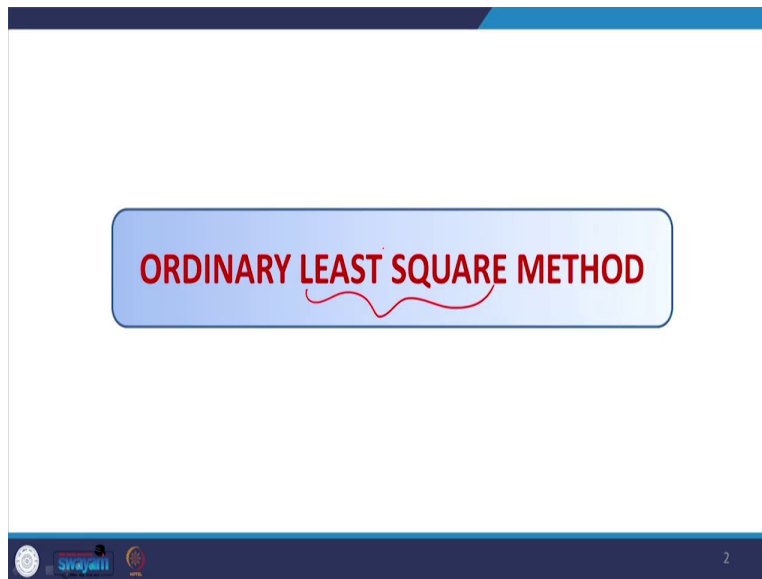


**Handling Large-Scale Unit Level Data Using STATA**  
**Professor. Pratap C. Mohanty**  
**Department of Humanities and Social Sciences**  
**Indian Institute of Technology Roorkee**  
**Lecture 29**  
**Linear Regression Analysis in Stata-II**

Welcome once again friends to this NPTEL module on Handling Large-Scale Unit Level Data Using STATA. We have already discussed the very basics of regression analysis, the linear regress analysis in the previous lecture with their very important properties related to population trend line and the sample out of it and their statistics. Now, we are going to discuss some further details related to the linear regression analysis. To start with the linear regression analysis, most important technique is through ordinary least square method. There are important interpretations to it.

(Refer Slide Time: 01:17)



We need to discuss about the least square approach of the error term, why least square, because just averaging the square of the error term, error term I think I have already shown you, if we simply take the error term which are deviating from the standard trend line, from the standard trend line if there are so many errors positive and negative errors are there from the trend line, if I simply take an average, average might be 0. But if the average is 0 that cannot guarantee the estimation. Basically, average might be 0, but if you square the error term, the minimum most error term if it is attach then we can able to estimate correctly.

The error term, by squaring the error term instead of just taking the average if that gets to 0, we cannot able to estimate anything, because it has already converted to a 0 value or we cannot derive the equation for estimation. But by squaring the errors it is not going to be 0. Every time the error term we are going to square that means a positive value will be boiled down. So, error term whatever is there we are converting that to a positive value.

So, positive value if it is minimized then that is called the best method to estimate and once the model ensures that you have reached into a minimum most square of the error terms, you can able to estimate the best coefficient out of it. That is why it is called the ordinary least square method.

(Refer Slide Time: 03:02)

Abbreviated as **OLS**.  
 Most commonly used method to obtain some reasonably good estimate of  $\beta$ 's which help the most in studying Y as affected by X.  
  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are obtained by finding the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that minimize the sum of squared differences between Y and  $\hat{Y}$  :

$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Or,  $\sum \varepsilon_i^2 = \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$

Here,  $\sum \varepsilon_i^2$  is called Error Sum of Squares (ESS)

This also abbreviated with OLS in simple language. Then most commonly used method to obtain some reasonably good estimate of betas which held the most in studying Y as affected by X. So, in our example or in our previous equation, we mentioned that we are supposed to derive the estimated value of it that is in terms of beta not hat and beta 1 hat. We are referring to the estimated value and are obtained by finding the values of beta not hat and beta 1 hat that minimize the sum of square differences between Y and Y hat. Y hat basically the estimated value. So, this is nothing but the error term.

So, this is the  $\hat{Y}$  that gets defined with the help of beta not, the intercept coefficient as well as the slope coefficient. Sum of the error term, why I said error term has to be, sum of the error term if it is 0, minimizing to it, if it is near about 0, then we can able to estimate and this is the equation which I already said, sum of the error term, the squaring of the error basically  $Y_i$  minus beta not plus beta 1  $X_i$  is equal to this.

So this is sum of error is also called as error sum of squares. We are going to discuss what is this ESS and how this is going to be useful for understanding the best fit of that trend line.

(Refer Slide Time: 04:45)

**Note:** the sum is taken over all observations.

□ Interpretation of the slope and intercept:

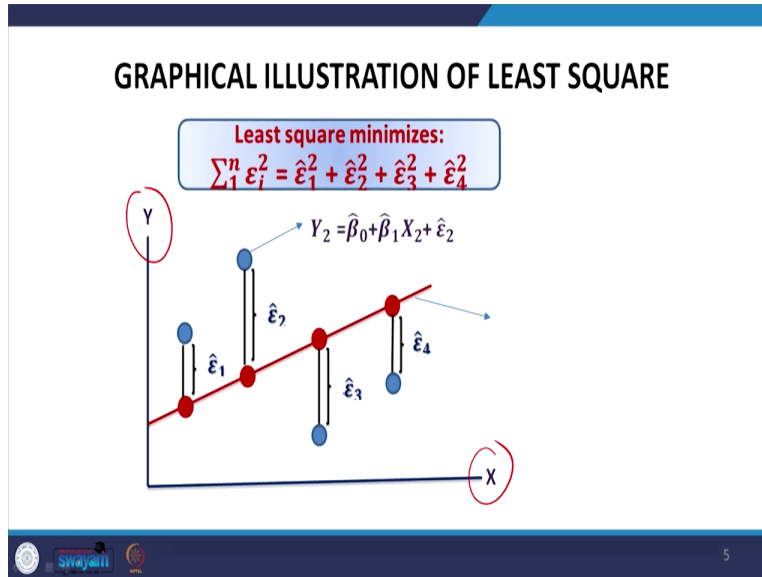
- $\hat{\beta}_0$  is the estimated average value of Y when the value of X is Zero.
- $\hat{\beta}_1$  is the estimated change in the average value of Y as a result of a 1-unit change in X.

swayamii 4

The sum is taken over all the observations. All the observations and their points are taken. The interpretation of the slope and intercept is also important. This is the slope, the coefficient of the intercept and this is the coefficient of the slope and beta not has the estimated, every time I am mentioning hat that indicates estimated average value of Y given the value of X when it is equal to 0.

That means, when X is 0. I told you already, beta not refers to autonomous variable, so intercept term. So, autonomous means it is not affected by change of 0, change of value of X. So, when there is a 0 change in X, what is the impact on Y. That is nothing but called beta not hat. So, beta 1 hat is the estimated change in the average value of Y as a result of 1 unit change of X that we already mentioned in our earlier slides or in our earlier lecture also.

(Refer Slide Time: 06:05)



Coming to the graphical interpretation of the least square method, why least square is important. So, the average, like in this case the model, if you take the least square that means, we can able to estimate the best fit line, but if you do not take square of it, it could be also negative. The average of the error could be negative or it could be positive. But if it is negative then the trend line is explaining nothing. So, no where it is saying it is within the space of X and Y. So, the error term has to be squared.

This is first error term at this particular point. There is numerous error term possible, but if I am writing down error term, error hat 1, error hat 2, 3, 4, like this, least square means sum of error square and so on. In this case 4 are there so we are writing 4.

(Refer Slide Time: 07:05)

### COEFFICIENT EQUATIONS

- ❑ Predicted equation:  
$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$
- ❑ Sample Slope:  
$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

*Cov*  
*Variance*
- ❑ Sample Y-intercept:  
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

---

- ❑ Abbreviated as **OLS**.
- ❑ Most commonly used method to obtain some reasonably good estimate of  $\beta$ 's which help the most in studying Y as affected by X.
- ❑  $\hat{\beta}_0$  and  $\hat{\beta}_1$  are obtained by finding the values of  $\hat{\beta}_0$  and  $\hat{\beta}_1$  that minimize the sum of squared differences between Y and  $\hat{Y}$  :  
$$\min \sum (Y_i - \hat{Y}_i)^2 = \min \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$$

Or,  $\sum \varepsilon_i^2 = \sum (Y_i - (\beta_0 + \beta_1 X_i))^2$

Here,  $\sum \varepsilon_i^2$  is called Error Sum of Squares (ESS)

What are the coefficient equations? We have predicted equations like beta not hat and beta 1 hat  $X_i$ . The sample slope based on this predicted equation is that how beta 1 is defined that is basically derived from the square of the error term which we already mentioned a couple of minutes back in this equation.

So, beta 1 can be derived from that, basically that is called the sum of  $X_i - \bar{X}$  minus  $Y_i - \bar{Y}$  sum of square between X Y and divided by sum of square between the independent variables. So, sum of square that is basically this is called the variance term, this is called covariance term.

So, covariance, basically, the relationship between two, that is a dependent variable and independent variable with respect to the total variance of the independent variable. So, basically, what is the unit change of the variance of the independent variable on the relationship between the 2 variables. So, that is the beauty of  $\beta_1$ . That really explains the change in the nature of X and Y.

According to the sample Y intercept, where that is basically indicated with  $\beta_0$  that simply when we take the average of it since this only boils down to mean of  $\bar{X}$  because this is going to be constant. So,  $\beta_0$  times number of  $\beta_1$  we have that will be some simply averaged. So, Y average minus  $\bar{X}$  average multiplied with its  $\beta_1$  coefficient is going to estimate our intercept.

So, that we will explain. I am not going to explain everything about the details of BLUE properties or other detail properties, because this is going to be the explanation in econometric class and I think any standard econometrics book if you unfold and go by the pages that explain the BLUE properties, best linear unbiased estimator why  $\beta_0$  has to be unbiased, why  $\beta_1$  has to be unbiased. So, that explanation you should follow on your own. We are not explaining. In fact, I am giving very important concept in this lecture. Next class we will have an applied version of all these coefficients and its estimated value.

Coming to the most interesting part of our lecture is understanding the variation. So, the estimated  $\beta_1$  value gives the relationship between the variance, relationship between X and Y with respect to the total variance of X. How the unit variance of X lead to the relationship variance between X and Y, that is all about understanding the variance. But, the variation where the variance has a role has to be very clearly emphasized. When we are saying variance somewhere we are also addressing the variation. The variation is explained largely by the error term. But what is that variation?

(Refer Slide Time: 10:51)

### MEASURES OF VARIATION

□ Total variation is composed of two parts:

$$\text{TSS} = \text{RSS} + \text{ESS}$$

TSS- Total sum of square.  $\sum(Y_i - \bar{Y})^2$

RSS- Residual sum of square.  $\sum(Y_i - \hat{Y}_i)^2$

ESS- estimated/ explained sum of square.  $\sum(\hat{Y}_i - \bar{Y})^2$

Where,

$\bar{Y}$  - average value of the dependent variable.

$\hat{Y}_i$  - predicted value of Y for the given value of  $X_i$ .

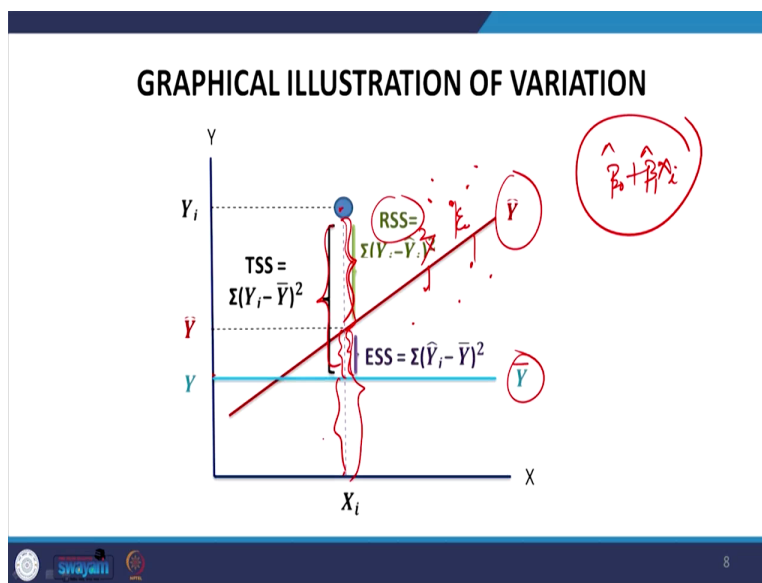
$Y_i$  - observed values of the dependent variable.

Measures the variation of  $Y_i$  values around their mean  $\bar{Y}$ .

Variation attributable to factors other than the relationship between X and Y.

Explained variation attributable to the relationship between X and Y.

7



So, the total variation is composed of two parts. So, let us decompose the total variation. 1 is in terms of residual sum of square and another is called explained sum of square. So, TSS stands for total sum of squares. It is very very interesting to understand. Look at this first then you come back to this.

So, you look at this, what we are trying to explain is the following. So, the total, suppose this is a point. We are estimating the value is somewhere here. What is the TSS in this case? Given a

particular point, what is the TSS for our understanding? The TSS basically the  $Y_i$  the particular point which we are referring here, corresponding to the  $X_i$  value that is  $X_i$ .

In this case, our total variance will be over, how to know the total variance if you have a standard average of that distribution of the  $Y_i$  dependent variable. So, the average, if it is indicated here this is the  $Y$  line. So, when we know the average value, it is projected or pointed out constant throughout the  $X$  because it is average, so the total change of individual  $Y$  is explained by this point.

This is the total change, isn't it? So, this is basically  $Y_i$  minus  $Y$  bar and that is the total change given the  $X_i$ . But how to explain that, the  $Y_i$  is deviating from its average value is correspond by the  $X_i$  variable, how to know it. If you can bifurcate, if you can decompose in terms of  $X$ , you can able to estimate the exact changes through  $X$ .

In order to do so, we have to take the help of estimated values. What is that? Estimated values means the estimated parameter, estimated line. The estimated line is explained by  $Y$  hat.  $Y$  hat that is basically  $\beta_0 + \beta_1 X_i$  or  $\beta_0$  hat plus  $\beta_1$  hat.  $\beta_0$  hat we have already explained number of times. So, this is  $X_i$ . So,  $Y_i$  hat once I estimated, so we can have a point here, like here we have a point. From this point, this is basically the estimated difference.  $Y_i$  over the estimated value that is called the residual sum of square.

So, how the individual  $Y$  point is deviating from the estimated value. What are the vertical differences? We said several times these difference, if any important points are there, any points is here, any point is there, basically this is from the trend line from the estimated value. So the error or the epsilon, these are called epsilons. The epsilon and its square is measuring the RSS. So,  $Y_i$  minus  $Y_i$  hat and its square and its sum of square is going to give us the value.

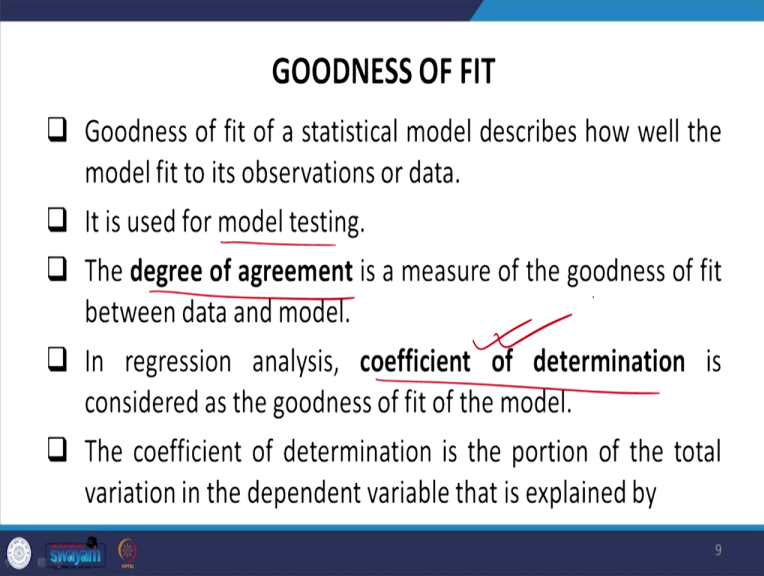
if I only take instead of square, we can get it clarified, if I only take without square, if square is not there, then that may boil down to 0 and whole equation may boil down to 0. It is not going to give any estimate value. So, square minimizes the error value and the best minimum error if you can do that is going to give a best unbiased estimator. So, blue properties somewhere validating our coefficients, where we can get it clarified, I am not spending more time on it.



What is also important, this part is explained by the error term. So, error sum of square or the residual sum of square we have explained. What about other part? This portion is called explained sum of square. Explained sum of square means what? This we know that this is called from this to this, this to this, this to this is called this is our difference between  $\bar{Y}$ , this is total  $\hat{Y}_i$ , this is totally  $\hat{Y}_i$  minus  $\bar{Y}$ .

$\hat{Y}_i$  minus  $\bar{Y}$  the gap is called estimated value, estimated sum of square. So, we are going to explain all those things in detail. I think the slides has tried to explain these things in detail you can go through and clarify it accordingly.

(Refer Slide Time: 15:34)



### GOODNESS OF FIT

- Goodness of fit of a statistical model describes how well the model fit to its observations or data.
- It is used for model testing.
- The degree of agreement is a measure of the goodness of fit between data and model.
- In regression analysis, coefficient of determination is considered as the goodness of fit of the model.
- The coefficient of determination is the portion of the total variation in the dependent variable that is explained by

Let us come to the goodness of fit. Goodness of fit is very very important in regression analysis throughout, not just in this lecture, in others subsequent lectures you will be very careful enough to understand that goodness of fit. Basically we are referring to the trend line every time. If our trend line is perfectly fit to the data or to the variation, if the variation is very close that means it has a very good goodness of fit or our fit is very good.

So, goodness of fit of a statistical model describes how well the model fits to its observations or data. It is used for model testing also and especially for the panel data and for the time series data, they are very careful enough about this value. Sometimes for individual cross sectional data, they do not check it in detail, but we will discuss later on. But, I just wanted to mention that

the degree of agreement is a measure of the goodness of fit between data and model. The model is basically the fit and the data are the points.

If the degree of agreement is good then we have a better estimate. In regression analysis, coefficient of determination is considered as the goodness of fit of the model. So, coefficient of determination we are going to discuss. I think in statistics where you study the standard deviation, standard variance, that time also some book and papers explain the coefficient of determination somewhere aligning with the argument we are floating. The coefficient of determination is the portion of the total variation in the dependent variable

(Refer Slide Time: 17:33)

Variation in the independent variable.

- ❑ This measure was coined by a geneticist **Sewall wright** in **1921**.
- ❑ It is denoted by  $r^2$  in simple linear regression model and  $R^2$  in multiple linear regression model.  $r^2 \rightarrow R^2$
- ❑ Explanation through venn diagram:

(A) (B) (C)

10

which is explained by the variation in independent variable to the dependent variable is basically called coefficient of determination. This measure was originally coined by a geneticist Sewall Wright in 1921.

It is also denoted by small r square, if it is a simple linear regression, we already clarified. But if it is a multiple linear regression that is denoted by capital R square, the difference between small r square and this is 1 of the quiz question and every time it is important. So, do not get confused these are the same answer, but this is our small r square written in case of simple and that is capital R square in case of multiple regression analysis or model.

A diagrammatic, Venn diagram explanation for the R square value that is the goodness of fit. The coefficient of determination we are trying to explain in diagram A, B and C, you can easily see perceive 1 important interpretation that diagram A as even a 2 variable that is dependent Y and independent X variable as if there is no relationship.

Whereas in case of C, there is 100 percent relationship. So, that means, where is the best fit we derive that is in the C model that is equal to 1. If you wanted to understand the coefficient of determination or the best fit, goodness of fit, C is the best one. The number will be 1, whereas in case of B model, it varies from 0 to 1. In the first case it is 0, A case it is 0, C it is 1, so B stands in between.

(Refer Slide Time: 19:34)

$r^2 = 0$  in figure (A), there is no relation between Y and X.  
  $r^2 = 1$  in figure (C), variation in Y is totally due to the variation in X.  
  $r^2 = \epsilon(0,1)$  in figure (B).  
  $r^2 = \frac{ESS}{TSS} = \frac{\text{Estimated sum of squares}}{\text{Total sum of squares}}$   
  $r^2 = 1 - \frac{RSS \text{ (Residual sum of squares)}}{TSS}$   
  $R^2$  therefore lies between **0 and 1**. the closer it is to 1, the better is the fit. The closer it is to 0, the worse is the fit.

$TSS = RSS + ESS$

11

This is what we are explaining in this slide, r square equal to 0 in case of figure A, r square equal to 1 in case of figure C, so r square lies between 0 and 1 in case of figure B. Then what is the r square in total in terms of equation this is simply ESS and TSS we already explained, explained sum of square divided by total sum of square.

How much explained sum of square is address out of the total sum of square. Total sum of square basically the particular data point as compared to its population parameter, it is not population parameter, the data point with respect to the average of that population parameter. That is only

counting the total sum of square. Total sum of square, when the deviation is calculated with respect to the data point and the average of that data point then that is called total sum of squares.

Out of that I already explained that it composed of residual sum of square and estimated sum of square. explained sum of square, so not estimated sum of square it is called explained sum of square. So, the explained sum of square, in that case we are estimating the explained part through the trend line then that of the trend line is the point for us minus the average of that dependent variable, that we have already explained.

So, what do you mean by r square can also be interpreted when we say it is ESS upon TSS and this is TSS, we have already explained, this is RSS plus ESS. Then the next equation follows that 1 minus RSS will explain the same r square. The residual sum of square, you can explain through RSS or you can also explain through ESS. So r square, therefore, lies between 0 and 1. So, the closer it is to 1 that means the model is fit. Closer it is to 0, it is the worse fit. So, that means your model is not correct.

(Refer Slide Time: 22:21)

- ❑ A disadvantage attached to the  $R^2$  measure is, it is an **increasing function of the number of regressors**. It means more the number of regressors, higher the value of  $R^2$ .
- ❑ To avoid this problem, another measure is used to assess the goodness of fit of a model, called **Adjusted  $R^2$** .
- ❑ **Adjusted  $R^2$**  explicitly takes into account the number of regressors included in the model.
- ❑ Denoted as  $\bar{R}^2$  (R-bar square).
- ❑ Computed from the  $R^2$ :

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k}$$

The disadvantage attached to this R square measure is that it is an increasing function of the number of regressors. So, number of regressors if it is increasing that means if you have more regressor, more explanation will be there to the model and so R square value is expected to be

higher. So, R square if it is higher, usually in the cross sectional data we have more regressors, if it is less regressor then R square value you will get very very less.

To avoid this problem another measure is used to assess the goodness of fit when the number of regressor increases though that defines a better fit to the model, but there are some disadvantages also. Like we are not attaching our degree of freedom, reduces when we have more number of regressors, because the number of parameter also increases.


When the number of parameter to be estimated increasing accordingly the degrees of freedom reduces, our estimate is going to be problematic. So, instead of R square, adjusted R square is very important for the interpretation, R square is not the right explanation when you have more regressors.

So, adjusted R square take into account number of regressors included in the model. So, basically we are explaining more than 1 or more regressions in a model. So, this is denoted by R square bar. So, make it very clear that whenever R square bar is written that means we are referring to adjusted R square. In STATA also this notation is going to be very useful and I will suggest that whenever you calculate or estimate the result, please refer to R square and try to interpret the R square bar.

So, how to compute that, that is basically adjusted, this is perfectly fine. Only we are addressing one important explanation, but what is this?  $1 - R^2$  also important, we are taking the ratio of the degrees of freedom,  $n - 1$  divided by  $n - k$ ,  $k$  is the number of parameter to be estimated. If you take the ratio like this, we are referring to the adjusted R square.


(Refer Slide Time: 24:53)

- The word adjusted here means **adjusted for the degrees of freedom** which depends on the number of regressors (k) in the model.
- If  $k > 1$ ,  $\bar{R}^2 < R^2$ , the number of regressor in the model increases, the  $\bar{R}^2$  becomes increasingly smaller than the  $R^2$ .
- The  $R^2$  is always positive, but the  $\bar{R}^2$  can sometimes be negative.
- $\bar{R}^2$  often used to compare two models that have the same dependent variable.

 13

- A disadvantage attached to the  $R^2$  measure is, it is an **increasing function of the number of regressors**. It means more the number of regressors, higher the value of  $R^2$ .
- To avoid this problem, another measure is used to assess the goodness of fit of a model, called **Adjusted  $R^2$** .
- Adjusted  $R^2$**  explicitly takes into account the number of regressors included in the model.
- Denoted as  $\bar{R}^2$  (R-bar square).
- Computed from the  $R^2$ :

$$\bar{R}^2 = 1 - (1 - R^2) \frac{n-1}{n-k}$$

 12

The word adjusted here means adjusted for the degrees of freedom which depends on the number of regressors that is k. So, if k is greater than 1, obviously your number of parameter increases and that is in the denominator, so the R square bar value is going to be lesser than that of the R square. So, that is less than this. The number of regressor in the model increases, the R squared bar becomes increasingly smaller than that of R square. So, the R square is always positive, this is very very important.




So, R square that is the coefficient of determination is going to be positive always, but R square bar, adjusted R square can be also negative because of the case that we are dividing the degrees of freedom and also there are some negative portions. So, when that exceeds than that of the 1 it could be also negative, but in rarest cases where we will experience negative coefficient. Further details we can explore during our experimentation with data. R square bar often used to compare two models that have the same dependent variable.

(Refer Slide Time: 26:16)

### STANDARD ERROR OF ESTIMATES

- $\beta$ s are random variables, as their values vary from sample to sample.
- In statistics, the variability of a random variable is measured by its variance  $\sigma^2$ , or its square root the standard deviation  $\sigma$ .
- In regression analysis, the **standard deviation** of an estimator is called the **standard error**.
- In LRM, an estimate of the variance of the error term  $\epsilon_i$ ,  $\sigma^2$  is computed as:

$$\hat{\sigma}^2 = \frac{\sum \epsilon_i^2}{n-k} = \frac{RSS}{df}$$

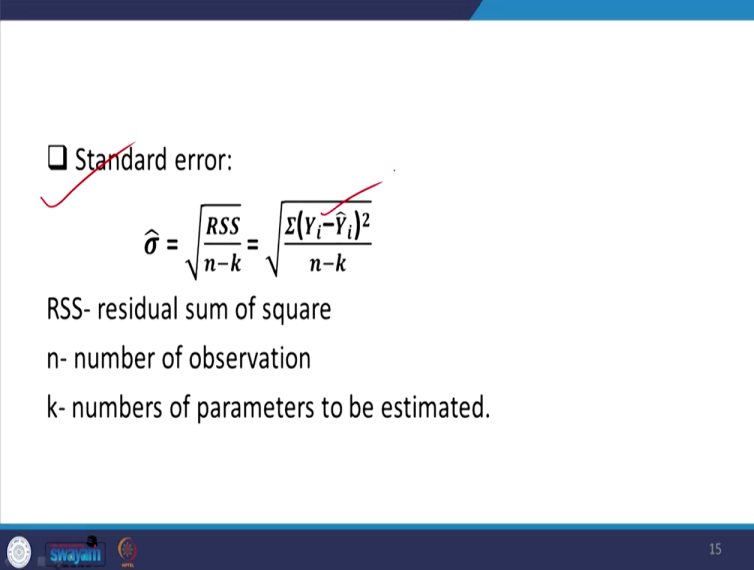



14

Coming to the standard error of the estimates. So, besides that this is also important for our models as every coefficient is attached with a standard error and where the interpretation is very, very important. Betas are random variables as their values vary from sample to sample. In statistics, the variability of a random variable is measured by its variance that is sigma square and its square root of basically called standard deviation.

So, in regression analysis, the standard deviation of an estimator is called the standard error. So, this is also a quiz question most often asked. I should suggest you to emphasize and read further. The linear regression model, an estimate of the variance in this model, the estimate of the variance of the error term that is epsilon is sigma square which is computed as sigma square hat then because we are estimating it.

So, sum of sigma square, error square, divided by the number of the degree of freedom, is adjusted with the sample size or the cases n minus k that is basically residual sum of square divided by the degrees of freedom. If you do it, we are getting the standard error, the estimate of the error term in this model.

(Refer Slide Time: 28:15)



□ Standard error:

$$\hat{\sigma} = \sqrt{\frac{RSS}{n-k}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-k}}$$

RSS- residual sum of square  
n- number of observation  
k- numbers of parameters to be estimated.

15

So, try to remember that RSS divided by the degrees of freedom is the answer for standard error. But that is standard variance. If you take the square root of it, that is basically called standard error. Standard deviation of the standard variance is called standard error. This is the equation we wanted to estimate.

So, error if you remember, RSS if you remember the deviation from the trend line to that of the actual data point if that is estimated and we divide by not just n, if you just divide n that will be not be an unbiased estimator, we have to adjust with the n minus k degrees of freedom to it then only it will be estimating correctly. I think that we have mentioned, this is what I was trying to say.






(Refer Slide Time: 28:49)

### TESTING HYPOTHESES

**T-test of significance**

- If we want to test the hypothesis that the (population) regression coefficient ( $\beta_k = 0$ ), we use t-test:
$$t = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)}$$
- this t value has  $n-k$  degree of freedom. N is the number of observation and k is number of parameters to be estimated.
- T-table gives the probability of obtaining such a t value or greater.
- If the probability of obtaining such a value is small, say 5% or less, we can reject the null hypothesis that  $\beta_k = 0$ . it means the estimated t value is statistically significant. i.e., significantly different from zero.



16

Some explanation to be made at this moment to clarify the hypothesis testing, because hypothesis testing requires null hypothesis in order to be accepted or rejected. So, usually, in our explanation we have to reject the null hypothesis, because null hypothesis usually considered to be the ideal condition and ideal condition if it is there, if it is proved, then there is no testing made. So, null hypothesis is to be rejected and the alternative hypothesis has to be proved. So, some significance level I think you might have heard in different papers that how it is interpreted we are going to discuss in a short while.

So, T-test largely if it is n is less than 30 but usually what happens our regression analysis, our estimator reduce the estimator to every k, every 1 less than 30. So T-test in maximum cases is applied because it is applied in case of small sample cases. If we want to test the hypothesis that the regression coefficient that is beta k is equal to 0 in case of null hypothesis context we use t-test. T-test usually applied in case of small sample cases.

And this T-test has n minus k degrees of freedom and how t is determined, likewise z distribution, t distribution that is, estimated value of beta divided by standard error or the standard deviation of the variance term or the RSS term, residual sum of square of the estimator that is beta k. So likewise z distribution, I think if you remember standard z it is always t distribution is also following the same pattern only, every time we are dividing by its standard error or standard deviation. So, it is basically the standard errors divided.

So, this t value has  $n - k$  degrees of freedom. So,  $n$  is the number of observation and  $k$  is the number of parameters to be estimated. So, every time  $k$  we are referring to the number of parameters to be estimated. In case of sample, do not get confused that in case of sample estimation we talk about statistics not the parameter.

The t-table gives the probability of obtaining such a t value or greater. We have a t-table in every statistics book or even in Google search you will get a standard t values, standard t-tables are there at every degrees of freedom. You can compare your estimation with that of the standard table at certain level of degrees of freedom.

Based on our comparison I think those are clarification we should get from our statistics and econometrics books. We are only giving you some basic guidance and then we will quickly start applying the STATA module. If the probability of obtaining such a value is small, say 5 percent or less, we can reject the null hypothesis that is  $\beta_k = 0$ , null hypothesis in this case the estimated value of that beta coefficient of the relationship between these is 0, basically there is no relationship, if I am saying  $\beta_k = 0$  basically X and Y having no relationship.

Usually in the ideal condition there is no such relationship. We have to prove that, yes, there exist a relationship and with our significance level we can able to prove that your null hypothesis is rejected, your alternative that is it has a relationship, alternative hypotheses validated. What I wanted to mention that if it is 5 percent or even less, we can reject the null hypothesis. It means that the estimated t value is statistically significant and significantly different than that of 0.

(Refer Slide Time: 32:43)

- ❑ Some common probability values are 10%, 5% and 1%, known as **level of significance** ( $\alpha$ ) or **type I error**.
- ❑ Stata provides t value along with its p-value.
- ❑ A low p-value suggest that the estimated coefficient have significant impact on the regressand.
- ❑ **Confidence interval:**
  - ❑ stata also compute 95% confidence interval for individual regression coefficient.
  - ❑ Provides a range of values that has a 95% chance of including the true population value.

So, some common standard limit of rejection or acceptance, any way you can interpret, the common probability values those are defined and accepted are 10 percent, those are 10 percent, 5 percent and 1 percent level. So, alpha values, these are also called type 1 error, alpha values to define the significance, that is also called level of significance. One important clarification I wanted to mention that usually that means 10 percent error or 5 percent error or 1 percent error is acceptable in any model.

So, then now, if it is exceeding 10 percent error, then we will certainly say that your model is not a best fit or it is not a good model. Your coefficient is not estimating correctly. It is not proving your hypothesis. So basically, any point you can also test your model, level of significance could be, like here it is 0.10, here it is 0.05, here it is 0.01 at 1 percent level.

there are some cut off points for better interpretation that does not mean any other value is not important, you can also test at 0.06, you can test also at 0.02 or any level till 0.01 that is 10 percent level you can test. So, there is no hardship, but it is little complicated because those standard values are difficult to derive.

So, STATA provides t value along with its p-values as well. So, that is the important aspects in STATA we are going to refer in our next class. Low p-value, the probability limit value, suggest

that the estimated coefficient have significant impact on the regression. So, low p-value we are going to define in our interpretation.

Similarly, we are also supposed to compare with the confidence interval. So, confidence interval defines the distribution of the data. So accordingly, once the interval is defined what are the extreme limit, the alpha limits are there and on that we can able to take a decision in terms of the significance level. So, STATA also computes a 95 percent confidence interval for individual regression coefficient. It provides a range of values that has a 95 percent chance of including the true population value. So, we will interpret those with the help of data.

(Refer Slide Time: 35:20)

This 95% is called the **confidence coefficient (cc)**, computed as  $100(1-\alpha)$ .

The  $1-\alpha$  confidence interval for population coefficient is calculated as:

$$\Pr[\beta_k \pm t_{\alpha/2} \text{se}(\beta_k)] = [1-\alpha]$$

Where, Pr- probability

$t_{\alpha/2}$  value of t statistics obtained from t distribution (table).

$\alpha/2$ - level of significance with appropriate degrees of freedom.

$\text{se}(\beta_k)$ - standard error of  $\beta_k$ .

$[\beta_k - t_{\alpha/2} \text{se}(\beta_k)]$  is lower limit and  $[\beta_k + t_{\alpha/2} \text{se}(\beta_k)]$  is upper limit of the interval.

18

This 95 percent is called the confidence interval, computed at 100 that is 1 minus alpha, at 100 level that is 1.1 minus alpha. If 95 percent basically means 5 percent error, we are going to estimate. So, it is 1 minus alpha, 1 minus 0.95 that is 0.05. So, 1 minus alpha is a confidence interval for population coefficient calculated, that is the probability limit.

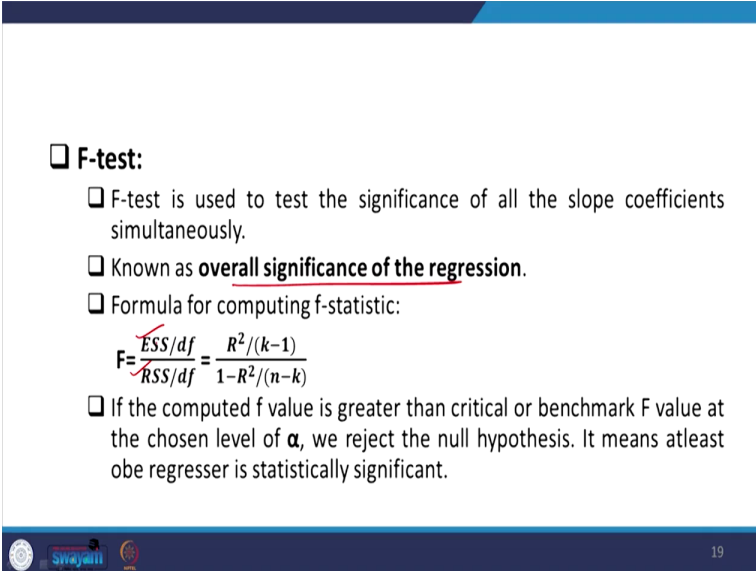
How You can get the details from the statistics books? It defines as, if it is a 2 tailed test, something we also discussed from the beginning of our lectures, if it is a 2 tailed test then the probability limit with its standard error that is defined with beta k plus minus t value alpha divided by 2. 2 we refer to if alpha percentage, we know that at 5 percent level or even 10 percent level. If it is 5 percent level then t value we need to compare with 0 point, instead of

0.05, we are comparing at 0.025. So, we need to divide it into two categories with a standard error with a probability limit that will defined our confidence interval that is 1 minus alpha.

So, t alpha by 2 is nothing but the t statistics obtained from the t distribution from the table. But be careful that what is this tailed test? Usually, it is two tailed test we refer, because the data may be distributed in a normal shape, so there is possibility of both sides.

So, alpha by 2 is nothing but the level of significance maybe 5 percent, maybe 1 percent with appropriate degrees of freedom. Standard error is noted with se of bk and if that limit is lower the limit and basically when you subtract the beta value to its tabulated value with its alpha significance level with their standard error minus if you are subtracting it you will get the lower limit of the confidence interval, lower limit as compared to upper limit. We will explain with the help of data also. If you plus it from the mean value its deviation explained by the error term that will define the upper limit of the confidence interval. We will explain this later.

(Refer Slide Time: 38:11)



The slide contains the following text:

- **F-test:**
  - F-test is used to test the significance of all the slope coefficients simultaneously.
  - Known as overall significance of the regression.
  - Formula for computing f-statistic:
$$F = \frac{ESS/df}{RSS/df} = \frac{R^2/(k-1)}{1-R^2/(n-k)}$$
  - If the computed f value is greater than critical or benchmark F value at the chosen level of  $\alpha$ , we reject the null hypothesis. It means atleast one regressor is statistically significant.

At the bottom of the slide, there are logos for Swayam and other institutions, and the number 19 in the bottom right corner.

Similarly, the F-test is also used to understand the significance of the slope coefficients, like F-test is used to test the significance of all the coefficients simultaneously. It is also known as overall significance of the regression. The formula, in that case, we said RSS was important, we are saying RSS with its standard error we defined in case of t, but we are comparing the

explained sum of square as compared to its residual sum of square and their respective degrees of freedom are also divided.

If the computed F value is greater than critical or benchmark F value at the chosen level of alpha, we reject the null hypothesis. It means that at least 1 regressor is statistically significant. We are going to discuss these details in due course of our time. I know that these are simply formula. Unless we do not test it with the help of the real life data, it might be difficult for your better interpretation.

Our purpose is to explain through STATA. So this lecture so far gave you a clear background of what are the basic test, basic necessary steps we must follow and we have already guided. With this, I think I should not stretch further. We will continue and carry forward our STATA application of the OLS in the next class at least half an hour we will spend on understanding the applications. So, thank you very much.