THE UNIVERSITY OF AUCKLAND
NEW ZEALAND
Te Whare Wānanga o Tāmaki Makaurau

WEEK 3

3.8 TREND, SCATTER AND OUTLIERS by Chris Wild

In the last video, we used scatterplots to display relationships between two numeric variables. We introduced interpreting scatterplots in terms of trend and scatter. We also saw examples with obvious clusters of points, triggering the question, "Why are these sets of individuals so different?"

But usually, it is most useful to view scatterplots in terms of trend plus scatter with the occasional outlier. We'll start by looking at a range of trend plus scatter scatterplot behaviours. The focus is simply on experiencing some of the sorts of patterns that occur-- trend and scatter.

Trend and scatter.

Trend and scatter and two outliers.

Trend, scatter and one outlier.

We had a quick look at making predictions using a scatterplot in the last video, but now we're going to go more deeply into some of the issues involved. When should a doctor start worrying that a liver length is abnormal for a foetus of a given age?

These are the gestational ages and liver lengths of 258 normal human foetuses, unborn babies, determined by ultrasound. Our outcome variable is the liver length and the predictor variable is the gestational age of the foetus, which is an estimate of the time since conception.

We want to use the graph to predict the range of liver lengths we'd expect for a healthy foetus from its age and the population this data was drawn from. We'll do this for two gestational ages, 18 weeks and 35 weeks.

So at 18 weeks, what liver length would we expect?

I think you'll agree it makes a lot of sense to put on a trend curve, and then do this. It suggests the liver length should be about 21 millimetres. And doing the same thing at 35 weeks, suggests a liver length of about 49 millimetres.

But hang on a minute. When we look at the scatter, we see that there is a reasonable amount of variation and liver lengths of healthy 18-week olds and a lot of variation in the liver lengths of normal 35-week olds. We clearly need to take that into account.

Well, let's draw an envelope that includes almost all of the scatter.

Now let's use that in the obvious way.

This suggests that at 18 weeks, the liver length of a healthy foetus would almost always be between 16 and 26 millimetres. At 35 weeks, we'd expect it to be between 35 and 62 millimetres. Beyond that range, the doctor might start worrying about the foetus's development.

It's the trend curve and the scatter around that trend that has allowed us to do all of this.

This simple example conveys many of the biggest ideas about the prediction problem, even though we've informally drawn the lines by eye. We can think of the trend here as a sort of summary of the strongest pattern in the data.

This picture shows that where there is a lot of scatter about the trend, we need to predict our outcome using a wide bracket of values. Where the scatter is small, our range of predicted values will be much narrower and probably more useful.

Predictions from data can only work well if the data you have is representative of the way things behave in the setting in which you want to make the prediction. Data collected from major London hospitals is unlikely to work well for predicting foetus development in the poorer regions of India, for example, because foetuses there will be much smaller.

In a similar vein, predictions that use data from the past to predict the future can only work well if the historical relationships between variables or patterns still hold. Major structural changes to the economic environment, for example, can make the way things behaved in the past, largely uninformative, about the way they will behave in the future.

Here's one last example on the prediction theme. This data came from an experiment in which a drop of urine is dropped on a Petri dish or plate. And sometime later, the number of visible bacterial colonies growing on the plates was counted. This was done several times at various concentrations.

How many colonies would you get at a concentration of 12?

Well, that's pretty obvious. It should be about 40. But when they actually experimented out near 12, this is what they found.

Whoops. At high concentrations the urine becomes toxic and, rather than continuing to rise, the numbers of colonies growing actually dropped right off. This underlies the dangers of predicting what is going to happen in a region where you have no data.

Take a look at these two plots. They are identical except for the superimposed trend line. Which trend line do you think best summarises what you see? Most people instinctively pick the left hand picture. Remember, however, that the purpose of the trend we draw is to predict the outcome y value from the predictor x value.

We've added some vertical dotted lines. You'll see that if you use the steeper line to predict the points, then when x is small, your prediction is almost always too small. And when x gets large, your prediction is almost always too big. The flatter line, on the other hand, is a much better predictor as it is always going through the middle of the points it's trying to predict.

When you're looking at a trend curve on a scatterplot, imagine a set of vertical lines like this, and check visually whether the curve is going straight through the middle of the points in a column. The trend summarises the main pattern we see in the data. This is a notion of averaging going on. The height of the trend curve at a predictor value of 5, say, should be telling me about the average outcome value when the predictor is at or near 5. Finally, I'll leave you with these questions to remind you of the ideas we've just covered.