Controlled ITS & CausalImpact

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

So far...

Interrupted Time Series

- Suitable when we have long time series, no control units
- Try to predict future **counterfactual** Y_t^0 from past (pre-intervention) data Y_{t-s}^0 from the treated unit

Synthetic Control

- Suitable when we have **many** control units
- Try to predict **counterfactual** Y_t^0 for the treated unit using (a weighted average) of data from other untreated units $C_{j,t}^0$

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

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

So far...

Interrupted Time Series

- Suitable when we have long time series, no control units
- Try to predict future **counterfactual** Y_t^0 from past (pre-intervention) data Y_{t-s}^0 from the treated unit

Synthetic Control

- Suitable when we have **many** control units
- Try to predict **counterfactual** Y_t^0 for the treated unit using (a weighted average) of data from other untreated units $C_{j,t}^0$

Synthetic Control

$Time$	Y_t	A_t	Y_t^0	Y_t^1	C_{1t}^0	C_{2t}^0	…	C_{jt}^0
1	7	0	7	NA	2	9	…	6
2	9	0	9	NA	6	9	…	8
3	6	0	6	NA	4	3	…	5
4	5	0	5	NA	2	1	…	4
5	6	0	6	NA	1	2	…	7
6	2	1	$\widehat{Y_6^0}$	2	3	6	…	7
7	3	1	$\widehat{Y_7^0}$	3	2	5	…	6
8	1	1	$\widehat{Y_8^0}$	1	4	6	…	5
…	…	…	…	…	…	…	…	4
T	2	1	$\widehat{Y_T^0}$	2	3	4	…	6

Estimate Weights

$$Y_t = \sum_{j=1}^{J} \widehat{w_j} C_{jt} \qquad t < T_0$$

$$\widehat{Y_t^0} = \sum_{j=1}^{J} \widehat{w_j} C_{jt} \qquad t > T_0$$

Impute counterfactual

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

This Lecture

Methods which **combine** interrupted time series and (synthetic) control analysis

Try to predict future counterfactual Y_t^0 directly from:

- Pre-intervention data Y_{t-s}^0 **from the treated unit**

AND

- Post-intervention data from **one or more** other **untreated units** C_j^0

Two parts:

1. Controlled Interrupted Time Series (CITS); suitable with relatively few control/covariate time series
2. (Synthetic) CITS with the CausalImpact package; many control time series

# Time-Points \ # Control Units	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

Controlled Interrupted Time Series

An extension of Interrupted Time Series when we have access to one or more **control time series**

Typically observations of the same process, for a different unit, which is **correlated with** (i.e. predictive of) the **target time series**, but which **does not experience the intervention**

Similar criteria as the synthetic control and DiD method “control” units

Basic Idea:

Build a time series / forecasting model, but include control time series as contemporaneous (same-time-moment) predictors

$Time$	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

Controlled Interrupted Time Series

$Time$	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

Fit a forecasting model
C as time-varying predictor

$$\widehat{Y_t} = f(Y_{t-1}, Y_{t-2}, \ldots Y_{t-s}) + \alpha * C_{1t}$$

Controlled Interrupted Time Series

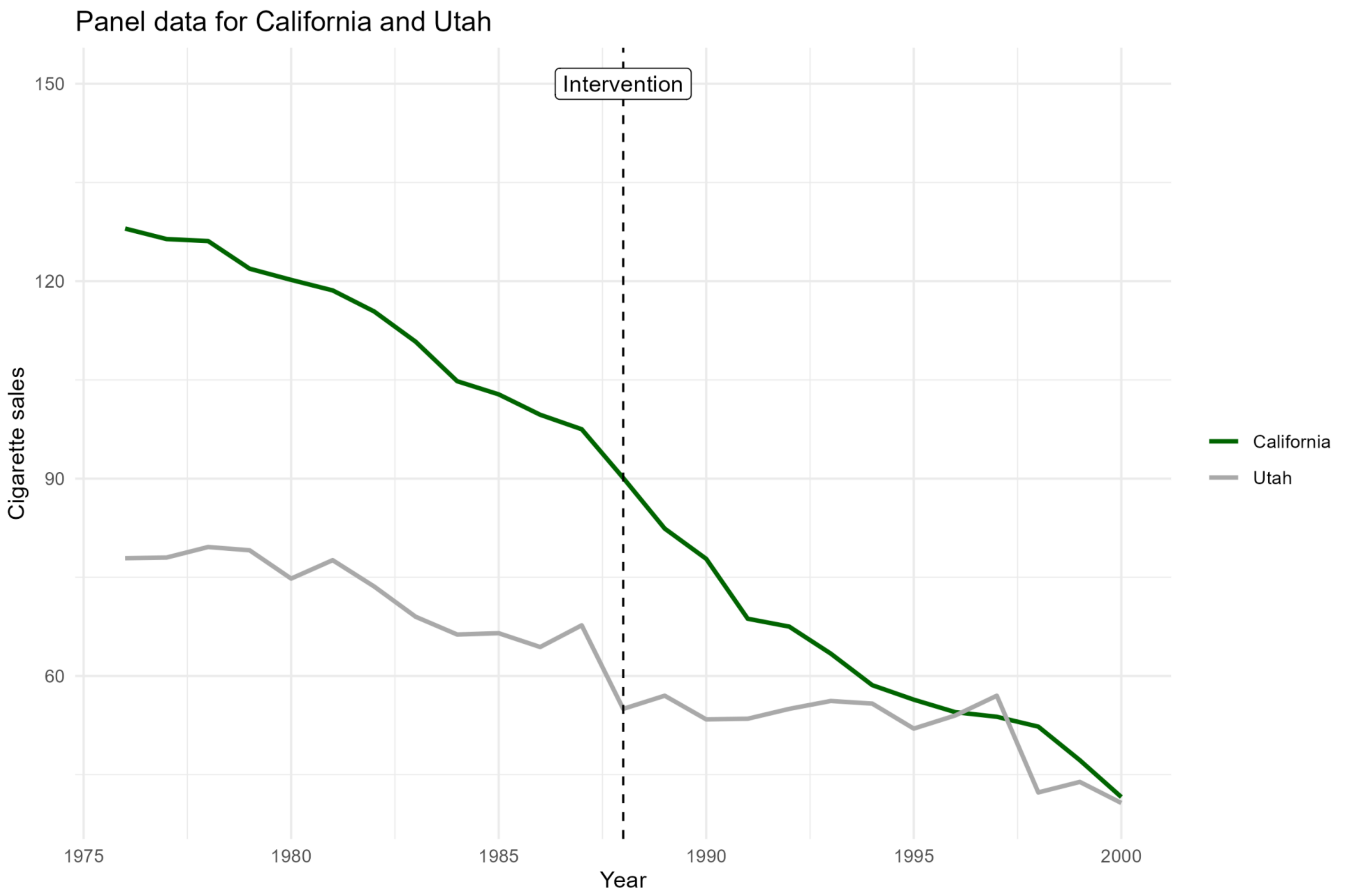
$Time$	Y_t	A_t	Y_t^0	Y_t^1	C_{1t}
1	7	0	7	NA	2
2	9	0	9	NA	6
3	6	0	6	NA	4
4	5	0	5	NA	2
5	6	0	6	NA	1
6	2	1	$\widehat{Y_6^0}$	2	3
7	3	1	$\widehat{Y_7^0}$	3	2
8	1	1	$\widehat{Y_8^0}$	1	4
…	…	…	…	…	…
T	2	1	$\widehat{Y_T^0}$	2	3

Fit a forecasting model
C as time-varying predictor

$$\widehat{Y_t} = f(Y_{t-1}, Y_{t-2}, \dots Y_{t-s}) + \alpha * C_{1t}$$

$$\widehat{Y_t^0} = f(Y_{t-1}, Y_{t-2}, \dots Y_{t-s}) + \alpha * C_{1t}$$

Impute counterfactual



Panel data for California and Utah
Intervention
150
120
90
60
Cigarette sales
1975
1980
1985
1990
1995
2000
Year
California
Utah

Controlled Interrupted Time Series

```r
prop99_ts <-
  prop99_cigonly |>
  select(cigsale_California, cigsale_Utah, cigsale_Illinois) |>
  mutate(years = 1970:2000) |>
  as_tsibble(index = "years")

# divide into pre and post-intervention, as a time
prop99_pre  <- prop99_ts[1:19,]
prop99_post <- prop99_ts[20:31,]

# fit data
fit_arima <-  prop99_pre |>
  model(
    timeseries = ARIMA(cigsale_California), # no regression!
    regression = ARIMA(cigsale_California ~ cigsale_Utah + cigsale_Illinois)
  )
```

Key Assumptions

This method inherits the key assumptions of ITS and Synthetic Control

Do you remember what they are?

1. **No interference:** California receiving treatment does not effect the potential outcome value of Utah
2. **Choose an appropriate time series model (!!!)**
3. **Time-invariant relationships:** Some form of “model invariance” over time

    (i.e. changes are attributable only to the intervention)

    Often phrased as some form of “stationarity” assumption

But what if we have many control time series? **AND** a long time-series pre treatment?

- Many potential states who did not have a law change
- Many different "products" that did not receive a new type of advertisement

Same basic principles apply, however, you may need to be clever in statistical terms:

- allow a **general enough model** to capture complex dependencies,
- try to avoid **overfitting** by keeping the end model simple

CausalImpact

# Time-Points \ # Control Units	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

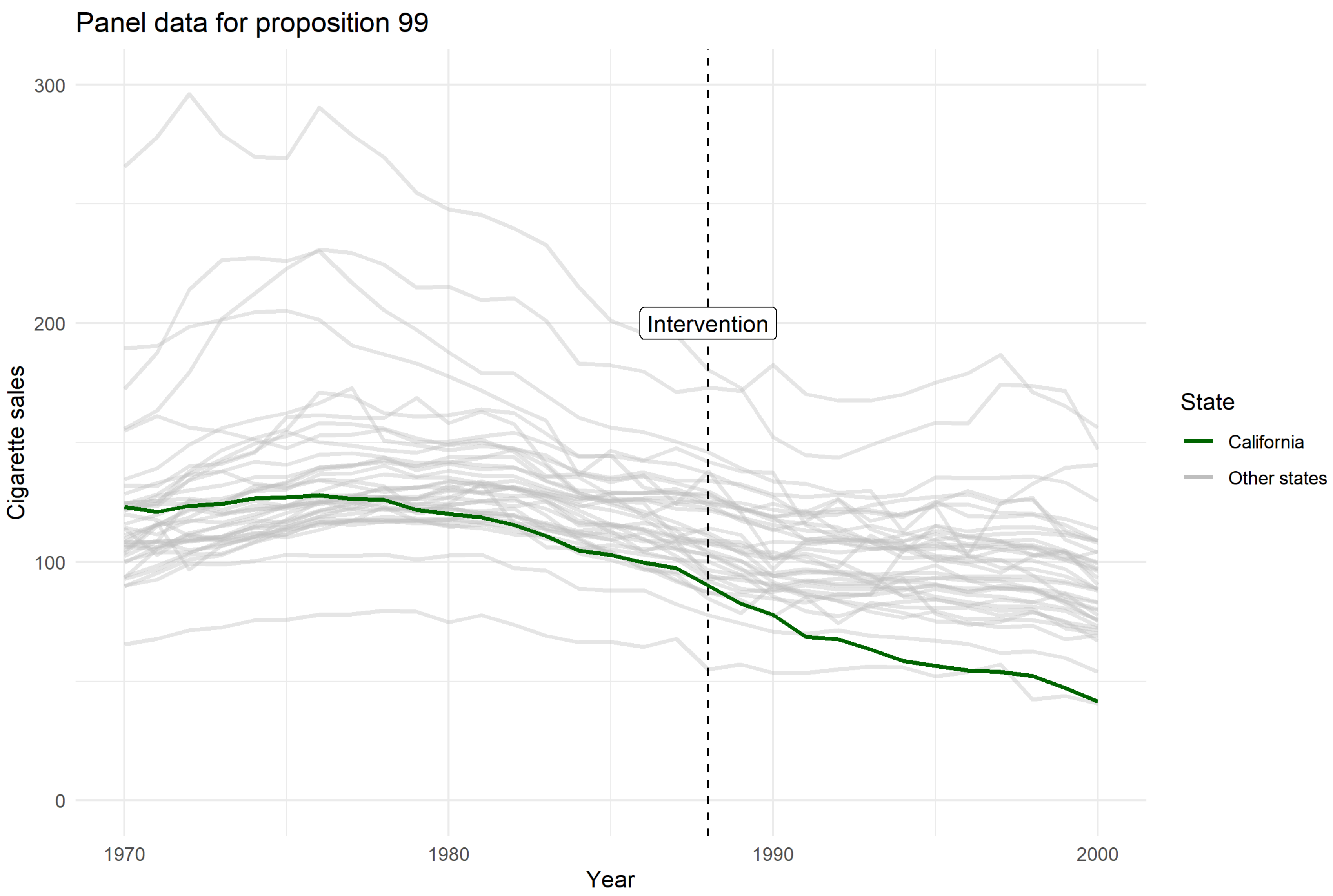
CausalImpact

CausalImpact is an **R** and **python** packaged developed by Google

Performs what <u>could</u> be described as "Synthetic Controlled Interrupted Time Series"

Basic Idea:

- Exactly the same principles as Controlled ITS
- The model has a time-forward forecasting part and a "control unit" regression part
- Behind-the-scenes uses **Bayesian estimation** to build the forecasting model
- A subset of units are included in the control unit part, with different weights; *similar* to a synthetic control analysis (but differing in many other details)
- CausalImpact package takes care of model building + selection behind the scenes (!)



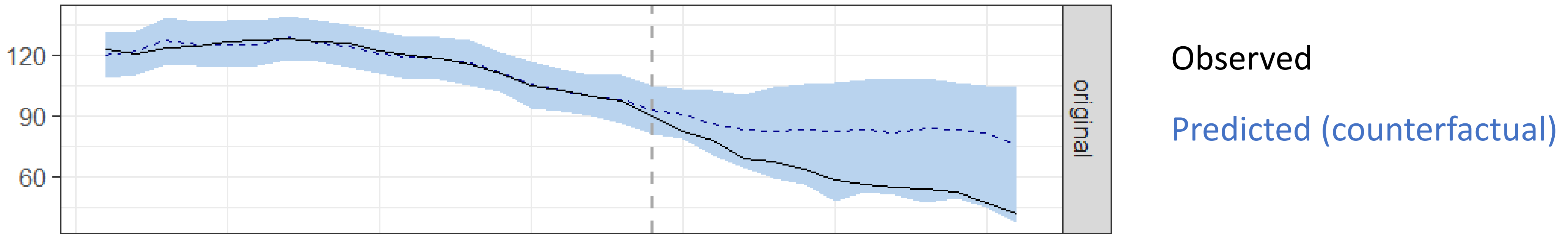
Panel data for proposition 99
Cigarette sales
300
200
100
0
Intervention
1970
1980
1990
2000
Year
State
California
Other states

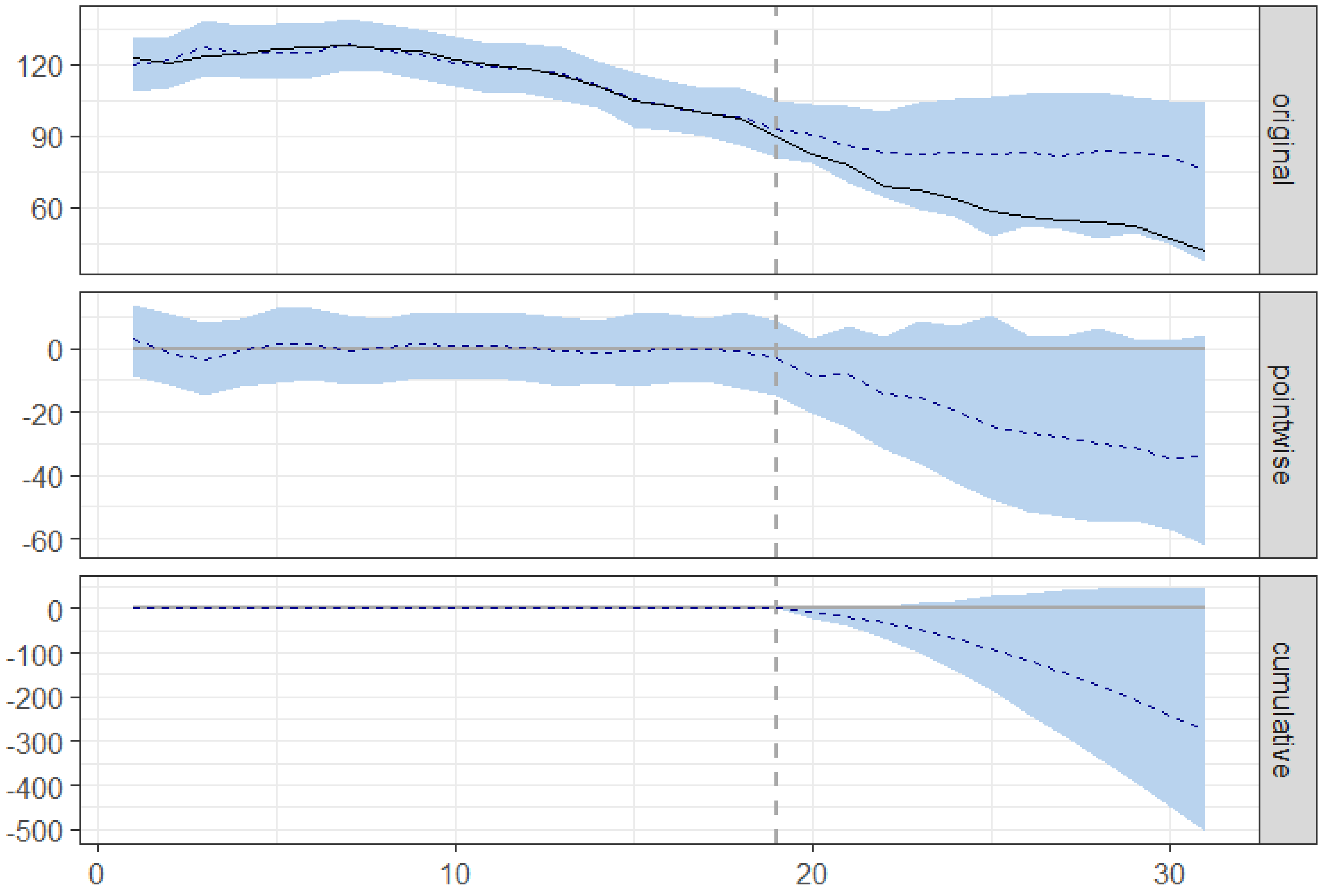
CausalImpact in action

```r
# using only cigarette sales from other states as potential covariates
pre_idx <- c(1, 19) # the first 18 years (1970 - 1988) are pre-intervention
post_idx <- c(20, 31) # the years after that (1989 - 2000)

impact_cigsale <- CausalImpact(
  data = prop99_cigonly,
  pre.period = pre_idx,
  post.period = post_idx
)

# then, plot the causal impact model
plot(impact_cigsale)
```


120
90
60
original
Observed
Predicted (counterfactual)



Observed

Predicted (counterfactual)

Causal Effect estimate at each t

Sum of causal effect estimates at all previous time points

Behind the scenes

CausalImpact uses **Bayesian Estimation**

- Bayesian structural time-series models (*bsts* package in R)
- Control units are “chosen” by using spike-and-slab priors
- Bayesian estimation means it iss easy to quantify **uncertainty** (i.e. get confidence intervals) around estimates of the causal effect, and other interesting metrics related to that

Behind the scenes

Beware: Bayesian estimation requires the user to specify many *priors*

- Controls things like model complexity and which part of the model (forecasting vs control units) will be dominant
- These choices are hidden from the user with **defaults**.
  - This is nice when you want to get something running, but in practice you really need to investigate how sensitive your conclusions are to sometime arbitrary choices, like small changes in these specifications
- In general: the package hides many model specification and selection choices from you. Good for usability, **bad for critical evaluation**. Always check how robust your estimates are!

Practical: fpp3, causalimpact

Work in your groups!
Take a break from 16:15 to 16:30

Break