Robert_S3_L10

Thu, 2/17 4:01PM 14:07

SUMMARY KEYWORDS

goals, shots, simple linear regression, score, equal, intercept, sample standard deviation, data, prediction, correlation, net, beta, beta coefficients, information, linear regression, coefficient, player, number, negative, variable

SPEAKERS

Robert McKeown

R Robert McKeown 00:05

Hello and welcome. In this video, we're going to apply what you learned about linear regression to a real world example. Before we get into that, I want to just go over the important conditions under which linear regression is going to make a accurate prediction about your variable of interest. To be accurate, linear regression requires three things that the data be quantitative, that the relationship be linear, and that there'll be no outliers. This should be familiar to you because it's the same conditions under which correlation is an accurate measure of the co-movement between two series or two different variables. As an example, I've collected data from NHL hockey players, specifically looking at shots on net and goals. So there is a wisdom out there among hockey fans, that there's a connection between how many shots a player gets and how many goals they score. If you look at the diagram, in front of you on the screen, you can see our scatterplot of goals and shots right here. So we have some data, each observation, each row of our column of data is going to have two pieces of information, it's going to tell us how many shots a player took how many shots on net they had. And it's going to tell us how many goals that same player had. So if we look here, we can see that we've got one player had 100 shots on net. And maybe they had it looks like they had six goals, five or six goals. And so we have, I believe it's 300 observations of hockey players, we restricted it. Got some more information. If you're interested in the data, you can look at our came from down here. Now we want to make a simple linear regression. In order to predict the number of goals an NHL forward will score if they had 75 shots on net. Now, because of the way that we taught, I've told you how to calculate the beta coefficients. The only information we actually need to perform this is in this table right here. I reproduce the table on a fresh sheet so that we can perform some calculations. Now if we want to do a simple linear regression, the first piece of information that we need is going to be our beta one coefficient. Remember that our prediction, Y hat, which is number of goals, is going to be equal to beta naught or beta zero, plus beta one times X, the number of shots on net we've got our correlation or sample correlation, multiplied by our sample standard deviation of Y goals scored divided by our sample standard deviation of shots on net. So here, we've got Y is equal to goals. X is equal to shots. Everything we need to calculate beta one is in the table above. And so I can write beta one is going to be equal to the correlation. Notice the correlation for goals with shots is equal to the correlation of shots with goals produced twice. So we've got 0.675 multiplied by the sample standard deviation of goals

3.97 divided by the sample standard deviation of shots on using a calculator or spreadsheet software like Excel. We're going to find that beta one. Maybe I'll rewrite it down here beta one is going to be equal to 0.11. Now that we've calculated our this is our slope coefficient, right, this is our slope coefficient right here. We're ready to calculate our intercept. And R intercept is equal to the average value, or the mean value of Y minus beta one times the mean value of x beta one times the mean value of X. Again, we have all the information we need in this table up here, we've got y bar is the average number of goals, which is 7.59. minus beta one, multiplied by our average number of shots, which is 72. And using a calculator or spreadsheet program to make it much easier, much faster, we're going to have negative 0.426. Rounding to three decimal places. Here are the beta coefficients. These are the two pieces of information that we need, in order to make a prediction using simple linear regression. The estimating equation is going to be y hat. This is goals is equal to negative 0.426 plus 0.111 times X. Now we can do a little analysis here. Our prediction is that for every shot a player takes for each X for one X, the player is going to score 0.111. Goals. They're going to score 0.111 goals. That's the prediction, or maybe the what we might also want to call the expected goals. What about this term here? What does that mean? Well, in simple linear regression, the intercept doesn't carry a lot of valuable information. In fact, you can see that this is kind of nonsensical. If X is equal to zero, then Y hat is equal to negative 0.426. If a player has zero shots, we expect them to score a negative number of goals. But of course, it's not possible to score negative goals. So our intercept is kind of nonsensical. It's kind of nonsense. And when we're working with real world data, it's not unusual to have an intercept, that doesn't really make a lot of sense. On the other hand, this 0.111, this is the ball. I won't say expected. I'll say it's the predicted goals per shot. So each time the player takes a shot, we expect them to score 0.1111 goals. Now, of course, you can't score a fraction of a goal. But you could think of it maybe like a probability, something along those lines in this in this situation. What does it all mean, when we tie this together? Well, why don't we graph this line that we've come up with? Right here, we're going to graph this, and we're going to graph it, graph it through the scatterplot that I showed you on the previous slide. When we do this, we can see that as a player takes more shots. We expect them to score more goals. There is a positive relationship between shots on net and goal scored. Or as Wayne Gretzky said, you miss 100% of the shots that you don't take. So the idea is you better try and do something because if you don't try, you know what the result is going to be. It's not gonna now question now we've got that our estimating equation up here, but we were asked to predict how many goals a player would score if they had 75 shots on net. And so we're going to have X is equal to 75. axis represents our shots. Y hat represents our goals. So we're gonna have negative 0.43 plus point one one. Well, I did another one from before, like we did before, times 75. We're going to have negative point four, three, plus 8.325. And we're going to have rounded 7.9 goals. So player who has 75 shots on net is expected to have 7.9 goals, we would predict that they would have 7.9 goals now, or we could round up, we could round up to eight, since it's not possible have 7.9. If we wish we could round up to eight. And that could be our answer as well. Now some other things of interest, remember that the correlation was equal to 0.675. So the correlation between goals scored, and shots on net 0.675. That means that we had an r squared a correlation determination equal to 0.675 squared, which is about equal to 0.455. How do we use the correlation determination the R squared here? Well, it's basically telling us that 45.5% of the variation in goal scored is can be explained by shots on that by the variability in shots on net. That is good in the sense that shots on net has some predictive power for goals scored in that number, but it also means that there's 54.5%. That's not explained by the model. So this model is perhaps useful, but it's definitely far from perfect. In fact, I might have got ahead of myself here, as we look at the next slide. Question number two, how much of the variability in goals scored is explained by the variability in shots taken? And the answer is going to be the r squared, which is equal to 0.675 squared, and that was equal to our 45.5%. Question number three for each additional shot

taken? What is the predicted change in goals scored? And the answer is 0.11, which is equal to beta one. That's our coefficient in front of the independent variable X. Every time player takes a shot, that you would expect them to score point one one goals. Now, of course, you don't actually expect them to score point one one goals because it's not possible yet their score you don't. But this captures that idea that the more you shoot, the more shots on net that you have, the more goals you're going to score.