Estimating the causal effect

Basic methods

In this part

- Running example: California proposition 99 data
- Pre-post estimator
- Difference-in-differences estimator

Proposition 99

Proposition 99

- Most 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.*

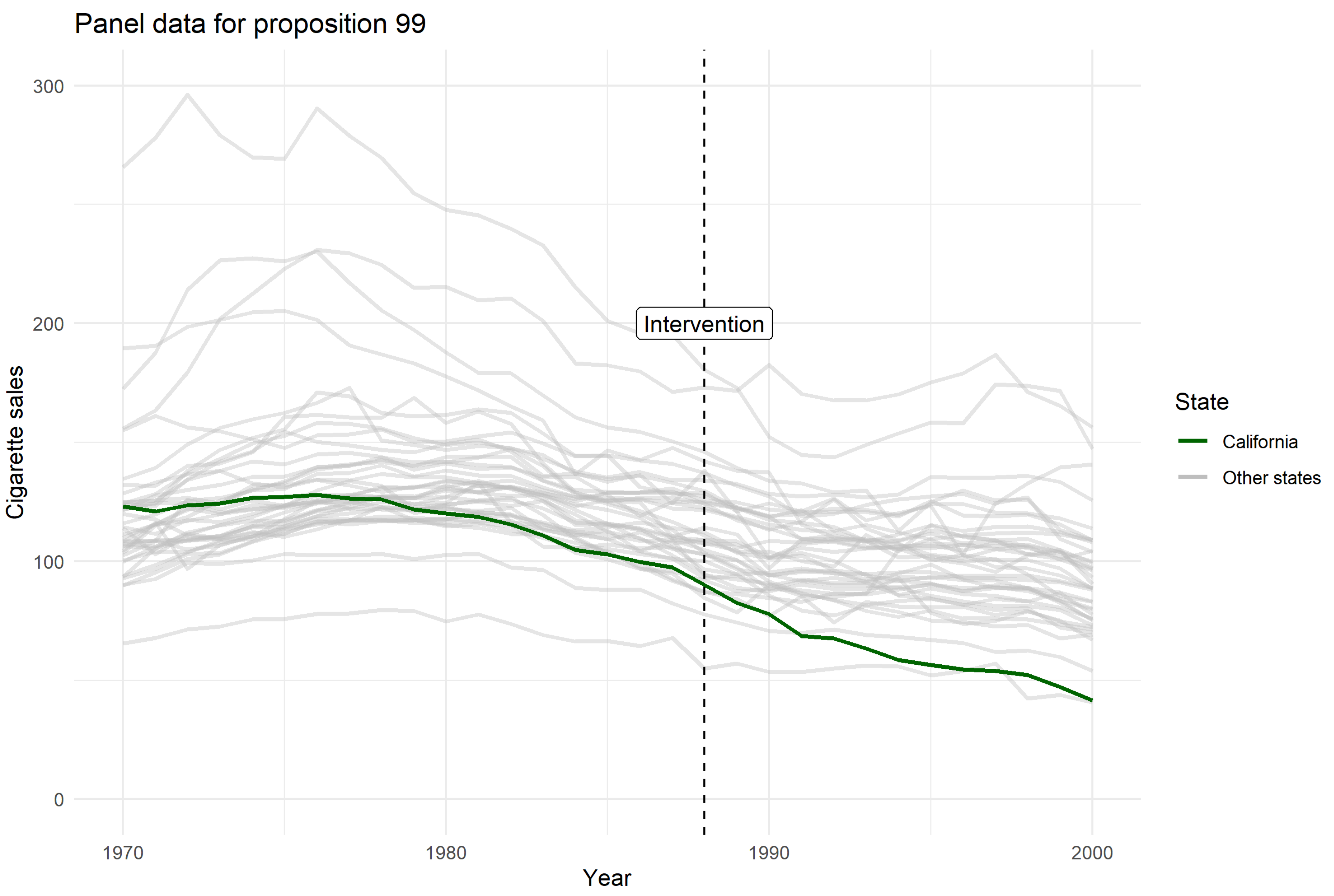
- In 1988, the state of California imposed a 25% tax on tobacco cigarettes
- Total savings in personal health care expenditure until 2004 is $86 billion (Lightwood et al., 2008)

Proposition 99

- We prepared a dataset for this workshop:

`proposition99.rds`

- Panel 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:**

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

- This works well with our prepared dataset

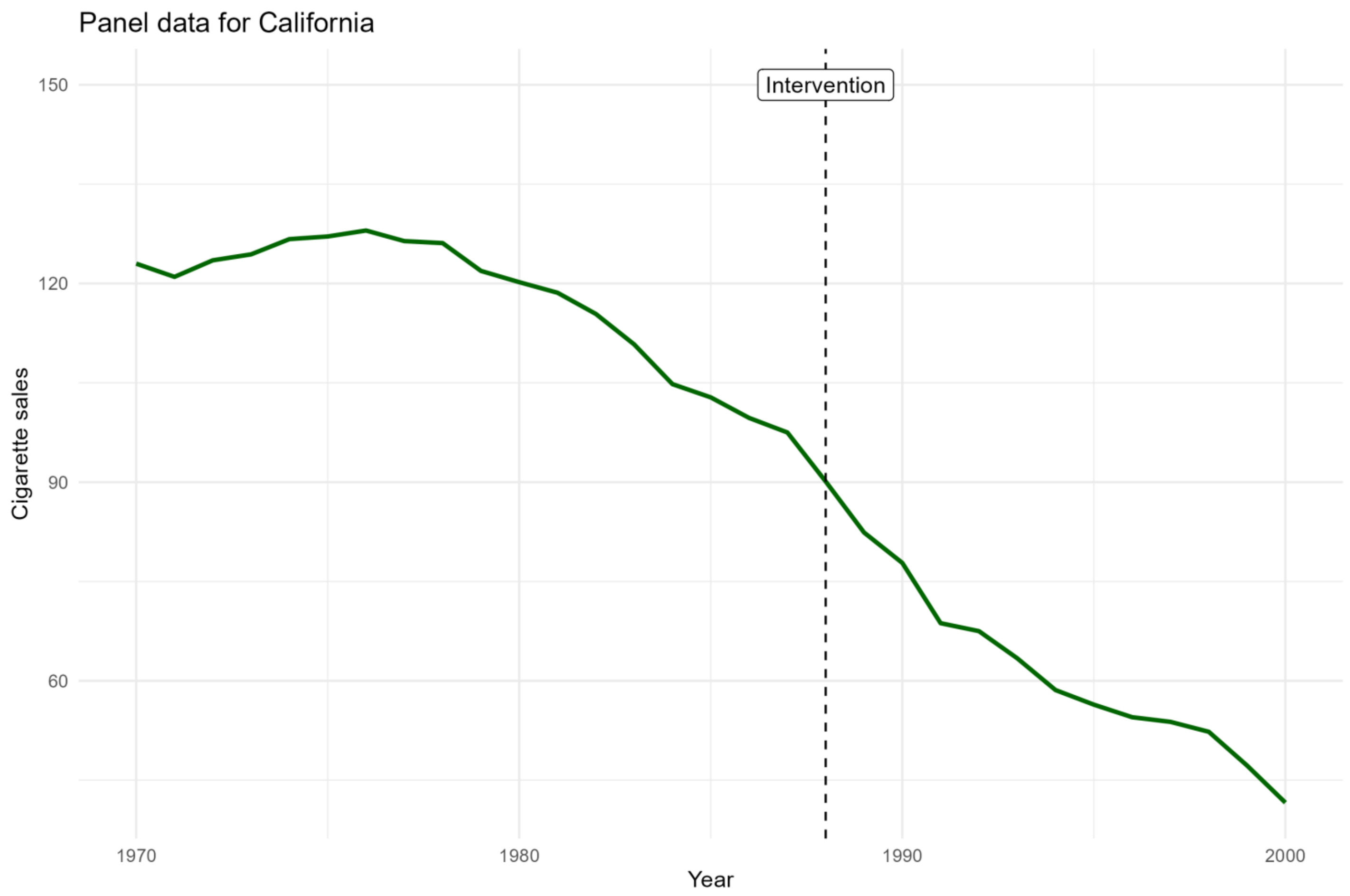
Proposition 99

```
# 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
```

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

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

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 (!)

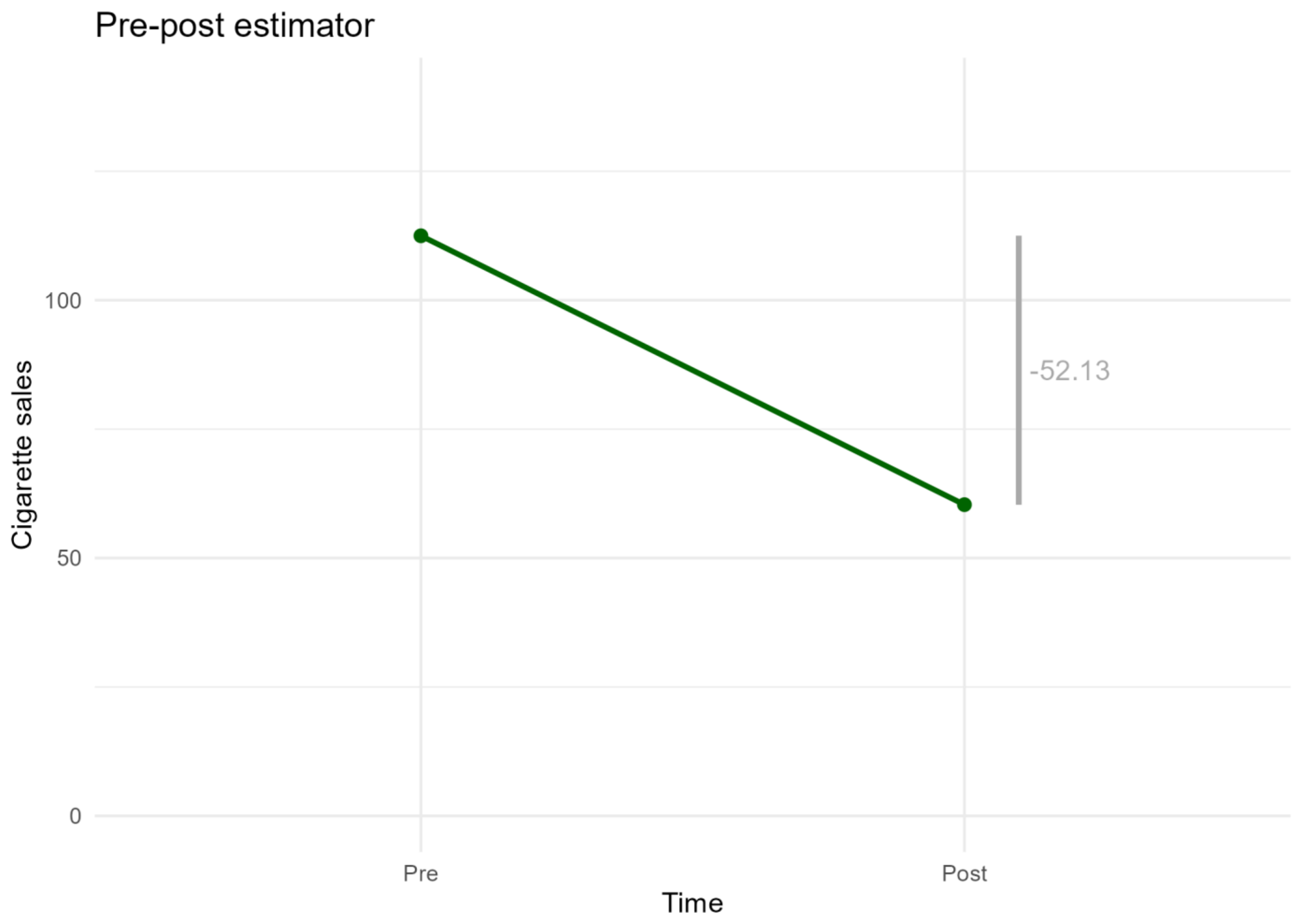
Pre-post estimator

- 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 $\widehat{CE}$

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

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 a 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

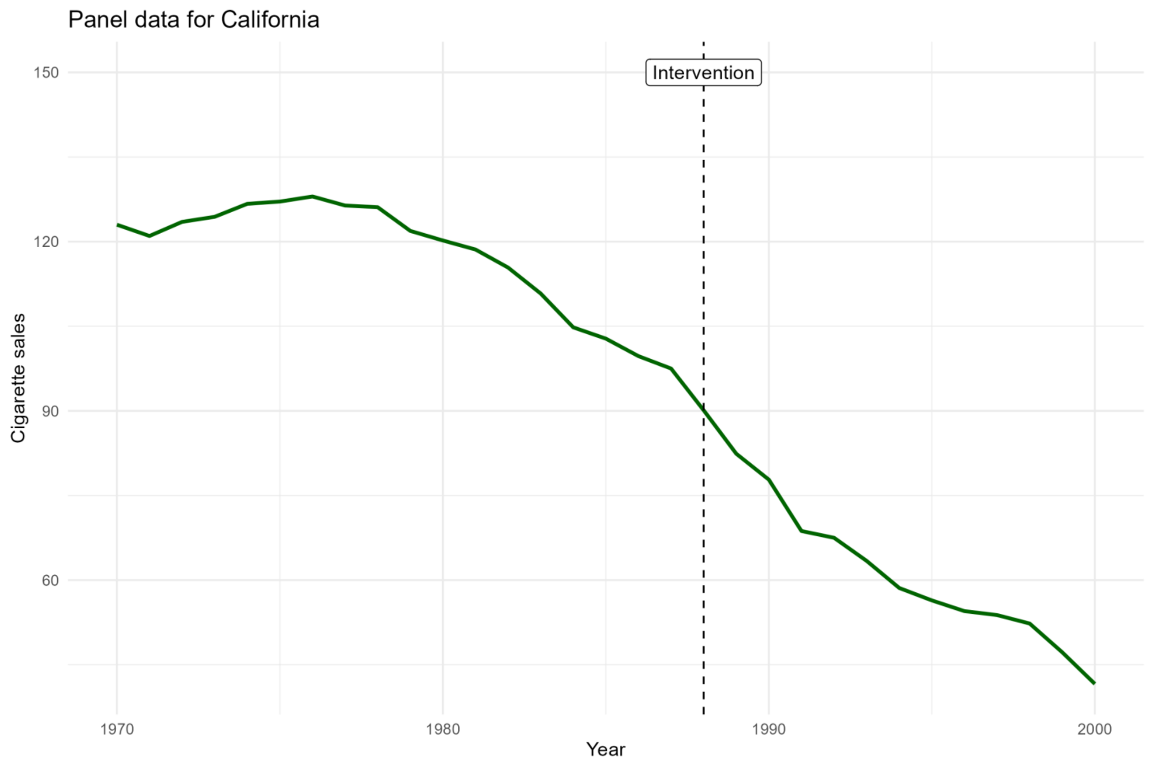
Most important / strict assumption:

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.

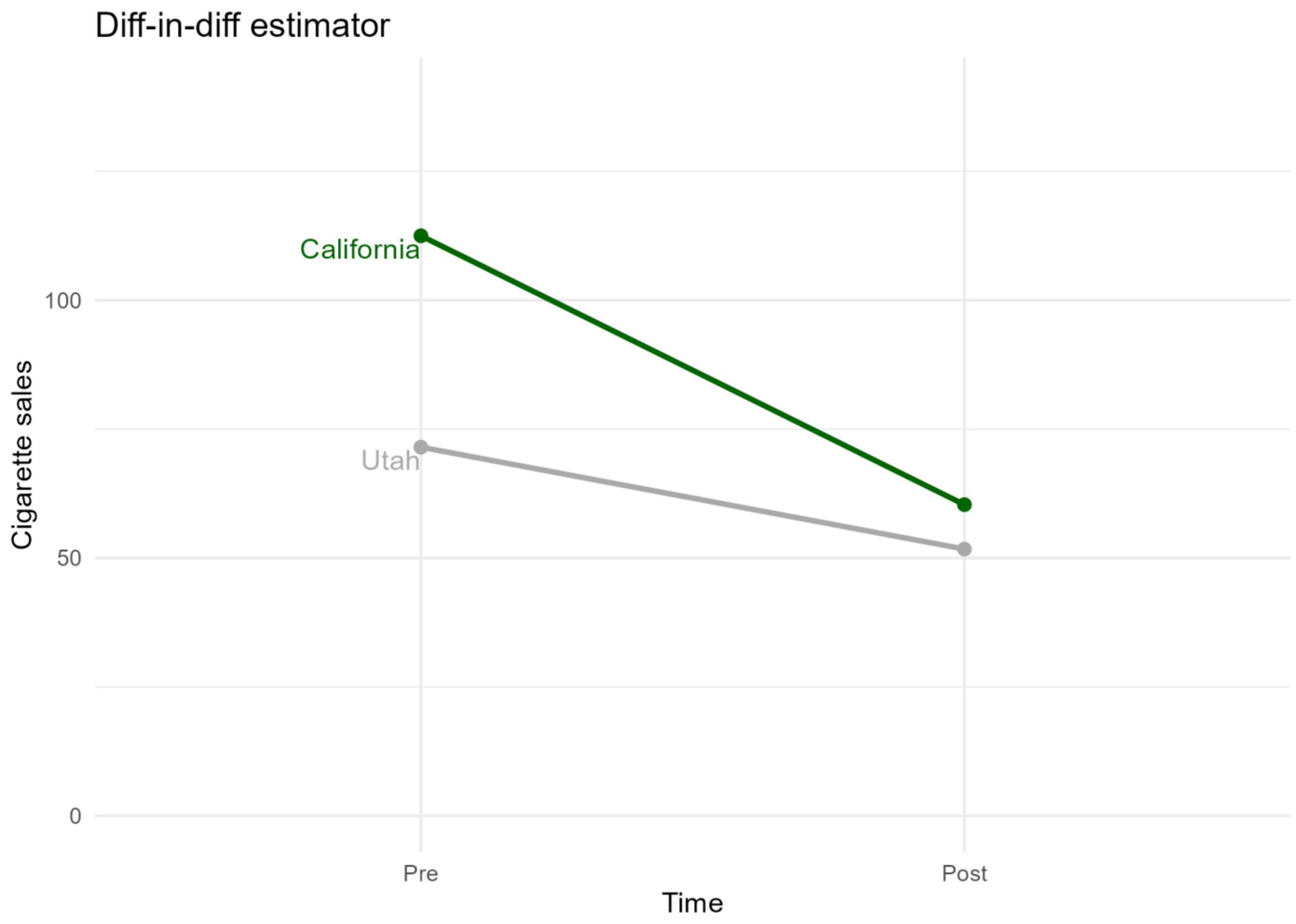
- Used a lot in economics
- There is a lot of discussion around this topic
- We will explain the basic method here
- There are a lot of possible extensions!

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")))
```

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:

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

- Difference in differences!

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

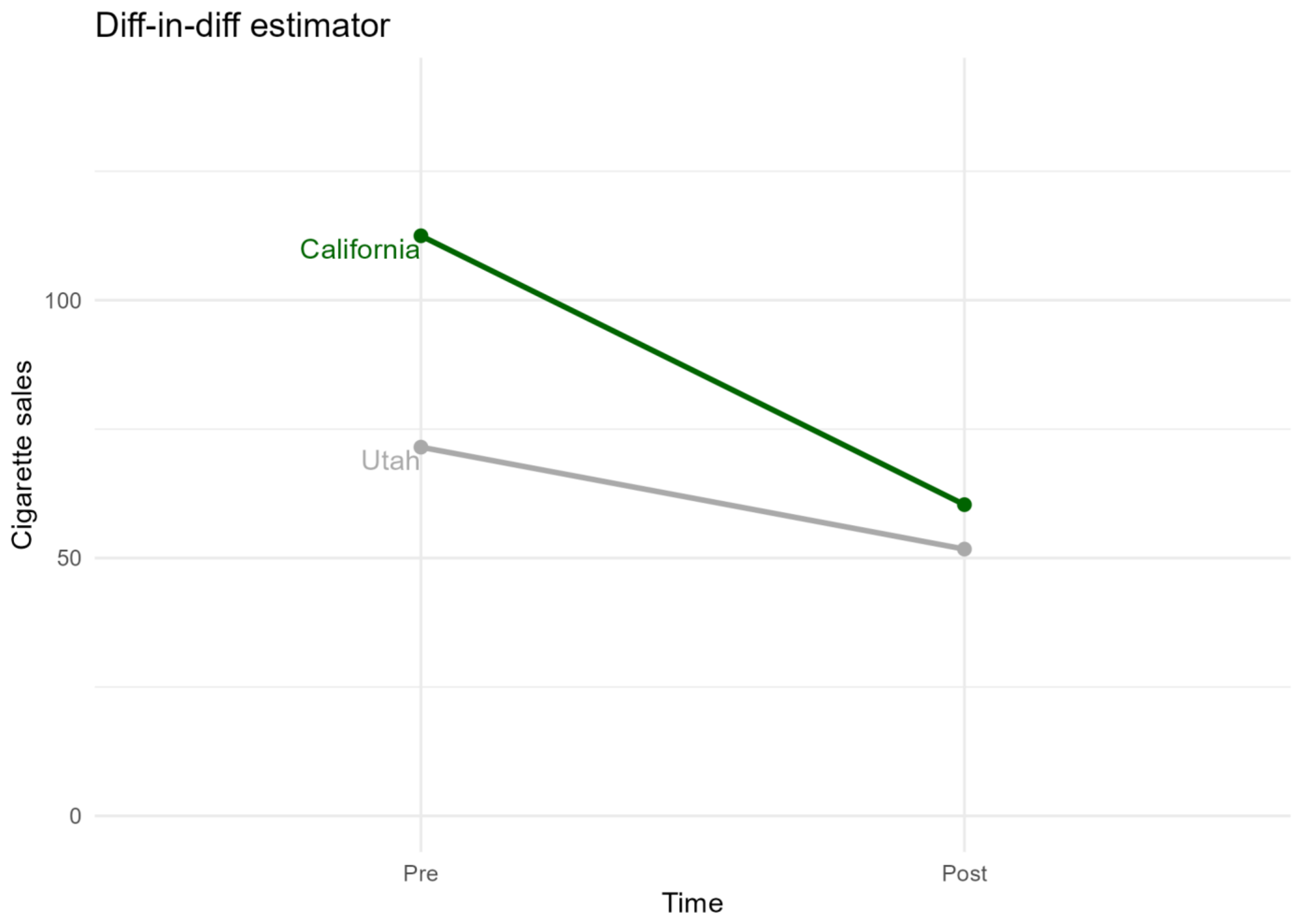
Difference-in-differences

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

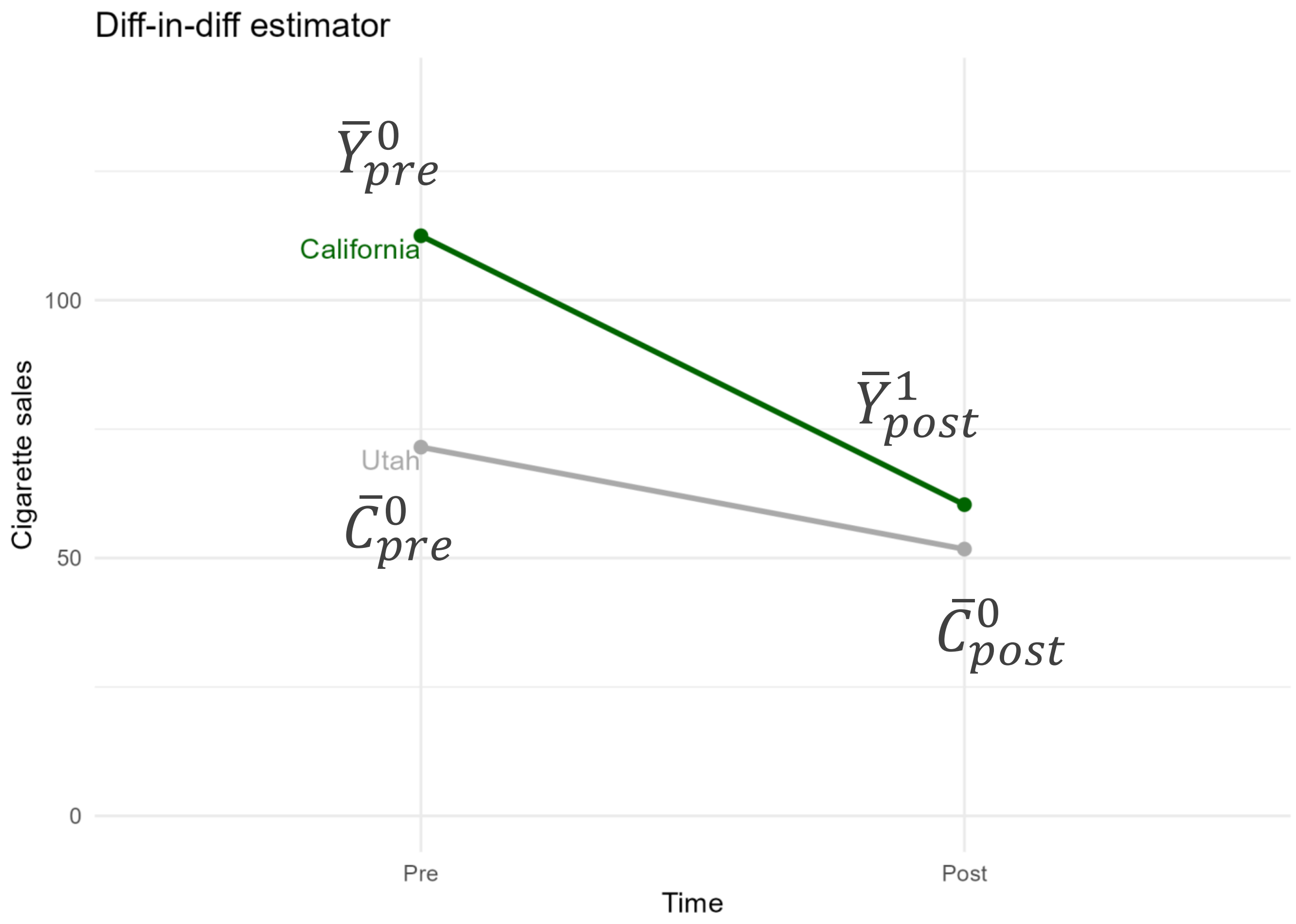
	state	Pre	Post
	<fct>	<dbl>	<dbl>
1	California	112.	60.4
2	Utah	71.5	51.7

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

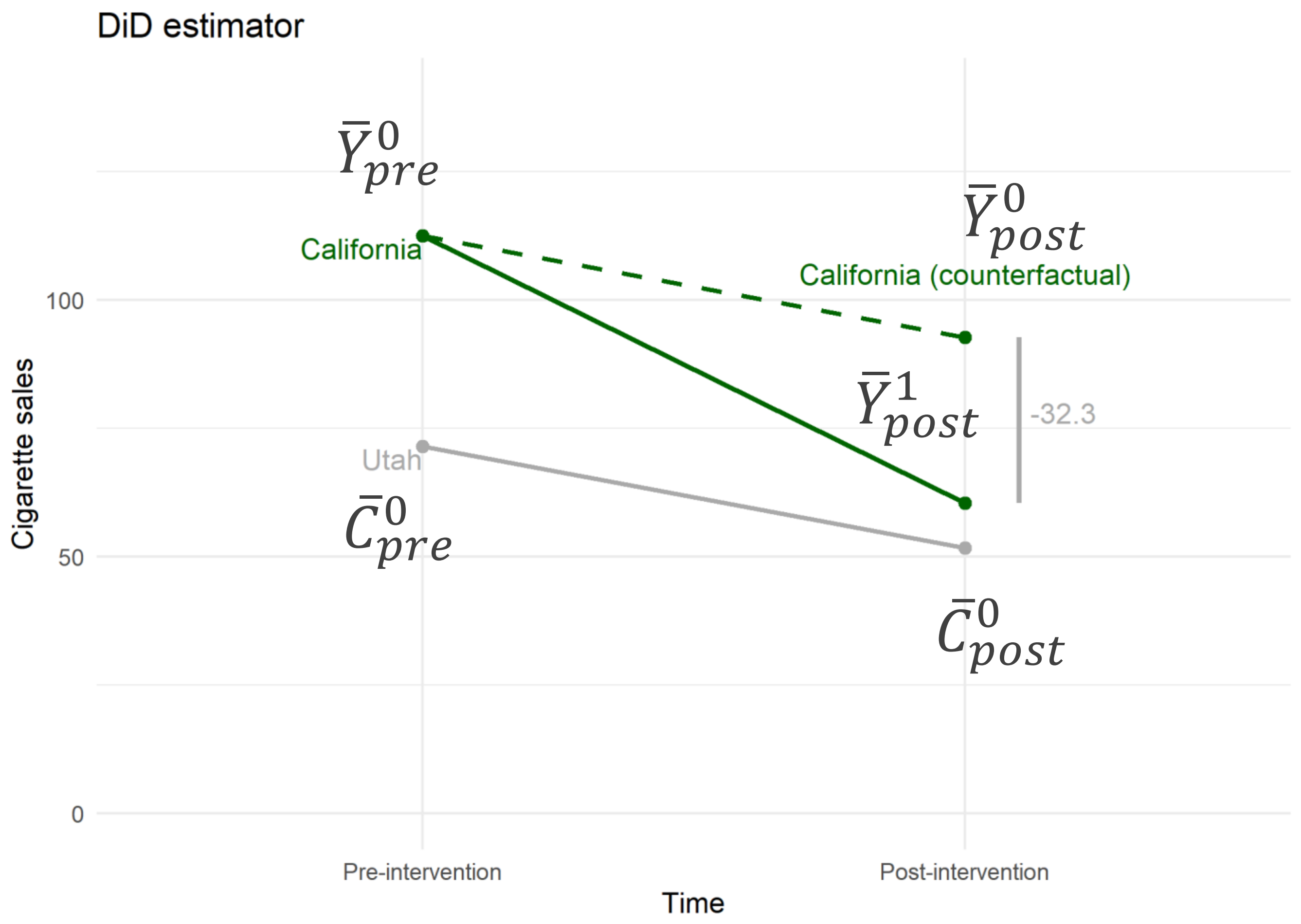
Difference-in-differences



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

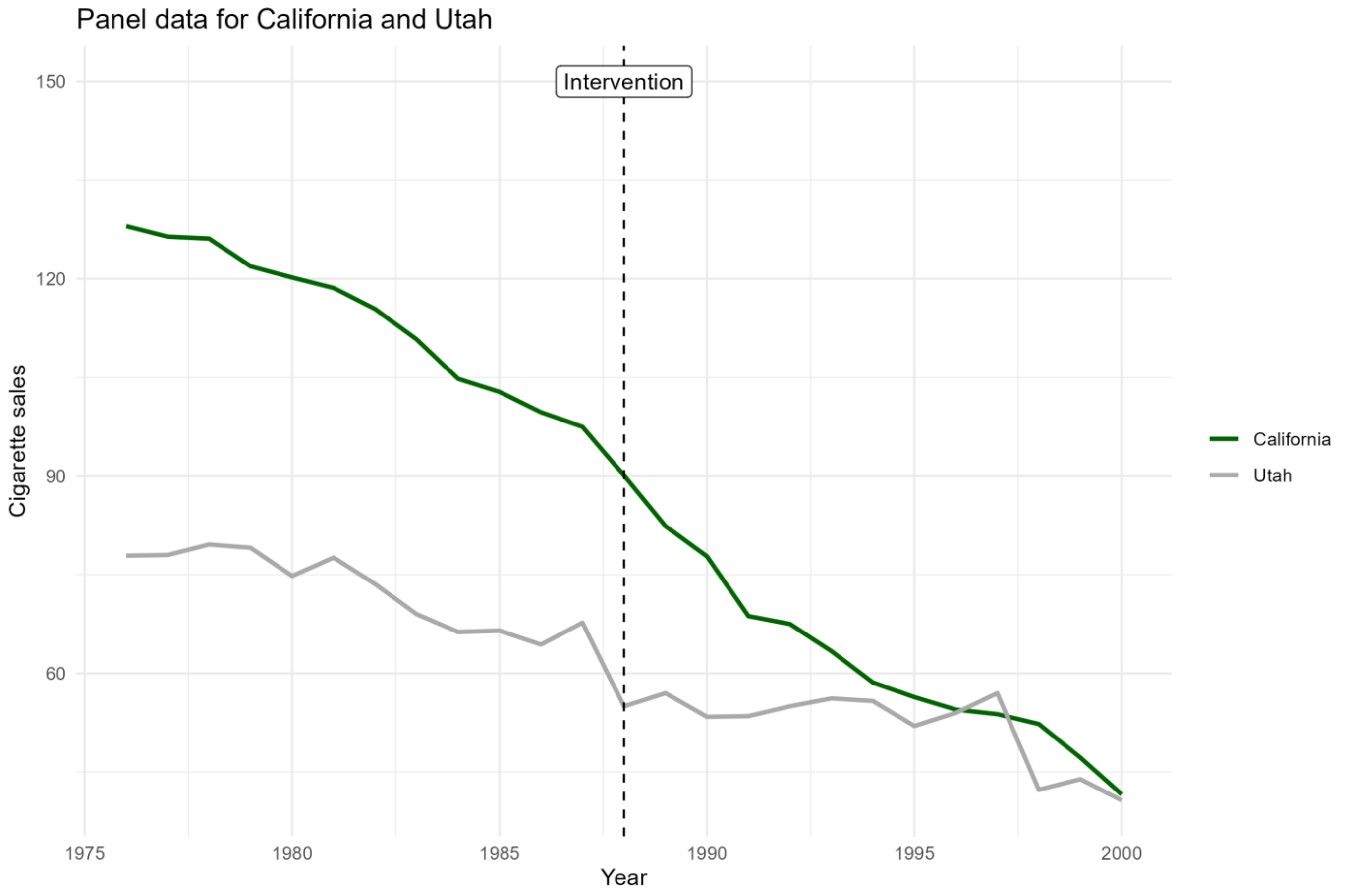
$\bar{C}_{post} = \bar{C}^0_{post}$

The control does not receive any intervention effect

Most important assumptions

- Can we assume parallel trends?

- At least superficially plausible ☺



Practical: data, pre-post, DiD

Work in pairs!
Take a break from 10:30 to 10:45

Break