Essentials of Data Science with R Software - 2
Sampling Theory and Linear Regression Analysis

Prof. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Linear Regression Analysis
Lecture - 44
Simple Linear Regression Analysis
Test of Hypothesis and Confidence Interval Estimation with R

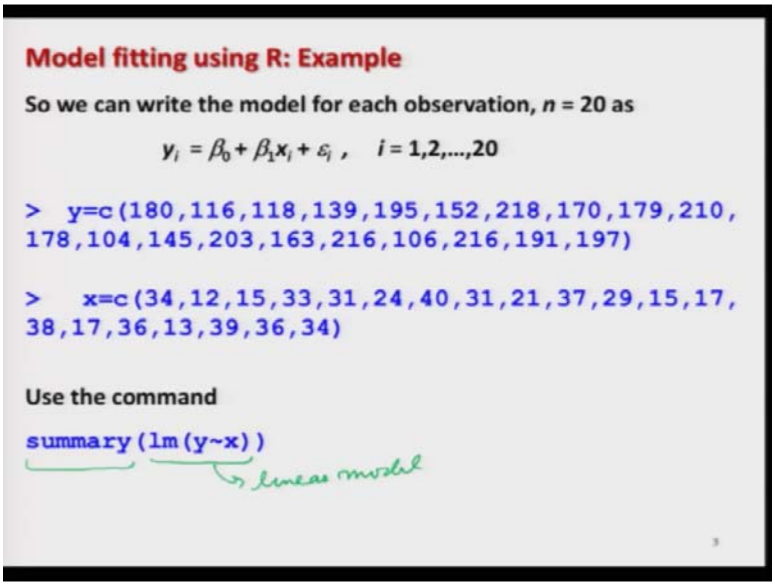
Hello friends, welcome to the course Essentials of Data Science with R Software 2, where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. In this lecture we are going to continue with our module on Linear Regression Analysis and we are going to learn the topic of Simple Linear Regression Analysis with R software.

So, you can recall that in the earlier lecture I had considered the test of hypothesis and confidence interval for the regression coefficients, slope parameter intercept term and for the variance σ^2 . So, we had seen conceptually and in theory that how these expressions are obtained. And we also constructed the confidence interval for the slope parameter intercept term and variance.

Now, in this lecture my objective is that I would try to show you all those things in the R software. So, what I will do? I will try to consider same example which I had considered earlier on a simple linear regression analysis and I will try to show you that how you can extract the information on the test of hypothesis and confidence interval under the command lm. You can recall that we have used the command lm linear models to find out the simple linear regression model in the earlier example.

So, we will continue with the same example and I will try to show you that how you can obtain such values. So, let us begin our lecture, ok. So, you can recall that I had earlier considered this example in which we had collected the data on 20 students on their marks obtained and the corresponding values of number of hours of study. So, these marks are indicated by y and the number of hours per week of study they are indicated by x. So, we have got here 20 pairs of observations on x i and y i.

(Refer Slide Time: 02:25)



Model fitting using R: Example

So we can write the model for each observation, $n = 20$ as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \,, \quad i = 1,2,\ldots,20$$

```
> y=c(180,116,118,139,195,152,218,170,179,210,
178,104,145,203,163,216,106,216,191,197)

> x=c(34,12,15,33,31,24,40,31,21,37,29,15,17,
38,17,36,13,39,36,34)
```

Use the command

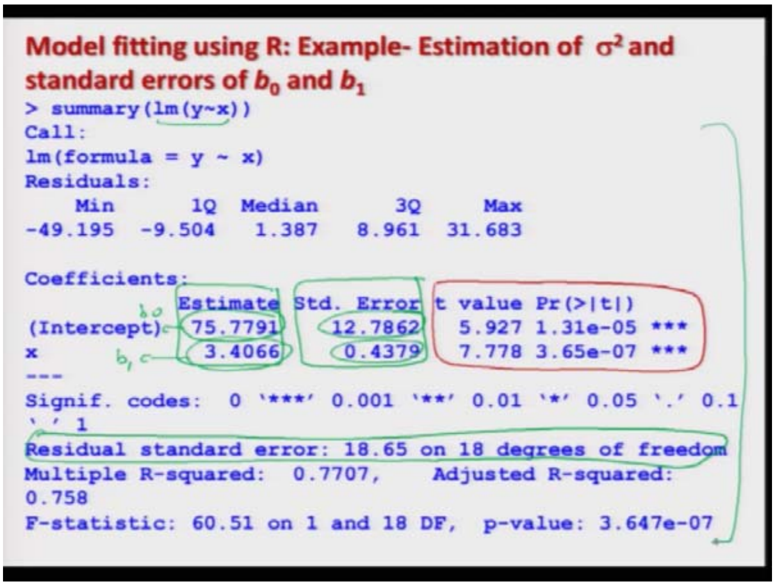
```
summary(lm(y~x))
```

linear model

And then we had entered this data and we had used the command lm y tilde x to obtain the linear model, the fitted linear model. And then we had earlier also used the command summary on this function. So, and then we had lots of outcome, so, again I am going to use the output of the summary command.

And I will show you that the information on test of hypothesis and confidence interval it is impeded in the outcome of summary command. And how to read it, how to get it that is my objective to show you here.

(Refer Slide Time: 03:00)



Model fitting using R: Example- Estimation of σ^2 and standard errors of b_0 and b_1

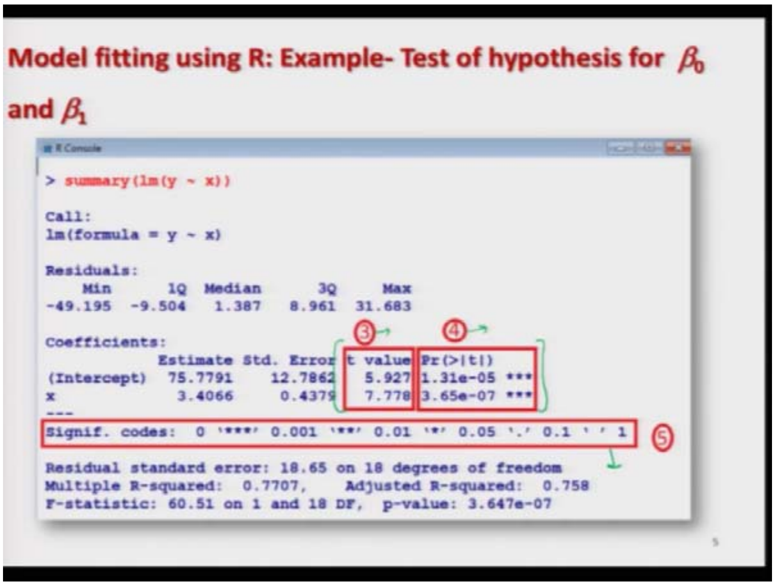
```
> summary(lm(y~x))
Call:
lm(formula = y ~ x)
Residuals:
    Min      1Q  Median      3Q     Max
-49.195  -9.504   1.387   8.961  31.683

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.7791    12.7862   5.927 1.31e-05 ***
x             3.4066     0.4379   7.778 3.65e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
' ' 1
Residual standard error: 18.65 on 18 degrees of freedom
Multiple R-squared:  0.7707,     Adjusted R-squared:
0.758
F-statistic: 60.51 on 1 and 18 DF,  p-value: 3.647e-07
```

So, if you remember if I try to do this execute this summary command on this function lm y tilde x on the R console we get this type of outcome. There are many many values and I promise you that as I am going further into the lectures I will try to show you the interpretation of all other parts.

So, if you remember we already have done this part. In this part this was the estimate of intercept term $\hat{\beta}_0$ or say b_0 value of b_0 and this was the estimate of slope parameter say b 1 and after this we also consider this part this was trying to give us the standard error of intercept term and this 0.4379 was the standard error of b 1. And we also had consider this part. So, up to now we have covered this part of this software outcome

Now, we are going to first consider on this part which I am highlighting in color red right.

(Refer Slide Time: 04:27)

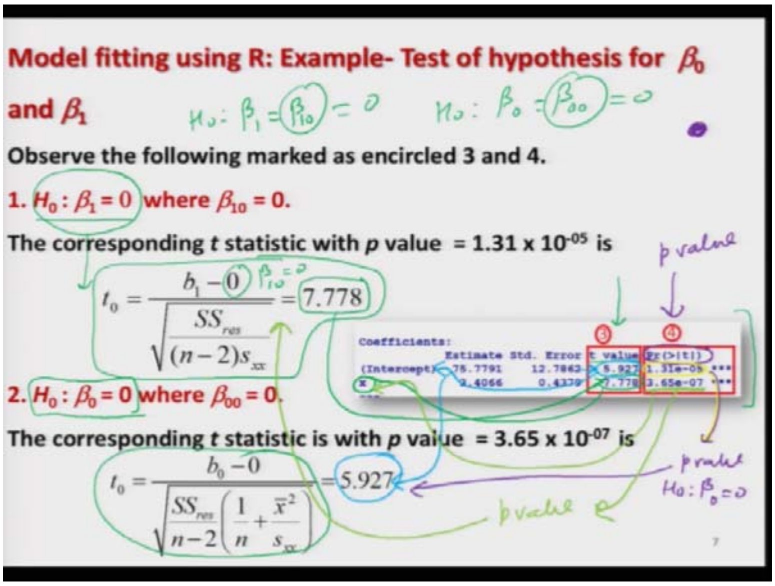


Model fitting using R: Example- Test of hypothesis for β_0 and β_1

```
> summary(lm(y ~ x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max 
-49.195  -9.504   1.387   8.961  31.683 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  75.7791    12.7862   5.927 1.31e-05 ***
x             3.4066     0.4379   7.778 3.65e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.65 on 18 degrees of freedom
Multiple R-squared:  0.7707,    Adjusted R-squared:  0.758 
F-statistic: 60.51 on 1 and 18 DF,  p-value: 3.647e-07
```

So, let me try to show it on this on the; a screenshot. So, I am going to first consider here this section which I have denoted by number 3 this section here which I am denoted by number 4 and after that I will try to give you the interpretation of this part which I have denoted at section 5. So, now, I will try to take one by one this part and I will try to show you.

(Refer Slide Time: 04:52)



Model fitting using R: Example- Test of hypothesis for β_0 and β_1

$H_0: \beta_1 = \beta_{10} = 0$ $\quad H_0: \beta_0 = \beta_{00} = 0$

Observe the following marked as encircled 3 and 4.

1. $H_0: \beta_1 = 0$ where $\beta_{10} = 0$.

The corresponding t statistic with p value = 1.31 x 10^{-05} is

$$t_0 = \frac{b_1 - 0}{\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}} = 7.778 \quad \beta_{10} = 0$$

2. $H_0: \beta_0 = 0$ where $\beta_{00} = 0$.

The corresponding t statistic is with p value = 3.65 x 10^{-07} is

$$t_0 = \frac{b_0 - 0}{\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n} + \frac{\bar{x}^2}{s_{xx}}\right)}} = 5.927$$

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.7791    12.7862   5.927 1.31e-05 ***
x             3.4066     0.4379   7.778 3.65e-07 ***
```

So, now I have taken here this part over here and I will try to explain you this number 3 and number 4 part on this slide right. So, here I am going to consider the test of hypothesis for β_1 and β_0 . You can recall that when we consider the test of hypothesis when we have taken for the slope parameter β one was taken as β_1 0.

Now, what are we going? We are going to assume that this $\beta_{10} = 0$. So, we are essentially going to test H_0 : $\beta_1 = 0$. And similarly for the intercept term also we had consider this H_0 $\beta_0 = \beta_{00}$. So, this β_{00} is going to be considered as 0.

So, we are interested in testing the hypothesis H_0 : β_0 = or $\beta_0 = 0$. The reason that why we are taking 0 because 0 has got some nice interpretation which is useful in the variable selection and in making a decision whether the variable is important in some sense or not right.

But this I will try to consider later on at this moment my objective is to show you that, what is happening in the R software? So, whenever you are trying to consider the soft the outcome of any software related to the linear regression analysis. Usually they will be giving you the outcome of the null hypothesis where the corresponding parameters are equated to be 0, right. Something like H_0 : β_0 equal to 0 some and some and say β_1 equal to 0.

So, we remember that when I want to test the H_0 :β_1 equal to 0 then we had considered the test hypothesis to be like this β_1 - β_{10} which is equal to here 0 divided by standard error. And the outcome of this statistics which is 7.778 this is shown here try to see the movement of my pen this is here.

So, if you try to consider the column number which is named as 3 then this is here the value 7.778 which is the t value and this t value is corresponding to the slope parameter right. And similarly if you try to consider the null hypothesis H_0 : $\beta_0 = 0$ then we had seen that the corresponding statistics was given by this.

And in the software this outcome is going to be indicated by this value here. Try to see if you come to here the just try to see the movement of my pen in blue color here you have a value 5.927. So, this value here 5.927 is the t value corresponding to the intercept term that is H_0 β not equal to 0 and which is the value of this statistic, right ok.

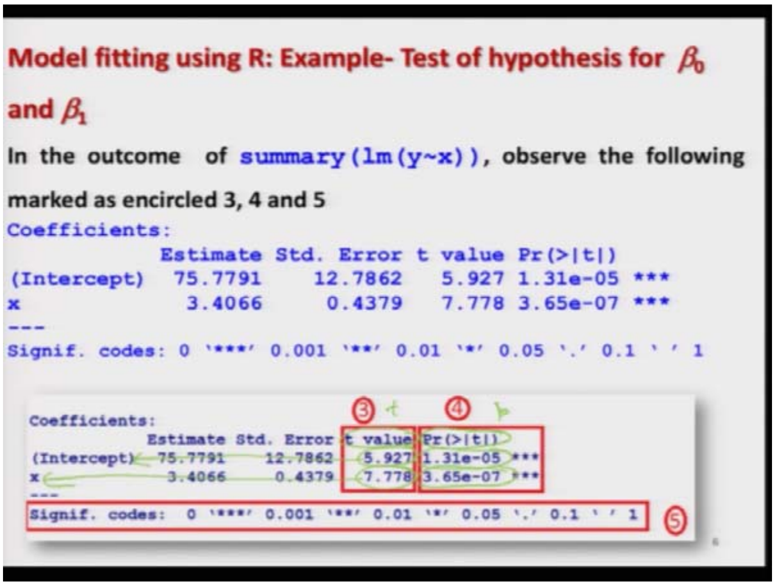
So, now you can see here that the column which is called as here t value that is trying to give you the corresponding t values of different null hypothesis of different regression coefficients ok. Now I try to use a different color pen you can see here this is my color of pen so, you can now watch it.

So, I come here on column number now 4. You can see here in the column number 4 it is written P r inside parenthesis greater than absolute value of t. So, this is the probability that the t value is greater than absolute value of t. So, this is essentially your p value So, this value here which is 1 point you can see here 1.31.

Now you can see it clearly here 1.31 x 10^{-5} this is the here this is here the p value corresponding to the null hypothesis H_0 intercept term that is $\beta_0 = 0$. So, this is the p value corresponding to this t_0 . And similarly, if I try to take here different color this second value here 3.65 into 10 to the power of - 7 this is also here a p value and this value p value is corresponding to you can see here this is corresponding to the regression coefficient associated with x.

So, this is the p value corresponding to the null hypothesis H_0 : $\beta_1 = 0$ right.

(Refer Slide Time: 10:40)



Model fitting using R: Example- Test of hypothesis for β_0 and β_1

In the outcome of `summary(lm(y~x))`, observe the following marked as encircled 3, 4 and 5

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.7791    12.7862   5.927 1.31e-05 ***
x             3.4066     0.4379   7.778 3.65e-07 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

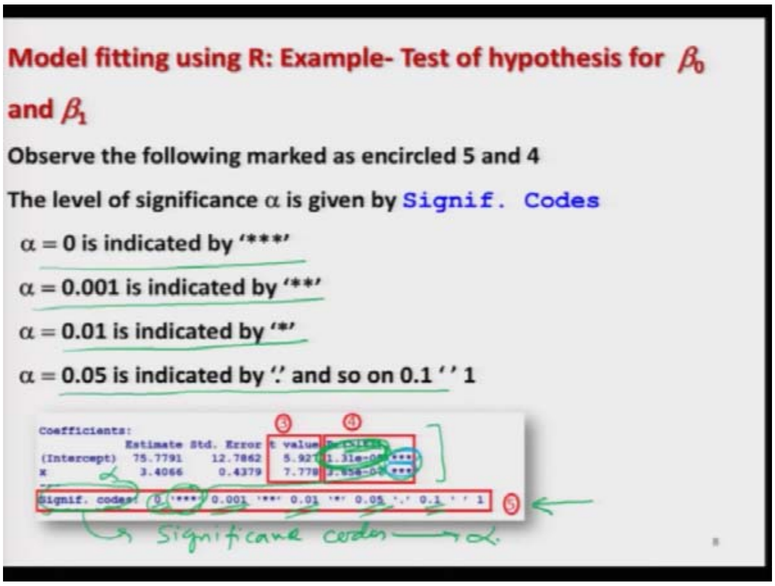
```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.7791    12.7862   5.927 1.31e-05 ***
x             3.4066     0.4379   7.778 3.65e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So, this is how you can see? That how the things are happening for example, now you can come here now you can see here very clearly on this screen that this is the t value and this is here the p value this is t value and this is here the p value And you can see here the corresponding with respect to on the left hand side.

So, the first row is indicating the intercept term and the second row is indicating for the slope parameter. So, now, you can see here that in the first row this is the value of t statistics corresponding to the null hypothesis H_0 intercept term equal to 0 and this is the corresponding p value.

And similarly the second row this is the value of t statistics corresponding to H_0 slope parameter equal to 0 and this is here the p value corresponding p value So, now you can see here that the obtaining the value of t statistics etc. in software is very simple. The only thing is this you should know, what are these value? How to identify that the corresponding values for intercept term or for slope parameter?

(Refer Slide Time: 11:48)



Model fitting using R: Example- Test of hypothesis for β_0 and β_1

Observe the following marked as encircled 5 and 4

The level of significance α is given by Signif. Codes

$\alpha = 0$ is indicated by '***'

$\alpha = 0.001$ is indicated by '**'

$\alpha = 0.01$ is indicated by '*'

$\alpha = 0.05$ is indicated by '.' and so on 0.1 ' ' 1

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.7791    12.7862   5.927 1.31e-0[illegible] ***
x             3.4066     0.4379   7.779 [illegible]e-0[illegible] ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

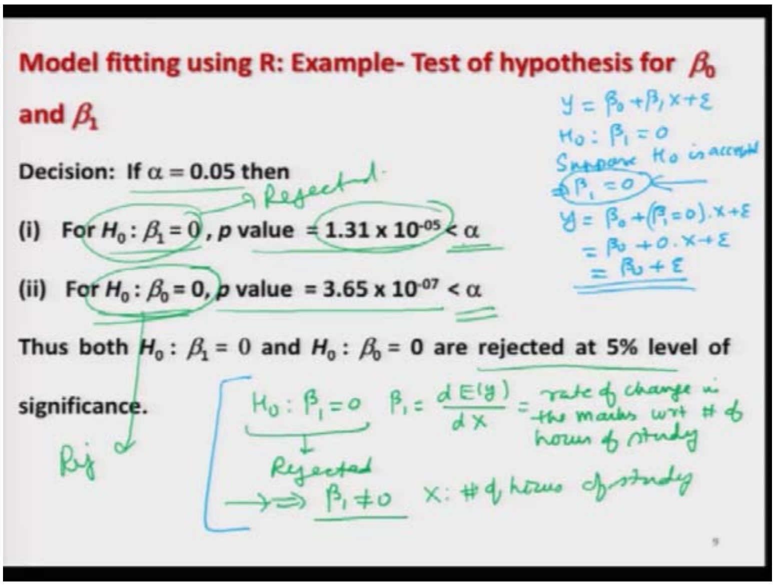
Now I come to this part number here 5. You can see here it is written here Signif dot c o d e s codes. So, this is actually significance codes right. So, this is essentially the value of actuallyα level of significance right. So, you can see here this is taking the so, I can write down here this is actuallyα So,α = 0, α = 0.001α = 0.01, α = 0.5 and so on so and α = 0.1.

So, and these stars are indicating the corresponding values over here you can see here right. So, that is essentially trying to tell you that if you try to look at this outcome here you are looking I will try to make it in the blue color you can see here this for 3 stars. So, 3 stars are indicating that the level of significance what is being used here this is the corresponding to theα where the indication is by 3 stars so this is right.

So, actually soα equal to 0 is indicated by 3 starα equal to 0.001 is indicated by 2 stars andα equal to 0.01 is indicated by single star andα = 0.05 is indicated by just here dot and so on right. So, this is how it is trying to indicate that what is the level of significance? What is being used? And then the value ofα is being used in the test of hypothesis which has been used by the p value right.

So, these stars are simply indicating the level of significance what is being used right?

(Refer Slide Time: 14:03)



Model fitting using R: Example- Test of hypothesis for β_0 and β_1

Decision: If $\alpha = 0.05$ then

(i) For $H_0 : \beta_1 = 0$, p value $= 1.31 \times 10^{-05} < \alpha$

(ii) For $H_0 : \beta_0 = 0$, p value $= 3.65 \times 10^{-07} < \alpha$

Thus both $H_0 : \beta_1 = 0$ and $H_0 : \beta_0 = 0$ are rejected at 5% level of significance.

Rejected
$y = \beta_0 + \beta_1 x + \varepsilon$
$H_0 : \beta_1 = 0$
Suppose H_0 is accepted
$\beta_1 = 0$
$y = \beta_0 + (\beta_1 = 0).x + \varepsilon$
$= \beta_0 + 0.x + \varepsilon$
$= \beta_0 + \varepsilon$
Rej
$H_0 : \beta_1 = 0$ $\beta_1 = \frac{d E(y)}{dx}$ = rate of change in the marks wrt # of hours of study
Rejected
$\Rightarrow \beta_1 \neq 0$ x: # of hours of study

So, for example, in this case you can see here I have written here very clearly that ifα = 0.05 then if you want to test the hypothesis H0 β1 equal to 0 then p value is here. 1.31 into 10 to the power of minus 5 which is actually given here you can see here, right. And which is smaller thanα So, I have to take a call so the rule was reject H0 when p value is less thanα So, I can say here that H0 β1 equal to 0 is rejected right atα level of significance.

Similarly, for the intercept term also the corresponding p value is 3.65 x 10^{-07} which is again less thanα So, I can conclude here that the null hypothesis H_0 : $\beta_0 = 0$ is also rejected atα equal to 5 percent level of significance.

And you can also see here that that they have got the interpretation also, you are trying to test for example, let me try to explain you are trying to test H_0 $\beta_1 = 0$. What is you are here β_1, β_1 is the rate of change in the value of y with respect to X right. So, you are trying to find out here the rate of change in the marks with respect to number of hours of study right.

So, now looking at the conclusion that H_0 β_1 equal to 0 is rejected this is indicating that β_1 is not equal to 0, β_1 is not equal to 0; that means, the corresponding value of here X is an important variable. And it is affecting the outcome and that is obvious that the marks of the students are going to be affected by the number of hours of studies.

And this test of hypothesis is also indicating the same conclusion that the variable which you have considered as the number of marks obtained the number of marks they are depending on the number of hours of study. And hence the variable X which is the number of hours of study is an important variable that is affecting the outcome means how to understand it? How to show it?

So, let me try to take here the opposite story and I try to write down in a blue color so that you can identify it. Our model was $y = \beta_0 + \beta_1 X + \varepsilon$. Now you try to test the hypothesis H_0 $\beta_1 = 0$. And suppose it is accepted H_0 is accepted; that means, $\beta_1 = 0$ in the population; that means, my model will be revised as $\beta_0 + (\beta_1 =0)$ times $x + \varepsilon$. Which is equal to $\beta_0 + 0$ into $x + \varepsilon$ which is $\beta_0 + \varepsilon$.

So, now you can see here that in this model there is no role of x. And also if you try to see if you are trying to accept the hypothesis that H_0 $\beta_1 = 0$. So, β_1 is trying to measure the rate of change in the marks with respect to the number of hours of study. So, you are saying that $\beta_1 = 0$; that means practically there is no change in the marks of the student when the number of hours of study changes.

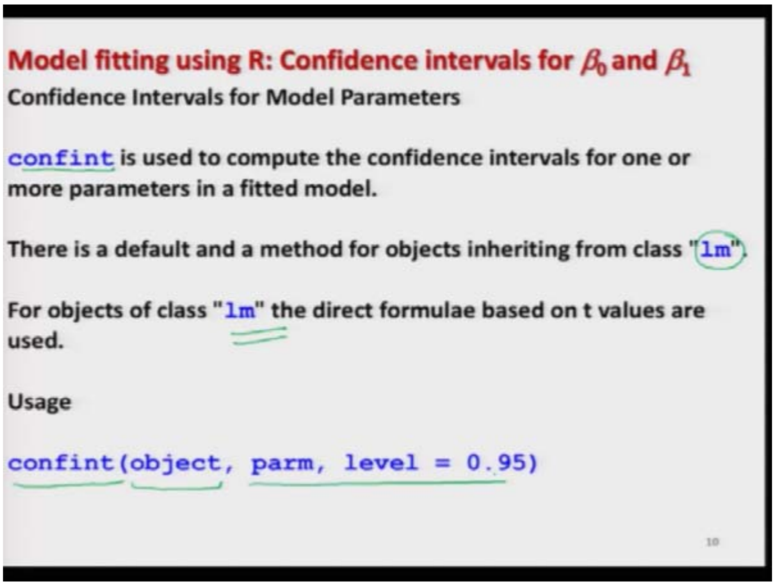
So that means, the marks are independent of the number of hours of study so whether the student's studies or not the marks are not going to be affected. Well we all know that this is a wrong conclusion this should not happen and that is what exactly your test of hypothesis is indicating here when you say that H_0 β_1 equal to 0 is rejected that is what you want right ok.

Similarly, now if I come to the interpretation of say H_0 :β ≠ 0 then this is also getting rejected. That mean the intercept term is playing an important role in the model and it is not equal to 0, right. So that means, we have to consider and it is correct.

And if you go logically also that is also correct if a student is not studying at all, but suppose the student is attending classes he is doing assignment etc. then yes surely he will get some marks in the in the exam that cannot be equal to 0 I mean, that is that we know from our experience.

So, whatever are the conclusions which we have obtained on the basis of test of hypothesis they are also getting confirmed with what is happening in the real life in the real experiment. So, now you can see here that the model which we had obtained which have that has also been verified through the test of hypothesis and you can believe that up to now the model is doing well ok.

(Refer Slide Time: 20:26)



Model fitting using R: Confidence intervals for β_0 and β_1

Confidence Intervals for Model Parameters

`confint` is used to compute the confidence intervals for one or more parameters in a fitted model.

There is a default and a method for objects inheriting from class "`lm`".

For objects of class "`lm`" the direct formulae based on t values are used.

Usage

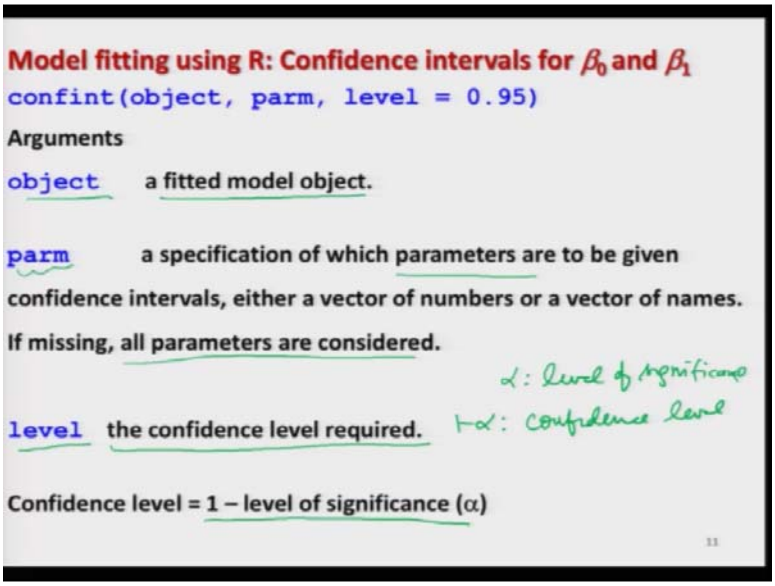
```
confint(object, parm, level = 0.95)
```

10

Now, if you want to find out the confidence interval for that we have a command here conf i n t that is the short form of confidence interval c o n f i n t. This is actually used to compute the confidence interval for one or more parameters in a fitted model. And this is actually this command has to be used along with actually the command lm. whatever is the output of the lm this conf i n t command will extract the confidence interval from the outcome of lm, right.

So, the command here is you try to write conf i n t then inside right inside the parenthesis you try to write down the object; the object act means the outcome of the lm command and then you try to give your parameters and the level.

(Refer Slide Time: 21:30)



Model fitting using R: Confidence intervals for β_0 and β_1

```
confint(object, parm, level = 0.95)
```

Arguments

`object` a fitted model object.

`parm` a specification of which parameters are to be given confidence intervals, either a vector of numbers or a vector of names. If missing, all parameters are considered.

α : level of significance

`level` the confidence level required. $1-\alpha$: confidence level

Confidence level = 1 – level of significance (α)

11

One thing what you have to keep in mind here that there are some different interpretations here which you have to keep in mind, means; obviously, when I say object this is a fitted model means I will try to show you with the example also. And when I am trying to say here p a r m this is indicating the parameters.

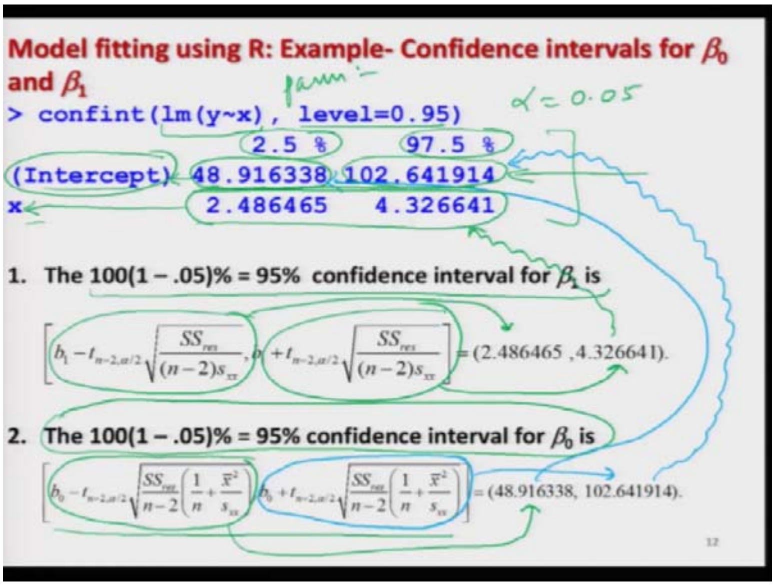
So, we would like to tell the software whether we want the confidence interval for intercept term or slope parameter and in multiple linear regression model, when we have more than one regression coefficient that what are the parameters of which we want to have the confidence interval right.

And if you do not write anything against this p a r m then all the parameters whatever are involved they will be considered and you will get the confidence interval for all the case.

For example, in the simple linear regression model if you do not use p a r m then you will get the confidence interval for slope parameter as well as for the intercept term.

And then there is a option here level l e v e l this is indicating the confidence level. Remember one thing ifα is the level of significance then, 1 minus α is the confidence level. So, in this function the input has to be given in terms of confidence coefficient. So, that is the value of 1 minusα or 1 -level of significance.

(Refer Slide Time: 23:10)



Model fitting using R: Example- Confidence intervals for β_0 and β_1

parm =

$\alpha = 0.05$

```
> confint(lm(y~x), level=0.95)
                2.5 %     97.5 %
(Intercept) 48.916338 102.641914
x            2.486465   4.326641
```

1. The 100(1 – .05)% = 95% confidence interval for β_1 is

$$\left[b_1 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}},\ b_1 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}\right] = (2.486465\ ,4.326641).$$

2. The 100(1 – .05)% = 95% confidence interval for β_0 is

$$\left[b_0 - t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)},\ b_0 + t_{n-2,\alpha/2}\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n}+\frac{\bar{x}^2}{s_{xx}}\right)}\right] = (48.916338,\ 102.641914).$$

12

So, now first I try to show you the outcome on this screen and then I will try to show you on the R console. So, if you remember we had used this command lm (y ~ x). So, this will give us the fitted linear model. Now, I am not using here p a r m command or option because I want the confidence interval for both the parameters.

Remember confidence interval is obtained only for the regression parameter not for the σ^2 ; for σ^2 we have to do it separately. So, at this moment with this conf i n t command I am going to consider the confidence interval only for the regression parameters.

So, I am trying to fix here level is equal to 0.95; that means, myα is 0.05. So, the confidence coefficient is 1 minusα which is 95 percent. Now, if you try to see the outcome it will look like this. So, you can see here first row here is 2.5 percent then 97.5

percent. And then in the first column there is intercept term then x then in the second column we have some values like this. So, let us try to consider one by one all the things.

So, first I try to consider the second row which is here right. So, you can see here there; there are here 2 values one is this and second is this right. Now, this is trying to give you the confidence interval for intercept term. So, now, if you try to look in the bottom of the slide from the theory we had obtained the 100 (1 - α) which is 0.5 which is 95 percent confidence interval for the intercept term β_0 by these 2 commands right.

And the values which are obtained here or I say if you try to compute this quantity manually then you will get the value 48.91 and so on. Similarly, if you try to compute the second value here the upper confidence limit you will get the value here 102.64 and so on.

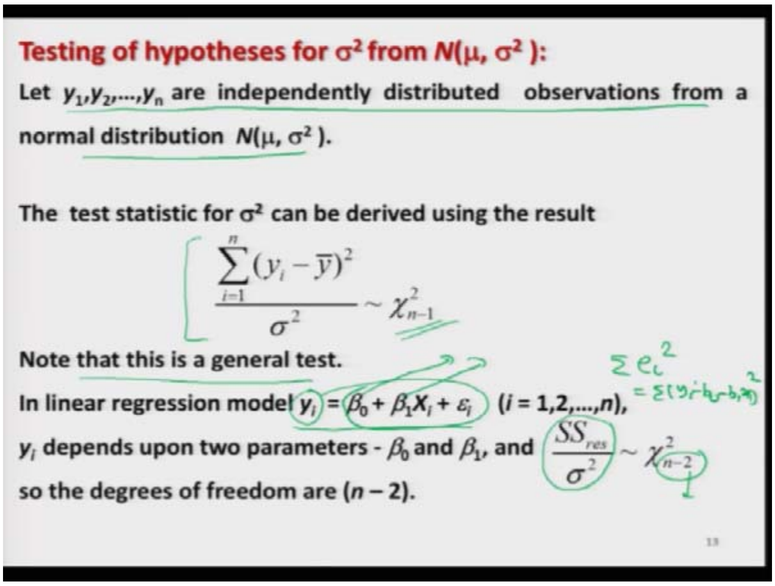
And now if you try to see these values in the software outcome you can see here that this value is given here and this 2nd value it is given here right. So, you can see here that these values 48.91 and 102.64 and so on, they are the confidence interval for the intercept term right.

And similarly, if you come to the 3rd row which is here like this is the confidence interval for the slope parameter β_1. And if you remember we had obtained the 100 1 minusα percent or here$\alpha = 0.05$ so 95 percent confidence interval for the slope parameter β_1 as like this the lower limit was this and upper limit was this.

So, if you try to compute it manually on the basis of given set of data you will get here this value 2.48 for the lower confidence interval and the upper confidence interval to be 4.32 and so on. And this is the same outcome which is given here. So, you can see here this is how you will obtain the values of the confidence limits and so you can obtain the confidence interval for all the parameters.

And if you try to give here the use the option here p a r m parameter then you can specify what you want if you want the confidence interval only for intercept term or only for slop parameter or some collected slope parameters then you will get only that outcome, right.

(Refer Slide Time: 27:52)



Testing of hypotheses for σ^2 from $N(\mu, \sigma^2)$:

Let $y_1, y_2, \ldots, y_n$ are independently distributed observations from a normal distribution $N(\mu, \sigma^2)$.

The test statistic for σ^2 can be derived using the result

$$\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Note that this is a general test.

$\sum e_i^2$

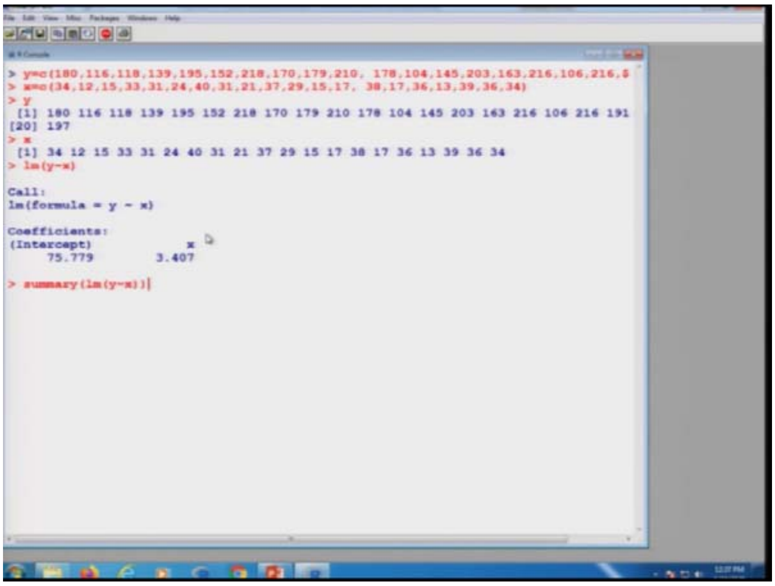
$= \sum (y_i - b_0 - b_1 x_i)^2$

In linear regression model $y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ $(i = 1,2,\ldots,n)$, y_i depends upon two parameters - β_0 and β_1, and $\frac{SS_{res}}{\sigma^2} \sim \chi^2_{n-2}$ so the degrees of freedom are $(n-2)$.

13

So, this is about identifying the confidence interval from the outcome of R software. Now I consider the test of hypothesis for the σ^2 , but before that I can show you all these things on the R console ok.

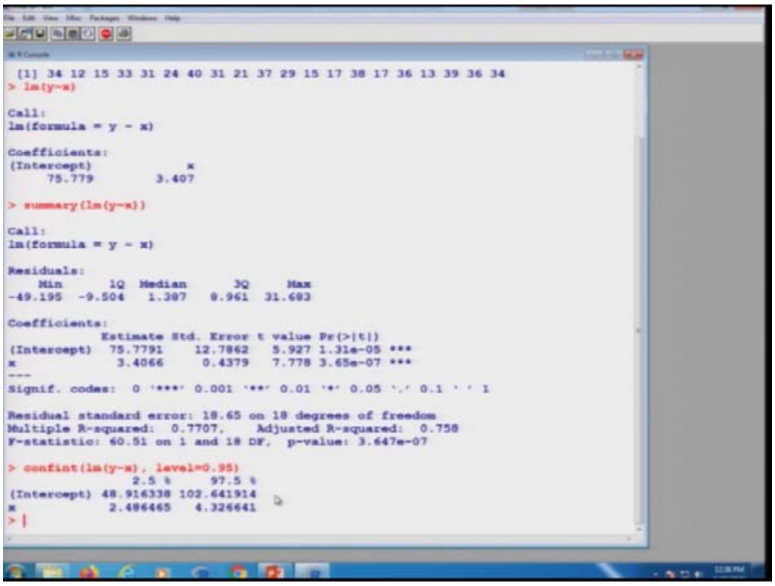
(Refer Slide Time: 28:17)



```
> y=c(180,116,118,139,195,152,218,170,179,210, 178,104,145,203,163,216,106,216,$
> x=c(34,12,15,33,31,24,40,31,21,37,29,15,17, 38,17,36,13,39,36,34)
> y
 [1] 180 116 118 139 195 152 218 170 179 210 178 104 145 203 163 216 106 216 191
[20] 197
> x
 [1] 34 12 15 33 31 24 40 31 21 37 29 15 17 38 17 36 13 39 36 34
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     75.779        3.407

> summary(lm(y~x))
```

So, I try to copy this command so, that I can save my time and yeah, I already have entered these values of x and y so, you can see here y is here x are like this. And if you try to use here the command y lm(y ~ x) this will come like this and if you try to use the summary command on this lm y tilde x it will come out to be like this, right.

(Refer Slide Time: 28:43)



```
 [1] 34 12 15 33 31 24 40 31 21 37 29 15 17 38 17 36 13 39 36 34
> lm(y~x)

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     75.779        3.407

> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-49.195  -9.504   1.387   8.961  31.683

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  75.7791    12.7862   5.927 1.31e-05 ***
x             3.4066     0.4379   7.778 3.65e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.65 on 18 degrees of freedom
Multiple R-squared:  0.7707,    Adjusted R-squared:  0.758
F-statistic: 60.51 on 1 and 18 DF,  p-value: 3.647e-07

> confint(lm(y~x), level=0.95)
                2.5 %     97.5 %
(Intercept) 48.916338 102.641914
x            2.486465   4.326641
> |
```

So, it is not giving you the confidence interval. But now if you try to the same command which I used here then you can see here this is the same outcome which you have obtained here right. So, you can see that it is not difficult right. And if you want to look at that this t values you can see here where I am highlighting right.

This is the t value corresponding to intercept term, this is the t value corresponding to β_1, this is the p value corresponding to the intercept term and this is the p value corresponding to the slope parameter. And remember this is for H_0 when the given value is assumed to be 0 right ok.

So, now you have seen these things so, now, let me come back to the confidence interval for σ^2 . So, first we assume here that $y_1, y_2, \ldots, y_n$ they are independently distributed observation from a normal distribution because for the validity of chi square you need the observation from normal distribution.

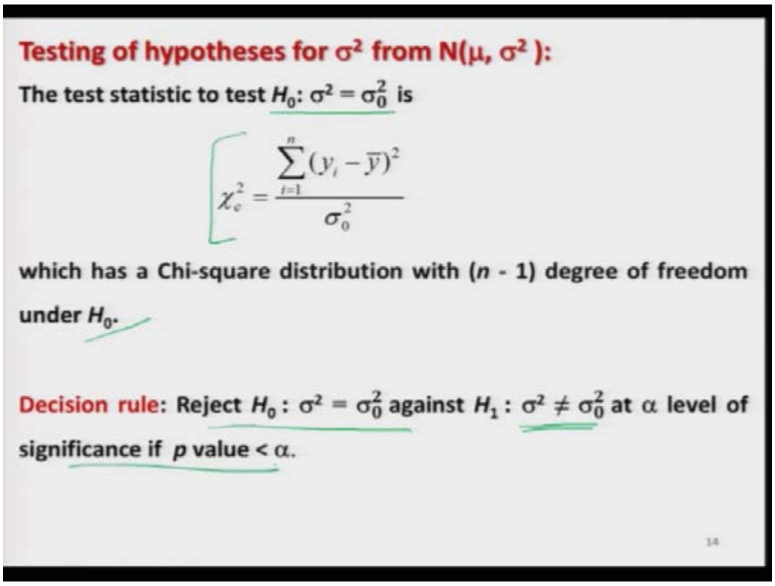
And you can recall that we had earlier discussed that the test statistic for σ^2 was found using the result $\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2}$ which follows a chi square distribution. And actually this is a very general test what I am going to show you here? Why I am calling it general? You will see that many more things can also be found from the same test. In the context of linear regression model what we are considering here we have this model.

So, your y is are obtained or they are assumed to follow this model right. And you can see here that y_i depends upon two parameters β_0 and β_1; means, unless and until you know β_0 and β_1 you cannot move further in the sense for example, if you want to find out sum of square due to residuals this was defined as $\sum_{i=1}^{n} e_i^2$ which was $\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2$.

So, unless and until you know the values of β_0 and β_1 you cannot obtain the sum of square due to residual. So, it depends on two parameters and that is the reason that here the degrees of freedom are n -2. So, this 2 is actually indicating that an order to know this quantity you should know 2 values.

So, the degrees of freedom are reduced from n to n -2 well the concept of degrees of freedom has several definition and different topics can give a different interpretation, but well I just told you so that you do not get confused because you will see that this degrees of freedom will getting; will be getting changing when I come to the multiple linear regression model. And there I will try to give you a more formal procedure that how we are determining this degrees of freedom.

(Refer Slide Time: 32:06)



Testing of hypotheses for σ^2 from $N(\mu, \sigma^2)$:

The test statistic to test $H_0: \sigma^2 = \sigma_0^2$ is

$$\chi_c^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{\sigma_0^2}$$

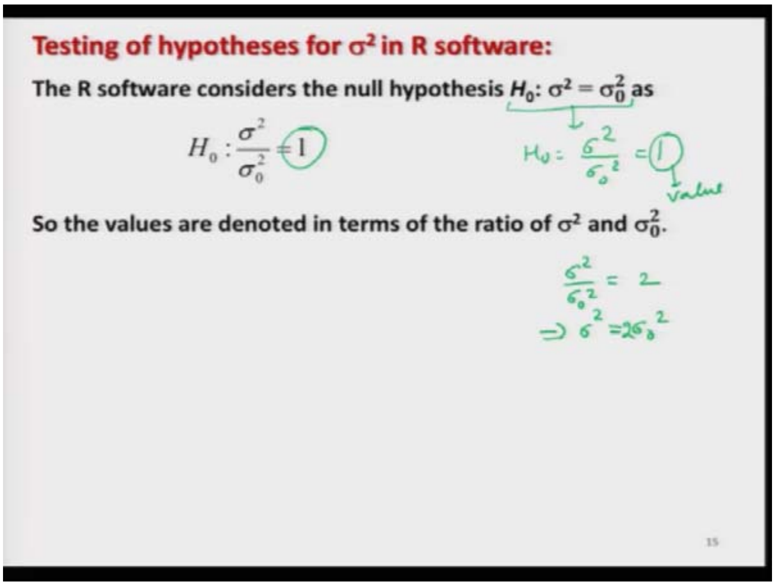
which has a Chi-square distribution with (n - 1) degree of freedom under H_0.

Decision rule: Reject $H_0 : \sigma^2 = \sigma_0^2$ against $H_1 : \sigma^2 \neq \sigma_0^2$ at α level of significance if p value < α.

14

You had seen earlier that the test statistics for testing $H_0 : \sigma^2 = \sigma_0^2$ was given by this chi square c which follows a chi square distribution with n -1 degrees of freedom under H_0 . So, now in this case the rule is the same that we are going to reject H_0 against H_1 which is a two sided hypothesis atα level of significance if p value is smaller than alpha.

(Refer Slide Time: 32:34)



Testing of hypotheses for σ^2 in R software:

The R software considers the null hypothesis $H_0: \sigma^2 = \sigma_0^2$ as

$$H_0 : \frac{\sigma^2}{\sigma_0^2} = \boxed{1}$$

$$H_0 : \frac{\sigma^2}{\sigma_0^2} = \underset{\text{value}}{\boxed{1}}$$

So the values are denoted in terms of the ratio of σ^2 and σ_0^2.

$$\frac{\sigma^2}{\sigma_0^2} = 2$$

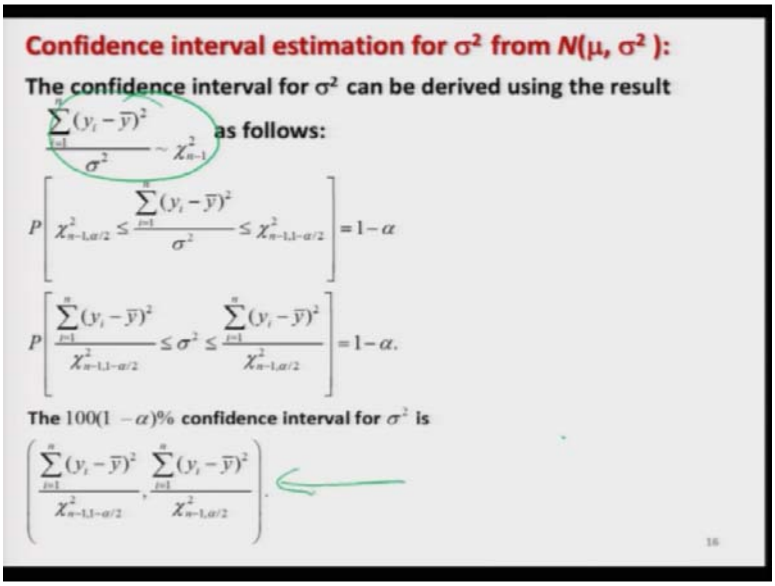
$$\Rightarrow \sigma^2 = 2\sigma_0^2$$

15

What you have to keep in mind that the package which we which we are the command which we are going to use to obtain the confidence interval and test of hypothesis for σ^2 , that requires to express the H_0 in a different way. We have done in the theory that it is our null hypothesis is H_0 $\sigma^2 = \sigma_0^2$. This thing can also be written as $\sigma^2/\sigma_0^2 = 1$ right.

So, this software requires that it is trying to replace this one by some value that can be 1 that can be 2 for example, if i say if $\sigma^2 = \sigma_0^2 = 2$ so; that means, it is going to test the hypothesis that $H_0 : \sigma^2 = 2\,\sigma_0{}^2$.

(Refer Slide Time: 33:48)



Confidence interval estimation for σ² from N(μ, σ²):

The confidence interval for σ² can be derived using the result $\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2} \sim \chi^2_{n-1}$ **as follows:**

$$P\left[\chi^2_{n-1,\alpha/2} \le \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma^2} \le \chi^2_{n-1,1-\alpha/2}\right] = 1-\alpha$$

$$P\left[\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\chi^2_{n-1,1-\alpha/2}} \le \sigma^2 \le \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\chi^2_{n-1,\alpha/2}}\right] = 1-\alpha.$$

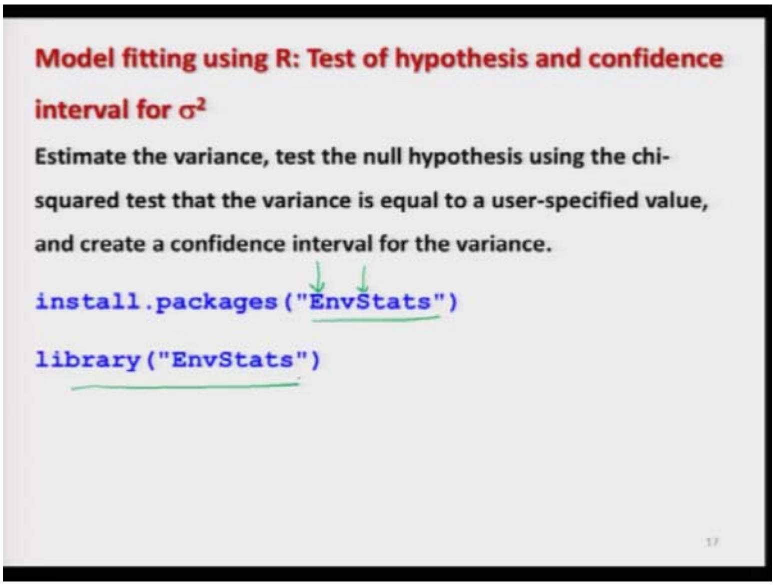
The 100(1 − α)% confidence interval for σ² is

$$\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\chi^2_{n-1,\alpha/2}}\right).$$

16

So, when you want to test the equality of σ^2 and $\sigma_0{}^2$ then in the software you have to provide the value as 1 right. Which is the ratio of σ^2 and the given value of σ^2 as $\sigma_0{}^2$ right. And you can recall that we had obtained the confidence interval also in the case of σ^2 using this statistics; statistics and based on that we had obtained the confidence interval like this so this is a quick revision for you, right.

(Refer Slide Time: 34:07)



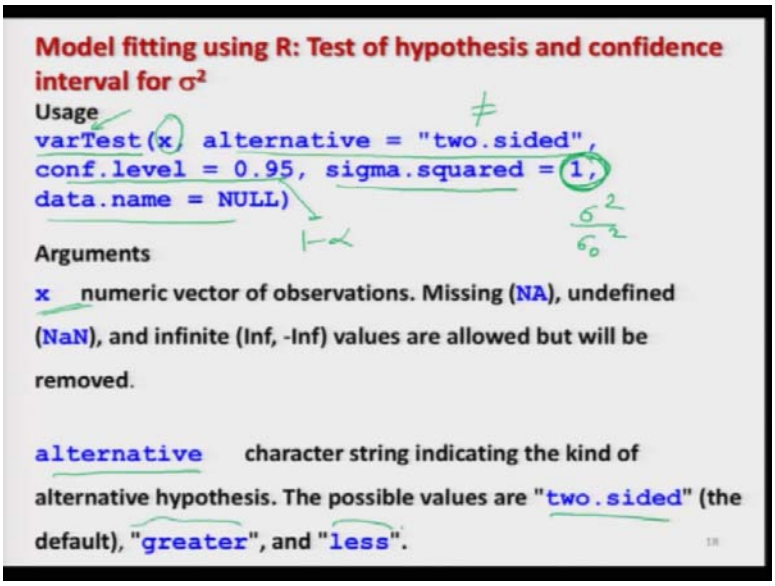
Model fitting using R: Test of hypothesis and confidence interval for σ^2

Estimate the variance, test the null hypothesis using the chi-squared test that the variance is equal to a user-specified value, and create a confidence interval for the variance.

```
install.packages("EnvStats")
library("EnvStats")
```

Now, I am going to use my software. So, in order to do the inferences for σ^2 to test the hypothesis and for confidence interval we need to install a package whose name is E n v S t a t s you have to keep a note that here this E and this S they are in capital letters right. So, you need to install this package and then you have to load this package.

(Refer Slide Time: 34:43)



Model fitting using R: Test of hypothesis and confidence interval for σ^2

Usage

```
varTest(x, alternative = "two.sided",
conf.level = 0.95, sigma.squared = 1,
data.name = NULL)
```

Arguments

`x` numeric vector of observations. Missing (`NA`), undefined (`NaN`), and infinite (Inf, -Inf) values are allowed but will be removed.

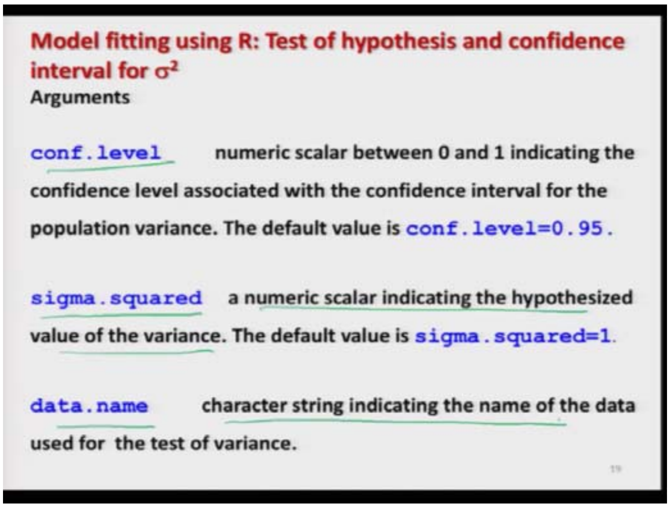
`alternative` character string indicating the kind of alternative hypothesis. The possible values are "`two.sided`" (the default), "`greater`", and "`less`".

And then the command here is varTest v a r T e s t, but again you have to be careful that here this capital T is used. So, in the test the first T is in capital letter and then you have to first give the observations on which you want to conduct this hypothesis. And then you have to choose the option alternative is equal to within double quotes two dot sided so that will indicate that your hypothesis is of type not equal to otherwise there are other options, but our interest here is in two sided.

And again you have to give the confidence level which is the value of here one minus level of significance which isα right. And once now you are trying to say here this σ^2 is equal to this 1. This 1 is actually the value of σ^2 upon σ_0 or σ_0^2 . So, this is the ratio which you have to be careful right and then you have to give the data name which we are not using because we are supplying otherwise you can give it.

So, that is what I am trying to explain here x is the numeric vector alternative is the character which is trying to indicate whether your alternative hypothesis is two sided or type of greater than or less than right.

(Refer Slide Time: 36:11)



Model fitting using R: Test of hypothesis and confidence interval for σ^2

Arguments

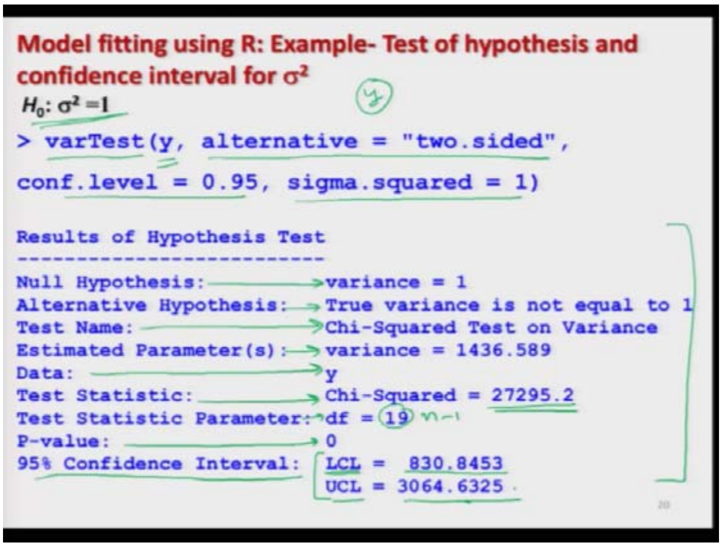
conf.level numeric scalar between 0 and 1 indicating the confidence level associated with the confidence interval for the population variance. The default value is conf.level=0.95.

sigma.squared a numeric scalar indicating the hypothesized value of the variance. The default value is sigma.squared=1.

data.name character string indicating the name of the data used for the test of variance.

And similarly this confidence level is the value between 0 and 1 which is equal to 1 - α, 1 minus level of significance. And sigma.squared is the numeric scalar indicating the hypothetical value of the variance, ok. And data dot name this is the character string indicating the name of the data to be used for the test of variance.

(Refer Slide Time: 36:31)



Model fitting using R: Example- Test of hypothesis and confidence interval for σ^2

H_0: $\sigma^2 = 1$

```
> varTest(y, alternative = "two.sided",
conf.level = 0.95, sigma.squared = 1)

Results of Hypothesis Test
--------------------------
Null Hypothesis:             variance = 1
Alternative Hypothesis:      True variance is not equal to 1
Test Name:                   Chi-Squared Test on Variance
Estimated Parameter(s):      variance = 1436.589
Data:                        y
Test Statistic:              Chi-Squared = 27295.2
Test Statistic Parameter:    df = 19
P-value:                     0
95% Confidence Interval:     LCL =  830.8453
                             UCL = 3064.6325
```

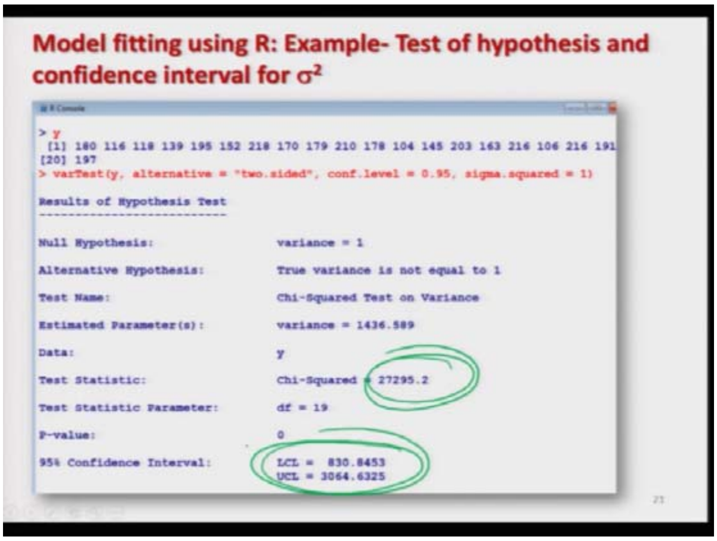
So, now we already have the data on the marks of the student which we had stored here I say y. So, I will try to use this command v a r capital T and then small letter e s t this data on y and my alternative here is two sided. So, you can see here my null hypothesis is now reframed as H_0 σ^2 = 1 ok.

So, confidence level now here is 0.95 and σ^2 σ dot^2 equal to 1; and if you try to see this is the outcome that you will get. So, here it is the first line is trying to give you that what is your null hypothesis; that means, variance is equal to 1 then it is trying to give you the alternative hypothesis which is not equal to 1 then it is trying to give you, what is the name of the test? This is the chi square test on variance because there are different types of chi square test ok.

And what is here estimated parameter it is trying to estimate the variance right. On the basis of given set of data then what is the data which you have used here this is y what is the test statistics this is chi square and its value is coming out to be 27295.2 right. And what are the corresponding degrees of freedom? The degrees of freedom were n - 1 n is 20 so this is 19 and p value which you have which you would like to know it is 0.

So, once again you can see here that this that you have to compare the p value with the value ofα and you have to take an appropriate decision. And then the 95 percent confidence interval is given by here like this so, the lower confidence limit is 830.8453 and the upper confidence limit this is UCL this is 3064.6325 right.

(Refer Slide Time: 38:30)

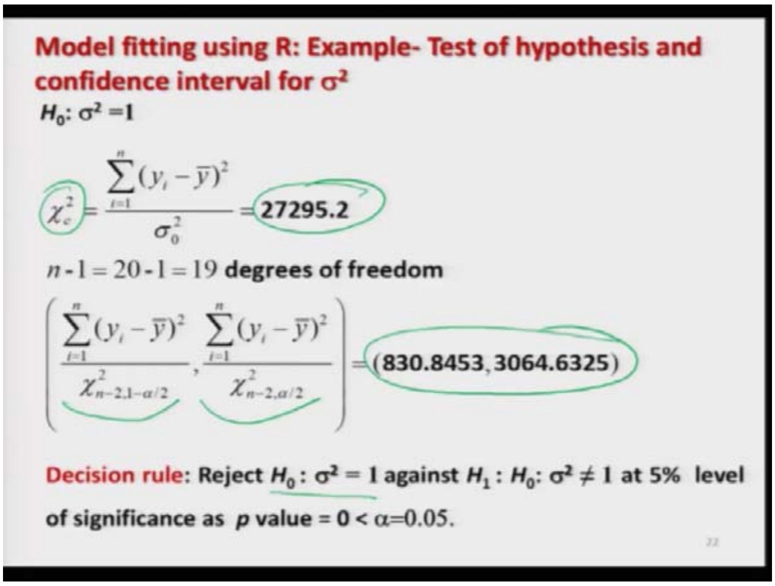


Model fitting using R: Example- Test of hypothesis and confidence interval for σ²

```
> y
 [1] 180 116 118 139 195 152 218 170 179 210 178 104 145 203 163 216 106 216 191
[20] 197
> varTest(y, alternative = "two.sided", conf.level = 0.95, sigma.squared = 1)

Results of Hypothesis Test
--------------------------

Null Hypothesis:                 variance = 1

Alternative Hypothesis:          True variance is not equal to 1

Test Name:                       Chi-Squared Test on Variance

Estimated Parameter(s):          variance = 1436.589

Data:                            y

Test Statistic:                  Chi-Squared = 27295.2

Test Statistic Parameter:        df = 19

P-value:                         0

95% Confidence Interval:         LCL =  830.8453
                                 UCL = 3064.6325
```

21

(Refer Slide Time: 38:36)



Model fitting using R: Example- Test of hypothesis and confidence interval for σ²

H_0: $\sigma^2 = 1$

$$\chi_c^2 = \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\sigma_0^2} = 27295.2$$

$n-1 = 20-1 = 19$ **degrees of freedom**

$$\left(\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\chi^2_{n-2,1-\alpha/2}}, \frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{\chi^2_{n-2,\alpha/2}}\right) = (830.8453, 3064.6325)$$

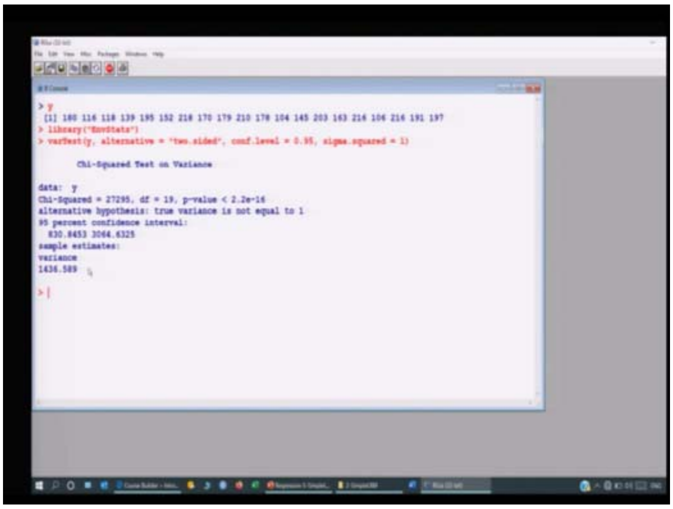
Decision rule: Reject H_0 : $\sigma^2 = 1$ against H_1 : H_0: $\sigma^2 \neq 1$ at 5% level of significance as p value = 0 < α=0.05.

22

So, and then here is the screenshot of the same thing what I have shown you here and now based on that you can take a conclusion how? This chi square c value this is coming out to be 27295.2 which is here you can see here this value right. And then we have here lower confidence limit upper confidence limit which are coming out to here like this which are indicated here right.

So, based on that you can take a conclusion since p value is equal to 0 which is smaller thanα So, your H_0 : σ^2 = 1 is rejected right.

(Refer Slide Time: 39:20)



```
> y
 [1] 180 116 118 139 195 152 218 170 179 210 178 104 145 203 163 216 106 216 191 197
> library("EnvStats")
> varTest(y, alternative = "two.sided", conf.level = 0.95, sigma.squared = 1)

        Chi-Squared Test on Variance

data:  y
Chi-Squared = 27295, df = 19, p-value < 2.2e-16
alternative hypothesis: true variance is not equal to 1
95 percent confidence interval:
  830.8453 3064.6325
sample estimates:
variance
1436.589

> |
```

So, this will give you an idea about the variability now let me come to the R console and try to show you it on the. So, you can see here my y that I already had entered is given like this and my this package environment stats that is already on my computer so I am just uploading it and then I am using the command here for the variance test. You can see here this is now here the outcome.

So, you can see here this is here your 95 percent confidence interval and sample estimates are here given to be here like this. Well, you can see here there is some difference in the structure of outcome that is because of the different versions of R, but the outcome is the same.

So, this is giving you value; value of here the sample estimates of variant and this is here the confidence interval and this is here you can see here the p value is here given to be like this right. And yeah the value of chi square statistics is coming out to be here like this right. So, you can see here that it is not really actually difficult to get it done right ok.

So, now let me finish this lecture. So, now, we have learnt how to obtain the point estimates of regression parameter as well as variances; variance σ^2 . How to obtain the confidence interval, and how to conduct the test of hypothesis for the regression coefficient as well as for the σ^2 ? We also have now understood how to obtain the fitted model fitted value residuals and those things.

So, now I will stop with the chapter on simple linear regression modeling with the brief description of these concepts. Well these concepts are brief, but they are very important their real use will come when we try to use them in the next chapter on multiple linear regression model.

So, my advice to all of you will be that you try to take any data set from the book from the chapter of simple linear regression model just take a small data set that you can enter from your hands and you can see what is happening you try to make plots, try to estimate them, try to obtain confidence interval, try to conduct test of hypothesis and see what you are thinking is that really happening?

And more you practice you will have a more understanding of the process and the data that how looking at the values of the software outcome you can make a correct interpretation which is the soul of statistics. It is just like by looking at an X ray or say MRI report or ECG report the doctor's different doctors can make different types of conclusion.

And we always say that the doctors learned by their experience and a doctor is a better if he is more experienced so more they practice they become a better doctor same is with you also now. More you practice with these data sets they will give you insight into the data what is happening as we said that the data is deaf and dumb data cannot raise the hand and can tell you I have this information I have this type of model, but this is the tool by which you can extract the information.

Now, you have now you know data had never told you that what is the relationship between x and y, but now you know what is the relationship. And what are the different type of information which is contained inside the data? Which has been obtained not by looking at the horoscope or the lines of the hands, but they are based on pure statistical tool.

And now I do not need to explain you the importance of statistical modeling in data sciences everybody knows. So, these small things a small concept will help you in taking bigger decisions. So, you practice and I will see you in the next lecture with a new topic on multiple linear regression model, till then good bye.