Difference in Differences

EC 320: Introduction to Econometrics

Emmett Saulnier

Spring 2022

Prologue

Housekeeping

Course evaluations are open

- If more than half of the class fills it out, I'll give 2 percentage points of extra credit
- I value your feedback! (but please be respectful)

Last problem sets

- Analytical problem set 7 due tomorrow (5/27)
- Computational problem set 7 due Wednesday (6/1)

Cholera in the 1850's

John Snow used difference-in-differences in 1855 to prove that Cholera was being spread by contaminated water.

(Not this Jon Snow)

Cholera in the 1850's

John Snow used difference-in-differences in 1855 to prove that Cholera was being spread by contaminated water.

Beliefs at the time

It was assumed at the time that Cholera was spread through things like...

- Rotting material and bad smells
- Low elevation
- Bad breeding
- Poverty

Great examples of **bad econometrics**

What might be an "ideal" experiment to test Snow's hypothesis?

Cholera in the 1850's

A change in water sources

Many companies served London's water needs. One of them, the Lambeth Company, changed the location of their water source

- **1849**: Thames River *downstream* from London 💩
- **1854**: Thames River *upstream* from London

Our "treatment" is people having their water pumped from a **clean source**. Lambeth switching their water supply is a natural experiment

1. Calculate change in Cholera rates **before** and **after** the switch for people served by Lambeth
2. Calcuate the change in Cholera rates for some control group over the same time period
3. Take the difference between (1) and (2) (*Hence the name, DiD*)

Cholera in the 1850's

Supplier	Death Rate 1849	Death Rate 1854
Non-Lambeth Only (Dirty)	134.9	146.6
Lambeth + Others (Mix clean + dirty)	130.1	84.9

Change for **Lambeth**: $y_{L,1854} - y_{L,1849} = 84.9 - 130.1 = -45.2$

Q: Is this all the evidence we need to conclude Cholera comes from water?

A: No, it could reflect anything that changed between 1849 and 1854!

Change **elsewhere**: $y_{NL,1854} - y_{NL,1849} = 146.6 - 134.9 = 11.7$

Cholera rates were actually **increasing** in places not served by Lambeth over the same time period!

Estimated treatment effect is $\Delta \bar{y}_L - \Delta \bar{y}_{NL} = -45.2 - 11.7 = -56.9$

Cholera in the 1850's

Change **elsewhere**: $y_{NL,1854} - y_{NL,1849} = 146.6 - 134.9 = 11.7$

- This is our estimate for the **counterfactual**
- "Ideally" we'd know Cholera rates in 1854 for places served by **Lambeth** if they hadn't switched to clean water
- But we only observe data from what actually happened

Thus, the change **elsewhere** is our approximation of what *would have happened* in areas served by **Lambeth** if they did not switch.

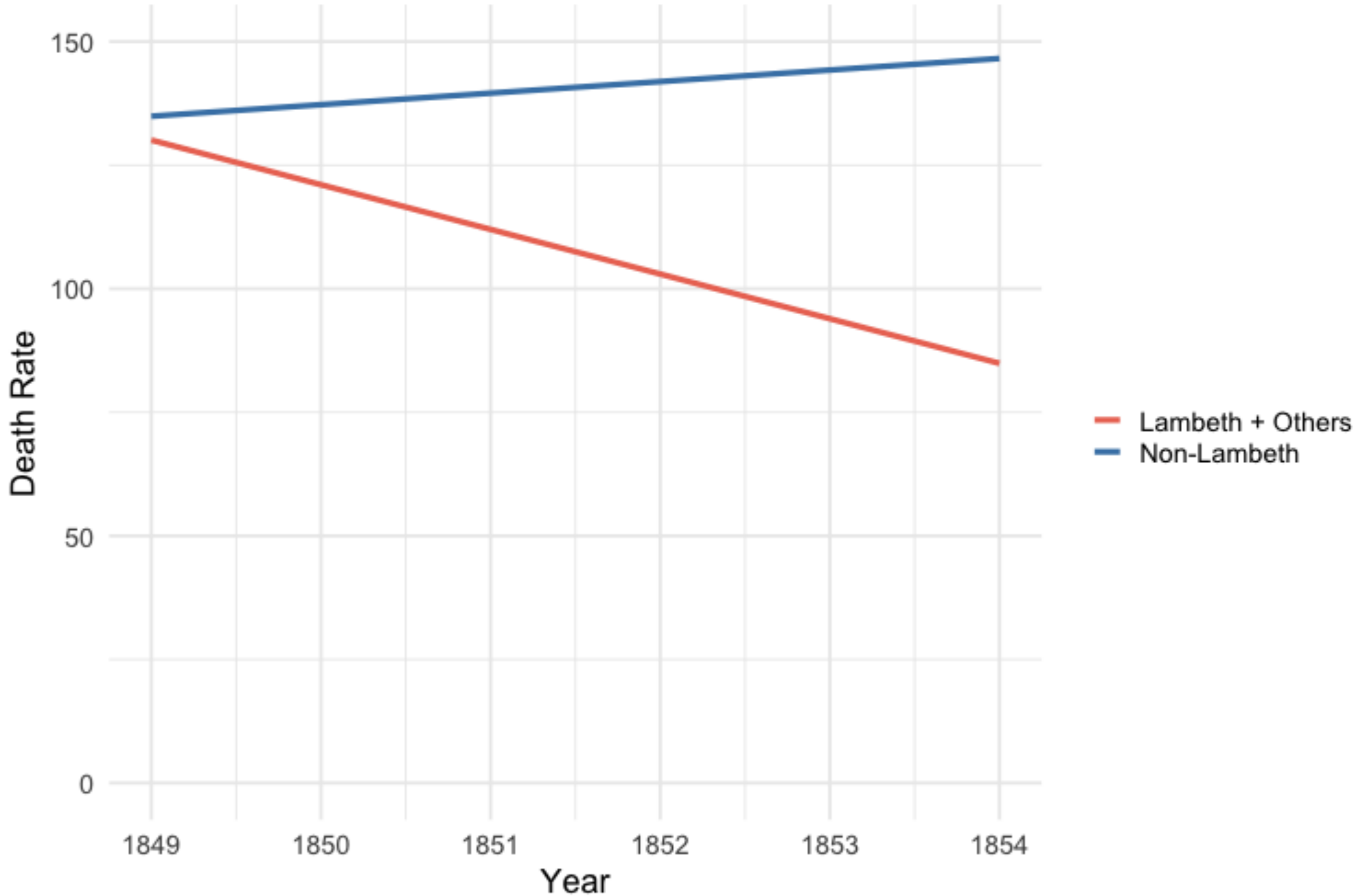
What are we assuming?

The **Parallel Trends Assumption**.

> That the difference between **Lambeth** and **non-Lambeth** areas would have remained the same if **Lambeth** hadn't switched their water source.

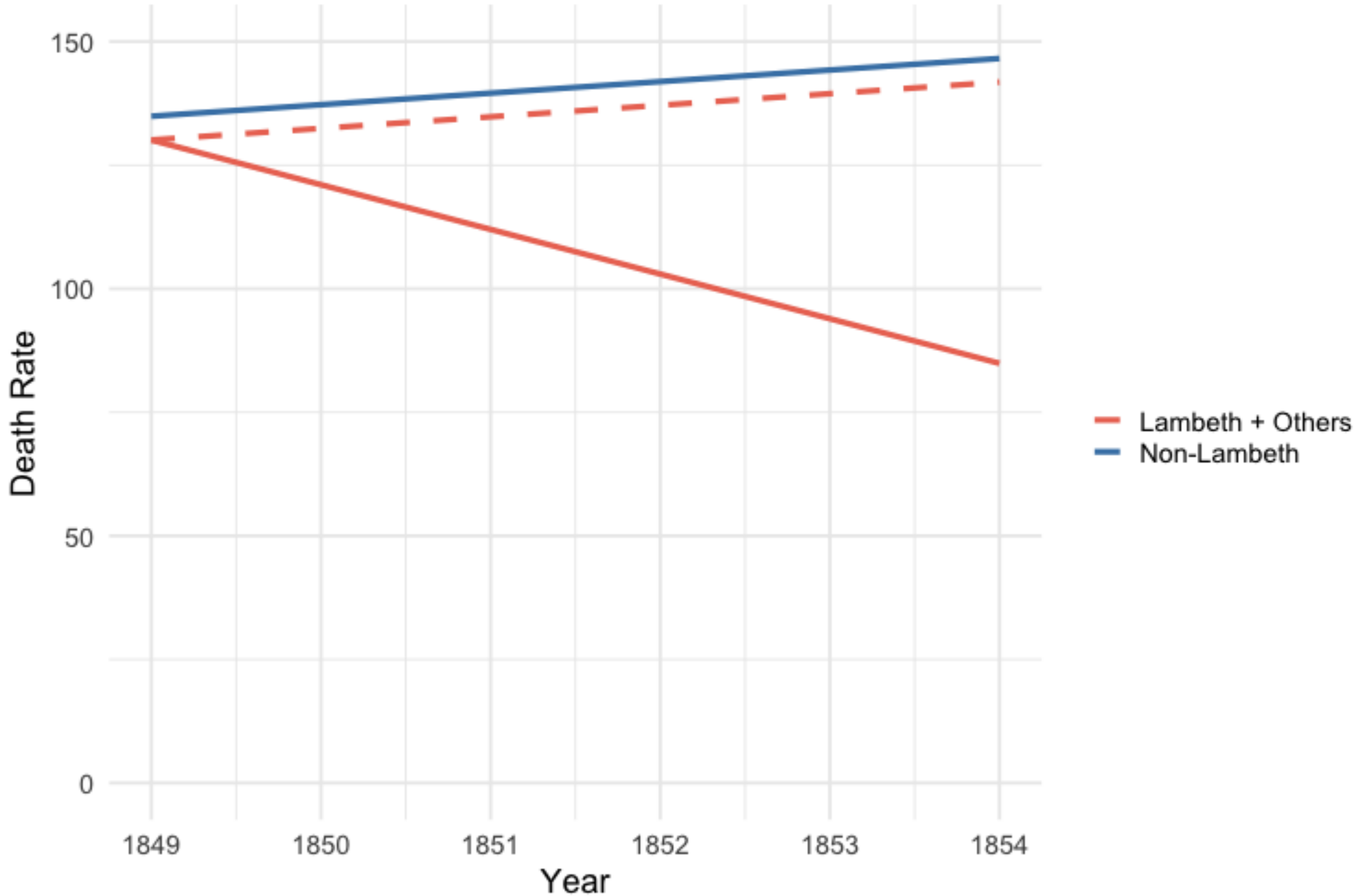
Cholera in the 1850's

Actual outcomes



Cholera in the 1850's

With our counterfactual



Difference in Differences (DiD)

Non-Random Treatment

We've talked about binary categorical variables. An important application of **dummy variables** is to study the impact of a treatment.

The estimation of treatment effects is important in a wide range of fields. **Examples:**

- impact of cash transfers on child health
- effect of class size on student achievement

In clinical trials of health interventions, a common question is whether use of a medicine will improve health outcomes.

Non-Random Treatment

Ideally treatment effects are evaluated using **randomized controlled trials**.

We capture treatment with a dummy variable and use a simple statistical model

$$Y_i = \beta_0 + \beta_1 D_i + u_i$$

for $i = 1, \ldots, n$, where $D_i = 1$ if treated and $D_i = 0$ if not treated, and where y_i is the outcome for individual i.

Q: What are the interpretations of the coefficients β_0 and β_1?

Random assignment: If D_i and u_i are **uncorrelated (A3.)**, then OLS estimates will be unbiased.

Our effect is the difference in mean outcomes for treated and untreated groups, $\hat{\beta}_1 = \bar{y}_1 - \bar{y}_0$, called the **differences estimator**

Non-Random Treatment

Often we are analyzing some *treatment* that takes place at a particular time, and we have data **before and after** the treatment took place.

If the policy intervention affected everyone and we have data both in pre- and post-policy periods, might consider using the **differences estimator**

> **Problem:** Assumes that there are not other reasons for the change in mean outcomes before and after the treatment

In this case the dummy variable D_i in effect stands for the impact of **all factors**, including the treatment, that are different between the pre- and post- periods.

Non-Random Treatment

Since **other factors** may be partially responsible for the change in mean outcomes, the difference estimator will likely be biased (example of exogeneity failing)

This point emphasizes importance of a control group that does not receive the treatment. Sometimes policies have actual randomization, **(randomized field experiments)**.

> **Example:** Oregon Health Plan (OHP) experiment of 2008. Oregon decided to expand coverage of its version of Medicaid to groups not previously eligible, but this expanded coverage was rationed by a lottery to reduce the cost of the program: somewhat under half of those who registered were randomly selected and invited to apply for expanded coverage.

Since treatment was randomly assigned, $E[u_i|D_i] = 0$

DiD Concept

Most of the time, we are not so lucky to have real randomization. Instead we often have to use **natural experiments**.

> **Setting:** There is a discrete change in policy that affects only part of the population, and that there are both pre- and post-treatment observations for both those who received treatment and those who did not.

Since the policy change is not distributed randomly, mean outcomes for the treated and control group may differ for a variety of reasons.

The **difference-in-difference (DiD)** method allows for differences between the treatment and control groups by measuring the treatment effect based on the **change** in outcomes before and after treatment for the treatment group **relative** to the control group.

DiD Structure

Suppose we have data on specific individuals $i = 1, \dots, n$ in each of two time periods $t = 1, 2$. A simple statistical formulation of the model is then:

$$Y_{it} = \beta_0 + \beta_1 T_t + \beta_2 D_i + \beta_3 (T_t \times D_i) + u_{it}$$

for $i = 1, \dots, n$, and $t = 1, 2$. (Note that we have $2n$ data points).

Here $T_t = 1$ if $t = 2$ and $T_t = 0$ if $t = 1$, while $D_i = 1$ if individual i receives treatment and $D_i = 0$ if individual i does not receive treatment.

DiD Structure

In the **pre-treatment period**, $t = 1 \implies T_t = 0$, we have

$$E(y_{i1}|D_i = 0) = \beta_0 \qquad \textbf{and} \qquad E(y_{i1}|D_i = 1) = \beta_0 + \beta_2$$

We are allowing for the possibility that there was imperfect control, in the sense that treated and untreated may have different pre-treatment means.

In the **post-treatment period**, $t = 2 \implies T_t = 1$, we have

$$E(y_{i2}|D_i = 0) = \beta_0 + \beta_1 \quad \textbf{and} \quad E(y_{i2}|D_i = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$$

The **treatment effect** corresponds to β_3: This is because

$$\begin{aligned} \text{Diff Treated: } & E(y_{i2}|D_i = 1) - E(y_{i1}|D_i = 1) = \beta_1 + \beta_3 \ \textbf{ and} \\ \text{Diff Control: } & E(y_{i2}|D_i = 0) - E(y_{i1}|D_i = 0) = \beta_1 \end{aligned}$$

DiD Example

"Monetary Intervention in the Great Depression" by **Gary Richardson and William Troost (JPE, 2009)**.

Consider the 1930s bank failures. Central Bank can prevent bank runs and bank failures by acting as lender of last resort to insolvent banks.

The twelve different regional Federal Reserve Districts reacted in the 1930s to potential bank failures in different ways.

- **Atlanta FRB (sixth district)** and **St Louis FRB (eighth district)** followed radically different policies on bank runs
- Makes for a natural experiment, comparing outcomes for bank failures in Mississippi banks in the two districts

DiD Example

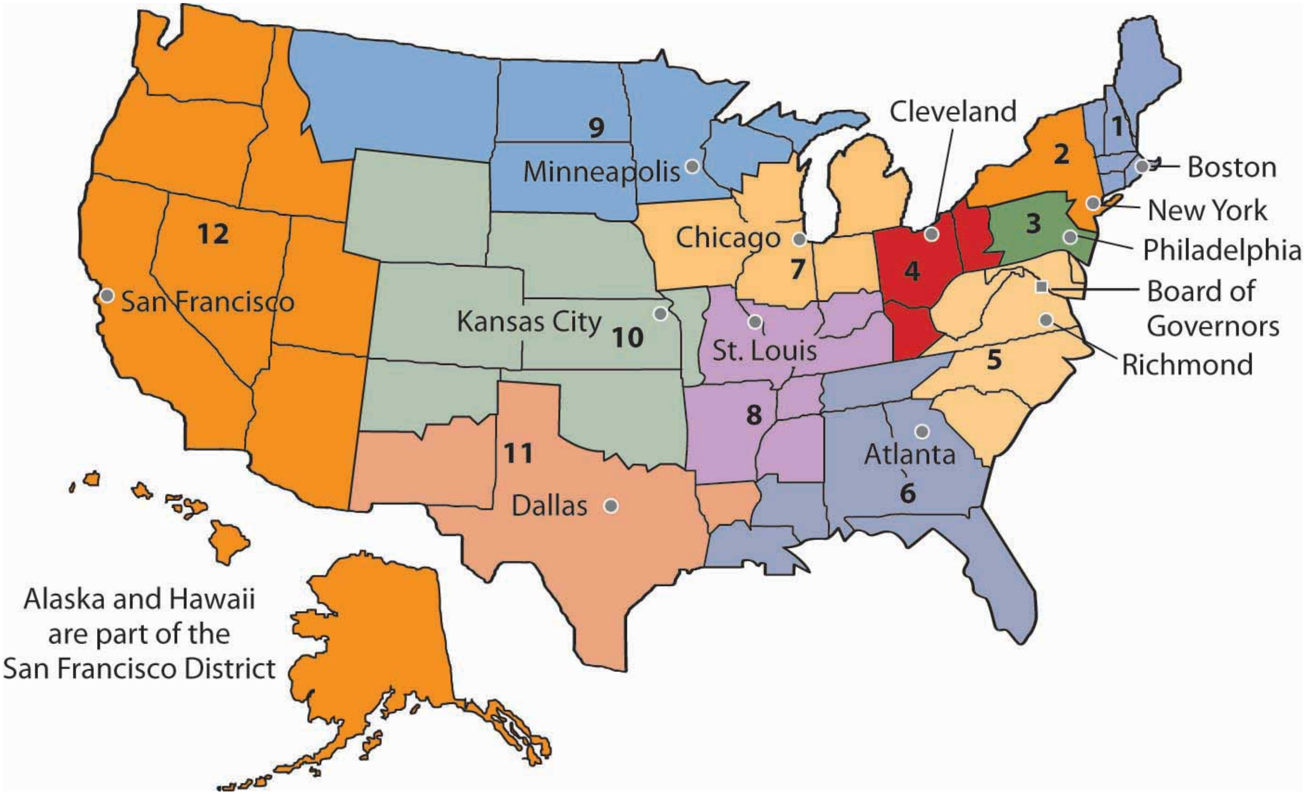
Banking simplified...

- Banks take deposits from savers and then lend that money out to others
- They do not hold enough cash on hand to give everyone their savings back if everyone tried to withdraw at once
- Bank runs happen when people are afraid the bank is going to fail
  - Worry about bank failure => Some people widthdraw their money
  - More withdraws => more likely for the bank to fail
  - More likely for bank to fail => More people withdraw

Differential policies by fed district

The **Sixth District** made credit to banks widely available to ensure they had cash to cover withdraws, while **Eighth District** restricted lending to banks with the idea that it would "weed out" bad banks

DiD Example



Mississippi is cut in half, the **northern part is St. Louis district** and the **southern part is Atlanta district**.

We can use this as a natural experiment!

DiD Example

R&T analyze this using **Difference-in-differences**.

Bank Losses by Policy

Variable	8th District	6th District	Diff., 6th-8th
No. of Banks open 1930	165	135	-30
No. of Banks open 1931	132	121	-11
Changes in banks open	-33	-14	19

Estimated to have saved **19** banks, 14% of **Sixth District** in 1930.

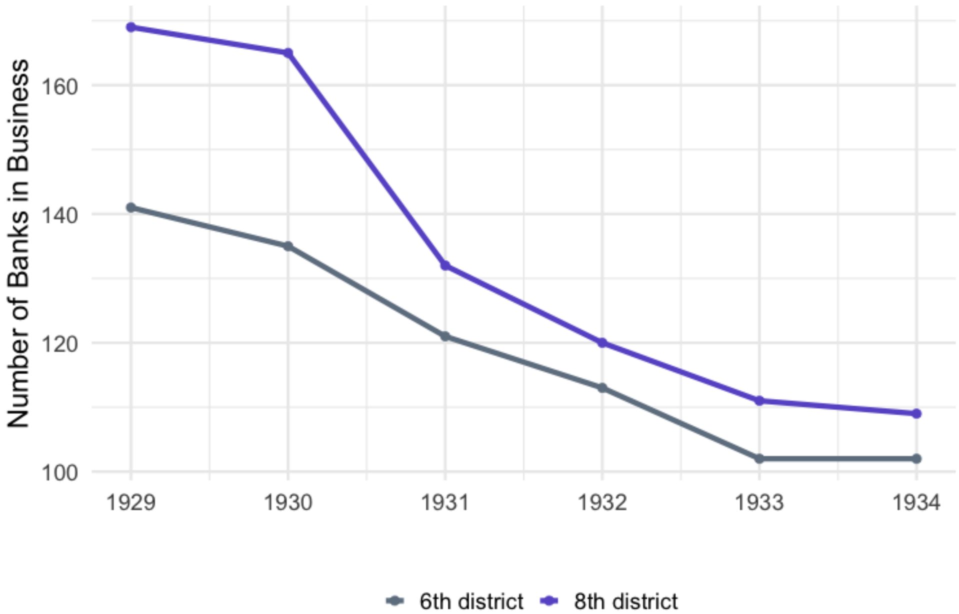
Note: It is equivalent to think of the DiD estimator as

- The difference in the change in the treated and the change in the untreated
- The change in the difference between treated and untreated pre/post treatment

DiD Required Assumption

A central issue is the **validity** of the control group, which is determined by the **parallel trends** assumption:

> In the absence of the *"treatment"*, would the number of open banks in the two districts have been expected to evolve in the same way?



DiD Interpretation

Now with a DiD Regression

$$Y_{dt} = \beta_0 + \beta_1 T_t + \beta_2 D_d + \beta_3 (T_t \times D_d) + u_{dt}$$

One can use this data to get an estimate of the treatment effect $\hat{\beta}_3$ and a standard error $SE(\hat{\beta}_3)$ based on these twelve data points:

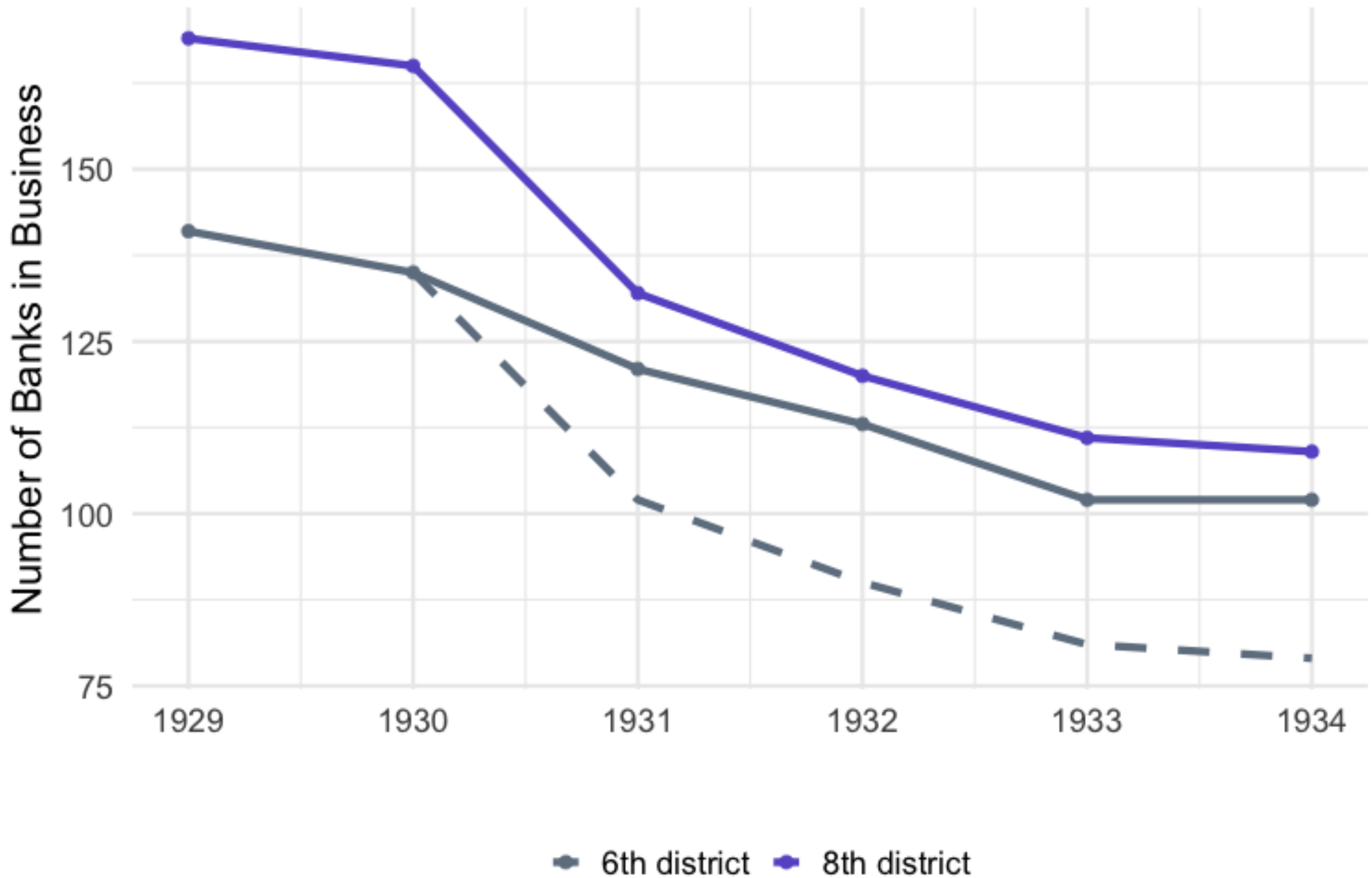
$$Y_{dt} = 167 - 49T_t - 29D_d + 20.5(T_t \times D_d) + u_{dt}$$

Here T_t is coded 1 for 1931 or later, and $D_d = 1$ for the 6th District.

$\hat{\beta}_3$ has $t = 1.9$ which is significant at 10% level using a two-tailed test.

This is about the best we can do with twelve data points.

DiD Interpretation



The dashed line represents our assumed **counterfactual**.

DiD Interpretation

Let's do the same **DiD** with other outcome variables!

	1929	1933	Difference (1933-1929)
Panel A. Number of Wholesale Firms			
Sixth Federal Reserve District (Atlanta)	783	641	-142
Eighth Federal Reserve District (St. Louis)	930	607	-33
Difference (Sixth-Eight)	-147	34	**181**
Panel B. Net Wholesale Sales ($ million)			
Sixth Federal Reserve District (Atlanta)	141	60	-81
Eighth Federal Reserve District (St. Louis)	245	83	-162
Difference (Sixth-Eight)	-104	-23	**81**

Notes: This table presents a DiD analysis of Federal Reserve liquidity effects on the number of wholesale firms and the dollar value of their sales, paralleling the DiD analysis of lquidity effects on bank activity.

Seems like more liberal credit availability saved a lot of wholesale firms.