Supplementary Material for *Bracketing in the Comparative Interrupted Time-Series Design to Address Concerns about History Interacting with Group: Evaluating Missouri's Handgun Purchaser Law*

Raiden B. Hasegawa

University of Pennsylvania

Daniel W. Webster

Johns Hopkins University

Dylan S. Small

University of Pennsylvania*

October 25, 2018

*Address for Correspondence: Dylan Small, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104 (E-mail: *dsmall@wharton.upenn.edu*).

1 Inferences Under Different Sampling Assumptions

The standard difference-in-difference estimator using a control group c, $\hat{\beta}_{dd.c}$, is

$$\begin{aligned}\hat{\beta}_{dd.c} &= \{\hat{E}[Y_1|G=t,S_1=1]-\hat{E}[Y_0|G=t,S_0=1]\} \\ &\quad -\{\hat{E}[Y_1|G=c,S_1=1]-\hat{E}[Y_0|G=c,S_0=1]\}.\end{aligned}$$

When the samples of (i) $Y_1|G=t,S_1=1$, (ii) $Y_0|G=t,S_0=1$, (iii) $Y_1|G=c,S_1=1$ and (iv) $Y_0|G=c,S_0=1$ are independent, then the standard error of $\hat{\beta}_{dd.c}$ is

$$\begin{aligned}&SE(\hat{\beta}_{dd.c}) = \\ &\{SE(\hat{E}[Y_1|G=t,S_1=1])^2+SE(\hat{E}[Y_0|G=t,S_0=1])^2 \\ &+SE(\hat{E}[Y_1|G=c,S_1=1])^2+SE(\hat{E}[Y_0|G=c,S_0=1])^2\}^{1/2}. \qquad (1)\end{aligned}$$

We use (1) to make inferences for our study of the effect of the repeal of Missouri's PTP law, where the $\hat{E}$ and corresponding SEs are obtained from the CDC's WONDER system.

Let κ_{tt} be the % change in the treated group's mean outcome in the after period compared to its mean counterfactual outcomes in the after period in the absence of treatment,

$$\kappa_{tt} = 100 \times \frac{E[Y_1^{(1)}|G=t,S_1=1]-E[Y_1^{(0)}|G=t,S_1=1]}{E[Y_1^{(0)}|G=t,S_1=1]}.$$

An estimate of κ_{tt} using control group c and assuming the parallel trends of standard-in-differences is

$$\hat{\kappa}_{tt.c} = 100 \times \frac{\hat{\beta}_{dd.c}}{\hat{E}(Y_0|G=t,S_0=1)+\{\hat{E}(Y_1|G=c,S_1=1)-\hat{E}(Y_0|G=c,S_0=1)\}}.$$

We approximate the standard error of $\hat{\kappa}_{tt.c}$ using the Delta method.

The model (1) can be extended to allow for observed covariates, clustering and multiple time

points using a regression framework[1]. The difference-in-difference estimator may be computed by regressing the observed outcome Y on a time period dummy, a group dummy and a treatment variable. Observed covariates $\mathbf{X}_{ip}$ that could vary by time can be incorporated into the model and then the difference-in-difference regression estimator can be computed by regressing Y on the observed covariates, a time period dummy, a group dummy and a treatment variable. The model assumptions then need to hold only conditionally on the observed covariates. The CITS can be applied to settings with more than two time periods. A full set of time period dummies can be added to model (1). The effect of the treatment over time can be allowed to vary by interacting the treatment dummy with time.

Within each group, there may be clusters of units, e.g., different countries that had the same policy reform. For such settings, we can extend model (1) to the following[2] where the index cip denotes the ith unit in cluster c at time period p:

$$Y_{cip}^{(d)} = h(\mathbf{U}_{cip}, p) + \beta d + \eta_{cp} + \epsilon_{cip}, \tag{2}$$

where η_{cp} represents an effect shared by members of cluster c in period p, e.g., an economic shock that is specific to a country c in period p. Under an assumption that the η_{cp} are independent and identically distributed (i.i.d.) normal random variables, Donald and Lang[2] showed that if we compute the mean in each cluster at each time period, and regress these cluster/period means on fixed effects for each cluster, a time period dummy and a treatment variable, then the t statistic for the treatment variable ($\frac{\hat{\beta}-\beta}{SE(\hat{\beta})}$) has a t distribution with the number of clusters minus two degrees of freedom. Using this approach, we do not need to have individual data but only summary data for each cluster. Other approaches to inference that allow for the η_{cp} to be non-i.i.d. such as autocorrelated within group, have been developed.[3,4].

Note that the presence of at least two clusters in at least one group enables us to make inferences that allow for shared effects η_{cp}. When there is only one cluster in each group, e.g., we are comparing just two countries, one in which a policy reform was implemented and one in

which it was not, then there are zero degrees of freedom to estimate the variance of the η_{cp} so inferences cannot be drawn that allow for η_{cp} to be nonzero using data from entirely within the sample. For such settings, it may be possible to get information from outside the sample to get a plausible estimate of the variance of the η_{cp}[5,2].

2 Proof of (9)

Suppose $h(\mathbf{U}, 1) - h(\mathbf{U}, 0)$ is a bounded increasing function of $\mathbf{U}$. Then from (5) and the property that bounded increasing functions of stochastically ordered random variables preserve order, it follows that

$$E[\hat{\beta}_{dd.uc}] \leq \beta \leq E[\hat{\beta}_{dd.lc}]. \tag{3}$$

Similarly, if $h(\mathbf{U}, 1) - h(\mathbf{U}, 0)$ is a bounded decreasing function of $\mathbf{U}$,

$$E[\hat{\beta}_{dd.lc}] \leq \beta \leq E[\hat{\beta}_{dd.uc}]. \tag{4}$$

(9) follows from (3) and (4).

3 Proof for Section 2.3

Here we prove that (10) has probability $\geq 1-\alpha$ of containing both $\min(\theta_{lc.t}, \theta_{uc.t})$ and $\max(\theta_{lc.t}, \theta_{uc.t})$ under the assumption that the two sided CIs are constructed in the usual way by taking the union of two one-sided $1 - (\alpha/2)$ confidence intervals. The result is basically derived by inverting multiparameter hypothesis tests about the minimum or maximum of two parameters[6,7]. Let $q = \min(\theta_{lc.t}, \theta_{uc.t})$ and $r = \max(\theta_{lc.t}, \theta_{uc.t})$. The probability that (10) does not contain both $\min(\theta_{lc.t}, \theta_{uc.t})$ and $\max(\theta_{lc.t}, \theta_{uc.t})$ is bounded by the probability that q is less than the lower endpoint of the interval plus the probability that r is greater than the upper endpoint of the interval. The probability that q is less than the lower endpoint of the interval is the probability that

both one-sided tests $H_0^l : \theta_{lc.t} \leq q$ vs. $H_1^l : \theta_{lc.t} > q$ and $H_0^u : \theta_{uc.t} \leq q$ vs. $H_1^l : \theta_{uc.t} > q$ give p-values $\leq \alpha/2$, which has probability at most $\alpha/2$ since each individual event has probability at most $\alpha/2$. Similarly, the probability that r is greater than the upper endpoint of the interval is the probability that both one-sided tests $H_0^{l'} : \theta_{lc.t} \geq r$ vs. $H_1^{l'} : \theta_{lc.t} < r$ and $H_0^{u'} : \theta_{uc.t} \geq r$ vs. $H_1^{l'} : \theta_{uc.t} < r$ give p-values $\leq \alpha/2$, which has probability at most $\alpha/2$ since each individual event has probability at most $\alpha/2$. Thus, the probability that (10) does not contain both $\min(\theta_{lc.t}, \theta_{uc.t})$ and $\max(\theta_{lc.t}, \theta_{uc.t})$ is bounded by α.

4 Modeling Time-varying Confounders

We model a setting with time-varying confounders as follows. We maintain the assumptions in Section 2.1 except for (4). We let $\mathbf{U}$ contain all variables that affect the outcome that differ in distribution between the groups (treated, upper control, lower control) in the before period and let ϵ_0 summarize the effect of factors in the before period that do not differ in distribution between the groups. We can model the average effect of the factors in ϵ_0 as an intercept in the $h(\mathbf{U}, 0)$ function so that $E(\epsilon_0 | S_0 = 1, G = g) = 0$ holds for all groups $g = lc, uc, tc$. The effect of factors that do not differ in distribution between the groups in the after period as well as the effect of time-varying confounders in the after period are summarized in ϵ_1. Some of these time-varying confounders may be variables in $\mathbf{U}$ that have changed their level over time. Let $\mathbf{U}_0 \equiv \mathbf{U}$ be the value of the variables in $\mathbf{U}$ in the before period and $\mathbf{U}_1$ be their value in the after period, where $\mathbf{U}_0 = \mathbf{U}_1$ for a unit only in the population in the after period (with $\mathbf{U}$ defined this way, the validity of (3) needs to be considered carefully). Then, assuming that the average effect of the factors in ϵ_1 that do not differ between the groups in the after period is modeled as an intercept in $h(\mathbf{U}, 1)$, we have

$$E(\epsilon_1 | G = g, S_1 = 1) = E[h(\mathbf{U}_1, 1) - h(\mathbf{U}_0, 1) | G = g, S_1 = 1].$$

Then for (11) to hold, we need to have

$$E[h(\mathbf{U}_1,1)-h(\mathbf{U}_0,1)|G=uc,S_1=1] \geq E[h(\mathbf{U}_1,1)-h(\mathbf{U}_0,1)|G=t,S_1=1]$$
$$\geq E[h(\mathbf{U}_1,1)-h(\mathbf{U}_0,1)|G=lc,S_1=1] \quad (5)$$

A set of sufficient conditions for (5) to hold when $\mathbf{U}$ is univariate and the assumptions in Section 2.1 hold is the following: (a) $S_0 = S_1 = 1$ for all units so that all units are in the study population in both periods; (b) $U_1 - U_0$ is independent of U_0 given G; (c) the function $h(U,1)$ is convex in U so that h has increasing differences in the sense that for u, $u^{'}$, $u^{''}$, $u^{'''}$ such that $u - u^{'} = u^{''} - u^{'''}$ and $u > u^{''}$, the following inequality holds: $h(u,1) - h(u^{'},1) \geq h(u^{''},1) - h(u^{'''},1)$, and (d) $U_1 - U_0|G = lc \preceq U_1 - U_0|G = t \preceq U_1 - U_0|G = uc$. The proof that this set of sufficient conditions implies that (5) holds is as follows. Let D_{lc} be a random variable with the distribution of $U_1 - U_0|G = lc$ where D_{lc} is independent of U_0 given G. Then from (c) and (5), it follows that

$$E[h(U_0+D_{lc},1)-h(U_0,1)|G=t] \geq E[h(U_0+D_{lc},1)-h(U_0,1)|G=lc]. \quad (6)$$

Now let D_t be a random variable with the conditional distribution of $U_1 - U_0|G = t$ and D_{uc} be a random variable with the conditional distribution of $U_1 - U_0|G = uc$ where D_t and D_{uc} are independent of U_0 given G. Then from (d) and h being an increasing function, it follows that $E[h(U_0 + D_t)|G = t] \geq E[h(U_0 + D_{lc})|G = t]$. Combining this with (6), we have

$$E[h(U_0+D_t,1)-h(U_0,1)|G=t] \geq E[h(U_0+D_{lc},1)-h(U_0,1)|G=lc]$$

which is equivalent to

$$E[h(U_1,1)-h(U_0,1)|G=t] \geq E[h(U_1,1)-h(U_0,1)|G=lc]. \quad (7)$$

Similarly from (d) and h being an increasing function, it follows that $E[h(U_0+D_{uc})|G=uc] \geq E[h(U_0+D_t)|G=uc]$, and from (c) and (5), it follows that

$$E[h(U_0+D_t,1)-h(U_0,1)|G=uc] \geq E[h(U_0+D_t,1)-h(U_0,1)|G=t],$$

and combining these, we have that

$$E[h(U_0+D_{uc},1)-h(U_0,1)|G=uc] \geq E[h(U_0+D_t,1)-h(U_0,1)|G=t]$$

which is equivalent to

$$E[h(U_1,1)-h(U_0,1)|G=uc] \geq E[h(U_1,1)-h(U_0,1)|G=t]. \tag{8}$$

Combining (7) and (8) gives us the desired conclusion.

Proof that (9) still holds as long as when (i) in (6) holds, (11) holds or when (ii) in (6) holds, (12) holds. When there are time varying confounders, we have that $E[\hat{\beta}_{dd.lc}]$ is the expression on the right hand side of (7) plus $E(\epsilon_1|G=t,S_1=1)-E(\epsilon_1|G=lc,S_0=1)$ and $E[\hat{\beta}_{dd.lc}]$ is the expression on the right hand side of (8) plus $E(\epsilon_1|G=t,S_1=1)-E(\epsilon_1|G=uc,S_0=1)$. When (i) in (6) holds, the expression on the right hand side of (7) is $\geq \beta$ and the expression on the right hand side of (8) is $\leq \beta$. Combining the facts in the last two sentences, we have that if (i) in (6) and (11) holds, $E[\hat{\beta}_{dd.uc}] \leq \beta \leq E[\hat{\beta}_{dd.lc}]$ and if (ii) in (6) and (12) holds, $E[\hat{\beta}_{dd.lc}] \leq \beta \leq E[\hat{\beta}_{dd.uc}]$.

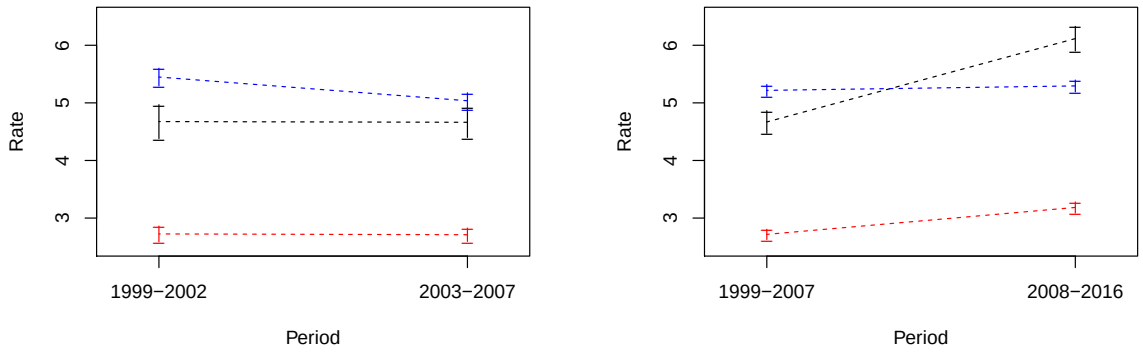
5 Test of Model/Assumptions by Examining the Groups' Relative Trends in the Before Period

We can test whether the violating pattern (iii) is present in the before period using an intersection-union test[6,7], which find evidence (say p-value < 0.05) for (iii) if there is evidence (p-value $< .05$) for both (a) the difference between the upper control group and the counterfactual treated group is larger in the second part of the before period than the first part and (b) the difference between the counterfactual treated group and the lower control group is smaller in the second part than the first part; for the firearm homicide data, splitting the before period into the two parts, 1999-2002 and 2003-2007, (a) gives a p-value of 0.96 and (b) gives a p-value of 0.5, so there is not evidence for (iii) being violated. Pattern (iv) can be tested in a similar way and for the firearm homicide data, there is not evidence for pattern (iv) holding (p-values of 0.04 and 0.5). Ideally, this testing procedure should have sufficient power to reduce the chance of proceeding with the analysis when the assumptions of the model don't, in fact, hold to an acceptable level. When sample sizes are beyond the control of the investigator or, for example, when dealing with complete counts of firearm homicides where variability depends on the rate itself rather than sampling error, increasing the level of the test can achieve some improvement in power. The p-value is ≥ 0.5 for the test of each alternative, that (iii) holds and that (iv) holds. Hence, α would have to be increased beyond 0.5 to affect the conclusions about the plausibility of our model assumptions.

Alternatively, the presence of violating patterns (iii) and (iv) can be assessed visually without requiring a formal testing procedure. In the left panel of Figure 1 we plot the relative trends of the population-weighted firearm homicide rates for the upper (dashed blue) and lower (dashed red) groups and the counterfactual treated group (dashed black) over the before period. The vertical bars indicate 95% CIs. Visually, there is no strong evidence that pattern (iii) or (iv) is present. The difference between upper controls and counterfactual Missouri and between

counterfactual Missouri and the lower controls both get smaller in the latter part of the before period. We can also partially assess whether this pattern might hold over the entire study period, our primary concern, by addressing how the upper and lower control trends compare between the before period and the entire study period. In the right panel of Figure 1 we plot the relative trends of the two control groups and treated group over the entire study period. The dashed black lines are not comparable between panels because the left panel is a counterfactual trend whereas the trend in the right panel is subject to treatment (i.e. PTP repeal). However, we can assess the comparability of the pattern of the control group trends between the two panels. They appear similar, with a slight narrowing of the difference in population-weighted firearm homicide rates over time.

Figure 1: (Left Panel): Relative trends of the population-weighted firearm homicide rates for the upper (dashed blue) and lower (dashed red) groups and the counterfactual treated group (dashed black) over the before period. The vertical bars indicate 95% CIs. (Right Panel): Relative trends of the population-weighted firearm homicide rates for the upper (dashed blue) and lower (dashed red) groups and the treated group (dashed black) over the entire period. The vertical bars indicate 95% CIs.



When paired with the test described above, visual inspection can answer questions about our model assumptions that our intersection-union tests do not address directly: If we find evidence that pattern (iii) or (iv) is present, are the violations substantial enough to arrest the planned analysis or should we still proceed but with increased caution? If the test doesn't find evidence

of a violation is that because our assumptions hold, at least approximately, or is it due to large standard errors and/or low power? We recommend that testing and visual inspection should be used in conjunction when assessing the plausibility of the model assumptions.

If one does find evidence for pattern (iii) or (iv) holding in the before period, and if one thinks there has been a structural shift such that the model (1)-(4) and assumptions (5)-(6) only start to hold in the latter part of the before period but continue to hold in the after period, one could just use the latter part of the before period. This is similar to the scenario in a difference-in-difference model when there is evidence of a diverging trend during an earlier portion of the pre-intervention period, researchers can restrict the analysis to include only the latter part of the before period with the hope that parallel trend assumption is more likely to be valid[8]. However, the finding of pattern (iii) or (iv) in the before period suggests caution.

6 Analysis Using After Period of 2008-2013

For the period of 2008-2013, Missouri's age-adjusted firearm homicide rate was 5.5, the upper control group's age-adjusted firearm homicide rate was 5.0 and the lower control's age adjusted firearm homicide rate was 2.9. Using an after period of 2008-2013, difference-in-difference estimates for the upper and lower control groups are shown in Table 1. Using an after period of 2008-2013, the interval (10) that has a $\geq$ 95% chance of containing the effect of the repeal on the firearm homicide rate is $[0.2, 1.4]$, corresponding to a 5% to 31% increase in firearm homicides, providing evidence that the repeal increased firearm homicides.

Table 1: Difference-in-difference estimates of effect of repeal of Missouri's permit-to-purchase handgun licensing requirement on firearm homicide rates per 100,000 persons using after period of 2008-2013

Control Group	Estimate	95% CI	Corresponding % Change Estimate	95% CI
Upper Controls	1.0	[0.6, 1.4]	22%	[14% ,31%]
Lower Controls	0.6	[0.2, 1.0]	17%	[5% ,19%]

7 Comparison with the Synthetic Control Method

Abadie et al.[9] proposed constructing a synthetic control group which is a linear combination of multiple control groups that matches the before period outcomes of the treatment group. The synthetic control method provides asymptotically unbiased estimates of the causal effect of treatment assuming that the unmeasured confounders can be represented by a factor model with the factors' effects in each time period being linear with a time-specific slope, whereas our bracketing method only provides bounds under this assumption. However, this assumption is strong and is not generally satisfied in our model (1)-(4). In the following section we provide a simple example that satisfies the assumptions of our model but for which the estimate returned by the synthetic control method will be biased.

If the types of interaction between history and group in the after period that are of concern have occurred in the before period (e.g., a similar recession occurred in the after period as the before period), then the synthetic control method's matching of the before period outcomes might enable it to match the treated group's counterfactual trajectory in the after period in the absence of treatment. However, if the types of interaction are different (e.g., there is a more severe recession in the after period or the interactions between poor health and the macroeconomy have been altered by other policy changes), then the synthetic control's matching in the before period does not provide much reassurance unless one has a basis for strong functional form assumptions such as the factors representing the unmeasured confounders' having a linear effect in each time period. In contrast, the bracketing method relies on assumptions such as (6) that the unmeasured confounders' effect is increasing (or decreasing) in importance over time over the whole range of the unmeasured confounders that can be assessed using subject matter knowledge without making strong functional form assumptions.

8 Example of How Synthetic Control Model Assumptions Are Violated in Our Model

For example, suppose U has an exponential distribution in each group with scale 0.2, 0.5 and τ in the lower control, upper control and treated groups respectively where $0.2 < \tau < 0.5$ and $h(U,0) = U$, $h(U,1) = exp(U)$. Then the synthetic control linear combination is $\frac{\tau-0.2}{0.3}\times$ lower control group + $\frac{0.5-\tau}{0.3}\times$ upper control group. For the after period, the linear combination of the mean outcomes for the synthetic control linear combination is $\frac{\tau-0.2}{0.3} \times 1.25 + \frac{0.5-\tau}{0.3} \times 2$ while the treated group's counterfactual mean outcome in the absence of treatment is $\frac{-1/\tau}{-1/\tau+1}$, and $\frac{\tau-0.2}{0.3} \times 1.25 + \frac{0.5-\tau}{0.3} \times 2 < \frac{-1/\tau}{-1/\tau+1}$ for all $0.2 < \tau < 0.5$. Thus the synthetic control group's after period mean is always less than than the counterfactual after period mean for the treatment group in the absence of treatment.

References

[1] G. W. Imbens and J. M. Wooldridge, "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, vol. 47, no. 1, pp. 5–86, 2009.

[2] S. G. Donald and K. Lang, "Inference with difference-in-differences and other panel data," *The Review of Economics and Statistics*, vol. 89, no. 2, pp. 221–233, 2007.

[3] M. Bertrand, E. Duflo, and S. Mullainathan, "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics*, vol. 119, no. 1, pp. 249–275, 2004.

[4] C. B. Hansen, "Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects," *Journal of Econometrics*, vol. 140, no. 2, pp. 670–694, 2007.

[5] J. L. Blitstein, P. J. Hannan, D. M. Murray, and W. R. Shadish, "Increasing the degrees of freedom in existing group randomized trials: The df* approach," *Evaluation Review*, vol. 29, no. 3, pp. 241–267, 2005.

[6] E. Lehmann, "Testing multiparameter hypotheses," *The Annals of Mathematical Statistics*, pp. 541–552, 1952.

[7] R. L. Berger, "Multiparameter hypothesis testing and acceptance sampling," *Technometrics*, vol. 24, no. 4, pp. 295–300, 1982.

[8] K. G. Volpp, A. K. Rosen, P. R. Rosenbaum, P. S. Romano, O. Even-Shoshan, Y. Wang, L. Bellini, T. Behringer, and J. H. Silber, "Mortality among hospitalized medicare beneficiaries in the first 2 years following ACGME resident duty hour reform," *Journal of the American Medical Association*, vol. 298, no. 9, pp. 975–983, 2007.

[9] A. Abadie, A. Diamond, and J. Hainmueller, "Synthetic control methods for comparative case studies: Estimating the effect of California's tobacco control program," *Journal of the American Statistical Association*, vol. 105, no. 490, pp. 493–505, 2010.