**Week 1 - Regression**

0:00  
In this video I'm going to go through regression, but first I'm going to just make a quick note on independent and dependent variables and the difference between them.

0:11  
So an example is the value of a house and how that changes over time.

0:15  
So you've got 2 variables in this case, the value of the house and the other variable being time.

0:20  
So I'm calling value of the house V at time T Since they both change both of these things, they're both variables.

0:27  
So value is a dependent variable since the value of a house depends on the time you bought it, what the conditions were at the time.

0:35  
Time though is independent in that time doesn't depend on the value of the house.

0:39  
Time is just changing all the time, it's just changing with the universe.

0:45  
So time is independent of value, but value then is dependent on time.

0:50  
So when we plush, we typically leave the horizontal axis to represent the independent variable.

0:55  
So often you see time variable T along the X axis and the vertical axis to represent the dependent variable.

1:01  
You know, value of the house, share price, ECG, voltage value over time or whatever it is.

1:09  
Also, I'm going to have a quick look at what a linear relationship is.

1:12  
And I mentioned just in one of the last videos in this course, we're mostly looking at linear relationships to make our predictions.

1:18  
The reason being that computers are good at doing linear algebra.

1:22  
They do it really, really quickly.

1:24  
A linear relationship.

1:25  
If something is linearly, linearly related to something else, a change in the value of X always produces the same proportionate.

1:33  
And this is important to the word proportionate change in the value of Y.

1:37  
So considering the following table, what is the value of Y when X is 3?

1:41  
So if we look, we've got 1528.

1:45  
What is the value of this right here?

1:48  
What value of X will give you or what will, what value of X be when Y is 44?

1:56  
So we could actually work this out right now, and I'm sure you may remember again, it was junior St maths when I did this.

2:02  
The equation of the line Y is equal to MX plus C.

2:05  
You plug in some values you've and you try and get M versus slope.

2:08  
If I remember, slope is the rise over the run.

2:10  
So Y 2 -, y one over X2 minus X1.

2:13  
Again, we could take any 2 values here, calculate the slope, and from that, once you have the slope, then you can calculate the C, the offset.

2:21  
So once we are convinced that some relationship does exist, we can establish the precise nature of that relationship and use it to predict values of one variable that will correspond to some other variable in the world.

2:36  
So we can see here we've got a really linear relationship between this variable on X and what's varying in Y, the variable on Y.

2:44  
The exactness of the relationship can be seen in the diagram.

2:47  
So here we've got a really precise linear relationship.

2:50  
The plotted points lie exactly along this imaginary straight line.

2:55  
If we were to draw a straight line through each of them, we could make predictions for values that aren't in the table.

2:59  
So you know, what is the value of Y when X is equal to 10 and so on.

3:06  
Like I said, this all boils down to in this very simple case with two variables and a linear relationship.

3:12  
There's a formula for this previous graph and it's a very famous Formula, One that's taught really early.

3:18  
Y is equal to MX plus C, and in this case, Y in the table that I showed you previously is equal to three X + 2.

3:23  
So if you look at this, so 3X is 3 + 2 will give us 53X is 9 + 2 will give us 11.

3:32  
And if we have a look at it here, 9 there is going to give us, sorry, it's a different graph.

3:36  
Mine will give us 11.

3:37  
And similarly we can reverse it backwards as well.

3:42  
This very early taut equation of a line is the basis for regression and lots of machine learning analysis.

3:48  
And it can be extended then into more than just two dimensions where instead of a line you have a plane.

3:53  
But anyways, we're just going to look at the simple case for now.

3:56  
Regression is where we try to zone in on continuous values rather than trying to predict a class or a category.

4:02  
So when you're predicting a class or a category, be it, you know, dog or cash, there's nothing really in between.

4:07  
It's just one or the other that you're trying to predict to.

4:09  
We're trying to bunch all the cats together and all the dogs together with continuous values is we're trying to like make predictions about this variable.

4:18  
An example that I'm going to go through is how to cook a Turkey.

4:22  
So there's a linear rule of thumb when trying to cook a Turkey and it's put it in a preheated oven for 20 minutes.

4:28  
So that's our kind of C or sorry, 20 minutes per pound.

4:31  
So that's MX or XS are a variable.

4:34  
Weight M is here plus 20 minutes.

4:37  
That's our constant.

4:39  
So where did this idea come from?

4:41  
Who came up with this linear relationship at the beginning?

4:44  
Is it correct?

4:45  
And also what could happen if what was wrong?

4:49  
So I guess you could get food poisoning.

4:50  
But anyway, so examples.

4:53  
So following on with this example, this is a linear model where with the time is equal to the M, the weight plus C It's a linear aggression analysis of different weights of turkeys.

5:08  
So somebody probably sat down, cooked loads of turkeys of different weights, checked when they were all cooked with an expert or something, drew a line and then came up with to see if, checked if the line was linear, and then went back and devised this equation.

5:23  
Where T is the cooking time, W is the weight, M is the slope, and C is the Y intercept.

5:28  
And we use 20 minutes for the slope and 20 minutes for the Y intercept when that is when the weight is measured in pounds.

5:34  
However, this guy Kanofsky disagreed and threw his own modelling.

5:41  
Use the following model.

5:42  
So T is equal to the waste to the power of 2/3 / 1 five.

5:48  
So let's have a look at this first yeah, first thing I'll say is that I've been with people while they're trying to cook turkeys at Christmas and different times and this equation can be tough enough for people to get their heads around working out on a calculator.

6:03  
And this then will be a different level.

6:05  
But we can look at the difference between the two.

6:07  
The first thing with this linear relationship, even if you had a Turkey, an imaginary Turkey, or a Turkey of like close to 0, like a tiny Turkey that was only like a few grammes, you'd still have to cook it for 20 minutes.

6:18  
So this is the Y intercept between 1 LB.

6:20  
This is Panofsky's line.

6:23  
The models are very similar.

6:24  
And then when you get up to a really big Turkey of 16 lbs, you have a difference of what is it there?

6:32  
Like that's 90 minutes difference.

6:34  
So Panoski was a physicist, I think he's a mathematician.

6:37  
He spent quite a bit of time doing other stuff as well, but also models.

6:41  
The question is which is safer, You know, which would taste better.

6:45  
I'm by the way, not recommending how you cook turkeys this Christmas or anything like that.

6:49  
Just definitely whatever way you do is just use a meat thermometer.

6:53  
Don't want to be anyone getting food poison or anthem based on this advice.

6:57  
But with the above example, we see the result of a linear model, but not how it was arrived at.

7:02  
So we can see that someone created this model of the 20 minutes plus 20 minutes per pound.

7:10  
Someone came up with that, but we don't know how that happened.

7:13  
So to do it, what you'd have to do if you wanted to come up with that linear relationship or you were the one that invented it, you'd have to collect some data and fit the best line to it.

7:22  
In the above case, this would mean cooking.

7:24  
If you were to try and develop this linear relationship or any relationship, you'd have to cook lots of turkeys of different weights.

7:31  
Cook them all using exactly the same method and then have to measure the outcome, you know, to measure when they're cooked.

7:37  
So first you figure out when do you define your Turkey as cooked, maybe using a meat thermometer or some sort of expert cook who could tell that's cooked Turkey by going through it.

7:46  
You collect all the right answers, as in what's the right time to cook a particular Turkey?

7:50  
And you plot the X, which is the weight versus the Y, the time.

7:54  
And you have a look at the data.

7:55  
You know, you see, OK, is it actually linear uses to build a model, figure out the parameters.

8:01  
So say if you had this data again and you had cooked all these turkeys and you, you know, turkeys of 124, six, 8 lbs.

8:08  
And yet all the times it looked like the blue line, you could see it's a linear relationship.

8:12  
But if it came out more looking like the orange line, then you could see that a slightly different, not necessarily a linear equation.

8:18  
And maybe the easiest way to do it still would be to approximate using a linear straight line.

8:25  
So given the data on the left, if you had a 9 LB Turkey, how long would you cook it for?

8:28  
It's two different models there.

8:30  
So, and if you had collected all this data, it would be a task for supervised learning or unsupervised learning.

8:35  
So the answer is because you've got all these right answers already and you've got an expert or a meat thermometer to verify.

8:41  
It's essentially a supervised learning problem.

8:45  
If we assume a linear model, what are the parameters that we need to learn?

8:48  
So we need to figure out these two parameters our we know our weight, we're trying to figure out the time we need to figure out our M and figure out our C So imagine you had this data where W is our input variable, T is the model which can be considered a regression target, which is the time and you need to learn M&C.

9:09  
So when we're doing this, say we have got M turkeys that we've cooked already and we know they're all and we have recorded the weight and the time we have one training example.

9:18  
Weight and time is just denoted W&T, while the ith training example is denoted WI and TI.

9:29  
You'll notice in the previous slides that you know, there's all these different letters being used.

9:33  
And it gets super confusing that we use variable names usually that match the problem, as in West for weight and T for time.

9:42  
But we're not always, first of all, going to be dealing with linear models.

9:45  
And it also gets confusing if you've got lots of different models and you've got lots of different letters.

9:48  
What's the dependent variable?

9:50  
What's the independent variable?

9:51  
What's the input?

9:52  
What's the output?

9:53  
So what we're going to do through the course as much as possible is use notation that can be scaled up to larger models.

9:58  
There's no set notation, although there is pretty much conventions for sure, but nothing really defines.

10:04  
So with whatever, with or wherever possible, we're going to use what's in site, site get learned, which is that the weights, the parameters that we're going to train, we're going to use W for that.

10:16  
And normally Y is always our output and X is our input.

10:20  
So here we've got 2 parameters in a linear model.

10:22  
Sometimes in the kind of traditional algebraic form of the equation, it's MX there plus C.

10:29  
But what we're going to do with psych, at learning, with machine learning is give kind of standard W for our parameters.

10:36  
And even in one of the equations we had the one second, what was it?

10:41  
The time, you know, different letters.

10:43  
But we're going to use this notation as much as possible.

10:45  
Parameters listed from West 0 W one.

10:48  
Eventually we might have W2 W 3 W 4 for different forms of regression.

10:54  
Just a warning on the use of linear regression is that you're always better to eyeball the data before you can assume anything, especially making the assumption that your data follows as some sort of a linear relationship between your variables.

11:06  
Linear regression is based on the example that the data points are scattered about a line.

11:10  
However, often data points can be scattered across something else, like a curve that if you have a set of data points like this, you can still fit a line to it.

11:18  
But what's happening in fact is that as we go along this axis, these data is this data is starting to drop.

11:23  
So you can still compute the coefficients of your line W0 and W1 and it'll give you this line.

11:29  
But you can see it's an inappropriate fish because it's actually something different going on, something non linear happening within our source data.

11:39  
Well, yeah, the curve has shown that they're discrete, decreasing after a certain point of X everything, the value is starting to drop.

11:45  
So the criterion for finding a regression line and even whether to use linear regression in the beginning, draw the scatter plot, have a look, eyeball your data, and if the data does not appear to be scattered across a line, just don't use linear regression.

11:58  
There's different types of regression then known different options.

12:02  
That's just a simple linear regression model, what we've been looking at over the last few slides where there's only one independent variables and one independent variable.

12:09  
But the real world and the real world relationships don't always work like that.

12:13  
So there's polynomial regression, which you might look at where you've got these terms which can be squared into the different powers, Linear regression where there's multiple.

12:21  
So the the relationship is still linear, but there's got multiple features, multiple inputs determining in your output.

12:28  
So Y isn't just determined in the case of the Turkey, the cooking time isn't only determined by the weight of the Turkey, but you could have other variables in in there that would affect Turkey cooking dump.

12:39  
I actually don't know but just pick that example there.

12:41  
Like the temperature of the oven would be another variable I think in that Y is equal to MX plus C.

12:46  
It assumes the temperature of the oven, but that could be another variable.

12:48  
To get the perfect the cooked Turkey, you could have maybe the humidity.

12:53  
I don't know what else would increase Turkey times.

12:55  
There could be a few other variables in here, which may be the temperature of the room or the temperature that the Turkey came out of the freezer.

13:02  
All of these things would have an influence on your cooking time and you could build a much more complex model, but that would drive people crazy at Christmas.

13:10  
But it's just like the.

13:12  
Yeah, rather than just having one variable, like your output can be depended on multiple variables like stock price.

13:17  
Again, we will look at some of these other forms of regression later in the course, but just always remember that W in all of these cases, your weights are always linear, even though some of the terms maybe nonlinear.

13:31  
So we always write these weights there can write these weights rather as a vector.

13:34  
So W is a set of W 0, W 1 all the way up to West.

13:39  
And that's our vector.

13:43  
The next thing to determine is how, how do you actually get the best fish or how do you determine the best fish?

13:51  
And this is often known as our measure of loss or kind of error.

13:55  
So we can see here in this relationship that there is a straight line, but there's also points that aren't on the line.

14:02  
Now there's lots of different straight lines that you could draw.

14:04  
There's one that you could draw that's a bit less shallow of a slope, another one that you could draw with a bit more of a steep slope.

14:11  
And how do we determine that this line right here is the very best 1?

14:14  
So what we do is we try and determine the error between our line and all the points.

14:19  
So we look at every individual point and we look at the distance from that point to the line, distance from this point to the line.

14:25  
And then we sum up the distance over all the points from our predicted value to the actual value.

14:32  
Again, this is our predicted value for this value of X.

14:35  
And that's our actual value.

14:36  
It's our predicted value, sorry, our predicted values on the line.

14:39  
And this is the actual value.

14:41  
And we look at the difference and we try add up all those differences and try and make those differences as small as possible.

14:46  
So your best fit is 1, where there's as little difference between the actual values and the predicted values along the line as possible.

14:57  
The prediction's not going to be exactly the same as we fed the data at the model.

15:00  
There's always going to be a bit of an error between the prediction at a particular value and the actual recorded value.

15:04  
Most of the time anyways.

15:06  
Especially, you know, if we think of the example that we put up with the friends versus minutes online, there was a curve and a relationship there, but there was lots of variances because the reality is, is that this relationship between friends and daily minutes online is too simplistic.

15:18  
There's much more variables that affect how much time someone spends online.

15:23  
So there's some amount of error on the model versus the real data, and we try and figure out what that error is and try and minimise the error in order to create the best model.

15:32  
The function will make a prediction of each observed data point.

15:36  
That's what our function MX plus C or W0 and W1, if you use that terminology, makes a prediction for every value of X.

15:44  
It'll make a prediction for our dependent variable.

15:49  
The observation is noted by Y.

15:51  
So that's what we actually observed from the real world.

15:54  
But the prediction is known as this Y hash.

15:58  
The line with the smallest amount of error is the model that we want.

16:01  
So as I said, the best fish is the one that minimises the error between the prediction and the observation.

16:08  
The way we figure that out, as I said, is there's a few different ways of doing it.

16:13  
Simplest way is a mean squared error.

16:14  
But all of these functions where where you're trying to calculate the error is known as the error function or the loss function.

16:20  
What we're trying to do usually is measure the difference between what the model predicts and what the actual values are.

16:26  
There's lots of ones to choose from and typical example is mean square error.

16:31  
We can see here is again, it's just taking the difference between the predicted and the actual value and then summing up all the M measurements, sum of them all up.

16:39  
This 1 / 2 M is just kind of chosen to make the mats look a bit nicer.

16:42  
But even if you don't use this, this is just a kind of a scaling constant.

16:46  
But this right here is the core, the kernel of the equation where it's just adding up all the error and setting the error to be your loss.

16:52  
So your L here, and you're trying to minimise that.

16:55  
You're trying to tune the parameters W zero and W 1 so that that are M&C so that that loss is as small as possible.

17:08  
So we've got these parameters and we're trying to find the parameters that give us the smallest L, quite easy for linear regression.

17:16  
And there's a known formula.

17:17  
We can use some algebra to do it as well, but with more complicated models it's a lot trickier.

17:25  
This idea works in all ML algorithms.

17:27  
That's what an ML algorithm is.

17:29  
Machine learning algorithm is trying to minimise this loss where you've got lots like billions of parameters that I said that gives us the smallest loss.

17:36  
But the technique for finding these parameters might be a bit longer in more complicated models.

17:40  
We're not just using basic algebra this as well.

17:43  
There's lots of different methods, gradient descent and all sorts of optimization and search algorithms to search the whole parameter space.

17:49  
Because if you've 220 billion parameters and all those parameters can range independently of each other, it's like you really need to have good algorithms to try and and narrow that down and converge towards a solution.

18:01  
It's really true of all kind of algorithms, including neural networks.

18:05  
One of the methods that we use is called gradient descent.

18:08  
We'll cover that later in the course that does lots of iterations and you're slowly creeping towards the best parameters.

18:13  
There is limitations to that, but when we use the dot fit method, which we will see in psych it learn or intensive flow, that's essentially what's happening is that is doing a gradient descent method to creep slowly towards the best solution.

18:30  
It's just you're trying to minimise that loss function bit by bit.

18:35  
So what we're going to do is, or what we'll do is within site kitlearn or if you wanted to build a model.

18:40  
So we have a set of data that's a pair of information X and yx being the inputs, as I said before, and Y being the predicted outputs, X.

18:48  
Yeah, here it is all the corresponding outputs and a particular Yi is the result of a particular XI.

18:54  
So what you can do is from SK Learn, you import your linear model.

18:58  
You can create your objects, which is your linear regression, and then you can do the fish on it where you're passing the inputs and the outputs that you know already.

19:08  
And then what you can do is with a different set of X or the same set of XS, you can actually predict the values.

19:12  
And you can look at how those predictions differ from Y and see how good the model is at fishing your data.

19:20  
You can calculate your Azure error metrics and all that to try and figure out how that is.

19:25  
So that's just a quick run through of linear regression, a well of regression in general and some of the other mathematical notation and conventions that we'll be using within this module.