

Week 10

Questions of Causation

Paul Laskowski and Alex Hughes

January 12, 2023

UC Berkeley, School of Information

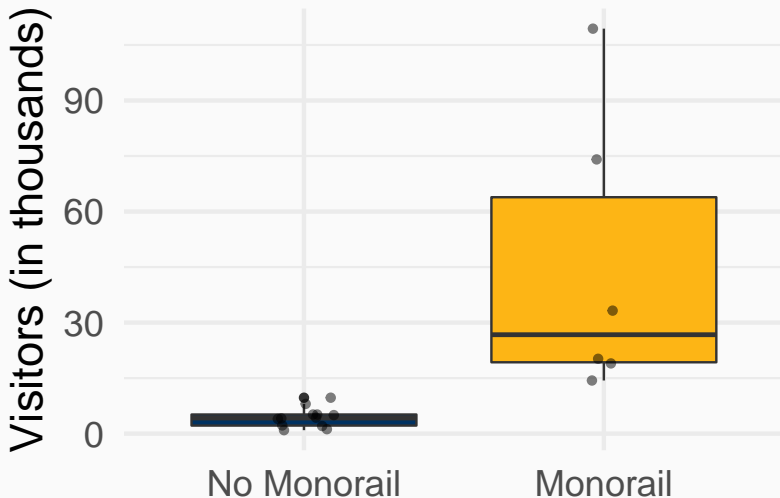
Questions of Causation

SOME BUSINESS QUESTIONS

- What will happen to coffee sales if we buy a new roaster?
- Will profits be higher if we design a new jet or upgrade our existing one?
- Will more people visit our amusement park if we add a monorail?

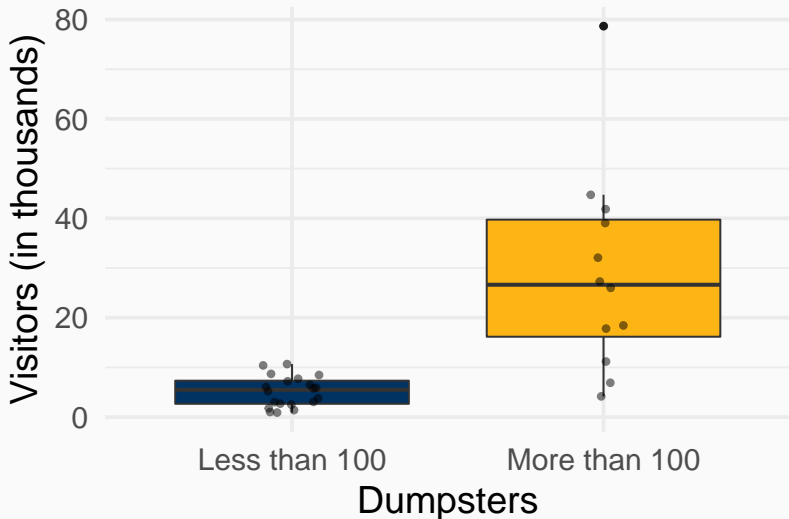
AMUSEMENT PARK DAILY VISITORS

Monorails Increase Visitors?



AMUSEMENT PARK DAILY VISITORS (CON'T)

Dumpsters Increase Visitors?



Correlation \neq Causation

Explanatory modeling: How can we test or estimate an effect in a causal theory?

Unit Plan

PLAN FOR THE WEEK

Three sections

1. What is explanatory modeling?
2. The one-equation structural model
3. Common violations of the one-equation model
 - Confounding, omitted variable bias
 - Outcome on the RHS
 - Simultaneity bias

PLAN FOR THE WEEK (CONT.)

At the end of this week, you will be able to:

- Recognize major strategies for estimating causal effects
- Understand the assumptions behind the one-equation structural model
- Reason about common violations of the one-equation structural model

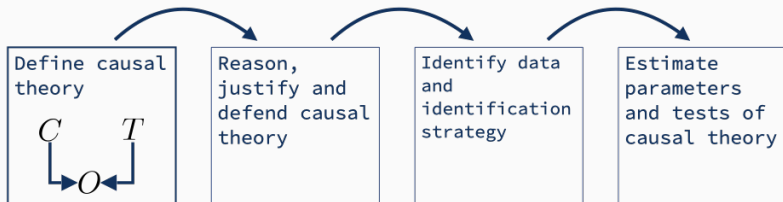
What Is Explanatory Modeling?

TOWARD EXPLANATION

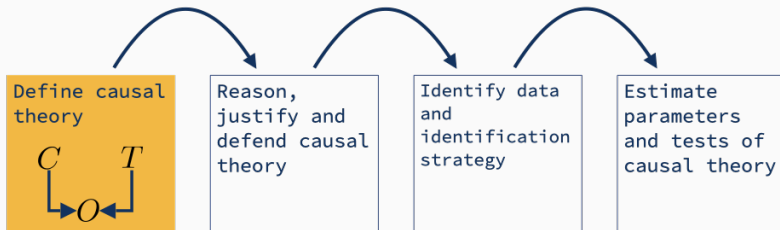
What extra assumptions are needed for OLS regression coefficients to be causal?*

* Misleading question

THE EXPLANATORY MODELING WORKFLOW



THE EXPLANATORY MODELING WORKFLOW



WHAT IS CAUSAL THEORY?

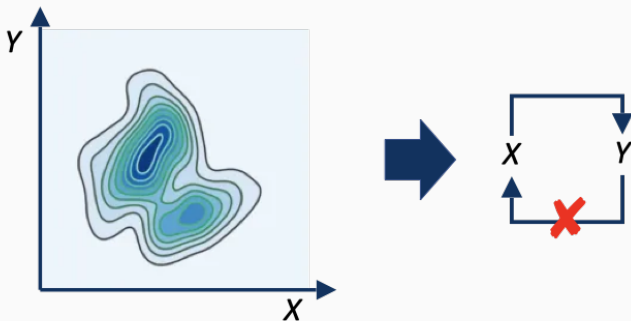
Causal theory

A *causal theory* is a statement of beliefs about what concepts *do* and what concepts *do not* cause other concepts.

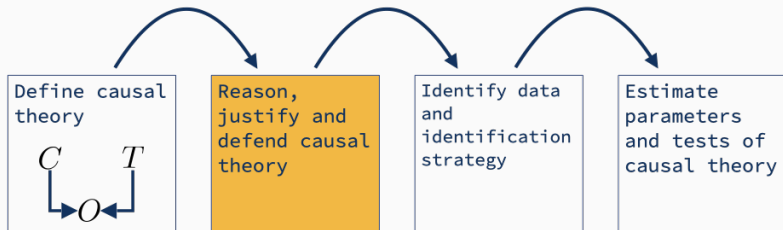
- Objective: narrow the range of causal explanations for associations we find in data.

WHAT IS CAUSAL THEORY? (CONT.)

- Joint distributions and cumulative density functions cannot identify causal information
- If we begin with causal statements, we can use logic to reach causal conclusions



HOW TO REASON ABOUT A CAUSAL THEORY

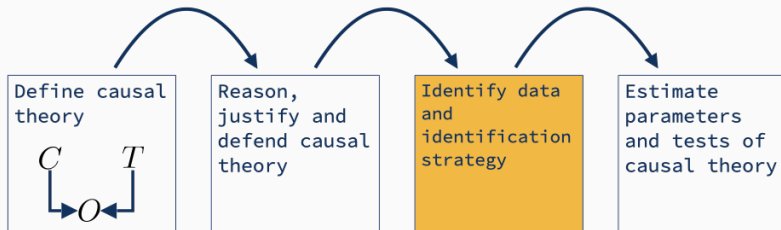


HOW TO REASON ABOUT A CAUSAL THEORY (CONT.)

Creating and eliminating possible causal paths

- *Time structure*
 - If X happens after Y, then X cannot have caused Y.
- *Domain Knowledge*
 - Germ theory of infections disease
 - Often formed through past experiments
- *Effectively "random" events*
 - Coin flips
 - Tropical storms
 - pseudorandom generators

HOW TO IDENTIFY AN IDENTIFICATION STRATEGY, PART I



HOW TO IDENTIFY AN IDENTIFICATION STRATEGY, PART II

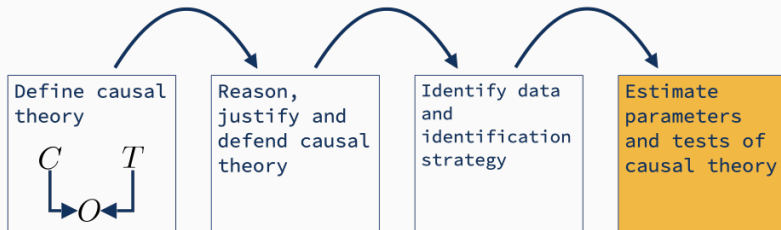
Goal: Produce a consistent estimate of the strength of the causal relationship given:

1. Causal theory
2. Data

No estimator provides estimates that *always* have a causal interpretation

- OLS Regression
- Regression discontinuity
- Diff-in-Diff
- Two-Stage Least Squares

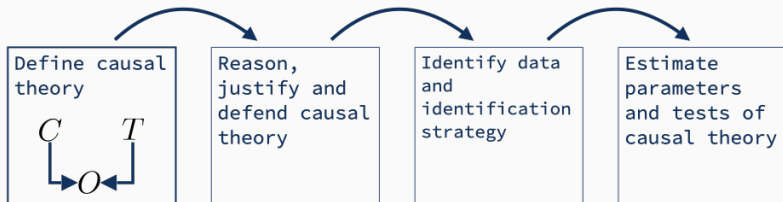
HOW TO IDENTIFY AN IDENTIFICATION STRATEGY, PART III



HOW TO ESTIMATE PARAMETERS

- Estimate model and interpret coefficients
- Return to reasoning about the causal model and possible violations

THE EXPLANATORY MODELING WORKFLOW



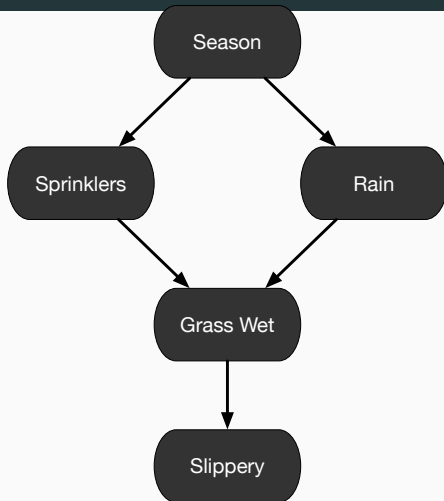
WE SHOULD HAVE AN ACTIVITY HERE

Note: We should either have a reading, or applied activity here.

- Reading activity

Pearl and Structural Equation Models

TOWARD A FLEXIBLE CAUSAL FRAMEWORK

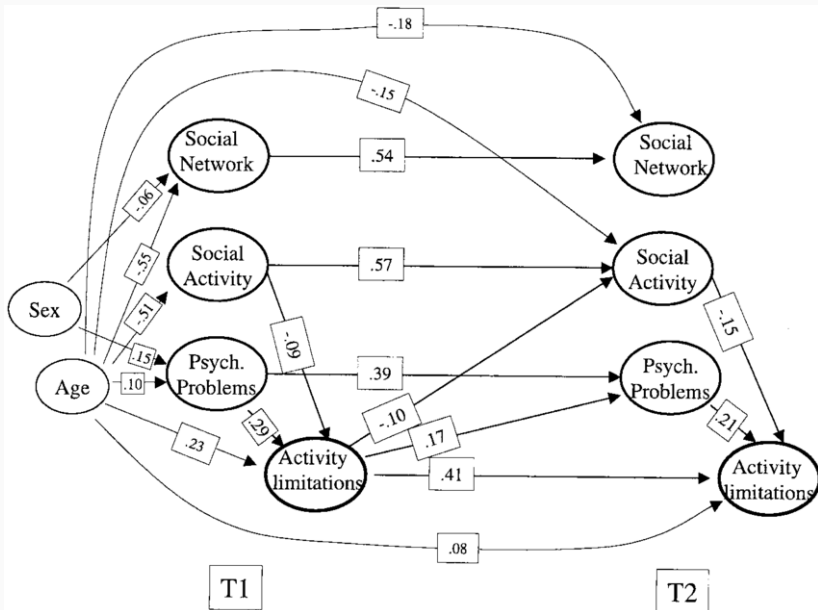


Causality (2000, 2009)



Judea Pearl

HOW TO REASON ABOUT A CAUSAL THEORY



STRUCTURAL EQUATION MODEL (SEM) BASICS

Endogenous variables

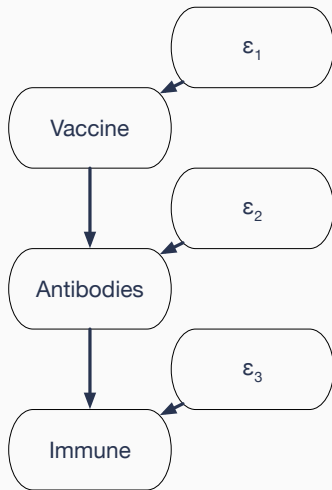
- V : Vaccine
- A : Antibodies
- I : Infection

Background variables

- ϵ : Outside causes

Structural equations

- $V = f_V(\epsilon_1)$
- $A = f_A(V, \epsilon_2)$
- $I = f_I(A, \epsilon_3)$



Pearl and Structural Equation Models

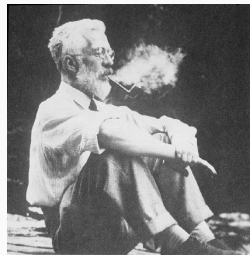
Reading: Alleged Dangers of Cigarette Smoking

READING: ALLEGED DANGERS OF CIGARETTE SMOKING

Note: This is a reading call. We're just placing it here for organization.

Read the two-page article, published in the BMJ in 1957 written by Ronald A. Fisher. Some context. R.A. Fisher is

- Perhaps the most influential statistician *of all time*
- At the very least, up there with Bayes, Neyman, and the canon.
- The student interested in a longer-form profile of this content can read the following article written by Pricenomics. [\[Link here\]](#).



Of course, smoking causes lung cancer – Fisher was dogmatic.

Pearl and Structural Equation Models

Evaluation and Execution of a Structural Equation Model

EXAMPLE: EXECUTION OF AN SEM

Note: This is a whiteboard, we're just placing it here for organization.

- What causes lung cancer?
- Coffee → Alertness → Work
- Interest → Awareness → Purchase

EXECUTION OF AN SEM

Step one

- Draw values of $\epsilon_1, \epsilon_2, \epsilon_3$.
- Assume $\epsilon_1 \dots \epsilon_k$ are independent

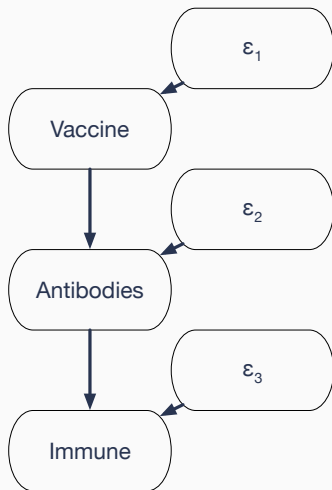
Step two

- Assign endogenous variables their values

$$V = f_V(\epsilon_1)$$

$$A = f_A(V, \epsilon_2)$$

$$I = f_I(A, \epsilon_3)$$



Statistical Implications of a Causal Graph

HOW TO APPLY AN SEM

Causal models require that we write down, clearly, the assumptions about the causal process *in the data generating process*.

- Evaluate how closely our data matches our *theory* about the world
- Choose an appropriate estimator for the data and theory

CAUSAL GRAPH DEFINITION

Causal graph

A **causal graph** is a graph that describes the causal pathways among a subset of all variables.

Causal graphs encode our theory about the causal structure:

1. If there is an arrow from X to Y , then X has a direct causal effect on Y .
2. If there is **no** arrow from X to Y , X has **no** direct causal effect on Y .
3. If X and Y have a common cause Z , then Z must be in the diagram, even if we cannot measure it.

EXAMPLES OF CAUSAL GRAPHS

Example: Direct and indirect effects

- $V \rightarrow A \rightarrow I$. Vaccines have a direct causal effect on Antibodies, but no direct effect on Infection.

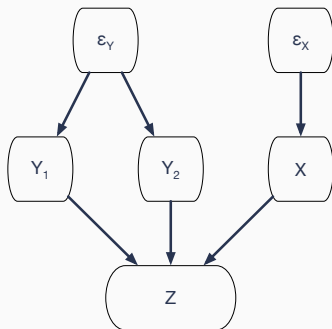
Example: Common Causes

- Education \rightarrow Wage. Do we need to include motivation?

STATISTICAL IMPLICATIONS OF A CAUSAL GRAPH

Theorem: independence

If X and Y have no common ancestors in an acyclic SEM, they are independent.



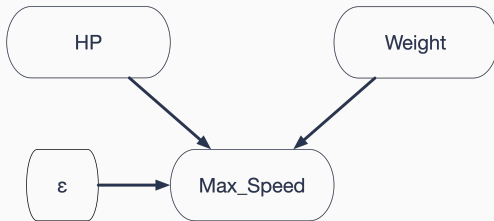
The One-Equation Structural Model

THE SIMPLEST CAUSAL GRAPH

One causal relationship:

- A single outcome: *Max_Speed*
- A set of background variables that have a causal effect on the outcome: *HP, Weight*
- An error term that also has a causal effect on the outcome: ϵ

THE ONE-EQUATION STRUCTURAL MODEL



$$Max_Speed = \beta_0 + \beta_1 HP + \beta_2 Weight + \epsilon \quad (S)$$

Where $E[\epsilon] = 0$

THE ERROR TERM IN STRUCTURAL EQUATIONS

To a statistician: ϵ is the difference between the target and the prediction

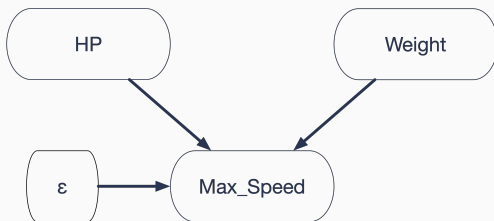
To an explanatory modeler: ϵ is unmeasured factors that have a causal effect on the outcome

THE ERROR TERM IN STRUCTURAL EQUATIONS (CONT.)

Thought experiment: Write down any missing variable that can affect the outcome.

$$\begin{aligned} \text{Max_Speed} = & \beta_0 + \beta_1 \text{HP} + \beta_2 \text{Weight} \\ & + \underbrace{\beta_3 \text{Air_Resistance} + \beta_4 \text{Tires} + \dots}_{\epsilon} \end{aligned}$$

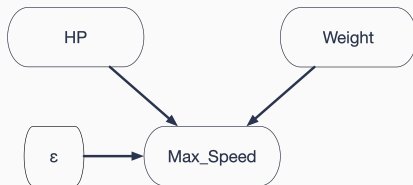
ASSESSING THE CAUSAL GRAPH



Two things we look for:

1. Are there any causal pathways back from *Max_Speed* to *HP* and *Weight*?
2. Are there any common ancestors of *HP* and *Max_Speed* or of *Weight* and *Max_Speed*?

ESTIMATION IN THE ONE-EQUATION MODEL

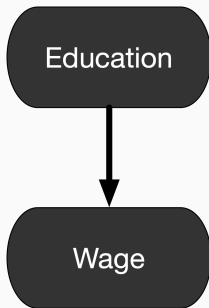


Applications for the One-Equation Model

AN IMPORTANT QUESTION

**When is the one-equation structural
model valid?**

CONFOUNDING VARIABLES IN OBSERVATIONAL DATA

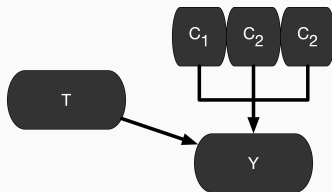


WHEN IS THE ONE-EQUATION MODEL CREDIBLE?

- True experiments
- Some natural experiments
- Differenced panels

THE TRUE EXPERIMENT

- Treatment T is randomly assigned (e.g. coin flip) \implies no incoming paths *other than the coin*.
- Controls C_1, C_2, C_3 are either measured or determined before treatment
 - No paths from T to controls, or controls to T
- Outcome Y measured after T



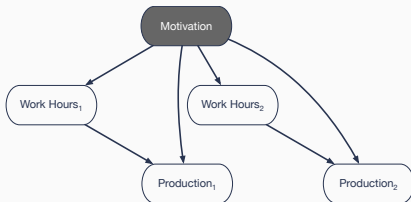
\implies OLS consistent estimates effect of T on Y .

SOME NATURAL EXPERIMENTS

Natural experiment: A scenario in which we can exploit naturally occurring variation to estimate structural parameters

- Often through instrumental variables, regression discontinuity, or other advanced techniques
- May enable OLS to *identify* causal quantities if treatment is random
 - The Vietnam War lottery
 - Tropical cyclones
 - Forest fires
 - Network outages

DIFFERENCED PANELS



$$\begin{aligned} & Production_1 = \beta_0 + \beta_1 Work_Hours_1 + \beta_2 Motivation + \epsilon_1 \\ - & \left[Production_2 = \beta_0 + \beta_1 Work_Hours_2 + \beta_2 Motivation + \epsilon_2 \right] \\ \hline \Delta Production = & \quad \beta_1 \Delta Work_Hours \quad + (\epsilon_1 - \epsilon_2) \end{aligned}$$

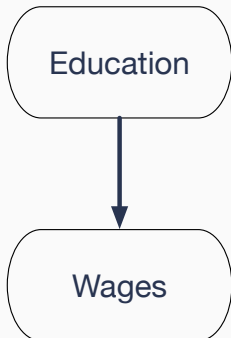
Violations of the One-Equation Structural Model

Omitted Variables

OMITTED VARIABLES

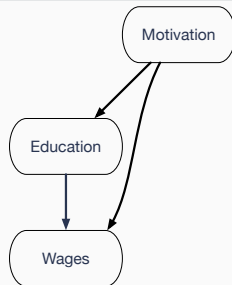
Assumed Model

Education causes wages



True Model

Motivation causes both education and wages



OMITTED VARIABLES

We fit

$$W = \tilde{\beta}_0 + \tilde{\beta}_1 E + \tilde{\epsilon}$$

True structural equation

$$W = \beta_0 + \beta_1 E + \beta_2 M + \epsilon$$

- We are interested in β_1 .
- What is the bias, $E[\tilde{\beta}_1 - \beta_1]$?

OMITTED VARIABLE BIAS IN SIMPLE REGRESSION

We fit

$$W = \tilde{\beta}_0 + \tilde{\beta}_1 E + \tilde{\epsilon}$$

True structural equation

$$W = \beta_0 + \beta_1 E + \beta_2 M + \epsilon$$

Regress M on E :

$$M = \delta_0 + \delta_1 E + \nu$$

Consider two quantities

- β_2 is the effect of M on W .
- δ_1 represents how related M and E are.

Omitted Variable Bias: $\tilde{\beta}_1 - \beta_1 = \beta_2 \delta_1$

OMITTED VARIABLE BIAS IN MULTIPLE REGRESSION

We fit

$$Y = \tilde{\beta}_0 + \tilde{\beta}_1 X_1 + \tilde{\beta}_2 X_2 + \dots \\ + \tilde{\beta}_{k-1} X_{k-1} + \tilde{\epsilon}$$

True structural equation

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots \\ + \tilde{\beta}_{k-1} X_{k-1} + \beta_k X_k + \epsilon$$

Regress X_k on other X 's:

$$X_k = \delta_0 + \delta_1 X_1 + \dots + \delta_{k-1} X_{k-1} + \nu$$

Omitted Variable Bias: $\tilde{\beta}_1 - \beta_1 = \beta_k \delta_1$

ESTIMATING OMITTED VARIABLE BIAS

$$\text{Omitted Variable Bias} = \beta_2\delta_1$$

How much does
omitted variable
affect outcome?

How related are
measured and
omitted variables?

We fit: $\widehat{Wage} = \tilde{\beta}_0 + \tilde{\beta}_1 Education$; Omitted: *Motivation*

ASSESSING OMITTED VARIABLE BIAS

Which is worse: Bias toward zero or bias away from zero?

Proof of Omitted Variable Bias

THE OMITTED VARIABLE BIAS IN SIMPLE REGRESSION

We fit

$$W = \tilde{\beta}_0 + \tilde{\beta}_1 E + \tilde{\epsilon}$$

True structural equation

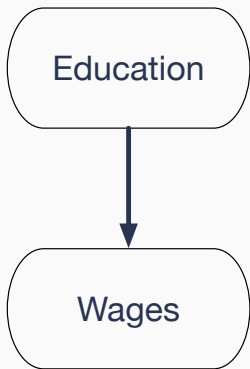
$$W = \beta_0 + \beta_1 E + \beta_2 M + \epsilon$$

Reverse Causality

REVERSE CAUSALITY

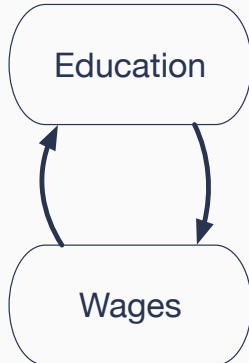
Assumed model

Education causes wages



True model

Education causes wages
and wages cause
education

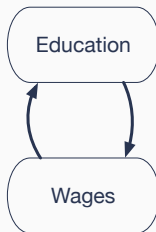


AN SEM VERSION

True structural equations:

$$W = \beta_0 + \beta_1 E + \epsilon_1 \quad (1)$$

$$E = \gamma_0 + \gamma_1 W + \epsilon_2 \quad (2)$$



Observations

- E is a descendant of ϵ_1 .
- $\implies E$ and ϵ_1 are dependent.
- $\implies (1)$ is not the BLP.

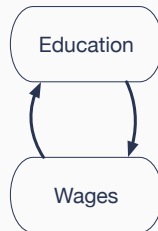
Since OLS estimates the BLP, it can't estimate (1).

UNDERSTANDING FEEDBACK

True structural equations:

$$W = \beta_0 + \beta_1 E + \epsilon_1$$

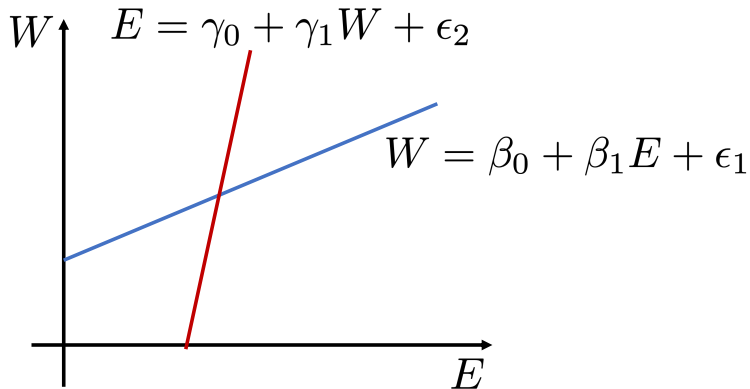
$$E = \gamma_0 + \gamma_1 W + \epsilon_2$$



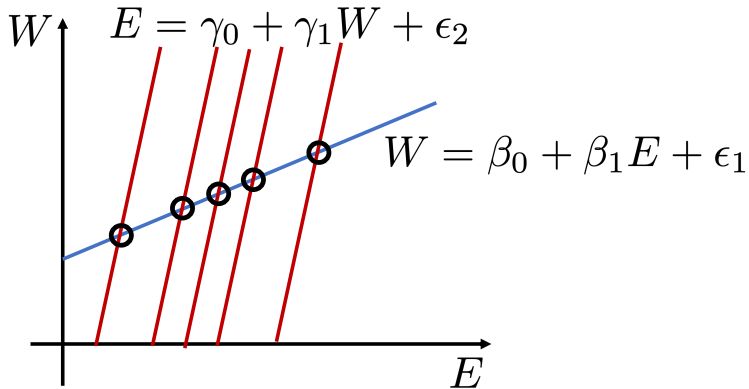
Suppose $\beta_1 > 0$

- Positive feedback $\gamma_1 > 0$
 - $\tilde{\beta}_1 > \beta_1$
- Negative feedback $\gamma_1 < 0$
 - $\tilde{\beta}_1 < \beta_1$

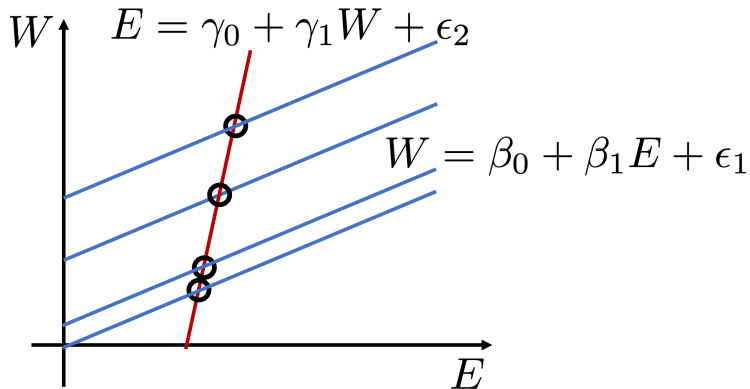
UNDERSTANDING THE SEM EQUILIBRIUM



UNDERSTANDING THE SEM EQUILIBRIUM



UNDERSTANDING THE SEM EQUILIBRIUM

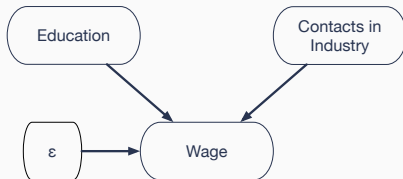


Outcome Variables on Right-Hand Side

OUTCOME VARIABLES ON THE RIGHT-HAND SIDE

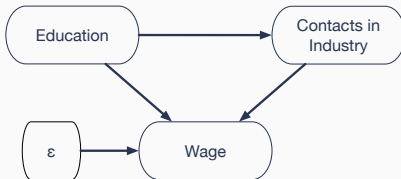
Assumed model

- Education causes wages
- Contacts in industry cause wages

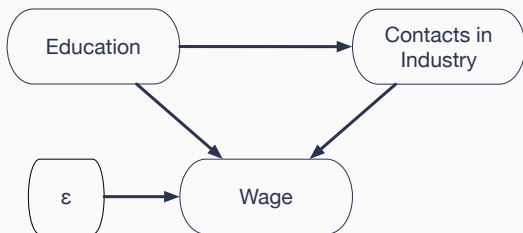


True Model

- Education causes wages
- Contacts in industry cause wages
- Education creates contacts in industry



ESTIMATING THE STRUCTURAL PARAMETERS



Structural Equation: $W = \beta_0 + \beta_1 E + \beta_2 C + \epsilon$ (S)

ϵ and E have no common ancestors.

$\implies \epsilon$ and E are independent.

$\implies \text{cov}(E, \epsilon) = 0$

\implies OLS is consistent for β_1

INTERPRETING THE STRUCTURAL COEFFICIENT

β_1 - The expected increase in Wage, from getting an extra year of education, holding the number of industry contacts constant.

TAKE AWAY

**Do not put outcome variables on the
right hand side.**

Explanatory Modeling Wrap-Up

TAKE AWAYS

- Explanatory modeling takes place inside a causal theory.
- The one-equation structural model is usually wrong.
- In special circumstances, advanced techniques can overcome omitted variables and reverse causality.
 - To learn more, try the instrumental variables and simultaneous equations chapters in *Introductory Econometrics* (Wooldridge).