12 Causality

In causal estimation one of the primary goals is to estimate the strength of the causal effect of one variable on another. In other words: We are interested in finding out how much the value of one variable influences the value of another variable. This will allow us to answer causal questions such as: "If I were to modify variable x, how much would I expect variable y to change?". This is not always an easy problem, and the answer most often depends on additional knowledge that is not evident in the statistical data alone.

Example 12.1 Bullet holes in air planes[1]

During the second world it was noticed that air planes returning from battle had more bullet holes in the fuselage than in the engine. At first, this finding suggested that more armour should be added to the fuselage to protect them where they seemed to get hit the most. However, a statistician who examined the case came to the opposite conclusion: They should add more armour to the engine. While at first this might seem backward, the argument is clear: If we assume that the bombers are equally likely to get hit everywhere, the fact that the air planes that return have more bullet holes in the fuselage suggests that planes that are hit in the engine are more likely to crash and therefore not return. The planes that do return are the ones that are hit where it does not matter so much.

Often causal estimation is discussed in the language of experimental design, where we imagine a hypothetical study where some units are given a treatment x and we measure the outcome y. The treatment could be a binary variable (some units recieve the treatment, others get no treatment) or it could be a

[1] This example is based on a true story as told by Jordan Ellenberg in the book "How Not to Be Wrong: The Power of Mathematical Thinking" regarding the statistician Abraham Wald who published his results in the 1943 report "A Method of Estimating Plane Vulnerability Based on Damage of Survivors"

continuous variable (for example the size of a dose of medicine). Similarly, the outcome could be binary or continuous.

12.1 Statistical dependence

Since causal estimation is concerned with measuring the causal dependence between variables, we begin by discusion how to measure statistical dependence. Here, it is easier to begin by defining when two variables are *not* associated, i.e. there is no statistical dependence.

Definition 12.1 Statistical independence

Two variables x and y are said to be statistically independent if there is no association between the variables: Knowing something about x tells us absolutely nothing about y and vice versa. Technically we say that the joint probability of x and y factorizes as the product of the individual distributions $p(x)$ and $p(y)$,

$$p(x, y) = p(x)p(y).$$

Example 12.2 Independent coin flips

If we flip a fair coin we get a random outcome of either heads or tails, and each outcome occurs with 50% probability. If we flip the coin twice and call the two outcomes x and y, we will observe one of the following sequences, each of which occur with probability 25%:

x	y	$p(x, y)$
heads	heads	0.25
heads	tails	0.25
tails	heads	0.25
tails	tails	0.25

In this case x and y are statistically independent: Knowledge about the outcome of one of the coin flips tells us nothing about the outcome of the other. We can verify this using the definition of statistical independence. Using the individual probabilities of the outcomes

$$p(x = \text{heads}) = p(x = \text{tails}) = 0.5$$
$$p(y = \text{heads}) = p(y = \text{tails}) = 0.5$$

we can verify that $p(x, y) = p(x)p(y) = 0.25$ for any value of x and y.

12.1.1 *Linear dependence*

Obviously, if two variables are not statistically independent, we say that they are statistically dependent. In that case, information about one of the variables tells us something about the other variable. The dependence between two variables x and y can either be positive or negative: If observing some particular value of x makes it more likely to observe some particular value of y, we say there is a positive dependence. If some particular value of x makes some particular value of y more unlikely, the dependence is negative.

Example 12.3 Dependent coin flips

Let us consider flipping a coin to get a random outcome of either heads or tails. Each of these possible outcomes occur with probability 50%. Let us call the outcome x, and let us define another variable y to be the *opposite* possible outcome. In other words, if x is tails, the y is heads and vice versa. The possible combined outcomes we can see and their probability would then be given by:

x	y	$p(x, y)$
heads	tails	0.5
tails	heads	0.5

Clearly, x and y are not independent, since they are always the opposite of each other. We can verify this by checking the definition of statistical independence, namely that the joint probabiliy is equal to the product of the individual marginal probabilities, $p(x, y) = p(x)p(y)$, for any values of x and y. As in the previous example we have that the probability of each individual outcome of x and y is 50%

$$p(x = \text{heads}) = p(x = \text{tails}) = 0.5$$
$$p(y = \text{heads}) = p(y = \text{tails}) = 0.5$$

However, the joint probability $p(x = \text{heads}, y = \text{tails}) = 0.5$

$$p(x = \text{heads}, y = \text{tails}) = 0.5 \\ \neq \\ p(x = \text{heads})p(y = \text{tails}) = 0.5 \cdot 0.5 = 0.25$$

This proves that x and y are not independent. In this case there is a negative dependence between observing $x = \text{heads}$ and $y = \text{heads}$, because if we see $x = \text{heads}$ it makes it less likely (impossible actually) to see $y = \text{heads}$.

One particular way that we can measure the degree of dependence between two variables is by their *covariance* and the related *correlation coefficient*. Both of these are measures of the *linear* dependence between two variables[2]. Covariance is defined in a similar way as variance is defined for a single variable. Where the variance is the average squared deviation from the mean, the covariance is the average product of the difference of the mean for the two variables.

Definition 12.2 Covariance and correlation

Covariance

The covariance between two variables is a measure of the linear dependence between two variables. It can be computed as:

$$\text{cov}(x, y) = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y).$$

Correlation coefficient

The correlation coefficient between two variables is their covariance normalized by the product of their standard deviations, and is computed as:

$$\rho_{x,y} = \frac{\text{cov}(x, y)}{\sigma_x \cdot \sigma_y}.$$

The correlation coefficient is a number between -1 and 1.

[2] Note that it is possible for two variables to have a non-linear dependence while having zero covariance; thus, if the covariance between two variables is zero, it does not necessarily mean that the variables are independent; however, if they are independent the covariance will be zero.

Example 12.4 Covariance and correlation coefficient

Consider two data sets which consist of five numbers each:

$$x = \{1, 8, 4, 10, 2\}, \quad y = \{7, 5, 4, 3, 6\}.$$

Recall from Example **??** that the means and standard deviations of these two data sets are:

$$\mu_x = 5, \quad \mu_y = 5, \quad \sigma_x \approx 3.46, \quad \sigma_y \approx 1.41.$$

Using the means, we can compute the covariance between x and y as

$$\begin{aligned}\operatorname{cov}(x, y) &= \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)\\ &= \frac{1}{5}\Big((1-5)(7-5) + (8-5)(5-5) + (4-5)(4-5)\\ &\qquad + (10-5)(3-5) + (2-5)(6-5)\Big) = -4\end{aligned}$$

The fact that the covariance is negative means that when x tends to be relatively high (greater than its mean) then y tends to be relatively low (smaller than its mean).

Finally, we can compute the correlation coefficient, which measures the degree of correlation as a number between -1 and 1:

$$\rho_{x,y} = \frac{\operatorname{cov}(x, y)}{\sigma_x \cdot \sigma_y} \approx \frac{-4}{3.46 \cdot 1.41} \approx -0.82$$

12.2 Correlation is not causation

If two variables are strongly correlated, we might think that there is a causal link between them; however, we should be very careful when we try to explain *why* the correlation occurs, because there could be many equally good explanations and we might not have enough information to determine the correct one.

Causal relation It could be the case, that there is a causal relation between the variables: This would mean that the changes observed in one of the variables is caused by changes in the other variable. But since the measure of correlation is symmetrical, we cannot say which variable is the cause and

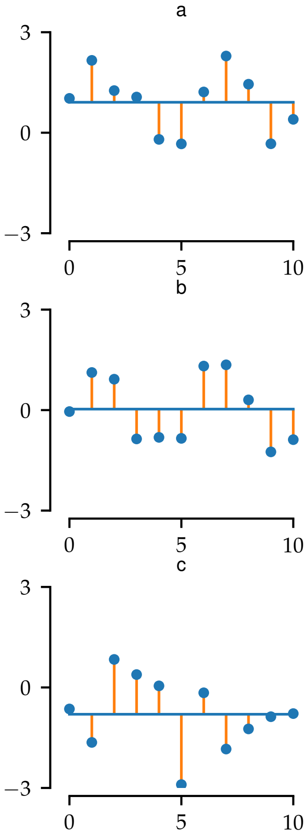

Figure 12.1: Examples of correlations. The plots shows the mean and 11 obervations of three different variables. The data in the two top panes are corelated: When one variable is above its mean the other also tends to be above its mean and vice versa. The bottom data is uncorrelated with the data in the top two panes.

which is the effect. To determine the direction of the causal link requires further information or assumptions.

Bidirectional relation The cause-effect relation might also go both ways, as for example in a predator-prey relation: As the number of predators increases, the number of prey will decrease (as they are eaten), which in turn leads to fewer predators (as they starve) and so on.

Confounding Another possibility could be that there is some third, unobserved variable, (known as a confounder) that is a common cause of both of our observed variables, and the observed correlation between our variables simply occurs because they are both influenced by this common cause.

Spurious correlation Finally, the observed correlation might simply be a random coincidence, a phenomenon known as *spurious* correlation.

Definition 12.3 Correlation and causation

Correlation is a statistical measure that describes the strenght of the linear relationship between two variables.

Causation means that one variable (the effect) is the result of another variable (the cause). If there is a causal relationship between two variables, and we were to manually modify the cause-variable, the effect-variable would be affected as well.

Example 12.5 Income earned and hours worked

Let us say that we have measured two variables "income earned" and "hours worked", and we have estimated a strong positive correlation (people who work more hours earn more income and vice versa). Is it likely that there is a direct causal effect? Well yes, we could imagine that if we increased the number of hours worked the income would increase as well, for example if we are talking about people working on hourly wages. So the "hours worked" could possibly have a causal effect on "income earned".

A causal effect the other way around is perhaps not so easy to imagine: If we increase their income, people would probably not start working more hours.

But before we make any conclusions we should think things through: Imagine a scenario where everybody have a fixed monthly wage and no overtime pay, and imagine that people with a high wage tend to put in more overtime without compensation. In that case "hours worked" and "income earned" would be correlated. But since no-one gets paid for their overtime, changing the number of hours worked will have no causal effect on the income.

The bottom line is that it requires strong assumptions about how the world works to go from an observed correlation to a causal explanation.

12.2.1 *Confounding variables*

In many situations the direction of the causal relation between two variable is fairly obvious, for example if the cause logically preceeds the effect in time or there is some other well established theory about their causal relation. In that case we often refer to the hypothesized cause as the "treatment" and the effect as the "outcome". But even in that situation we should be very cautious about using a measure of correlation to say something about the strength of the causal relationship between treatment and outcome. One reason is that there could be *confounding* variables that influence the measured correlation by affecting both the treatment and the outcome (see Figure 12.2).

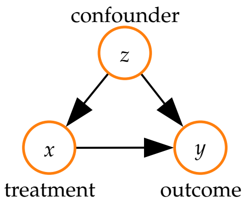

Figure 12.2: Causal diagram: The treatment is a direct cause of the outcome, but the confounder is a cause of both the treatment and the outcome.

Definition 12.4 Confounding variable

A *confounder* is a variable (often unobserved) that influence both the treatment and the outcome. In the presence of unobserved confounding variables, the observed correlation cannot be taken as evidence for the strength of a hypothesized causal relation.

When an unobserved confounder is present, the observed correlation between the treatment and the outcome could be completely different from the causal relation. It might even be the case that there is a positive observed correlation even if the underlying causal relation is negative.

Example 12.6 The best AI teacher

Let us say that a university course in AI runs two times per year. In the spring semester the course is taught by Teacher A, and in the fall it is taught by Teacher B. We now observe the number of students who pass the standardized exam in the two semesters:

- 95 out of 100 students (95%) pass the course with Teacher A in the spring.
- 72 out of 80 students (90%) pass the course with Teacher B in the fall.

We can think of the teacher as the treatment and the proportion of students who pass the exam as the outcome: It seems that Teacher A does a better job than Teacher B, and it would appear that taking the course with Teacher A more likely leads to passing the exam.

Now, assume we asked the teachers about their classes and they gave us the following further information.

- In the spring, 80 of the students were experienced graduate students 78 of whom passed the exam, and the remaining 20 were inexperienced freshmen 17 of whom passed.
- In the fall, 10 of the students were graduate students who all passed and 70 were freshmen of whom 62 passed.

If we examine how well the two types of students do, we might break down the numbers as follows:

	Graduates	Freshmen	Total
Teacher A	78/80≈98%	17/20=85%	95/100=95%
Teacher B	10/10=100%	62/70≈89%	72/80=90%

Now we can see that while Teacher A overall has the best passing rate, the passing rate for Teacher B is better *both for the graduates and the freshmen*.

In this example, the type of student is a confounder. The passing rate with Teacher A in the spring is higher in part because more experienced students take the course in the spring. While this more detailed analysis might suggest that Teacher B is best after all, there is no guarantee that there are not other unobserved confounders, which might lead us to reverse our conclusion again.

12.2.2 *Randomized trials*

If we want to interpret an observed correlation as a causal relation, we need to make sure that we take into account all possible confounding variables. In an observational study, where we gather data from some "real life" situation, there is no general way to find out which confounders that are present. If we did have access to measurements of all the confounders, we could in principle adjust for them to estimate the strength of the causal effect. One way to do this would be to divide our observations into subgroups in which the confounder is practically constant, and perform our analysis on each subgroup separately, as we did for the different types of students in Example 12.6. However, this approach relies strongly on the assumption that we have measured all relevant confounders.

Another approach, which completely sidesteps the issue with possible confounders is to conduct a *randomized trial*. The idea in a randomize trial is that the decision about who gets the treatment (or how large a treatment) is completely randomized, which by definition guarantees that there are no possible confounders. Since the decision about the treatment is completely specified by a randomized procedure, there can be no other hidden causes that affect the treatment, and thus there is no confounders. In a randomized trial, it is suddenly easy to estimate the strength of a causal relation, because any observed correlation immediately can be interpreted as causal.

Problems

1. What is the covariance and correlation coefficient between x and y in Example 12.2 concerning the dependent coin flips? Hint: Consider a sample that includes every possible outcome.

2. I have observed that my average speed (in my car) is 62 km/h when I use my summer tires, and 58 km/h when I use my winter tires. If the type of tire is the treatment and the average speed of the car is the outcome, what are possible confounders?

3. What do you think are the most probable explanations for the following observed relations

   (a) Dancing at parties is correlated with throwing up.

   (b) Snow is correlated with road accidents.

   (c) Cheese consumption is correlated with the risk of dying by becoming entangled in bedsheets.

   (d) Smoking is correlated with cancer.

   (e) Moderate alcohol consumption is correlated with increased life expectancy.

Solutions

1. $\operatorname{cov}(x, y) = -0.25$ and $\rho_{xy} = -1$.

2. One possible confounder could be the weather (winter tires are used in winter weather, which also might cause me to slow down.)

3. While there is not necessarily any correct answer here, possible explanations could be

   (a) A common cause could be alcohol consumption.

   (b) This is most likely direct causal relation.

   (c) This is probably a spurious correlation.

   (d) This is most surely a direct causal relation.

   (e) Who knows — the jury is still out on this one.