Applied Econometrics

Prof. Tutan Ahmed
Vinod Gupta School of Management
Indian Institute of Technology-Kharagpur

Lecture - 42
Error Term, Coefficient of Determination, Regression Coefficient (Contd.)

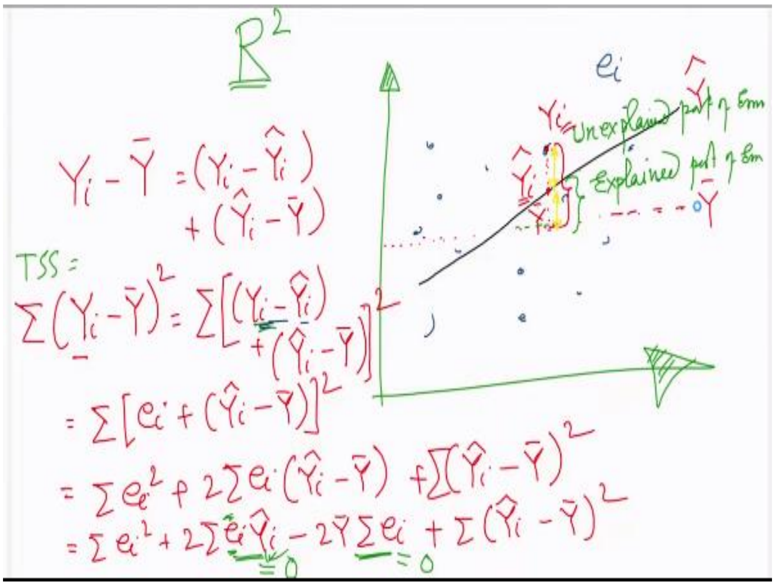
(Refer Slide Time: 00:27)

Concepts Covered

Error Term in a Regression Equation
Coefficient of Determination and Correlation Coefficient
Properties of the Regression Coefficient

Welcome back to the lecture on Applied Econometrics. We are in module II econometric modeling we are talking about and in this lecture we are going to talk about coefficient of determination and its relationship with correlation coefficient. So let us do some hands on here.

(Refer Slide Time: 00:50)



R^2

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})$$

TSS =

$$\sum (Y_i - \bar{Y})^2 = \sum [(Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y})]^2$$

$$= \sum [e_i + (\hat{Y}_i - \bar{Y})]^2$$

$$= \sum e_i^2 + 2\sum e_i (\hat{Y}_i - \bar{Y}) + \sum (\hat{Y}_i - \bar{Y})^2$$

$$= \sum e_i^2 + 2\underbrace{\sum e_i \hat{Y}_i}_{=0} - 2\bar{Y}\underbrace{\sum e_i}_{=0} + \sum (\hat{Y}_i - \bar{Y})^2$$

So when I say, coefficient of determination, we simply remember that coefficient of determination is nothing but the R square. Now what is R square? So let us draw something which is very common, when we actually talk about R square. Let us say that we have all these different data points scattered on this plane. And I need to sort of draw a regression line, okay, and I actually end up drawing a regression line.

And let us say this is my regression line, okay? Now I have, this is my original data point here. These are all my original data points. And let us say I have also drawn a line, which is kind of the average, which is actually the average of all these different points. Let us say these are all my Y's, this is my Y i, and this is my Y bar. So this Y bar is representing the average of all these different Y points.

And this line is actually representing the estimated Y which we usually say Y hat, okay. Now say, I am talking about this point. And I want to measure the error that is accounted for by this model or that is not accounted for by this model. So what is the extent of deviation of Y i from this Y bar. So this if the Y bar line is here, so the extent of deviation is going to be this line.

Now this line, I can actually write this point as Y i hat, because I am talking about this particular Y i. So I write it Y i hat and this is of course my Y bar, right? It is the same Y bar, same value along the line. Now the total deviation from the main line is going to be Y i - Y bar. Now I can write it as, if I just divide it into this two points, let me use a different color.

If I divide it into these two points, so I can write it as Y i – Y i hat, this is one. And then I can add Y i hat – Y bar, right? Now to get the deviation, we take square and then sum it up, that is the total deviation that you get. So I will do the same here. I will just square and sum it up. And here also I will do the same thing. So let us say I will use the third bracket Y i - Y i hat + Y i hat – Y bar and to the whole square, right?

Now for this term, for the first term here, this is this part is actually explained. So this is, my model is able to explain so much of the error and this we call explained part. And this part is called unexplained part of the error, okay? Explain part of error. And this is unexplained part of error. Now the unexplained part of error we sometimes write as either e i or e y, whichever way we want to write.

So let me use the notation of e i here. And I will call this part as unexplained part of the error. So I will just write the equation using that notation. So which means this is going to be my e i plus, let me use a third bracket here instead of first bracket, e i + Y i hat - Y bar whole square. Now if I expand it what I get, I get e i square plus 2 e i Y i hat - Y bar + Y i hat - Y bar and whole square.

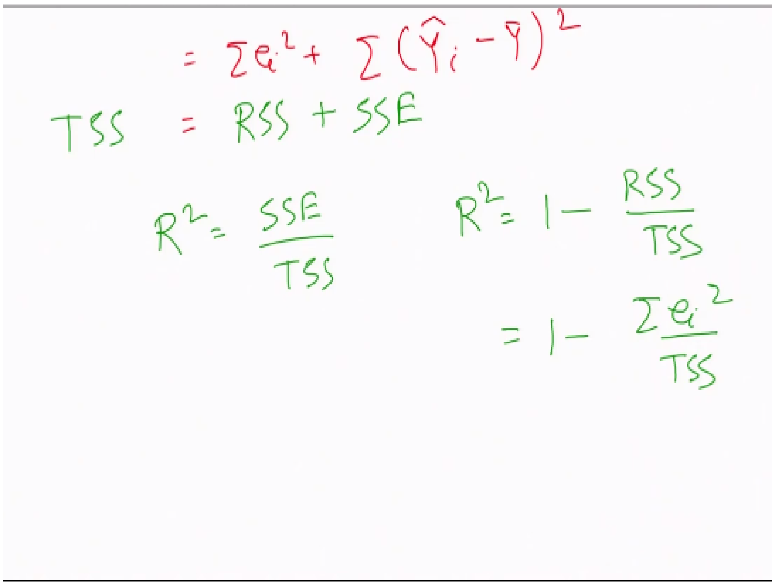
And of course, I will have a summation sign here. Now something very important here. The important part here is that if I expand this, let me expand actually e i square plus 2 summation e i Y i hat minus 2, now the Y bar is constant, so I can bring it out of the summation sign and here I have my e i and then I have this full term Y i hat minus Y bar whole square.

Now let us look at these two terms which are really important here. So the second term, summation of e i, that will be 0, that we know. Because it is the sum total of error term. There are positive errors and negative errors. So when you add them up, it is going to be zero. What about this term? What about this one? This is a very interesting thing that you have to kind of think.

So e i and Y i hat, so basically the e i is the random component. So the moment your model is explaining the error part, the explained error part, so whatever is left out is the unexplained error part. So unexplained error part which is this one, this is the random component of the error. So when I take so if I if you look carefully, the e i and Y i hat essentially if you take the summation of that, it is basically giving you the covariance between e i and Y i hat.

Now for a random variable like e i or a random component e i which can take any value and that ideally should not have any correlation with Y i. So given that logic this has to be equal to zero.

(Refer Slide Time: 06:50)



$$= \sum e_i^2 + \sum (\hat{Y}_i - \bar{Y})^2$$

$$TSS = RSS + SSE$$

$$R^2 = \frac{SSE}{TSS} \qquad R^2 = 1 - \frac{RSS}{TSS}$$

$$= 1 - \frac{\sum e_i^2}{TSS}$$

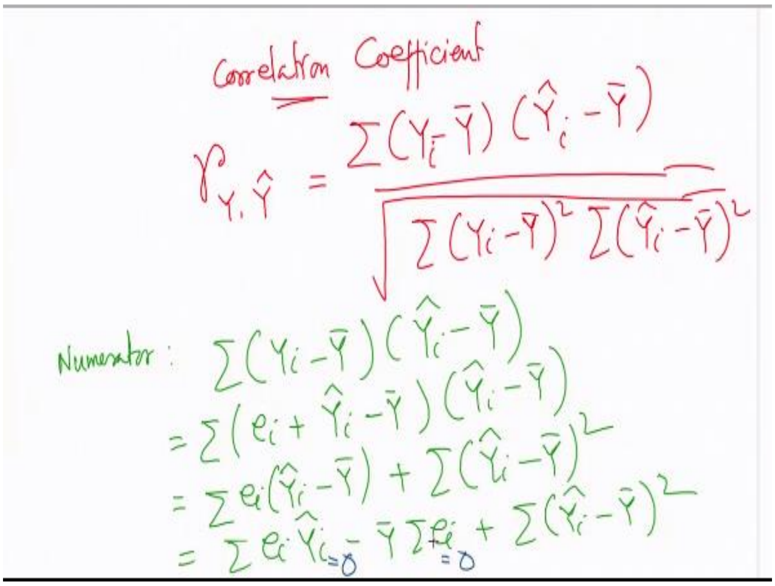
Now if that is the case, then what I will have end of the day is that I will have equal to summation of e i square plus summation of, summation of Y i hat minus Y bar square. Now, so I had this, essentially this is my total sum square. I can actually write the TSS is equal to this. And this part is going to be, this is the unexplained part of the error. So you can write, you know sometimes we write residual sum of square RSS.

And this part we can write explained sum of square, okay. So we can write SSE, is sum of square explained, okay. So that way you can write TSS is equal to RSS and SSE. And when I talk about R square, we know the definition of R square. R square is nothing but the portion of the error which is explained by your model. So essentially that would mean R square equal to is SSE by TSS, okay.

Or you can alternatively write R square is equal to which is 1 minus what is not explained. So which means RSS by TSS. Or you can also write RSS as nothing but summation of e i square by TSS. So this is about the R square term. So we are kind of familiar with R square. The good value of R square is basically saying that my model is good. Low value of R square means my model is not that good.

Now that is about the coefficient of determination. Now what to do with correlation coefficient? So we will just look into that part now.

(Refer Slide Time: 08:31)



Correlation Coefficient

$$r_{Y,\hat{Y}} = \frac{\sum (Y_i-\bar{Y})(\hat{Y}_i-\bar{Y})}{\sqrt{\sum (Y_i-\bar{Y})^2 \sum (\hat{Y}_i-\bar{Y})^2}}$$

Numerator: $\sum (Y_i-\bar{Y})(\hat{Y}_i-\bar{Y})$

$$= \sum (e_i + \hat{Y}_i - \bar{Y})(\hat{Y}_i-\bar{Y})$$

$$= \sum e_i(\hat{Y}_i-\bar{Y}) + \sum (\hat{Y}_i-\bar{Y})^2$$

$$= \underbrace{\sum e_i \hat{Y}_i}_{=0} - \bar{Y} \underbrace{\sum e_i}_{=0} + \sum (\hat{Y}_i-\bar{Y})^2$$

Now correlation coefficient, the way we define it, here what I am going to do is I am going to take correlation coefficient between the actual Y, this is my actual Y and my estimated Y, okay. So essentially what I am trying to see is through the correlation coefficient, I want to see the relationship between Y and Y estimated, okay?

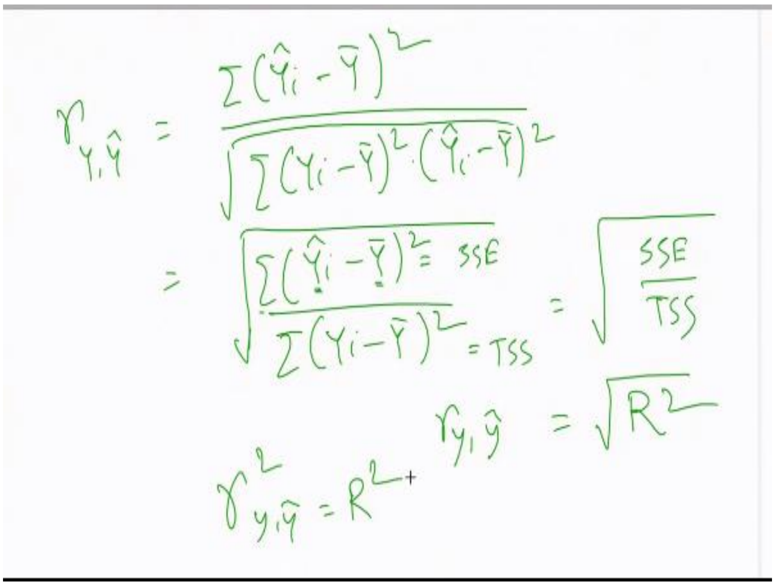
Essentially what I want to see is how good my Y estimated is, you know like performing so far as the estimation of actual Y is concerned. So that is what we are trying to understand. So now if I expand this, what I will write is summation Y - Y bar. So let me write i of course, and Y i hat minus Y bar. So remember the Y bar is nothing but this Y bar we are talking about.

And of course, I will have to have this term in the denominator, summation of Y i - Y bar square the summation of Y i hat - Y bar square, all right. So that is by definition. That is coming from the definition of correlation coefficient. Now let us work with the numerator first, let us work with the numerator first. So the numerator. The numerator goes like this. It is equal to summation Y i - Y bar into Y i hat - Y bar.

Now what I am going to do? I am going to use the same concept, same notations that I used here. I am going to use this term as e, okay? And I am going to divide this Y i - Y bar as e i + Y i hat - Y bar, okay? So let us just do that. So that means this first part is going to be this. It is nothing but my e i + Y i hat - Y bar. And then it is multiplied by Y i hat - Y bar, right?

Now if I actually work on this, it will be like summation e i Y i hat minus Y bar plus summation Y i hat minus Y bar whole square. Now if I further expand, what I will have is e i Y i hat minus, again I will take Y i out e i plus summation Y i minus Y bar square. Now going from the, you know the derivation we have done previously, going by the same logic, this one is going to be 0, and this one is also going to be 0.

(Refer Slide Time: 11:40)



$$r_{Y,\hat{Y}} = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sqrt{\sum(Y_i - \bar{Y})^2 \cdot (\hat{Y}_i - \bar{Y})^2}}$$

$$= \sqrt{\frac{\sum(\hat{Y}_i - \bar{Y})^2 = SSE}{\sum(Y_i - \bar{Y})^2 = TSS}} = \sqrt{\frac{SSE}{TSS}}$$

$$r_{y,\hat{y}}^2 = R^2 + \qquad r_{y,\hat{y}} = \sqrt{R^2}$$

So essentially, my numerator is going to be, I will use the next page. My numerator is going to be summation of Y i hat - Y bar whole square. That is my numerator, right? And my denominator, so this is the numerator. It has, I have got that and my denominator is as it is. So let me write down the denominator. And that is basically summation of Y i - Y bar whole square into Y i hat – Y bar whole square.

And that is nothing but my r Y Y hat, okay. Now if I do the further cancelling out between numerator and denominator what I will be left with is this. So big square root and what I will have is I will have summation of Y i hat Y bar square by summation of Y i minus Y bar square okay. So that is what we will have.

Now if you look at this numerator and denominator carefully, the numerator is nothing but it is saying to the extent your model is able to explain the error from the Y bar. And whereas, this Y i minus Y bar is the extent of error your actual data, I mean your actual data is generating. So basically this is the total sum of square and this is the sum of square explained, okay. So SSE by TSS.

So this is nothing but SSE by TSS, which is my R square, okay. So what I get is the correlation coefficient between Y and Y hat is nothing but R square. Or I can write r square Y, Y hat is equal to capital R square. Now that is the mathematical part, we have derived that. But how do I really explain the how we have just come, how we have just derived right now.

So the expansion will go like this, the correlation coefficient between Y and Y hat is nothing but how well your Y hat is able to explain your Y, okay? So the moment, essentially what you are seeing is the how they are covarying. So if they are covarying together, essentially they are explaining, the Y hat is explaining the Y, okay? Whereas the R square is also doing the same thing.

It is how well your the line is actually able to explain the variation. So essentially, these two are the same thing. So that is basically what you need to remember. So one thing that we need to keep in mind whenever we learn new concepts, how we can actually relate this concept. So this is how, if you ever get confused between correlation coefficient and coefficient of determination, so this is how we will remember the concepts.