Causality & Difference in Differences

Social Media and Web Analytics

Lachlan Deer
Tilburg University
Updated: 2022-04-14

Learning Goals for this Week

- Explain the the difference between correlation and causation
- Understand the difference between regression assumptions and causal assumptions
- Explain the terms Randomized Control Trial and Natural / Quasi Experiment
- Define the term 'Difference in Differences'
- Estimate treatment effects using Difference in Differences
- Reflect on assumptions underlying causal claims from Difference in Difference estimates

Causality

Why Causality?

- Many questions we want answers to are **causal**
- When we talk about marketing, we often want to know why something happens
  - Did demand/revenue/... change because of ?
  - And by how much?
- We also care about non-causal questions (prediction, descriptive evidence)
  - But our comparative advantage should be causality

Why Causality as a Marketing Analyst?

- Causality should be a marketing analyst's **comparative advantage**
  - Plenty of fields do statistics, many probably do it better
  - Few fields worry about causality and the *why* questions the way we (should) do
- We can design more effective marketing strategies if we can identify causal effects
  - Which will generate a boost in KPIs
- **Skill to acquire**: Understanding when to make causal claims and when not
  - Your value to a future employer sky rockets if you can do this well

What is Causality?

X causes Y if …

- We intervene and change X and nothing else
- Then Y changes as a result

Examples of Causal Relationships

Obvious:

- Turning on a light switch causes a light to be on
- Fireworks raise the noise level

Not so obvious:

- TV Advertising increases product demand
- Tweets about movies increase demand for it at theatres

Remark: The **size** these effects are **much smaller** than you probably think

Examples of Non-Causal Relationships

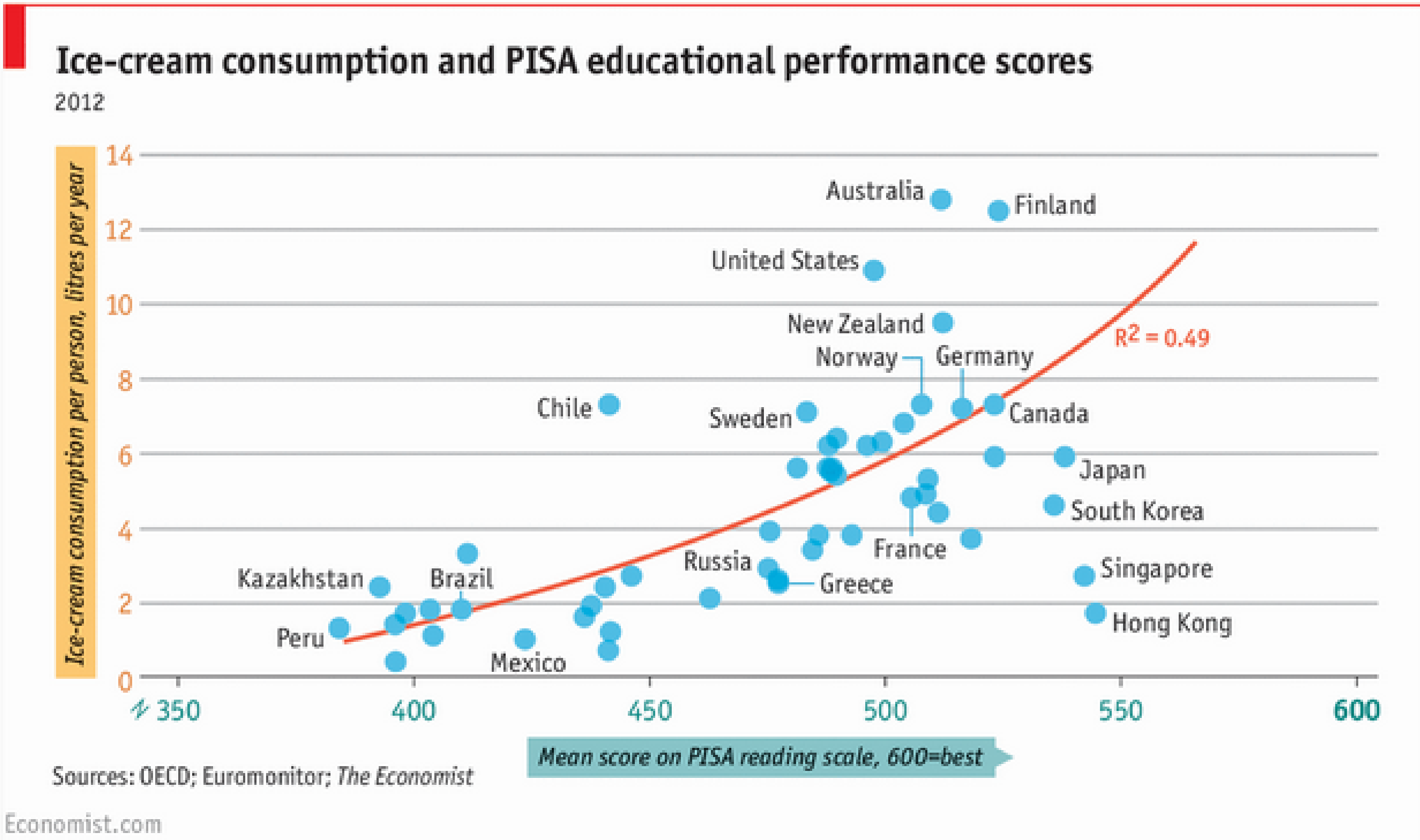
Obvious:

- Number of people wearing shorts at the beach and ice cream consumption
- Roosters crowing followed by sunrise

Some not so obvious:

- School vending machines and obesity
- Search engine advertising and revenue (in the short term!)

Correlation is not Causation



Why Correlation is not Causation

(Some) possible reasons **A** might not cause **B**:

- **The opposite is true**
  - B actually causes A
- The two are correlated, but **there's more to it**:
  - A and B are correlated, but they're actually caused by C
- There's **another variable involved**:
  - A does cause B as long as D happens
- There is a **"chain" reaction**:
  - A causes E, which leads E to cause B
  - ... but you only saw that A causes B from your own eyes
- It's due to **chance**

The Difficulty of Causal Inference

Can we tell when correlation $\implies$ causation?

- Answer 1: It's *hard*
- Answer 2: It is possible, but we *need assumptions*

What kind of assumptions?

- "What would have beens" - i.e. (approximate) counterfactual outcomes
- "As good as random" - i.e. no selection on unobservables
  - Known as "conditional independence"
  - Intuition: Given some control variables, differences in variable we care about are only due to randomness
  - No unobserved factors driving variation in variable of interest

Even then:

- At *best* we'll estimate an **average causal effect**

Regression and Causality

Regression assumptions on their own

$$\neq \quad \textbf{causal interpretations of } \beta$$

- **Regression assumptions**: Unbiasedness, Variance of estimates
- "**Causal Inference assumptions**": Can an unbiased estimate be interpreted causally
  1. Valid counterfactual outcomes
  2. Conditional independence

Note: Cannot test these assumptions 'statistically'

Experiments in Marketing Analytics

Recent trend: use **'experiments' to estimate causal effects**

- Why? Clear counterfactual outcomes, reasonable to assume conditional independence

Experiments in Marketing!?

Yes. Two kinds ...

- **Randomised Control Trial (RCT)**
  - Researcher randomly assigns observational units to treatment group, control group
- **Natural Experiments / Quasi-Experiments**
  - "Nature" divides population into treatment and control in a way that is "as good as random"

Both approaches: Compare changes over time between groups

- How? ... that's what is coming next

Difference in Differences

What is Difference in Differences?

Want to answer the following question:

What is the effect of some marketing intervention on those who were effected by it?

- Call the intervention a **treatment**
- The treatment takes one of two values:
  - treatment = 1 if an observation is effected by the treatment
  - treatment = 0 if an observation is not effected by the treatment
- Observations are **treated at random**
- The treatment effects an **outcome**:

Estimator I: Before vs After?

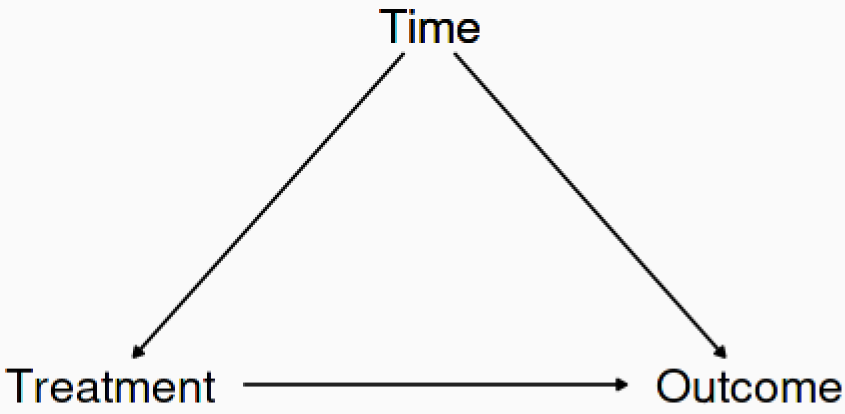
- We have data on observations **before** and **after** a treatment is introduced
- Let $\bar{y}$ denote averages

Proposed estimator I: **Before vs After for Treatment Group**

$$\text{Treatment Effect} = \bar{y}_{\text{after}} - \bar{y}_{\text{before}}$$

This **will not work**. Why?

- Time: things change over time for reasons unrelated to treatment



Estimator I: Before vs After?

Can't we control for time via (say) regression!?

- **No**
  - **treatment** occurrence and **time** are perfectly correlated
- Observation is either:
  - Before and Untreated, or
  - After and Treated.
- If control for time, you're comparing people with the same values of Time ...
- ... who must also have the same values of Treatment!

$\Longrightarrow$ Estimator won't work

Estimator II: Treatment vs Control

- We have data on observations for **treated** and **untreated** after the treatment is introduced
- Let $\bar{y}$ denote averages

Proposed estimator II: **Treated vs Untreated in the After Period**

$$\text{Treatment Effect} = \bar{y}_{\text{treated}} - \bar{y}_{\text{untreated}}$$

This **will not work**. Why?

- Treatment group might naturally vary from control group

$\implies$ Difference between them could be due to:

- The intervention, or
- Uncontrolled differences between the two groups

$\implies$ Estimator won't work

Difference in Differences

- Previous estimators: one difference (one minus sign)
    - **They don't work**

Why?

- Estimator I: confounded by time differences
- Estimator II: confounded by group differences

What if we could combine ideas from both?

$\implies$ that is what difference in differences does

Cool! How?

Difference in Differences: Notation

Assumption: The effect of time is constant between treated and control groups

We need four averages:

1. Control group, before intervention starts

$$\bar{y}^{\text{control}}_{\text{before}} = \beta_0$$

2. Control group, after intervention starts

$$\bar{y}^{\text{control}}_{\text{after}} = \beta_0 + \beta_1$$

3. Treatment group, before intervention starts

$$\bar{y}^{\text{treatment}}_{\text{before}} = \beta_0 + \beta_2$$

4. Treatment group, after intervention starts

$$\bar{y}^{\text{treatment}}_{\text{after}} = \beta_0 + \beta_2 + \beta_1 + \delta$$

$\Longrightarrow$ the (average) treatment effect is δ

This looks easier in a table...

The Difference in Difference Table

	Before	**After**
Control	β_0	$\beta_0 + \beta_1$
Treatment	$\beta_0 + \beta_2$	$\beta_0 + \beta_2 + \beta_1 + \delta$

The Difference in Difference Table

	Before	After	After - Before
Control	β_0	$\beta_0 + \beta_1$	β_1
Treatment	$\beta_0 + \beta_2$	$\beta_0 + \beta_2 + \beta_1 + \delta$	$\beta_1 + \delta$
Treatment - Control			δ

'Double Differencing' $\implies$ estimate δ

I call this DiD estimate using averages **simple DiD**

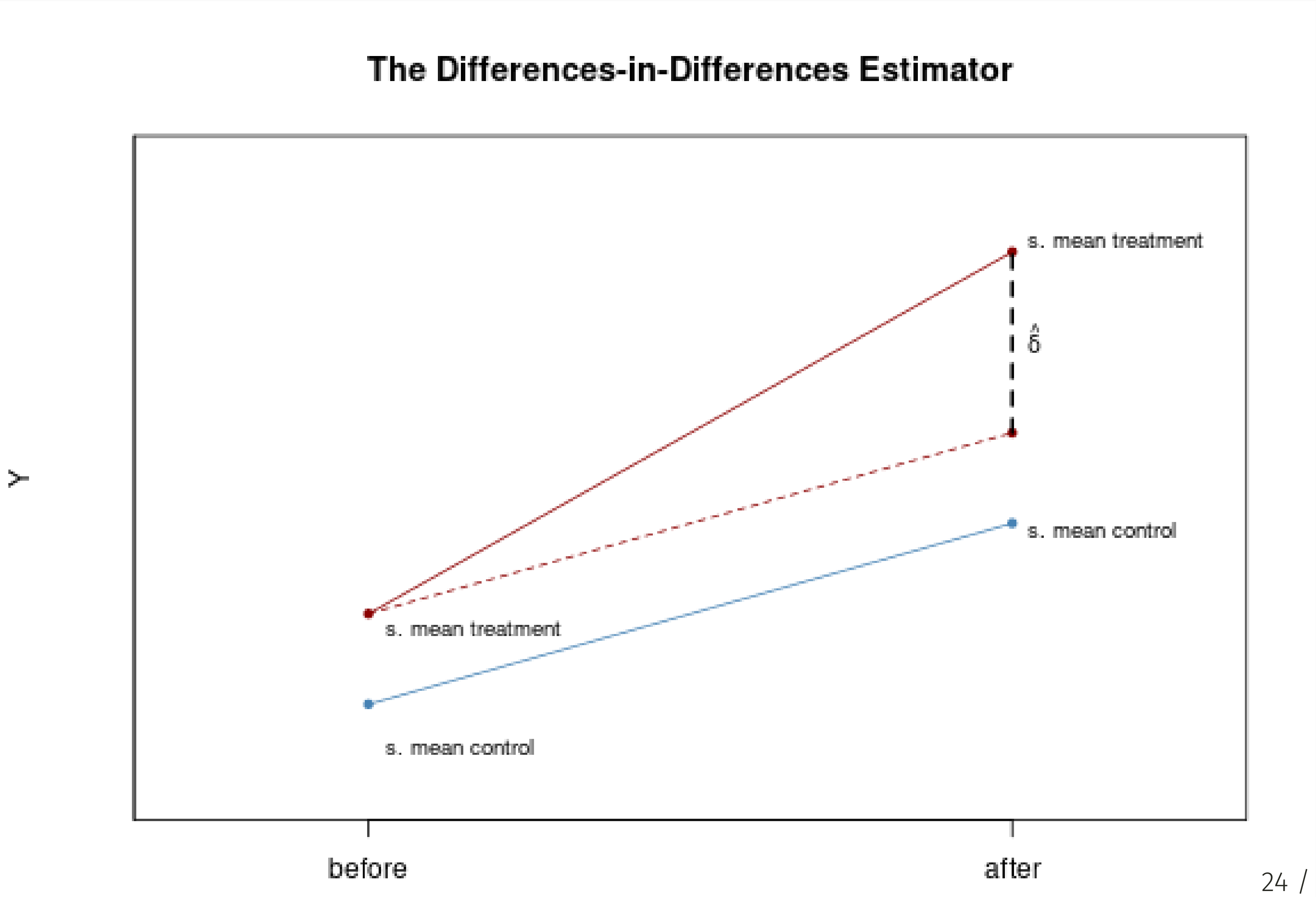
The Difference in Difference Table

	Before	After	After - Before
Control	β_0	$\beta_0 + \beta_1$	
Treatment	$\beta_0 + \beta_2$	$\beta_0 + \beta_2 + \beta_1 + \delta$	
Treatment - Control	β_2	$\beta_2 + \delta$	δ

'Double Differencing' $\implies$ estimate δ

I call this DiD estimate using averages **simple DiD**

Difference in Difference Graphically



Difference in Difference in R

How can we do this in R?

Let's first create some data:

- years: 2002 - 2010
- treatment for some observations in year 2007
- treatment effect: 2

```
# Create our data
diddata <- tibble(year = sample(2002:2010,10000,replace=T),
                  group = sample(c('TreatedGroup','UntreatedGroup'),10000,replace=
  mutate(after = (year >= 2007)) %>%
  #Only let the treatment (i.e. Treatment) be applied to the treated group
  mutate(Treatment = after*(group=='TreatedGroup')) %>%
  mutate(Y = 2*Treatment + .5*year + rnorm(10000)) %>%
  select(-Treatment) %>%
  mutate(treatment = case_when(
    group == "TreatedGroup" ~ TRUE,
    TRUE ~ FALSE
    )
  )
```

Difference in Difference in R

Now, compute averages by group and treatment status

```
means ←
  diddata %>%
  group_by(group,after) %>%
  summarize(Y=mean(Y)) %>%
  ungroup()

print(means)
```

```
## # A tibble: 4 × 3
##   group          after      Y
##   <chr>          <lgl> <dbl>
## 1 TreatedGroup   FALSE 1002.
## 2 TreatedGroup   TRUE  1006.
## 3 UntreatedGroup FALSE 1002.
## 4 UntreatedGroup TRUE  1004.
```

Difference in Difference in R

As a 'table'

```
did_table <-
  means %>%
  pivot_wider(names_from = after,
              values_from = Y
              )
print(did_table)
```

```
## # A tibble: 2 × 3
##   group          `FALSE` `TRUE`
##   <chr>            <dbl>  <dbl>
## 1 TreatedGroup     1002.  1006.
## 2 UntreatedGroup   1002.  1004.
```

Difference in Difference in R

Compute Treatment Effect, $\hat{\delta}$

```r
#Before-after difference for untreated, has time effect only
bef_aft_untreated <- filter(means,group=='UntreatedGroup',after==1)$Y -
                     filter(means,group=='UntreatedGroup',after==0)$Y
#Before-after for treated, has time and treatment effect
bef_aft_treated <- filter(means,group=='TreatedGroup',after==1)$Y -
                   filter(means,group=='TreatedGroup',after==0)$Y
#Difference-in-Difference! Take the Time + Treatment effect,
#                          and remove the Time effect
did <- bef_aft_treated - bef_aft_untreated

print(paste("Diff in Diff Estimate: ", did))
```

```
## [1] "Diff in Diff Estimate:  1.97404126317736"
```

Is Our Estimate Causal

We need **two assumptions** for causality:

1. A **valid counterfactual outcome** to compare treated group to
    - The control group gives us this
2. **Conditional Independence**: treatment assignment "as good as random"
    - We randomly assigned the treatment to some observations

$\implies$ **Difference in difference can give is causal estimates of the average treatment effect!**

Difference in Differences as a Regression

DiD as a Regression

$$y_{it} = \beta_0 + \beta_1 After_t + \beta_2 Treated_i + \delta After_t \times Treated_i + \varepsilon_{it}$$

where:

- $After_t$ = 1 in the period after treatment occurs, zero otherwise
- $Treated_i$ = 1 if the individual is ever treated, zero otherwise

DiD as a Regression

$$y_{it} = \beta_0 + \beta_1 After_t + \beta_2 Treated_i + \delta After_t \times Treated_i + \varepsilon_{it}$$

- β_0 is the prediction when $Treated_i = 0$ and $After_t = 0$
  - $\rightarrow$ the Untreated Before mean!
- β_1 is the *difference between* Before and After for $Treated_i = 0$
  - $\rightarrow$ Untreated (After - Before)
- β_2 is the *difference between* Treated and Untreated for $After_t = 0$
  - $\rightarrow$ Before (Treated - Untreated)
- δ is *how much bigger the Before-After difference* is for $Treated_i = 1$ than for $Treated_i = 0$
  - $\rightarrow$ (Treated After - Before) - (Untreated After - Before) = Treatment Effect!

Let's see that in action with `R`

DiD as a Regression

```r
reg_did <- lm(Y ~ after*treatment, data = diddata)

tidy(reg_did, conf.int = TRUE)
```

```
## # A tibble: 4 × 7
##   term                    estimate std.error statistic p.value conf.low conf.high
##   <chr>                      <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept)              1.00e+3    0.0226 44409.      0      1.00e+3 1002.
## 2 afterTRUE                2.29e+0    0.0341    67.1     0      2.22e+0    2.35
## 3 treatmentTRUE           -9.34e-3    0.0320    -0.292   0.770 -7.21e-2    0.0534
## 4 afterTRUE:treatmentTR…   1.97e+0    0.0484    40.8     0      1.88e+0    2.07
```

Advantages of Regression Approach

1. **Get standard error of the estimate**
    - Assess whether effect is statistically significant
    - *Should cluster standard errors*
    - (see this week's reading for suggestions on how)
2. **Can add extra control variables into the regression**
    - Either as 'usual' controls and/or as fixed effects
    - Particularly useful for Natural / Quasi Experiments
    - (see this week's reading)
3. **Can use $log(y)$ as dependent variable**
    - $\rightarrow \hat{\delta}$ is the percentage change in y due to the treatment

Hidden Assumptions, Caveats, etc

Hidden-ish Assumption: Parallel Trends

I briefly mentioned this in passing...

We must assume that Time effects treatment and control groups equally

- Otherwise controlling for time (i.e. `after`) won't work

This is called the **parallel trends** assumption

- Again, *if the Treatment hadn't happened to anyone*, the differences between the treatment and control would stay the same

Checking for Parallel Trends

Like many assumptions - its **untestable**

- Though we can **'check' whether patterns in the data are suggestive its OK**
- Here's one way:
  - Are *prior trends* are the same for Treated and Control groups
  - Generally, compute average of outcome by group over time
  - (needs multiple pre-treatment periods)
  - Was the gap changing a lot during that period? If not, suggestive we're OK

"As good as random" Redux

Remember our two assumptions for causality:

1. **Valid counterfactual outcomes**
    - Control Group solves this one for us
2. **Conditional independence**: nothing unobserved is causing selection into treament group
    - Trickier ...
    - Randomised Control Trial $\longrightarrow$ You're more than likely gonna be OK
    - Natural / Quasi Experiment - have you got a credible proxy for random assignment?
    - Profession's thoughts: Large, visible, unexpected shocks

Threats to Validity

Internal Validity: statistical inference made about causal effects are valid for the considered population

External Validity: inferences and conclusion are valid for the study's population and can be generalized to other populations and settings

Threats to Internal Validity

- **Failure to Randomise**
- **Failure to Follow Treatment Protocol**
- **Attrition**
- **Experimenter Demand Effects**
- **Small Sample Sizes**

Threats to External Validity

- **Non-representative sample**
- **Non-representative Marketing Intervention / Policy**
- **General Equilibrium Effects**

A Warning!

- DiD's popularity is relatively recent, so we're still learning a lot about it!
  - Most relevant has to do with **staggered roll out DiD**
- The regression version of DiD doesn't *necessarily* need to have treatment applied at *one* particular time
  - Treatment could be gradually implemented over time
- Nothing we've explicitly said would prevent us from using the regression DiD right!?
  - Well... that's what we thought for a long time.
  - And you'll see many of published studies doing this.
  - BUT it turns out to actually **bias results by quite a lot**
- There are more complex, newer estimators for staggered roll out case,
  - Too much for this class

Recap

Recap

- Many marketing questions require causal answers
- Establishing causality is goes beyond finding (partial) correlations in data
- RCT and Natural/Quasi Experiments introduce "as good as random" allocation to a treatment / marketing intervention
- Can use Difference in Difference to estimate causal effects of above experiments

Acknowledgements

Material in this set of slides borrows from the great work of others:

- Nick C Huntington Klein's course on Causality and Analytics
- Ed Rubin's Econometrics III
- Alan Spearot's class notes from Econ 113 in Fall 2014
- Hanck et al's Econometrics with R
- Goldfarb & Tucker's Conducting Research with Quasi-Experiments: A Guide for Marketers

License & Citation

Suggested Citation:

```
@misc{smwa_2022_lecture02,
      title={"Social Media and Web Analytics: Causality & Difference in Differences"},
      author={Lachlan Deer},
      year={2022},
      url = "https://github.com/tisem-digital-marketing/smwa-lecture-03"
}
```

This course adheres to the principles of the Open Science Community of Tilburg University. This initiative advocates for transparency and accessibility in research and teaching to all levels of society and thus creating more accountability and impact.