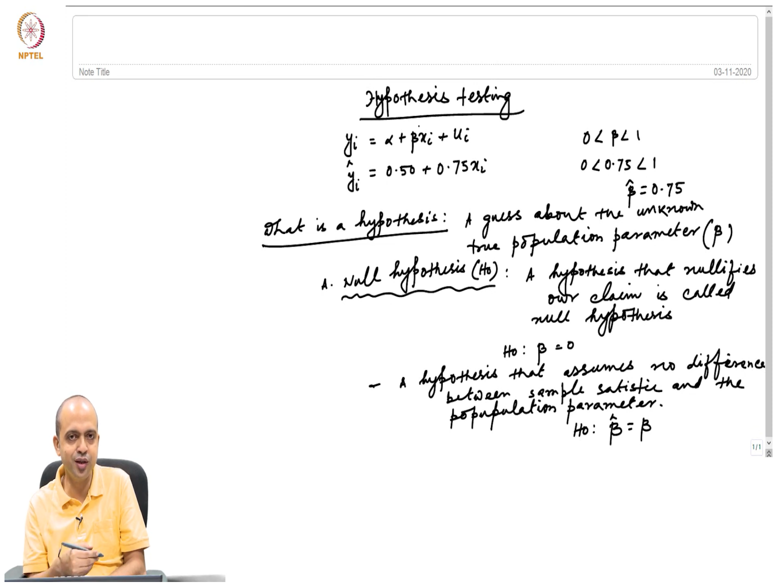
Introduction to Econometrics
Professor Sabuj Kumar Mandal
Department of Humanities and Social Sciences
Indian Institute of Technology Madras
Goodness of fit measure, ANOVA and Hypothesis Testing Part 4

(Refer Slide Time: 00:14)



Welcome, so in our last class we were discussing about how to estimate a model using our statistical software Stata right? And we learnt how to estimate the model and then how to interpret the coefficients. We have also learnt about the goodness of fit measure and we also discussed about our ANOVA table.

So, that means when your objective is to explain the variation in your dependent variable with the help of the independent ones, how will you estimate the coefficients and interpret those coefficients and how to check the goodness of fit.

Now, today what we will do? As we said yesterday in our last class, today we will discuss about hypothesis testing which is very very important. So, please be very sincere and careful. Sincerely you have to listen to this and understand right, what is hypothesis testing part. So, this is hypothesis testing that we are going to discuss.

So, we will discuss this with the help of the same example. Let us say that this is our model, $Y_i = \alpha + \beta x_i + u_i$ and when you estimate the model, your estimated model is $\hat{Y}_i = 0.50 + 0.75 x_i$. Now, from this model you can see that 0.75 is actually less than 1 but greater than 0 which we were expecting from the Keynesian consumption function hypothesis.

Now, why hypothesis testing is required? As we said, we want to ensure that this value of $\hat{\beta}$ (0.75), we are getting not by a chance factor but we would like to ensure that there is a statistical mechanism that will ensure that even if you collect 100 samples from the same population, more than 90 percent of the cases you will get a beta hat value like this.

So, that means it is not by a chance factor. In that context, how will you ensure the statistical significance? That means so far what we got this 0.758 is only mathematical significance. Mathematically, this 0.75 is obviously different from 0. But how to check the statistical significance? And that is the context where hypothesis testing is actually required.

So, what is a hypothesis first of all? Hypothesis is actually a guess about the true population parameter. So, in this context, what is a true population parameter? α and β, let us say β here, about the true population parameter β. So, what is our guess? We are guessing that $\hat{\beta}$ is actually greater than 0. That means income has significant impact on consumption.

That is the guess we are making about the true population parameter. Guess about the unknown true population parameter. Since it is unknown, we are making this case. Now, there are two types of hypothesis. One is called null hypothesis denotes by H knot. What is null hypothesis? Null hypothesis, the definition says that a hypothesis that nullifies our claim is called null hypothesis. So, from the name itself, you are getting this.

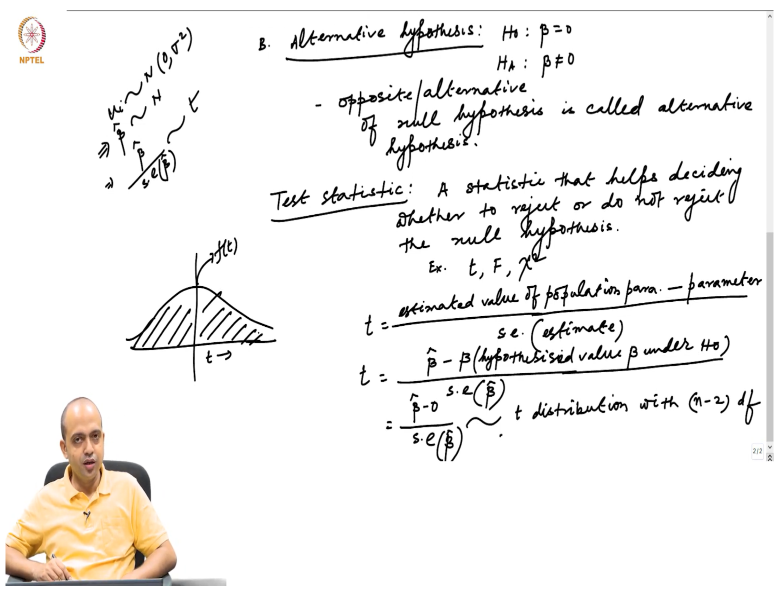
Null hypothesis is a hypothesis that nullifies our claim. Now, in this context, what is our claim? Our claim is that income has significant impact on consumption. That means if you nullify, what will be the hypothesis? Income does not have any significant impact on consumption. Then how will you make your null hypothesis? In terms of this notation? We will say that beta equals to 0. That is one hypothesis.

So, your claim is beta greater than 0, that means income has significant impact on consumption. And then you nullify and you will get your null hypothesis which says H_0 is β equals to 0. This is one way of understanding null hypothesis or defining null hypothesis. There is another way of defining null hypothesis which says that a hypothesis that assumes no difference between sample statistic and the population parameter.

That means how will you define your null? $\hat{\beta}$ equals to β, that is no difference. Now, this way of defining null hypothesis is little difficult to understand. So, I will make you understand the second definition of null hypothesis in a later part of our discussion. For the time being, you just understand a null hypothesis is a hypothesis which nullifies our claim.

You try to identify your claim, that my claim is income has significant impact on consumption, then you nullify. If you nullify, then what will happen? You will say that income does not have any impact on consumption. That means β equals to 0. If β equals to 0, then x does not have any impact on y, very simple.

(Refer Slide Time: 09:41)



And then, the other type of hypothesis is called alternative hypothesis. So, second type of hypothesis is alternative hypothesis. So, when your null is β equals to 0, so alternative is denoted as Ha which says β not equals to 0. So, that means opposite or alternative of null hypothesis is called alternative hypothesis. This is what is hypothesis.

Now, in hypothesis testing, we need to understand certain concepts that will help us understanding hypothesis testing procedure. First of all, the concept that we would like to discuss is test statistic. What is the definition? A statistic that helps deciding whether to reject or do not reject the null hypothesis. Now, why we are rejecting or do not rejecting the null? Because you see, when I say H_0 beta equals to 0, if you reject the null hypothesis, then only you can establish your claim, is it not?

So, that is why our objective here in this hypothesis testing is either to reject or do not reject the null hypothesis. If we can reject the null hypothesis, then we can establish our claim, when beta equals to 0 is rejected, that means we are saying that income has impact on consumption. If we are not able to reject our null hypothesis, then we will say that income does not have any impact on consumption. And how are you going to decide that? With the help of test statistic.

So, the value of the test statistic will help us deciding either to reject or do not reject the null. For example, there are several examples are there, like t-statistic, then f-statistic, chi square statistic which you might have studied in your statistics course. Here, when our objective is to check the significance of a particular variable, we will use the t-statistic and how the t-statistic is defined? T-statistic is defined in this way.

The definition is - estimated value of population parameter minus the parameter divided by standard error of the estimate. That means in terms of notation, estimated value of the population parameter is beta hat minus, what the parameter, that means, beta divided by standard error of the estimate, that means standard error of beta hat, beta hat minus beta divided by standard error of beta hat. This is how we define our t-statistic.

Now, from this definition of the test statistic, we can understand the alternative definition of null hypothesis also. What is a null hypothesis we say, the alternative definition says? That there is no difference between the sample statistic and the population parameter. Now, here how we are constructing? The numerator is estimated value of the population parameter which is nothing but the sample statistic beta hat minus the population parameter beta.

So, since your null is saying that there is no difference and here our numerator is constructed based on the difference between $\hat{\beta}$ and β, that means higher the difference between these two, larger would be the value of t and more information we would be able to gather to say that, yes actually there is a significance difference between population parameter and the sample statistic.

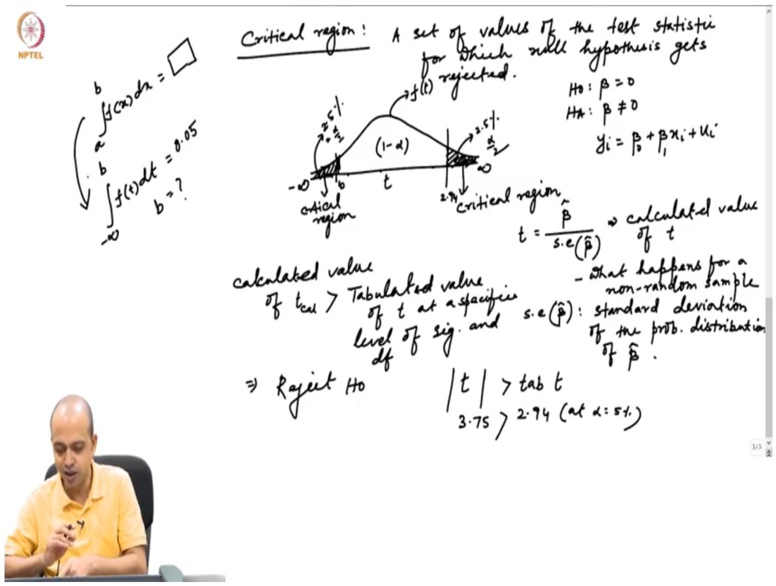
Now, the population parameter is unknown, that is why here we will say that instead of β, we will say that the hypothesized value of β under H_0. So, the original definition is β- β, but then β is unknown population parameter. Unless you put some value for β, you cannot get a value for t statistic. That is why, in place of β, what we will do, hypothesized value of β under the null.

And what is the value of hypothesized value of β under the null? That is 0. That is why t statistic is defined as simply ($\hat{\beta}$ -0)/s.e. $\hat{\beta}$, or simply $\hat{\beta}$ /s.e. $\hat{\beta}$. And this follows t distribution with n minus 2 degrees of freedom. This follow a t distribution with n minus 2 degrees of freedom. Now, how does this t distribution look like? t distribution looks like a normal distribution.

Why this is so? Because we assume u_i follows a normal distribution and when u_i follows a normal distribution, let us say, 0 mean and sigma square variance, that ensures that $\hat{\beta}$ also follows a normal distribution. And when $\hat{\beta}$ follows a normal distribution obviously, standard error of $\hat{\beta}$ will also follow a distribution which is similar to normal but with little flatter tail because we are dividing that with standard error of $\hat{\beta}$.

So, u_i follows a normal distribution. That is why β is also following a normal distribution. So, when we divide the β by standard error of $\hat{\beta}$, the result and distribution which is actually t, becomes little flatter than the normal distribution. But it is a normal distribution which has the properties of the normal distribution, basically symmetric around this mean value. This is how the test statistic is defined.

(Refer Slide Time: 19:37)



Now, the next concept that we need to know is critical region. How we define this? A set of values of test statistic for which null hypothesis get rejected. Now, I will explain this critical region using the diagram. This is your t distribution and critical region, that means there are two areas in the left side as well as in the right side of the distribution. So, in this diagram, what we are doing, this is the density and here, we are putting several values of the test statistic t.

And the critical region lies in the two tails of the distribution. For the time being, let us say that the size of this critical region is 5 percent, that means here it is 2.5 percent, here also it is 2.5 percent. This is basically called alpha by 2 and this is also alpha by 2. So, that means this region is basically 1 minus alpha. So, a set of values of t for which null hypothesis get rejected is called critical region. And these two are actually the critical region.

This is the critical region, this is also critical region. That means, what I am saying that, now why the critical region lies in both the sides of the distribution? Because your null hypothesis is beta equals to 0 and alternative hypothesis is beta actually not equal to 0. Since alternative is beta not equals to 0, that is why it can be either on the left side or on the right side of the distribution.

That is why when your alternative hypothesis is a composite one, like not equals to 0, your critical regions will lie on both sides of this. And since the t distribution is symmetric, the area is actually the same when you are considering the right side as well as in the left side. So, that means it is $\alpha/2$, this is also $\alpha/2$. This alpha is actually, is little different from the α what we have used in our econometric model, this is not the intercept, this is different.

That is why you may change your model like $Y_i = \beta_0 + \beta_1 x_i + u_i$. So, now instead of alpha I have used the beta 0. So, this alpha is actually different. Now, how will you decide whether to use reject or not reject? That means you can calculate your test statistic immediately after estimating your model, standard error of $\hat{\beta}$. So, that will give you, this is called calculated value of t.

Now, look at the definition of t, it is $t = \frac{\hat{\beta}}{s.e.\hat{\beta}}$ and that is why you are going to get a distribution. Now, what will happen if our sample is not random? If our sample is not random, that means what will happen? What happens for a non-random sample? Now, if you get a non-random sample from the population you can get one such non-random sample.

And if you do not get multiple samples from the population, if there is no randomness in the sample, sampling part, then what will happen? You will get only your $\hat{\beta}$. What is the standard error of $\hat{\beta}$? Standard error of beta hat is basically the standard deviation of the probability distribution of $\hat{\beta}$. So, if I define standard error of $\hat{\beta}$ is basically standard deviation of the probability distribution of $\hat{\beta}$. So, in a non-random sample, there is only one such sample you can collect. You will get only one $\hat{\beta}$.

So, distribution of $\hat{\beta}$ is not a possibility when you are going for non-random sampling. That means you can estimate your $\hat{\beta}$ with that non random sampling. But you cannot go for hypothesis testing, because hypothesis testing requires test statistic to be computed which again requires standard error of $\hat{\beta}$ to be computed, which is nothing but the standard deviation of the probability distribution of $\hat{\beta}$ which we can get only when the sample is random.

But that does not mean that I will be collecting actually so many samples from the given population, but that randomness should be there, so that, by the assumption distributional assumption of the error term, we can ensure that a distribution is possible to get for the estimated parameter $\hat{\beta}$ and standard error of $\hat{\beta}$ is also possible to get, right?

Now, if your calculated value of the t statistic so defined is greater than tabulated value of t at a specific level of significance and degrees of freedom, then we will say that reject your null. Now, this calculated value can be either positive or negative depending on the sign of your $\hat{\beta}$. But our decision making rule is always same, which is calculated value of the t is the tabulated value of t at a specific level of significance and degrees of freedom. That is why I will say that this is t modulus is greater than tabulated t. Now, this tabulated value of t, if you go to your statistics book, you will get a table for t.

That means, that table will tell you at a specific level of significance, what is level of significance? I will explain. For the time being, you just keep in mind that at a specific level of significance and degrees of freedom, that mode value of t which is calculated by beta hat, in this formula should be greater than the tabulated. If you go to that table, t table, if you specify your level of significance and degrees of freedom, you will get a value.

For example, here let us say this tabulated value is 2.94 when your alpha is 5 percent. And your tabulated value is let us say 3.75. Let us say this is 3.75 and this is 2.94 at alpha equals to 5 percent. That means I am saying the size of the critical region is 5 percent.

Since 3.75 is greater than 2.94, that means you are actually entering into this critical region, either this side or this side, depending on the side, but your decision making rule would be the same. Now, the question is, how will you get this value? 2.94? How do you generate the calculated, tabulate value of t?

How do the statisticians prepare this type of table? The idea is very simple. They have actually applied the concept of definite integral. Here, what we are happening, if you recall, in our definite integral, generally we used to get two limits and then the function fx dx and we used to get a specific value. If you know fx, if you know a and b, you will get a value.

Now, following the area property of the definite integral, what we are doing here is the reverse integration. That means you know that the range of this is minus infinity to plus infinity. So, I am specifying a problem like this, minus infinity to, let us say this is b and this fx or let us say ft dt equals to 5 percent that means 0.05.

So, what I have to determine here? What is b. So, you know the lower limit of the integral, the upper limit you need to decide b, so that this area becomes 5 percent or 0.05. That is the idea is followed by the statistician and you get several values of t the moment you specify the area. Instead of 5 percent, you can make it 10 percent, you can make it 1 percent, you can make it 4 percent and immediately, you will get a value if you do the procedure of reverse integration.

This is the idea how you, the statistician, they have prepared the table and the decision making rule says that your calculated value of the t which is defined by $\frac{\hat{\beta}}{s.e.\hat{\beta}}$. Then you have to compare that with the tabulated value and immediately, you will see whether that is greater than the tabulated one or not.