

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 1

(Refer Slide Time: 00:14)

So about the goodness of fit measure, suppose this is your X and this is your Y and this is basically your Y bar and this is the line that you have fitted, this is the SRF which is $Y_i = \alpha + \beta X_i$. Now, you take any value of X, let us say this is your X value, so obviously, you will get an observed Y that is called Y_i and this is X_i . Your line says that this should be your predicted consumption which let us say that I am denoting by \hat{Y}_i . Now, if you recall what I said, just before, sometimes back, that the entire purpose of econometric analysis is to explain the variation in Y and that variation in $Y_i - \bar{Y}$ which is around its mean value.

If you take summation and square it up, then it is known as total sum of square. I will write here.

So, $\sum_{i=1}^n (Y_i - \bar{Y})^2$ is called total sum of square or TSS. Now, out of this total variation in Y which

is given by $\sum (Y_i - \bar{Y})^2$, how much your model is able to explain? Your model is able to explain

only this part which is $\hat{Y}_i - \bar{Y}$. How will you denote that? This is $\hat{Y}_i - \bar{Y}$. Let us again take the

summation and square it up. That means $\sum (\hat{Y}_i - \bar{Y})^2$ is actually called Explained Sum of Square or ESS because that is what you have explained. And then, this is the portion that you are not able to explain which is called $Y_i - \hat{Y}_i$. Take square of $Y_i - \hat{Y}_i$ and that is the portion that you

are not able to explain. So, this is, $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$ is called Residual Sum of Square. So, that means this much, you are not able to explain-Residual sum of square or RSS.

That means from this we got TSS, we got ESS, we got RSS and the diagram easily shows the TSS. This entire portion $Y_i - \bar{Y}$ is actually a summation of two components. So, from the diagram, I can easily write that TSS is equal to ESS plus RSS. This is one important relationship. Now, another concept, our major focus was on goodness of fit. Let us now define goodness of fit.

Our total variation was $\sum (Y_i - \bar{Y})^2$. Out of this total variation I am able to explain $\hat{Y}_i - \bar{Y}$.

Now, if I take a percentage of total variation in Y_i , that is explained by the model which is $\hat{Y}_i - \bar{Y}$, a proportion of this with respect to the total variation, that is a proportion of ESS with respect to TSS is actually the goodness of fit.

Let us take a proportion ESS by TSS. Out of total variation, how much your model is going to explain? This proportion is known as R^2 or goodness of fit measure. That means R^2 basically says, so if the R^2 is let us say 0.5090 in our example it means 50.90 percent of total variation in Y_i in consumption is actually explained by your model. That is the meaning of R^2 . That means ESS by TSS ratio is 0.5090. The ratio of ESS and TSS out of the total variation in Y_i which is $(Y_i - \bar{Y})^2$, how much your model is going to explain, that proportion is known as R^2 . It means that my model is going to explain 50.90 percent of total variation in Y_i around its mean value. So now we have learnt the three sum of squares - TSS, ESS and RSS. And we will now go back to Stata and we will see how Stata has reported those.

(Refer Slide Time: 08:50)

The screenshot shows the Stata command window with the following output:

```

reg consumption income

Source      SS          df           MS       Number of obs   =    10
Model      852.72727      1      852.72727   Prob > F         =  0.0000
Residual   337.27272      8      42.1590909   R-squared        =  0.9621
Total      1190.00000     9      132.2222222   Adj R-squared    =  0.9573
           8890     9  987.7777778   Root MSE        =  6.493

consumption   Coef.   Std. Err.   t    P>|t|   [95% Conf. Interval]
income        -5090909   .8357428   34.24   0.000   -4266678   -5915134
_cons         24.45455   6.413617   3.81   0.005   9.464256   39.24483
    
```

Below the command window, a man in a blue and white checkered shirt is visible, sitting at a desk.

Now, see, look at here, they are source, this table, this particular table, here you see the R^2 , sorry R^2 is not 0.5090, I have written wrong.

(Refer Slide Time: 09:06)

The slide contains handwritten notes and a graph titled "Goodness of fit".

Equations:

- $TSS = ESS + RSS$
- $\frac{ESS}{TSS} = R^2 = 0.9621$
- $0 \leq R^2 \leq 1$
- Case 1: $R^2 = 1, \hat{Y}_i = Y_i$
- Case 2: $R^2 = 0, \hat{Y}_i = \bar{Y}$

Graph: A scatter plot with a regression line. The y-axis is labeled 'Y' and the x-axis is labeled 'X'. The regression line is labeled $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$. The total variance is shown as $\sum_{i=1}^n (Y_i - \bar{Y})^2$. The explained variance is $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$. The residual variance is $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$.

Definitions:

- Total Sum of Squares (TSS): $\sum_{i=1}^n (Y_i - \bar{Y})^2$
- Explained Sum of Squares (ESS): $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$
- Residual Sum of Squares (RSS): $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$

Below the graph, the same man in a blue and white checkered shirt is visible, sitting at a desk.

This R^2 is actually 0.9621. So, that means 96.21 variation in Y is actually explained by your model which is really good. And you have to keep in mind that R^2 lies between 0 and 1. So, when R^2 is actually 1 that means we can write these two cases. Case 1, let us say R^2 equals to 1. That means your \hat{Y}_i is actually your Y_i . So, this point is actually whatever your model predicted and whatever your actual Y is the same. So, that means your model can perfectly predict sum 1

consumption. That is a theoretical possibility. So, it is not possible to estimate a model for which \hat{Y}_i is equal to Y_i . There is no deviation between the predicted and observed. It is a theoretical possibility.

So, when R^2 equals to 1 means - \hat{Y}_i is equal to Y_i . Similarly, what does then R^2 equals to 0 mean? When R^2 equals to 0, that means your model is not able to explain anything. Now, in terms of the diagram, that means there is no explained sum of square. You are not able to explain anything. That is why it is 0. So obviously your like will then tend towards downwards and it will converge to the average line. So, that means your \hat{Y}_i is equal to \bar{Y} . So, that means this regression line will tend downward and it will converge to the average line. So, that means the average value of Y_i is the best prediction for this model since I do not have any explanatory power. Anyway, these are the two extreme possible cases, theoretical possibility. In reality, we do not get R^2 equal to 0 or 1, rather R^2 will always lie between 0 and 1.

(Refer Slide Time: 12:00)

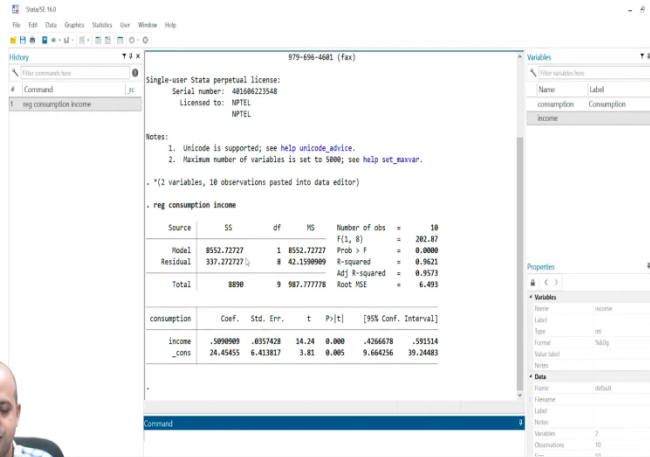
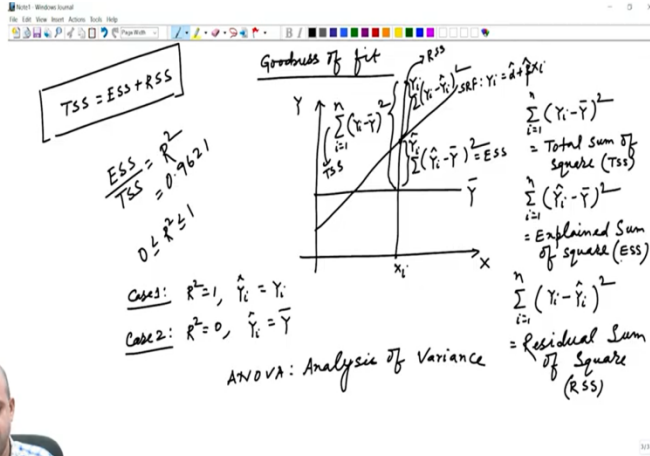
The screenshot displays the Stata command window and results for the regression model `reg consumption income`. The ANOVA table is as follows:

Source	SS	df	MS	Number of obs = 10	F(1, 8) = 262.87
Model	8532.72727	1	8532.72727		Prob > F = 0.0000
Residual	337.272727	8	42.1590909		R-squared = 0.9621
Total	8870	9	987.777778		Adj R-squared = 0.9573
					Stock MSE = 6.493

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
consumption					
income	-.0090909	.0357428	14.24	0.000	-.4266678 -.581514
_cons	24.45455	6.413817	3.81	0.005	9.664256 39.24483

Now, here this table if you look, we have mentioned source and then there are two models residual and total and then you see this particular table is called ANOVA table. ANOVA means analysis of variance.

(Refer Slide Time: 12:23)



So, in ANOVA table, this TSS, ESS and RSS what Stata is giving, is called ANOVA table. A-N-O-V-A this is called Analysis of Variance. See, this diagram is also a diagrammatic representation of ANOVA only. This my total sum of square, total variation in Y. Out of this, I am able to explain this. That is why this is called my ESS, and this is called RSS and this is TSS. So, this diagram also shows the analysis of variance only or ANOVA. Now, what is ESS here? The term what we have written ESS in Stata language it is called Model Sum of Square. So, ESS in Stata is known as model sum of square. So, your model is able to explain how much and the ESS model sum or explain sum of square value is 8552.77. 8552.77 is the model sum of square.

And what is the residual sum of square? Residual sum of square is 33.27 and if you add these two, then the total sum of square is 8890. So, this is called analysis of variance which we have understood using the simple diagram. Then, they have also given degrees of freedom.

(Refer Slide Time: 14:33)

The slide contains the following handwritten notes and diagrams:

- Goodness of fit** diagram showing a regression line $\hat{Y}_i = \hat{\alpha} + \hat{\beta}x_i$ and the regression equation $Y_i = \alpha + \beta x_i$. The total sum of squares is $\sum_{i=1}^n (Y_i - \bar{Y})^2 = TSS$, the explained sum of squares is $\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 = ESS$, and the residual sum of squares is $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = RSS$.
- Formulas: $TSS = ESS + RSS$, $ESS = R^2 \cdot TSS = 0.9621$, and $0 \leq R^2 \leq 1$.
- Cases: Case 1: $R^2 = 1, \hat{Y}_i = Y_i$; Case 2: $R^2 = 0, \hat{Y}_i = \bar{Y}$.
- ANOVA: Analysis of Variance
- Degrees of freedom: $\sum_{i=1}^n (Y_i - \bar{Y})^2 \rightarrow$ Total number of observations - No. of linear restriction $df = (n-1)$

Now, how do you know the degrees of freedom? For any particular measure, for example, let us

say I am trying to understand the degrees of freedom for this $\sum_{i=1}^n (Y_i - \bar{Y})^2$. I am trying to understand the degrees of freedom here and how do you define degrees of freedom? Total number of observations minus number of linear restrictions. That is how the degrees of freedom concept is defined.

See, here, what is my total number of observations? Look, it is i running from 1 to n . That means I have n number of observations. And out of n number of observations, how many restrictions I have put? I have given only one restriction in terms of \bar{Y} . Then degrees of freedom for this is defined as n minus 1. Why this is so?

That means 1 observation is lost due the restriction imposed, that is why n minus 1 number of observations are able to move freely to define this particular measure. That is the concept of degrees of freedom, to define this particular measure. What is degrees of freedom? For a particular measure, it is defined as total number of observations minus number of linear

restriction. Here, total number of observation is n and number of restriction is 1, that is why it is degrees of freedom is n minus 1.

(Refer Slide Time: 17:22)

The slide contains the following handwritten notes:

- df for RSS $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$
- to compute RSS, it is necessary to compute \hat{Y}_i
- to compute \hat{Y}_i , we need to compute $(\hat{\alpha}, \hat{\beta})$
- df for RSS = $(n-2)$
- $$Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + u_i$$
- $$RSS = (n-k)$$
- df for ESS = $\frac{df TSS - df RSS}{(n-1) - (n-2)} = \frac{(n-1) - (n-k)}{1} = k$

The video inset shows a man in a blue and white checkered shirt speaking.

Similarly, if you try to understand the degrees of freedom for RSS, now RSS how we have

defined? It is defined as $\sum_{i=1}^n (Y_i - \hat{Y}_i)^2$. That means from here what I can understand is to compute

RSS, it is necessary to compute \hat{Y}_i . And to compute \hat{Y}_i , we need to compute $\hat{\alpha}, \hat{\beta}$ right? That

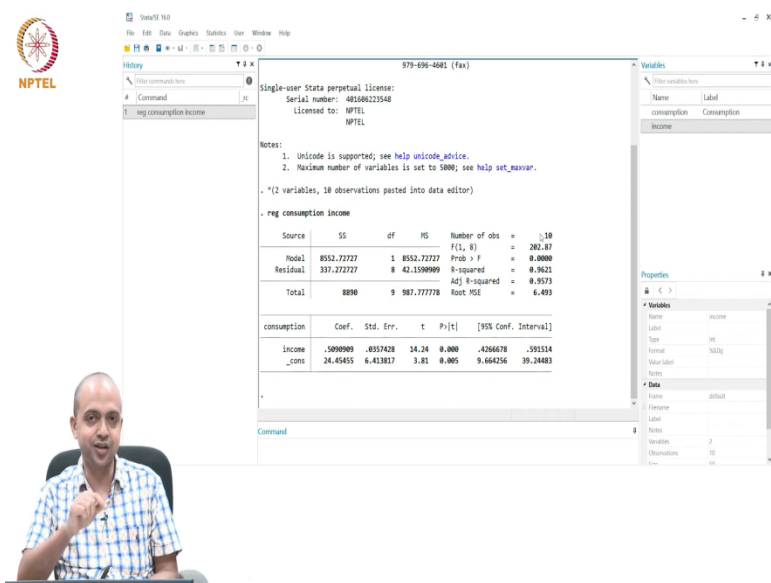
means if you have n number of total observations, while estimating $\hat{\alpha}, \hat{\beta}$, you will lose two observations and that is why degrees of freedom for RSS would be n minus 2 where 2 is the total number of parameters to be estimated. Now, in this model you have only two parameters.

That means you have only one explanatory variable, you have two parameters to be estimated.

For a generalized model, let us say I am doing $Y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} + u_i$. Here, I am trying to estimate actually k number of parameters. So, that is the reason I would say that when you have k number of parameters, then the degrees of freedom for RSS would be (n-k) for a generalized model.

That means n minus k is the degrees of freedom. Then what should be the degrees of freedom for ESS? Degrees of freedom for ESS would be df for TSS - df for RSS. So that means in this case, in a two variable model, it would be (n-1) - (n-2), so that means 1. This is how you can calculate the degrees of freedom. And for a general model then it would become n-1, so this would become (n-1) - (n-k), so it would become k degrees of freedom for TSS.

(Refer Slide Time: 21:17)



The screenshot shows the Stata interface with the following output:

```
Single-user State perpetual license:
Serial number: 481606223548
Licensed to: NPTEL
NPTEL

Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 5000; see help set_maxvar.

*(2 variables, 10 observations pasted into data editor)

. reg consumption income

Source      SS          df           MS          Number of obs =      10
-----+-----+-----+-----+-----
Model    8552.72727      1    8552.72727    Prob > F = 0.0000
Residual 337.272727      8    42.1590909    R-squared = 0.9621
Total    8900           9    987.777778    Adj R-squared = 0.9573
                          Root MSE = 6.493

consumption  Coef.  Std. Err.   t    P>|t|   [95% Conf. Interval]
-----+-----+-----+-----+-----
income     -5090989   .8357428   14.24  0.000   -4266678   -591514
_cons      24.45455   6.419817   3.81   0.005   9.664256   39.24483
```

Right, now you see in Stata, total number of observation is 10 that is why degrees of freedom for TSS, total sum of square is $n - 1$, 9. And what is the degrees of freedom for RSS? Since I have only one explanatory variable, that means I am estimating two parameters. So, $10 - 2$ which is 8. And $9 - 8$ equals to 1. That is how we can calculate the degrees of freedom for TSS, RSS and ESS.

So, first we will calculate the degrees of freedom for TSS which is always $n - 1$. And the degrees of freedom for RSS is actually $n - k$ for the generalized model and in this case, it is $n - 2$. Once you know the degrees of freedom for TSS, degrees of freedom for RSS, just take the difference between these two, you will get the degrees of freedom for your ESS, model sum of square.

And then, the stata is always defined that MS, mean sum of square, which is nothing but the SS divided by degrees of freedom. So, if you divide this, 8550.72 divided by 1, is this. Similarly, if you divide this by 8, it will come 42.15. So, it is simply SS by its degrees of freedom. And if you take the ratio of these two MS, then that will lead to another important statistic which is called F statistic.

MS of model, MS of residual if you take, then that will give you a statistic which is called F and that value will turn out to be 202. What is implication of that F statistic? That we will discuss later once you understand the hypothesis testing. Right, so with this, we are just closing our

discussion on today. So, today we discussed about the meaning of regression, what are the three important properties of beta hat that we estimate.

Then we have also learnt how to estimate the model, how to interpret the coefficient and then, we have also learnt the goodness of fit measure and using a simple diagram, we understand the analysis of variance, three important components that is there in the analysis of variance, total sum of square, which you are going to explain through your model defined as sum of Y_i minus \bar{Y} whole square.

And then, the other two components is explain sum of square and in Stata, it is actually model sum of square and then the residual sum of square. And we have also learnt how to interpret the ANOVA from the diagram to this table what Stata is reporting. Now, in our next class, we will try to understand the interval estimation portion, that means the hypothesis testing, to ensure that this 0.50 what I get here, that is not out of a chance, rather it is statistically different from 0. So in our next class, we will discuss about hypothesis testing. Thank you.