Robert_S3_L16

Thu, 2/17 4:02PM 14:39

SUMMARY KEYWORDS

goals, outlier, shots, coefficient, correlation, trendline, scatterplot, slope, squared, select, regression line, click, beta coefficient, rid, true, change, axes, analysis, further investigation, format

SPEAKERS

Robert McKeown

R

Robert McKeown 00:05

Our last series of questions is asking us to work with the sheet goals. We have a bunch of questions to address. And we can do these maybe a little more quickly than it might appear given just how many questions there are laid out here. Now, as we click on the goal sheet, we can see we've got shots on net goal scored and goal scorer with outliers. This is similar to what we did in previous lectures. And we've got goals and goals true. And so we're supposed to do analysis of regressing and looking at the correlation between goals and shots. And then we're supposed to look at goals true, and shots as well. So why don't we start by creating scatter plots, and we'll create a scatter plot, we'll create two scatter plots. So I'm going to click on some whitespace over here, and I'm going to go to Insert. And I'm going to choose a scatter plot chart. And I'm going to right click on the chart and scroll down to select data. I'm going to go to series here. And I'm going to add, and we are asked for a series name, and we'll call it goals scored. We're asked for our x axis values, and those are going to be shots. And then we're asked for our series Y values, I like to delete this, whatever that equal one is supposed to mean and get rid of it. And then we're going to our y value is going to be goals. That's what we're going to be trying to predict. And then I'm going to press OK. Now we've got we've got goals scored here, the horizontal axis, I believe these are our shots. And I'll press OK. And now we've got a scatterplot of goals scored against shots on that. And remember what we did last time, first, I will make this larger so that it's easier to see and read the text. Next, we might want to change and format the axes. So the minimum shots is 100. And the maximum is actually just 260. Next, we can get rid of these gridlines that are so problematic. Now we've got a nice whitespace. And now we can add in that regression line by clicking on Chart Elements, this plus item here, select trendline. And we can go to more options, I can make the line thicker. And I can change it from a dash to a solid line. And then maybe I'll pick a different color so that it stands out. Also, we could add in by clicking up here where it says trendline options, we can choose to display the equation. So there's our regression line. And there's our R squared. And remember, the question was asking us, what is our beta naught? Well, our beta naught here, just over there for now, okay our intercept is 8.1651. And our slope is 0.0653. How else could we calculate these suppose that we didn't see them? Well, there are some commands in Excel that we can use. If we want to know what the beta one coefficient is. I can press equal and type in slope. And all I have to do is select the Y values. Those are our goals, what we're trying to predict. And what we're using to predict to goals that's shots. And I close

the parentheses, hit enter. And look lo and behold, we have the slope coefficient that we got on this graph. There's also an intercept command. And if I want the intercept, I can do the same thing. Select the goals. Select the shots. Alright, remember the shots are the independent variable, the goals are the dependent variable. And there you have it, we can, we can get these coefficients, our beta coefficients quite easily. Now, we've got an R squared, that's our correlation, determinan- determination is 0 point...0.184. And if we want to know what our correlation is, and we called that are in a previous lecture, it's just going to be the square root of our correlation determination. And so our correlation here will format to two decimal places 0.43. And maybe we could do the same thing for our correlation determination. If I wanted to, I could highlight this whole column, I'll go to format cells, and change the formatting for the whole thing. Now looking at this, we see that it doesn't seem like the correlation determination is very large, the correlation is 0.43. That's something but not very large. When we look at our graph here, and I hope the regression line is not I guess we can get rid of that. We could label our axes. But keep going. We can see that we've got a big outlier out here. And why is that an outlier? Well, maybe we want to add a data data label to that point, you can see that it's goal scored is five, but the shots on net were more than 250. And it's quite a bit further from the line than any other observation. And so it looks like we have a outlier here. Now, for the second part of the question, we're being asked to do a similar analysis that we did here, but with goals true. And the only difference between goals and goals true is this outlier that we identified. And it turned maybe it turns out when you look at the data, that there was a mistake in the inputting of the data, and the goal should actually be 36. So let's redo our analysis. Let's run another regression using this goal is true, and see how it compares to this data that had the the outlier. So I'm going to click on an open space. I'm going to click Insert, I'm going to go to charts, I'm going to choose scatter. Got a lovely, empty white space. I'm going to right click on it, I'm going to go to Select Data, I'm going to add a series under legend entries. The name of this series called goals true. And the x- x axis values are going to be whoops, it's not quite right. Get rid of that the X axis values are going to be shots. That's what we're using to predict goals. And make sure we delete that one that's in there where it says select Y values. Oops, I almost made a mistake. We're going to select goals true enterprise press OK. And looks looks like our title got wrong this is the series name should be goals true. And we will click OK. Not sure why it's so truncated. And here is our scatterplot of goals true against shots. I will make things a little bit bigger for us here. So we can read the values more clearly. Next, I want to format the axis no one has less than 100 shots so you can start at 100 I want to get rid of the gridlines get rid of the gridlines and now we're ready to add a trendline Linear, but we want better options than that, we can make it thicker. And we can have a solid line our regression line information, so you can right click on the trendline and then go to trendline options. And we'll display both the equation and the R squared. And if you take a look at our equation here, it's quite a bit different than our equation when we had that one outlier, you can see that the R squared has increased quite a bit, it's two and a half times larger. Instead of explaining only 18% of the variability. It's now explaining 43% of the variability. And also the coefficient, the beta one coefficient on shots is higher, instead of one shot, increasing your expected goals by 0.065. A one shot is increasing a player's expected goals by 0.101. Rounding to three decimal places, or 0.1007. And so this analysis demonstrates to you the power of outliers, outliers can drastically change the result of a regression. Because they're so powerful, and they can really change the results of your regression, we have to be careful about how we deal with them. If you know you'll, there's no one method, you don't always want to take them out. You don't always want to leave them in, they require further investigation, there are some, it's going to be more art than science, there are some techniques you can use to deal with outliers. But here, I'm not going to get into them just going to demonstrate to you the power of outliers in changing results, which I think this this exercise demonstrates very clearly by changing the R squared and the slope coefficient by so much. To answer the last part of the question, which involves creating a scatterplot of the

residuals against the predicted goal score, I went ahead and and calculated the predicted goals and the residuals using both goals and goals true. And so we can create two scatterplot diagrams. Here's the first one. And here is second. And we can take a look at the two of them. Now might be a little bit it should be fairly obvious that this is the outlier. Notice that the scale both are a maximum of 15. Rez to these are the residuals using goals true. The residuals fall between negative 10 and a little bit above positive 10. When we're looking at the residuals from just regular goals the residuals fall from a little bit greater than 10, all the way to minus 20. And so for this player, the bottle is overestimating the number of goals by 20. And so it's pretty clear in this diagram, that there's a outlier here and that there are no outliers in this diagram here. And so we can see that this outlier is causing a lot of distortions in the beta coefficient estimates and consequently in the correlation and correlation determination