Day 2: Policy evaluation and impact assessment

Regression and DAGs

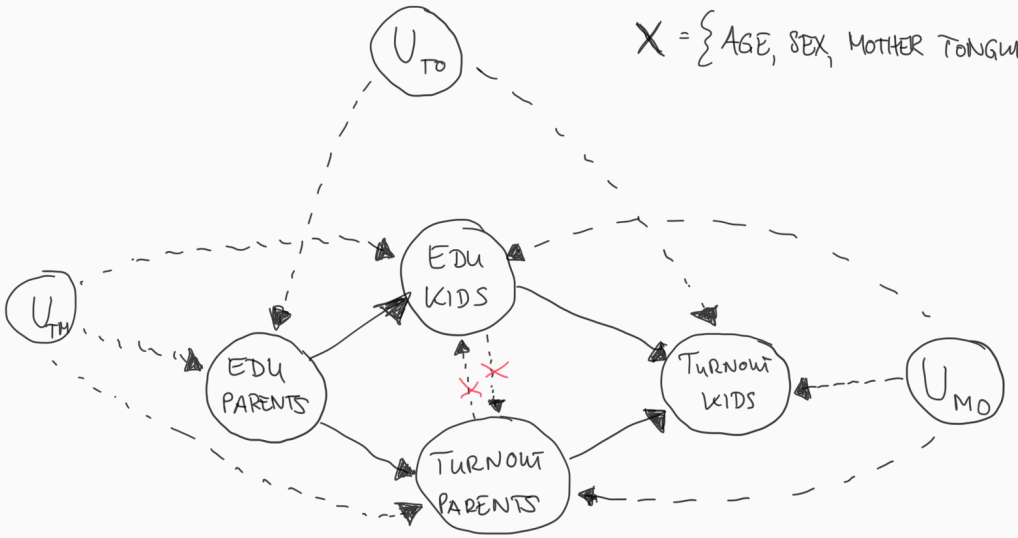
Simon Munzert
Hertie School

Table of contents

Causal reasoning with DAGs

Your most important toolset for causal reasoning

Hertie School



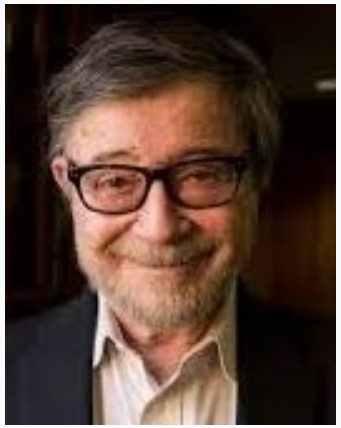
Origins of causal graphs

A pioneer of causal graphs

- Path diagrams were originally developed by **Sewall Wright** (geneticist) in the early 1920s.
- Sewell's father, Philip Wright, used them to demonstrate the use of instrumental variables regression (Pearl and Mackenzie, *Book of Why*)
- Path analysis gained popularity in social science, particularly sociology and psychology, in the form of structural equation modeling.

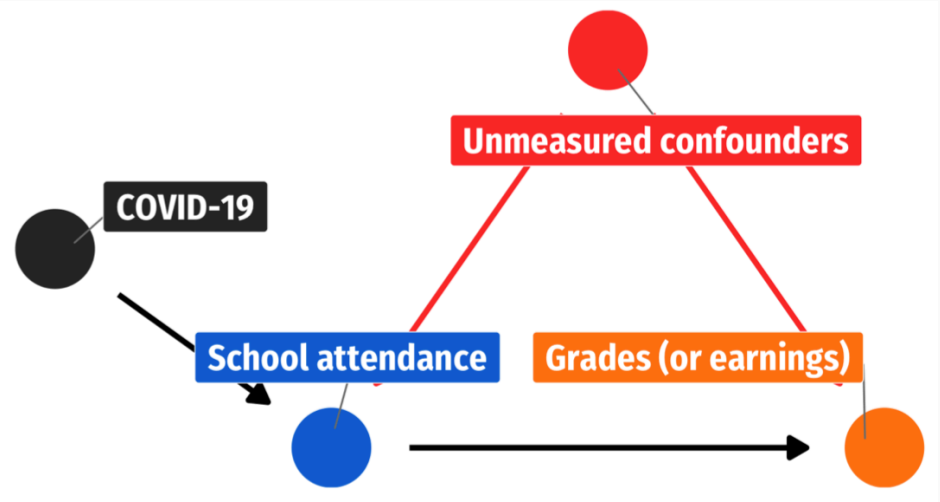
A godfather of causality

- Up until today, causal graphs are not very popular in economics, where the potential outcomes framework (POF) dominates
- Computer scientist **Judea Pearl** developed causal inference framework based on directed acyclic graphs (DAGs); widely considered godfather of causal graphs analysis



Why causal graphs are useful

- Graphs are often used informally to express beliefs about relations among variables in an intuitive manner
- Graphs allow **non-mathematicians** to draw rigorous conclusions about the nature of statistical associations
- Informal use can be expanded by adopting formal rules so that they meet the criteria for **Directed Acyclic Graphs** (Pearl 2009 [2000])
- They are practical for **choosing "control" variables** for regression/matching, assessing natural experiments, etc.
- They provide a **unified framework** to think about causal inference, sample selection, measurement error, and other methodological problems



Basic functioning of causal graphs

Variables

- Variables are also called **nodes**.
- Each variable may take multiple values, e.g. "treated" and "not treated".
- They are scale-agnostic (can be categorical, continuous).
- Usually one of the variables is treated as outcome variable of interest (often "Y").
- Variables that are unobservable or unmeasured but that play an important role in the graph should be added too (usually marked as hollow dots or circles)

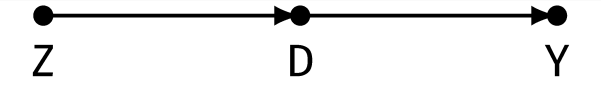
Arrows

- Connections joining variables are called **edges**.
- If an edge is an **arrow**, this implies a causal relationship between variables.
- The arrow makes a qualitative, not a quantitative statement. The effect can be positive or negative and of arbitrary size.

Basic functioning of causal graphs

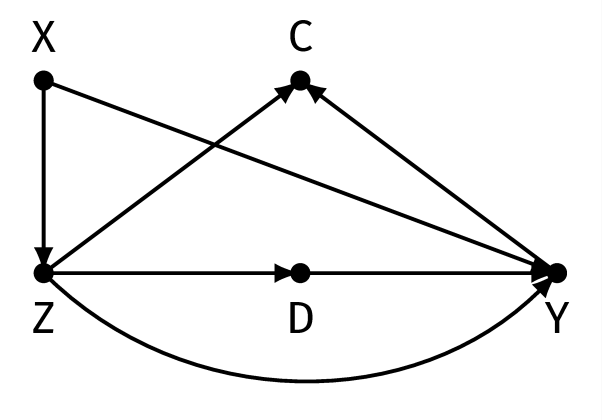
Endogeneity and exogeneity

- A variable without a parent is **exogenous**, otherwise it is **endogenous**.
- Describing a variable as truly exogenous (here: Z) is a strong statement. It means it has no cause that is of relevance here.



Paths

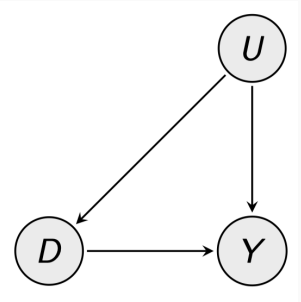
- A **path** is a sequence of arrows that links one node to another, regardless of the direction of arrowheads.
- Retracing of arrows or going through the same variable twice is not possible.
- A **causal path** is a path traced out entirely along arrows tail-to-head.

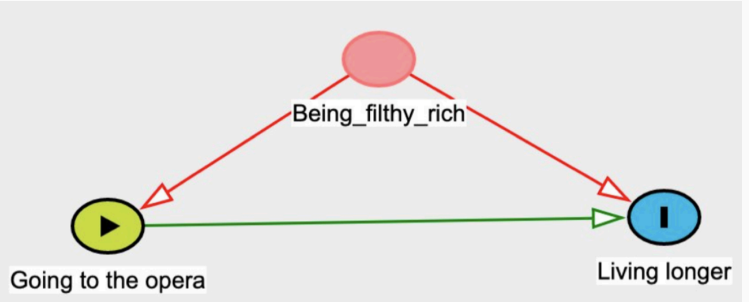


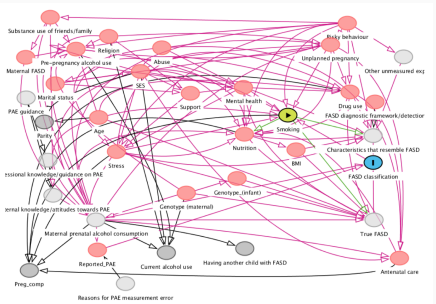
What are DAGs?

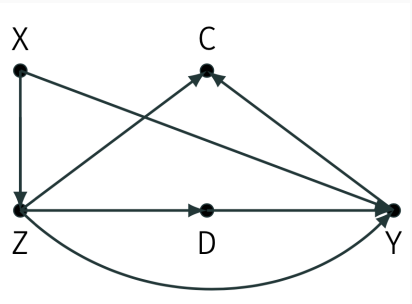
- **Directed Acyclic Graphs** (DAGs) are a type of causal graph.
- Acyclic means that there are no loops in the graph.
- DAGs encode the researcher's **qualitative** causal assumptions about the data-generating process in the population
- Constructing them requires **theoretical and empirical knowledge** or assumptions (whatever you have should go into it - theory, model, observation, experience, prior studies, intuition)

Different DAG styles





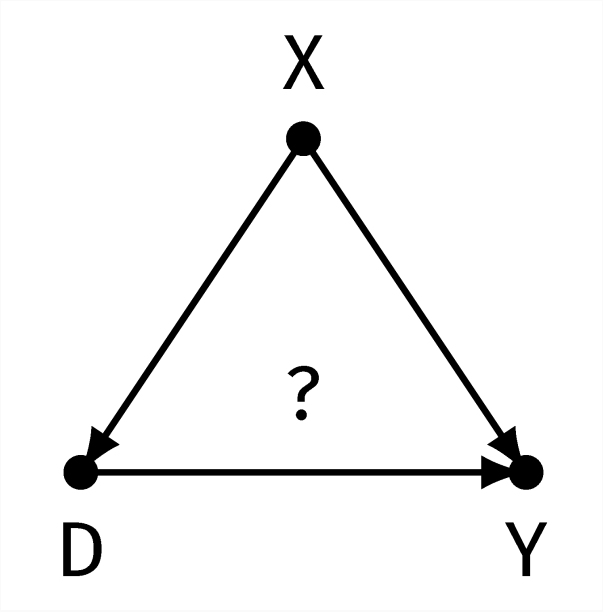




Basic patterns in DAGs

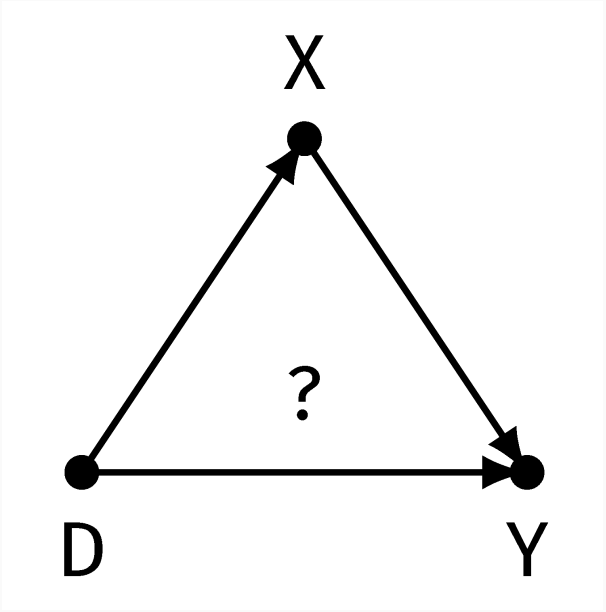
X is **confounder**.

The "fork" pattern.



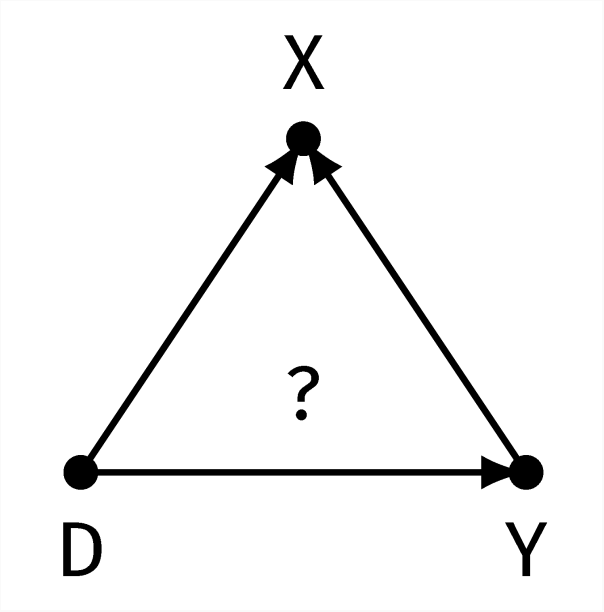
X is **mediator**.

The "chain" pattern.



X is **collider**.

The "reversed fork" pattern.

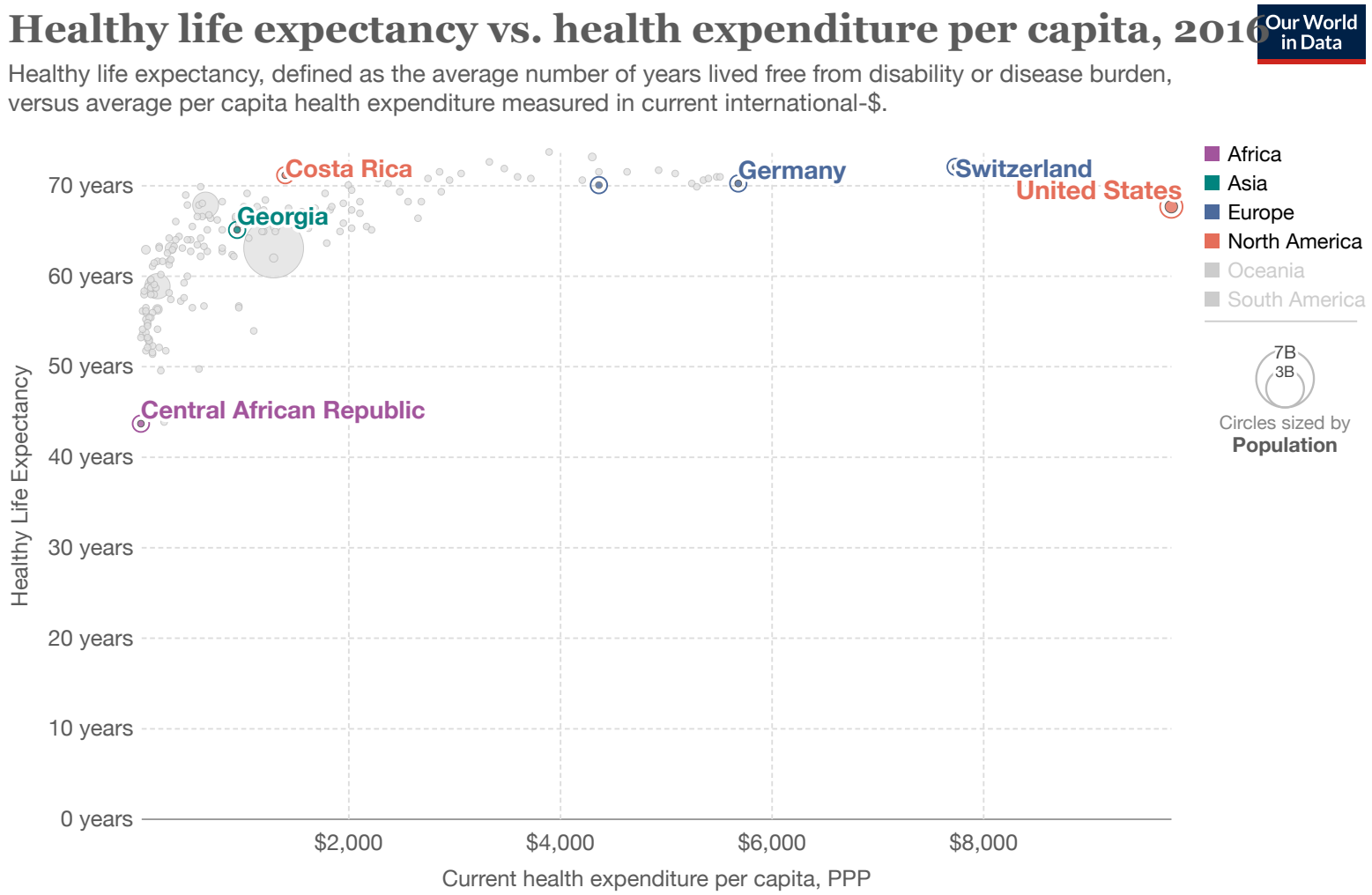


Dissecting more complex graphs into these basic patterns can help to understand the causal structure of the data-generating process. This will be **key to understanding how to adjust for confounding** in regression analysis.

Causal inference with regression

Example: health expenditure and life expectancy

Hertie School



Example: health expenditure and life expectancy

Hertie School

JAMDA

journal homepage: www.jamda.com

Original Study

How Important Are Health Care Expenditures for Life Expectancy? A Comparative, European Analysis

Wim J.A. van den Heuvel PhD [a,*], Marinela Olaroiu MD, PhD [b]

[a] SHARE Research Institute, University Medical Centre, University of Groningen, Groningen, The Netherlands
[b] Foundation Research and Advice in Care for Elderly (RACE), Spaubeek, The Netherlands

Keywords:
Health care expenditure
life expectancy
quality of care
life style
social protection

A B S T R A C T

Objectives: The relationship between health care expenditures and health care outcomes, such as life expectancy and mortality, is complex. Research outcomes show different and contradictory results on this relationship. How and why health care expenditures affect health outcomes is not clear. A causal link between the two is not proven. Without such knowledge, effects of increase/decrease in health care expenses on health outcomes may be overestimated/underestimated. This study analyzes the relationship between life expectancy at birth and expenditures on health care, taking into account expenditures of social production and education, as well as the quantity and quality of health care provisions and lifestyles.
Design: This is a cross-sectional study, analyzing national data of 31 European countries. First, the bivariate correlation between the dependent variable and independent variables are calculated and described. Next a forward linear regression analysis is applied.
Measurement: The data are derived from standardized, comparative data bases as available in the Organisation for Economic Co-operation and Development and Eurostat. Health care expenditures are assessed as a percentage of the Gross Domestic Product (GDP).
Results: Health care expenditures are not the main determinant of life expectancy at birth, but social protection expenditures are. The regression analysis shows that in countries that spend a high percentage of their GDP on social protection, that have fewer curative beds and low infant mortality, whose citizens report fewer unmet health care needs and drink less alcohol, citizens have a significant longer life expectancy.
Conclusion: To realize high life expectancy of citizens, policy measures have to be directed on investment in social protection expenditures, on improving quality of care, and on promoting a healthy life style.

Table 1
Bivariate Pearson Correlations Between LEaB and Expenditure Indicators, in 2013

	Expenditure Indicators as % of GDP in 2013 on		
	Health Care	Social Protection	Education
LEaB	0.700*	0.747*	0.549*

*Significant at $P < .01$ level.

Table 2
Bivariate Pearson Correlations Between LEaB and Quantitative Health Care Indicators, in 2013

	Quantitative Health Care Indicators Per 100,000 Inhabitants Number in 2013				
	Curative Beds	Long-Term Beds	Practicing Doctors	General Practitioners	Nursing and Caring Personnel
LEaB	−0.578†	−0.231	0.139	0.302	0.447*

*Significant at $P < .05$ level.
†Significant at $P < .01$ level.

Table 5
Final Model of Forward Linear Regression Analysis

Indicators	Standardized Beta	Significance Level	Adjusted Explained Variance
Social protection	0.416	.000	84%
Number curative beds per 100,000 inhabitants	−0.193	.042	
Infant standardized mortality rate	−0.233	.013	
% unmet health care needs	−0.232	.008	
Alcohol consumption in liters per inhabitant	−0.280	.004	

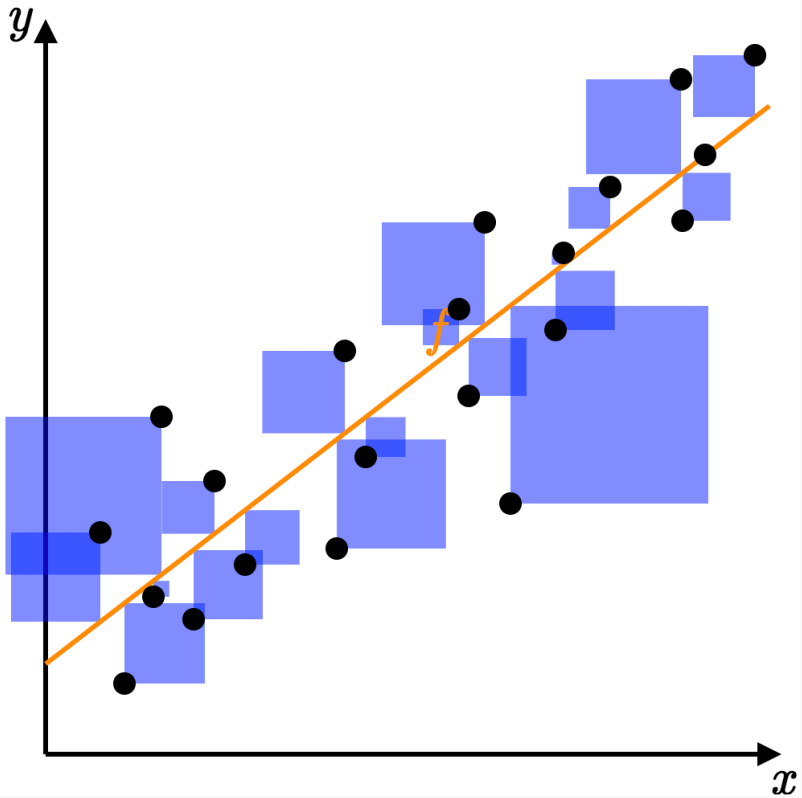
Regression mechanics

Ordinary Least Squares

- **OLS** is a method for estimating the unknown parameters in a linear regression model.
- It addresses a simple mechanical problem: how to minimize the sum of the squared differences between the observed and predicted values?
  $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$
- The solution is to find the values of the coefficients that minimize the sum of squared residuals.

Relevance for policy analysis

- Crude but simple model of a relationship between X and Y: a linear fit
- "What is our best guess for Y (outcome, KPI) given particular value of X (policy)?"



Regression mechanics

Bivariate regression

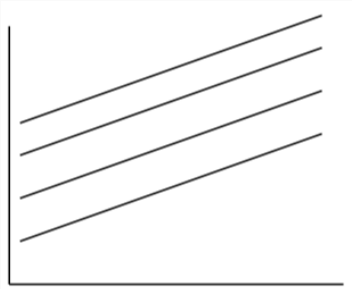
- In bivariate regression, the formula for the slope is $\hat{\beta}_1 = \frac{cov(x,y)}{var(x)} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$
- The intercept can then be derived as $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- Compare with correlation: bivariate regression gives us more precise information on the strength of a relationship

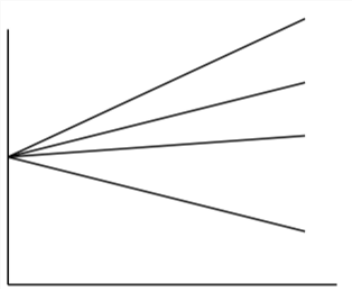
Multiple regression

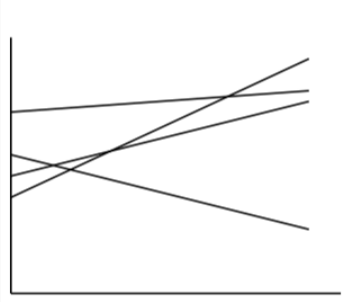
- With multiple regression, we choose the line that minimizes: $\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik})^2$
- This allows for "controlled comparisons": we can isolate the statistical effect of one variable while holding others "constant"
- Whether this is a causal effect depends on the causal structure of the data-generating process (see later)

Intercept and slope in action

Which parameter(s) are changing?







Example

- What's the effect of education on income?
- Income measured in hourly wage
- Education measured in years
- "Slider/dimmer" logic

	Hourly wage
Education	0.541***
	(0.053)
Intercept	−0.905
	(0.685)
N	526
R^2	0.165
Adjusted R^2	0.163

*p < .1; **p < .05; ***p < .01

Example

- And what's the effect of gender on income?
- Income measured in hourly wage
- Gender measured binary (female vs. non-female)
- "Switch" logic

	Hourly wage
Female	−2.211***
	(0.276)
Intercept	6.835
	(0.665)
N	526
R^2	0.125
Adjusted R^2	0.121

*p < .1; **p < .05; ***p < .01

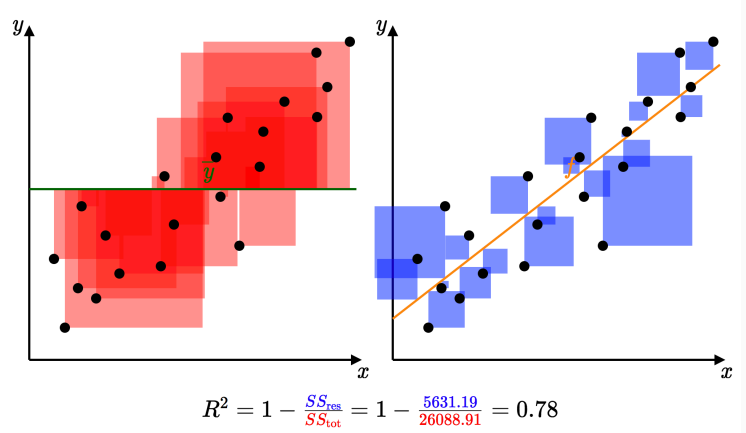
Example

- And what's the effect of education, gender, and ethnicity on income?
- Predictor variables measured as before
- Education measured in years

	Hourly wage
Education	0.505***
	(0.051)
Female	−2.275***
	(0.279)
Nonwhite	−0.119
	(0.460)
Intercept	0.650
	(0.681)
N	526
R^2	0.259
Adjusted R^2	0.255

R-Squared

- Some people are obsessed with R^2. Is that justified?
- R^2 is a PRE (proportional reduction of error) measure.
- It compares the **total sum of squares (TSS)** from the mean to the **explained sum of squares (ESS)** improvement from the OLS fit.
- R^2 is the proportion of the variance in the dependent variable that is predictable from the independent variable(s):
  $R^2 = \frac{ESS}{TSS} = \frac{\sum(\hat{y}_i - \bar{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$
- Here is an interactive visualization.
- Adjusted R^2 (aka $\bar{R}^2$) corrects for the number of parameters estimated in a model but loses the nice interpretation of R^2.



Why R-Squared is of limited use for impact evaluation

- Usually the interest is not in predicting the outcome variable but in estimating the causal effect of a treatment on an outcome.
- R^2 can be high even if the treatment effect is zero.

Variable selection for regression

Back to the original problem

- We want to estimate the causal effect of a treatment on an outcome.
- Observational data is often not that useful evidence for that purpose because correlations can be misleading.
- Correlational patters should **not** serve as input for evidence-based policymaking.

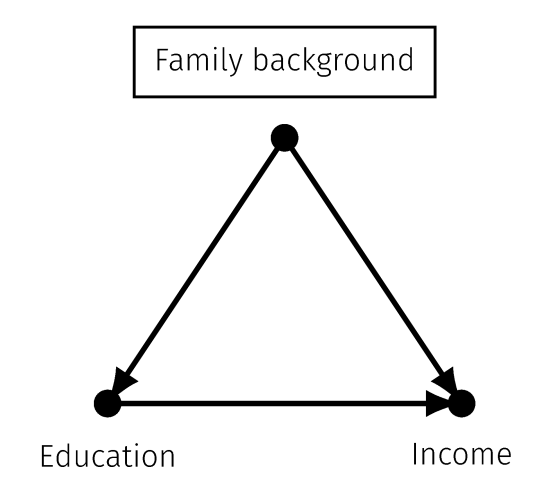
Is regression-based inference with observational data pointless?

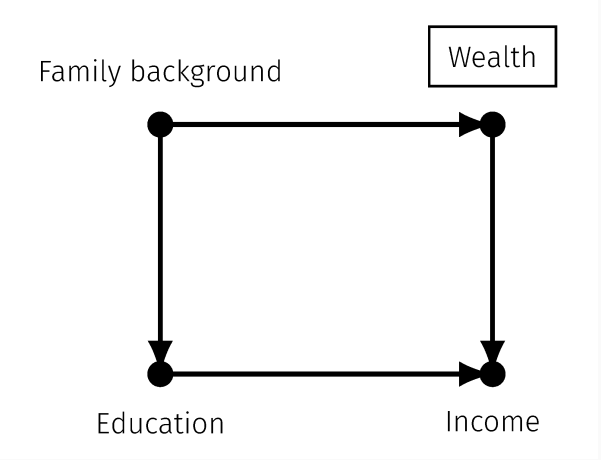
- Not necessarily. If we **control for all relevant confounders**, we can still isolate the effect of interest.
- This gives us three important tasks:
  1. Identifying the causal structure between all variables affecting the relationship between treatment and outcome of interest.
  2. Identifying all variables that need to be adjusted for to estimate the causal effect, and those that must not be adjusted for (mediators, colliders).
  3. Measure what is necessary and run the regression.
- **This is where DAGs come in**: they help us deal with the first two tasks.

Confounders

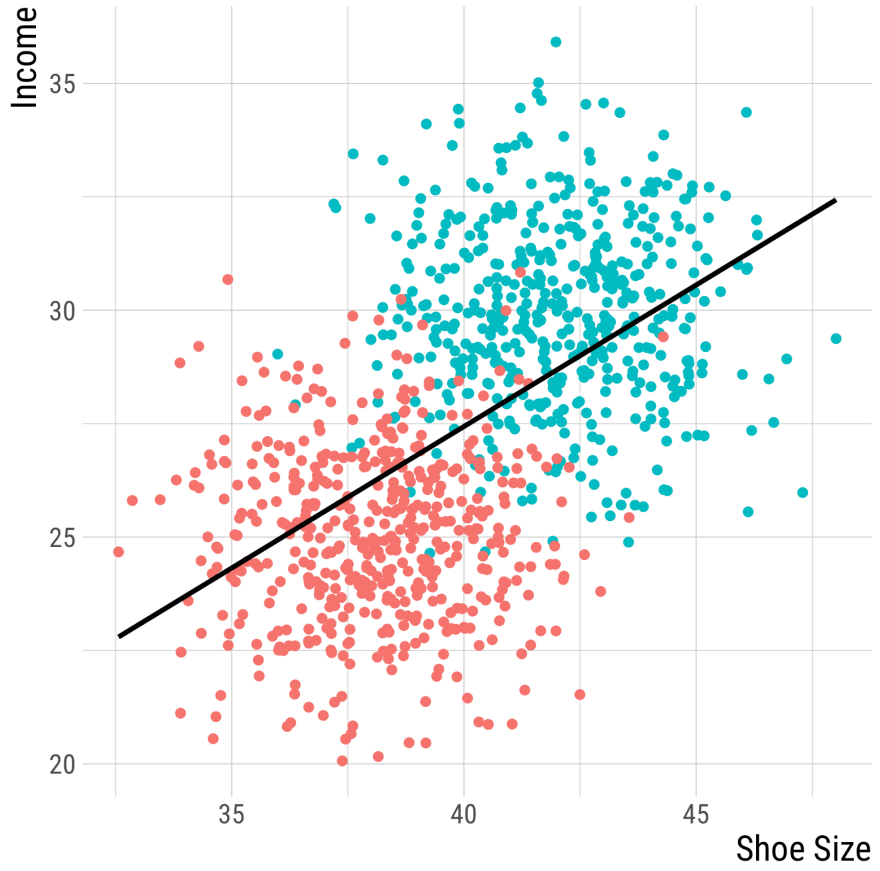
Dealing with confounders

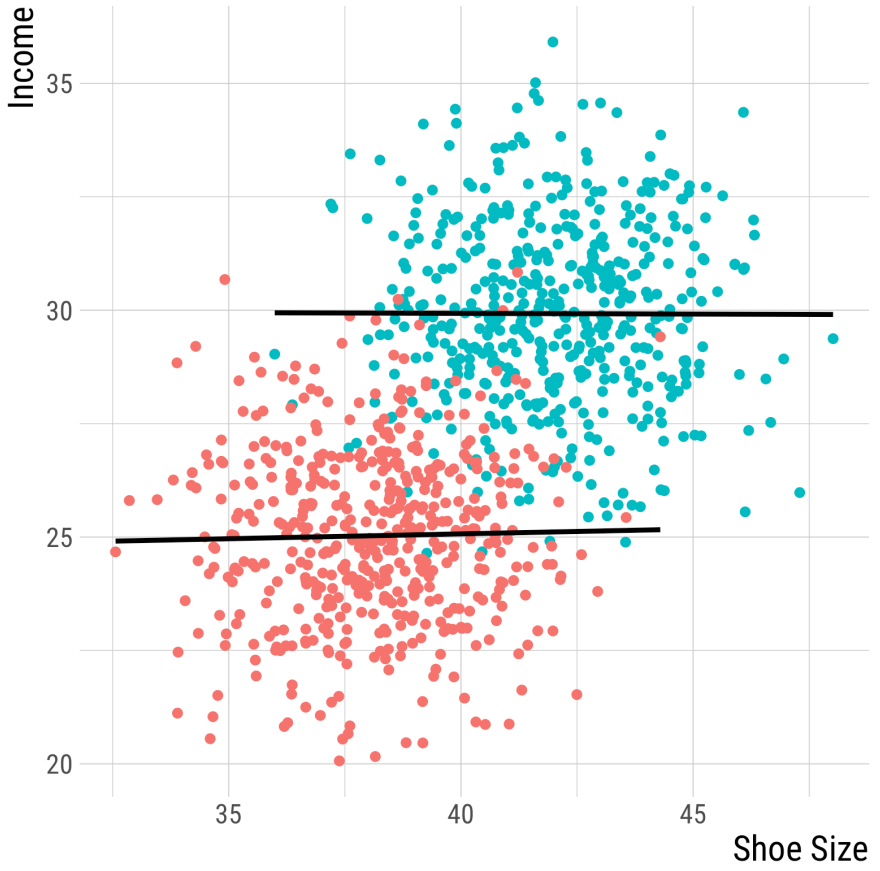
- A confounder induces statistical association between its effects.
- **Conditioning** on a confounder (or a descendant of a confounder) on the path **blocks the path**.
- In DAGs, conditioning is usually indicated with a **box around the variable**.
- Failing to condition on a relevant confounder induces non-causal statistical association or **omitted variable bias**.





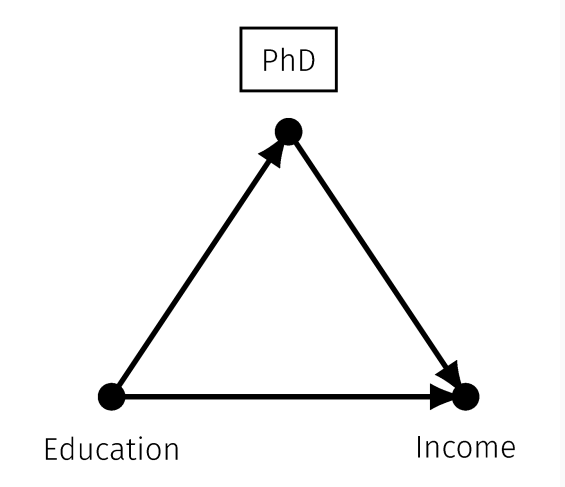
Conditioning on a confounder: example





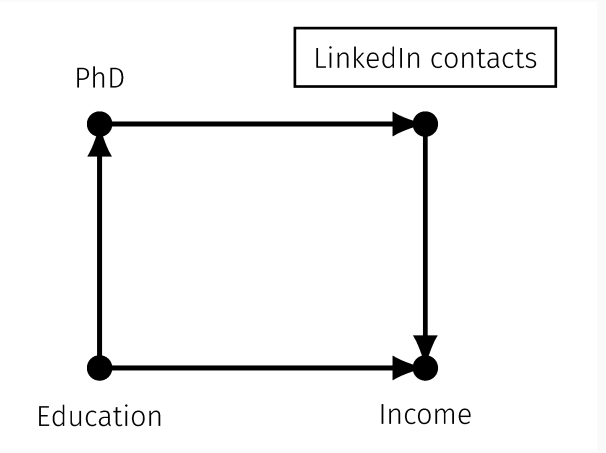
Dealing with mediators

- Chains of mediation D $\rightarrow$ M $\rightarrow$ Y let us express how exactly a treatment impacts the outcome.
- Mediators express a **mechanism** by which one variable affects another.
- Example: Coffee (D) leads to better stats exam results (Y) because the caffeine (M) stimulates the nervous system and prevents drowsiness.



Beware of post-treatment bias

- Conditioning on a mediator on the path blocks the path.
- Often, we are interested in total effects (e.g., ATE) including both direct and indirect routes.
- Conditioning on a mediator may thus lead to **overcontrol or post-treatment bias**.
- "Controlling away" for consequences of the treatment!



Variable selection - post-treatment bias

Post-treatment bias: avoidable

- Causal effect (CE) of party ID on the vote
  - Do control for race
  - Do not control for short-term voting intentions
- CE of nuclear desaster on attitudes towards nuclear energy
  - Do control for political context
  - Do not control for media coverage

Post-treatment bias: unavoidable?

- CE of democratization on civil war: **control for GDP**?
  - Yes, since GDP $\rightarrow$ democratization
  - No, since democratization $\rightarrow$ GDP
- CE of education on income: **control for IQ**?
  - Yes, since IQ $\rightarrow$ education
  - No, since education $\rightarrow$ IQ

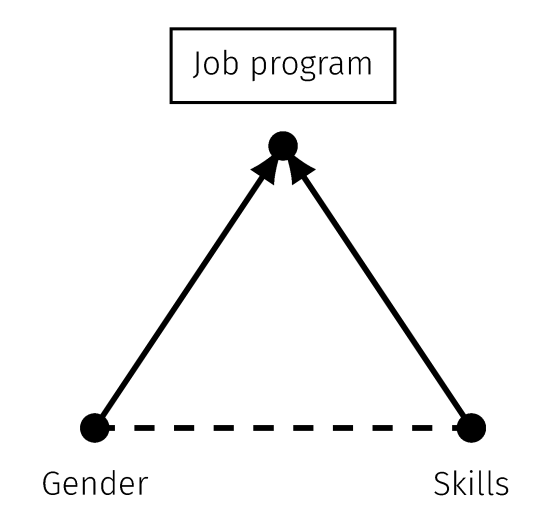
Colliders

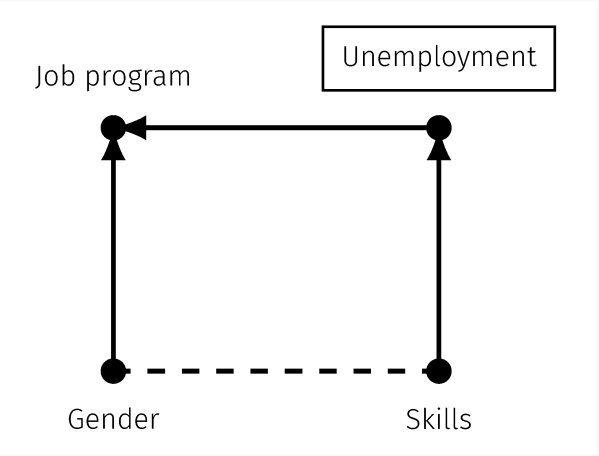
What are colliders?

- Colliders are variables that are caused by two other variables.
- They are the opposite of mediators: they block the path between two variables.

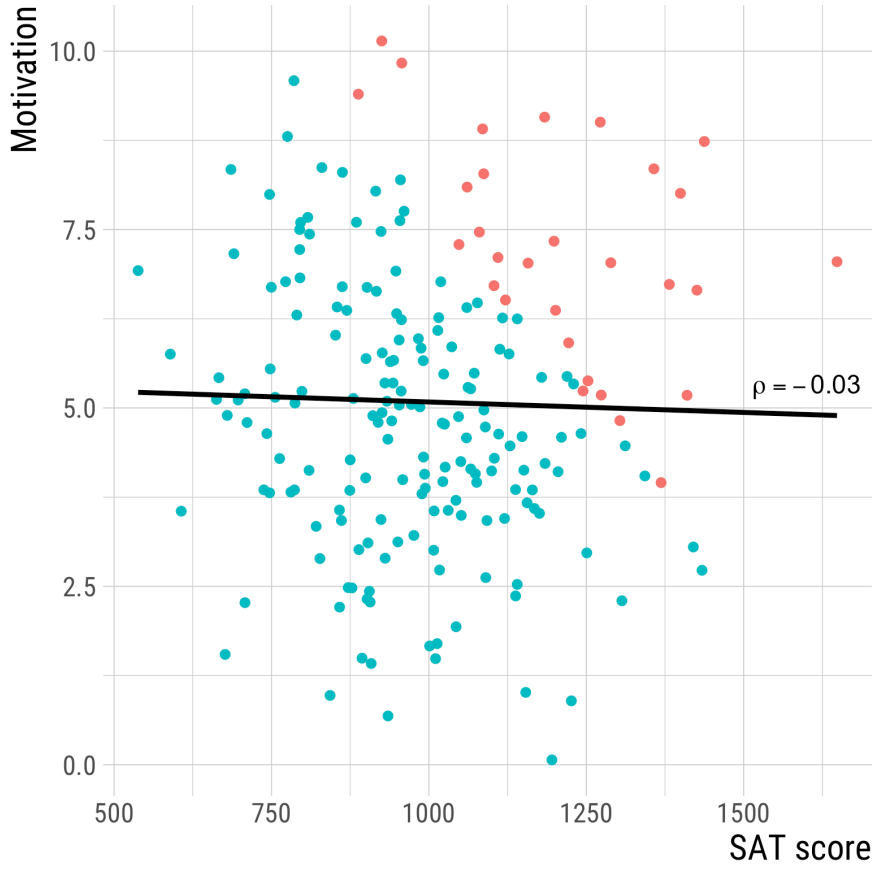
Conditioning on colliders

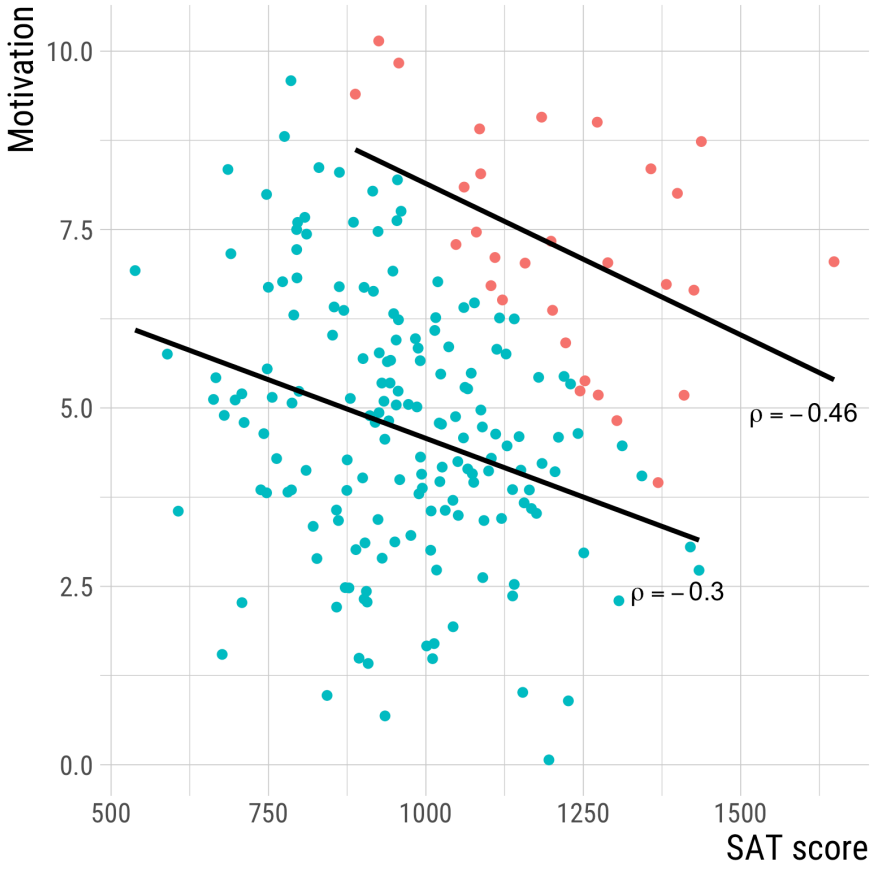
- Conditioning on a collider (or a descendant of a collider) opens the path between its causes.
- This can lead to **spurious association** between the causes, which is called **collider bias** or **endogenous selection bias**.
- This type of bias is much less well known than omitted variable bias but can be fatal in causal inference.





Conditioning on a collider: example

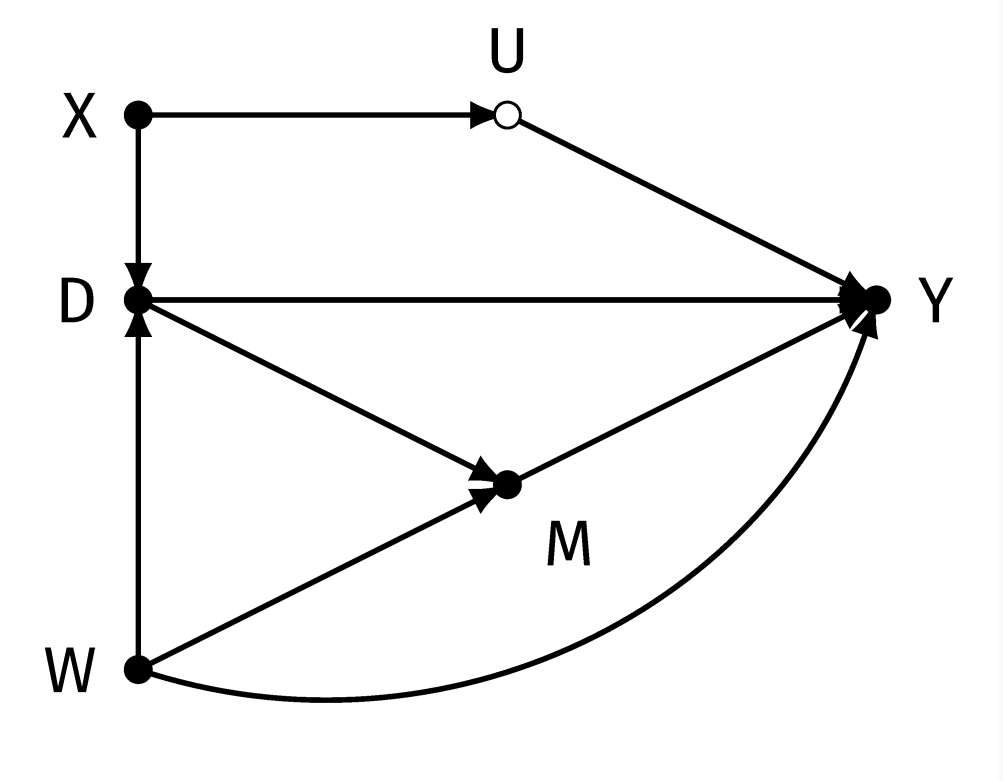




Example: Y motivation, D SAT score, X admission to college (red: admitted, blue: rejected)

Selecting relevant covariates

1. Draw a DAG using your theoretical and empirical knowledge
2. Find the causal and non-causal paths between treatment and outcome
3. Identify conditions that satisfy the back-door criterion with regards to treatment and outcome
4. Include these variables into the model specification
5. Withstand the temptation to give any other coefficient than that for the treatment a causal interpretation - the status of covariates is path-specific!



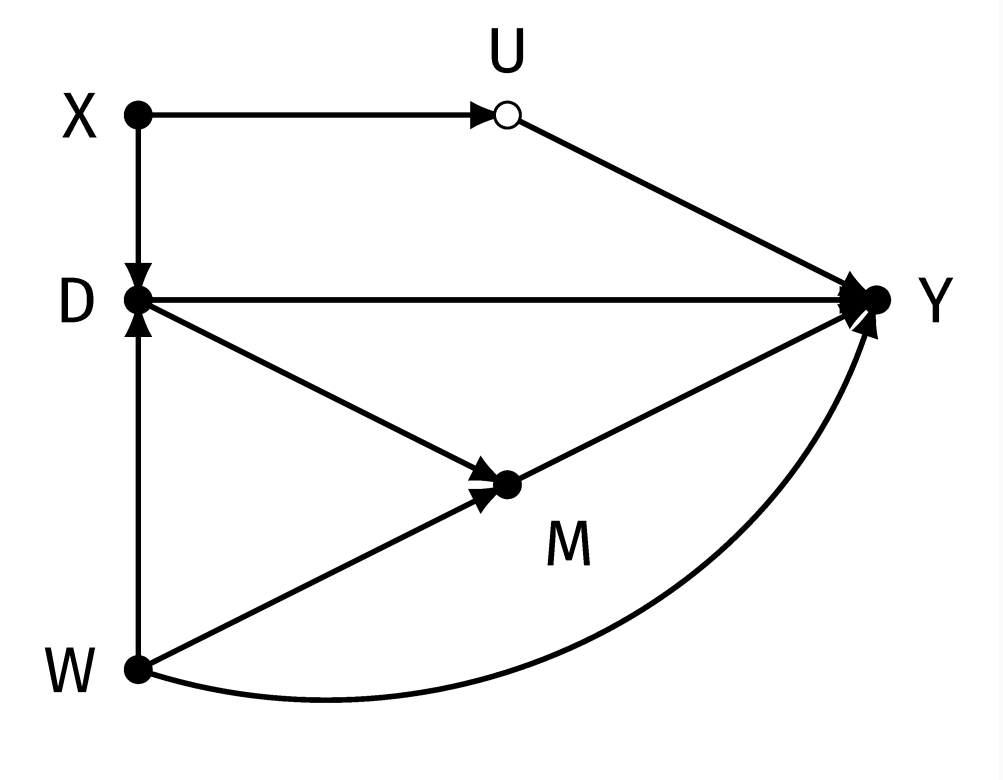
The back-door criterion

A set of observed variables Z satisfies the back-door criterion relative to the total causal effect of a treatment on an outcome if

- Z blocks all back-door paths from treatment to outcome and
- No variable in Z lies on or descends from a causal path from treatment to outcome

Example

- What variables should we control for to identify the effect of D on Y?



"Does going to the opera make you live longer?" revisited

Hertie School

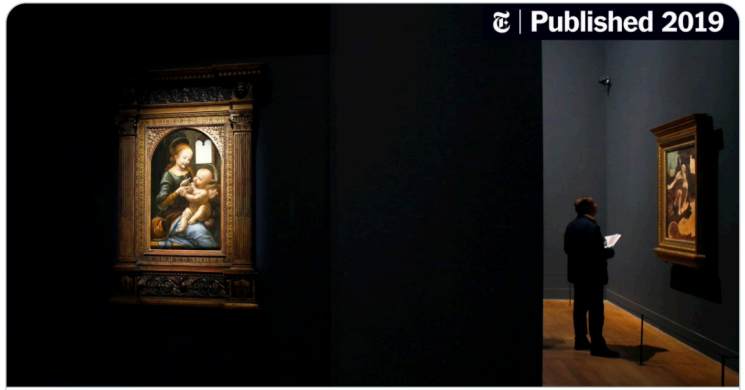
From the paper:

> "We adjusted models for **demographic variables** (age, sex, marital status, ethnicity, educational qualifications, wealth, employment status, and occupational status); **health related variables** (eyesight, hearing, depressive symptoms, other psychiatric conditions, diagnosis of cancer, lung disease or cardiovascular disease, history of any other long-term condition, smoking, alcohol consumption, sedentary behaviours, mobility, problems in undertaking activities of daily living, osteoporosis, and cognition); and **social covariates** (loneliness, number of close friends, living alone, frequency of civic engagement, frequency of social engagement, and whether participants had a hobby or pastime)."

"Does going to the opera make you live longer?" revisited

NYT Health
@NYTHealth

Want to live longer? Try going to the opera. Researchers in Britain have found that people who reported going to a museum or concert even once a year lived longer than those who didn't.



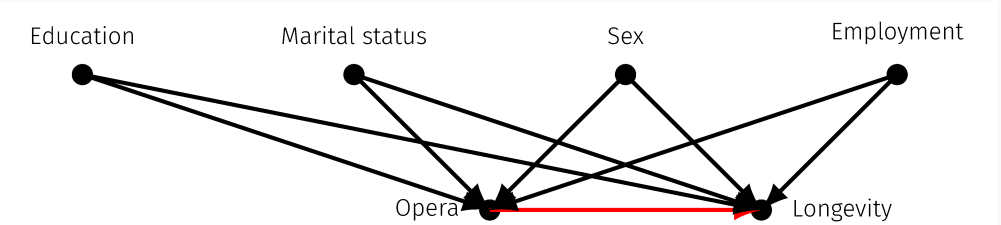
Another Benefit to Going to Museums? You May Live Longer (Published 2019)
Researchers in Britain found that people who go to museums, the theater and the opera were less likely to die in the study period than those who didn't.
nytimes.com

3:19 PM · Dec 22, 2019 · SocialFlow

318 Retweets **1,705** Quote Tweets **1,239** Likes

What's the underlying causal reasoning here?

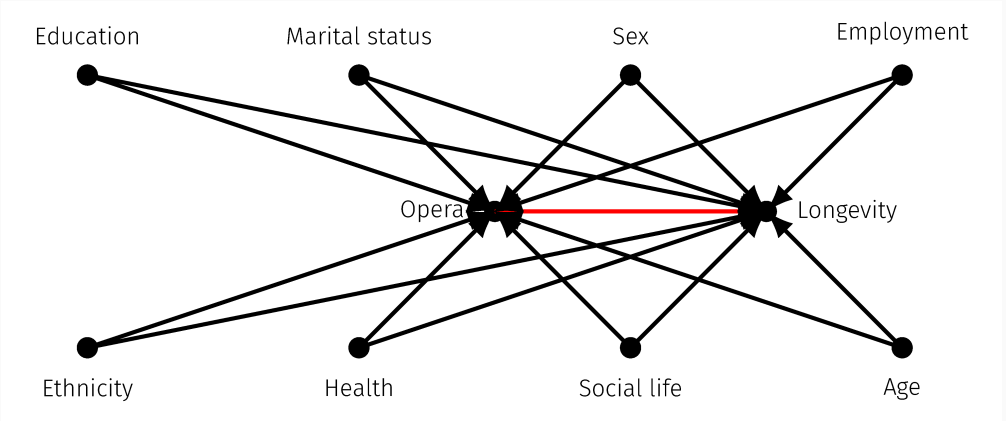
- What's our idea of how all variables are related?



"Does going to the opera make you live longer?" revisited

What's the underlying causal reasoning here?

- What's our idea of how all variables are related?
- But that's not nearly the end of the story.
- Think about other variables (observed and unobserved). What would this imply for our causal inference?



Lessons learned

On regression for impact evaluation

- The regression way of estimating causal effects is a powerful tool is (1) to find the factors responsible for a priori differences between groups (confounders), and to include these variables in the equation, (2) in the hope that an unbiased estimate is obtained
- DAGs are a powerful tool to find the set of confounders we need to control for, but they're not a panacea

On selecting relevant covariates

- Specifying regression models for causal inference is a **theoretical**, not a statistical problem
- That's why our robot overlords are not going to take over any time soon - inferring from experience, drawing on knowledge of related studies, thinking hard about the missing components of the treatment selection mechanism - that's what we (as scientists, policy-makers) are good at!
- But don't make it a mechanical exercise

> "Regression models make it all too easy to substitute technique for work (...) Regression models often seem to be used to compensate for problems in measurement, data collection, and study design. By the time the models are deployed, the scientific position is nearly hopeless." (Freedman, 1991)