**Data Analytics with Python**
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 31
Estimation Prediction of Regression Model Residual Analysis: Validating Model
Assumptions - 1

Dear students in the previous class we have explained I have explained the confidence interval for the x coefficient that is the b. For the b we have found what was the lower limit and upper limit. In this class we will find the confidence interval for y and prediction interval.

(Refer Slide Time: 00:43)
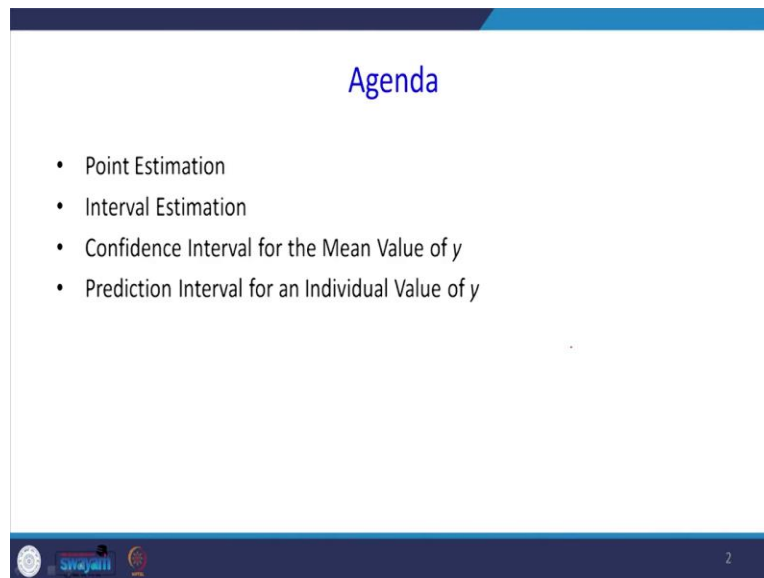
So, today's class agenda is we list will explain what is the point estimate and interval estimate; and confidence interval for the mean value of y and prediction interval for the individual value of y.

(Refer Slide Time: 00:43)

We will take one problem then I first I will solve this problem with the help of Python then I will explain what is the meaning of this confidence interval and prediction interval. Data were collected from a sample of 10 ice cream vendors located near college campuses for the ith observation our restaurant in the sample xi is the size of the student population and y is the quarterly sales of ice cream. The values of xi and yi for 10 restaurants in the sample are summarized in the table this is given x 1.

(Refer Slide Time: 01:29)

So, what the problems is the independent variable is student population dependent variable is sales there was a 10 data set like this.
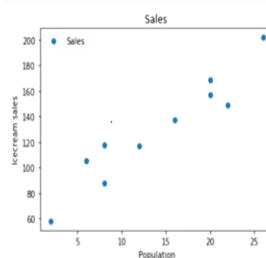
(Refer Slide Time: 01:40)

For given data set first we will run the regression model so import pandas pd import matplotlib as mpl import stats model dot formula dot api as sm from sklearn dot linear underscore model input linear regression from scipy import stats import Seaborn as sns, import numpy as np import matplotlib dot pyplot as plt. First we load the data we will treat the data there is a pd dot breed underscore this is a path where I have stored my excel file, so this is the data.

So, what is it there is a 10 dataset 10 restaurants this is a student population this student populations in terms of 1000 sales also in terms of 1000 of it, for a product called ice cream first we will plot the scatter plot.
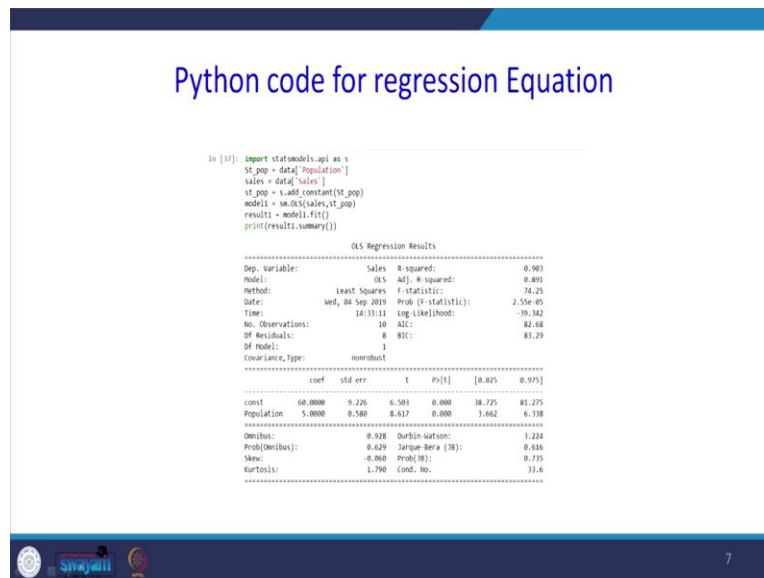
(Refer Slide Time: 02:29)

So, data dot plot population, sales style equal to 'o', so why label is ice cream sales title is sales when we they show this versus, so what is happening there seems to be some positive trend when the student population is more there is a more number of sales. We will find a regression model for this.

(Refer Slide Time: 02:56)

So import startsmodels dot api as s, St_pop equal to that is a student population equal to data I am going to in the population I am going to stay a store variable called St_pop, sales equal to data sales St_pop = s.add_constant because I need to have the constant in the regression equation. So, model one equal to sm.OLS(sales, sale is our dependent variable St_pop) is our independent variable result 1 equal to model one dot fit. So, print result one dot summary.

So what we are getting here and you look at this, this is the constant value. So, the sales equal to I can write sales equal to 60+5 this is our independent variable say population st underscore. It is a population. So, what is the meaning of interpretation of this file if the student population is increased by one unit the sales will increased by 5 units look at this R square R square is very good that is 90.3 I will explain meaning of adjusted R square in coming class.
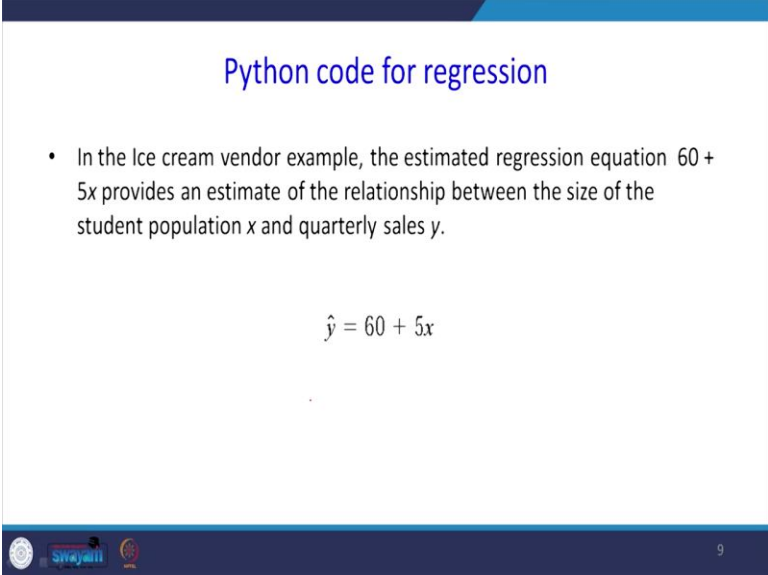
Then we have to remember this is there is a standard at it is 0.58 the t value is 8.68 this is a problem the probability is 0.00 this is lower limit this is upper limit. There is another way we can write it otherwise directly we can get y-intercept and x coefficient from a sklearn dot linear

underscore model input linear regression x equal to data ['population'] dot values reshaped (-1, 1), y equal to data['sales'] dot value.reshape (-1, 1), reg = LinearRegression(), reg.fit is x, y so linear regression is copy underscore x equal to true fit underscore intercept true, equal to true, n underscore jobs equal to one normalize equal to false.

What is the meaning of fit underscore intercept sometime if you put false so sometimes when you fit a regression line suppose it is coming like this, so there is a y intercept is a this much distance is y intercept. Sometime you need not have the y intercept for that time for that time you have to use write false. The another one is normalized equal to false so there are y and there are x value you if you normalize x-value and y-value then you run the regression they will get a standardized regression coefficient.

Now we are not equal to false is written so we are not going to normalize the data set so reg dot intercept underscore 0, reg coefficient underscores 0, 0 so this is your intercept this is your x coefficient. The previous also you look at the previous slide there also here got the 60 and 5 same result.
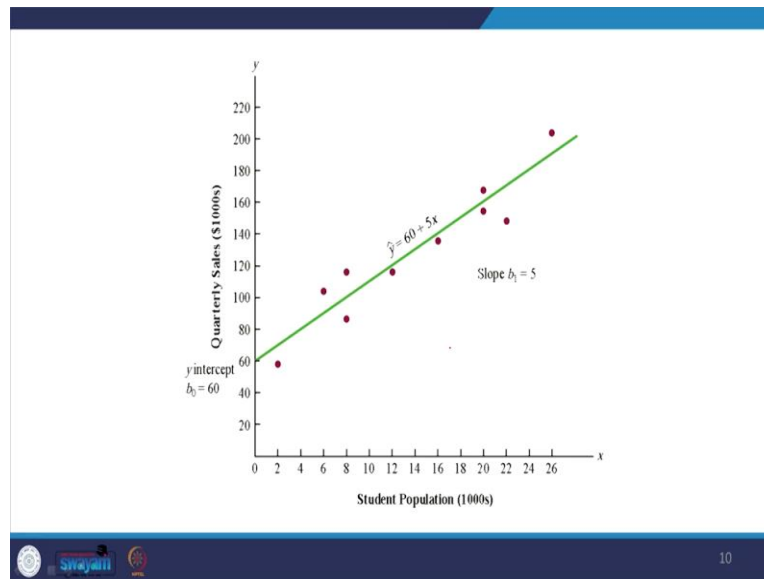
(Refer Slide Time: 06:08)

So, what we can do in the ice cream at our example the estimated regression equation is 60 + 5x provides an estimate of the relationship between the size of the student population x and quarterly sales y. so, this is our regression equation.

So, this is our regression equation so the y intercept the slope is 5 the y intercept is 60.

Then we will see what is the point estimate, we can use the estimated regression equation to develop your point estimate of the mean value of y for a particular value of x or to predict an individual value of y corresponding to a given value of x. So, whatever value which you are predicting is the mean value there is another we can predict an individual value. For instance suppose your manager want to want see a point estimate of the mean quarterly sales for all restaurants here you have to see all restaurants located nearby a college campus with the 10,000 students.

(Refer Slide Time: 07:23)

So if you say student population is 10 what will happen when you substitute to 10 it is 110. So, that is your point estimate for the mean quarterly cells of all restaurant located near campus is the 10,000 students a one lakh 10,000 dollar. So, even regression equation also we can use the predict function reg.predict when you put the input value that is x value you can get y value is 110.

(Refer Slide Time: 07:45)

So, what is a point estimate now suppose the manager want to predict the sales of an individual restaurant located in area college with the 10,000 students. In this case we are not interested in the mean value of all restaurants located near compass of the 10,000 students we are just

interested predicting quarterly sales of one individual restaurant as it turns out the point estimate for an individual value of y is the same as the point estimate for the mean value of y.

Hence we would predict quarterly sells sales 60 + 5 the 10 is our input it is 110000 dollars so what I am saying for you find estimate the value of confidence interval and the value of prediction interval is same you see that you may see the similarity also here and when you go for hypothesis testing see x-bar + or - Z Sigma by root n, right. So, this x bar is nothing but our point estimate so whatever value after substituting 10 we are getting 110 we are getting that is only point estimate so point estimate is not the reliable one so we need to have interval estimate.
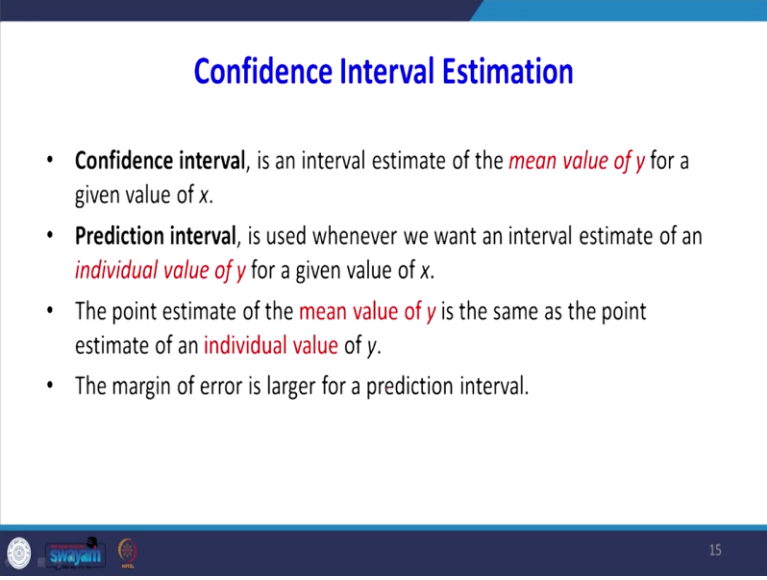
So, interval estimate in the hypothesis testing context x bar + Z Sigma by root n is the upper limit x bar - Z Sigma by root n lower limit. How we are going to find out upper limit lower limit in the regression context I will explain in the next slide.

(Refer Slide Time: 09:17)
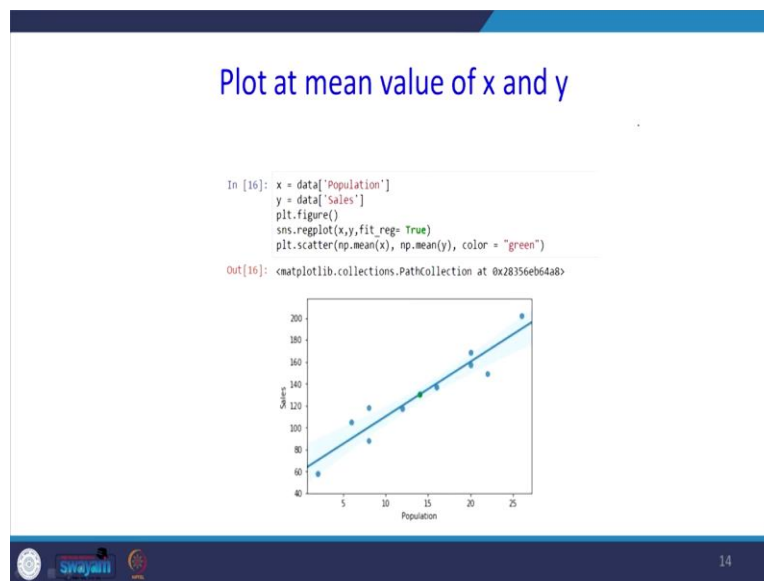
First we will plot it so what will happen plot at mean value of x and y so x equal to data population y equal to data sales so x is the population y is the sales value plot dot figure, sns dot reg plot x, y fit underscore regression is true plot dot scatter np dot mean value of x, np dot mean value of y so we got this regression equation. You see that if you want to draw the best regression equation that has to pass through x bar, y bar. So, that is why, so this point is mean of x so this point is mean of y.

(Refer Slide Time: 10:06)

So what is a confidence interval estimation, confidence interval is an interval estimate for the mean value of y for a given value of x. But the prediction interval is used whenever we want an interval estimate of an individual value of y right this is an individual value of y that is a mean value of y for a given value of x. For example y it is a mean value so what we are predicting is expected value of $E(y) = a + bx$ so whatever value after substituting x we are getting into the mean value.

So what will happen the margin of error is larger for your prediction interval. So, the prediction interval the margin of error will be larger for your confidence interval the margin of error will be smaller.

(Refer Slide Time: 10:56)

Confidence Interval Estimation

x_p = the particular or given value of the independent variable x

y_p = the value of the dependent variable y corresponding to the given x_p

$E(y_p)$ = the mean or expected value of the dependent variable 'y' corresponding to the given x_p

$\hat{y} = b_0 + b_1 x_p$ = the point estimate of $E(y_p)$ when $x = x_p$

$$60 + 5(10) = 110.$$

So, confidence interval of estimation, for example take xp equal to the particular or given value of independent variable x, yp is the value of the dependent variable y corresponding to the given xp, so expected value of yp is nothing but mean or expected value of dependent variable y corresponding to the given xp, so y hat equal to b 0 + beta 1 xp, is the point estimate of expected value of yp when x equal to xp so that is why 60 + 5*10 is 110.

(Refer Slide Time: 11:47)

Confidence Interval Estimation

In general, we cannot expect $\hat{y}_p$ to equal $E(y_p)$ exactly.

If we want to make an inference about how close $\hat{y}_p$ is to the true mean value $E(y_p)$, we will have to estimate the variance of $\hat{y}_p$.

The formula for estimating the variance of $\hat{y}_p$ given x_p, denoted by , is $s^2_{\hat{y}_p}$

$$s^2_{\hat{y}_p} = s^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2} \right]$$

In general we cannot expert y hat p is equal to expected value of y p exactly if you want to make an inference about how close y hat p is to the true mean value of expected value of y p we will have to estimate the variance of y hat p. The formula for estimating the variance of y hat p at

given x p is denoted by s square, y hat p so this y hat p is nothing but s^2 $((1 / n) + ((xp - x\ bar)^2 / \Sigma(xi - x\ bar)^2))$, this is the variance of predicted y.

(Refer Slide Time: 12:36)

So, the confidence interval is you see that the confidence interval is we are writing y hat p + or - so this yes y hat p is the variance of this y at p right so what you are done previously in the hypothesis testing example x + or - z Sigma by root n. So, instead of x bar we are writing y hat p + or - so instead of z we are writing t α/2 this standard error and so write we are writing Sy hat p, so that was the formula is equal to s^2 (1 divided by n) this can be derived a very easily.

So I am not deriving you can refer any book for this. so, the variance of y cap is equal to s^2 $((1 / n) + ((xp - x\ bar)^2 / \Sigma(xi - x\ bar)^2))$, We can substitute this value here the s square is nothing but the standard error. So, we can substitute s square value n is 10 xp is 10 because that is a value of x so x bar is 14 whole square when you substitute 2 you are getting $110 + or - 11.415$.

(Refer Slide Time: 13:44)

Confidence Intervals for the Mean sales *y* at given values of student population *x*
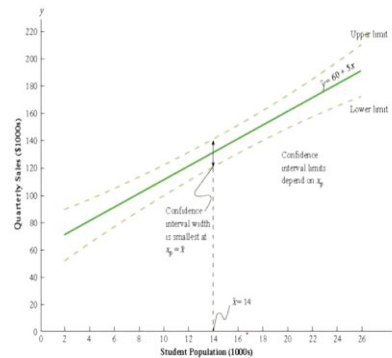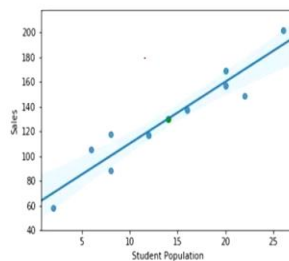
So, the Green Line shows green dotted line shows the upper limit the down one is shows the lower limit. You see that it is the confidence interval is not a straight line it is somewhat curved one so what is happening when x bar equal to 4 the interval now it is a very narrow. What will happen that is a special case.

(Refer Slide Time: 14:09)

Now we will plot this confidence interval okay what is happening here you see that when this Point, see it is not the straight line it is somewhat curved one. The confidence interal is very narrow when there is x equal to x bar we will see that a special case.

(Refer Slide Time: 14:27)

The estimated standard deviation of y hat p is smallest when xp equal to x bar. So, what will happen in the previous equation when you substitute xp equal to x bar so this term will become 0. So, remaining is s divided by 1 by n you see that this is similar to our the result of central limit theorem. The variance of a sampling distribution is Sigma by root n it is similar to that.

(Refer Slide Time: 14:58)

Now we will go for prediction interval for an individual value of y instead of estimating the mean value of sales for all restaurants located near campus of the trend of students we want to estimate sales on individual restaurant located near a particular college with the 10,000 students. So, when you go for predicting y value for an individual restaurant there are two component of variance has to be added one component is the variance of individual y values about the mean

value of y p that is given by s square, the variance associated with using y hat p estimate is expected value of y p and estimate of which is given by s square of y hat p.

So what is happening here if you want to go for a prediction interval these two variances has to be added one variances for y another variances for y hat p right.

(Refer Slide Time: 16:03)

Prediction Interval for an Individual Value of y

$$s_{ind}^2 = s^2 + s_{\hat{y}_p}^2$$
$$= s^2 + s^2\left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]$$
$$= s^2\left[1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\Sigma(x_i - \bar{x})^2}\right]$$

You see that so that is where s square individual is s square + s square y hat p so when you add it the s square is common so we will get this formula. So, for this formula we will substitute the value when you substitute it you see it is a 14.69.

(Refer Slide Time: 16:17)

Prediction Interval for an Individual Value of y

$$\hat{y}_p \pm t_{\alpha/2}s_{ind}$$

$$t_{\alpha/2}s_{ind} = 2.306(14.69) = 33.875,$$

$$110 \pm 33.875$$

Now the value of t alpha by two and you look at the then when you substitute 14.69 you will get this was the this 33.875 is the margin of error so 110 + or − 33.875 will get so this one.

(Refer Slide Time: 16:37)
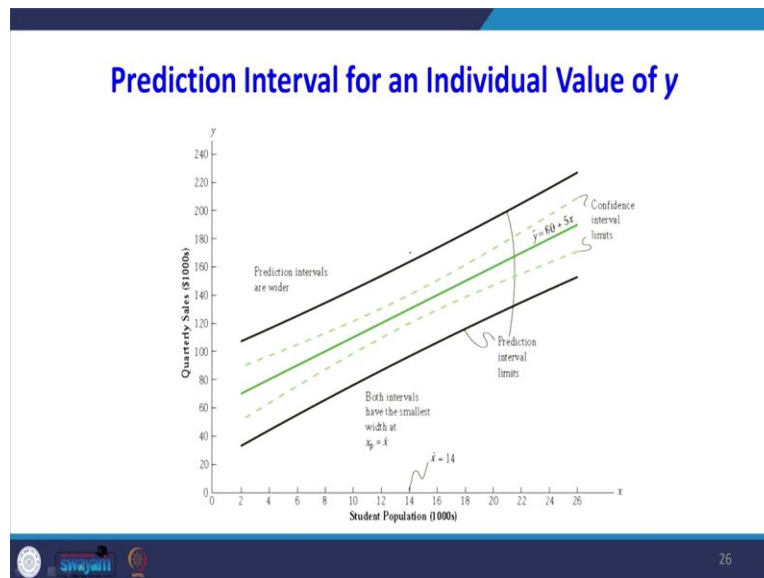
So, in the black line shows the prediction interval the Green Line shows the confidence interval both are not the straight line. So, when you look at this one see the prediction line is having margin of error is more compared to the confidence interval.

(Refer Slide Time: 16:58)
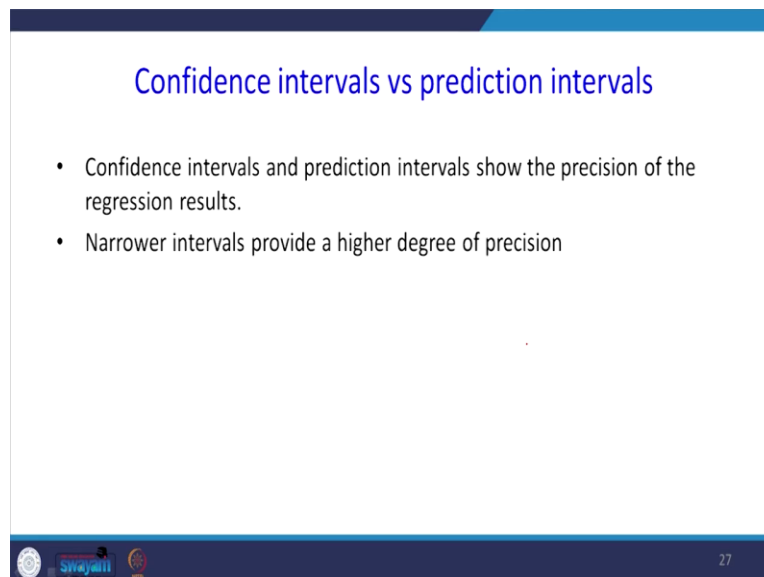
Now confidence interval versus prediction interval confidence intervals and prediction intervals show the precision of the regression result narrower intervals provide a higher degrees of

precision. So, what after doing regression analysis when you plot the confidence and prediction interval it has to be narrow if it is wide means that model is not the good model.

(Refer Slide Time: 17:23)

Now we will use Python to plot this prediction interval and confidence interval. So, for that purpose from statsmodels dot s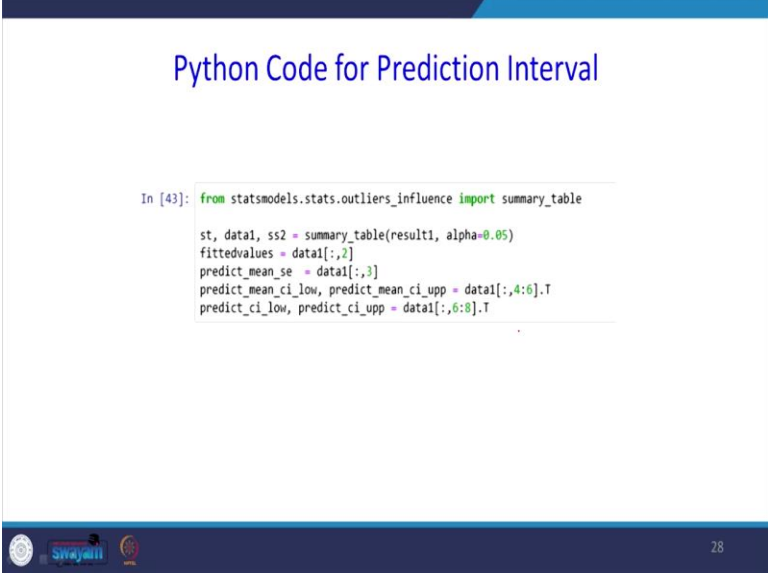tats dot out layer underscore influence import summary table st command data 1, ss 2 equal to summary underscore table result 1, alpha equal to 5% fitted value is equal to data colon, second that means we are referring the third column predict underscore mean underscore ac equal to date data 1 colon, 3 that is we are referring fourth column predictor underscore mean ci interval that is mean ci interval means your confidence interval lower limit confidence interval upper limit.

So, that was because in their summary table that is in the summary table we are referring the fourth to sixth column, dot t predict underscore see a confidence table low-protein predict underscore ci upper limit data to 6 : 8. Actually what do you have is what is happening here we are getting in the summary table all the result so we are calling 4 to 6, 6 to 8, 3 2 to get a particular value that is the reason here.

(Refer Slide Time: 18:41)

Python Code

```
In [44]: predict_mean_ci_low
Out[44]: array([ 51.03868339,  75.2931351 ,  87.10977127,  87.10977127,
                109.56629808, 129.56629808, 147.10977127, 147.10977127,
                155.2931351 , 171.03868339])

In [45]: predict_mean_ci_upp
Out[45]: array([ 88.96131661, 104.7068649 , 112.89022873, 112.89022873,
                130.43370192, 150.43370192, 172.89022873, 172.89022873,
                184.7068649 , 208.96131661])

In [46]: predict_ci_low
Out[46]: array([ 32.89834155,  54.8817226 ,  65.60291394,  65.60291394,
                 86.446108  , 106.446108  , 125.60291394, 125.60291394,
                134.8817226 , 152.89834155])

In [47]: predict_ci_upp
Out[47]: array([107.10165845, 125.1182774 , 134.39708606, 134.39708606,
                153.553892  , 173.553892  , 194.39708606, 194.39708606,
                205.1182774 , 227.10165845])
```

29

You see that this is the predict underscore main underscore ci_low so we are getting the confidence interval for the lower limit here predict mean underscore ci for upper limit. So, this was predict for this is a prediction interval this is for the confidence interval the first 2 things for the confidence interval the next bottom 2 is for the prediction interval. So, lower limit upper limit this is the lower limit see a ci _ low, ci _ upper is the upper limit.

(Refer Slide Time: 19:19)

Python Code

```
In [48]: X = s.add_constant(x)
         fig, ax = plt.subplots(figsize=(8,6))
         ax.plot(x, y, 'o', label="data")
         ax.plot(X, fittedvalues, 'r-', label='OLS')
         ax.plot(X, predict_ci_low, 'b--')
         ax.plot(X, predict_ci_upp, 'b--')
         ax.plot(X, predict_mean_ci_low, 'g--')
         ax.plot(X, predict_mean_ci_upp, 'g--')
         ax.legend(loc='best');
         plt.show()
```
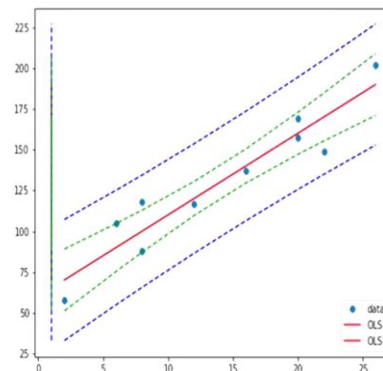
30

So, this picture shows you see that x equal to s dot ad underscore constant fig, ax equal to plot dot subplot fig size equal to 8, 6, ax dot plot x, y ou label equal to data ax dot plot x, featured values are hypen, label Wireless ax plot dot x, predict underscore ci low it is in the dotted line by ax dot plot x, predict underscore ci underscore upper limit b hypen hypen ax dot plot x, predict

underscore mean ci low g a x dot plot x predict underscore main ci upper limit so the location is the best plt dot show.

So, when you run this command you will get this kind of here model. So, the green one shows the confidence level the blue one shows the prediction interval. In this picture when you look at the 'r-' represents the red color b represents blue color g represents green color the hypen represents what kind of pattern we need to have in the in the picture. Now what we have done in this class we have explained what is point interval and what is confidence interval and what is prediction interval.

So the point interval is same for particular value of x for both confidence and prediction interval. So, the another point which you have learnt in this lecture is that the confidence interval is not the straight line it is curved line similarly the prediction interval. The another one is the prediction interval is having more margin of error when compared to confidence interval after that what you have done with help of Python I have run this code to show you how to plot this confidence and prediction interval, thank you very much.