

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 - 51
Multiple Linear Regression Analysis
Goodness of Fit and Implementation in R Software

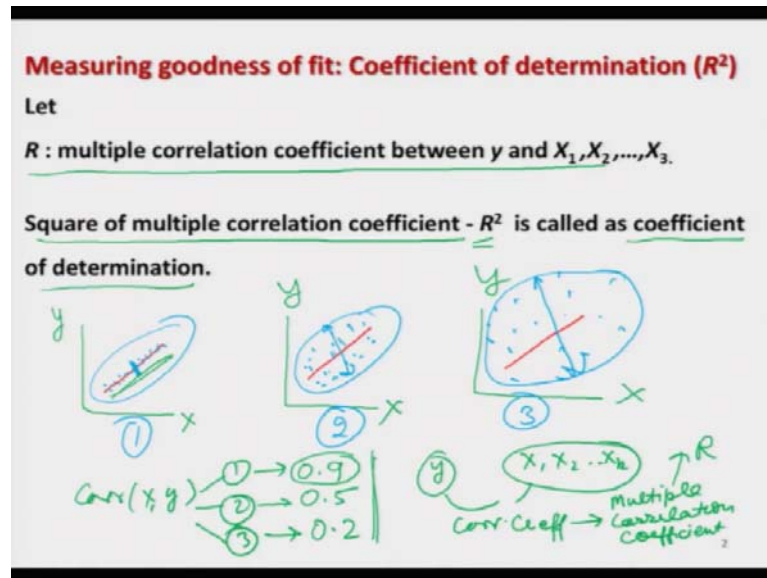
Hello friends, welcome to the course Essentials of Data Science with R Software-2, where we are trying to learn the topics of Sampling Theory and Linear Regression Analysis. In this module on Linear Regression Analysis we are going to continue with the Multiple Linear Regression Analysis with R Software. So, now, we have completed the estimation of parameter, that is model fitting. We have completed confidence interval estimation, point estimation, test of hypothesis and analysis of variance.

So, now after doing all these steps, suppose you get a data you estimate your parameters either using point estimation or interval estimation. Using ordinary least square estimator or maximum likelihood estimator whatever you want whatever you feel is appropriate. You have obtained the fitted model, you have conducted the test of hypothesis using analysis of variance as well as on the individual regression coefficients and you have identified, that which are the variables which are important.

Now, you have finally, obtained the final model using those important variables. So, now, you have a reason to be happy that you have obtained a good model. But, my question is how can you say that the model is good or bad? So, in this lecture I am going to introduce a statistic which is called as coefficient of determination or multiple correlation coefficient, which gives you an idea whether the model which you have fitted on the basis of selected independent variable is this really good or bad.

So, the question comes how would you say the whether the model is good or whether the model is bad? How to say? So, for that first we need to understand the concept and you will see this is a very simple thing and you already have seen this outcome in the software, right. So, let us try to begin our lecture with here this thing.

(Refer Slide Time: 02:23)



Now, before I go into the details of anything let me ask you a more simple question. Suppose, you have only two variables, one is here X another here is y . And suppose you have three possible data sets and you fit here a line like this. Suppose the line is the same. Now, you try to look at the concentration of the data around the line, here the points are very close to the line, here there are more scattered points on the X and y they are the same and so on.

And here the scatteredness is more something like this. Now, in what case, let me call it as case number 1, 2 and here 3, do you think that the model will be good. So obviously, you can see that in the case number 1, the points are lying very close to the line so that means, you can imagine that the fitted model should be good. Whereas, in the 2nd case, the points are quite away from the fitted line.

So, but in the 3rd case, they are quite away most of the points are far away from the fitted line and you can see here is the variation like of this amount, in the 2nd case the variation is this amount and the 3rd case the variation is even very small here only this thing, right.

So, now the question is this now visually you can see and once you are trying to use the multiple linear regression model, then these type of scatter plots are difficult to obtain. So, ultimately you need something a value which can indicate you whether the fitted

model is like 1, 2 or case 3. So, we try to understand how to develop the goodness of fit. So, if you try to use the concept of correlation coefficient, then we tried to find out the correlation coefficient between X and y. So, in case number 1, you can see that the points are very very close to the line. So, you can expect the point that the correlation is 0.9 assume suppose. In the case number 2, the points are quite far away from the line. So, you can assume the correlation coefficient is suppose close to 0.5 and in the 3rd case, the correlation coefficient is will be something 0.2.

So, you can see here that you are saying that whichever model enrich the correlation coefficient is higher that can be considered as a good model because the points are lying very close to the fitted line, right. But now, you have a different situation, this is a correlation coefficient which is between one independent and one dependent variable.

But now, what is the situation you have here one dependent variable but there are more than one independent variable and you want to find out the correlation coefficient between a variable and a group of variable. So, how to find it out? So, in order to measure such a correlation coefficient we have a concept of multiple correlation coefficient.

So, multiple correlation coefficient measures the correlation coefficient between a variable y and a group of variable say X_1, X_2, \dots, X_k , ok and this is in the case of linear regression model, this multiple correlation coefficient is denoted by say capital R, ok. So, if R is the multiple correlation coefficient between y and X_1, X_2, \dots, X_k then we try to take the square of this multiple correlation coefficient that is R^2 and we call it as coefficient of determination.

And this value coefficient of determination can be used to just the goodness of fit of the model. That means, on the basis of given set of data whatever model you have obtained, the value of R^2 will indicate whether the model is good or bad. How to do it? This is what we are trying to now understand, right.

(Refer Slide Time: 07:07)

Measuring goodness of fit: Coefficient of determination (R^2)
 The value of R^2 commonly describes that how well the sample regression line fits to the observed data. This is also treated as a measure of goodness of fit of the model.

So, the value of R^2 commonly describes how well the sample regression line fits to the observed data. And this is treated as a measure of goodness of fit of the model, right. I so, I hope I am now clear.

(Refer Slide Time: 07:22)

Measuring goodness of fit: Coefficient of determination (R^2)
 Assuming that the intercept term is present in the model as

$$y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_k X_{ik} + \varepsilon_i, i = 1, 2, \dots, n$$

 Observe that the number of explanatory variable is $k - 1$ and an intercept term.
 R^2 is defined as

$$R^2 = \frac{y'X(X'X)^{-1}X'y}{y'y}, \quad 0 \leq R^2 \leq 1$$

Handwritten notes and derivations:

- $-1 \leq R \leq 1$
 $0 \leq R^2 \leq 1$
- $SS_{total} = SS_{regression} + SS_{res}$
- $SS_{res} = \text{as small as possible}$
- $SS_{regression} = \text{as large as possible}$
- $\frac{SS_{regression}}{SS_{total}} + \frac{SS_{res}}{SS_{total}} = 1$
- $R^2 = \frac{SS_{regression}}{SS_{total}} = 1 - \frac{SS_{res}}{SS_{total}}$
- $= 1 - \frac{e'e}{\sum_{i=1}^n (y_i - \bar{y})^2}$

So, now, we try to understand the structure of this R square. Well, I am not going to give you the proofs but I will try to show you the interpretation petition and use. So, we consider here a model $y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \dots + \beta_k X_{ik} + \varepsilon_i, i = 1, 2, \dots, n$.

One very important thing which you always have to keep in mind without forgetting that when we are trying to define the coefficient of determination, the model should have the intercept term. We assume that the intercept term is present in the model and if the intercept term is not present in the model please do not use it, right.

I will try to come back on this topic and I will try to give you more details, but in the beginning itself I am telling you that you always have to consider a model with an intercept term, only then you can use the value of R^2 . Otherwise you do not use the value of R^2 do something else, ok. So now, in this case if you try to see I have taken here X_k so; that means, the total number of explanatory variables are here X_2, X_3, \dots, X_k which are actually $k - 1$ in number and there is an intercept term.

So, the total number I am keeping as here k , right. So, now, the coefficient of determination R^2 is defined as, you can see here first in this thing which is more simple to understand. Now you have done the regression analysis and in its analysis of variance. So, under analysis of variance do you remember you had done the relationship sum of square due to total is equal to sum of square due to regression + sum of square due to error that is residuals.

So, what we try to do here, we try to you could see in this case if you try to see a model will be good, only if SS residual that is the contribution of random error should be as small as possible. And what you want if a model is good then the sum of regression should be as large as possible. So, what we try to do here, I try to define here say SS regression divided by SS total, I try to divide the entire equation by $SS_{\text{total}} + SS_{\text{res}}$, residual divided by SS_{total} is equal to 1, right.

So obviously, now the sum of square due to total has been partitioned into two components. So, each of this individual proportion will lie between 0 and 1. And that will indicate that if in the model this ratio, this ratio sum of square due to regression divided by total, if this is large then the model is good and if the ratio of sum of square due to residual.

And sum of square due to total is small then also the model is good. Whereas, on the opposite side, if sum of square due to residual divided by sum of square due to total is high that mean the model is bad because the model is heavily dependent on the random

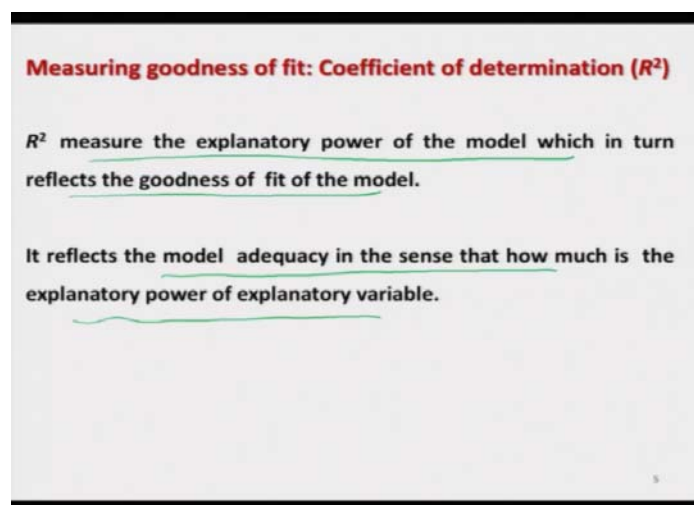
error which are beyond your control, right. And in that case, you can see if the sum of square due to regression divided by sum of square equal due to total is large then; obviously, the second component SS_{res} divided by SS_{total} becomes say smaller.

So, this goodness of fit actually is defined in terms of analysis of variance components, as sum of square due to regression divided by sum of square due to total. And if you remember, you had obtained the expression for the sum of square due to regression and sum of square due to total. So, if you try to use those components who are here you can get this expression, right, ok.

And this same quantity can be expressed at $1 - \frac{SS_{res}}{SS_{total}}$ and if you try to replace the values of SS_{total} you can obtain here like this. So, one thing you have to keep in mind that here I am trying to express this $y'y$ here as $\sum_{i=1}^n (y_i - \bar{y})^2$, right.

Because that is your sum of square due to total. So, essentially you are trying to centre the observations, right. And this value of R^2 this will be lying between 0 and 1. You can see there are different reasons because number one R is a multiple correlation coefficients. So, this R^2 will always lie between 0 and 1, because you know that if the correlation coefficient R lies between say here -1 and 1 . So, R^2 will be lie between 0 and 1, right.

(Refer Slide Time: 12:47)



So, now this has got very simple and interesting interpretations, right. This R^2 actually measures the explanatory power of the model and which in turn actually reflects the goodness of fit of the model. It reflects the model adequacy in the sense that how much is the explanatory power of the explanatory variables. That, how much stronger are our explanatory variables which are capable of explaining the variation in the fitted model, right.

Ideally you assume that whatever process you have consider all the independent variable, whatever you have chosen they are all going to explain the variation in the outcome, right.

(Refer Slide Time: 13:28)

Measuring goodness of fit: Coefficient of determination (R^2)

The limits of R^2 are 0 and 1, i.e., $0 \leq R^2 \leq 1$

- $R^2 = 0$ indicates the poorest fit of the model.
- $R^2 = 1$ indicates the best fit of the model.
- $R^2 = 0.95$ indicates that 95% of the variation in y is explained by the explanatory variables.
- In simple words, the model is 95% good.
- Similarly any other value of R^2 between 0 and 1 indicates the adequacy of fitted model.

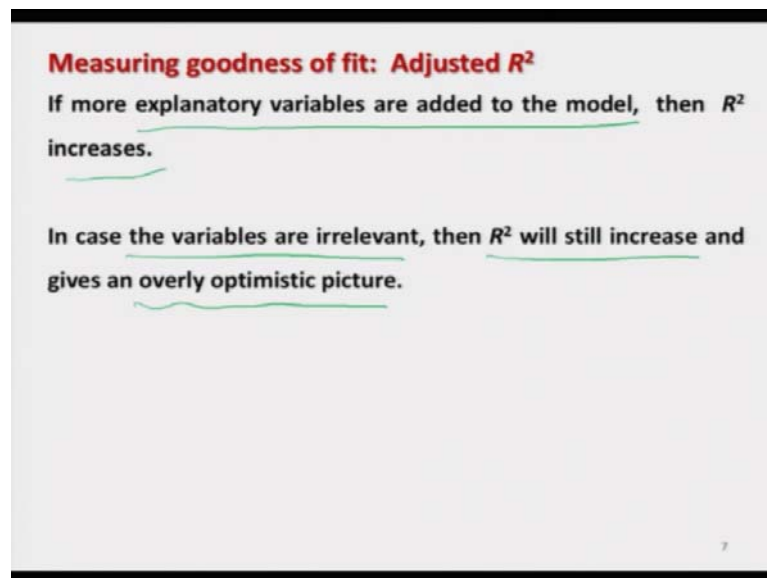
And this R^2 has very simple interpretations. The limits of R^2 are 0 and 1. So, $R^2 = 0$. Now you can very easily understand it that with the correlation coefficient is 0, that means, there is no relationship. So, if you are getting the value of R^2 close to 0, then this is going to indicate the worst fit of the model, poorest fit of the model.

And if you are getting R^2 equal to 1, then this indicates the best fit of the model, as if now you have achieved the model that was created by God it looks like this. And similarly, if you are getting any other value of R^2 between 0 and 1, for example, if you are getting R^2 equal to 0.95, in very simple words I can say that R^2 equal to 0.95 is indicating that the model is 95 percent nearly 95 percent good. Why?

Because R^2 equal to 0.95 is indicating that 95 percent of the variation in y is being explained by the chosen explanatory variable X_2, X_3, \dots, X_k . So, that is indicating that the success of the choice of your explanatory variable.

If your explanatory variables are good they are relevant then your R^2 is going to be higher and that is obvious if you try to give bad input, you will get a bad output. Garbage in, garbage out, right. Similarly, if you try to take another value of R^2 between 0 and 1, that will try to indicate the adequacy of the fitted model, ok.

(Refer Slide Time: 15:10)



Now, before I go further go let me try to adjust few things. The first thing is this we have assumed that whenever you are trying to use the R^2 then the model should have an intercept term in. Now the question is why do I need it? Actually, when we try to define the R square, well you see the R^2 is not coming from the sky but we are trying to find out the expression for the multiple correlation coefficient and we are trying to squaring it.

So, when we try to find out the multiple correlation coefficient under the multiple linear regression model. So, there are certain conditions which have to be satisfied and those conditions are going to be satisfied only when we have an intercept term in the model and these conditions are actually coming from the normal equation.

That you once when you try to find out the estimators, for example, you have you are fitting this model on the basis of ordinary least square estimator estimation. So that

means, whatever OLSEs you have obtained, this R^2 have to satisfy those conditions. So, those conditions are going to be satisfied only when there is an intercept term in the model. So, that is why R^2 is usable only when there is an intercept term in the model.

The next question comes suppose there is a situation where we have where we are compiled to take a model without intercept term? Then in those cases you can choose all other things, but do not choose R square. That is my simple answer. Now, your next question will what are the things which you can choose? Well, people have tried to define some ad-hoc measures for the model without intercept term.

But, they are not really dependable you cannot depend on them much. So, in those cases we try to look for other types of characteristics of the model and we try to take a final column, that is the first thing. The second thing is which I would like to address, now you can understand that if R^2 equal to 1, that is the best fit because there is no random variation.

But, when you come to real life, do you think that can you get a realistic model where the $R^2 = 1$. I wish you get it, but at least in my life up to I have not got any such model where R^2 is so high. Because I personally believe that if R^2 is so high, you are close to God and you are trying to control the process which God is controlling.

But, many times you will see that people are reporting the values of correlation coefficients, multiple correlation coefficient to be very high. So, there you have to be careful. I am not saying at all that they are wrong or, right I am not commenting on that, but I am simply asking you to be careful when you try to use it.

Now, if you ask me that what is the value of R^2 which is acceptable and should not be as low or should not be very high. Then I will say this is the decision which a data scientist has to take. That by looking at the experience so you try to fit your model, try to do some forecasting, try to prediction and try to see the residuals. If there are residuals are good then you can possibly assume that model is reasonably well fitted. There can be different types of problems in the model try to remove them.

And finally, whatever data you have try to fit the model and you can believe on it this is the best possible value that you are getting. But, definitely if it is too low, it is not

advisable to use the model, but it is possible that there is some other relationship. You are assuming here a linear relationship and remember multiple linear regression model is going to measure the degree of linear relationship only. If there is a non-linear relationship, then R^2 cannot be used.

So, now in these slides I will try to show you that R^2 is a very important measure it is very helpful but there are several limitations, right. So, let us try to see the possible issues with the R square, ok. So, now before I move further you can see here, in this expression of R^2 try to see my pen in red color there is also here $e'e$.

Now, if you remember e was your residual. So, that is what I was trying to say that if your if in a model residuals are higher then you cannot say that the model is good. And this thing is also reflected in the definition of R^2 , right, ok. One of the deficiency in this or the characteristic of R^2 is that if more number of explanatory variables are added in the model, then R^2 increases.

Now the question is this, R^2 will always increase as soon as you add more independent variable. Now, there are two situation that you are trying to add some relevant variable or you are trying to add some irrelevant variable.

So obviously, when you are trying to add some irrelevant variable, then the model will be getting bad, but R^2 will still increase and that will possibly indicate that the model is getting better and better and it will give us an overly optimistic picture. So, you have to be very careful when you are trying to add a variable in the model. That you have to add only an important variable which you already have checked.

(Refer Slide Time: 21:06)

Measuring goodness of fit: Adjusted R^2

With a purpose of correction in overly optimistic picture, adjusted R^2 , denoted as \bar{R}^2 or $\text{adj } R^2$ is used which is defined as

$$\bar{R}^2 = 1 - \frac{SS_{\text{res}} / (n - k)}{SS_{\text{total}} / (n - 1)}$$

Anova table

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

In order to take care of this problem this R^2 has been modified. And a version of R^2 which is called as adjusted R^2 is proposed. This adjusted R^2 is indicated by \bar{R}^2 or this is called as adj that means, adjusted R square. So, actually this R square, this adjusted R^2 is a function of R^2 and which has been given by like this.

So, you can see here, this sum of square due to residual is divided by $n - k$ and a sum of square due to totally divided by $n - 1$. Now you are intelligent enough to understand what are these things. Now you can recall your Anova table. And try to recall that what are the degrees of freedom of sum of square due to residual and total.

So, essentially we are dividing the sum of square due to residual and total by the respective degrees of freedom and we are trying to define a modified version of R square, which has a property that it will take care. That if some variables are added in the model then the value of adjusted R^2 will increase, but the magnitude of increase will be much lower than the corresponding increase in the value of R square.

So, now if you try to substitute either the value of R^2 then your \bar{R}^2 or the adjusted R^2 can be express here like this. So, you can see here this is here the function $1 - R^2$. So, actually this R^2 and adjusted R^2 both are used to judge the goodness of the fit model. I will try to show you in the software outcome but this adjusted R^2 also has some problem, right.

(Refer Slide Time: 22:50)

Measuring goodness of fit: Adjusted R^2

The adjusted R^2 will decline if the addition of an extra variable produces too small a reduction in $(1 - R^2)$ to compensate for the increases .

Another limitation of adjusted R^2 is that it can be negative also.

For example if $k = 3, n = 10, R^2 = 0.16$ then

$$\bar{R}^2 = 1 - \frac{9}{7} \times 0.84 = -0.08 < 0$$

which has no interpretation.

And the property of adjusted R^2 is that this adjusted R^2 will decline if the addition of an extra variable produces too small a reduction in the value of $1 - R^2$ to compensate for the increase. Because you can see here, that here we have got the term here $1 - r$ square, ok.

One limitation of adjusted R^2 is that it can be negative also. This is a very serious drawback because this is a squared value. So, how it can be negative, but it is possible for example, if I suppose take these values k equal to 3, $n = 10$ and suppose $R^2 = 0.16$, then you can compute \bar{R}^2 which will come out to be -0.08 .


Well, this is possible theoretically but in practice if such type of condition arises, then my question to you will be that why you are not looking at the value of R square? This is 0.16, this is extremely low. So, that is that value itself is indicating that possibly a linear model cannot be fitted over here and some other model needs to be to be here.

So, now once you are trying to unnecessarily fit and inappropriate model, then how the model can be well fitted, right. But in such cases theoretically I say yes this is possible and in this square will have no interpretation.

(Refer Slide Time: 24:29)

Measuring goodness of fit: Limitations

1. If constant term is absent in the model, then R^2 can not be defined. In such cases, R^2 can be negative. Some ad-hoc measures based on R^2 for regression line through origin have been proposed in the literature.
2. R^2 is sensitive to extreme values, so R^2 lacks robustness.



10

Now, some limitation of this R square, as well as this \bar{R}^2 . First, I already have told you that if constant term is absent in the model then R^2 cannot be defined. And in such cases you will find that in sometimes in real life the R^2 can be negative which is not possible because this is a square of the multiple correlation coefficient.

So, in order to handle such condition as I said some ad-hoc measures based on R^2 for regression line passing through origin have been proposed in the literature, but they are not very good or they are not very much dependable. So, it is difficult to just the goodness of it in such a case and R^2 is sensitive extreme value.

So, for example, if you have a data set which are lying over here along the line and suppose there is some value which is coming out to be here. So, and if you try to add this data into your model and try to refit your model, your line will go from somewhere here and then the R^2 will change drastically. So, that is what I am trying to say that R^2 is sensitive to the extreme values. So, R^2 is not robust.

(Refer Slide Time: 25:42)

Measuring goodness of fit: Limitations

3. Consider a situation where we have following two models:

$$y_i = \beta_1 + \beta_2 X_{i2} + \dots + \beta_k X_{ik} + u_i, \quad i = 1, 2, \dots, n$$

$$\log y_i = \gamma_1 + \gamma_2 X_{i2} + \dots + \gamma_k X_{ik} + v_i$$

The question is now which model is better?

For the first model,

$$R_1^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

and for the second model, an option is to define R^2 as

$$R_2^2 = 1 - \frac{\sum_{i=1}^n (\log y_i - \log \hat{y}_i)^2}{\sum_{i=1}^n (\log y_i - \log \bar{y})^2}$$

As such R_1^2 and R_2^2 are not comparable.

Now, suppose I try to take here two models and try to show you here another type of trouble. One thing what you have to keep in mind that whenever you start whenever you start preparing a model then you have only one thing in your hand, data on X and y. The data they does not speak and tell you that I have got this relationship or I have got that a relationship or I have these many important variable or my this variable is not important.

Data is simply different term. So, now, you are the only one who has to think about the form of the model and before that some time you also have to change the explanatory variable. So, that you can make the model to be linear. So, suppose you get some data and two different persons try to fit two different models like this.

First person, try to fix fit the data model on the given data X_1, X_2, \dots, X_k and another person tries to take the log of y_i . And then the model is fitted using the log y and then correspondingly the model parameters I am trying to indicate by different symbol $\gamma_1, \gamma_2, \dots, \gamma_k$. Now, the question here is if you want to find out which of this fitted model is better?

When you can use any least square estimation or maximum likelihood estimation you can obtain the values of β hats and γ hats using the same classical way. And after that your question comes out to be which model is better? Think about it, right. So obviously,

you will say that, ok I will try to find out the value of R^2 and whichever values higher I will say that the that model is better fitted.

So, if you try to find out the R^2 for the first model this is straightforward, this can be obtained from here whatever we have discussed. Now, the question is for the second model. So, one option one ad-hoc solution is this I can define the R^2 for the second model called as R_2 square.

Something like by instead of using here y_i and \hat{y}_i , I will try to use here log of y and log of \hat{y}_i and here you have y and y_i and y ba \bar{y} r. So, I will be using here log of y_i and log of \bar{y} , but they are not comparable actually and it is not even appropriate to just use a log of y_i in place of y_i or log of \hat{y}_i in place of \hat{y}_i , or even use log of \bar{y} in place of \bar{y} .

(Refer Slide Time: 28:31)

Measuring goodness of fit: Limitations

If still, the two models are needed to be compared, a better proposition to define R^2 can be as follows:

$$R_1^2 = 1 - \frac{\sum_{i=1}^n (y_i - \text{anti log } \hat{y}_i^*)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

where $y_i^* = \widehat{\log y}$. Now R_1^2 and R_2^2 on comparison may give an idea about the adequacy of the two models.

So, this R_1^2 and R_2^2 which you have obtain here they are not really comparable. Well, if you want to solve it in an ad-hoc way one, option is this, ok you can take here the anti log and you can define here another type of R_3^2 which is based on the classical value of R^2 which is here, right. And you can define here this R_3^2 the only thing is you have to take here anti log of y_i^* which is y_i^* is the estimator of log of y , right log of y_i .

Now, means R_1^2 and R_2^2 on comparison they may. So, now, I can see here that instead of comparing R_1^2 and R_2 square, you can compare R_1^2 and R_3^2 here, right. But, but definitely

this will give you only a sort of ad-hoc comparison and it is very difficult to find it out where which one is the correct model based on the values of R square.

(Refer Slide Time: 29:17)

Example

Observations on 20 students are collected

Let

y : Marks of students (Max. marks: 250)

X1 : Number of hours per week of study,

X2 : Number of assignments submitted per month,

X3 : Number of hours of play per week

Student no.	y	X1	X2	X3
1	180	34	3	15
2	116	12	1	13
3	118	15	3	11
4	139	33	1	10
5	195	31	5	17
6	152	24	1	15
7	218	40	5	18
8	170	31	5	13
9	179	21	2	20
10	210	37	3	19
11	178	29	4	16
12	104	15	1	10
13	145	17	1	16
14	203	38	5	16
15	163	17	1	19
16	216	36	3	20
17	106	13	1	11
18	216	39	5	18
19	191	36	5	15
20	197	34	1	19

Now, I try to take a simple example, same example which I have considered couple of times earlier and I will try to show you how you can obtain these values in the software. So, I am taking the same data set here we have 20 observations on the students and they are the observations on the marks obtained by the students, the number of hours per week the student has studied, number of assignment which a student has submitted per month and the number of hours of play which the student has done and these variables are denoted by y, X1, X2 and X3.

So, for example, if you try to take here 1st student, this means that the student has got 180 marks out of 250 marks. The student studied for 34 hours in a week and the student submitted 3 assignments in a month and the student played for 15 hours in a week. And somehow, and this is the way all the 20 observations have been obtained here.

(Refer Slide Time: 30:23)

Model fitting with R:
Measuring goodness of fit - R^2 and Adj. R^2
Test of hypothesis in fitting linear models
`lm` is used to fit linear models.
`summary` is used to get the results about R^2 and Adj. R^2 along with other results.

Usage
`summary(lm(formula, data,...))$r.squared` gives R^2
`summary(lm(formula, data,...))$adj.r.squared` gives Adjusted R^2

Now, I try to use the same command to fit the model `lm` which we already have used and then I try to obtain the summary of the object which is created from the `lm`. So, whatever is the outcome of this `lm`, I try to use here the `summary` and `summary` will have the outcome of R^2 as well as adjusted R square. But, if you want to just extract the values of R^2 and adjusted R square.

Then also we can do it for that the command is this. Try to use the `summary` command on the `lm` object and then use this dollar sign and try to write down the command `r dot squared` and similarly if you want to extract the values of adjusted R square, use the same command for `summary lm` and then try to join it with the dollar sign and write the command `adj dot r dot squared`. So, that will give you the value of adjusted R square.

(Refer Slide Time: 31:33)

Model fitting with R: Example-
Measuring goodness of fit - R^2 and Adj. R^2
The model for each observation, $n = 20$ as
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i, \quad i = 1, 2, \dots, 20$$

Data
`X1=c(34, 12, 15, 33, 31, 24, 40, 31, 21, 37, 29, 15, 17, 38, 17, 36, 13, 39, 36, 34)`
`X2=c(3, 1, 3, 1, 5, 1, 5, 5, 2, 3, 4, 1, 1, 5, 1, 3, 1, 5, 5, 1)`
`X3=c(15, 13, 11, 10, 17, 15, 18, 13, 20, 19, 16, 10, 16, 16, 19, 20, 11, 18, 15, 19)`
`y=c(180, 116, 118, 139, 195, 152, 218, 170, 179, 210, 178, 104, 145, 203, 163, 216, 106, 216, 191, 197)`

So now, how to get it done? I will try to show you first on the screen and then I will try to come on the R software. So, you can see here I am trying to consider this model for which I have collected data I have enter the data X1, X2, X3 and y on my R console.

(Refer Slide Time: 31:48)

```
Model fitting with R: summary command- R2 and Adj. R2
> summary(lm(y~X1+X2+X3))
Call:
lm(formula = y ~ X1 + X2 + X3)
Residuals:
    Min       1Q   Median       3Q      Max
-2.04524 -0.25493  0.09177  0.37276  1.47180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.76945     1.03065   8.509 2.47e-07 ***
X1           1.99675     0.03228  61.850 < 2e-16 ***
X2           3.91840     0.16679  23.493 7.90e-14 ***
X3           6.10603     0.07234  84.405 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.922 on 16 degrees of freedom
Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
F-statistic: 1.07e+04 on 3 and 16 DF, p-value: < 2.2e-16
```

And after that I try to use the summary command, I already have demonstrated it couple of times. So, I am skipping the details. But now, you have to identify where is this outcome. You can see here in the bottom where I moving my pen you can see here what is this thing.

This is giving you the value of multiple R squared which is 0.9995 and this adjusted R² which is 0.9994. So, one important information for you will always see that the value of adjusted R squared is always smaller than the value of R square.

But, definitely this difference will be very small and in case if this difference is large you can expect that something is wrong in the system and you need to lo, ok into the data and then you have to find it out, right. So, this is here the screen shot and you have to consider on this box which I have made in red color, right.

(Refer Slide Time: 33:53)

Model fitting with R: summary command - R^2 and Adj. R^2

```
R Console
> summary(lm(y~X1+X2+X3))

Call:
lm(formula = y ~ X1 + X2 + X3)

Residuals:
    Min       1Q   Median       3Q      Max
-2.04524 -0.25493  0.09177  0.37276  1.47180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.76945    1.03065   8.509 2.47e-07 ***
X1           1.99675    0.03228  61.850 < 2e-16 ***
X2           3.91840    0.16679  23.493 7.90e-14 ***
X3           6.10603    0.07234  84.405 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.922 on 16 degrees of freedom
Multiple R-squared:  0.9995,    Adjusted R-squared:  0.9994
F-statistic: 1.07e+04 on 3 and 16 DF, p-value: < 2.2e-16
```

So, this is trying to give you the multiple R squared and the value of adjusted R squared. So, you can see that it is close to 0.995 and say 0.994 and so that is why you can say believe that the model is good. But now, let me confess it here that why the cases coming out to be so high because I have taken the data artificially. My objective was to show you that whatever is happening in the data whether the same thing is happening in the statistical software and this outcome as well as the statistical tool.

Whether the statistical tool are capable enough to diagnose the same thing what is present in the data or not. So, that much I can show you, but if the data has any problem then the problem can be due to different reasons for that we have to investigate more, right. So, that is why this values coming out to be very high in practice yeah it is difficult and the second thing is this I have taken a very small controlled data set with only three variables. So, that is why you are getting this value to be here higher, ok.

(Refer Slide Time: 34:08)

Model fitting with R: summary command- R^2 and Adj. R^2
Observe the following outcome

Multiple R-squared: 0.9995, Adjusted R-squared: 0.9994

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 0.9995$$
$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k} \right) (1 - R^2) = 0.9994$$

Residual standard error: 0.922 on 16 degrees of freedom
Multiple R-squared: 0.9995, Adjusted R-squared: 0.9994
F-statistic: 1.07e+04 on 3 and 16 DF, p-value: < 2.2e-16

18

So, now if you try to see how this value have been obtained. You have the value of multiple R squared which is obtained here by to this expression. The ratio of sum of square due to regression and sum of square due to total and the value of adjusted R square this is obtained by this formula of \bar{R}^2 , right and this is indicated here.

(Refer Slide Time: 34:34)

Model fitting with R: summary command- R^2 and Adj. R^2
Observe the following outcome

```
> summary(lm(y~X1+X2+X3))$r.squared  
[1] 0.9995016  
> summary(lm(y~X1+X2+X3))$adj.r.squared  
[1] 0.9994082
```

$$R^2 = \frac{SS_{regression}}{SS_{total}} = 0.9995016$$
$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k} \right) (1 - R^2) = 0.9994082$$

19

So now, I am sure that it is clear that how are you going to get the value of this R square. Now, if you want to extract only the value of R^2 or say adjusted R square, then you have

to use the summary command and you have to add here a dollar sign and in the command r dot squared and you will get here only the value of R square.

And similarly, if you want to extract the value of adjusted R square. So, this value will be obtained by using the summary command and on the lm and then use the dollar sign and then use the command adj dot r dot squared and it will give you directly the value of adjusted R² which are obtained by these two expression.

(Refer Slide Time: 35:23)

```

> summary(lm(y~X1+X2+X3))

Call:
lm(formula = y ~ X1 + X2 + X3)

Residuals:
    Min       1Q   Median       3Q      Max
-2.04524 -0.25493  0.09177  0.37276  1.47180

Coefficients:
(Intercept)  8.76945  1.03065  8.503  2.47e-02 ***
X1           1.99675  0.03228  61.850  < 2e-16 ***
X2           3.91840  0.16672  23.482  1.90e-14 ***
X3           6.10603  0.07234  84.405  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.922 on 16 degrees of freedom
Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
F-statistic: 1.07e+04 on 3 and 16 DF, p-value: < 2.2e-16

> summary(lm(y~X1+X2+X3))$r.squared
[1] 0.9995016
> summary(lm(y~X1+X2+X3))$adj.r.squared
[1] 0.9994082

```

So, you can see here it is not difficult to find these values, right. And here is the screen shot but before we try to go to the R console, let me try to explain you here what you have completed in this outcome which is the summary command giving you all the information and which part is left. So, let me try to say here first line.

This part you know what is this thing, this part you know what is this? This is simply the formula then they there are the values of residuals what are these thing that you also know, what are residuals. And then these are the values which are obtained from the values of residuals which are minimum value, 1st quartile, median, 3rd quartile and the maximum value.

So, you also this part also now there are coefficients. So, you understand this is intercept term, this is first variable X1, then X2. then X3. What are these estimates? You also know you have to simply follow my pen, right, you cannot see me, ok. These estimates

are trying to give you this is intercept term, this is ordinary least square estimator of β_1 , this is for β_2 and this is for β_3 .

So, these estimates are the ordinary least square estimator of β vector and they can also be maximum likelihood estimate because they have the same value, right. What are these thing? This is standard error. So, this is standard error of intercept term b_0 , this is standard error of b_1 , this is for b_2 and this is for b_3 .

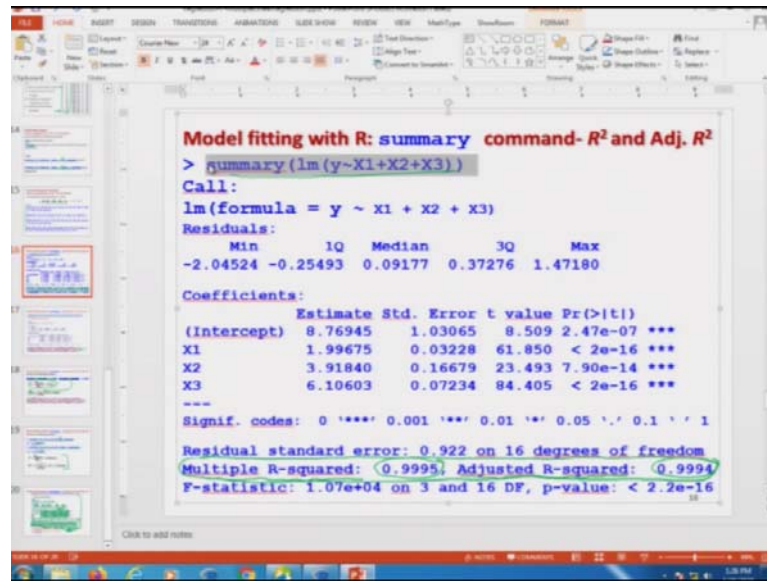
What are this thing t value? Now, you know these are the t statistics corresponding to $H_0 \beta_j = 0$. So, this is corresponding to intercept term this is for β_1 , β_2 and β_3 . What are this thing? These are the corresponding p values. So, you know all these four values.

What about this significance code? You know that these are the values of α . After that the last in this line here residual standard error, you know this is the value of say s then these are the degrees of freedom corresponding to the distribution of s square. Then, also you know what is now multiple R^2 and you also know what is adjusted R squared. Then you also know what is this F statistics, this is relative to your anova, right.

This is the final F statistics which is for the say m s regression divided by m s res, right. So, this is on 3 and 16 degrees of freedom and you can see here this is the p value here. So, it is less than 0.5 at α equal to 5 percent level of significant. So, you can say in general that H_0 say β_1 be or say $\beta_1 = \beta_2 = \beta_3$, this hypothesis is rejected.

So, now you know this value also. So, now, you see means you have completed the entire outcome. So, now, you understand each and everything in this outcome. So, that will finish your multiple linear regression model. But, before that I will try to do these things on the R console.

(Refer Slide Time: 36:10)

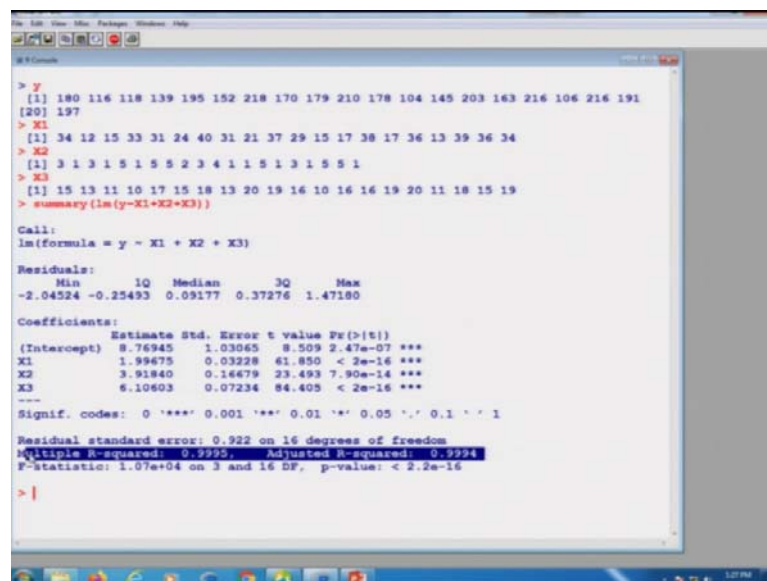


```
Model fitting with R: summary command- R2 and Adj. R2
> summary(lm(y~X1+X2+X3))
Call:
lm(formula = y ~ X1 + X2 + X3)
Residuals:
    Min       1Q   Median       3Q      Max
-2.04524 -0.25493  0.09177  0.37276  1.47180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.76945    1.03065   8.509 2.47e-07 ***
X1           1.99675    0.03228  61.850 < 2e-16 ***
X2           3.91840    0.16679  23.493 7.90e-14 ***
X3           6.10603    0.07234  84.405 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.922 on 16 degrees of freedom
Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
F-statistic: 1.07e+04 on 3 and 16 DF, p-value: < 2.2e-16
```

(Refer Slide Time: 36:11)



```
The R GUI: User: Administrator, Packages: Windows, 10/10/2014
R Console
> y
[1] 180 116 118 139 195 152 218 170 179 210 178 104 145 203 163 216 106 216 191
[20] 197
> X1
[1] 34 12 15 33 31 24 40 31 21 37 29 15 17 38 17 36 13 39 36 34
> X2
[1] 3 1 3 1 5 1 5 5 2 3 4 1 1 5 1 3 1 5 5 1
> X3
[1] 15 13 11 10 17 15 18 13 20 19 16 10 16 16 19 20 11 18 15 19
> summary(lm(y~X1+X2+X3))

Call:
lm(formula = y ~ X1 + X2 + X3)

Residuals:
    Min       1Q   Median       3Q      Max
-2.04524 -0.25493  0.09177  0.37276  1.47180

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.76945    1.03065   8.509 2.47e-07 ***
X1           1.99675    0.03228  61.850 < 2e-16 ***
X2           3.91840    0.16679  23.493 7.90e-14 ***
X3           6.10603    0.07234  84.405 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

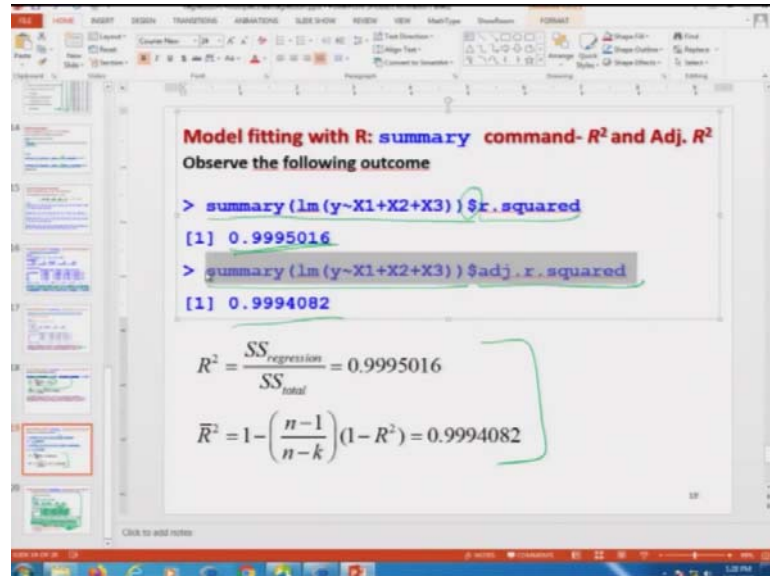
Residual standard error: 0.922 on 16 degrees of freedom
Multiple R-squared:  0.9995, Adjusted R-squared:  0.9994
F-statistic: 1.07e+04 on 3 and 16 DF, p-value: < 2.2e-16
> |
```

So, you can see here I already have entered the values of y, X1, X2 and X3 and so I try to find out here is a summary command. So, you can see here these are the values here, right. I will try to highlight it here. You can see this is the where I am highlighting again and again try to opt see this value, right.

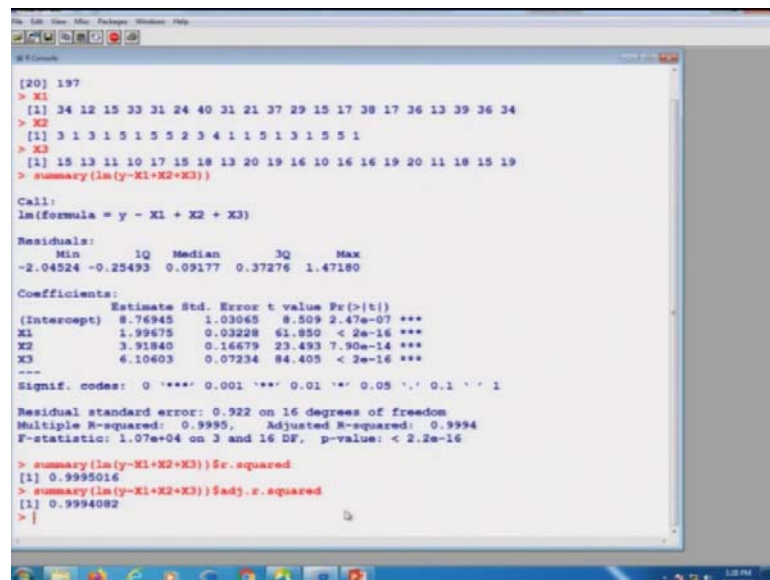
So, these are the values of multiple R² and say adjusted R² which are called in the software. And if you really want to find out only the value of R squared and say adjusted

R squared you have to use cookies command here and you can see here this is the value of 0.995016 and this is here the same thing here.

(Refer Slide Time: 40:05)



(Refer Slide Time: 40:09)



This is, this particular value has been extracted. And if you want to find out the value of adjusted R squared. So, you have to use this command and you can see here this value here is 0.999 this is the same value which is obtained here, right. So, you can see here obtaining these values is not difficult at all.

So, now we have done each and every component of the outcome of a R software when it is when we are trying to fit a multiple linear regression model. But now, there is a difference in your understanding, whatever is the outcome, now you are understanding how it is coming, what is the value which it is trying to estimate you know it and if there is a problem if you feel that there is some problem which is bound to happen when you are trying to deal with data science.

You are dealing with large number of variables, large number of huge data sets that is itself a big challenge that how to handle such a big data set for fitting the multiple linear regression model. So, my advice is this first try to divide your data into homogeneous groups. Now how you can do it? For that, you have to go back to your sampling theory and you need to know what is your certified sampling.

Then if you try to fit a model for every group, within a group then possibly you will get a good model and if you try to fit a model for the entire big data the chances are very less that you will get a very good model, because the variation in such a big data is expected to be very high, unless and until you are fortunate that your data has very less variation.

So, this is my only an advice this may hold or this may not hold in real data set that you have to see. And you will have large number of independent variable that is itself a big issue because every experimenter is trying to choose only the important variable, but the main thing problem is this what multiple linear regression is model is thinking as good variable and what my experimenter is thinking as a good variable these two opinions can be different.

And multiple linear regression model considers only those variable good, which the statistics tells him or what is it, right. So, these are the challenges which will come to you, but since now you understand each and everything by looking at the different value, different types of component you can possibly go back to your data lo, ok into the data, go back to your model and try to lo, ok at the different component and possibly you will be able to get a good model.

And when you are trying to look into the real life data, the data may have different types of problem. The data may violate all the basic assumptions what you have made for multiple linear regression model and based on that you have to use different types of tool

there are graphical tools, there are analytical tools. We have different types of regression also logistic regression, portion regression and so on.

Well, those things are not possible for me to cover here in this course, but definitely with this background I can promise you it will not be difficult for you, if you try to read this topics this chapter from the book. But, now you will read them with a different perspective possibly up to now many of you most of you why might be thinking why should I read the theory, why should I understand the theory.

But, now getting these concepts you understand that these concept will help you in taking a final call whether your data is good or bad and whether you find a model is good or bad. So, on the next turn I will try to see you with a new topic on variable selection and multiple linear regression model is now done.

So, you try to have now look into the all the lectures, all the topics because all the concepts are going to be used at the same time when you are trying to deal with the real data in a multiple linear regression model fitting. So, you enjoy the course, you try to have a practice, try to take data sets and try to experiment with it, try to learn and I will see you in the next lecture once again. Till then, Good bye.