Bayesian Networks and Causal Inference
Lecture Notes Problem Answers
Chapter 1

Daan Brugmans
S1080742

September 8, 2023

Problem 1.1

For my bachelor's degree, I did a graduation internship about the feasibility of training machine learning models on code, natural language relevant to the code and the relationships between the two. For this project, I focused on the *Prediction* task of data science: the goal of the internship was to develop something that could predict if a piece of code and a piece of natural language were related to one another.

An example of a *Description* task for this project could be generating a set of graphs describing the distribution of the data. Although this does require that numerical features have already been made based on the natural language data, a visualization of the distribution of these features could provide insight into the contents of the data.

An example of a *Causal inference* task for this project could be determining which factors in the natural language data had what effects on the corresponding code. Within the business domain of my internship project, the natural language data would describe the design and requirements for a piece of software, and the code would be an implementation of that. Already there is some sort of causality taking place: the code is the result of the natural language data. One might construct a Bayesian network to visualize the relationships between different aspects of the natural language data, like user stories, requirements, test cases, etc., to determine what effects they have on the resulting code. That should be an example of causal inference and with that knowledge, one might be able to determine how to update the design process so that it may lead to better code.

Problem 1.2

- Income and marriage have a high positive correlation, because the union of two persons in marriage also results in the union of two persons' incomes. An individual in the marriage does not see their earnings rise just because they marries someone; earnings only increase on the level of the union.
- As the amount of fires increases, the demand for firefighters might rise. This demand is a direct result of the amount of fires occurring: the increase in fires is the cause and the increase in firefighters is the effect. Therefore, changing the amount of firefighters will not have an effect on the amount of fires, as the relationship only goes one way.
- People that hurry do so *because* they are late. The claim about the data formulates this the other way around. Instead of being too late for meetings being the effect of hurrying, it is hurrying that is the effect of being too late for meetings.

	X coefficient
Y ∼X	4.270047
Y ∼X + Z1	0.863585
Y ∼X + Z1 + Z2	-1.049287

Problem 1.3

The source code for this problem may be found in the "Problem Chapter 1.R" file.

The following is a table of the X coefficients I got for one of generated data sets:

Without any covariates, the X coefficient of the linear model is over 4, clearly a positive trend.

However, when the Z1 covariate is introduced, the X coefficient goes down: although it is still positive, it is now only 0.8. The sign has not fully flipped yet, but it is heading in that direction.

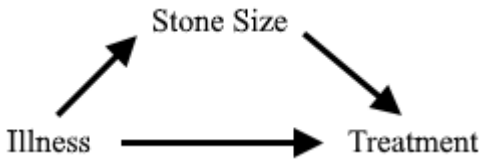
When the Z2 covariate is also introduced, the X coefficient becomes less than -1. The introduction of the two covariates has caused the sign of the linear model to flip.

Problem 1.4

...
(Can't figure out how to build a dataset that gives the desired averages. I have tried giving Kim less data points than Pem, but having those few data points be higher on average. This did not do the trick for me.)

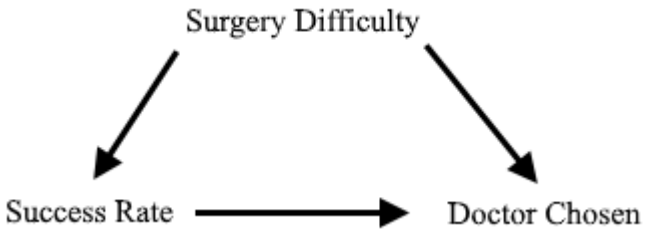
Problem 1.5

- In this case, stone size has a direct effect on the kind of treatment used. This means that stone size is a mediator for the treatment chosen. This may be represented like so:



  Because stone size is a mediator, we should *not* condition on it and use the aggregate data.

- The chosen doctor for the surgery depends on two variables: the success rate of each doctor and the difficulty of the operation. The success rate itself is partly determined by the difficulty of the operation: generally, the more difficult an operation is, the lower the success rate of that operation will be. Abiding by the idea that the difficulty of an operation affects the success rate, and not the other way around, the relationship between the doctor chosen, surgery difficulty and success rate may be visualized as follows:



  In this case, surgery difficulty is a confound for the success rate and the doctor chosen. Therefore, we should condition on surgery difficulty and use the segregated data.