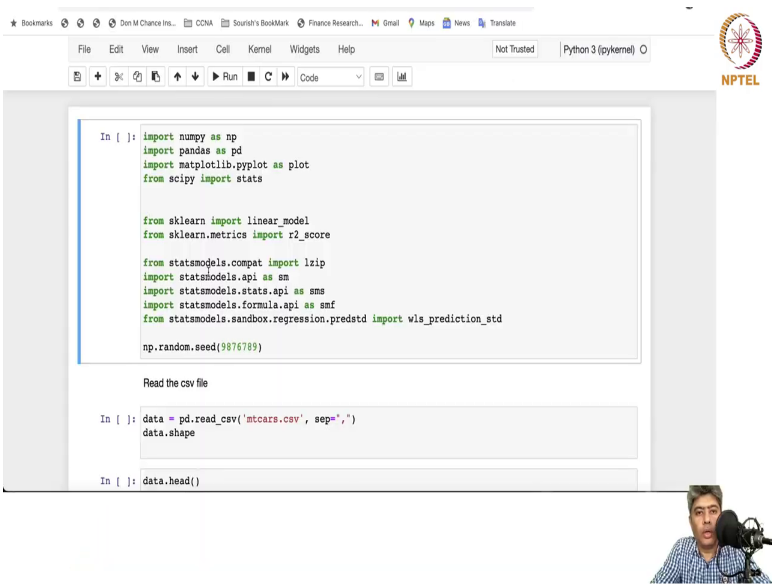
Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Chennai Mathematical Institute

Lecture - 25
Hands-on with Python Part - 3

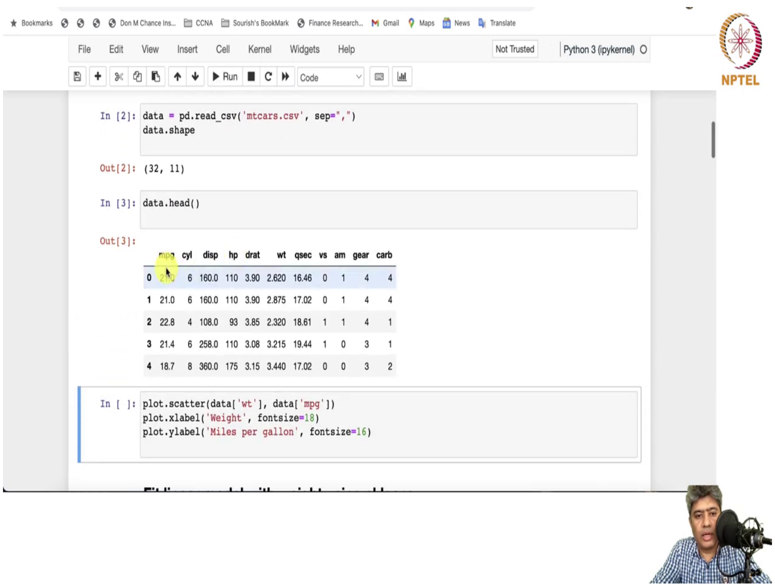
Hello guys, welcome back to hands on for lecture 7, in this hands on we are going to do some regression analysis using Python.

(Refer Slide Time: 00:33)



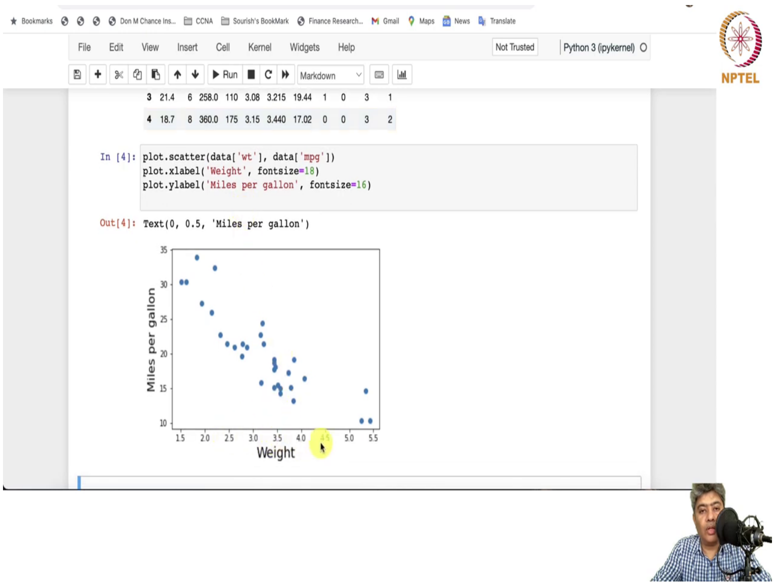
So, let me start my Jupyter Notebook ok; so, in this we are going to use the mtcars dataset and here is the bunch of packages that I have called you know here I have bunch of packages numpy, pandas, matplot, library, scipy from sklearn, we imported linear models; from sklearns matrics, we calculated r2 score. Then some other models from statmodels stats api. So, all these things we have computed, let me run it has no problem.

(Refer Slide Time: 01:30)



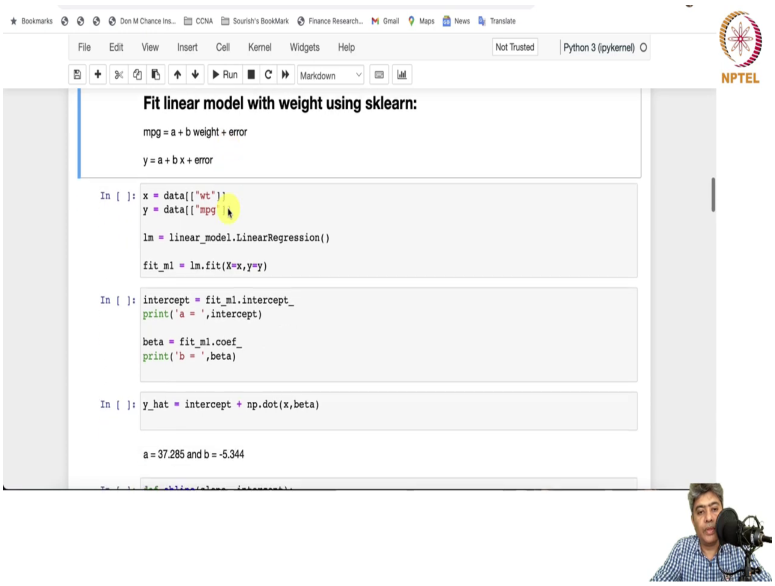
So, next is read the mtcars dataset, we created and it has 32 rows, 11 column as expected and head. So, it has first column is miles per gallon, second column is cylinder, third column is displacement, fourth column is horsepower, rears ratio, weight qsec, v shape or knot. automatic or manual, how many gears it has and carburettor.

(Refer Slide Time: 02:01)



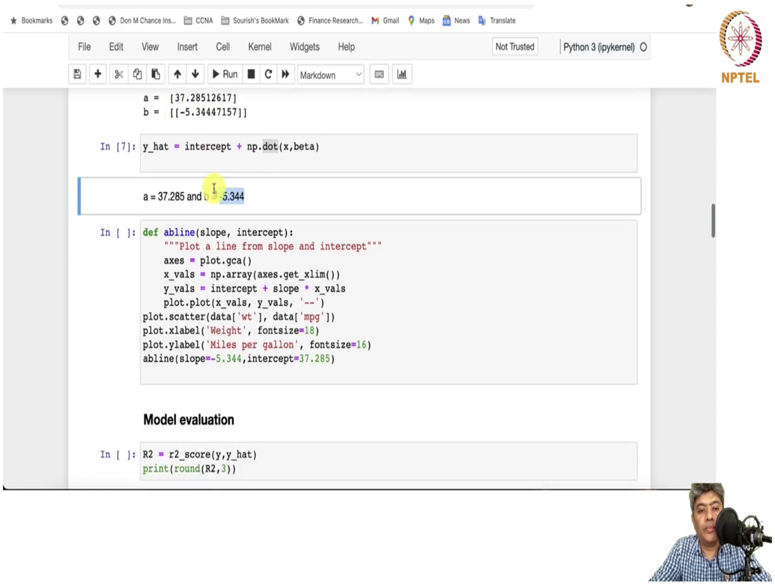
So, the first we are going to plot weight versus miles per gallon and as expected on the x axis we put a weight and a on the y axis we put gallon miles per gallon. And it there is a negative relationship what we are seeing is as the weight of the car increases efficiency of the car drops. So, first we are going to fit a linear model with weight using a sklearn.

(Refer Slide Time: 02:30)



So, first; so, we are going to fit miles per gallon as a function of a plus b times weight plus error. So, x is in the weight and y as the miles per gallon and from the linear model we are calling the linear regression module. Defining it as a lm and then we are going to fit lm fit and x we have to give this x and for all y we have to give this y, and that will be fitted as a first model ok.

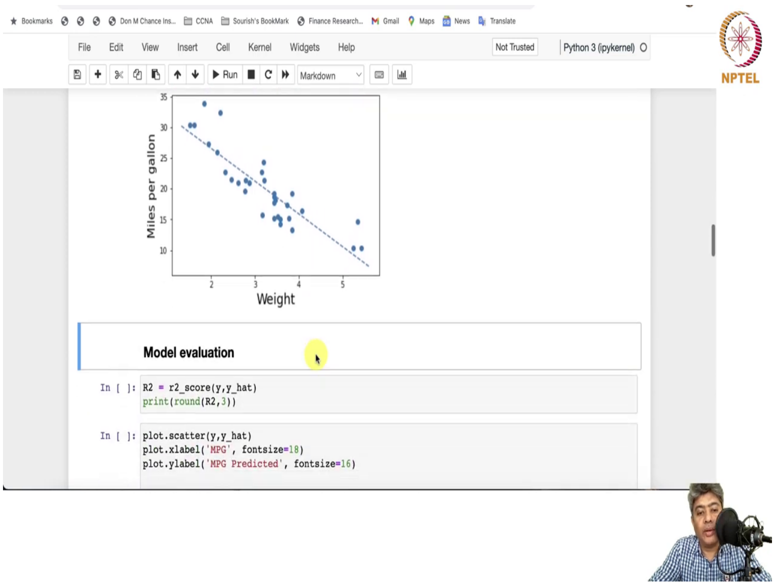
(Refer Slide Time: 03:18)



So, it has done it has done and let from the fit_m1, here it is; we are extracting the intercept and calling it intercept and the coefficient beta and we take it as b equal to a equal to b equal to. So, we got a equal to 37.2 a and b equal to negative 5.344 and we are going to calculate y hat equal to intercept, plus in dot operation we are going to do x into beta.

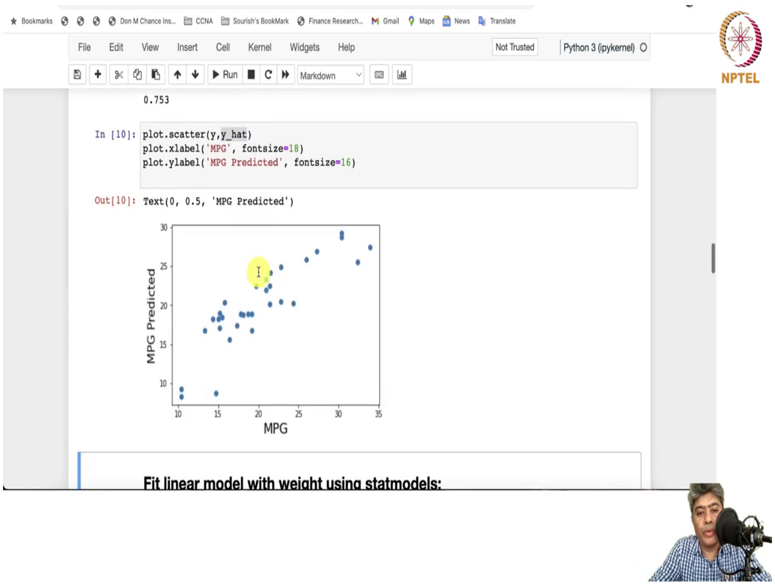
So, but it is a dot operation from; so, from numpai, we are going to call the dot operation. So, we ran it we took a equal to 37.285 and b equal to negative 5.344 from here. And then we are going to draw a straight line abline sometimes called through the plots. So, this we have written this small piece of function which, if you go give the slope value and the intercept value, it will plot the line from slope and intercept over the scatter plot.

(Refer Slide Time: 04:39)



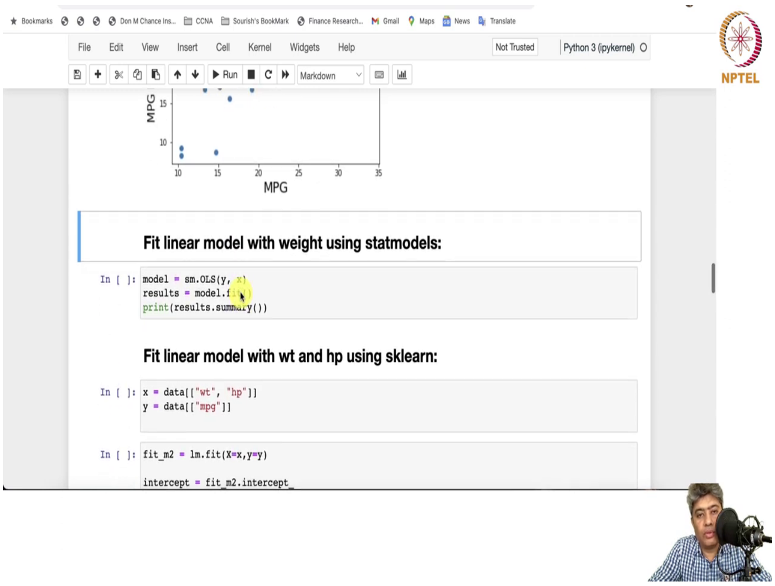
So, we run this; so, over the scatter plot we draw a straight line, and this is the best fitted OLS line.

(Refer Slide Time: 04:50)



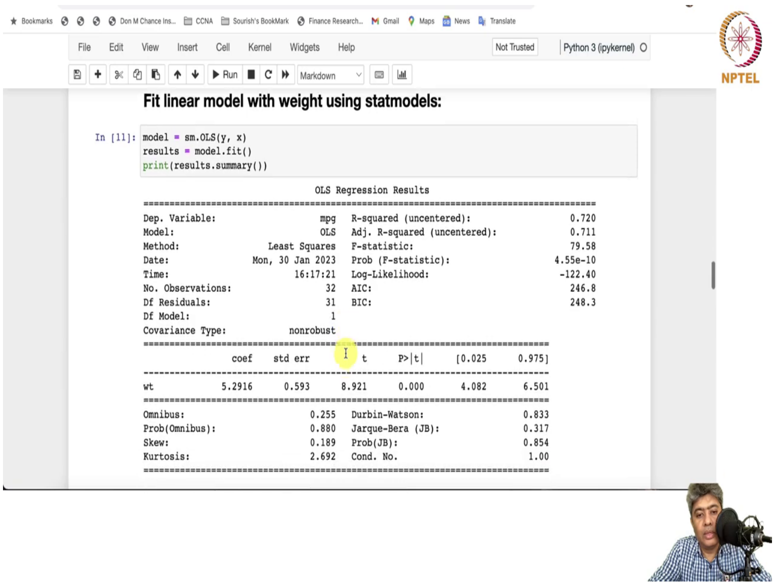
Now, we are fitting a model evaluation; now, we are fitting r square, r square is 0.753. And then from the we have y and y hat; we plot y and y hat, y is the miles per original miles per gallon and y hat is the predicted miles per gallon or estimated miles per gallon. So, we can see that they are positively correlated; that means, the cars which has lower efficiency it is predicted as lower. But definitely there are some variability, more time it will be better the model prediction is we can say.

(Refer Slide Time: 05:36)

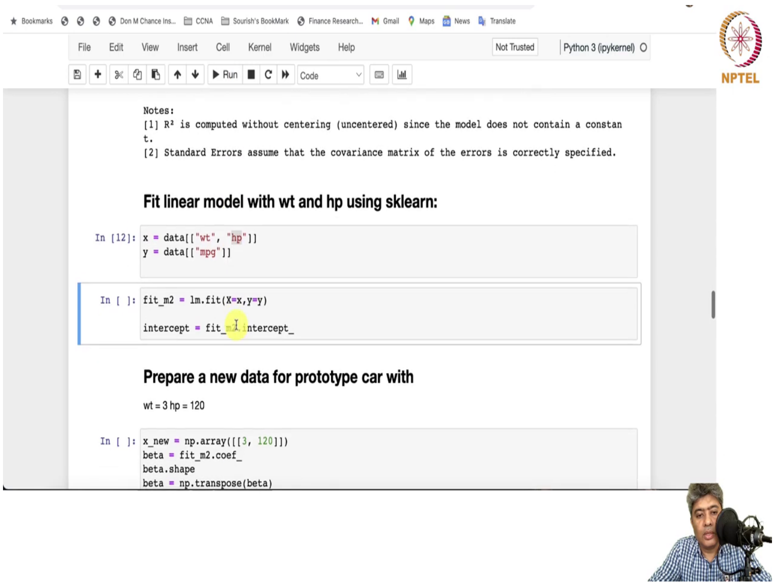


So, now fit linear model using weight using stat model; so, that was we fit this model using; so, when we fit this model in the above yeah, here this was done using sk learn. Now, we are going to fit the same model using stat models; so, sm OLS, sm dot OLS.

(Refer Slide Time: 06:10)

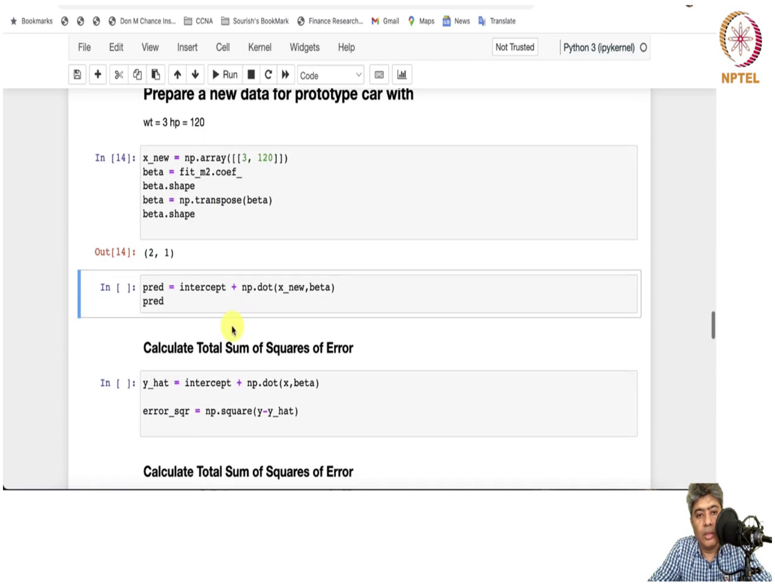


(Refer Slide Time: 06:20)



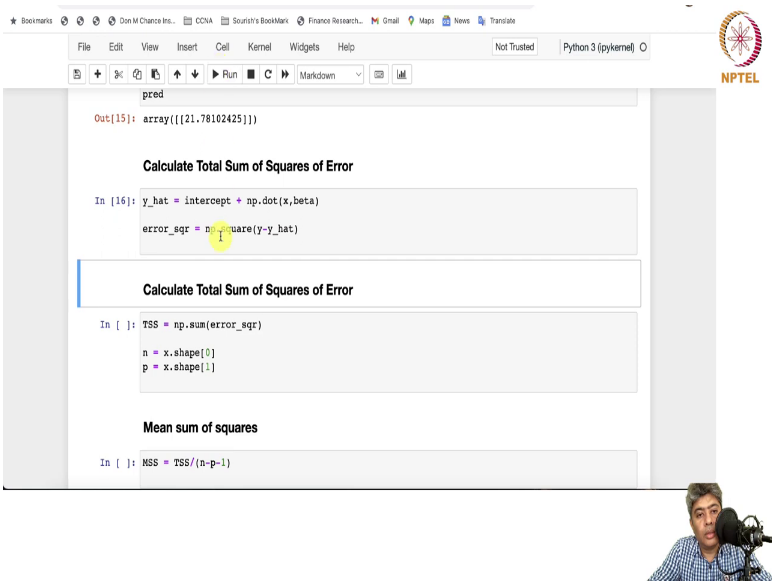
We if we run that, it gives print out more like a you know like a r kind of print out. Now, along with weight we want to put the horsepower; so, in the x we put along with weight we put the horsepower and we fit the model. Now, we are preparing a new data with a new prototype car with weight equal to 3 and horsepower equal to 120.

(Refer Slide Time: 06:47)



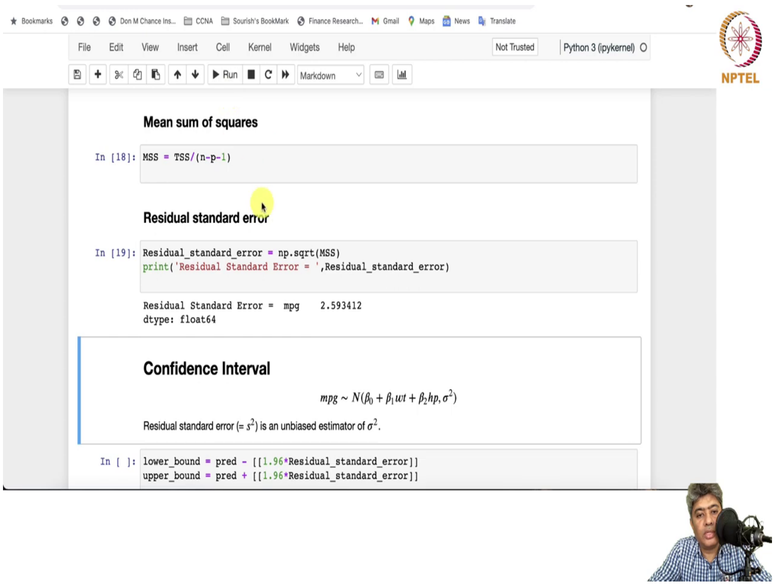
So, in that we run; so, we got the alpha then beta; now, beta we have 2 coefficients; so, that is why it is 2 cross 1.

(Refer Slide Time: 07:00)



Now, predicted values will be for that for a new prototype car which has 3 weights. And how weight equal to 3 meat and horsepower equal to 120 units. We will have a miles per gallon 20 mile 1.78. So, we can calculate the total sum of square; first you calculate y hat then y minus y hat and that. If you square it that will you give you error sum of squares ok, and then if you just take sum that will give you the total sum of squares.

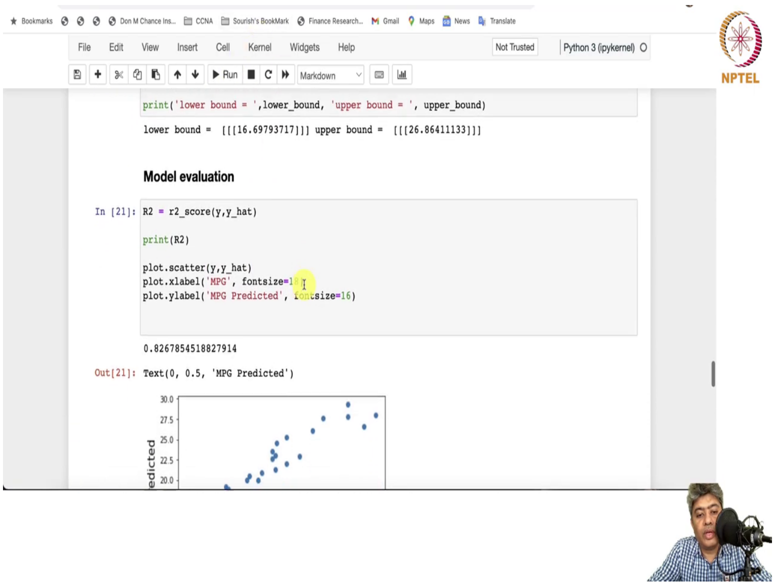
(Refer Slide Time: 07:51)



And then if you just calculate divide that by n minus p minus 1 that will give you the mean sum of square. So, the residual standard error is just take the square root of mean sum of square that will give you the residual standard error ok. Now, so, the residual standard error for this model is 2.59. Now, miles per gallon; that means, for a normal beta naught plus beta 1 times weight plus beta 2 times horsepower comma sigma square.

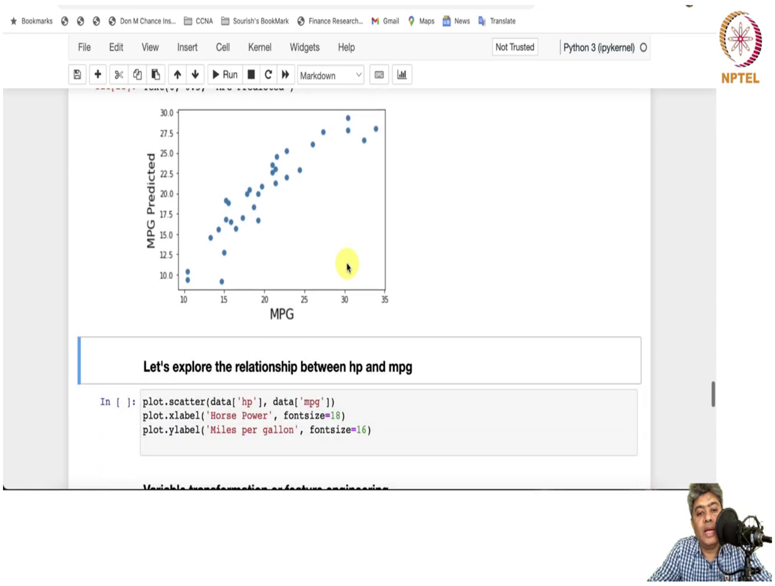
So, residual standard error is an unbiased estimator of the sigma square. So, you can calculate the lower bound as predicted value minus 1.96 times residual standard error and upper bound would be predicted value plus 1.96 times residual standard error and we print the lower bound and the upper bound.

(Refer Slide Time: 08:53)



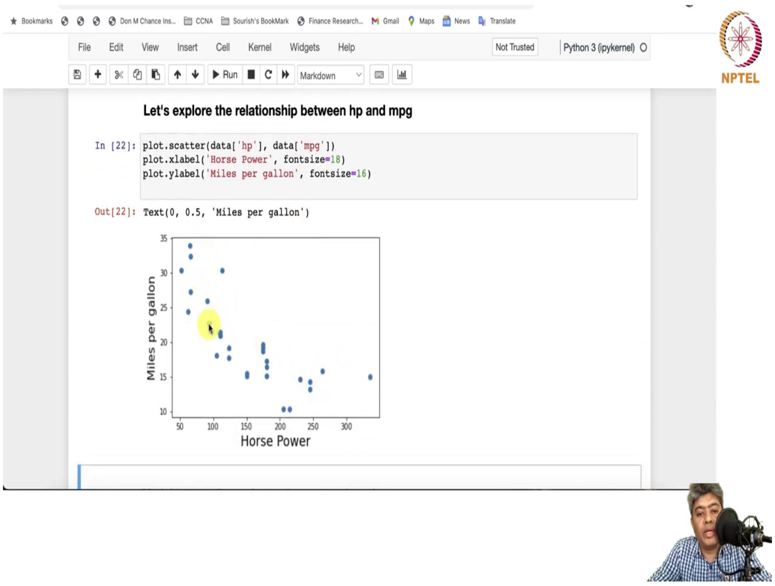
So, the lower bound will be 16.69 and upper bound will be 26.86.

(Refer Slide Time: 09:03)



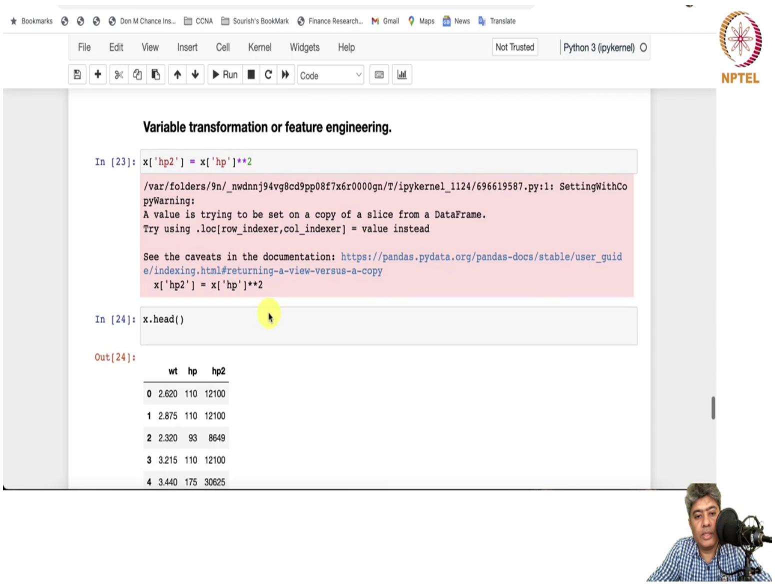
So, for next we fit a you know r square is 82.67 percent; so, along with weight when we put horsepower, the r square jump to 82.6 percent and it becomes slightly tighter.

(Refer Slide Time: 09:25)



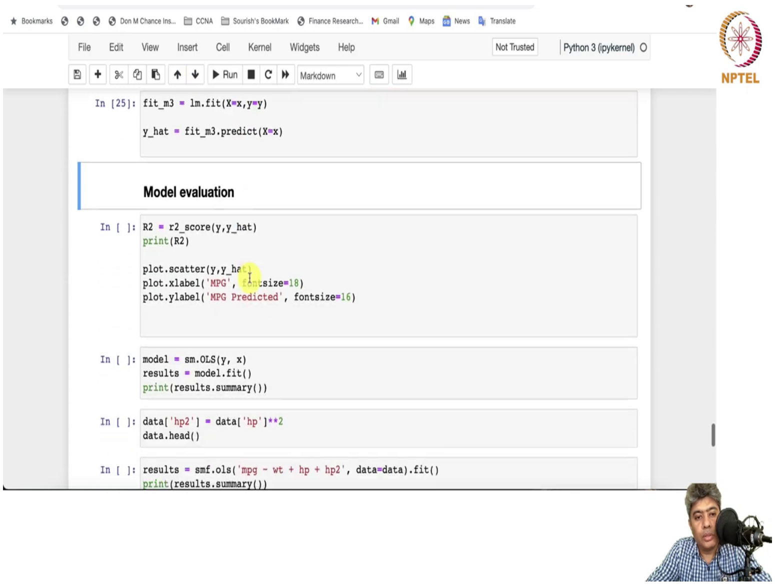
Now, let us explore the relationship between horsepower and miles per gallon. So, what we are seeing the relationship is maybe there is a quadratic relationship.

(Refer Slide Time: 09:40)



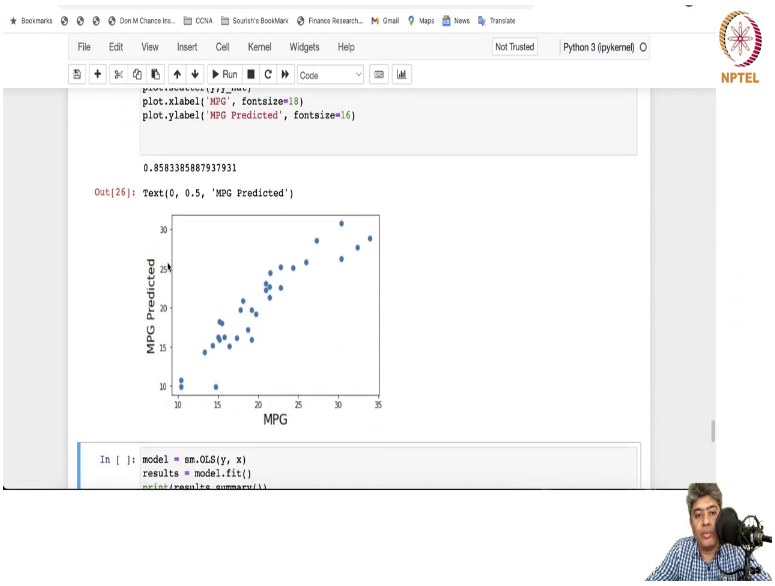
So, we have to do some feature engineering; so, we add a new column by simply taking the square of the column. And then x hat is x hat is to have weight horsepower and horsepower square.

(Refer Slide Time: 09:57)



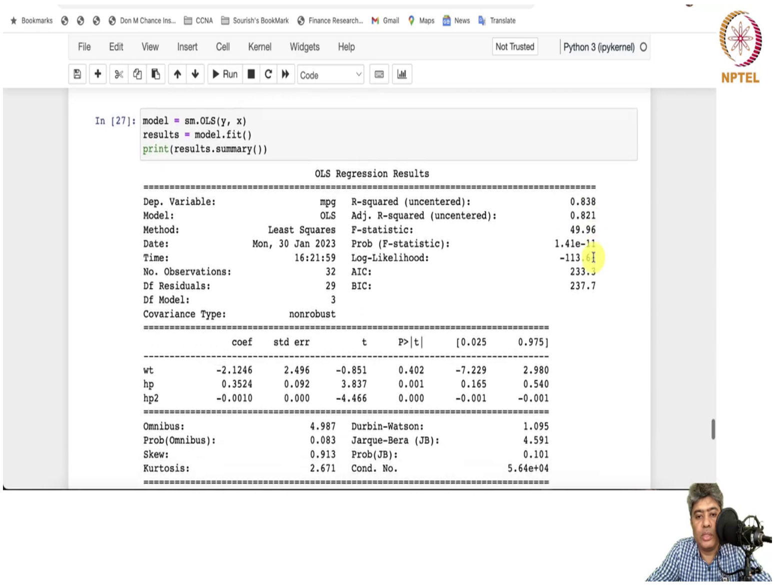
Now, we can fit the third model by just providing the proper x and y and use the predict to get the predicted values.

(Refer Slide Time: 10:13)



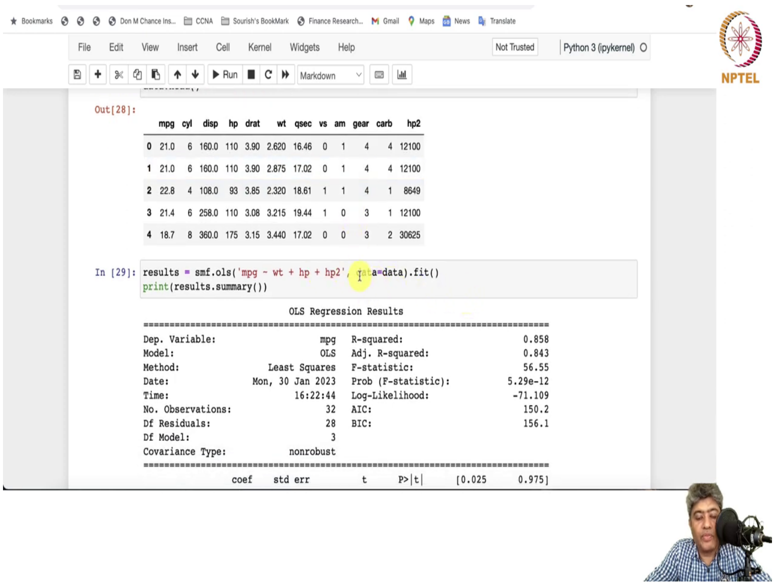
So, the model evaluation when we do the model evaluation, what we find that the whole and you know the r square be has gone up to 85.8 percent and it is becoming more tighter.

(Refer Slide Time: 10:31)



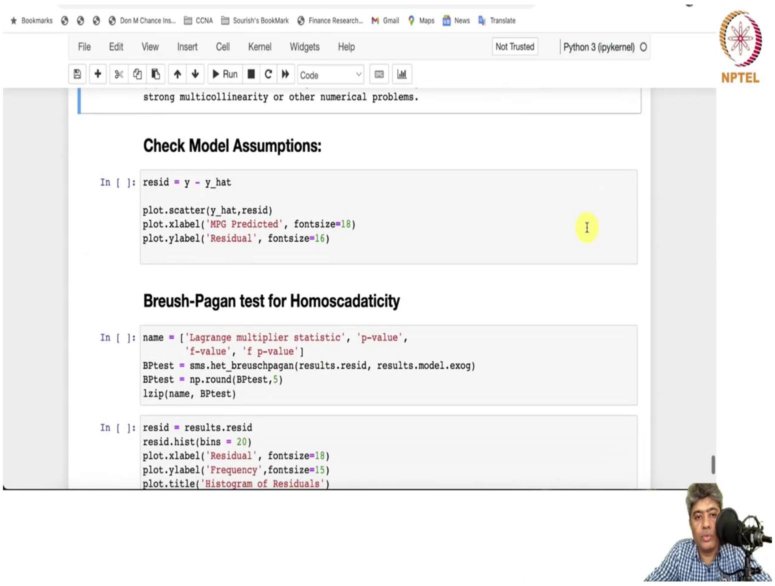
Now, and we fit the OLS way r square slightly different 83 percent we have just did r square is 82 percent. So, this is the sklearn, remember that sklearns calculation is 85.8 percent; whereas, stat models is giving us a slightly different 83 percent ok.

(Refer Slide Time: 11:09)



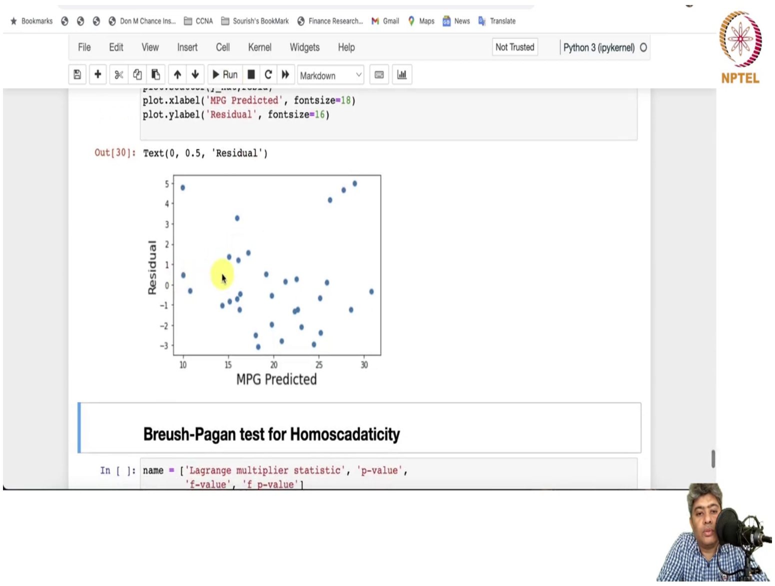
And this is our data set because you know dataset and then when we fit the OLS in this way interesting. When we fit the OLS in this way what we have seen that r square is not 85.5 percent. So, maybe this is the right way of fitting the model ok; so, previous paper is fitting you know the old way ok.

(Refer Slide Time: 11:38)

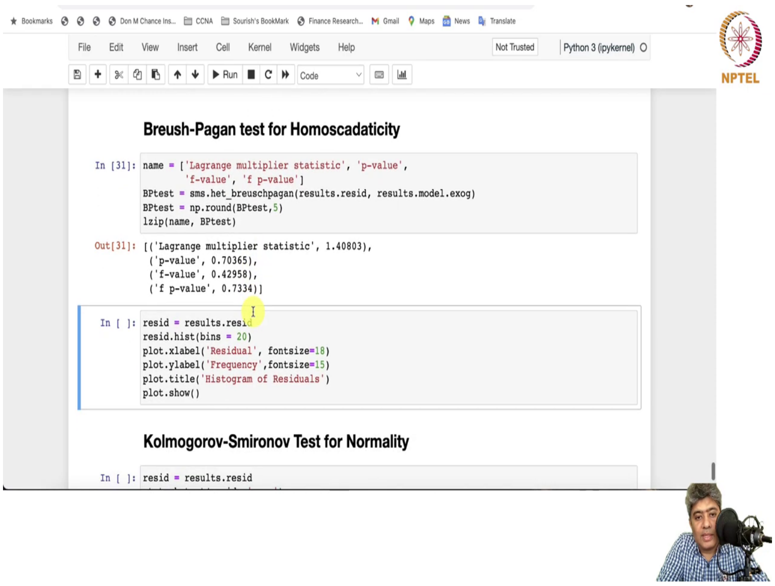


And now, we have to do some we have to do some model as checking the model assumptions. If you want to check the model assumptions what you do? First you define the residuals and plot the residuals against the predicted value.

(Refer Slide Time: 11:55)

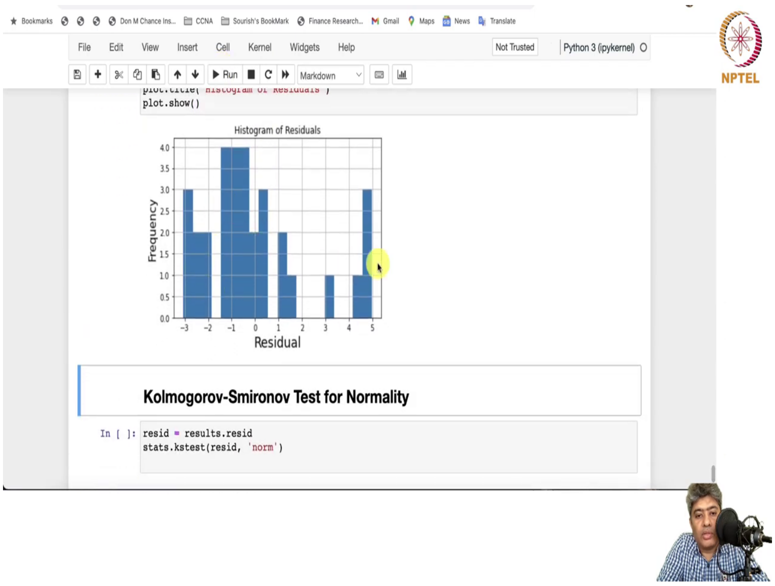


(Refer Slide Time: 12:05)

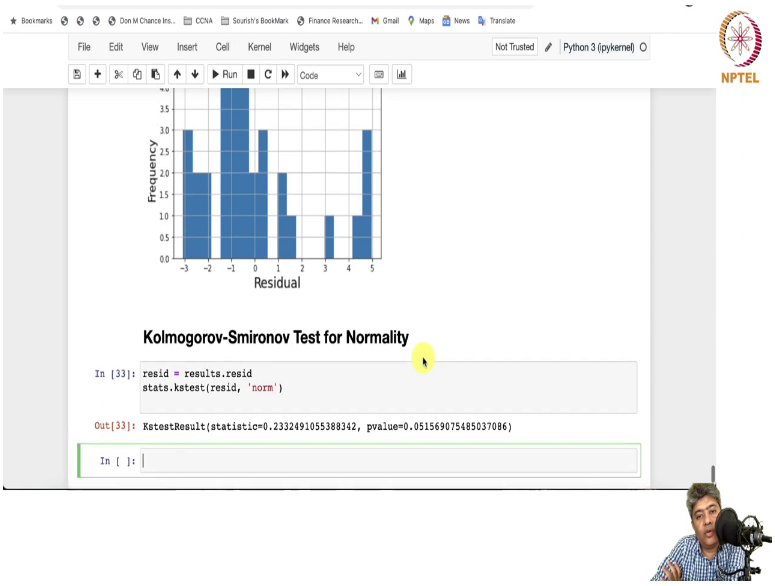


So, we see that there is not much going on; so, we can do a Breush Pagan test for a homogeneity and the p value is quite high. So, from the Breush Pagan test we cannot reject homoscadaticity, we can say safely that it is homoscadaticity behaviour, it has a homoscadaticity behaviour.

(Refer Slide Time: 12:23)



(Refer Slide Time: 12:29)



The histogram is sparse; we can run a Kolmogorov Smironov test the p value is marginally small. So, in the previous also we saw that Kolmogorov Smironov test was marginally smaller giving us a indication that there could be questionably the normality could be questionable though other test could be fine; so, with this we will I will stop.