Assignment 6: Panel Data II – Part 2 (Staggered DiD)

Applied Quantitative Methods for the Social Sciences II

Spring 2026

```r
library(did)
library(dplyr)
library(ggplot2)
library(fixest)

data(mpdta)
```

1. Data structure and visualization

a) Number of counties and treatment cohorts:

```r
length(unique(mpdta$countyreal))
```

```
## [1] 500
```

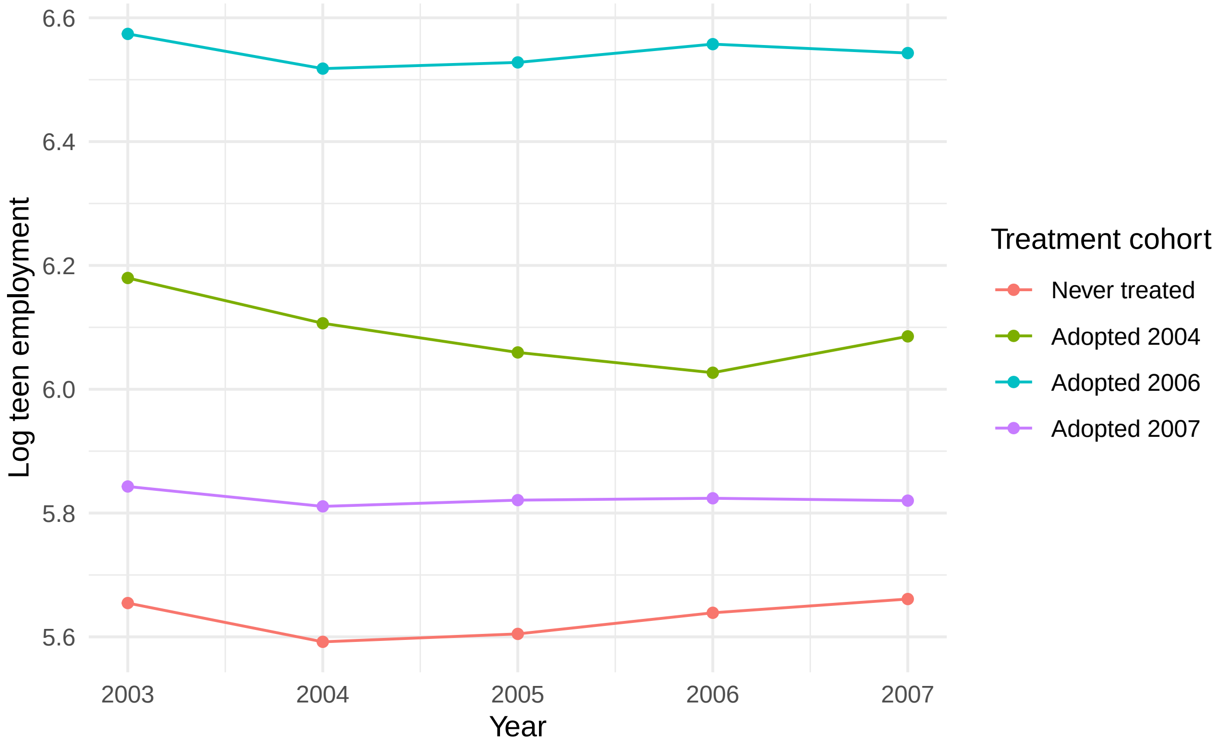
```r
table(mpdta$first.treat)
```

```
##
##    0 2004 2006 2007
## 1545  100  200  655
```

There are 500 counties observed over five years (2003–2007). The `first.treat` variable shows four distinct values: 0 (never treated), 2004, 2006, and 2007 — representing four treatment cohorts. This is a staggered adoption design: counties adopted a minimum wage policy in different years. Simply comparing treated vs. untreated counties ignores this timing variation and can produce biased estimates, because counties that adopted earlier may differ systematically from those that adopted later, and treatment effects may accumulate or fade over time in ways that a single treated/untreated dummy cannot capture.

b) Average log teen employment by cohort over time:

```r
mpdta_avg = mpdta %>%
  mutate(cohort = factor(first.treat,
    levels = c(0, 2004, 2006, 2007),
    labels = c("Never treated", "Adopted 2004",
               "Adopted 2006", "Adopted 2007"))) %>%
  group_by(year, cohort) %>%
  summarise(mean_lemp = mean(lemp, na.rm = TRUE))

ggplot(mpdta_avg, aes(x = year, y = mean_lemp, color = cohort)) +
  geom_line() +
  geom_point() +
  theme_minimal() +
  labs(x = "Year", y = "Log teen employment", color = "Treatment cohort")
```



```
ggsave("mpdta_cohort_trends.pdf", width = 7, height = 4)
```

Pre-treatment, all cohorts track relatively similar trends in log teen employment, which is encouraging for parallel trends. After treatment, the cohorts that adopted the minimum wage (particularly in 2004) tend to have lower teen employment relative to the never-treated counties, consistent with a negative employment effect of minimum wages. The 2007 cohort has only a within-sample post-treatment period in 2007, so its post-treatment pattern is limited.

2. Naive TWFE vs. Callaway-Sant'Anna estimator

a) Naive TWFE with a time-varying treatment indicator:

```
mpdta = mpdta %>%
  mutate(treated_post = as.integer(first.treat > 0 & year >= first.treat))

m_twfe = feols(lemp ~ treated_post | countyreal + year,
               data = mpdta, cluster = ~countyreal)
summary(m_twfe)
```

```
## OLS estimation, Dep. Var.: lemp
## Observations: 2,500
## Fixed-effects: countyreal: 500,  year: 5
## Standard-errors: Clustered (countyreal)
##               Estimate Std. Error  t value Pr(>|t|)
## treated_post -0.036549   0.013265 -2.75526 0.006079 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.124223     Adj. R2: 0.991505
##                  Within R2: 0.004169
```

The coefficient on `treated_post` is the TWFE estimate of the average effect of minimum wage adoption on log teen employment. This model pools all treated counties into a single group and implicitly assumes that the treatment effect is constant across cohorts and over time. If effects differ by cohort or grow/shrink after treatment — a condition called treatment effect heterogeneity — TWFE can be badly biased and can even recover negative weights, yielding estimates that do not correspond to any interpretable treatment effect.

b) Callaway-Sant'Anna estimator:

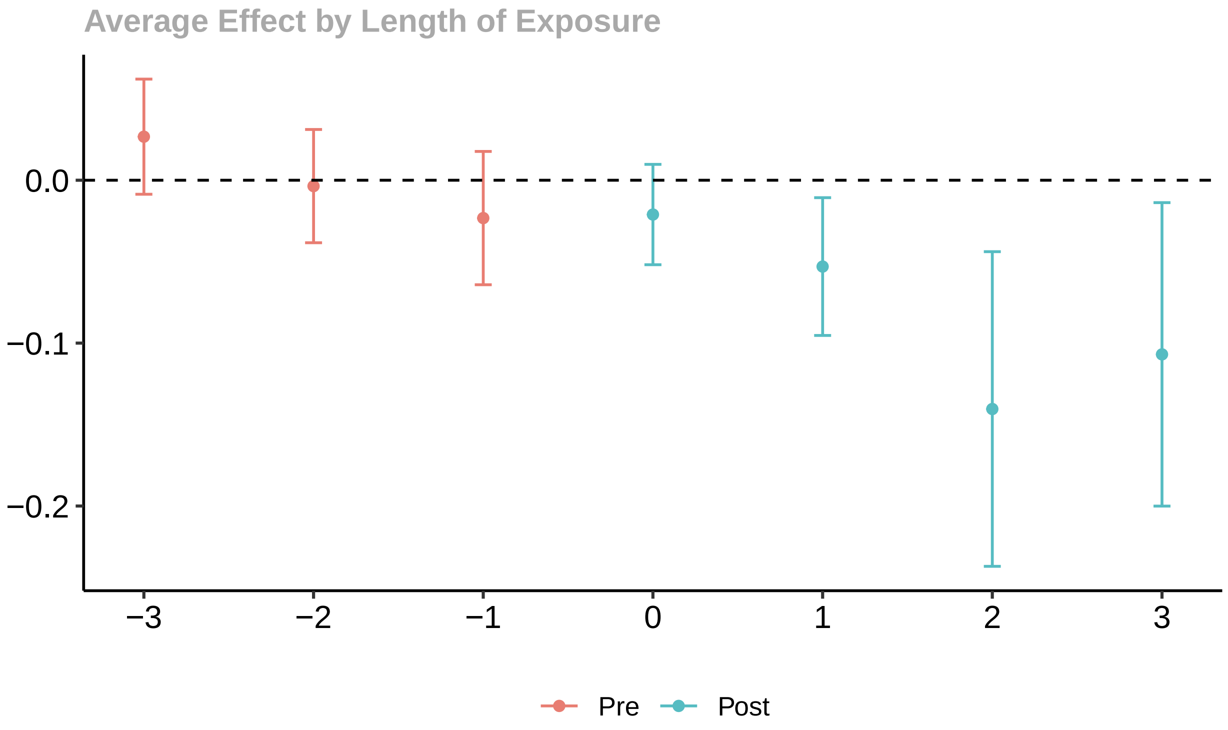
```
cs_out = att_gt(
  yname         = "lemp",
  gname         = "first.treat",
  idname        = "countyreal",
  tname         = "year",
  xformla       = ~ lpop,
  data          = mpdta,
  control_group = "nevertreated")

aggte(cs_out, type = "simple")
```

```
##
## Call:
## aggte(MP = cs_out, type = "simple")
##
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Differences with Multiple Time P
##
##
##      ATT    Std. Error     [ 95%  Conf. Int.]
##  -0.0418        0.0123    -0.0659     -0.0176 *
##
##
## ---
## Signif. codes: `*' confidence band does not cover 0
##
## Control Group:  Never Treated,  Anticipation Periods:  0
## Estimation Method:  Doubly Robust
```

The Callaway-Sant'Anna overall ATT aggregates group-time average treatment effects using only clean comparisons: each cohort is compared to never-treated counties before and after its adoption year. Compare this estimate to the naive TWFE coefficient — they may differ substantially, illustrating that TWFE can produce a distorted average when cohort-specific effects are heterogeneous.

c) Dynamic event-study estimates:

```
cs_dyn = aggte(cs_out, type = "dynamic")
ggdid(cs_dyn)
```



```r
ggsave("mpdta_event_study.pdf", width = 7, height = 4)
```

The event-study plot shows estimates for each period relative to the cohort's treatment date (negative values are pre-treatment periods, positive values are post-treatment). The pre-treatment estimates should be close to zero and statistically indistinguishable from zero; if they are, this supports the parallel trends assumption — treated and control counties were on similar trajectories before treatment. The post-treatment estimates show the dynamic treatment effect: the immediate impact and how it evolves in subsequent periods. A growing negative effect over time would indicate that minimum wage increases progressively reduce teen employment.

3. Pre-testing the parallel trends assumption

a) Formal joint pre-test with bootstrapped SEs and uniform confidence bands:

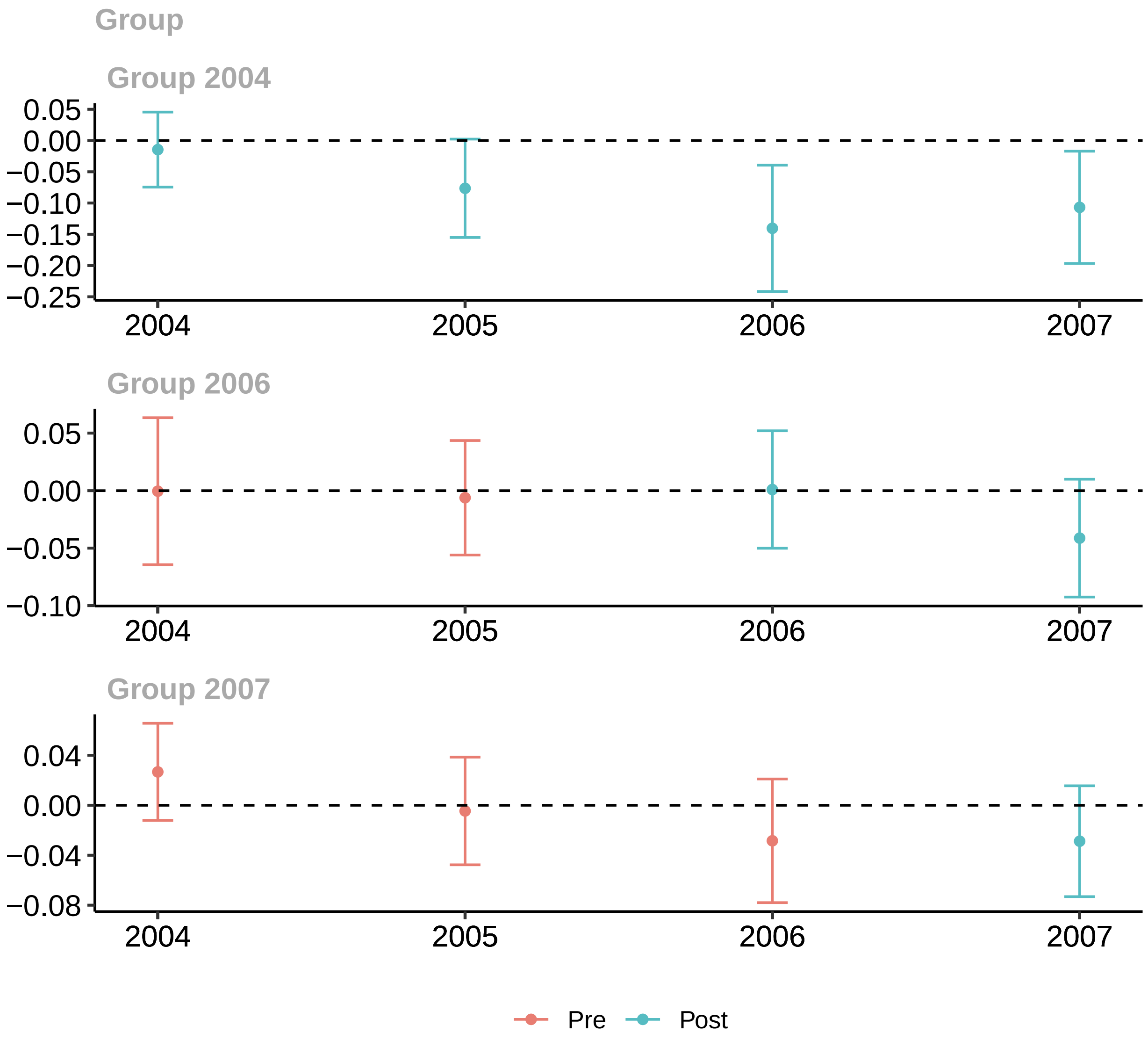
```r
cs_out_bt = att_gt(
  yname         = "lemp",
  gname         = "first.treat",
  idname        = "countyreal",
  tname         = "year",
  xformla       = ~ lpop,
  data          = mpdta,
  control_group = "nevertreated",
  bstrap        = TRUE,
  cband         = TRUE)

summary(cs_out_bt)
```

```
##
## Call:
## att_gt(yname = "lemp", tname = "year", idname = "countyreal",
##     gname = "first.treat", xformla = ~lpop, data = mpdta, control_group = "nevertreated",
##     bstrap = TRUE, cband = TRUE)
##
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Differences with Multiple Time P
##
## Group-Time Average Treatment Effects:
##  Group Time ATT(g,t) Std. Error [95% Simult.  Conf. Band]
##   2004 2004  -0.0145     0.0225       -0.0746      0.0455
##   2004 2005  -0.0764     0.0295       -0.1551      0.0023
##   2004 2006  -0.1404     0.0378       -0.2414     -0.0395 *
##   2004 2007  -0.1069     0.0336       -0.1967     -0.0171 *
##   2006 2004  -0.0005     0.0239       -0.0644      0.0634
##   2006 2005  -0.0062     0.0186       -0.0560      0.0436
##   2006 2006   0.0010     0.0191       -0.0501      0.0520
##   2006 2007  -0.0413     0.0192       -0.0925      0.0100
##   2007 2004   0.0267     0.0146       -0.0122      0.0656
##   2007 2005  -0.0046     0.0161       -0.0477      0.0385
##   2007 2006  -0.0284     0.0185       -0.0780      0.0211
##   2007 2007  -0.0288     0.0166       -0.0732      0.0156
## ---
## Signif. codes: `*' confidence band does not cover 0
##
## P-value for pre-test of parallel trends assumption:  0.23267
## Control Group:  Never Treated,  Anticipation Periods:  0
## Estimation Method:  Doubly Robust
```

The `summary()` output reports a p-value for the pre-test of the parallel trends assumption. The null hypothesis is that all pre-treatment group-time ATT(g,t) are jointly equal to zero — i.e., that treated and control counties were on the same trend before treatment. A large p-value (e.g., > 0.05) means we fail to reject this null, which is consistent with parallel trends. In the `mpdta` data the p-value is typically well above 0.05, providing no statistical evidence against the parallel trends assumption in the pre-period.

b) Group-time ATT plot including pre-treatment periods:

```
ggdid(cs_out_bt)
```



```r
ggsave("mpdta_att_gt.pdf", width = 10, height = 6)
```

Each panel shows a different treatment cohort. Negative event-time values (left of zero) are pre-treatment periods; positive values are post-treatment. If parallel trends holds, the pre-treatment ATT(g,t) estimates should scatter around zero with confidence intervals that include zero. Comparing to the aggregated event-study plot from Section 2.2c, this plot is more granular: it shows the cohort-specific pre-trends rather than a single pooled pre-trend, making it easier to spot whether one particular cohort drives any pre-trend concern.

c) Pre-testing has an important limitation: failure to reject parallel trends in the pre-period does not guarantee that the assumption holds in the post-treatment period. The pre-test only examines observable pre-treatment trajectories; any divergence that begins exactly at treatment — whether due to confounders or anticipation effects — would not be detected. The pre-test is therefore a necessary but not sufficient diagnostic for credible DiD identification.

4. Comparing control group specifications

a) CS estimator using not-yet-treated counties as controls:

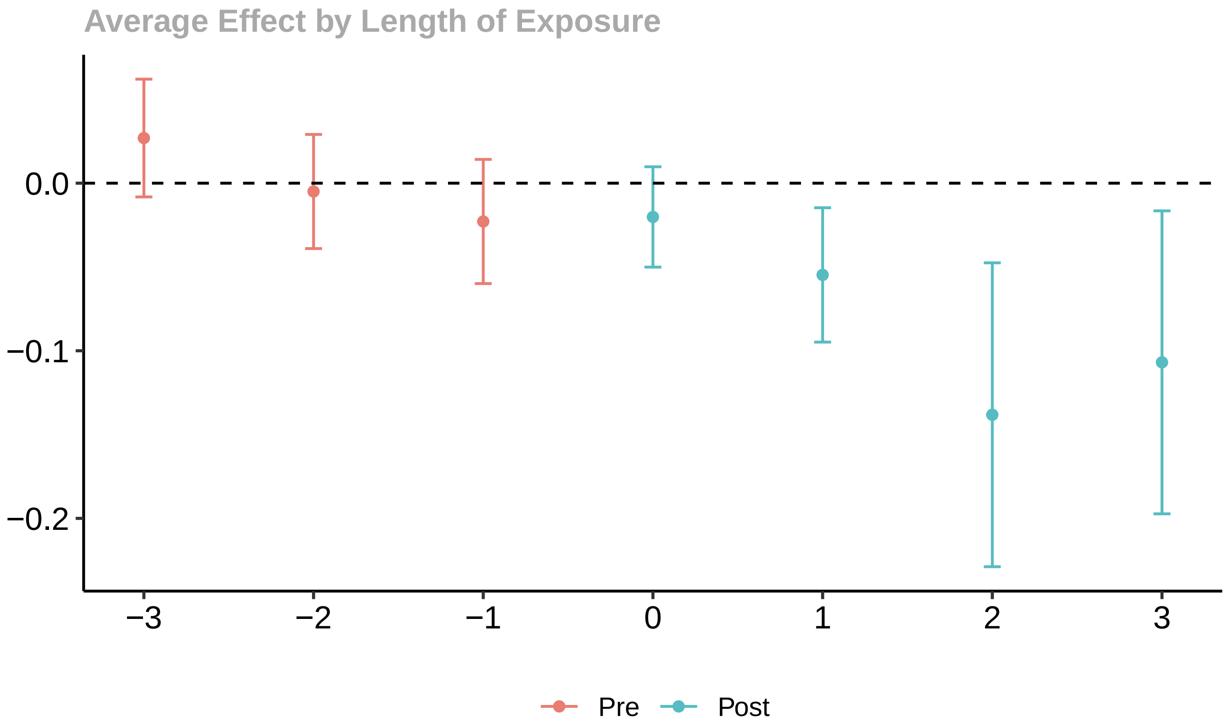
```r
cs_out_nyt = att_gt(
  yname         = "lemp",
  gname         = "first.treat",
  idname        = "countyreal",
  tname         = "year",
  xformla       = ~ lpop,
  data          = mpdta,
  control_group = "notyettreated")

aggte(cs_out_nyt, type = "simple")
```

```
##
## Call:
## aggte(MP = cs_out_nyt, type = "simple")
##
## Reference: Callaway, Brantly and Pedro H.C. Sant'Anna.  "Difference-in-Differences with Multiple Time Pe
##
##
##      ATT    Std. Error     [ 95%  Conf. Int.]
##  -0.0414        0.0113    -0.0635     -0.0192 *
##
##
## ---
## Signif. codes: `*' confidence band does not cover 0
##
## Control Group:  Not Yet Treated,  Anticipation Periods:  0
## Estimation Method:  Doubly Robust
```

The not-yet-treated specification uses counties that will eventually receive treatment but have not yet done so at each point in time, in addition to never-treated counties. This yields a larger effective control group. The overall ATT estimate is typically close to the never-treated estimate in the `mpdta` data, suggesting the control group choice does not dramatically alter the conclusions here.

b) Event-study for the not-yet-treated specification:

```
cs_dyn_nyt = aggte(cs_out_nyt, type = "dynamic")
ggdid(cs_dyn_nyt)
```



```
ggsave("mpdta_event_study_nyt.pdf", width = 7, height = 4)
```

Comparing to the never-treated event study, the pre-trends and post-treatment patterns should look broadly similar. Minor differences arise because the not-yet-treated control group includes counties whose future treatment may already be affecting their current outcomes (anticipation effects), which could contaminate the comparison if treated counties begin adjusting before the official treatment date.

c) The never-treated control group is more conservative: it avoids any contamination from anticipation effects among future-treated units, but may leave fewer control observations in settings where most units eventually get treated. The not-yet-treated control group offers more power and is preferred when the never-treated group is small or unrepresentative, but requires the additional assumption that not-yet-treated units do not anticipate their own future treatment. In this dataset, with a reasonably large never-treated group, either specification is defensible and the estimates are close.

5. Discussion: why does TWFE fail in staggered settings?

a) In a staggered DiD setting, the naive TWFE estimator implicitly uses already-treated units as part of the control group for units that receive treatment later. This is the “forbidden comparison”: if a county that adopted the minimum wage in 2004 is used as a control for a county adopting in 2006, its outcome in 2006 already reflects (potentially growing) effects of its own treatment. When treatment effects are heterogeneous — varying across cohorts or accumulating over time — these forbidden comparisons contaminate the TWFE estimate, assigning negative weights to some group-time comparisons and potentially reversing the sign of the overall estimate. The result is a weighted average where the weights do not correspond to any meaningful population quantity.

b) The TWFE estimate from question 2.2a and the Callaway-Sant’Anna overall ATT from question 2.2b may differ in both magnitude and, in some datasets, sign. The event-study pre-trends from question 2.2c and the formal pre-test from question 3a are the key diagnostics: if the pre-treatment estimates are all close to zero and the joint pre-test p-value is large, the parallel trends assumption is supported, and the Callaway-Sant’Anna estimate is more credible because it uses only valid comparisons (never-treated counties as controls) and allows for cohort-specific treatment effects. The TWFE estimate should be treated with skepticism whenever the staggered design involves heterogeneous timing and potentially heterogeneous effects, as is typical in minimum wage research where economic conditions at adoption differ across states and years.