Exercises

Contents

1 Association vs causation

1. NEED EXERCISES ON ASSOCIATION

2. Table 1 shows the data collected in a study of 12 individuals. The goal was to estimate the effect of daily low-dose aspirin ($A = 1$) versus no aspirin ($A = 0$) on the risk of heart disease ($Y = 1$). The table also shows the values of the potential outcomes that would have been observed under treatment (Y_1) and under no treatment (Y_0) with aspirin.

ID	A	Y	Y_1	Y_0
1	0	1	0	1
2	0	1	0	1
3	0	0	0	0
4	0	0	0	0
5	0	1	1	1
6	0	1	1	1
7	1	0	0	1
8	1	0	0	1
9	1	0	0	0
10	1	0	0	0
11	1	1	1	1
12	1	1	1	1

Table 1: Potential outcome data

   (a) Compute the causal risk difference, the causal risk ratio, and the causal odds ratio.

   (b) Compute the causal effects listed in a) for the subpopulation that was factually exposed.

   (c) Compute the causal effects listed in a) for the subpopulation that was factually unexposed.

3. Consider the question 'is there a causal effect of A on Y?' applied to each of the pairs (A, Y) displayed in Table 2. For each pair,

A	$A=1$	$A=0$	Y
Aspirin	Daily use (150 mg)	No use	Coronary heart disease
Low salt diet	Yes	No	Stroke
sex	Female	Male	Lung cancer

Table 2: Causal research questions

(a) do you think the causal question above is appropriate for meaningful inference (i.e. is the causal effect of interest well defined?).

(b) If not, can you propose a more precise causal question (possibly involving a modified version of A) that reduces the vaguenesss of the original question?

4. In this course (as well as in the research field of causal inference in general) we assume that the potential outcome for each subject under each treatment is deterministic (i.e. non-random). This implies, for example, that with respect to a binary exposure A ($A=0$ for untreated and $A=1$ for treated) and a binary outcome Y ($Y=0$ for the favorable outcome and $Y=1$ for the unfavorable outcome), each subject *must* belong to one of the four groups displayed in Table 3. Subjects in the

Y^a	'healthy'	'harmed'	'protected'	'doomed'
$A=0$	0	0	1	1
$A=1$	0	1	0	1

Table 3: Principal stratification

first group do not develop the outcome, regardless of whether they are treated or not - we may call them 'healthy'. Subjects in the second group develop the outcome if and only if they are treated - we may call them 'protected'. Subject in the third group develops the outcome if and only if they are *not* treated - we may call them 'harmed'. Subjects in the fourth group develop the outcome regardless of whether they are treated or not - we may call them 'doomed'. This classification of subjects based on joint potential outcomes is usually referred to as 'principal stratification'. Many people feel that the deterministic nature

of potential outcomes and principal stratification is unrealistic (this is especially true for statisticians, who are used to model everything in a random fashion). What is your opinion? Does your answer depend on what particular exposure and outcome we consider? Does your answer depend on what we mean with the word 'subject'?

5. A respectable researcher claims that she can compute individual causal effects in her research area. This claim implies that the fundamental problem of causal inference - the impossibility of observering a subject's outcome under two different values of the exposure - does not apply to her research. Do you think that such a claim can ever be correct? If yes, give an example of a study design that would allow for the computation of individual causal effects. If no, motivate.

6. Consider your research project. Is your project motivated by a causal question (i.e. is your aim to estimate a causal effect)? If so, can you formulate your research question in terms of potential outcomes?

2 Estimation of causal effects

1. Consider the data in Table 1. Are treated ($A = 1$) and untreated ($A = 0$) exchangeable?

2. Suppose that you are given the first three columns of Table 1, i.e. ID, A, and Y, but not the potential outcomes Y_0 and Y_1.

   (a) Given the first three columns, can you rule out the causal null hypothesis CRD=0? If not, fill in example values of Y_0 and Y_1 for each subject so that CRD=0.

   (b) Given the first three columns, can you rule out that CRD=1? If not, fill in example values of Y_0 and Y_1 for each subject so that CRD=1.

   (c) Given the first three columns, which are the possible values for CRD? Given no data at all, which are the possible values for CRD?

3. Table 4 shows the data from a study to compute the causal effect of antiretroviral therapy A on death Y in subjects infected with HIV.

Individuals were classified as treated ($A = 1$) if they received antiretroviral therapy, and as untreated ($A = 0$) otherwise. Death is coded as $Y = 1$. The variable L represents CD4 count (1:low; 0:high). Assume that the treated and untreated are exchangeable, given L.

	$L = 0$		$L = 1$	
	$Y = 1$	$Y = 0$	$Y = 1$	$Y = 0$
$A = 1$	20	30	108	252
$A = 0$	40	10	24	16

Table 4: HIV data

(a) Compute the causal effect of A on Y, given $L = 0$ and $L = 1$.

(b) Compute the causal effect of A on Y.

(c) Can you compute the causal treatement effect for those who actually received the treatment? Is this a relevant parameter?

4. Consider your research project. If your aim is to estimate a causal effect, do you think that exposed and unexposed are exchangeable with respect to the outcome of interest in your study population? If not, which covariates do you think you need to adjust for in order to achieve exchangeability?

3 Directed Acyclic Graphs

1. Consider your research project. Draw a DAG that describes your study. What do you need to adjust for? Anything 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 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) which affects 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) Suppose that we carry out this test using *finite* samples. Discuss how the statistical power of the test (i.e. the chance of rejecting the null hypothesis, given that it is false) depends on the strength of each separate arrow on your DAG.

(d) 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 questions, 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*). It is now well established that long-term use of hormone replacement therapy (HRT) causes breast cancer (BC). Suppose that we want to investigate whether HRT also affects the prognosis (i.e. survival) for BC patients. We randomize each subject in a large cohort to HRT ($A = 1$) or no HRT ($A = 0$). The cohort is followed over time and each BC case ($Z = 1$) is recorded. For those subjects who develop BC, the survival time (Y) is recorded as well. Subjects who don't develop BC during follow up ($Z = 0$) are ignored in the analysis.

   (a) Draw a DAG that represents this study.

(b) Suppose that we observe an association between A and Y, either positive or negative, for those subjects who develop BC ($Z = 1$). Does this association prove that A has an effect on Y?

(c) Suppose that A does not have any effect on Y. Would you then expect A and Y to be independent, positively associated, or negatively associated?

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.

7. Use d-separation to find all (conditional) independencies between the variables on the DAG in Figure 1.

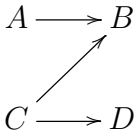

Figure 1:

8. For each of exercies a), b) and c) below, draw a DAG that contains four variables A, B, C, and D. Each DAG should imply the (conditional) independencies listed in the corresponding exercise, and *only* these independencies.

   (a)

   $$A \amalg C|B$$

   (b)

   $$A \amalg C|B, D$$

   (c)

   $$A \amalg C|B, D$$

   and

   $$B \amalg D|A$$

4 Time varying exposures

1. Consider the DAG in Figure 2. Suppose that the observed number of subjects having each combination of (A_0, L, A_1, Y) are as in Table 5.

   (a) Given the population proportions of (A_0, L, A_1, Y), which of the arrows in Figure 2 can you test the presence of? Carry out all the feasible tests.

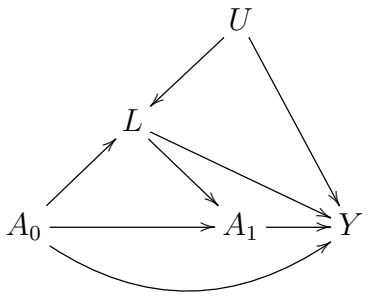

Figure 2:

A_0	L	A_1	$Y=0$	$Y=1$
0	0	0	168	132
0	0	1	42	33
0	1	0	32	68
0	1	1	8	17
1	0	0	102	48
1	0	1	68	32
1	1	0	48	102
1	1	1	32	68

Table 5: Data for the DAG in Figure 2.

(b) Use the G-formula to compute the direct effect of A_0 on Y, at both $A_1 = 0$ and $A_1 = 1$. Given your answer to the previous question, can you test for a direct effect of A_0 on Y (i.e. not mediated through A_1) using a 'standard' associational test? If so, do it and compare to the result of the G-formula.

5 Regression models

1. Load the data set 'point exposure.dta'. It contains the data set used in the lecture to illustrate regression models for point exposures.

   (a) Replicate the results showed in the lecture.

   (b) Suggest one alternative outcome model and one alternative exposure model, which make weaker assumptions than the models used in the lecture. Fit your suggested models, and use them to estimate standardized effects. Do your results differ qualitatively from the results in the lecture?

2. Load the data set 'time varying exposure.dta'. It contains the data set used in the lecture to illustrate regression models for time varying exposures.

   (a) Replicate the results showed in the lecture.

   (b) Suggest one alternative MSM which makes weaker assumptions than the model used in the lecture. Fit your suggested MSM. Interpret your results. Do they differ qualitatively from the results in the lecture?

   (c) Suppose that we are specifically interested in whether there is a direct effect of A_0 on Y, not mediated through A_1. Suggest a MSM which allow us to estimate this direct effect. Fit your suggested model. Conclusion?