Robert_S3_L15

Thu, 2/17 4:02PM 20:51

SUMMARY KEYWORDS

residuals, scatterplot, trendline, graph, predicted, line, income, columns, regression line, beta coefficients, axis, click, cells, regression, x axis, sample standard deviation, variable, change, y hat, variances

SPEAKERS

Robert McKeown

R Robert McKeown 00:06

We're now ready to work on question number five, we're going to work with the data in the education sheet. And we're going to perform a linear regression of income on education, or we're going to try to use education to predict income. We're then asked to identify the beta folk coefficients, correlation, and the correlation of determination. So let's go over to the education sheet. Now, we had some work on this sheet already, and I copied some information that we'll need, we have a correlation coefficient, and we have the correlation of determination the r squared. So one way to perform a regression would be to do the following, we could calculate our beta one and our beta zero, we know that the formula for beta one is the correlation multiplied by the standard deviation, and we'll use the sample standard deviation of the Y variable. So that's income income is what we want to predict. So we have the correlation multiplied by the sample standard deviation of income, which we're trying to predict, divided by the sample standard deviation of education, which is in cells B 3 to be 102. Seems a lecture so make sure we have that right. I'll hit enter. After looking at this for a moment, it seems like the Y hat is too large. So maybe there's a mistake somewhere in our calculation. If I look at this equation for the intercept, and I click on it, notice that I called on column D, which is not the correct column, I should have called on column B. So let's change the D's to B's. Hit Enter. And now we have a new intercept, that's negative $8,556. Remember that the intercept, we don't want to take it too seriously. It's doesn't carry a lot of social science intuition. Now, if we look at our y hat estimate, it seems like it might be a more realistic expectation for someone who has only a high school degree. Looking at the formula, I want to make some changes to it, I don't want the columns can, we don't really want the columns to change, and they won't change by just drawing down. But the numbers, the row numbers are going to change. So we want to make sure that if we're going to use this formula and other cells, it keeps drawing on our beta one coefficient, and our beta naught coefficient. So I'm going to add a dollar, sign to J 3, and I'm going to add a dollar sign to the three in I 3. And as long as we just copy and paste this cell downwards, it should not call it should not change, and continue to draw on the cell that contains our beta coefficients. And so here are all our numbers, I could format the cells and turn it into a currency. If we wish, maybe we don't need the cents. So we'll just have it in dollars, now looks very uniform. And there's that dollar sign at the front. So here are our Y hats. Now we can graph our results. And there's a few ways to do that. Based on what we have, I'm

just going to copy and paste the education and income into these K and L columns like so the reason why I'm doing that is because when I ask Excel to create this graph, it's going to take the first value as the X axis and the following values will be on the Y axis. And I'm going to highlight these columns, I'm going to go to insert, I'm going to choose scatterplot, with lines, and we kind of get a very, very messy graph. So maybe I'll try and do that slightly different way. Let's see what it looks like with dots. And so the scatter plot, the blue dots in this chart, are the actual values. And the orange hat is the line the regression line of least fit. So our little yellow dots here, these are the predicted ones where my cursor is going along. And then we've got these blue dots that are the actual observations scattered all over the place. Now this turns out to be not a very great way to create our graph. So let me actually delete it. And let's just go create a scatterplot of our original data. So I'm going to insert, I'm going to go to two charts, I'm going to insert a scatter plot, and I'm going to insert it just with the dots, like so. Now to get started, I've got a lot of whitespace over here, let's get rid of that. So I'm going to click on the X axis, click on Format Axis. And then over here, for where it says minimum, I'm going to make the minimum be, say 12. Or maybe 11, maybe 11 is a better minimum. So we've got a little bit of space to the left, but not too much. And the maximum value is 23. That seems okay to me. And the rest of it looks pretty good. I do want to get rid of these gridlines. So we can get rid of the gridlines. If I select them, and then right click them. Where it says automatic here, I'll just choose no line. And then I'll click on the horizontal ones, and right click on that. And then I'll say no line there. So now we've got a nice, nice white space clean looking scatterplot diagram. And what else could I do? Maybe I'll add, maybe we could add axis titles. And so on the y axis we have income. And on the X axis, we have years of education. And we want to maybe make the text larger so that all of us can read it. And I'll just click on that button. And things get bigger, we don't really need this title. I'll delete the title. Now that we've got our axes labeled, we don't really need a title for the graph. Now I did oh, here we go, I open the graph, it seems to have shrunk down the x axis to a maximum of 23, which is what looks pretty good to me. Now, what about a regression? Well, we can actually add the regression line using the options that are available on our chart. So if I click on the inner chart, the area where the scatterplot is, and I click on this plus button for chart elements, we can add a trendline. And this trendline is going to be linear by default. And you can see the trendline there is this dotted line in blue. We can why I clicked on it, and then I right clicked and I went to format. And when we format it, we have some options, we want linear, we're not going to use the other alternatives. And we want to display the equation on the chart and there is our regression line, and it matches our beta coefficients. And so we can see there, we could have calculator a beta coefficients, just by creating a scatterplot. Now this line, the format the trend line, it looks a little little soft, it's kind of small, and dots, let's see if we can change that. If I select this paint can work, which is called fill and line, if I slip, select that I've got an option for width here. And we can make it a thicker line. There should be we can change the dash type, we could make it thick, like just a solid line, and I'm going to change the color. There's black, black seems pretty reasonable, maybe green is a little bit nicer. But everyone at home you should be able to see that there is a solid line here, I'm not sure if green was really an improvement. At least not that shade of green. There that's a little better. And you can see where our predictions are going to be. And you can get a sense of the variance as well you can see that the observations are not always that close to our regression line. Now if I go back to our regression line, and I right click on it, and I choose Format trendline. I can also display our art squared. And you can see that our R squared here is the same that we've got that we calculated earlier in a previous video. And if we were take the square root of that, the square root of that is going to give is our correlation coefficient of 0.6. And so this is how you can perform a regression analysis using Excel. There are other ways, some of which I'm going to show you in our next upcoming videos. But this is a great place to start, create a scatterplot, and then use the trendline option, and you'll get a nice linear regression. Now, the last part of our question asks us to calculate the residuals. So let's go

ahead and let's calculate some residuals. And then we're going to look at them on a scatterplot. So we already calculated the Y hats. So if we want to know what our residuals are, we take the actual value of income, C 3, subtracted by our Y hat, predicted value. And there we get a residual of 37,769. So in this case, our regression line underestimated the income that this individual was going to make based on them having 12 years of education. And we can copy or fill that equation through all the rows or all the cells in the end column. And we're going to get all the residuals now we can double check and make sure Oh, it seems like I forgot one. No, I did. We can check to make sure that we didn't make a mistake. And the sum of the residuals is equal to zero. So that's good. That's what we would suspect. Now that we've got our residuals, the next part of the question is asking us to graph our residuals against our predicted income, which is Y hat. And so we can do that. Now what do we want on the x axis? What do we want on the Y axis? I guess it doesn't really matter too much. So I'll just go with the order that they're in. So we should have our y hat on the x axis and our residuals on the Y axis. I'm going to insert, go to insert and choose chart scatter. And we don't need any lines connecting it. And there we have our scatterplot. I noticed that they're all dotted around this residual line. What can we do to make this graph easier for us to look at? Well, we could go to select it and go to home and then add or change the size of the font so that it's easy to see, easier to see. We can get rid of these gridlines that are not not very nice to look at. I don't think they seem to just be kind of noisy. And we could format the axis if we wanted to and have major unit have fewer major units. Wonder what would happen if we had 120. And we sort of get rid of it. And remember, what what are we looking for here, we're looking to see a cloud. This looks pretty cloudy to me. So if I'm looking at this, after doing a regression, that seems to me, like it's pretty good, there might be a few little observations down there, and maybe a few up there. But so I would say that this residual scatterplot looks pretty healthy, which means that our linear regression was the appropriate model, the underlying relationship is likely linear. And we don't see any obvious outliers that might be skewing our result. We also don't see any shapes, and we don't see a cone shape, which would suggest that the variances are not constant, the variances are changing. And we don't see any curves and things that might suggest that the underlying relationship is not linear. We're now on Question seven, and we're going to perform a linear regression of crime on temperature. So looking at whether we can use temperature to predict crime, and we're going to use Excels analysis tool pack to do the regression. So let's head over to the sheet crime. Here we are, we've got some older information that we're not going to need. And so what I'm going to do is I'm going to highlight these cells, or I should say the columns D to K. And I'm going to right click on it, and I'm going to go to hide, so I didn't delete them, I've just hidden them, so that we don't have to worry about them for now. Now, you want to make sure that by going to File and options that you have installed the correct addons. So I've got analysis tool pack. If you've got it down here, just click on it and press go. And you'll be able to install it. Since I've already got mine installed right here, Analysis Toolpak, I'm going to hit cancel here. And I'm going to go to data. And data analysis appears on the top far right, it may just be an icon, occasionally, that's how it appears. And I'm going to click on that. And remember, we use this to create a histogram previously, this time, we're going to go to regression. And it wants us to know the Y values, and the Y values are going to be crime. And that's going to go from C 3 all the way to C 102. And the x input range, what is our X value, it's going to be temperature. So it's going to be A 3 all the way to A 102. And if we add in the first title, we might even get it, we add in that first cell, we might even get it to automatically add in a whoops might even get it to automatically add in the title for us. Now, let's not worry about these options, we're gonna have it appear on a new sheet and we're gonna call the new sheet crime to or better yet, crime underscore reg for regression. Maybe we want the residuals so that we can take a look at them. We could also ask it to plot some things for us. So we might want a residual plot. We're interested in looking at the residuals. And we might want a fitted line plot, if we want to see what the regression looks like compared to the actual values. So

let's select those options. And let's press OK. So it looks like it was kind of unhappy with me. It didn't like the titles. So let's change the first row to three instead of two. Let's see if this works. And all of a sudden, it created a new sheet called Crime underscore reg. And it created these cool plots for us that we can see here. And we can see a few things of interest here. So some of this, you don't, we haven't covered. So we don't need this information here. So I'll just clear that out. We didn't discuss anything about ANOVA. So we don't need that. And we haven't learned about any of this stuff yet. But you will learn some of it in the next module. But we did learn about the coefficients. So we've got our this is the intercept, our beta naught coefficient and the X axis variable. That's our beta one. And it also predicted the Y hats for us predicted Y's and our residuals down here. Looking at the graph, oops, you can see that maybe we want to make this look a little bit nicer. We could change the X axis variable, so that there's not so much whitespace. Maybe we'll go from 17 to 37. And we probably don't need the legend here. So that will give us a little bit more room to look at. And we've got our X variable X variable one. This is simple linear regression. So there's always just the one single X value. And we can see our fitted line going through the scatterplot. And it looks pretty good. It looks like they're about as many observations above as there are below, maybe something strange going on up here. We could add in the trendline if we wanted to. And we could add in the trendline for the predicted Y and that would just give us a dotted line along there. If we added in the line for just the Y's we get the same thing because of course, our fitted line is equal to the regression line. If we want to take a look at the residuals we can look at the residuals first blush it looks pretty good. It's Pretty nice and cloudy. The residuals are around zero, which we would expect, and there's no discernible pattern here. As we relate the residuals and the x variable. Of course, we could change the formatting of the axis again to make it a little bit easier to look at. And that just sort of, it looks relatively like a cloud. Got some observations here, but they're not very far away from zero. So I don't see any problems in terms of the I've seen any shapes or patterns in the residuals