Pooled Synthetic Control Estimates for Continuous Treatments: An Application to Minimum Wage Case Studies

Arindrajit Dube* Ben Zipperer†

October 6, 2014

Abstract

We apply the synthetic control approach in a setting with multiple cases and continuous treatments. Using minimum wage changes as an application, we propose a simple distribution-free method for pooling across cases using mean percentile ranks, which have desirable small sample properties. We invert the mean rank statistic in order to construct a confidence interval for the pooled estimate, and we test for the heterogeneity of the treatment effect using the distribution of estimated ranks. We also offer guidance on model selection and match quality—issues that are of practical concern in the synthetic control approach generally and when pooling across many cases. Using 32 cases of state minimum wage increases between 1979 and 2013, we do not find a statistically significant effect on teen employment, with the mean elasticity close to zero. There is also no indication of heterogeneous treatment effects. Finally, we discuss some important practical challenges, including the ability to find close matches and the choice of predictors used for constructing a synthetic control.

*University of Massachusetts Amherst, and IZA
†Washington Center for Equitable Growth

1 Introduction

The synthetic control offers a data driven method for choosing control groups that is valuable for individual case studies (Abadie, Diamond and Hainmueller, 2010, hereafter ADH). This increasingly popular technique generalizes the difference-in-difference approach and also provides a semi-parametric version of the lagged dependent variables model, offering a way to control for time-varying heterogeneity that complicates conventional regression analysis. For a single state that receives a policy treatment, the synthetic control is the weighted average of untreated units that best predicts the treated state in the pre-treatment period, and the estimator is the post-treatment difference between the treated state and its synthetic control. Whereas conventional regression designs equally weight all units (conditional on covariates), units comprising the synthetic control typically receive unequal weights. Matching on pre-treatment outcomes allows the synthetic control approach to provide unbiased estimates for case studies even when there are multiple unobserved time factors, unlike the conventional difference-in-difference model which imposes a single factor assumption.

A growing number of papers have used the synthetic control approach to study topics as diverse as the impacts of anti-smoking legislation (ADH), immigration laws (Bohn, Lofstrom and Raphael 2013), and minimum wages (Sabia, Burkhauser and Hansen 2012). However, to date the applications have largely been restricted to estimating the effect of individual treated cases or to presenting numerous such estimates separately. For example, Billmeier and Nannicini (2013) use synthetic controls to investigate the effects of 30 country-level episodes of trade liberalization on their GDP. While the authors helpfully organize their presentation of synthetic and actual GDP trends by region and time period, the presentation of 30 individual pictures obscures their argument that later episodes of liberalization failed to boost GDP. Some episodes appear to raise, lower, or have no effect on growth, and it becomes difficult for the reader to gauge the magnitude of estimates or to draw statistical inference. Perhaps due to issues of space, for only 16 of the 30 case studies do the authors display figures relating to statistical inference, perhaps due to issues of space. Using synthetic controls, Campos et al. (2014) estimate a mean effect of EU integration on GDP, but the authors do not perform statistical inference on either the mean or individual case study estimates.

In this paper, we present a method for pooling synthetic control estimates in a setting with continuous and recurring treatment: state-level minimum wage changes. Because the intensity of the treatment varies across cases, we convert the estimates to elasticities by scaling them by the sizes of the minimum wage changes and then average these elasticities across events. A key contribution of the paper shows how the mean of the percentile ranks of the effects in the treated states *vis-à-vis* donor (or potential control) states can be used to judge the statistical significance for a pooled estimate, the Hodges Jr. and Lehmann (1963) point estimate. Pooling the estimates using their ranks is particularly useful since the exact distribution of the sum (or mean) of the ranks under the null is known, enabling us to perform exact inference that is valid for small samples. Additionally, we invert the mean rank statistic to construct the confidence interval for the pooled estimate. Our approach of pooling across treated units is closely related to the van Elteren (1960) stratified rank

sum test. It is also a natural extension of the placebo-based inference used by ADH for a single case study, where the distribution of a test statistic under the null is constructed by randomly permuting the treatment status of the donor units. Our inferential procedure has some similarity to Conley and Taber (2011); operating under the classic difference-in-difference context, they also use information from control units to form an empirical distribution under the null, and invert the test statistic to construct confidence intervals that are valid with a small number of treated cases. Finally, Dube, Kaplan and Naidu (2011) also use a average rank-based test that is valid for small samples in the context of financial market event studies.

Since percentile ranks of the estimates have a known distribution under the null hypothesis, exact inference is feasible not only for the mean but also distributional statistics as well. In this paper we use the Kolmogorov-Smirnov and Anderson-Darling tests to determine whether the distribution of ranked effects indicates heterogeneous treatment effects. One concern when pooling across events is that the quality of match between the treated and synthetic control unit may be poor in some instances. We assess the role of match quality by trimming on pre-intervention goodness of fit as determined by the mean squared prediction error (*MSPE*). A final contribution of the paper concerns the choice of predictor variables, since there is little guidance on this issue in the existing literature. We use a cross-validation criterion of minimizing *MSPE* among donor units to choose between alternative sets of predictors.

The minimum wage offers an interesting setting for applying the synthetic control estimator, since states receiving treatment have important differences that appear to vary over time, thereby confounding the standard fixed effects panel estimator (Allegretto, Dube, Reich and Zipperer, 2013). Since the synthetic control method depends on isolated treatment events with well-defined pre- and post-treatment periods, the approach can only utilize a limited amount of minimum wage variation available to conventional regression techniques. We select those events with no minimum wage changes two years prior to treatment and with at least one year of post-treatment data, which we consider to be the minimal requirement for measuring the policy's impact. Of the 215 state minimum wage changes during our 1979-2013 study period, only 32 meet our minimal criteria; on average these events have a 20 quarters of data prior to the intervention and 9 quarters afterward. While this is a limited number of events, we show that pooling across these 32 cases provides us with sufficient statistical power to detect economically relevant effects posited in the literature.

Our results show that the minimum wage changes were binding: 27 out of the 32 cases have positive effects on average teen wage, with a median elasticity of 0.22 and mean of 0.32. The pooled estimate is statistically significant at the one percent level using our mean rank test. Turning to teen employment, we find 15 positive elasticities and 17 negative ones. Both the median (-0.019) and mean (-0.039) estimates are small in magnitude. The mean percentile rank is 0.497 and the associated pooled Hodges-Lehman elasticity is -0.029. With a 5 percent confidence interval, we rule out a pooled employment elasticity more negative than -0.153. We also show that the distribution of the ranked employment estimates is consistent with the sharp null of zero effect everywhere, as opposed to an averaging of true positive and true negative effects across events. To address concerns

about match quality, we show that our results are similar after trimming our case studies to those with better pre-treatment fit. We do note that the treated states are generally some of the highest wage areas, making it difficult to find a convex combination of donors to closely match the average wage level in the pre-intervention period. However, this does not affect our ability to match their overall employment dynamics prior to the intervention.

Three papers in the minimum wage literature have particular relevance to the application of synthetic controls. An early precursor to synthetic controls is the study of California's 1988 minimum wage change by Card (1992). Card compares California with an aggregate control formed by four southern states and one metro area that failed to raise their minimum wages during the 1987-1989 period. Similar to the synthetic control approach, the aggregated control in Card (1992) roughly matches California on several pre-treatment labor market and demographic characteristics. However, Card's selection of the donor states is heuristic and not based on a solution to the explicit optimization problem underlying the contemporary synthetic control approach.

More recently, Sabia et al. (2012) uses the synthetic control approach to study the impact of the 2005 New York minimum wage increase. They ignore four other candidate treatment events that also began the same year in Florida, Minnesota, New Jersey, and Wisconsin. It is not clear what guided the authors' selection of New York as the sole case; in our results for these five states, we find that the New York event is associated with the most negative employment estimate. Sabia et al. (2012) also crucially fail to use *any* pre-treatment outcomes as predictors. Although any characteristics unaffected by the policy treatment are valid predictors under the synthetic control approach, some combination of pre-intervention outcomes should be included. Intuitively, if the synthetic control fits a sufficiently large set of pre-intervention outcomes, it is able to account for any number of time-varying factors.[1] As a result of omitting pre-intervention outcomes, the authors obtain an invalid synthetic control: specifically, employment paths for actual and synthetic New York *never* coincide during the entire 2000-2004 pre-treatment period.[2]

Neumark, Wascher and Salas (2013) use a matched panel estimator loosely based on the synthetic control method. They do not actually pool synthetic control estimates: rather, they use weights calculated using synthetic control to create matches and then estimate a panel regression with this sample. Allegretto et al. (2013) discusses in detail the serious problems with this approach. Most fundamentally, as the authors acknowledge, there is no econometric basis for this estimator. For example, they use residuals from an OLS regression of employment on the minimum wage as the predictor for calculating synthetic control weights. However, these are not valid predictors for a synthetic control study.[3] Additionally, their donors are sometimes themselves receiving treatment,

[1] Formally, the unbiasedness of the synthetic control estimator relies specifically on pre-treatment outcome balance between the treated unit and the weighted combination of donors (see Appendix B of ADH). The choice of exactly which pre-treatment outcomes and other characteristics to select as predictor variables is not obvious, *a priori*. We provide guidance for these decisions in section 3.2.

[2] See Figure 3 of Sabia et al. (2012). Relatedly, the authors try to account for the poor pre-intervention fit by using a *difference-indifference* with respect to the synthetic control. However, this is quite different from the synthetic control estimator, which requires the pre-intervention values in the treated and synthetic control units to be close.

[3] Besides being *ad hoc*, the use of OLS residuals as predictor variables does not have a heuristic justification. In the best case scenario, if the OLS estimates are unbiased, then the true and estimated residuals are uncorrelated with

violating a key assumption of the synthetic control approach. They also use a very short pre-intervention window (4 quarters) to calculate synthetic control matches, which makes finding a good match difficult. Finally, they use only a single quarter of post-intervention data to measure the policy impact—making this the shortest-run estimate in the minimum wage literature of which we are aware.

In contrast to these prior applications, we select 32 different events using clear (and reasonable) criteria for case selection, estimate synthetic controls for each treatment using a data-driven choice of predictors, and pool across these estimates using rigorous statistical procedures that are valid for small samples.

The rest of the paper is structured as follows. Section 2 reviews the synthetic control method and explains our proposed rank-based inference for the pooled estimate. Section 3 discuss our sample and the choice of predictor variables. Section 4 presents our findings, including the mean effect and the test of heterogeneity, as well as issues of match quality. Section 5 concludes.

2 Synthetic controls and multiple case studies

2.1 Synthetic control estimators

Consider the case of a single treated state ($i = 1$) that raises its minimum wage at time $t = t'$, where the outcome of interest Y_{it} is teen employment. Denoting the intervention as D, the synthetic control approach assumes a data generating process such that the observed outcome Y_{it} is a sum of the effect from the treatment, $\alpha_{1t}D_{it}$, and the counterfactual outcome, Y_{it}^N:

$$Y_{it} = \alpha_{it}D_{it} + Y_{it}^N = \alpha_{it}D_{it} + \boldsymbol{\theta}_t\mathbf{Z}_i + \boldsymbol{\lambda}_t\boldsymbol{\mu}_i + \boldsymbol{\delta}_t + \epsilon_{it}.$$

Here $\boldsymbol{\delta}_t$ is an unknown common factor with constant factor loadings across units, $\mathbf{Z}_i$ is a $(1 \times r)$ vector of observed covariates unaffected by the intervention, and $\boldsymbol{\theta}_t$ is a vector of unknown parameters. Like the standard fixed effects model, there is a common time factor $\boldsymbol{\delta}_t$. However, there is an additional $\boldsymbol{\lambda}_t\boldsymbol{\mu}_i$ term as well. Here $\boldsymbol{\lambda}_t$ is a vector of unobserved time-varying factors and $\boldsymbol{\mu}_i$ are the unknown factor loadings. Since the factor loadings can vary across states, treatment and control states need not follow parallel trends, conditional on observables. If we knew the true factor loadings $\boldsymbol{\mu}_1$ for the treated state, we could construct an unbiased control by taking untreated "donor" states whose factor loadings average to $\boldsymbol{\mu_1}$. Since we do not observe the factor loadings, the synthetic control procedure constructs a vector of weights $\mathbf{W}$ over J donor states such that the weighted combination of donor states closely matches the treated state in pre-intervention outcomes. This weighted combination of donors is called the synthetic control; as shown in ADH, the average factor loadings of the synthetic control thus constructed matches the loadings of the treated state.

Formally, for the treated state, define the $(k \times 1)$ vector of pre-treatment characteristics as

all covariates—making them uninformative as variables to construct a reliable control group. In contrast, if the OLS estimates are biased, so are the estimated residuals, making them potentially worse than uninformative.

$\mathbf{X}_1 = \left(\mathbf{Z}_1', Y_i^{K_1}, \ldots, Y_i^{K_L}\right)$, where $k = r + L$ and $Y_i^{K_l}$ are L linear combinations of pre-treatment outcomes. Analogously, define the $k \times J$ matrix $\mathbf{X_0}$ containing the same characteristics for the J donor states. The synthetic control procedure chooses donor weights $\mathbf{W}$ to minimize the distance between pre-treatment characteristics $\mathbf{X}_1$ and $\mathbf{X}_0$ of the treated state and untreated states. The distance equals the mean square prediction error (*MSPE*)

$$\sum_{m=1}^{k} v_m (X_{1m} - \mathbf{X}_{0m}\mathbf{W})^2$$

over k pre-treatment characteristics, and where v_m measures relative importance of the m-th predictor. Given the optimal weights w_j^* for each of the $j = 2, \ldots, N$ donors, the synthetic control at any time t is simply the weighted combination of donor employment $\sum_{j=2}^{N} w_j^* Y_{jt}$. The estimate of the employment impact α_{1t} is therefore difference between employment in the treated state Y_{1t} and the synthetic state $\sum_j w_j^* Y_{jt}$ at any post-treatment time $t \geqslant t'$:

$$\hat{\alpha}_{1t} = Y_{1t} - \sum_{j=2}^{N} w_j^* Y_{jt}.$$

When the intensity of treatment varies across events, elasticities offer a simple way of comparing across events. Moreover, the use of elasticities also connects our findings with other estimates in the minimum wage literature. For this reason, our key estimates of interest are the employment or wage elasticities of the minimum wage, defined as follows. For a single treatment event, we construct the synthetic control $\sum_j w_j^* Y_{jt}$ for the treated outcome Y_{1t}. In the post-intervention period $t = t', \ldots, T$, the average percent difference between the treated and synthetic control outcomes is given by

$$\hat{\beta}_1 = \frac{\frac{1}{T}\sum_{t=t'}^{T}\left(Y_{1t} - \sum_j w_j^* Y_{jt}\right)}{\frac{1}{T}\sum_{t=t'}^{T}\sum_j w_j^* Y_{jt}}.$$

Writing the percent minimum wage increase as

$$\Delta MW = \frac{MW_{t'} - MW_{t'-1}}{MW_{t'-1}}$$

we define the post-treatment elasticity η_1 to be the ratio

$$\hat{\eta}_1 = \frac{\hat{\beta}_1}{\Delta MW}.$$

As we describe below, it will be useful for placebo-based inference to construct analogous elasticities η_j for each of the donor states. Specifically, for each of the donor states $j = 2, \ldots, N$ we calculate the post-treatment difference β_j, this time using the remaining $N - 2$ donor states as donors for the synthetic control of state j. The placebo elasticity η_j is scaled by the actual minimum wage increase in treated state: $\eta_j = \beta_j / \Delta MW$.[4]

[4] As we discuss in Section 3, since some states change their minimum wage multiple times during the post-treatment

When there are multiple treatment events, we calculate separate event-specific elasticities $\eta_{e1} = \beta_{ej}/\Delta MW_e$ for the events $e = 1, \ldots, E$. Note that this construction of elasticities allows us to incorporate the fact that treated states vary both in their outcome levels and in their minimum wage treatment intensities. To aggregate across events we simply take the mean or median of these estimated elasticities. The mean treatment effect, for example, is equal to the mean elasticity

$$\overline{\eta} = \frac{\sum_e \hat{\eta}_{e1}}{E}.$$

2.2 Inference using the rank test with single and multiple events

We follow ADH in using placebo-based inference from randomly permuting the treatment status in donor states in order to assess the statistical significance of a single treated state's estimated elasticity. For each event, we estimate η_{ej} for every *donor* state j (excluding the actually treated state but using the same minimum wage change) and determine whether the elasticity η_{e1} for the treated state lies in the tails of the resulting placebo distribution formed by $\hat{\eta}_{ej}$ for $j = 2, \ldots, N_e$.

Equivalently, we summarize the relative position of the treated state's elasticity among the placebo distribution by using the percentile rank statistic $p_{e1} = \hat{F}_e(\eta_{e1})$, where $\hat{F}_e$ is the empirical CDF of the elasticities $\hat{\eta}_{ej}$ from event e.[5] Since the percentile rank is (approximately) uniformly distributed on the unit interval, we determine whether the rank of the treated state p_{e1} lies in the tails of the uniform distribution. Using a statistical significance level of five percent, we reject the null of $\eta_{e1} = 0$ precisely when $p_{e1} < 0.025$ or $p_{e1} > 0.975$. We note that the number of available donors limits the range of confidence levels we can implement for a single treated event. For example, many of our events have only twenty donors; in these cases we can only assess a ten percent level of significance. Using multiple events allows us to assess stronger levels of statistical confidence.

The above approach suggests a natural way of conducting inference in a pooled case study approach by constructing a test statistic $\overline{p}$ which is the the mean of the percentile ranks of individual events:

$$\overline{p} = \frac{\sum_{e=1}^{E} p_e}{E}.$$

The exact distribution of $\overline{p}$ can be calculated using the Irwin-Hall distribution of the sum of E independent uniform random variables. The sum of the ranks, $s = E \cdot \overline{p}$, has the the CDF

period, we simply define the minimum wage change to be the largest percent change between the post- and pre-treatment periods. We define the elasticity η_1 using the ratio of means in β_1 rather than the post-treatment mean of the percent changes $\frac{1}{T}\sum_{t=t'}^{T}\left(\frac{Y_{1t}-\sum_j w_j^* Y_{jt}}{\sum_j w_j^* Y_{jt}}\right)$ to avoid the possibility that the resulting elasticity has a different sign than the post-treatment mean of level changes in the numerator of β_1.

[5] To calculate the percentile p_{ei} of the ranked position r_{ei} of the estimated elasticity η_{ej} for state i in event e, we use the Weibull-Gumbel rule (see Hyndman and Fan, 1996): $p_e = r_{e1}/(N_e + 1)$, where N_e equals one plus the number of donor states, ensuring that the median effect within an event receives the rank 0.50 when the total number of states N_e is odd.

$$\Gamma(s;E) = \frac{1}{E!}\sum_{k=0}^{\lfloor s \rfloor}(-1)^k \binom{E}{k}(s-k)^{E-1}$$

where $\lfloor . \rfloor$ is the floor function.[6] Under the sharp null hypothesis of zero effect everywhere, the average of E ranks, $\overline{p}$, is distributed with mean 0.5. If $G(x;E) = \Gamma(x \cdot E; E)$ denotes the CDF of the mean of E uniformly distributed variables random variables, then for a statistical significance level of five percent, we reject the null hypothesis $\overline{\eta} = 0$ precisely when $G(\overline{p};E) < 0.025$ or $G(\overline{p};E) > 0.975$.

While the central limit theorem tells us that the distribution of the mean rank will converge to an appropriately scaled normal distribution, for small E we should prefer to use the exact distribution. Table A1 shows various percentiles of this distribution for $E = 1, \ldots, 35$. At 32 treatment events—the maximum number of case studies we will have in our study—a two-sided 5% significance test requires the mean rank to fall below 0.400 or above 0.600. We note that this method is closely related to the van Elteren (1960) stratified rank sum test, where the rank of each treatment is estimated using placebos associated with the stratum (i.e., event). The only substantive difference is that we use the percentile ranks of each treatment from each stratum, p_{e1}, instead of the ranked position r_{e1}, for transparency of the calculations; this choice potentially impacts the critical values when the number of observations (states) varies across strata (events) and the number of observations is also small. However, in practice, there is very little difference if we calculate the critical values taking into account the number of observations in each stratum used to calculate the ranks.[7] For concision, in the rest of this paper we will often we refer to the percentile rank as simply the "rank."

While there are alternative ways of doing pooled inference, we note some advantages to our approach. First, the rank-based pooling is a natural generalization of the single-case study based inference in ADH, who use the rank of the treatment effect for individual events. Second, the mean (or sum of) ranks has a known distribution under the sharp null, allowing for exact inference. This avoids reliance on large sample properties, and also avoids the empirical estimation of distribution of the statistic under the null—as would be the case were we, for example, to conduct inference for the mean elasticity. Third, and relatedly, the use of the mean rank $\overline{p}$ diminishes the impact of outliers as compared to the mean elasticity $\overline{\eta}$, which may be a particular concern given a small number of events. Fourth, within the class of rank sum tests, the ranks could be estimated without regard to strata, as in the case of the Wilcoxon (1945) rank sum test. However, stratification accounts for event-wise heteroscedasticity, which may be of particular concern given varying window lengths across events.

In section 4.5, we relax the approximation that the event ranks are independently and uniformly distribution by accounting for the finite number of donors, some of which overlap across events.

[6] See `http://en.wikipedia.org/wiki/Irwin-Hall_distribution`.

[7] Simulations of the mean of 32 percentile ranks calculated by the Weibull rule (with the appropriate number of donors for each event) result in 95% critical values 0.403 and 0.597, in contrast to 0.400 and 0.600 from the mean of 32 continuous uniforms. The associated rejection rate using the Weibull-rule-based distribution with continuous uniform critical values is 4.3 percent instead of 5.0 percent. See Table 10 for details.

We show that in our case this makes little difference to the calculation of the critical values, or the resulting confidence intervals for the treatment effect. For comparison, we also calculate the confidence interval using randomization inference on the mean effect (elasticity), as opposed to the mean rank; this too produces similar results.

One limitation of our approach is that we are testing the sharp null that effect is zero everywhere, as opposed to the *average* effect being zero. However, we address this concern in Section 4.3 by testing for heterogeneous treatment effects.

2.3 Inverting the rank test to form confidence intervals

We also invert the individual-event and mean rank statistic to estimate confidence sets.[8] These confidence sets show values of the elasticities which imposed as the null cannot be rejected as being equal to the estimated effect. For a single treatment event with estimated elasticity $\hat{\eta}_{e1}$, we use the percentile rank $p_{e1} = \hat{F}_e(\hat{\eta}_{e1})$ as the test statistic to determine statistical significance: we cannot reject the null hypothesis $\eta_{e1} = 0$ at the five percent level precisely when $0.025 \leq \hat{F}_e(\eta_{e1}) \leq 0.975$. Inverting this test, we ask for what values of τ does the adjusted response $\eta_{e1} - \tau$ appear free from treatment: when does $0.025 \leq \hat{F}_e(\hat{\eta}_{e1} - \tau) \leq 0.975$? The 95 percent confidence interval is the set of τ not rejected using the critical values 0.025 and 0.0975.

In the framework of multiple treatment events, we can apply a similar procedure to construct Hodges Jr. and Lehmann (1963) confidence intervals for the pooled effect, using the mean rank $\overline{p}$ as the test statistic to be inverted. We first calculate the adjusted responses $\eta_{e1} - \tau$ for all events $e = 1, ..., E$, and re-calculate event-specific ranks $\hat{F}_e(\hat{\eta}_{e1} - \tau)$. Define the mean adjusted rank

$$\overline{p}(\tau) = \frac{\sum_e \hat{F}_e(\eta_{e1} - \tau)}{E}.$$

The 95 percent confidence interval for the pooled effect is the set of τ such the mean adjusted rank $\overline{p}(\tau)$ lies within the critical values given by the mean of E uniform distributions. In other words, we find values τ such that $0.025 < G(\overline{p}(\tau); E) < 0.975$. Figure 1 illustrates this procedure for the estimated mean elasticity $\overline{\eta} = c$. The confidence interval is $(c - b, c + a)$ because $G(\overline{p}(c - (c + a)); E) = 0.05$ and $G(\overline{p}(c - (c - b)); E) = 0.95$.

Collapsing these confidence intervals yields the Hodges-Lehman point estimate, which we also refer to as the pooled estimate. In the case of a single event, the mean, median, and pooled effects are trivially the same, and so are the confidence intervals. In the case of multiple events, the mean, median and Hodges-Lehman point estimate and confidence intervals need not correspond. This is especially the case when outlying estimates of individual treatment events heavily influence the mean estimate. The robustness to outliers is one reason we prefer using the Hodges-Lehman confidence interval, as it is ultimately based on ranked location. Our primary estimates report the mean percentile rank, the pooled Hodges-Lehman point estimate, and the Hodges-Lehmann confidence

[8] Although ADH do not explicitly construct these confidence sets in the case of their single treatment event, they follow directly from their inferential procedure.

intervals. We also report the median and mean elasticities because of their natural interpretations.

Our inference assumes that the ranks of the treated states across events are independent draws. There are two potential concerns with this assumption, but overall we do not believe they represent major problems in our case. First, some events are from the same state, which may bring up a concern that the ranks of the events are not independent draws. However, while Y_{it} may be serially correlated, the same need not be true for $\hat{\eta}_{eit}$ across two events e' and e'' from the same state i in time periods t' and t''. If the synthetic control estimator is unbiased, and it successfully matches pre-treatment outcomes of both events, the post-treatment gap would from the two events are (by construction) uncorrelated: $E(\hat{\eta}_{e'1}, \hat{\eta}_{e''1}) = 0$.

The second and more serious concern is that the because the minimum wage increases often occur around the same time, two states with minimum wage increases may share many of the same potential donors. As a result, the ranks determined by the placebo distributions are not truly independent across treatment events. For two events e' and e'', the set of placebo estimates $\hat{\eta}_{e'qt'}$ and $\hat{\eta}_{e'qt''}$ from donor q may be correlated, in particular when $t' = t''$. In the extreme case, the donors and hence the placebo estimates $\hat{\alpha}_{e'qt'}$ may be identical. This induces a correlation in the ranks $\hat{F_{e'}}(\hat{\eta}_{e'1})$ and $\hat{F_{e''}}(\hat{\eta}_{e''1})$ even though $E(\hat{\eta}_{e'1}, \hat{\eta}_{e''1}) = 0$. However, in reality the overlap in donor pool is only partial, which mitigates this problem. As a way to bound the bias in our inference, we calculate critical values using placebo-law interventions that match the timing and donor overlap patterns of the actual 32 treatments in our sample. The results suggest that donor overlap has no substantial impact on critical values, justifying our use of the mean of independent uniform distributions.[9]

3 Minimum wage treatment events and empirical specification

3.1 Sample periods and timing of treatment

The synthetic control estimator requires a set of untreated or donor units for each treatment event. Since the vast majority of states were affected by the federal minimum wage increases, federal increases are not suitable for use with the synthetic control method: there are very few untreated donors that can be drawn from to construct a synthetic control for affected states. For example, 45 states changed their minimum wage at some point during the year of the 2007 federal minimum wage increase, leaving only 5 states as potential donors to form synthetic controls.

To maximize the number of treatment events, we consider the entire 1979-2013 period available using Current Population Survey (CPS) data. We focus on teen employment and wages, as many 16- to 19-year olds have wages near the minimum. During this period, almost 38 percent of teens received wages within 10% of the statutory minimum wage, compared with about 5% of workers aged 20 to 64. While there is considerable debate regarding the size of teen employment effects, we

[9] A Monte Carlo simulation of placebo-law interventions obtains 95% critical values for the mean percentile rank of 32 events of 0.392 and 0.607, in contrast to 0.400 and 0.600 using the mean of 32 independent uniform random variables. Using the latter critical values with the placebo-law simulation distribution implies a rejection rate of 6.9 percent instead of 5.0 percent. See Appendix for details.

expect to find significantly positive effects on teen wages. The high incidence of minimum wage workers among teens makes them the most frequently studied group in the minimum wage literature (e.g., Neumark et al. 2013, Allegretto et al. 2013). For outcome variables we calculate quarterly state-level teen employment-to-population ratios and average wages using the CPS.[10] Although annual state means would contain less noise, they would correspondingly limit the number of pre- and post-treatment observations; moreover, not all minimum wage increases occur during the same part of the calendar year.

The top panel of Figure 2 shows all quarterly minimum wage changes during the study period.[11] During this period the federal minimum wage increased nine times, indicated by the vertical lines in the Figure. Aside from federal minimum wage changes, 33 states in this period raised their minimum wage 215 times. Many states increase their minimum wage frequently, often on an annual basis. To utilize the synthetic control method, we limit the sample of usable treatment events to those with well-defined pre- and post-treatment periods. We select those events with no minimum wage changes two years prior to treatment and with at least one year of post-treatment data. We also limit the sample to minimum wage increases of at least 5 percent, and to treatment events with at least 10 potential donors or untreated states. These restrictions yield the 32 treatment events in the top panel of Figure 2 labeled in dark text.

The eligible events have valid pre- and post-treatment periods of varying length. West Virginia, for example, has many years of data prior to its minimum wage change in 2006q3 available but only one year of post-treatment data. By contrast, California's treatment in 2001q1 allows only two years of clean pre-treatment data but many years of post-treatment data. Also, California's post-treatment period includes an additional minimum wage increase in 2002q1. To simplify choices, for each event we select its "maximal" pre-treatment period available from 8-32 quarters; having done so, we then select each event's maximal post-treatment window from 4-12 quarters. The bottom panel of Figure 2 illustrates these pre- and post-treatment selections in blue and red, respectively, with circles indicating the times of treatment. Two features stand out. First, while the pre-treatment period contains no minimum wage increases by definition, the post-treatment period includes multiple minimum wage changes—states that raise their minimum wage often do so again within the next year or two. Table 1, which lists all 32 treatment event configurations that form the basis for our primary specifications, shows that most events include multiple minimum wage increases. There are four events whose post-treatment period includes four minimum wage increases. For this reason our treatment intensity definition incorporates the maximum minimum wage in the post-treatment period.

Second, Figure 2 there are three states in the 2000s with recurring minimum wage changes

[10] For employment outcomes we use the Unicon CPS extracts for the monthly Basic Survey (`http://unicon.com`). Wage data is only available in the outgoing rotation group subset of these data; for wage data, we use the NBER Merged Outgoing Rotation Group extracts (`http://www.nber.org/morg/annual/`). We calculate wages as hourly earnings or, if these are not reported, weekly earnings divided by usual weekly hours. State-quarter-level averages use the sampling weights.

[11] We thank Sylvia Allegretto for providing monthly historical minimum data, which we convert to quarterly averages.

where the post-treatment period of one minimum wage change overlaps with the pre-treatment period of a later minimum wage change: Hawaii, Rhode Island, and Vermont. For example, Hawaii's post-treatment period for its 2002q1 treatment overlaps with the pre-treatment period of Hawaii's 2006q1 treatment. Delayed effects from the former 2002q1 treatment could in principle violate the assumption that, for the latter 2006q1 event, Hawaii's pre-treatment period is absent from treatment. On the other hand, the pre-treatment period of Hawaii's 2006q1 is absent from treatment using our original definition that it contains no minimum wage changes. For our primary specifications we will include all 32 events, but we will also describe results excluding the three events of Hawaii 2006q1, Rhode Island 2006q1, and Vermont 2004q1.

There is a trade-off between window length and the number of events and donors. Allowing relatively short pre- and post-treatment periods maximizes the number of treatment events but, at the same time, may reduce the quality of the estimated counterfactual, as there is less pre-treatment data informing the selection of synthetic controls. On the other hand, lengthy pre-treatment periods limit both the number of events and potential donors, thereby reducing the credibility the resulting estimates. When we limit our treatment events to those with more restrictive pre- and post-treatment window lengths, we sharply reduce the number of case studies, as Table 2 illustrates. The first line in Table 2 is our primary configuration: 32 events with at least 8 and 4 quarters of respective pre- and post-treatment data. Requiring pre-treatment and post-treatment windows of at least 16 and 8 quarters, respectively, curtails the number of case studies to 17. In terms of restricting the donor availability, the configurations in Table 2 show only a small amount of variation. As we limit the pool of case studies to the most restrictive window configurations, the mean minimum wage treatment rises a small amount, from about a 19 percent increase to an increase of about 23 percent. Our primary results use the maximal window configuration with 32 events, but we explore how the alternate window configurations affect our results in Section 4.3.

3.2 Specifying predictor variables

Any characteristics unaffected by the policy intervention are valid predictors under the synthetic control approach, including demographic and industrial compositions or other economic attributes of the region. However, the unbiasedness of the estimator relies on the predictors including some linear combination of the pre-treatment values of the outcome of interest. There are two related questions when it comes to these predictors. First, exactly which variables should one include in the set of predictors? Second, what weight should one place on each of those predictor variables when estimating the donor weights? ADH provides a simple answer to the second question of how best to determine the weights on specific predictors within a set, which we describe first. Then we tackle the more challenging question of what predictors—and specifically what pre-treatment outcomes—one should include in this set.

For a given event e, the optimal donor weights are defined as

$$\mathbf{W}_e^*(\mathbf{V}_e) = \underset{\mathbf{W}_e}{\operatorname{argmin}} \sum_m^k v_m \left(X_{e1m} - \mathbf{X}_{e0m}\mathbf{W}_e\right)^2. \tag{1}$$

The optimal weights depend on the predictor importance matrix $\mathbf{V}_e = \{v_{em}\}$ selected by the researcher. For example, $\mathbf{V}$ might weight each predictor equally. Instead we follow the suggestion in ADH to select $\mathbf{V}_e^*$ such that the resulting synthetic control best fits pre-treatment outcomes. We therefore solve the joint (nested) optimization problem given by equation 1 and the equation

$$\mathbf{V}_e^* = \underset{\mathbf{V}_e}{\operatorname{argmin}} \sum_{t<t_e'} \left(Y_{e1t} - Y_{e0t}\mathbf{W}(\mathbf{V}_e)\right)^2$$

which minimizes pre-treatment fit. Our results use the "optimal" choice of weights $\mathbf{W}_e^*(\mathbf{V}_e^*)$ instead of alternatives such as manually specifying weights for predictors, or using the computationally less intensive methods available to users.[12]

But exactly which sets of pre-treatment outcomes and other characteristics should the researcher choose as predictors? When computationally feasible, perhaps the simplest strategy is to include every pre-treatment outcome in the predictor set $\mathbf{X}$. In their study of the effects of Arizona's 2007 Legal Arizona Workers Act, Bohn et al. (2013) employ this strategy with annual CPS data, using every pre-treatment value of the 1998-2006 non-citizen Hispanic share of the population, in addition to other industrial and demographic shares.

Within the pre-intervention sample, one cannot do any better in terms of pre-intervention *MSPE* than to include every pre-intervention outcome. However, this will not be true when predicting out of the pre-intervention sample, which is ultimately the object of interest. Matching on higher frequency pre-intervention data may actually produce less reliable synthetic controls. For example, our study uses quarterly CPS data, we risk matching on noise when using as predictors every quarterly pre-treatment value of teen employment-to-population ratios or average wages. As a result, we also consider the predictor set $\mathbf{X}$ consisting of annualized averages of the pre-treatment outcome.[13]

Different sets of predictors may result in different synthetic controls, and there is little explicit guidance in the synthetic control literature to assess predictor choice. We consider four different

[12] We implement the synthetic control approach in Stata using the synth package with nested optimization and allopt starting point checks for robustness: `http://www.mit.edu/~jhainm/synthpage.html`. There is a option for using a less computationally intensive but less reliable "regression-based" predictor weights. In our experience, the regression-based weights can produce worse fit, and the nested optimization uses regression-based weights as an initial set of values for optimization. The optimization process always converges to a solution for the 32 actual treatments, but it fails to converge for a small subset of donor-based placebo treatments – in only these cases do we resort to the regression-based predictor weights. Failure to converge on an employment solution occurs 0.01, 0.06, 0.05, and 0.01 percent of the time for synthetic control models 1, 2, 3, and 4, respectively. For wages, these failure rates are 1.73, 0.01, zero, and 1.53 percent.

[13] Here, annualized averages refer to the mean of the first through fourth quarter before treatment, the mean of the fifth through the eighth quarter before treatment, *etc.* For Minnesota's 2005q3 treatment, say, these refer to the 2004q3-2005q2 mean, the 2003q3-2004q2 mean, *etc.*

predictor sets $\mathbf{X}$, which vary according to whether we include every quarterly or annualized pre-treatment outcome, and whether we include other pre-treatment average demographic, labor market, industry shares.[14] These predictor sets are summarized at the bottom of Table 3. Using teen employment as an example outcome, predictor set 1 is all quarterly pre-treatment values of teen employment-to-population ratios. Predictor set 2 is all annualized pre-treatment employment-to-population ratios. Predictor set 3 includes all annualized pre-treatment employment *and* wage outcomes. Predictor set 4 adds to predictor set 3 the pre-treatment demographic, labor market, and industry shares described above. We note that when every quarterly pre-intervention outcome is included in the predictor set, inclusion of other predictor variables is redundant when weights on those predictors are calculated optimally using nested optimization. For this reason, it only makes sense to include variables such as industry or demographic shares when using annualized and not quarterly pre-treatment outcomes.

To identify the "best" choice for $\mathbf{X}$, we use a cross-validation procedure to choose from different sets of predictors. Recall that in creating synthetic controls for each event, the pre-intervention observations of donors effectively form a "training sample" upon which we select synthetic control donor weights as well as predictor weights for a given set of predictors. Here, when choosing the most reliable set of predictors, we use the post-intervention observations of the donors as our "validation" sample to evaluate prediction error associated with a given set of predictors. For a given predictor set $\mathbf{X}$, we calculate the post-treatment mean-squared prediction error (*MSPE)* for each donor j given by

$$MSPE_{ej} = \frac{1}{T_e} \sum_{t=t'_e}^{T_e} \left(Y_{jte} - \sum_q w^*_{eq} Y_{eqt} \right)^2$$

where t'_e begins the post-treatment period in event e, and where q indexes the available $N-2$ donors for (untreated) state j. We define the average *RMSPE* to be the square root of the mean of this quantity across all donors for all 19 events. The optimal model will yield the smallest post-treatment *RMSPE*, so predictor sets $\mathbf{X}$ with higher average *RMSPE* in the post-treatment period indicate models with worse performance in the sample of untreated donors.[15]

Table 3 reports the average donor *RMSPE* in the training sample for both the post-treatment and pre-treatment periods across four candidate specifications for predictors. Predictor set 1, which uses quarterly pre-treatment outcomes, naturally obtains the best pre-treatment fit to quarterly employment or wages when compared to predictor sets 2 through 4, which try to fit quarterly frequency data using annualized pre-treatment outcomes. Incorporating both annualized outcomes and additional controls improves pre-treatment fit relative to using only one annualized outcome: for employment, pre-treatment *RMSPE* falls from about 0.040 in specification 2 to about 0.035 in

[14] The demographic and labor market variables are the pre-treatment means of white, black, female, and married shares of the teen population, the teen population share, the share of the overall population with a college degree, and the overall unemployment rate. Industry variables are the employment shares in agriculture & mining, construction, manufacturing, wholesale & retail trade, transportation & utilities, information/finance/professional/business services, education & health services, leisure/hospitality/personal services, and public administration.

[15] We do not use the treated states for this cross-validation exercise because use of the post-treatment period in these states would require us to also have a valid estimate of the treatment effect.

specifications 3 and 4.

While using every pre-treatment outcome by definition maximizes goodness-of-fit in the pre-intervention sample, the same need not hold out of sample. In terms of post-treatment fit, the specification 4 is actually mildly preferable to specification 1. Using both annualized outcomes and demographic, labor market, and industry shares in specification 4, the post-treatment *RMSPE* for teen employment is about 0.0472, compared to the *RMSPE* of about 0.0478 for quarterly predictors. For wages, post-treatment *RMSPE* falls more—from 0.7911, when using quarterly outcomes in specification 1, to about 0.7758 in specification 4. The observed reduction in *RMSPE* —although admittedly small—is consistent with our *a priori* concerns about noise in the aggregations of quarterly CPS data, leading us to select specification 4 as our primary configuration. Yet because the small measured reduction in *RMSPE* makes our preference for this model somewhat weak, we explore the robustness of estimates across all sets of predictors in section 4.3.

4 Synthetic control estimates of minimum wage effects

4.1 Donors selected by synthetic control

Conditional on other covariates, conventional regression effectively assigns equal weights to the states the researcher selects as potential controls. By contrast, the synthetic control approach selects a convex combination of donor states based on that combination's pre-intervention fit to the treated state. For our sample of treatment events, we observe that the synthetic control procedure on average assigns greater weights for nearby donors, suggesting that nearby states generally form better counterfactuals than do distant states. To illustrate this, Table 4 compares average per donor weights for those donors near to and further away from the treated state. For each treated state, some donors reside within the same Census region as the treatment, whereas other donors lie outside that region. We first calculate the sum, across events, of all weights for these within-region donors, and then we divide this sum by the total number of within-region donors. The first entry in Table 4 is the resulting within-region per donor weight, 0.050, when the outcome of interest is the teen employment-to-population ratio. For outside-region donors, the per donor weight is 0.027. Calculating per donor weights in this way adjusts for the fact that the number of potential donors within or outside a given area varies across treatments.

The primary statistic of interest in Table 5 is the ratio of within-area to outside-area per donor weights: the relative per donor weight. For the employment-to-population ratio, the relative weight is 1.836, indicating that donors within the same Census region as the treated state are, on average, assigned weights almost twice as high as donors from outside the the same Census region. Relative weights tend to increase as we restrict the relative distance band. Same-Census-division donors – a finer aggregation level – receive even more weight, with relative weights of about 3.0 and 2.5, for teen employment and wages, respectively.[16] Donors within 1000 miles receive 1.3 to 1.5 times as

[16]The US Census Bureau partitions the country into four Census regions and nine Census divisions: https://www.census.gov/geo/www/us_regdiv.pdf.

much weight, and donors within 500 miles receive about 2.0 times as much weight. On the whole, the evidence shows that nearby donors form better synthetic controls.

4.2 Primary results

We begin with reporting the estimates for each of the 32 treatment events in Table 5. First, the results appear to indicate a positive impact of minimum wage increases on average teen wages. While the wage elasticity estimates range from -0.188 to 1.337, we find that 27 out of the 32 estimates are positive and almost half (15) are strictly greater than 0.25. As described earlier, the reported rank is the percentile rank of the treated state's elasticity relative to the placebo distribution. Six of the 32 estimated wage effects are statistically significant at the 10 percent level.

Turning to teen employment, the estimated elasticities range from -1.999 to 0.829, although 14 of the 32 events have employment effects less than 0.2 in magnitude. Consistent with a constant zero treatment effect, or heterogeneous effects centered around zero, 15 out of the 32 employment estimates are positive. There are two statistically significant employment effects: both Massachusetts (-0.456) and Oregon (-1.081) have negative elasticities that are significant at the 10 percent level. Highlighting the imprecision of individual case studies, we find that the confidence intervals are wide, with an average spread of 1.669 (1.875) for employment (wage) elasticities.

The presence of occasionally very large estimates is partly due to the non-normality of the distribution of synthetic control estimates. To show this, Figure 5 compares probability densities of the *donor* employment and wage elasticities to normal probability densities. For both employment and wage outcomes, the placebo distribution formed by the donors is clearly non-normal: although centered relatively close to zero (about -0.01 and 0.01 in employment and wages, respectively), extreme values give the placebo distribution fatter tails. The estimated kurtosis is 5.82 for donor employment elasticities and 58.37 for wage elasticities, compared to the value of 3.0 for any normally distributed sample. The especially severe departure from normality in wage estimates is partly due to extreme estimates in this space with poor *pre*-intervention fit, as discussed in Section 4.3. Shapiro-Wilk tests clearly reject the null of normality in both cases, with p-values close to zero. In the presence of such fatter tails, the placebo-based confidence intervals are wider than those formed under large sample assumptions.

The imprecision of individual estimates highlights the gains from pooling case studies. Table 6 presents our preferred aggregated results as both the mean elasticity and median elasticity across events. As discussed earlier, we also present the mean ranks, and the associated Hodges-Lehmann confidence intervals; both the median estimate and the Hodges-Lehman interval are less swayed by potential outliers, a concern that is highlighted by the presence of fatter tails. The median and mean employment elasticities for the 32 treatment events are relatively small: -0.019 and -0.039, respectively. Across treatment events, the mean employment rank is 0.497, essentially what would be expected under the null of a zero treatment effect. The pooled Hodges-Lehman effect is small at -0.003 and statistically insignificant, as the mean rank falls between the cutoffs (0.400, 0.600) derived from the 2.5th and 97.5th percentiles of the mean of 32 uniformly distributed variables.

The associated 95 percent confidence interval is (-0.153, 0.104). Although somewhat wide, pooling across the 32 events nonetheless allows us to draw economically meaningful inference, and rules out a substantial portion of the old "consensus" estimate of -0.1 to -0.3. (Brown 1999).

These small-to-zero aggregated employment effects contrast sharply with those for wages, where the median and mean elasticity are 0.220 and 0.317, respectively. The pooled wage elasticity of 0.265 is statistically significant at the 1 percent level, as the mean rank is 0.763. The associated confidence interval rules out wage effects smaller than 0.174 and larger than 0.382.

Figure 3 illustrates these aggregate effects by showing the time path of the mean annualized employment and wage elasticities, both before and after the minimum wage increase.[17] The top panel shows the mean annualized employment elasticities ranging from 8 years prior to the minimum wage increase (i.e., quarters -32 through -29) to 3 years afterward (i.e., quarters 8 through 11). The middle panel shows the analogous estimates for wages. The bottom panel shows the number of treated states used for the estimation of the elasticity of each 4-quarter bin, as well as the associated proportionate change in the minimum wage.

For employment, all pre-treatment point estimates but the 8th year lead (i.e., quarters -32 through -29) are small in magnitude, adding validity to our research design. After the minimum wage increase, employment nominally falls, but the elasticity remains less than 0.1 in magnitude. For wages, pre-treatment elasticities are centered around zero until about the first two years prior to treatment (quarters -8 through -5), at which point we detect a statistically significant elasticity of about 0.1 on employment. Positive pre-treatment elasticities for wages suggest that the synthetic control research design may not be as reliable for wage impacts – partly because wages for minimum wage increasing states are generally higher than potential donor states, making good matches difficult. At the same time we do find a sharp increase in teen wages at the time of and after the minimum wage increase. The Hodges-Lehmann point estimate for the teen wage elasticity lies between 0.2 and 0.4 in the post-treatment period. Approaching 0.40, the pooled wage elasticity is high after three years of treatment, but this is not inconsistent with the fact that nearly 38 percent of teens during the 1979-2013 period earned within 10 percent of the minimum wage.

Before exploring match quality and robustness issues, we take stock of our baseline estimates in Figure 4. The Figure shows all 32 individual employment elasticities (vertical axis) and wage elasticities, along with the mean and pooled effects. Overall, while the estimates appear noisy, there is very little relation between the magnitude of the wage elasticity and employment elasticity. In particular, the dotted line shows the locus of unitary elastic labor demand ($\eta^{emp}/\eta^{wage} = -1$), where the wage effects of the minimum are completely offset by the employment effects, ignoring any changes in hours. Of the 32 treatment events, 21 lie clearly above this locus, as do the mean

[17] Specifically, we annualize actual treated state and synthetic control outcomes by taking the event-specific mean of these values at every pre- and post-treatment four-quarter interval. The percent difference between these values, divided by the actual minimum wage increase, forms the event-specific elasticity at each time interval. The figure displays the mean elasticity across events at each time interval. Performing the analogous calculation for the donors, we then construct event-time-specific percentile ranks, which we invert to calculate Hodges-Lehmann point estimates and 95% confidence intervals, where the latter use mean uniform cutoffs from Appendix Table A1 with the appropriate number events.

and pooled effects. Discounting issues of statistical precision, the point estimates in the Figure seem inconsistent with the idea that negative employment effects are more likely when there is a more binding minimum wage: only when wage elasticities fall below the pooled estimate of 0.265 do we observe events with employment elasticities below the elastic demand locus.

4.3 Accounting for match quality

Pre-treatment fit to actually treated state is the key criterion for the credibility of a synthetic control analysis. The extent to which a synthetic control matches the treatment unit in the pre-treatment period indicates how well it accounts for time-varying confounders. For a single case study, the pre-treatment match quality is usually apparent: for example, the synthetic control for New York in in Sabia et al. (2012) never coincides with the actual treated state. However, when pooling across many cases, it may be difficult to evaluate and account for match quality merely by inspection. Some of the the synthetic controls for the 32 treatment events in this paper also suffer from poor pre-treatment fit, but our pooling of estimates does not account for differences in match quality.

To assess this issue more systematically, we progressively exclude events with particularly poor pre-treatment fit and examine how this affects our post-treatment elasticities. For each treatment event, we calculate a pre-treatment *RMSPE* between the synthetic and actual treatment outcomes—this is our measure of pre-treatment fit. We also calculate an estimated pre-treatment elasticity, defined just as our conventional treatment effect η_1 except calculated over the pre-treatment period (and scaled by the actual minimum wage increase). Next, we trim our sample of case studies on pre-treatment fit and examine how the trimming affects the pooled pre-treatment and post-treatment elasticities.[18]

Figure 7 shows how pre- and post-treatment elasticities vary after trimming up to 11 events (about one-third of our sample). The top panel shows that the post-treatment Hodges-Lehmann point estimate remains relatively for both employment and wages. One exception is the mean wage effect, which rises from to just below 0.40 after removing 11 events with the worst match quality. We discuss how the mean wage effect is susceptible to three large elasticities (greater than 1.0) in the next section. In the bottom panel, pre-treatment elasticites for employment remain close to zero. For wages, as we eliminate events with the worst match quality, pre-treatment elasticities fall only slightly. The pooled pre-treatment wage elasticity is always statistically significant at the 5 percent level, indicating that our research design spuriously detects some pre-trends in wages. These spurious effects are nonetheless very small, always less than 0.05.

While the foregoing trimming of treatment events aims to improve the identification of the pooled treatment effect, there is also a concern that poor match quality for donor-based (placebo) synthetic controls biases our inference. In particular, donors whose synethetic controls have poor pre-treatment fit are not informative for assessing the post-treatment rank of the treated state.

[18]Note that a reduction in the pre-treatment *RMSPE* can occur either from a reduced pre-treatment variance or a pre treatment bias. Therefore, an improved pre-treatment fit does not automatically guarantee a smaller pre-treatment elasticity, which is the measure of bias.

In the context of a single treatment event, ADH address this concern by limiting inference to the subset of donors with better pre-treatment synthetic control fit relative to the treated state. Specifically, using the ratio $\gamma_j = MSPE_j/MSPE_1$ of donor-to-treated synthetic control fit during the pre-treatment period, ADH limit randomization inference to subsets of donors with lower values of this ratio. Following this guidance, we explore how mean ranks and the associated confidence intervals change when we limit donors to those with event specific ratios $\gamma_{ej} = MSPE_{ej}/MSPE_{e1}$ less than 20, 10, 5, and 2.

Table 7 presents the pooled effects from this exercise. Restricting donors has almost no effect on the mean rank or Hodges-Lehmann confidence intervals for the pooled employment elasticity, even when we limit the number of donors to those with $MSPE$ ratios of less than 2, removing nearly one-quarter of donor states from the full sample. For wages, removing donors with the worst relative pre-treatment fit has removes some extreme donor elasticities: moving from the full sample to the subset of donors with a $MSPE$ ratio of less than 20, the maximum donor of elasticity of 8.170 and kurtosis of 58.4 fall to 3.439 and 12.0, respectively. But as with employment, inference for wages remains relatively unchanged.

4.4 Robustness to window configuration length and predictor sets

Researchers using synthetic controls face choices about the exact length of pre- and post-treatment windows: more lengthy pre-treatment windows provide more pre-treatment predictor information but also reduce the number of available treatment events. Similarly, synthetic control-based research designs require decisions about the exact set of predictor variables. In this section, we consider how the aggregated results change when modifying window configuration length and the predictor variable set.

We explore two issues using alternative configuration lengths: first, what happens to our estimates when we allow for longer lagged effects? Second, we examine how our estimates vary when we only consider events with a longer pre-intervention period to fit the model. To consider lagged effects, Table 8 begins by showing employment effect estimates for the subsets of events with longer post-treatment periods. When we restrict the sample to those with at least 3 years of post-treatment data (12 quarters), the mean and Hodges-Lehmann point estimates stay very close zero. Only in the case of restricting the sample to post-treatment periods of at least 10 quarters, where the mean employment elasticity is -0.06, are we unable to reject a pooled elasticity greater than -0.20 in magnitude. In short, although we cannot reject moderately sized lagged effects of, say, -0.10, we are unable to detect any presence of lagged effects through the third year of treatment. Table 8 also shows that our employment effect estimates are similar for events with longer pre-intervention periods. After requiring events to have longer pre-intervention windows, pooled employment elasticities remain small and range from -0.003 to 0.072.

Table 8 reports suggestive evidence of lagged wage effects, as the pooled elasticity rises monotonically from 0.265, to 0.303, and then to 0.392, moving from the full sample to events to those with two and then three years of post-treatment data, respectively. Mean elasticities for these subsets

seem rather high, ranging from 0.317 to 0.505, and sometimes come close to the upper bound of the Hodges-Lehmann confidence intervals. These large elasticities are substantially influenced by three events with wage elasticities greater than 1.0 (MA 1986q3, ME 1985q1, NH 1987q1). As we require samples with longer pre-intervention windows, the mean and pooled wage elasticity estimates rise to 0.379 and 0.289, respectively, but then begin fall sharply to about 0.250 and 0.188 and below when requiring at least six years of treatment. These requirements drop the aforementioned extreme elasticites and mechanically lower our pooled estimates. While the magnitude of the wage effect shifts depending on window configuration requirements, we find statistically significant teen wage effects of the minimum in all configurations except the most restrictive pre-treatment configuration of 8 years, which reduces our sample to only 4 events.

We additionally consider how alternative predictor sets affect pooled estimates of the teen employment and wage effects. Testing four candidate models, we found in Section 3.2 a weak preference for the most saturated model with annualized outcomes, including both annualized wage and employment outcomes and other labor market controls. This set of predictors provided all of the above estimates in this paper. Table 9 shows pooled estimates for all of these candidate models in columns 1 through 4. For both teen employment and teen wages, mean elasticities, ranks, and pooled elasticities change little across predictor sets. Hodges-Lehmann estimates for employment range from -0.003 to 0.073; for wages, the pooled elasticities lie between 0.265 and 0.294, all of which are statistically significant at the one percent level. With predictor set 4 (our preferred specification), confidence intervals are somewhat tighter.

One concern with the synthetic control estimator is that by matching on the *levels* of wages and employment, we do not positively weight donors that match the treated state's *trend.* For sufficiently many pre-treatment periods, this is less of a concern, as matching levels or matching trends will produce similar results. But for a small number of pre-treatment periods, synthetic control estimates may differ. To assess the magnitude of this problem Table 9 presents two additional models, where we de-mean the data, as in a difference-in-difference specification. Specifically, for each donor and treated state in a treatment event, we subtract the pre-intervention mean of each predictor and match on the deviations from this pre-treatment mean.[19] Because we do not know the asymptotic properties of this "centered" synthetic control estimator, we view these results as a robustness check.

We calculate synthetic control estimates for "centered" models using two predictor sets. Column 5 uses quarterly outcomes and is the "centered" version of the model in column 1. Column 6 uses the most saturated annualized model — the "centered" version of our preferred specification in column 4. Employment estimates become slightly more negative: the mean and pooled elasticities rise from -0.039 and -0.003 in our preferred specification, to -0.039 and -0.067 in column 5, to -0.065 and -0.087 in column 6. These "centered" results are somewhat more consistent with a small disemployment effect, but the mean ranks are close to 0.45 and 0.43, and at the five percent level we cannot rule out small positive employment effects. For wages we find evidence of smaller impacts,

[19] We only use the pre-treatment mean rather than the entire state mean, to avoid any complications arising from the minimum wage increase affecting the overall mean through the post-treatment period.

with the pooled elasticity falling from 0.265 to 0.211 to 0.163 over specifications 4 through 6. The mean rank for wage effects remains above 0.66 and statistically significant at the one percent level in all specifications. On the whole, the "centered" specifications are largely consistent with our preferred specification's pattern of small employment effects with sizable and statistically significant effects on teen wages.

4.5 Alternative methods of inference

For multiple case studies, the mean percentile rank is an intuitive test statistic and is also a natural extension of the donor-based randomization inference proposed in ADH. The results presented thus far assume that under the sharp null of zero treatment effects, the mean percentile rank is distributed as the mean of independent uniform distributions. There are two sets of potential problems with this assumption. First, because some of our events contain as little as 20 donors, the mean percentile rank is too discrete to be assumed to be uniform. In what follows, we assess how our results change using an alternate "discrete" null distribution of percentile ranks. The second and potentially more serious set of problems is that because some of our treatment events occur at the same time period, they may have the same donors. The donor overlap means that the estimated ranks from two events may be correlated across events, violating the assumption that the ranks are independent. Moreover, recurring treatments may also induce serial correlation in the donor ranks across time. We assess the extent of these problems through a Monte Carlo simulation of synthetic control-based elasticities using placebo interventions, where the timing, treatment overlap, and donor overlap mimics our actual 32 treatment events.

We compare our primary results with these two alternative rank-based methods of conducting inference on the pooled Hodges-Lehmann employment elasticity. These three procedures represent three different ways of conducting rank-based randomization inference on the pooled Hodges-Lehmann employment elasticity; for each of these we estimate the Hodges-Lehman confidence interval. In contrast, our fourth method constructs a confidence interval for the mean effect using randomization inference on the means of donor elasticities (as opposed to the ranks). These four methods provide researchers a toolkit of possibilities for conducting inference with synthetic controls with multiple events. First we describe how we construct three rank-based counterfactual distributions used to test the sharp null $\eta_{e1} = 0$ for all N events, along with the mean elasticity randomization inference method. Then we discuss the results using these four methods .

The first appraoch is the baseline one used throughout this paper, which assumes that the percentile ranks p_{e1} of these elasticities *vis-à-vis* the donor states are distributed continuously as independent uniform variables on $[0, 1]$. We use one million simulations of the mean of N uniforms to calculate the 95% critical values for this Irwin-Hall distribution. Table A1 lists these critical values for $N = 1 \ldots 35$. These are the preferred critical values for the mean percentile rank used throughout the paper. For the full sample of 32 events, the 5% critical values of this distribution are 0.400 and 0.600.

The second method recognizes that in practice the percentile ranks are calculated for a finite,

event-specific number N_e of donor states. For this second method, we relax the assumption of continuous approximation, and calculate percentile ranks with the Weibull-Gumbel rule $p_e = r_e/(N_e + 1)$ by simulating the uniform integer ranks $r_e \in [1, N_e]$ with the appropriate number of donor states present in our data. The distribution of the mean of these percentile ranks across 32 treatment events forms the counterfactual of the mean percentile rank.

The third method further relaxes the assumption that the ranks are independently distributed. As described above and in Section 2.3, the set of placebo elasticities is not independently distributed because some distinct treatments occur around the same time period. For example, the five states that raise their minimum wages during 2005 share many of the same donors. Therefore these treatments are associated with similar donor elasticities, inducing a correlation in donor ranks across this set of events. To account for donor overlap we use a Monte Carlo simulation using placebo laws, calculating synthetic controls for teen employment-to-population ratios in each of the 50 states, using the remaining 49 states as potential donors. These synthetic controls are constructed using the exact timing and pre-/post-treatment window length present in our sample of 32 events.[20]

To form the counterfactual distribution of mean percentile ranks accounting for donor overlap, in this third approach we randomly permute state identifiers of the placebo law sample and then merge the shuffled outcomes and associated synthetic controls to the actual 32 treatment events. The resulting dataset shares the exact timing and structure of donor overlap in our actual sample of 32 events, as well as the actual sample's event-specific pre- and post-treatment window configurations. The dataset also retains the same structure of recurring treatments. For each event, and for each "treated" and donor state, we calculate the percentile ranks of the placebo-treatment effects, defined as the post-treatment percent difference of employment-to-population ratios between the state and its synthetic control. Finally, we calculate the mean percentile rank of the 32 "treated" states. We iterate this procedure one million times. Note that although the placebo-law sample contains actually treated states, the resulting distribution remains a valid counterfactual because all states have the same probability of treatment assignment, before permuting state identifiers.

Table 10 describes all three methods of rank-based inference associated with our sample's 32 treatment events. The first column of results lists the 2.5 and 97.5 percentiles of each counterfactual distribution. These 95% critical values are very similar across methods, with the third method $(0.392, 0.607)$ having the largest deviation from the uniform-based distribution $(0.400, 0.600)$. In the second column, we calculate the 5% rejection rates for each counterfactual distribution using the 95% critical values of uniform-based distribution. By construction, this is exactly 5.0% for the first method using uniform-based distribution. For the second method, the rejection rate is similar but slightly lower, at about 4.3%. Finally, for the third method accounting for overlap using a Monte Carlo simulation, the rejection rate rises to about 6.9%, suggesting that donor overlap causes the

[20] For example, Alaska's treatment event in 2003q1 has a pre-treatment window of 20 quarters and a post-treatment window 12 quarters (see Table 1). For the Monte Carlo simulation we construct synthetic controls for all 50 states using the same date of treatment and the same pre-/post-treatment window lengths. We repeat this for all 32 treatment events. For the synthetic control procedure, we use the preferred predictor set used throughout most of the paper (predictor set 3 in Table 3)

uniform-based distribution to over-reject somewhat, although the bias is modest.

To see how these methods affect estimates of the confidence intervals of the aggregate effects, columns 3-5 show the Hodges-Lehmann point estimate and associated confidence interval. Here, we form confidence intervals as described in Section 2.3 by inverting the mean rank, but we use the critical values from the respective distributions for each method. The 95% Hodges-Lehman elasticity confidence of (-0.153, 0.104) with the first method remains essentially unchanged when using the critical values accounting for the discreteness of the ranks. Using the critical values based on the Monte Carlo simulation accounting for donor overlap, the 95% confidence interval widens a small amount to (-0.164, 0.112). The over-statement in precision due to ignoring overlap therefore appears to be very small in terms of Hodges-Lehmann confidence intervals: about 0.01 in elasticity space. These results provide a strong justification for simply using the Irwin-Hall distribution.

In the fourth row of Table 10 we compare the above results using the Hodges-Lehmann pooled effect and rank-based inference to randomization inference for the mean effect. To form randomization-based confidence intervals around the mean employment elasticity estimate of -0.039, we first construct the counterfactual distribution of the mean of 32 donor elasticities: we select, at random, one donor from each of the treatment events, and calculate the mean of the synthetic control-based elasticities. Repeating this process one million times, we use the resulting 2.5 and 97.5 percentiles of this distribution to calculate the 95% randomization inference confidence interval. For the mean effect of -0.039, the 95% randomization confidence interval is (-0.179, 0.137). Slightly wider than than our preferred confidence interval (-0.153, 0.104) around the Hodges-Lehmann point estimate, the randomization inference confidence interval for the mean, based on means of donor elasticities, is somewhat more influenced by large elasticities in the donor space.

Relatedly, one reason to prefer conducting inference using the mean rank is that under the independence assumption, we know its exact distribution under the sharp null of no effects. The sample mean rank is, therefore, a pivotal statistic. This virtue cannot be claimed by the sample mean elasticity, whose distribution under the null is unknown. We must empirically estimate the mean elasticity under the null using the mean of donor elasticities. A second, and related, reason to prefer rank-based inference is that it allows for conceptually simple tests of the distribution of effects. For example, given the independence assumption, the percentile ranks of treated states should follow a uniform distribution under the sharp null. In the next section we extend this distributional analysis of the percentile ranks to test of heterogenous effects.

4.6 Heterogeneity

Thus far we have focused on the average effects of the treatment, and found employment estimates that are close to zero. However, it is possible that such an average effect is composed of causal effects of differing signs. For example, if the low wage labor market is characterized by monopsonistic competition, employment effects there could be positive in some cases and negative in others (Card and Krueger 1995, Burdett and Mortensen 1989, Manning 2003). Conversely, the spread in the estimated elasticities could simply be due to sampling error, with a true effect of zero everywhere.

Here we take advantage of the fact that the distribution of the percentile ranks is of a known form under the sharp null hypothesis of a zero effect—it is uniformly distributed over the unit interval. If the distribution of empirically estimated percentile ranks differs sufficiently from the theoretical distribution, this constitutes evidence against the sharp null. This possibility can arise either from a non-zero constant effect everywhere, or heterogeneous effects across events.

We use the Kolmogorov-Smirnov test of equality of distributions to determine both the presence of any minimum wage effect and heterogeneous minimum wage effects. Under the sharp null of zero effects everywhere, $\eta_e = 0$, the percentile ranks of the nineteen treatment effects should be uniformly distributed. The empirical CDF of the actual percentile ranks is given by

$$\hat{H}(x) = \frac{1}{E} \sum_{e} I(p_{e1} \leq x)$$

where I is the indicator function. The one-sample Kolmogorov-Smirnov test of the null hypothesis $\hat{H}(x) = x$ has as its test statistic the maximum distance between the empirical CDF and the uniform CDF (see Gibbons and Chakraborti 2003):

$$D = \sup_{x} \left| \hat{H}(x) - x \right|.$$

The top panel of Figure 6 shows the uniform CDF and the empirical CDF for employment and wages. Visually, the percentile ranks of the employment effects are very similar to what would be expected under the sharp null, whereas the wage effects have a different percentile rank distribution. The Kolmogorov-Smirnov test p-values confirm this impression with $p = 0.726$ for employment and $p = 0.000$ for wage elasticity percentile ranks, leading us to reject the sharp null of zero effect only in case of wages.

The above test is against a sharp null of zero effect. To detect heterogeneous effects that average to something other than zero, we extend the procedure above to test for a constant effect equal to the mean effect: $\eta_e = \overline{\eta}$. This is particularly relevant for wage elasticities whose average is positive and substantial in magnitude. Under this sharp null, after centering the 32 elasticities around the mean effect, the adjusted percentile ranks $\widetilde{p}_{e1} = \hat{F}_e(\eta_{e1} - \overline{\eta})$ should be uniformly distributed. The new Kolmogorov-Smirnov test statistic is

$$\widetilde{D} = \sup_{x} \left| \frac{1}{E} \sum_{e=1}^{E} I(\widetilde{p}_{e1} \leq x) - x \right|.$$

The bottom panel of Figure 6 shows the distribution of adjusted percentile ranks for employment and wage elasticities after centering the effects around their means of -0.039 and 0.317, respectively. With Kolmogorov-Smirnov p-values of 0.502 for employment and 0.337 for wages, the re-calculated ranks appear to be uniformly distributed, revealing no evidence of heterogeneous effects in our sample of treatment events.

The Kolmogorov-Smirnov test can be relatively insensitive to distributional differences near

the tails, because there, by construction, the distributional deviations converge to zero (as the empirical distributions converge to 0 and 1). We may expect precisely this kind of heterogeneity if most minimum wage increases have small employment impacts, but a few sufficiently binding increases lead to large employment effects. As an alternative test of uniformly distributed ranks $\widetilde{p}_{e1}$, we use the Anderson-Darling test statistic A^2, which places relatively more weight on the tails of the distribution (see Conover 1999):

$$A^2 = -n - \sum_{e=1}^{E} \frac{2e-1}{E} \left(\log(\tilde{p}_{e1}) + \log(1 - \widetilde{p}_{(E+1-e)1}) \right).$$

For completeness we perform this test both for heterogeneous effects (using the de-meaned percentile ranks $\widetilde{p}_{e1}$) and for a constant zero effect (using the original percentile ranks p_{e1}). As shown in Figure 6, the resulting Anderson-Darling tests p-values are very similar to our previous tests:

In summary, we find that the minimum affects teen wages but no indication of heterogeneous treatment effects for either teen employment or the average teen wage. For the events analyzed in this paper, the synthetic control-based estimates are consistent with a constant positive wage elasticity coupled with a zero employment elasticity of the minimum wage.

5 Conclusion

The appeal of using a data-driven method to choose control groups has led to the increased popularity of the synthetic control method. Since there may be a multitude of case studies one can investigate, the ability to pool across events is useful in many contexts. In this paper we propose and implement and way to pool across synthetic control case studies. We use a variant of the rank sum test to conduct exact inference appropriate for small samples. We also invert the mean rank to provide the Hodges-Lehman confidence interval for the pooled effect.

Although constructing confidence intervals of estimates using permutation-based inference is relatively straightforward in the case of a single treatment event, estimates from individual case studies are often imprecise. Pooling across them allows one to draw economically meaningful inference with, in our case, just 32 events. Our mean employment elasticity is -0.039, and with a 95 percent level of confidence, we reject teen employment elasticities more negative than -0.153. The Hodges-Lehmann pooled wage elasticity of 0.265 is statistically significant at the 1 percent level.

Within the minimum wage literature, our estimates are similar to those from border discontinuity designs and estimates controlling for time-varying heterogeneity. Using CPS data on teens for 1990-2012, Allegretto et al. (2013) find a similar wage elasticity (0.167) and also small employment elasticities: 0.002 with spatial controls and -0.025 with lagged dependent variables. While we cannot rule out moderate negative employment effects of the minimum wage (say, an elasticity around -0.1), our pooled estimates are consistent with small teen employment effects and substantially larger effects on teen wages.

A substantial limitation of pooled synthetic control-based case studies concerns window length.

Due to the nature of minimum wage variation in the United States, increasing the post-treatment window requirements quickly limits the number of available case studies and potential donors. Although we find no lagged effects three years after the treatment, focusing on these events cuts our sample in half. Similar limitations apply to increasing the pre-treatment window in the hope of obtaining better quality matches. In contrast, a clear advantage of a more conventional regression-based approach is the greater ease of considering lagged effects. Lagged effects may be a particularly relevant concern in minimum wage studies when examining whether the short-term and medium-term employment responses differ. Nevertheless, Dube, Lester and Reich (2010) find similar results of small employment impacts even when considering longer lags up to 16 quarters.

Finally, the method we propose for rank-based inference need not be applied to synthetic control estimates, but instead can be used in other difference-in-difference settings with multiple treatments. Calculating and conducting inference on pooled effects is natural once the researcher identifies treatment events and potential controls. In particular, the mean percentile rank is an intuitive test statistic with a known distribution that can be derived for a relatively small number of events, including the case where both the number of treated and control units is small. Moreover, the appeal of a rank-based inferential method extends to the testing of heterogeneous treatment effect more generally, only requiring the comparison of the empirical ranks of the treatment effects from the set of treated units against a known distribution.

References

Abadie, Alberto, Alexis Diamond, and Jens 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.

Allegretto, Sylvia, Arindrajit Dube, Michael Reich, and Ben Zipperer. 2013. "Credible Research Designs for Minimum Wage Studies." *IZA Discussion Paper No. 7638.*

Billmeier, Andreas and Tommaso Nannicini. 2013. "Assessing Economic Liberalization Episodes: A Synthetic Control Approach." *The Review of Economics and Statistics*, 95 (3): 983–1001.

Bohn, Sarah, Magnus Lofstrom, and Steven Raphael. 2013. "Did the 2007 Legal Arizona Workers Act Reduce the State's Unauthorized Immigrant Population?" *Forthcoming in The Review of Economics and Statistics.*

Brown, Charles. 1999. "Minimum Wages, Employment, and the Distribution of Income." in Orley C. Ashenfelter and David Card, eds., *Orley C. Ashenfelter and David Card, eds.*, Vol. 3, Part B of *Handbook of Labor Economics*, Elsevier, pp. 2101 – 2163.

Burdett, Kenneth and Dale T. Mortensen. 1989. "Equilibrium Wage Differentials and Employer Size." *Northwestern University CMSEMS Discussion Papers no. 860.*

Campos, Nauro F., Fabrizio Coricelli, and Luigi Moretti. 2014. "Economic Growth and Political Integration: Estimating the Benefits from Membership in the European Union Using the Synthetic Counterfactuals Method." *IZA Discussion Paper No. 8162.*

Card, David. 1992. "Do Minimum Wages Reduce Employment? A Case Study of California, 1987-1989." *Industrial and Labor Relations Review*, 46 (1): 38–54.

Card, David and Alan B. Krueger. 1995. *Myth and Measurement: The New Economics of the Minimum Wage*, New Jersey: Princeton University Press.

Conley, Timothy G. and Christopher R. Taber. 2011. "Inference with "Difference in Differences" with a Small Number of Policy Changes." *The Review of Economics and Statistics*, 93 (1): 113–125.

Conover, W.J. 1999. *Practical Nonparametric Statistics*, New York: John Wiley & Sons.

Dube, Arindrajit, Ethan Kaplan, and Suresh Naidu. 2011. "Coups, Corporations, and Classified Information." *The Quarterly Journal of Economics*, 126 (3): 1375–1409.

Dube, Arindrajit, T. William Lester, and Michael Reich. 2010. "Minimum Wage Effects across State Borders: Estimates Using Contiguous Counties." *The Review of Economics and Statistics*, 92 (4): 945–964.

Gibbons, Jean Dickinson and Subhabrata Chakraborti. 2003. *Nonparametric Statistical Inference*, Basel, Switzerland: Marcel Dekker.

Hodges Jr., J.L. and E.L. Lehmann. 1963. "Estimates of Location Based on Rank Tests." *Annals of Mathematical Statistics*, 34 (2): 598–611.

Hyndman, Rob J. and Yanan Fan. 1996. "Sample Quantiles in Satistical Packages." *The American Statistician*, 50 (4): 361–365.

Manning, Alan. 2003. *Monopsony in Motion: Imperfect Competition in Labor Markets*, Princeton University Press.

Neumark, David, William Wascher, and J.M. Ian Salas. 2013. "Revisiting the Minimum Wage-Employment Debate: Throwing Out the Baby with the Bathwater?" *Industrial and Labor Relations Review.*

Sabia, Joseph J., Richard V. Burkhauser, and Benjamin 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.

van Elteren, P.H. 1960. "On the Combination of Independent Two-Sample Tests of Wilcoxon." *Bulletin of the International Statistical Institute*, (37): 351–61.

Wilcoxon, Frank. 1945. "Individual Comparisons by Ranking Methods." *Biometrics Bulletin*, 1 (6): 80–83.

Figures

Figure 1: Forming confidence intervals by inverting the mean rank statistic

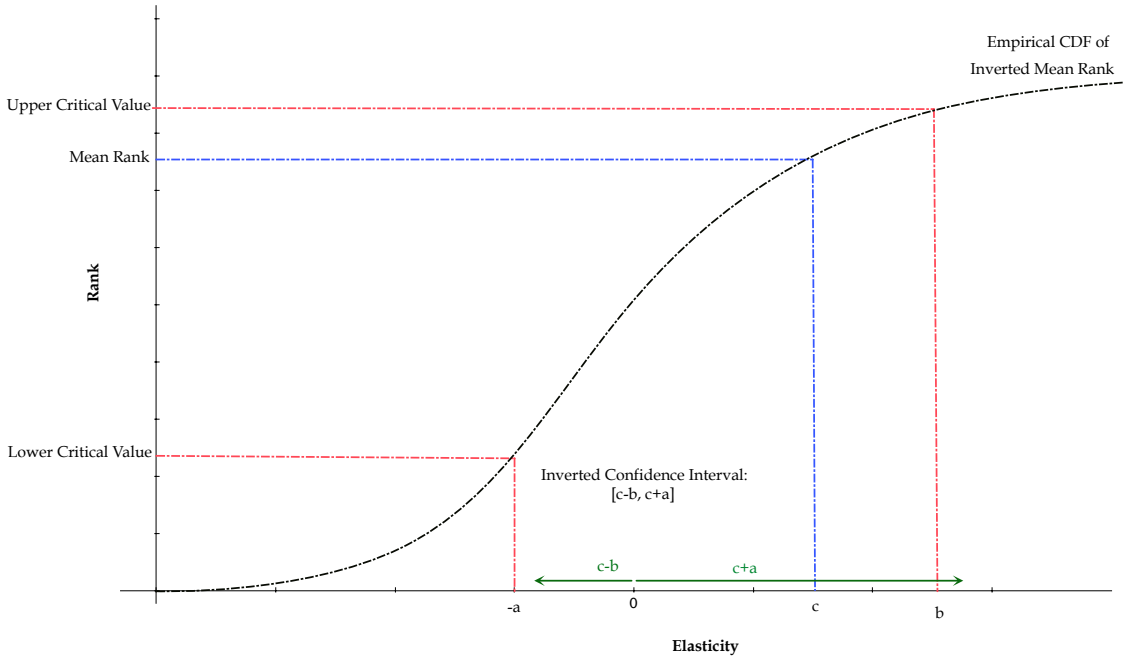
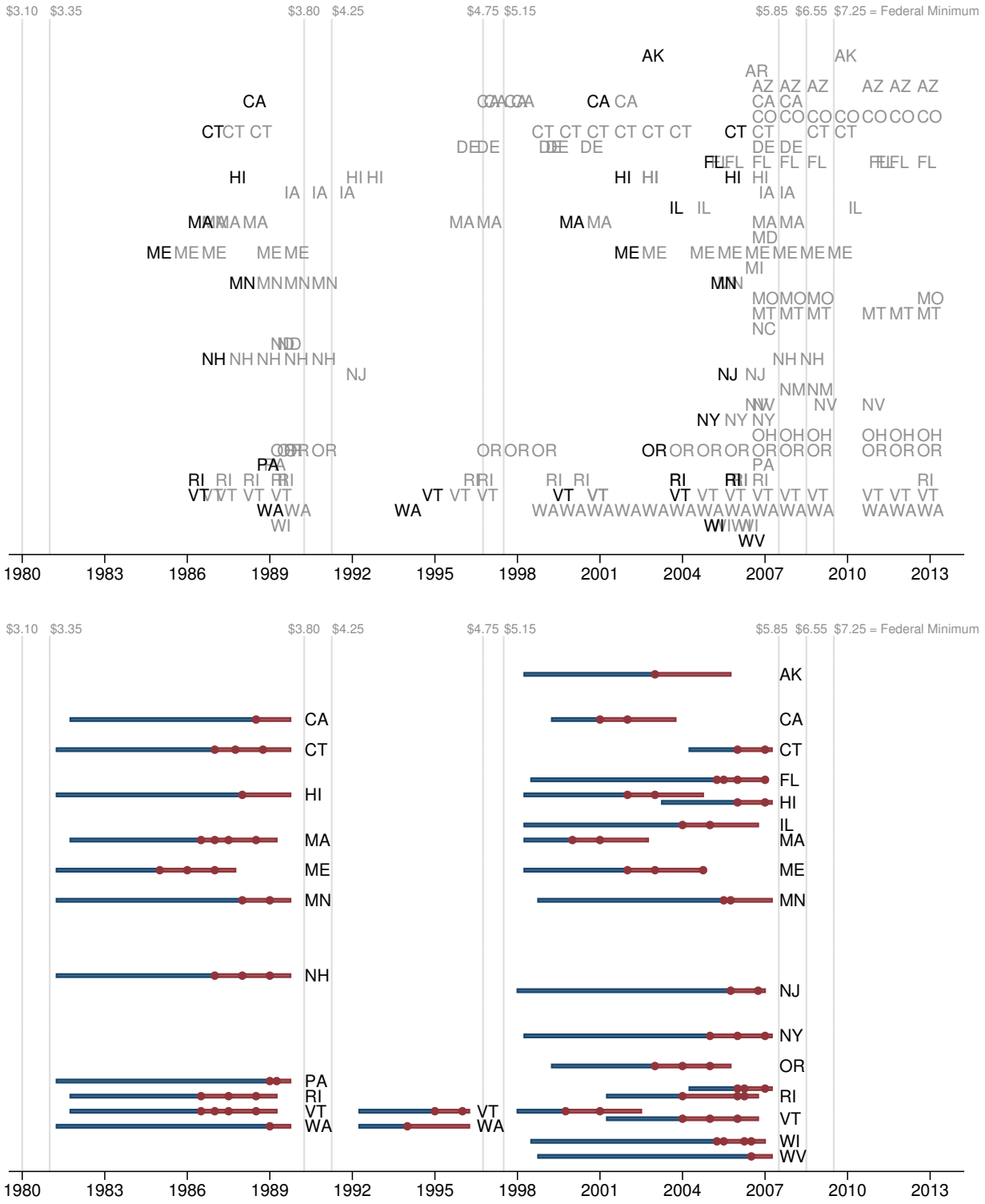
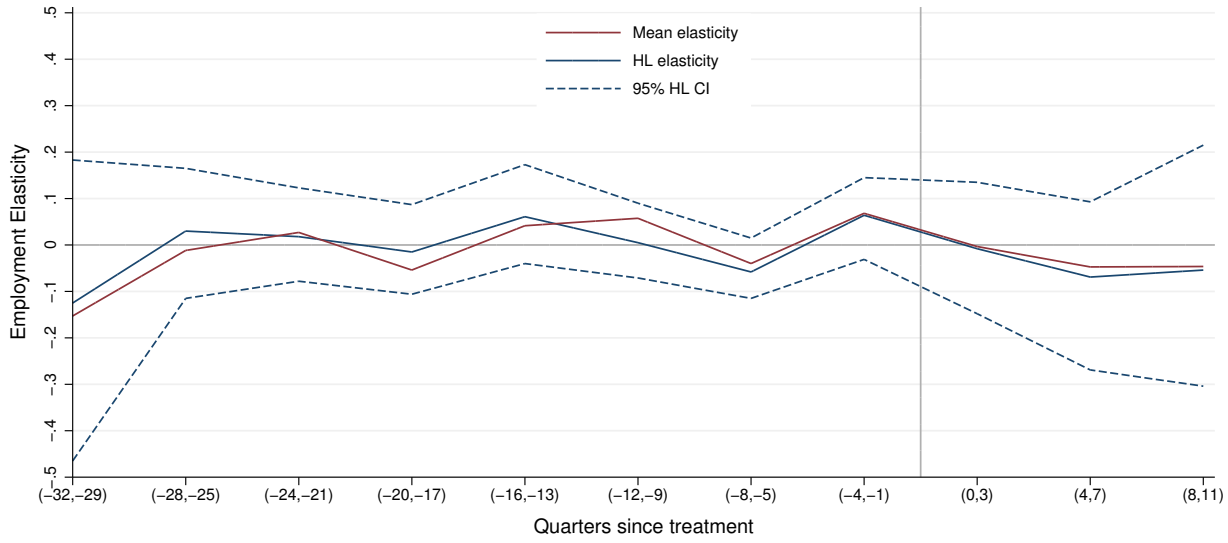


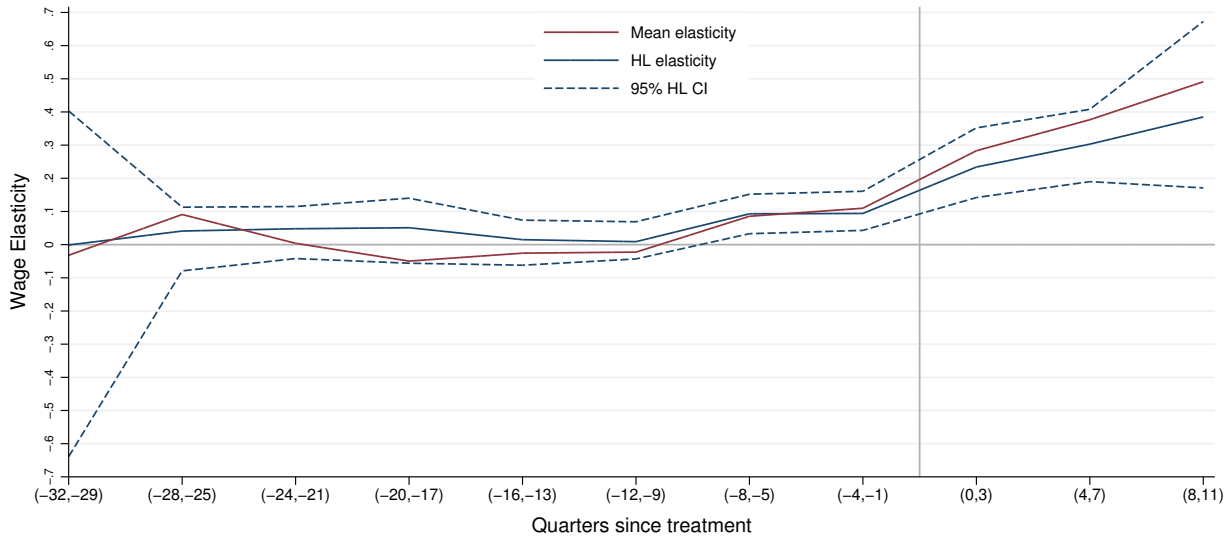
Figure 2: Quarterly minimum wage changes and usable treatment events during 1979-2013



Notes: The top panel displays all quarterly state-level minimum wage changes during the 1979-2013 period, with vertical lines representing the federal minimum wage changes, where bolded states indicate the first minimum wage rise of the 32 usable treatment events. The bottom panel shows the pre-treatment (blue) and post-treatment (red) windows of the 32 treatment events, with red circles showing the minimum wage increases during the post-treatment periods.

Figure 3: Mean annualized elasticities, minimum wage changes, and number of donors, by time





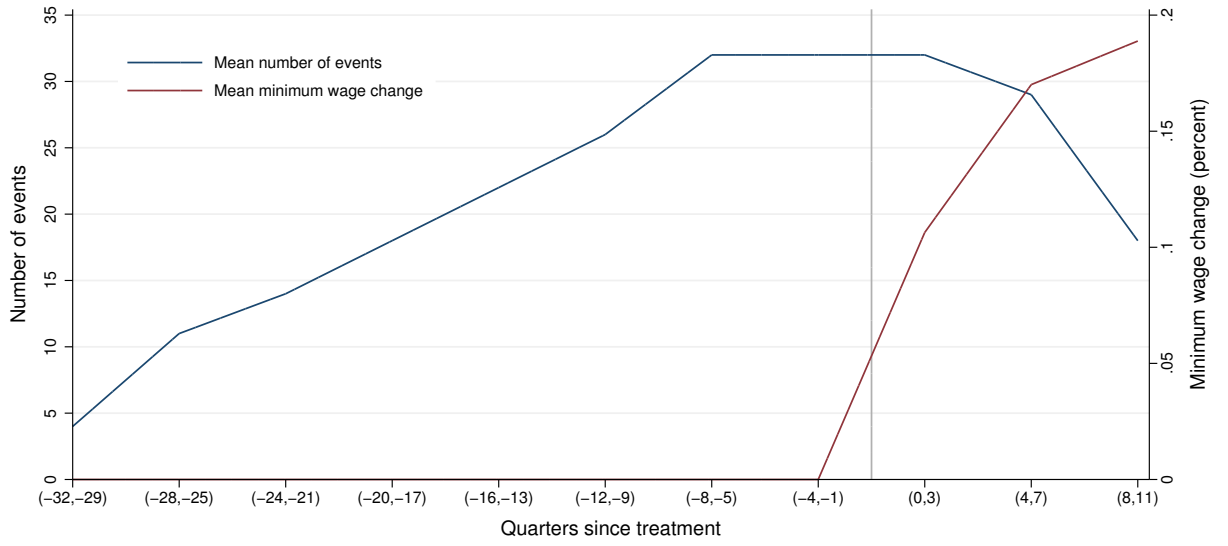


Figure 4: Event-specific and aggregated elasticities

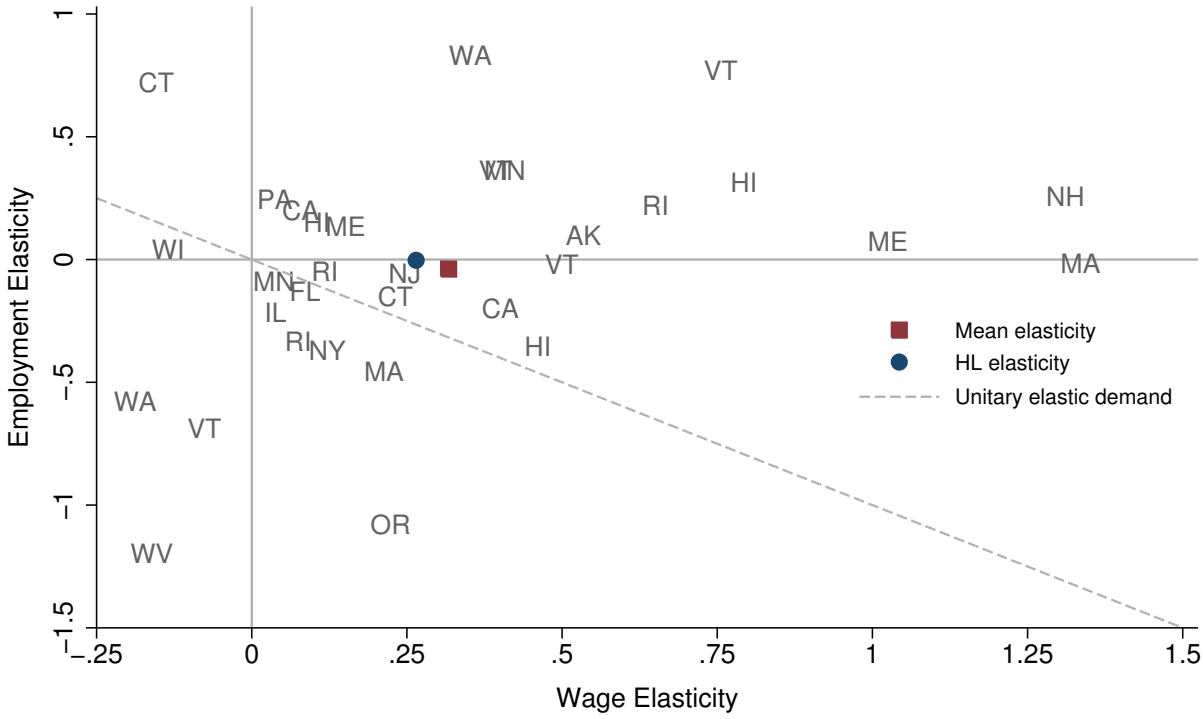
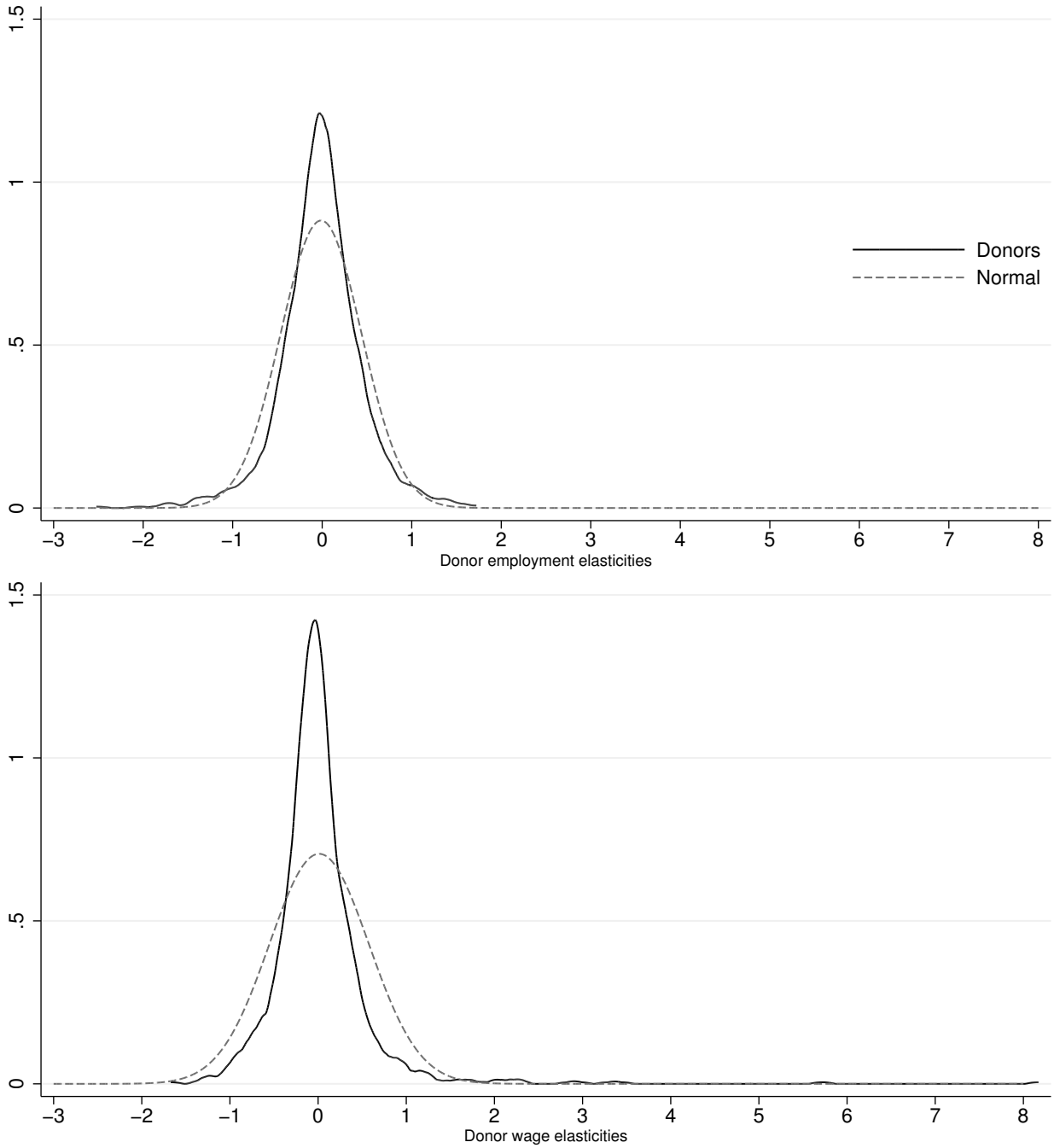


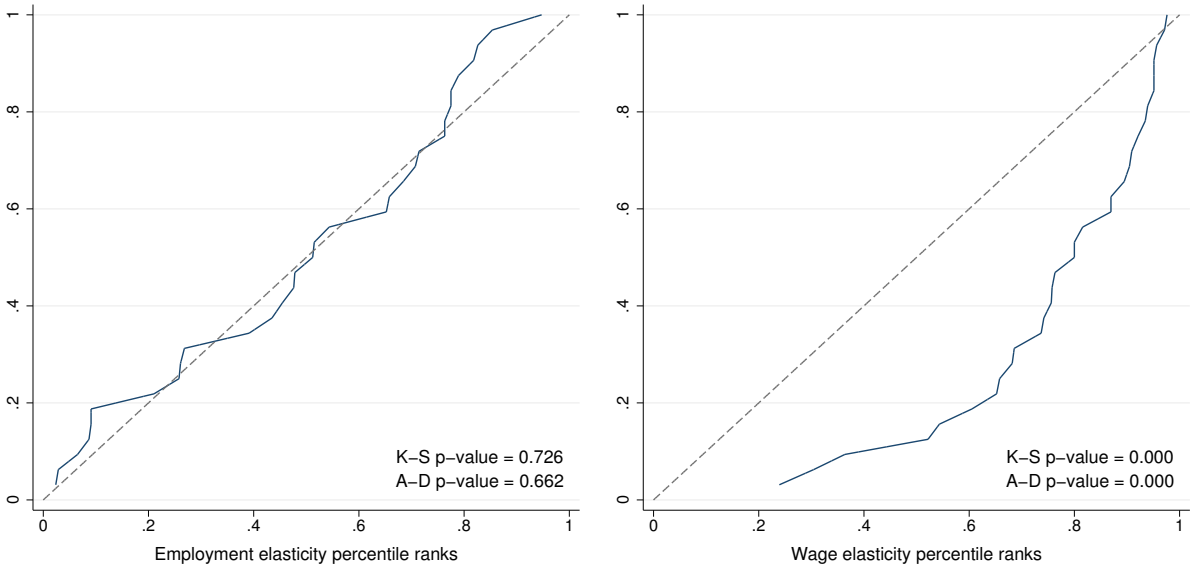
Figure 5: Probability density functions for the employment and wage elasticities of donors



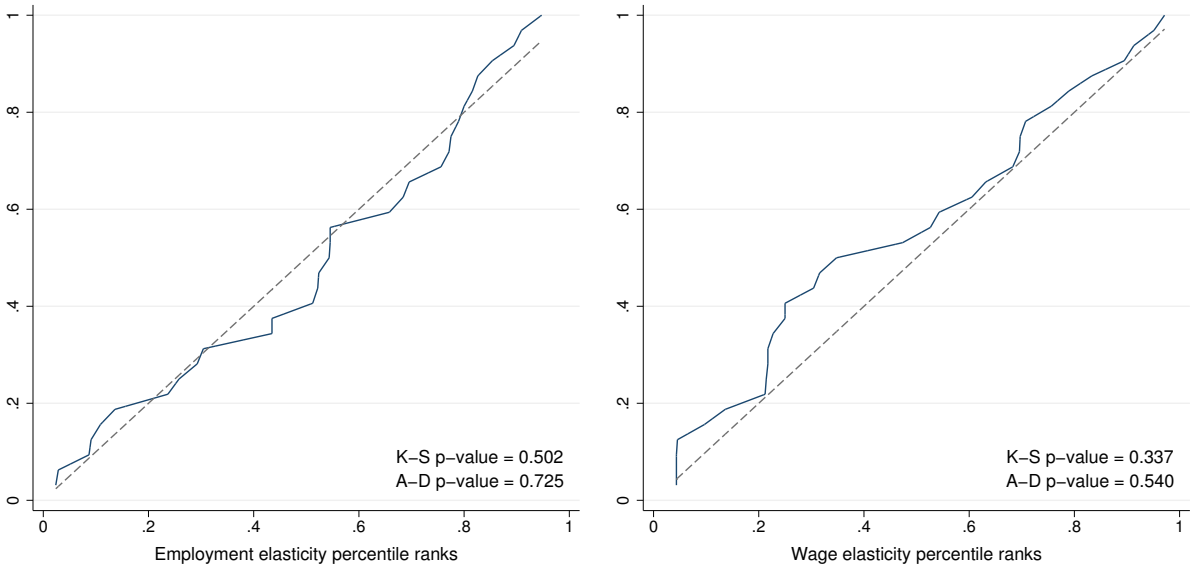
Notes: The donor employment (wage) elasticity distributions have mean -0.012 (0.012), standard deviation 0.452 (0.565), and kurtosis 5.82 (58.37). For each outcome, the illustrated normal distributions have the same mean and variance. Shapiro-Wilk normality test p-values are 0.000 for both outcomes.

Figure 6: Cumulative distribution of estimated percentile ranks

Test of the sharp null hypothesis that $\eta_e = 0$.

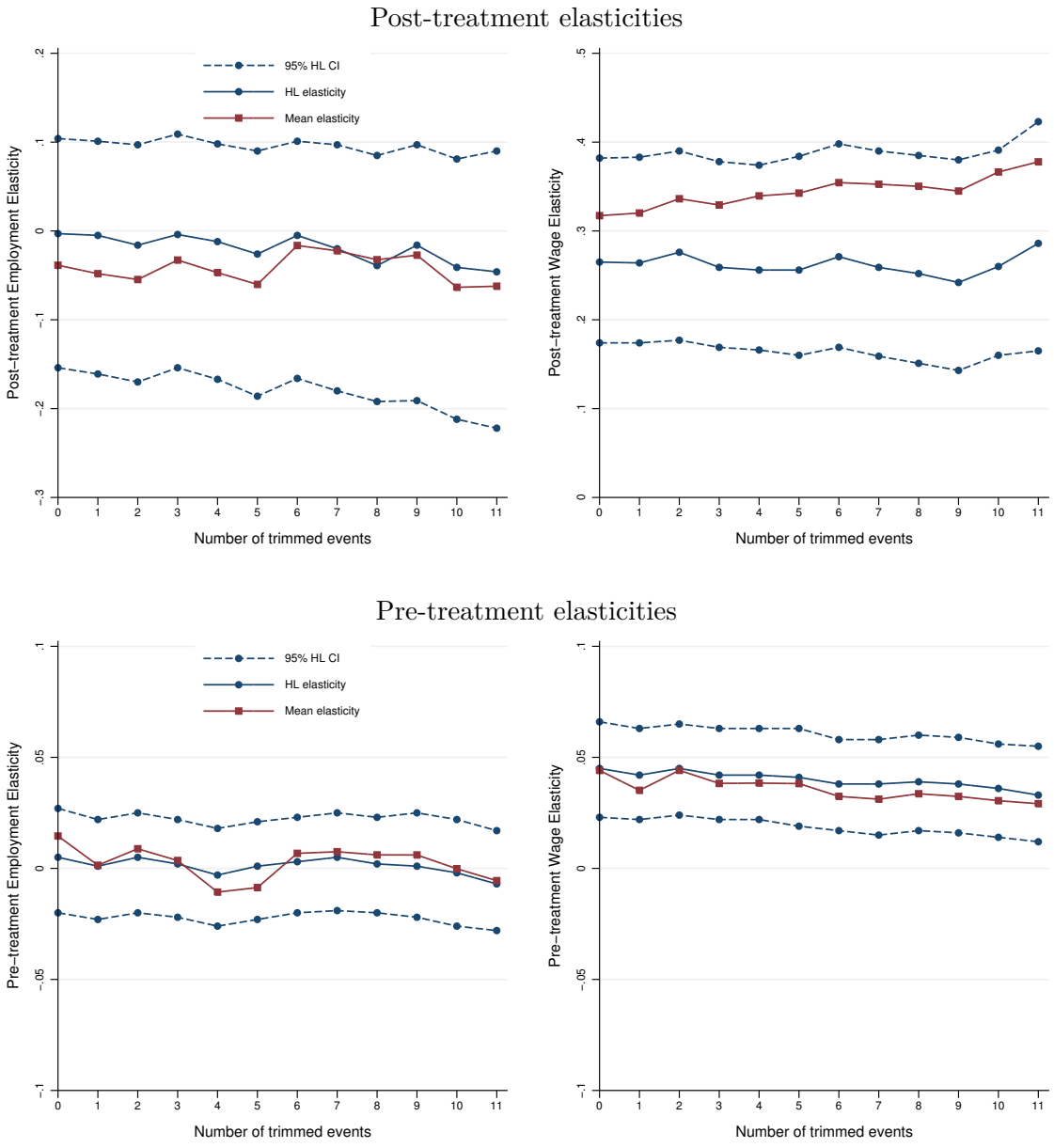


Test of the sharp null hypothesis that $\eta_e = \bar{\eta}$.



Notes: The solid line is the cumulative distribution of percentile ranks of the 32 treatment effects, and the dashed line is the uniform CDF. P-values are for the Kolmogorov-Smirnov and Anderson-Darling tests of equality of these two distributions.

Figure 7: Employment and wage elasticities, trimming events on pre-treatment match quality



Notes: Mean and HL elasticities and confidence intervals re-calculated after successive trimming of $N = 0, \ldots, 11$ events with the highest pre-treatment RMSPE for the actually treated state.

Tables

Table 1: Extrema of the distribution of the mean of N uniformly distributed random variables

Event		Window length		Donors	MW increases	
		Pre	Post		Number	Percent
AK	2003q1	20	12	33	1	0.265
CA	1988q3	28	6	36	1	0.269
CA	2001q1	8	12	39	2	0.174
CT	1987q1	24	12	36	3	0.269
CT	2006q1	8	6	21	2	0.077
FL	2005q2	28	8	21	4	0.295
HI	1988q1	28	8	36	1	0.149
HI	2002q1	16	12	38	2	0.190
HI	2006q1	12	6	21	2	0.160
IL	2004q1	24	12	29	2	0.262
MA	1986q3	20	12	39	4	0.119
MA	2000q1	8	12	40	2	0.286
ME	1985q1	16	12	44	3	0.090
ME	2002q1	16	12	38	3	0.233
MN	1988q1	28	8	36	2	0.149
MN	2005q3	28	8	20	2	0.194
NH	1987q1	24	12	36	3	0.090
NJ	2005q4	32	6	21	2	0.388
NY	2005q1	28	10	20	3	0.388
OR	2003q1	16	12	33	3	0.115
PA	1989q1	32	4	36	2	0.104
RI	1986q3	20	12	39	3	0.194
RI	2004q1	12	12	31	3	0.154
RI	2006q1	8	6	21	3	0.096
VT	1986q3	20	12	39	4	0.090
VT	1995q1	12	6	44	2	0.118
VT	1999q4	8	12	40	2	0.190
VT	2004q1	12	12	31	3	0.160
WA	1989q1	32	4	36	1	0.149
WA	1994q1	8	10	44	1	0.153
WI	2005q2	28	8	21	4	0.262
WV	2006q3	32	4	20	1	0.136

Notes: "Pre" and "Post" are the respective pre-treatment and post-treatment window lengths. Percent minimum wage increase is the percent increase from the pre-treatment minimum to the maximum post-treatment minimum.

Table 2: Summary of treatment events for more restrictive window configurations

Minimum Window				Number of donors			Percent MW change		
Pre-	Post-	Events	Treated states	Min.	Mean	Max.	Min.	Mean	Max.
8	4	32	19	20	32.5	44	0.077	0.187	0.388
8	8	23	16	20	34.0	44	0.090	0.194	0.388
8	12	16	11	29	36.6	44	0.090	0.180	0.286
16	4	22	19	20	32.1	44	0.090	0.200	0.388
16	8	17	14	20	32.8	44	0.090	0.197	0.388
16	12	11	10	29	36.7	44	0.090	0.174	0.269
24	4	14	13	20	28.9	36	0.090	0.222	0.388
24	8	9	8	20	28.3	36	0.090	0.229	0.388
24	12	3	3	29	33.7	36	0.090	0.207	0.269
32	4	4	4	20	28.2	36	0.104	0.195	0.388

Notes: Each row describes the subset of all 32 events in Table 1 with at least the specified number of pre- and post-treatment quarters.

Table 3: Average pre- and post-treatment RMSPE for donors, by model specification

	(1)	(2)	(3)	(4)
		Employment		
Pre-treatment	0.0313	0.0400	0.0370	0.0347
Post-treatment	0.0478	0.0507	0.0490	0.0472
		Wage		
Pre-treatment	0.7084	0.7758	0.7574	0.7356
Post-treatment	0.7911	0.7810	0.7832	0.7758
Predictors				
Quarterly outcomes	Y			
Annualized outcomes		Y	Y	Y
Both annualized outcomes			Y	Y
Industry & Other controls				Y

Notes: The average RMSPE is the square root of the mean of all donors' MSPEs across all treatment events, for either the pre- or post-treatment period. Predictor categories are either all quarterly pre-treatment outcomes (Quarterly), annualized pre-treatment outcomes, both annualized employment and wage pre-treatment outcomes, or pre-treatment means of industry shares and demographic/labor market variables.

Table 4: Donor weights and distance to treated states

Donor relation to treatment	Weights per donor		Relative weight
	Inside	Outside	
		Employment	
Same region	0.050	0.027	1.836
Same division	0.087	0.029	3.039
Within 0 - 500 miles	0.054	0.028	1.932
Within 0 - 1000 miles	0.036	0.028	1.281
		Wage	
Same region	0.051	0.027	1.872
Same division	0.074	0.029	2.544
Within 0 - 500 miles	0.054	0.028	1.954
Within 0 - 1000 miles	0.039	0.026	1.491

Notes: The inside (outside) weight per donor is equal to the sum across all treatment events of the weights assigned to donors inside (outside) the specified area, divided by the total number of inside (outside) donors. Relative weight is the ratio of inside-to-outside weights per donor. Distance in miles is the distance between population-weighted state centroids.

Table 5: Employment and wage elasticities, by event

Event		Employment			Wages		
		Elasticity	Rank	90% CI	Elasticity	Rank	90% CI
AK	2003q1	0.097	0.714	(–0.280, 0.498)	0.540	0.971*	(0.369, 0.985)
CA	1988q3	0.198	0.763	(–0.355, 0.526)	0.083	0.658	(–1.615, 0.598)
CA	2001q1	–0.200	0.268	(–0.768, 0.398)	0.405	0.951*	(0.008, 0.758)
CT	1987q1	–0.150	0.211	(–0.593, 0.329)	0.237	0.816	(–0.711, 0.794)
CT	2006q1	0.718	0.826	(–1.004, 2.757)	–0.149	0.522	(–2.361, 0.834)
FL	2005q2	–0.129	0.435	(–0.637, 0.381)	0.095	0.870	(–0.086, 0.312)
HI	1988q1	0.312	0.763	(–0.658, 0.928)	0.805	0.921	(–1.481, 1.523)
HI	2002q1	0.149	0.775	(–0.415, 0.705)	0.117	0.800	(–0.299, 0.400)
HI	2006q1	–0.354	0.261	(–1.096, 0.676)	0.473	0.870	(–0.472, 0.978)
IL	2004q1	–0.214	0.258	(–0.661, 0.235)	0.054	0.742	(–0.328, 0.382)
MA	1986q3	–0.015	0.512	(–1.149, 0.998)	1.337	0.951*	(0.436, 1.996)
MA	2000q1	–0.456	0.024*	(–0.834,–0.157)	0.215	0.905	(–0.205, 0.513)
ME	1985q1	0.072	0.543	(–1.003, 1.496)	1.027	0.935	(–0.011, 1.982)
ME	2002q1	0.133	0.775	(–0.328, 0.588)	0.153	0.800	(–0.187, 0.384)
MN	1988q1	0.362	0.789	(–0.607, 0.979)	0.409	0.763	(–1.878, 1.127)
MN	2005q3	–0.089	0.455	(–0.762, 0.750)	0.036	0.682	(–0.292, 0.573)
NH	1987q1	0.256	0.658	(–1.070, 1.693)	1.314	0.895	(–1.528, 2.986)
NJ	2005q4	–0.056	0.478	(–0.377, 0.423)	0.254	0.957*	(0.118, 0.476)
NY	2005q1	–0.372	0.091	(–0.726, 0.065)	0.126	0.909	(–0.049, 0.274)
OR	2003q1	–1.081	0.029*	(–1.889,–0.267)	0.225	0.686	(–0.247, 1.031)
PA	1989q1	0.245	0.684	(–1.197, 2.768)	0.043	0.605	(–8.128, 1.394)
RI	1986q3	0.218	0.707	(–0.479, 0.842)	0.663	0.951*	(0.108, 1.069)
RI	2004q1	–0.051	0.515	(–0.727, 0.679)	0.130	0.758	(–0.329, 0.777)
RI	2006q1	–0.333	0.391	(–1.719, 1.306)	0.087	0.652	(–1.693, 0.878)
VT	1986q3	0.769	0.854	(–0.742, 2.120)	0.762	0.756	(–0.441, 1.641)
VT	1995q1	–0.689	0.087	(–1.310, 0.216)	–0.070	0.543	(–1.147, 0.681)
VT	1999q4	–0.022	0.476	(–0.528, 0.518)	0.506	0.976*	(0.183, 0.990)
VT	2004q1	0.361	0.818	(–0.292, 1.065)	0.400	0.939	(–0.044, 1.024)
WA	1989q1	0.829	0.947	(–0.181, 2.595)	0.352	0.737	(–5.368, 1.297)
WA	1994q1	–0.580	0.065	(–1.393, 0.022)	–0.188	0.239	(–1.022, 0.257)
WI	2005q2	0.038	0.652	(–0.533, 0.613)	–0.127	0.304	(–0.330, 0.117)
WV	2006q3	–1.199	0.091	(–2.207, 0.166)	–0.162	0.364	(–1.117, 0.836)

Notes: * indicates significance at the 10% level using the percentile rank of the elasticity within the event-specific donor-based placebo distribution. Inverting this rank obtains 90% CIs.

Table 6: Employment and wage elasticities, pooled

				Hodges-Lehmann	
	Median elasticity	Mean elasticity	Mean rank	Elasticity	95% CI
Employment	–0.019	–0.039	0.497	–0.003	(–0.153, 0.104)
Wages	0.220	0.317	0.763***	0.265	(0.174, 0.382)

Notes: Critical values for the mean percentile rank are derived from the mean of 32 uniform distributions.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table 7: Employment and wage confidence intervals, by pre-treatment donor MSPE ratios

	Donor states		Donor state elasticities				Treated state elasticities	
MSPE Ratio	Number	Fraction	Min	Max	SD	Kurtosis	Mean rank	95% CI
				Employment				
.	1039	1.000	–2.522	1.722	0.452	5.816	0.497	(–0.153, 0.104)
20	1038	0.999	–2.038	1.722	0.445	5.199	0.497	(–0.153, 0.104)
10	1033	0.994	–2.038	1.722	0.442	5.115	0.497	(–0.153, 0.104)
5	987	0.950	–2.038	1.722	0.443	5.125	0.494	(–0.161, 0.103)
2	802	0.772	–2.038	1.722	0.449	5.141	0.502	(–0.155, 0.115)
				Wages				
.	1039	1.000	–1.671	8.170	0.565	58.371	0.763***	(0.174, 0.382)
20	978	0.941	–1.671	3.439	0.426	11.969	0.774***	(0.186, 0.387)
10	954	0.918	–1.671	3.439	0.415	12.290	0.777***	(0.186, 0.387)
5	912	0.878	–1.671	3.439	0.413	12.594	0.777***	(0.178, 0.387)
2	759	0.731	–1.671	3.439	0.417	13.796	0.764***	(0.176, 0.389)

Notes: Rows show results for samples where donors limited to those with donor-to-treated MSPE ratios less than X, where X=. indicates the full sample.
* p < 0.1, ** p < 0.05, *** p < 0.01.

Table 8: Employment and wage elasticities, by minimal window length

Window length					Hodges-Lehmann	
Pre-	Post-	Events	Mean elasticity	Mean rank	Elasticity	95% CI
			Employment			
.	4	32	–0.039	0.497	–0.003	(–0.153, 0.104)
.	6	29	–0.038	0.489	–0.007	(–0.161, 0.097)
.	8	23	–0.026	0.495	–0.005	(–0.165, 0.121)
.	10	18	–0.060	0.461	–0.062	(–0.242, 0.105)
.	12	16	–0.009	0.509	0.015	(–0.191, 0.161)
8	.	32	–0.039	0.497	–0.003	(–0.153, 0.104)
12	.	26	–0.014	0.533	0.038	(–0.114, 0.165)
16	.	22	0.017	0.554	0.050	(–0.099, 0.196)
20	.	18	0.061	0.559	0.048	(–0.094, 0.218)
24	.	14	0.002	0.520	0.015	(–0.153, 0.206)
28	.	11	0.013	0.559	0.047	(–0.129, 0.260)
32	.	4	–0.045	0.550	0.072	(–0.587, 0.693)
			Wages			
.	4	32	0.317	0.763***	0.265	(0.174, 0.382)
.	6	29	0.342	0.784***	0.267	(0.176, 0.384)
.	8	23	0.402	0.805***	0.303	(0.184, 0.439)
.	10	18	0.446	0.832***	0.350	(0.208, 0.485)
.	12	16	0.505	0.865***	0.392	(0.256, 0.541)
8	.	32	0.317	0.763***	0.265	(0.174, 0.382)
12	.	26	0.357	0.776***	0.289	(0.168, 0.407)
16	.	22	0.379	0.776***	0.290	(0.158, 0.435)
20	.	18	0.379	0.770***	0.289	(0.145, 0.481)
24	.	14	0.251	0.730***	0.188	(0.060, 0.358)
28	.	11	0.174	0.706**	0.173	(0.021, 0.319)
32	.	4	0.122	0.666	0.256	(–0.310, 0.646)

Notes: Window length of X restricts the events to the subset with a pre- or post-period greater than X quarters, where $X = .$ is eight pre-treatment or four post-treatment quarters. Critical values for the mean percentile rank are derived from the mean of appropriate number of uniform distributions.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 9: Employment and wage elasticities, by alternative predictor sets

	(1)	(2)	(3)	(4)	(5)	(6)
			Employment			
Mean elasticity	–0.008	–0.008	0.066	–0.039	–0.039	–0.065
Mean rank	0.508	0.506	0.552	0.497	0.454	0.434
HL elasticity	0.009	0.013	0.073	–0.003	–0.067	–0.087
HL 95% CI	(–0.128, 0.147)	(–0.169, 0.135)	(–0.080, 0.209)	(–0.153, 0.104)	(–0.179, 0.074)	(–0.200, 0.045)
			Wage			
Mean elasticity	0.338	0.402	0.393	0.317	0.254	0.175
Mean rank	0.757***	0.775***	0.775***	0.763***	0.682***	0.664***
HL elasticity	0.290	0.290	0.294	0.265	0.211	0.163
HL 95% CI	(0.198, 0.416)	(0.189, 0.417)	(0.210, 0.421)	(0.174, 0.382)	(0.098, 0.326)	(0.070, 0.269)
Predictors						
Quarterly outcomes	Y				Y	
Annualized outcomes		Y	Y	Y		Y
Both annualized outcomes			Y	Y		Y
Industry & Other controls				Y		Y
Centered					Y	Y

Notes: Predictor categories are either all quarterly pre-treatment outcomes (Quarterly), annualized pre-treatment outcomes, both annualized employment and wage pre-treatment outcomes, or pre-treatment means of industry shares and demographic/labor market variables. Models 5 and 6 use quarterly and annualized predictors, respectively, that are de-meaned by the pre-treatment mean.

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table 10: Alternative methods of inference for employment elasticty

	Rank-based inference		Aggregated employment elasticity		
	95% crit. values	5% rej. rate	Type	Point estimate	95% CI
Method 1: Rank, uniform iid.	(0.400, 0.600)	0.050	Hodges-Lehmann	–0.003	(–0.153, 0.104)
Method 2: Rank, discrete iid	(0.403, 0.597)	0.043	Hodges-Lehmann	–0.003	(–0.151, 0.102)
Method 3: Rank, rand. inference	(0.392, 0.607)	0.069	Hodges-Lehmann	–0.003	(–0.164, 0.112)
Method 4: Elasticity, rand. inference	-	-	Mean	–0.039	(–0.179, 0.137)

Notes: The baseline method 1 assumes mean rank is distributed as the mean of 32 uniform variables on [0,1]. Method 2 calculates the rank as the mean of 32 percentile ranks using the Weibull-Gumbel percentile rank calculation with the appropriate number of donor states for each event. Method 3 further relaxes the assumption that the ranks are independent across events by accounting for the donor overlap across events. In particular, method 3 uses a monte-carlo simulation using placebo laws by permuting state identifiers randomly, and then estimating the distribution of mean ranks from that iteration. The details for the simulation are described in the text. Method 4 conducts randomization inference using the mean effect (elasticity) as opposed to the mean rank; this method uses the mean of 32 donor employment elasticities, choosing one donor per event, for 1 million iterations. Rejection rates for each method are shown when using the critical values from method 1.

Appendix Tables

Table A1: Extrema of the distribution of the mean of N uniformly distributed random variables

	Percentile					
N	0.5	2.5	5.0	95.0	97.5	99.5
1	0.005	0.025	0.050	0.950	0.975	0.995
2	0.050	0.111	0.158	0.842	0.888	0.950
3	0.103	0.176	0.223	0.777	0.823	0.896
4	0.147	0.220	0.261	0.738	0.780	0.852
5	0.182	0.249	0.287	0.713	0.751	0.819
6	0.206	0.271	0.305	0.694	0.729	0.793
7	0.227	0.288	0.320	0.680	0.712	0.774
8	0.244	0.301	0.332	0.668	0.699	0.757
9	0.258	0.312	0.341	0.658	0.687	0.743
10	0.269	0.322	0.350	0.650	0.678	0.731
11	0.280	0.330	0.357	0.643	0.670	0.720
12	0.289	0.337	0.363	0.637	0.663	0.711
13	0.297	0.344	0.368	0.632	0.656	0.703
14	0.304	0.349	0.373	0.627	0.651	0.696
15	0.311	0.354	0.377	0.623	0.646	0.689
16	0.317	0.359	0.381	0.619	0.641	0.683
17	0.322	0.363	0.385	0.615	0.637	0.679
18	0.327	0.367	0.388	0.612	0.633	0.673
19	0.331	0.370	0.391	0.609	0.630	0.669
20	0.335	0.374	0.394	0.606	0.626	0.665
21	0.339	0.377	0.396	0.604	0.623	0.660
22	0.343	0.379	0.398	0.601	0.620	0.657
23	0.346	0.382	0.401	0.599	0.618	0.654
24	0.349	0.384	0.403	0.597	0.615	0.650
25	0.352	0.387	0.405	0.595	0.613	0.647
26	0.355	0.389	0.407	0.593	0.611	0.645
27	0.358	0.391	0.408	0.591	0.609	0.642
28	0.360	0.393	0.410	0.590	0.607	0.639
29	0.363	0.395	0.412	0.588	0.605	0.637
30	0.365	0.397	0.413	0.587	0.603	0.635
31	0.367	0.398	0.415	0.585	0.601	0.632
32	0.369	0.400	0.416	0.584	0.600	0.630
33	0.371	0.401	0.417	0.583	0.598	0.628
34	0.373	0.403	0.418	0.581	0.597	0.627
35	0.375	0.404	0.420	0.580	0.595	0.625

Notes: Simulated using one million iterations of the mean of N uniformly distributed variables on $[0, 1]$.