7.2: Modelling Continuous Relationships

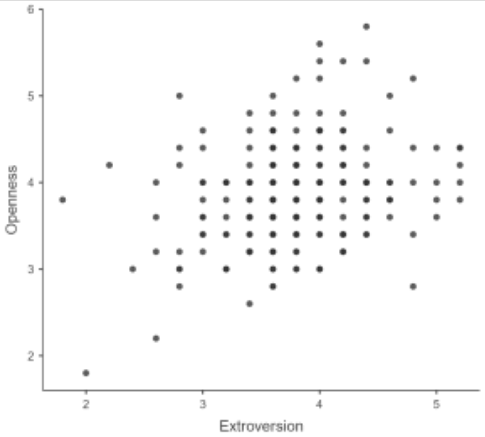

Figure 7.2.1. Scatterplot of extraversion and openness to experience

Figure 7.2.2 shows us the results:

Correlation Matrix

Correlation Matrix

		Extroversion	Openness
Extroversion	Pearson's r	—	
	p-value	—	
Openness	Pearson's r	0.28	—
	p-value	< .001	—

Figure 7.2.2. Correlation matrix for extraversion and openness to experience (screenshot from jamovi)

The correlation value of 0.28 between extroversion and openness to experience seems to indicate a reasonably moderate positive relationship between the two. The p-value above shows that the likelihood of an *r* value this extreme or more is quite low under the null hypothesis, so we would reject the null hypothesis of $r = 0$. Note that this test assumes that both variables are normally distributed.

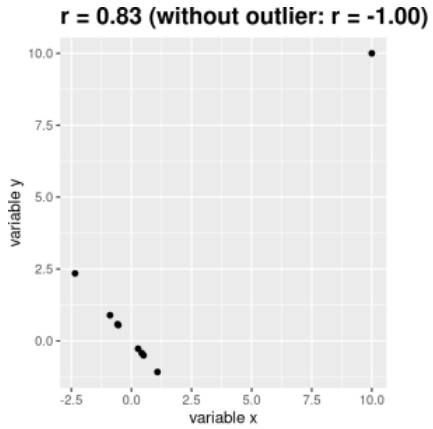

Figure 7.2.3. A simulated example of the effects of outliers on correlation. Without the outlier, the remainder of the data points have a perfect negative correlation, but the single outlier changes the correlation value to be strongly positive. Image by Poldrack, licensed under CC BY-NC 4.0

One way to address outliers is to compute the correlation on the ranks of the data after ordering them, rather than on the data themselves; this is known as the Spearman correlation. Whereas the Pearson correlation for the example above is 0.28, the Spearman correlation is 0.25, showing that the rank correlation reduces the effect of the outlier and reflects the negative relationship between the majority of the data points. Getting the Spearman correlation is really easy in jamovi, you just click this as an additional option for your results.

Correlation and Causation

When we say that one thing *causes* another, what do we mean? There is a long history in philosophy of discussion about the meaning of causality, but in statistics, one way that we commonly think of causation is in terms of experimental control. That is, if we think that factor X causes factor Y, then manipulating the value of X should also change the value of Y.

In medicine, there is a set of ideas known as *Koch's postulates* which have historically been used to determine whether a particular organism causes a disease. The basic idea is that the organism should be present in people with the disease, and not present in those without it – thus, a treatment that eliminates the organism should also eliminate the disease. Further, infecting someone with the organism should cause them to contract the disease. An example of this was seen in the work of Dr. Barry Marshall, who had a hypothesis that stomach ulcers were caused by a bacterium (*Helicobacter pylori*). To demonstrate this, he infected himself with the bacterium, and soon thereafter developed severe inflammation in his stomach. He then treated himself with an antibiotic, and his stomach soon recovered. He later won the Nobel Prize in Medicine for this work.

Often we would like to test causal hypotheses but we can't actually do an experiment, either because it's impossible ("What is the relationship between human carbon emissions and the earth's climate?") or unethical ("What are the effects of severe neglect on child brain development?"). However, we can still collect data that might be relevant to those questions. For example, we can potentially collect data from children who have been neglected as well as those who have not, and we can then ask whether their brain development differs.

Let's say that we did such an analysis, and found that neglected children had poorer brain development than non-neglected children. Would this demonstrate that neglect *causes* poorer brain development? No. Whenever we observe a statistical association between two variables, it is certainly possible that one of those two variables causes the other. However, it is also possible that both of the variables are being influenced by a third variable; in this example, it could be that child neglect is associated with family stress, which could also cause poorer brain development through less intellectual engagement, food stress, or many other possible avenues. The point is that a correlation between two variables generally tells us that something is *probably* causing something else, but it doesn't tell us what is causing what.

Figure 7.2.4 shows the causal relationships between study time and two variables that we think should be affected by it: exam grades and exam finishing times.

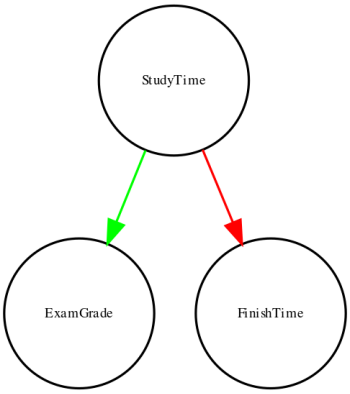

Figure 7.2.4. A graph showing causal relationships between three variables: study time, exam grades, and exam finishing time. A green arrow represents a positive relationship (i.e. more study time causes exam grades to increase), and a red arrow represents a negative relationship (i.e. more study time causes faster completion of the exam). Image by Poldrack, licensed under CC BY-NC 4.0

However, in reality, the effects on finishing time and grades are not due directly to the amount of time spent studying, but rather to the amount of knowledge that the student gains by studying. We would usually say that knowledge is a *latent* variable – that is, we can't measure it directly but we can see it reflected in variables that we can measure (like grades and finishing times). Figure 7.2.5 shows this:

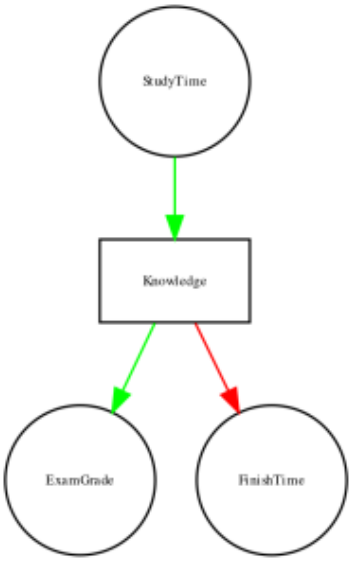

Figure 7.2.5: A graph showing the same causal relationships as above, but now also showing the latent variable (knowledge) using a square box. Image by Poldrack, licensed under CC BY-NC 4.0

$$y = x * \beta_x \beta_0 \epsilon$$

The βx"> β_x value tells us how much we would expect y to change given a one-unit change in x. The intercept β_0 is an overall offset, which tells us what value we would expect y to have when $x = 0$; you may remember from our early modelling discussion that this is important to model the overall magnitude of the data, even if xX"> never actually attains a value of zero. The error term ϵ∈"> refers to whatever is left over once the model has been fit; we often refer to these as the *residuals* from the model. If we want to know how to predict y (which we call $\hat{y}$) after we estimate the beta values, then we can drop the error term:

$$\hat{y} = x * \hat{\beta}_x + \hat{\beta}_0$$

Figure 7.2.6. The linear regression solution for extroversion and openness to experience is shown in the solid line

The value of the intercept is equivalent to the predicted value of the y variable when the x variable is equal to zero. The value of beta is equal to the slope of the line – that is, how much y changes for a unit change in x. This is shown schematically in the dashed lines, which show the degree of increase in openness to experience for a single unit increase in extroversion.

The relation between correlation and regression

There is a close relationship between correlation coefficients and regression coefficients. Remember that Pearson's correlation coefficient is computed as the ratio of the covariance and the product of the standard deviations of x and y:

$$\hat{r} = \frac{covariance_{xy}}{s_x * s_y}$$

whereas the regression beta for x is computed as:

$$\hat{\beta}_x = \frac{covariance_{xy}}{s_x * s_x}$$

Based on these two equations, we can derive the relationship between $\hat{r}$and latex]\hat{\beta}[/latex]:

$$covariance_{xy} = \hat{r} * s_x * s_y$$

$$\hat{\beta}_x = \frac{\hat{r} * s_x * s_y}{s_x * s_x} = r * \frac{s_y}{s_x}$$

That is, the regression slope is equal to the correlation value multiplied by the ratio of standard deviations of y and x. One thing this tells us is that when the standard deviations of x and y are the same (e.g. when the data have been converted to Z scores), then the correlation estimate is equal to the regression slope estimate.

Regression to the Mean

The concept of *regression to the mean* was one of Galton's essential contributions to science, and it remains a critical point to understand when we interpret the results of experimental data analyses. Let's say that we want to study the effects of a reading intervention on the performance of poor readers. To test our hypothesis, we might go into a school and recruit those individuals in the bottom 25% of the distribution on some reading test, administer the intervention, and then examine their performance on the test after the intervention. Let's say that the intervention actually has no effect, such that reading scores for each individual are simply independent samples from a normal distribution. Results from a computer simulation of this hypothetical experiment are presented in Table 7.2.1.

Table 7.2.1. Reading scores for Test 1 (which is lower, because it was the basis for selecting the students) and Test 2 (which is higher because it was not related to Test 1).

test	score
Test 1	88
Test 2	101

If we look at the difference between the mean test performance at the first and second test, it appears that the intervention has helped these students substantially, as their scores have gone up by more than ten points on the test! However, we know that in fact the students didn't improve at all, since in both cases the scores were simply selected from a random normal distribution. What has happened is that some students scored badly on the first test simply due to random chance. If we select just those subjects on the basis of their first test scores, they are guaranteed to move back towards the mean of the entire group on the second test, even if there is no effect of training. This is the reason that we always need an untreated *control group* in order to interpret any changes in performance due to an intervention; otherwise, we are likely to be tricked by regression to the mean. In addition, the participants need to be randomly assigned to the control or treatment group, so that there won't be any systematic differences between the groups (on average).

Chapter attribution

This chapter contains material taken and adapted from *Statistical thinking for the 21st Century* by Russell A. Poldrack, used under a CC BY-NC 4.0 licence.

Screenshots from the jamovi program. *The jamovi project* (V 2.2.5) is used under the AGPL3 licence.