Robert_S3_L12

Thu, 2/17 4:01PM 19:47

SUMMARY KEYWORDS

residuals, prediction, model, outliers, linear regression, simple linear regression, predict, histogram, variance, observations, estimate, predicted values, variability, shape, larger, normal distribution, information, accurate predictions, goals, remember

SPEAKERS

Robert McKeown

R Robert McKeown 00:06

Hello, and welcome to this video on linear regression. In a previous video, you learned how to calculate the residual. Remember, the residual is a measure of the air and your model. Since it's a measure of error, we can analyze these residuals to see if the model is accurate. And also to see if any of the underlying assumptions or conditions of the model are not existent in the data. So for example, we might find that the relationship between the two variables is not linear. And if the relationship is not linear, then we shouldn't be using a linear regression. So let's take a closer look to how we can use these residuals to better understand the accuracy and appropriateness of our model. A simple method to analyze the residuals of model estimation is to create a scatterplot. And you want to create a scatter plot of the residuals on the Y axis against the predicted values Y had on the X axis. So we think back to our hockey example, we have residuals on the Y axis, and we have predicted goals, Y hat on the X axis. And we want to look at the residuals. And we're hoping to see that there is no useful information in the diagram. And I'll explain, and we'll talk and look at examples of what we mean when we say that there's no information, what does an omission a lack of information look like? And in words, what it will look like is we should not see any patterns, we shouldn't see a trend, we shouldn't see shapes. And we don't want to see outliers. Now, when it comes to outliers, if we see outliers in the data, when we're estimating our model, we'd also expect to see outliers in the residual. So if we have outliers in the data, we shouldn't be surprised when we see outliers in the residual. If we see patterns, shapes, trends, these can signify various specific types of problems. They can tell us something about the relationship between the variables, we've got the experience to interpret them. And we're going to look at a number of examples of these kinds of patterns that we might see in the residuals and what they might mean and what might be the underlying problem that the model is not able to deal with. Here we're looking at a graph from our example before, we've got the, the scatterplot. With the residuals from the linear regression, we estimate it using data from NHL hockey players. So we've got the residuals on the Y axis. Looking at the diagram, we can see a nice cloud of observations. So our residuals are kind of clouded, especially this area here looks particularly good. They have maybe an outlier out here. And we know from previously that this is a hockey player named Austin Matthews, who plays for the Toronto Maple Leafs, and he is a fantastic shooter and goal scorer. And so he's pretty exceptional. And his observation is exceptional is a

bit of an outlier. But it's not too bad. He's not too far away from the rest of the cloud. And so mostly, this, this scatterplot is telling us that the model did a reasonably good job of estimating goals based on the number of shots that a player takes. Now let's look at some potential residuals that have interesting shapes. And if we look here, we can see a very interesting shape. It looks kind of like a sort of a rainbow. If we had multi colors, we might call that a rainbow. It's bowed outward in the middle. And it's got a curve shape, so it doesn't look like a horizontal line or horizontal cloud of observations. that we would like to see. And we'd like to see lots of observations grouped around zero with fewer observations as we move farther away from zero. So we would like to see that horizontal line or, or cloud shape in our scatterplot, of residuals. And what this is kind of telling us is that if you give me a value for predicted goals, say, say 12, we could kind of predict, we could predict that the residual is going to be positive, right? This area is kind of high, right? It's above the zero line. So we could predict that it's positive. And so we could kind of predict, we can predict that our model is going to under estimate, the number of goals scored, if the estimated number of goals scored is 12. And similarly, if the number of predicted goals is say, five, then we can predict that our prediction is going to over estimate the number of goals scored. And it's in this sense that we know the kind of mistake our prediction is going to make, based on the prediction that it's making. That tells us that there is information in these residuals in these predictions. That has not been accounted for in the model. So if the model predicts we're going to score 12 goals, we know that it's actually probably going to score a little bit more than that. Because our model is under estimating the actual number of goals scored. Here we have a similar situation, except in instead of looking at maybe a rainbow or a hill, we're looking at a valley. In this case, if the predicted value is say, negative 17, and here we're looking at a hypothetical situation, then we expect the residual to be positive. And consequently, we can conclude that this model is under estimating the number of printer the predicted score, the predicted score. And I can't use a hockey example anymore, because we've got negative values on our on our X axis. But our analysis is the same. This model, it's residuals still contain useful information. And so there must be some sort of problem with the model we've chose and its predictions are not going to be reliable. They're not going to be accurate predictions. Here we see another collection of residuals that do not have that cloud shape that knowing that don't include any information, we've got kind of a zigzagging swerving kind of line or what we might call a wave. But it's definitely an interesting shape. And there's a pattern there. That gives us information about how our predictions are faring, we can guess based on the predictive value, whether that predicted value is going to overestimate or underestimate the actual value. Here is a different kind of shape, that's actually going to lead us to a new condition under which linear regression will give us an accurate result. It's also a very important condition. If you go into further more advanced studies of statistics and you start looking at multiple regression. In this case, we see that the residuals have a kind of cone shape. They're becoming larger as the predictive value becomes larger. And in statistical terms, we say here that the variance the variability now the variance is a measure of the variability, but the variance is becoming larger as the predicted value becomes columns larger. And this is a new problem that we haven't seen before. And it really just tells us that really tells us that the rule to follow is the variance of the residuals should be constant. For simple linear regression to make accurate predictions. There are ways to deal with a situation that you might learn about in more advanced statistic classes. But the answer is not to use simple linear regression. So we can use simple linear regression, unless the variance in the residuals is constant. Here is a similar situation, we've got the residuals acted in reverse, they've got a very high variance to begin with. And then the variance is starting to decrease as the predicted values become larger. And this is also a violation of our new condition that the variance be or the variance of the residual, be constant. Another way to graphically analyze the residuals is to create a histogram. So if you look at the slide, and the screen in front of you, you can see that we have a histogram of the residuals, the value of the

residuals is represented by the bins, the size of the bins, and it's in frequencies. So we've got the number of residuals, that our values are there and the band that is shown before you. And this is from our example, in a previous video, using NHL shots and goals. What we like to see in the residuals is that they are clustered around zero, we expect to have many residuals close to zero. And we expect to have fewer residuals as we move away from zero, so lots of residuals up here. And then fewer residuals. As we get farther away from zero, remember that the sum of the residuals is always equal to zero, so that the average residual is equal to zero. And the mean or I should say the mean residual is equal to zero. And that's why or that's part of the reason why we have many residuals clustered close to zero in relative terms. Now one way to help us analyze the histogram is to superimpose a normal distribution around the residuals. And so I've asked a computer program to do that for us, you can see that represented here, in terms of this black line. And if we think we want the histograms bars to kind of fit closely to these lines, now it's kind of hard to see, in my, my opinion, this looks pretty good. We've got a decently normal distribution of residuals, the bin sizes can make things a little bit tricky. But in my estimation, this looks fairly good and that we've got lots of observations close to zero and fewer, very few observations out here at what we call the tails. And this is the upper tail. And this is the lower tail. Here is an example of a histogram that doesn't look quite as healthy as the one we saw before. We've got our normal distribution. And we'd like using the normal distribution, you might remember because it can be defined completely by its mean, and its standard deviation or variance. And we've got lots of observations around zero in fact, we had too many observations around zero more than we would expect. We also have What seemed to be quite a few observations out here and out here. In fact, you can even see that on the histograms scale, there's a plus five is the highest residual we have above zero. And the lowest residual we have, it's actually looks like it could even be, or it's very close to minus 10. So, a little concerning, maybe some concerns there, we'd have to take a closer look, the histograms don't tell as strong a story as the scatter plots. And here is just a written description of what I said on the previous slide, you can see that we have more than the usual number of observations around zero. And we have some large, large values, relatively large values for some of the residuals on the tails. To summarize linear regression, remember to only use simple linear regression to make a prediction when the underlying relationship between variables is linear. If the underlying relationship is quadratic, if it's exponential or logarithmic, then linear regression is not going to give you the best predictions. You want to identify outliers and investigate why they are unusual. There are different things we can do to try and deal with outliers. It's more of an art than a science, and something that we would have to learn and discuss in a future course, to learn how to deal with these things. For now, you should identify outliers, and understand that they can skew and change the estimation, the predictions that our model makes, they can either make us our model over predict or under predict if the outlier is highly influential. Once you've used the linear regression, you've tried to make a prediction using linear regression. You want to check the residuals and you want to check the residuals for unusual patterns and shapes. We saw some examples of that, that kind of rainbow shape or the little hill and the or the little valley, or waves. And you want to check for changing variability, we would like to see that the residuals have a constant variance or constant variability throughout the data or through through for all the predicted values. And a last note, that's a summary of linear regression. Remember, having a higher R squared is useful, it means that more of the variability in your X axis variable is explaining the Y axis variable. Generally speaking, that means your prediction is going to be stronger. But a high R squared is only helpful when the above conditions have been satisfied. Otherwise, R squared is not going to be an accurate measure of the strength of your vitals predictions. You could have higher squared and still not have accurate predictions. Hello, everyone. I hope you enjoyed this module on correlation and linear regression. Remember, there's a lot of things to keep in mind. But this is a really useful tool in the social sciences. Throughout Business and Economics, for

understanding the big data and the data that we have about the world around us. We can really understand relationships, everything from crimes, to sports, to how the macro economy affects individual households. I'm going to sign off and say goodbye, and I'll look forward to see you and seeing you in our next lecture videos where we jump into Excel and we estimate some linear regression models together.