Differences-in-differences (DiD)

Dummy-dummy interactions can be used for something called "Differences-in-differences" (DiD) estimation.

Example: increasing the minimum wage (image by Stable Diffusion)

- In 1992, New Jersey's minimum wage rose from \$4.25 to \$5.05 per hour.
- Card and Krueger (1994) surveyed 410 fast-food restaurants before and after the increase, and asked about things like the number of employees.

Download Card and Krueger data:

```
did <- read.csv("https://rtgodwin.com/data/card.csv")
```

Some variables to look at for now:

`EMP` – number of full-time employees

`TIME` – a dummy equal to 0 for before the wage increase, 1 for after the increase

`STATE` – a dummy equal to 0 for Pennsylvania, equal to 1 for New Jersey

Difference in the number of employees before and after the wage increase:

```
mean(did$EMP[did$STATE == 1 & did$TIME == 1]) -
      mean(did$EMP[did$STATE == 1 & did$TIME == 0])
[1] 0.4666667
```

The difference is not significant:

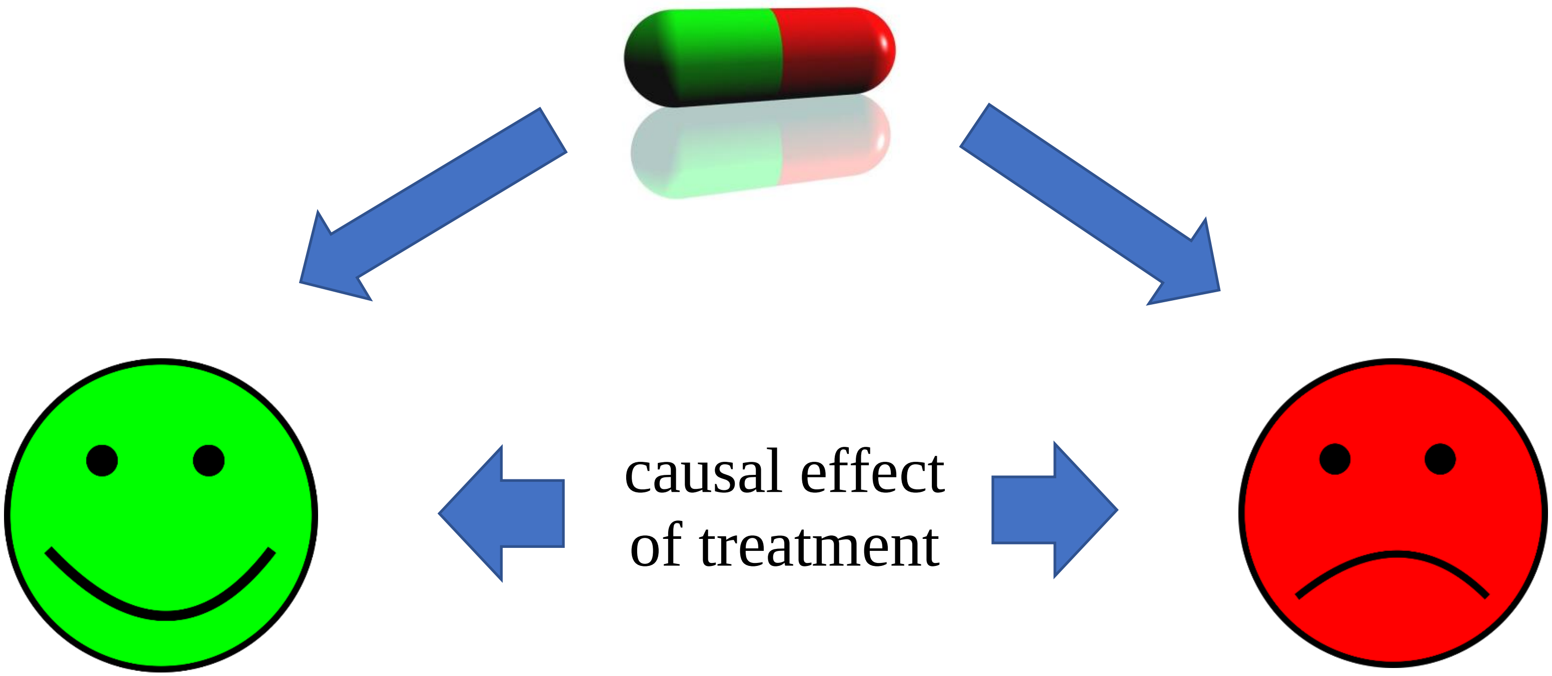
```
dids <- subset(did, STATE==1)
summary(lm(EMP ~ TIME, data=dids))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.4306     0.5289  38.627   <2e-16 ***
TIME          0.4667     0.7480   0.624    0.533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.298 on 616 degrees of freedom
Multiple R-squared:  0.0006315,   Adjusted R-squared:  -0.0009909
F-statistic: 0.3892 on 1 and 616 DF,  p-value: 0.5329
```

So, the causal effect of the increase in minimum wage on employment is estimated to be an increase of 0.47 workers on average, but this increase is not statistically significant.

What is the problem with calling this a "causal effect"?

Next: "The Fundamental Problem of Causal Inference"

<u>Fundamental problem of causal inference</u>



Potential outcome under treatment

Potential outcome under no treatment

Suppose we want to know the *difference* that a cause (treatment) makes.

That is, we want to know:

$$E[y_1 - y_0]$$

- y_1 – outcome with treatment
- y_0 – outcome without treatment

Treatment is broadly defined:

- Treatment with a drug - (y_1 and y_0 blood pressure with/without the drug)
- Addictions treatment (methadone) – (y_1 and y_0 probability of success)
- Health insurance - (y_1 and y_0 the number of visits to the doctor with or without insurance)
- Education (y_1 and y_0 the wage with/without an education)
- Job training
- Monetary policy
- Student debt
- Information
- **Increase in minimum wage** (y_1 and y_0 the employment rate)

Fundamental problem of causal inference

Because an "individual" can't be in both states (treated and untreated), we can't observe both y_1 and y_0.

We can never observe a causal effect!

- One of the two outcomes will occur, and is factual.
- The other outcome(s) is imagined, or counterfactual.
- We only ever observe either y_1 **or** y_0.

Maybe we could observe a causal effect?

Wooldridge calls it a problem of "missing data".

How could we observe the missing data?

- Time travel
- Parallel universe

Barring the above, we have to think in *counterfactuals* and try to find ways to estimate what the unobserved outcome (y_1 or y_0) would have looked like so that we can calculate $y_1 - y_0$.

Estimation of a causal effect

Unit	Treated:	Outcome under treatment y_1	Outcome under no treatment y_0
1	yes	✓	?
2	yes	✓	?
3	no	?	✓
4	no	?	✓

causal effect estimate

Back to minimum wage example

EMP (y)	number of full-time employees
TIME	**0** for **before** the wage increase
	1 for **after** the increase
STATE	**0** for Pennsylvania (no wage increase – "control")
	1 for New Jersey (wage increase – "treatment")

The naïve approach is to take the difference between New Jersey's employment before and after the wage increase:

$$\bar{y}_{at\ TIME=1} - \bar{y}_{at\ TIME=0} = 0.4667$$

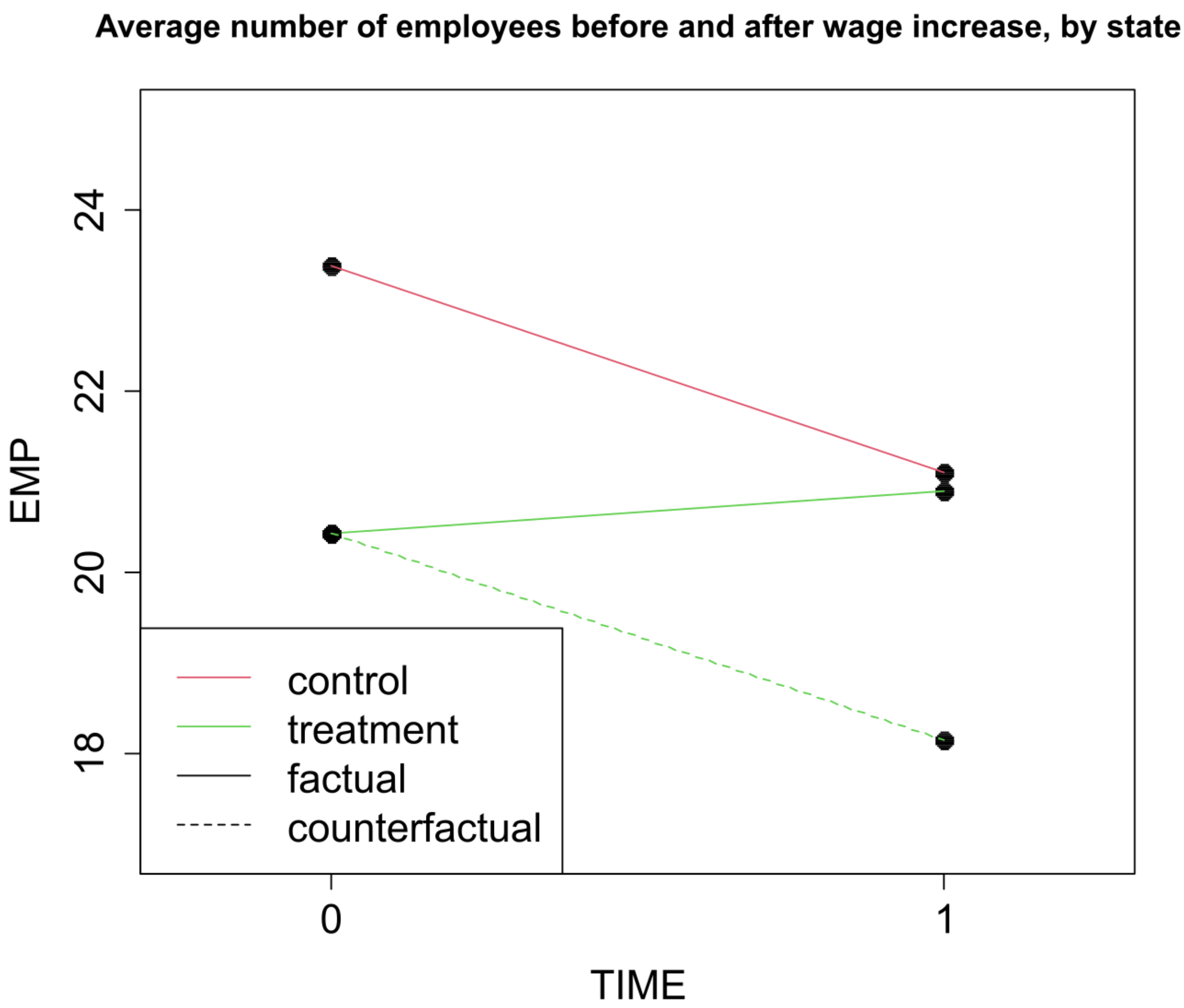
But for this to be the causal effect, need to assume that the level of employment would have stayed constant over the 6 months!

```
dids <- subset(did, STATE==1)
summary(lm(EMP ~ TIME, data=dids))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  20.4306     0.5289  38.627   <2e-16 ***
TIME          0.4667     0.7480   0.624    0.533
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.298 on 616 degrees of freedom
Multiple R-squared:  0.0006315,   Adjusted R-squared:  -0.0009909
F-statistic: 0.3892 on 1 and 616 DF,  p-value: 0.5329
```

Table 1: Average employment by STATE and TIME

	TIME = 0	TIME = 1	Difference
New Jersey STATE = 1 (treatment)	20.431	20.897	0.466
Pennsylvania STATE = 0 (control)	23.380	21.096	-2.283
Difference	-2.949	-0.199	**2.750**

- Parallel trends assumption: the *difference* in employment that occurred for the control group would have also occurred for the treatment group (if they hadn't have been treated): -2.283
- The *difference* in employment that actually did occur under treatment was 0.466
- The *difference-in-difference* is 0.466 – (-2.283) = **2.750**


Average number of employees before and after wage increase, by state
EMP
24
22
20
18
TIME
0
1
control
treatment
factual
counterfactual

We can get the DiD estimator by differencing the sample means between groups. But often, we want to include other "X" variables in the model in order to avoid OVB. If we estimate the model:

$$EMP = \beta_0 + \beta_1 TIME + \beta_2 STATE + \beta_3 (TIME \times STATE) + \epsilon$$

Then b_3 is the DiD estimator!

- Other "X" variables can be added to the model
- $TIME \times STATE$ is an **interaction term**
- β_1 is the effect of $TIME$ for the control group
- β_2 is the difference in EMP at $TIME = 0$
- β_3 is the difference in the effect of $TIME$ between the two groups

$$EMP = \beta_0 + \beta_1 TIME + \beta_2 STATE + \beta_3 (TIME \times STATE) + \epsilon$$

Plug in values for the dummies to get the interpretation of the β:

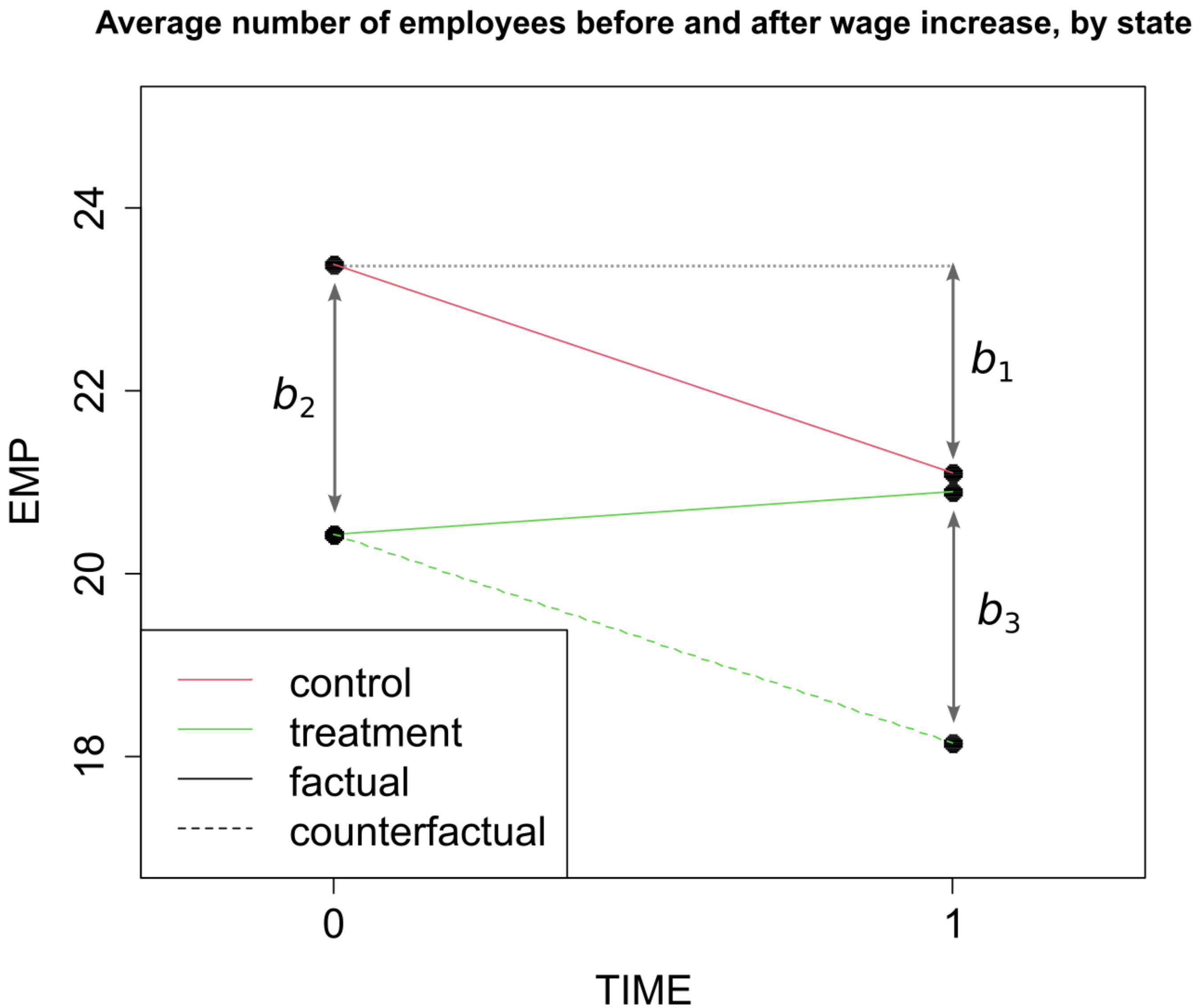
<table>
<tr><th>$TIME$</th><th>$STATE$</th><th>EMP</th><th>difference</th></tr>
<tr><td>0</td><td>0</td><td>β_0</td><td rowspan="2">β_1
(for control)</td></tr>
<tr><td>1</td><td>0</td><td>$\beta_0 + \beta_1$</td></tr>
<tr><td>0</td><td>1</td><td>$\beta_0 + \beta_2$</td><td rowspan="2">$\beta_1 + \beta_3$
(for treatment)</td></tr>
<tr><td>1</td><td>1</td><td>$\beta_0 + \beta_1 + \beta_2 + \beta_3$</td></tr>
</table>

Difference over time for control: β_1

Difference over time for treatment: $\beta_1 + \beta_3$

Difference-in-difference: $(\boldsymbol{\beta_1} + \boldsymbol{\beta_3}) - \boldsymbol{\beta_1} = \boldsymbol{\beta_3}$

```
summary(lm(EMP ~ TIME + STATE + I(TIME * STATE), data = did))

Coefficients:
                Estimate Std. Error t value Pr(>|t|)    
(Intercept)       23.380      1.098  21.288   <2e-16 ***
TIME              -2.283      1.553  -1.470   0.1419    
STATE             -2.949      1.224  -2.409   0.0162 *  
I(TIME * STATE)    2.750      1.731   1.588   0.1126    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.511 on 764 degrees of freedom
Multiple R-squared:  0.007587,   Adjusted R-squared:  0.00369
F-statistic: 1.947 on 3 and 764 DF,  p-value: 0.1206
```


Average number of employees before and after wage increase, by state
EMP
24
22
20
18
TIME
0
1
b_1
b_2
b_3
control
treatment
factual
counterfactual