Homework for Chapter 6: Causal Diagrams

1. In a conversation with your friend, you mention a study you read that found that being tall causally makes you more likely to earn above $100,000 per year. Your friend says the study must be wrong, since they know several tall people who make much less than that, and several short people who do earn that much. Does your friend’s reasoning make sense or not, and why?

My friend’s reasoning behind his statement that the study must be wrong does not make sense in Statistics. The people my friend knows are not necessarily representative of the target population of the study. The people my friend knows does not constitute a random sample, but at most a convenient sample with which we cannot isolate the variation of the variable (height) we are interested in from other possible confounders such as education, family background, and locality. My friend’s sample size is also too small to be externally valid to draw the conclusion against the study’s finding.

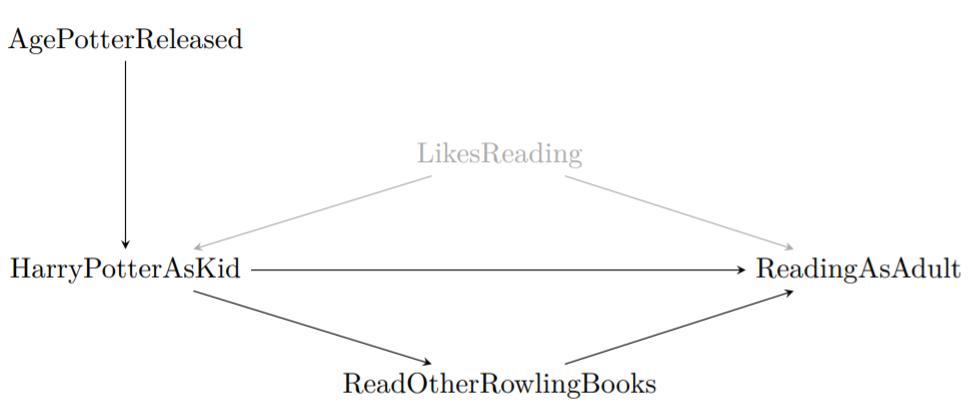
1. Consider the below diagram, which reproduces Figure 6.3:  
   Diagram

   Description automatically generated  
   In this diagram...
   1. Which variables have a direct effect on Money?

TerryInRoom and CoinFlip.

* 1. Which variables have an indirect effect on Money?

Terry’sMood.

1. You are interested in the question “Does reading Harry Potter as a child make you read more as an adult?” and draw the diagram below.   
   
   1. What direct effects should be included when trying to answer your research question of interest?

To answer our research question of interest based on the given causal diagram, we should be interested in any parts of it that allow HarryPotterAsKid to cause ReadingAsAdult. The direct effect in the causal diagram above that should be included is HarryPotterAsKid->ReadingAsAdult, whether variation in reading Harry Potter or not during childhood causes variation in one’s reading volume in the adulthood.

* 1. What indirect effects should be included when trying to answer your research question of interest?

Similarly, the indirect effect in the causal diagram above that should be included is HarryPotterAsKid->ReadOtherRowlingBooks->ReadingAsAdult, whether variation in reading Harry Potter or not during childhood causes variation in one’s preference of reading other J.K. Rowling’s books which then affects one’s reading volume in the adulthood.

* 1. What is a likely alternative explanation of why we might see a relationship between reading Harry Potter and reading more as an adult?

According to the given causal diagram, a likely alternative explanation of why we might see a relationship between reading Harry Potter and reading more as an adult is that LikesReading causes both HarryPotterAsKid and ReadingAsAdult, and so HarryPotterAsKid and ReadingAsAdult might be related because LikesReading causes both to rise (people who naturally likes reading may be more likely to read Harry Potter in the childhood and to continue the reading habit till the adulthood).

1. The figure in Question 3 has LikesReading included as an unobserved variable. In a few sentences each, explain:
   1. Why do we bother to include variables on our diagrams if we can’t observe them?

Because LikesReading is an important part of the data generating process: One doesn’t read Harry Potter in their childhood or keep reading as an adult at random, but rather make those decides based on their constant preference of reading. Therefore, though unobservable, it still needs to be considered on our diagrams.

* 1. Why might we think that LikesReading is an unobserved or latent variable?

Because LikesReading is not a variable that we can directly observe or measure but are rather estimated or inferred through some proxy measurement such as measuring other observed variables like one’s monthly book consumption and weekly time spent on reading.

1. Consider the research question “do government fire-safety advertisements reduce forest fires?”.
   1. Draw a causal diagram with these features: (a) ForestFires is caused by FireSafetyAds and ParkRangers (who can help catch fires early and put them out). (b) Both FireSafetyAds and ParkRangers are caused by GovtBudget (the government has to pay for this stuff!). (c) GovtBudget is caused by TimeSinceLastFire— a recent fire might get more money sent to the parks department.

FireSafetyAds

ForestFires

GovtBudget

TimeSinceLastFire

ParkRangers

* 1. Suggest one omitted arrow or variable that should be on the diagram.

One omitted variable is weather. A period of windy and dry weather could increase the probable occurrence of forest fires and affect TimeSinceLastFire (the recency of the last forest fire is dependent on the weather for the period).

1. Think of a research question in your field/major.
   1. Sketch out the possible data generating process.

The research question I think of is “does higher exposure in live-streaming in a month increase the monthly active players of a game (GMAP)?”. The GMAP can be directly affected by the GameQuality, LiveStreamingExposure, Promo, FreeTime: A good quality of a game will naturally attracts more players to play the game; High exposure in live-streaming may make more people know and try the game; Promotions like in-game purchase sales in certain period would attract more players; The free time that people have to spend on entertainment also affect GMAP. Game quality is affected by a game’s affordances, visuals, music, story setting, interaction, system, and server, etc. Different live-streaming platforms with varying traffics media, live-streamers with distinct influences on the platforms, time and length of the live-streaming of the game all affect the live-streaming exposure of a game. A good game quality may also cause more live-streamers to broadcast the game and the more free time the people have, the more likely the game’s live streaming reaches more audience.

* 1. What is the cause of interest? The outcome?

The cause of interest is the game’s exposure in live streaming in a month (LiveStreamingExposure). The outcome is the amount of monthly active players of the game (GMAP).

* 1. What other variables are in play? GameQuality, Promo, FreeTime,
  2. Draw a causal diagram depicting the relationships between all of the variables?

Likes

LiveStreamingLength

LiveStreamingTime

FreeTime

System

Interaction

StorySetting

Server

GMAP

LiveStreamingExposure

GameQuality

Music

Visuals

Affordances

Promo

LiveStreamer

LiveStreamingPlatform

* 1. Can there be any unobserved or latent variables? Include them in the diagram.

The GameQuality is an unobserved variable. People’s preference of the game can also be an unobserved variable (which can be affected by GameQuality).

1. Consider this research question: Does the inclusion of “free shipping” cause people to buy items from an online store more?
   1. List six variables that should be included in a causal diagram.

Returning Fees, Online Product Information, Product Price, Delivery Time, Rate of Internet Access, Product Warranty

* 1. Is it feasible to collect data on all the variables that you listed in part a? Can the variables be measured easily?

It is feasible to collect data on the variables that I listed in part a because Returning Fees, Product Price, Delivery Time, and Rate of Internet Access are measurable quantitative variables whose values can be obtained from shopping and official websites. The variables Online Product Information and Product Warranty are qualitative variables that are also available on the shopping websites. Though they can be hard to measure, we can boil them down to quantitative variables, for example, instead of trying to describe all the details of the product information and warranty policy, by asking “how many items of parameters are given in the product description” and “how long is the product warranty” and summarizing them instead.

1. Define *causality*. In a few sentences, why is causality interesting and important?

Causality refers to a causal relationship that the pattern of changes in one variable X cause changes in the distribution of another variable Y. The direction of influence is from cause X to its effect Y and their relationship is nonspurious, namely the association between the variables cannot be explained by an extraneous variable. Causality is interesting and important because besides some descriptive research questions, many of the research questions we are interested in are causal in nature. Even non-causal analyses and research questions have a causal question lurking underneath. Cause-and-effect connections allow us to draw inferences and externalize our findings to interpret the larger universe. We depend on causality every day to explain what happened, make realistic predictions, and to take corresponding actions to try to affect the future.

1. Which of the following describes a representation of a data generating process (DGP) including variables in that DGP and the causal relationships between them? d
   1. Causality
   2. Direct and indirect effect
   3. Latent variable
   4. Causal diagram