Assignment 6

Grant Jackson

2024-11-18

```r
setwd("~/Applied Microeconometrics/Data")

tva <- read_csv('tva.csv')
```

```
## Rows: 13675 Columns: 18
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (1): county_code
## dbl (16): year, tva, treat, post, ln_agriculture, ln_manufacturing, agricult...
## lgl  (1): county_has_no_missing
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Question 1: Different DiD Estimation Methods

```r
tva_short <- tva |>
  filter(year == 1940 | year == 1960)

tva_short
```

```
## # A tibble: 5,470 x 18
##    county_code  year   tva treat  post ln_agriculture ln_manufacturing
##    <chr>       <dbl> <dbl> <dbl> <dbl>          <dbl>            <dbl>
##  1 01001        1940     0     0     0           8.39             6.66
##  2 01001        1960     0     0     1           7.12             7.26
##  3 01003        1940     0     0     0           8.27             7.20
##  4 01003        1960     0     0     1           7.59             8.14
##  5 01005        1940     0     0     0           8.75             7.45
##  6 01005        1960     0     0     1           7.44             7.46
##  7 01007        1940     0     0     0           7.77             6.17
##  8 01007        1960     0     0     1           6.28             7.47
##  9 01009        1940     0     0     0           8.72             6.17
## 10 01009        1960     0     0     1           7.61             7.53
## # i 5,460 more rows
## # i 11 more variables: agriculture_share_1920 <dbl>,
## #   agriculture_share_1930 <dbl>, manufacturing_share_1920 <dbl>,
## #   manufacturing_share_1930 <dbl>, ln_avg_farm_value_1920 <dbl>,
## #   ln_avg_farm_value_1930 <dbl>, white_share_1920 <dbl>,
## #   white_share_1930 <dbl>, white_share_1920_sq <dbl>,
## #   white_share_1930_sq <dbl>, county_has_no_missing <lgl>
```

Method 1

```
ybar_treat_post <- tva |>
  filter(year == 1960 & tva == 1) |>
  pull(ln_manufacturing) |> mean()

ybar_treat_pre <- tva |>
  filter(year == 1940 & tva == 1) |>
  pull(ln_manufacturing) |> mean()

ybar_control_post <- tva |>
  filter(year == 1960 & tva == 0) |>
  pull(ln_manufacturing) |> mean()

ybar_control_pre <- tva |>
  filter(year == 1940 & tva == 0) |>
  pull(ln_manufacturing) |> mean()

did_manual <- (ybar_treat_post - ybar_treat_pre) - (ybar_control_post - ybar_control_pre)

print(paste("Manual DiD Estimate:", did_manual))
```

```
## [1] "Manual DiD Estimate: 0.277418918170987"
```

Method 2

```
reg_did <- feols(
  ln_manufacturing ~
    i(tva, year == 1960, ref = 0) | county_code + year,
  data = tva_short, vcov = "hc1"
)


print(reg_did)
```

```
## OLS estimation, Dep. Var.: ln_manufacturing
## Observations: 5,470
## Fixed-effects: county_code: 2,735,  year: 2
## Standard-errors: Heteroskedasticity-robust
##                       Estimate Std. Error t value   Pr(>|t|)
## tva::1:year == 1960 0.277419   0.046994 5.90324 4.0043e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.312065     Adj. R2: 0.947108
##                  Within R2: 0.011817
```

Method 3

```
first_diff <- tva_short |>
  mutate(
    .by = county_code,
    delta_ln_manufacturing =
      ln_manufacturing[year == 1960] -
      ln_manufacturing[year == 1940]
```

```r
  ) |>
  filter(year == 1960)

did_fd <- feols(
  delta_ln_manufacturing ~ i(tva, ref = 0),
  data = first_diff,
  vcov = 'hc1'
)

print(did_fd)
```

```
## OLS estimation, Dep. Var.: delta_ln_manufacturing
## Observations: 2,735
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error  t value   Pr(>|t|)
## (Intercept) 0.750600   0.012373 60.66309  < 2.2e-16 ***
## tva::1      0.277419   0.046994  5.90324 4.0043e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.62413   Adj. R2: 0.011456
```

We can verify that the methods all produce the same estimate.

Question 2: Requirements for Causal Interpretation

Key assumptions needed for causal interpretation of these estimates are:

1. Parallel Trends Assumption = Treated and control counties would have followed the same trend in the absence of the treatment(tva)
2. No Spillover Effects = TVA treatment didn't affect manufacutring employment in control counties
3. No Anticipation Effects = Behavior of counties didn't change before treatment because of anticiaption of TVA
4. Stable Unit Treatment Assumption = Each country's outcomes only depends on its own treatment status

Question 3: Pre-trends and 1950 Estimates

```r
calc_did <- function(start_year, end_year) {
  temp_data <- tva |>
    filter(year %in% c(start_year, end_year)) |>
    mutate(
      .by = county_code,
      delta_ln_manufacturing =
        ln_manufacturing[year == end_year] -
        ln_manufacturing[year == start_year]
    ) |>
    filter(year == end_year)

  model <- feols(
    delta_ln_manufacturing ~ i(tva, ref = 0),
    data = temp_data,
    vcov = "HC1"
  )
```

```r
  return(model)
}

did_1920 <- calc_did(1940, 1920) # pre-trend
did_1930 <- calc_did(1940, 1930) # pre-trend
did_1950 <- calc_did(1940, 1950)

print("Pre-trend 1920-1940:")
```

```
## [1] "Pre-trend 1920-1940:"
```

```r
print(did_1920)
```

```
## OLS estimation, Dep. Var.: delta_ln_manufacturing
## Observations: 2,735
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -0.851150   0.036988 -23.01152 < 2.2e-16 ***
## tva::1      -0.461013   0.159185  -2.89608 0.0038086 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.88286   Adj. R2: 0.003251
```

```r
print("Pre-trend 1930-1940:")
```

```
## [1] "Pre-trend 1930-1940:"
```

```r
print(did_1930)
```

```
## OLS estimation, Dep. Var.: delta_ln_manufacturing
## Observations: 2,735
## Standard-errors: Heteroskedasticity-robust
##              Estimate Std. Error   t value  Pr(>|t|)
## (Intercept) -1.000817   0.031758 -31.51361 < 2.2e-16 ***
## tva::1      -0.335189   0.141376  -2.37090  0.017814 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.62126   Adj. R2: 0.002216
```

```r
print("Effect 1940-1950:")
```

```
## [1] "Effect 1940-1950:"
```

```r
print(did_1950)
```

```
## OLS estimation, Dep. Var.: delta_ln_manufacturing
## Observations: 2,735
## Standard-errors: Heteroskedasticity-robust
##             Estimate Std. Error  t value  Pr(>|t|)
## (Intercept) 0.438941   0.008443 51.98915 < 2.2e-16 ***
## tva::1      0.071753   0.032131  2.23313  0.025621 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 0.425933   Adj. R2: 0.00135
```

Question 4: Event Study

```r
tva_event <- tva |>
  mutate(
    event_time = if_else(
      tva == 0,
      -10,  # for untreated group
      year - 1950  # for treated group
    )
  )

event_study <- feols(
  ln_manufacturing ~
    i(tva, i.event_time, ref = 0, ref2 = -10) |
    county_code + year,
  data = tva_event,
  vcov = 'hc1'
)

print(event_study)
```

```
## OLS estimation, Dep. Var.: ln_manufacturing
## Observations: 13,675
## Fixed-effects: county_code: 2,735,  year: 5
## Standard-errors: Heteroskedasticity-robust
##                          Estimate Std. Error   t value  Pr(>|t|)
## tva::1:event_time::-30 -0.461013   0.152354 -3.025927 0.0024845 **
## tva::1:event_time::-20 -0.335189   0.137928 -2.430182 0.0151072 *
## tva::1:event_time::0    0.071753   0.076748  0.934921 0.3498498
## tva::1:event_time::10   0.277419   0.085143  3.258275 0.0011243 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## RMSE: 1.0067     Adj. R2: 0.79181
##                Within R2: 0.004365
```

We can verify the methods from questions 3 and 4 produce the same results.

Conditional Parallel Trends

Question 1: Argument for Including Baseline Manufacturing Share

Including baseline manufacturing share as co-variate is important because:

1. Captures pre-existing industrial development
2. Areas with different initial manufacturing levels could have different growth movement
3. Helps control for mean reversion in manufacturing employment
4. Accounts for potential convergence patterns in regional development

Question 2: Regression Adjustment Estimator

```r
covariates <- c(
  "agriculture_share_1920", "agriculture_share_1930",
  "manufacturing_share_1920", "manufacturing_share_1930",
  "white_share_1920", "white_share_1930"
)
```

```r
reg_adj <- feols(
  delta_ln_manufacturing ~ i(tva, ref = 0) +
    agriculture_share_1920 + agriculture_share_1930 +
    manufacturing_share_1920 + manufacturing_share_1930 +
    white_share_1920 + white_share_1930,
  data = first_diff
)

etable(reg_adj)
```

```
##                                          reg_adj
## Dependent Var.:          delta_ln_manufacturing
##
## Constant                      0.4905*** (0.0840)
## tva = 1                       0.2362*** (0.0468)
## agriculture_share_1920          -0.0695 (0.0965)
## agriculture_share_1930         0.2513** (0.0948)
## manufacturing_share_1920    -0.5200*** (0.1509)
## manufacturing_share_1930     -1.048*** (0.1748)
## white_share_1920                 0.1219 (0.1604)
## white_share_1930                 0.2542 (0.1578)
## ________________________ ______________________
## S.E. type                                   IID
## Observations                              2,735
## R2                                      0.12176
## Adj. R2                                 0.11951
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question 3: DRDID Panel Regression

```r
drdid_data <- tva_short |>
  arrange(county_code, year)

y1 <- drdid_data$ln_manufacturing[drdid_data$year == 1960]
y0 <- drdid_data$ln_manufacturing[drdid_data$year == 1940]
D <- drdid_data$tva[drdid_data$year == 1960]
covariates <- as.matrix(drdid_data[drdid_data$year == 1960,
                                   c("agriculture_share_1920",
                                     "agriculture_share_1930",
                                     "manufacturing_share_1920",
                                     "manufacturing_share_1930",
                                     "white_share_1920",
                                     "white_share_1930")])

covariates <- cbind(1, covariates)

drdid_panel <- DRDID::reg_did_panel(
  y1 = y1,
  y0 = y0,
  D = D,
  covariates = covariates,
  boot = TRUE,            # Use bootstrap for inference (optional)
```

```r
  nboot = 999             # Number of bootstrap repetitions
)

summary(drdid_panel)
```

```
##  Call:
## DRDID::reg_did_panel(y1 = y1, y0 = y0, D = D, covariates = covariates,
##     boot = TRUE, nboot = 999)
## ------------------------------------------------------------------
##  Outcome-Regression DID estimator for the ATT:
##
##    ATT     Std. Error  t value    Pr(>|t|)  [95% Conf. Interval]
##   0.244      0.0407     5.9979       0        0.1572     0.3307
## ------------------------------------------------------------------
##  Estimator based on panel data.
##  Outcome regression est. method: OLS.
##  Boostrapped standard error based on 999 bootstrap draws.
##  Bootstrap method: weighted .
## ------------------------------------------------------------------
##  See Sant'Anna and Zhao (2020) for details.
```

Question 4: Doubly-robust Estimator

```r
# Perform Doubly-Robust DiD
drdid_results <- DRDID::drdid_panel(
  y1 = y1,
  y0 = y0,
  D = D,
  covariates = covariates,
  boot = TRUE,     # Bootstrap for inference
  nboot = 999      # Number of bootstrap repetitions
)

summary(drdid_results)
```

```
##  Call:
## DRDID::drdid_panel(y1 = y1, y0 = y0, D = D, covariates = covariates,
##     boot = TRUE, nboot = 999)
## ------------------------------------------------------------------
##  Locally efficient DR DID estimator for the ATT:
##
##    ATT     Std. Error  t value    Pr(>|t|)  [95% Conf. Interval]
##   0.2234     0.0428     5.2204       0        0.1405     0.3064
## ------------------------------------------------------------------
##  Estimator based on panel data.
##  Outcome regression est. method: OLS.
##  Propensity score est. method: maximum likelihood.
##  Boostrapped standard error based on 999 bootstrap draws.
##  Bootstrap method: weighted .
## ------------------------------------------------------------------
##  See Sant'Anna and Zhao (2020) for details.
```