Correlation and Regression: Regression

Before you Watch

This video describes the concept of comparing two numerical variables. This builds on the video Random Variables, where variable type is described.There you can revisit how to identify whether a variable is numerical. A numerical variable is a variable where the possible responses are numbers which reflect the quantity, or amount, of something. Take for example height, in centimeters, or weight, in kilograms. This also builds on the Correlation Video.

The Video Content

This video explains how to determine the equation of the line of best fit that can be constructed for the relationship between two numerical variables.

Whilst two numerical variables may be related, as discussed in the video on Correlation, when the response to one variable (known as the independent or predictor variable) causes a response in another variable, known as the dependent or outcome variable we say we have a causal relationship and we may wish to quantify the relationship so that if you are given the value for the predictor you can determine the value of the outcome variable, on average.

Is the length of stay in hospital related to the age of the patient? Do older people, for example, spend longer in hospital, on average compared to younger patients?

We may wish to be able to predict length of stay based on a patient's age.

Or, as described in the earlier video on Correlation we may consider the relationship between Weight (in kg) and Cholesterol, and determining average cholesterol based on weight. Regression can assist with this task.

Here we are interested in predicting cholesterol based on weight so weight is the Independent variable and cholesterol is the Dependent variable. We can link these two variables together using an equation. You tell me a person's weight and I'll tell

you the average cholesterol.

Scatterplots enable a visual assessment of relationships between two numerical variables. They are a graphical display with a horizontal line (known as the x-axis) and a vertical line (known as the y-axis),

The independent variable, Weight, is displayed on the horizontal or x axis, and the dependent variable, Cholesterol, is displayed on the vertical or y axis. Plotting the data we have on 20 patients we see a positive association.

We could draw a line of best fit through the points like this. We could use this line to estimate cholesterol for a given weight. For example, if a person weighs 80kg then this line would estimate the cholesterol to be 217. The line of best fit has an equation which connects Weight and Cholesterol.

Using the observed data, the line of best fit through the data is obtained from the equation of the line that essentially minimises the distances of each point from the line of best fit (referred to as residuals) - HERE ARE TWO EXAMPLES OF THE DISTANCES, or residuals, WE WANT TO BE AS SMALL AS POSSIBLE. Most statistical computing packages will calculate the equation of this line of best fit for you.

Finding the equation of the regression line here is referred to as simple linear regression. Simple since there is only one predictor. Linear because the relationship can be well-represented by a straight line.

In this case we obtain the regression equation

Cholesterol = -111 + 4 .1 Weight

So, if you tell me a person’s weight is 80 kg. Then

Cholesterol = 217

Let’s assess the equation in more detail. First, we really should state it as Expected cholesterol equals -111 plus 4.1 lots of weight. Since we are estimating the cholesterol result (on average) for a given weight.

In the equation, the value 4.1 has an interpretation. Many of you may have heard of the gradient or slope of a line. Well this number also has an interpretation in this context.

If Weight is 80 kg, then Expected cholesterol is 217. If Weight is 81 kg, then Expected cholesterol is 221.1.

We increased weight by 1kg, by how much did the expected cholesterol change? It changed by 4.1. This value !

The coefficient of Weight, 4.1 is the slope of the line.

This means that as weight increases by 1 kg, cholesterol is expected to increase by 4.1 units, on average. This can be very useful information in the clinical setting when providing patients with advice on lowering cholesterol. Another example of how such an equation could be useful in the clinical setting is an anaesthetist who bases dosage on a person's weight.

The other component of the equation is the y-intercept of the line. This is the constant in the equation. Here it is -111. This is known as the y intercept. If weight was 0, so the y-intercept is the estimated value when the predictor is 0. Of course it doesn't always make sense to consider this value. For example in this situation, we cannot have someone weighing 0 kg.

These models can be extended to include several predictors and would be called multiple linear regression. Such regression models apply to where the relationship between variables is linear, for example where each predictor contributes as it is, for example weight not for example by having to take the square of weight.

If the relationship between variables is nonlinear, like in this graph, then other nonlinear regressions can be performed.

Now What?

This video introduced how to determine the line of best fit to describe the relationship between two numerical variables. You can look to further develop your understanding of this concept by looking to the Other Links listed below. Alternatively, ensure that you are also confident with the material covered in Association Between Two Continuous Variables: Correlation as this video is closely related to this one.

But, when am I going to use this?

Statistics is essential in the study of systems of situations where there is an unpredictable random element. This includes a huge number of situations, such as any system that involves living things, which always have a degree of unpredictability. In fact, any studies that involve people involve statistics: for example, medical processes, education and economics. Other areas statistics can be applied to include quality control, stars, nature, or how wear and tear affects machinery.

Other Links

A YouTube video made by the Statistics Learning Centre in NZ gives an overview of data types using animation that is easy to understand and fun. Note that numerical data types has been given alternative names. Interval and ratio data are used instead of discrete and continuous. It is a slightly different approach but the same principle.

https://www.youtube.com/watch?v=hZxnzfnt5v8

If you really want to understand types of data, along with appropriate statistics and graphs, you can learn on our new Snack-size course. Takes about an hour, and lots of fun!

http://www.statslc.com/snack/