Data Analytics with Python

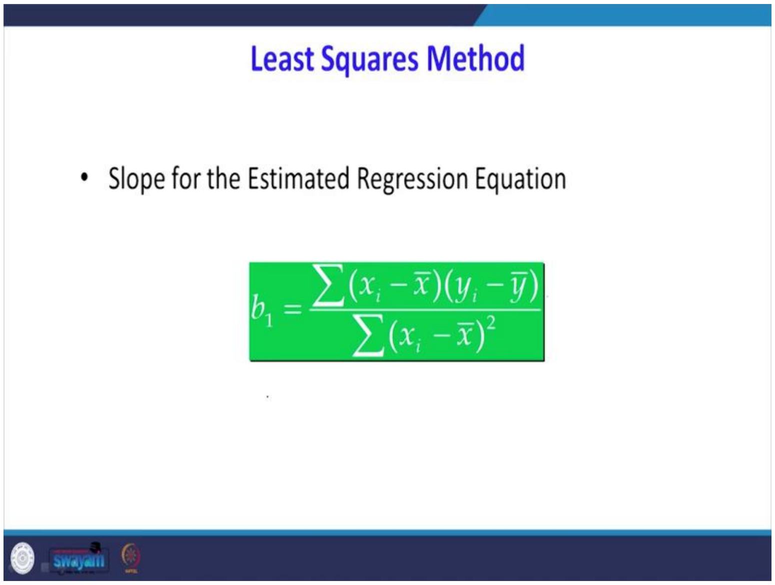
Prof. Ramesh Anbanandam
Department of Management Studies
Indian Institute of Technology – Roorkee

Lecture – 29
Linear Regression - II

Dear students in the previous class I have derived what is the formula for Slope of a linear regression equation and y intercept and also I have explained the concept of least square method. In the formula of slope then I explained slope is nothing but covariance of the two variable if it is the 1 independent variable and one dependent variable simple linear regression. Covariance divided by variance, variance of independent variable.

Then we also got another important result in the previous class that if you want to draw a best line that line has to pass through its average of x, y value that that lines to pass through average value of x that x bar and average value of y bar class what you are going to do.

(Refer Slide Time: 01:22)



Least Squares Method

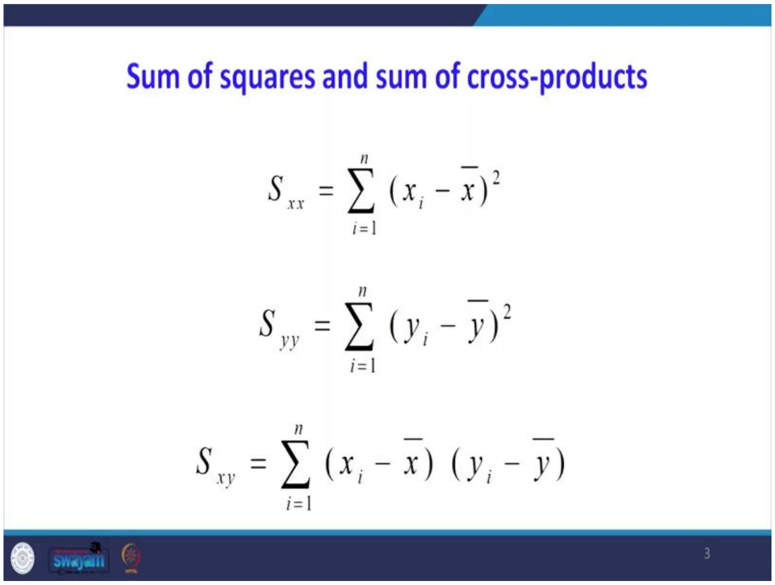
- Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

I have taken the problem with help of a small problem I going to find out the slope and y intercept then I will explain the practical meaning of this y intercept and slope this formula b1 nothing, but the slope of the line is writing y = b 0 + b 1 x to the b1 is nothing but the slope is covariance divided by variance, we know that the formula for covariance is Σ (x – x bar) . (y - y bar) divided n -1 divided by the variance.

The variance is Σ (x – x bar)^2 / (n -1) because the numerator and the denominator n -1 is same so when you cancel that the remaining formulas, Σ (x – x bar) . (y - y bar) divided by Σ (x – x bar)^2.

(Refer Slide Time: 02:22)



Sum of squares and sum of cross-products

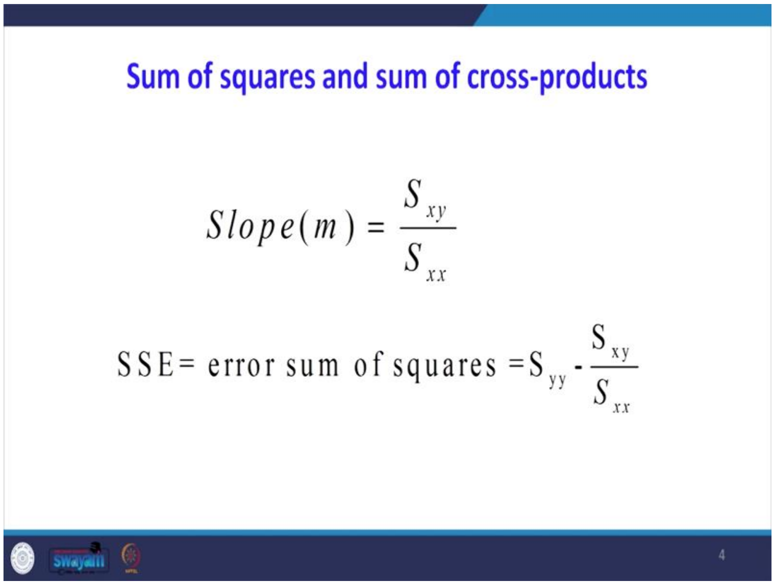
$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$$

$$S_{yy} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$S_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})\ (y_i - \overline{y})$$

If you are using calculator for examination purpose this is the very very very useful notations convention. So by using if you know Sxx the meaning of this is Sxx is Σ (x – x bar)^2 the meaning of yy is, it is a convention is Σ (y – y bar)^2 if I write S xy that is nothing but S xy equal to Σ (x – x bar) . (y - y bar).

(Refer Slide Time: 03:01)



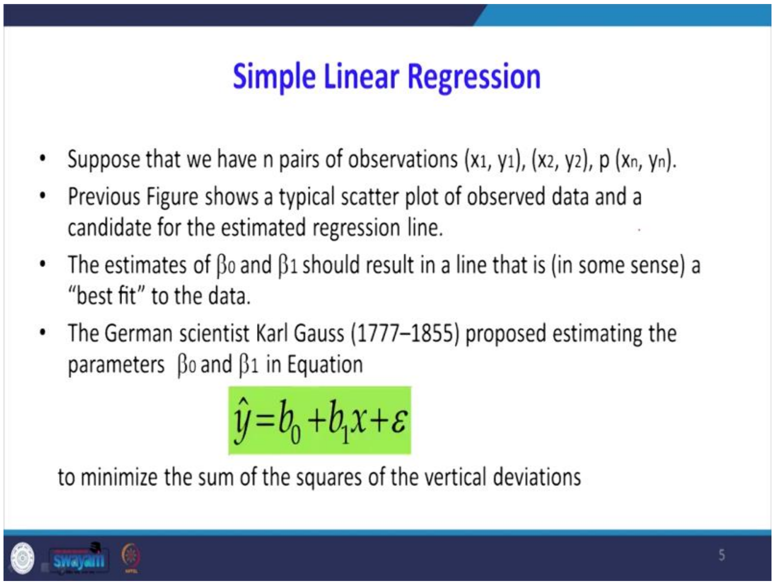
Sum of squares and sum of cross-products

$$Slope(m) = \frac{S_{xy}}{S_{xx}}$$

$$\text{SSE= error sum of squares} = S_{yy} - \frac{S_{xy}}{S_{xx}}$$

The formula for slope is nothing but slope m equal to S xy divided by S xx. If I want to know error sum of square, I will explain what is the meaning of error sum of square. Now you take this formula what is error sum of square equal to S yy – (S xy divided by S xx) suppose why this formula so useful suppose if you know this 3 term S xx, S yy, and S xy you can find out the slope. And you can find out the error sum of square and this is convenient also later to find out the coefficient of determination. I will explain the next lecture.

(Refer Slide Time: 03:52)



Simple Linear Regression

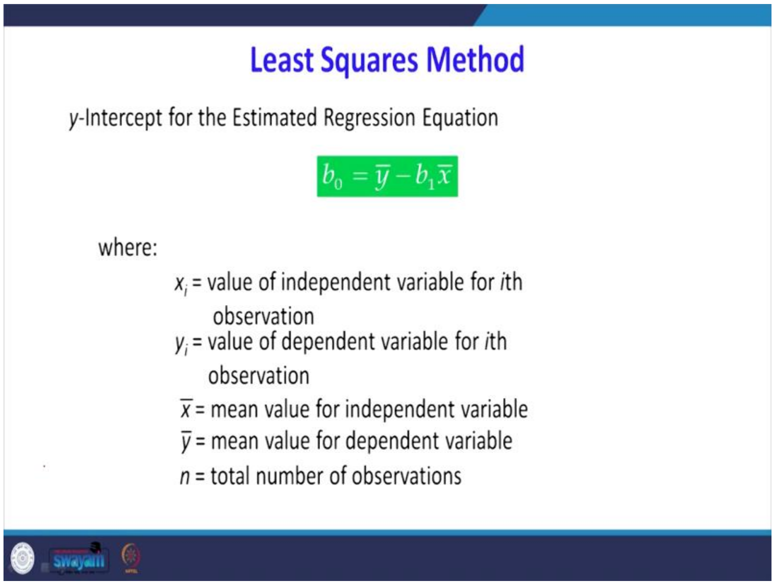
- Suppose that we have n pairs of observations (x_1, y_1), (x_2, y_2), p (x_n, y_n).
- Previous Figure shows a typical scatter plot of observed data and a candidate for the estimated regression line.
- The estimates of β_0 and β_1 should result in a line that is (in some sense) a "best fit" to the data.
- The German scientist Karl Gauss (1777–1855) proposed estimating the parameters β_0 and β_1 in Equation

$$\hat{y} = b_0 + b_1 x + \varepsilon$$

to minimize the sum of the squares of the vertical deviations

So what is simple linear regression suppose that we have n pairs of observation se(x1, y1),(x2, y2) like this (xn, yn).

(Refer Slide Time: 04:07)



Least Squares Method

y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1 \bar{x}$$

where:

x_i = value of independent variable for *i*th observation

y_i = value of dependent variable for *i*th observation

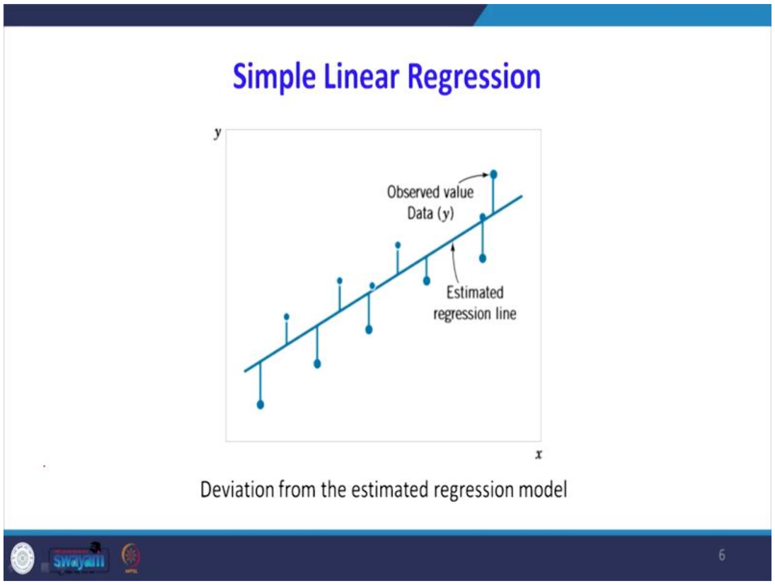
$\bar{x}$ = mean value for independent variable

$\bar{y}$ = mean value for dependent variable

n = total number of observations

Previously we have seen the formula for slope. Now the formula for y- intercept to be zero equal to y bar - b 1 x bar, explain how this formulas come here x1 is value of independent variable for the ith observation, yi is the value of dependent variable for right observation x bar is the mean of mean value for independent variable, y bar is mean value of value for dependent variable n actually we not using here n is the total number of observations to find out the mean value.

(Refer Slide Time: 04:46)



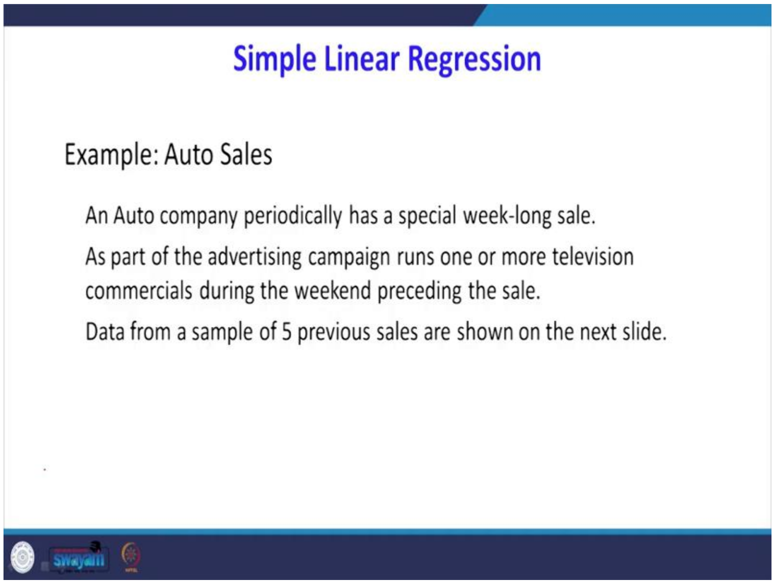
Simple Linear Regression

Deviation from the estimated regression model

Simple Linear Regression is you this is your x axis. This is your y-axis. So whatever value which is said that is observed value actual value this line shows the critical value. So this line generally written as b 0 + b 1 x the final objective in a regression equation is to find out what is the slope of this line and y intercept of a line because if you know slope and y intercept, so this was your y intercept, if you know, y intercept and slope then you can construct the regression equation.

The previous class already explain this is your error one. This is error 2 this is error 3 so the concept of least square method is the sum of the square of the error has to be minimised. So that idea is taken care to find out the value of slope and y intercept.

(Refer Slide Time: 05:57)



Simple Linear Regression

Example: Auto Sales

An Auto company periodically has a special week-long sale.

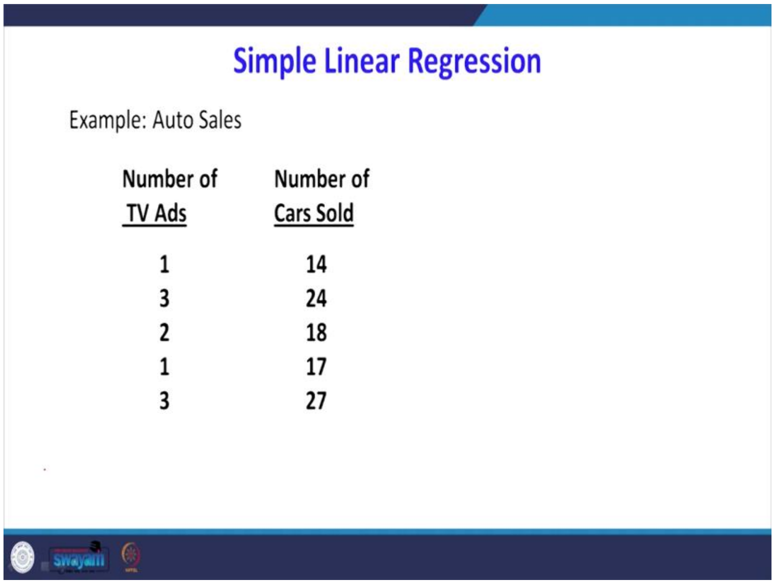
As part of the advertising campaign runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown on the next slide.

We will take one example simple linear regression why it is called simple linear regression only one independent variable is there. If there are more than one independent variable that you will call it as multiple linear regression small problem and explain how to construct a regression equation and how to use this formula of slope and y intercept. An auto Company periodic special week-long sale as a part of the advertising campaign Company runs one or more television commercials during the weekend preceding the sale.

Data from a sample of 5 previous sales are shown in the next slide actually the company before introducing a new product. They go for television advertisement. This problem says, is there any effect of television advertisement on the sales of the car?

(Refer Slide Time: 06:56)



Simple Linear Regression

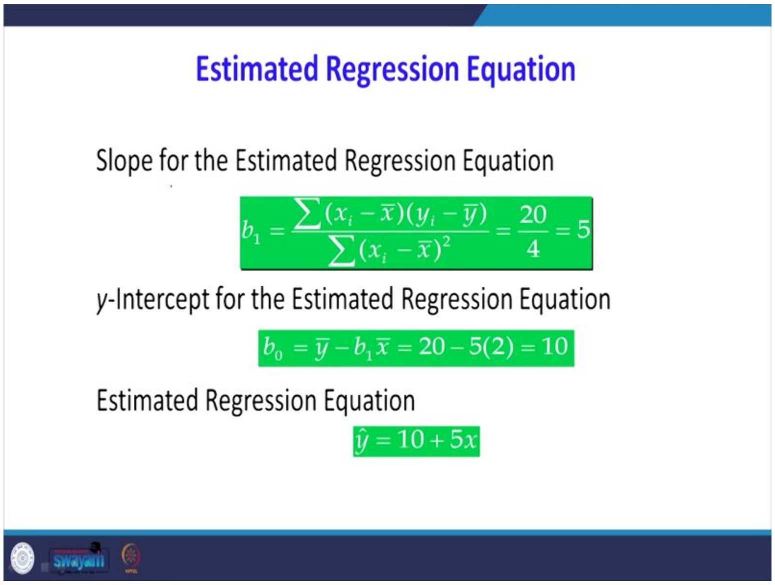
Example: Auto Sales

Number of TV Ads	Number of Cars Sold
1	14
3	24
2	18
1	17
3	27

The data says number of TV ads 1 the number of cars sold 14 when the number of TV ads is 3 number of cars sold is 24, number of TV ads is 2 number of cars sold is 18 number of TV ad is 1 number of cars sold is 17 number of TV ads is 3 number of cars sold 27. In this the dependent variable is number of cars sold this generally we will call it as y is dependent variable. The independent variable is x that is nothing but number of TV ads.

So we have to know effect of this number of ads on the number of cars sold. Generally what is perception when you the frequency of ads is more than be more sales.

(Refer Slide Time: 07:51)



Estimated Regression Equation

Slope for the Estimated Regression Equation

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{20}{4} = 5$$

y-Intercept for the Estimated Regression Equation

$$b_0 = \bar{y} - b_1\bar{x} = 20 - 5(2) = 10$$

Estimated Regression Equation

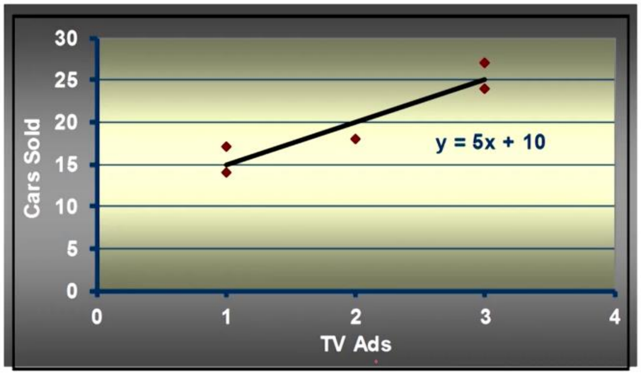
$$\hat{y} = 10 + 5x$$

The first task is the Slope of the estimated regression equation. So what we have to do first we have to find out x bar and y bar then for each value of x and find out x - x bar and for each value of y you find out y - y bar then you have to multiply that and you have to sum that multiplication that lead to 20. Then sigma of x - x bar you have to square that then you to submit that is 4. So the slope is 5 the y intercept for estimated regression equation is b 0 is y bar – b 1 x 1 bar already we know b1.

So you take b1 value here y bar you know it is 20 the x bar is 2 this is 10 so the estimated regression equation is 10 + 5 x, you are to be very careful. This is estimated regression equation. It is not why the value of y is the estimated value it is not the actual value it is nothing but the mean value. So, y equal to 10 + 5 x so how to interpret this, when the value of x is increasing 1 unit y will be increased by 5 units so keeping other things constant when x is increased by 1 unit the sales will increase by 5 times not the 15 times it is not right 5 into 1 equal to 15. We have seen the rate of increment of x and rate of increment of y.

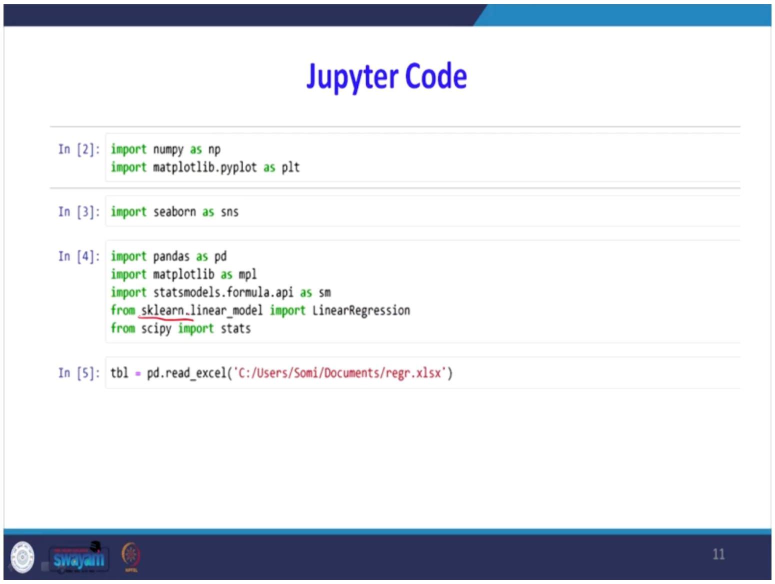
(Refer Slide Time: 09:35)

Scatter Diagram and Trend Line



So, that this show that y = 5 x + 10. So here the 10 is y intercept when you extend this. This is a y intercept. Ok 5 is the slope. So suppose if the TV number of TV Advertisement is a 5 now we have taken up to 4 suppose this is 5 you can put to here 5, 25 + 10, 35. This way; this is a trend line is a regression line.

(Refer Slide Time: 10:08)

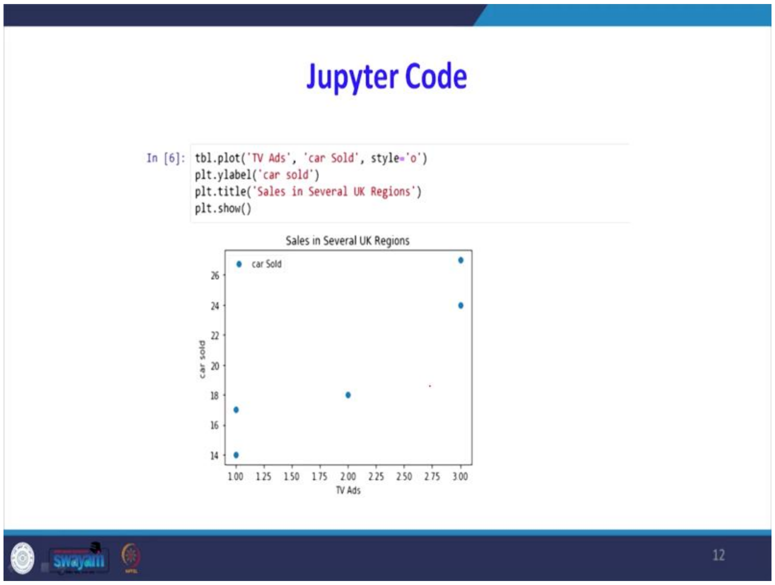


Jupyter Code

```
In [2]: import numpy as np
        import matplotlib.pyplot as plt

In [3]: import seaborn as sns

In [4]: import pandas as pd
        import matplotlib as mpl
        import statsmodels.formula.api as sm
        from sklearn.linear_model import LinearRegression
        from scipy import stats

In [5]: tbl = pd.read_excel('C:/Users/Somi/Documents/regr.xlsx')
```

This one will do with the help of python import numpy as np import matplotlib.pyplot as plt import seaborn as sns import Pandas as pd import matplotlib as mpl import statsmodels.formula.api as sm, from sklearn and see that this one is sklearn that is the library for running linear regression sklearn.linear_model import linear regression, from scipy import stats first I have entered this value in excel and going to save that filename object called tb1, tb1 equal to pd.read_excel. This was the path where I have stored my excel file.

(Refer Slide Time: 11:16)

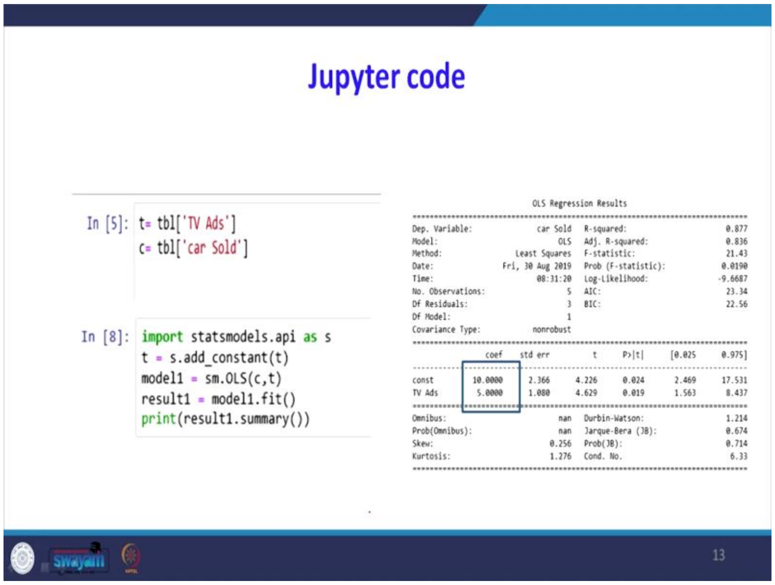


Jupyter Code

```
In [6]: tbl.plot('TV Ads', 'car Sold', style='o')
        plt.ylabel('car sold')
        plt.title('Sales in Several UK Regions')
        plt.show()
```

First task is your to plot the scatter plot scatter plot is we have to see only we can go for correlation here. Also, we can discuss scatter plot will say rough idea about what will happen when the value of x increases and how it is affected y what is happening? You see that you want some point when the number of TV ads increasing the car sales also increasing so tbl.plot (‘TV ads’,’cars sold’, style = ‘o’), plt.ylabel(“ cars sold”). Plt.title(‘ sales in UK regions’), plt.show(). The that will show the this graph.

(Refer Slide Time: 12:01)



Jupyter code

```
In [5]: t= tbl['TV Ads']
        c= tbl['car Sold']

In [8]: import statsmodels.api as s
        t = s.add_constant(t)
        model1 = sm.OLS(c,t)
        result1 = model1.fit()
        print(result1.summary())
```

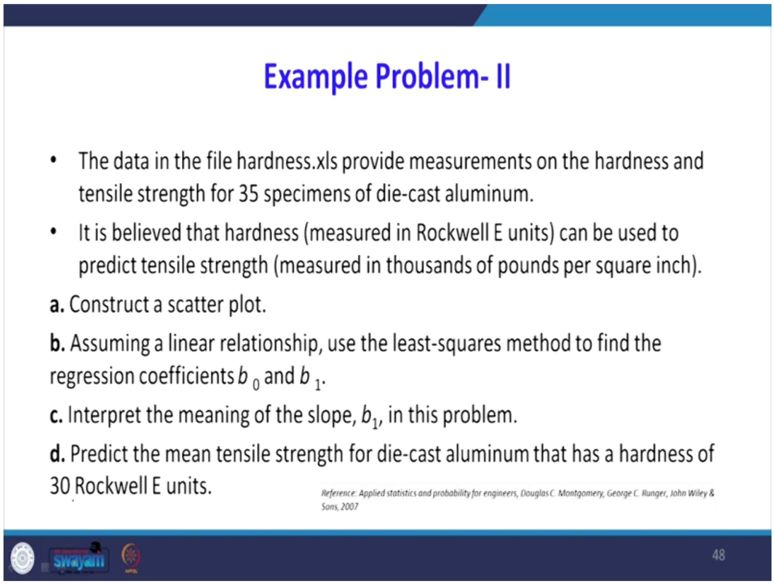
```
                            OLS Regression Results
==============================================================================
Dep. Variable:               car Sold   R-squared:                       0.877
Model:                            OLS   Adj. R-squared:                  0.836
Method:                 Least Squares   F-statistic:                     21.43
Date:                Fri, 30 Aug 2019   Prob (F-statistic):             0.0190
Time:                        08:31:20   Log-Likelihood:                -9.6687
No. Observations:                   5   AIC:                             23.34
Df Residuals:                       3   BIC:                             22.56
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         10.0000      2.366      4.226      0.024       2.469      17.531
TV Ads         5.0000      1.080      4.629      0.019       1.563       8.437
==============================================================================
Omnibus:                          nan   Durbin-Watson:                   1.214
Prob(Omnibus):                    nan   Jarque-Bera (JB):                0.674
Skew:                           0.256   Prob(JB):                        0.714
Kurtosis:                       1.276   Cond. No.                         6.33
==============================================================================
```

Next one I am going to save that TV ads in variable quantity equal to t = tv ads in and c = car sold here. The car sold is dependent variable TV ads is independent variable import statsmodels.api as s, so t = s.add_constant(t) because we need to have the t so model1 I am saying model1 = sm.OLS(c,t) , ols means ordinary least square method c, t. What is c here? c is your dependent variable t is your independent variable result1 equal to model1.fit() so print(result1.summary()).

This is was the output of your linear regression equations. So, look at this most importantly is it that the coefficient so how to write this one y equal to so the constant is 10 + 5 TV ads. There are many terms are here model is ols methos is least square when it was conducted number of observations 5 number of residuals residual is nothing but you error here your r square is 0.877, I will explain the meaning of 0.877 in the next class.

Then the adjusted r square this value of adjusted r squared interpreted for multiple regression equation. I will explain what is this F statistics in next class then there are many fitness index is there. So these standard error this is the t value I will explain t value and one more thing there in look at the probability value suppose Alpha equal to 5 percentage the probabilities less the 0.05 then you can say it is significant I will explain this also in the next class. So this is the output of our regression equation. We will take another problem on regression analysis.

(Refer Slide Time: 14:33)



Example Problem- II

- The data in the file hardness.xls provide measurements on the hardness and tensile strength for 35 specimens of die-cast aluminum.
- It is believed that hardness (measured in Rockwell E units) can be used to predict tensile strength (measured in thousands of pounds per square inch).

a. Construct a scatter plot.

b. Assuming a linear relationship, use the least-squares method to find the regression coefficients b_0 and b_1.

c. Interpret the meaning of the slope, b_1, in this problem.

d. Predict the mean tensile strength for die-cast aluminum that has a hardness of 30 Rockwell E units.

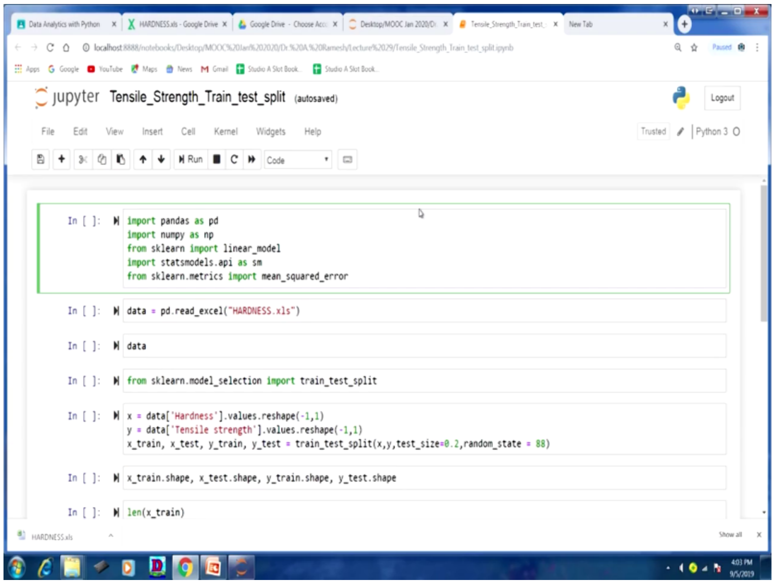
Reference: Applied statistics and probability for engineers, Douglas C. Montgomery, George C. Runger, John Wiley & Sons, 2007

48

The problem is the data in the file. I have a file called hardness.xlsx, provides measurement on the hardness and tensile strength for 35 specimen of die cast aluminium. It is believed that hardness that is measured in Rockwell E unit can be used to predict the tensile strength measured in 1000 of Pounds per square inch. So, what are the things used to do is construct a scatter plot assuming a linear relationship. Use the least square method to find the regression coefficient for b0 and b1 interpret the meaning of slope b1 in this problem.

Predict the main tensile strength for the die cast aluminium that has hardness of 30 Rockwell E unit. Today is a tensile strength is given hardness is given for this data set. We are going to construct a regression equation. So will switch to Python I will tell you how to do that.

(Refer Slide Time: 15:27)



jupyter Tensile_Strength_Train_test_split (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Trusted
Python 3
Run
Code
HARDNESS.xls
Show all

```
In [ ]: import pandas as pd
        import numpy as np
        from sklearn import linear_model
        import statsmodels.api as sm
        from sklearn.metrics import mean_squared_error

In [ ]: data = pd.read_excel("HARDNESS.xls")

In [ ]: data

In [ ]: from sklearn.model_selection import train_test_split

In [ ]: x = data['Hardness'].values.reshape(-1,1)
        y = data['Tensile strength'].values.reshape(-1,1)
        x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2,random_state = 88)

In [ ]: x_train.shape, x_test.shape, y_train.shape, y_test.shape

In [ ]: len(x_train)
```

For this import pandas as pd, import numpy as np, from sklearn import linear_model, import statsmodels.api as sm and from sklearn.matrix import mean_squared_error. first I will load the data object called data. So run It import pandas as pd, import numpy as np, from sklearn import linear_model, import statsmodels.api as sm and from from sklearn.matrix import mean_squared_error the file I have stored in a object called data.

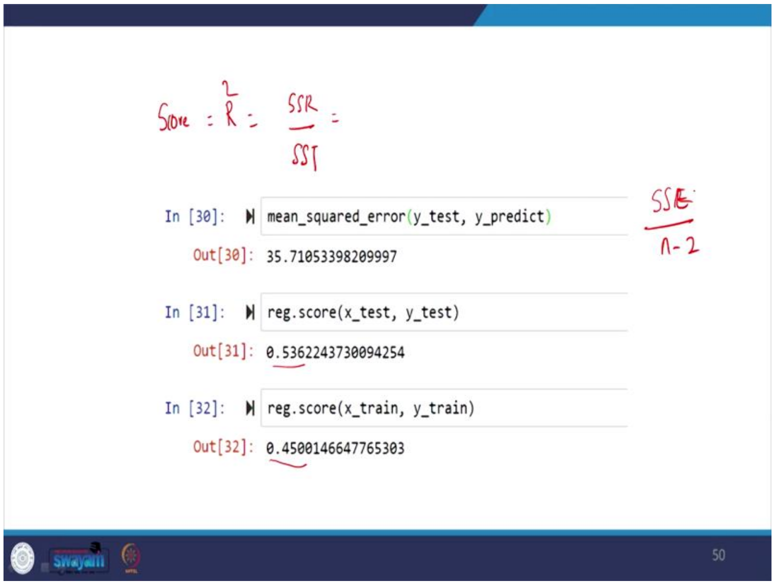
So the data source this is the tensile strength and hardness. What are going to do? There are 35, data set that we are going to split into two categories only for training and other only for testing. For that purpose from sklearn.model_selection import train_test_split, x equal to data so x value is going to be hardness dot values dot where reset command is used to convert one dimensional array into 2 dimensional array, then y equal to data that is a tensile strength.

So, x is independent variable, y is dependent variable. The next one is x underscore train, x underscore test, y underscore train, y underscore test equal to train underscore test underscore split x, y test underscore size equal to 20%, so this 20% What is the meaning is there 20% the data will be kept for testing our model remaining 80% of the data will be used for building our regression model. So random underscore stat equal to you can give any number so that when you repeat this program again, you will get the same answer.

Because the 20% of the data is randomly chosen out of 35 data set. So if you use this shape command now you can see I will run this one, so you are getting 28, 1, so 28 data sets for training 7 data status for testing ok. So, we can see the length also also can see that the 28 for training 7 is for testing this is the train data sets. Now will go for constructing the regression model from a scale and linear underscore model input linear regression, so regression equal to Linear regression when you run they will run this model.

So now we will see what is y interested? Now y intercept is 7.045 so the regression coefficient is 1.9974. Now will predict for the test data set will predict what is the y value? So this is your predicted y value when giving x data set as an input. Now we will find out what is the mean square error the mean squared error, the mean squared error is 35, if the error in smaller that model is good model. The next one is the fitness of this regression model is nothing 0.53 by taking x underscore data set an independent variable and y underscore data set is dependent variable. The next is when you set for training data set it is 0.45 explain the meaning of this score data.

(Refer Slide Time: 19:21)



$\text{Score} = R^2 = \frac{SSR}{SST} =$

$\frac{SS\cancel{E}}{n-2}$

```
In [30]:  mean_squared_error(y_test, y_predict)

Out[30]:  35.71053398209997

In [31]:  reg.score(x_test, y_test)

Out[31]:  0.5362243730094254

In [32]:  reg.score(x_train, y_train)

Out[32]:  0.4500146647765303
```

This the meaning of score is meaning of score is nothing but your r square. This r square is nothing but coefficient of determination. So, the coefficient of determination is your SSR divided by SST, SSR is regression sum of square, SST is total sum of square in your problem the r square is see that the reg.score (x_test , y_test) is the r square is 53%. What is meaning of this 53 is the 53% of the variable variability of y is explained with the help of this is dependent variable for the training data set.

It is for the training dataset it is 0.45. that is the 45% of the variability of y is explain with help of independent variable x this mean square mean squared error is nothing but SSE divided by n-2 that is mean square error. If it is the lesser value the model is good fit. Otherwise, it is not good. Now will see the another concept in the regression model that is called machine learning in machine learning this is one category of supervised learning. The machine learning techniques are classified into two categories under supervised learning method and unsupervised learning method.

So the regression is example for supervised learning because in advance we are labeling what is independent variable and what is dependent variable if it is unsupervised learning. So, we cannot label in advance that what is going to be dependent and what is independent variable. So, in the supervised learning is nothing but the regression analysis. That in the context of machine learning will call it is supervised learning in statistics will call it is simple regression.

Dear students in the previous class derive the formula for y intercept and slope. Then I have taken one sample problem with the help of sample problem explain how to use that formulas and also called with the help of python. I have taken another problem also in that problem. The data set is divided into two parts, 1 part is for Training the for building the data set the using training data set the other part of the data set that is for test data set.

So the test data set was used for validating the model which we have constructed. Thank you very much.