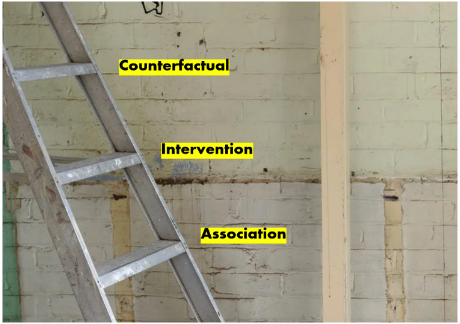
Chapter 2: Ladder of Causation

Prashant Shekhar, PhD
Assistant Professor of Data Science
Department of Mathematics
Embry-Riddle Aeronautical University, FL, USA
Email: shekharp@erau.edu

Outline

Introduction

- The Ladder of Causation, introduced by Judea Pearl (Pearl, Mackenzie, 2019), is a helpful metaphor for understanding distinct levels of relationships between variables – from simple associations to counterfactual reasoning.
- Pearl's ladder has three rungs. Each rung is related to different activity and offers answers to different types of causal questions. Each rung comes with a distinct set of mathematical tools.



Introduction

Rung	Action	Question
Association (1)	*Observing*	How does observing X change my belief in Y?
Intervention (2)	*Doing*	What will happen to Y if I do X?
Counterfactual (3)	*Imagining*	If I had done X, what would Y be?

Figure: Three rungs of the Ladder of Causation:

Examples:

- **Association (Observing)**: What is the probability of fever given that I have a headache ?
- **Intervention (Doing)**: Will my fever be cured if I take the drug X ?
- **Counterfactual (Imagining)** If I had taken drug Y instead, what would have happened to my fever ?

Structural Causal Models (SCMs)

- **Endogenous variables**: They represent the observed variables in our model. Endogenous variables are always children of at least one other variable in a model. In SCMs, they are represented as solid circles.
- **Exogenous variables**: They represent unobserved noise variables. They are not descendants of any other variable. In SCMs, they are represented as dotted circles.

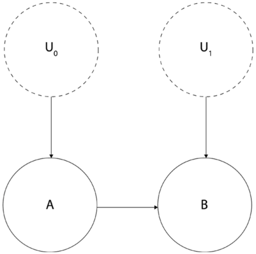


Figure: A graphical representation of a SCM. Circles or nodes represent variables. Lines with arrows or edges represent relationships between variables

Some more graph terminology

When we talk about graphs, we often use terms such as parents, children, descendants, and ancestors:

- We say that the node X is a **parent** of the node Y and that Y is a **child** of X when there's a direct arrow from X to Y.
- If there's also an arrow from Y to Z, we say that Z is a **grandchild** of X and that X is a **grandparent** of Z.
- Every child of X, all its children and their children, their children's children, and so on are **descendants** of X, which is their **ancestor**.

Association

Using SCM as a data generation mechanism and using Associations

1. Using the SCM on previous slide to generate data with a non-zero probability of buying book A, given we bought book B.
2. Using the following distributions
   - $U_0 \sim U(0, 1)$
   - $U_1 \sim N(0, 1)$
   - $A := 1_{\{U_0 > 0.61\}}$
   - $B := 1_{\{(A + 0.5U_1) > 0.2\}}$
3. In other words:
   - U_0 is a continuous random variable between 0 and 1.
   - U_1 is a normal random variable with mean 0 and standard deviation 1.
   - $A = 1$ if $U_0 > 0.61$, otherwise its 0.
   - $B = 1$ if $A + 0.5U_1 > 0.2$, otherwise its 0.

CODE

Refer to the [Python notebook] for implementation of:

- Data generation mechanism
- Association based inference

Why Associations

- Associations are useful. They allow us to generate meaningful predictions of potentially high practical significance in the absence of knowledge of the data-generating process.
- We used an SCM to generate hypothetical data for our bookstore example and estimated the strength of association between book A and book B sales, using a conditional probability query.
- Conditional probability allowed us to draw conclusions in the absence of knowledge of the true data-generating process, based on the observational data alone (note that although we knew the true SCM, we did not use any knowledge about it when computing the conditional probability query; we virtually forgot anything about the SCM before generating the predictions).
- That said, associations only allow us to answer rung one questions.

Intervention

Whats intervention

- The idea of intervention is very simple. We change one thing in the world and observe whether and how this change affects another thing in the world.
- This is the essence of scientific experiments.
- To describe interventions mathematically, we use a special do-operator. We usually express it in mathematical notation in the following way:

$$P(Y = 1 | do(X = 0))$$

- The preceding formula states that the probability of $Y = 1$, given that we set X to 0. The fact that we need to change X's value is critical here, and it highlights the inherent difference between intervening and conditioning.

Intervening vs Conditioning

- Conditioning only modifies our view of the data, while intervening affects the distribution by actively setting one (or more) variable(s) to a fixed value (or a distribution).
- Intervention changes the system, but conditioning does not.

Intervention Cont..

When we intervene in a system and fix a value or alter the distribution of some variable – let's call it X – one of three things can happen:

1. The change in X will influence the values of its descendants (assuming X has descendants and excluding special cases where X's influence is canceled, for example $f(x) = x - x$)
2. X will become independent of its ancestors (assuming that X has ancestors).
3. Both situations will take place (assuming that X has descendants and ancestors, excluding special cases)

Note that none of these would happen if we conditioned on X, because conditioning does not change the value of any of the variables – it does not change the system.

Some things to know

Pearson's correlation

Given two vectors: $x = [x_1, x_2, ..., x_n]$ and $y = [y_1, y_2, ...y_n]$, Pearson's correlation coefficient (r_{xy}) is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

And the corresponding $p - value$ is calculated using the following t-statistic as:

$$t = \frac{r_{xy} * \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}}$$

Here n is the number of data points. By default, if the p-value < 0.05, then the result (r_{xy}) is considered valid and useful

Some things to know

P-value example

Assume

$$x = [1, 2, 3, 4, 5], \quad y = [5, 10, -2, 3, 20]$$

- Using the formula for Pearson's correlation we get $r_{xy} = 0.43559$
- To calculate $p - value$, calculating the t-statistic ($\text{n} = 5$)

$$t = \frac{r_{xy} * \sqrt{n-2}}{\sqrt{1 - r_{xy}^2}} = 0.83816$$

- Now, you can either use the t-distribution table or you can use python to compute the p-value as follows:

$$\text{p-value} = 2 * (1 - stats.t.cdf(np.abs(t), df = n - 2)) = 0.46346$$

Some things to know

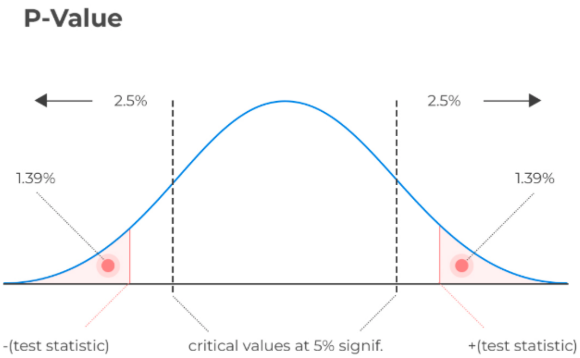
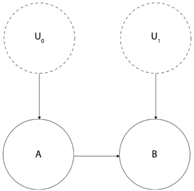


Figure: How to visualize p-value on a t-distribution PDF. Fig. Adapted from HERE

- Cumulative Distribution Function (CDF) is the integrated version of Probability Density Function (PDF). P-values below 5% indicate that the result is believable and good.

Intervention example

Again assuming the SCM



with the following structural equations:

- $U_0 \sim N(0,1)$
- $U_1 \sim N(0,1)$
- $A := U_0$
- $B := 5A + U_1$

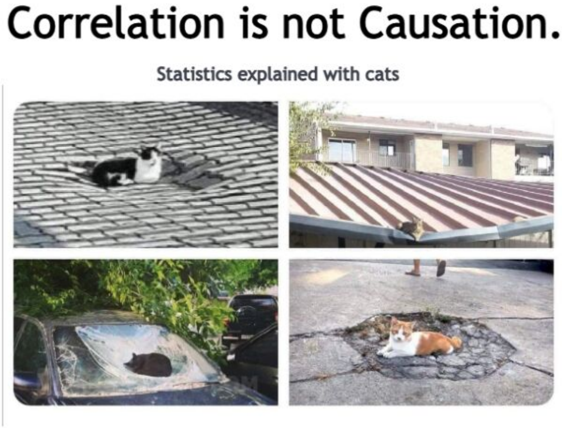
- Here we set A and B to be continuous variables (as opposed to the model in the previous section, where A and B were binary.
- Note that this is a new example, and the only thing it shares with the bookstore example is the structure of the graph.

CODE

Refer to the [Python notebook] for implementation of:

- Intervention on variable A (leads to change in B due to causal link A $\rightarrow$ B)
- Intervention on variable B (breaks the causal link between A and B)

Correlation vs Causation

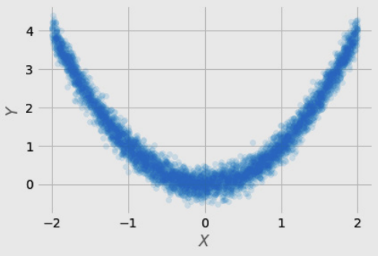


Correlation vs Causation

You might have heard the phrase that correlation is not causation. That's approximately true.

The following structural equations:

- $X := U(-2, 2)$
- $Y = X^2 + 0.2xN(0, 1)$



CODE

Refer to the [Python notebook] for implementation of:

- Correlation computation for above data (Its surprisingly close to 0)
- Risks of biases in sampling (makes understanding causal relationships more difficult)

Counterfactuals

Counterfactuals: Introduction

These are scenarios that were never observed and hence we dont have any data for that. For example:

- If I had chosen a different job, what would have been my financial status ?
- If I lived in a different city, how would my health be?

Its like thinking about a parallel universe, where the course of events was different. Contrary to interventions, counterfactuals can never be observed.

Counterfactuals vs Interventions example

- Would John have bought the chocolate last Friday had he seen the ad last week (counterfactuals)
- Show John an ad and see whether he buys (intervention)

Counterfactuals cont..

Running example

Imagine you had a coffee this morning and now you feel bad in your stomach. Would you feel the same or better if you hadn't had your coffee?

Similar questions that can be answered by intervention

1. What is the probability that people similar to you react to coffee the way you reacted, given similar circumstances?
2. What is the probability that you'll feel bad after drinking a coffee in the morning on a day similar to today, given similar circumstances?

However, we need counterfactuals for our problem

Given an alternative world that is identical to ours and only differs in the fact that you did not drink coffee this morning (plus the necessary consequences of not drinking it), what is the probability that you'd feel bad in your stomach?

Counterfactuals cont..

The quantity we want to estimate:

$$P(Y_{x=0} = 1 | X = 1, Y_{X=1} = 1) \tag{1}$$

Here:

- X = 1/0: Whether coffee was consumed (1) or not (0)
- $Y_{X=1} = 1$: Had coffee and felt bad.
- $P(Y_{X=0} = 1)$: probability of not having coffee and still feeling bad.

Hence (1) can be read as: *probability of feeling bad if coffee wasn't consumed, given that you had your coffee and you feel bad.*
Please note:

- Everything on the right side of the conditioning bar comes from the actual observation.
- The expression on the left side of the conditional bar refers to the alternative, hypothetical world.

Computing counterfactuals step-by-step

Judea Pearl and colleagues (Pearl, Glymour and Jewell, 2016) proposed a three-step framework for computing counterfactuals:

1. **Abduction**: Using evidence to calculate values of exogenous (unobserved) variables
2. **Modification (originally called an action)**: Replacing the structural equation for the treatment (action/variable of interest) with a counterfactual value.
3. **Prediction**: Using the modified SCM to compute the new value of the outcome under the counterfactual

Continuing coffee example

- **T**: Treatment, i.e. drinking coffee
- **U**: 1 (coffee sensitivity) and 0 (lack thereof)
- **Y**: 1 (felt bad) and 0 (felt normal)
- Additionally, let's assume that we know the causal mechanism for reaction to coffee. The mechanism is defined by the following SCM:

$$T := t$$

$$Y := TU + (T-1)(U-1)$$

- We know the outcome under the actual treatment, $Y_{T=1} = 1$ (you drank the coffee and you felt bad), but we don't know your characteristics (U). Can we do something about it?

Continuing coffee example

- **STEP 1 (Abduction)**: Here our exogenous variable is U. We will solve for it using the structural equation of Y(denoted as u).

$$u = \frac{T + Y - 1}{2T - 1}$$

Assigning values T = 1 (drank coffee) and Y = 1 (felt bad)

$$u = \frac{1 + 1 - 1}{2 - 1} = 1$$

The value we obtained for U reveals that you're a coffee-sensitive person. Now, we have all the elements necessary to compute counterfactual outcomes at our disposal – a causal model and knowledge about your personal characteristics.

Continuing coffee example

- **STEP 2 (Modification)**: We will fix the value of our treatment at the counterfactual of interest, (T = 0):

$$T := 0$$

$$Y := 0U + (0 - 1)(U - 1)$$

- **STEP 3 (Prediction)**: To make a prediction, we need to substitute U with the value(s) of your personal characteristic(s) that we computed before:

$$Y := 0 * 1 + (0 - 1)(1 - 1) = 0$$

Hence, you wouldn't feel bad if you hadn't had your coffee!!