Universiteit Utrecht

Leveraging register data to estimate causal effects of policy interventions

Workshop ODISSEI

Oisín Ryan & Erik-Jan van Kesteren

About us

Erik-Jan van Kesteren

- Background in statistics / social science
- Assistant professor @ methodology & statistics UU
- Social Data Science team lead @ ODISSEI (consortium of universities)

SoDa

Some stuff I work on:

Latent variables, high-dimensional data, optimization, regularization, visualisation, Bayesian statistics, multilevel models, spatial data, generalized linear models, privacy, synthetic data, high-performance computing, software development, open science & reproducibility

About us

Oisín Ryan

- Background in statistics / social science
- Currently: Postdoc @ methodology & statistics UU
- From July: Assistant Professor @ Data Science and Biostatistics, Julius Center, UMC Utrecht
- Co-ordinator Special Interest Group in Causal Data Science UU/UMCU

Website: oisinryan.org

Some stuff I work on:

Causal inference, causal discovery, time-series analysis, computational modeling and complex systems, Bayesian statistics, multilevel models, open science & reproducibility, R programming

Today’s Goal

A brief survey and practical introduction to the

- Core concepts
- Key assumptions
- Different statistical methods

used to evaluate the **causal effects** of **policy interventions**

Disclaimer:

We take a “wide” instead of “deep” view

Many details / extensions / advanced topics omitted!

causalpolicy.nl

Today’s plan: morning

- Introduction + Practical (105 minutes)
  - Policy Interventions and Causal Inference
  - Pre-Post Analyses and Difference-in-Difference
- Break (15 minutes)

- Interrupted Time Series (45 minutes)
- Practical (30 minutes)

- Lunch around 12:00 ; re-start at 13:00

Today’s plan: afternoon

- Synthetic Control Methods (45 minutes)
- Practical (45 minutes)
- Break (15 minutes)

- Controlled ITS and CausalImpact (45 minutes)
- Practical (45 minutes)
- Break (15 minutes)

- Discussion session (30 minutes)

- Finish around 17:00

Context: “Policy Evaluations”

Many social science **research questions** concern evaluating what **the effect** of implementing a particular **policy** or **intervention** was on some outcome of interest

Examples:

- What was the effect of raising the maximum speed limit on road deaths?
- What effect did introducing students loans have on post-graduation debt levels?
- Did introducing an after-school programme in disadvantaged neighbourhoods lead to improved educational outcomes in children from that neighbourhood?

Context: “Policy Evaluations”

Sometimes referred to as “policy evaluation” research or “comparative case studies”

Basic Structure:

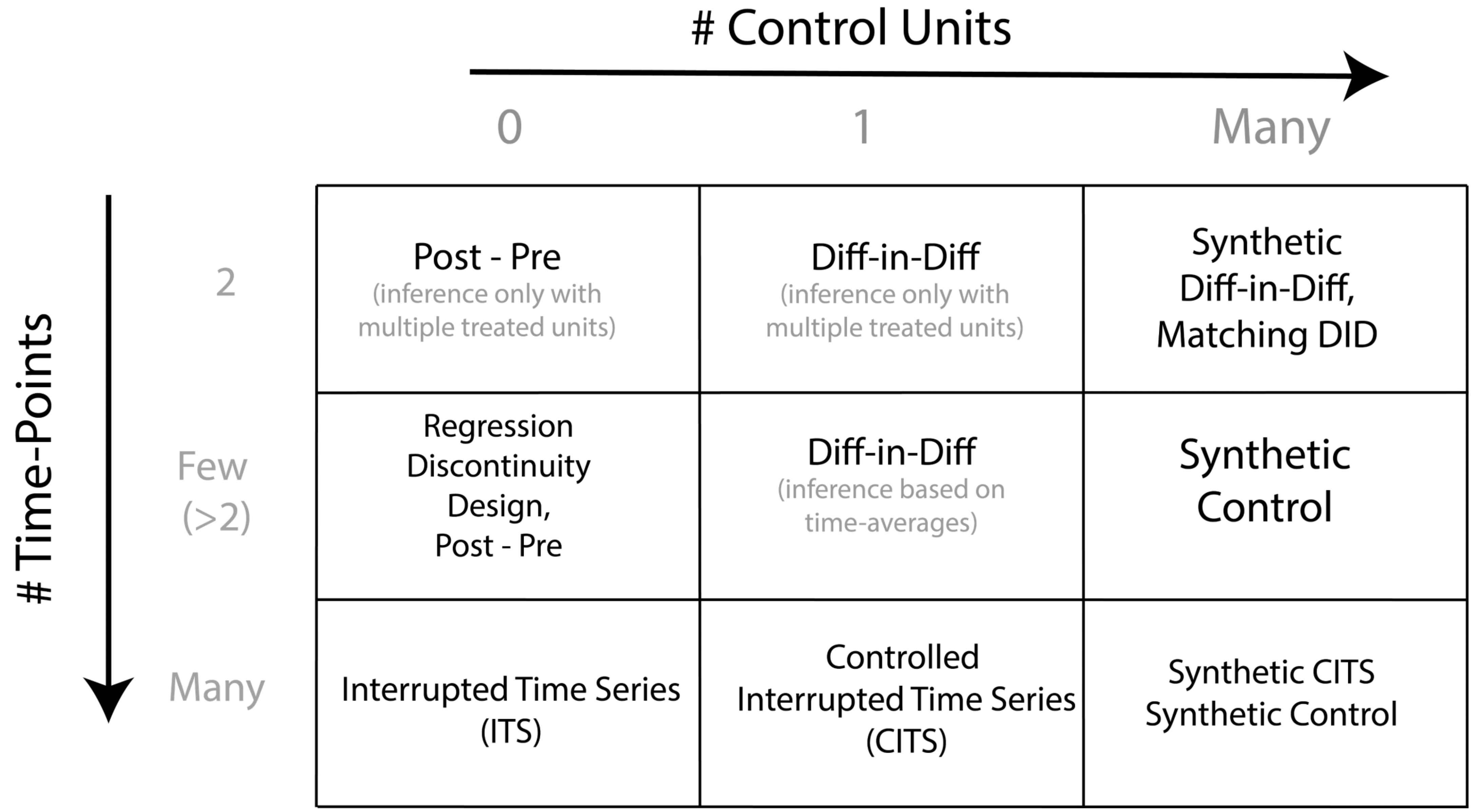
- We have some **unit** (or units) which we observe **before** and **after** some intervention or action
- Did the intervention produce a change in the outcome for that unit?

Methods for Policy Evaluation

Many different methods have been developed to answer these types of research questions

These methods differ in terms of:

- The **amount** and **type** of information they use
    - Amount of time-points and amount of potential “control” units
- The specific **statistical approach** they take
- The types of **assumptions** they make



		# Control Units		
		0	1	Many
# Time-Points	2	Post - Pre (inference only with multiple treated units)	Diff-in-Diff (inference only with multiple treated units)	Synthetic Diff-in-Diff, Matching DID
	Few (>2)	Regression Discontinuity Design, Post - Pre	Diff-in-Diff (inference based on time-averages)	Synthetic Control
	Many	Interrupted Time Series (ITS)	Controlled Interrupted Time Series (CITS)	Synthetic CITS Synthetic Control

Causal Inference: A primer

Potential Outcomes

Causal Inference is (broadly) concerned with using **data** to estimate what the effect is of **intervening or changing** the value of one or more **variables**.

Using the **potential outcomes** framework, we can define causal inference as a *missing data problem*

Potential Outcomes

Let Y_i represent your headache level (high is a very bad headache, low is no headache), and let A_i be whether you take aspirin or not (A =1 you take it, A = 0 you don’t)

You only want to take an aspirin if your headache level **after taking aspirin** is lower relative to what your headache would be **if you wouldn’t take aspirin**

There are **two possible versions** of the outcome variable

- Y_i^1 your headache level **if you would take aspirin**
- Y_i^0 your headache level **if you would not take aspirin**

Causal Effects

We can define the **causal effect** of taking aspirin on your headache levels as the difference in potential outcomes

$$CE_i = Y_i^1 \; - Y_i^0$$

The **fundamental problem of causal inference:** You only ever observe one of the potential outcomes!

Data and Potential Outcomes

ID	Y
1	7
2	9
3	6
4	5
5	6
6	2
7	3
8	1
…	…
I	2

Data and Potential Outcomes

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
I	2	1	NA	2

Data and Potential Outcomes

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
I	2	1	NA	2

Causal Inference

In cross-sectional settings, we typically aim to make inferences about the **average causal effect.** This is known as a **causal estimand:**

$$ACE = E[Y^1] - E[Y^0]$$

In a **Randomized Controlled Trial,** we often use the difference in treated and untreated groups as an **estimator** of this causal effect:

$$\widehat{ACE} = E[Y|\, A = 1] - E[Y\,|A = 0]$$

Causal Inference

In cross-sectional settings, we typically aim to make inferences about the **average causal effect.** This is known as a **causal estimand:**

$$ACE = E[Y^1] - E[Y^0]$$

In a **Randomized Controlled Trial,** we often use the difference in treated and untreated groups as an **estimator** of this causal effect:

$$\widehat{ACE} = E[Y|\,A = 1] - E[Y\,|A = 0]$$

Causal Inference

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
I	2	1	NA	2

Causal Inference

In cross-sectional settings, we typically aim to make inferences about the **average causal effect.** This is known as a **causal estimand:**

$$ACE = E[Y^1] - E[Y^0]$$

In a **Randomized Controlled Trial,** we often use the (sample) difference in treated and untreated groups as an **estimator** of this causal effect:

$$\widehat{ACE} = E[Y|\,A = 1] - E[Y\,|A = 0]$$

Causal Inference

ID	Y	A	Y^0	Y^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\cdots$	…
I	2	1	NA	2

Causal Inference Assumptions

This type of **inference** about causal effects from **observed data** is only possible under certain **conditions** or **assumptions**

Exchangeability

- If we were to reverse treatment assignment we would observe the same group differences. Information is exchangeable between groups
- Basically: absence of **confounder variables**
  - E.g. People who have bad headaches choose to take the aspirin
- **RCTs** are powerful because **randomization** ensures exchangeability. But in principle this kind of inference is possible from non-RCT designs
- In practice we need **conditional exchangeability**; to control for **confounders!**

Causal Inference Assumptions

This type of **inference** about causal effects from **observed data** is only possible under certain **conditions** or **assumptions**

Stable Unit Treatment Value (also known as SUTVA)

- **No Interference**: The potential outcomes of one unit does not depend on the treatment assigned to another unit.
    - No “spillover”: My taking an aspirin does not influence your headache levels
- **Consistency:** Only one version of treatment, treatment is unambiguously defined.
- I can directly observe one of the potential outcomes. $Y_i = Y_i^1$

Causal Inference Assumptions

These two generic assumptions essentially always appear in causal inference problems, and as we will see, we will have to deal with concerns around **confounders** and **no interference** repeatedly today

Other assumptions or conditions may also be needed depending on the specific **design** and **analytic approach you take**

Causal Inference and Policy Evaluations

Todays Topic

Policy evaluation is a special case of causal inference

We typically have **one unit** observed **repeatedly over time**

At some point in time (T_0) an **intervention** takes place

Pre-intervention we observe Y_t^0 and **post-intervention** Y_t^1

$Time$	Y_t	A_t
1	7	0
2	9	0
3	6	0
4	5	0
5	6	0
6	2	1
7	3	1
8	1	1
…	…	…
T	2	1

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
T	2	1	NA	2

Causal Effects of Policies

We want to estimate the **causal effect of the policy intervention**

We think about this as the difference between

(a) the **observed outcome** after the policy was introduced

(b) What the outcome **would have been** without the intervention

$$CE_t = Y_t^1 - Y_t^0$$

where $t > T_0$ (i.e., the post-intervention time period)

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
I	2	1	NA	2

Running Example: Proposition 99

Proposition 99

- A famous example in causal inference literature

  Abadie, A., Diamond, A., & Hainmueller, J. (2010). Synthetic control methods for comparative case studies: ***Estimating the effect of California's tobacco control program****. Journal of the American statistical Association, 105(490), 493-505.*

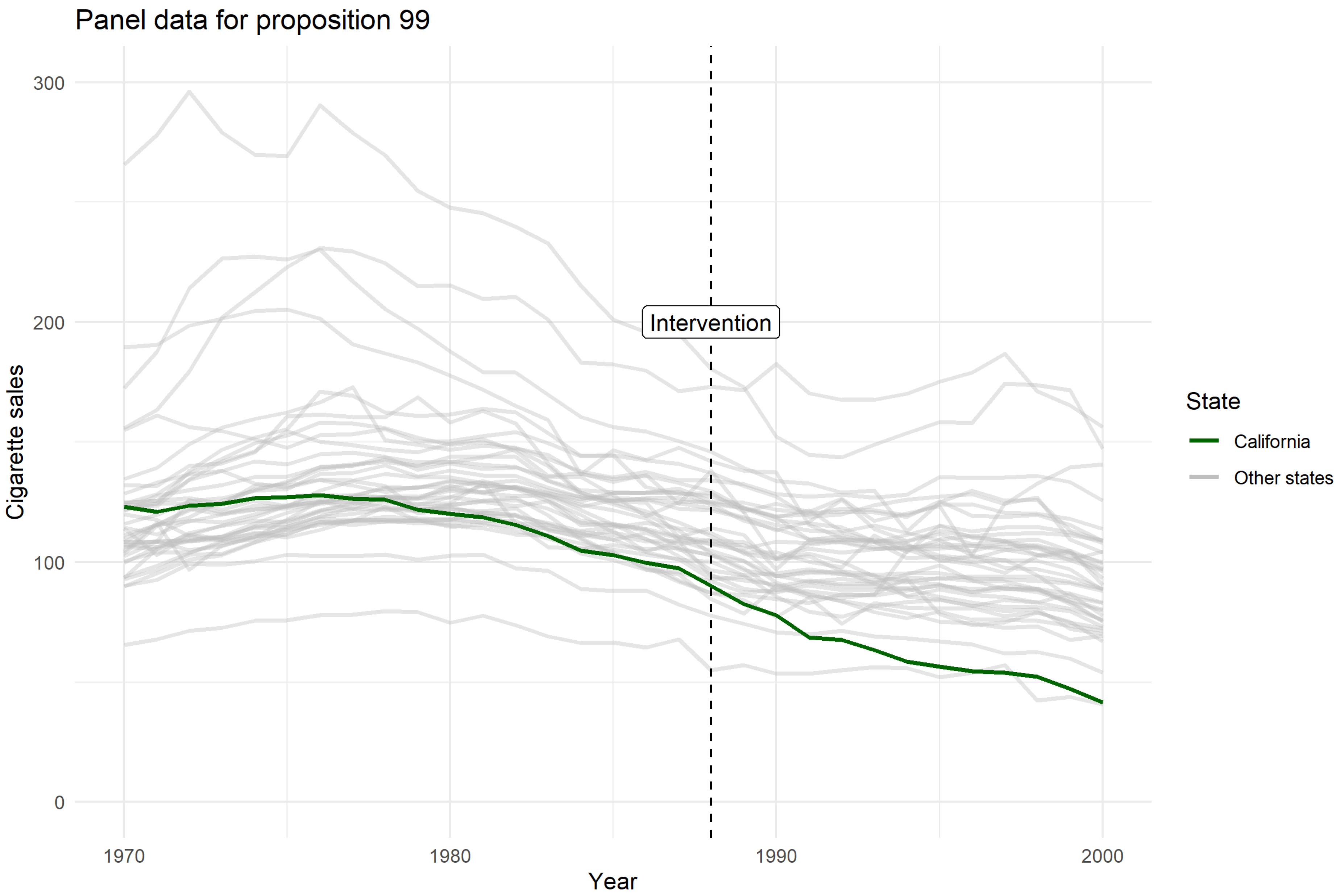
- In 1988, the state of California imposed a 25% tax on tobacco cigarettes
- Total savings in personal health care expenditure until 2004 is $86 billion (Lightwood et al., 2008)

Proposition 99

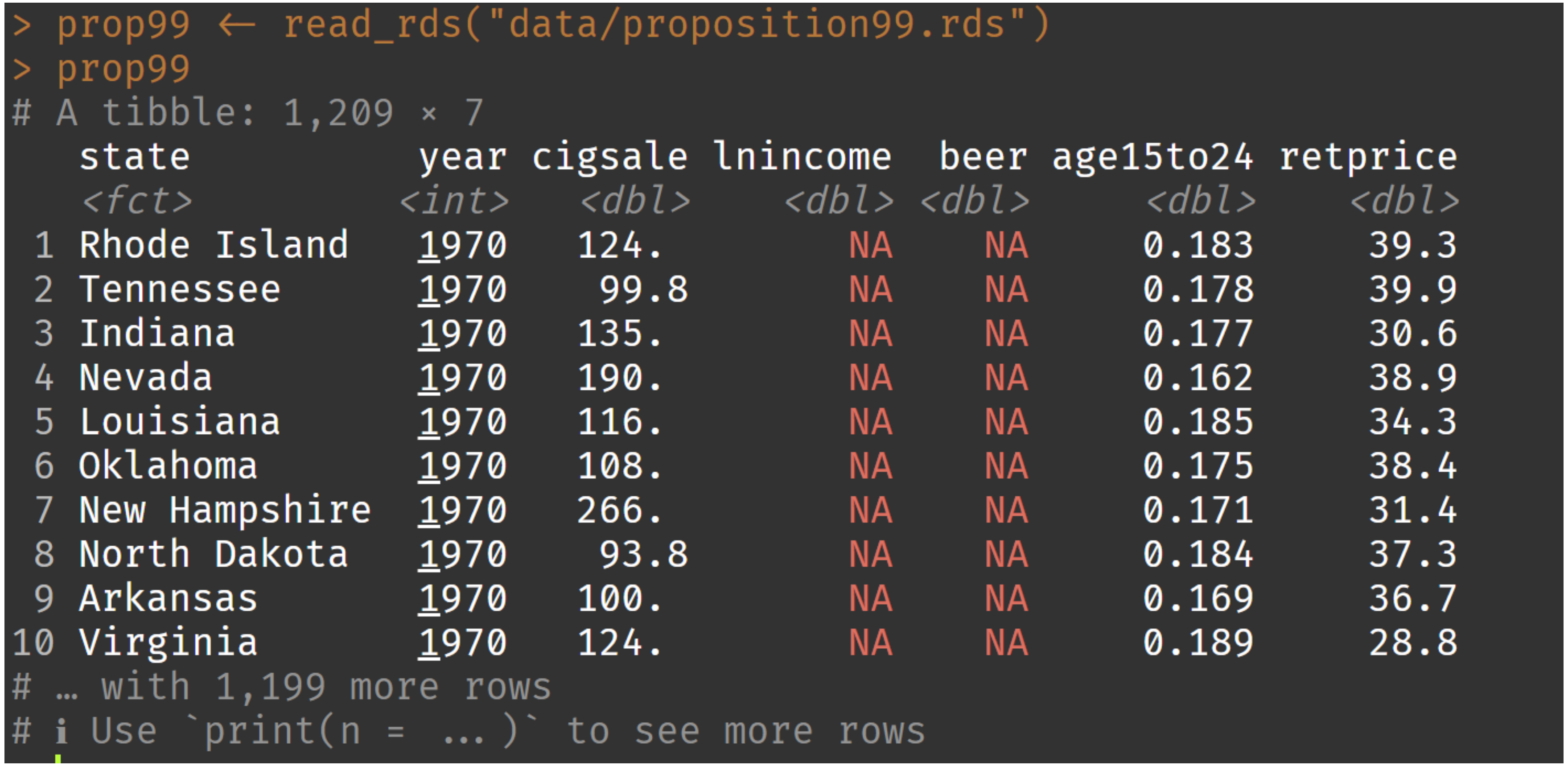
- We prepared a dataset for this workshop:

  `proposition99.rds`

- Panel (i.e. longitudinal) dataset
- Can be downloaded from the website
- Let’s explore!


Panel data for proposition 99
Intervention
Cigarette sales
300
200
100
0
1970
1980
1990
2000
Year
State
California
Other states

Proposition 99



```
> prop99 ← read_rds("data/proposition99.rds")
> prop99
# A tibble: 1,209 × 7
   state          year cigsale lnincome  beer age15to24 retprice
   <fct>         <int>   <dbl>    <dbl> <dbl>     <dbl>    <dbl>
 1 Rhode Island   1970   124.        NA    NA     0.183     39.3
 2 Tennessee      1970    99.8       NA    NA     0.178     39.9
 3 Indiana        1970   135.        NA    NA     0.177     30.6
 4 Nevada         1970   190.        NA    NA     0.162     38.9
 5 Louisiana      1970   116.        NA    NA     0.185     34.3
 6 Oklahoma       1970   108.        NA    NA     0.175     38.4
 7 New Hampshire  1970   266.        NA    NA     0.171     31.4
 8 North Dakota   1970    93.8       NA    NA     0.184     37.3
 9 Arkansas       1970   100.        NA    NA     0.169     36.7
10 Virginia       1970   124.        NA    NA     0.189     28.8
# … with 1,199 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Proposition 99

state: 39 different states, used in Abadie et al. (2010)

year: 1970 until 2000

cigsale: packs of cigarettes per 100 000 people

lnincome: natural log of mean income

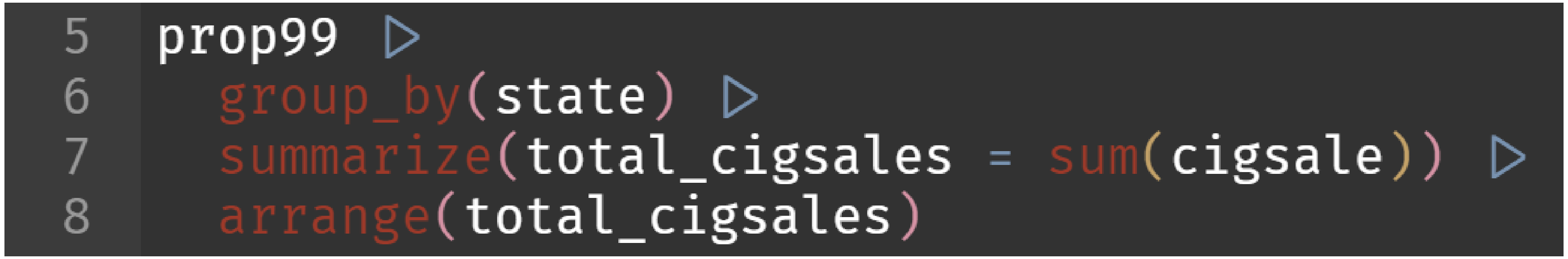
beer: beer sales per 100 000 people

age15to24: proportion of people between 15 & 24

retprice: retail price of a box of cigarettes

Proposition 99

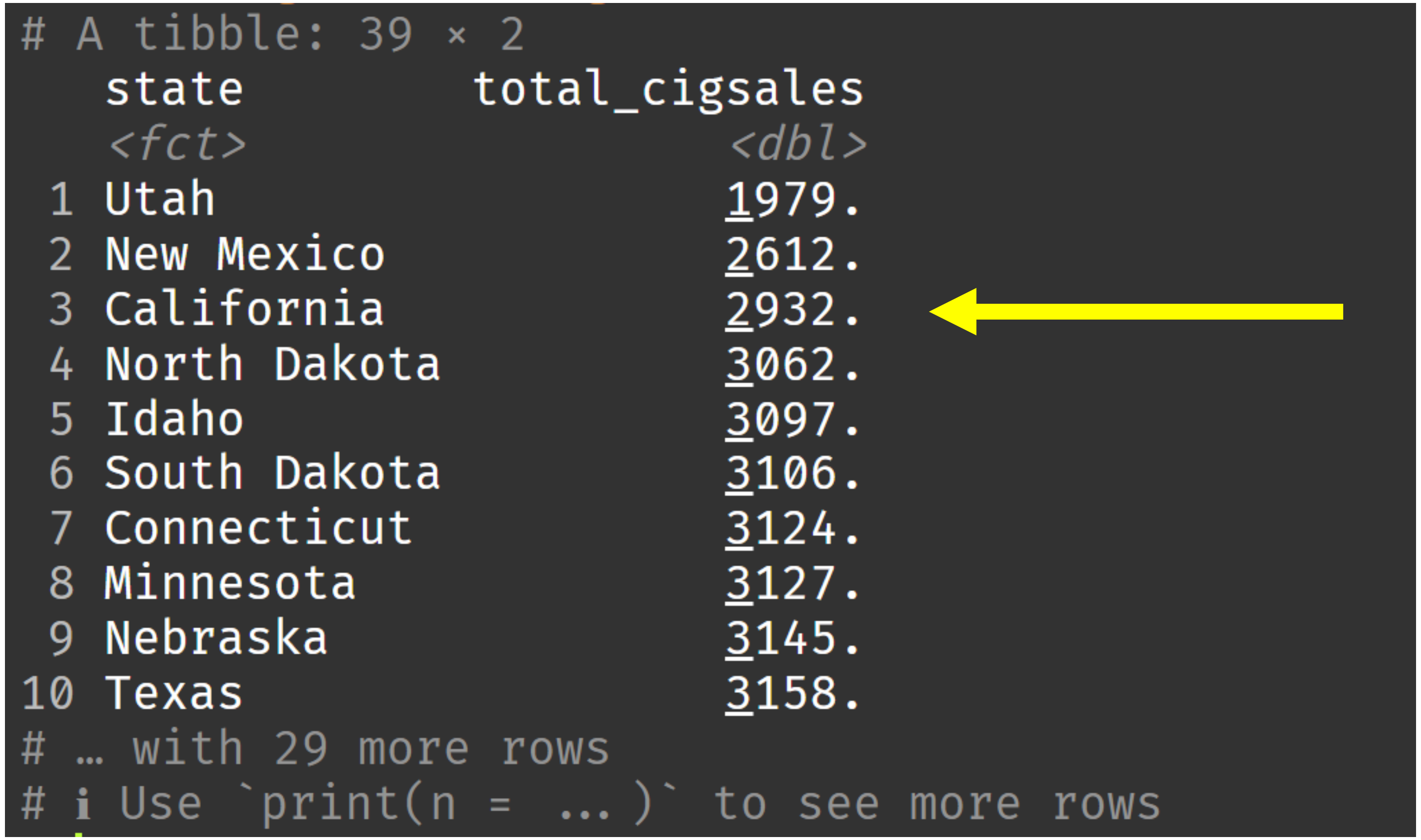
- Which state sold the least cigarettes per capita?
- We make use of **tidyverse:**



```r
prop99 |>
  group_by(state) |>
  summarize(total_cigsales = sum(cigsale)) |>
  arrange(total_cigsales)
```

- This works well with our prepared dataset

Proposition 99



```
# A tibble: 39 × 2
   state          total_cigsales
   <fct>                   <dbl>
 1 Utah                    1979.
 2 New Mexico              2612.
 3 California              2932.
 4 North Dakota            3062.
 5 Idaho                   3097.
 6 South Dakota            3106.
 7 Connecticut             3124.
 8 Minnesota               3127.
 9 Nebraska                3145.
10 Texas                   3158.
# … with 29 more rows
# ℹ Use `print(n = ...)` to see more rows
```

Practical: set-up and data

Work in pairs/groups!

Exercises 1 – 3

causalpolicy.nl

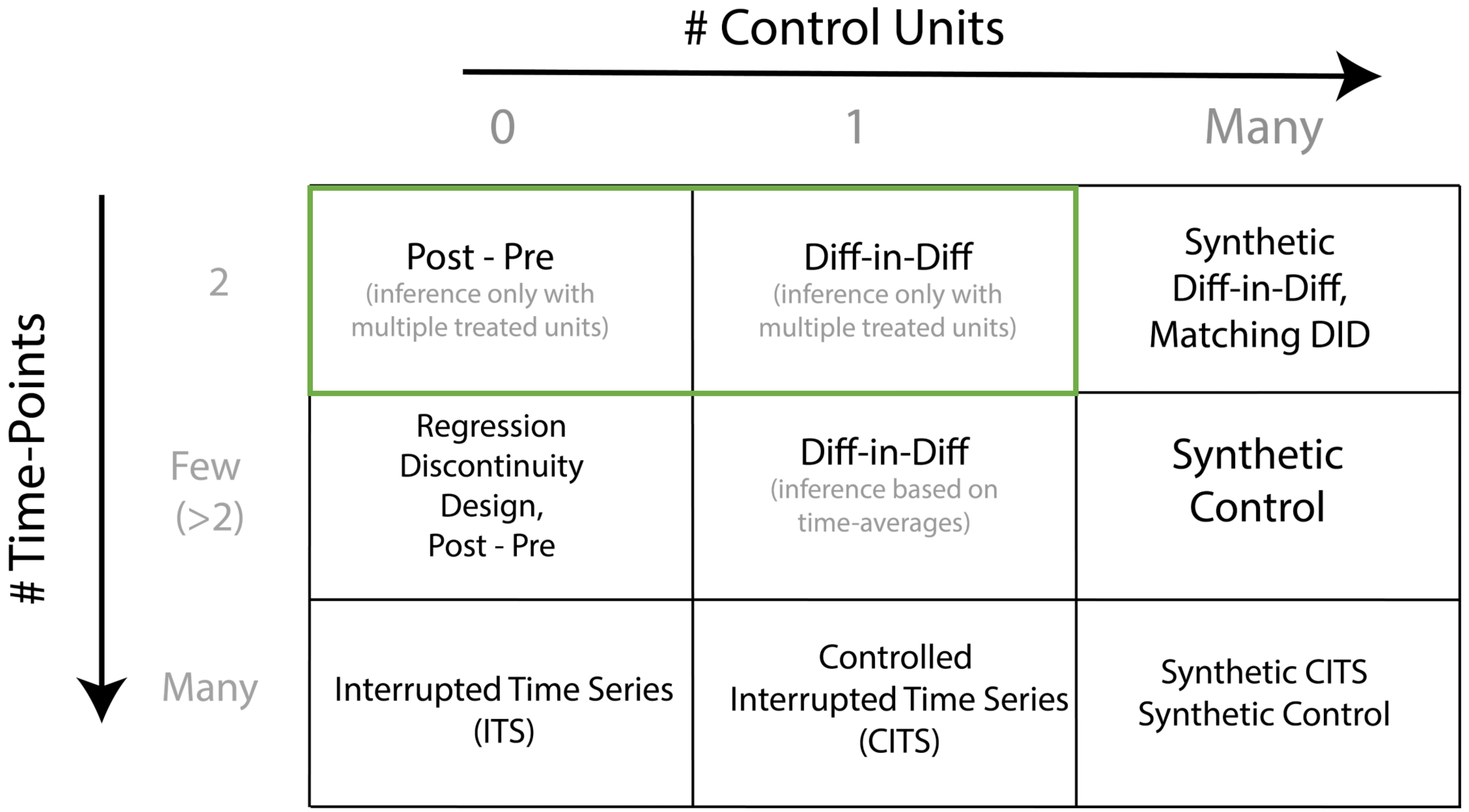
Estimating the causal effect

Basic methods

Control Units

Time-Points

	0	1	Many
2	Post - Pre (inference only with multiple treated units)	Diff-in-Diff (inference only with multiple treated units)	Synthetic Diff-in-Diff, Matching DID
Few (>2)	Regression Discontinuity Design, Post - Pre	Diff-in-Diff (inference based on time-averages)	Synthetic Control
Many	Interrupted Time Series (ITS)	Controlled Interrupted Time Series (CITS)	Synthetic CITS Synthetic Control

# Control Units

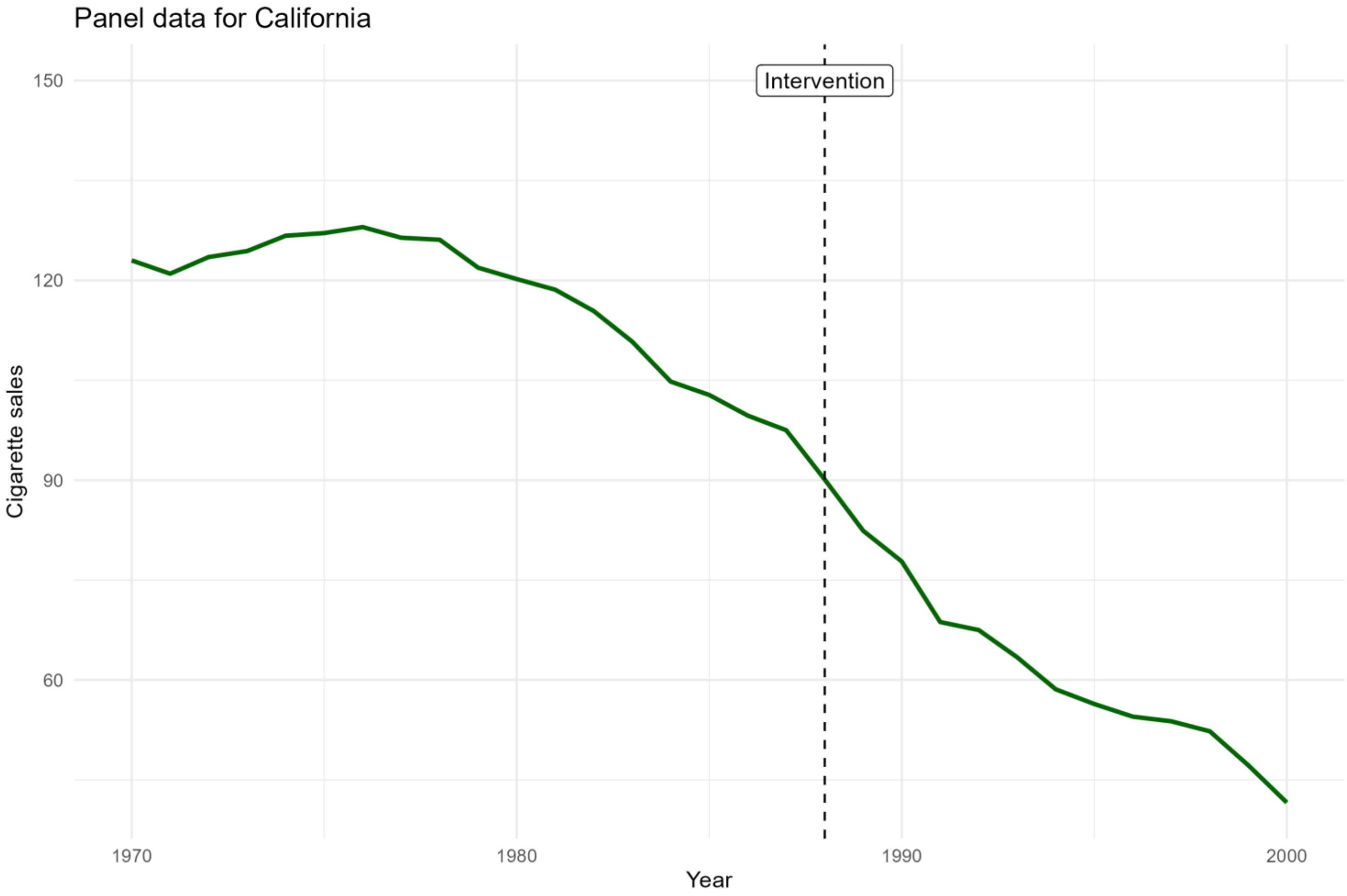
Time-Points

	0	1	Many
2	Post - Pre (inference only with multiple treated units)	Diff-in-Diff (inference only with multiple treated units)	Synthetic Diff-in-Diff, Matching DID
Few (>2)	Regression Discontinuity Design, Post - Pre	Diff-in-Diff (inference based on time-averages)	Synthetic Control
Many	Interrupted Time Series (ITS)	Controlled Interrupted Time Series (CITS)	Synthetic CITS Synthetic Control

Pre-Post Estimator

Pre-post estimator

We use only the cigarette sales time series for California



Pre-post estimator

- We want to estimate the following quantity:

$$\overline{CE}_{post} = \bar{Y}^1_{post} - \bar{Y}^0_{post}$$

- But we cannot observe $\bar{Y}^0_{post}$!
- Solution: replace $\bar{Y}^0_{\boldsymbol{post}}$ by $\bar{Y}^0_{\boldsymbol{pre}}$, which is observable

$$\overline{CE}_{post} = \bar{Y}^1_{post} - \bar{Y}^0_{pre}$$

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
T	2	1	NA	2

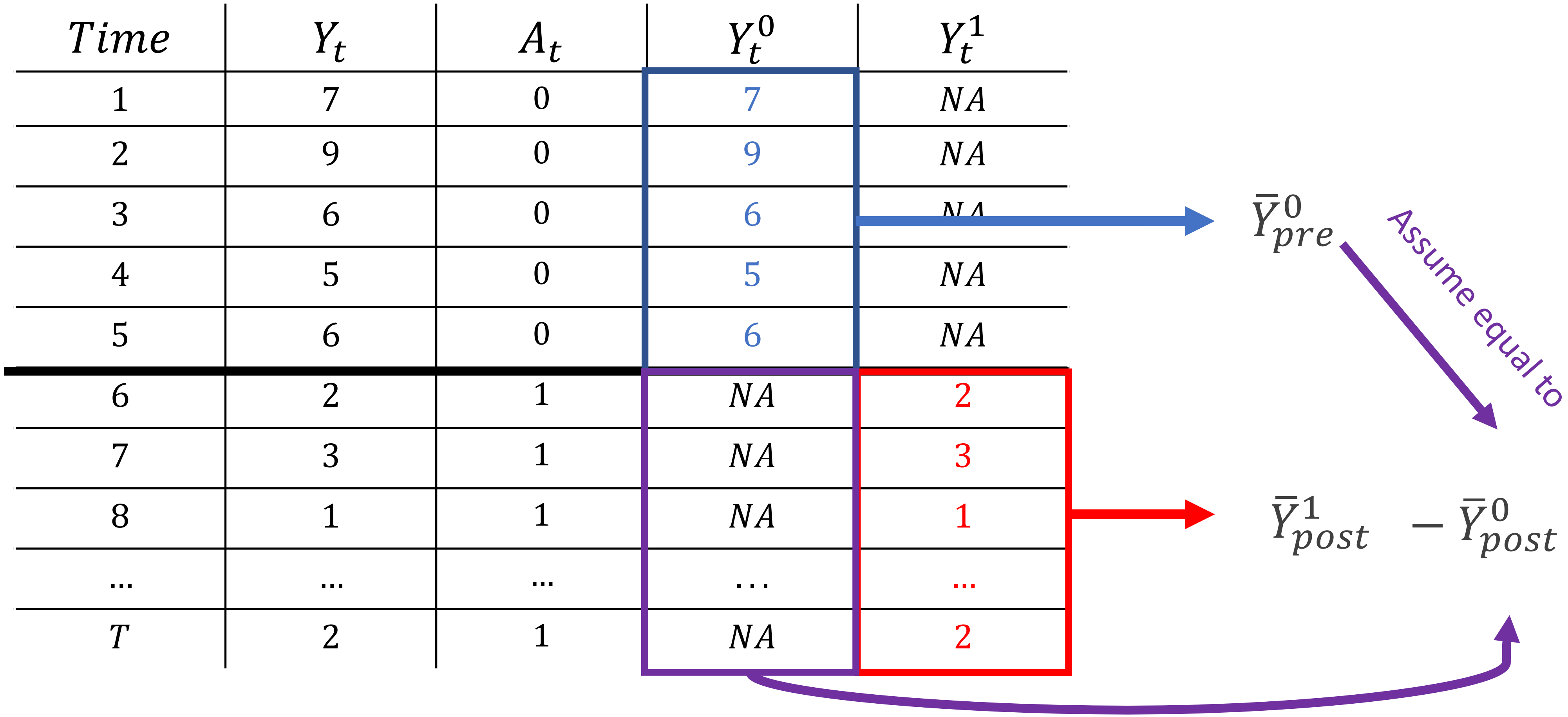
Pre – Post analysis

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
T	2	1	NA	2

$$\bar{Y}^0_{pre}$$

$$\bar{Y}^1_{post}$$

Pre – Post analysis



$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	$\dots$	…
T	2	1	NA	2

$\bar{Y}_{pre}^0$

Assume equal to

$\bar{Y}_{post}^1 \quad - \bar{Y}_{post}^0$

$$\overline{CE}_{post} = \bar{Y}_{post}^1 - \bar{Y}_{post}^0$$

Pre-post estimator

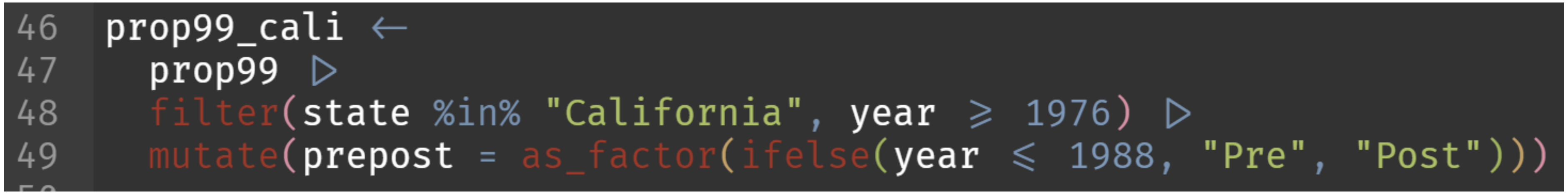
- Estimate the mean before the intervention $\bar{Y}_{pre}$
- Estimate the mean after the intervention $\bar{Y}_{post}$

$$\widehat{CE}_{post} = \bar{Y}_{post} - \bar{Y}_{pre}$$

- We can choose to consider equal time before and after the intervention (!)

Pre-post estimator

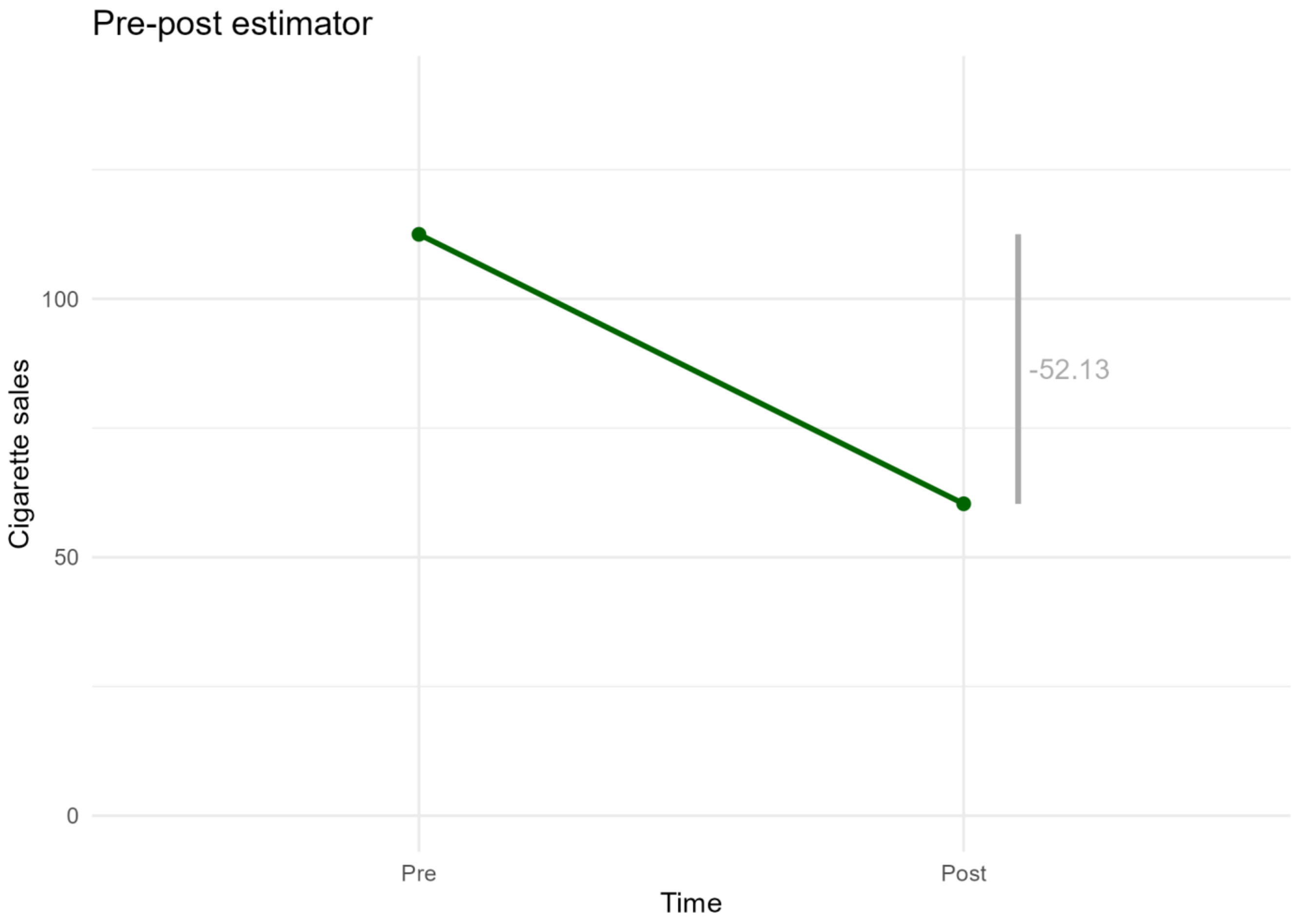
- Filter & compute pre-post factor variable



```r
prop99_cali <-
  prop99 |>
  filter(state %in% "California", year >= 1976) |>
  mutate(prepost = as_factor(ifelse(year <= 1988, "Pre", "Post")))
```

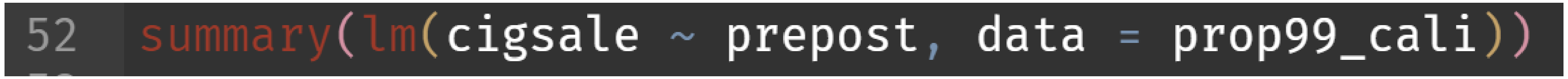
- Compute the pre-post difference

Pre-post estimator



Pre-post estimator

- But what about uncertainty?
- Use linear regression / OLS to compute $\widehat{CE}$

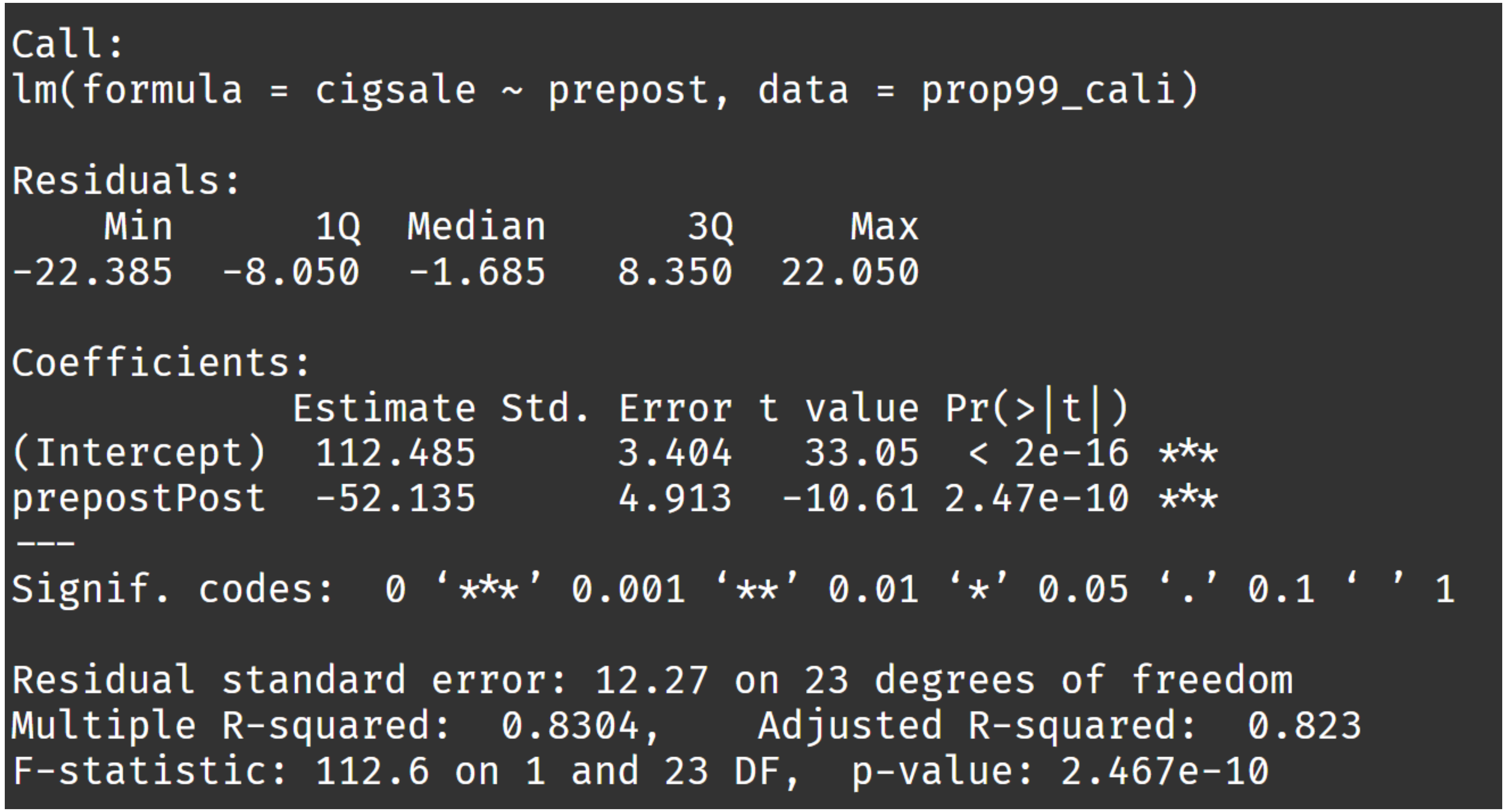


```r
summary(lm(cigsale ~ prepost, data = prop99_cali))
```

$Time$	Y_t	A_t	Y_t^0	Y_t^1
1	7	0	7	NA
2	9	0	9	NA
3	6	0	6	NA
4	5	0	5	NA
5	6	0	6	NA
6	2	1	NA	2
7	3	1	NA	3
8	1	1	NA	1
…	…	…	…	…
T	2	1	NA	2

Pre-post estimator

Result:



```
Call:
lm(formula = cigsale ~ prepost, data = prop99_cali)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.385  -8.050  -1.685   8.350  22.050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  112.485      3.404   33.05  < 2e-16 ***
prepostPost  -52.135      4.913  -10.61 2.47e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.27 on 23 degrees of freedom
Multiple R-squared:  0.8304,	Adjusted R-squared:  0.823 
F-statistic: 112.6 on 1 and 23 DF,  p-value: 2.467e-10
```

Standard errors assume no autocorrelation (!)

Pre-post estimator

The causal effect of the tax increase on cigarette sales is an average yearly decrease of 52 packs of cigarettes per 100000 people

- Interpretation depends on choices in analysis
- In this case: effect averaged over 1989 – 2000
- Be precise – define your causal estimand $\overline{CE}_{post}$

Pre-post estimator

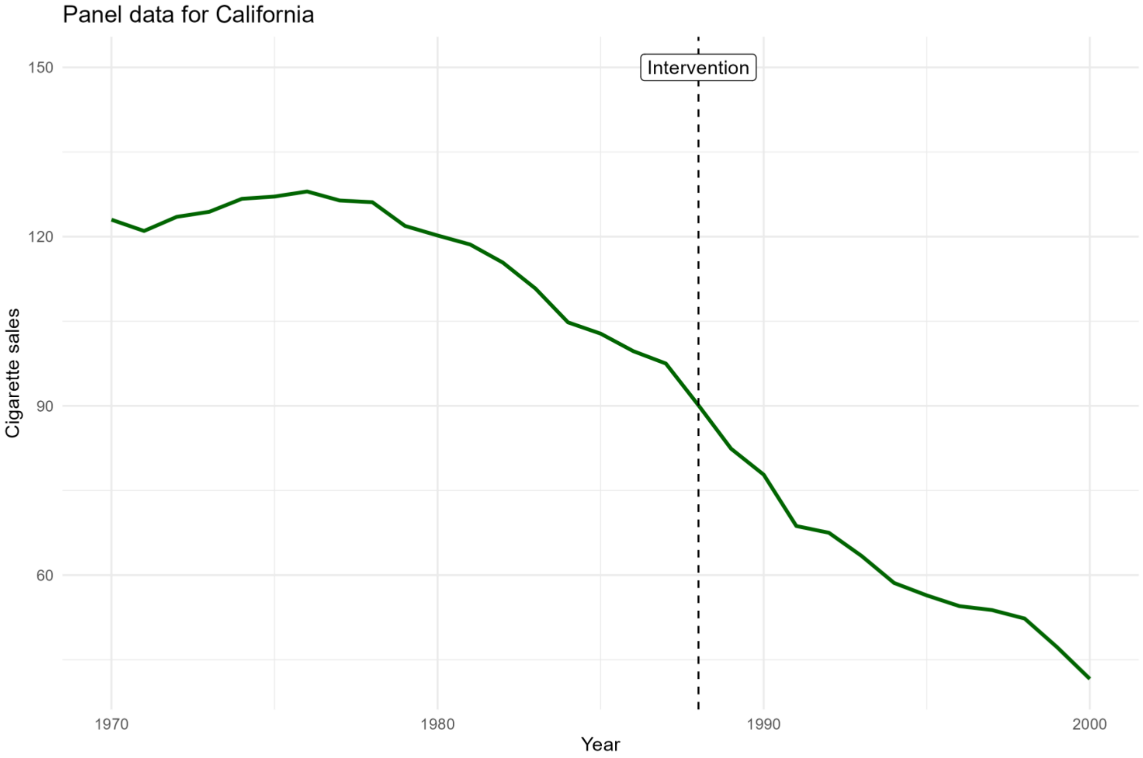
Most important / strict assumption:

No trend in time

- Remember: we assumed $\bar{Y}^0_{post} = \bar{Y}^0_{pre}$
- We assume the pre-post difference is caused by intervention **only**
- If trend exists, then the effect of trend and of intervention cannot be distinguished

Pre-post estimator

- Is there a trend in time, independent of the intervention?
- How much of pre-post difference is caused by intervention?



Difference-in-Differences

Difference-in-differences

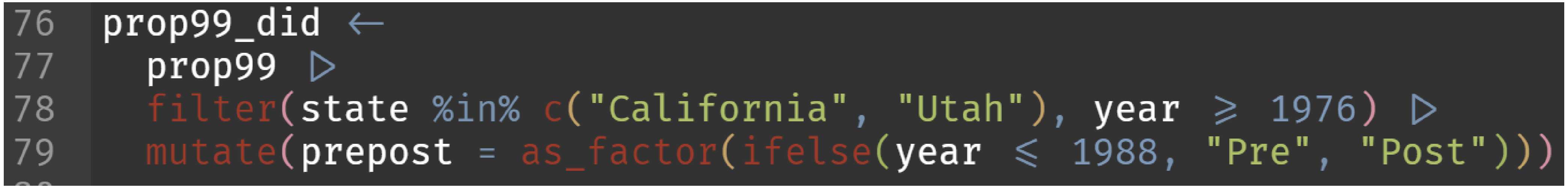
„transparent and often at least superficially plausible”

Angrist, J. D. and Krueger, A. B. (1999). Empirical strategies in labor economics. In Handbook of labor economics, volume 3, pages 1277–1366. Elsevier.

- Used a lot in economics
- There is a lot of discussion around this topic
- We will explain the basic method here
- There are a lot of possible extensions!

Difference-in-differences

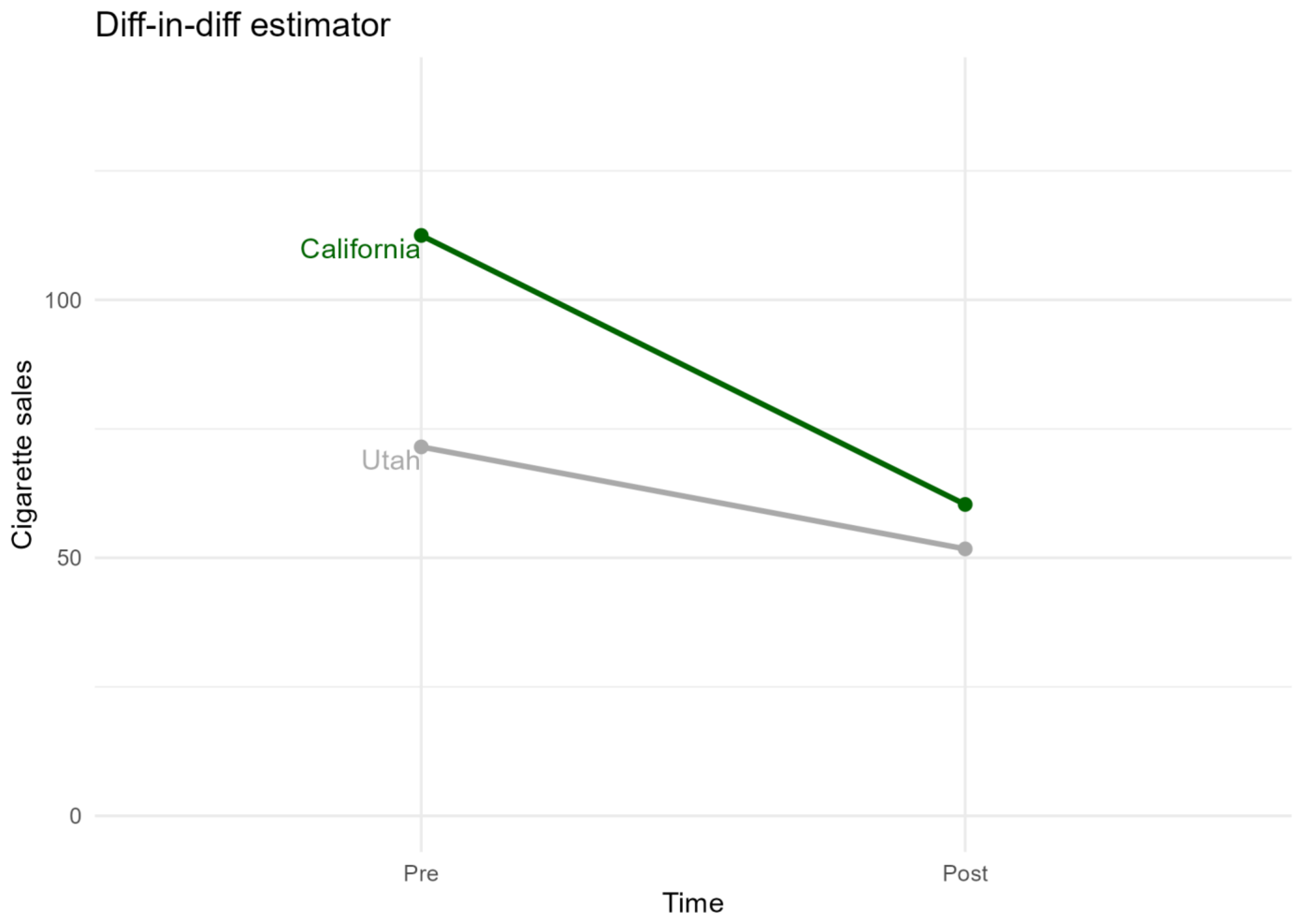
- Like before:
  - Measure outcome pre- and post-intervention
  - Choose what time period to consider
- Unlike before:
  - Also measure pre & post outcome C for a ***control unit***
  - The control should not have received the intervention



```
prop99_did <-
  prop99 |>
  filter(state %in% c("California", "Utah"), year >= 1976) |>
  mutate(prepost = as_factor(ifelse(year <= 1988, "Pre", "Post")))
```

$Time$	Y_t	A_t	Y_t^0	Y_t^1	C_{1t}
1	7	0	7	NA	2
2	9	0	9	NA	6
3	6	0	6	NA	4
4	5	0	5	NA	2
5	6	0	6	NA	1
6	2	1	NA	2	3
7	3	1	NA	3	2
8	1	1	NA	1	4
…	…	…	…	…	…
T	2	1	NA	2	3

Difference-in-differences



Pre-post estimator

- Like before, we estimate the following quantity:

$$\overline{CE}_{post} = \bar{Y}^1_{post} - \bar{Y}^0_{post}$$

- Now, we assume there is an effect of time: $\beta \cdot Time$
- We can represent unobservable $\bar{Y}^0_{post}$ as

$$\bar{Y}^0_{post} = \bar{Y}^0_{pre} + \beta \cdot Time$$

Pre-post estimator

- But the trend $\beta \cdot Time$ is also unobservable!
- Solution: assume equal trends for Utah and California

$$\beta \cdot Time = (\bar{C}^0_{post} - \bar{C}^0_{pre})$$

- Thus, our model for the counterfactual is:

$$\bar{Y}^0_{post} = \bar{Y}^0_{pre} + (\bar{C}^0_{post} - \bar{C}^0_{pre})$$

Pre-post estimator

- Plugging this into the causal effect equation:

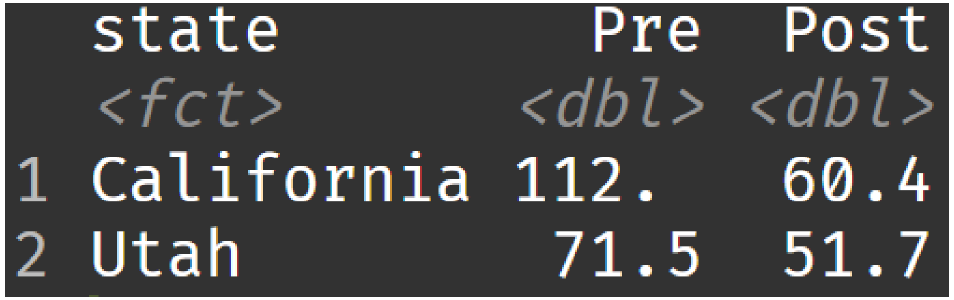
$$\overline{CE}_{post} = \left(\bar{Y}^1_{post} - \bar{Y}^0_{pre}\right) - \left(\bar{C}^0_{post} - \bar{C}^0_{pre}\right)$$

- Difference in differences!

$$\widehat{CE}_{post} = \left(\bar{Y}_{post} - \bar{Y}_{pre}\right) - \left(\bar{C}_{post} - \bar{C}_{pre}\right)$$

Difference-in-differences

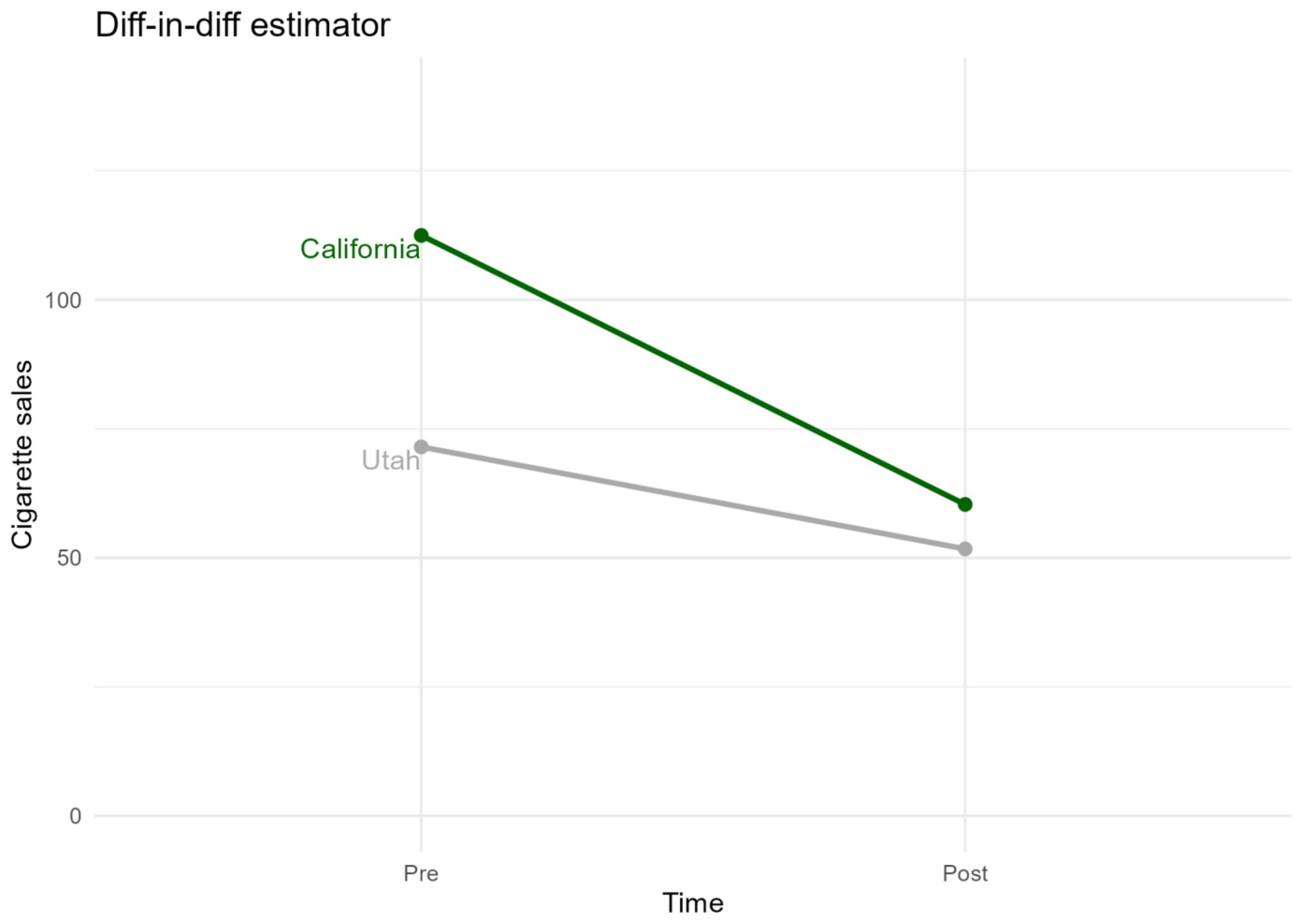
CE = (Cali_post – Cali_pre) – (Utah_post – Utah_pre)



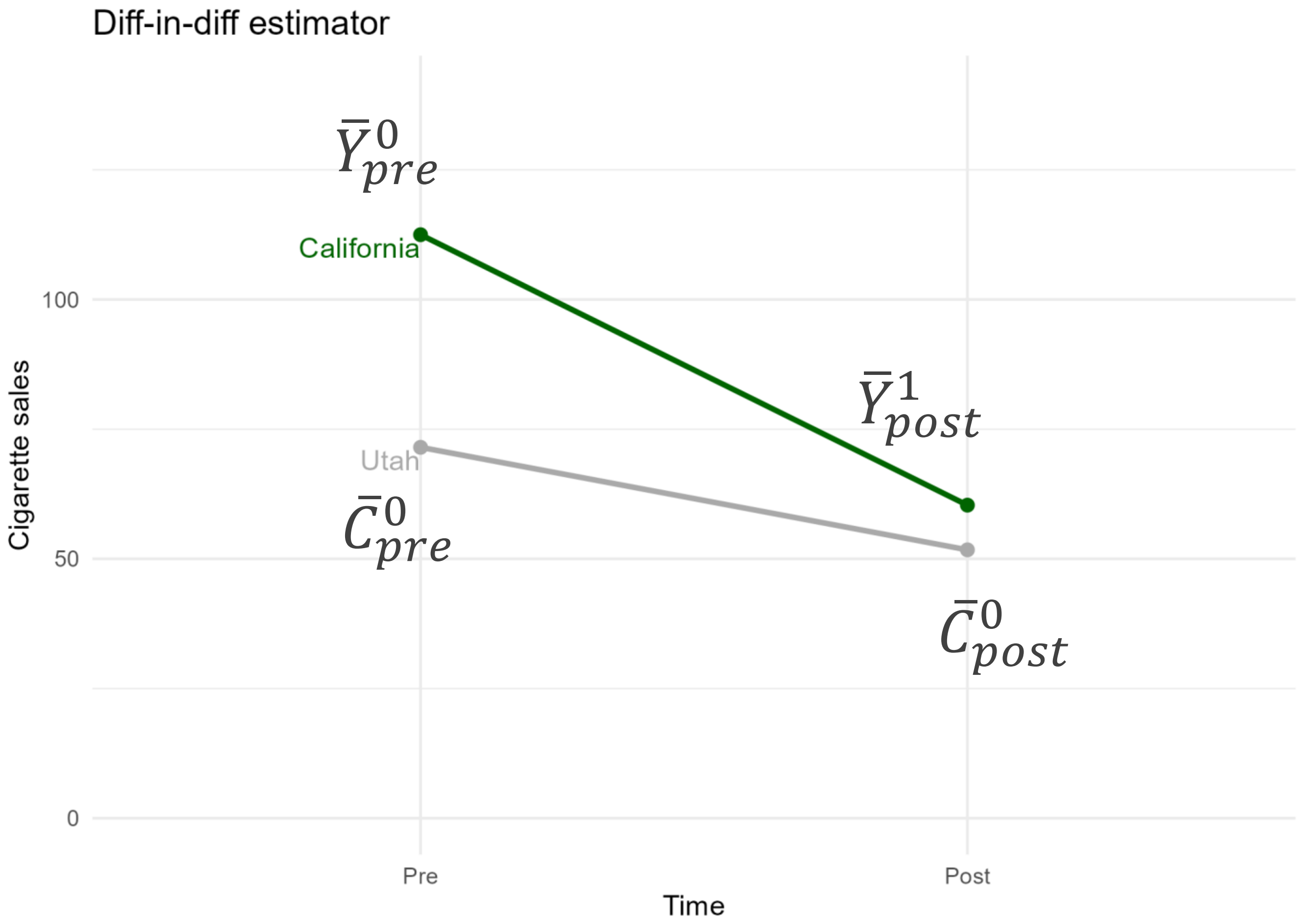
```
  state        Pre  Post
  <fct>      <dbl> <dbl>
1 California 112.   60.4
2 Utah        71.5  51.7
```

(60.4 - 112)-(51.7 - 71.5) = -32.3

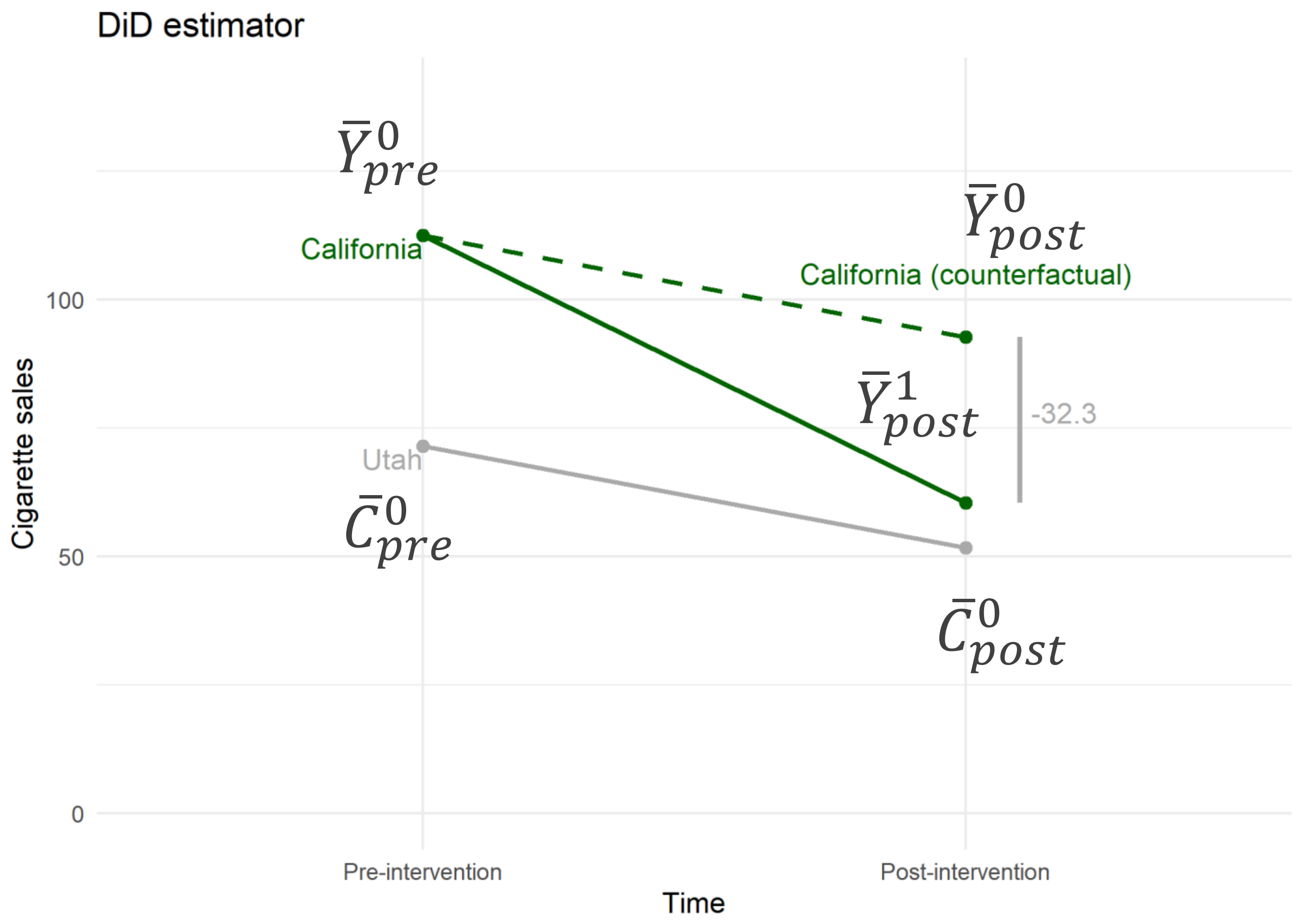
Difference-in-differences



Difference-in-differences

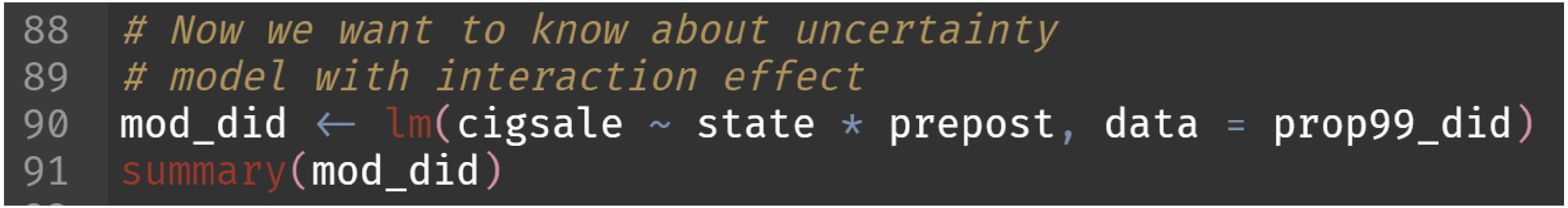


Difference-in-differences



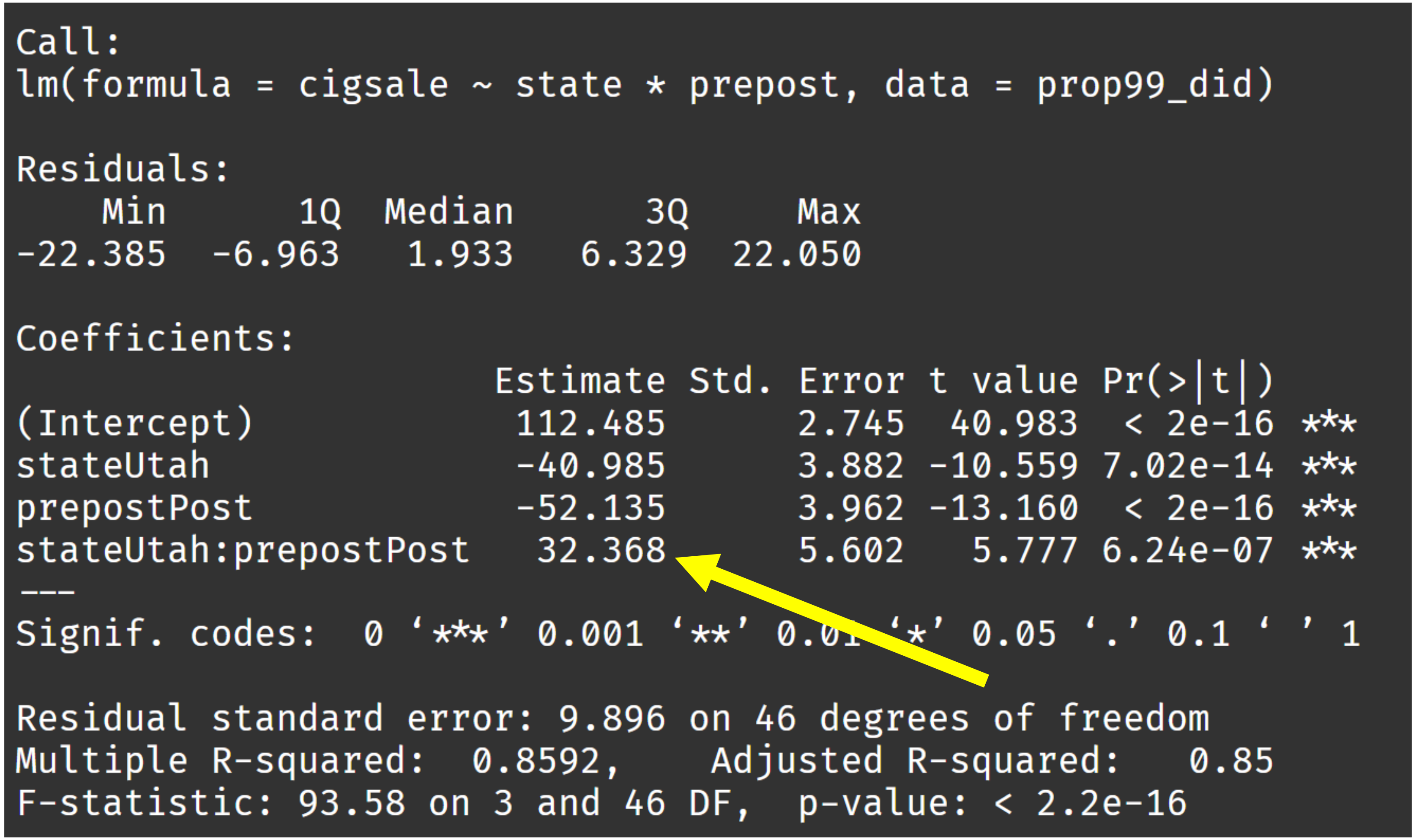
Difference-in-differences

- But what about uncertainty?
- Use linear regression / OLS to compute $\widehat{CE}$



```r
# Now we want to know about uncertainty
# model with interaction effect
mod_did <- lm(cigsale ~ state * prepost, data = prop99_did)
summary(mod_did)
```

Difference-in-differences



```
Call:
lm(formula = cigsale ~ state * prepost, data = prop99_did)

Residuals:
    Min      1Q  Median      3Q     Max 
-22.385  -6.963   1.933   6.329  22.050 

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)    
(Intercept)            112.485      2.745  40.983  < 2e-16 ***
stateUtah              -40.985      3.882 -10.559 7.02e-14 ***
prepostPost            -52.135      3.962 -13.160  < 2e-16 ***
stateUtah:prepostPost   32.368      5.602   5.777 6.24e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.896 on 46 degrees of freedom
Multiple R-squared:  0.8592,	Adjusted R-squared:   0.85 
F-statistic: 93.58 on 3 and 46 DF,  p-value: < 2.2e-16
```

Standard errors assume no autocorrelation (!)

Most important assumptions

Parallel trends

$\beta \cdot Time = (\bar{C}^0_{post} - \bar{C}^0_{pre})$

Time effect is the same for the treated and the control unit

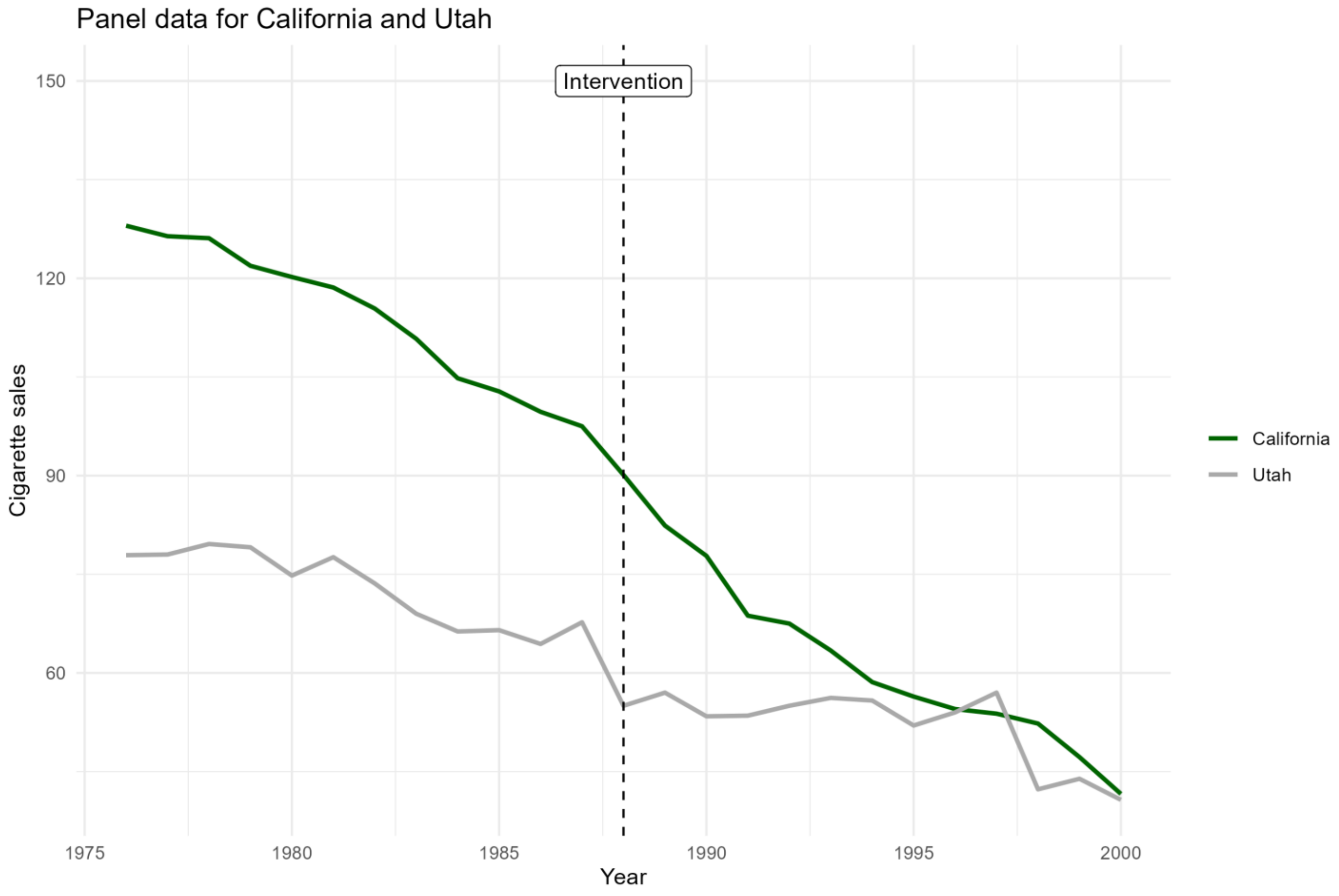
No interference / spillover

$\bar{C}_{post} = \bar{C}^0_{post}$

The control does not receive any intervention effect

Most important assumptions

- Can we assume parallel trends?
- At least superficially plausible ☺



Practical: pre-post & DiD

Work in pairs/groups!
Take a break from 10:45 to 11:00

Break