Quantitative Research Methods IV - 17.806

Recitation, Week 8.

Topic: Longitudinal Data II and Survival Analysis I.

Benjamín Muñoz

April 7, 2023

MIT

Massachusetts
Institute of
Technology

Table of contents

1. Longitudinal Causal Inference

2. Math Review

3. Survival Analysis: Basics

1/ Longitudinal Causal Inference

Panel Causal Inference

- **Multiple Tools:**
  1. Fixed Effects.
  2. Difference-in-Differences.
  3. Synthetic Control Method.
  4. Model-Based Counterfactual Estimators: FEct, IFEct and MC (see Liu, Wang & Xu, 2022).
  5. Panel Matching, Trajectory Balancing.

- **Quantity of Interest:** Average Treatment Effect on the Treated (ATT).

$$ATT = \mathbb{E}[Y(1)_i - Y(0)_i | D_i = 1]$$

- **Counterfactual:** using information from both the past and others units.

Descriptive Analysis

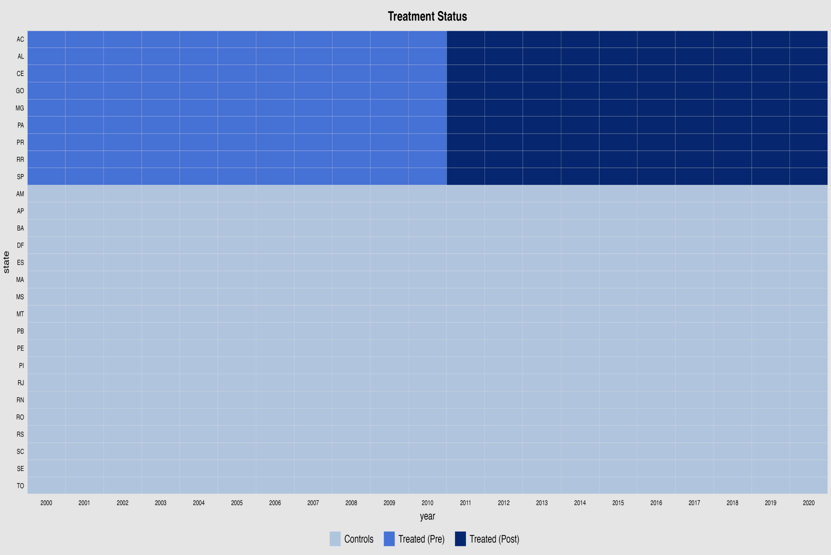
- **Caveat:** treatment assignment mechanism may be complicated.
- Always perform a descriptive analysis of treatment patterns!
- Use the `panelView` package. More information here.

R Code

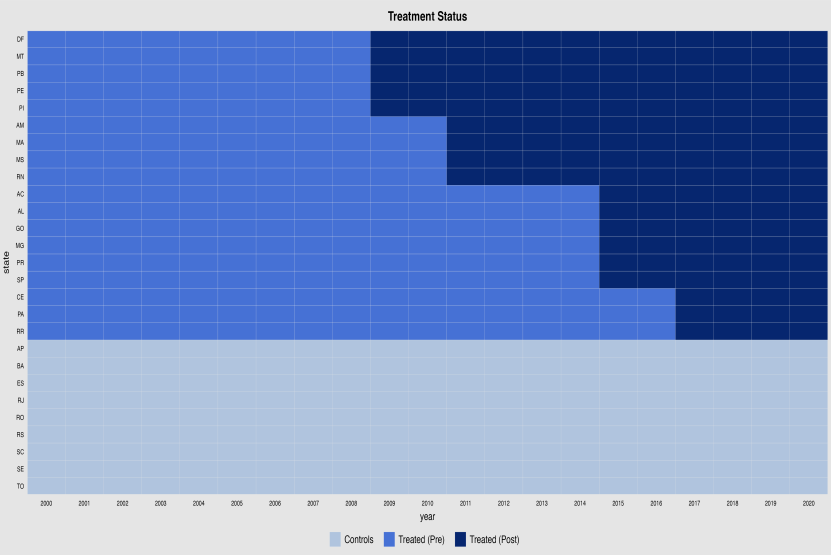
```
### Load packages
library(panelView)

### Create Plot
panelview(data = df, formula = Y ˜ D,
          index = c(”state”, ”year”),
          type = ”treat”,
          outcome.type = ”continuous”,
          treat.type = ”discrete”,
          pre.post = TRUE, by.timing = TRUE)
```

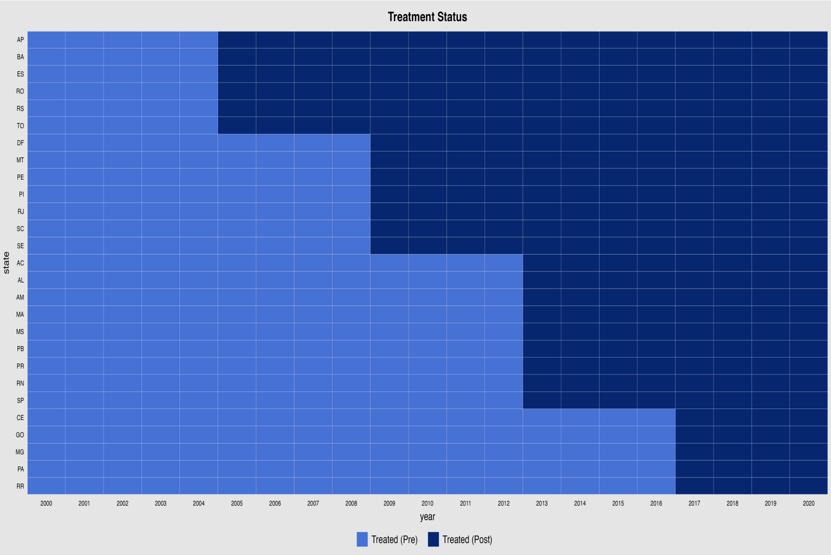
Descriptive Examples



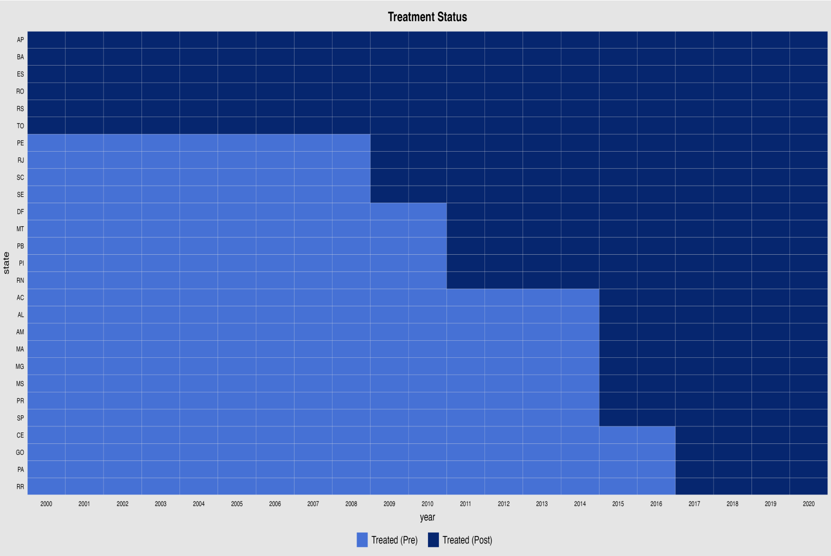
Descriptive Examples



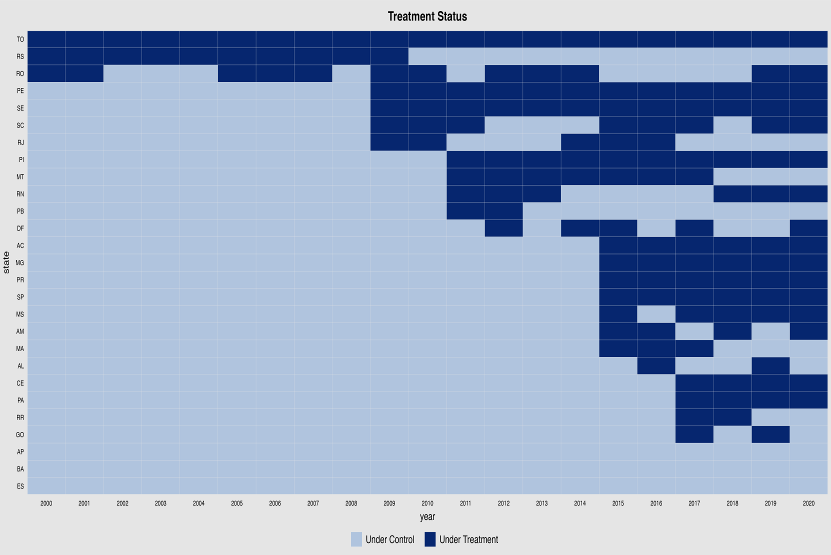
Descriptive Examples



Descriptive Examples



Descriptive Examples



DiD

- Things to consider:
  - Treatment Measurement Level.
  - Group Size
  - Treatment Timing + Reversal.
  - Balanced/Unbalanced Nature.
- **Key Assumption:** Parallel Trends.

$$\mathbb{E}[Y_{i2}(0) - Y_{i1}(0)|G_i = 1] = \mathbb{E}[Y_{i2}(0) - Y_{i1}(0)|G_i = 0]$$

- Typical Estimator: $\mathrm{Y_{it}} = \theta_\mathrm{t} + \eta_\mathrm{i} + \alpha \mathrm{D_{it}} + \beta \mathrm{X_{it}} + \upsilon_\mathrm{it}$
- Linear Form Assumption.
- 2×2 DiD $=$ TWFE
- Multi-Period DiD $\neq$ TWFE. Multi-Period DID $=$ Weighted TWFE $\rightsquigarrow$ Negative weights.

Goodman-Bacon Decomposition

- TWFE is a weighted average of all possible 2×2 DiD estimators that compare timing groups to each other. **Key Idea:** $\sum W_i = 1$.

- Size of Weights: timing, group sizes, and the variance of the treatment in each pair .

- $\text{plim}_{N \to \infty} \beta^{\widehat{\text{TWFE}}} = \text{VWATT} + \text{VWCT} - \underbrace{\Delta ATT}_{\text{Change in Treatment Effects over Time}}$

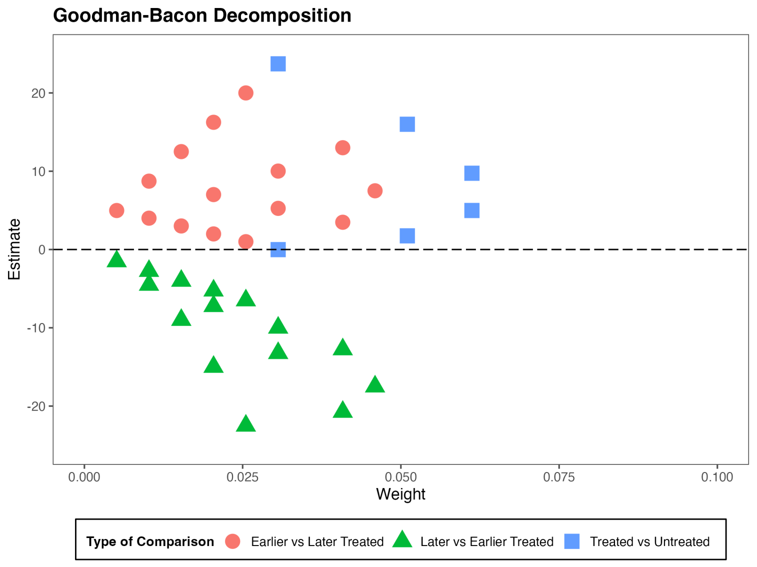
R Code

```
### Load packages
library(bacondecomp)

### Run TWFE
fit_tw <- lm(l_homicide ˜ post + factor(state) + factor(year),  data = bacondecomp::castle)

### Run Goodman-Bacon Decomposition
df_bacon <- bacon(formula = l_homicide ˜ post,  data = bacondecomp::castle,
id_var = "state",  time_var = "year")
```

Goodman-Bacon Decomposition



Callaway and Sant'Anna

- Notation:
  - $\mathcal{T}$ periods where t can be $t = 1, \ldots, \mathcal{T}$.
  - G_g is a binary variable denoting whether a unit is first treated in period g.
  - C is a binary variable denoting a unit that is never treated.
- One estimand is the group-time average treatment effect:

$$ATT(g,t) = \mathbb{E}\left[Y_t(1) - Y_t(0) | G_g = 1\right]$$

- Key Identifying Assumption is parallel trends based on never treated units: For all $t = 2, \ldots, \mathcal{T}$ such that $g \leq t$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | C = 1\right]$$

Callaway and Sant’Anna

- Alternative: parallel trends based on not-yet treated units: For all $t = 2, \ldots, \mathcal{T}$ such that $g \leq t$,

$$\mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | G_g = 1\right] = \mathbb{E}\left[Y_t(0) - Y_{t-1}(0) | D_s = 0, G_g = 0\right]$$

R Code

```
### Load packages
library(did)

### Run Callaway and Sant'Anna estimator
attgt <- att_gt(yname = "Y", tname = "Time", idname = "Unit",
gname = "Cohort", xformla = NULL, data = sim_df,
control_group = "nevertreated", biters = 5000, cores = 2)
```

2/ Math Review

Probability

Law of the Unconsciou Statistician (LOTUS)

$$\mathbb{E}[g(X)] = \sum_{x \in R_X} g(x) P_X x$$

Probability: Random Vectors

$$\text{Joint PMF: } P_{X,Y}(x, y) = P(X = x, Y = y)$$

$$\text{Marginal PMF: } P_X(x) = \sum_{y \in R_Y} P_{X,Y}(x, y_j)$$

$$\text{Conditional PMF: } P_{X|Y}(x|y) = \frac{P_{X,Y}(x, y)}{P_Y(y)}$$

Calculus

Integration by Parts

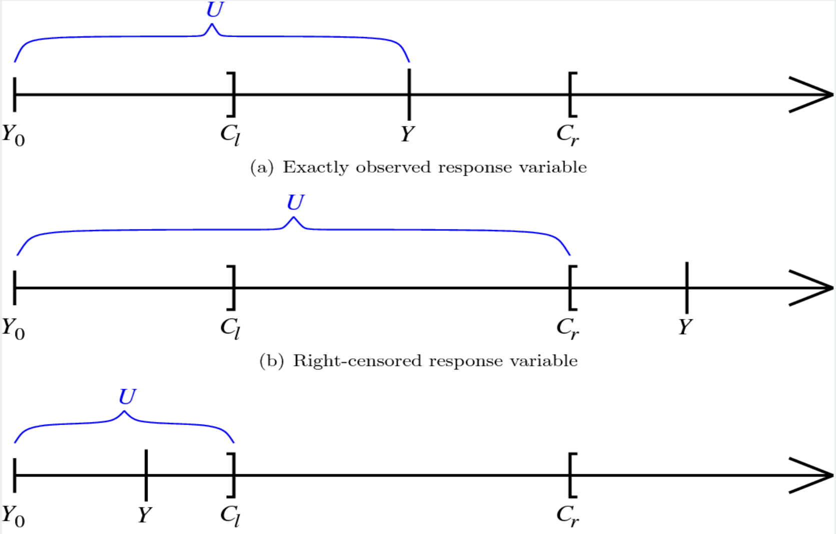
$$\int f(x)g'(x)df = f(x)g(x) - \int f'(x)g(x)dx$$

$$\int udv = uv - \int vdu$$

Trick: how to choose u and v.

3/ Survival Analysis: Basics

Basics



(a) Exactly observed response variable

(b) Right-censored response variable

(c) Left-censored response variable

Basics

- Survival function: Probability of surviving at least up to time y

$$S(y) \equiv \Pr(Y_i > y) = 1 - F(y)$$

  - $S(0) = 1$ and $S(\infty) = 0$; monotonically decreasing
  - Area under $S(y)$ is the average survival time:

$$\mathbb{E}(Y_i) = \int_0^\infty S(t)dt$$

  - Key relationships with density and probability:

$$f(y) = -\frac{d}{dy}S(y) \quad \text{and} \quad S(y) = \int_y^\infty f(t)dt$$
$$\Pr(y \leq Y_i < y + h) = S(y) - S(y + h)$$

Basics

- Hazard function: Instantaneous rate of leaving a state at time t conditional on survival up to that time

$$\lambda(y) \;\equiv\; \lim_{h \downarrow 0} \frac{\Pr(y \leq Y_i < y + h \mid Y_i \geq y)}{h} \;=\; \frac{f(y)}{S(y)}$$

- One-to-one relationship with survival function:

$$\lambda(y) \;=\; -\frac{d}{dy} \log S(y) \quad \text{and} \quad S(y) \;=\; \exp / (- \int_0^y \lambda(t) dt)^{\circ}$$