Lecture – 44
Regression Analysis Model Building – 1

We have seen so far a simple linear regression model and multiple linear regression model. In this class, we are going to see how to construct a regression model by considering different independent variables, that is model building using regression analysis.

(Refer Slide Time: 00:41)

What we are going to do. Model building is the process of developing an estimated regression equation that describes the relationship between a dependent variable and one or more independent variables. What is the meaning of model building? There are different independent variable is there and there is some dependent variable. We are going to find out how to construct a regression model by considering all the independent variable. Whether we have to consider all independent variables or which variable has to be dropped or which variable has to be added.

The major issues in model building are finding the proper functional form of the relationship and selecting the independent variables to include the model. Two concept is there. One is what kind of relationship is going to be found. One is whether it is linear or nonlinear and the other one is how to select the appropriate independent variables. How to select the appropriate independent variables.

Suppose, we collect data for one dependent variable y and k independent variables. The independent variables x1, x2 and so and xk. Objective is to use these data to develop an estimated regression equation that provides the best relationship between the dependent and independent variables. So general form of linear regression model is $y = \beta_0 + \beta_1 z_1 + \beta_2 z_2$ plus and so on plus $\beta_p z_p$ plus error term. Here, the zj, j = 1, 2 up to p is a function of x1, x2, the variables for which the data are collected. In some cases, Zj may be a function of only one x variable, only one independent variable.

(Refer Slide Time: 02:37)

If it is only one independent variable, that is called a simplest first-order model with one predictor variable is this. What is happening here, the value of z is taken only x1. There will be an error term. This is the simplest linear regression model.

Now, we will go for modeling curve linear relationships. This problem is taken from this book, statistics for business and economics, 11th edition by David Anderson, Sweeney and Williams. To illustrate, let us consider the problem facing a company called Reynolds, a manufacturer of industrial scales and laboratory equipment. Managers at Reynolds wants to investigate the relationship between the length of the employment of their salespeople and the number of electronic laboratory scales sold.

So what they want to know, their length of employment of salespeople that is how many years they are working in that company versus number of electronic laboratory scales sold, how much they have sold. Generally, what is the assumption, a person who is having a lot of experience will sell more product. The table in the next slides gives number of scales sold by 15 randomly selected salespeople for the most recent sales period and the number of months each salesperson has been employed by the firm.

Data

Scales Sold	Months Employed
275	41
296	106
317	76
376	104
162	22
150	12
367	85
308	111
189	40
235	51
83	9
112	12
67	6
325	56
189	19

So, this slide shows the data, scales the product sold, months employed. So, here months employed is going to be our independent variable, scales sold is going to be our dependent variable. This is y and this is our x.

(Refer Slide Time: 04:27)

Importing libraries and table

```
In [1]:  import pandas as pd
         import numpy as np
         import matplotlib.pyplot as plt
         import statsmodels.api as sm

In [9]:  tbl1 = pd.read_excel('Reynolds.xlsx')
         tbl1

Out[9]:
```
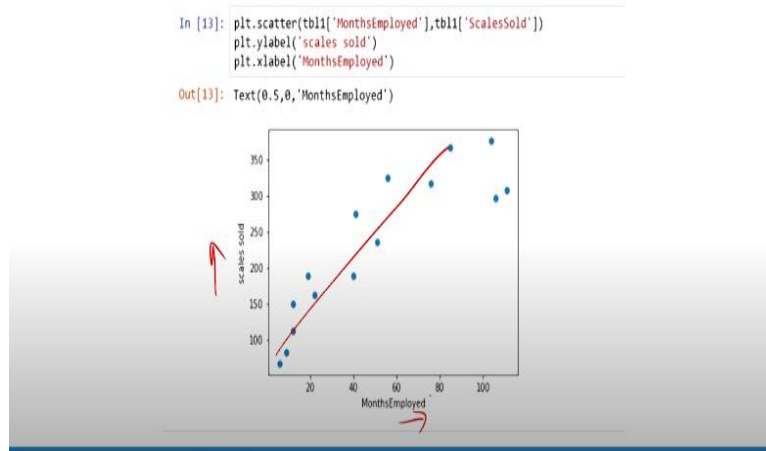
	Scales Sold	MonthsEmployed
0	275	41
1	296	106
2	317	76
3	376	104
4	162	22
5	150	12
6	367	85
7	308	111
8	189	40
9	235	51
10	83	9
11	112	12
12	67	6

With the help of Python, first we will construct a simple linear regression equation; let us see what is happening. I have brought the screenshot of Python programming. At the end of the class, I will run these codes. You can type these commands in your PC and you can verify the answer. Import pandas as pd, import numpy as np, import matplotlib.pyplot as plt, import statsmodels.api as sm. The data, which I have stored in the file name called Reynolds.xlsx. I have read this data. This was the dataset. There is y, this is the x.

(Refer Slide Time: 05:17)

SCATTER DIAGRAM FOR THE REYNOLDS EXAMPLE

```
In [13]: plt.scatter(tbl1['MonthsEmployed'],tbl1['ScalesSold'])
         plt.ylabel('scales sold')
         plt.xlabel('MonthsEmployed')

Out[13]: Text(0.5,0,'MonthsEmployed')
```


Let us do a Regression Analysis. First, going for Regression Analysis, we will go for a scatter plot. For drawing the scatter plot, plt. scatter tb1, that is my first variable months employed, this variable, second variable is going to be in y-axis scales sold. When it is plotted, it seems to be there is a positive trend. What is the meaning of positive trend? When the number of months employed is increasing and the sales also increasing. It says that the person who is experienced the salesperson can sell more products when compared to an inexperienced person.
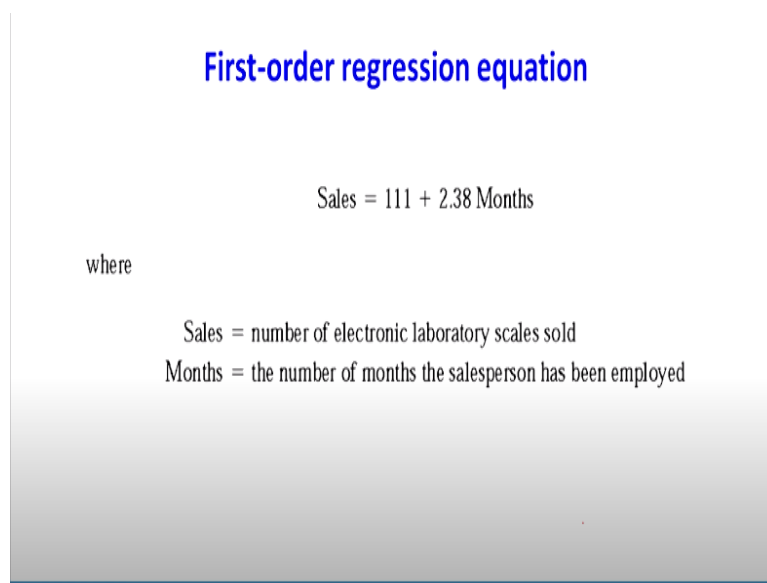
(Refer Slide Time: 06:01)
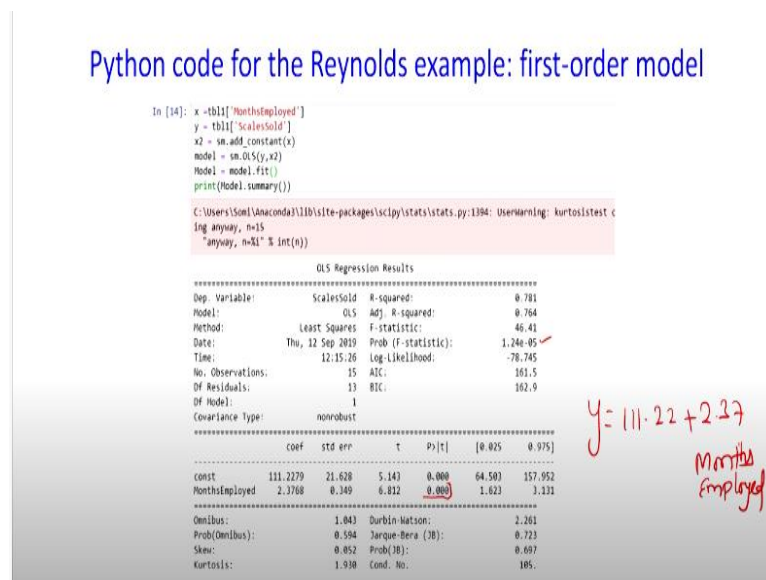
So, this was the code for running Regression. So x = tb1, that is months employed is independent variable, ytb1 ScalesSold is my dependent variable. I am going to add x2 equal to sm.add_constant, so that in my Regression model, I will get a constant. So, model = sm.OLS. OLS is ordinary least square method, y is dependent variable, x2 is independent

variable because in the x2, I am going to have the x variable also. So, model equal to model.fit, so print Model.summary. So, how to interpret this one.

Look at this constant, say y = 111.22 + 2.37 months employed. We will test the significance of the model. First, we will look at the F statistics and the corresponding p value. Here, p value is 1.24 into 10 to the power - 5, very low. As a whole model, this model is significant. Then look at the significant of individual variables. The month employed is independent variable. When we look at the p value, this also less than 5, so we can say the months employed is a significant variable.

(Refer Slide Time: 07:39)

First-order regression equation

$$Sales = 111 + 2.38\ Months$$

where

$$Sales = \text{number of electronic laboratory scales sold}$$
$$Months = \text{the number of months the salesperson has been employed}$$

This was my model. Here, the sales is that is y variable, number of electronic laboratory scales sold, months equal to number of months the salesperson has been employed.

(Refer Slide Time: 07:51)

Standardized residual plot for the Reynolds example: first-order model

What will happen? First we will plot, the residual plot because as I told you in my previous classes, it is not only the R square fp value and individual significance value is important, the same time, we have to check the residual of that Regression Model. First, we will find the residual E = Model. resid_pearson. So, this was my residual. So, for the x2 that is my independent variable, I have predicted the y hat value. This is my predicted y value.

(Refer Slide Time: 08:34)

Standardized residual plot for the Reynolds example: first-order model
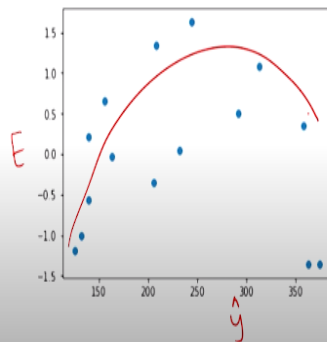
Now, I am going to make a plot between in x-axis, we have taken y hat, in y-axis, this is the error. When you look at this picture, in x-axis, it is our y hat. In y-axis, standardized residual. When I look at this standardized residual, it is not coming in the rectangular shape. You see that, there is a possibility of certain kind of curvilinear relationship. You may not agree it is exactly a curve linear relationship because the number of dataset is less. If there are more

number of dataset, we can exactly say. So, what is happening, it is not in the rectangular shape, it is suggesting that there may be curvilinear relationship between x and y.

(Refer Slide Time: 09:25)

Need for curvilinear relationship

- Although the computer output shows that the relationship is significant (p-value .000) and that a linear relationship explains a high percentage of the variability in sales (R-sq 78.1%), the standardized residual plot suggests that a curvilinear relationship is needed.

Very important point, why we have to go for curvilinear relationship. Although the computer output shows that the relationship is significant, because the p value is less than 0.05 and that the linear relationship explains the higher percentage of variability, that is R-square, so R-square is 78.1. The standardized residual plot suggest that the curvilinear relationship is needed.

So, what is the point which I wanted to say, not only R-square, not only the significant value. Apart from that, we have to draw the different residual plot to verify whether the model is correct or not. When we are plotting the standardized residual model, it is suggesting for a nonlinear relationship.

(Refer Slide Time: 10:19)

Second-order model with one predictor variable

- Set $Z_1 = x_1$ and $Z_2 = X^2$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

So, we are going for what kind of nonlinear relationship we are going to have it. Z1 is x1, no problem. Then Z2, I am going to square that X value. So, this squared value, that is x1 squared, is taken as a new independent variable. Previously only x1 was there, this square of x1 is a new independent variable. So, this is a general linear model, there is independent variables are having some non-linear patterns. That is x1 squared.

(Refer Slide Time: 10:54)

New Data set

- The data for the MonthsSq *independent variable is obtained by squaring the values of* Months.

```
In [29]: X_sq = (x**2)
         X_sq

Out[29]: 0      1681
         1     11236
         2      5776
         3     10816
         4       484
         5       144
         6      7225
         7     12321
         8      1600
         9      2601
         10       81
         11      144
         12       36
         13     3136
         14      361
         Name: MonthsEmployed, dtype: int64
```

So, what I am going to do is to prepare a new dataset that is square of x. For that purpose, x_sq, I am naming that way is equal to X double star 2, so that is a squared. So, this squared value is going to be taken as another independent variable.

(Refer Slide Time: 11:11)

Python output for the Reynolds example: second-order model

You see that x_new is np.column_stack, see this x squared. I wanted to have the constant, so x_new2 equal to sm.add_constant x, new. So, the model 2 equal to sm.OLS y, x_new2, so model.fit, then print summary. So, now what has happened. Look at this point, so there are two independent variable, one is x1 and x2. So, one variable is x1 is the month, the x2 is the squared value. Look at the R-square. The R-square is previously 0.7, now it is improved.

So, our model is good. Look at the significance of each variable. One is x1, another one is squared value of x1, so both are significant value. This also less than 0.05, this also less than 0.05. So, what has happened, when you introduce a squared value, then the model is significant. Not only the significance is enough for us to decide, the model is good or not, we should go for error analysis.
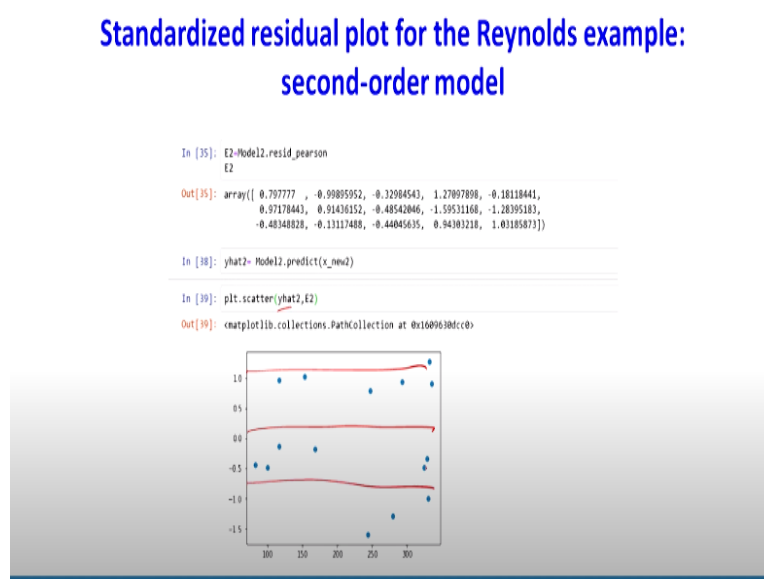
(Refer Slide Time: 12:26)

Second-order regression model

$$Sales = 45.3 + 6.34\,Months - .0345\,MonthsSq$$

$MonthsSq =$ the square of the number of months the salesperson has been employed

So, what is that model which we have created. 45.3 + 6.34 months, then minus 0.0345. That is squared value of month. Now, look at the standardized residual plot of new variable.

(Refer Slide Time: 12:41)

Now, look at the standardized residual plot of new variable. So, here what happened, I have found the error term here. In the error term, for the model which you have created. So, I have predicted y hat. So, now I am drawing a graph of standardized residual. So what is happening, it is kind of a linear residual relationship but not only that you can see that there is a possibility of getting a rectangular shape. So that we can say our model is improved.

(Refer Slide Time: 13:15)
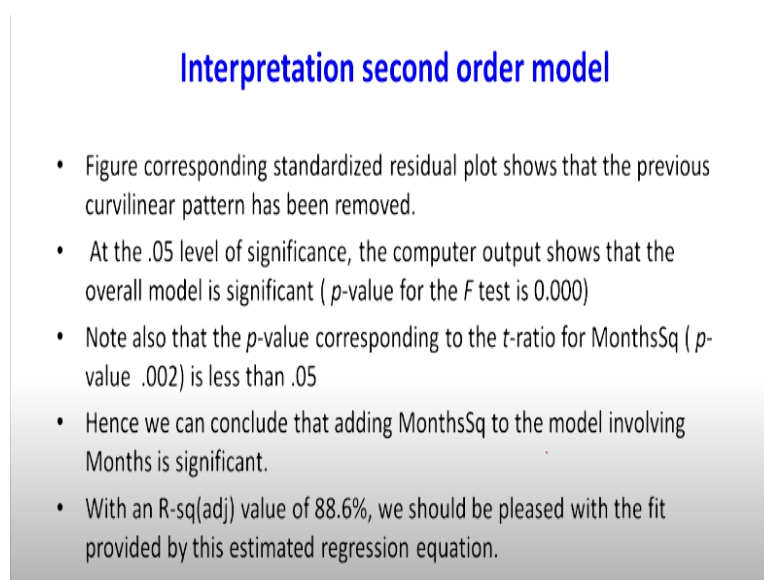
How to interpret the second order model? The figure corresponding to standardized residual plot shows that the previous curvilinear pattern has been removed at 0.05 level of significance, our Python output shows that the overall model is significant because the p

value for the f-test is 0.000. Note also that the p value corresponding to the t-ratio of Months Sq is 0.002. This also significant. Hence, we can conclude that adding months square as a new variable to the model also significant. With an adjusted R-sq of 88.6%, we should be pleased with the fit provided by this estimated regression equation where there is a nonlinear relationship is there.

(Refer Slide Time: 14:22)


Meaning of linearity in a general linear regression model. In multiple regression analysis, the word linear in the term general linear model refers to the fact that beta0, beta1, up to beta p all have exponents of 1. What is the meaning of this one is, see suppose a model is there, beta0+ beta 1x1 + beta 2x2. When I say linear model, this coefficient of beta1, beta2, beta0, these are linear. There is a x1, y relationship. We are not discussing about relation x and y.

When we say linear, the coefficient beta0, beta1, beta2 are linear. It does not imply the relation between y and x is linear. Indeed, we have seen one example, how linear equation general linear model can be used to model a curvilinear relationship. Previously, we have done one model y = b0 + b1x1 + b2x1 square. Actually, this x1 square is nonlinear but we have called it as linear model because this b0, b1 and b2 is linear. Now, I will run the Python code for this model which I have explained.

(Refer Slide Time: 15:55)
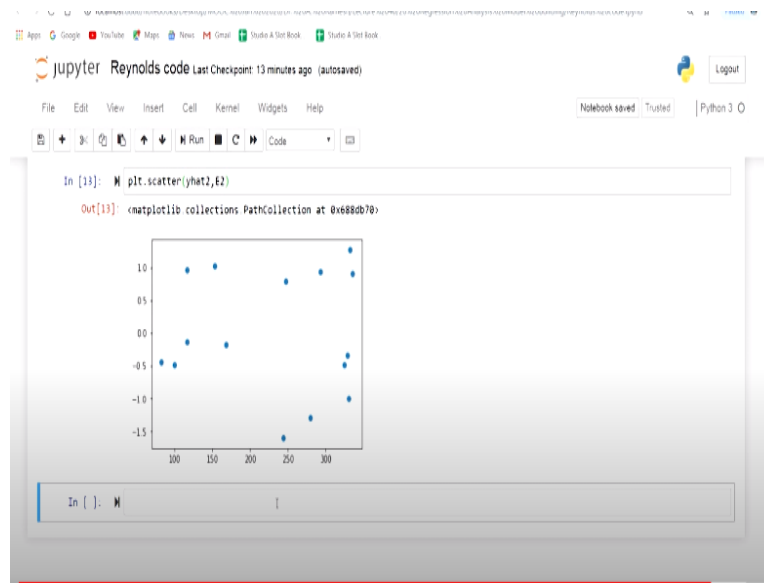
Now, in the Python environment, I will tell you how to do the curvilinear relationship. First, I have imported the necessary libraries pandas, numpy, matplotlib, statsmodels, by running this then I am going to import the data. The data which I have stored in the excel file. The file name is Reynolds. This is the data. This data shows the MonthsEmployed dependent variable, ScalesSold is our dependent variable.

First, we will go for a scatter plot between these two variable. Now, this scatter plot shows that there is a positive relationship between the MonthsEmployed and the ScalesSold. This implies a person having more experience can sold more products. Now, we will go for a simple linear Regression equation. X equal to MonthsEmployed tbl1, y is our dependent variable tbl1 scales sold, x2 = sm.add_constant x. Thus I need to have a constant and model = sm.OLS, so model.fit and printing the summary.

So, it shows that when you look at that first one is the R-squared, it is equal to 0.781. The fp value is 1 into 10 to the power - 5, it is very low, so that overall model is significant. Now, look at the independent variable that is our month employed and the corresponding p value is 0.00, so the MonthsEmployed independent variable is also significant independent variable. Now, we can construct a Regression equation. That is, y = 111.27 + 2.37 months employed.

Now everything is ok, that is not important. Apart from this, we have to draw the residual plots, we have to look at the behavior of the residual plots. That will say whether or model is correct or not. So, I am plotting the residual, so this is my residual value, then the residual

plot what is going to be there in x-axis, I am going to have the y predicted value, in y-axis we are going to have standardized residual value. This is my y hat, y predicted value.

Now, I am going to draw the scatter plot between y hat and standardized residual. Look at this one, there is a curvilinear relationship. It is not the straight line. It is not coming in the rectangular shape, so it is suggesting that instead of going for linear relationship, you go for non-linear or curvilinear relationship that may be the better data model for the given set. So what we are going to do.

We have one independent variable that we are going to square that. You may ask why we have to square. You can go for cube also, you can go for power 3, power 4, power 5, but at the beginning we start with the power 2. So this was my squared value. Now, this variable is taken as another independent variable. Now, we are having two independent variable, one independent variable is MonthsEmployed, another independent variable square of that MonthsEmployed.

Now that variable also look at this one. X_sq that is taken as another independent variable. Now, we will run the Regression equation. Now, what is happening. When you look at this, the R square is previously 0.7, now it is 0.9, so the R-square value is improved. So the model is a good model. The adjusted R-square also 0.886. So, the model is good. You look at the P-value of F statistics.

When you look at the P value, this is very low, less than 0.05, that means the overall model is significant. Now, look at the two independent variable, one is x1, another one is x2 square. Here, the x2 is nothing but the square of the first independent variable. When you look at the p-value for the first variable, it is 0.00, for the second variable it is 0.002. That is where the squared term. So both the p values are less than 0.05, we can conclude that both independent variables are significant.

It is not enough to check only the individual significant, overall significant and R square. Apart from this, we have to go for residual plot. So, when you go for residual plot for our second model, so this is the standardized residual, now we will go for predicted value of new y, that is y hat 2, and now we will plot it. Now, it shows there is no curvilinear relationship.

Now, we can say that in this model, we can go for a simple linear relationship when you go for curvilinear relationship that is the best model for a given data.

In this lecture, we have seen how to do a curvilinear regression model. In our previous lecture, I have explained how to do simple linear regression and multiple linear regression model, but in this lecture I have given you an example when we should go for curvilinear relationship between x and y. We have taken one example, in that example, we have taken our first simple linear relationship between x and y.

Then we look at the residual plot, that residual plot suggested that we should go for non-linear relationship, so we have squared that independent variable. Again, we have constructed a new regression model. In that, we have realized that when we look at the residual plot of the new model, we realize that the curvilinear model is the better model for the given data when compared to simple linear relationship.

In the next class, we will go for interaction, how to do if there is interaction between two independent variable x1 and x2, how to do that kind of regression model that we will see in the next class.