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

Lecture - 41
Error Term, Coefficient of Determination, Regression Coefficient

Hello and welcome back to the lecture on Applied Econometrics. We are in module II on econometric modeling. And the particular topics that we are going to cover in this lecture are error term, coefficient of determination and regression coefficient. Now error term is something that we are going to deal with pretty extensively going forward.

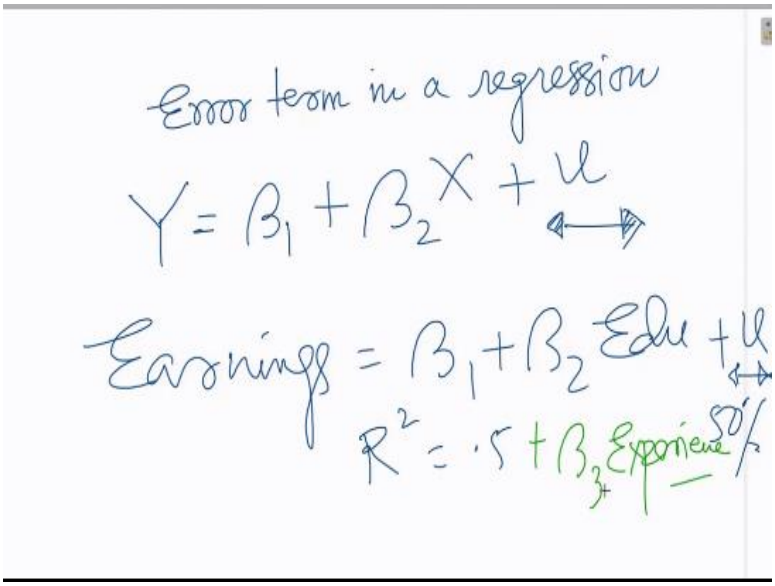
And this the introduction where we are going to talk about where the error is coming from in the regression equation, and how we are going to deal with it. Of course, the dealing part will come later. But mostly we will talk about where from it is coming. The second topic that we are going to talk about is coefficient of determination. We are going to see how coefficient of determination is actually related with correlation coefficient.

We will try to sort of prove the relationship mathematically. And then we will talk about the properties of regression coefficient.

(Refer Slide Time: 01:12)

So let us start with the error term in regression equation. Now when we say error in a regression equation, we write a regression equation; in general, we write something like this.

(Refer Slide Time: 01:22)



So in general, a regression equation would look like this. So if my Y is dependent variable, and let us say I have a beta 1 coefficient, beta 2, x and there is an error term, right? Now this error term is something that is very important for a regression equation. We need to find ways to sort of identify where the error is coming from. And of course, we build up different tools and techniques to reduce the error term.

Now the error term, if in general, to get an idea about how important this error term is, when for example, we run a regression, say we usually use a regression on earnings, say from education and some error terms. So if we run, this is a kind of very common question that we use. And we use other explanatory variables as well.

If we have the R square value for this model, around 0.5, or you know even less than 0.5, we consider that model to be a good model. So that means 50% of the variation in the model is not, a variation in the data is not explained by the model, which is accounted for by the error term. But even then, the 50% of error, we consider this model to be a reasonably good model.

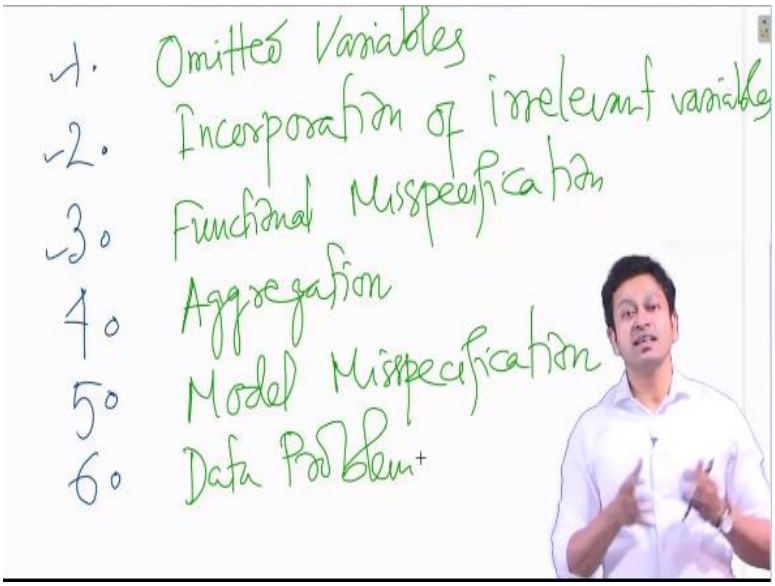
So that means the error term is that big in your regression equation. So now we need to understand where this error term is coming from. Let us say in your regression

equation, the previous regression equation, where we spoke about earning as an outcome of education, let us say you also introduce some other external variable, let us say, I will use a different color. Let us say you use for example, experience, okay?

Now if we run a regression actually we will see that your regression equation is actually the R square value is improving, because the experience is something that accounts for your earnings. But there are so many other variables, which may actually influence earnings, but you will not have the way to capture them. For example, it could be my talent.

My, you know there are so many talents I have, which are influencing my earning, and that cannot be captured by education or by experience, right? There could be other variables like my family background, my social network. So all these things, which matters in terms of getting a job, right? So all these things that cannot really be accounted for in this equation.

(Refer Slide Time: 04:01)



So these variables, which we cannot account for are called the problem because of that is or omission of relevant variables, let us say or omitted variable basically, omitted variables. So omitted variables, this is a problem. So when we actually omit relevant variables. Now another problem could be when in a model, we actually end up adding more and more variables, and the model looks robust, but it actually is not.

So many variables, which are not relevant to the model, but still, you end up incorporating. So that will also create this error term. So let us say incorporation of irrelevant variables. So this is another source of error. Third, the third source of error could be a functional misspecification. So what I mean by functional misspecification is that the way these variables are related.

Here I have kind of assumed that education, experience all are linearly related with earnings. But there could be some other forms of these variables which might actually influence, which might actually have a more direct relationship with earning. For example, I might have a variable called experience square. I can actually take experience square and I will actually see that is the standard notation.

We will actually see that that is influencing earning even better, right? We in economics, we often use a Cobb-Douglas production function. So there again we will see when we will talk about model specification, how the functional specification is actually determining the robustness of the model. So that could be another issue or another source of error.

The other source of error is that we are actually when we were actually learning the regression equation we are actually aggregating all the different individuals. And they have their different you know reasons and you know like different abilities, which are influencing their earning potential. Now the aggregation always omits all the, you know like nitty gritties.

And because of the aggregation we actually end up getting some error. So aggregation is another reason why we actually have this, we get this errors. Now all these different items I have mentioned here, we can actually sum it up in something called model misspecification. Now in model misspecification, you have all these different reasons that can actually come under the umbrella.

Model misspecification can also be caused because of the assumptions that we have made. For example, in a time series data, suppose what you have done is, you actually have taken say you have forgotten to take into account the past period influence, okay.

And instead of that you have just regressed y with all the x variables. So in that case, your model is misspecified because you have missed the structure here.

So that could be another reason where you can get the errors. Another very critical problem for social scientists, people who collect data from the field is the quality of data that you get, or the kind of variables that you can collect data on. So I will tell you a story. Let me just actually write down. So let us say, the data problem. So what happens here is that suppose you want to collect data on certain indicator, for example, let us say your income, okay?

Or let us say the question we discussed about ability. So how do you collect data on ability? So that is a really difficult thing to actually collect. But one practical problem that, for example, in India national sample survey data, when they collect the consumption data as a proxy to income, so it often creates problem in the sense that you know what time you are actually collecting data,? Who is your respondent?

So I will tell you a story about of data collection, you know experience. It is a long way back, and I think, two decades, three decades back. And that time, data was being collected in Calcutta, where people were, you know like, the surveyors used to go to the home, in the houses in the morning. And morning, usually, you have all the male people who are in who are still in home. They have not gone to their work yet.

So they were actually interviewing the male candidates. There was an apprehension about the quality of the data. So what was done is that on the same households, again after a couple of months, the survey was conducted. And this time it was in the afternoon. And there is a remarkable difference in terms of the finding that people got.

It actually, what we saw is that the earnings of the households in the morning from the morning data, we found that is actually very low, or relatively lower than the earnings that you found from the afternoon data. Now why is that? So that is very tricky problem. And I would actually want you to pause the video for a moment and think about it. Now, but I will give you the answer.

The reason is, what was found out is that in the morning, it was the male candidates in the household, they were, you know reporting. And they had for some reason, they were actually undermining their earnings, okay. Whereas in the afternoon, it was the female candidates in the household, they were reporting and they had, for some reason they were actually inflating or actually giving the, you know true picture of whatever assumptions they make.

So even this small, nuanced, you know reasons can actually make a lot of differences in the data quality. So data this, because of the quality of data, we can often have this error that we see in the regression equation.