Robert_S3_L11

Thu, 2/17 4:01PM 13:39

SUMMARY KEYWORDS

residual, model, data, goals, predicted, beta coefficients, prediction, calculate, scored, overestimated, equal, regression line, outliers, green dots, equation, y hat, sum, number, values, shots

SPEAKERS

Robert McKeown

R Robert McKeown 00:06

Now that you've seen how to calculate a linear regression, you calculated the beta coefficients, and then you made a prediction about Y, the Y value based on a value of X. Now we're going to turn our attention to looking at whether that prediction is a good prediction, we can use what are called the residuals to evaluate whether the linear regression was accurate or not. And so in this video, I'll explain to you what the residuals are. And we'll go through an example or actually, we'll continue the example of the shots on net, and the goal scored from the NHL. Every model is imperfect. When you're working with real world data, you're never going to have a model that is so perfect, that the correlation of determination will be 100%. And since every model is imperfect, there's always going to be an error. And we call this error the residual. And for one observation, we call it a residual. The residual is related to the model itself. And so we can think of it this way, imagine you've got the data, specifically, those Y variables, the series of values that are associated with the Y series. So in our previous example, that would be goal scored, we could say that goal scored is going to be equal to how we model the goal score, that would be like our Y hat, plus some residual. And if we add a residual to the models prediction, it's always going to be equal to the data. And we're going to use this concept of this equation here to find and calculate the residual, and we're going to infer it from this equation. So we can say that the residual is going to be equal to the values in the data minus the predicted value or the predictions of what those values would be, according to our model. More simply, we're going to define the residual as the letter E. Sometimes, instead of using E, you might see an R, R for residual, and you might also see a U. Here on these slides in this course, we'll use E. And so E, the residual is equal to the observed value of Y minus its predicted value. And of course, the predicted value of Y, Y hat is just equal to and so each data point is going to have a predictive value. And it's also going to have a residual, these residuals are useful, we can use them to measure the accuracy of the model. And in this course, in this module, we won't do too much of that we'll do a little bit, we can also use the residuals to check the conditions. The conditions for linear regression to be accurate, were that the relationship between the two variables was indeed linear, we had to be sure that there were not outliers. And the data had to be a quantitative variable. We're also a little bit later a future video, we'll add in some new conditions to that as well. Now before we go any further, let's make sure that we really understand what a residual is, and how it relates to the work that we've already done together. If we look at the graph, that a simplified graph, with just a limited number of players from the

NHL, we've got shots on net on the x axis and goals scored on the Y axis. And of course, this is a scatter plot. And the green dots represent the data, while this purple line here represents the regression line. And that regression line is our Y hat values. So for any X on this axis, the model is going to predict a point on this line. As I mentioned, the dots the green dots are the data there the series of values associated with the y variable. So in this case, they're going to be a bunch of goals scored and shot The shots are keyed to the x axis down here. And the goals are key to the Y axis. If we want to find the residual on this graph, each of these green dots, each of these observations, or data points, is going to have a residual, so fairly interested. And this data point right here. This hockey player had up, well, let's just say about 122 shots on net. If we want to know the prediction, their prediction of the model would be that they had, let's call this 19 goals. So the model is predicting they would have 19 goals, how many did they actually have? They had 15 goals, and they had 122 shots. And so if we want to know the residual, the residual is going to be the distance between the green dot and the fitted line. What is called the regression line on this diagram. So using the numbers that are approximating, because I'm just looking at the diagram, our residual here is going to be equal to the actual goals for which is 15. minus the predicted which looks approximately 19. Maybe I can do curvy lines to make an approximation. And so at this point, the residual for this observation is going to be equal to negative four. The predicted is higher than the actual. What we actually see in the data. And in this case, the residual is negative. Let's answer some questions together. The first one says that if a player takes 120 shots, predict the number of goals. And so we can do that using our linear equation, our estimated linear equation. And so Y hat is going to be equal to negative 0.43 plus point one one times 120. And using Excel, I find that the expected number of goals is 12.78, using these rounded numbers that are available to us on the slides. Now part B is asking us to calculate a residual. Notice that both the players, both the actual player and this hypothetical player have 120 shots except that the hypothetical player only scored eight goals. So the residual E is going to be equal to eight minus 12.78, which is going to be equal to negative 4.78. We have a negative residual. And Part C is asking us did the model overestimate or underestimate the number of goals that the player would have? And the answer is the model overestimated overestimated the goals scored by how much by 4.78. So a negative residual implies that the model overestimated the actual number of goals predicted a higher number of goals than were actually scored by this player. Just as we saw in our previous example, a negative residual implies an overestimation. A positive residual, on the other hand, is going to imply that the actual value is larger than the predicted value. Consequently, our model is under estimating the true value. Some other things to keep in mind when you're looking at your regression results. Remember that the regression line is giving you information about the average of values. That's what the averages is how we calculate the intercept. And therefore, it's having an effect on our predicted value. And so we're not going to these models are not going to capture outliers very well. So we have to be very careful when it comes to outliers, and deciding how to treat them. As previously stated, the accuracy in the models prediction is captured in the residuals. So you saw that we could look at the sign on the residual, and that would tell us whether it was an over or under estimation, we could look at the actual value itself. And it could possibly give us some sense of the magnitude of the over or under estimation, we might consider four goals not to be very much, but maybe eight goals is a lot. Remember that the sum of the residuals is always equal to zero. If you add them all up, you're gonna end up with zero, just like when I was just like when we were looking at deviations from the mean, the sum of all the deviations from the mean was also equal to zero. That means, if we want to get a measure of the size of the residuals, we're going to have to transform them. And we're going to transform them very similarly to how we treated standard deviation, we're going to square these residuals. And we might want to add them up. We're not going to do that in this module. But that is a common technique, that will give us an idea of the variance of the sum of squared residuals, that's really good to get as close to the variance calculation. The last

point I'd like to make is about the regression line, which has many names. Sometimes it's called the line of best fit. And it's also called the least squares line. And what it actually does is eliminate minimizes this equation right here. So if you're wondering where those coefficient calculations came from, it came from a procedure that minimizes this sum of squared residuals. And in a future course, if you're interested, or if you want to look it up online, you're welcome to, but we won't be going in any more detail, about how to calculate the those beta coefficients will just take you can take the equation as given for now. statistics can be complicated. There's a lot of notation to keep in mind. There's a lot of terminology to keep in mind. And there's a lot of rules and different things that you're expected to understand. Try not to be overwhelmed. Focus on the things that you do understand, focus on how to calculate those beta coefficients. Understand that a linear regression is used basically fitting a straight line through the data so that it minimizes the sum of the squared residuals. And in the next module, we'll look at some graphs of residuals. To better understand how accurate our linear model is, indeed, whether we even have the correct model at all.