Interrupted Time Series
& Regression Discontinuity

The story so far

The **proposition 99** data has a number of pre- and post-intervention observations (i.e. time points)

So far we **computed averages** and estimated

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

Interrupted Time Series

Interrupted Time Series:

- Instead of taking averages, use pre-intervention data Y^0_{pre} to **forecast/predict** Y^0_{post}

- Once we have predictions $\widehat{Y}^0_{post}$, we compare those to the observed Y^1_{post}

- I.e. we use pre-intervention data to **impute** the missing counterfactual

This means we can in principle estimate

$$\widehat{CE}_t = Y^1_t - \widehat{Y^0_t}$$

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

Interrupted Time Series

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

Fit a forecasting Model

$$\widehat{Y}_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \beta * Time$$

Interrupted Time Series

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	$\widehat{Y_6^0}$	2
7	3	1	$\widehat{Y_7^0}$	3
8	1	1	$\widehat{Y_8^0}$	1
…	…	…	…	…
T	2	1	$\widehat{Y_T^0}$	2

Fit a forecasting Model

$$\widehat{Y}_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \beta * Time$$

Make forecasts

Interrupted Time Series

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	$\widehat{Y_6^0}$	2
7	3	1	$\widehat{Y_7^0}$	3
8	1	1	$\widehat{Y_8^0}$	1
…	…	…	…	…
T	2	1	$\widehat{Y_T^0}$	2

Fit a forecasting Model

$$\widehat{Y}_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots \beta * Time$$

Make forecasts

$$\widehat{CE}_t = Y_t^1 - \widehat{Y_t^0}$$

Interrupted Time Series

Point forecasts allow us to compute point estimates of our causal effect

$$\widehat{CE}_t = Y_t^1 - \widehat{Y_t^0}$$

We can quantify our **uncertainty** about the causal effect based on our **uncertainty** around our (model-based) forecasts

Building a forecasting model

Much of the challenge of this approach is in choosing an appropriate **forecasting** model

These can be very simple or very complex, e.g.:

- If we forecast with the **mean** we are very close to the post – pre analysis
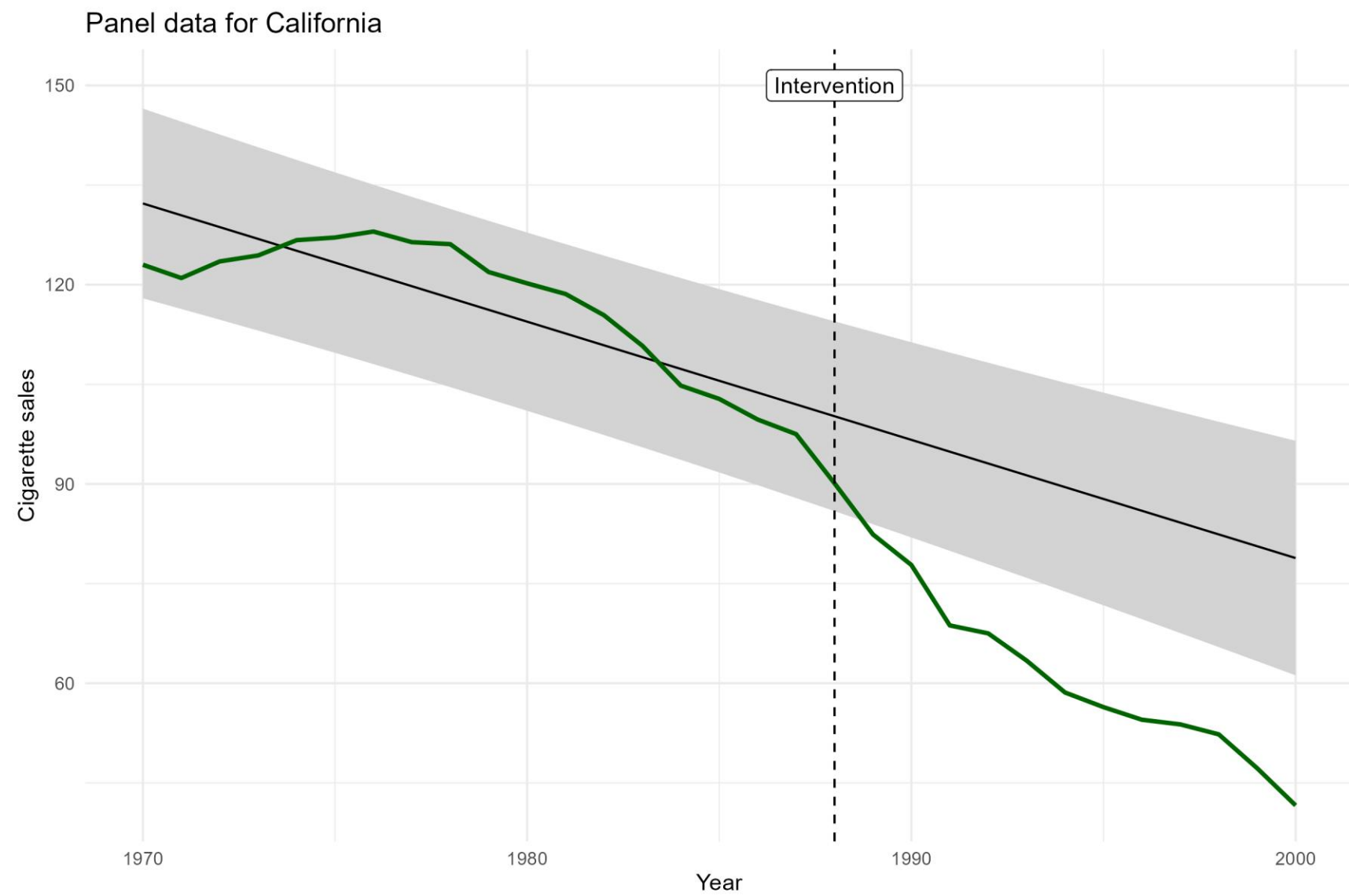
$$Y_t = \mu_{pre} + e_t$$

Building a forecasting model

Much of the challenge of this approach is in choosing an appropriate **forecasting** model

These can be very simple or very complex, e.g.:

- If we forecast with the **mean** we are very close to the post – pre analysis

$$Y_t = \mu_{pre} + e_t$$

- If we forecast with a **growth curve** we would model the overall time trend

$$Y_t = \beta_0 + \beta_1 \, Time + e_t$$

Forecasting with growth curves

```r
# predict pre-intervention sales by year
fit_growth <- lm(
  formula = cigsale ~ year,
  prop99_ts |> filter(prepost == "Pre")
)

# predict values for the post-intervention period
pred <- predict(
  object = fit_growth,
  newdata = prop99_ts,
  interval = "prediction"
)
```

Panel data for California

Building a forecasting model

Much of the challenge of this approach is in choosing an appropriate **forecasting** model

These can be very simple or very complex, e.g.:

- If we forecast with the **mean** we are very close to the post – pre analysis

$$Y_t = \mu_{pre} + e_t$$

- If we forecast with a **growth curve** we would model the overall time trend

$$Y_t = \beta_0 + \beta_1\,Time + e_t$$

- We can also forecast by using **time-series models** that model *autocorrelation*

$$Y_t = \phi_1 Y_{t-1} + e_t \qquad Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + e_t \qquad Y_t - Y_{t-1} = \gamma e_{t-1} + e_t$$

*e.g. ARIMA models can account for autocorrelation and time trends*

Fitting time-series models fpp3

Prep data as tsibble object

```r
prop99_ts <-
  prop99 |>
  filter(state == "California") |>
  select(year, cigsale) |>
  mutate(prepost = factor(year > 1988, labels = c("Pre", "Post"))) |>
  as_tsibble(index = year) |>
  mutate(year0 = year - 1989)
```

Fit a model (automatic model selection on AR, differencing, and MA components)
and produce forecasts

```r
fit_arima <-
  prop99_ts |>
  filter(prepost == "Pre") |>
  model(timeseries = ARIMA(cigsale, ic = "aicc"))

# create forecasts
fcasts <- forecast(fit_arima, h = "12 years")
```
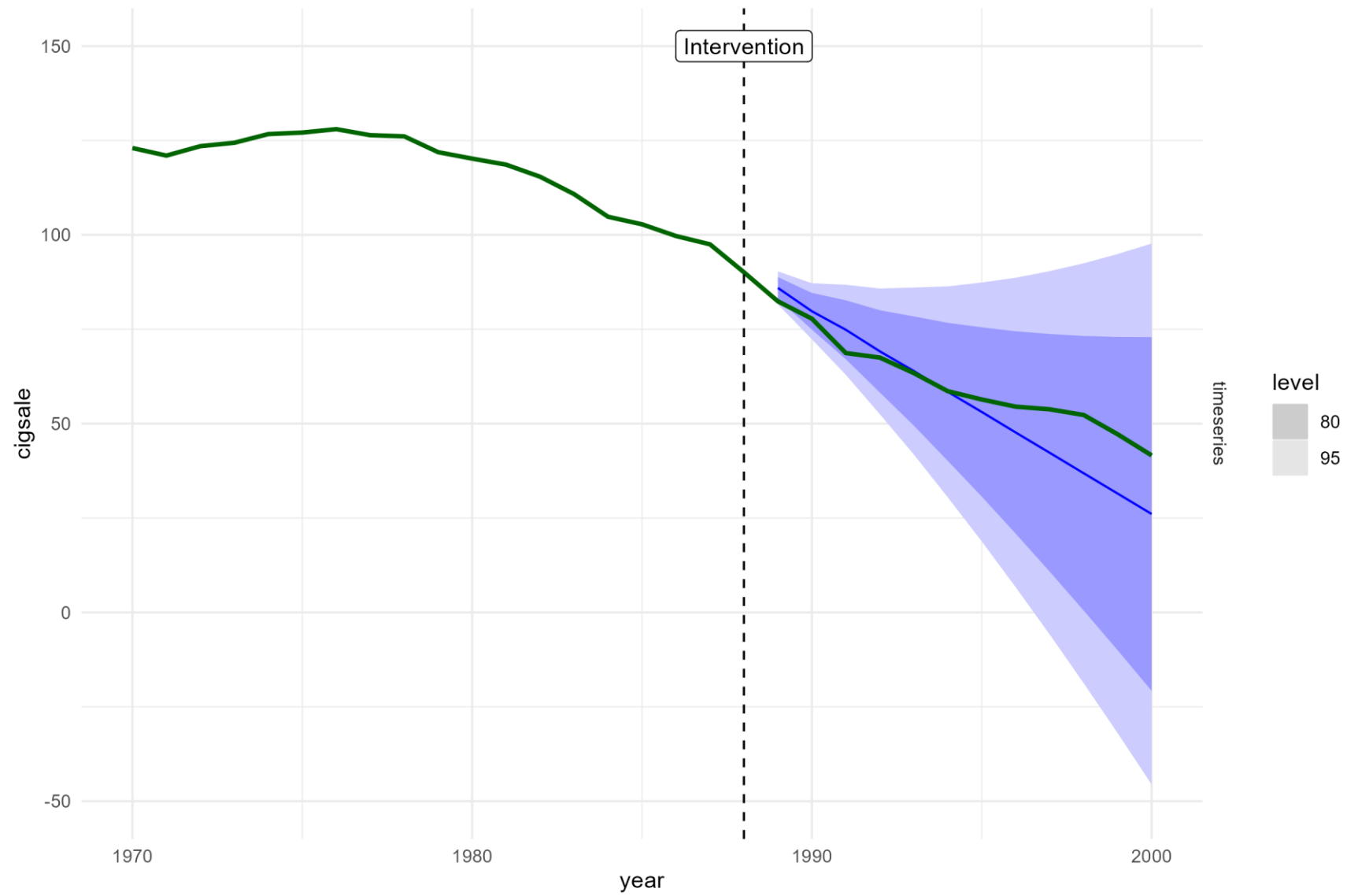
Key Assumptions

Our inferences about the causal effect are entirely dependent on being able to fit **an appropriate forecasting model**

- i.e. one that correctly captures the trend and autocorrelation structures in the data

In practice, this may be **very difficult**

Key Assumptions

Data driven approaches can be applied, but may only be feasible with **a large amount of pre-intervention training data**

- We use information criteria for model selection

- See also: cross-validation

In addition, different forecasting models come with their own assumptions,

- E.g. **constant trend** or **time-invariant relationships**

Poor forecasts = Poor estimates (and uncertainty) of causal effects

Key Assumptions

When comparing to the **pre-post design**;

- We relax the no-trend assumption: we model any trend / serial dependence

No-confounding assumption:

- We still assume that any changes can be attributed to the intervention

- And not, e.g., something else that happened around the same time

- To tackle that we need control units + other assumptions

Regression Discontinuity (RDD)

Closely related technique, but used in many other contexts

E.g., instead of "Time" we may have "Income"; if above X, eligible for social welfare.

In a RDD analysis you fit **piecewise** **growth-curve** type model such as

$$Y_t = \beta_0 + \beta_1 A_t + \beta_2 \; Time + \beta_3 * Time \; * A_t + e_t$$

In this model the effect of the intervention is parameterized by the change in **level** β_1 and the change in **trend** β_3 after the intervention

Hypothesis tests on these parameters are used as hypothesis tests about the presence / absence of a causal effect

Regression Discontinuity in Practice

```r
fit_rdd <- lm(cigsale ~ year0 + prepost + year0:prepost, prop99_ts)
summary(fit_rdd)
```

```
Call:
lm(formula = cigsale ~ year0 + prepost + year0:prepost, data = prop99_ts)

Residuals:
     Min       1Q   Median       3Q      Max
-10.0953  -2.9216  -0.7423   3.9069   8.1100

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        98.4158     2.4746  39.770  < 2e-16 ***
year0              -1.7795     0.2170  -8.199 8.36e-09 ***
prepostPost       -20.0581     3.7471  -5.353 1.18e-05 ***
year0:prepostPost  -1.4947     0.4846  -3.084  0.00467 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.182 on 27 degrees of freedom
Multiple R-squared:  0.9732,	Adjusted R-squared:  0.9702
F-statistic: 326.4 on 3 and 27 DF,  p-value: < 2.2e-16
```
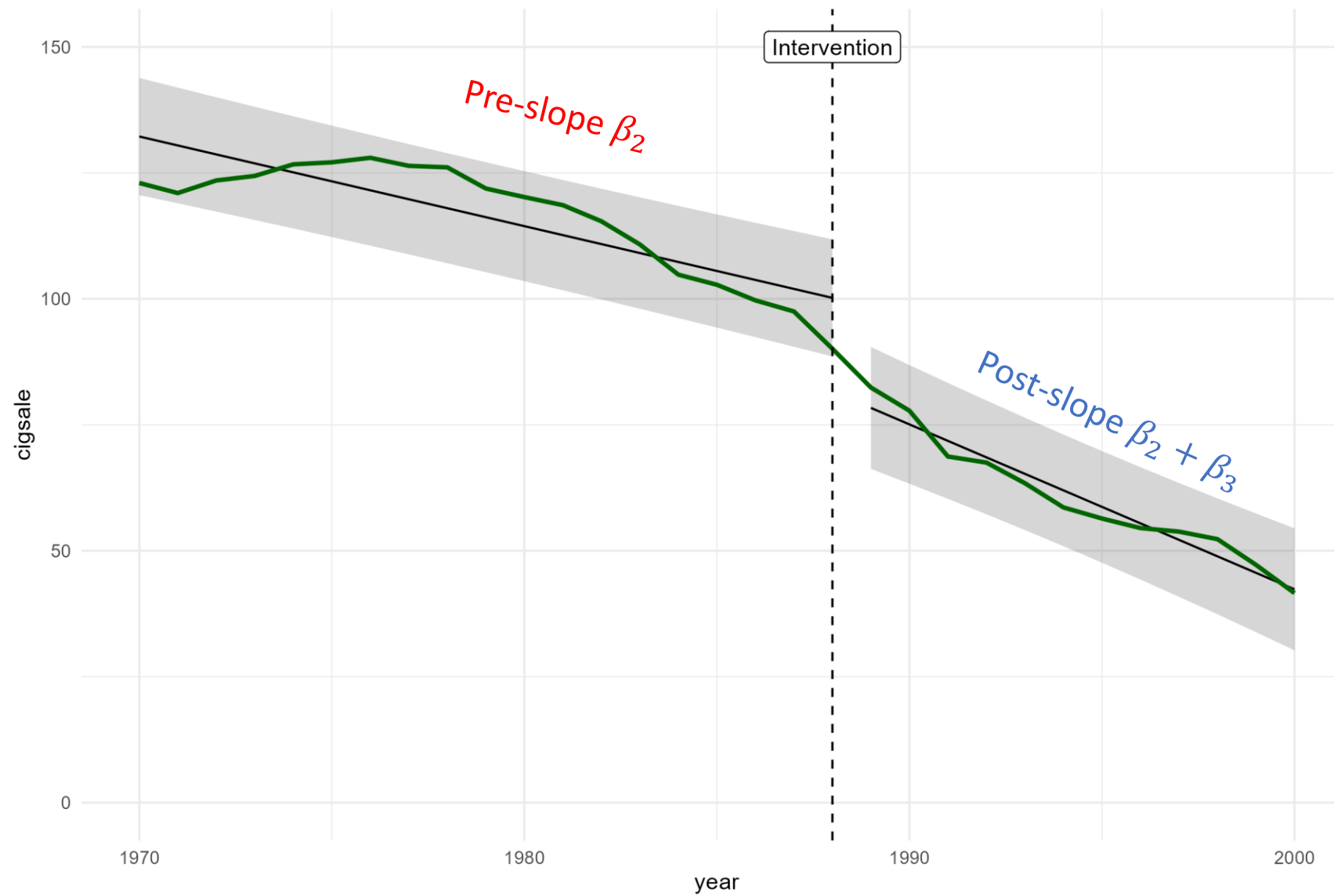
Regression Discontinuity

Basic Idea:

Directly **model** whatever changes you think happen to the target process

- Instead of making forecasts/predictions of the counterfactual directly

Advantages

- More direct. Inference about CE based on significance tests on "change" parameters
- Many extensions and theory to deal with, e.g., "sharp" vs "fuzzy" designs

Disadvantages

- Strongly rely on correct model specification and interpretability; specify *where* or *how* the intervention has an effect

Practical

Work in your groups!

Lunch