# Exercises

## Contents

# 1 Exercises

1. Consider your own research project. Draw a DAG that describes the study. Which variables do you need to adjust for? Are there any variables that you should not adjust for?

2. (*Non-compliance in randomized experiments*). A common feature of randomized experiments is that subjects do not always adhere to their assigned treatment. In this exercise we investigate why this feature may be problematic. We consider a study in which each subject is randomly assigned either to a new treatement or a standard treatment. The study is unblinded, i.e. the participants are aware of which treatment they are asigned to. Some subjects who are assigned to the new treatment eventually decide to take the standard treatment and vice versa. Thus, we distinguish between *treatment assignment*, which we denote with $R$ (0 for 'assigned to standard treatement', 1 for 'assigned to new treatement'), and *treatment actually taken*, which we denote with $A$ (0 for 'taking standard treatement', 1 for 'taking new treatement'). For each subject, a binary outcome $Y$ is measured (0 for 'unfavorable outcome', 1 for 'favorable outcome').

   (a) Draw a DAG that represents this study.

   (b) One possible way to analyze this data is to compare the outcome for those who actually took the new treatment ($A = 1$) versus those who actually took the standard treatment ($A = 0$). In the literature, this analysis is usually referred to as the 'as-treated' (AT) analysis. Use the DAG to explain why the AT analysis may be problematic from a causal inference point of view.

   (c) An alternative approach is to compare the outcome for those who were randomized to the new treatment ($R = 1$) versus those who were randomized to the standard treatment ($R = 0$). This analysis is usually referred to as the 'intention-to-treat' (ITT) analysis. Use the DAG to explain the rationale behind the ITT analysis.

3. (*Intrumental variables*). Many observational studies suffer from confounding. In this exercise we investigate a method of 'confounding adjustment' which, under certain assumptions, has the remarkable property of producing causal inference even in the presence of unmeasured

confounding. Let $A$ be the exposure of interest, let $Y$ be the outcome of interest, and let $U$ be all unmeasured variables (confounders) that affect both $A$ and $Y$. Let $Z$ be a measured variable which have the following properties: a) $U$ does not affect $Z$, b) $Z$ does not affect $U$, c) $Z$ and $U$ don't have common causes, d) $Z$ affects $A$, e) $Z$ has no effect on $Y$, apart from an indirect effect mediated through $A$. A variable $Z$ which have properties a)-e) is called an *instrumental variable*.

(a) Draw a DAG that connects $A$, $Y$, $U$, and $Z$.

(b) Show that an observed association between $Z$ and $Y$ implies that $A$ has a causal effect on $Y$ (that is, we can test whether $A$ has a causal effect on $Y$ by testing whether $Z$ an $Y$ are associated).

(c) Try to come up with a real epidemiological scenario which may be represented by your DAG, to a reasonable degree of approximation.

4. (*Measurement errors*). Suppose that we want to test whether high protein diet is beneficial with respect to various health indicators, compared to an ordinary diet. Each subject in a large cohort is randomized to high protein diet ($A = 1$) or ordinary diet ($A = 0$). After 6 months each subject is asked to fill in a detailed questionaire. One question concerns weight loss during the last 6 months. Unfortunately we have reason to believe that not all subjects are totally honest when answering this specific question. Thus, we distinguish between true weight loss ($Y$) and reported weight loss ($Y^*$).

   (a) Draw a DAG that represents this study.

   (b) Given your DAG, can you use the association between $A$ and $Y^*$ to *test* for a causal effect of $A$ on $Y$?

   (c) If your answer is 'no' to the previous question, what additional assumptions would you have to make (i.e. how would you have to modify your DAG) in order for you to give an affirmative answer? If your answer is yes, can you then see any problem at all with the fact that the study suffers from measurement errors?

5. (*Post treatment selection bias; due to Hernandez-Diaz et al 2006*). Birth weight is a strong predictor of neonatal and infant mortality. Probably for that reason, and because birthweight data are readily

available, investigators have frequently stratified on birth weight when evaluating the effect of other risk factors (e.g., maternal smoking, multiple pregnancies, placenta previa) on infant mortality. This stratification often produces a rather counter-intuitive result: infants who are exposed to the particular risk factor (e.g. infants whose mother smoked during pregnancy) have a lower mortality rate than infants who are not exposed to the risk factor. This phenomenon is known as the 'birth weight paradox', and it has been a source of controversy for decades.

(a) Draw a DAG that represents the causal relations betwen smoking during pregnancy, birth weight, and infant mortality.

(b) Use the DAG to discuss possibly explanations of the birth weight paradox.

6. Read the article 'Cigarette smoking and the incidence of Parkinson's disease in two prospective studies' (*Annals of Neurology* 2001; 50:780-786).

   (a) Summarize the designs and methods of the study.

   (b) Summarize the results of the study, i.e. what associations did the authors find?

   (c) In the first column of page 784, a paragraph begins with 'The key question is whether this strong inverse association reflects a truly protective effect of smoking on the risk of developing PD.' Draw at least one DAG that is consistent with the explanations given in this paragraph.

   (d) In the first column of page 785, a paragraph begins with 'There are also several versions of the argument claiming the existence of a causal effect of PD on smoking behavior...' Draw at least one DAG that is consistent with the explanations given in this paragraph.

   (e) In the first column of page 785, a paragraph begins with 'Confounding...' Draw at least one DAG that is consistent with the explanations given in this paragraph.

   (f) In the first column of page 784, a paragraph begins with 'The information bias...' Draw at least one DAG that is consistent with the explanations given in this paragraph.

(g) In the second column of page 784, a paragraph begins with ‘There are a number of variations of the hypothesis that selection bias...’ Draw at least one DAG that is consistent with the explanations given in this paragraph.

# 2 Solutions

1. Discuss in class.

2. (a) See Figure 1. Here, $U$ represents all (usually unmeasured) factors that affect both treatment taken and the outcome. For instance, subjects with a bad health may be more eager to take the treatment, and also more likely to have an unfavorable outcome.

$$
\begin{tikzcd}
& & U \arrow[dl] \arrow[dr] & \\
R \arrow[r] & A \arrow[rr] & & Y
\end{tikzcd}
$$

Figure 1: DAG for a randomized experiment with non-compliance.

   (b) An association between $A$ and $Y$ may be explained by the non-causal path $A \leftarrow U \rightarrow Y$.

   (c) If the arrow from $A$ to $Y$ is missing, then there would be no association between $R$ and $Y$, since the path $R \rightarrow A \leftarrow U \rightarrow Y$ is blocked at $A$. Hence, if we observe an association between $R$ and $Y$, then we can say that the arrow from $A$ to $Y$ exists.

3. (a) See Figure 2.

$$
\begin{tikzcd}
& & U \arrow[dl] \arrow[dr] & \\
Z \arrow[r] & A \arrow[rr] & & Y
\end{tikzcd}
$$

Figure 2: DAG for an instrumental variable setting.

   (b) If the arrow from $A$ to $Y$ is missing, then there would be no association between $Z$ and $Y$, since the path $Z \rightarrow A \leftarrow U \rightarrow Y$ is blocked at $A$. Hence, if we observe an association between $Z$ and $Y$, then we can say that the arrow from $A$ to $Y$ exists.

(c) Discuss in class.

4. (a) See Figure 3. The arrow from $Y$ to $Y^*$ reflects the fact that true weight loss (hopefully) influences reported weight loss. $U$ is the set of all factors that affect both $Y$ and $Y^*$. For instance, young people may loss more weight when put on a high protein diet, but may also tend to overestimate (or underestimate!) their true weight loss, as compared to old people. The arrow from $A$ to $Y^*$ represents the influence of diet on reported weight loss. For instance, people who have followed the high protein diet may feel more confident that they have lost weight than people who have followed the ordinary diet (a placebo effect).

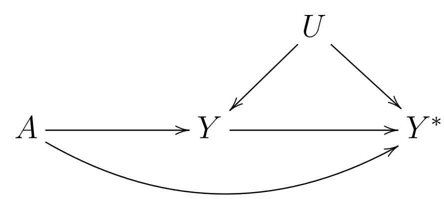


Figure 3: DAG for a randomized study with measurement error in the outcome.

(b) No; an association between $A$ and $Y^*$ could be explained by the path $A \rightarrow Y^*$.

(c) If the arrow from $A$ to $Y^*$ is absent, then an association between $A$ and $Y^*$ proves that the arrow from $A$ to $Y$ exists. However, the association between $A$ and $Y^*$ is probably weaker then the association between $A$ and $Y$. Thus, even though we can test for a causal effect, we cannot estimate its mganitude. Also, measurement errors are likely to reduce the power of the study.

5. (a) See Figure 4. Here, $U$ is the set of all factors (e.g. genetics and lifestyle) that may influence the child's birth weight and the child's risk of death.

(b) The 'birth weight paradox' may be explained by the non-causal path SDP $\rightarrow$ BW $\leftarrow$ $U$ $\rightarrow$ Mortality; this path becomes open when stratifying on birth weight.

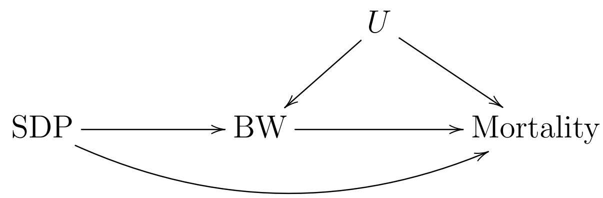


Figure 4: DAG for the causal relations between smoking during pregnancy (SDP), birth weight (BW), and infant mortality.

6. Discuss in class.