Recent Advances in Diff-in-Diff

Joel Ferguson

11/01/2021

Treatment effect heterogeneity in Diff-in-Diff

Over the past few years there has been a massive improvement in our understanding of Diff-in-Diff and the tools available for estimating treatment effects using a Diff-in-Diff research design. Like the expansion of our understanding of IV, regarding its interpretation as estimating a Local Average Treatment Effect (we'll cover this in a few weeks), these advances came from thinking about treatment effect heterogeneity and how are we summarizing it with a single estimate. To see why treatment effect heterogeneity matters, it's worth revisiting the most basic form of diff-in-diff estimation.

In a canonical diff-in-diff, like Card and Kreuger (1994), there's a single treated group and a single post period. Letting $i \in \{T, C\}$ denote the treated and control groups respectively, that canonical diff-in-diff estimator is

$$\tau^{DiD} = \mathbb{E}[Y_{T,1}(1) - Y_{T,0}(0)] - \mathbb{E}[T_{C,1}(0) - T_{C,0}(0)]$$

Under the identifying assumption (referred to as the "**parallel trends**" assumption) that $\mathbb{E}[Y_{T,1}(0) - Y_{T,0}(0)] = \mathbb{E}[T_{C,1}(0) - T_{C,0}(0)]$, we get

$$\tau^{DiD} = \mathbb{E}[Y_{T,1}(1) - T_{T,1}(0)]$$

This gives us our first clue that treatment effect heterogeneity matters: canonical diff-in-diff identifies the ATT; we would need to restrict treatment effect heterogeneity between the treatment and control groups to guarantee that diff-in-diff identifies the ATE. However, in this setting there's nothing wrong with the widely-used OLS estimator, as shown below.

```
N <- 5000
N_T <- floor(N/2)
N_C <- N-floor(N/2)
D <- c(rep(1,N_T),rep(0,N_C)) # First half assigned to Treatment, second half to control

Y_0 <- c(rnorm(N_T,5),rnorm(N_C)) # Treated units have period 0 outcomes ~ N(5,1), control ~ N(0,1)
Y_1D0 <- c(rnorm(N_T,6),rnorm(N_C,1)) # Both treated and control have untreated POs for period 1 that a
Y_1D1 <- c(rnorm(N_T,8),rnorm(N_C,1)) # Treated units have treated POs in period 1 that are 2 greater t

ATE <- mean(Y_1D1-Y_1D0)
ATE
```

```
## [1] 0.9896843
```

```r
ATT <- mean(Y_1D1[D==1]-Y_1D0[D==1])
ATT
```

```
## [1] 1.987423
```

```r
Y_1 <- Y_1D1*D+Y_1D0*(1-D)
df <- tibble(D,Y_0,Y_1) %>%
  pivot_longer(cols=c("Y_0","Y_1"),
               names_to="t",
               names_prefix = "Y_",
               values_to="Y") %>%
  mutate(t=as.numeric(t))

OLS <- lm(Y~D*t,
          data=df)
summary(OLS)
```

```
##
## Call:
## lm(formula = Y ~ D * t, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4523 -0.6764 -0.0074  0.6609  3.7727
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.008951   0.019922  -0.449    0.653
## D            4.963424   0.028173 176.174   <2e-16 ***
## t            0.986628   0.028173  35.020   <2e-16 ***
## D:t          2.040947   0.039843  51.224   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9961 on 9996 degrees of freedom
## Multiple R-squared:  0.9115, Adjusted R-squared:  0.9115
## F-statistic: 3.433e+04 on 3 and 9996 DF,  p-value: < 2.2e-16
```

Where things start to go wrong is when we have heterogeneous treatment effects *and* groups of units being treated at different times. The classic estimating equation in this setting estimated via OLS gives us the **two-way fixed effects** (TWFE) diff-in-diff estimator

$$Y_{i,t} = \alpha_i + \delta_t + \beta^{TWFE} D_{i,t} + \varepsilon_{i,t}$$

There are two main issues with this estimator. The first is similar to issues we've seen with OLS (in section) already: OLS puts relatively more weight on observations with a lot of variance in the treatment. In this setting, groups treated near the middle of the time horizon are excessively weighted, meaning their treatment effects will be relatively overrepresented in the estimated treatment effect.

```r
N <- 3000
Ts <- 100
```

```r
N1 <- floor(N/3)
N2 <- floor(N/3)
N3 <- N-2*floor(N/3)

Y0G1 <- as.vector(sapply(c(1:Ts),function(x) rnorm(N1,x))) # Group 1 has untreated potential outcomes t
Y0G2 <- as.vector(sapply(c(1:Ts), function(x) rnorm(N2,x+5))) # Same for Group 2, but they start at 5 r
Y0G3 <- as.vector(sapply(c(1:Ts), function(x) rnorm(N3,x+10))) # And group 3 starts at 10

Y1G1 <- as.vector(sapply(c(1:Ts),function(x) rnorm(N1,x))) # Group 1 has no expected treatment effect
Y1G2 <- as.vector(sapply(c(1:Ts), function(x) rnorm(N2,x+15))) # Group 2 has an expected treatment effe
Y1G3 <- as.vector(sapply(c(1:Ts), function(x) rnorm(N3,x+11))) # And group 3 has an expected treatment

ATE <- mean(c(Y1G1,Y1G2,Y1G3)-c(Y0G1,Y1G2,Y1G3))
ATE
```

```
## [1] 0.0008711167
```

```r
DG1 <- rep(0,N1*Ts) # Group 1 is never treated
DG2 <- c(rep(0,N2*floor(Ts/2)),rep(1,N2*(Ts-floor(Ts/2)))) #Group 2 is treated half the time
DG3 <- c(rep(0,N3*(Ts-1)),rep(1,N3)) # Group 3 is treated in the

ATT <- mean(c(mean(Y1G2[DG2==1]-Y0G2[DG2==1]),mean(Y1G3[DG3==1]-Y0G3[DG3==1])))
ATT
```

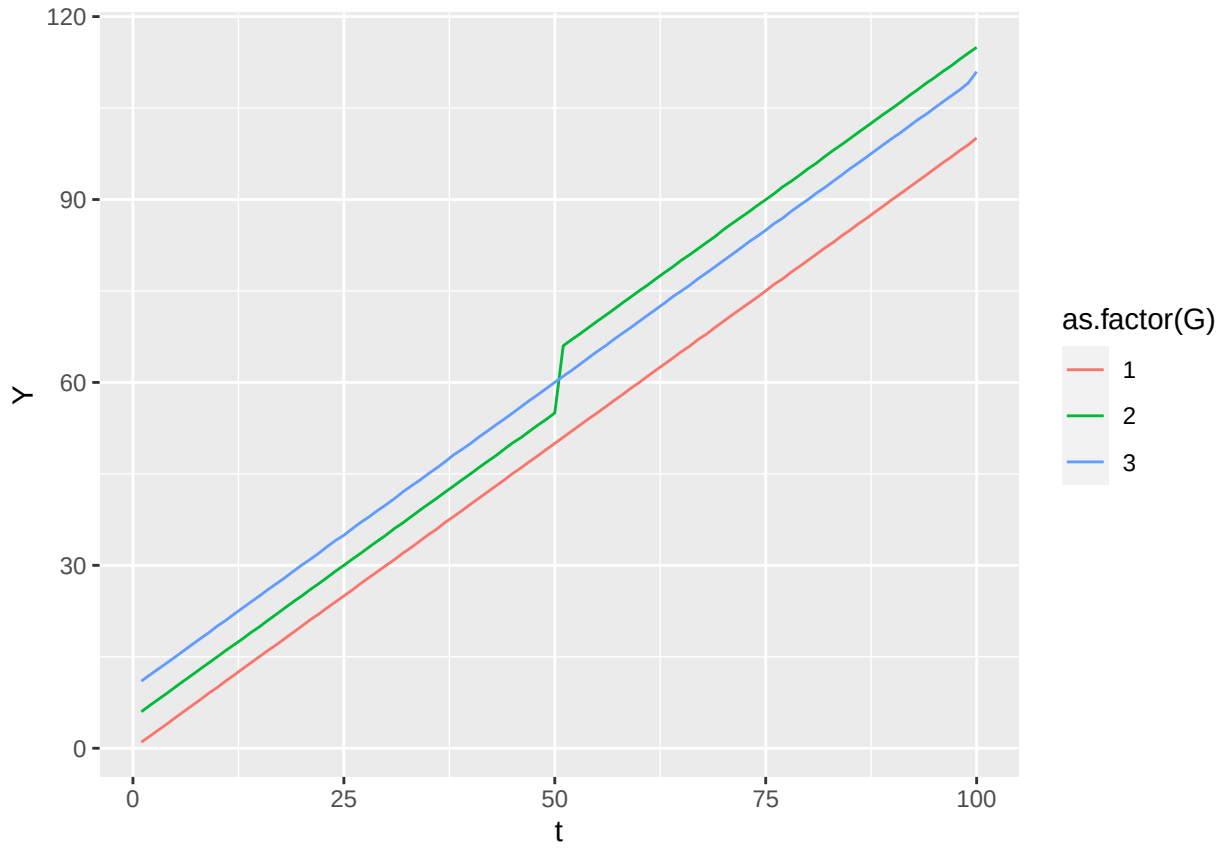
```
## [1] 5.491303
```

```r
df <- tibble("D"=c(DG1,DG2,DG3),
             "Y0"=c(Y0G1,Y0G2,Y0G3),
             "Y1"=c(Y1G1,Y1G2,Y1G3),
             "G"=c(rep(1,N1*Ts),rep(2,N2*Ts),rep(3,N3*Ts)),
             "t"=c(rep(c(1:Ts),each = N1),rep(c(1:Ts),each = N2),rep(c(1:Ts),each = N3))) %>%
  mutate(Y=D*Y1+(1-D)*Y0)

Gmeans <- df %>%
  group_by(G,t) %>%
  summarise(Y=mean(Y))
```

```
## `summarise()` regrouping output by 'G' (override with `.groups` argument)
```

```r
ggplot(Gmeans)+
  geom_line(aes(x=t,y=Y,group=G,color=as.factor(G)))
```




```
TWFE <- feols(Y~D|
                G+t,
              data=df)
summary(TWFE)
```

```
## OLS estimation, Dep. Var.: Y
## Observations: 300,000
## Fixed-effects: G: 3,  t: 100
## Standard-errors: Clustered (G)
##   Estimate Std. Error t value   Pr(>|t|)
## D  9.72648   0.268907 36.1704 0.00076348 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.08289     Adj. R2: 0.998764
##                 Within R2: 0.820461
```

In this case OLS severely overestimates the ATT. Loosely speaking, OLS is doing a good job of estimating the period-specific ATTs, but it gives all of them equal weight, whereas the only one that gives us any idea of how group 3 contributes to the ATT is the one from the last period.

The second issue is referred to as the "negative weighting problem." At a high level, this issue arises because OLS doesn't differentiate between "good" comparisons, ones where we compare a group that's becoming treated to a group that's remaining untreated, and "bad" comparisons, one where we compare a group that's becoming treated to a group that's remaining *treated*. Consider a setting similar to a canonical diff-in-diff, but instead of having a group that never receives treatment, we have a group that is always treated. In a

way, this group is also a control since its treatment status never changes. And indeed, as long as treatment effects don't change within the control group over time we can still do diff-in-diff without issue in this setting

$$\tau^{DiD} = \mathbb{E}[Y_{T,1}(1) - Y_{T,0}(0)] - \mathbb{E}[T_{C,1}(1) - T_{C,0}(1)] = \mathbb{E}[Y_{T,1}(1) - Y_{T,1}(0)] - \mathbb{E}[(Y_{C,1}(1) - Y_{C,1}(0)) - (Y_{C,0}(1) - Y_{C,0}(0))]$$

The final expectation in the expression above shows that if treatment effects don't evolve over time $\mathbb{E}[(Y_{C,1}(1) - Y_{C,1}(0)) - (Y_{C,0}(1) - Y_{C,0}(0))] = 0$ we recover the ATT. However, we haven't made any assumptions about the evolution of treated potential outcomes, and in fact we often are explicitly interested in their evolution over time. As such, without an alternative restriction on the evolution of potential outcomes (one that restricts treatment effect heterogeneity), we don't even necessarily identify the ATT. This is referred to as the "negative weighting problem" because, as Andrew Goodman-Bacon showed in his seminal 2021 *Journal of Econometrics* article, $\hat{\beta}^{TWFE}$ can be written as a weighted average of all valid canonical diff-in-diff estimates in the data. For "bad" comparisons, like the one we just analyzed, the weight assigned is negative[1].

Dealing with Heterogeneity in Diff-in-Diff

So the TWFE estimator is not good, as we've seen. Ideally, an alternative would deal with both issues: the reweighting of group-specific ATTs by treatment timing, and "bad" comparisons. There are many proposed solutions, my favorite was developed by Borusyak, Jaravel, and Spiess (2021).The way it derives point estimates is incredibly intuitive and accords very well with how we think diff-in-diff should work. In particular, it simply predicts untreated potential outcomes using estimates of unit and time period treatment effects, then takes a weighted average of differences between observed outcomes and predicted untreated outcomes.

```
bjs_Y0 <- feols(Y~1|G+t,
                data=df[df$D==0,])
df$Y0_hat <- predict(bjs_Y0,newdata=df)

bjs_est <- df %>%
  mutate(ATT=Y-Y0_hat) %>%
  filter(D==1) %>%
  group_by(G) %>%
  summarize(ATT=mean(ATT))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
mean(bjs_est$ATT)
```

```
## [1] 5.422074
```

The BJS estimator does much better than our TWFE estimator here because it separately calculates the ATT by group and then averages them[2]. However, it should be clear that it also avoids bad comparisons because we only use untreated observations to predict untreated outcomes.

[1] The estimates for these "bad" comparisons are also the negative of what we derived. The issue is more that the parallel trends assumption doesn't restrict evolution of *treated* potential outcomes over time.

[2] Actually, BJS lets you average them up however you want!