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

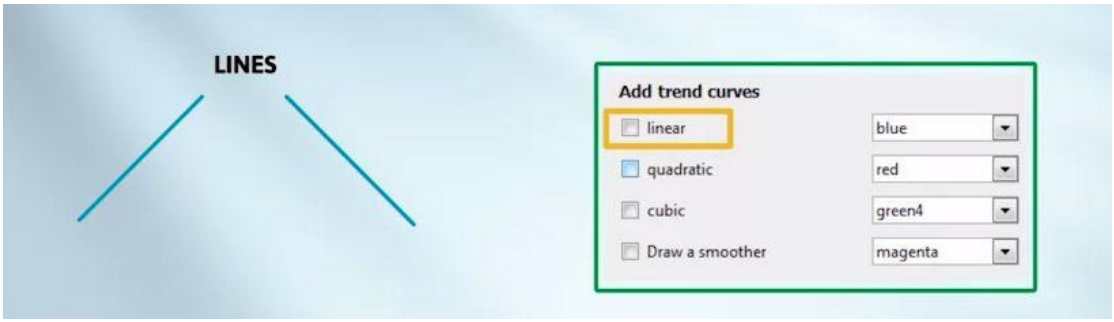
WEEK 4

4.2 LINES, CURVES AND SMOOTHERS by Chris Wild

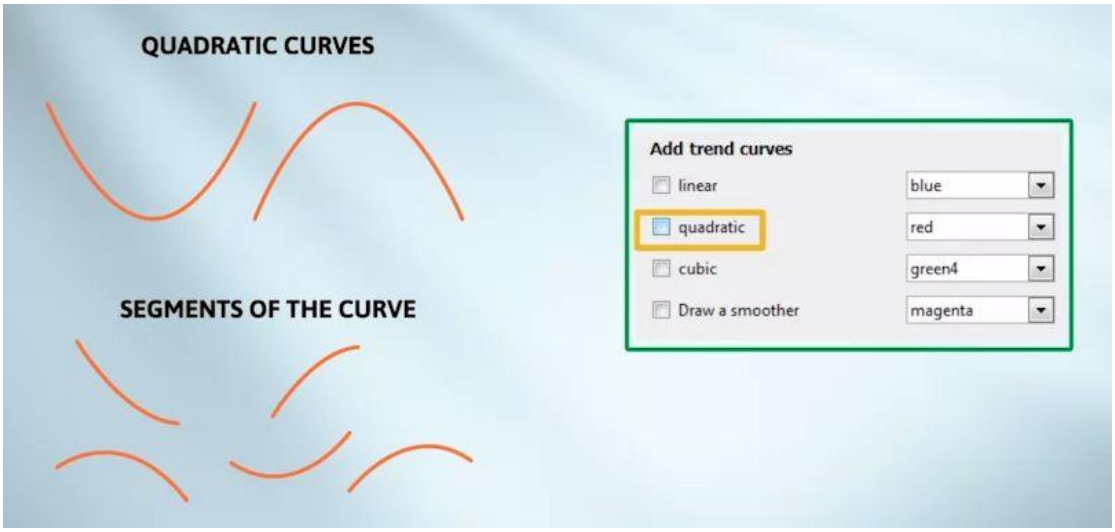
Last week, we introduced scatterplots for visualising relationships between two numeric variables. We added trend lines or curves to the plot as a summary of the main pattern we saw in the data.

But so far, all our curves have been freehand curves, drawn by eye to make us really engage with seeing trends. This is important as the untrained eye can be misled. Of course, in real data analysis, trends are put on to scatterplots by software. But we can't just trust software. We need to be able to look critically and decide whether the software has done the right thing.

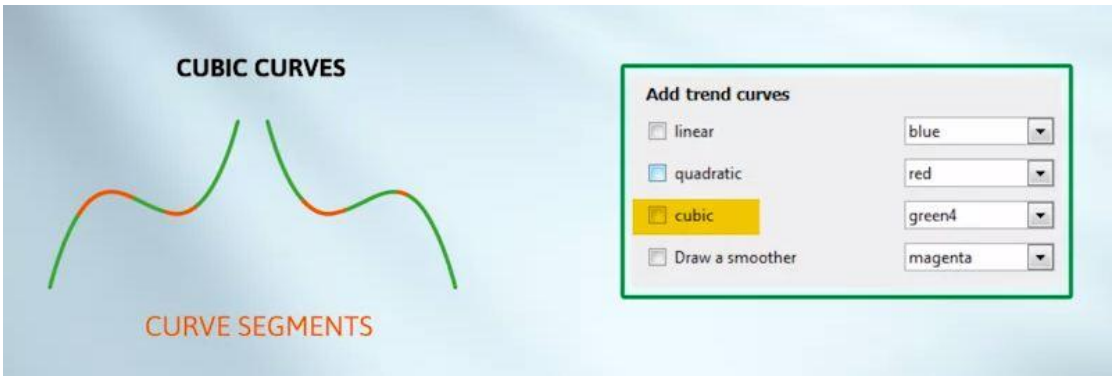
iNZight, for example, gives these choices.



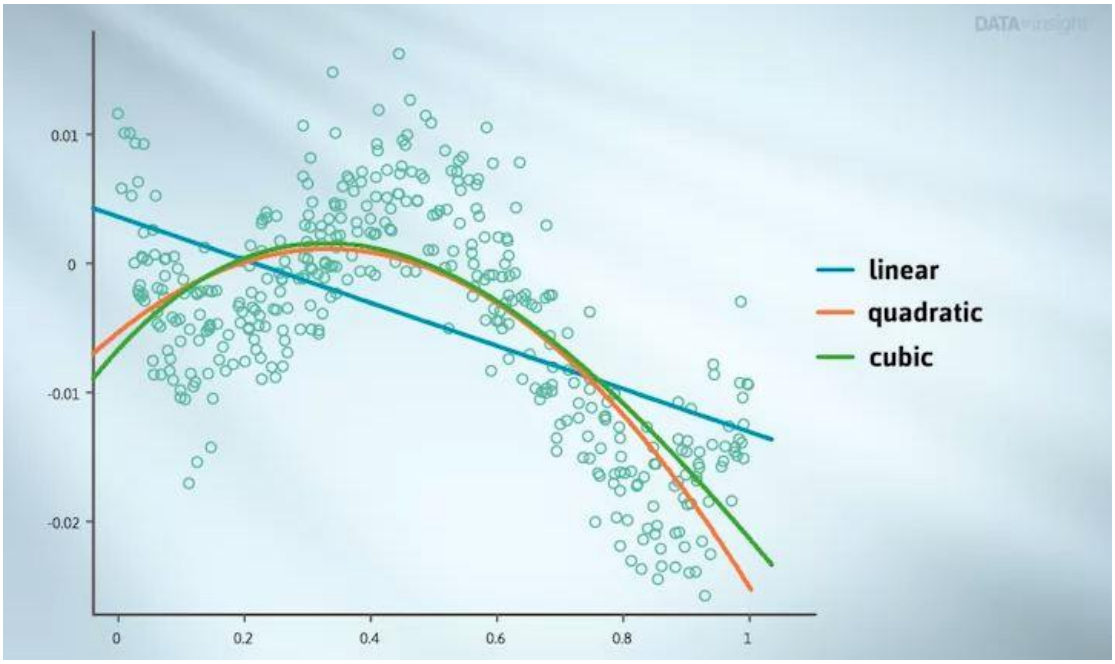
Linear is used to capture a trend that looks like a straight line.



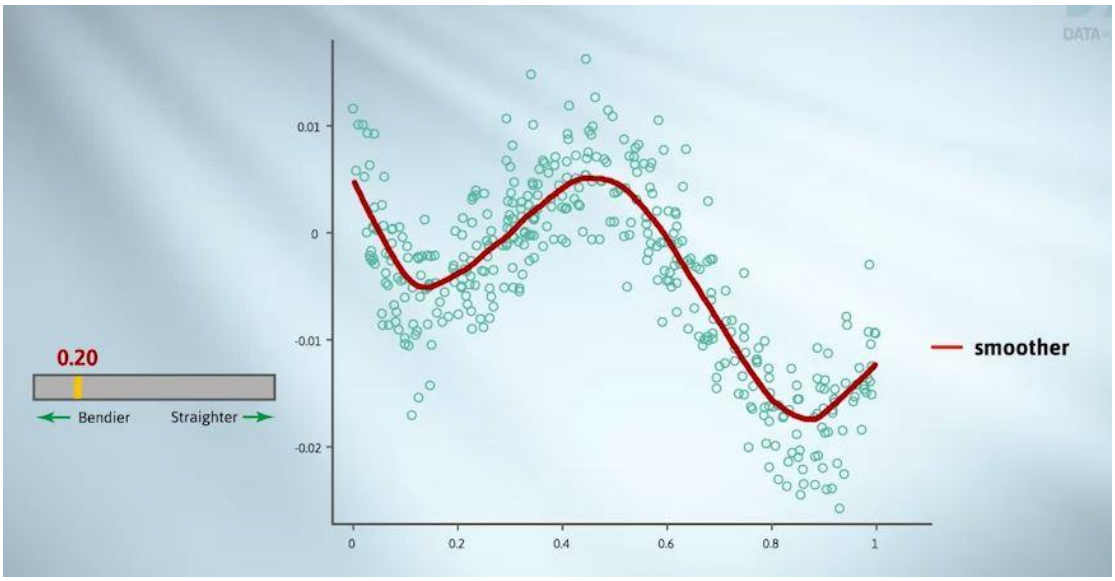
Quadratic curves could be used to capture trend-curve shapes that look like any segment of one of the two curves on the left.



Cubic curves are more flexible because they can take up to two bends. Any segment of these two curves may be chosen by the software.



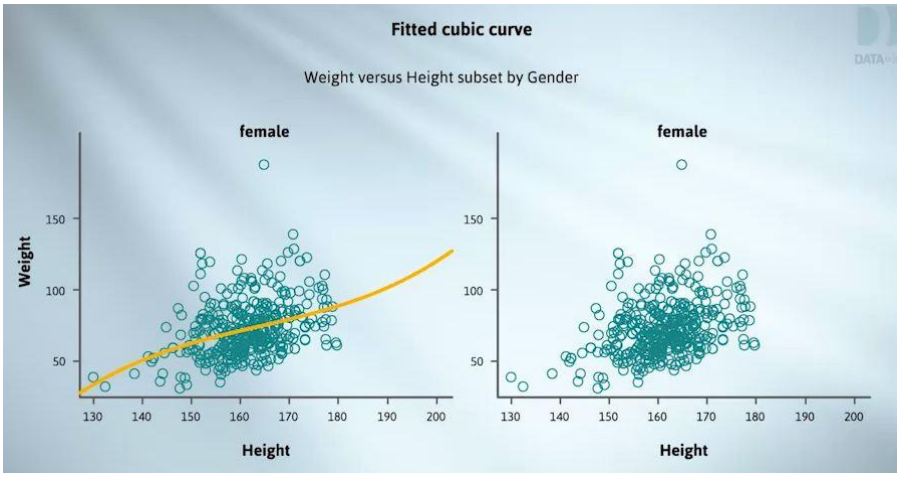
When you ask for a particular form of curve, the software will draw the curve of that type that best predicts the data. We'll talk more about that soon. But the sets of shapes we've seen is not sufficient to capture all of the trends we see in data. An alternative is to put on a smoother.



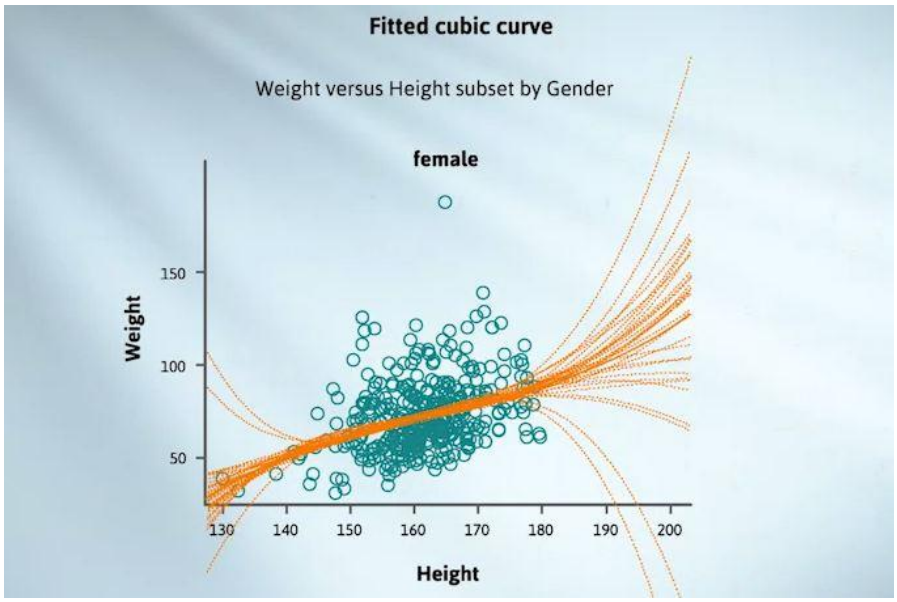
Best of the smoother curves

Smoothers are even more flexible and take on an even greater variety of shapes. You can control how flexible a smoother is using a slider. Whoops, gone too far. It's become too wiggly. We need to find a balance. That looks pretty good. Getting worse, better. That's definitely the best I've seen. I'll stop there.

In my own data exploration, I usually use lines when they look like they'll do a good job and smoothers when the trend looks non-linear. With curve-fitting methods, we should not take the behaviour at the right and left-hand edges of the plot too seriously because it's often determined by precise positions of just a few data points.



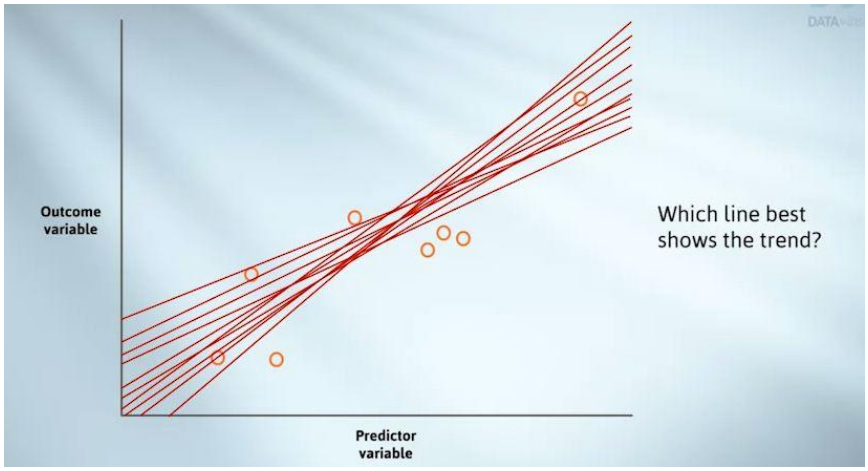
The left-hand plot has a smooth to summarise the trend. The right-hand plot is what you get when you click the "Inference Information" button in iNZight.



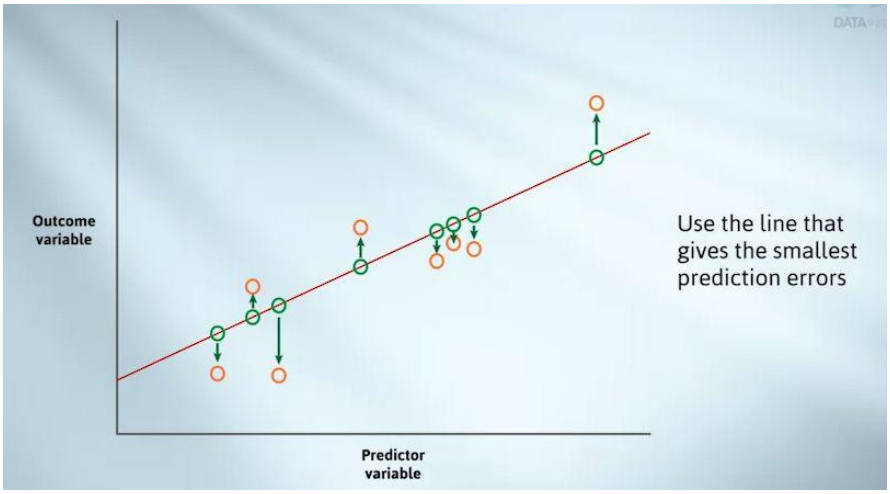
A set of curves like this is often described as a nest of curves. We'll find out where they came from later in the course.

The trend is most reliably captured where the curves in the nest are close together, and least reliably captured where they're spread far apart, as they are here towards the edges. But that's getting ahead of ourselves.

We'll now address the question of how the computer decides precisely what line or curve to put on a scatterplot. We'll explain this in the context of choosing the best line. Here's a small data set. Which of all these possible lines best captures the trend?



First, we'll have to decide what we mean by best. Let's draw a candidate line.



Candidate line

For every data point, there's a corresponding point on the line. This is the prediction this line would make for that data point. The predicted point and the real observation seldom coincide, leading to a set of prediction errors. These prediction errors are the vertical arrows on the plot.

We want to choose the line that makes the smallest prediction errors in some overall average sense. Most software uses the least squares method, which minimises the sum of the squares of the prediction errors. The same idea works for curves. So if we ask for a quadratic curve, the software will draw the quadratic curve which makes the smallest average prediction errors.
This concludes this video.