

Assignment #6

ECON 5783 — University of Arkansas

Prof. Kyle Butts

These assignments should be completed in groups of 1 or 2 but submitted individually. My preference is for you to use Rmarkdown files to have your code, results, and your answers to the questions intermixed. Since I am not requiring you to code in R for these assignments, you can use latex or microsoft word to write up answers alternatively. Not that in these cases, I would like you to upload your code separately.

Coding Exercise: Kline and Moretti (2014, QJE)

This exercise will have you work with replication material from Kline and Moretti (2014, QJE). The dataset can be loaded from `tva.csv`. The paper analyzes the effects of a large-scale New Deal intervention called the Tennessee Valley Authority.

The data set contains census data on US counties measured every decade (year). The treatment variable, D_i , equals 1 if the county is in the Tennessee Valley Authority (`tva`). For an outcome variable, we focus on the log manufacturing employment in a county (`ln_manufacturing`).

Difference-in-differences estimation

1. First, create a data frame that contains the years 1940 (pre-period) and 1960 (post-period). Verify that these three ways to estimate a difference-in-differences estimate are identical:
 - i. Calculate the difference-in-differences estimate by hand, i.e. taking the four sample means of `ln_manufacturing`.
 - ii. Then, calculate this using a regression with county and year fixed-effects with `feols`.
 - iii. Last, estimate this using first-differenced data (i.e. regressing $y_{i,1960} - y_{i,1940}$ on an intercept and TVA_i).
2. Comment in words what you would need to believe for this estimate to be interpreted as causal
3. Now, estimate 2×2 difference-in-differences estimates for 1950 and pre-trend estimates using 1920 and 1930. *Note: for 1920 and 1930, still use 1940 as your 'base period', i.e. $y_{i,1920} - y_{i,1940}$*

4. Now, create a variable for event time $\times D_i$. This variable should take the value $t - 1950$ for treated units (i.e. 1920 \rightarrow -30, 1930 \rightarrow -20, etc.) and the value -10 for untreated units. Then, run a regression of y_{it} on these event-study indicators (dropping -10) and county and year fixed-effects. Verify these estimates are the same as the previous question

Conditional Parallel Trends

In the context of the Tennessee Valley Authority, this program was targetted towards an area with lower access to electricity and far less manufacturing than the rest of the US. “Regional convergence”, an economic theory of development, would suggest this region would have faster counterfactual manufacturing growth (in the absence of treatment) than the rest of the country.

Kline and Moretti, therefore, wanted to use a set of covariates to better model these non-parallel trends. The assumption is that comparing units with the same values of X_i would be on the same trends. We will use the following covariates: `agriculture_share_1920`, `agriculture_share_1930`, `manufacturing_share_1920`, `manufacturing_share_1930`, `white_share_1920`, and `white_share_1930`.

1. In words, make an argument why you would include baseline manufacturing share as a covariate.
2. Let’s return to the dataset subset to 1940 and 1960. Create a variable containing first-differenced outcome $\Delta y_i = y_{i,1960} - y_{i,1940}$ and call it `delta_ln_manufacturing`. Then subset the data to just 1960 (see the next page for an example of how to do this).

With this first-differenced data, perform the regression adjustment estimator (refer back to previous labs) of Δy_i on D_i conditional on X_i

3. Now, let’s try this with the DRDID package and the `DRDID::reg_did_panel()` function
4. Last, try the `DRDID::drdid()` function to perform a doubly-rboust estimator (Augmented IPTW Estimator).

Question 1 Hint:

```
df |>
  mutate(
    .by = county_code,
    delta_ln_manufacturing =
      ln_manufacturing - ln_manufacturing[year == 1940]
  ) |>
  filter(year == 1960)
```