The Employment Effects of the Minimum Wage: A Selection Ratio Approach to Measuring Treatment Effects

David Slichter*

November 13, 2015

Abstract

This paper studies the employment effects of the minimum wage using a novel empirical strategy which can allow the researcher to identify treatment effects when more than one control group is available but each such control group is imperfect. Expanding on previous researchers who have compared regions which increase the minimum wage with nearby regions which do not change the minimum wage, I compare border counties in which the minimum wage increases to the set of neighboring counties, the set of neighbor-of-neighboring counties, etc. The key innovation is to model the ratio of the bias of these comparisons. The model I select uses the relative similarity of control groups to the treated group on observables as a guide to their relative similarity on unobservables. Crucially, models of this type have a testable implication when there are enough control groups. Using data from the United States, I find that recent minimum wage increases have produced modest or zero disemployment effects for teenagers.

JEL classification: J38, C21, C29

Keywords: Minimum wage, control groups, selection ratio, regression discontinuity

*Department of Economics, University of Rochester. The most recent version of this paper is available here: http://bit.ly/1kLgpGT. I am indebted to my advisor Gregorio Caetano for his support and advice throughout this project, and to Carolina Caetano, Josh Kinsler, Ronni Pavan, and Neşe Yildiz for their generous help and guidance. I am also grateful to Mark Bils, Yongsung Chang, Bin Chen, Richard DiSalvo, Mike Elsby, Umair Khalil, Mike Lovenheim, Nirav Mehta, Ryan Michaels, Sulagna Mookerjee, Ismael Mourifié, and seminar participants at the University of Rochester and the 2015 Econometric Society World Congress for their helpful comments. All remaining errors are my own.

1 Introduction

The employment effects of the minimum wage continue to be a source of contention among economists. For example, a recent poll asked top economists whether "raising the [United States] federal minimum wage to $9 per hour would make it noticeably harder for low-skilled workers to find employment." 34% of the economists agreed, 32% disagreed, and 24% were uncertain (IGM 2014). The lack of consensus is not for lack of effort. Neumark and Wascher (2007) review over 100 empirical studies, and a number of additional papers have appeared in the years since.

Perhaps the most fundamental obstacle to consensus estimates of the effects of the minimum wage is the lack of an obviously appropriate control group which can be used as a counterfactual for regions which change their minimum wage. In the absence of random assignment, researchers have resorted to a variety of measures, including but not limited to econometric adjustments using time series methods (e.g. Kaitz 1970),[1] looking for heterogeneous effects of law changes (e.g. Card 1992a, Katz and Krueger 1992),[2] and using nearby regions as a control group.

This last method of using regional variation as a sort of "natural experiment" was suggested by Card and Krueger (1994), who measured the effects of a minimum wage law change in New Jersey by comparing changes in employment in fast food establishments on either side of the border between New Jersey and Pennsylvania. More recently, two papers by Dube, Lester, and Reich (2010, henceforth DLRa; and 2014, henceforth DLRb) generalize this research design by comparing employment in neighboring pairs of counties where one county experiences a minimum wage increase and the other does not. (In practice, this means comparing counties which lie on opposite sides of a state border.)

The logic in favor of spatially proximate control groups is simple. Labor markets in nearby places are likely to be similar, in both observable and unobservable ways. Therefore using nearby control observations may reduce any bias resulting from unobserved labor market shocks.

However, there are also arguments against the use of this kind of control group. For one, the use of the smaller set of control observations substantially reduces sample size. More fundamentally, though, few researchers believe that Pennsylvania is really a perfect counterfactual for New Jersey. While nearby control observations might be more similar on unobservables, it is not immediately obvious how much more similar they might be.

Some researchers have argued whether or not spatially proximate control observations are likely to be similar to the treated group on *unobservables* by looking at

[1] See Brown, Gilroy, and Cohen (1982) for a more detailed review of early evidence, which tended to concentrate on time series variation because minimum wages were mostly determined at the national level between the passage of the Fair Labor Standards Act in 1938 and the 1980s, and Card and Krueger (1995a) for criticism that this evidence might be contaminated by specification search.

[2] An early example is Lester (1946).

whether they are similar to the treated group on *observable* variables. The idea behind this exercise is that, if nearby regions do experience similar labor market shocks, this would manifest itself as similarity on both observed and unobserved dimensions. DLRa, DLRb, and Allegretto, Dube, and Reich (2011, henceforth ADR) all offer evidence that nearby control regions are more similar on observables than distant control regions. However, none of these papers establishes that there is a perfect match on observables; they each settle for arguing that the match on observables is better than when more distant control regions are used. Neumark, Salas, and Wascher (2014a; henceforth NSW) argue that the difference in observables between treated border counties and their neighbors on the other side of the border is in fact substantial.

In this paper, I introduce a new identification strategy which can identify the effects of the minimum wage even if using border counties does not eliminate differences on unobservables. I am able to relax this assumption by using additional control groups composed of the set of neighbors-of-neighbors of treated counties, the set of neighbors-of-neighbors-of-neighbors of treated counties, etc. The identifying assumption will invoke the logic of the previous paragraph: that control groups which are more similar to the treated group on observables are more similar on unobservables too.

My main results suggest that recent minimum wage increases in the United States have produced only modest disemployment effects, if any, in the year after implementation. The point estimate is that an increase in the minimum wage by 10% decreases teen employment by only .4% a year later. This estimate is not significantly different from zero, but the confidence interval excludes many estimates in the minimum wage literature (see Neumark and Wascher 2007). As expected, I also find that the minimum wage leads to an increase in the monthly earnings of employed teenagers. The earnings effect is estimated to be larger than the disemployment effect, in contrast with some other papers which suggest that the minimum wage decreases average earnings of low-skilled workers (see Neumark and Wascher 2008).

These results are robust to concerns about cross-border employment spillovers. The results also do not appear to be attenuated by anticipatory responses, and the estimates are not sensitive to small violations of the identifying assumption.

The identification strategy is sufficiently general that it may be useful in the future to other researchers who face a choice of imperfect control groups. Examples include comparing a treated student to untreated classmates, schoolmates, other children in the same district, etc.; comparing competition winners to runners-up, second runners-up, etc.; or comparing an observation in time t to the same observation in $t-1$, $t-2$, etc. Appendix E contains an empirical example illustrating that the method can be used to estimate treatment effects in a regression discontinuity (RD) design, including for populations away from the discontinuity.

The identification argument and the key identifying assumption are described in detail in Section 3. First, I show that, when a treated group can be compared to more than one control group, the average treatment effect on the treated is identified whenever the *relative* bias of different control groups is identified. (*Bias* refers to the difference between the outcomes for the control group and the outcomes that the treated group would have had without being treated; this is also sometimes called *selection*.) As an example, if the neighbors had an employment gain of 4% relative to the treated group, the neighbors-of-neighbors had a gain of 6% relative to the treated group, and we knew that the bias of the neighbors was half the bias of the neighbors-of-neighbors, then we could infer that the effect of treatment was to depress employment by 2%.

Next, I introduce an identifying assumption which models the similarity of each control group to the treated group in terms of its similarity on observables, and show that this assumption is a model of the relative bias of the control groups. Therefore this assumption identifies the average treatment effect on the treated.

I additionally show that models of the relative bias of control groups have a testable implication when the researcher has enough control groups. The testable implication results from a standard overidentification argument: If we can identify the treatment effect with only two control groups, and we have more than two control groups, then any pair of control groups must lead to the same treatment effect estimate. So, in the example given previously, if we found that a third control group had a gain of 10% in employment relative to the treated group, then the model must imply that the bias from this third control group is twice the bias from the second control group and four times the bias from the first.

Since the identifying assumption for studying the minimum wage is a model of the relative bias, it also contains a testable implication. This implication can be illustrated graphically as well as tested formally. This test does not have the power to reject all violations of the identifying assumption; however, I show that the identifying assumption is equivalent to two other assumptions, one of which is fully testable. The assumption which is not testable is that a control group with the same average value of the observables as the treated group would also have the same average value of the unobservables.

This paper fits into two large literatures. First, an enormous body of work has studied the effects of the minimum wage, especially on employment, hours, labor market fluidity, poverty, and human capital accumulation. Comprehensive reviews are provided by Brown et al. (1982), Card and Krueger (1995b), Neumark and Wascher (2007), and Neumark and Wascher (2008).[3] My contribution is to measure effects using a spatial research design while still allowing the possibility that neighboring regions are not well-matched on unobservables.

Second, this paper fits into a literature on measuring treatment effects when no perfect control group is available. Various prior papers have used subjectively appealing control groups to identify treatment effects while simply accepting the possibility

[3] Additional papers in the period since 2008 include but by no means are limited to Addison et al. (2009), Sabia (2009), Brochu and Green (2013), Neumark and Wascher (2011), Sen, Rybczynski, and Van De Waal (2011), Sabia, Burkhauser, and Hansen (2012), Giuliano (2013), Gorry (2013), Matsudaira (2014), Sen and Ariizumi (2013), Ropponen (2010), Bárány (2015), and Yannelis (2014).

that unobservable differences remain (e.g. Currie and Thomas 1995, Hagedorn, Karahan, Manovskii, and Mitman 2013). While the applicability of the key assumptions may vary depending on context, the identification strategy used here may allow future researchers to consistently estimate treatment effects instead. Econometric methods such as changes-in-changes (Athey and Imbens 2006), interactive fixed effects (Bai 2009), and synthetic controls (Abadie and Gardeazabal 2003, Abadie, Diamond, and Hainmueller 2010) offer useful alternatives for measuring treatment effects in a panel context when common trends do not appear to hold. This paper differs from them in utilizing information about unobservables that might be captured by the membership of control observations in subjectively appealing control groups. There is also a substantial literature on spatial econometrics (see LeSage and Pace 2009 for an introduction);[4] this paper's approach is distinguished through its focus on control groups in the aggregate – a focus which also allows the selection ratio method to tackle problems such as the RD example which are outside the domain of spatial econometrics. The literature on RD estimation for populations away from the discontinuity is discussed in Appendix E. Altonji and Mansfield (2015) also provide a model in which observables can proxy for unobservables when observations are aggregated, though they consider a different type of aggregation. Finally, Altonji et al. (2005) also offer an assumption in which observables are assumed to be informative about features of the unobservables. However, while I follow their definitions of observables and unobservables, the content of their assumption is unrelated to mine; in short, they make an assumption about the magnitude of a regression coefficient, and make no reference to the possibility of using more than one control group, while I make an assumption about how the mean of unobservables varies across control groups, and no assumption about the magnitude of any coefficient. An example in Appendix C illustrates this difference.

The paper proceeds as follows. Section 2 gives background information on the data and some related literature. Section 3 introduces the econometric model and explains the assumption required for identification and estimation. Section 4 presents the results. Section 5 discusses the results. Section 6 concludes.

2 Background and Data

I use data from the Quarterly Workforce Indicators (QWI), which began in 1990 in four states and by 2004 achieved a sample of all of the contiguous 48 states except for Massachusetts. Observations are counties. Data is quarterly and includes information by county on employment, average wages, and job flows by age bracket.

I will refer to a county i as being *treated* in time t if, at some point during quarter t, the minimum wage increases. All other counties are therefore referred to as *untreated* or *control* counties. I use the term *border* counties to refer to the set of

[4] See also LeSage (2008) for a brief introduction, and Gibbons and Overman (2012) for critiques of causal inference in spatial models.

treated counties which neighbor an untreated county and untreated counties which neighbor a treated county. Note that a county i can be treated in one quarter and untreated in another; in fact, every county experiences a minimum wage increase at some point in my sample.

The magnitude and timing of minimum wage law changes are shown in Table 1. We can see that the minimum wage increases by an average of approximately 10% in treated counties. Treated counties which neighbor an untreated county experience the same average increase in the minimum wage, but are less likely to experience their increase due to a national minimum wage law. (Minimum wage increases can come from changes to state-specific laws or national laws; national laws are binding for most states, so it is not surprising that a low fraction of counties which are treated by national law changes would border a county which is not.) The dispersion of increases is also similar for border and non-border treated counties. In the data, there are six instances in which a state increased the minimum wage by more than 20% in one quarter, but no instances in which a state increased the minimum wage by more than 30%. There is also one instance in which a state decreased the minimum wage, which I do not count as an instance of treatment.

The mean values of some key variables are given in Table 2. Standard deviations are listed in parentheses. The additional columns show the descriptive statistics for the populations of treated and control observations, including populations restricted to border counties.

Similarity on observables At the population level, we can see that treated and control border counties resemble each other more than treated and control observations do in general. This suggests that perhaps border counties might be more similar on unobservables as well. This argument has been made in several previous papers as well.

DLRb consider the average absolute difference on various observables for pairs of counties. When the pair consists of a treated border county and an untreated county contiguous to that same treated county, the average absolute difference on each observable variable is smaller than when the treated border county is compared to a non-contiguous untreated county. (Note that this non-contiguous untreated county might be contiguous to a different treated county.) However, the reduction in the average absolute difference is modest; there is no covariate for which DLRb report that the absolute difference for contiguous counties is less than half the size on average of the absolute difference for non-contiguous counties.

ADR document that states which change their minimum wage tend to have experienced recent declines in employment, and that these recent declines are better matched by nearby states than by distant states. Allegretto et al. (2013) provide additional arguments along the same lines.

However, NSW point out that the choice to use only border counties is the same as placing a weight of 1 on neighboring control observations and a weight of 0 on all non-neighboring observations, which they argue is unreasonable on the grounds that many

Table 1: Characteristics of treated counties

Variable	All treated	Border treated
Change in log of min wage	.096 (.051)	.100 (.065)
Dummy: Law changes in 1st month of quarter	.48 (.50)	.70 (.46)
Dummy: Law changes in 2nd month of quarter	.39 (.49)	.24 (.43)
Dummy: Law change in 1st quarter	.26 (.44)	.50 (.50)
Dummy: Law change in 2nd quarter	.06 (.24)	.08 (.27)
Dummy: Law change in 3rd quarter	.58 (.49)	.37 (.48)
Dummy: Treated due to federal law change	.62 (.49)	.26 (.44)
Number of observations	14,800	2,328

Entries present sample means with standard deviations reported in parentheses.

Table 2: Distribution of key variables

Variable	All	All treated	Treated border	All control	Control border
Log of total employment,	5.80	5.96	6.10	5.77	6.08
teen workers	(1.61)	(1.60)	(1.64)	(1.61)	(1.61)
Log of avg. monthly	6.12	6.21	6.17	6.11	6.16
earnings, teen workers	(0.29)	(0.28)	(0.27)	(0.29)	(0.29)
Gain in log emp.,	-.005	-.044	-.079	.001	-.061
$t-2$ to $t-1$	(.230)	(.208)	(.206)	(.233)	(.198)
Gain in log emp.,	.055	.019	.102	.061	.125
$t-3$ to $t-2$	(.242)	(.241)	(.267)	(.242)	(.264)
Gain in log emp.,	-.023	-.063	-.056	-.016	-.058
$t-4$ to $t-3$	(.206)	(.179)	(.176)	(.201)	(.201)
Gain in log emp.,	.004	.120	.062	-.015	.032
$t-5$ to $t-4$	(.227)	(.241)	(.245)	(.219)	(.246)
Gain in log emp.,	-.006	-.043	-.079	0.001	-.059
$t-6$ to $t-5$	(.227)	(.206)	(.198)	(.229)	(.194)
Gain in log emp.,	.056	.028	.108	.061	.121
$t-7$ to $t-6$	(.240)	(.236)	(.257)	(.240)	(.252)
Gain in log emp.,	-.023	-.068	-.057	-.015	-.054
$t-8$ to $t-7$	(.205)	(.177)	(.178)	(.208)	(.195)
Gain in log emp.,	.002	.112	.058	-.016	.031
$t-9$ to $t-8$	(.226)	(.240)	(.242)	(.218)	(.251)
Number of observations	96,828	13,522	2,218	83,306	2,365

Entries present sample means with standard deviations reported in parentheses. Observations with missing lags of employment are dropped for this table. Log employment gain between t and t' is defined as the natural log of teen employment at the start of quarter t minus the log of teen employment at the start of t'. Total employment is defined as employment at the start of the period. The first two variables are the averages within the period in which treatment occurs. The set of "all control" observations includes only control counties during a quarter in which some county is treated.

non-neighboring observations are much better-matched on observables. Neumark, Salas, and Wascher (2014b) argue that, using a synthetic control method to find appropriate counterfactuals for each treated observation, the weight placed on direct neighbors of each treated county is not unusually high. Since the synthetic control weightings are determined by observables (including past patterns of employment), this statement is equivalent to arguing that contiguous counties are not especially well-matched on observables.

An important caveat to the argument of some of these papers is that what really matters for recovering reasonable estimates is the comparability of the control *group* to the treated *group*. This is a weaker requirement than the requirement that every individual treated *observation* have a well-matched counterpart in the control group. As a metaphor, houses on one side of a street can collectively be like houses on the other side of the street, even if individual houses are often quite different from the individual house across the street. This caveat helps to explain why I will find a greater relative similarity of nearby control *groups* than these papers find for nearby control *observations*.

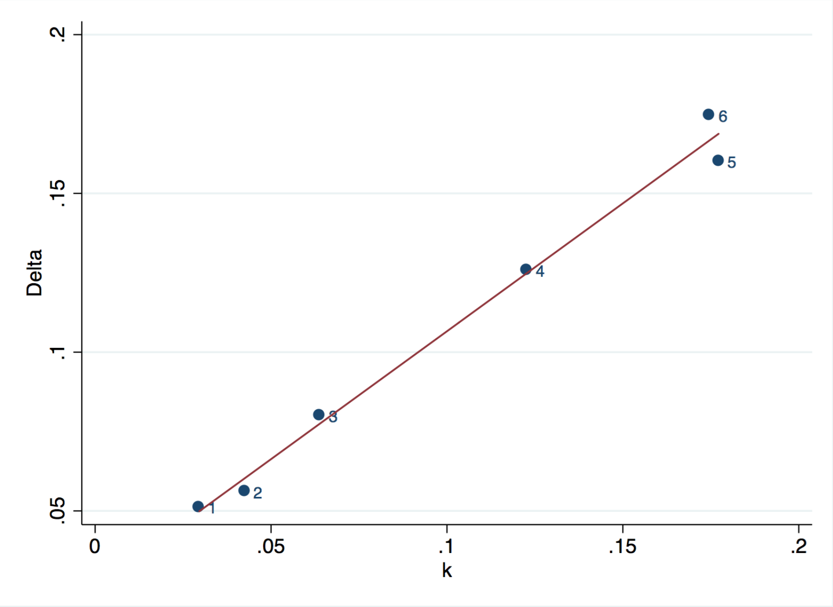
It bears repeating that, in each of these papers, the arguments being presented are not fundamentally about the observables; if we were only concerned about differences on the observables, it would suffice to control for those differences. The more important argument is whether border counties are similar on *un*observables, with greater similarity on observables suggesting greater similarity on unobservables.

Therefore, the conclusions which are suggested from the comparison of descriptive statistics are that (a) if there are unobservable confounders, they might be better matched in a comparison of border counties, since the observables are better matched, and (b) if there are unobservable confounders, they are likely not exactly matched in a comparison of border counties, since the observables are not exactly matched.

Other control groups Table 2 shows that the difference in the average values of recent trends in employment are more similar when comparing treated and control border counties than when comparing all treated observations to all control observations. We might also be interested in how this similarity evolves as we allow control observations to be taken from slightly greater distances.

One way to represent the evolution of this similarity is to plot differences on an index of observables by control group. In Figure 1, each point represents a comparison of the border treated counties to a control group. The point labeled 1 represents the comparison of this treated group to neighboring control observations, the point labeled 2 represents the comparison of the same treated group to neighbors-of-neighboring control observations (not including those observations which are already in the first control group), etc. The x-coordinate of each point is the difference on an index of observables between the treated group and that control group, and the y-coordinate is the average difference between the treated group and that control group on a variable $Sep4_{it} := ln(Sep_{i,t+4}) - ln(Sep_{i,t-1})$, where Sep_{it} is the number of separations for teen workers in county i and period t. That is, $Sep4$ is a measure

Figure 1: Increase in log of separations, $t-1$ to $t+4$



of the growth in separations which happens between the quarter right before t and one year later – a plausible effect of minimum wage laws (e.g. DLRb). The index of observables takes the eight previous quarters of change in $ln(Sep)$ and weights them according to their coefficient in a regression of $Sep4_{it}$ on those lags and time fixed effects.

Three things stand out. First, consistent with intuition, we can see that the nearer control groups have x-coordinates closer to 0, indicating that they are more similar on the index of observables. Second, while the nearer control groups are substantially better than the more distant ones, they are still different from the treated group. Third, we can see that the points in this graph appear to be well-approximated by a line. Interestingly, this linear relationship reappears as well for variables other than separations.

Section 3 develops a theory which explains this linearity and exploits it to identify treatment effects. This theory will allow for the identification of the average treatment effect on the treated even when there are unobservable differences between the treated group and each control group.

3 Identification

This Section describes the econometric model and explains the identifying assumption. The structure of the argument in this Section is as follows. First, we will see that, when more than one control group is available, then the average treatment effect on the treated is identified under any assumption which identifies the relative bias of the control groups. Second, I will present an assumption that allows identification of the relative bias of control groups, and therefore of the average treatment effect on the treated. Third, I discuss the testable implication of the identifying assumption. Finally, we will see that the identifying assumption produces moment conditions which allow for the use of the generalized method of moments (GMM) for estimation of the average treatment effect on the treated.

3.1 The ratio of biases identifies ATT

Let D be a binary random variable representing treatment status (e.g. equal to one if a county has increased its minimum wage) and let Y be the outcome of interest (e.g. changes in teen employment). Following the notation of the treatment effects literature (e.g. Imbens and Angrist 1994), let $Y_i(1)$ be the potential outcome for observation i if i is treated, and let $Y_i(0)$ be i's potential outcome if untreated. That is,

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0).$$

Assume Y and D are observed. Then $Y_i(1)$ is observed when $D_i = 1$ and $Y_i(0)$ is observed when $D_i = 0$. Therefore we can identify the difference in means between treated observations and untreated observations:

$$E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0).$$

A standard decomposition (e.g. Angrist and Pischke 2008) is to break this difference in means into the sum of the average treatment effect on the treated (ATT) and selection:

$$\begin{aligned}\underbrace{E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0)}_{\text{difference in means}} &= \underbrace{E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 1)}_{\text{ATT}} \\ &\quad + \underbrace{E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0)}_{\text{selection}}.\end{aligned}$$

I denote the difference in means as Δ, average treatment on treated as ATT, and selection (also called bias) as B to simplify the notation of the previous equation:

$$\Delta = ATT + B.$$

Multiple control groups Now suppose that the control observations belong to two different control groups, such that each observation i with $D_i = 0$ is a member of either control group 1 or control group 2.

For control groups $g \in \{1, 2\}$, define Δ_g and B_g as follows:

$$\Delta_g := E(Y_i(1) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0, i \in g)$$

$$B_g := E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0, i \in g).$$

Notice that ATT does not need to be indexed by group. This is because the definition of ATT does not condition on the distribution of potential outcomes for observations with $D_i = 0$.

The following system of equations holds for control groups 1 and 2:

$$\Delta_1 = ATT + B_1$$
$$\Delta_2 = ATT + B_2.$$

By defining $k_2 = \frac{B_2}{B_1}$, we then get the following system of equations:

$$\Delta_1 = ATT + B_1$$
$$\Delta_2 = ATT + k_2 B_1.$$

Δ_1 and Δ_2 are identified. Suppose for the moment that k_2 is identified. Then the above system of two equations has only two unknowns (ATT and B_1), and we can solve for those unknowns (provided $k_2 \neq 1$). Therefore, identifying the selection ratio k_2 is sufficient to identify ATT and B_1.

Now suppose that there are more than two control groups. Following the same definitions of Δ_g and B_g, for $m > 2$ control groups, we have the system of equations

$$\begin{aligned}
\Delta_1 &= ATT + B_1 \\
\Delta_2 &= ATT + k_2 B_1 \\
&\vdots \\
\Delta_m &= ATT + k_m B_1,
\end{aligned}$$

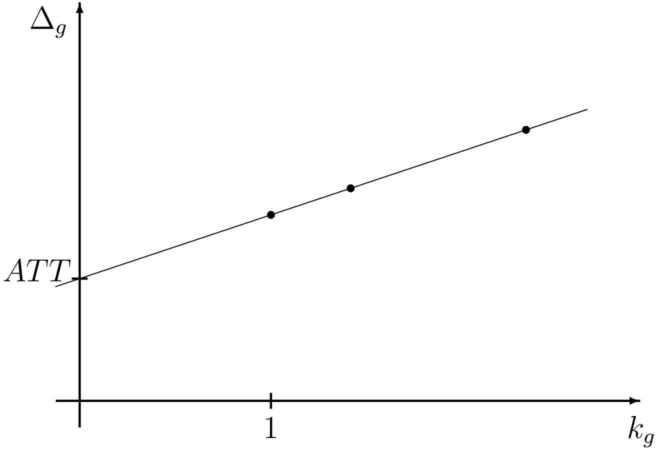
where $k_g := \frac{B_g}{B_1}$.

If k_g is identified for each g, then any two of these equations would suffice to identify ATT, and therefore ATT and B_1 are overidentified.

Naturally the task of identifying the selection ratio k_g for each g will require additional assumptions. However, such assumptions have a testable implication: Any subset of the control groups should produce the same estimate of ATT, up to sampling error.

In the results (Section 4), in addition to running a statistical test of overidentifying restrictions, I will also illustrate the overidentification test graphically. Notice that the above system of equations implies that Δ_g is a linear function of k_g, with a slope of B_1 and intercept of ATT. Therefore, any model of k_g must produce values of k_g which are exactly linear with respect to Δ_g (up to sampling error). Figure 2 illustrates this point.

Figure 2: Overidentified



Each point represents one control group g, with coordinates (k_g, Δ_g). For group 1, $k_1 = \frac{B_1}{B_1} = 1$.

3.2 Modeling the selection ratio

Suppose that the true structural model is

$$Y = \alpha_0 + \delta_0 D + X\beta_0 + V,$$

where Y is the outcome of interest, D is treatment, X is a vector of covariates possibly including lags and fixed effects, and V is unobserved.

We are interested in learning the value of δ_0, which represents the average treatment effect on the treated. For simplicity, δ_0 is not indexed by i, despite the use of potential outcomes notation in the previous section, but the results which follow all hold when the structural coefficient is in fact δ_{0i} and $\delta_0 = E(\delta_{0i} \mid D_i = 1)$, i.e. δ_0 is the average treatment effect on the treated.

One obstacle to identifying δ_0 is that V might be correlated with D and X. To capture this potential correlation, we can write V in terms of D and X. For simplicity, we begin by assuming a linear conditional expectation function:

$$E(V \mid D, X) = \alpha_V + \delta_V D + X\beta_V.$$

Appendix A describes the more general case.

We can then define $\eta = V - E(V \mid D, X)$ such that η is mean-zero and uncorrelated with D or X, and so

$$V = \alpha_V + \delta_V D + X\beta_V + \eta.$$

Therefore, the conditional expectation of Y is

$$\begin{aligned} E(Y \mid D, X) =& \alpha_0 + \delta_0 D + X\beta_0 + E(V \mid D, X) \\ =& (\alpha_0 + \alpha_V) + (\delta_0 + \delta_V)D + X(\beta_0 + \beta_V). \end{aligned}$$

and we can write

$$Y = (\alpha_0 + \alpha_V) + (\delta_0 + \delta_V)D + X(\beta_0 + \beta_V) + \eta,$$

with the mean-zero error term η being uncorrelated with D and X.

Define the unobservable $U = \alpha_V + \delta_V D + \eta$. That is, U is the same as V, but removing the portion $X\beta_V$ which is already captured in controlling for X. An analogous unobservable is used in Altonji, Elder, and Taber (2005). We can write

$$Y = \alpha_0 + \delta_0 D + X\beta + U,$$

where $\beta := \beta_0 + \beta_V$.

To simplify the notation, let

$$B_g^X := E(X\beta \mid D_i = 1) - E(X\beta \mid D_i = 0, i \in g)$$

and let

$$B_g^U := E(U \mid D_i = 1) - E(U \mid D_i = 0, i \in g).$$

That is, B_g^X is a measure of the similarity of group g to the treated group on an index of observables, and B_g^U is a measure of the similarity of group g to the treated group on the unobservable U.

Then we can write the relationship between these two quantities as a new assumption, which we will later see identifies the selection ratio and therefore the average treatment effect on the treated δ_0.

Assumption 1. *(Proportional Unobservables)* For some constant c and for all control groups g, $B_g^U = cB_g^X$.

To help understand the content of Assumption 1, we can equivalently write it as two intermediate assumptions, which will be discussed in detail.

Assumption 1a. *(Linearity of Unobservables)* For any two control groups g and g' and some constant c,

$$E(\eta \mid D_i = 0, i \in g) - E(\eta \mid D_i = 0, i \in g') = c[E(X\beta \mid D_i = 0, i \in g) - E(X\beta \mid D_i = 0, i \in g')]$$

Assumption 1b. *(Intercept)* For some control group g and the same constant c as in Assumption 1a,

$$\delta_V - E(\eta \mid D_i = 0, i \in g) = c[E(X\beta \mid D_i = 1) - E(X\beta \mid D_i = 0, i \in g)]$$

Theorem 1. *Assumptions 1a and 1b together are equivalent to Assumption 1.*

Proof. See Appendix A. □

The following three remarks explain the content of Assumptions 1a and 1b.

Remark 1 (Assumption 1a)**.** Assumption 1a states that, across control groups, the group-level average of the unobservable η is a linear function of the group-level average of observables. It can equivalently be stated as assuming that the group-level average of the unobservable U is a linear function of the group-level average of observables; or, as the assumption that B_g^U is a linear function of B_g^X. See Appendix A for details. Assumption 1a is testable, for reasons which will be described shortly.

Assumption 1a is related to the implicit assumption made by the previous minimum wage literature that control groups which are more similar to the treated group on observables are likely to be more similar to the treated group on unobservables too. However, Assumption 1a takes a stronger stance in two respects. First, it converts the observables to a single index. This is broadly consistent with two tendencies in the literature: The "observables" are discussed as a single object, and balance on observables is taken to be more important for observables which seem prominently related to Y (see e.g. DLRa, DLRb, NSWa, NSWb, ADR). The decision to aggregate the observables using regression coefficients also follows Altonji, Elder, and Taber (2005).

The second respect in which Assumption 1a takes a stronger stance is that it assumes the functional form by which differences on the index of observables map to differences on unobservables. A linear functional form is perhaps the simplest relationship we could postulate, but it is also somewhat arbitrary and not obvious ex ante. Appendix C contains an example model which would generate this linearity, but there is no reason to expect linearity to hold in general across all possible applications of this method. However, just as there exist both informal visual and formal econometric tests of functional form choices in regressions, this functional form assumption can be tested both informally (through a graph) and formally whenever there are more than two control groups.

Remark 2 (Assumption 1b)**.** Assumption 1b imposes a single point that the line of fit between group-level observables and group-level unobservables must pass through. Unlike Assumption 1a, Assumption 1b will not be testable. However, as I will discuss momentarily, it follows more closely from previous arguments used in the literature, and substitutes for less plausible identifying assumptions which were previously used.

Assumption 1b imposes that any control group g with $B_g^X = 0$ would have bias $B_g = 0$ (and also $B_g^U = 0$); see Appendix A for a proof. That is, as differences on observables between control groups and the treated group vanish to zero, so too do differences on unobservables.

In the minimum wage context, this assumption follows from the arguments made in the previous literature and discussed in Section 2 in which control groups are treated as suspect for being different on unobservables if they fail a balancing test (i.e. a test of equality of observables compared with the treated group) and are treated as likely to provide a reasonable counterfactual (i.e. $\Delta_g = ATT$) if they do pass a balancing test. Any group which passed a balancing test would have $B_g^X = 0$, so these previous arguments imply that a group g with $B_g^X = 0$ would have $B_g = 0$.

In making this assumption, we are able to relax assumptions which have previously been used in the minimum wage literature. Comparing with studies that use border counties, we can relax the assumption that there are no differences on unobservables when using a group which has $B_g^X \neq 0$. This seems clearly less plausible in the minimum wage setting than Assumption 1b, which only requires that there are no differences on unobservables when $B_g^X = 0$.

Remark 3 (Example: Selection on observables)**.** Note that

$$\begin{aligned}E(U \mid D = 1) =& E(\alpha_V + \delta_V D + \eta \mid D = 1)\\ =& \alpha_V + \delta_V + E(\eta \mid D = 1)\\ =& \alpha_V + \delta_V + E(\eta \mid D = 0)\\ =& \delta_V + E(U \mid D = 0)\end{aligned}$$

where the middle step holds because η was constructed to be uncorrelated with D. Therefore we have that $\delta_V = 0$ is equivalent to $E(U \mid D = 1) = E(U \mid D = 0)$, which is a case of selection on observables.

This is possible under two scenarios. First, it could be the case that $E(U \mid D_i = 0, i \in g)$ varies across control groups, but that the composition of control groups is such that the average across all control observations is coincidentally exactly equal to $E(U \mid D = 1)$.

The second, more plausible, possibility is that $E(U \mid D_i = 0, i \in g) = E(U \mid D = 1)$ for all groups g. This implies that $B_g^U = 0$ for all g. Then $B_g^U = cB_g^X$ for $c = 0$, and thus Assumption 1 holds.

Assumption 1 also allows cases where selection on observables does not hold, i.e. when $c \neq 0$. This can occur when both $X\beta$ and the unobservable η are correlated with group. This may seem counterintuitive, since η was defined to be uncorrelated with X, but there is no contradiction in two variables each being correlated with a third while not being correlated with each other. In fact, any argument for a control group's similarity to the treated group on unobservables which is based on similarity on observables implicitly appeals to this possibility. Appendix C contains an illustrative theoretical example in which Assumption 1 holds but selection on observables does not. In addition, Appendix E contains an empirical example using a regression discontinuity design in which selection on observables is rejected but Assumption 1 is not.

3.3 Identification and testability theorems

The claim of identification is completed by showing that Assumption 1 is a model of the relative bias of different control groups.

Theorem 2. *If Assumption 1 holds, then for any g and g' such that $B_{g'} \neq 0$, we have $\frac{B_g}{B_{g'}} = \frac{B_g^X}{B_{g'}^X}$.*

If Assumption 1 holds and we observe two control groups g and g' such that $B_g \neq B_{g'}$, then ATT is identified.

Proof. See Appendix A. □

Testability Recall that the potentially testable implication of selection ratio models was that the true Δ_g and k_g are by definition exactly linear, and therefore any model of k_g must reproduce this linearity (up to sampling error). Under Assumption 1, B_g^X is the model of relative bias, so intuitively it should therefore be linear with respect to Δ_g. In fact, we can be precise about what part of Assumption 1 is responsible for the linearity of B_g^X with respect to Δ_g:

Lemma 1. *Assumption 1a holds if and only if B_g^X and Δ_g are exactly linear.*

Proof. See Appendix A. □

Theorem 3. *Assumption 1 implies that B_g^X and Δ_g are exactly linear.*

Proof. Assumption 1 implies Assumption 1a, which implies that B_g^X and Δ_g are exactly linear. □

The statistical test used to test Assumption 1a will be discussed in the section on the GMM estimation of the model.

Remark 4 (Relaxing linearity of $E(V \mid D, X)$)**.** In introducing the model, I assumed for convenience that $E(V|D, X)$ was linear in D and X. Under regularity conditions, the previous theorems all follow even if this is not the case. See Appendix A for details.

3.4 GMM estimation and testing

Assumption 1 implies a set of moment conditions which can then be used to solve for the coefficient of interest, as well as to test the model when enough control groups are available.

Since Assumption 1 assumes that $B_g^U = cB_g^X$ for some c; since $B_g = B_g^X + B_g^U$; and since $\Delta_g = ATT + B_g$; then we have

$$\Delta_g = ATT + (1 + c)B_g^X,$$

where ATT and c are unknown, but Δ_g is the solution to moment conditions, B_g^X is a solution to moment conditions given β, and β is the solution of moment conditions.

In particular, β is the coefficient on X in a regression of Y on D and X, and is therefore the solution to the following usual OLS moment conditions:

$$\begin{aligned}E(Y - \alpha - \delta D - X\beta) =&0\\ E[D(Y - \alpha - \delta D - X\beta)] =&0\\ E[X(Y - \alpha - \delta D - X\beta)] =&0.\end{aligned}$$

We will not make further use of the other coefficients α and δ, but they must be included in order to estimate β.

Next, $\Delta_g = E(Y \mid D = 1) - E(Y \mid D_i = 0, i \in g)$. That is, Δ_g is the population regression coefficient in a regression of Y on D in the population of i such that either $D_i = 1$ or $D_i = 0, i \in g$. Let I_g be a dummy which is equal to 1 for any i such that $D_i = 1$ or $D_i = 0, i \in g$, and 0 otherwise. In the restricted population with $I_g = 1$, Δ_g solves the following moment conditions:

$$\begin{aligned}E(Y - \alpha_g^{\Delta} - \Delta_g D) =&0\\ E[D(Y - \alpha_g^{\Delta} - \Delta_g D)] =&0,\end{aligned}$$

where once again α_g^{Δ} is a parameter that we will not reuse. We can then write moment conditions which hold for the unrestricted population, and not just those with $I_g = 1$. For each group g, we have

$$\begin{aligned}E[I_g(Y - \alpha_g^{\Delta} - \Delta_g D)] =&0\\ E\left(I_g[D(Y - \alpha_g^{\Delta} - \Delta_g D)]\right) =&0.\end{aligned}$$

Finally, $B_g^X = E(X\beta \mid D = 1) - E(X\beta \mid D_i = 0, i \in g)$, and is therefore the regression coefficient on D if $X\beta$ were regressed on D in a population with $I_g = 1$. Following the exact same logic as with Δ_g, we have the following moment conditions for each group g:

$$\begin{aligned}E[I_g(X\beta - \alpha_g^{X\beta} - B_g^X D)] =&0\\ E\left(I_g[D(X\beta - \alpha_g^{X\beta} - B_g^X D)]\right) =&0\end{aligned}$$

Armed with moment conditions for B_g^X and Δ_g, we can now impose the assumption that they are linearly related to each other across groups g by imposing that $\Delta_g = ATT + aB_g^X$ for some slope a and intercept ATT which will be estimated. Therefore we can collect the previous moment conditions, substituting $ATT + aB_g^X$ for Δ_g (where $a = 1 + c$), to get the full set of moment conditions:

$$\begin{aligned}E(Y - \alpha - \delta D - X\beta) =&0\\ E[D(Y - \alpha - \delta D - X\beta)] =&0\\ E[X(Y - \alpha - \delta D - X\beta)] =&0\end{aligned}$$

and, for each g,

$$\begin{aligned}E[I_g(Y-\alpha_g^{\Delta}-(ATT+aB_g^X)D)] =&0\\ E\left(I_g[D(Y-\alpha_g^{\Delta}-(ATT+aB_g^X)D)]\right) =&0\\ E[I_g(X\beta-\alpha_g^{X\beta}-B_g^XD)] =&0\\ E\left(I_g[D(X\beta-\alpha_g^{X\beta}-B_g^XD)]\right) =&0\end{aligned}$$

The model is now the solution to moment conditions and can be estimated using the generalized method of moments (GMM) (Hansen 1982) or alternatives such as empirical likelihood (Owen 1988, Qin and Lawless 1994, Newey and Smith 2004).

Model test In the results (Section 4) as well as in the RD example (Appendix E), I test the linearity of Δ_g and B_g^X (which is equivalent to Assumption 1a and is therefore implied by Assumption 1) using the J-test (Hansen 1982).

To see what the J-test tests, we can see what the restrictions are that we have placed on the data such that not all the moment conditions might be able to hold simultaneously. First, under the usual rank conditions in which OLS is identified, β is exactly identified by the first set of moment conditions. Each Δ_g and B_g^X was exactly identified as well before we replaced each Δ_g with $ATT + aB_g^X$; Δ_g was identified regardless of the value of β or B_g^X, and given any value of β, B_g^X was exactly identified as well.

Therefore the potentially binding constraint is the restriction that Δ_g can be substituted by $ATT+aB_g^X$. This substitution makes it such that B_g^X is the solution to two sets of moment equations, and the moment conditions can only all hold if there exist some constants ATT and a such that B_g^X can satisfy both the moment conditions for Δ_g and from the definition of B_g^X. Whenever such constants exist, then Δ_g can be written as a constant plus B_g^X times another constant, which is to say that Δ_g is a linear function of B_g^X.

As described earlier, the linearity of Δ_g with respect to B_g^X holds if and only if Assumption 1a holds. Therefore we can test Assumption 1a by testing whether all of these moment conditions can hold simultaneously, as Hansen's J-test does.

Remark 5 (Selection on observables)**.** Recall that selection on observables implies that $c = 0$, and therefore that $a = 1$. Two tests of selection on observables are therefore possible. First, we can estimate the model and then test whether $a = 1$. Second, we can estimate the model fixing $a = 1$ and then use the J-test. Simulations results suggest that the former option has more power.

4 Results

I will first show OLS results, then the results using the selection ratio method.

Table 3: OLS estimates of employment effects to $t + 4$

	(1)	(2)	(3)	(4)
treated	-.0287*	-.0157	-.0134*	-.0095
	(.0138)	(.0116)	(.0055)	(.0070)
Time FE	Y	Y	Y	Y
County FE	Y	Y	N	N
Lags	N	N	Y	Y
Border only	N	Y	N	Y

**,* indicate significance at 1, and 5 percent, respectively. OLS estimates. Standard errors clustered at the state level are reported in parentheses. The dependent variable is the gain in log of teen employment from the start of period $t - 1$ to the start of period $t + 4$. Lags are the eight prior periods of changes in log of teen employment. In Column 4, only border counties are used in the regression. In Column 2, in order to estimate the county fixed effect, all years of data from counties which are ever border counties are used in the regression.

4.1 OLS results

A baseline estimate of the employment effects of changing the minimum wage can be obtained by estimating a simple two-way fixed effects model:

$$Y_{it} = \alpha + \delta D_{it} + \gamma_i + \omega_t + \varepsilon_{it}.$$

Here, D is a dummy for whether county i experiences an increase in the minimum wage in quarter t, and ω_t and γ_i are time and county fixed effects respectively. I define Y to be the growth in the log of teen employment which occurs between the start of period $t-1$ and the start of period $t+4$, i.e. at the start of the quarter which occurs one year after the minimum wage was increased. I use $t - 1$ as the starting point to allow the possibility that employers may begin to reduce the size of their workforce before the minimum wage increase begins.

The results are shown in the first two columns of Table 3, first for the full sample and then for the set of border counties alone. We can see that the results are sensitive to the decision to restrict the sample. This is consistent with ADR's argument that employment trends are systematically different in regions near where minimum wage changes from trends in more distant regions.

To account for the possibility of violations of the common trends assumption, I next include information on recent trends in employment. While the minimum wage is more likely to be increased in states with downward employment trends (e.g. Addison et al. 2012), Meer and West (2015) caution against the direct use of a linear trend term, since the trend term will be estimated off of data after the law change, and since the effects of a law change may be expected to appear as a new trend in employment rather than as a discrete jump to a new level. To accommodate these points, I estimate the regression again, but controlling for lagged employment changes for eight periods prior to the change in the minimum wage law, i.e. controlling for the change in log of teen employment between the start of $t-2$ and $t-1$, between the start of $t-3$ and $t-2$, etc.

In this specification, I now find a point estimate of more modest disemployment effects, at the low end of the -.1 to -.3 range that was sometimes described as the "consensus" estimate in the early minimum wage literature (Brown et al. 1982). This estimate is also less sensitive to the restriction on the control group; the coefficients are not statistically different from each other when we use the sample of border counties instead of the full sample. The results are shown in the third and fourth columns of Table 3.

The estimates of the coefficient of interest are not sensitive to the inclusion of higher-order (quadratic, cubic) versions of the lags. However, we might be concerned that the coefficient of interest is nonetheless sensitive to the existence of some spatially-varying unobservables. The selection ratio method explicitly models this possibility.

4.2 Selection ratio results

I apply the selection ratio method developed in Section 3 to study the employment effects of the minimum wage using the following definitions of variables. Observations are a county i in quarter t. As before, let $D_{it} = 1$ if the minimum wage changes in county i at some point during quarter t. (In practice, this means that the minimum wage increases.) I will use a variety of outcome variables Y, each involving a change over some period of time in the log of either employment, average monthly earnings (conditional on employment), number of hires, or number of separations for the population of workers age 14-18. I do not show results for ages 19-21; they are broadly similar, but somewhat smaller in magnitude.

Finally, as in the last OLS specification, the vector of controls X contains both time period fixed effects and eight quarters of lags of changes in the outcome of interest. So, when I study the effects of an increase in the minimum wage in county i at time t on Y where Y is defined to be a change in the log of teen employment over any period of time, X will contain time period fixed effects plus the increase in log of teen employment in county i which occurred during quarter $t-8$, the increase in $t-7$, etc. The results are not sensitive to the inclusion of additional lags.

I take the treated group to be the set of treated counties which neighbor an untreated county. The main results are similar in sign and magnitude if the treated group is taken to be the entire set of treated observations, but I consider the estimates to be less informative because a greater degree of extrapolation is required to project Δ_g for a group with $B_g^X = 0$, which increases the standard errors and makes the results more sensitive to small violations of the linearity assumption (Assumption 1a).

The first control group consists of untreated counties which neighbor a treated county; the second control group consists of untreated counties which neighbor a neighbor of a treated county, and are not in the first control group; etc. The main analysis uses six control groups defined in this way; this number was selected before obtaining any results, but the main results are not sensitive to the use of additional control groups.[5]

Estimates are reported as the effects of an increase in the minimum wage. Recall from Section 2 that the average minimum wage increase in the data is approximately 10%; therefore, if we follow the convention in the literature of assuming that effects take the form of an elasticity, each effect should be multiplied by ten to recover an estimate of the elasticity.

Earnings effects I begin by showing that minimum wage law increases did in fact have an effect on the average monthly earnings of employed teenagers. The first outcome variable is the increase in the log of average monthly wage for employed teenagers between the end of quarter $t-1$ and the end of quarter t, where t is the quarter during which the minimum wage changes. The point estimate is that increases in the minimum wage caused an immediate increase of 1.6% in monthly earnings.

Figure 3 shows estimates of Δ_g and B_g^X for each control group. Points are numbered by the distance between members of that control group and the treated group, i.e. neighbors are 1, neighbors-of-neighbors are 2, etc. We can see from the graphs that Δ_g and B_g^X do seem to be linearly related, and the J-test fails to reject.

The evolution of earnings effects is shown in Table 4. I use the same X and D in each estimation, but change the Y variable. In the first column, Y is the increase in log of average monthly earnings between the end of $t-1$ and the end of t. In the second column, Y is the increase between the end of $t-1$ and the end of $t+1$; in the third column, it is the gain from the end of $t-1$ to the end of $t+2$; etc.

Figure 4 shows estimates of Δ_g and B_g^X for each control group when the outcome variable is the gain in the log of earnings between $t-1$ and $t+4$, the quarter one year after the minimum wage is increased.

Employment effects Next, I turn to employment effects. Treatment is defined as before, and the covariates are time fixed effects and the increases in employment in county i from the start of $t-2$ to the start of $t-1$, from the start of $t-3$ to the start of $t-2$, etc. through eight quarters.

[5] I selected this number because I suspected that there would be little additional variation in B_g^X beyond this control group for the key observables.

Figure 3: Increase in log earnings, $t-1$ to t

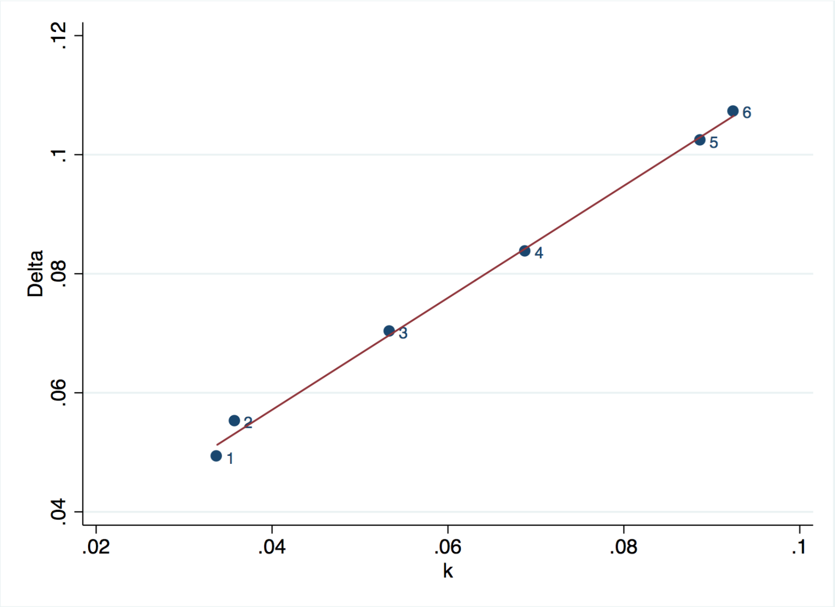


Figure 4: Increase in log earnings, $t-1$ to $t+4$

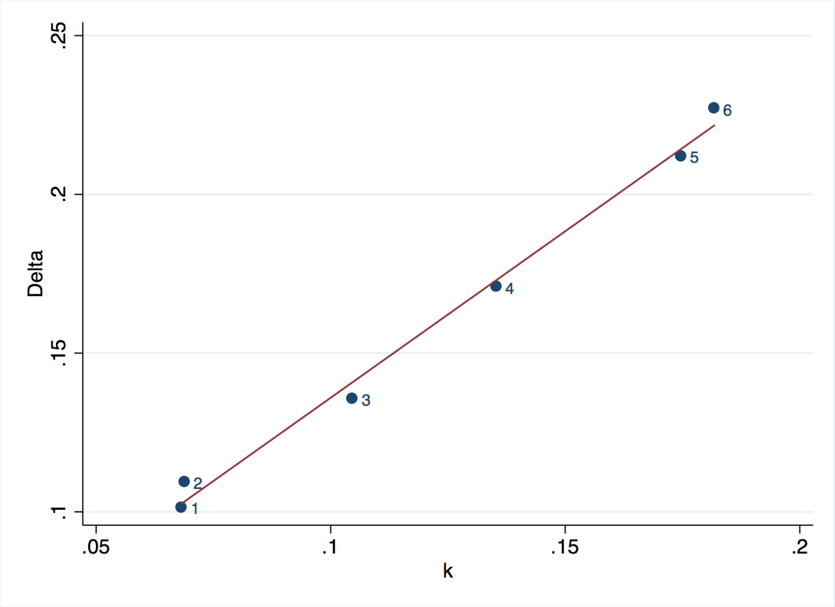


Table 4: Earnings effects

	t	$t+1$	$t+2$	$t+3$	$t+4$
ATT	.0160**	.0127*	.0129**	.0154**	.0104
	(.0042)	(.0056)	(.0043)	(.0038)	(.0057)
a (1=selection	1.034	.983	1.144	.741	1.035
on observables)	(.051)	(.099)	(.110)	(.347)	(.068)
J-test p-value	.25	.84	.99	.32	.53

**,* indicate significance at 1, and 5 percent, respectively, where statistical significance is from 0 for *ATT* and from 1 for a. GMM estimates, clustered by state. See Appendix B for details of the estimation. The dependent variable is the gain in log of average monthly earnings for employed teens from the start of period $t-1$ to the start of the period listed at the top of each column. Controls are time period dummies and the eight prior periods of changes in log of teen monthly earnings.

As before, I allow for anticipatory effects of the minimum wage law increase by measuring employment gains relative to the start of period $t-1$ rather than the start of period t. However, it is worth noting that this concern about anticipatory effects of the minimum wage does not seem to be important in practice. Figure 5 shows graphical results for the effects of treatment on the increase in employment from the start of $t-1$ to the start of t. As can be seen in the first column of Table 5, the effect of a minimum wage law change on employment in that period is measured to be zero.

The full set of GMM estimates of employment effects are shown in Table 5. Each column in the GMM results represents a different outcome variable, though the set of covariates are kept the same. The first column shows the effect on the gain in teen employment from $t-1$ to t, the second column is the gain from $t-1$ to $t+1$, etc. Figure 6 shows the relationship between the estimated values of Δ_g and B_g^X when the outcome variable is the employment gain in the year following the minimum wage increase, i.e. through period $t+4$. We can see once again that the J-test fails to reject and the linearity assumption seems reasonable. The measured employment effects are small and not statistically significant.

Flows DLRb report that increases in the minimum wage lead to large decreases in both separations and hires, which they argue is consistent with a job ladder model of the labor market. I perform the same analysis for separations and hires as for

Figure 5: Increase in log employment, $t-1$ to t

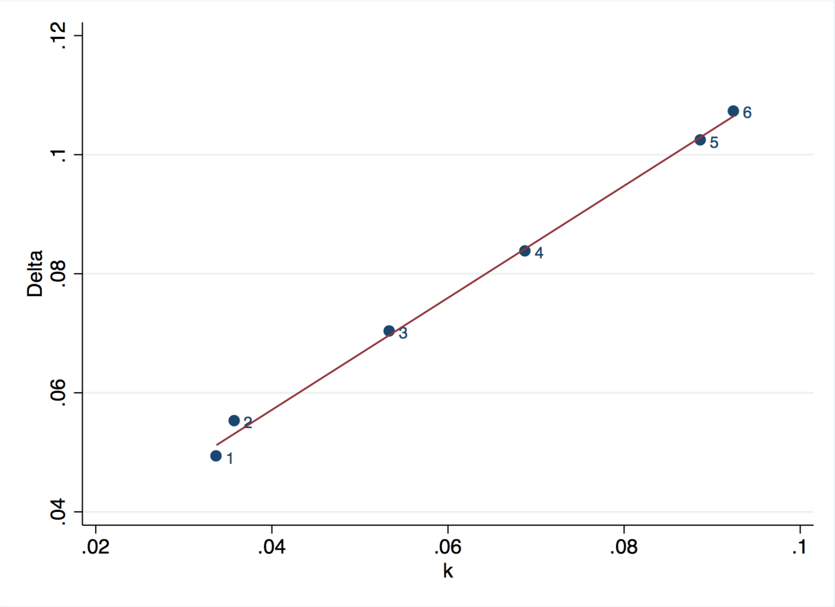


Figure 6: Increase in log employment, $t-1$ to $t+4$

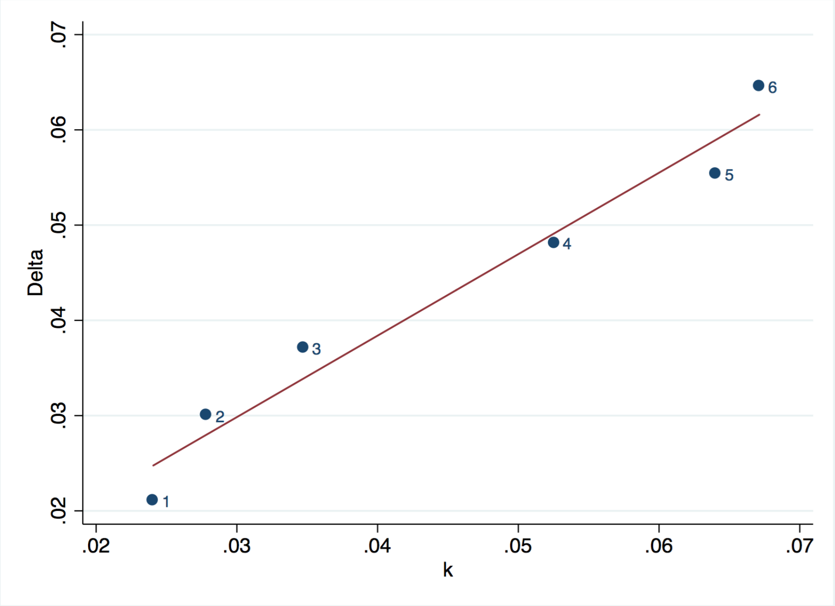


Table 5: Employment effects

	t	$t+1$	$t+2$	$t+3$	$t+4$
ATT	-.0006	-.0033	-.0030	-.0006	-.0041
	(.0051)	(.0046)	(.0055)	(.0054)	(.0090)
a (1=selection	1.022	.991	1.144	1.043	.970
on observables)	(.043)	(.123)	(.201)	(.273)	(.117)
J-test p-value	.65	.34	.74	.32	.18

**,* indicate significance at 1, and 5 percent, respectively, where statistical significance is from 0 for *ATT* and from 1 for a. GMM estimates, clustered by state. See Appendix B for details of the estimation. The dependent variable is the gain in log of teen employment from the start of period $t-1$ to the start of the period listed at the top of each column. Controls are time period dummies and the eight prior periods of changes in log of teen employment.

employment, but substituting changes in separations (or hires) for each variable that was previously a measure of employment.

As seen in Tables 6 and 7, the results for separations and hires are consistent with the possibility of some effect on labor market flows. However, the standard errors are sufficiently large that the results are not statistically significant, even though they are not inconsistent with the previous findings of DLRb.

5 Discussion

The results suggest several conclusions.

First, selection on observables is generally not rejected in these models when the observables are taken to be eight quarters of lags of changes in the outcome variable prior to the quarter in which the law changed. This regression is subtly different from regressions in which more complex versions of treatment are allowed (e.g. allowing a continuous treatment variable containing the log of the change in the minimum wage, or allowing the outcome to vary as a function of all recent changes in the minimum wage law and not just whether the law was changed in a single period) but it increases confidence in the possibility that rich controls for recent trends can allow researchers to overcome the problem that minimum wage increases are not randomly assigned. Additionally, the lack of variation in unobservables across the control groups supports the idea that the primary value of using border counties in minimum wage research designs is simply through the diminished difference in recent

Table 6: Effects on separations

	t	$t+1$	$t+2$	$t+3$	$t+4$
ATT	-.0155	-.0235*	-.0113	-.0180	-.0065
	(.0099)	(.0099)	(.0074)	(.0109)	(.0179)
a (1=selection	.991	.916	1.018	1.299	.870
on observables)	(.045)	(.359)	(.044)	(.357)	(.078)
J-test p-value	.30	.73	.87	.59	.19

**,* indicate significance at 1, and 5 percent, respectively, where statistical significance is from 0 for *ATT* and from 1 for a. GMM estimates, clustered by state. See Appendix B for details of the estimation. The dependent variable is the gain in the log of total separations for teen workers between period $t-1$ and the period listed at the top of each column. Controls are time period dummies and the eight prior periods of changes in log of teen separations.

Table 7: Effects on hires

	t	$t+1$	$t+2$	$t+3$	$t+4$
ATT	-.0222	-.0055	-.0011	-.0165	-.0304
	(.0193)	(.0135)	(.0105)	(.0090)	(.0264)
a (1=selection	.763	1.074	1.003	.786	.741
on observables)	(.363)	(.060)	(.030)	(.401)	(.262)
J-test p-value	.68	.25	.86	.30	.28

**,* indicate significance at 1, and 5 percent, respectively, where statistical significance is from 0 for *ATT* and from 1 for a. GMM estimates, clustered by state. See Appendix B for details of the estimation. The dependent variable is the gain in log of teen hires between period $t-1$ and the period listed at the top of each column. Controls are time period dummies and the eight prior periods of changes in log of teen hires.

employment trends.[6]

However, the evidence to support selection on observables for longer-term outcome measures is much weaker. When using changes in employment or flows over periods longer than a year after an increase in the minimum wage, the plots of Δ_g against B_g^X become increasingly noisy. This increased noise diminishes the power to reject Assumption 1a, such that it is difficult to consider the failure to reject in the J-test as strong evidence in favor of this assumption. The evidence in favor of selection on observables in the short run is suggestive of the possibility that selection on observables works for long run outcomes as well, but only to the extent that the set of relevant omitted variables is the same for each regression.

Second, the measured employment effects are modest. While the standard errors are sufficiently large that we cannot rule out elasticities of $-.1$ or $-.2$, we can nonetheless reject many of the disemployment effects seen in the literature (see Neumark and Wascher 2007). The small amount of variation on unobservables suggests that, under modest violations of Assumption 1, it would still be unlikely that the employment elasticity is $-.4$, or any larger disemployment effect.

Third, the net effect of minimum wages on earnings for teenagers is likely to be positive. This stands in contrast with Neumark and Wascher (2008), who, reviewing the literature, claim that "although minimum wages compress the wage distribution, because of employment and hours declines among those whose wages are most affected by minimum wage increases, a higher minimum wage tends to reduce rather than to increase the earnings of the lowest-skilled individuals." For this to hold true in the data, it would need to be the case that the decrease in the number of employed teens was larger than the increase in average teen earnings conditional on being employed. Although a hypothesis test fails to reject that the disemployment effect is as large as the earnings effect after one year, my point estimates of earnings increases conditional on employment are larger in magnitude than the disemployment effect estimates, suggesting a net transfer to teens.

5.1 Potential threats to identification

I consider several threats to the research design.

Spillovers: One possible issue with border designs is that treatment in one county might affect the outcome in a neighboring county. We might particularly be concerned that neighboring counties belong to the same labor market. However, it seems less likely that employment spillovers would extent to distances of greater than a county, and the main estimates are all robust to dropping the nearest control group. In fact, robustness of estimates to dropping a control group is a general feature of the selection ratio method, since the overidentification test is precisely a test that any subset of the control groups would produce the same estimate up to sampling error.

[6] DLRa, ADR, and Meer and West (2015) make similar suggestions.

Simultaneous law changes: Another concern might be that there is some unobservable which does not change across any control group but which differs between the treated group and all of the control groups. In this case, Assumption 1a may hold, but Assumption 1b would be suspect, since taking a control group spatially close enough to match on observables would not result in matching on this unobservable. Therefore, $B_g^X = 0$ would not imply that $B_g = 0$.

A candidate for this sort of unobservable is other state policies, which vary between the treated group and each control group (since treatment is determined at the state level) and which may not vary substantially across control groups (since e.g. neighbors and neighbors-of-neighbors of a treated observation tend to lie within the same state). However, because the outcome is a change in teen employment, then an unobservable must be a variable which affects trends in employment after the law change; yet if this unobservable were a policy change, it is also likely to appear in prior trends in employment, which are in X. Therefore this concern is likely to be limited to state-specific policy changes which tend to coincide with minimum wage increases, but which generally occur either simultaneous to or directly after minimum wage increases. While I cannot rule out this possibility, prior research is often also susceptible to the same concern and has not uncovered any obvious policy which fits this description.

Finite sample issues: Since this method is novel, it is natural to be concerned that its finite sample properties might be undesirable. To allay this concern, I show in Appendix D that the method performs well in simulations. In addition, I show in Appendix E that the method also produces reasonable estimates of a regression discontinuity design.

Limitations There are several limitations to the results described here.

First, the concept of a single minimum wage which applies equally to all establishments is a simplification. Minimum wage laws have historically contained various exceptions, such as for sufficiently small firms, for recently hired workers, for workers of various ages and (though not recently) genders, and for workers in various industries.[7] To the extent that there is not simply a single minimum wage, this complicates any attempt to interpret the employment effects of the minimum wage. The effect on wages does suggest, however, that the measured minimum wage rules have some bite.

Second, I do not make a serious attempt to engage with the problem of dynamics, beyond tracing out the effects over the range of time that I feel is warranted. Previous authors (e.g. Baker et al. 1999, Meer and West 2015) have argued that employment effects may increase over time since a law is passed, in which case the estimates I have presented do not reflect long-term effects. Sorkin (2015) points out that the

[7] Neumark and Wascher (2008) describe the history of minimum wage laws, and who was affected by them, in some detail. Interestingly, Neumark and Wascher describe how the choice of covered populations in the United States has historically been heavily shaped by attempts to reduce the probability that minimum wage laws would be ruled unconstitutional.

effects of minimum wage increases on employment through substitution from labor to capital might be attenuated when firms view the change in the real minimum wage as temporary, and presents a model in which temporary minimum wage increases produce small employment effects while permanent minimum wage increases produce large employment effects. I make no attempt to determine whether recent minimum wage increases in the United States should properly be seen as short-run or long-run changes to the real minimum wage.

Third, as noted before, states which increase their minimum wage do so by an average of approximately 10%. However, there is some variation in the size of the increases, with some increases by less than 5% and a few by over 20%. I adopt the convention in the literature of assuming that effects take the form of an elasticity, and therefore interpret the selection ratio results as estimates of the effects of a 10% increase in the minimum wage. However, it is possible that the causal effects on the log of employment do not scale linearly with the log of the minimum wage (e.g. Baker et al. 1999), in which case the results should be interpreted as an average of the effects of the various law change sizes seen in the data.

6 Conclusion

I study the effects of the minimum wage on employment using a new method that identifies treatment effects using a series of imperfect control groups. I find that, for the kind of minimum wage law increases seen in the recent history of the United States, the employment effects for teenagers were most likely small or zero. Obviously I am unable to produce evidence on the potential employment effects of minimum wage law changes substantially larger than those seen in the data, as are now sometimes being proposed in the United States; these employment effects may differ substantially from the results in this paper if, for example, the small effects seen in this paper reflect flexibility in wages arising from adjustment mechanisms (like worker effort or fringe benefits) which may not be able to scale up effectively.

While some of the assumptions required for the method to be useful may be unusually believable in the context of the minimum wage, the large number of empirical problems with a similar structure (treated group and a series of imperfect control groups) suggests that the method may be useful to researchers studying other problems. This is illustrated through the inclusion of an additional application in Appendix E, demonstrating that the same identification strategy can be used to measure treatment effects in a regression discontinuity design, including for populations away from the threshold.

References

[1] Abadie, A. and J. Gardeazabal (2003). The Economic Costs of Conflict: A Case Study of the Basque Country, *American Economic Review*, 93(1), 112-132.

[2] Abadie, A., Diamond, A., and J. Hainmueller (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program, *Journal of the American Statistical Association*, 105(490), 493-505.

[3] Addison, J., M. Blackburn, and C. Cotti (2009). Do Minimum Wages Raise Employment? Evidence from the U.S. Retail-Trade Sector, *Labour Economics*, 16(4), 397-408.

[4] Addison, J.T., Blackburn, M.L., and C.D. Cotti (2012). The Effect of Minimum Wages on Labour Market Outcomes: County-Level Estimates from the Restaurant-and-Bar Sector, *British Journal of Industrial Relations*, 50(3), 412-435.

[5] Allegretto, S., Dube, A., and M. Reich (2011). Do Minimum Wages Really Reduce Teen Employment? Accounting for Heterogeneity and Selectivity in State Panel Data, *Industrial Relations*, 50(5), 205-240.

[6] Altonji, J.G., Elder T.E., and C.R. Taber (2005). Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools, *Journal of Political Economy*, 113(1), 151-184.

[7] Altonji, J.G. and R.K. Mansfield (2014). Group-Average Observables as Controls for Sorting on Unobservables When Estimating Group Treatment Effects: The Case of School and Neighborhood Effects, NBER working paper no. w20781.

[8] Angrist, J.D. and J.S. Pischke (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

[9] Angrist, J.D. and M. Rokkanen (2012). Wanna Get Away? RD Identification Away from the Cutoff, NBER working paper no. w18662.

[10] Athey, S. and G.W. Imbens (2006). Identification and Inference in Nonlinear Difference-in-Differences Models, *Econometrica*, 74(2), 431-497.

[11] Bai, J. (2009). Panel Data Models with Interactive Fixed Effects, *Econometrica*, 77(4), 1229-1279.

[12] Baker, M., Benjamin, D., and S. Stanger (1999). The Highs and Lows of the Minimum Wage Effect: A Time-Series Cross-Section Study of the Canadian Law, *Journal of Labor Economics*, 17(2), 318-350.

[13] Bárány, Z, (2015). The Minimum Wage and Inequality, *Journal of Labor Economics*, forthcoming.

[14] Bancroft, T.A. (1944). On Biases in Estimation Due to the Use of Preliminary Tests of Significance. *Annals of Mathematical Statistics*, 15, 190204.

[15] Black, S.E. (1999). Do Better Schools Matter? Parental Valuation of Elementary Education, *Quarterly Journal of Economics*, 114(2), 577-599.

[16] Breusch, T.S. and A.R. Pagan (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation, *Econometrica*, 47(5), 1287-1294.

[17] Brochu, P. and D. A. Green (2013). The Impact of Minimum Wages on Labour Market Transitions, *Economic Journal*, 123(12), 1203-1235.

[18] Brown, C., Gilroy, C., and A. Kohen (1982). The Effect of the Minimum Wage on Employment and Unemployment, *Journal of Economic Literature*, 20(2), 487-528.

[19] Card, D. (1992). Using Regional Variation in Wages to Measure the Effects of the Federal Minimum Wage, *Industrial and Labor Relations Review*, 46, 22-37.

[20] Card, D. (1992). Do Minimum Wages Reduce Employment? A Case Study of California, 1987-89, *Industrial and Labor Relations Review*, 46, 38-54.

[21] Card, D. and A.B. Krueger (1994). Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania, *American Economic Review*, 84(4), 772-793.

[22] Card, D. and A.B. Krueger (1995a). Time-Series Minimum-Wage Studies: A Meta-analysis, *American Economic Review Papers and Proceedings*, 85(2), 238-243.

[23] Card, D. and A.B. Krueger (1995b). *Myth and measurement.* Princeton, NJ: Princeton University Press, 1995.

[24] 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.

[25] Chamberlain, G. (1987). Asymptotic Efficiency in Estimation with Conditional Moment Restrictions, *Journal of Econometrics*, 34(3), 305-334.

[26] Currie, J. and D. Thomas (1995). Does Head Start Make a Difference?, *American Economic Review*, 85(3), 341-364.

[27] Danilov, D. and J.R. Magnus (2004). On the Harm that Ignoring Pretesting Can Cause, *Journal of Econometrics*, 122, 27-46.

[28] Dube, A., Lester, T.W., and M. Reich (2010). Minimum Wage Effects Across State Borders: Estimates Using Contiguous Counties, *Review of Economics and Statistics*, 92(4), 945-964.

[29] Dube, A., Lester, T.W., and M. Reich (2014). Minimum Wage Shocks, Employment Flows, and Labor Market Frictions, *Journal of Labor Economics*, forthcoming.

[30] Dong, Y. and A. Lewbel (2014). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models, *Review of Economics and Statistics*, forthcoming

[31] Giuliano, L. (2013). Minimum Wage Effects on Employment, Substitution, and the Teenage Labor Supply: Evidence from Personnel Data, *Journal of Labor Economics*, 31(1), 155-194.

[32] Gorry, A. (2013). Minimum Wages and Youth Unemployment, *European Economic Review* 64, 57-75.

[33] Hagedorn, M., Karahan F., Manovskii I., and K. Mitman (2013). Unemployment Benefits and Unemployment in the Great Recession: The Role of Macro Effect, NBER working paper no. 19499.

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

[35] Hansen, B. (2014). *Econometrics*, self-published.

[36] IGM poll of economists (2014): http://www.igmchicago.org/igm-economic-experts-panel/poll-results?SurveyID=SV_br0IEq5a9E77NMV.

[37] Imbens, G.W. and J.D. Angrist (1994). Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62(2), 467-475.

[38] Imbens, G.W. and T. Lemieux (2008). Regression Discontinuity Designs: A Guide to Practice, *Journal of Econometrics*, 142(2), 615-635.

[39] Jackson, C.K. (2010) Do Students Benefit From Attending Better Schools?: Evidence From Rule-based Student Assignments in Trinidad and Tobago, *Economic Journal*, 120(549): 1399-1429.

[40] Kaitz, H. (1970) Experience of the Past: The National Minimum, *Youth unemployment and minimum wages*. Bulletin 1657. U.S. Department of Labor, Bureau of Labor Statistics, 30-54.

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

[42] Lee, D.S. and T. Lemieux (2010). Regression Discontinuity Designs in Economics, *Journal of Economic Literature*, 48(2), 281-355.

[43] Matsudaira, J. (2014). Monopsony in the Low-Wage Labor Market? Evidence from Minimum Nurse Staffing Regulations, *Review of Economics and Statistics* 96(1), 92-102.

[44] Meer, J. and J. West (2015). Effects of the Minimum Wage on Employment Dynamics, *Journal of Human Resources*, forthcoming.

[45] Neumark, D., Salas, J.M.I., and W. Wascher (2014a). Revisiting the Minimum Wage-Employment Debate: Throwing Out the Baby with the Bathwater?, *Industrial and Labor Relations Review*, 67, 608-648.

[46] Neumark, D., Salas, J.M.I., and W. Wascher (2014b). More on Recent Evidence on the Effects of Minimum Wages in the United States, *IZA Journal of Labor Policy*, 3(24).

[47] Neumark, D. and W. Wascher (2007). Minimum Wages and Employment, *Foundations and Trends in Microeconomics*, 1-182.

[48] Neumark, D. and W. Wascher (2008). *Minimum wages*. MIT Press.

[49] Neumark, D. and W. Wascher (2011). Does a Higher Minimum Wage Enhance the Effectiveness of the Earned Income Tax Credit?, *Industrial and Labor Relations Review*, 64(4), 712-746.

[50] Newey, W.K. and R.J. Smith (2004). Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators, *Econometrica*, 72(1), 219-255.

[51] Owen, A.B. (1988). Empirical Likelihood Ratio Confidence Intervals for a Single Functional, *Biometrika*, 75(2), 237-249.

[52] Qin, J. and J. Lawless (1994). Empirical Likelihood and General Estimating Equations, *Annals of Statistics*, 22, 300325.

[53] Ropponen, O. (2010). Minimum Wages and Employment: Replication of Card and Krueger (1994) Using the CIC Estimator, HECER Discussion Paper no. 289.

[54] Rubin, D. B. (1986). Comment: Which Ifs Have Causal Answers, *Journal of the American Statistical Association*, 81(396), 961-962.

[55] Sabia, J. (2009). The Effects of Minimum Wage Increases on Retail Employment and Hours: New Evidence from Monthly CPS Data, *Journal of Labor Research*, 30(1), 311-328.

[56] Sabia, J., R. Burkhauser, and B. Hansen (2012). Are the Effects of Minimum Wage Increases Always Small? New Evidence from a Case Study of New York State, *Industrial and Labor Relations Review*, 65(2), 350-376.

[57] Sen, A., K. Rybczynski, and C. V. D. Waal (2011). Teen Employment, Poverty, and the Minimum Wage: Evidence from Canada, *Labour Economics*, 18(1), 36-47.

[58] Sen, A. and H. Ariizumi (2013). Teen Families, Welfare Transfers, and the Minimum Wage: Evidence from Canada, *Canadian Journal of Economics*, 46(1), 338-360.

[59] Sorkin, I. (2015). Are There Long-Run Effects of the Minimum Wage?, *Review of Economic Dynamics*, 18(2), 306-333.

[60] Stigler, G.J. (1946). The Economics of Minimum Wage Legislation, *American Economic Review*, 36(3), 358-365.

[61] Wing, C. and Cook, T. D. (2013). Strengthening the Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison, *Journal of Policy Analysis and Management*, 32(4), 853877.

[62] Yannelis, C. (2014). The Minimum Wage and Employment Dynamics: Evidence from an Age Based Reform in Greece, *Working Paper*.

Appendix

A Proofs

A.1 Proof of Theorem 1

Assume that Assumptions 1a and 1b are true. From Assumption 1a, we have that, for any control groups g and g',

$$E(\eta \mid D_i = 0, i \in g) - E(\eta \mid D_i = 0, i \in g') = c[E(X\beta \mid D_i = 0, i \in g) - E(X\beta \mid D_i = 0, i \in g')].$$

First, observe that the right-hand side of this equality is equal to $c(B^X_{g'} - B^X_g)$. Second, observe that the left-hand side is equal to $[\alpha_V + E(\eta \mid D_i = 0, i \in g)] - [\alpha_V + E(\eta \mid D_i = 0, i \in g')] = E(U \mid D_i, i \in g) - E(U \mid D_i, i \in g') = B^U_{g'} - B^U_g$. Therefore from Assumption 1a we have that, for any g and g',

$$B^U_{g'} - B^U_g = c(B^X_{g'} - B^X_g).$$

This establishes the linearity of B^U and B^X with a slope of c. It remains to show that the intercept is equal to 0.

Under Assumption 1b, we have that

$$\delta_V - E(\eta \mid D_i = 0, i \in g) = c[E(X\beta \mid D_i = 1) - E(X\beta \mid D_i = 0, i \in g)]$$

for some g. That is,

$$\delta_V - E(\eta \mid D_i = 0, i \in g) = cB_g^X.$$

Now,

$$\begin{aligned}\delta_V - E(\eta \mid D_i = 0, i \in g) =& [E(U \mid D_i = 1) - \alpha_V] - [E(U \mid D_i = 0, i \in g) - \alpha_V] \\ =& E(U \mid D_i = 1) - E(U \mid D_i = 0, i \in g) \\ =& B_g^U.\end{aligned}$$

Therefore the statement is that for some group g, $B_g^U = cB_g^X$ for the same c. Then it cannot be the case that the line relating B_g^U with B_g^X across all groups g has a nonzero intercept. We conclude that Assumption 1 holds.

Now the converse. Suppose that Assumption 1 holds. Then $B_g^U - B_{g'}^U = cB_g^X - cB_{g'}^X = c(B_g^X - B_{g'}^X)$. Assumption 1a follows. Additionally, since $B_g^U = cB_g^X$ holds for all g, then it follows trivially that it holds for some g. Therefore Assumption 1b holds as well.

A.2 Proof of Theorem 2

First, observe that, for any g, $B_g = B_g^X + B_g^U$. This is because, from the definition of B_g,

$$\begin{aligned}B_g =& E(Y_i(0) \mid D_i = 1) - E(Y_i(0) \mid D_i = 0, i \in g) \\ =& E(\alpha_0 + X\beta + U \mid D_i = 1) - E(\alpha_0 + X\beta + U \mid D_i = 0, i \in g) \\ =& E(X\beta \mid D_i = 1) - E(X\beta \mid D_i = 0, i \in g) + E(U \mid D_i = 1) - E(U \mid D_i = 0, i \in g) \\ =& B_g^X + B_g^U.\end{aligned}$$

Then $B_g = B_g^X + cB_g^X = (1 + c)B_g^X$ for any g. It follows that $\frac{B_g}{B_{g'}} = \frac{B_g^X}{B_{g'}^X}$ whenever $B_{g'} \neq 0$.

Now, completing the claim of identification: Following the system of equations argument from Section 3.1 and using that $\frac{B_g}{B_{g'}} = \frac{B_g^X}{B_{g'}^X}$, then identification of ATT follows when we can identify $\frac{B_g^X}{B_{g'}^X}$ for two control groups g and g' such that $B_g^X \neq B_{g'}^X$. The antecedent of the theorem posits the existence of two control groups g and g' such that $B_g^X \neq B_{g'}^X$, so it remains only to show that B_g^X is identified for each group g.

Now, $B_g^X := E(X\beta \mid D_i = 1) - E(X\beta \mid D_i = 0, i \in g)$. We observe X, D, and membership in each control group. Therefore it remains to show that β is identified. But we have that

$$E(Y \mid D, X) = (\alpha_0 + \alpha_V) + (\delta_0 + \delta_V)D + (\beta_0 + \beta_V)X$$

and $\beta = \beta_0 + \beta_V$. Therefore β is identified as the coefficient on X in $E(Y \mid D, X)$, which is identified whenever Y is observed as well. This finishes the proof of identification.

A.3 Proof of Theorem 3

If Assumption 1a holds, then $B_g^U = a + cB_g^X$ for all g and some constants a and c. (See the proof of Theorem 1 for a demonstration that Assumption 1a implies linearity of B_g^U and B_g^X.) Then, since $B_g = B_g^X + B_g^U$, we have $B_g = a + (1+c)B_g^X$. Finally, $\Delta_g = ATT + B_g$. Therefore, $\Delta_g = (ATT + a) + (1+c)B_g^X$, where ATT, a, and c are all constants. In other words, Δ_g is linear in B_g^X.

The converse follows by allowing a and c to be whatever constants rationalize the equation $\Delta_g = (ATT + a) + (1+c)B_g^X$ and following the same algebraic steps in reverse.

A.4 Relaxation of linear CEF

It is not necessary to assume a linear $E(V \mid D, X)$. Instead, it suffices to write that the best linear predictor of V using D and X is $\alpha_V + \delta_V D + \beta_V X$. That is, out of the set of linear functions $g(D, X)$, the function $g^*(D, X) = \alpha_V + \delta_V D + \beta_V X$ minimizes $E(V - g^*(D, X)$. Some such α_V, δ_V, and β_V will exist whenever the variances of V, D, and X are all finite and D and X are not perfectly multicollinear (see Hansen 2014, chapter 2.18). Since D is binary and therefore has finite variance, then it suffices to assume a rank condition and that X and Y have finite variance.

We now define η to be the difference between V and the best linear prediction of V. Then η is still uncorrelated with D and X and is still mean-zero (see Hansen 2014). Furthermore, by plugging in for V in the structural equation for Y, we can still write

$$Y = (\alpha_0 + \alpha_V) + (\delta_0 + \delta_V)D + (\beta_0 + \beta_V)X + \eta. \tag{1}$$

The definitions of β, U, B_g^X, and B_g^U, Assumptions 1, 1a, and 1b, and the proofs of Theorems 1 and 3 all follow without modification. The proof of Theorem 2 also follows so long as β is identified. Therefore it remains to establish that β is still identified.

Claim: β is identified as the coefficient on X in the best linear prediction of Y given D and X.

Proof. First, we must determine whether a best linear predictor exists. Since we have assumed that a best linear predictor of V given D and X exists, then we have assumed that X and V have finite variance. It follows that Y has finite variance as well, since

$$Y = \alpha_0 + \delta_0 D + \beta_0 X + V$$

from the structural equation, and D, X, and V all have finite variances (and therefore finite covariances as well). Then Y, D, and X all have finite variances and D and X are not perfectly multicollinear, so therefore there exists a best linear predictor of Y given D and X.

The next step is to show that the best linear predictor of Y given D and X is $(\alpha_0 + \alpha_V) + (\delta_0 + \delta_V)D + (\beta_0 + \beta_V)X$.

Observe that, if this is the best linear predictor, then from Equation 1, η is the prediction error.

Let the true best linear predictor be $\alpha^* + \delta^* D + \beta^* X$. Then the prediction error $\eta^* := Y - (\alpha^* + \delta^* D + \beta^* X)$ is

$$\eta^* = \eta + (\alpha_0 + \alpha_V - \alpha^*) + (\delta_0 + \delta_V - \delta^*)D + (\beta_0 + \beta_V - \beta^*)X.$$

Since η is uncorrelated with D and X, and D and X are not perfectly correlated with each other, then if it were the case that $\delta^* \neq \delta_0 + \delta_V$ or $\beta^* \neq \beta_0 + \beta_V$, then η^* would be correlated with D or X. This is a contradiction if η^* is the error in a best linear prediction on D and X. Therefore $\delta^* = \delta_0 + \delta_V$ and $\beta^* = \beta_0 + \beta_V$. Then we have that $\eta^* = \eta + (\alpha_0 + \alpha_V - \alpha^*)$. We know that η^* and η are both mean-zero, so therefore it must be the case that the constant $\alpha_0 + \alpha_V - \alpha^*$ is equal to zero. Therefore we have that $\alpha^* = \alpha_0 + \alpha_V$. This establishes that the best linear predictor of Y given D and X is

$$Y = (\alpha_0 + \alpha_V) + (\delta_0 + \delta_V)D + (\beta_0 + \beta_V)X.$$

Finally, because Y, D, and X are observed, then the coefficient on X in this best linear prediction is identified (see Hansen 2014). This concludes the proof that β is identified. Theorem 2 follows. □

B Modified GMM conditions

For ease of computation, I estimate the model using a slightly modified version of the moment conditions presented in the body of the paper. The computational advantage is the reduction of the number of moment conditions when the number of control groups is large. For example, when estimating the RD model with control groups divided by individual percentiles of margin of victory, there are 50 control groups and, in the original formulation of the moment conditions, four moment conditions per group, plus moment conditions to estimate β. To retain the comparison between these more complicated cases and the cases emphasized in the body of the text, I use the computationally simpler GMM conditions for each estimation that I perform, including the simulation.

The derivation of the modified moment conditions is as follows. From the basic equations, we have that

$$ATT + k_g B - \Delta_g = 0$$

I will refer to observation i as "used in comparison g" if i is either treated or a member of control group g. Let I_i^g be defined to equal 1 if i is used in comparison g and 0 otherwise. Using the parallel convergence assumption and the alternative normalization discussed above, we additionally have that

$$k_g = E(X\beta_i \mid I_i^g = 1, D_i = 1) - E(X\beta_i \mid I_i^g = 1, D_i = 0)$$

Similarly, we have that

$$\Delta_g = E(Y_i \mid I_i^g = 1, D_i = 1) - E(Y_i \mid I_i^g = 1, D_i = 0)$$

To keep the notation simple, restrict for the moment to considering observations in comparison g. Substituting into the original formula, we have that

$$ATT + [E(X_i\beta \mid D_i = 1) - E(X_i\beta \mid D_i = 0)]c - E(Y_i \mid D_i = 1) + E(Y_i \mid D = 0) = 0$$

Let $W \equiv X\beta$. Then this in turn gives that

$$ATT + \left[\frac{E(W_iD_i)}{E(D_i)} - \frac{E(W_i(1-D_i)}{E(1-D_i)}\right]c - \left[\frac{E(Y_iD_i)}{E(D_i)} - \frac{E(Y_i(1-D_i)}{E(1-D_i)}\right] = 0$$

For simplicity of notation, let $p_g = E(D_i \mid I_i^g = 1)$. Note that each p_g can be trivially estimated prior to the use of GMM. Then the equation written above is equal to

$$ATT + \left[\frac{E(W_iD_i)}{p_g} - \frac{E(W_i(1-D_i)}{1-p_g}\right]c - \left[\frac{E(Y_iD_i)}{p_g} - \frac{E(Y_i(1-D_i)}{1-p_g}\right] = 0$$

Moving the expectations to the outside of the equation gives the moment condition

$$E\left(ATT + \frac{cW_iD_i}{p_g} - \frac{cW(1-D_i)}{1-p_g} - \frac{Y_iD_i}{p_g} + \frac{Y_i(1-D_i)}{1-p_g}\right) = 0$$

Of course, this moment condition only holds for $I_i^g = 1$. But it only takes a simple modification to then create a moment condition which holds for all i. Since $I_i^g = 0$ for all observations which are not in comparison g, the following moment condition holds across the entire population:

$$E\left[I_i^g\left(ATT + \frac{cW_iD_i}{p_g} - \frac{cW(1-D_i)}{1-p_g} - \frac{Y_iD_i}{p_g} + \frac{Y_i(1-D_i)}{1-p_g}\right)\right] = 0$$

Therefore, we have one moment condition for each comparison g. We can then estimate ATT and σ using GMM. While they are also the solution to moment conditions, to simplify computation, I also estimate p_g and W prior to implementing GMM. Furthermore, Hansen's J-test (Hansen 1982) gives a test of the internal consistency of the model. Like the moment conditions described in the body of the paper, these moment conditions could of course also be used to estimate the parameters of interest using an empirical likelihood approach (Owen 1988, Qin and Lawless 1994, Newey and Smith 2004).

Testing selection on observables As discussed before, we have $c = 1$ under selection on observables. When we impose this restriction, the moment condition for each comparison is now

$$E\left[I_i^g\left(ATT + \frac{W_iD_i}{p_g} - \frac{W(1-D_i)}{1-p_g} - \frac{Y_iD_i}{p_g} + \frac{Y_i(1-D_i)}{1-p_g}\right)\right] = 0$$

With only one unknown, this model is overidentified (and therefore amenable to the use of the J-test) as soon as there are two moment conditions, i.e. two comparison groups.

C Illustrative Example

A classic example in which two variables can be correlated with the same variable but uncorrelated with each other is that, when a coin is flipped twice, the result of each coin flip is correlated with the total number of heads, yet not correlated with the result of the other coin flip. The example I construct follows very similar logic in a simple spatial context, with the parallel to the classic example being that treatment status is determined by the sum of independent realizations of an observable X and an unobservable U, and therefore correlated to both X and U even though they are not correlated with each other.

Suppose that counties lie at the integers along a number line and are indexed by n. We observe random variables X, D, and Y, with the interpretation that Y is the outcome, D is the treatment, and X is a (single) covariate. These variables are in turn determined by unobserved random variables U, V, Z^U, and Z^X according to equations below.

$$X_n = Z_n^X + \phi Z_{n-1}^X$$
$$U_n = Z_n^U + \phi Z_{n-1}^U$$

for some $\phi \neq 0$, where Z_n^U and Z_n^X are i.i.d. and drawn from the same distribution. Note that this means that X and U have the same distribution and are uncorrelated with each other. Also note that the construction of the variables generates a correlation in the values of X for neighboring observations but not for non-neighboring observations, and likewise for U.

Finally, D and Y are determined by the following equations:

$$D = 1\{X + U > 0\}$$
$$Y = \alpha + \delta D + \beta X + \gamma U + V,$$

where $E(V \mid D, X, U) = 0$ and where β, $\gamma \neq 0$. We are trying to recover δ, the effect of D on Y.

We define two control groups, one consisting of control observations which neighbor treated observations, and the other consisting of all other control observations.

Claim: In this model, proportional unobservables holds but selection on observables does not.

Proof. Selection on observables does not hold, since U is correlated with D conditional on X (it is correlated with D unconditionally, and not correlated with X) and since U enters into the Y equation. Therefore U is an omitted variable and regressing Y on D and X gives biased and inconsistent estimates of δ.

Proportional unobservables holds from the fact that X and U have identical distributions and roles in determining D, and therefore also group membership. Therefore the difference on the observable X between the treated group and each control group is exactly equal to the difference on the unobservable U. That is, $E(X \mid D_n = 1) - E(X \mid D_n = 0, n \in g) = E(U \mid D_n = 1) - E(U \mid D_n = 0, n \in g)$ for each control group g. (If there were an inequality, we could arbitrarily relabel X as U and obtain the reverse inequality, which would be a contradiction.) Some simple algebra completes the proof.[8] □

Remark 6 (Altonji, Elder, and Taber)**.** I previously mentioned that the proportional unobservables assumption is similar in tone to Altonji, Elder, and Taber (2005), who also use an assumption which can also loosely be described as the assumption that observables are a guide to the unobservables. In this model, they would be bounding γ to lie within a particular range. In this model, proportional unobservables holds regardless of the sign or magnitude of γ. Altonji, Elder, and Taber make no reference to the existence of more than one control group, and therefore no assumption about the relative value of U and X in those control groups.

Remark 7 (Generality)**.** I chose X and U to have the same distribution for convenience of exposition. However, proportional unobservables can hold under less restrictive circumstances. Specifically, under the assumption that X and U both enter linearly in the Y equation, it will hold whenever $E(X \mid D_n = 1) - E(X \mid D_n = 0, n \in g) = c[E(U \mid D_n = 1) - E(U \mid D_n = 0, n \in g)]$ for each control group g and some constant c. This equality holds, for example, when Z^U and Z^X are both normally distributed, with any mean or variance.[9]

D Simulation results

I consider a data-generating process in which Assumption 1 is imposed by construction.

[8] Let $c_g = E(X_n \mid D_n = 1) - E(X_n \mid D_n = 0, n \in g)$. Then $B_g = \beta c_g + \gamma c_g = (\beta + \gamma) c_g = \frac{\beta+\gamma}{\theta}(\theta c_g)$, where θ is the coefficient on X in a projection of Y on D, a constant, and X, such that $\theta c_g = B_g^X$.

Then for groups g and g', $\frac{k_g}{k_{g'}} = \frac{\sigma_g}{\sigma_{g'}} = \frac{\theta c_g(\beta+\gamma)/\theta}{\theta c_{g'}(\beta+\gamma)/\theta} = \frac{\theta c_g}{\theta c_{g'}} = \frac{s_g^X}{s_{g'}^X}$. This equality is the proportional unobservables assumption.

[9] Proof available upon request.

Table 8: Simulation: Estimates

	OLS	GMM: *ATT*	GMM: *a*
Mean	6.40	5.03	1.05
S.D.	(1.74)	(0.43)	(.07)

First line: mean of estimates over 1000 simulations. Second line: standard deviation of estimates. The true *ATT* is equal to 5. 200 observations per control group. First column is OLS estimate of the coefficient on D in a regression of Y on D and each X. Second column is an estimate of *ATT* using the GMM method outlined in Appendix B. Third column is an estimate of the slope a using GMM.

Table 9: Simulation: Hypothesis tests

Significance level:	OLS	GMM	J-test
$p = .1$	.60	.09	.07
$p = .05$	.56	.04	.04
$p = .01$	.49	.01	.01

Fraction of rejections over 1000 simulations. Each row tests a true null at the listed significance level. First column: Test that coefficient on D when Y is regressed on D and X is equal to 5. Second column: Test that $ATT = 5$ in GMM estimation. Third column: J-test of Assumption 1a in the GMM model.

In each simulation, a scalar λ_g for each control group is drawn from a standard uniform distribution. Observations are assigned to be treated ($D = 1$) with probability .5 and to each control group with equal probability. The data-generating process is

$$Y = 5D + \Sigma_{j=1}^{10} X_j + U + V,$$

where, for each X_j of the ten covariates $X_1, ..., X_{10}$, $X_j \sim N(0,1)$ for treated observations and $X_j \sim N(-3\lambda_g, 1)$ for untreated observations, $U \sim N(0,6)$ for treated observations and $U \sim N(-50\lambda_g, 6)$, and $V \sim N(3,8)$.

The distribution of point estimates are shown in Table 8, with OLS results included for comparison. The true value to be estimated is 5. We can see that OLS produces biased estimates, but the GMM estimates are very close to the true value.

Additional results are given in Table 9. These results show the probability of rejection of various (true) null hypotheses. The first columns test the null hypothesis that the effect of D on Y is 5, first in OLS and then in GMM. The last column uses the J-test to test Assumption 1, which holds by construction.

The results show that the GMM method performs well even when OLS fails. The J-test may be modestly undersized, however.

E Regression discontinuity demonstration

In this appendix, I show that the same selection ratio method (a) replicates traditional regression discontinuity estimates, which are considered persuasive by many researchers, and (b) can be used to measure treatment effects for populations away from the discontinuity. Because traditional RD techniques are considered highly persuasive by many researchers, I interpret (a) as evidence that my approach successfully measures treatment effects.

A common claim in both the academic literature and popular media is that incumbency confers an advantage in elections. When attempting to empirically assess the effects of incumbency, though, there is a clear endogeneity problem stemming from the fact that the same factors which caused a candidate to win the previous election (such as candidate quality and the partisan lean of a district) would also help that candidate to win the next election. Furthermore, some of these factors might be unobserved. Lee (2008) resolves this issue in the context of United States House of Representatives elections by using the fact that assignment of parties to incumbent status is a discontinuous function of their margin of victory (with a positive margin of victory creating incumbency and negative margin of victory generating no incumbency).

To resolve the concern that incumbents may only decide to run for re-election if their prospects are strong, Lee defines the research question to be as follows: What is the effect of the Democratic candidate winning the election in a Congressional district in election t on the Democratic candidate's margin of victory (which is allowed to be negative, and is measured in percent of the vote) in the same district in the next election, $t + 1$?[10] That is, the research question does not condition on the winning candidate in t running in election $t + 1$.

First, I will show how the selection ratio model can be adopted to estimate RD models. Second, I show that OLS recovers a treatment effect which is inconsistent with the results of any reasonable RD estimation exercise. Third, I show that applying the selection ratio approach with a parallel convergence assumption – using the same OLS specification – *does* recover an estimate of the local treatment effect which is within the range of traditional RD estimates.

[10] Lee also considers the effect of a Democratic win in period t on the Democrat's probability of winning in period $t + 1$, but I focus only on the margin of victory for simplicity.

E.1 Adaptation to RD context

The treatment D is a binary variable $DemWin$ for whether the Democratic candidate won the election in period t, and the outcome Y is the Democrat's margin of victory in the following election, denoted $DemMargin_{t+1}$. The running variable is the Democrat's margin of victory in period t, $DemMargin_t$, with positive values of $DemMargin_t$ assigning $DemWin = 1$ and negative values assigning $DemWin = 0$.

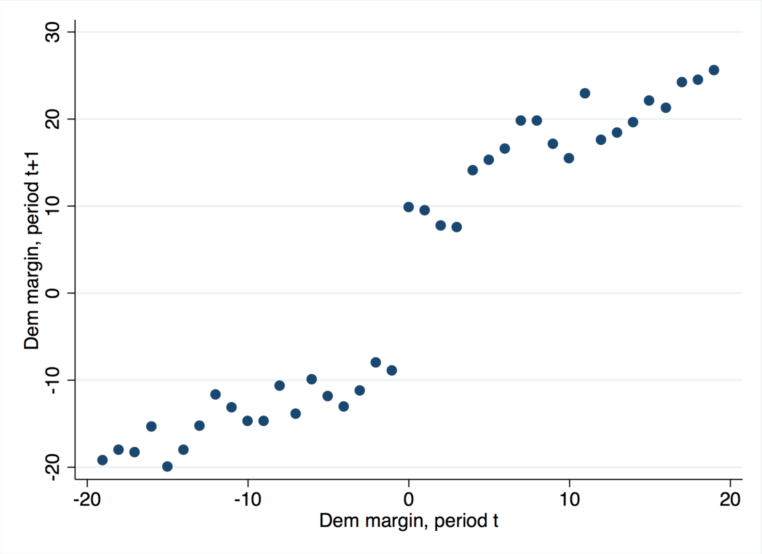
In the minimum wage context, I fixed a treated group and then constructed control groups according to their geographical distance from the nearest treated observation. In this case, I will fix a treated population and compare it to a series of control groups defined by the value of the running variable.

Notice that the basic setup required for a selection ratio approach – the existence of a treated group and a series of imperfect control groups – is satisfied in this context, even if the treated group is a set of observations where the election was not marginal. In other words, if the assumptions from Section 3 hold, it is still possible to identify treatment effects for populations away from the discontinuity. This is a departure from traditional RDD methods, which, with a few recent exceptions (Jackson 2010, Angrist and Rokkanen 2012, Wing and Cook 2013, Dong and Lewbel 2014, Rokkanen 2015) focus on estimating the average treatment effect for the population of individuals with values of the running variable right at the discontinuity. (This average treatment effect is usually referred to as the local average treatment effect, or LATE.)

Finally, the proportional unobservables assumption seems about equally plausible in this context as in the minimum wage context. In both cases, it seems intuitively plausible that taking control groups sufficiently close to the treated group (either spatially or along the running variable) should cause differences on both observables and unobservables to vanish; in fact, designs which use neighboring spatial regions are sometimes referred to as a "spatial RDD." In both cases, the claim that in the limit, both observable and unobservable differences would be absorbed lends itself to believing Assumption 1b, which states that a control group that is matched on observables would be matched on unobservables as well. The remaining part of the proportional unobservables assumption is Assumption 1a, which is testable.

Before proceeding further, it is worth briefly comparing this approach to existing approaches to estimating RD models. Current practice is to impose a functional form (or bandwidth, in the case of nonparametric estimation) on the data in order to estimate the treatment effect, and then to argue for the reasonableness of the model by showing that the estimation technique does not find discontinuities on the observables (e.g. Imbens and Lemieux 2008, Lee and Lemieux 2010). When applied in the RD context, my approach inverts this procedure by first assuming that a control group which was well-matched on observables would be well-matched on unobservables (the implicit assumption when a test of continuity on observables is taken as evidence for continuity on unobservables) and then extrapolating to this case using the functional form dictated by Assumption 4.

Figure 7: Binned scatterplot replication of Lee (2008)



E.2 OLS and traditional RD estimates

I use data from United States House of Representatives elections between 1942 and 2008. My version of this data is borrowed from Caughey and Sekhon (2011); see their paper for a discussion of how the data is cleaned.

First, I reproduce Lee's RD. Without any line of fit, the discontinuity appears in a binned scatterplot to be just over 15 percentage points. (Bins are by percentage point of margin of victory.)

Caughey and Sekhon raise the objection that some observables appear to be non-random even quite close to the threshold. This objection is disputed (Eggers et al. 2014), and I find that, using traditional methods, the existence of a significant measured effect in a placebo test for the effects of winning election t on margin of victory in $t-1$ depends on the specification used to estimate the discontinuity. To account for this potential violation, I also present traditional RD estimates of the effect of a Democrat winning a district in election t on the Democratic candidate's margin of victory in election $t+1$ net of the measured effect on the Democratic candidate's margin of victory in election $t-1$.

Tables 10, 11, and 12 show traditional RD estimates. Table 10 shows estimates from local linear regression over the range of plausible bandwidths. Each cell is an estimate of the discontinuity under a particular bandwidth and outcome variable. Tables 11 and 12 show polynomial estimates under a variety of choices of the order of the polynomial (cubic, quartic, and quintic) and the set of data used in the estimation (elections decided by 30 percentage points or fewer, 40 points or fewer, and 50 points or fewer). While the results are somewhat sensitive to the specification, the general

Table 10: Local polynomial results

Outcome variables	(1) bwidth=2	(2) bwidth=3	(3) bwidth=4	(4) bwidth=5	(5) bwidth=6	(6) bwidth=7
$DemMargin_{t+1}$	17.13	16.01	15.45	15.71	16.64	17.01
	(3.24)	(2.69)	(2.32)	(2.06)	(1.88)	(1.73)
$DemMargin_{t+1}$ $-DemMargin_{t-1}$	9.38	10.12	10.52	11.86	14.02	15.12
	(4.70)	(3.90)	(3.37)	(3.00)	(2.74)	(2.53)

Each cell is a local linear estimate of the effect of incumbency in a regression discontinuity design. Values are the estimated effect of the Democratic candidate winning in election t on the Democratic candidate's margin of victory (in percentage points) in $t+1$, in the first row, and on the difference between the Democratic candidate's margin in $t+1$ and $t-1$, in the second row. Standard errors in parentheses. Columns denote bandwidths, in percentage points.

theme is that estimates are roughly in the range of 15 percentage points.

I then estimate the effect of incumbency using OLS. I implement the following regression:

$$DemMargin_{it+1} = \beta_0 + \delta DemWin_{it} + X_{it}\beta + u_{it}$$

where i indexes districts and t time. The vector X of covariates includes Congressional Quarterly ratings of district competitiveness, ideology scores for the incumbent, the district's margin of victory for Democratic candidates for Congress and the presidency (averaged across the decade), dummies for a Democratic governor and Secretary of State, incumbency status and relative experience of the candidates in the period t election, and shares of black, urban, foreign, and government workers in the district.

Results are displayed in Table 13. The key result is that the effect of incumbency is estimated to be quite high: Incumbency is estimated to convey an advantage of 40 points in the next election! This is in stark contrast with the RD estimates. Additional specifications produce similar results.

E.3 Results with selection ratio approach

To estimate the effect of incumbency with the selection ratio approach, I bin elections according to the margin of victory of the Democratic candidate. I begin by constructing a treated group of districts where the Democratic candidate won by 0-5 points, which I then compare to control groups composed of observations where Republicans won by 0-5 points; 10-15 points; etcetera, up to 45-50 points. Then I

Table 11: Polynomial regression results for $DemMargin_{t+1}$

	maxdist=30	maxdist=40	maxdist=50
order=3	16.20	17.03	16.00
	(2.15)	(1.91)	(1.77)
order=4	15.52	16.83	18.25
	(2.58)	(2.30)	(2.12)
order=5	15.04	14.41	14.91
	(3.02)	(2.69)	(2.47)
Observations	4,994	6,467	7,561

Table 12: Polynomial regression results for $DemMargin_{t+1} - DemMargin_{t-1}$

	maxdist=30	maxdist=40	maxdist=50
order=3	12.46	15.13	16.64
	(3.51)	(3.06)	(2.78)
order=4	8.20	11.27	13.55
	(4.39)	(3.82)	(3.45)
order=5	6.05	6.88	8.94
	(5.25)	(4.58)	(4.14)
Observations	4,767	6,194	7,240

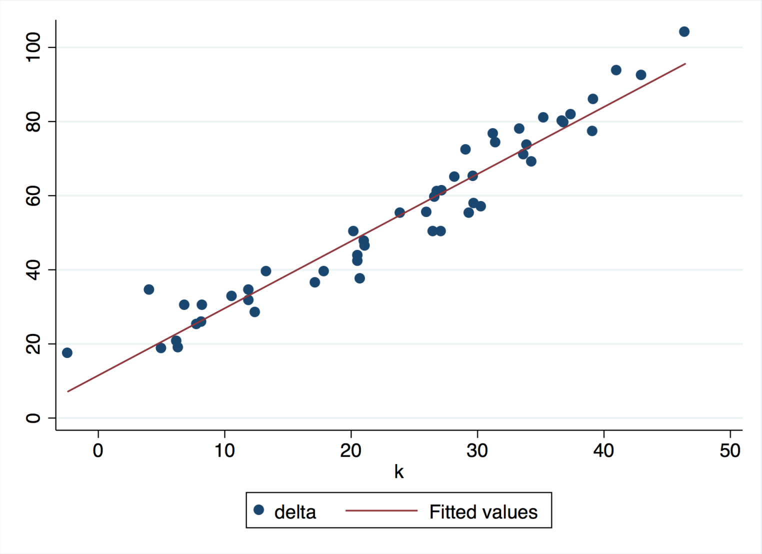
Table 13: OLS results, clustered by year

	(1)
Variables	DemMargin$_{t+1}$
DemWin	39.51**
	(3.268)
CQRating3	6.013**
	(1.253)
IncDWNOM1	8.668
	(4.860)
DifPVDec	19.49*
	(7.201)
DemMargin$_{t-1}$	0.299**
	(0.0415)
GovDem	-2.356**
	(0.766)
VtTotPct	-0.0140
	(0.0210)
UrbanPct	-0.00714
	(0.0341)
BlackPct	0.223**
	(0.0741)
GovWkPct	0.299
	(0.344)
ForgnPct	0.216
	(0.105)
RExpAdv	-6.827
	(3.804)
DExpAdv	-1.939
	(2.939)
DemInc	-4.756
	(4.131)
DemOpen	-0.566
	(3.760)
NonDInc	2.423
	(2.714)
SoSDem	4.178**
	(1.037)
Constant	-17.38**
	(5.084)
Observations	4,691
R-squared	0.643

Robust standard errors in parentheses

** $p<0.01$, * $p<0.05$

Figure 8: Selection ratio approach: $\widehat{\Delta}_g$ vs. $\widehat{B}_g^X$



repeat the same analysis, but fixing a treated group of districts where the Democrat won by 5-10 points; 10-15; and 15-20; and, inverting treatment to be defined as a Republican win, I repeat the same analysis with the parties reversed. I also perform the same analysis binning the treated group in the same way but allowing control groups to be binned by individual percentage points of margin of victory, i.e. districts where a Republican won by 0-1 points, 1-2 points, etc. Results are not sensitive to the binning choice – which is logical, given the linearity assumption.

With the groups defined, I then impose the proportional unobservables assumption. Y is the Democrat's (Republican's) margin in election $t+1$, D is a dummy for the Democrat (Republican) winning in period t, and X is the vector of controls used in the regression specification previously described (which produced a 40 percentage point effect estimate).

Figure 8 shows the graph of $\widehat{\Delta}_g$ against $\widehat{B}_g^X$ under the definition of the treated group being districts where Democrats won by 0-5 points and control groups are binned in five-point intervals, using elections decided by 50 or fewer points.[11] Points are labeled by the control group they represent, with 1 being the nearest control group and 10 being the most distant.

There are two important takeaways from these scatterplots. First, the points do fall roughly on a line, up to reasonable sampling error. This supports the model of unobservables. The J-test fails to reject the parallel convergence model for elections decided by 50 or fewer points, though the p-value is suspiciously good (.9996).

[11] The linear relationship weakens beyond 50 groups. This seems to occur at least partially because of small sample size; there are comparatively few elections which are contested but lopsided.

Table 14: Selection ratio results by population

	-(15-20)	-(10-15)	-(5-10)	-(0-5)	0-5	5-10	10-15	15-20
ATT	15.79	19.79	19.44	18.87	16.22	20.38	17.16	14.78
	(4.45)	(4.23)	(3.94)	(3.52)	(3.27)	(3.56)	(3.88)	(1.68)
a (1=selection	1.64	1.63	1.63	1.64	1.67	1.68	1.68	1.68
on observables)	(.14)	(.14)	(.14)	(.14)	(.15)	(.15)	(.15)	(.15)

GMM estimates. See Appendix B for details of the estimation. The dependent variable is the winning party candidate's margin (in percentage points) in the next election. Controls are the regressors listed in Table 13. Each column represents a treated population, defined by an interval of Democratic margin of victory in period t. Negative numbers represent districts where the Republican won in t; for these columns, treatment is defined as a Republican win in period t.

Second, a simple glance at the graph shows that the model of unobservables seems to yield a treatment effect estimate which is quite similar to RD estimates, and dissimilar to OLS estimates.

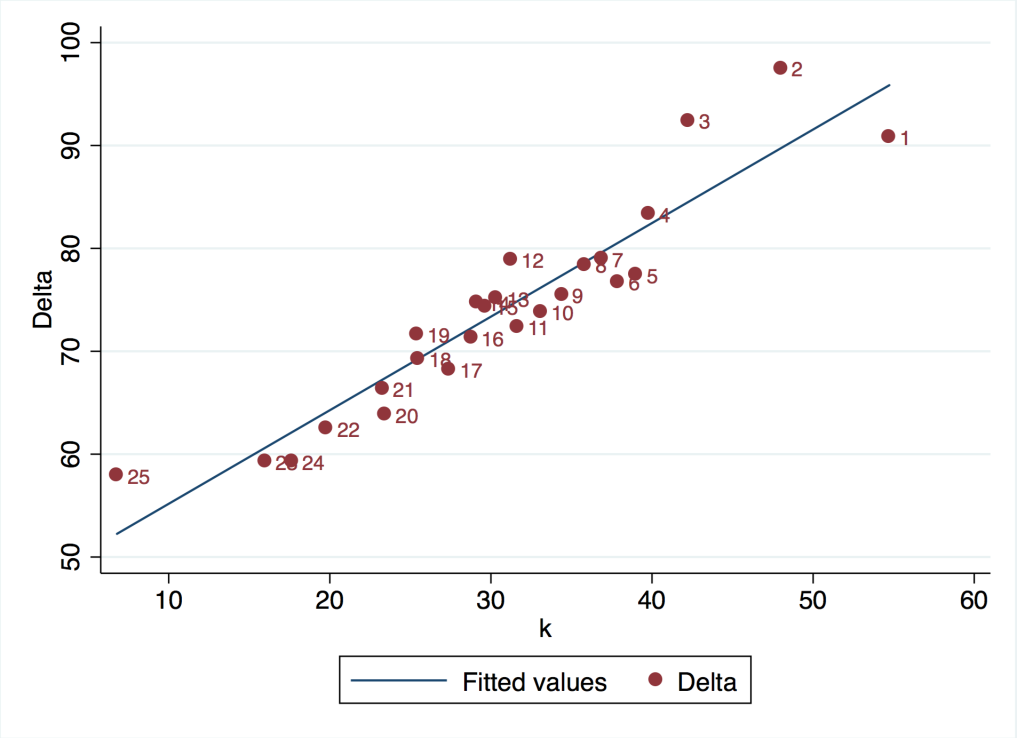
Results from the GMM estimations are shown in Table 14. In Table 14, each column represents estimation with a different choice of the treated group, with positive numbers representing a Democratic win in period t such that 0-5 are districts narrowly won by the Democrat and $-(0-5)$ are districts narrowly won by Republicans. Standard errors are shown in parentheses.

The GMM estimates are generally just over 15 points, even though it inputs exactly those same OLS results to create a weight for the index of observables. The estimated slope coefficient a is over 1.6, suggesting a large deviation from selection on observables, and we can comfortably reject that $a = 1$. (As we have seen, under selection on observables, a would be equal to 1.)

The importance of the definition of groups A key feature of this example is that, because of the way groups are defined, it is reasonable to believe that comparisons with almost no selection on observables would have almost no selection on unobservables as well. To illustrate the importance of this assumption, consider an alternative approach to defining control groups: quantiles of propensity scores.

To create multiple control groups, I divide the control observations into 25 quantiles by their estimated propensity score. (The propensity score is estimated using a logit regression of $DemWin$ on the same vector of covariates as before.) I use the universe of districts in which Democrats win as the set of treated observations in each comparison. I then estimate the relative selection of each group using the same parallel convergence assumption as before, using the exact same vector X and

Figure 9: $\widehat{\Delta}_g$ vs. $\widehat{B}_g^X$, propensity score case



the exact same regression to generate β. When the groups are defined this way, the graph of $\widehat{\Delta}_g$ against $\widehat{B}_g^X$ is once again fairly linear, but this time leading to an intercept of over 40 percentage points! Results are shown in Table 15.

I include this result as an illustration of the importance of Assumption 1b, and by extension the definition of groups. When control groups are defined solely by the value of observables, then there is no reason to believe that a control group which is well-matched on observables would also be well-matched on unobservables; otherwise, selection on observables would hold. It is not surprising, then, that this method yields a very similar result to simply estimating the effect of incumbency using selection on observables. Similarly, it is not surprising that variation across groups defined using observables is insufficient to reject the assumption of selection on observables.

This highlights situations in which the selection ratio method may be useful in the future: when the variable used to define control groups is potentially sufficient in the limit to control for any endogeneity problem by itself (such that confounding due to observables and unobservables are absorbed at the same time), but the control groups are defined in such a way that existing methods cannot make use of information about group membership.

Table 15: Selection ratio results: Propensity score groups

ATT	46.07 (2.30)
a (1=selection on observables)	.91 (.07)

F Additional concerns

As shown in Section 3, Assumption 1a is testable because of it is equivalent to the linearity of B_g^X and Δ_g. However, this test might incorrectly fail to reject in finite samples even when the assumption is not true. This appendix considers two reasons why it might be undesirable to claim that Assumption 1a is true whenever we fail to reject that B_g^X and Δ_g are linear in the data.

F.1 Potential for overfitting

One concern is that a determined researcher may attempt the J-test a number of times, choosing a different vector X of covariates each time. If each different vector of covariates gave an independent opportunity to find linearity of Δ_g and B_g^X, then it is quite possible that the researcher could always find some combination of observables (e.g. through the inclusion of higher-order terms and interactions) such that linearity would appear to be satisfied and the J-test would fail to reject the model. Therefore it is important to ask to what extent the linearity of Δ_g and B_g^X can be manipulated through the choice of X.

In practice, I find that the amount of such manipulation which is possible under reasonable specifications is quite limited in the minimum wage case. While there is no metric by which to quantify the potential for overfitting, a more detailed look at how overfitting might occur may give practitioners a sense of the extent to which it may be possible to overfit. In particular, I will compare the potential for overfitting to the problems of specification search in OLS and in difference-in-differences.

First, observe that the choice of X does not affect the estimate of Δ_g, which is just the raw difference in mean outcomes between the treated group and each control group. Therefore we only have scope to manipulate the appearance of linearity by altering values of B_g^X.

Analogy: OLS Most practitioners are familiar with situations in which an OLS coefficient of interest remains approximately stable across a range of specifications. Suppose we were estimating δ in the following equation:

$$Y = \alpha + \delta D + X\beta + \varepsilon.$$

Then

$$\delta = [E(Y \mid D = 1) - E(Y \mid D = 0)] - [E(X\beta \mid D = 1) - E(X\beta \mid D = 0)].$$

The first half of this expression does not change as we change the set of covariates. Therefore, when estimates of δ are stable across specifications, this means that the difference $E(X\beta \mid D = 1) - E(X\beta \mid D = 0)$ is also stable across specifications. Of course the object B_g^X is simply equal to the difference $E(X\beta \mid D = 1) - E(X\beta \mid D = 0, i \in g)$, and therefore will likely be difficult to manipulate in situations where the OLS estimate of δ is difficult to manipulate.

Furthermore, even if some amount of manipulation of B_g^X is possible by changing the set of covariates, a very specific manipulation is required to align each B_g^X linearly with respect to Δ_g. That is, an overfitting researcher must select X to manipulate the estimated δ in several regressions at the same time: The regression of Y on D net of $X\beta$ using only the treated group and the first control group; the same regression, using only the treated group and the second control group; etc.

This manipulation is complex. For example, consider that a necessary but insufficient condition for linearity is that Δ_g should be monotonic in B_g^X up to sampling error – and of course Δ_g cannot be manipulated through the choice of X.

So the challenge for an econometrician willing to engage in specification search until he simply finds linearity of Δ_g and B_g^X and can proceed to finding results is likely to be considerably more complex than the challenge of an econometrician who is using OLS and is willing to engage in specification search until she finds the treatment effect estimate that she prefers.

Analogy: Difference-in-differences The difficulty of manipulating the apparent linearity of Δ_g and B_g^X through the choice of covariates is also analogous to the difficulty of manipulating the apparent plausibility of the common trends assumption in difference-in-difference models through the use of covariates. In a difference-in-difference model, we have the equation

$$Y_{it} = \alpha + \delta D_{it} + \omega Z_i + X_{it}\beta + \gamma_t + \varepsilon_{it},$$

where γ_t is a time fixed effect and Z_i is a dummy for whether observation i is ever treated. The required common trends assumption is that ω has a stable value across time periods. Suppose we tested this stable value by separately estimating an ω_t for each period before treatment begins, as can be done explicitly and is often implicitly done by showing pre-trends net of the effect of covariates or by running placebo tests for effects in periods prior to the start of treatment. Then

$$\omega_t = [E(Y \mid Z = 1, t) - E(Y \mid Z = 0, t)] - [E(X\beta \mid Z = 1, t) - E(X\beta \mid Z = 0, t)].$$

Once again, the first half of this expression is fixed in the data. Therefore the manipulation required to pass the test of stability of ω_t is to select X such that $E(X\beta \mid Z = 1, t) - E(X\beta \mid Z = 0, t)$ is stable across time.

Note that this can equivalently be described as selecting X to generate β such that $[E(X \mid Z = 1, t) - E(X \mid Z = 0, t)]'\beta$ is stable across t, while in the case of proportional unobservables, we are concerned with matching $[E(X \mid D = 1) - E(X \mid D = 0, i \in g)]'\beta$ across g. In other words, in both instances, we have some fixed difference on the average value of X and any other observables in two populations, and an overfitting econometrician is attempting to select the set of observables X to multiply the difference in expectations to some number.

In the case of difference-in-differences, an overfitting econometrician gets one value of ω_t for free (since ω_t is always identical across time periods when there is only one time period) and then must use the choice of X to manipulate the exact values of ω_t to some exact value (up to sampling error) for the other $T - 1$ periods, where T is the number of pre-treatment periods. In the case of fitting linearity for the proportional unobservables assumption, the overfitting econometrician is instead given two values of B_g^X for free (since two points are always on some line) and then must choose X to fit exact values (up to sampling error) of B_g^X for the remaining $G - 2$ groups, where G is the total number of control groups.

F.2 Distortions from pre-testing

Assumption 1a – the choice of a linear specification for the relationship between group-level observables and group-level unobservables – is not strongly implied by theory. I have suggested that this assumption can still potentially be used for the reason that, while not obviously true ex ante, the content of the assumption is directly observable and testable. In the absence of sample-splitting, this type of argument requires that the same data be used for selecting an econometric model as for estimating it. In other contexts, this double use of the same sample can bias the estimates derived from the model, since the model might be more likely to be rejected when it would have yielded certain results (e.g. Bancroft 1944, Danilov and Magnus 2004). Further work will be required to determine with greater certainty whether pre-testing could theoretically distort the results from the selection ratio approach. However, I can offer some preliminary evidence on whether pre-testing is likely to be a serious concern.

Simulation evidence For pre-testing to distort the results, it must be the case that certain estimates are more likely when the linearity assumption fails to reject than when it is rejected. I use a simulation (described previously) to investigate the possibility that such a correlation exists for Type I errors in the pre-test.

To search for a relationship, I first regress the GMM estimate for each simulation on the J-test p-value for that simulation. There is no significant relationship. To check for differences in the dispersion of estimates for different p-values of the J-test,

Table 16: Tests for correlation of J-test and parameter estimates

	(1) OLS	(2) B-P	(3) Means: .1 reject	(4) Means: .05 reject	(5) Means: .01 reject	(6) K-S: .1 reject	(7) K-S: .05 reject	(8) K-S: .01 reject
p-value:	.51	.26	.52	.51	.80	.92	.62	.39

Each entry is a p-value for the null that model rejection is unrelated to estimates. The first column regresses point estimates on the J-test p-value and tests that the coefficient is equal to 0. The second column is the Breusch-Pagan test for heteroskedasticity in this regression. The third through sixth columns are tests of equality of the mean of point estimates between instances where the model is rejected by the J-test at each threshold and instances where the model is not rejected. The final three columns are Kolmogorov-Smirnov tests of equality of the distribution of estimates between instances where the model is and is not rejected by the J-test.

I use the Breusch-Pagan (1979) test of homoskedasticity and fail to reject that the dispersion of estimates is the same across all values of the J-test.

Then, refining this procedure to specifically compare cases where linearity is rejected at conventional significance levels to cases where it is not, I run three regressions of the GMM estimate on a dummy for rejection of the null hypothesis of linearity, with the three regressions varying by the size used in the linearity test (.01, .05, or .1). In each case, I do not find any significant relationship between the GMM estimate and the result of the J-test. Finally, I perform a two-sample Kolmogorov-Smirnov test of equality of distributions for the GMM estimates when linearity is rejected and is not rejected. Once again (and regardless of whether the .01, .05, or .1 threshold is used in testing the linearity assumption), I fail to reject that the distribution of estimates is the same. The results are shown in Table 16.

The same battery of tests also fails to recover any statistically significant relationship between the estimated standard errors and the results of the J-test.

Therefore the simulation results provide no indication that there is any relationship between the results of the linearity test and the GMM estimates.

Analogy: Difference-in-differences The common trends assumption is also frequently not principled before looking at the data, and therefore might be subject to similar concerns about pre-testing. As described in the discussion of overfitting, the common trends assumption holds when a single value of

$$\omega = [E(Y \mid Z = 1, t) - E(Y \mid Z = 0, t)] - [E(X\beta \mid Z = 1, t) - E(X\beta \mid Z = 0, t)]$$

fits across time periods t. Therefore sampling error in $[E(Y \mid Z = 1, t) - E(Y \mid Z = 0, t)]$, in $[E(X \mid Z = 1, t) - E(X \mid Z = 0, t)]$, or in β can lead to a mistaken assessment of whether ω_t is stable before treatment begins. Similarly, the proportional unobservables assumption requires that a single value of

$$ATT = [E(Y \mid D_i = 1) - E(Y \mid D_i = 0, i \in g)] - a[E(X\beta \mid D_i = 1) - E(X\beta \mid D_i = 0, i \in g)]$$

fits across groups g. (This equation is simply rewriting $ATT = \Delta_g - B_g$ with an expression for B_g.)

We can see that the objects to be estimated have similar structure and (since any error in ω usually shifts difference-in-differences estimates by the same amount)[12] the same relationship with the estimated treatment effect except for the inclusion of the slope term a. Note that a will be a function of the same expected value terms across groups, with a being overestimated or underestimated depending on the overestimation or underestimation of those same expected values. Since the effect of a on ATT is differentiable, and the effects of overestimates or underestimates of these expected values on a is also differentiable, then for sufficiently small estimation error, the effects on our estimate of ATT is still roughly linear in sampling error in these averages. Naturally this argument does not apply when sampling error is sufficiently large that deviations from these linearity arguments are not second-order.

[12] This is true if all untreated periods occur before treatment begins and we are assessing the common trends assumption using the stability of ω_t across those periods.