CompSci 295, Causal Inference

Rina Dechter, UCI

Lecture 5: Linear Structural Causal Models

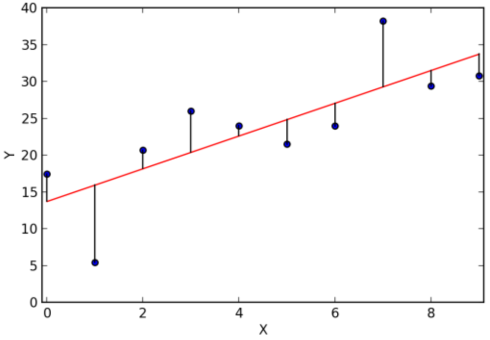
Slides: Daniel Kumor, Elias Bareinboim
(reading: Primer 3, Causality 5)

Lecture Outline

1. Linear Regression
2. Introduction to Linear Structural Causal Models
3. Examples of when regression can and cannot be used to find causal effects.
4. Modern algorithmic approaches to identification in linear SCM

Regression

- *Predict the value of Y based on X*
- *Used in Machine Learning too*
- *How to create a regression line?*
    - Plot data values of X, Y
    - “Fit” them to y = mx + b
    - The least square regression is the line that minimize the sum of the squared error average $\sum (y - b - mx)^2$
    - Need to find b and m
        - What do they represent on the graph?



Regression Coefficient

- *R_{YX} is slope of regression line of Y on X*
- *$m = R_{YX} = \sigma_{XY}/\sigma_X^2$*
  - $R_{YX} = R_{XY}$?
  - When is it?
- *Slope gives correlation*
  - Positive number → positive correlation
  - Negative number → negative correlation
  - Zero → independent or non-linear

$$\sigma_{XY} \triangleq E[(X - E(X))(Y - E(Y))]$$

The covariance σ_{XY} is often normalized to yield the *correlation coefficient*

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Multiple Regression

- $y = r_0 + r_1 \cdot x + r_2 \cdot z$
- How do we visualize?
- 3d plane
- What happens if we hold x at a value?
- $r_1 \cdot x$ becomes a constant
- r_2 is now the 2d slope of slice along X-axis
- What happens if we hold z at a value?
- $r_2 \cdot z$ becomes a constant
- r_1 is now the 2d slope of slice along Z-axis

Partial Regression Coefficient

- *Symbol for regression coefficient of Y on X?*
  - R_{YX}
- *Symbol for regression coefficient of Y on X when holding Z constant?*
  - $R_{YX \cdot Z}$
  - Called **partial regression coefficient**
- *What happens when R_{YX} is positive and $R_{YX \cdot Z}$ is negative?*
- *What are partial regression coefficients in $y = r_0 + r_1 \cdot x + r_2 \cdot z$?*
  - r_1 and r_2

$$Y = r_0 + r_1 X_1 + r_2 X_2 + \cdots + r_k X_k + \epsilon \tag{1.24}$$

then, regardless of the underlying distribution of $Y, X_1, X_2, \ldots, X_k$, the best least-square coefficients are obtained when ϵ is uncorrelated with each of the regressors $X_1, X_2, \ldots, X_k$. That is,

$$Cov(\epsilon, X_i) = 0 \quad \text{for } i = 1, 2, \ldots, k$$

To see how this *orthogonality principle* is used to our advantage, assume we wish to compute the best estimate of $X = \textit{Die 1}$ given the sum

$$Y = \textit{Die 1} + \textit{Die 2}$$

Writing

$$X = \alpha + \beta Y + \epsilon$$

$$E[X] = \alpha + \beta E[Y] \tag{1.25}$$

Further multiplying both sides of the equation by X and taking the expectation gives

$$E[X^2] = \alpha E[X] + \beta E[YX] + E[X\epsilon]. \tag{1.26}$$

The orthogonality principle dictates $E[X\epsilon] = 0$, and (1.25) and (1.26) yield two equations with two unknowns, α and β. Solving for α and β, we obtain

$$\alpha = E(X) - E(Y)\frac{\sigma_{XY}}{\sigma_Y^2}$$

$$\beta = \frac{\sigma_{XY}}{\sigma_Y^2}$$

Consider for example the problem of finding the best estimate of Z given two observations, $X = x$ and $Y = y$. As before, we write the regression equation

$$Z = \alpha + \beta_Y Y + \beta_X X + \epsilon$$

But now, to obtain three equations for α, β_Y and β_X, we also multiply both sides by Y and X and take expectations. Imposing the orthogonality conditions $E[\epsilon Y] = E[\epsilon X] = 0$ and solving the resulting equations gives

$$\beta_Y = R_{ZY \cdot X} = \frac{\sigma_X^2 \sigma_{ZY} - \sigma_{ZX} \sigma_{XY}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \tag{1.27}$$

$$\beta_X = R_{ZX \cdot Y} = \frac{\sigma_Y^2 \sigma_{ZX} - \sigma_{ZY} \sigma_{YX}}{\sigma_Y^2 \sigma_X^2 - \sigma_{YX}^2} \tag{1.28}$$

Lecture Outline

1. Linear Regression
2. Introduction to Linear Structural Causal Models
3. Examples of when regression can and cannot be used to find causal effects.
4. Modern algorithmic approaches to identification in linear SCM

Linear Structural Causal Models

Linear SCM are defined as a system of linear equations representing ground-truth:

$$Y := \sum_i \lambda_{x_i y} X_i + \mathcal{E}_y$$

1. All correlations between $\mathcal{E}$ are explicitly specified.
2. X_i are the direct causes of Y, and $\lambda_{x_i y}$ is the change in Y per X_i.
3. WLOG assume normalized data ($\mathbf{E}[X] = 0$ and $\mathbf{E}[XX] = 1$) to simplify math
4. Assume $\mathcal{E}_y \sim \mathcal{N}$, meaning that the distribution is fully specified by covariance matrix Σ (σ_{ij}).

Causal Inference In Linear Systems

Examples:

- What is the effect of birth control use on blood pressure after adjusting for confounders; or the **total** effect of an after-school study program on test scores;
- What is the **direct effect** or the unmediated by other variables, of the program on test scores.
- What is the effect of enrollment in an optional work training program on future earnings, when enrollment and earnings are confounded by a common cause (e.g., motivation).

- **Continuous variables:** We need to model with continuous variables. These traditionally been formulated as linear equation models .

- We will assume linear functions and Normal distributions of errors .

Linear systems are useful because

1. Efficient representation
2. Substitutability of expectations for probabilities
3. Linearity of expectations
4. Invariance of regression coefficients

Multivariate Gaussian can be expressed with expectation ond covariance on pairs of variables at most.
Also conditional probability can be captures by conditional expectation

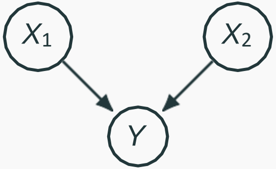
Non-Parametric to Linear

The only substantive change we are making is that the function f becomes linear:

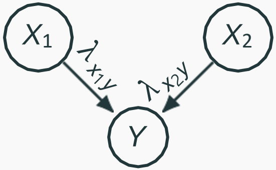
$$V_i \leftarrow f_i(pa_i, U_i) \quad \Rightarrow \quad V_i \leftarrow \sum_{j|V_j \in pa_i} \lambda_{ji} V_j + \mathcal{E}_i$$

1. λ_{ji} is called the "Structural Coefficient".
2. Instead of using U_i, we rename it to $\mathcal{E}_i$ by convention.
3. If we know all λ_{ji}, we can find the causal effect of V_j on V_i.

Example



$\Longrightarrow$
becomes



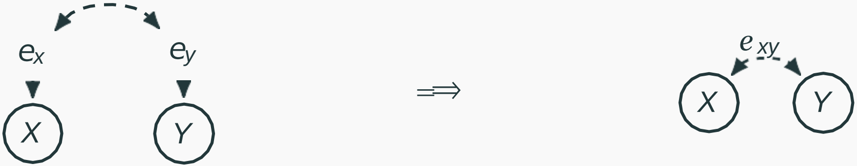
$$
\begin{aligned}
X_1 &= f_{x_1}(U_{x_1}) \\
X_2 &= f_{x_2}(U_{x_2}) \\
Y &= f_y(X_1, X_2, U_y)
\end{aligned}
$$

$$
\begin{aligned}
X_1 &= \varepsilon_{x_1} \\
X_2 &= \varepsilon_{x_2} \\
Y &= \lambda_{x_1y}X_1 + \lambda_{x_2y}X_2 + \varepsilon_y
\end{aligned}
$$

We can draw the structural coefficients directly on the graph, which then fully specifies the model.

Latent Confounding

The covariance between e_i and e_j is represented by e_{ij}, and is used as the value of a bidirected edge:

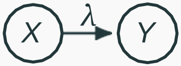


$$e_{xy} \equiv \mathbb{E}[e_x e_y]$$

e_{xy} is unobserved, since it is covariance of latent variables. It is mathematically useful, however, so we draw it on the graph just like structural coefficients.

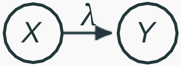
This is different from graph of non-parametric SCM, where a bidirected edge represents an explicit latent variable.

Linear SCM: Interventions



$$\mathbf{E}[Y \mid do(X = x)] = ?$$

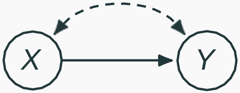
Linear SCM: Interventions



$$
\begin{aligned}
\mathbf{E}[Y \mid do(X = x)] &= \mathbf{E}[\lambda x + e_y] \\
&= \lambda x + \mathbf{E}[e_y] \\
&= \lambda x
\end{aligned}
$$

Identification In Linear SCM: The Problem Statement

- **Graph**: We are assuming that you have a hypothesized causal graph structure. In other words, you think you know what causes what, and which variables have an unknown common cause.
- **Observational Data**: You have a set of datapoints with measurements of all of the observable variables.
- **Goal: Structural Coefficients** You do NOT have knowledge of the underlying structural coefficients. These represent the actual causal effects that we want to find.

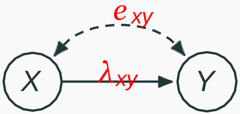


(x_1, y_1)

(x_2, y_2)

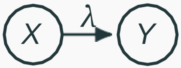
...

(x_n, y_n)



Connecting Observed with Unobserved

Remember that we assumed $e \sim N$, meaning that the distribution is fully specified by covariance matrix Σ (σ_{ij}).

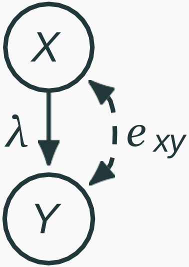


Remember, we normailize
The mean to 0 and variance to 1

$$\begin{aligned}
\sigma_{xy} &= \mathbf{E}[XY] \\
&= \mathbf{E}[X(\lambda X + e_y)] \\
&= \mathbf{E}[\lambda XX + Xe_y] \\
&= \lambda \mathbf{E}[XX] + \mathbf{E}[Xe_y] \\
&= \lambda 1 + 0 \\
&= \lambda
\end{aligned}$$

Connecting Observed with Unobserved

Solve for σ_{xy} in terms of the structural coefficients λ and e_{xy}
.

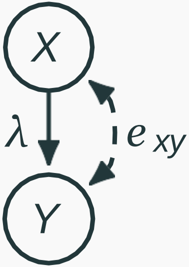


$\sigma_{xy} = ?$

Connecting Observed with Unobserved

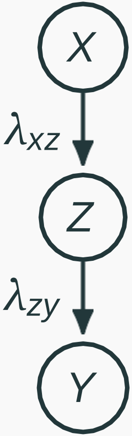
Solve for σ_{xy} in terms of the structural coefficients λ and e_{xy}

.



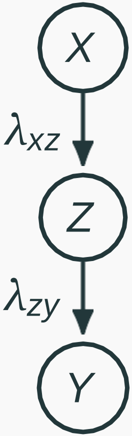
$$
\begin{aligned}
\sigma_{xy} &= \mathbf{E}[XY] \\
&= \mathbf{E}[X(\lambda X + e_y)] \\
&= \mathbf{E}[\lambda XX + X e_y] \\
&= \lambda \mathbf{E}[XX] + \mathbf{E}[X e_y] \\
&= \lambda 1 + \mathbf{E}[X e_y] \\
&= \lambda 1 + \mathbf{E}[e_x e_y] \\
&= \lambda + e_{xy}
\end{aligned}
$$

A Curious Property



$\sigma_{xy} = ?$

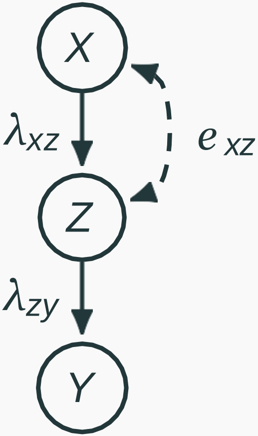
A Curious Property



$$
\begin{aligned}
\sigma_{xy} &= \mathbf{E}[XY] \\
&= \mathbf{E}[X(\lambda_{zy}Z + e_y)] \\
&= \mathbf{E}[\lambda_{zy}XZ + Xe_y] \\
&= \lambda_{zy}\mathbf{E}[XZ] + \mathbf{E}[Xe_y] \\
&= \lambda_{zy}\mathbf{E}[XZ] \\
&= \lambda_{zy}\mathbf{E}[X(\lambda_{xz}X + e_z)] \\
&= \lambda_{zy}\lambda_{xz}\mathbf{E}[XX] + \lambda_{zy}\mathbf{E}[Xe_z] \\
&= \lambda_{zy}\lambda_{xz}
\end{aligned}
$$

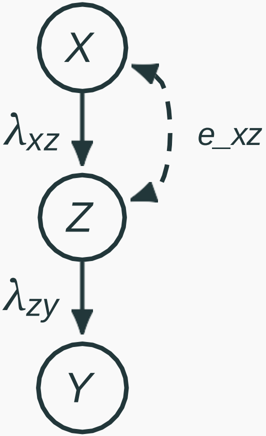
We replace X with e_x

A Curious Property



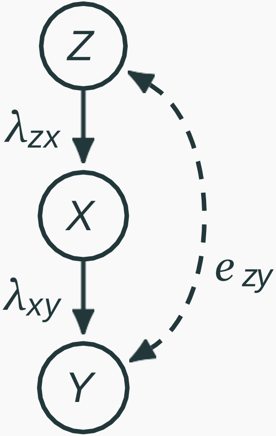
$$\sigma_{xy} = ?$$

A Curious Property



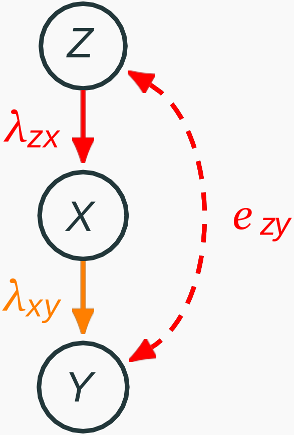
$$
\begin{aligned}
\sigma_{xy} &= \mathbf{E}[XY] \\
&= \mathbf{E}[X(\lambda_{zy}Z + e_y)] \\
&= \mathbf{E}[\lambda_{zy}XZ + Xe_y] \\
&= \lambda_{zy}\mathbf{E}[XZ] + \mathbf{E}[Xe_y] \\
&= \lambda_{zy}\mathbf{E}[XZ] \\
&= \lambda_{zy}\mathbf{E}[X(\lambda_{xz}X + e_z)] \\
&= \lambda_{zy}\lambda_{xz}\mathbf{E}[XX] + \lambda_{zy}\mathbf{E}[Xe_z] \\
&= \lambda_{zy}\lambda_{xz} + \lambda_{zy}\ e_{xz}
\end{aligned}
$$

There seems to be a relationship between covariances and paths in the graph.



$$
\begin{aligned}
\sigma_{xy} &= \mathbf{E}[XY] = \mathbf{E}[X(\lambda_{xy}X + e_y)] \\
&= \lambda_{xy}\mathbf{E}[XX] + \mathbf{E}[Xe_y] \\
&= \lambda_{xy} + \mathbf{E}[(\lambda_{zx}Z + e_x)e_y] \\
&= \lambda_{xy} + \lambda_{zx}\mathbf{E}[e_z e_y] + \mathbf{E}[e\_x, e\_y] \\
&= \lambda_{xy} + \lambda_{zx}\ e_{zy}
\end{aligned}
$$

There seems to be a relationship between covariances and paths in the graph.



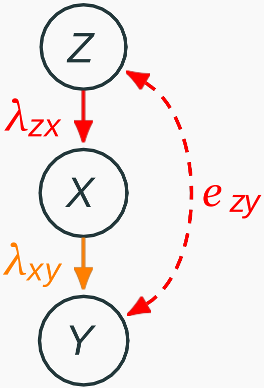
$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx}\, e_{zy}$$

The resulting terms correspond to paths between X and Y in the causal graph

Treks & Wright's Rule

The covariance between variables X and Y is the sum of paths between them in the causal graph, i.e. any non-self-intersecting path without colliding arrowheads ($\rightarrow\leftarrow$):

$$x \leftarrow \ldots \leftrightarrow \ldots \rightarrow y \qquad x \leftarrow \ldots \leftarrow w \rightarrow \ldots \rightarrow y \qquad x \leftarrow \ldots \leftarrow y \qquad x \rightarrow \ldots \rightarrow y$$



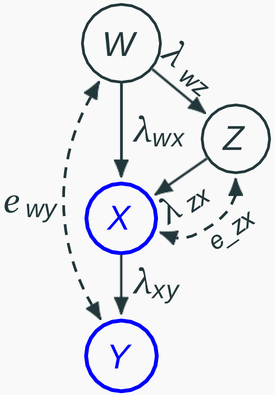
$$\sigma_{xy} = (X \xrightarrow{\lambda_{xy}} Y) + (X \xleftarrow{\lambda_{zx}} Z \xleftrightarrow{e\_zy} Y)$$

$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx}\ e_{zy}$$

Reading Covariances off the Graph

The covariance between variables X and Y is the sum of open paths between them in the causal graph, so paths with no colliding arrowheads ($\rightarrow\leftarrow$):

$$x \leftarrow \ldots \leftrightarrow \ldots \rightarrow y \qquad x \leftarrow \ldots \leftarrow w \rightarrow \ldots \rightarrow y \qquad x \leftarrow \ldots \leftarrow y \qquad x \rightarrow \ldots \rightarrow y$$

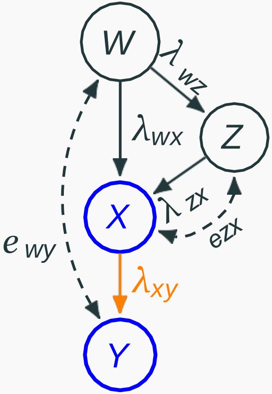


$$\sigma_{xy} = \lambda_{xy} + \lambda_{wx} e_{wy} + \lambda_{zx}\lambda_{wz} e_{wy}$$

Reading Covariances off the Graph

The covariance between variables X and Y is the sum of open paths between them in the causal graph, so paths with no colliding arrowheads ($\rightarrow\leftarrow$):

$$x \leftarrow \dots \leftrightarrow \dots \rightarrow y \qquad x \leftarrow \dots \leftarrow w \rightarrow \dots \rightarrow y \qquad x \leftarrow \dots \leftarrow y \qquad x \rightarrow \dots \rightarrow y$$

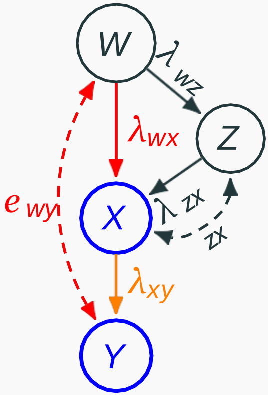


$$\sigma_{xy} = \lambda_{xy} + \lambda_{wx}\, e_{wy} + \lambda_{zx}\lambda_{wz}\, e_{wy}$$

Reading Covariances off the Graph

The covariance between variables X and Y is the sum of open paths between them in the causal graph, so paths with no colliding arrowheads ($\rightarrow\leftarrow$):

$$x \leftarrow \dots \leftrightarrow \dots \rightarrow y \qquad x \leftarrow \dots \leftarrow w \rightarrow \dots \rightarrow y \qquad x \leftarrow \dots \leftarrow y \qquad x \rightarrow \dots \rightarrow y$$

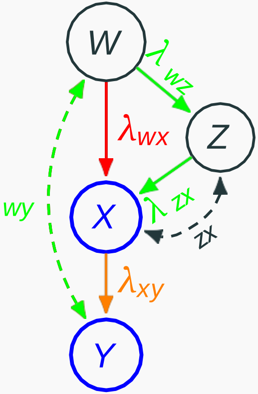


$$\sigma_{xy} = \lambda_{xy} + \lambda_{wx\ wy} + \lambda_{zx}\lambda_{wzEwy}$$

Reading Covariances off the Graph

The covariance between variables X and Y is the sum of open paths between them in the causal graph, so paths with no colliding arrowheads ($\to\leftarrow$):

$$x \leftarrow \ldots \leftrightarrow \ldots \rightarrow y \qquad x \leftarrow \ldots \leftarrow w \rightarrow \ldots \rightarrow y \qquad x \leftarrow \ldots \leftarrow y \qquad x \rightarrow \ldots \rightarrow y$$



$$\sigma_{xy} = \lambda_{xy} + \lambda_{wx}\ _{wy} + \lambda_{zx}\lambda_{wz}\ e_{wy}$$

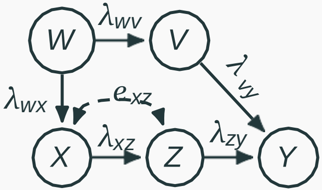
Wright's Rules (1921)

Wright's Rules [9]

$$\sigma_{xy} = \text{Sum of products of path coefficients}$$
$$\text{along all open paths between } X \text{ and } Y$$

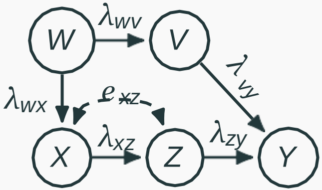
- σ_{xy} is only 0 when X and Y are d-separated.
- If there is an edge $X \xrightarrow{\alpha} Y$ in the model, then
  $\sigma_{xy} = \alpha +$ other paths between x and y.
  Thus $\sigma_{xy} = \alpha$ if X and Y are d-separated in G_{α} (graph where edge α is removed)
- Wright's rules are defined for acyclic models

One More Example



$$\sigma_{xy} = ?$$

One More Example



$$\sigma_{xy} = (\lambda_{xz} + e_{xz})\lambda_{zy} + \lambda_{wx}\lambda_{wv}\lambda_{vy}$$

Linear Regression

Example: The Medical Researcher

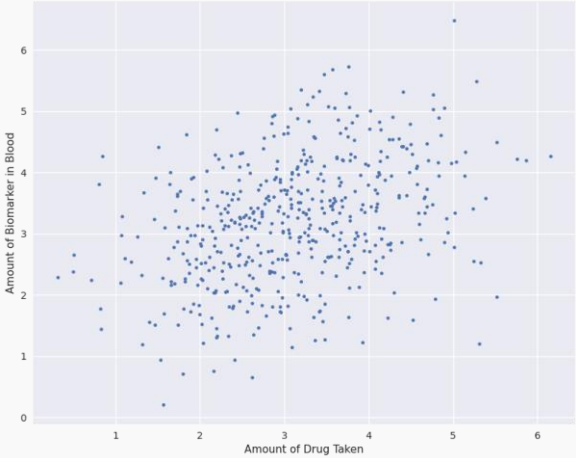
Suppose we are a medical researchers who are trying to determine if a new drug is helpful for curing a disease.

Example: The Medical Researcher

Suppose we are a medical researchers who are trying to determine if a new drug is helpful for curing a disease.

Our job is to make a treatment recommendation, which will be followed by doctors around the country.

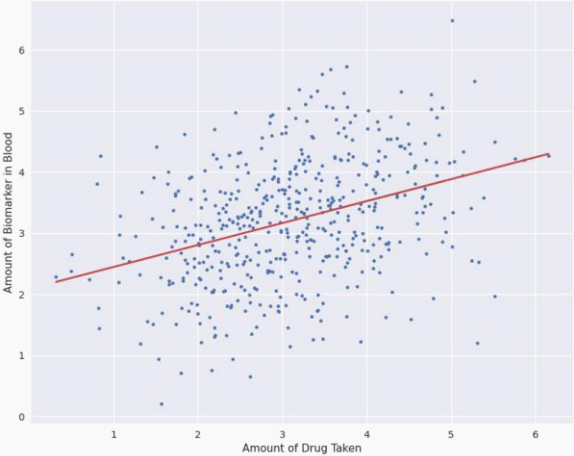
Step 1: Gather a Dataset



Start by gathering a dataset of patients who have taken the drug, including:

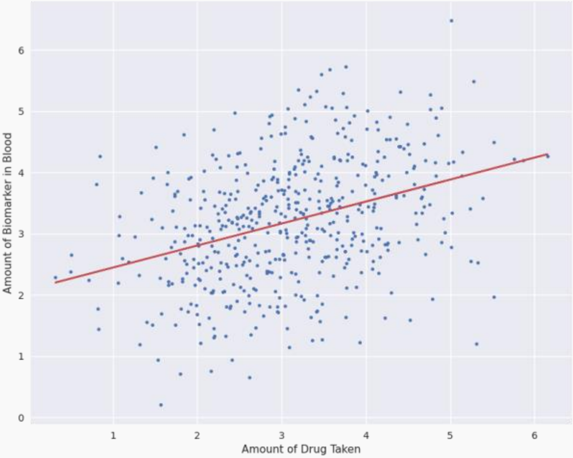
1. How much of the drug they took
2. The amount of a biomarker (antibodies) in their blood.

Step 2: Perform a Regression



Perform a regression $Y = \beta X + e$ on the data, with X as amount of drug taken, and Y the amount of biomarker, giving:

$$\beta = 0.375$$

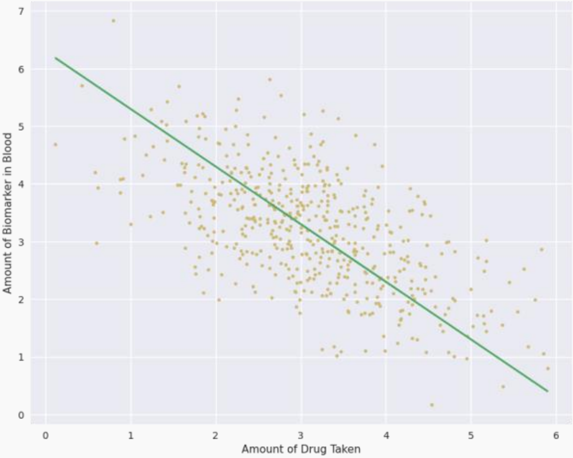


Perform a regression $Y = \beta X + e$ on the data, with X as amount of drug taken, and Y the amount of biomarker, giving:

$$\beta = 0.375$$

The drug seems to be beneficial, so you authorize its use.

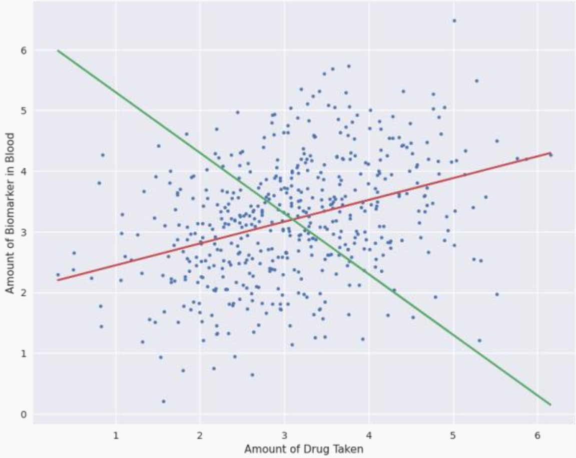
Step 3: The Drug is Given to Everyone



When the drug is given to everyone in the population, the result is a clear negative association, with slope -1.

This drug actually hurts people!

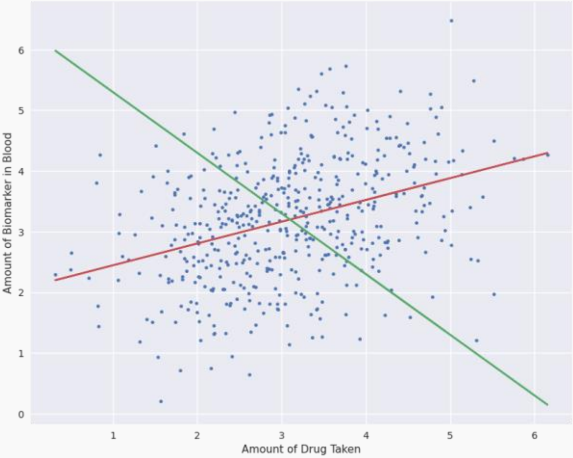
What’s Happening Here?



Why was this negative effect not visible in the original dataset?

- Maybe we didn’t gather enough data?

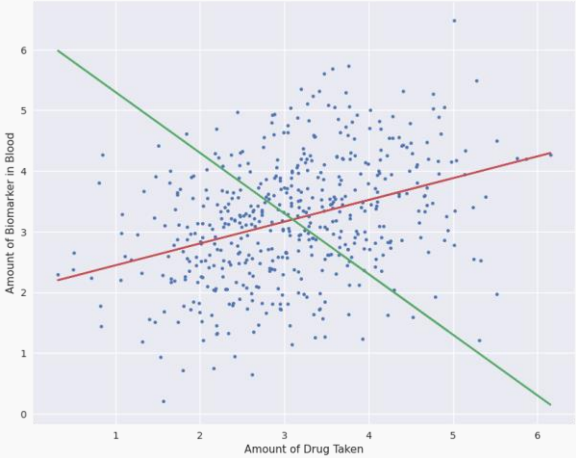
What's Happening Here?



Why was this negative effect not visible in the original dataset?

- ~~Maybe we didn't gather enough data?~~
- Why did the original regression "fail" here? (red line)

What's Happening Here?

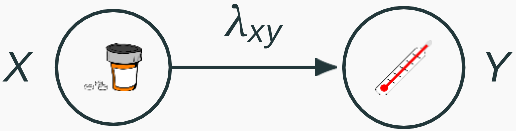


Why was this negative effect not visible in the original dataset?

- ~~Maybe we didn't gather enough data?~~
- Why did the original regression "fail" here? (red line)
- Is there a way to get the true causal effect? (green line)

Key Assumption: Lack of Confounding

The following world model is implicitly assumed when attributing causal meaning to the regression coefficient:



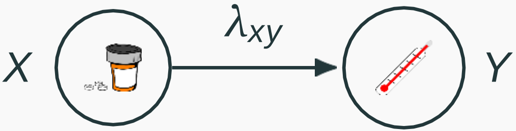
$$X := e_x$$

$$Y := \lambda_{xy} X + e_y$$

e_x, e_y independent

Key Assumption: Lack of Confounding

The following world model is implicitly assumed when attributing causal meaning to the regression coefficient:



$$X := e_x$$

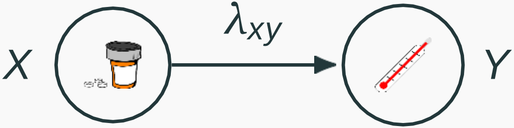
$$Y := \lambda_{xy} X + e_y$$

e_x, e_y independent

Regression $Y = \beta X + e$ gives correct $\beta = \lambda_{xy}$.

Key Assumption: Lack of Confounding

The following world model is implicitly assumed when attributing causal meaning to the regression coefficient:



$$X := e_x$$
$$Y := \lambda_{xy} X + e_y$$

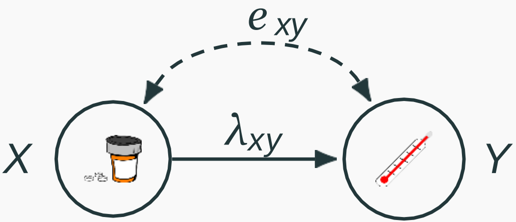
e_x, e_y independent

The covariance gives the same answer:

$$\sigma_{xy} = \mathbb{E}[XY] = \mathbb{E}[X(\lambda_{xy} X + e_y)] = \lambda_{xy} \overbrace{\mathbb{E}[XX]}^{1} + \overbrace{\mathbb{E}[X e_y]}^{0}$$

The Ground-Truth Model

If one is unable to ascertain the assumption of no confounding between X and Y, this is the corresponding graphical model:



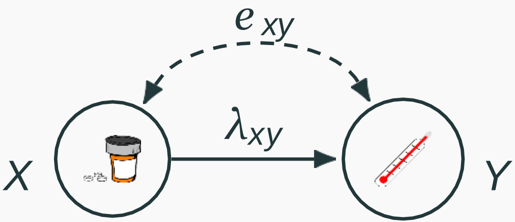
$$X := e_x$$

$$Y := \lambda_{xy} X + e_y$$

e_x, e_y correlated

The drug is expensive so mostly rich people are getting it.
But data not gathered...

If one is unable to ascertain the assumption of no confounding between X and Y, this is the corresponding graphical model:

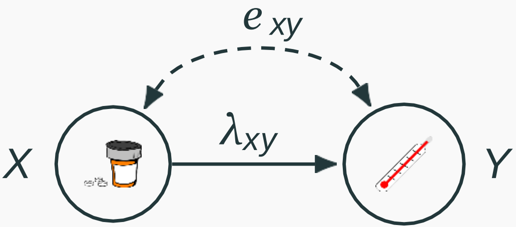


Regression $Y = \beta X + e$ gives *biased answer*

$$\begin{aligned}\sigma_{xy} &= \lambda_{xy}\mathrm{E}[XX] + \mathrm{E}[e_x e_y] \\ &= \boxed{\lambda_{xy} + e_{xy}}\end{aligned}$$

The Ground-Truth Model

If one is unable to ascertain the assumption of no confounding between X and Y, this is the corresponding graphical model:



It is provably impossible to disentangle the effect of the drug from the confounding.

That is, λ_{xy} is **not identifiable**

$$Y = \beta X + e$$

Here, β is the regression coefficient.

What does β represent?

What does Regression Compute?

Let's do least squares symbolically:

$$
\begin{aligned}
\mathbf{E}[(Y-\beta X)^2] &= \mathbf{E}[YY - 2\beta XY + \beta^2 XX] \\
&= \mathbf{E}[YY] - 2\beta\mathbf{E}[XY] + \beta^2\mathbf{E}[XX] \\
&= 1 + \beta^2 - 2\beta\mathbf{E}[XY] \\
&= 1 + \beta^2 - 2\beta\sigma_{xy}
\end{aligned}
$$

Minimizing:

$$
\begin{aligned}
0 &= \frac{\partial \mathbb{E}[(Y-\beta X)^2]}{\partial\beta} = \frac{\partial}{\partial\beta}\ \ 1 + \beta^2 - 2\beta\sigma_{xy} \\
&= 2\beta - 2\sigma_{xy} \\
\beta &= \sigma_{xy}
\end{aligned}
$$

The regression coefficient is just the covariance between x and y!

Regression Equation vs. SCM: Confusion of the Century

- **Regression Equation:**

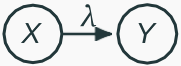
$$Y = \beta X + e \qquad \text{Assuming } e \perp X$$

When solved, $\beta = \sigma_{xy}$. We will call this value r_{yx} (solved value of linear regression of y on x). It makes no causal claims.

- **Structural Equation:**

$$Y = \lambda X + e_y$$

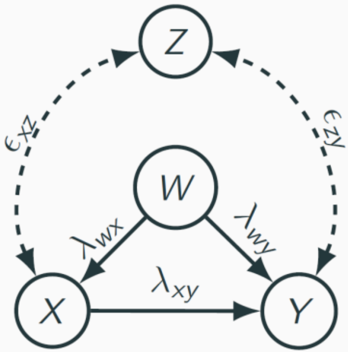
$$\mathbf{E}[Y | do(X)] = \lambda X$$



Makes claims about the interventional distribution which can be tested, and can be falsified.

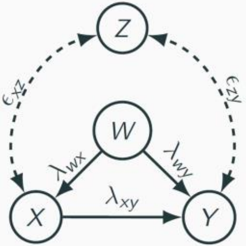
Be Careful With Regression

Remember: alpha, beta are regression
Coefficients and lambdas are causal



Be Careful With Regression

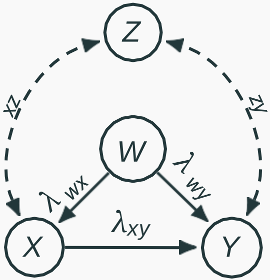
Remember: alpha, beta are regression
Coefficients and lambdas are causal



$$Y = \beta X + e$$

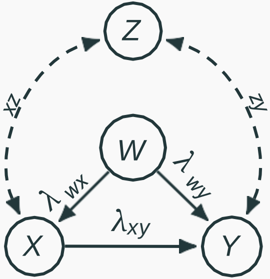
$$\beta = \sigma_{xy} = \lambda_{xy} + \lambda_{wx}\lambda_{wy}$$

Be Careful With Regression



$$Y = \beta X + \alpha W + \gamma Z + e$$

$$\beta = \lambda_{xy} - \frac{\epsilon_{xz}\epsilon_{zy}}{1 - \lambda_{wx}^2 - \epsilon_{xz}^2}$$

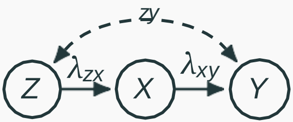


$$Y = \beta X + \alpha W + e$$

$$\beta = \lambda_{xy}$$

How to Use Regression Correctly?

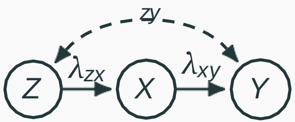
Single-Door Criterion



We want to find λ_{xy}.

$$r_{yx} = \sigma_{xy} = ??$$

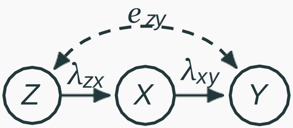
Single-Door Criterion



We want to find λ_{xy}. How can it be isolated?

$$r_{yx} = \sigma_{xy} = \lambda_{xy} + \lambda_{zx} \ e_{zy}$$

Single-Door Criterion: Multiple Regression



What if we find the least squares regression parameters of this model?

$$Y = \alpha X + \beta Z + e$$

$$\alpha = \lambda_{xy}$$
$$\beta = e_{zy}$$

Single-Door Criterion

Theorem Single-Door (Identification of Direct Effects) [Causality, Pearl]

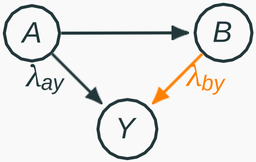
Let G be any path diagram in which λ is the path coefficient associated with the link $X \rightarrow Y$, and let G_λ denote the diagram that results when $X \rightarrow Y$ is removed from G. The coefficient λ is identifiable if there exists a set Z such that

1. Z contains no descendants of Y, and
2. Z D-separates X from Y in G_λ

Moreover, if Z satisfies these conditions, $\lambda = r_{yxz}$

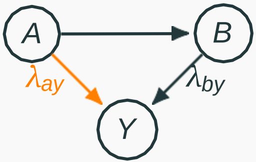
Here, we use the notation r_{yxz} to be the regression coefficient of x when performing regression y on x and z.

Example



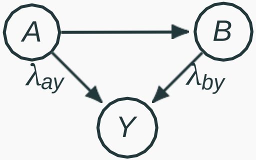
$$\lambda_{by} = ?$$

Example



$$\lambda_{by} = r_{yba}$$
$$\lambda_{ay} = ?$$

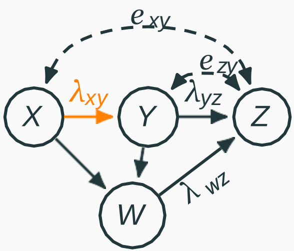
Example



$$\lambda_{by} = r_{yba}$$

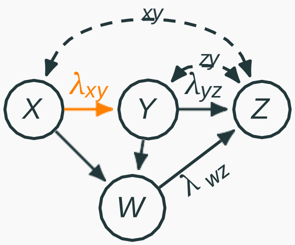
$$\lambda_{ay} = r_{yab}$$

Try It



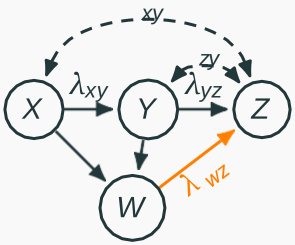
$$\lambda_{xy} = ?$$

Try It



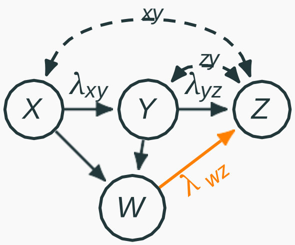
$$\lambda_{xy} = r_{yx}$$

Try It Again



$$\lambda_{wz} = ?$$

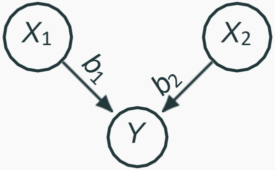
Try It Again



$$\lambda_{wz} = r_{zwyx}$$

Corollary: When are Multiple Parameters Useful?

When can we use multiple regression to solve for multiple coefficients *simultaneously?*



Back-Door Criterion

Theorem Back-Door (Identification of Total Effects) [8]

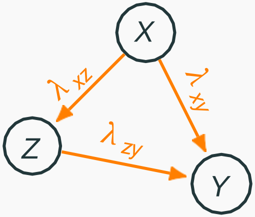
For any two variables X and Y in a causal diagram G, the total effect of X on Y is identifiable if there exists a set of measurements Z such that

1. No member of Z is a descendant of X, and
2. Z d-separates X from Y in the subgraph $G_{\underline{X}}$

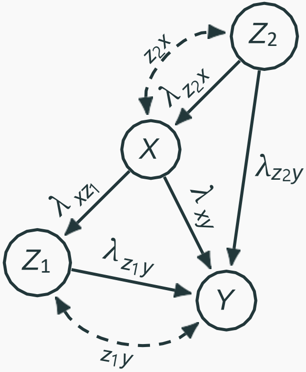
Moreover, if Z satisfies these conditions, the total effect of X on Y is given by r_{yxz}

Remember that $G_{\underline{X}}$ means delete all edges outgoing from X.

Why no Descendants of X?

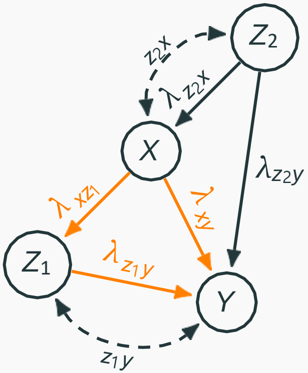


Example



What is the total effect of X on Y?

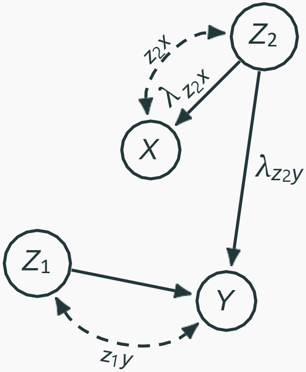
Example



What is the total effect of X on Y? $\lambda_{xz_1}\lambda_{z_1y} + \lambda_{xy}$

Can we find it using the back-door?

Example

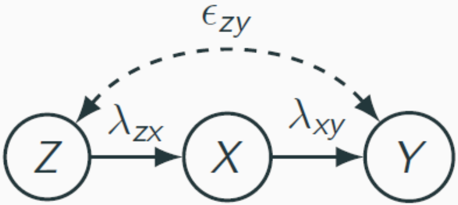


What is the total effect of X on Y? $\lambda_{xz_1}\lambda_{z_1y} + \lambda_{xy}$

Can we find it using the back-door? r_{yxz_2}

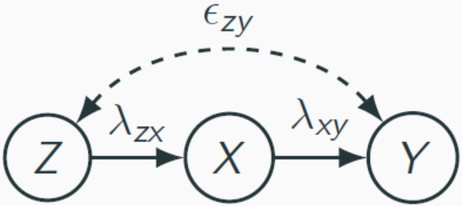
Algorithmic Identification Methods

The Equations of Linear Identification



$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & \lambda_{zx} \\ \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & 1 & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} \\ \lambda_{zx} & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} & 1 \end{bmatrix}$$

The Equations of Linear Identification

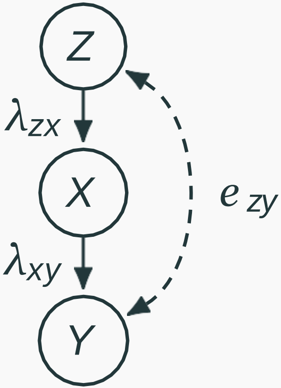


$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} & \sigma_{xz} \\ \sigma_{yx} & \sigma_{yy} & \sigma_{yz} \\ \sigma_{zx} & \sigma_{zy} & \sigma_{zz} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & \lambda_{zx} \\ \lambda_{xy} + \lambda_{zx}\epsilon_{zy} & 1 & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} \\ \lambda_{zx} & \lambda_{zx}\lambda_{xy} + \epsilon_{zy} & 1 \end{bmatrix}$$

The General Idea of Identification

Given a SCM and an observational dataset, is it possible to uniquely determine λ_{xy}?

Can λ_{xy} be solved in terms of σ?



$$\sigma_{xz} = \lambda_{zx}$$

$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx} \; e_{zy}$$

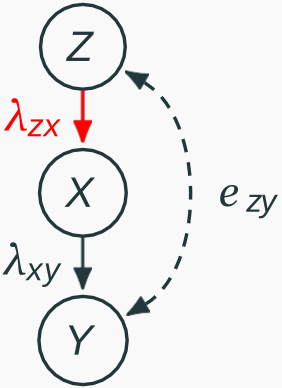
$$\sigma_{zy} = \lambda_{zx}\lambda_{xy} + e_{zy}$$

The σ are known, the λ, unknown

The General Idea of Identification

Given an SCM and an observational dataset, is it possible to uniquely determine λ_{xy}?

Can λ_{xy} be solved in terms of σ?



$$\sigma_{xz} = \lambda_{zx}$$

$$\sigma_{xy} = \lambda_{xy} + \lambda_{zx} e_{zy}$$

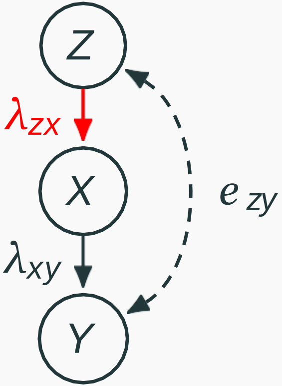
$$\sigma_{zy} = \lambda_{zx}\lambda_{xy} + e_{zy}$$

Know the value $\lambda_{zx} = \sigma_{xz}$

The General Idea of Identification

Given an SCM and an observational dataset, is it possible to uniquely determine λ_{xy}?

Can λ_{xy} be solved in terms of σ?



$$\sigma_{xz} = \lambda_{zx}$$

$$\sigma_{xy} = \lambda_{xy} + \sigma_{xz} \, e_{zy}$$

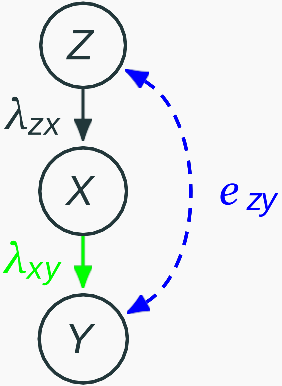
$$\sigma_{zy} = \sigma_{xz}\lambda_{xy} + e_{zy}$$

Substitute in other equations

The General Idea of Identification

Given an SCM and an observational dataset, is it possible to uniquely determine λ_{xy}?

Can λ_{xy} be solved in terms of σ?



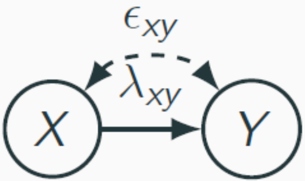
$$\sigma_{xz} = \lambda_{zx}$$

$$\sigma_{xy} = \lambda_{xy} + \sigma_{xz} e_{zy}$$

$$\sigma_{zy} = \sigma_{xz}\lambda_{xy} + e_{zy}$$

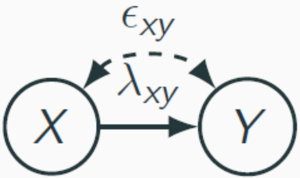
2 full-rank* linear equations in two unknowns.

A Familiar Graph



$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \epsilon_{xy} \\ \lambda_{xy} + \epsilon_{xy} & 1 \end{bmatrix}$$

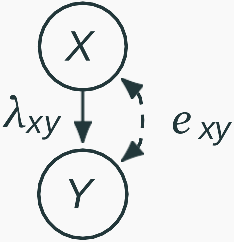
A Familiar Graph



$$\begin{bmatrix} \sigma_{xx} & \sigma_{xy} \\ \sigma_{yx} & \sigma_{yy} \end{bmatrix} = \begin{bmatrix} 1 & \lambda_{xy} + \epsilon_{xy} \\ \lambda_{xy} + \epsilon_{xy} & 1 \end{bmatrix}$$

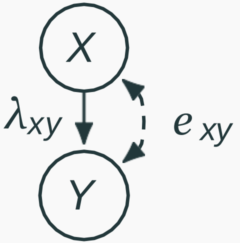
A Familiar Graph

Is it possible to solve for λ_{xy} here?



$$\sigma_{xy} = \lambda_{xy} + e_{xy}$$

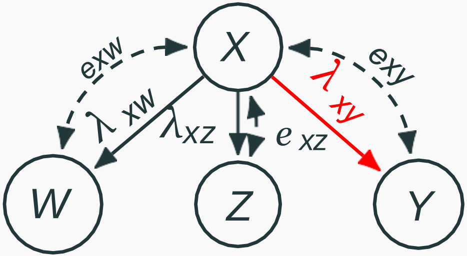
Is it possible to solve for λ_{xy} here?



$$\sigma_{xy} = \lambda_{xy} + e_{xy}$$

One equation in two unknowns: infinite number of values of λ_{xy} and e_{xy} give same covariance matrix!

Another Possibility

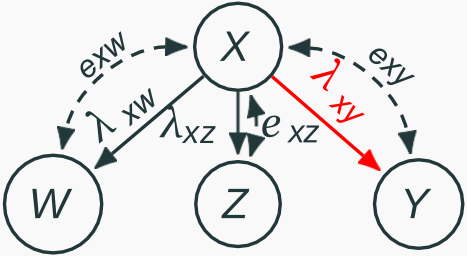


$$\sigma_{xw} = \lambda_{xw} + e_{xw} \qquad \sigma_{wz} = \lambda_{xw}\lambda_{xz} + \lambda_{xz}\, e_{xw} + \lambda_{xw}\, e_{xz}$$

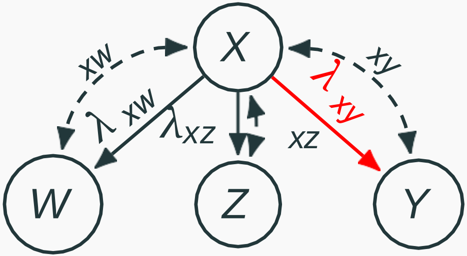
$$\sigma_{xz} = \lambda_{xz} + e_{xz} \qquad \sigma_{wy} = \lambda_{xw}\lambda_{xy} + \lambda_{xw}\, e_{xy} + \lambda_{xy}\, e_{xw}$$

$$\sigma_{xy} = \lambda_{xy} + e_{xy} \qquad \sigma_{zy} = \lambda_{xz}\lambda_{xy} + \lambda_{xz}\, e_{xy} + \lambda_{xy}\, e_{xz}$$

Another Possibility



$$0 = (\sigma_{xw}\sigma_{xz} - \sigma_{wz})\lambda_{xy}^2 + 2(\sigma_{xy}\sigma_{wz} - \sigma_{xw}\sigma_{xz}\sigma_{xy})\lambda_{xy}$$
$$+ (\sigma_{yw}\sigma_{xz}\sigma_{xy} + \sigma_{yz}\sigma_{xw}\sigma_{xy} - \sigma_{xy}^2\sigma_{wz} - \sigma_{yz}\sigma_{yw})$$



$$\lambda_{xy} = \frac{\dfrac{-(\sigma_{xy}\sigma_{wz} - \sigma_{xw}\sigma_{xz}\sigma_{xy}) + \sqrt{(\sigma_{xy}\sigma_{wz} - \sigma_{xw}\sigma_{xz}\sigma_{xy})^2 - (\sigma_{xw}\sigma_{xz} - \sigma_{wz})(\sigma_{yw}\sigma_{xz}\sigma_{xy} + \sigma_{yz}\sigma_{xw}\sigma_{xy} - \sigma_{xy}^2\sigma_{wz} - \sigma_{yz}\sigma_{yw})}}{(\sigma_{xw}\sigma_{xz} - \sigma_{wz})}}{\dfrac{-(\sigma_{xy}\sigma_{wz} - \sigma_{xw}\sigma_{xz}\sigma_{xy}) - \sqrt{(\sigma_{xy}\sigma_{wz} - \sigma_{xw}\sigma_{xz}\sigma_{xy})^2 - (\sigma_{xw}\sigma_{xz} - \sigma_{wz})(\sigma_{yw}\sigma_{xz}\sigma_{xy} + \sigma_{yz}\sigma_{xw}\sigma_{xy} - \sigma_{xy}^2\sigma_{wz} - \sigma_{yz}\sigma_{yw})}}{(\sigma_{xw}\sigma_{xz} - \sigma_{wz})}}$$

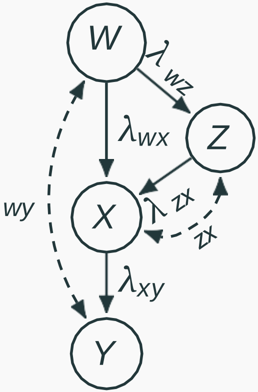
The 3 Cases of Linear ID

- **Identifiable** - Single value of λ_{xy} consistent with observational data
- **Not Identifiable** - Infinite values of λ_{xy} consistent with observations
- **Finite ID** - A finite number of possible values for λ_{xy} consistent with data

Identification in Linear SCM

Given an SCM and an observational dataset, is it possible to uniquely determine λ_{xy}?

Can λ_{xy} be solved in terms of σ?



$$\sigma_{wz} = \lambda_{wz}$$

$$\sigma_{wx} = \lambda_{wx} + \lambda_{wz}(\lambda_{zx} + e_{zx})$$

$$\sigma_{zx} = \lambda_{zx} + e_{zx} + \lambda_{wz}\lambda_{wx}$$

$$\sigma_{wy} = (\lambda_{wx} + \lambda_{wz}(\lambda_{zx} + {}_{zx}))\lambda_{xy} + e_{wy}$$

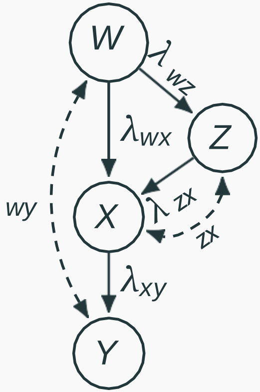
$$\sigma_{zy} = (\lambda_{zx} + e_{zx} + \lambda_{wz}\lambda_{wx})\lambda_{xy} + \lambda_{wz}$$

$$e_{wy}\ \sigma_{xy} = (\lambda_{wz}\lambda_{zx} + \lambda_{wx})\ e_{wy} + \lambda_{xy}$$

Identification in Linear SCM

Given an SCM and an observational dataset, is it possible to uniquely determine λ_{xy}?

Can λ_{xy} be solved in terms of σ?



$$\sigma_{wz} = \lambda_{wz}$$

$$\sigma_{wx} = \lambda_{wx} + \lambda_{wz}(\lambda_{zx} + e_{zx})$$

$$\sigma_{zx} = \lambda_{zx} + e_{zx} + \lambda_{wz}\lambda_{wx}$$

$$\sigma_{wy} = (\lambda_{wx} + \lambda_{wz}(\lambda_{zx} + e_{zx}))\lambda_{xy} + e_{wy}$$

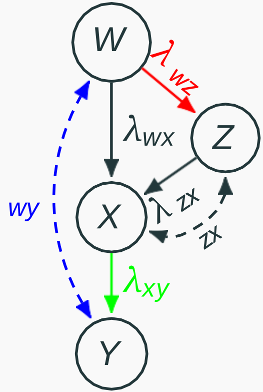
$$\sigma_{zy} = (\lambda_{zx} + e_{zx} + \lambda_{wz}\lambda_{wx})\lambda_{xy} + \lambda_{wz} e_{wy}$$

$$\sigma_{xy} = (\lambda_{wz}\lambda_{zx} + \lambda_{wx}) e_{wy} + \lambda_{xy}$$

Computer algebra approach doubly exponential in # params [2, 6]

Identification in Linear SCM

The goal of an ID algorithm is to *efficiently* find patterns reflecting solvable subsystems of equations/series of substitutions. **This is approached through graphical criteria.**



$$\sigma_{wz} = \lambda_{wz}$$

$$\sigma_{wx} = \lambda_{wx} + \lambda_{wz}(\lambda_{zx} + e_{zx})$$

$$\sigma_{zx} = \lambda_{zx} + e_{zx} + \lambda_{wz}\lambda_{wx}$$

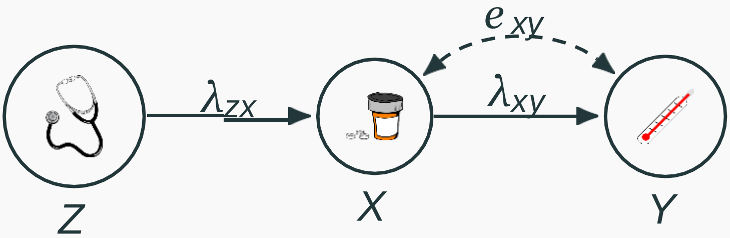
$$\sigma_{wy} = (\lambda_{wx} + \lambda_{wz}(\lambda_{zx} + zx))\lambda_{xy} + e_{wy}$$

$$\sigma_{zy} = (\lambda_{zx} + e_{zx} + \lambda_{wz}\lambda_{wx})\lambda_{xy} + \lambda_{wz}\, e_{wy}$$

$$\sigma_{xy} = (\lambda_{wz}\lambda_{zx} + \lambda_{wx})\, e_{wy} + \lambda_{xy}$$

The Starting Point: Instrumental Variables

Suppose that an independent doctor's recommendation was added to the original drug/biomarker dataset, which influences how much drug patients take.



$$Z := e_Z$$

$$X := \lambda_{zx} Z + e_x$$

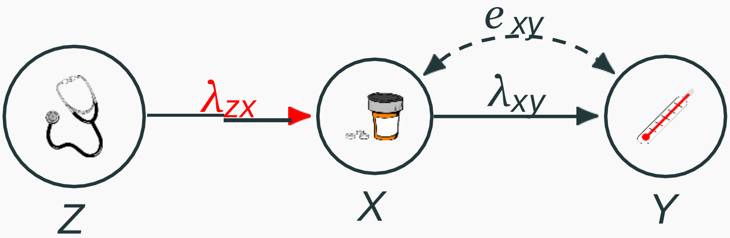
$$Y := \lambda_{xy} X + e_y$$

E_x, E_y correlated

Is λ_{xy} identifiable non-parametrically?

The Starting Point: Instrumental Variables

Suppose that an independent doctor's recommendation was added to the original drug/biomarker dataset, which influences how much drug patients take.



$\sigma_{zx} = \lambda_{zx}$

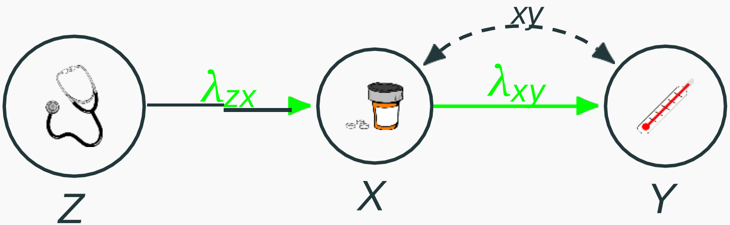
$Z := e_Z$

$X := \lambda_{zx} Z + e_x$

$Y := \lambda_{xy} X + e_y$

E_x, E_y correlated

Suppose that an independent doctor's recommendation was added to the original drug/biomarker dataset, which influences how much drug patients take.



$Z := E_Z$

$X := \lambda_{zx} Z + E_x$

$Y := \lambda_{xy} X + E_y$

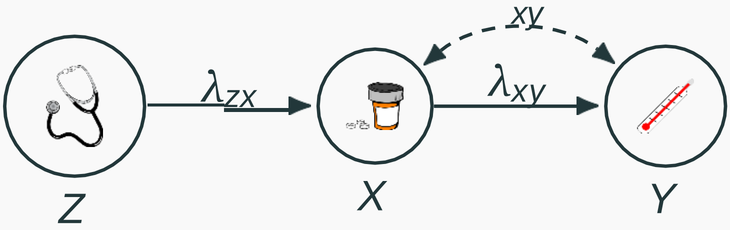
E_x, E_y correlated

$\sigma_{zx} = \lambda_{zx}$

$\sigma_{zy} = \lambda_{zx}\lambda_{xy}$

The Starting Point: Instrumental Variables

Suppose that an independent doctor's recommendation was added to the original drug/biomarker dataset, which influences how much drug patients take.



$Z := E_Z$

$X := \lambda_{zx} Z + E_x$

$Y := \lambda_{xy} X + E_y$

E_x, E_y correlated

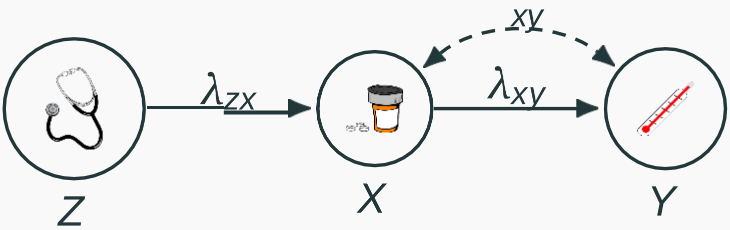
$\sigma_{zx} = \lambda_{zx}$

$\sigma_{zy} = \lambda_{zx}\lambda_{xy}$

$$\boxed{\lambda_{xy} = \frac{\sigma_{zy}}{\sigma_{zx}}}$$

The Starting Point: Instrumental Variables

Suppose that an independent doctor's recommendation was added to the original drug/biomarker dataset, which influences how much drug patients take.



A variable Z is an IV (p. 248 [Causality]) for λ_{xy} from X to Y if

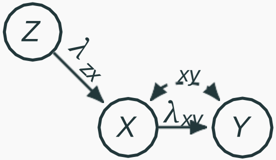
- Z is d-separated from Y in the subgraph $G_{\lambda_{xy}}$,
- Z is not d-separated from X in $G_{\lambda_{xy}}$

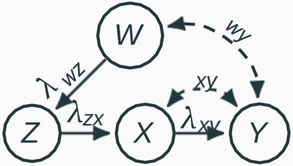
Conditional Instrumental Variables

Conditional IV Definition [3]

A variable Z qualifies as a conditional IV given a set W for structural coefficient λ_{xy} from X to Y if

- W contains only non-descendants of Y
- W d-separates Z from Y in the subgraph $G_{\lambda_{xy}}$,
- W does not d-separate Z from X in $G_{\lambda_{xy}}$



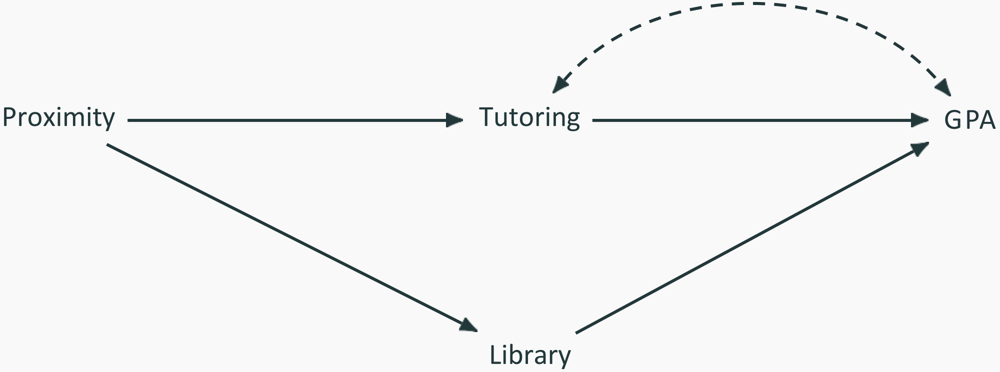


IV in Practice [1]

- **Goal:** Estimate effect of tutoring program on GPA
- The relationship between attending the tutoring program and GPA may be confounded: students attending the program may care more about their grades or may be struggling with their work.
- If students are assigned dormitories at random, the proximity of the dorm to the tutors is a natural candidate instrumental variable

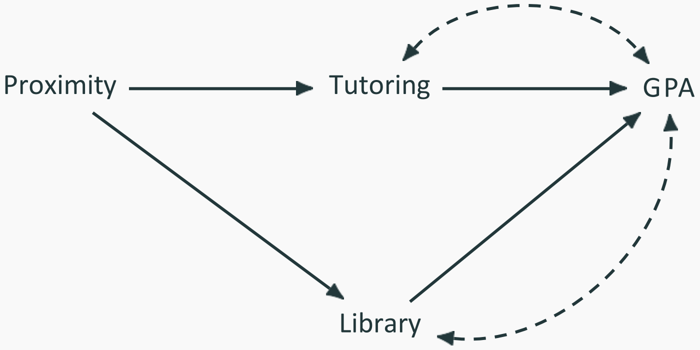
IV in Practice

What if the tutoring program is located in the college library? In that case, Proximity may also cause students to spend more time at the library, which in turn improves their GPA



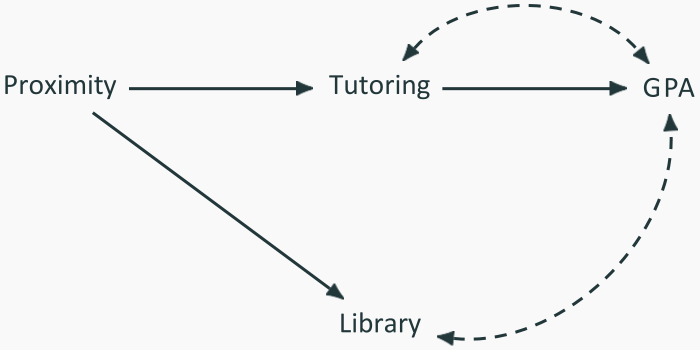
IV in Practice

Now, suppose the student’s “natural ability" affects his or her number of hours in the library as well as his or her GPA.



IV in Practice

Finally, suppose that Library Hours does not actually affect GPA because students who do not study in the library simply study elsewhere

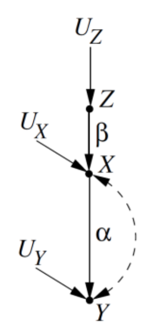


Summary on direct and total effects in SEM

Regression is essential for identification and causal effect computation.
To estimate causal effect we need to do a particular regression and specify:

What variables should be included
Which coefficient we are interested in.

As long as we have a Markovian system every structural parameter can be identified this way. We can use various regression equations. But when some variables are not measurable or errors are correlated G_alpha can be used.



- In this graph we cannot find the direct effect of X on Y via adjustment, because the dashed double-arrow arc. In this case, Z is an instrument with regard to the effect of X on Y that enables the identification of α

Definition: A variable is called an "instrument" if it is d-separated from Y in G_α and, it is d-connected to X.

Example: Z is an instrumental variable in the example.
We regress X and Y on Z separately, yielding

$$y = r_1 z + \epsilon \text{, and } x = r_2 z + \epsilon.$$

Since Z emits no backdoors, r_2 equals β and r_1 equals the total effect of Z on Y , $\beta\alpha$.
Therefore, the ratio r_1/r_2 provides the desired coefficient α.

This example illustrates how direct effects can be identified from total effects but not the other way around.

In nonlinear systems, on the other hand, the direct effect is defined through expressions such as (3.18), or

$$DE = E[Y \mid do(x, z)] - E[Y \mid do(x', z)$$

where $Z = z$ represents a specific stratum of all other parents of Y (besides X).

Even when the identification conditions are satisfied, and we are able to reduce the $do()$ operators (by adjustments) to ordinary conditional expectations, the result will still depend on the specific values of x, x', and z.

Moreover, the indirect effect cannot be given a definition in terms as do-expressions, since we cannot disable the capacity of Y to respond to X by holding variables constant. Nor can the indirect effect be defined as the difference between the total and direct effects, since differences do not faithfully reflect operations in non-linear systems.

References

1 https://en.wikipedia.org/wiki/Instrumental_variables_estimation#Selecting_suitable_instruments. Accessed: 2017-11-1.

2 M. Bardet. On the Complexity of a grobner Basis Algorithm. page 8, 2002.

3 C. Brito and J. Pearl. Generalized instrumental variables. In Proceedings of the Eighteenth conference on Uncertainty in artificial intelligence, pages 85–93, 2002.

4 B. Chen, J. Pearl, and E. Bareinboim. Incorporating knowledge into structural equation models using auxiliary variables. In Proceedings of the 25th International Joint Conference on Artificial Intelligence, pages 3577–3583, 2016.

5 C. Cinelli, D. Kumor, B. Chen, J. Pearl, and E. Bareinboim. Sensitivity analysis of linear structural causal models. In Proceedings of the 36th International Conference on Machine Learning, volume 97, pages 1252–1261, Long Beach, CA, 2019.

6 L. D. Garcıa-Puente, S. Spielvogel, and S. Sullivant. Identifying Causal Effects with Computer Algebra. page 8, 2010.

7 D. Kumor, B. Chen, and E. Bareinboim. Efficient identification in linear structural causal models with instrumental cutsets. In Advances in Neural Information Processing Systems 32, pages 12477–12486, Vancouver, Canada, 2019.

8 J. Pearl. Causality: models, reasoning and inference. 2003.

9 S. Wright. Correlation and causation. Journal of agricultural research, 20(7):557–585, 1921.

10 C. Zhang, B. Chen, and J. Pearl. A simultaneous discover-identify approach to causal inference in linear models. In Proceedings of the Thirty-fourth AAAI Conference on Artificial Intelligence, 2020.