Causal Inference Assignment - Part I

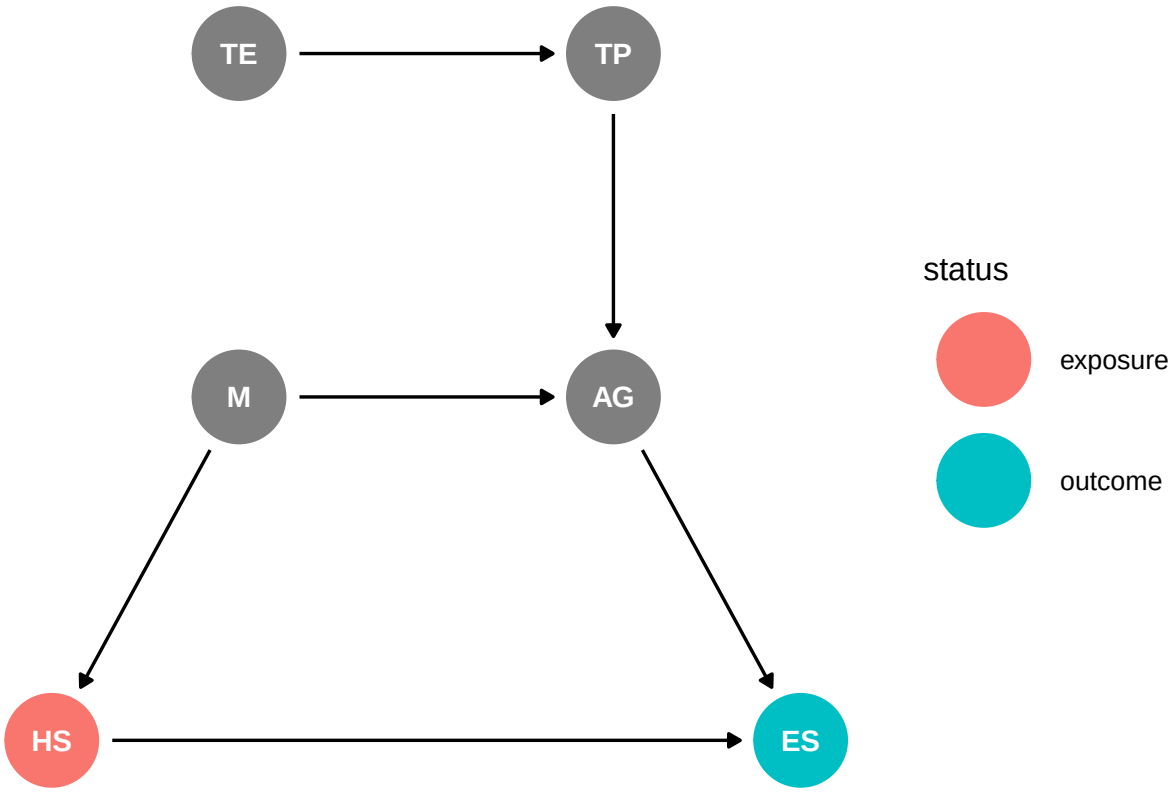
Lotte Mensink (9585842), Nina van Gerwen (1860852)

07/03/2022

1. Our Chosen Directed Acyclic Graph

For the assignment, we chose to make a Directed Acyclic Graph (DAG) about the causal relation between hours spent time studying for a mathematics exam, and final mathematics exam scores. The following variables are included in the DAG:
- ES: Mathematics exam scores. The exam scores range from 1 to 10 points.
- AG: GPA for mathematics. GPA ranges from 1 to 10 points.
- HS: Hours spent studying for the mathematics exam. The hours spent studying range from 10 to 40.
- M: Motivation to study for mathematics. The motivation to study ranges from 1 to 10, as measured by a motivation scale.
- TP: Teacher proficiency in mathematics. The teacher proficiency ranges from 1 to 20, as measured by a proficiency scale.
- TE: Years of teacher experience. Teacher experience ranges from 0 to 20 years.

The causal system is pictured in a DAG below.



The idea behind our DAG is that your exam scores for a course are dependent on two things:
a) the hours spent studying for the exam, and
b) your GPA for that course.
Both of these factors, however, are influenced by your motivation to study. Furthermore, your GPA also depends on the proficiency of your teacher for that course, which in turn is affected by the experience (s)he has.

2. Specifying the structural causal model

Assuming that all relationships between the variables are linear and that they have normally distributed residuals, we assumed the following structural causal model (SCM):

TE measures the years of teacher experience. The teacher experience ranges from 0 to 25 years. The mean teacher experience is 8.

$$TE := 8 + \epsilon_{TE} \text{ , where } \epsilon_{TE} \sim \mathcal{N}(0, 2.5)$$

,

TP measures teacher proficiency in mathematics. The teacher proficiency ranges from 1 to 20. Since mean teacher experience is 8, mean teacher proficiency is 10.

$$TP := 6 + .5TE + \epsilon_{TP} \text{ , where } \epsilon_{TE} \sim \mathcal{N}(0, 2)$$

,

M measures the motivation to study for mathematics. The motivation to study ranges from 1 to 10. The mean motivation is 5.

$$M := 5 + \epsilon_{M} \text{ , where } \epsilon_{M} \sim \mathcal{N}(0, 1)$$

,

HS measures the hours spent studying for the mathematics exam. The hours spent studying range from 10 to 40. Since mean motivation is 5, the mean for hours spent studying is 25.

$$HS := 10 + 3M + \epsilon_{HS} \text{ , where } \epsilon_{HS} \sim \mathcal{N}(0, 5)$$

,

AG measures average GPA for mathematics. The average GP ranges from 1 to 10 points. Since mean motivation is 5, and mean teacher proficiency is 10, mean GPA is 6.25.

$$AG := 3 + .35M + .15TP + \epsilon_{AG} \text{ , where } \epsilon_{AG} \sim \mathcal{N}(0, 1)$$

,

ES measures the mathematics exam scores. The exam scores ranges from 1 to 10 points. Since the mean for hours spent studying is 25, and mean average GPA is 6.25, the mean for exam scores is 5.69.

$$ES := 2.5 + .1HS + .1AG + \epsilon_{ES} \text{ , where } \epsilon_{ES} \sim \mathcal{N}(0, 1)$$

,

3. Generating data from our SCM

To generate data from our above specified SCM, we set a specific seed and generate data using the above formulas. For the residuals, we made use of the function rnorm() with the appropriate values.

```
# setting random seed for generating
set.seed(123)
# setting the sample size to 500
n <- 500
# generating the data based on the equations specified above
TE <- 8 + rnorm(n, 0, 2.5)
TP <- 6 + 0.5*TE + rnorm(n, 0, 2)
M <- 5 + rnorm(n, 0, 1)
HS <- 10 + 3*M + rnorm(n, 0, 5)
AG <- 3 + .35*M + .15*TP + rnorm(n, 0, 1)
ES <-  1 + .1*HS + .35*AG + rnorm(n, 0, 1)
data <- as.data.frame(cbind(TE, TP, M, HS, AG, ES))
```

A summary of the data is provided in below. We can observe that the minimums and maximums for all variables fall inside the ranges specified in the definition of our DAG. Furthermore, we can see that the means are roughly (with some error, due to the random error term) the same as in the definition of our DAG.

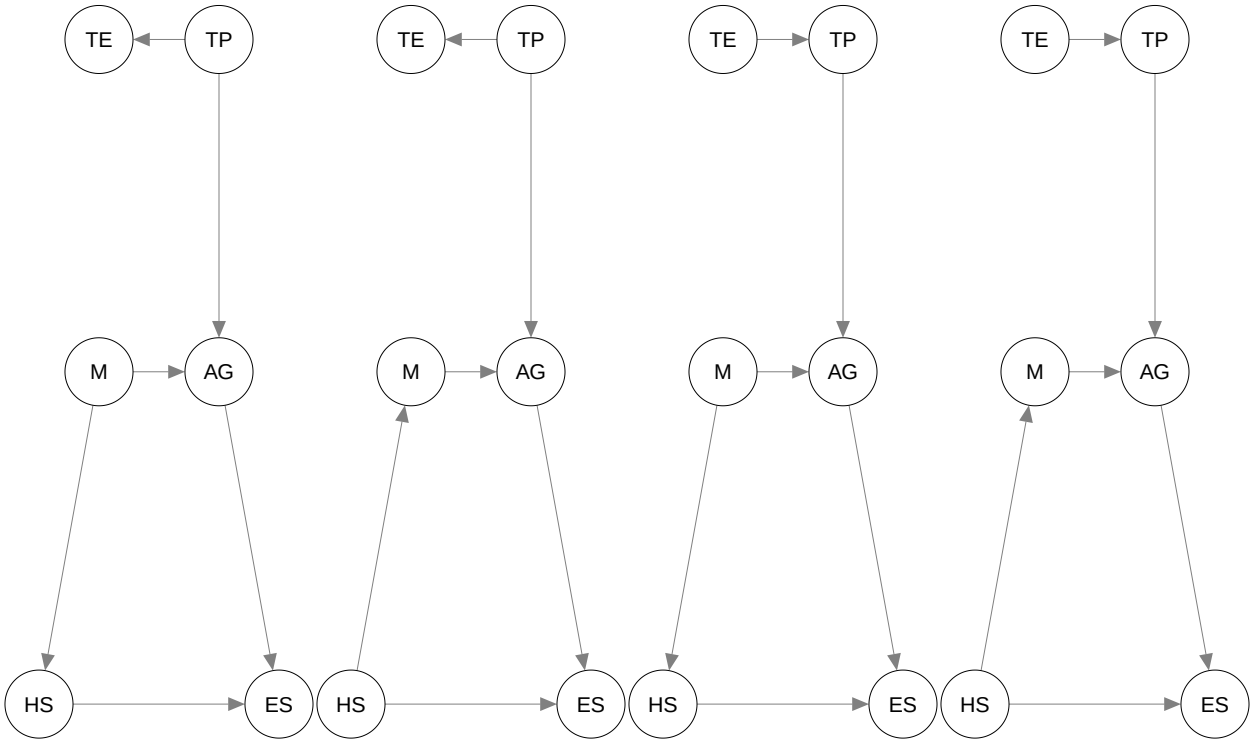
	TE	TP	M	HS	AG	ES
	Min. : 1.348	Min. : 3.430	Min. :2.305	Min. :11.41	Min. :2.593	Min. :2.539
	1st Qu.: 6.563	1st Qu.: 8.319	1st Qu.:4.317	1st Qu.:21.61	1st Qu.:5.480	1st Qu.:4.845
	Median : 8.052	Median :10.085	Median :5.060	Median :25.37	Median :6.167	Median :5.742
	Mean : 8.086	Mean :10.039	Mean :5.026	Mean :25.37	Mean :6.201	Mean :5.731
	3rd Qu.: 9.713	3rd Qu.:11.496	3rd Qu.:5.660	3rd Qu.:29.32	3rd Qu.:6.931	3rd Qu.:6.605
	Max. :16.103	Max. :17.413	Max. :8.390	Max. :38.98	Max. :9.242	Max. :9.844

4. Using the PC-algorithm on our generated data

A. Is our true DAG in the markov equivalence class?

In order to find the Markov Equivalence Class, we first use the data in combination with the PC algorithm to obtain the CPDAG. From the CPDAG, we can infer the DAGs in our Markov Equivalence class. The CPDAG, as pictured in Figure 1, has two undetermined arrows that could go either way. Therefore, we have four DAGs in our Markov Equivalence Class, as pictured below.

Inspecting the Markov Equivalence Class, we can see that our true DAG is provided by the PC-algorithm. The third DAG in the markov equivalence class shows the causal system as we had theorized and specified it.



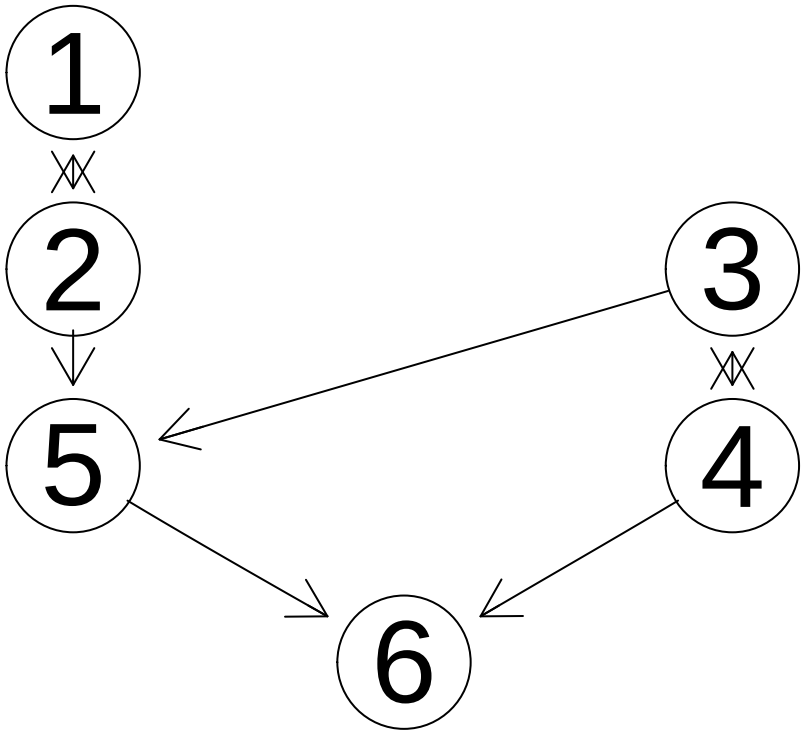
B. Provide the CP-DAG. To what extent did the procedure correctly recover which relationships were absent/present/directed?

Using the PC algorithm on our generated data correctly recovered all relationships that were present in the DAG! Furthermore, both colliders in the data set (GPA and exam scores, 5 and 6 respectively in the CP-DAG) also correctly had their directions recovered. This is to be expected as colliders have a distinct statistical dependence structure compared to mediators and confounders, which are statistically identical. However, getting the PC-algorithm to correctly uncover the CP-DAG took some trial and error.

When we began, we had chosen another regression coefficient for the relationship between GPA and exam scores (namely, 0.10). However, when we used the PC-algorithm with that regression coefficient, it did not correctly recover the relationship between GPA and exam scores. A plausible reason for this is that the effect of GPA on exam was too small to be uncovered with a dataset of only 500 observations. Hence, we simulated a larger data set (10000) and found that the algorithm did recover the correct CP-DAG.

In the end, we decided to change the SCM so that with a dataset of 500, the PC-algorithm still correctly uncovered the CP-DAG through increasing the effect of GPA on exam scores. Now, for every extra point in GPA, your exam score increases by 0.35. Furthermore, we changed to intercept to a more meaningful value. Namely, 1, the most likely grade someone would get had they not studied any hours and had a very low GPA.

Figure 1: Inferred CPDAG using PC Algorithm



5.Two variables for which the causal relationship is to be estimated

The two variables for which the causal relationship should be estimated are:
a) Cause variable: Hours Studied (HS)
b) Outcome variable: Exam Scores (ES)

6. Estimate the Causal Effect

A. What is the true causal effect of the cause on the outcome variable based on your SCM?

To determine the true causal effect of hours studied (HS) on exam scores (ES), we take a look at the regression equations we used to generate the data.

$$ES := 2.5 + .1HS + .1AG + \epsilon_{ES} \text{ , where } \epsilon_{ES} \sim \mathcal{N}(0, 1)$$

From the regression equation, we can observe that there is a direct effect from hours studied on exam scores, as quantified by the number 0.1. Thus, the true causal effect of the cause on the outcome variable is 0.1 in our structural causal model.

B. Based on the true DAG, what linear regression model should be used to estimate the causal effect correctly?

Based on our true DAG, we can infer the linear regression model we should use to correctly estimate the causal effect. To do this, we use d-separation. In determining which variables should and should not be controlled for in analyzing the causal effect between HS and ES, we use the following guidelines:
- Block all backdoor paths
- Don't open up any spurious paths
- Leave all directed paths you care about intact

When we look at our true DAG, we can observe that there is one backdoor path into hours studied (HS). This backdoor path should be blocked. We can block the backdoor path either by controlling for motivation (M) or average GPA (AG). It should be noted that average GPA is a collider between motivation and teacher proficiency, meaning that conditioning on this variable would open up a spurious path between M and TP. This would not change the causal effect between HS and ES, since it would block the path between these two variables. However, we choose to condition on M, rather than AG, because we do not want to open any spurious paths (even though it makes no difference in estimating the ACE). In theory, we could also control for the other variables, teacher experience (TE) and teacher proficiency (TP) as well, since this would not change the estimate of the ACE. However, we choose not to do this since we consider them directed paths we care about. Thus, we decide to condition on motivation only, to block the backdoor path into hours studied.

C. Estimate the causal effect. To what extent is the true effect recovered?

We estimate the causal effect by fitting a simple linear regression model with ES as the dependent variable, HS as a predictor, and M as a covariate, by executing the code below:

```
# estimate causal effect between ES and HS by controlling on M
ce_fit <- lm(ES ~ HS + M, data = data)
# obtain the coefficient estimate for HS: our ACE
round(ce_fit$coefficients[[2]], 3)
```

```
## [1] 0.113
```

The estimate that we obtain for the causal effect (0.113) is quite similar to the true causal effect, which is 0.100. They deviate 0.013 from each other. This is probably due to the random error term that was included during the generation of the data. One possible explanation for the fact that the regression coefficient is biased upwards a little bit, is that the mean for our dependent variable exam scores is also biased upwards a little bit (the expected value based on the regression equations is 5.69, whereas the mean in our dataset is 5.74).

7. Make a dichotomous cause variable

A. Dichotomize hours studied (HS)

We dichotomized our cause variable (hours studied) simply by assigning people who scored below the mean (i.e., low scorers) a 0 and assigning people who scored above the mean (i.e., high scorers) a 1.

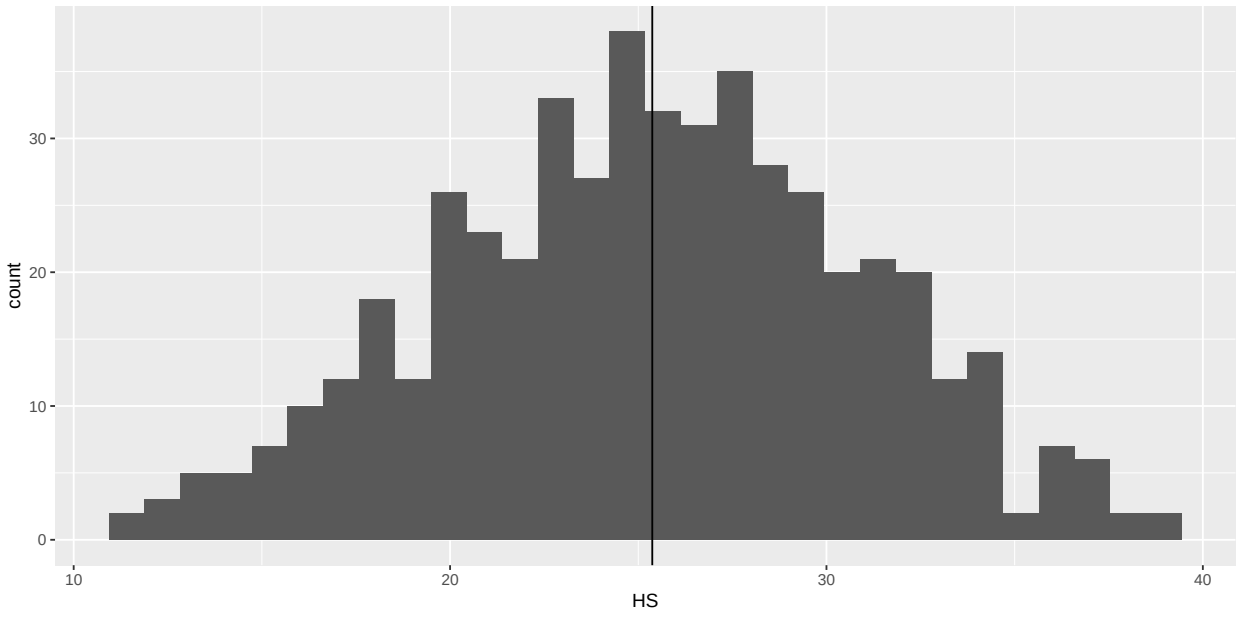
B. Estimating the causal effect with the dichotomous cause variable

We estimate the causal effect as we did for question 6, but now with our dichotomized cause variable, by executing the code below:

```
ce_fit_dich <- lm(ES ~ HS_Dichot + M, data = data)
round(ce_fit_dich$coefficients[[2]], 3)
```

```
## [1] 0.94
```

Estimating the causal effect of the dichotomized hours studied on exam scores through a linear regression, we find a larger estimate of the causal effect (0.940) compared to the continuous version (0.113). Whereas the original causal effect estimate was quite accurate, this estimate is way off. We expected the estimate of the causal effect to be further away from the true causal effect when using a dichotomized variable. This is because by dichotomizing the variable, we in essence get rid of a lot of information that the data provides. Instead of getting specific values for how long someone studied, we only now know whether someone studied less or more than the average amount studied. The histogram below illustrates the dichotomization of the HS variable. The values to the left of the bar are assigned a 0, the values to to right are assigned a 1. Most values are around the mean, and even though these values do not differ much from each other they are assigned either a 1 or a 0. We suspect that because this is theoretically unsensible thing to do, the estimate for the causal effect is very biased.



8. Preparing data for the second assignment

The dataframe has been saved as an .RData file and a .txt file has been added as a codebook. The codebook contains the following information:

The exam_data.rdata file contains the following variables, ordered alphabetically:

- AG: GPA for mathematics. GPA ranges from 1 to 10 points.
- ES: Mathematics exam scores. The exam scores range from 1 to 10 points.
- HS: Hours spent studying for the mathematics exam. The hours spent studying range from 10 to 40.
- HS_Dichot: A dichotomized version of HS with a 0 for people who scored below the mean and a 1 for people who scored above the mean.
- M: Motivation to study for mathematics. The motivation to study ranges from 1 to 10, as measured by a motivation scale.
- TE: Years of teacher experience. Teacher experience ranges from 0 to 25 years.
- TP: Teacher proficiency in mathematics. The teacher proficiency ranges from 1 to 20, as measured by a proficiency scale.

The variable which is the main cause of interest will be both HS and HS_Dichot, whereas the variable which is the main outcome of interest is ES.