Causal Inference II

Esra Suel

CASA0006: Data Science for Spatial Systems

Regression discontinuity slides based on slides from Ollie Ballinger

DAG slides are based on PWG Tennant (2022) "Introduction To Causal Inference And Directed Acyclic Graphs"

Outline

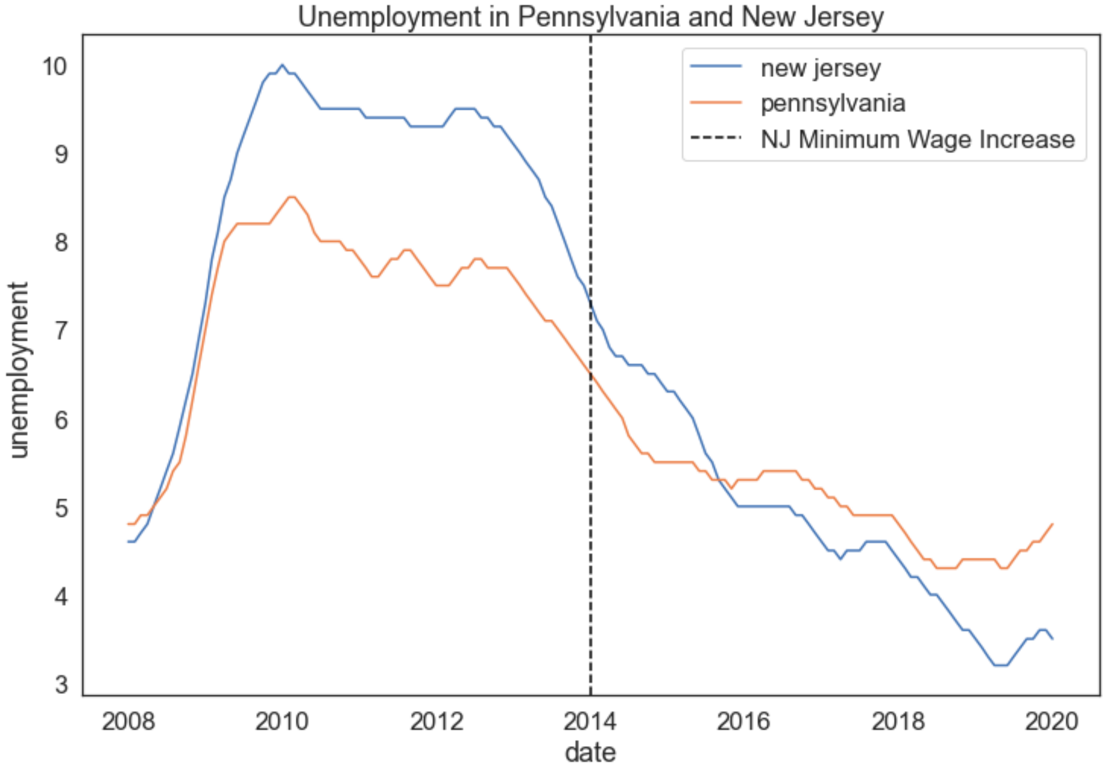
1. Regression Discontinuity
2. Fuzzy Regression Discontinuity
3. Directed Acyclic Graphs (DAGs)

Regression Discontinuity

Recap: Difference in Differences

$$Y_{it} = \beta_0 + \beta_1 Post_t + \beta_2 Treat_i + \beta_3 Treat_i \times Post_t + \varepsilon_{it}$$

	Control	Treatment
Before	β_0	$\beta_0 + \beta_2$
After	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$



Counterfactuals

Difference in differences requires a few things:

- A treatment located at a point in time
- Two distinct groups for which we have measurement pre-and post- treatment
- Parallel pre-treatment trends in the outcome variable y
- No simultaneous treatment occurring around our treatment of interest

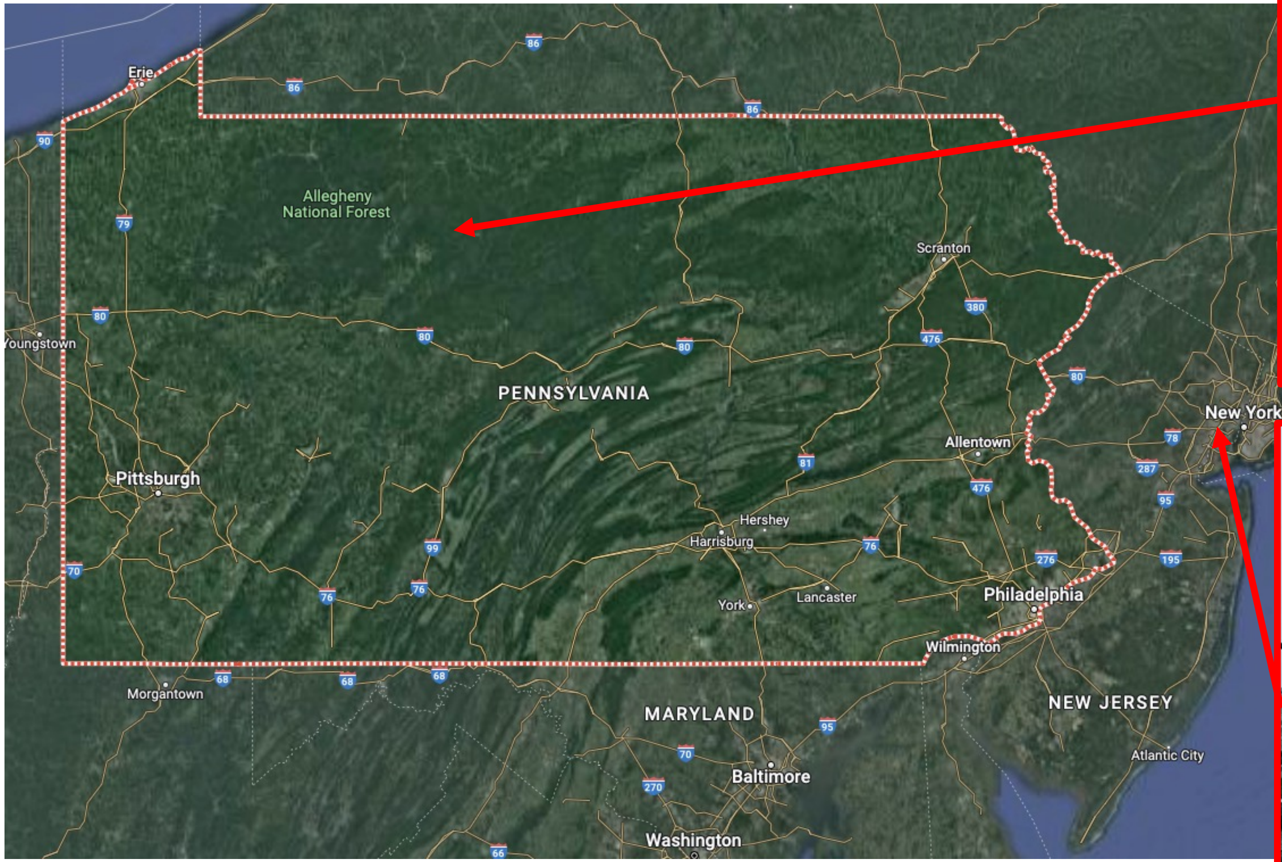
These allow us to construct a valid **counterfactual:**

- We can argue that the control group in the post-treatment period acts as a valid representation of the treatment group's behaviour in the absence of treatment. We can thus interpret the difference between the two as the causal effect of the treatment.

But this isn't the only way of constructing a valid counterfactual

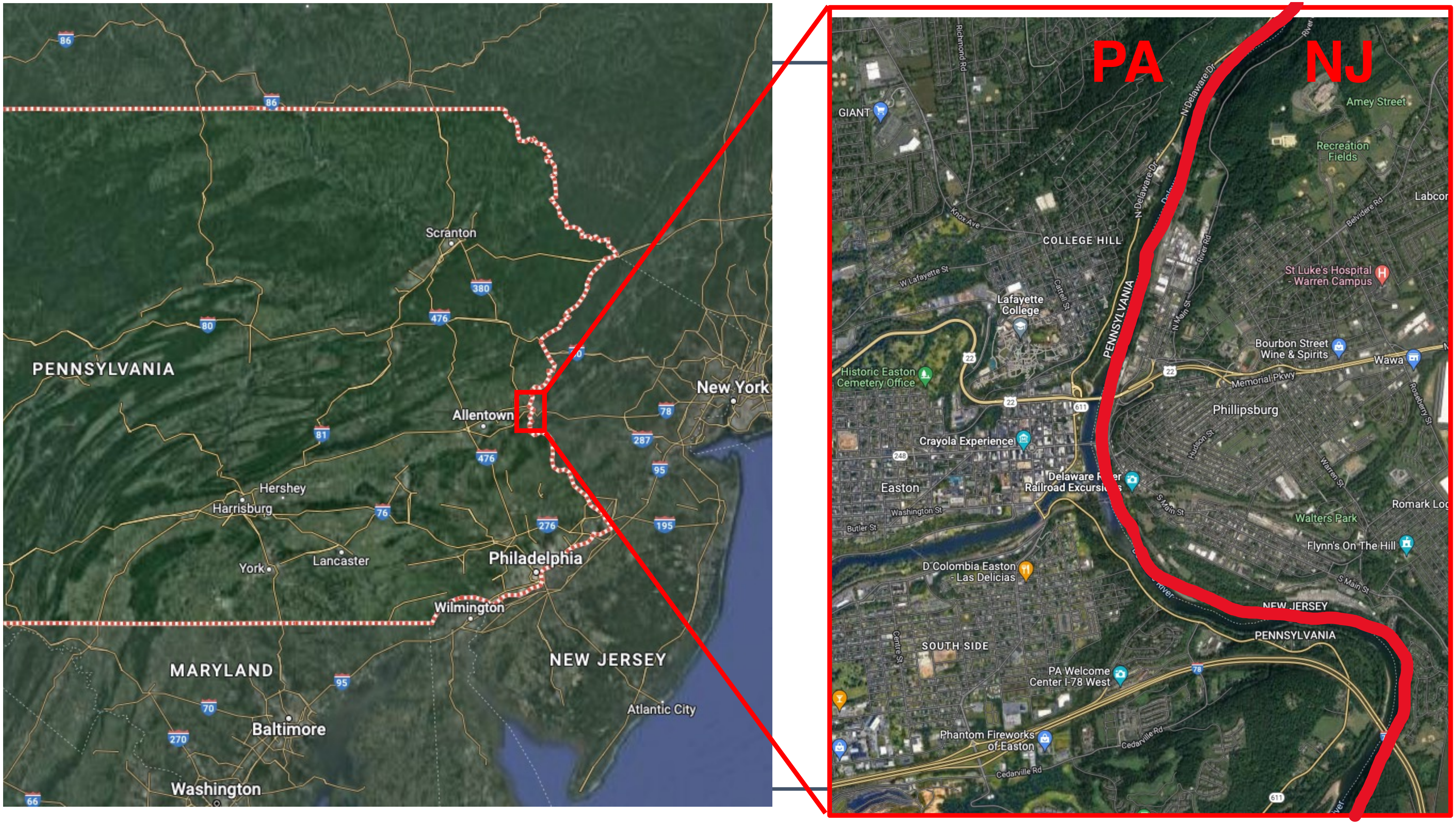
Counterfactuals

- What if we didn't have pre-treatment data for NJ and PA, but we had county level data on poverty for the period after the minimum wage was introduced.
- We can't do DiD, since we can't show that NJ and PA had similar trends in poverty pre-treatment. Maybe poverty in NJ was already falling rapidly before they introduced the minimum wage, while it was increasing in PA.
- We also can't just compare all the counties in PA against the counties in NJ, since there are huge states many differences between them that have nothing to do with the minimum wage law.



Tiona, PA

Jersey City, NJ


86
86
Scranton
380
476
80
PENNSYLVANIA
Allentown
New York
78
287
81
476
95
Hershey
Harrisburg
76
276
195
York
Lancaster
Philadelphia
Wilmington
MARYLAND
NEW JERSEY
95
70
Baltimore
270
Atlantic City
66
Washington
PA
NJ
GIANT
Amey Street
Recreation Fields
Labcor
COLLEGE HILL
St Luke's Hospital - Warren Campus
Lafayette College
PENNSYLVANIA
Bourbon Street Wine & Spirits
Wawa
Historic Easton Cemetery Office
Memorial Pkwy
Phillipsburg
Crayola Experience
Delaware River Railroad Excursions
Easton
Romark Log
Walters Park
Flynn's On The Hill
D'Colombia Easton - Las Delicias
NEW JERSEY
PENNSYLVANIA
SOUTH SIDE
PA Welcome Center I-78 West
Phantom Fireworks of Easton

Regression Discontinuity Design

- Regression Discontinuity Design (RDD) is a quasi-experimental evaluation option that measures the impact of an intervention, or treatment, by applying a treatment assignment mechanism based on a cutoff point in a continuous eligibility index.
- RDD estimates local average treatment effects around the cutoff point, where treatment and comparison units are most similar.

Source: World Bank

RDD Framework

Question: Do minimum wages increase unemployment?

Running variable:

- State– NJ and PA are hugely different beyond the fact that one implemented a minimum wage policy and the other didn’t

Exogenous cutoff:

- State border

Bandwidth:

- If we compare areas in PA and NJ within a small distance of the border between these states, they are probably going to be very similar in terms of demographics, economic structure, etc.

Conditions

1. A continuous eligibility index:

   A continuous measure on which the population of interest is ranked

   (e.g., test score, poverty score, age)

2. A clearly defined cutoff point:

   A point on the index above or below which the population is determined to be eligible for the program.

   Students with a test score of at least 80 of 100 are eligible for a scholarship

   Households with a poverty score less than 60 out of 100 are eligible for food stamps

   Individuals age 67 and older might be eligible for pension.

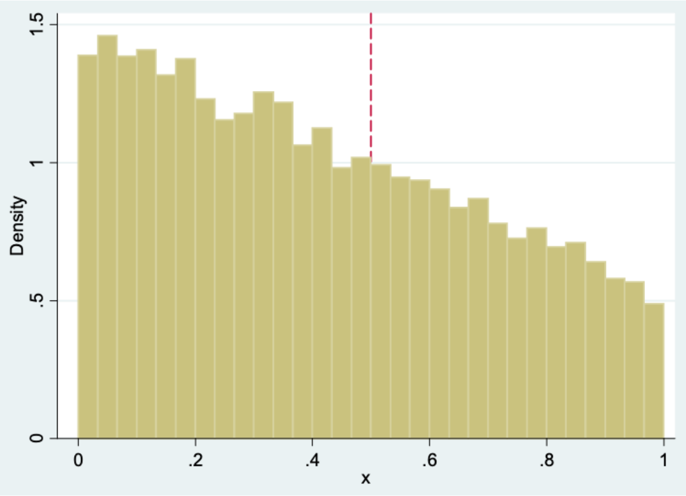
The cutoff points (thresholds) in these examples are 80, 60, and 67, respectively.

Assumption 1

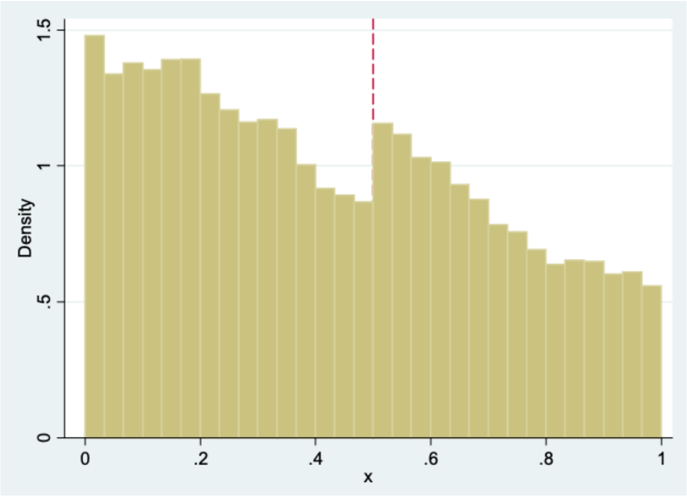
The eligibility index should be continuous around the cutoff point.

There should be no jumps in the eligibility index at the cutoff point or any other sign of individuals manipulating their eligibility index in order to increase their chances of being included in or excluded from the program.

Continuous distribution



Heaping around the cutoff



Assumption 2

Individuals close to the cutoff point should be very similar, on average, in observed and unobserved characteristics.

- In the RDD framework, this means that the distribution of the observed and unobserved variables should be continuous around the threshold.
- Even though researchers can check similarity between observed covariates, the similarity between unobserved characteristics must be assumed. This is considered a plausible assumption to make for individuals very close to the cutoff point, that is, for a relatively narrow window.

Original RDD paper

Thistlethwaite and Campbell (1960) study the effects of college scholarships on later students' achievements

Scholarships are granted based on whether a student's test score exceeds some threshold *c*

Consider the following variables:

- Binary treatment D is receipt of scholarship
- Covariate *X* is test score with threshold *c*
- Outcome Y is subsequent earnings
- Y_0 denotes potential earnings without the scholarship
- Y_1 denotes potential earnings with the scholarship

Source: Jonathan Mummolo

Original RDD Paper

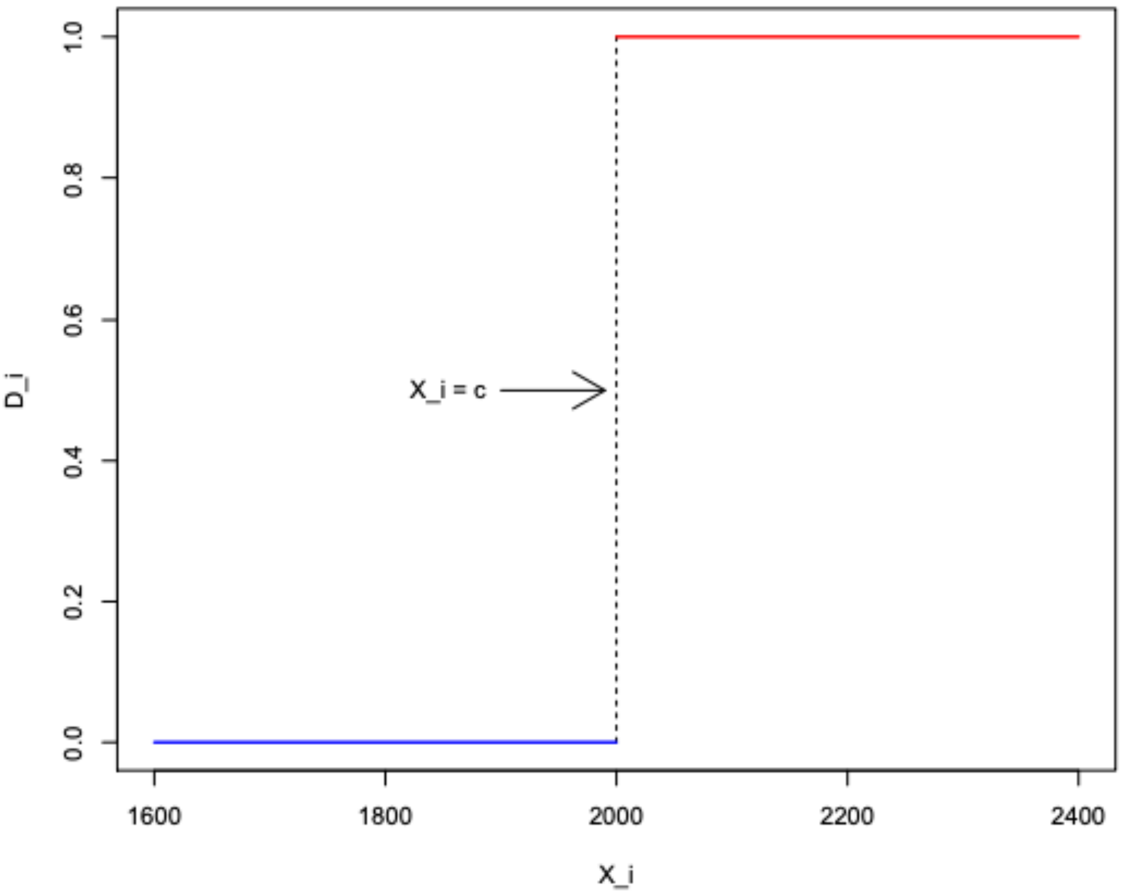
Assignment to the scholarship treatment Di is completely determined by the value of the SAT score $\mathrm{X_i}$ being on either side of the threshold c:

$$D_i = 1\{X_i > X_c\}$$

- X is called the forcing variable, because it “forces” units from control into treatment once X_i exceeds c
- X may be correlated with Y_0 and Y_1 so comparing treated and untreated units does not provide causal estimates (e.g. students with higher SAT scores obtain higher earnings even without the scholarship)
- If the relationship between X and the potential outcomes Y_1 and Y_0 is “smooth” around the threshold c, we can use the discontinuity created by the treatment to estimate the effect of D on Y at the threshold

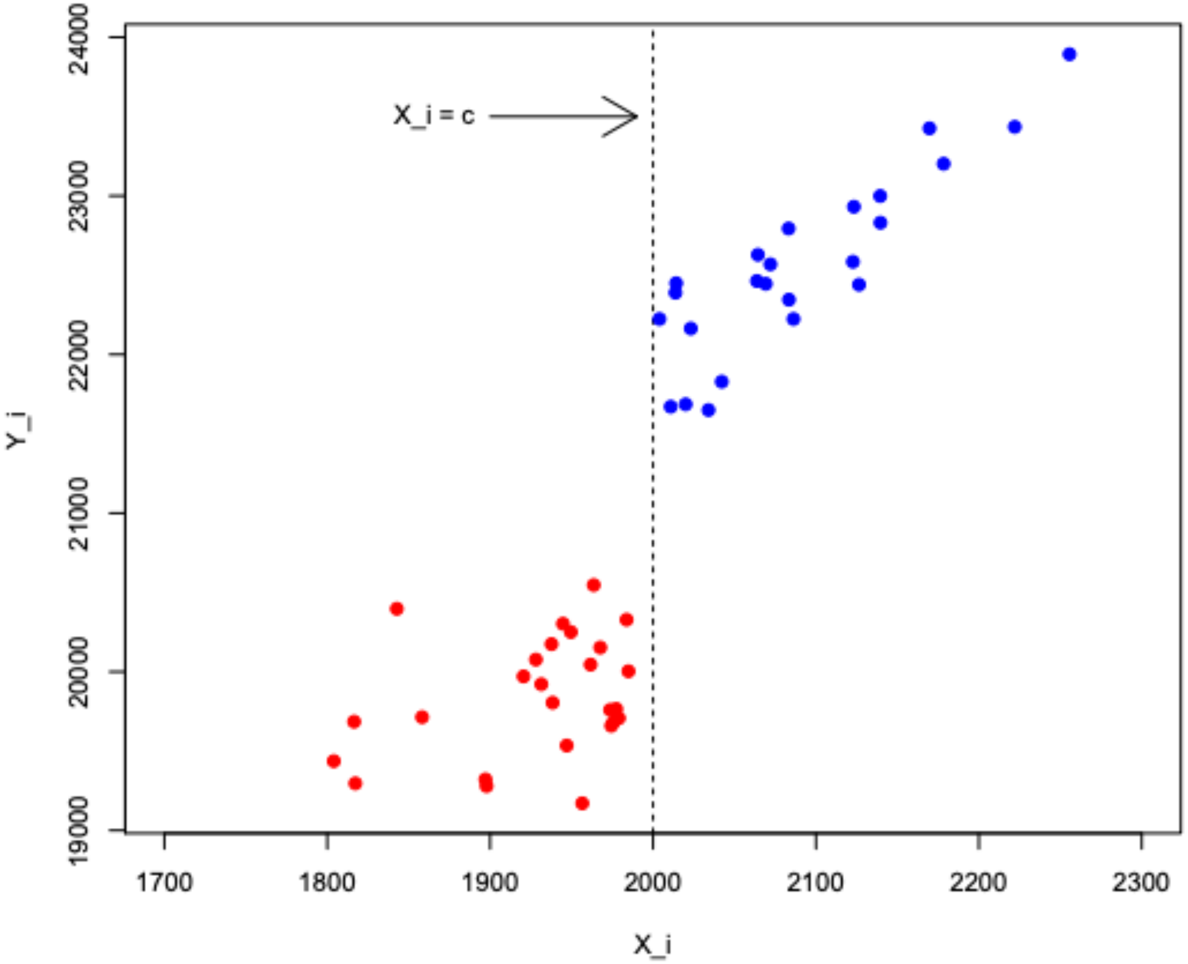
Treatment Assignment

$$D_i = \begin{cases} 1 \; if \; X_i > X_c \\ 0 \; if \; X_i < X_c \end{cases}$$



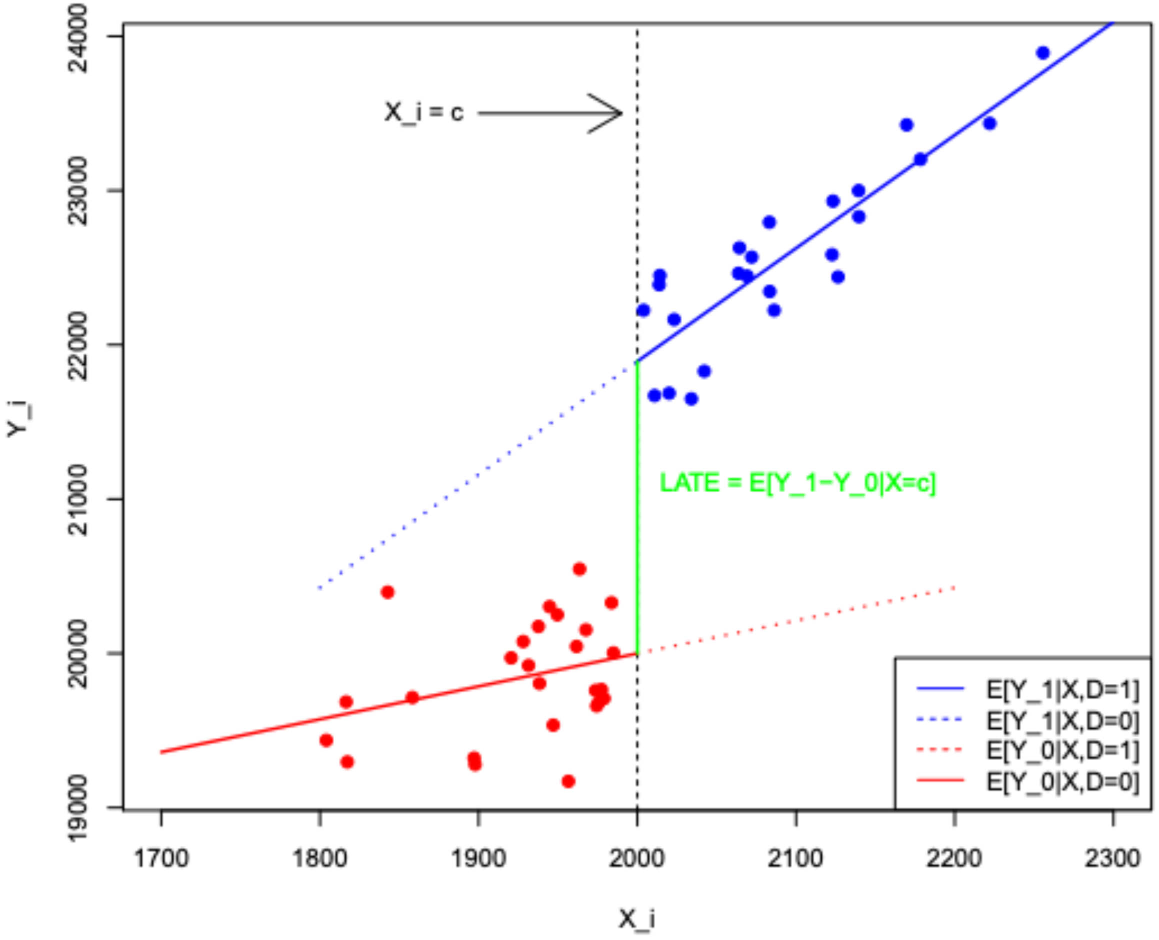
Observed Outcomes

There appears to be a jump in the observed values of the outcome variable (earnings) around the cutoff (scholarship eligibility)



Potential Outcomes

We can use the discontinuity created by the treatment to estimate the effect of D on Y at the threshold



Estimation

1. Trim the sample to a reasonable window around the cutoff c (discontinuity sample):
   - $c - h \leq X_i \leq c + h$, where h is some positive value that determines the size of the window h may be determined by cross-validation
2. Code the margin $\tilde{X}$ which measures the distance to the threshold:
   - $\tilde{X} = X - c$ such that

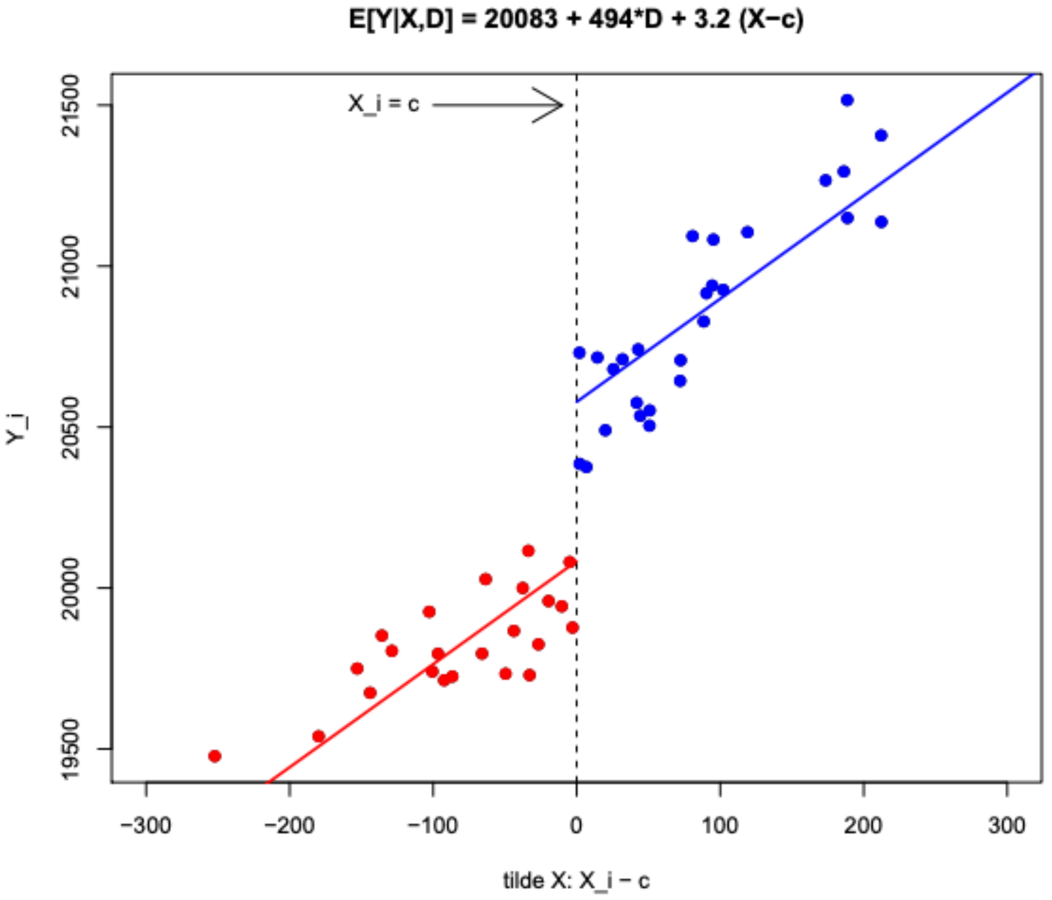
$$\tilde{X}_i = \begin{cases} \tilde{X} > 0 \; if \; X > c \;\; and \; thus \; D = 1 \\ \tilde{X} < 0 \; if \; X < c \;\; and \; thus \; D = 0 \end{cases}$$

3. Decide on a model for $E[Y \,|X]$
   1. linear, same slope for $E[Y_0|X]$ and $E[Y_1|X]$
   2. linear, different slopes
   3. non-linear

Linear, Same Slopes

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \beta_2 D_i + \varepsilon_i$$

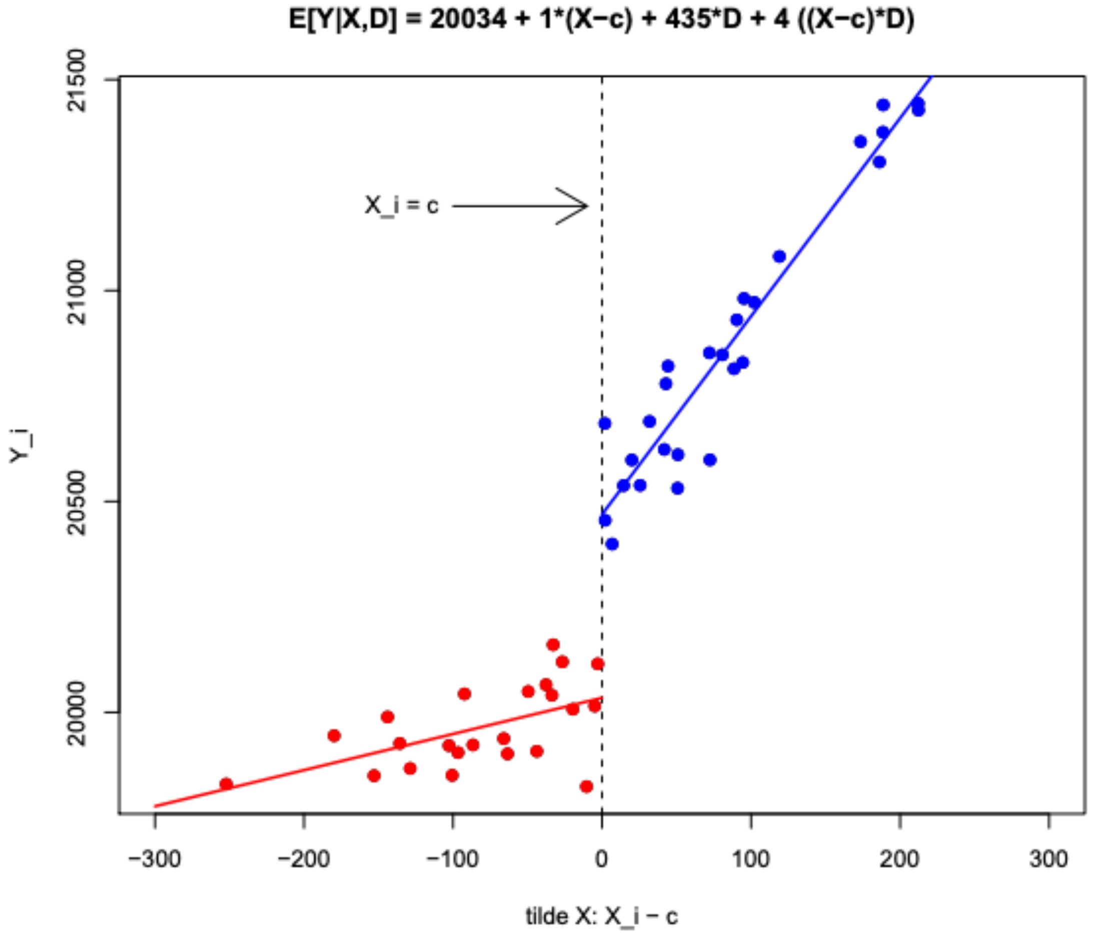
$$D_i = \begin{cases} 1 \; if \tilde{X} > 0 \\ 0 \; if \tilde{X} < 0 \end{cases}$$



Linear, Different Slopes

$$Y_i = \beta_0 + \beta_1 \tilde{X}_i + \beta_2 D_i$$

$$+\beta_3 (\tilde{X}_i \cdot D_i) + \varepsilon_i$$



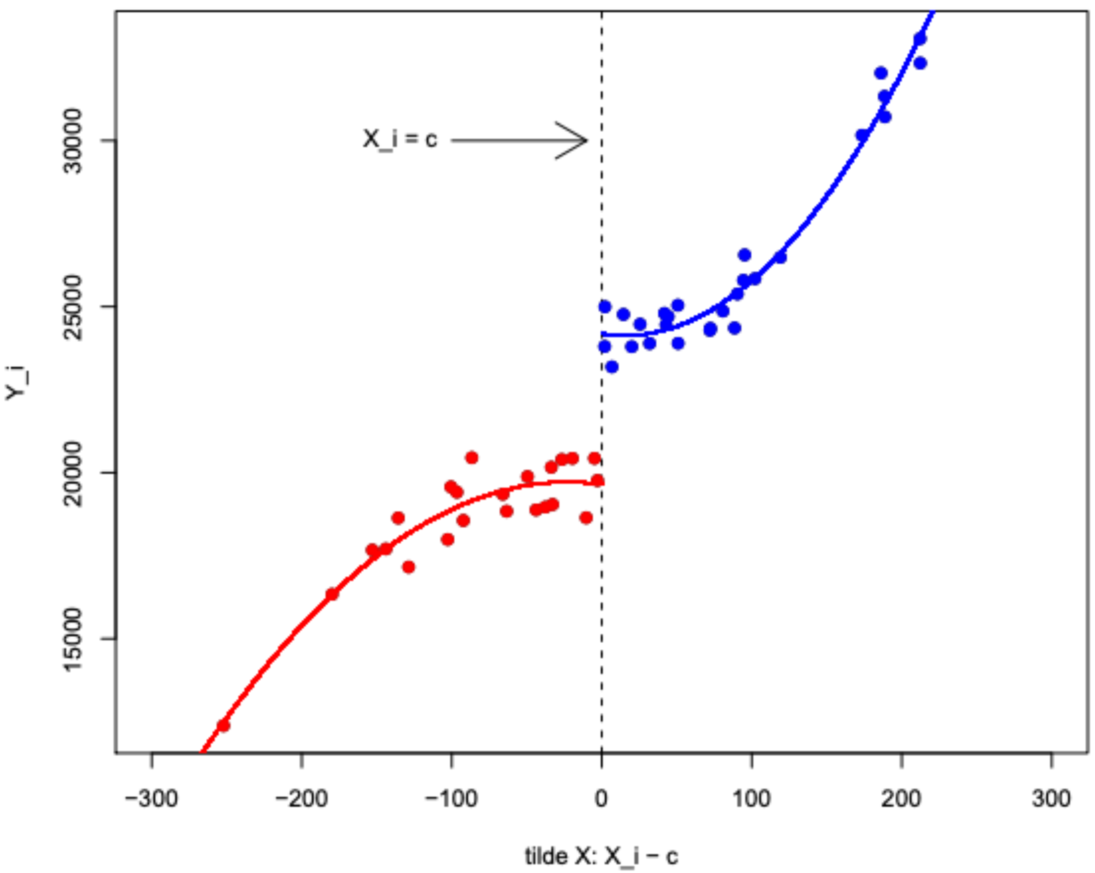
Non-Linear

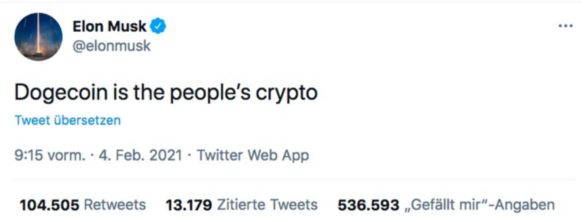
$$Y_i = \beta_0 + \beta_1\tilde{X}_i + \beta_2\tilde{X}_i^2 + \beta_3 D$$

$$+\beta_4\left(\tilde{X}_i \cdot D_i\right) + \beta_5\left(\tilde{X}_i^2 \cdot D_i\right)$$

$$+\varepsilon_i$$

E[Y|X,D]=19647−6*(X−c)−.1*(X−c)^2+4530*D−.9*((X−c)*D)+.4((X−c)^2*D)





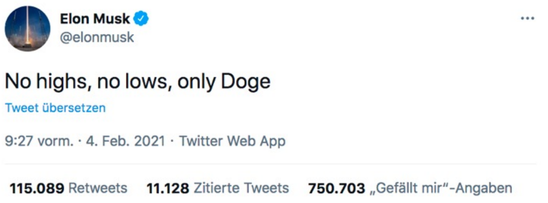
Elon Musk
@elonmusk

Dogecoin is the people's crypto

Tweet übersetzen

9:15 vorm. · 4. Feb. 2021 · Twitter Web App

104.505 Retweets **13.179** Zitierte Tweets **536.593** „Gefällt mir"-Angaben



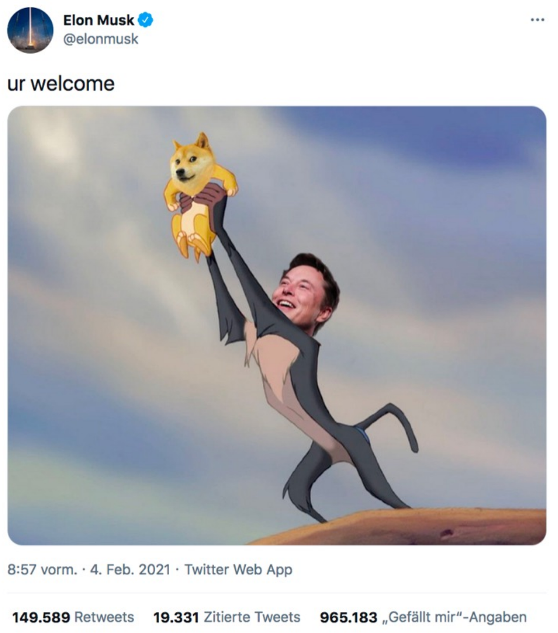
Elon Musk
@elonmusk

No highs, no lows, only Doge

Tweet übersetzen

9:27 vorm. · 4. Feb. 2021 · Twitter Web App

115.089 Retweets **11.128** Zitierte Tweets **750.703** „Gefällt mir"-Angaben

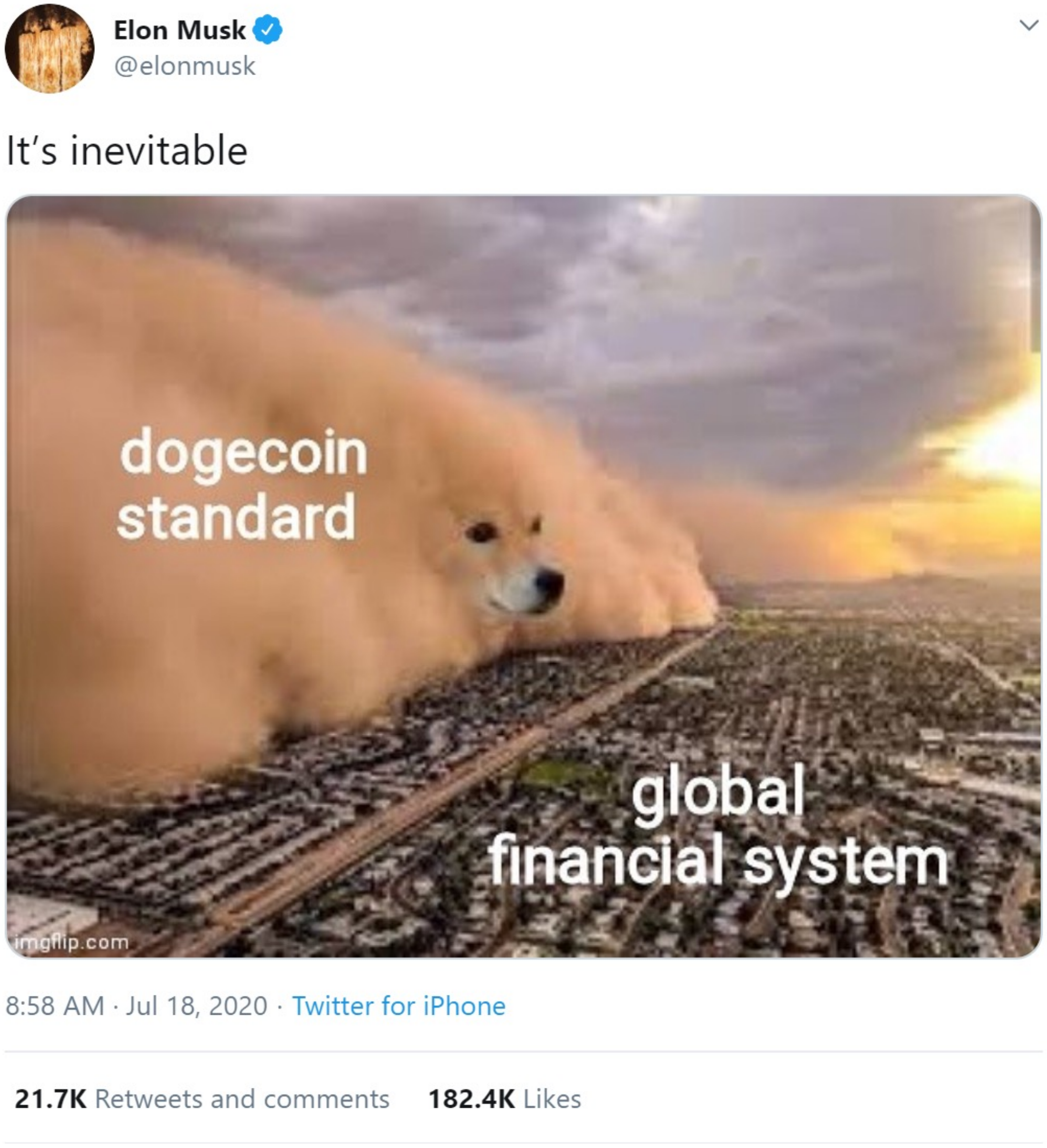


Elon Musk
@elonmusk

ur welcome

8:57 vorm. · 4. Feb. 2021 · Twitter Web App

149.589 Retweets **19.331** Zitierte Tweets **965.183** „Gefällt mir"-Angaben

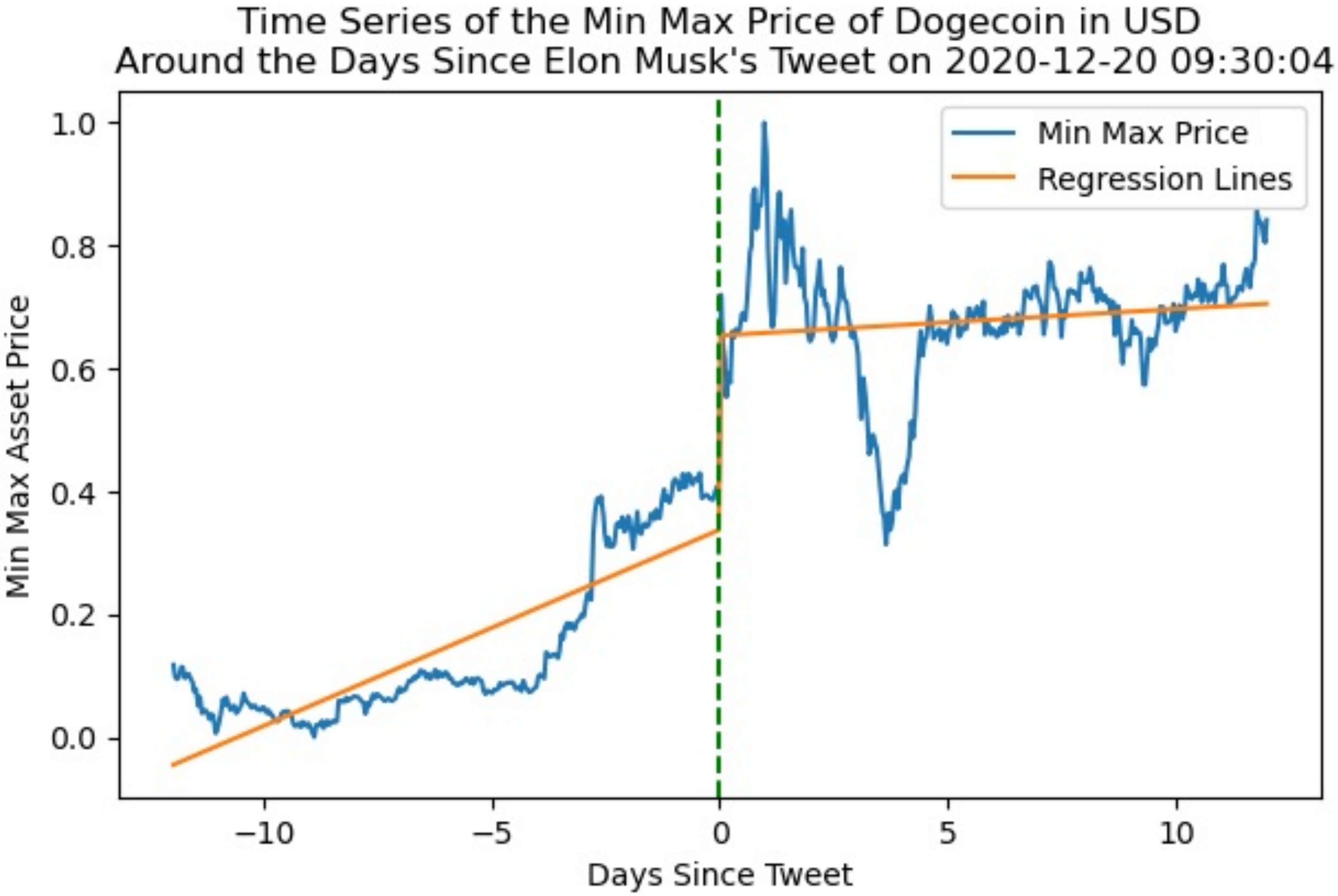


Elon Musk
@elonmusk

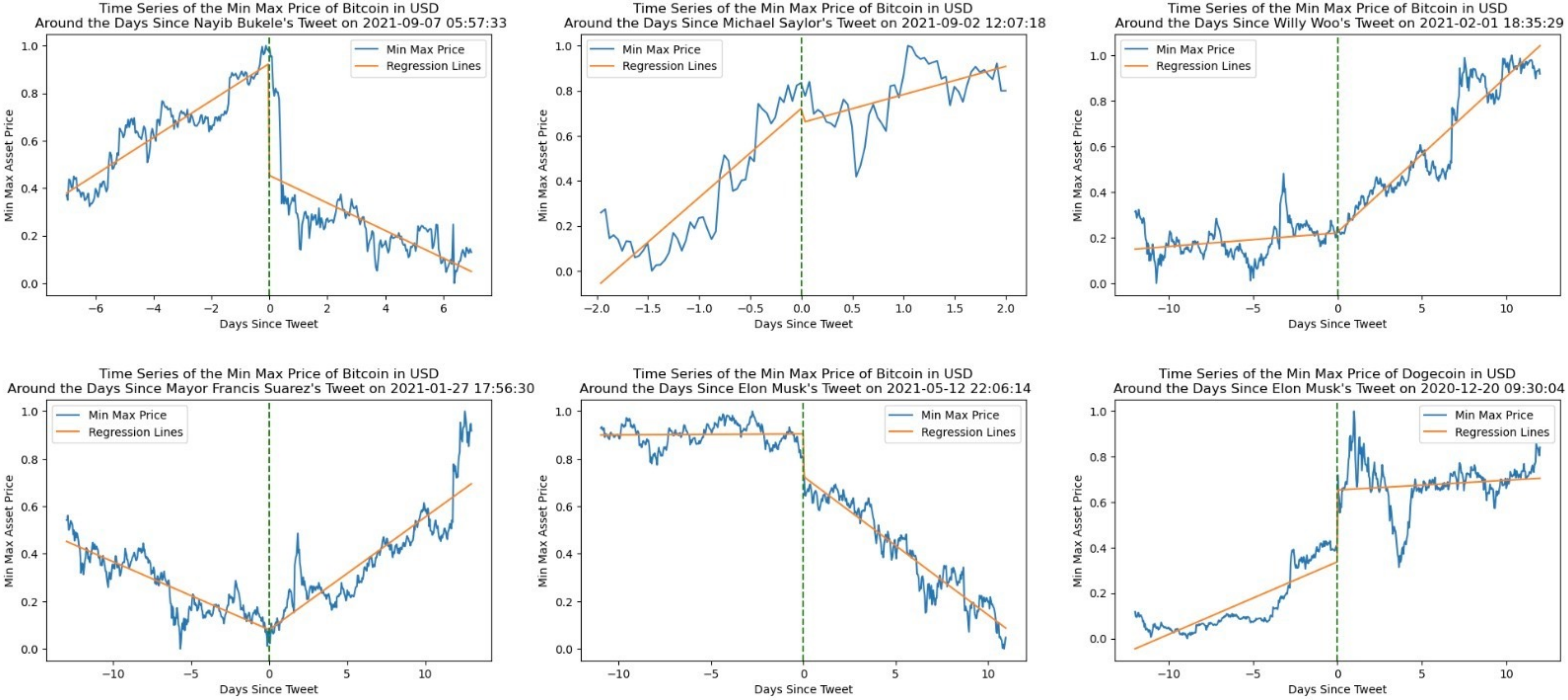
It's inevitable

8:58 AM · Jul 18, 2020 · Twitter for iPhone

21.7K Retweets and comments **182.4K** Likes


Time Series of the Min Max Price of Dogecoin in USD
Around the Days Since Elon Musk's Tweet on 2020-12-20 09:30:04
Min Max Price
Regression Lines
Min Max Asset Price
1.0
0.8
0.6
0.4
0.2
0.0
−10
−5
0
5
10
Days Since Tweet

Time Series of the Min Max Prices of each Tweet's Discontinuity Regression Model



Time Series of the Min Max Price of Bitcoin in USD
Around the Days Since Nayib Bukele's Tweet on 2021-09-07 05:57:33

Time Series of the Min Max Price of Bitcoin in USD
Around the Days Since Michael Saylor's Tweet on 2021-09-02 12:07:18

Time Series of the Min Max Price of Bitcoin in USD
Around the Days Since Willy Woo's Tweet on 2021-02-01 18:35:29

Time Series of the Min Max Price of Bitcoin in USD
Around the Days Since Mayor Francis Suarez's Tweet on 2021-01-27 17:56:30

Time Series of the Min Max Price of Bitcoin in USD
Around the Days Since Elon Musk's Tweet on 2021-05-12 22:06:14

Time Series of the Min Max Price of Dogecoin in USD
Around the Days Since Elon Musk's Tweet on 2020-12-20 09:30:04

Example 1

Question: Does a test-based scholarship increase future earnings?

Running variable:

- Test score– those with really low test scores are different from those with high test scores in ways unrelated to the scholarship (e.g. wealth/ability to pay for tutoring, etc.)

Exogenous cutoff:

- Fuzzy, test score cutoff (e.g., need 90% to qualify for scholarship)

Bandwidth:

- 90%±5%
- If we compare people who narrowly qualify for the scholarship to those who narrowly fail to qualify, the latter may be similar enough to the former to act as a counterfactual.

Example 2

Question: Do members of Party A curtail women's rights?

Running variable:

- Vote share– districts that are strongholds for Party A will be different in terms of their social, economic and cultural factors compared to strongholds of Party B.

Exogenous cutoff:

- Sharp, victory/loss (assuming a 2 party system, 50% vote share)

Bandwidth:

- 50%±4%
- If we compare districts where candidates of that party narrowly won against those where they narrowly lost, those districts are very similar in terms of their unobserved characteristics

Example 3

Question: Does an income-based cash transfer program decrease infant mortality?

Running variable:

- Income– those in extreme poverty face problems related to infant mortality that the rich do not (e.g. nutrition, access to healthcare)

Exogenous cutoff:

- Fuzzy, income-based assignment to treatment (e.g. need to make <£20k/yr)

Bandwidth:

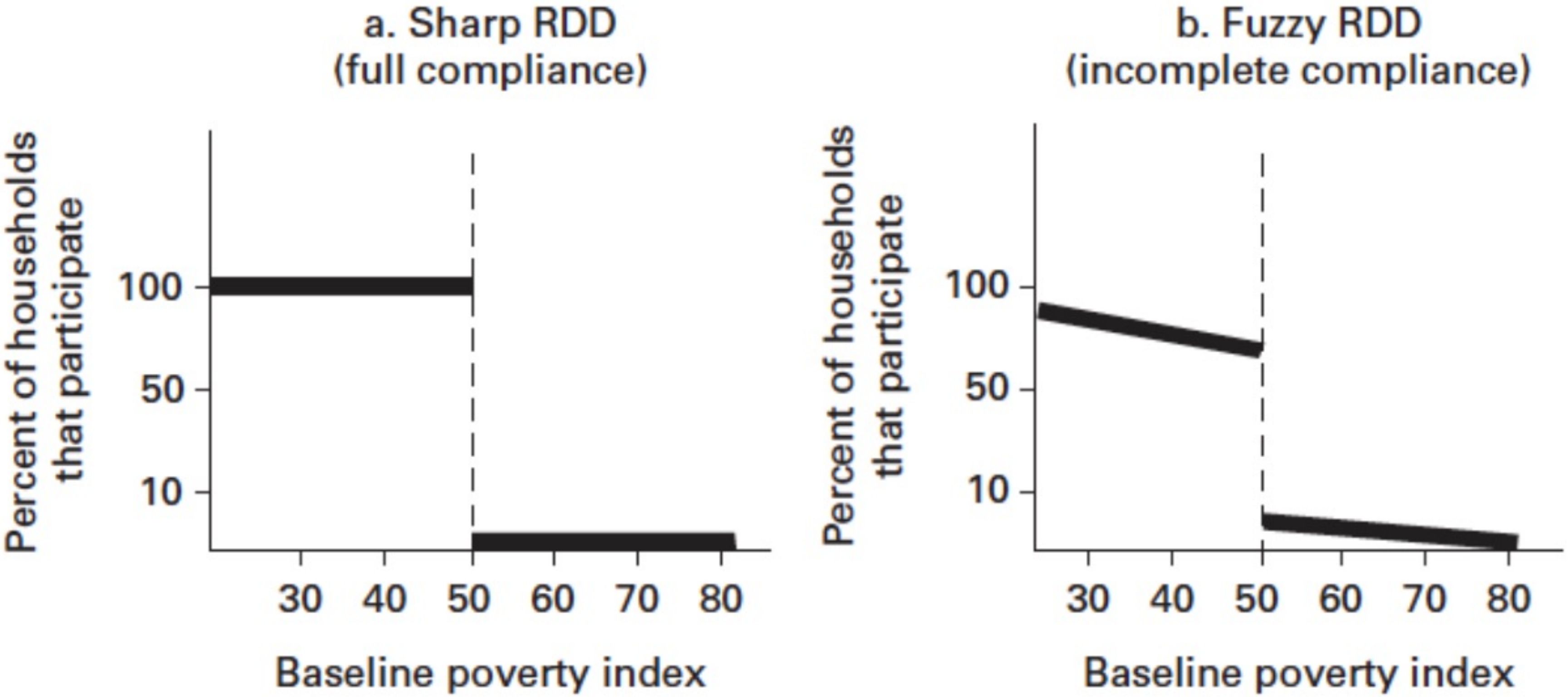
- £20k±3k
- If we compare people who narrowly qualify for a social program to those who narrowly fail to qualify, the latter may be similar enough to the former to act as a counterfactual.

Fuzzy Regression Discontinuity

Fuzzy Regression Discontinuity

- Threshold may not perfectly determine treatment exposure, but it creates a discontinuity in the probability of treatment exposure
- Incentives to participate in a program may change discontinuously at a threshold, but the incentives are not powerful enough to move all units from non-participation to participation
- We can use such discontinuities to produce instrumental variable estimators of the effect of the treatment (close to the discontinuity)

Sharp vs. Fuzzy RDD



Assigned versus observed treatment

Assume the treatment is offered to everybody above c, but not everybody might take it

Let $Z = 1\{X > c\}$ be a binary encouragement indicator that captures whether units are above or below the threshold c

Let D be the binary observed treatment indicator that captures whether individuals take the treatment or not

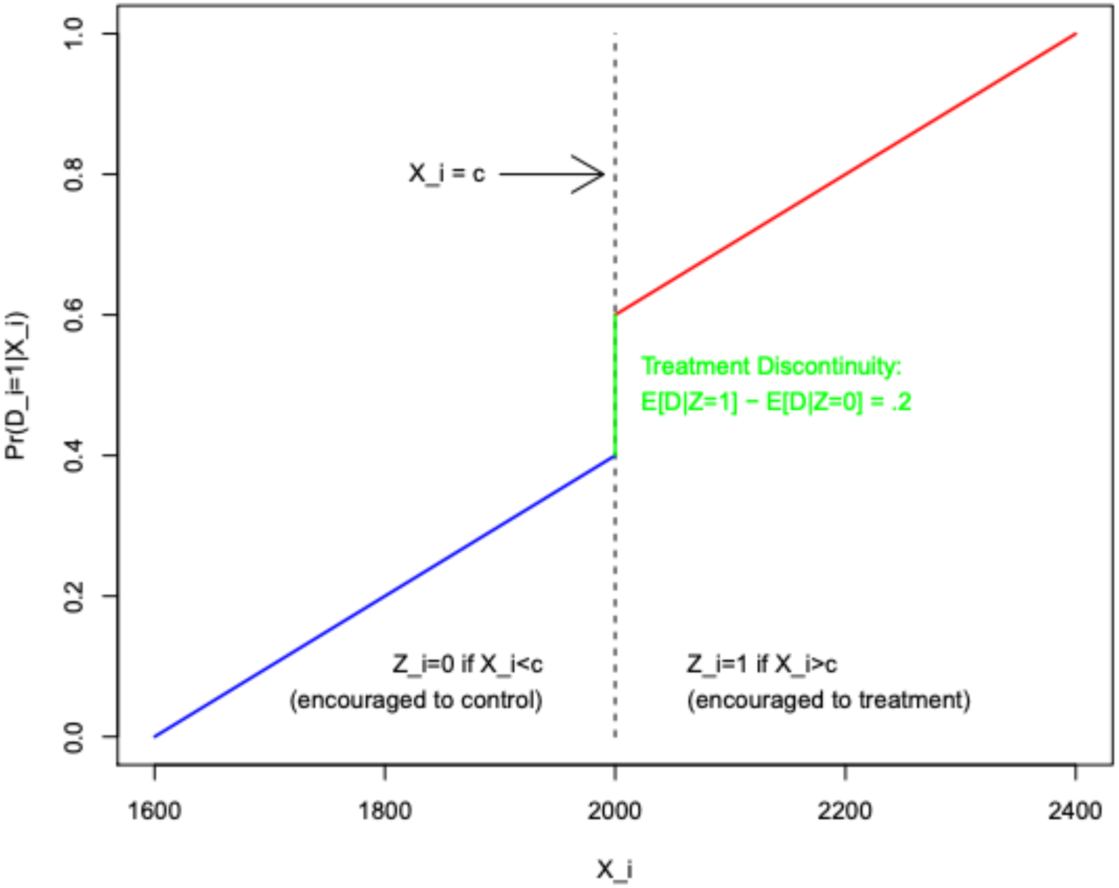
Observed treatment D_i can be modeled as a function of the running variable X_i and the binary encouragement indicator Z_i

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \varepsilon_i$$

First stage

Estimating uptake based on assignment to treatment group:

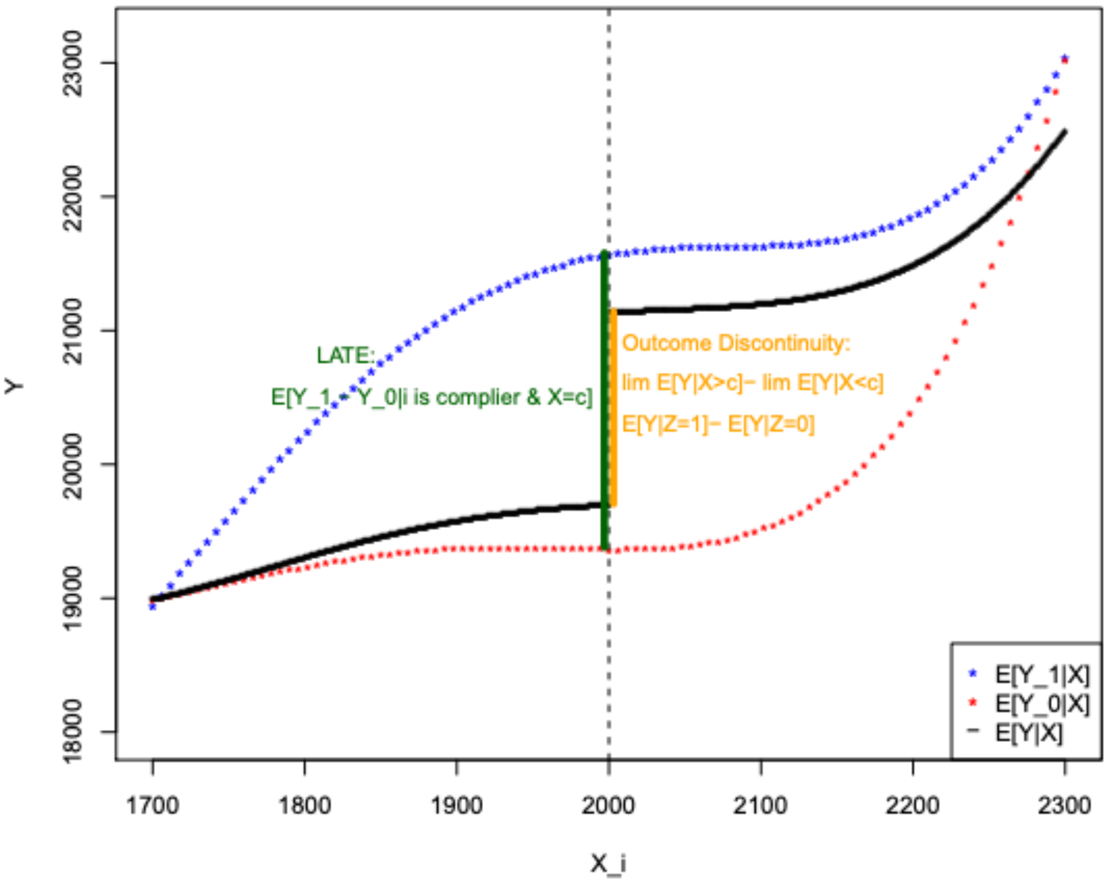
$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 Z_i + \varepsilon_i$$



Second stage

Estimating the outcome discontinuity using fitted values $\widehat{D}_i$ from the first stage:

$$Y_i = \beta_0 + \beta_1 X_i + \alpha \widehat{D}_i + u_i$$



Directed Acyclic Graphs (DAGs)

Slides are based on:

PWG Tennant (2022) "Introduction To Causal Inference And Directed Acyclic Graphs"

What did we cover so far?

- Randomized Control Trials (RCTs) are the gold standard for causal inference.

JAMA®

*“**Causal language** (including use of terms such as effect and efficacy) **should be used only for randomized clinical trials**. For all other study designs (including meta-analyses of randomized clinical trials), methods and results should be described in terms of association or correlation and should avoid cause-and-effect wording.”*

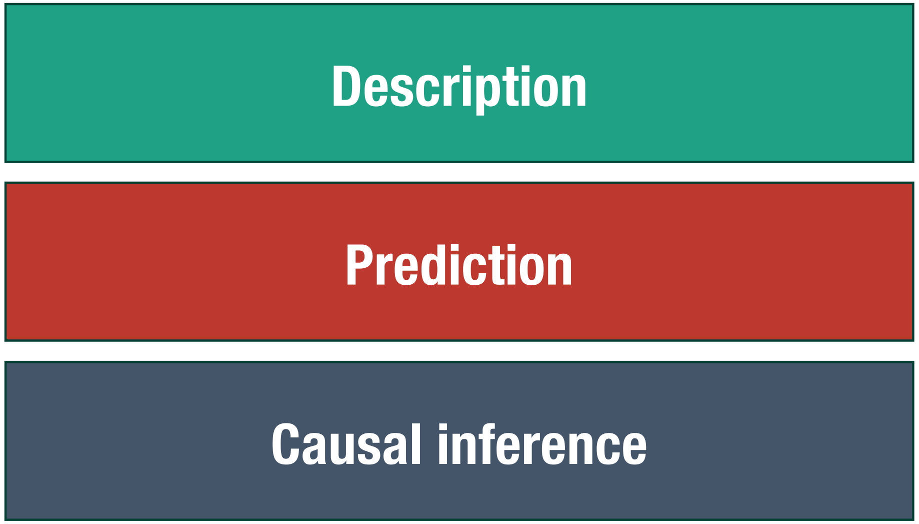
- Researchers would avoid causal language, and you’ll see:

  “link”, “risk factor”, ”correlation”, “association”, “predictor”

What did we cover so far?

- Randomized Control Trials (RCTs) are the gold standard for causal inference.
- If only observational data is available, exploring natural experiments can provide settings for causal analysis.
- Difference-in-Differences (DID) and Regression Discontinuity (RD) are designs that aim to approximate the randomized assignment of treatment and control groups, similar to what is achieved in RCTs.
- Next: Directed Acyclic Graphs (DAGs) to enhance our understanding of causal relationships and aid in the identification and adjustment of confounding variables in observational studies.

Different tasks in data science



Description

Prediction

Causal inference

Description

- Focused on summarising, describing, and/or visualising features
- Data driven – involves e.g., simple calculations, models, unsupervised learning

Questions

- What happened?
- Who was affected?
- What was the occurrence of Y in people with X.

What is the risk of death from COVID-19 among bald men?

Prediction

- Classification and regression
- Focus on pattern recognition and forecasting
- Data driven – involves e.g., statistical modelling, supervised learning

Questions

- What will happen?
- Who will be affected?
- Are people with X more likely to have Y

Are bald men more likely to die from COVID-19?

Causal Inference (counterfactual prediction)

- Focus on understanding
- NOT data driven – involves external knowledge brought into statistical modelling and supervised learning

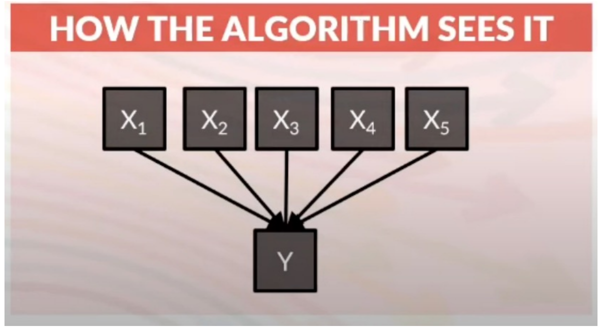
Questions

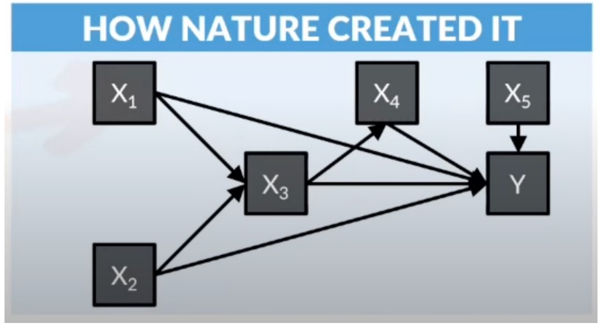
- What will happen if … ?
- Why are they affected?
- If we changed X, how would it change Y.

If bald man buys a wig, does this reduce his risk of death from COVID-19?

What the machine cannot learn

- Data driven algorithms are excellent at finding patterns in complex data, and very well suited for prediction.
- Causal inference requires identifying and estimating counterfactuals, which cannot be learned from data
- We must provide ‘external knowledge’ of (or control) the data generating process – the story behind how the data came into being



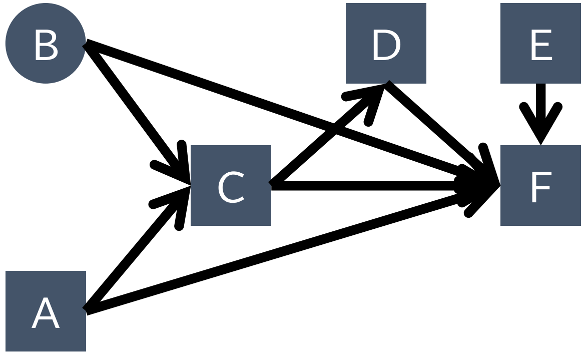


Pearl’s Structural Causal Model

- Provide a formal mathematical and philosophical framework for considering and estimating causal effects
- Key idea: you must formally identify what you want to estimate from external theory before you conduct your analysis

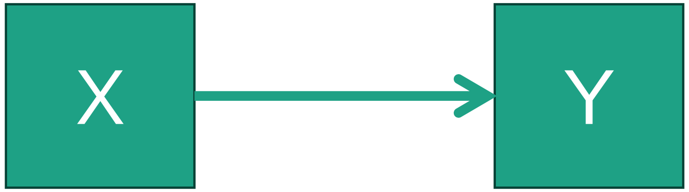
Causal Directed Acyclic Graphs (DAGs)

- Causal diagrams, such as directed acyclic graphs (DAGs) help us encode and represent our theory of the **data generating mechanism**
- DAGs are non-parametric graphical representations of (hypothesised) causal relationships between variables, where:
  - Variables are represented as '**nodes**'
  - Causal relationships are represented as directed 'arcs' (i.e. **arrows**)
  - There are no circular paths, hence the word '**acyclic**'



DAGs

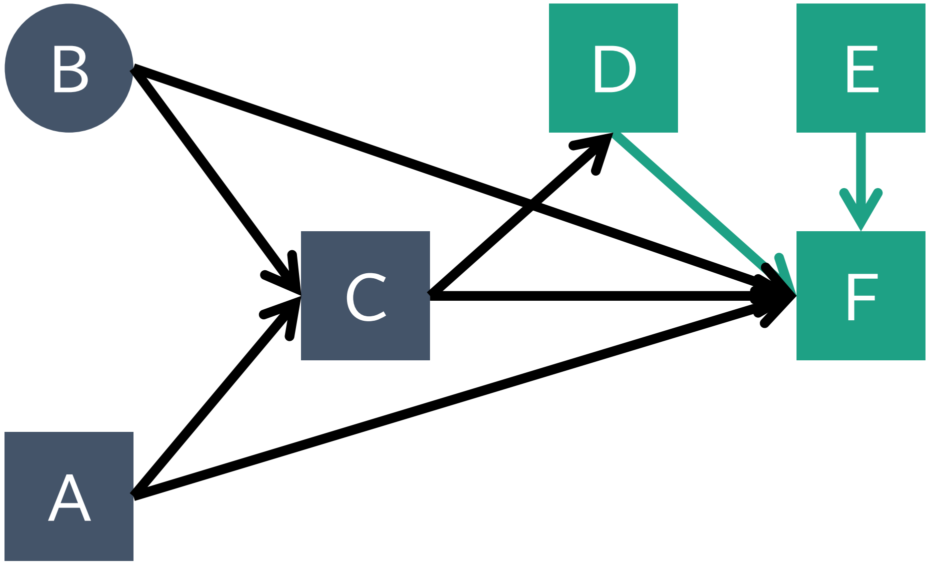
- DAGs are directed – causality is directed over time (e.g., a cause cannot occur after a consequence)



- Drawing an arc between X and Y, we state that we believe:
  - Changing X modifies the probability of Y (probabilistic reasoning)
  - If X had been different, Y would have been different (counterfactual reasoning)
  - Y “listens” to X
  - If we could wiggle X, it would wiggle Y

Paths

- A path exists between two variables if they are connected by one or more arrows (regardless of the direction)
- D → F ← E

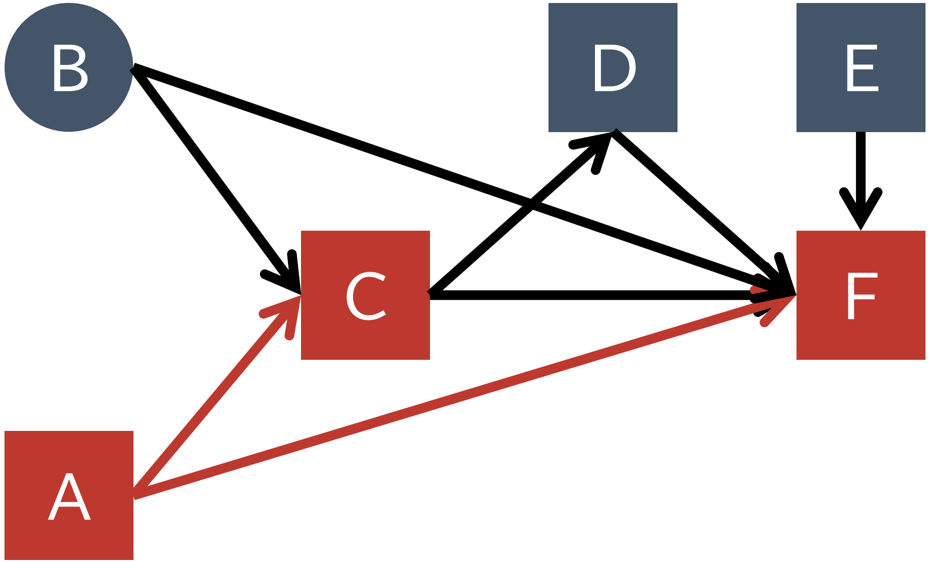


Open and closed paths

- A path may be open or closed
- Open path: transmits associations (correlations or dependencies)
- Closed path: do not transmit associations

Confounding paths

- A **confounding path** (a **backdoor path**): all the arrows run do not flow in the same direction – initially backwards, then forwards
- C ← A → F
- Without conditioning, confounding paths are open and will transmit dependencies between the variables cause by the confounder (e.g. A)



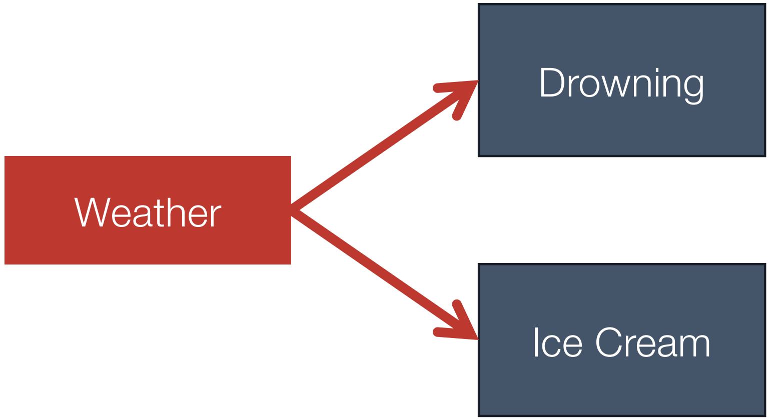
Conditioning (adjusting / controlling)

Conditioning is the process of estimating a statistic (e.g. coefficient in a regression model) at **fixed levels of one or more other variables**.

- **Restriction:** Estimating the effect in a sample with similar values of one or more variables (e.g. non-smokers only)
- **Stratification:** Estimating the effect in strata with similar values of one more other variables (e.g. non-smokers, ex-smokers, current smokers)
- **Covariate adjustment:** Estimating the effect while controlling for values of one or more other var.s (e.g. including smoking as a covariate in a regression model)
- **Matching** Estimating the effect in clusters with similar values of one or more other variables (e.g. participants are matched on smoking status at recruitment)

Example – confounding bias

Association between two variables due to a common cause

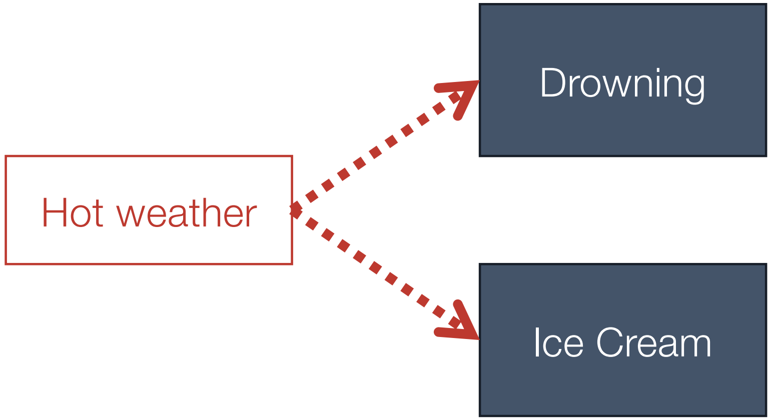


Weather causes ice-cream consumption and risk of drownings
Weather is a confounder, causing unconditional dependency through:

Ice-cream ← Weather → Drowning

Example – confounding bias

Association disappears if you only look at days with similar weather



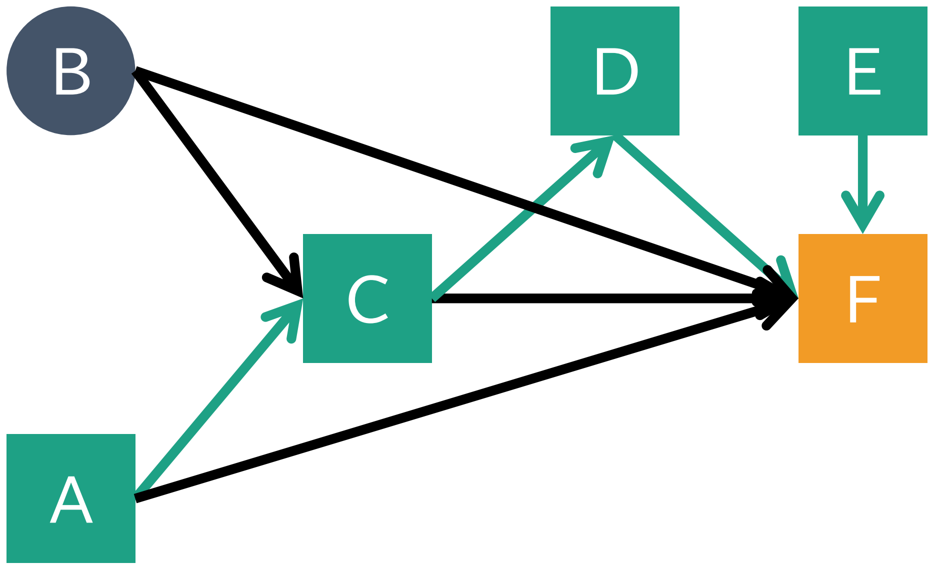
By conditioning on the confounder we close the path

Ice-cream ← Weather → Drowning

Collider paths

- A **collider path:** arrows do not flow in the same direction – initially forwards then backwards
- E.g. A → C → D → F ← E

- Without <u>conditioning</u>, collider paths are closed and will <u>NOT</u> transmit dependencies between either side of the collider (e.g. F)

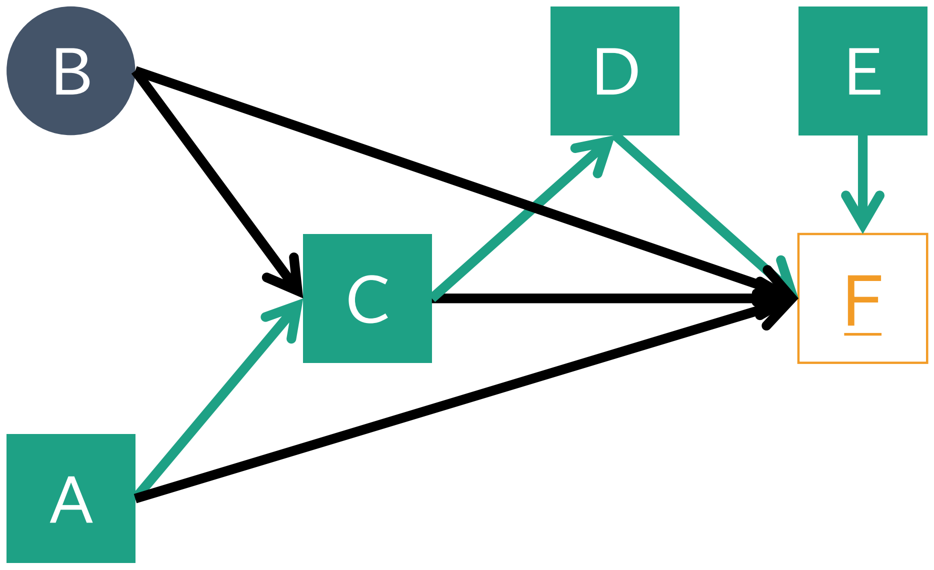


Collider paths

- A **collider path** can be opened by conditioning on the collider node(s) – or a proxy – transmitting conditional dependencies between variables that cause the collider (e.g. F)

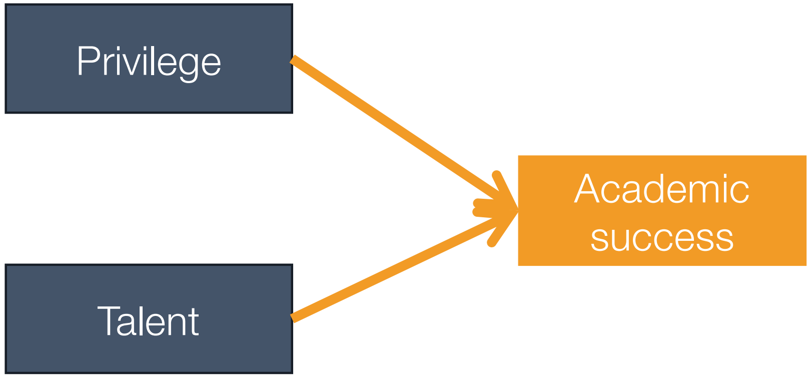
- Conditioning on F opens

  A → C → D → F ← E



Example: Collider Bias

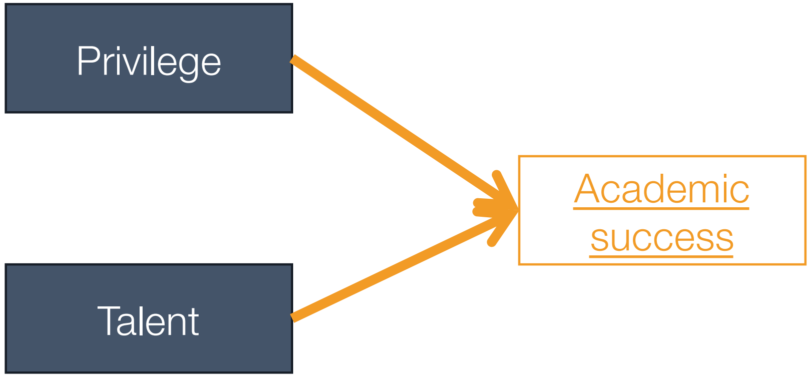
- Privilege and talent are two competing reasons why someone might achieve academic success



- In the general population, it is unlikely that privilege and talent are related

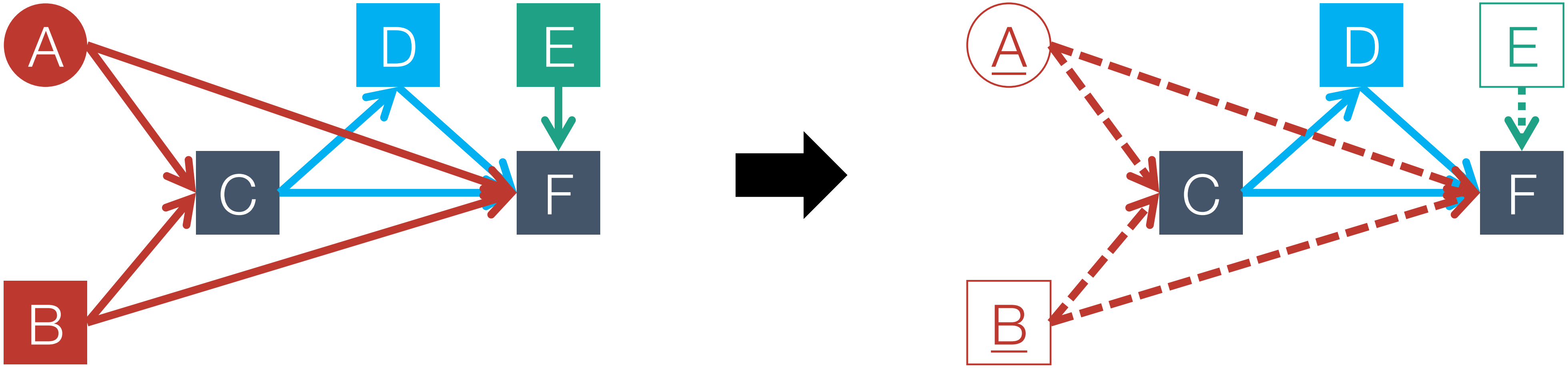
Example: Collider Bias

- If we condition on academic success – if you look at people who have achieved certain levels of academic success you can expect a conditional dependency – privilege will be inversely associated with talent.



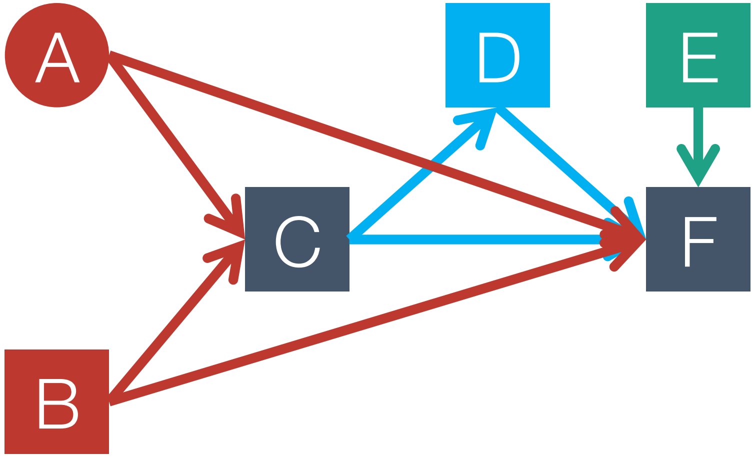
- Conditioning on success opens the backdoor path Privilege → Success ← Talent

How can we use DAGs to inform our approach to analysing data and interpreting models?



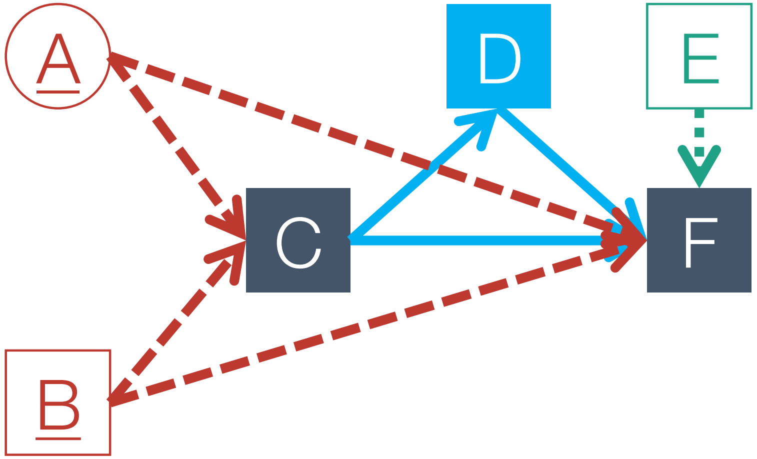
Causal effects

- To estimate the total causal effect of C on F (the *focal relationship'):*
  - We want all causal paths to be open
  - We want all confounding paths to be closed
  - So, we will need to identify and condition on all confounders



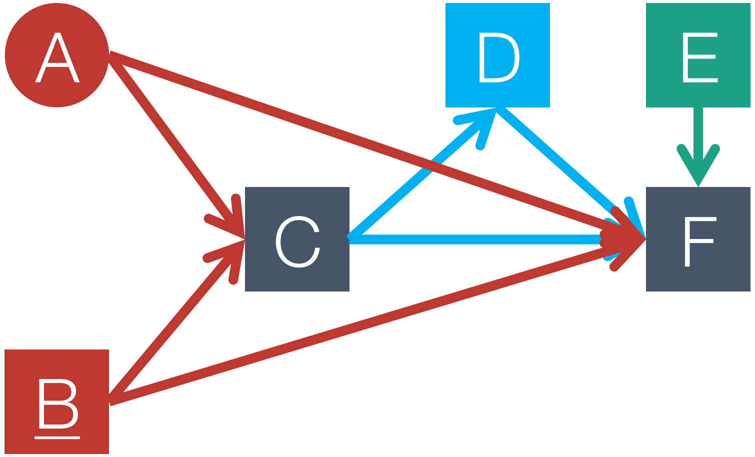
Causal effects

- To estimate the total causal effect of C on F (the *focal relationship'):*
  - We want all causal paths to be open
  - We want all confounding paths to be closed
  - So, we will need to identify and condition on all confounders



Contextual variables

- The role of each variable (and which require conditioning) is defined by their relationship to your focal relationship of interest
  - Considering total causal effect of C on F



Mediator
Channels part of the total causal effect of C on F

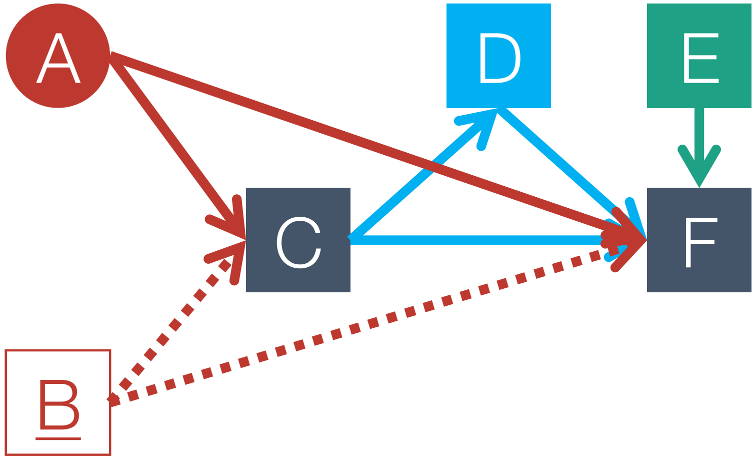
Competing exposure
Does not bias association, but source of heterogeneity in F

Confounders
Create indirect associations between C on F

Unobserved cofounding

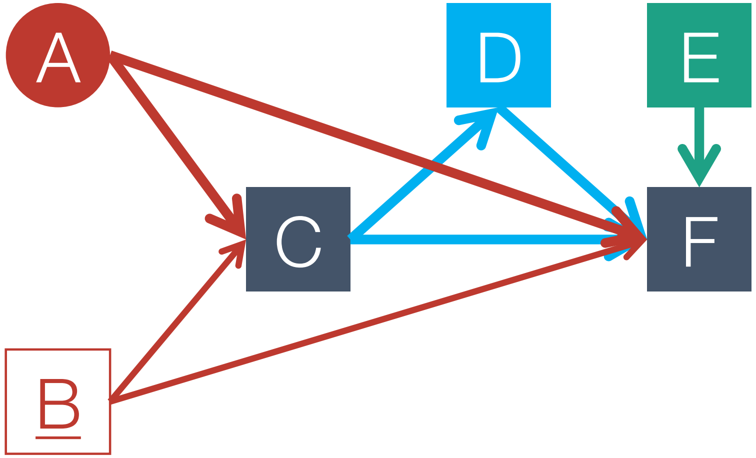
- Unobserved confounding is created by confounders that we have not measured and cannot condition on
    - If we measure it but don't condition on it, it's uncontrolled confounding

A is unobserved so cannot be conditioned, creating unobserved confounding



Residual cofounding

- There is always residual confounding after conditioning, because you can never measure a concept/variable perfectly



Estimating total causal effects

DO: condition on confounders to block confounding paths

DO NOT: condition on mediators as this will block causal pathways

OPTIONALLY: condition on competing exposure to improve precision of your estimates

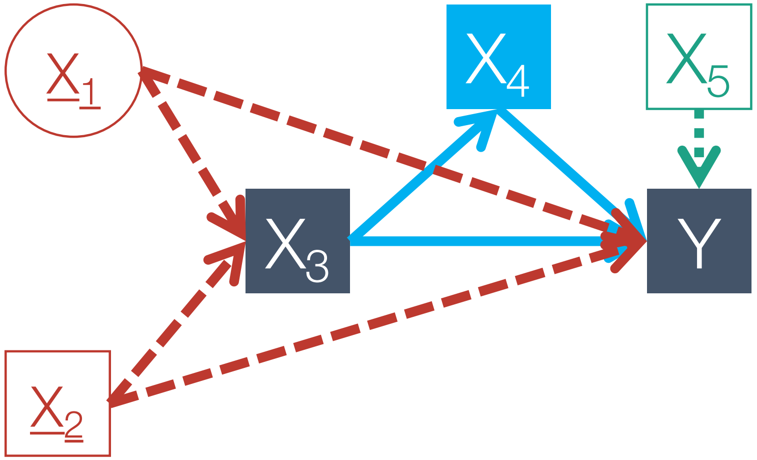
Estimating total causal effects

Total causal effect of X_3 on Y:

Model should include confounders X_1 and X_2

and competing exposures X_5

$Y \sim X_3 + X_1 + X_2 + X_5$



Estimating total causal effects

In the model of the total causal effect of X_3 on Y:

One should not interpret coefficients for other covariates X_1, X_2 , X_5

that would require different adjustment sets!

$Y \sim X_3 + X_1 + X_2 + X_5$

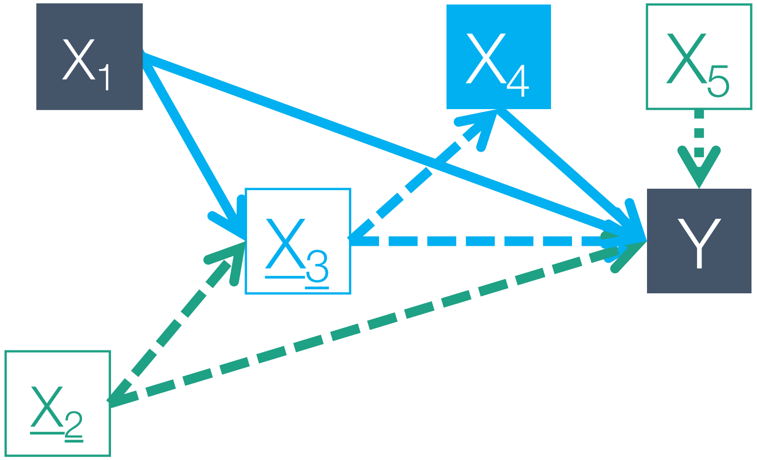


Table 2 Fallacy

DAGs process

1. Define your research question (state your focal relationship)
2. Consider and state your context
3. Draw your DAGs as early as possible, get help to revise
4. Include all relevant variables
5. DAGs in temporal order
6. Draw forward arcs, unless confident otherwise
7. Check & update your DAGs against your data
8. Use your DAGs to inform and interpret your model
9. Share your DAG

Summary

- Causal diagrams like DAGs are helpful to identify our assumptions about how different variables influence each other and plan an appropriate analysis
- By forcing us to make those assumptions explicit, they offer a huge advance in transparency over ‘black box’ or post-hoc approaches to data analysis
- By encouraging us to state our causal aims and what we want to estimate, and focusing on interval estimation, big step towards reproducible and replicable research
- DAGs are helpful in identifying, understanding, and avoiding common analytical pitfalls such as Table 2 fallacy

A word on causality in deep learning

Any Questions?