An identification problem in DID

Miao Cai

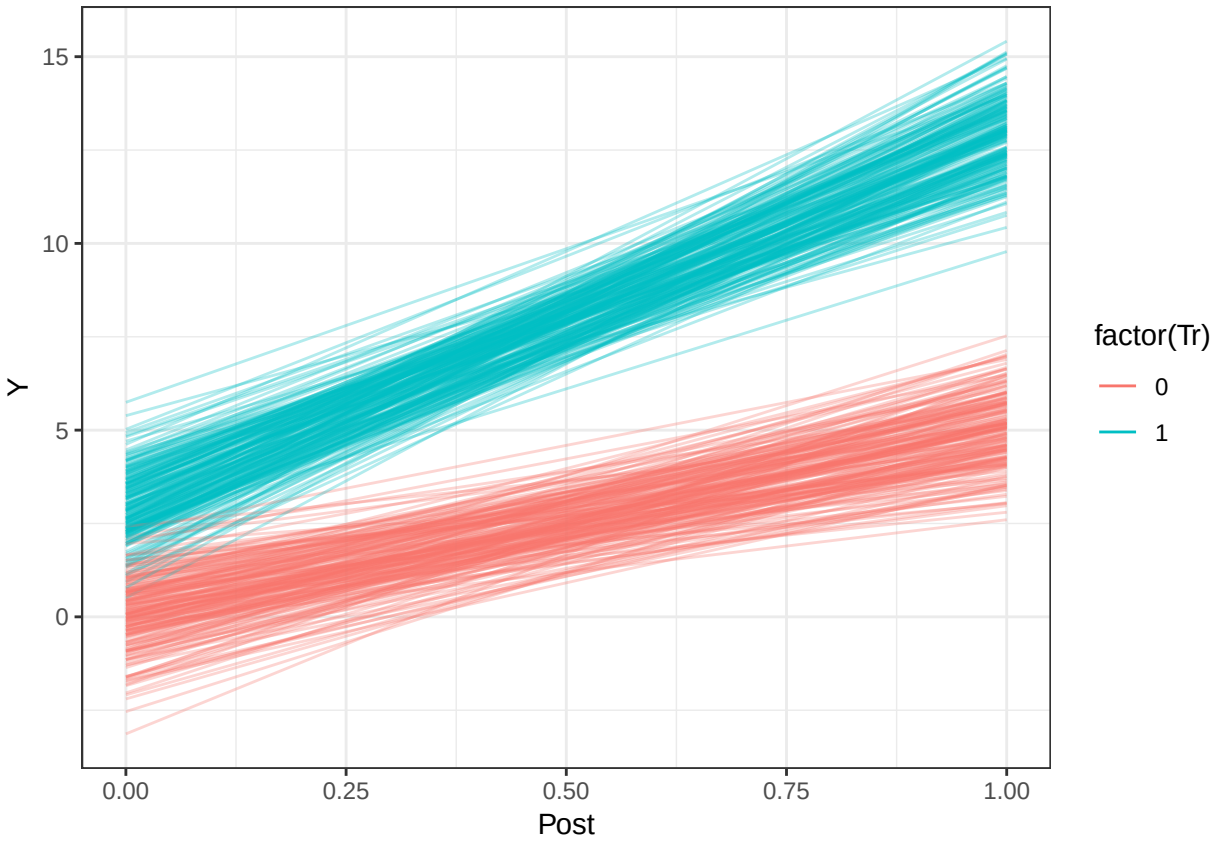
12/20/2018

In classic difference-in-difference (DID) settings, there are only two periods, before and after.

$$Y = \beta_0 + \beta_1 \cdot Tr + \beta_2 \cdot Post + \beta_3 T \cdot Post + \epsilon$$

Where T is an indicator of treatment or not, $Post$ is an indicator of pre or post periods, and β_3 is the DID estimator of treatment causal effect.

```
set.seed(666)

nunits = 400
nyear = 2

b3 = 5
b1 = 3
b2 = 5
b0 = 0

Group = rep(1:nunits, each = nyear)
Year = rep(2007:(2007+nyear-1), times = nunits)
Tr = rep(sample(0:1, nunits, replace = TRUE), each = nyear)
Post = ifelse(Year<2008, 0, 1)



Y = b0 + b1*Tr + b2*Post + b3*Tr*Post  + rnorm(nunits*nyear, 0, 1)

dat = data.frame(Y, Group, Tr, Post)

require(tidyverse)

dat %>%
  ggplot(aes(x = Post, y = Y, group = Group, color =
               factor(Tr))) +
  geom_line(alpha = 0.3)+
  guides(colour = guide_legend(override.aes = list(alpha = 1))) +
  theme_bw()
```



In modern econometrics, long panel data are more and more commonly used. Researchers typically add year dummy variables to allow for time trend fixed effects in their models.

$$Y = \beta_0 + \beta_1 \cdot Tr + \beta_2 \cdot Post + \beta_3 \cdot T \cdot Post + \beta_4 \cdot Year + \epsilon$$

However, identification may be a potential issue in this case since $Post$ and $Year$ dummies have a lot in common. In this demonstration, I will use simulation to test the identification issue in this scenario.

```
set.seed(666)

nunits = 400
nyear = 10

b0 = 1
b1 = 2
b2 = 3
b3 = 4

Group = rep(1:nunits, each = nyear)
Year = rep(2007:(2007+nyear-1), times = nunits)
Tr = rep(sample(0:1, nunits, replace = TRUE), each = nyear)
Post = ifelse(Year<2013, 0, 1)

byear = c(0, rnorm(nyear-1, 0, 1))
yyear = model.matrix( ~ YEAR-1, data=data.frame(YEAR = factor(Year))) %*% byear
```

```
Y = b0 + b1*Tr + b2*Post + b3*Tr*Post + yyear + rnorm(nunits*nyear, 0, 1)

dat = data.frame(Y, Group, Tr, Post, Year)

require(tidyverse)

dat %>%
  ggplot(aes(x = Year, y = Y, group = Group, color = factor(Tr))) +
  geom_line(alpha = 0.3) +
  guides(colour = guide_legend(override.aes = list(alpha = 1))) +
  theme_bw()
```

- *nobs*: 4000
- $Year$: 2007 - 2016, in total 10 years
- Tr: 0 or 1
- $Post$: 0 or 1
- $\beta_0 = 1$
- $\beta_1 = 2$
- $\beta_2 = 3$
- $\beta_3 = 4$

```
didfit = lm(Y ~ Tr + Post + Tr*Post + factor(Year), data = dat)
summary(didfit)
```

```
##
## Call:
## lm(formula = Y ~ Tr + Post + Tr * Post + factor(Year), data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2349 -0.6484  0.0036  0.6683  3.1994
##
## Coefficients: (1 not defined because of singularities)
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.99494    0.05365  18.545  < 2e-16 ***
## Tr                 2.01737    0.04037  49.969  < 2e-16 ***
## Post               4.16517    0.07730  53.884  < 2e-16 ***
## factor(Year)2008   2.03704    0.06988  29.149  < 2e-16 ***
## factor(Year)2009   1.31346    0.06988  18.795  < 2e-16 ***
## factor(Year)2010  -0.19983    0.06988  -2.859  0.00427 **
## factor(Year)2011   1.27712    0.06988  18.275  < 2e-16 ***
## factor(Year)2012   0.57595    0.06988   8.241 2.28e-16 ***
## factor(Year)2013  -1.27254    0.06988 -18.209  < 2e-16 ***
## factor(Year)2014   0.89008    0.06988  12.736  < 2e-16 ***
## factor(Year)2015   0.35510    0.06988   5.081 3.92e-07 ***
## factor(Year)2016        NA         NA      NA       NA
## Tr:Post            3.99777    0.06383  62.627  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9883 on 3988 degrees of freedom
## Multiple R-squared:  0.9239, Adjusted R-squared:  0.9237
## F-statistic:  4404 on 11 and 3988 DF,  p-value: < 2.2e-16
```

```r
namerow = rownames(summary(didfit)$coefficients)

estimates = round(summary(didfit)$coefficients[namerow %in% c("(Intercept)", "Tr", "Post", "Tr:Post"), 1
parameters = c(b0, b1, b2, b3)
```

- The parameters are 1, 2, 3, 4
- The estimates are 0.99, 2.02, 4.17, 4

It seems that there are identification issues here: the estimate of β_2, which is estimates the post-treatment fixed effect, is biased by one unit. The model cannot generate estimate for the last year's fixed effect.