Video: “Linear Models” (6:45 min)

Linear Models (0:08):

This short video introduces the basic ideas of linear modeling. We often observe a linear relationship between two variables, but measurements have noise. It’s unlikely that the points we measure will fall exactly on a straight line. We want to find the best line that is we want to optimize the fit. Once we have such a line or, linear model, we can use it to make predictions of new points and to assess the quality of our previous data. Let’s begin by introducing the basic ideas of linear modeling. Here we plotted some points, x y, in the plane. The x values are the independent variables and the y values are the dependent variables. We are looking for a model y = ax + b. It’s very unlikely that these points are going to fall exactly on a line. As you see here they are scattered and there are many possible lines that we could draw. Regardless of which line we draw there are going to be some errors. In fact, each point (xi, yi) satisfies the equation yi = axi + b + ei. The ei is the error specific to that point. The goal is to find an a and b, the same a and b for all the points that makes the overall error as small as possible. Now that’s a little vague, we have to be more specific about what we mean about overall error.

Finding the Best Fit (1:30):

Let's start with an example to make things more specific. Here we plotted to points in the same plane. There is one and only one line that goes through them, no problem here. The equation of that line is y = -0.5x + 2.5, so far so good. Now we add a third point, and now we have a problem. There is no one line that will fit all 3 points, so we will have to compromise. Let’s figure out how much error is here using this line to model the points. The original points are no problem, they are right on the line, so they have non error. The new point (2,0) is not on the line in fact it is quite far away. We’ll figure out the error. The error is the actual value of y in this case 0 minus the predicted value of y. We’ll have to plug into the equation to get that. To calculate the y predicted we evaluate at x = 2 and find that -0.5*2 + 2.5 is 1.5. The equation predicts 1.5 but the actual value is 0, that is an error of -1.5. Let’s summarize are results so far in a table. We have 3 points with x values of 1, 2, and 3 respectively. The corresponding y values are 2, 0, and 1 and when we do the predictions, we find that the first and the third point agree. However, the error for the second point is -1.5.If we just average the error at this point we find that + and - values cancelled out and we wouldn’t get a good estimate of error. Instead we will square the error and average that, this is called the mean squared error. In our case it turns out to be 0.75.

Can We Do Better? (3:25):

At this point you're probably wondering if this model is the best we can do and you probably suspect that a compromise line between the 3 points might be better, so let’s take a look at an example. Let’s try the line y = -0.5x + 2. This line has a y-intercept of 2, while the original had a y-intercept of 2.5. It’s more of a compromise line. When we evaluate at x = 1, we find the model predicts y will be 1.5, but it’s actually 2 so we make an error of 0.5 by using the model. Similarly, for x = 2 we find that the prediction error is -1 and for x = 3 we find an error of 0.5. We can calculate the mean squared error and it turns out to be 0.5. Recall that the mean squared error for the original line was 0.75, so the new model is better. In fact, it’s the best

model. It minimizes the mean squared error. Fortunately we don't have to try all line to find the best one, there is a formula for it and its often derived from statistic courses and calculus courses. We will just take their word for it. One additional comment, the mean squared error which its name implies is the square of values. The units of mean squared error is the square of the units of the original y. So if y was in millimeters, mean squared error is in millimeters squared. Usually, when we report errors we take the square root of the mean squared error which has its own special name, RMS for root mean squared error and it has the same units as y. In this case the RMS of our best fit line is 0.71 while the original line had an RMS of 0.87.

Measure of Quality of Fit (5:27):

Let's summarize, the RMS or root mean squared error the average amount of error made in using the model to predict the values. It's in the same units as the y value and it's calculated by taking the square root of the mean squared error. The mean squared error as its name implies is just the average of the squared errors. An R squared value will often be reported with a linear fit to indicate quality. R squared is the coefficient of determination, it's the amount of variance accounted for by a model. For linear model, R squared is just the square of the correlation, we can easily compute it. At this point you may be wondering which is better RMS or R squared. You should really compute them both because they tell you different things. RMS averages the average error and using the model to predict your data. It's in units of your data and the value should be small compared to your typical data points. On the other hand R squared is always a value between 0 and 1. It gives you a measure of the overall quality of the fit. R squared is a value that people are familiar with and it will allow them to evaluate the quality of your model and make comparisons across other models.

In MATLAB (6:50)

In MATLAB we can find linear models using the polyfit function. The return value of polyfit is a vector with the coefficients of the model. Here a is p(1) and b is p(2). Once we have are model, we can get the predicted values of y using the polyval function. The error is just the difference between y and their predictions. We can easily calculate the mean square error, the root mean squared error and R squared which is the correlation squared. Creating models, linear models in particular, is a powerful tool that's used throughout science, engineering, statistics, and many other fields. Linear models allow you to do predictions, assess errors, and even provide motivation for formulating new physical laws. It's a tool well worth knowing.