DIFF-IN-DIFF II

PMAP 8521: Program Evaluation for Public Service

October 21, 2019

Fill out your reading report on iCollege!

PLAN FOR TODAY

Interactions and regression

Diff-in-diff review

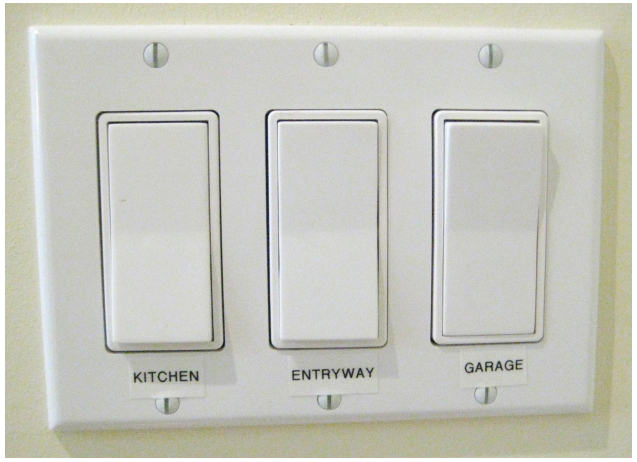
Standard error adjustments

Practice!

INTERACTIONS & REGRESSION

SLIDERS AND SWITCHES

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \epsilon$$



$$\begin{aligned}\hat{\text{happiness}} = \beta_0 &+ \beta_1 \text{Europe} + \beta_2 \text{Latin America} + \\ &\beta_3 \text{MENA} + \beta_4 \text{North America} + \\ &\beta_5 \text{South Asia} + \beta_6 \text{Sub-Saharan Africa} + \epsilon\end{aligned}$$

```
model_life_school_region <-
  lm(happiness_score ~ life_expectancy + school_enrollment + region,
     data = world_happiness)
```

term	estimate	std_error	statistic	p_value
intercept	-2.821	1.355	-2.083	0.04
life_expectancy	0.102	0.017	5.894	0
school_enrollment	0.008	0.01	0.785	0.435
regionEurope & Central Asia	0.031	0.255	0.123	0.902
regionLatin America & Caribbean	0.732	0.294	2.489	0.015
regionMiddle East & North Africa	0.189	0.317	0.597	0.552
regionNorth America	1.114	0.581	1.917	0.058
regionSouth Asia	-0.249	0.45	-0.553	0.582
regionSub-Saharan Africa	0.326	0.407	0.802	0.425

$$\hat{\text{happiness}} = \beta_0 + \beta_1 \text{life expectancy} + \beta_2 \text{school enrollment} +$$
$$\beta_3 \text{Europe} + \beta_4 \text{Latin America} + \beta_5 \text{MENA} +$$
$$\beta_6 \text{North America} + \beta_7 \text{South Asia} + \beta_8 \text{SSA} + \epsilon$$

INDICATORS & INTERACTIONS

Indicators (dummies)

Change in **intercept** for specific group

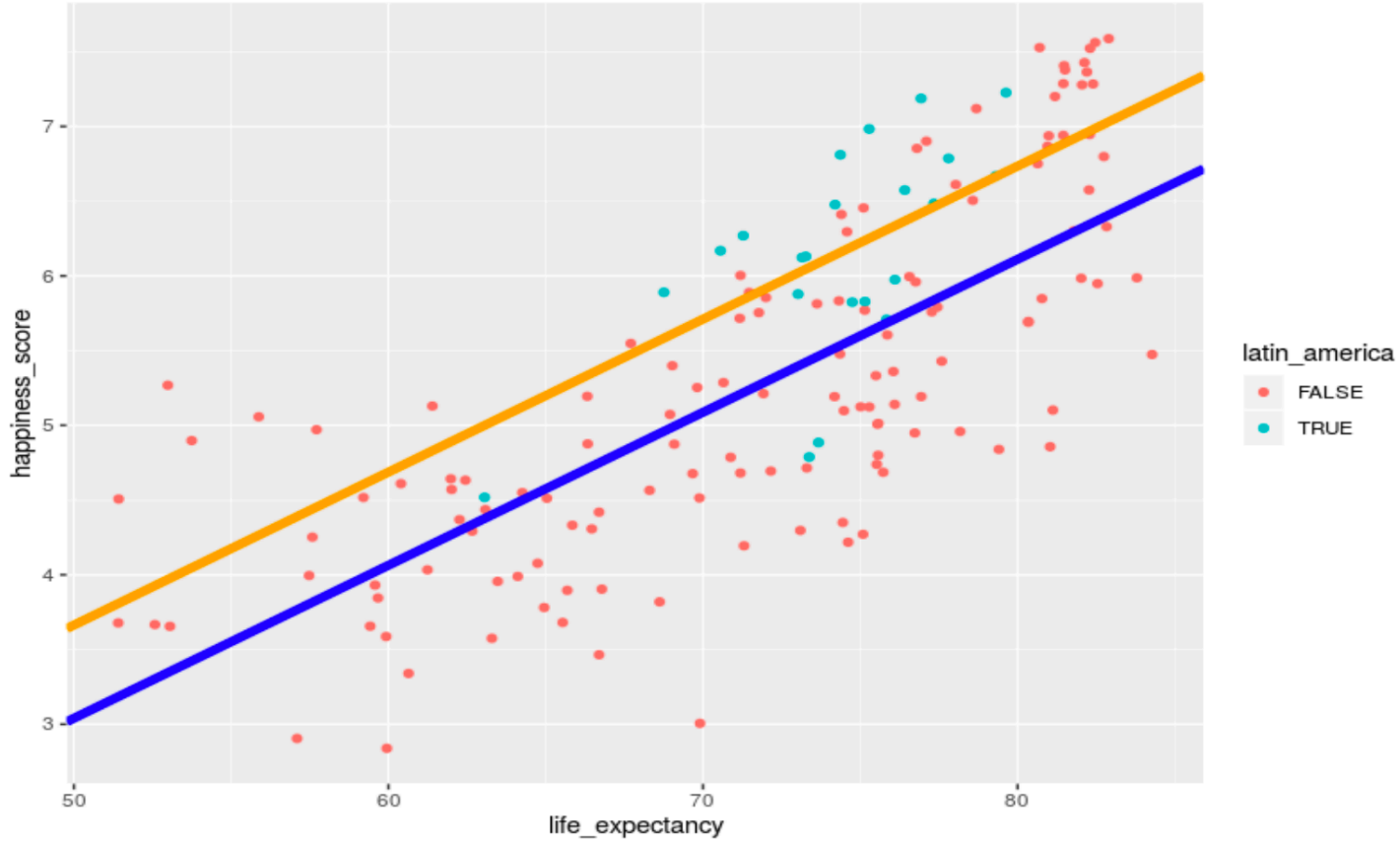
Interactions

Change in **slope** for specific group

```
model_life_la <-
  lm(happiness_score ~ life_expectancy + latin_america, data = world_happiness)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-2.0770858	0.536773852	-3.869573	1.613712e-04
life_expectancy	0.1023494	0.007449708	13.738707	1.954881e-28
latin_americaTRUE	0.6234255	0.172757872	3.608666	4.171373e-04

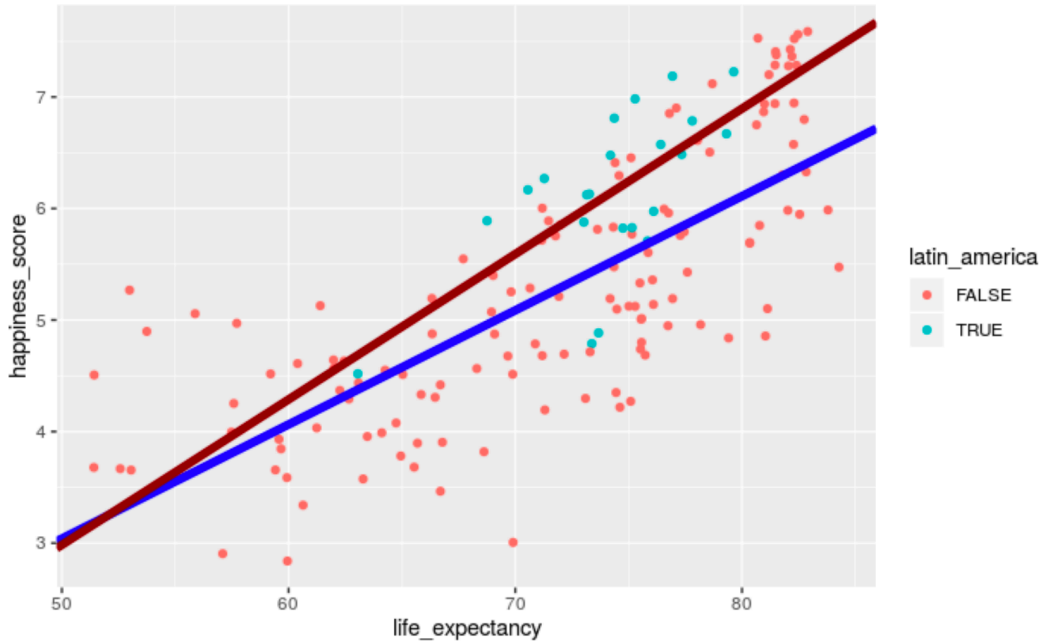
3 rows



```
model_life_la_int <-
  lm(happiness_score ~ life_expectancy + latin_america +
                       (life_expectancy * latin_america), data = world_happiness)
```

term <chr>	estimate <dbl>	std.error <dbl>	statistic <dbl>	p.value <dbl>
(Intercept)	-2.01948544	0.545386030	-3.7028551	2.983292e-04
life_expectancy	0.10154408	0.007570767	13.4126556	1.649813e-27
latin_americaTRUE	-1.51554651	3.364657434	-0.4504311	6.530456e-01
life_expectancy:latin_americaTRUE	0.02884127	0.045307973	0.6365606	5.253749e-01

4 rows



DIFF-IN-DIFF REVIEW

BEFORE VS. AFTER

Average fast food jobs in NJ

Before: 20.44

After: 21.03

Δ: 0.59

Does this show the causal effect?

TREATMENT VS. CONTROL

Average fast food jobs in states

PA_{after}: 21.17

NJ_{after}: 21.03

Δ: −0.14

Does this show the causal effect?

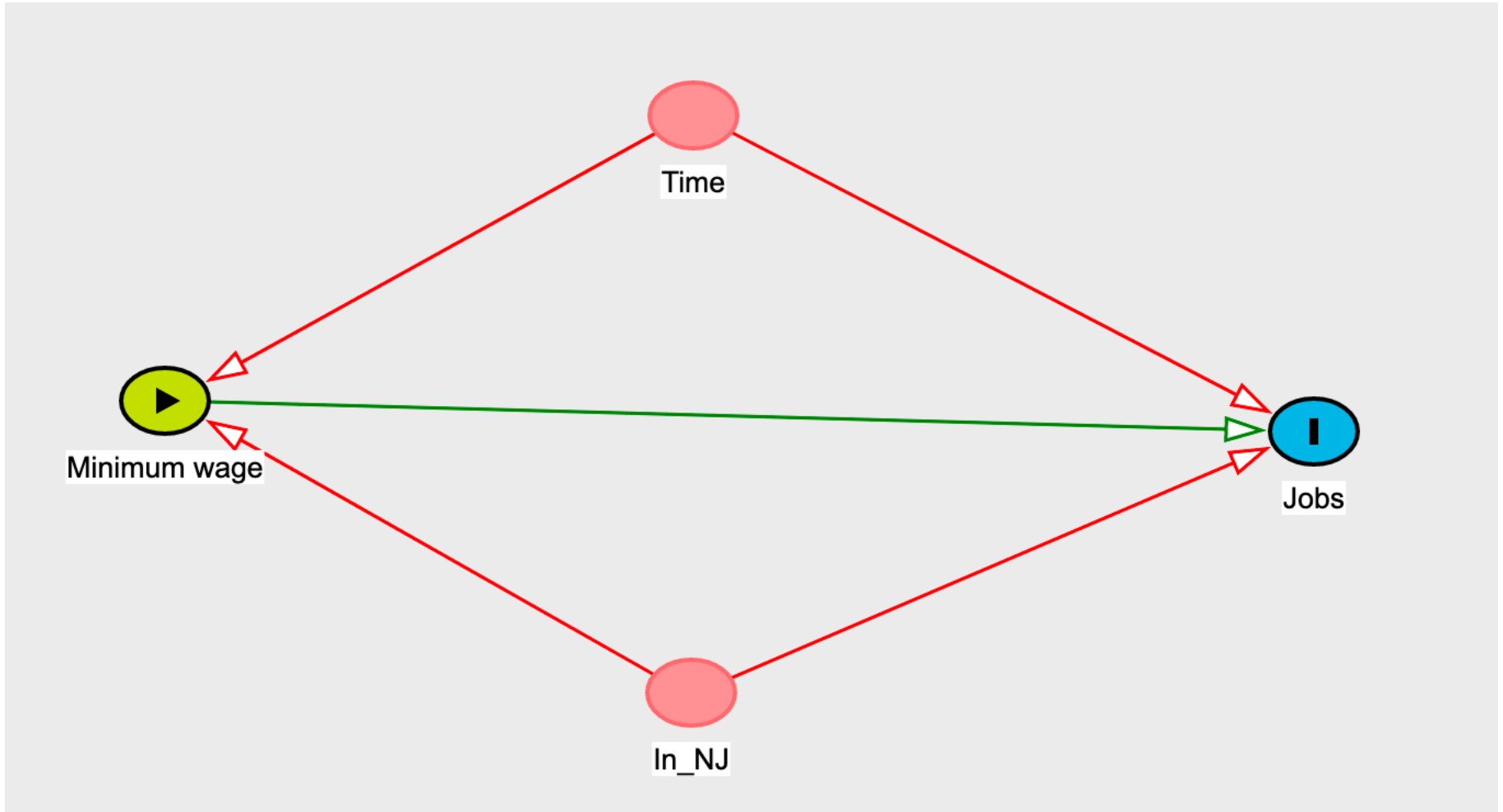
PROBLEMS

Comparing only before/after

Impossible to know if growth happened because of treatment or just naturally

Comparing only treatment/control

Impossible to know if any changes happened because of natural growth


Time
Minimum wage
Jobs
In_NJ

	Pre mean	Post mean
Treatment	A (not yet treated)	B (treated)
Control	C (never treated)	D (never treated)

	Pre mean	Post mean	Δ (post−pre)
Treatment	A (not yet treated)	B (treated)	B−A
Control	C (never treated)	D (never treated)	D−C

Growth!

	Pre mean	Post mean
Treatment	A (not yet treated)	B (treated)
Control	C (never treated)	D (never treated)
Δ (trtmt−ctrl)	A−C	B−D

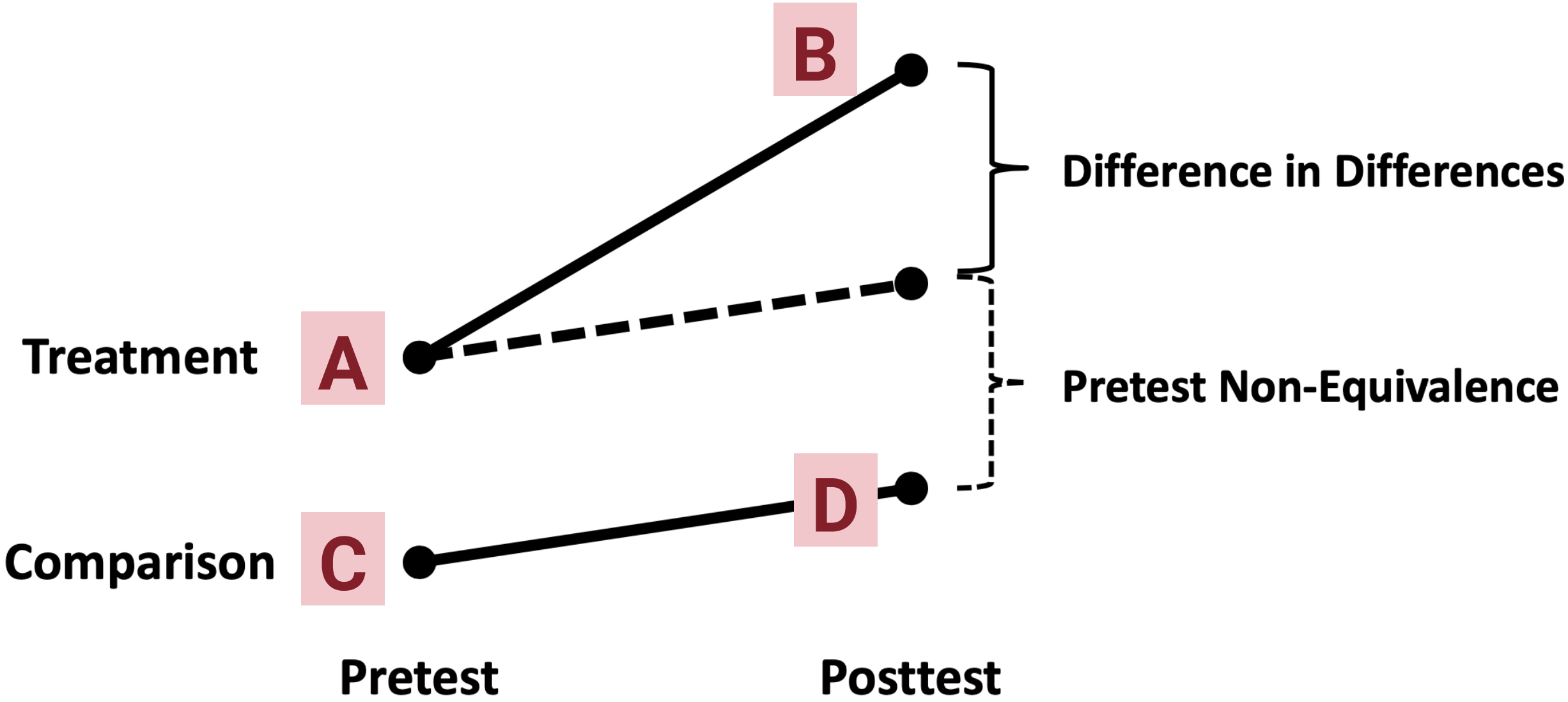
Within-group effects

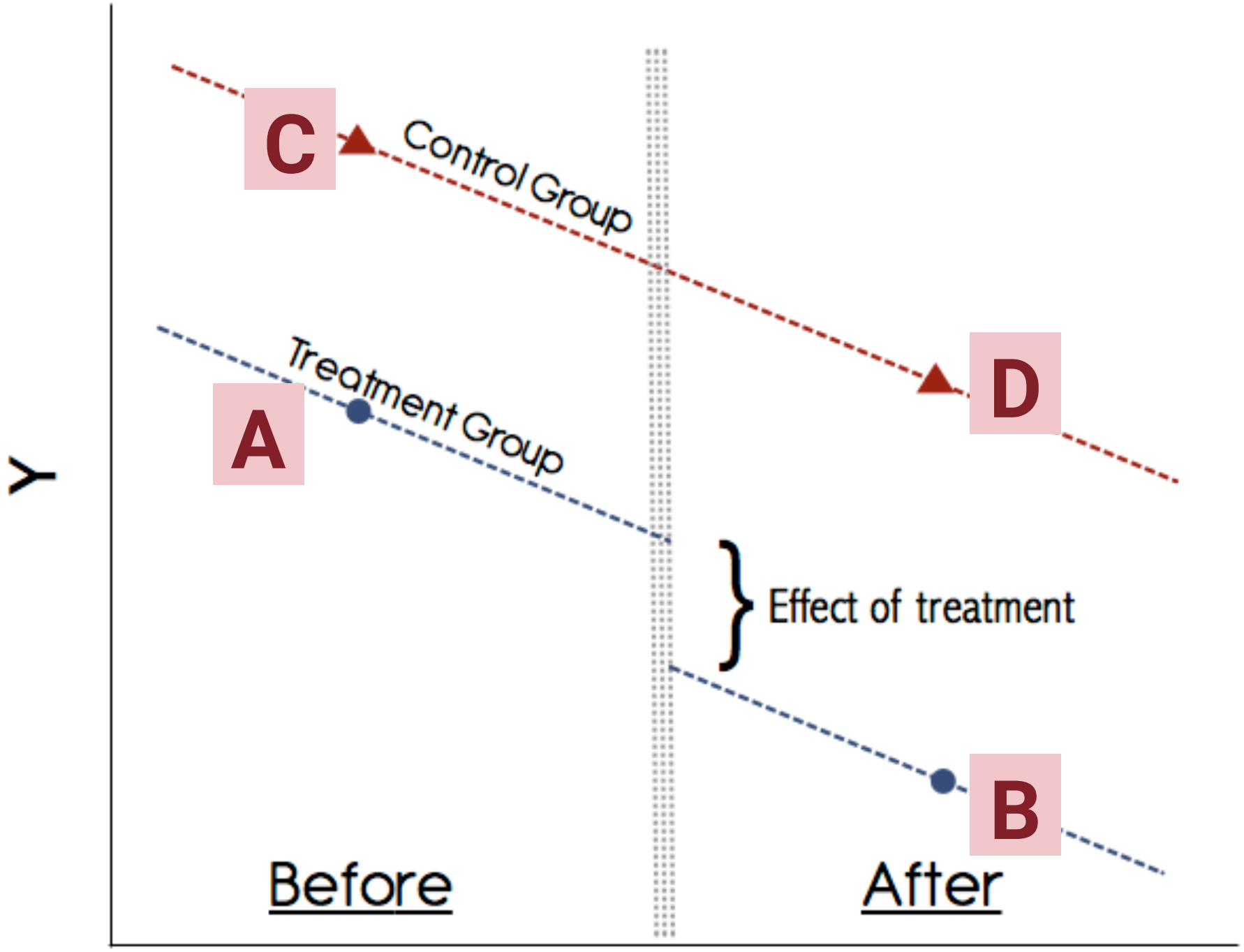
	Pre mean	Post mean	Δ (post−pre)
Treatment	A (not yet treated)	B (treated)	B−A
Control	C (never treated)	D (never treated)	D−C
Δ (trtmt−ctrl)	A−C	B−D	(B−A) − (D−C)

Growth of treatment − growth of control (DiD!)

$$\begin{aligned}\text{DD} =& (\bar{x}_{\text{treatment, post}} - \bar{x}_{\text{treatment, pre}}) \\ & - (\bar{x}_{\text{control, post}} - \bar{x}_{\text{control, pre}})\end{aligned}$$

	Pre mean	Post mean	Δ (post−pre)
NJ	A 20.44	B 21.03	B−A 0.59
PA	C 23.33	D 21.17	D−C −2.16
Δ (trtmt−ctrl)	A−C −2.89	B−D −0.14	(0.59) − (−2.16) = 2.76


B
Difference in Differences
Treatment
A
Pretest Non-Equivalence
D
Comparison
C
Pretest
Posttest


C
Control Group
D
Treatment Group
A
Y
Effect of treatment
B
Before
After

Finding all the group means is tedious though!

What if there are other backdoors to worry about?

Regression to the rescue!

$$Y_{it} = \alpha + \beta\ \text{Group}_i + \gamma\ \text{Time}_t + \delta\ (\text{Group}_i \times \text{Time}_t) + \epsilon_{it}$$

```
model <- lm(outcome ~ group + time + group * time)
```

Group = 1/TRUE if treatment

Time = 1/TRUE if after

$$Y_{it} = \alpha + \beta\ \text{Group}_i + \gamma\ \text{Time}_t + \delta\ (\text{Group}_i \times \text{Time}_t) + \epsilon_{it}$$

```
model <- lm(outcome ~ group + time + group * time)
```

α = Mean of control, pre-treatment

β = Increase in outcome across groups

γ = Increase in outcome across time

δ = Difference in differences!

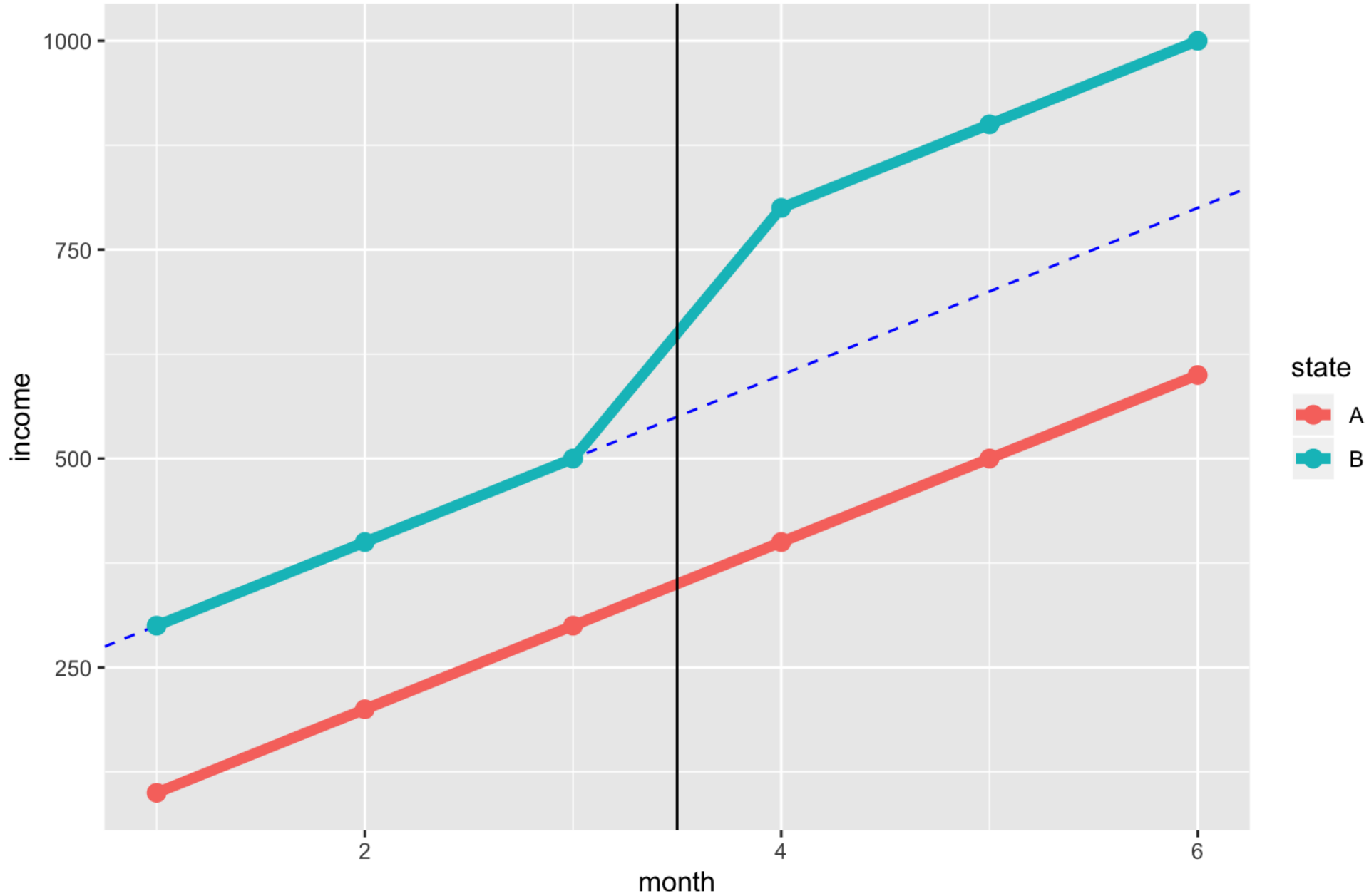
$$Y_{it} = \alpha + \beta\ \text{Group}_i + \gamma\ \text{Time}_t + \delta\ (\text{Group}_i \times \text{Time}_t) + \epsilon_{it}$$

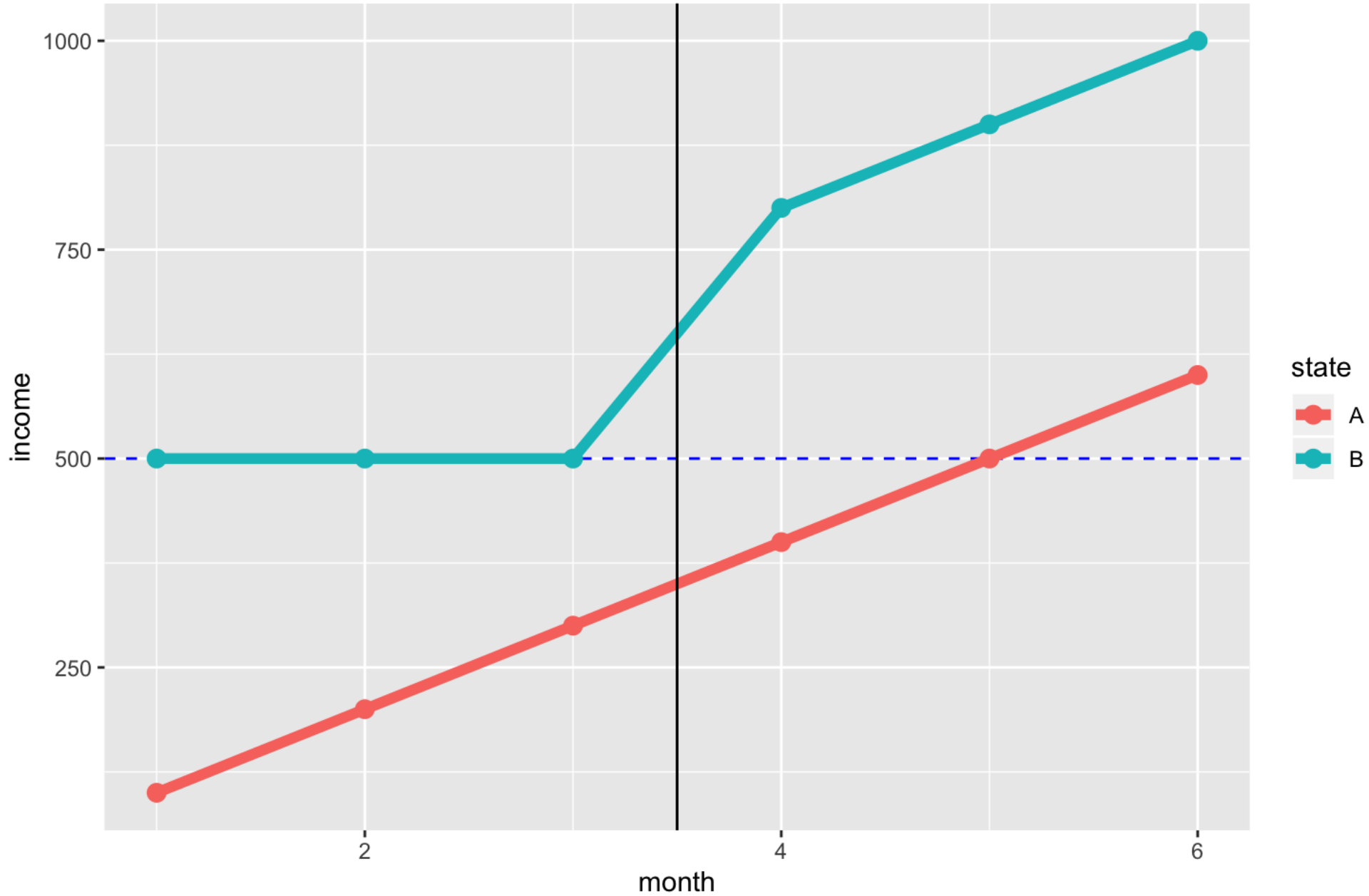
	Pre mean	Post mean	Δ (post−pre)
Treatment	α	α + γ	γ
Control	α + β	α + β + γ + δ	γ + δ
Δ (trtmt−ctrl)	β	β + δ	δ

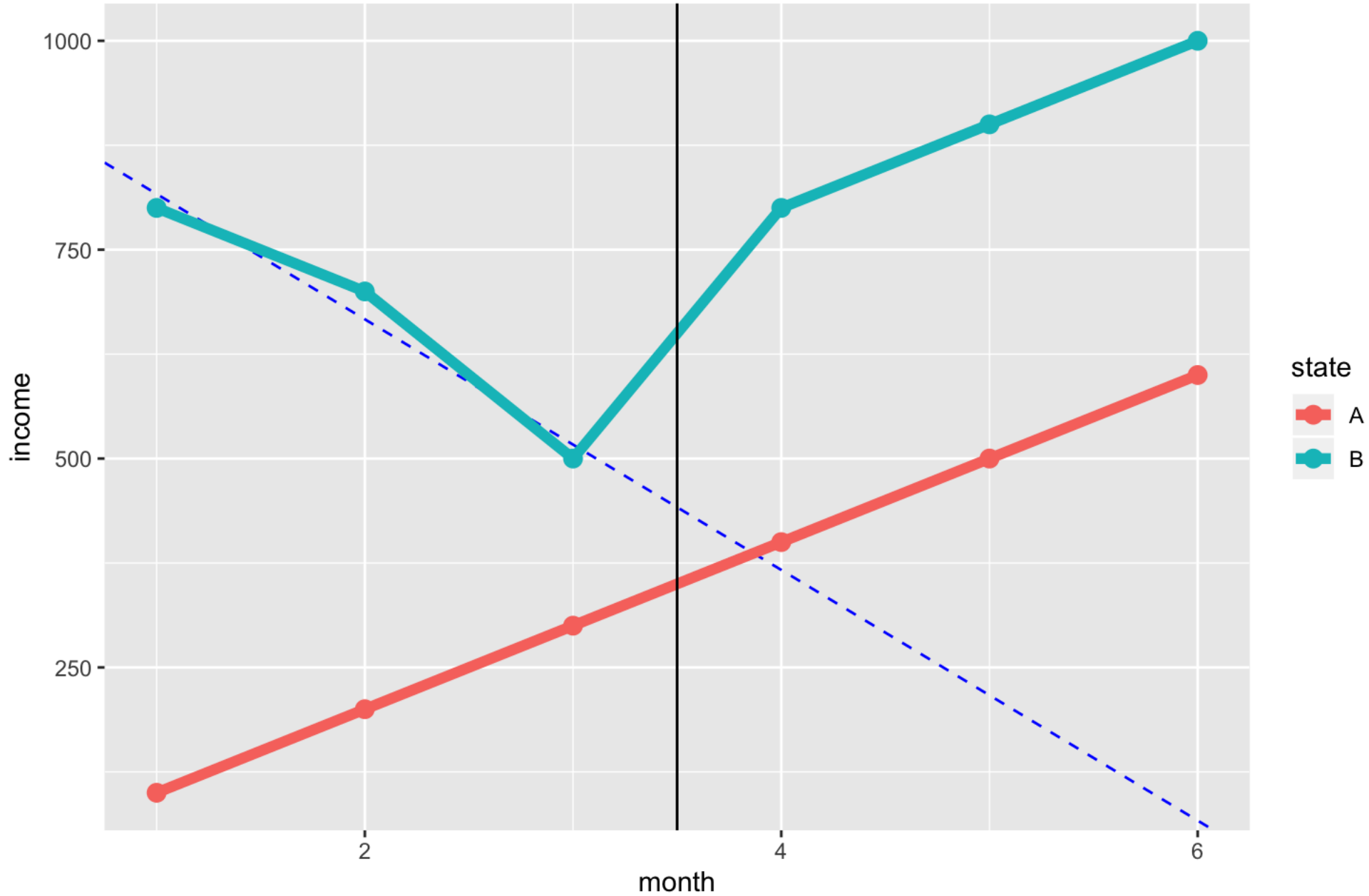
ASSUMPTIONS

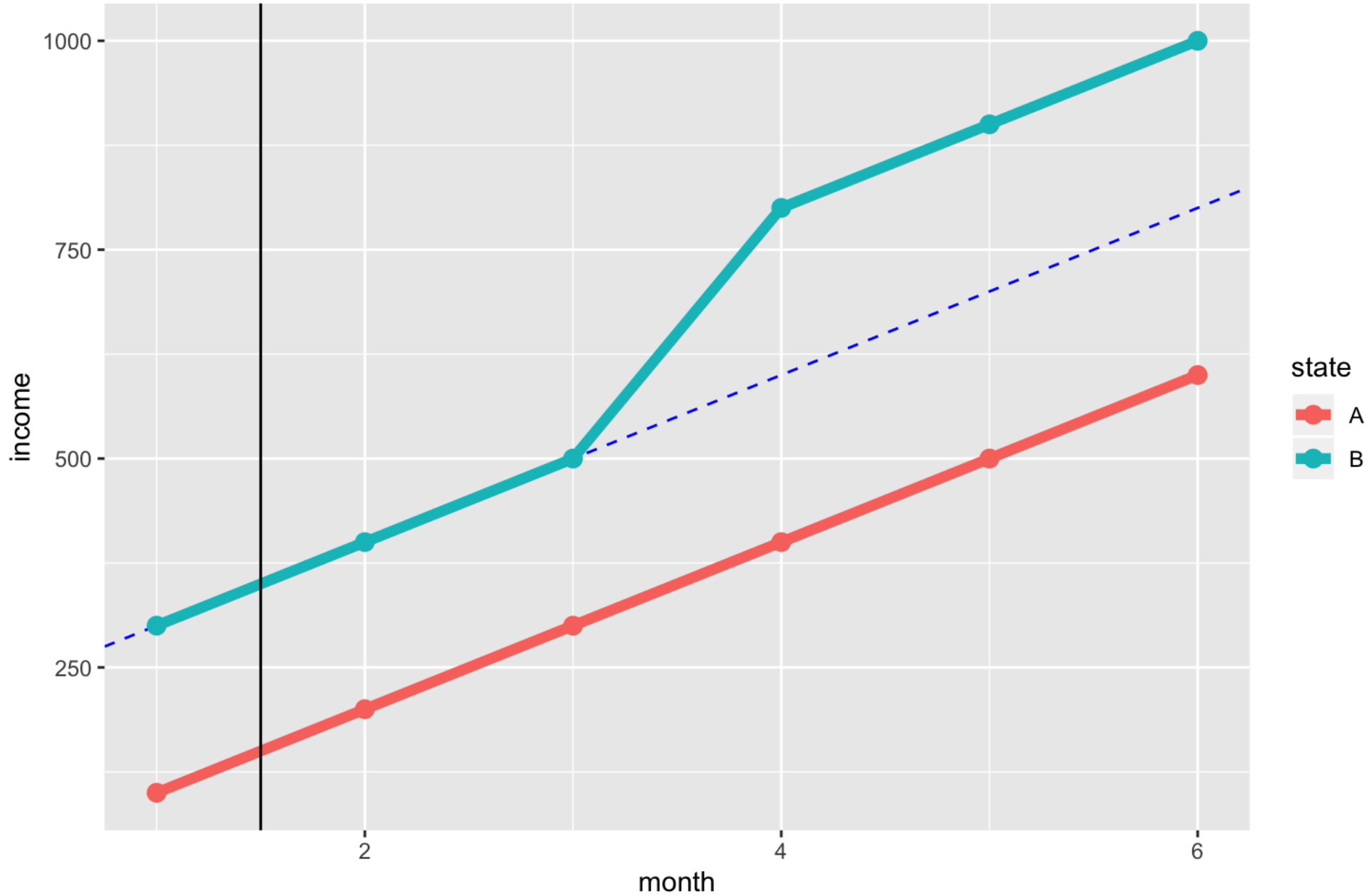
Parallel trends

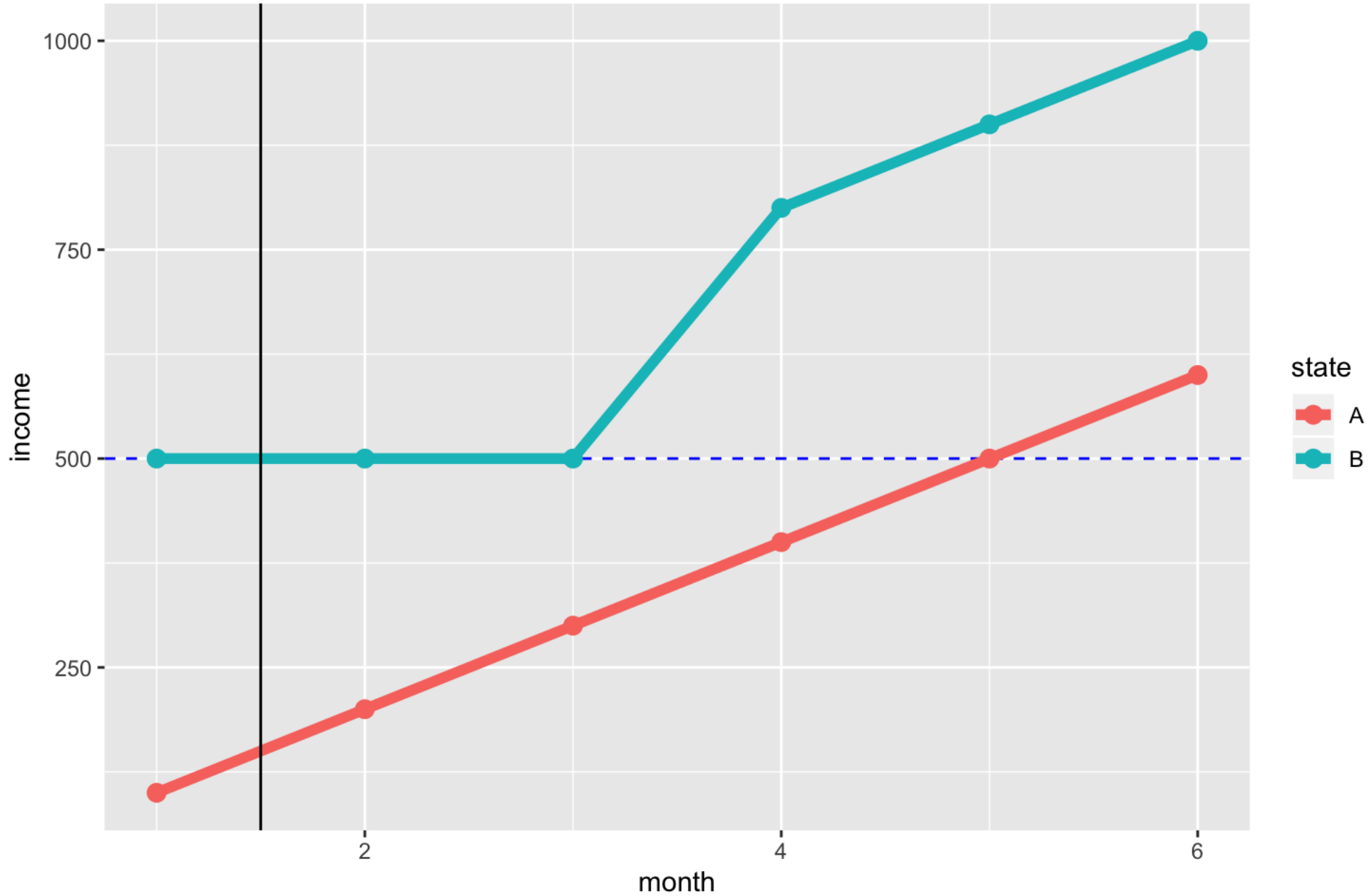
Treatment and control might have different values at first, but we assume treatment group would have changed like control in absence of treatment


1000
750
500
250
income
2
4
6
month
state
A
B


1000
750
500
250
income
2
4
6
month
state
A
B


1000
750
500
250
income
2
4
6
month
state
A
B


1000
750
500
250
income
2
4
6
month
state
A
B


1000
750
500
250
income
2
4
6
month
state
A
B

STANDARD ERROR ADJUSTMENTS

Diff-in-diff gives you an estimate (coefficient), but that number has nothing to do with inference

Is the number statistically significant?

Clusters can mess up your standard errors (and significance)!

Test scores are measured at student level, but treatment varies by classroom

Mosquito net use is measured at individual level, but treatment varies by village

Adjust your standard errors by accounting for clustering

```
lm(y ~ treatment + time + treatment * time,
   data = data)
```

```
library(estimatr)
lm_robust(y ~ treatment + time +
                treatment * time,
          data = data,
          clusters = state)
```