Smile: A Simple Diagnostic for Selection on Observables

Slichter, David

Binghamton University (SUNY)

25 April 2020

Smile: A Simple Diagnostic for Selection on Observables

David Slichter
Binghamton University (SUNY)[*]

April 2020

Abstract

This paper develops a simple diagnostic for the selection on observables assumption in the case of a binary treatment variable. I show that, under common assumptions, when selection on observables does not hold, designs based on selection on observables will estimate treatment effects approaching infinity or negative infinity among observations with propensity scores close to 0 or 1. Researchers can check for violations of selection on observables either informally by looking for a "smile" shape in a binned scatterplot, or with a simple formal test. When selection on observables fails, the researcher can detect the sign of the resulting bias.

1 Introduction

When studying the effect of a treatment on an outcome of interest, researchers often use estimators which require that treatment status is as good as randomly assigned conditional on a set of covariates. This assumption is known variously as *selection on observables*, *unconfoundedness*, or the *conditional independence assumption*. Ordinary least squares (OLS) regression and propensity score matching are examples of estimators which can be used to measure causal effects under selection on observables.

Estimators based on selection on observables can be attractive due to ease of implementation and because they often give more precise estimates than alternatives (see e.g. Young 2020). However, these advantages can be overshadowed by the difficulty of knowing whether selection on observables actually holds in any particular

[*] State University of New York at Binghamton, Economics Department, PO Box 6000, Binghamton NY 13902, slichter@binghamton.edu. Thank you to Gregorio Caetano, Hugo Jales, Umair Khalil, Sulagna Mookerjee, Solomon Polachek, Yulong Wang, and participants at the Stockman Conference and a LISER workshop for helpful comments. All remaining errors are mine.

context. This paper introduces a simple plausibility check for selection on observables in the case of a binary treatment.

Consider the following simple model. Suppose we observe an outcome Y, binary treatment D, and vector of covariates X. We are interested in the causal effect of D on Y. Suppose the structural equations for D and Y are as follows:

$$D = 1\{h(X) + U > 0\}$$
$$Y = \xi + X'\beta + \delta D + \gamma U + V,$$

where h is some function, U and V are unobserved variables, U is independent of X, and $E(V|D, X, U) = 0$. If $\gamma = 0$, then there is no confounding and the effect of D on Y can be recovered by regressing Y on D controlling for X. If γ is positive (negative), then this regression suffers from positive (negative) omitted variable bias.

We are interested in whether selection on observables ($\gamma = 0$) holds in the above model. The key intuition which allows us to assess this assumption is as follows. Among observations with propensity scores close to 1 (0), untreated (treated) observations are quite unusual; to have their treatment status, they must have an outlier value of the unobserved variables determining treatment (in the model above, U). The closer the propensity score is to 1 (0), the more extreme these outliers must be. Therefore, the difference in the unobservables between treated and untreated observations becomes very large at extreme propensity scores. It follows that, if the unobservables enter into the outcome equation (i.e., $\gamma \neq 0$), the resulting omitted variables bias becomes large in populations with extreme propensity scores.

In Section 2, I show that the implication of this increasingly large bias is that, when there is positive (negative) omitted variables bias, the implied treatment effect as a function of propensity score asymptotes to infinity (negative infinity) as the propensity score approaches both 0 and 1. The presence of asymptotes can be checked informally with a binned scatterplot. Additionally, I construct a simple formal test that can be run in a few seconds using existing commands in standard statistical software. Section 5 demonstrates that this approach has reasonable finite sample performance using simulations, while Section 6 contains an empirical example related to the effect of incumbency in United States House of Representatives elections.

It is important to note that the test comes with caveats, due to the fact that the structural model described above will not always apply. Practitioners should be sure to read Section 3 and review the list of situations generating false positives and false negatives in Section 4. Practitioners will generally have priors about whether the conditions described in these sections hold or not, but not full knowledge. Because of this, the test is properly understood as informative rather than decisive, i.e. it allows a practitioner to Bayesian update about the probability that selection on observables holds but will not lead to certainty of belief even in infinite samples. While it is nicer for a test to be decisive, informative tests with well-understood blind spots are useful enough to be ubiquitous in the applied literature – for instance, overidentification tests used to study the validity of instrumental variables (e.g. Sargan 1958, Hansen

1982), or the standard practice of inspecting for common pre-trends in difference-in-differences designs.

Some tests of the selection on observables assumption exist in the literature. One line of research offers tests of selection on observables when an instrument is available (e.g., Hausman 1978, Blundell and Horowitz 2007). There are also instrument-based tests of the conditional mean independence assumption (Donald, Hsu, and Lieli 2014, Chen et al. 2018), which is sufficiently closely related to selection on observables that the simple model above does not allow for any distinction between these cases. Another recent line tests selection on observables by exploiting bunching (Caetano 2015, Khalil and Yildiz 2017). Finally, there are tests based on distributional assumptions and functional form (e.g., Heckman 1979, Rivers and Vuong 1988).

It has also previously been noted that, when the error term in the treatment equation is normally distributed, bias in selection on observables designs with omitted variables is minimized at a propensity score of .5 (Black and Smith 2004). However, I am not aware of any prior work demonstrating the presence of asymptotes at extreme propensity scores, or applying this to construct a test of selection on observables.

Section 2 develops the key implication of the model. Section 3 discusses the sensitivity of the model's implications to its assumptions. Section 4 describes how to test selection on observables based on the model's implications. Section 5 shows simulation results for the formal test. Section 6 shows performance in an application. Section 7 concludes.

2 Model results

Assume that the structural model is as above. Let $P(X) := Pr(D = 1 \mid X)$. $P(X)$ is commonly referred to as the propensity score. Additionally, let $\Delta(p)$ be the difference in the mean of Y between treated and untreated observations at propensity score p. That is, let

$$\Delta(p) := E(Y|D = 1, P(X) = p) - E(Y|D = 0, P(X) = p).$$

Note that, if selection on observables holds, then $\Delta(p)$ is the average treatment effect for observations with $P(X) = p$ (Rosenbaum and Rubin 1983).

Let $\Delta_U(p)$ be the difference in the mean of U between treated and untreated observations at propensity score p:

$$\Delta_U(p) := E(U|D = 1, P(X) = p) - E(U|D = 0, P(X) = p).$$

The following lemma states that, as the propensity score approaches 0 or 1, then $\Delta_U(p)$ asymptotes.

Lemma 1. *Suppose U and $h(X)$ have full support and U has a finite first moment. Then* $\underset{p\to 1}{lim}\, \Delta_U(p) = \underset{p\to 0}{lim}\, \Delta_U(p) = \infty$.

Proof. See Appendix A.1. □

A brief summary of the proof is that $lim_{p\to 1}E(U|D = 1, P(X) = p)$ is bounded below by $E(U)$ while $lim_{p\to 1}E(U|D = 0, P(X) = p) = -\infty$. This proves the lemma for p approaching 1. Similarly, $lim_{p\to 0}E(U|D = 1, P(X) = p) = \infty$ while $lim_{p\to 0}E(U|D = 0, P(X) = p)$ is bounded above, which proves the lemma for p approaching 0.

That is, the infinite divergence between average values of U arises because the "weird" observations – untreated observations with high propensity scores and treated observations with low propensity scores – move to increasingly extreme values of the unobservables.

Now, simple math (see Appendix A.2 for proof) gives that

$$\Delta(p) = \delta + \gamma\Delta_U(p).$$

That is, the difference in mean outcomes at propensity score p equals the treatment effect δ plus an omitted variables bias term $\gamma\Delta_U(p)$. Lemma 1 shows that $\Delta_U(p)$ asymptotes to infinity at extreme propensity scores, so it follows that $\Delta(p)$ asymptotes as well whenever $\gamma \neq 0$. In particular, if γ is positive, the asymptotes are to positive infinity, while if γ is negative, the asymptotes are to negative infinity.

Theorem 1. *Suppose U and $h(X)$ have full support and U has a finite first moment.*

- *If $\gamma > 0$, then $\underset{p\to 1}{lim}\,\Delta(p) = \underset{p\to 0}{lim}\,\Delta(p) = \infty$.*
- *If $\gamma < 0$, then $\underset{p\to 1}{lim}\,\Delta(p) = \underset{p\to 0}{lim}\,\Delta(p) = -\infty$.*
- *If $\gamma = 0$, then $\underset{p\to 1}{lim}\,\Delta(p) = \underset{p\to 0}{lim}\,\Delta(p) = \delta$.*

Proof. See Appendix A.2. □

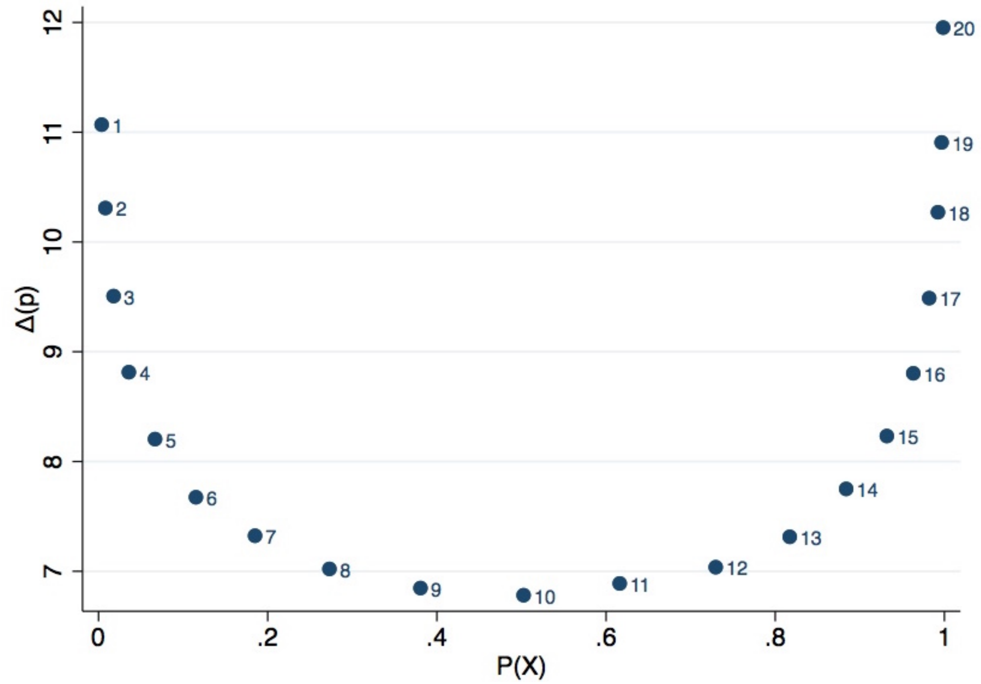
Theorem 1 gives a potentially testable implication of selection on observables: If selection on observables does not hold, implied treatment effects explode to infinite values at extreme propensity scores. In particular, if selection on observables does not hold, then the sign of the infinity is equal to the sign of the omitted variables bias. Otherwise, implied treatment effects remain constant at the true effect of treatment.

2.1 Illustration

To illustrate the implications of the model, I simulate data and illustrate the implied treatment effects. I first simulate the data using the following data generating process:

$$D = 1\{-3 + .3X + U > 0\}$$
$$Y = 4X + 2D + 3U + V,$$

Figure 1: Binned scatterplot with a smile



Simulation results: Plot of the estimated difference in means $\Delta(p)$ against the estimated propensity score across the 20 points of support of propensity score when $\gamma = 3$.

where X is a random integer between 1 and 20 drawn independently of U and V, and where

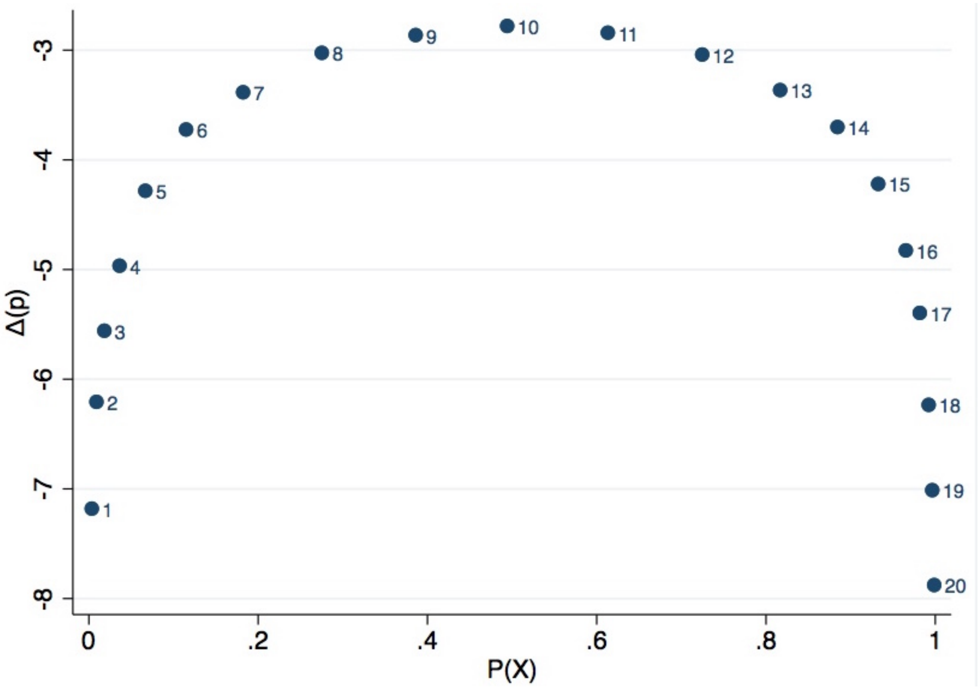
$$\begin{bmatrix} U \\ V \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right).$$

Using 1,000,000 observations, I bin by value of X and estimate both $P(X)$ and $\Delta(p)$ using sample averages. Figure 1 shows a graph of the results, with each point representing a value of X, and with the estimated $\Delta(p)$ on the vertical axis and the estimated $P(X)$ on the horizontal axis. Note the characteristic "smile" shape, consistent with upwards omitted variable bias.

A second simulation, illustrated in Figure 2, uses the same data generating process, but with the coefficient on U in the Y equation set to -3 instead of 3. Note the characteristic "frown," consistent with downwards omitted variables bias.

In each case, the implied treatment effect remains fairly stable across intermediate values of $P(X)$, but then accelerates rapidly at propensity scores beyond .1 or .9. The exact location of the point of acceleration is of course dependent on the normality of U, so this region of acceleration may not hold in all applications.

Figure 2: Binned scatterplot with a frown



Simulation results: Plot of the estimated difference in means $\Delta(p)$ against the estimated propensity score across the 20 points of support of propensity score when $\gamma = 3$.

3 Sensitivity to alternate assumptions

The model used to produce the previous section's results makes several restrictions which a practitioner would not necessarily make in a selection on observables setting. This section discusses relaxations of those assumptions, and how they would affect the conclusions drawn from Theorem 1.

Heterogeneous treatment effects The model posits a single treatment effect δ which is constant across propensity scores. Most researchers would interpret their estimate of δ as a kind of average treatment effect, but would not assume that this average is constant across all subpopulations. Hence, my model is making an additional assumption.

A simple modification of the model would be to allow treatment effects to vary with X. That is, we might instead write the Y equation as

$$Y = \xi + X'\beta + g(X)D + \gamma U + V,$$

where $g(X)$, the treatment effect, is now some function of X. We might also define $\delta(p) := E(g(X)|P(X) = p)$, the average treatment effect at propensity score p.

Lemma 1 still holds, and, following the proof of Theorem 1, $\Delta(p) = \delta(p)+\gamma\Delta_U(p)$. If $\underset{p\to 1}{lim}\ \delta(p)$ and $\underset{p\to 0}{lim}\ \delta(p)$ are each finite, then Theorem 1 is unaffected, up to replacing δ with some other finite numbers as the limits in the $\gamma = 0$ case.

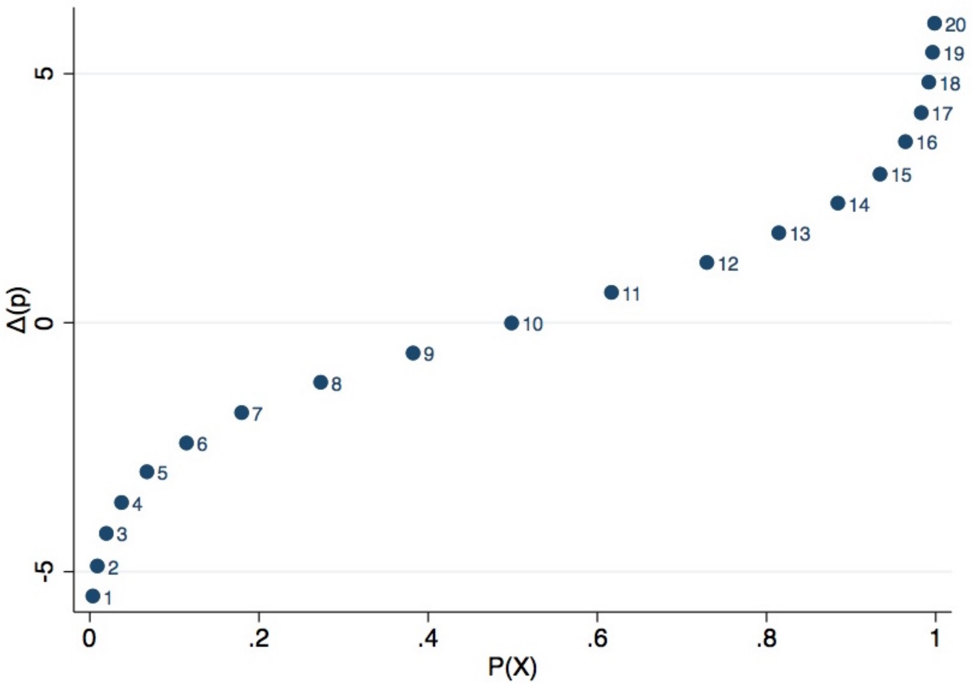
A more tricky case is if treatment effects go to infinity at one or both propensity score extremes. An infinite treatment effect added to a zero bias term will be infinite. An infinite treatment effect added to an opposite-signed infinite omitted variable bias can result in a sum which is either finite, the infinity of the treatment effect, or the infinity of the bias term.

This can result in erroneous conclusions based on Theorem 1 under two circumstances, each of which requires $\delta(p)$ to have a smile or frown shape. First, if $\delta(p)$ has a smile or frown shape and $\gamma = 0$, Theorem 1 would lead us to conclude that selection on observables does not hold. Second, when $\gamma \neq 0$ and $\delta(p)$ has a smile (frown) shape, it is possible to have a knife-edge case in which this cancels out the oppositely-oriented frown (smile) that results from the bias term, resulting in finite limits of $\Delta(p)$.

A final situation of interest is that infinite $\delta(p)$ can lead to data patterns which are not predicted under *any* value of γ by Theorem 1. This might occur, for example, if $\gamma = 0$, the D equation is a linear single index model, and the interaction of D and X appears in the Y equation, resulting in a "smirk" shape, illustrated in Figure 3 using a simulation.[1] The other ambiguous shape is an asymptote at one propensity score extreme but a finite limit at the other. These shapes can be rationalized both

[1] The parameters of the simulation are the same as in Section 2, but with γ set to 0 and replacing $\delta * D$ in the Y equation with $\delta * W * D$, where $W = -1 + .3X$.

Figure 3: Binned scatterplot with a smirk



Simulation results: Plot of the estimated difference in means $\Delta(p)$ against the estimated propensity score across the 20 points of support of propensity score when $\gamma = 0$ and treatment effects are increasing in X.

if selection on observables holds and if it doesn't, so it is probably best to treat such shapes as inconclusive.[2]

Assumptions on unobservables The model makes some important assumptions about the unobservables, which are used to (a) produce Lemma 1, and (b) produce Theorem 1 conditional on Lemma 1. The assumptions used for (a) are that U has full support, a finite first moment, and is independent from X, while (b) involved assuming that U enters linearly in the Y equation.

Violations of these assumptions would never result in a false rejection of selection on observables based on Theorem 1. This is because, under selection on observables, $\gamma = 0$ (a special case of U appearing linearly in the Y equation) and $\Delta_U(p)$ does not influence $\Delta(p)$ (hence Lemma 1 is irrelevant).

However, violations can sometimes lead to an incorrect failure to reject selection on observables, i.e. it is possible that implied treatment effects remain finite at ex-

[2] The smirk shape in Figure 3 could be fit, for example, with $\gamma > 0$ and $\delta(p)$ rapidly approaching $-\infty$ as p goes to 0.

treme propensity scores if selection on observables fails and these assumptions do not hold. This can happen either if (a) or (b) fails.

The assumptions used to guarantee (a) are sufficient but not necessary for Lemma 1. For example, even if U is not independent of X, it is clear that Lemma 1 will still hold so long as U has approximately the same distribution at extreme propensity scores as at non-extreme ones. But situations of concern would be ones where U is bounded, or where the distribution of U becomes very narrow at extreme propensity scores.

Similarly, the assumption used for (b) is sufficient but not necessary for Theorem 1, conditional on Lemma 1. Any scenario in which, at extreme propensity scores, sending U to ∞ sends Y to ∞ (or $-\infty$), and sending U to $-\infty$ sends Y to $-\infty$ (or ∞), would suffice to make violations of selection on observables detectable.

The following are two situations which might make (b) fail.

First, certain forms of non-linearity of U in the Y equation can cause trouble. One possibility is if the effect of U on Y is bounded, e.g. if we replaced γU with $\gamma \frac{e^U}{1+e^U}$. In this case, infinite differences in unobservables between treated and untreated observations do not translate to infinite differences in Y. Non-monotonicity can also produce inconclusive results, e.g. we get a smirk shape if U^2 enters into the Y equation instead of U.[3]

Note that, if Y is bounded (e.g., due to top-coding), then the effect of U on Y is necessarily bounded. In this case, implied treatment effects cannot possibly go to infinity. Nonetheless, provided that the boundedness is not so extreme as to censor a large fraction of observations, we would expect to see a spike in $\Delta(p)$ for p close to 0 and 1.

Second, translating Lemma 1 into Theorem 1 requires that U affects Y specifically at extreme propensity scores. Therefore, it is important to ask if the effect of U on Y might vary with X. Consider capturing this by replacing γ in the Y equation with $\gamma(X)$. Theorem 1 fails if $\gamma(X)$ is not always 0, but as $P(X)$ goes to 0 or 1, $\gamma(X)$ goes to zero sufficiently quickly that the bias term $\gamma(X)\Delta_U(p)$ does not explode even though $\Delta_U(p)$ is going to infinity.[4] Note also that, if $\gamma(X)$ has opposite signs at one propensity score extreme as the other, then we will observe an inconclusive smirk shape.[5]

[3] Explanation: The lower bound of U to produce $D = 1$ is going to infinity as p approaches 0, so the lower bound of U^2 is also going to infinity, and therefore $E(U^2|D = 1, P = p) - E(U^2|D = 0, P = p)$ goes to infinity as p goes to 0. Meanwhile, the upper bound of U to produce $D = 0$ is going to negative infinity as p goes to 1, so $E(U^2|D = 0, P = p)$ is going to infinity, so $E(U^2|D = 1, P = p) - E(U^2|D = 0, P = p)$ is going to negative infinity. This means that $\Delta(p)$ will go to opposite-signed infinities at the two extremes of propensity scores.

[4] This is because $\Delta(p) = \delta + \gamma(X)\Delta_U(p)$ (see Appendix A.2) and Lemma 1 still holds.

[5] An example of a modified set of assumptions which preserves Theorem 1 is as follows. Suppose that $\gamma(X) > 0$ for all X and $\gamma(X)$ does not go to zero as $P(X)$ goes to 0 or 1. Then $\underset{p\to 1}{lim}\, \Delta(p) = \underset{p\to 0}{lim}\, \Delta(p) = \infty$. Similarly, if $\gamma(X) < 0$ for all X and $\gamma(X)$ does not go to zero as $P(X)$ goes to 0 or 1, then $\underset{p\to 1}{lim}\, \Delta(p) = \underset{p\to 0}{lim}\, \Delta(p) = -\infty$.

A final important consideration is that U might actually be a vector of unobservables, and we may fail to detect violations of selection on observables if, even though the assumptions above hold for some unobservables, they do not hold for those unobservables which cause trouble. Suppose some unobservables have full support while others are bounded, and only the bounded unobservables enter into the Y equation. In this scenario, Lemma 1 applies only to the unproblematic unobservables, and therefore, while bias in the effect estimates might increase at extreme propensity scores, it would not explode to infinite values.

Summarizing, misleading results can happen when selection on observables fails when (i) the key unobservables causing endogeneity are bounded, (ii) the effect of those key unobservables on Y is bounded, or (iii) unobservables have a vanishing effect on Y at extreme propensity scores. An important intuitive question to ask to assess (i) and (iii) is whether the same sort of unobservables are likely to matter at extreme propensity scores as at other propensity scores.

Obviously, practitioners who do not know whether selection on observables holds are unlikely to have a completely confident answer to whether any of these is the case or not. However, to the extent that practitioners have some prior about whether these statements would be true conditional on selection on observables failing (e.g. because it clear what potential sources of confounding would be), and to the extent that this prior is separate from their prior about whether selection on observables is true, the practitioner may still find the exercise informative.

Simultaneity The model excludes Y from the D equation and therefore frames omitted variables bias as the key concern about selection on observables. However, it is not without loss of generality of the model that Y was excluded, and in fact simultaneity is a common reason for concern about selection on observables. Nonetheless, it is easy to see that the same intuition holds for the case of simultaneity: The differences between treated and untreated observations in whatever factors besides X explain D grow large at extreme propensity scores, so if Y itself is one of those factors, then differences in Y would be expected to grow large at extreme propensity scores, i.e. Theorem 1 will hold. This conclusion is, of course, subject to analogous restrictions to the ones mentioned above, simply substituting the combination of U and Y where the previous discussion mentions U alone.

4 Implementation

This section describes informal and formal diagnostics for selection on observables based on the implication derived in Section 2.

4.1 Informal visual test

A simple informal method is to simply plot implied treatment effects against propensity scores, and inspect for a smile or frown pattern. The steps are as follows:

1. Estimate the function $P(X)$.

2. Group observations into K bins on the basis of their propensity score.

3. Estimate the regression of interest. Let $\widehat{\beta}$ denote the estimated coefficient vector on the controls X.

4. Calculate the fraction of treated observations for each bin k and denote this as $\widehat{p_k}$. Calculate the difference in the average of $Y^{adj} := Y - X'\widehat{\beta}$ between treated and untreated observations for each bin k and denote this $\widehat{\Delta}_k$.

5. Plot $\widehat{\Delta}_k$ against $\widehat{p}_k$.

6. Check for a smile shape, in which $\widehat{\Delta}_k$ veers upwards as $\widehat{p}_k$ approaches 0 and 1. This indicates $\gamma > 0$, which means omitted variable bias is positive.

7. Check for a frown shape, in which $\widehat{\Delta}_k$ veers downwards as $\widehat{p}_k$ approaches 0 and 1. This indicates $\gamma < 0$, which means omitted variable bias is negative.

Note that this procedure estimates $\Delta(p)$ using a version of the outcome Y that contains some adjustment for the covariates. This is because, unless propensity score bins are extremely narrow, there will generally be some discrepancy in X between treated and untreated observations which would be eliminated if we had conditioned on an exact propensity score rather than a bin of propensity scores. We are interested in observing patterns of model-implied treatment effect heterogeneity, rather than picking up differences in the quality of covariate balance within propensity score bins.

4.2 Formal test

It is also possible to construct a formal test.

While the key implication from Section 2 is of vertical asymptotes in a conditional expectation function, in practice it is hard to directly test this by taking a limit. For one, effective sample sizes are likely to be small and vanishing in the limit, precisely because treated (untreated) observations with propensity scores close to 0 (1) are outliers. For another, taking a limit as propensity scores go to 0 or 1 requires the econometrician e.g. to confidently distinguish between observations with propensities of .02 and .01. This inevitably creates an important, non-transparent role for modeling assumptions; without strong modeling assumptions, sampling error makes it hard for a researcher to make such fine distinctions about propensity scores. Sample sizes for estimating treatment effects are already typically small at extreme propensity scores, precisely because treated (untreated) observations with propensity scores close to 0 (1) are outliers.

The following test accommodates the need for adequate sample size while also not relying too heavily on exact modeling of the propensity score. Suppose we are using regression to estimate the parameters of

$$Y_i = \alpha_0 + X_i'\alpha_1 + \alpha_2 D_i + \epsilon_i,$$

and we would like to test the selection on observables assumption in this regression. The results of Section 2 imply that, due to the role of unobservables, the residual ϵ_i will take large positive (negative) values for treated observations with low propensity scores, and large negative (positive) values for untreated observations with high propensity scores, in the presence of positive (negative) omitted variables bias.

The following procedure allows for a formal test:

1. Estimate propensity scores, e.g. with a logit or probit model.
2. Let $\widehat{P}_i$ denote the estimated propensity score for individual i. Construct
$$W_i = \begin{cases} 1 & \text{if } D_i = 1 \text{ and } \widehat{P}_i < .1 \\ -1 & \text{if } D_i = 0 \text{ and } \widehat{P}_i > .9 \\ 0 & \text{otherwise} \end{cases}$$
3. Regress Y on D, X, and W, i.e. estimate
$$Y_i = \beta_0 + X_i'\beta_1 + \beta_2 D_i + \beta_3 W_i + \nu_i.$$
4. Implement a standard t-test for whether β_3, the coefficient on W, is equal to 0. Reject selection on observables if and only if we reject that β_3 is 0.

Intuitively, this test proceeds by noting that some observations seem "weird." Observations with $W = 1$ have weirdly high unobservables U, while observations with $W = -1$ have weirdly low U. Therefore, W winds up being a proxy for U. The choice of cutoffs (.1 and .9) is a simple rule of thumb motivated by the fact that, in simulations across a variety of distributions for the error term in the D equation, large omitted variable bias effects seem to kick in for observations past this cutoff.

An alternative would to rely more heavily on a formal model of propensity scores, including something like an inverse Mills ratio instead of W.[6] If the model of propensity scores is correct, this alternate approach would have more power. However, there are two main advantages to the simpler construction of W above. First, the assumptions in Section 2 are weaker than the assumptions required for such an approach,

[6] This follows the tradition of "identification by functional form" as seen e.g. in Heckman (1979). For example, Olsen (1980) establishes identification in the case of a normally-distributed unobservable in the D equation without a distributional assumption for the error term in the Y equation.

and the key implication from Section 2 is only related to implied treatment effect heterogeneity at extreme propensity scores. Using an inverse Mills ratio instead would lead the test to focus a substantial part of its attention on observations with intermediate propensity scores, where we might feel that implied treatment effect heterogeneity speaks less persuasively to the plausibility of selection on observables.

The second main advantage relative to such an approach is the interpretability of the result. It is possible that $\beta_3 \neq 0$ even though selection on observables holds, simply due to treatment effect heterogeneity. The test above delivers an estimate of the amount of treatment effect heterogeneity (e.g. if $\widehat{\beta}_3 = .6$, then "treatment effects are estimated to be .6 larger at extreme propensity scores"). Practitioners can decide for themselves based on the details of the application whether this degree of heterogeneity is likely or not, and update their priors accordingly about the plausibility of selection on observables. By contrast, the coefficient on an inverse Mills ratio is not as easily interpretable, and therefore does not allow practitioners to cleanly map their priors about treatment effect heterogeneity into posterior beliefs about selection on observables.

Finally, this test can be implemented quickly by most researchers without any additional software or commands.

4.3 Summary of conditions yielding false negatives and false positives

The test above is based on Theorem 1. As discussed in Section 3, these assumptions may not always hold, and their violation may sometimes lead Theorem 1 not to hold.

Having now developed a formal test, I will briefly summarize the conditions which will lead the test to reject or fail to reject selection on observables.

Rejection A rejection of selection on observables can occur in three ways:

1. **Correct rejection.**
2. **Selection on observables holds, rejection due to sampling error.** Simulations in Section 5 find that the test has appropriate size when the primary model assumptions hold.
3. **Selection on observables holds, rejection due to heterogeneous treatment effects.** In Section 3, we saw that treatment effect heterogeneity can disrupt Theorem 1. The estimated coefficient on W tells both the sign and magnitude of treatment effect heterogeneity that we must be forced to accept in order to believe in selection on observables.

In practice, therefore, researchers who wish to argue against a rejection of selection on observables must argue in favor of treatment effect heterogeneity.

Fail to reject Failure to reject selection on observables can occur three ways:

1. **Correct failure to reject.**
2. **Selection on observables does not hold, failure to reject due to sampling error.** Across various simulations and the example, I find that the coefficient on W is generally on the order of one-half or one-third of the omitted variables bias in the estimate of δ. While this ratio obviously will vary across applications depending on factors like the distributions of U and X, it is at least probably imprudent to draw strong conclusions on the basis of the smile test alone that omitted variables bias is smaller than k if it is not possible to reject that β_3 could be as large as $k/3$.[7]
3. **Selection on observables does not hold, failure to reject due to incorrect assumptions about unobservables.** As discussed in Section 3, an incorrect failure to reject can occur when (i) the unobservables of concern are bounded, (ii) their effect on Y is bounded, or (iii) the effect of unobservables on Y disappears at extreme propensity scores. To the extent that practitioners have a sense of what sort of unobservables might threaten their research design, practitioners are likely to have some prior about these possibilities.
4. **Selection on observables does not hold, failure to reject due to knife-edge case.** As discussed in Section 3, it is possible that treatment effects are heterogeneous in a way that exactly cancels out the pattern of implied treatment effect heterogeneity that results from omitted variables bias. Similarly, the "smirk" data patterns that are labelled as inconclusive in Section 3 can lead to failures to reject in the test if a positive coefficient on W at one extreme is cancelled out by a negative coefficient at the other – though this possibility can be detected with a simple graph. Exact canceling would require quite a coincidence, but approximate canceling to the point where the test does not have power to detect the violation might be realistic in some applications.

5 Simulation Results

I next perform simulations to explore the performance of the formal test in finite samples.

The data-generating process follows the model from Section 1. I use a single control variable $X \sim N(0, 1)$. The function $h(X)$ is set to $.3X$. The coefficients in the Y equation are $\xi = 0$, $\delta = 3$, $\beta = -4$, and I draw V at random from a standard normal as well. I set γ variously equal to 0, .1, and .2 to simulate no omitted variables bias, smaller omitted variables bias, and larger omitted variables bias, respectively.

[7] W will have a very small coefficient relative to the degree of omitted variables bias, for example, if all observations lie, say, at propensity scores extremely close to .1 or .9. In this scenario, even the rule of thumb described above would be insufficiently prudent.

In one version of the simulation, I set U to be normally distributed with mean of 0 and standard deviation of 1. Propensity scores are estimated using a probit. In a second version, to simulate misspecification of the propensity score, I generate U as a mixture of three distributions: two t-distributions with 10 degrees of freedom, one centered on -1.5 and the other on 1.5, each receiving weight of .45 in the mixture, and a χ^2-distribution with 4 degrees of freedom, receiving a .1 weight. For comparability to the normal U case, I then normalize this distribution to have a mean of 0 and standard deviation of 1. This creates a bimodal and skewed omitted variable with a distribution that looks quite different from normal; see Figure 4. Then I estimate propensity scores using a probit, such that the propensity score estimates are misspecified.

I use three sample sizes: 2,000, 10,000, and 50,000 observations. These are unconventionally large sample sizes for a simulation, but they reflect that the method is unlikely to be appropriate for very small samples because the number of observations with $W \neq 0$ will inevitably be less than 10% of the sample (since less than 10% of observations with extreme propensity scores will have the unexpected treatment status), and generally less than that. For normal U, the settings in this simulation produce 29% of observations which a probit estimates to have extreme propensity scores, and 1.3% of all observations with $W \neq 0$. For non-normal U, these numbers are 26% and .8%. Therefore, at the smallest simulation sample size, there are typically only around 20 observations with $W \neq 0$ in a simulation.

I run 10,000 simulations at each combination of sample size and value of γ. The rejection frequencies at each p-value are listed in Table 1. I also report the average coefficient on D when Y is regressed on D and X, and the average coefficient on W when Y is regressed on D, X, and W, where these averages are taken across all simulations with a fixed value of γ.

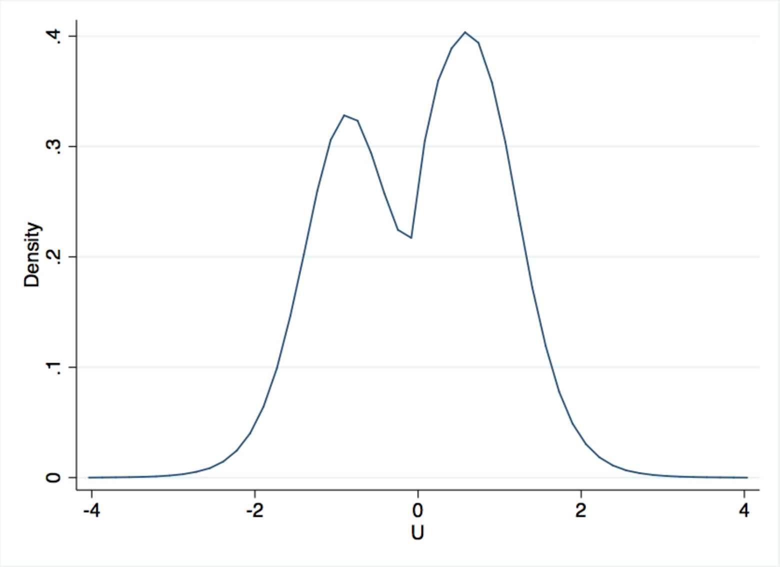
Unsurprisingly, the test has appropriate size – since, when $\gamma = 0$, this is simply a conventional significance test of a variable with a true 0 coefficient. Also unsurprisingly, the power of the test is substantially better with larger samples or more endogeneity to detect. Misspecification of the propensity score does not affect size but reduces power.

6 Application: Incumbency Effects

I demonstrate the formal test using an application from Lee (2008). Lee studies the incumbency advantage in United States House of Representatives elections using a regression discontinuity design, and finds that a Democratic candidate winning a House election in election t leads to an approximately 15 percentage point increase in the Democratic candidate's margin relative to their Republican opponent in the period $t + 1$ election.

Using elections data from Caughey and Sekhon (2011), I implement an OLS specification with a rich set of covariates, but which nonetheless produces a substantially

Figure 4: Distribution of non-normal U



Distribution of U in the second simulation, as estimated with a kernel density plot (Epanechnikov kernel, bandwidth of .0358) using $10,000,000$ simulated observations.

Table 1: Simulation results

		U normal			U non-normal		
N	Level	$\gamma = 0$	$\gamma = .1$	$\gamma = .2$	$\gamma = 0$	$\gamma = .1$	$\gamma = .2$
2,000	$p = .01$	.011	.011	.019	.011	.011	.016
	$p = .05$	.049	.058	.081	.048	.050	.060
	$p = .1$	.096	.114	.139	.100	.100	.114
10,000	$p = .01$	.010	.026	.087	.009	.017	.034
	$p = .05$	.053	.099	.225	.048	.066	.123
	$p = .1$	.099	.171	.330	.098	.118	.203
50,000	$p = .01$	.009	.114	.542	.009	.051	.215
	$p = .05$	.048	.267	.767	.050	.145	.432
	$p = .1$	.099	.380	.852	.096	.231	.556
Average $\widehat{\delta}$		3.00	3.16	3.33	3.00	3.17	3.33
Average $\widehat{\beta_3}$		.00	.06	.11	.00	.04	.09

Entries are fraction of simulations rejecting at the listed significance level in 10,000 simulations. The bottom two rows report the average coefficient on D when regressing Y on D and X (labelled as $\widehat{\delta}$), and the average coefficient on W when regressing Y on D, X, and W ($\widehat{\beta_3}$) across all simulations in that column.

Table 2: Incumbency effects

	$DemMargin_{t+1}$	$DemMargin_{t+1}$
$DemWin_t$	39.67***	34.83***
	(1.56)	(2.43)
W		10.40***
		(3.26)
Controls	Y	Y
N	4,701	4,701

Robust standard errors in parentheses. *** indicates significance at the 0.001 level. Both columns are estimated using controls including Congressional Quarterly rating, incumbent's DW-NOMINATE, Democratic vote shares in previous Congressional and presidential elections, district demographics, party affiliation of state officials, and party advantages in incumbency and candidate experience in time t.

higher incumbency advantage estimate than a regression discontinuity design – approximately 40 points, instead of 15. Here, Y is the Democratic candidate's margin of victory in $t+1$, D is a dummy for whether the Democratic candidate won in t, and X is a vector of covariates including the Democrat's margin in election $t-1$, the recent history of presidential votes in the district, incumbent ideology, voter demographics, party affiliation of state officials, and characteristics of Democratic and Republican nominees in period t. Propensity scores are estimated using a probit model.

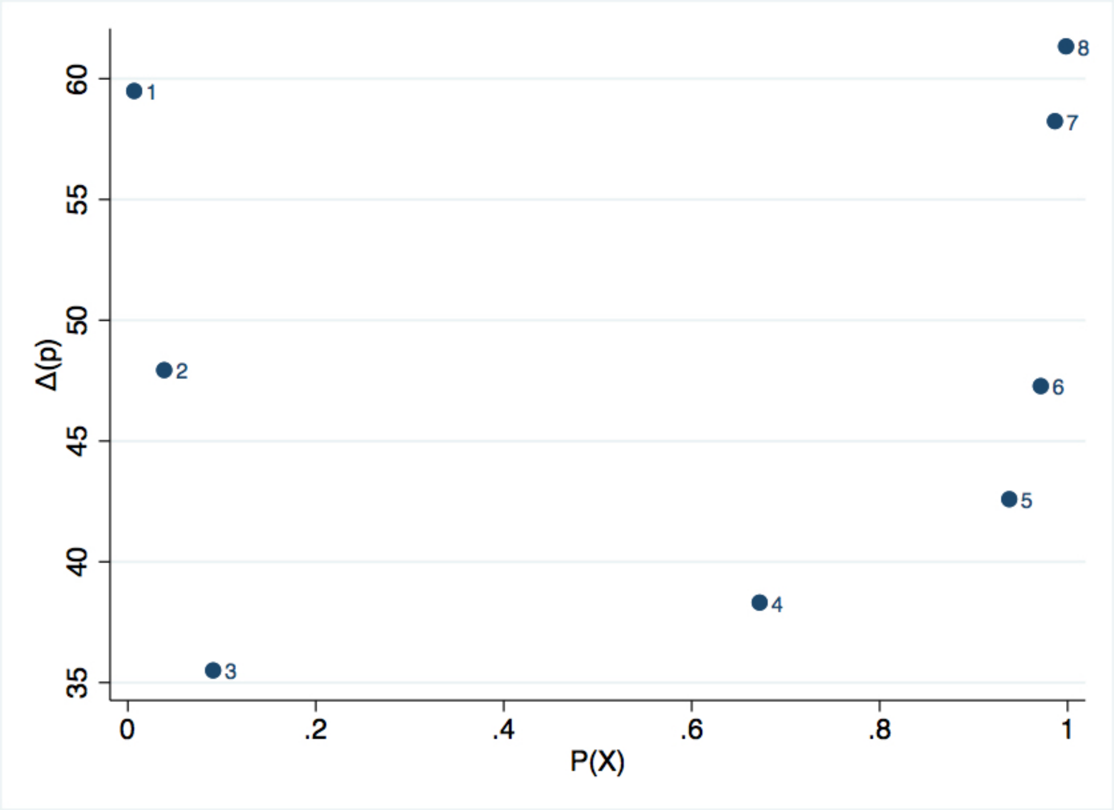
I first implement the informal diagnostic. As illustrated in Figure 5, the results show an upward smile, consistent with OLS estimates being biased upwards.

Next, I implement the formal test. Regression results with and without including W are shown in Table 2. The coefficient on W is estimated to be 10.40, with a standard error of 3.04. This produces a t-statistic of 3.42, and the test rejects selection on observables with a p-value of 0.0006.

An empirical researcher who wished to defend selection on observables would then be required to argue for a pattern of heterogeneity in treatment effects in which incumbency effects are 10 points larger in districts which are very prone to voting for one party.

A researcher who wished to argue for *downwards* omitted variables bias, such that OLS underestimates the effect of incumbency, would need to argue for either (a) greater treatment effect heterogeneity than described above, or (b) this degree of treatment effect heterogeneity (in order to explain the apparent pattern of upwards omitted variables bias) plus disappearing omitted variables bias in extreme

Figure 5: Smile graph: Incumbency effect



Binned scatterplot of estimated difference in mean outcomes, adjusted for covariates, against estimated propensity score. Outcome variable is Democrat's margin of victory in election $t + 1$, treatment variable is Democratic candidate winning in period t election, and controls include Congressional Quarterly rating, incumbent's DW-NOMINATE, Democratic vote shares in previous Congressional and presidential elections, district demographics, party affiliation of state officials, and party advantages in incumbency and candidate experience in time t.

observations.

7 Conclusion

This paper introduces a simple diagnostic for the selection on observables assumption in cases of a binary treatment variable and strong enough covariates to produce propensity scores approaching 0 and/or 1. This diagnostic can be implemented in a few seconds without new software. Simulations and an application support the reliability and practicality of the method for empirical use.

References

[1] Black, D.A. and J.A. Smith (2004). How Robust is the Evidence on the Effects of College Quality? Evidence from Matching, *Journal of Econometrics*, 121, 99-124.

[2] Blundell, R. and J. Horowitz (2007). A Non-Parametric Test of Exogeneity, *Review of Economic Studies*, 74, 1035-1058.

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

[4] Caughey, D. and J.S. Sekhon (2011). Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942-2008, *Political Analysis*, 19(4), 385-408.

[5] Chen, T., Ji, Y., Zhou, Y., and P. Zhu (2018). Testing Conditional Mean Independence Under Symmetry, *Journal of Business and Economic Statistics*, 36(4), 615-627.

[6] Donald, S.G., Hsu, Y., and R.P. Lieli (2014). Testing the Unconfoundedness Assumption via Inverse Probability Weighted Estimators of (L)ATT, *Journal of Business and Economic Statistics*, 32(3), 395-415.

[7] Hansen, L.P. (1982). Large Sample Properties of Generalized Method of Moments Estimators, *Econometrica*, 50(4), 1029-1054.

[8] Hausman, J.A. (1978). Specification Tests in Econometrics, *Econometrica*, 46(6), 1251-1271.

[9] Heckman, J.J. (1979). Sample Selection Bias as a Specification Error, *Econometrica*, 47(1), 153-161.

[10] Lee, D.S. (2008). Randomized Experiments from Non-Random Selection in U.S. House Elections, *Journal of Econometrics*, 142(2), 675-697.

[11] Khalil, U. and N. Yildiz (2017). A Test of the Selection-on-Observables Assumption Using a Discontinuously Distributed Covariate, *working paper.*

[12] Olsen, R.J. (1980). A Least Squares Correction for Selectivity Bias, *Econometrica*, 48(7), 1815-1820.

[13] Rivers, D. and Q.H. Vuong (1988). Limited Information Estimators and Exogeneity Tests for Simultaneous Probit Models, *Journal of Econometrics*, 39, 347-366.

[14] Rosenbaum, P.R. and D.B. Rubin (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects, *Biometrika*, 70(1), 41-55.

[15] Sargan, J.D. (1958) The Estimation of Economic Relationships Using Instrumental Variables, *Econometrica*, 26(3), 393-415.

[16] Young, A. (2020). Consistency without Inference: Instrumental Variables in Practical Application, *working paper.*

Appendix

A Proofs

A.1 Proof of Lemma 1

Since U is uncorrelated with X, then $E(U) = E(U|P(X) = p)$. Furthermore, $E(U|P(X) = p) < E(U|D = 1, P(X) = p)$ for all $p \in (0, 1)$. So

$$\begin{aligned}\Delta_U(p) &= E(U|D = 1, P(X) = p) - E(U|D = 0, P(X) = p) \\ &> E(U) - E(U|D = 0, P(X) = p).\end{aligned} \tag{1}$$

Note that $P(X) = Pr(h(X)+U > 0|X) = Pr(U > -h(X)|X) = 1-F_U(-h(X))$, where F_U is the cumulative distribution function of U. For untreated observations with propensity score $P(X)$, U must be no greater than $-h(X)$, and therefore $E(U|D = 0, P(X) = p)$ must be less than $-h(X)$. Since X and U have full support, then as p goes to 1, $-h(X)$ goes to $-\infty$. This yields that $lim_{p\to1}E(U|D = 0, P(X) = p) = -\infty$.

Combining previous steps, we have

$$\lim_{p\to1} \Delta_U(p) > E(U) - E(U|D = 0, P(X) = p) = \infty,$$

since $E(U)$ is finite. Therefore

$$\lim_{p\to1} \Delta_U(p) = \infty.$$

The proof that $\lim_{p\to0} \Delta_U(p) = \infty$ follows analogously from the facts that $E(U|D = 0, P(X) = p)$ is bounded above by $E(U)$ and $E(U|D = 1, P(X) = p)$ goes to infinity as p goes to 0.

A.2 Proof of Theorem 1

The following equalities hold, respectively, by the definition of $\Delta(p)$, the structural equation for Y, the propensity score theorem (Rosenbaum and Rubin 1983), which states that the distribution of X is independent of D conditional on $P(X)$, and by the definition of $\Delta_U(p)$

$$\begin{aligned}\Delta(p) =&E(Y \mid D = 1, P(X) = p) - E(Y \mid D = 0, P(X) = p) \\ =&E(\xi + X'\beta + \delta D + \gamma U + V \mid D = 1, P(X) = p) \\ &- E(\xi + X'\beta + \delta D + \gamma U + V \mid D = 0, P(X) = p) \\ =&\delta + \gamma\left[E(U \mid D = 1, P(X) = p) - E(U \mid D = 0, P(X) = p)\right] \\ =&\delta + \gamma\Delta_U(p).\end{aligned}$$

Then

$$\lim_{p \to 1} \Delta(p) = \delta + \gamma \lim_{p \to 1} \Delta_U(p),$$

and

$$\lim_{p \to 0} \Delta(p) = \delta + \gamma \lim_{p \to 0} \Delta_U(p).$$

The conclusion follows from Lemma 1.