability of low birth weight show less evidence of endogeneity. The test was significant for all but the smallest bandwidth, but the test statistic was always considerably smaller than in the birth weight case. Since LBW is a threshold variable, its determination has to be equal to or less complex (in terms of the number of possible confounders) than the actual birth weight level. Thus, it is expected that the endogeneity will be the same or lower than for the birth weight process.
Table 1: Test Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Birth Weight</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>203 178 184 186 202 204</td>
</tr>
<tr>
<td></td>
<td>(34) (18) (12) (11) (7) (7)</td>
</tr>
<tr>
<td>II</td>
<td>137 99 75 74 85 87</td>
</tr>
<tr>
<td></td>
<td>(30) (16) (11) (10) (6) (6)</td>
</tr>
<tr>
<td>LBW (%)</td>
<td></td>
</tr>
<tr>
<td>I</td>
<td>-4.7 -3.9 -4.3 -4.3 -4.4 4.4</td>
</tr>
<tr>
<td></td>
<td>(1.6) (0.9) (0.6) (0.5) (0.4) (0.3)</td>
</tr>
<tr>
<td>II</td>
<td>-2.8 -1.9 -1.5 -1.4 -1.5 -1.5</td>
</tr>
<tr>
<td></td>
<td>(1.6) (0.8) (0.6) (0.5) (0.3) (0.3)</td>
</tr>
</tbody>
</table>

Table 1 shows the values of $\hat{\theta}$ and the standard errors $\sqrt{\hat{\Omega}/nh}$ (between parentheses) for birth weight and the percent likelihood of low birth weight. Specification I uses no controls, and specification II uses the most complete control specification in Almond et al. (2005). The numbers under the bandwidths represent the proportion of the smoking mothers weighted positively by the kernel (out of 94,212 observations).

Figures 19 and 20 depict the main results from Table 1. Intuitively, the difference between the solid dot representing $\hat{E}[Y|X = 0]$ and the hollow dot cannot be due to smoking. However, it can be due to selection on observables. The $\times$ point depicts the estimated expected birth weight of a non-smoker with the same covariates as a woman that smokes just a marginally positive amount. The difference between this point and the solid dot $\hat{E}[Y|X = 0]$ is due entirely to differences in the covariates of non-smokers and those that smoke a marginally positive amount. This part of the discontinuity is due to selection on observables. The distance between the $\times$ point and the hollow dot is the expected difference in birth weight between a non-smoking mother that has the same covariates as a mother that smoked a marginally positive amount and a mother that did smoke a marginally positive amount. Since both mothers have the same covariates, the difference can only be explained by selection on unobservables.

Figure 20 has the same interpretation. In the case of the probability of LBW, selection on observables explains a comparatively larger proportion of the discontinuity. The part of the difference that is due to endogeneity (the difference between the $\times$ point and the hollow dot) is significant for most choices of bandwidth, albeit the test statistic is considerably smaller than in the birth weight case.
Figures 19 and 20: Dots represent average values among pregnant mothers for each level of daily cigarette consumption. The vertical lines represent the 95% confidence interval of the mean. High variability for higher amounts of smoking is just small sample variance. The hollow dot represents the estimate of \( \lim_{x \downarrow 0} \mathbb{E}[Y|X = x] \). The “×” point represents the estimate of \( \lim_{x \downarrow 0} \mathbb{E}[\mathbb{E}[Y|X = 0, Z]|X = x] \). Their difference is \( \hat{\theta} \). (Bandwidth is \( h = 3 \).)

**Remark 3.1.** A natural concern with this application is that the causal effect of smoking could be discontinuous at zero. An observationally equivalent concern is that the observed smoking variable may be too coarse and smoking 1 cigarette may have a very steep effect on birth weight in comparison with subsequent cigarettes,
so that what is perceived as a discontinuity is in fact a very large derivative. To entertain this possibility, suppose that the model is

\[ Y = g(X) + Z'\gamma + U, \]

where \( E[U|X, Z] = 0 \), \( g(0) = 0 \) (w.l.o.g) and \( g \) may be discontinuous at zero. The effect of a marginal amount of smoking is thus \( \lim_{x \downarrow 0} g(x) - g(0) \), and the next \( x \) cigarettes will have a combined effect of \( g(x) - \lim_{x \downarrow 0} g(x) \).

This is a selection on observables model that attributes the discontinuity at zero to the causal effect of smoking. This model is identifiable. \( \gamma \) can be estimated by regressing \( Y \) onto \( Z \) using only observations such that \( X = 0 \) (because \( E(Y|X = 0, Z) = Z'\gamma \)). Then, since \( E[Y - Z'\tilde{\gamma}|X = x] = g(x) \), \( g \) is estimated by regressing \( Y - Z'\tilde{\gamma} \) onto \( X \) using a local linear estimator with \( h = 3 \) and the epanechnikov kernel.

The estimates yield \( \lim_{x \downarrow 0} g(x) - g(0) = -137 \) grams, and \( g(10) - \lim_{x \downarrow 0} g(x) = -67 \) grams. This means that, if this model is correct, smoking just a minimal amount would have more than two times the effect of the next 10 cigarettes. It seems reasonable to presume instead that even if the smoking effects are discontinuous or very steep for small smoking amounts, at least part of the estimated effect is due to endogeneity.

**Remark 3.2.** There could be a concern with measurement error on the cigarettes variable, namely that mothers can misreport how much they smoke. Indeed, in the data there is a large concentration of observations at 5 and 15 cigarettes, and a much higher concentration at 10 and 20 cigarettes. However, medical research that relates self-reported smoking levels and smoking biomarkers such as urine and blood cotinine levels shows that the measurement error is much smaller than intuition suggests. The most recent studies with the best sets of controls conclude that less than 1% of the population misreport the amount smoked (e.g. Yeager and Krosnick (2010), Pickett et al. (2005), and Caraballo et al. (2001)).

Even if it is true that there is measurement error in the reporting of cigarettes, the test detects endogeneity of any kind, including the kind generated by measurement error. Thus, although measurement error will affect the power of the test, it will not invalidate it. In fact, endogeneity combined with measurement error may yield a more powerful test, for example if the mothers that underreport are better selected than the average non-smoker.
4 Conclusion

This paper develops a test of the exogeneity of a single variable, which does not require instrumental variables. In fact, it does not require the identifiability of the structural function even under the null hypothesis. Thus, this test can be used to validate a selection on observables approach.

The test depends on the continuity of the causal effect of the explanatory variable of interest. The power of the test derives from the discontinuity of the distribution of the unobservable confounders at a known point in the domain of the variable of interest. When the variable is endogenous, the discontinuities in the unobservables generate discontinuities in the expectation of the outcome variable conditional on the observables. The test consists of estimating such discontinuities, averaging them over a particular distribution of the covariates, and then testing for whether this average is equal to zero.

This paper also presents an application of the test to the problem of the estimation of the effect of maternal smoking on birth weight. A particular aggregation is suggested in the context of this application which has great practical advantages. The results suggest that there is substantial endogeneity remaining for the birth weight outcome, even after controlling for the most complete covariate specification in the literature. The evidence of endogeneity is considerably weaker for the probability of low birth weight outcome.

References


