Week 11

Questions of Causation

Paul Laskowski and Alex Hughes

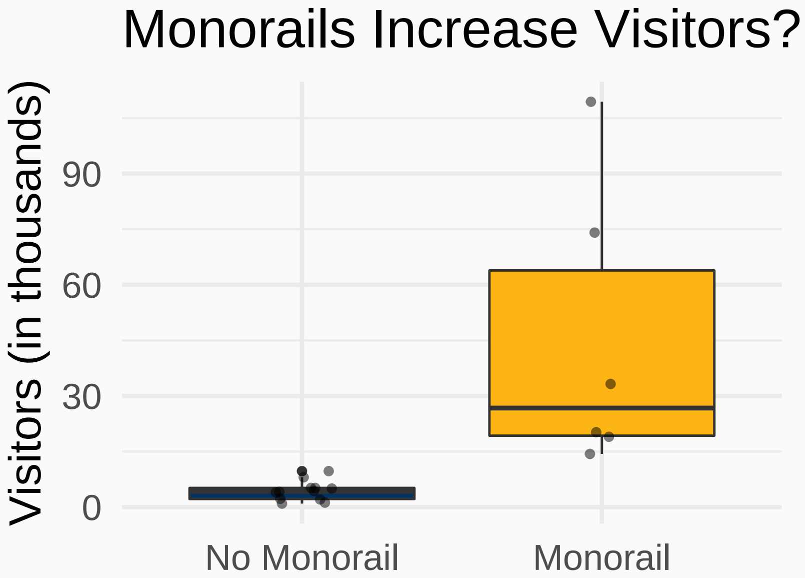
March 21, 2023

UC Berkeley, School of Information

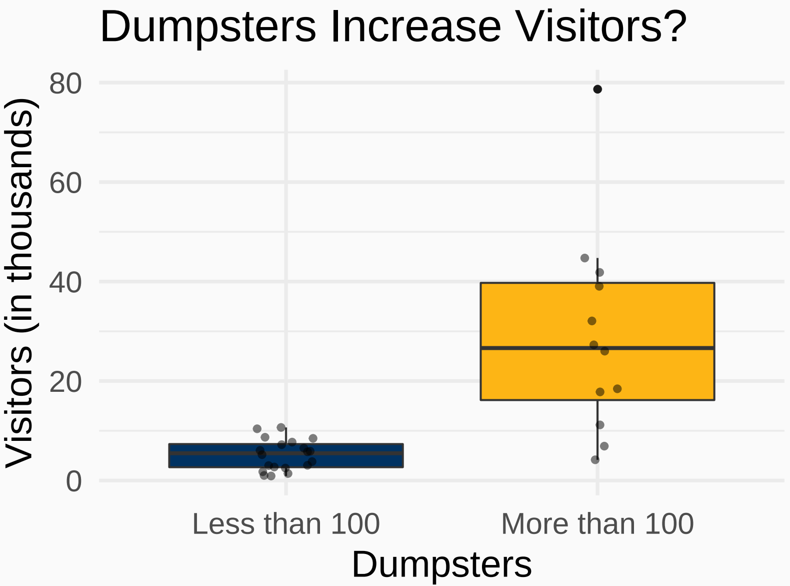
Questions of Causation

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?


Monorails Increase Visitors?
Visitors (in thousands)
90
60
30
0
No Monorail
Monorail

Amusement Park Daily Visitors (con't)



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

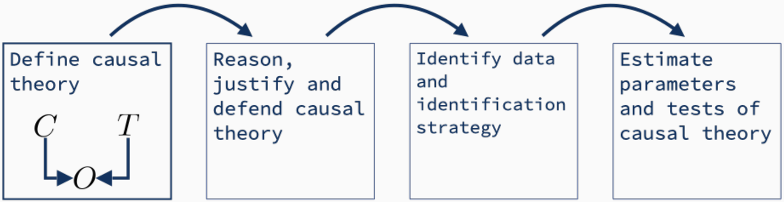
What Is Explanatory Modeling?

Toward Explanation

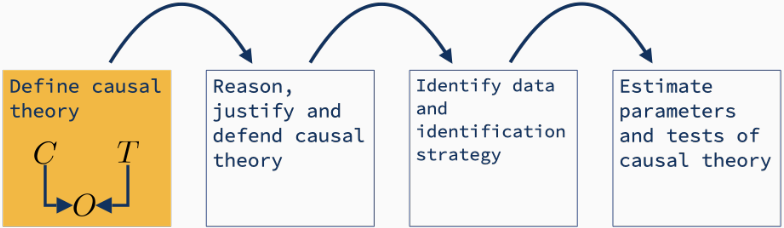
What extra assumptions are needed for OLS regression coefficients to be causal?*

* Misleading question

The Explanatory Modeling Workflow



The Explanatory Modeling Workflow



What Is Causal Theory?

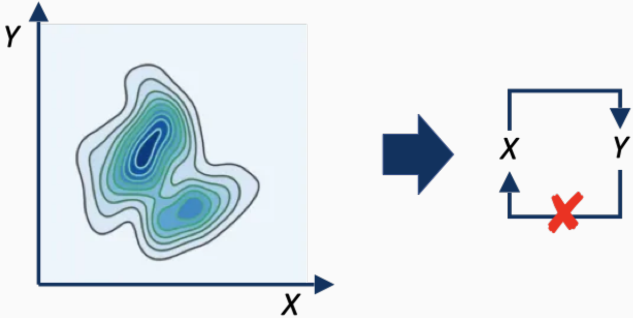
Causal theory

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

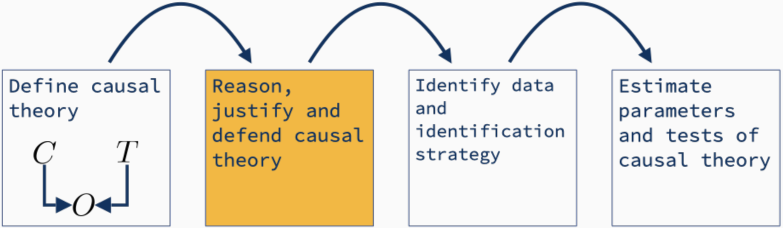
- Objective: narrow the range of causal explanations for associations we find in data.

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

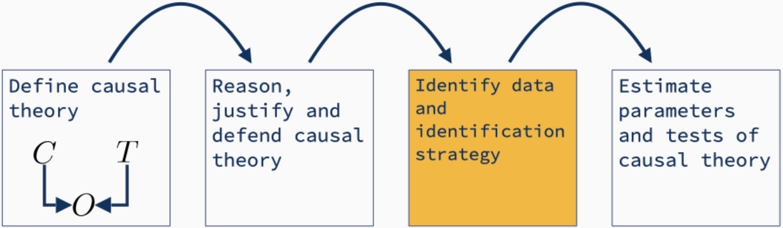


How to Reason About A Causal Theory (cont.)

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

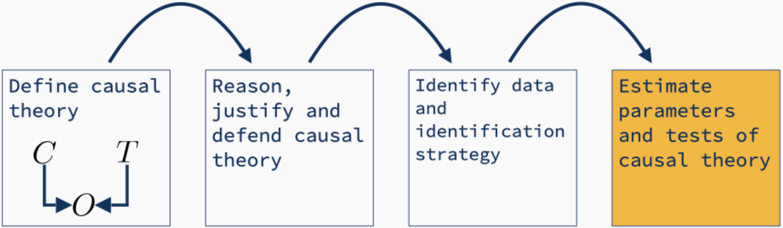
Goal: Produce a consistent estimate of the strength of the causal relationship given:

1. Causal theory
2. Data

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

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

How to Identify an Identification Strategy, Part III

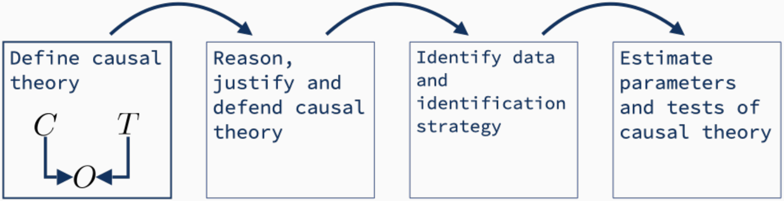


5

How to Estimate Parameters

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

The Explanatory Modeling Workflow



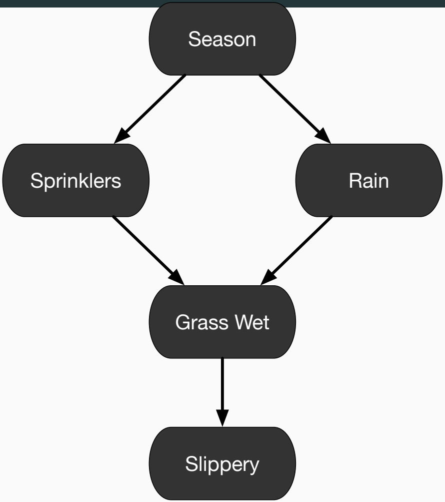
WE SHOULD HAVE AN ACTIVITY HERE

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

- Reading activity

Pearl and Structural Equation Models

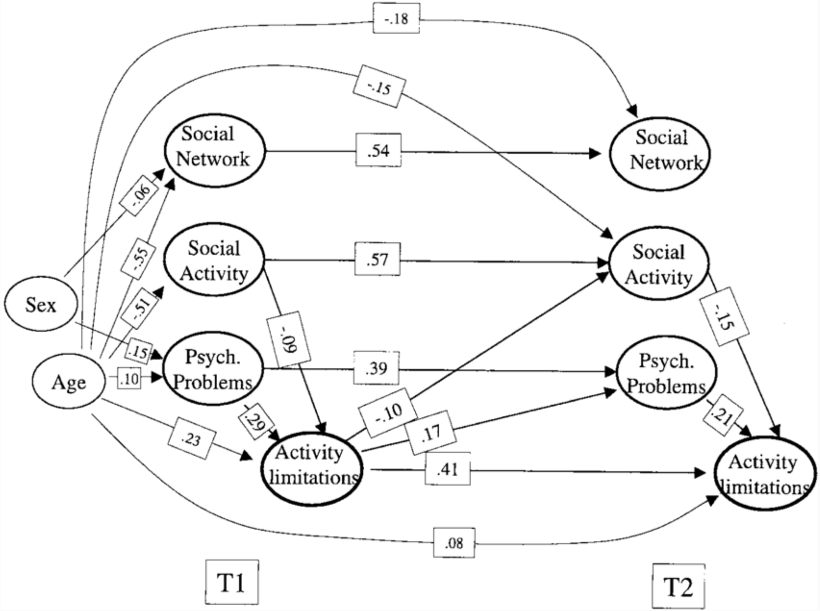
Toward a Flexible Causal Framework



Causality (2000, 2009)

Judea Pearl

How to Reason About a Causal Theory



Structural Equation Model (SEM) Basics

Endogenous variables

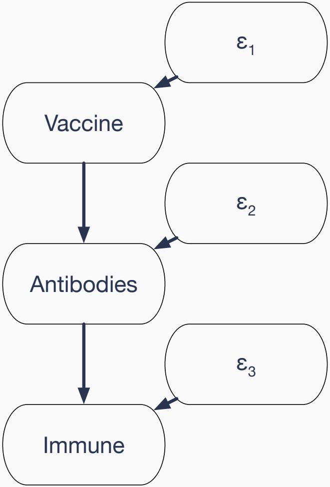
- V: Vaccine
- A: Antibodies
- I: Infection

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

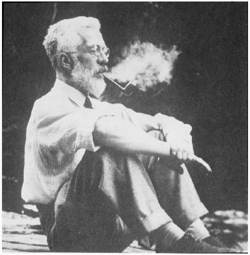
Reading: Alleged Dangers of Cigarette Smoking

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 $\rightarrow$ Alertness $\rightarrow$ Work
- Interest $\rightarrow$ Awareness $\rightarrow$ Purchase

Execution of an SEM

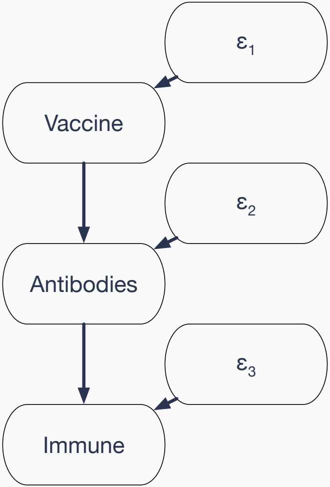
Step one

- Draw values of $\epsilon_1, \epsilon_2, \epsilon_3$.
- Assume $\epsilon_1 \ldots \epsilon_k$ 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

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

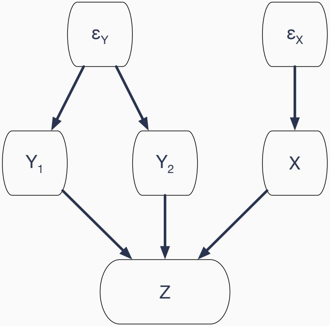
Example: Common Causes

- Education $\to$ 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.



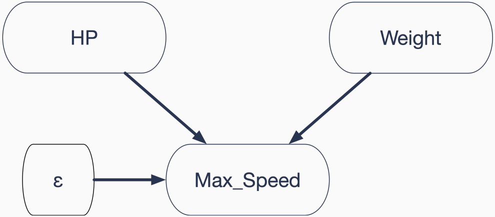
The One-Equation Structural Model

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



$$Max\_Speed = \beta_0 + \beta_1 HP + \beta_2 Weight + \epsilon \tag{S}$$

Where $E[\epsilon] = 0$

The Error Term in Structural Equations

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

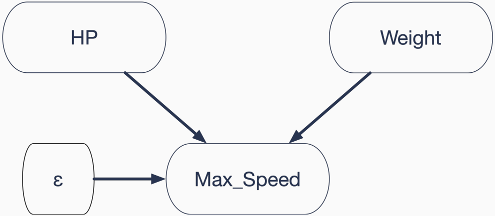
To an explanatory modeler: ϵ is unmeasured factors that have a causal effect on the outcome

The Error Term in Structural Equations (cont.)

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

$$\begin{aligned} Max\_Speed = &\beta_0 + \beta_1 HP + \beta_2 Weight \\ &+ \underbrace{\beta_3 Air\_Resistance + \beta_4 Tires + \ldots}_{\epsilon} \end{aligned}$$

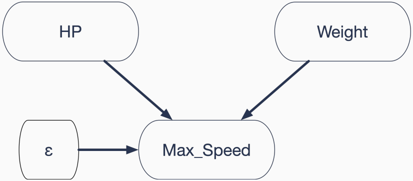
Assessing the Causal Graph



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

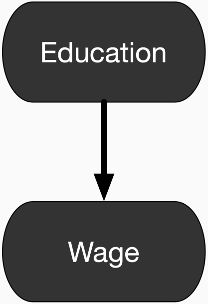


Applications for the One-Equation Model

An Important Question

When is the one-equation structural model valid?

Confounding Variables in Observational Data

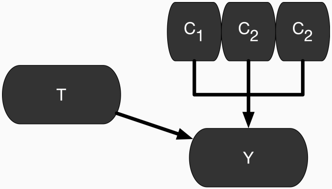


When Is the One-Equation Model Credible?

- 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



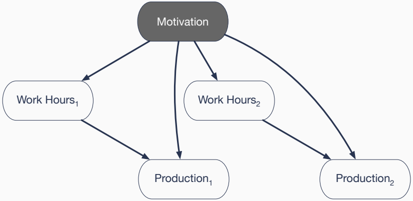
$\implies$ OLS consistent estimates effect of T on Y.

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



$$
\begin{array}{l}
\quad Production_1 = \beta_0 + \beta_1 Work\_Hours_1 + \beta_2 Motivation + \epsilon_1 \\
-\left[Production_2 = \beta_0 + \beta_1 Work\_Hours_2 + \beta_2 Motivation + \epsilon_2 \right] \\
\hline
\quad \Delta Production = \qquad \beta_1 \Delta Work\_Hours \qquad + (\epsilon_1 - \epsilon_2)
\end{array}
$$

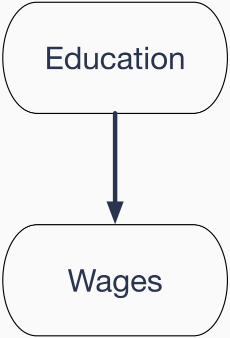
Violations of the One-Equation Structural Model

Omitted Variables

Omitted Variables

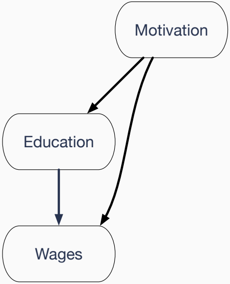
Assumed Model

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	True structural equation
$W = \tilde{\beta}_0 + \tilde{\beta}_1 E + \tilde{\epsilon}$	$W = \beta_0 + \beta_1 E + \beta_2 M + \epsilon$

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

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

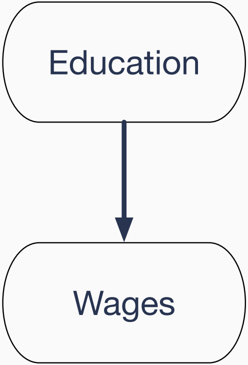
True structural equation

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

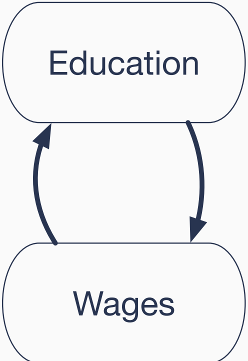
Reverse Causality

Reverse Causality

Assumed model
Education causes wages

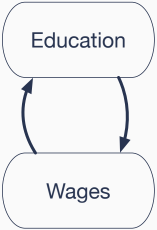


True model
Education causes wages *and* wages cause education



True structural equations:

$$W = \beta_0 + \beta_1 E + \epsilon_1 \qquad (1)$$
$$E = \gamma_0 + \gamma_1 W + \epsilon_2 \qquad (2)$$

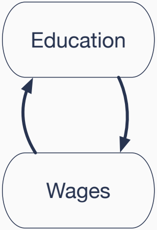


Observations

- 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



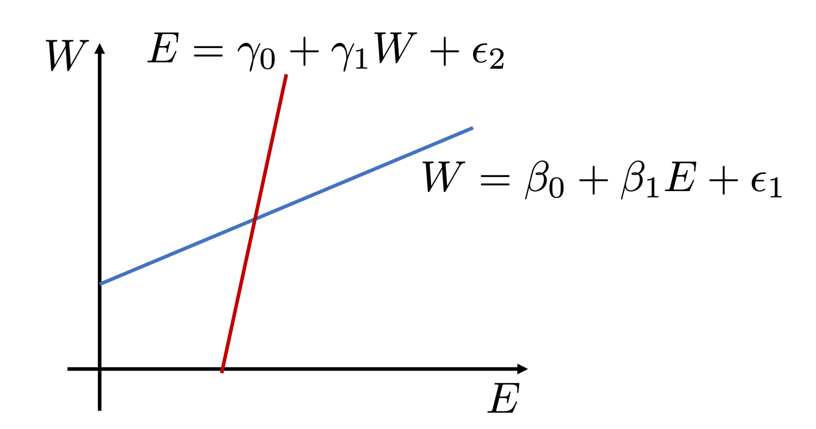
True structural equations:

$$W = \beta_0 + \beta_1 E + \epsilon_1$$
$$E = \gamma_0 + \gamma_1 W + \epsilon_2$$

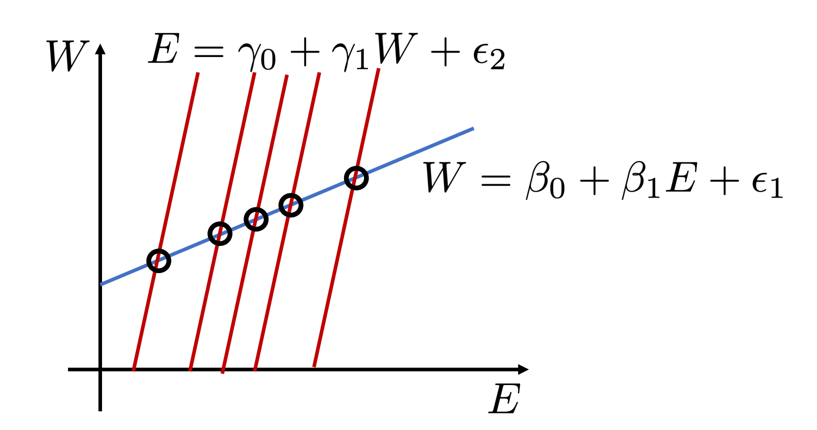
Suppose $\beta_1 > 0$

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

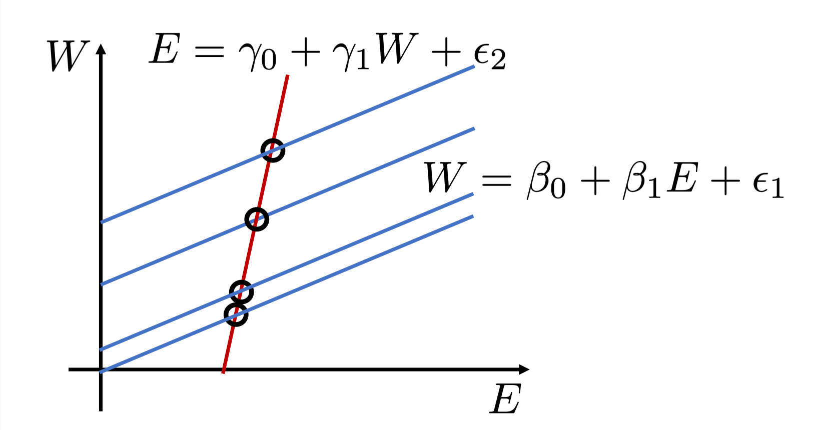
Understanding the SEM Equilibrium



Understanding the SEM Equilibrium



Understanding the SEM Equilibrium



Outcome Variables on Right-Hand Side

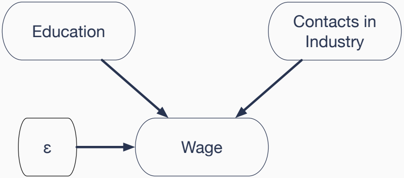
Outcome Variables on the Right-Hand Side

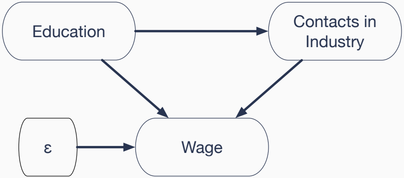
Assumed model

- Education causes wages
- Contacts in industry cause wages

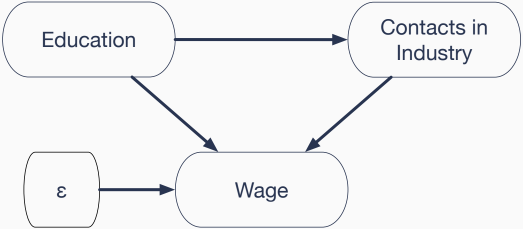
True Model

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





Estimating the Structural Parameters



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

ϵ and E have no common ancestors.

$\implies$ ϵ and E are independent.

$\implies$ $cov(E, \epsilon) = 0$

$\implies$ OLS is consistent for β_1

Interpreting the Structural Coefficient

β_1 - The expected increase in Wage, from getting an extra year of eduction, 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).