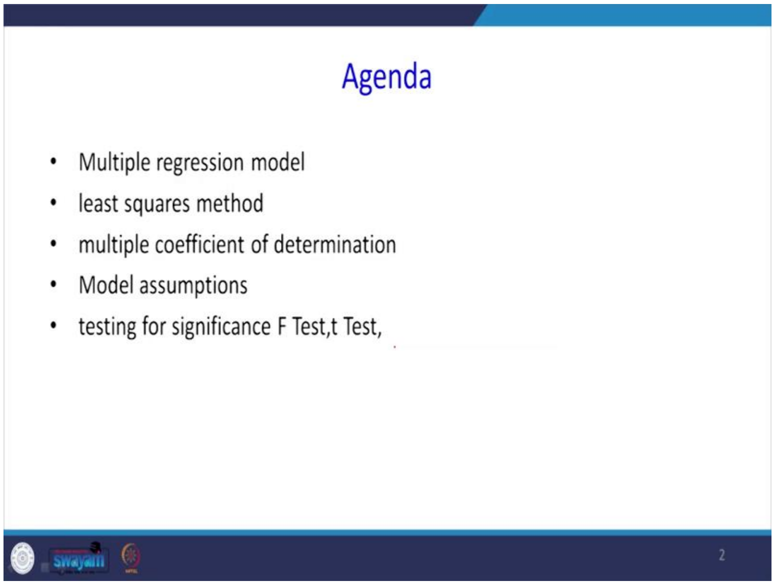
Data Analytics with Python

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

Lecture – 33
Multiple Regression Model - I

In the previous class we have studied about simple linear regression, in this class we are going to discuss about multiple regression models.

(Refer Slide Time: 00:35)

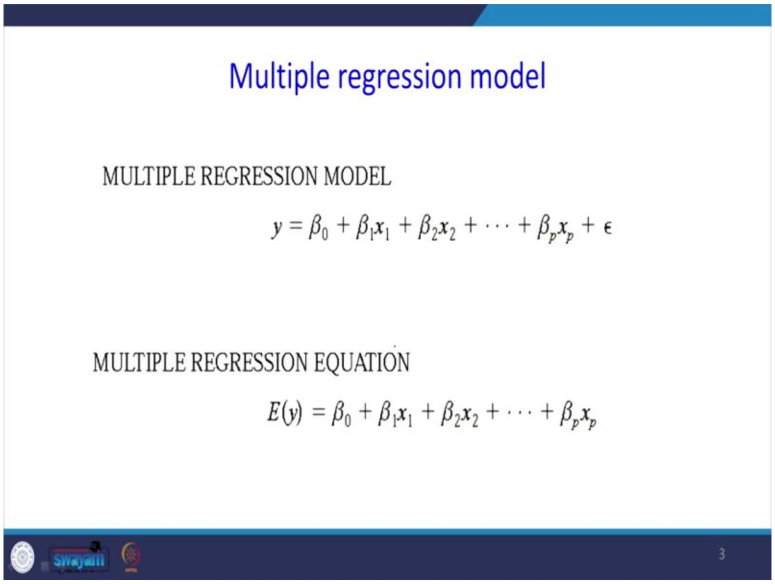


Agenda

- Multiple regression model
- least squares method
- multiple coefficient of determination
- Model assumptions
- testing for significance F Test,t Test,

The class agenda is I am going to explain what is multiple regression model then what is a least square method then multiple coefficient of determination. In the multiple coefficient of determination I am going to explain what is adjusted r-square also. Then what are the assumption in the multiple linear regression. Then I am going to test the significance of by using F test and t test.

(Refer Slide Time: 01:04)



Multiple regression model

MULTIPLE REGRESSION MODEL

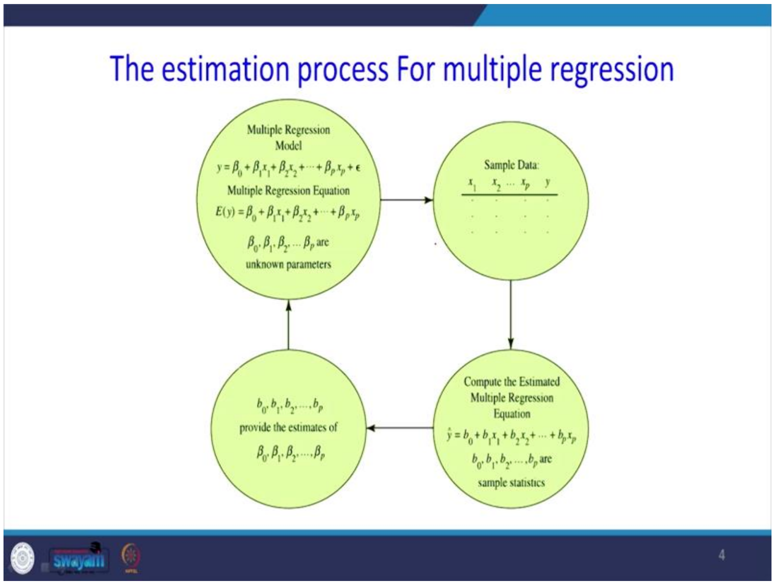
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$$

MULTIPLE REGRESSION EQUATION

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

What is a multiple regression model so multiple regression model is when there are more than one independent variable that is called multiple linear regression model. If it is only one independent variable it is linear regression model. When you take the expected value of this multiple regression model so we know that that assumption any regression equations that the expected value of error term is 0, so when you take expected value of y there would not be any error term that is that is a multiple regression equation. Here beta 1 beta 2 is the coefficient of x 1 x 2 and beta p a coefficient of x p.

(Refer Slide Time: 01:42)



The estimation process For multiple regression

Multiple Regression Model

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$

Multiple Regression Equation

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$

$\beta_0, \beta_1, \beta_2, \cdots \beta_p$ are unknown parameters

Sample Data:

x_1	x_2	$\cdots$	x_p	y
.	.		.	.
.	.		.	.
.	.		.	.

Compute the Estimated Multiple Regression Equation

$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p$

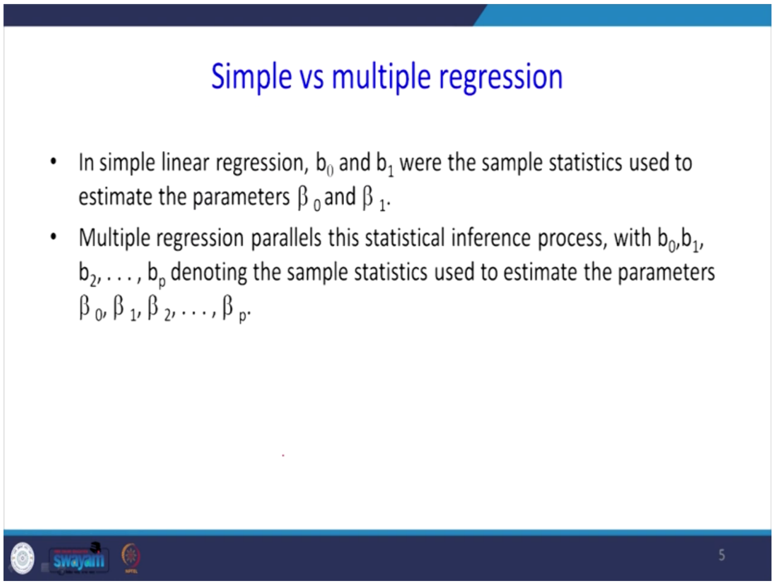
$b_0, b_1, b_2, \ldots, b_p$ are sample statistics

$b_0, b_1, b_2, \ldots, b_p$ provide the estimates of $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$

What is the estimation process for a multiple regression there is a multiple regression model y equal to beta 0 beta 1 beta 2 and beta p and x be an error term from this we can go for multiple regression equations where beta 0 beta 1 beta 2 are unknown parameters. To find out this unknown parameter from the population we are going to collect sample data for x 1 x 2 like this up to x p and sample data for y that is dependent variable. With the help of sample data we are going to construct your sample regression equation what is that compute to the estimator multiple regression equation that is y hat equal to b 0 + b 1 x 1 + b 2 x 2 and so on + b p x p where b 0 b 1 b 2 b p our sample statistics.

So with the help of sample statistics we are going to find out the population parameter that is beta 0 beta 1 beta 2 beta p then there will do a significant test then we will see that whether the beta 1 beta 2 is equal to 0 or not equal to 0 after testing that we will find out what is the actual value of beta 1 beta 2 at the population level. This is the process of doing a multiple regression model. This is similar to the simple linear regression model but what we have done in the simple linear regression model only x1 and y1 was taken only one independent variable is there but here more than one independent variable that is only difference all other concepts are same.

(Refer Slide Time: 03:21)



Simple vs multiple regression

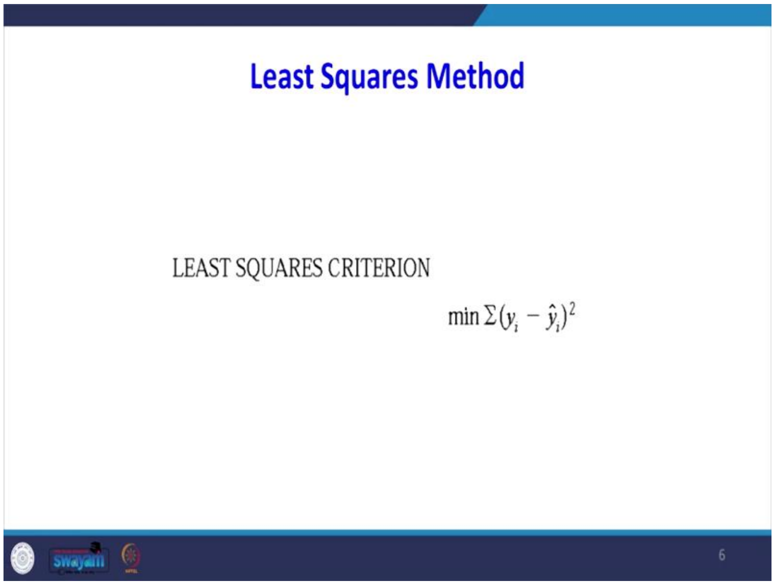
- In simple linear regression, b_0 and b_1 were the sample statistics used to estimate the parameters β_0 and β_1.
- Multiple regression parallels this statistical inference process, with b_0,b_1, b_2, . . . , b_p denoting the sample statistics used to estimate the parameters β_0, β_1, β_2, . . . , β_p.

5

So, what is simple versus multiple regressions. In simple linear regression b0 b1 bear the sample statistics used to estimate the parameter of beta0 and beta1 but in multiple regression the parallel is that the statistical inference process with b0 b1 b2 and bp denoting the sample statics are used to estimate the parameter of beta 0 beta 1 beta 2 and beta b. So, what is the meaning of this one is with the help of sample statistics b 0 beta 1 beta 2 we are going to predict the population parameter the beta 0 beta 1 and beta 2.

In simple regression there was only b 0 was there beta 1 was there only one independent variable in multiple regression more than one independent variable there is only difference.

(Refer Slide Time: 04:18)



Least Squares Method

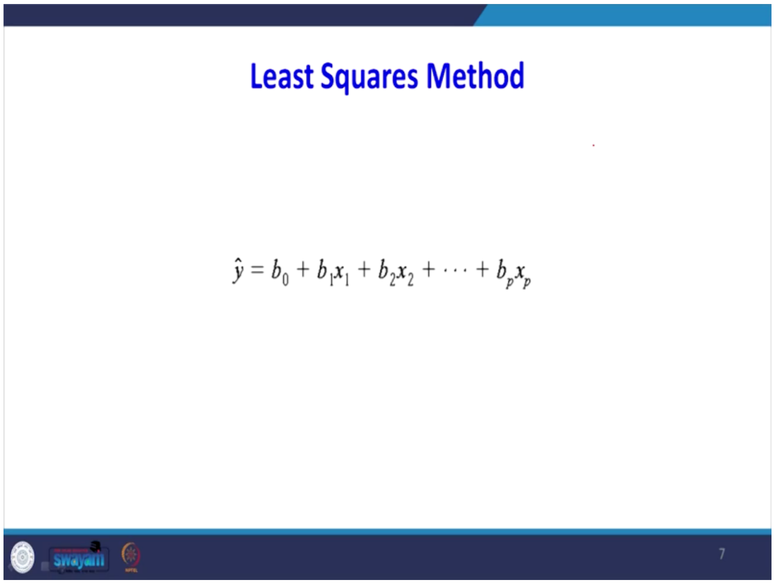
LEAST SQUARES CRITERION

$$\min \Sigma (y_i - \hat{y}_i)^2$$

Least square method in simple linear regression also I have derived the formula for b 0 b 1 by having the assumption that when we draw a line the error term that is the sum of the square of the error has to be minimized. But the y hat there in simple linear regression y hat I was b 0 + b 1 x 1 but in multiple regression this y hat I equal to b 0 + b 1 x 1 + b 2 x 2 and so on + b p x p, p is the number of independent variable.

So, all other procedure is same here also what we are going to do that there are but here it is a multi-dimensional picture we cannot draw a two-dimensional picture because we need it because there are more than one independent variable that is going to be a a multi-dimensional picture that we cannot explain with the help of a simple graph.

(Refer Slide Time: 05:17)



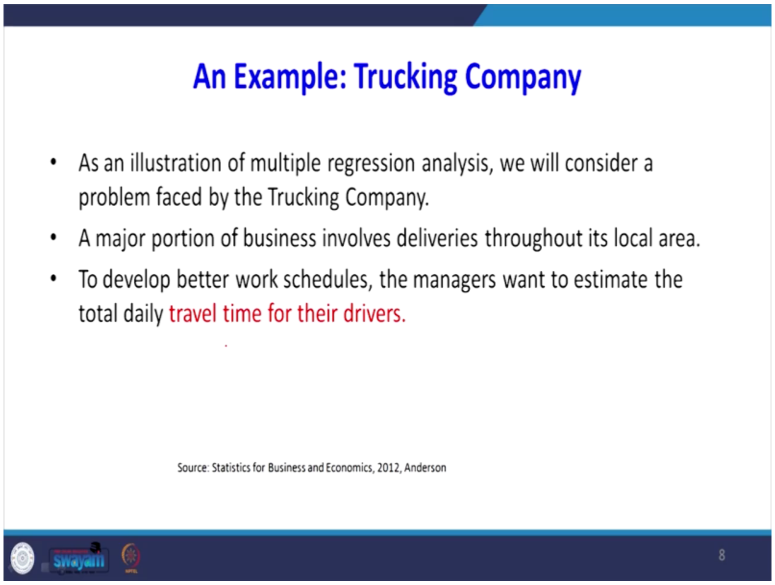
Least Squares Method

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_px_p$$

7

The least square estimate what happened to y hat equal to b 0 + beta 1 x 1 beta 2 x 2 up to beta p and x p because there would not be error term here because the expected value of the error term becomes 0. So, how to interpret the value of b 1 b 2 and b 3 how do you interpret the coefficient of b 1 is by keeping other variables constant if the x 1 is improved by one unit the y hat will be improved by b 1 units. It is a similar way for simple linear regression but here when you are interpreting one coefficient we have to assume that that we other coefficient for other independent variables are constant.

(Refer Slide Time: 06:03)



An Example: Trucking Company

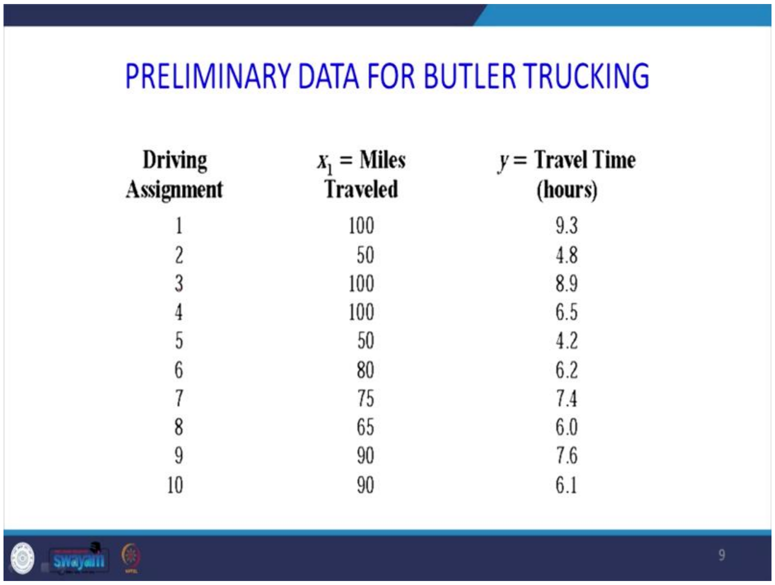
- As an illustration of multiple regression analysis, we will consider a problem faced by the Trucking Company.
- A major portion of business involves deliveries throughout its local area.
- To develop better work schedules, the managers want to estimate the total daily travel time for their drivers.

Source: Statistics for Business and Economics, 2012, Anderson

8

We will take an example this example problem is taken from statistics for Business and Economics is the other by Andersen. As an illustration of multiple regression analysis we will consider a problem faced by a tracking company the major portion of the business involves deliveries throughout the local area to develop a better work schedule the manager want to estimate total daily travel time for their drivers. So, they want to estimate this is going to be total daily travel time is going to be our dependent variable.

(Refer Slide Time: 06:43)

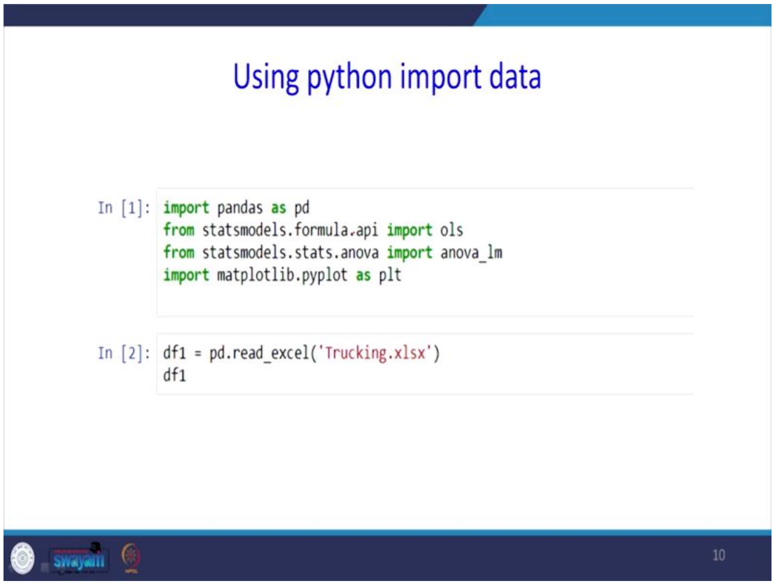


PRELIMINARY DATA FOR BUTLER TRUCKING

Driving Assignment	x_1 = Miles Traveled	y = Travel Time (hours)
1	100	9.3
2	50	4.8
3	100	8.9
4	100	6.5
5	50	4.2
6	80	6.2
7	75	7.4
8	65	6.0
9	90	7.6
10	90	6.1

There are 10 assignments there are 10 a semi drivers x 1 equal to miles traveled y equal to travel time there is a connection between x 1 and y what is the meaning of that 1 when the travel time we will increase distance traveled also high. So, y is the dependent variable x 1 is independent variable.

(Refer Slide Time: 07:03)



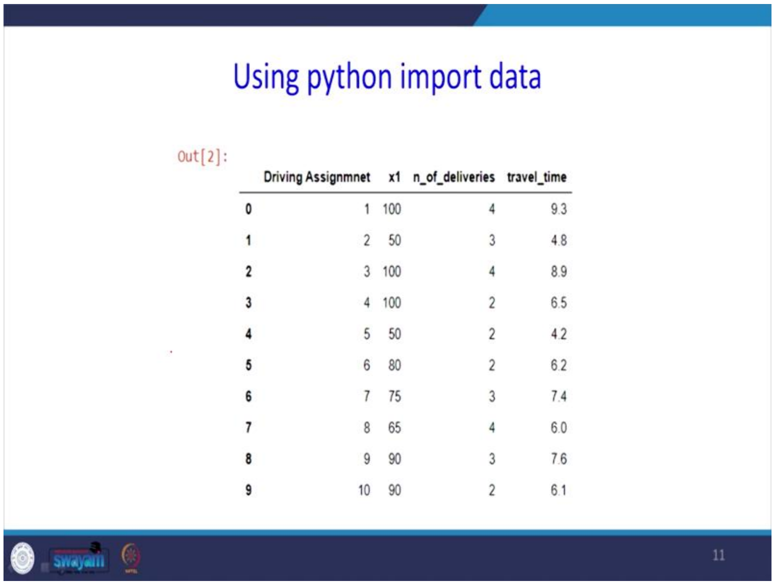
Using python import data

```
In [1]: import pandas as pd
        from statsmodels.formula.api import ols
        from statsmodels.stats.anova import anova_lm
        import matplotlib.pyplot as plt

In [2]: df1 = pd.read_excel('Trucking.xlsx')
        df1
```

I have brought the screenshot at end of this lecture I will run this quotes then you can understand it better when I will show that I will explain the screenshot import pandas as pd from stats model dot formula dot api import Wireless that is ordinary least square regression models from stats model dot stats dot anova input anova underscore lm because this library will be used to see the ANOVA table for a regression model.

Then import matplotlib dot pi plot as a plt the file name is which I have stored is it tracking that is an excel file I going to store this data into an object called df1, df1 equal to pd dot read underscore Excel that file name, so if you want to know what is the data set this is the data set.

(Refer Slide Time: 07:54)



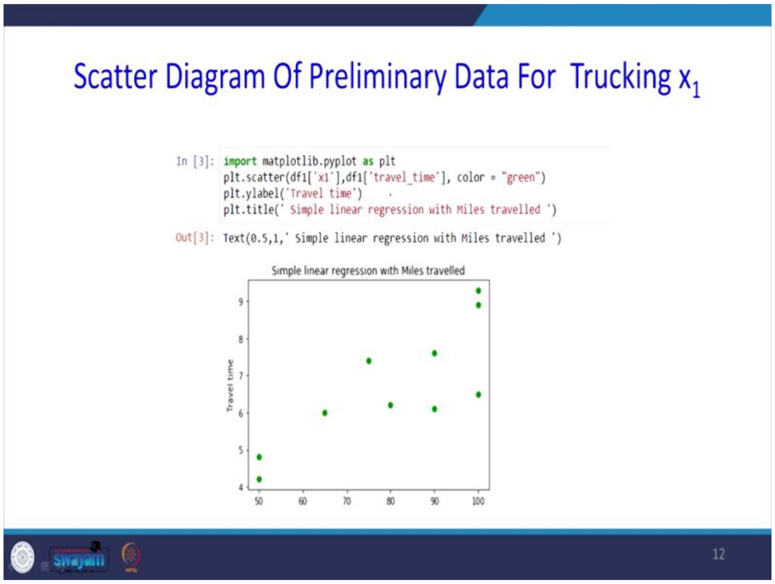
Using python import data

Out[2]:

	Driving Assignmnet	x1	n_of_deliveries	travel_time
0	1	100	4	9.3
1	2	50	3	4.8
2	3	100	4	8.9
3	4	100	2	6.5
4	5	50	2	4.2
5	6	80	2	6.2
6	7	75	3	7.4
7	8	65	4	6.0
8	9	90	3	7.6
9	10	90	2	6.1

So, in this data set there are 1 travel underscore time is dependent variable there are 2, independent variable one is x 1 and another is number of deliveries. The meaning of x 1 is miles traveled before going to regression first we ought to have an idea between this independent variable x 1 miles traveled and time dependent variable is there any connection.

(Refer Slide Time: 08:19)

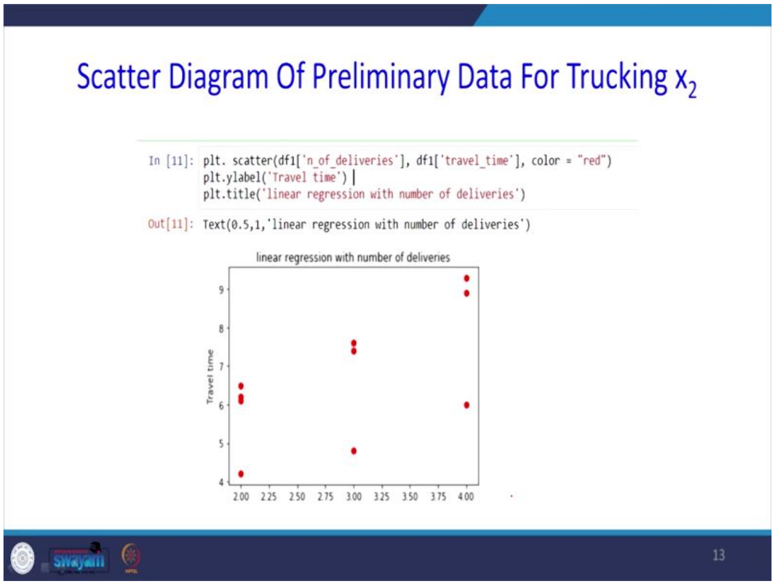


Scatter Diagram Of Preliminary Data For Trucking x_1

```
In [3]:  import matplotlib.pyplot as plt
         plt.scatter(df1['x1'],df1['travel_time'], color = "green")
         plt.ylabel('Travel time')
         plt.title(' Simple linear regression with Miles travelled ')

Out[3]:  Text(0.5,1,' Simple linear regression with Miles travelled ')
```

So the first step is first you have to draw the scatter plot so import the matplotlib dot pi plot as a PLT I am drawing the scatter plot df1 x 1 is in the x-axis travel underscore time in y-axis green color so label is travel time this one, so what is happening that there seems to be some relation between this miles traveled and the travel time that means the obviously when the miles traveled is more the travel time also will be more. This is between one independent variable and one dependent variable.

(Refer Slide Time: 08:55)

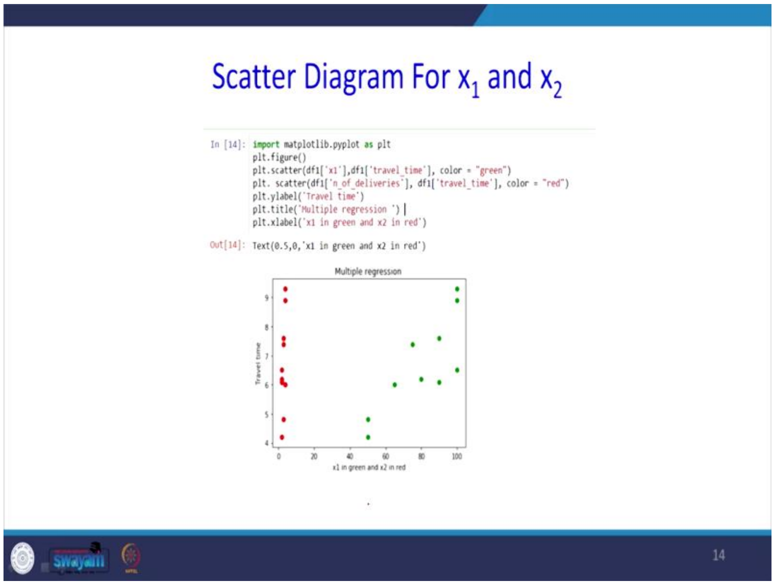


Scatter Diagram Of Preliminary Data For Trucking x_2

```
In [11]:  plt. scatter(df1['n_of_deliveries'], df1['travel_time'], color = "red")
          plt.ylabel('Travel time')
          plt.title('linear regression with number of deliveries')

Out[11]:  Text(0.5,1,'linear regression with number of deliveries')
```

Similarly we will take another variable number of deliveries as an independent variable then travel time as the dependent variable there also seems to be there is a positive correlation. Why it is required that if there is no correlation at all between that independent variable and dependent variable we need not do the regression analysis.

(Refer Slide Time: 09:20)



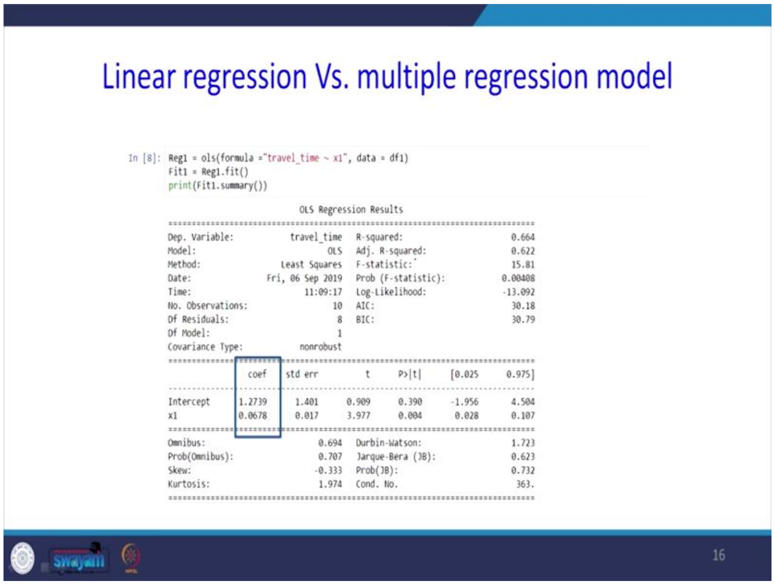
Scatter Diagram For x_1 and x_2

```
In [14]: import matplotlib.pyplot as plt
         plt.figure()
         plt.scatter(df1['x1'],df1['travel_time'], color = "green")
         plt. scatter(df1['n_of_deliveries'], df1['travel_time'], color = "red")
         plt.ylabel('Travel time')
         plt.title('Multiple regression ') |
         plt.xlabel('x1 in green and x2 in red')

Out[14]: Text(0.5,0,'x1 in green and x2 in red')
```

14

Now in this graph both the variable are taken together what is that vary the distance traveled and the number of deliveries this is the code for to show both variables in the same figure. So, what are I'm going to do first I am going to take one independent variable I am going to plot construct the regression equation then I am going to take both intermediate variables together then I go to construct a regression equation. The first taking for one independent variable this is a y hat equal to 1.27 + 0.0678 x 1 I will show you in the next slide how we got this answer.

(Refer Slide Time: 10:01)



Linear regression Vs. multiple regression model

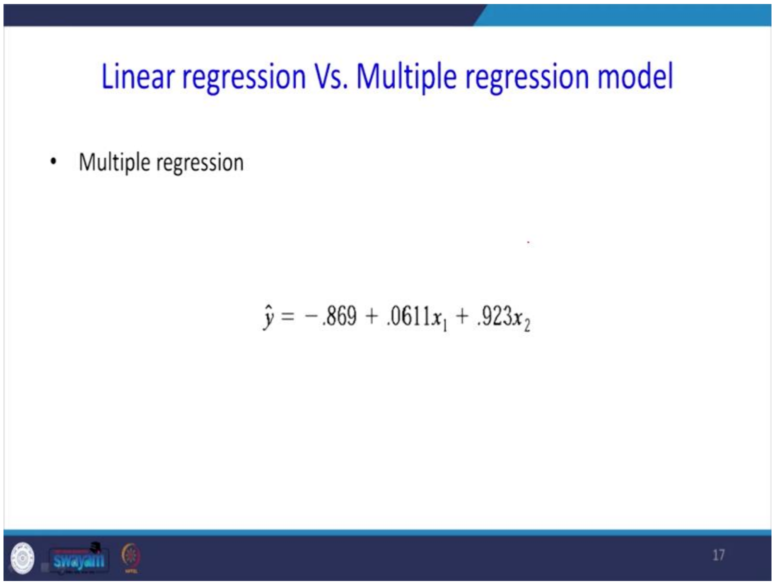
```
In [8]: Reg1 = ols(formula ="travel_time ~ x1", data = df1)
        Fit1 = Reg1.fit()
        print(Fit1.summary())

                            OLS Regression Results
==============================================================================
Dep. Variable:            travel_time   R-squared:                       0.664
Model:                            OLS   Adj. R-squared:                  0.622
Method:                 Least Squares   F-statistic:                     15.81
Date:                Fri, 06 Sep 2019   Prob (F-statistic):            0.00408
Time:                        11:09:17   Log-Likelihood:                -13.092
No. Observations:                  10   AIC:                             30.18
Df Residuals:                       8   BIC:                             30.79
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      1.2739      1.401      0.909      0.390      -1.956       4.504
x1             0.0678      0.017      3.977      0.004       0.028       0.107
==============================================================================
Omnibus:                        0.694   Durbin-Watson:                   1.723
Prob(Omnibus):                  0.707   Jarque-Bera (JB):                0.623
Skew:                          -0.333   Prob(JB):                        0.732
Kurtosis:                       1.974   Cond. No.                         363.
==============================================================================
```

16

So I am going to do a regression analysis that regression model I am going to say reg1 is equal to y LS formula the travel time is taken as the dependent variable x 1 distance traveled is taken as the independent variable. So, fit to 1 equal to reg one dot fit so print fit one dot summary so what is happening here we are getting the coefficient what is it coefficient the intercept is 1.2739 x1 is 0.0678, so how we can write it y hat equal to 1.2739 + 0.0678 x 1 variable this is an independent variable with you see that the same answer we are getting here.

So for here one more things I were to understand see the R square is 0.664 okay now the next one what I am going to do I am going to introduce another variable here after introducing the another variable I am going to see what is going to happen this r square. The r square says the goodness of the model the higher the r square the model is better what is the meaning of 66.4 here was 0.664, 66.4% of the variability of y can be explained with the help of this model.

(Refer Slide Time: 11:25)



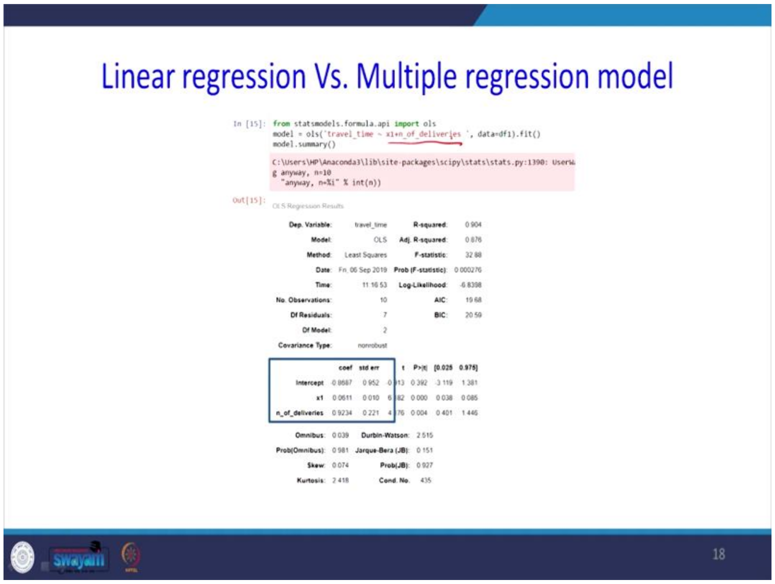
Linear regression Vs. Multiple regression model

- Multiple regression

$$\hat{y} = -.869 + .0611x_1 + .923x_2$$

Now what happening that I am going to bring another independent variable that is number of deliveries so when you bring another independent variable I will show you that model you see that model equal to ils travel underscore time tilde sign x 1 + n underscore of underscore deliveries so this is two independent variable if there are three you can write it plus that variables this is the way to do the multiple regression in Python.

(Refer Slide Time: 11:46)



Linear regression Vs. Multiple regression model

```
In [15]: from statsmodels.formula.api import ols
         model = ols('travel_time ~ x1+n_of_deliveries ', data=df1).fit()
         model.summary()

         C:\Users\HP\Anaconda3\lib\site-packages\scipy\stats\stats.py:1390: UserW
         g anyway, n=10
           "anyway, n=%i" % int(n))

Out[15]:
         OLS Regression Results

             Dep. Variable:     travel_time         R-squared:     0.904
                    Model:              OLS     Adj. R-squared:     0.876
                   Method:    Least Squares        F-statistic:     32.88
                     Date: Fri, 06 Sep 2019 Prob (F-statistic):  0.000276
                     Time:         11:16:53     Log-Likelihood:   -6.8398
         No. Observations:               10                AIC:     19.68
             Df Residuals:                7                BIC:     20.59
                 Df Model:                2
          Covariance Type:        nonrobust

                            coef  std err       t   P>|t|  [0.025  0.975]
                Intercept -0.8687   0.952  -0.913   0.392  -3.119   1.381
                       x1  0.0611   0.010   6.182   0.000   0.038   0.085
          n_of_deliveries  0.9234   0.221   4.176   0.004   0.401   1.446

                 Omnibus:  0.039    Durbin-Watson:  2.515
           Prob(Omnibus):  0.981 Jarque-Bera (JB):  0.151
                    Skew:  0.074         Prob(JB):  0.927
                Kurtosis:  2.418         Cond. No.    435
```

18

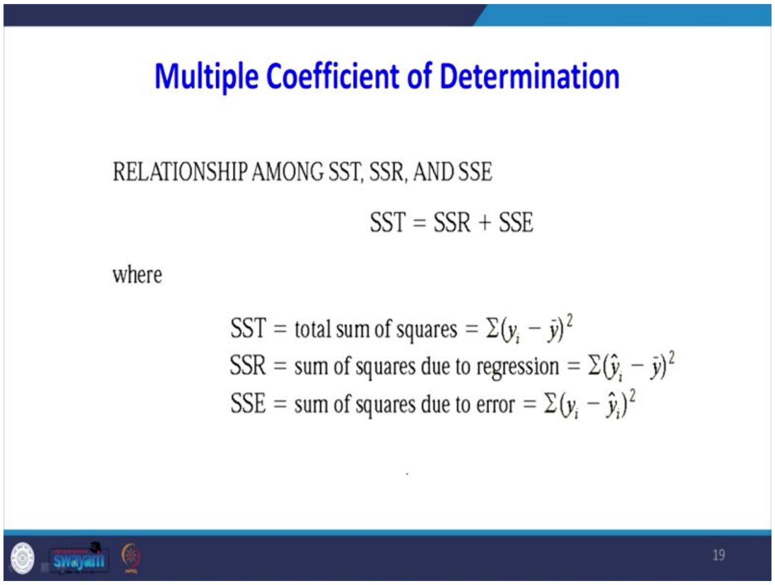
So, now what is happening here you look at the y intercept it is y equal to – 0.8687 + 0.0611 x 1 + 0.9234 x 2, here you can call it as x 2 is what is the meaning of x 2 number of deliveries okay so, what is this, this is important. We will verify this in the previous slide also we got the same thing – 0.869 + 0.0611 x 1 + 0.923 x 2 now look at this the previous r square now look at this now this r square after introducing new variable.

After introduce a new variable the r square is previously it was 0.6 something now it is increased to 0.90 so adding a new variable as helped to improve the explaining power of this regression model. Then I explain there is one more term adjusted r-square because in many previous lectures I am saying that I will do the next lecture but I am not able to do that one now in this lecture I will explain what is the meaning of adjusted r-square.

The other point you have to understand you look at the p-value for each independent variable. So, what is the null hypothesis for a year what is the null hypothesis H 0 equal to beta 1 equal to beta 2 equal to 0, so in all hypothesis for you look at the b values here see for x 1 it is a point 0 0 so we have to reject null hypothesis. When you reject null hypothesis beta 1 is not equal to 0, that mean there is a relation between x 1 and y 1.

Similarly look at the number of deliveries corresponding p-value is 0.04 that also less than 0.05 so that hypothesis beta to 0 also at we rejected that means at a population level there is a the relationship is significant what is the meaning of that one is even at the population level between x 2 and y there is a significant relationship is there.

(Refer Slide Time: 14:04)



Multiple Coefficient of Determination

RELATIONSHIP AMONG SST, SSR, AND SSE

$$SST = SSR + SSE$$

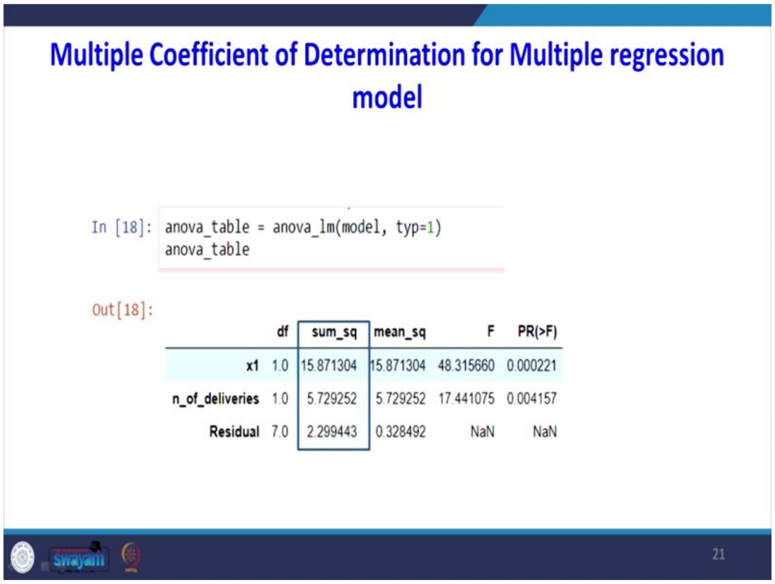
where

$$SST = \text{total sum of squares} = \Sigma(y_i - \bar{y})^2$$
$$SSR = \text{sum of squares due to regression} = \Sigma(\hat{y}_i - \bar{y})^2$$
$$SSE = \text{sum of squares due to error} = \Sigma(y_i - \hat{y}_i)^2$$

Relation among SST SSR in SSE we know that SST total sum of square equal to the regression sum of square + error sum of square, SST this I have explained in my previous lecture total sum of square is this way for your convenience I am drawing one more time this is your y bar this is your y hat so this is y, so this distance okay this distance is your SSR this distance is your SSE, so the total distance is SST.

So this total distance is SST, so, what is SST? SST is y i - y bar whole square Sigma what is SSR y hat I - y bar whole square what is SSE y - y hat whole square so when we have only one independent variable look at this here what is SST when you add this SST equal to summation of 15.87 + 8.02 so it will come around 89 SST. You see the residual sum of square so what is SSE? SSE is 8.02 when there is only one independent variable SSR is 15.871 to get this regression model output you have to use this one print ANOVA underscore LM the or to call the first regression model.

(Refer Slide Time: 16:08)



Multiple Coefficient of Determination for Multiple regression model

```
In [18]: anova_table = anova_lm(model, typ=1)
         anova_table
```

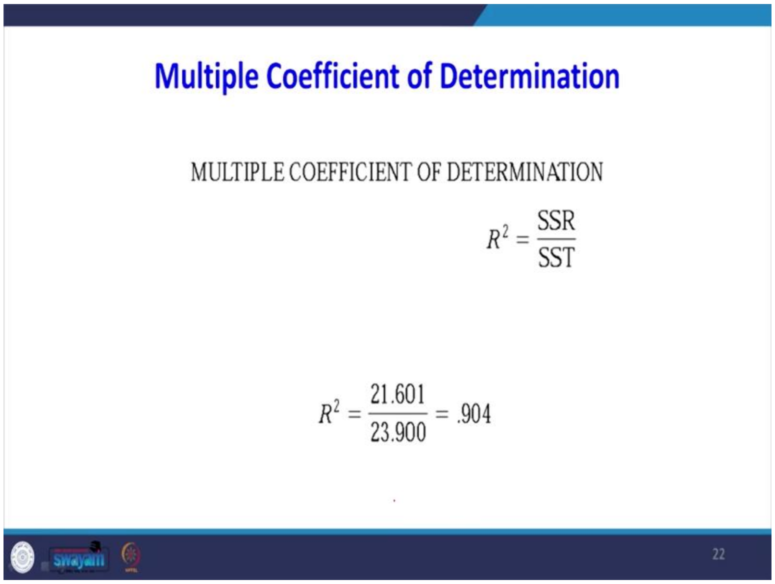
Out[18]:

	df	sum_sq	mean_sq	F	PR(>F)
x1	1.0	15.871304	15.871304	48.315660	0.000221
n_of_deliveries	1.0	5.729252	5.729252	17.441075	0.004157
Residual	7.0	2.299443	0.328492	NaN	NaN

The next slides we are going to bring another ANOVA table when there are two independent variables for that purpose and I want a score table equal to n our lm model one type one ANOVA table. Now you see that the SST is same SST is around 22 around 22 but look at SSE is 2.29 so error has been decreased. You see SSR, SSR is these two 15.87 + 5 approximately 20. Something so what is happening when you introduce a new variable the value of SSR is increased to 20 variously SSR only one independent variable SSR is 15.

So after introduced a new variable the 5 unit of variants is increased and at the other point is previously when there are only one independent variable is the error term is 8.02. Now the error is reduced to 2.29 so that is the advantage of using more number of independent variable so that we can have more accurate model.

(Refer Slide Time: 17:20)



Multiple Coefficient of Determination

MULTIPLE COEFFICIENT OF DETERMINATION

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

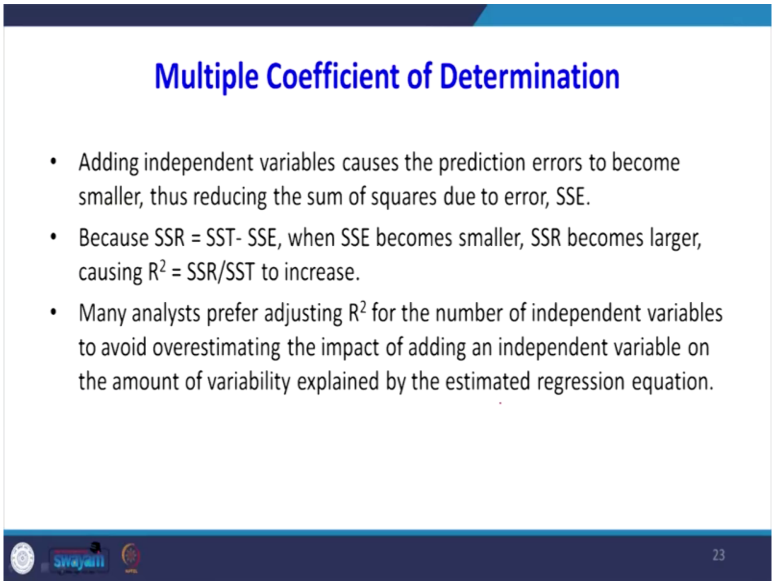
$$R^2 = \frac{21.601}{23.900} = .904$$

22

Now you will see what is multiple coefficient of determination when there is a simple linear regression model we have called it coefficient of determination. Now there is a multiple independent variable we are going to call it is multiple coefficient of determination it is SSR by SST. So, what is R square SSR, SSR is when you add this to 15.87 + 5.7, 21.6. SST is when you are all three 22.2 approximately 23.0.

So there is a 90.4% of the variability of y can be explained with the help of these two independent variable. So, the r square is increased so it is a good model when compared to simple linear regression model.

(Refer Slide Time: 18:07)



Multiple Coefficient of Determination

- Adding independent variables causes the prediction errors to become smaller, thus reducing the sum of squares due to error, SSE.
- Because SSR = SST- SSE, when SSE becomes smaller, SSR becomes larger, causing R^2 = SSR/SST to increase.
- Many analysts prefer adjusting R^2 for the number of independent variables to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation.

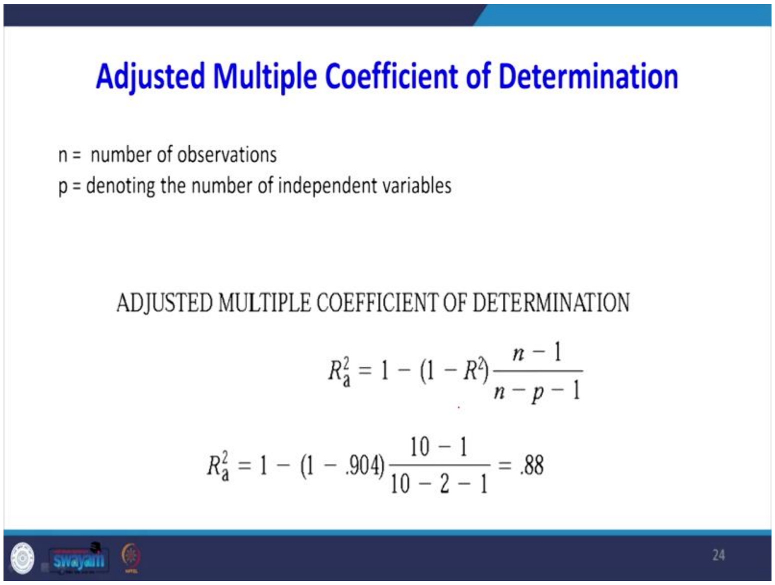
23

So, now we will go for another concept adjusted R square what is the purpose of adjusted R square. So, adding independent variable causes the prediction errors to become smaller, so we know that see SST equal to SSR + SSE so when you add independent variable prediction error become smaller what will happen this error will become smaller so what will happen this when SSE becomes smaller SSR will become bigger one because SSR equal to SST - SSE when SSE becomes smaller SSR become larger.

So, causing R square to increase whenever you add any independent variable SSR will increase SSE will decrease due to that SSR will increase due to that the R square will increase. Many analysts prefer adjusting R square for number of independent variable to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation.

So what is happening instead of using R square we are going for adjusted R square. The advantage of adjusted R Square is whether the added new variable is it is really as an explaining variable or it is a noise variable otherwise the added a new variable how much it is helping to explain the variance of the existing model.

(Refer Slide Time: 19:38)



Adjusted Multiple Coefficient of Determination

n = number of observations

p = denoting the number of independent variables

ADJUSTED MULTIPLE COEFFICIENT OF DETERMINATION

$$R_a^2 = 1 - (1 - R^2)\frac{n - 1}{n - p - 1}$$

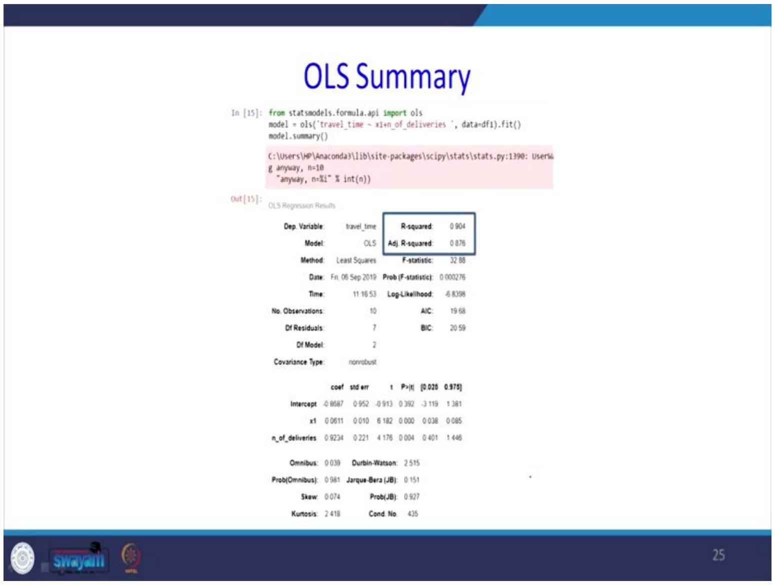
$$R_a^2 = 1 - (1 - .904)\frac{10 - 1}{10 - 2 - 1} = .88$$

So, what is a formula for adjusted R square s previously what was the formula for R square see that R square equal to SSR divided by SST explained variance divided by overall variance. So, this explained variance the regression sum of square can be written this way SST – SSR SST - SSE because what is happening this SSR this SSR represents regression sum of square for all independent variables. So, when you add a new variable you cannot know the contribution of that new variable into the SSR we are going to split this SSR into two term that is SST - SSE so now this will become 1 - SSE divided by SST.

But what we have to do we have to write the degrees of freedom because what is the meaning of adjusted is this adjusting for degrees of freedom. so, when SSE what is the degrees of freedom SSE the degrees of freedom is n - p - 1 what is the n, n is the total number of data set p is number of independent variable - 1 here it will become n - 2 divided by SST you write SST as it is. It is n – 1, so when you simplify this you will get this method.

So here what is the n, n is number of observations what is the p it is number of independent variables and you substitute here R a equal to 1 - 1 - R square n - 1 divided by n - p - 1 when you expand this R square otherwise you write R square equal to SSR by SST you will end up with this relationship this is adjusted R square is 0.88.

(Refer Slide Time: 21:36)



OLS Summary

```
In [15]: from statsmodels.formula.api import ols
         model = ols('travel_time ~ x1+n_of_deliveries ', data=df1).fit()
         model.summary()

         C:\Users\HP\Anaconda3\lib\site-packages\scipy\stats\stats.py:1390: UserW
         g anyway, n=10
           "anyway, n=%i" % int(n))

Out[15]:
         OLS Regression Results

           Dep. Variable:      travel_time          R-squared:      0.904
                   Model:              OLS      Adj. R-squared:     0.876
                  Method:    Least Squares         F-statistic:     32.88
                    Date: Fri, 06 Sep 2019  Prob (F-statistic):  0.000276
                    Time:         11:16:53      Log-Likelihood:   -6.8398
        No. Observations:               10                 AIC:     19.68
            Df Residuals:                7                 BIC:     20.59
                Df Model:                2
         Covariance Type:        nonrobust

                          coef  std err       t   P>|t|  [0.025  0.975]
              Intercept -0.8687   0.952  -0.913  0.392  -3.119   1.381
                     x1  0.0611   0.010   6.182  0.000   0.038   0.085
        n_of_deliveries  0.9234   0.221   4.176  0.004   0.401   1.446

                Omnibus: 0.039    Durbin-Watson: 2.515
         Prob(Omnibus): 0.981  Jarque-Bera (JB): 0.151
                   Skew: 0.074          Prob(JB): 0.927
               Kurtosis: 2.418          Cond. No.  435
```

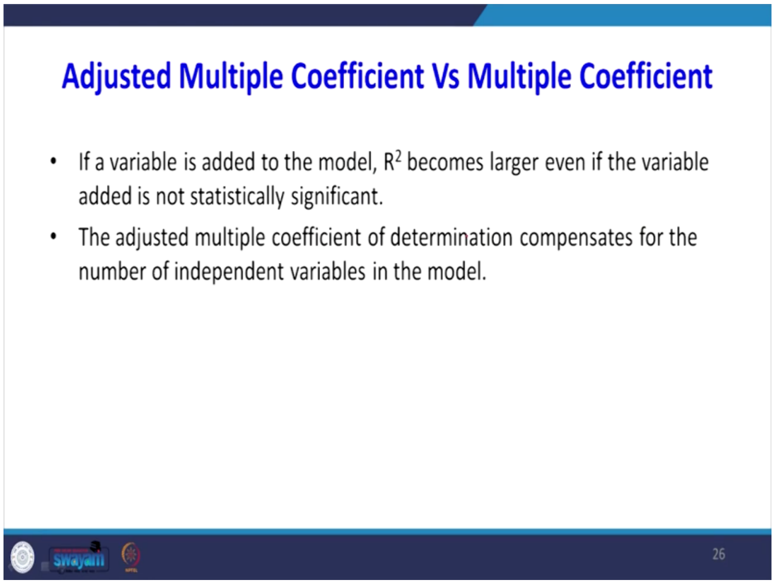
You look at this that is the meaning of 0.8, so another importance of this adjusted r-square is sometime you see what will happen I am writing here R square adjusted R square what will happen whenever you introduce new variable the value of R square will increase adjusted R square also will increase. So, I will explain what is the meaning of R square and I just R square assume that there is a one dependent variable there are many independent variable that independent variable is x 1 x 2 x 3 and x 4.

Now what I am doing here I am going to build a regression model. So, first what I will do first I will take y then I will write regression equation in terms of x1 so what will happen R Square increase and will also adjusted R square. Now taking y is a dependent variable I am going to bring 2 independent variable R square increases adjusted R square also will increase. So, what will happen if the x2 is really helping to explain the variance of the y some time suppose say variable x 3 x 1, x 2 this x 3 variable is the noise variable.

Noise variable means it will not help to explain the variability why it is going to disturb the existing relationship. So, what will happen R square will increase adjusted R square will start decreasing. So, this is the hint for us that the variable which you have added is not helping to explain the model instead of that it is deteriorating the existing model. So, x 3 should not be added that is the meaning of this adjusted R square most of the time.

If the value of R square adjusted R square is similar that means that we have no need to increase any further variable into the model that means you have reached the good model.

(Refer Slide Time: 23:30)



Adjusted Multiple Coefficient Vs Multiple Coefficient

- If a variable is added to the model, R^2 becomes larger even if the variable added is not statistically significant.
- The adjusted multiple coefficient of determination compensates for the number of independent variables in the model.

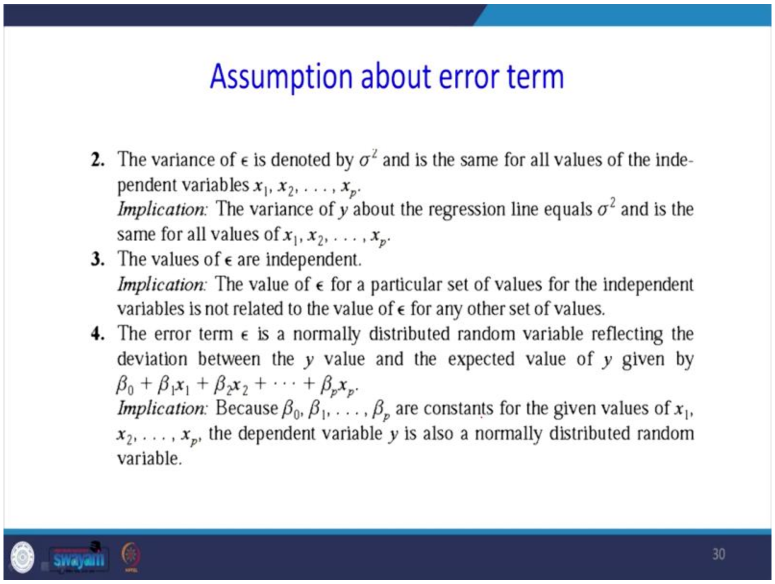
If there is a gap for example R square is 0.9 adjusted R square is 0.3 that there is a possibility of adding more independent variable into the model. Now let us see adjust and multiple coefficient was this multiple coefficient if your variable is added to the model yeah that is the point which I am saying previous slide, if your variable is added to the model R square become larger even if the variable added is not statistically significant it is very important.

The adjusted multiple coefficient of determination that is adjusted R square compensate for the number of independent variable. So, it is adjusted means it is adjusted for the number of independent variable otherwise adjusted for it is a degrees of freedom. If the value of R square is smaller and the model contains a large number of independent variable are just total coefficient of determination can take negative value.

It is a very important point here the interpretation of R square and adjusted R square is not same. The R square is that how much variability of y is explained but the adjusted R square is not the same interpretation. What will happen many time adjusted R square may become negative okay you should be very careful on that. Then we will go for checking model assumptions so as I told you in the beginning of the class y equal to this is the regression model when you there will be error term when you go for regression equation there would not be error term because when you go for expected value of y beta 0 + beta 1 x 1 + beta 2 x 2 and so on.

And there would not be error term because the expected value of error is 0. We will go for some assumption what is the first assumption the error term epsilon is a random variable with mean or expected value of 0 what is implication for the given value of x 1 x 2 and up to x p the expected or average value of y is given by this way you look at this when you go for expected value of y there is no error term. This equation represents the average of all possible values of y that might occur for the given value of x 1 x 2 up to x p by expected value of y.

(Refer Slide Time: 25:58)



Assumption about error term

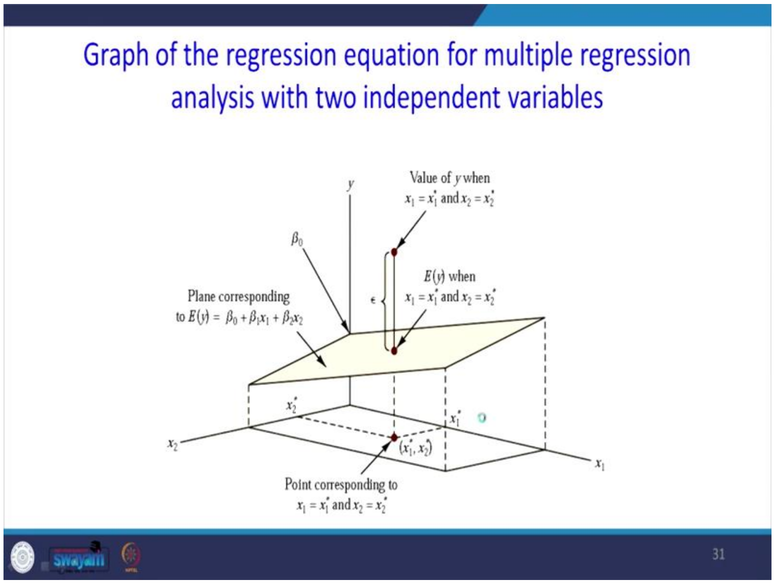
2. The variance of ϵ is denoted by σ^2 and is the same for all values of the independent variables $x_1, x_2, \ldots, x_p$.
   Implication: The variance of y about the regression line equals σ^2 and is the same for all values of $x_1, x_2, \ldots, x_p$.
3. The values of ϵ are independent.
   Implication: The value of ϵ for a particular set of values for the independent variables is not related to the value of ϵ for any other set of values.
4. The error term ϵ is a normally distributed random variable reflecting the deviation between the y value and the expected value of y given by $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$.
   Implication: Because $\beta_0, \beta_1, \ldots, \beta_p$ are constants for the given values of $x_1, x_2, \ldots, x_p$, the dependent variable y is also a normally distributed random variable.

We will go for second assumption the variance of epsilon is denoted by Sigma square and is the same for all values of the independent variable x 1 x 2 x p what is implications the variance of y about the regression line equal to Sigma square and is the same for all values of x 1 x 2 x p. if it is different we will call it is there is effect of heteroscedasticity. Why this point is required if you want to compare the variance of x 1 x 2 up to x p should be same then only there is a meaning for comparison.

The third assumption is the value of epsilon are independent. What is implications the value of epsilon for a particular set of values for independent variable is not related to the value of epsilon for any other set of values. Another way the error terms are independent when you plot that error term there should not be any pattern whether it is increasing or decreasing pattern that is the meaning of this third assumption. Then fourth assumption the error term epsilon is normally distributed random variable reflecting the deviation between y value and the expected value of y given by beta 0 + beta 1 x 1 + beta 2 x 2 up to beta p x p.

What is implications because of beta 0 beta 1 beta b are constant for given values of x 1 x 2 x b the dependent variable y also normally distributed random variable because what will happen the error term it should be independent but it should follow a normal distribution with equal variance if it is not equal variance then it will go to the second assumptions also get violated.

(Refer Slide Time: 27:39)



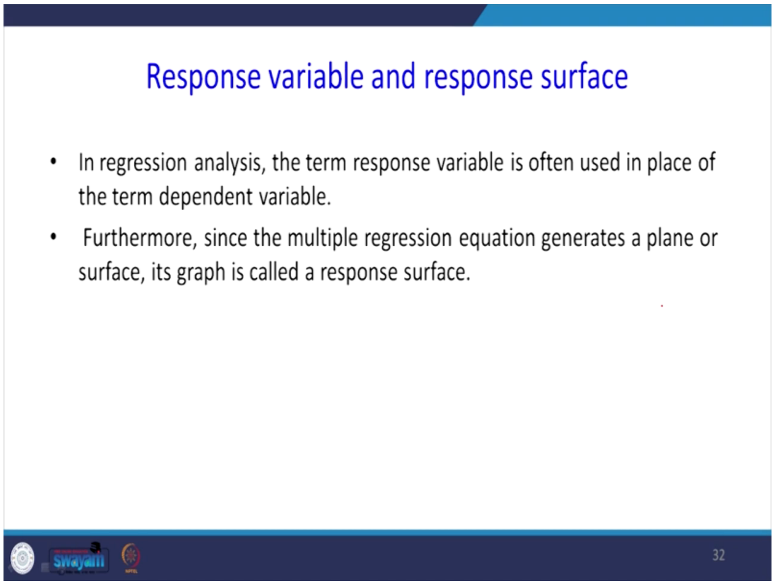
Graph of the regression equation for multiple regression analysis with two independent variables

Value of y when $x_1 = x_1^*$ and $x_2 = x_2^*$
y
β_0
$E(y)$ when $x_1 = x_1^*$ and $x_2 = x_2^*$
Plane corresponding to $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
ϵ
x_2^*
x_1^*
x_2
(x_1^*, x_2^*)
x_1
Point corresponding to $x_1 = x_1^*$ and $x_2 = x_2^*$
31

Now look at this graph of a regression equation for multiple regression analysis with 2, independent variable x 1 is in one independent variable x 2 is another independent variable. See this is the mean value of x 1 this is mean value of x 2 you see this is a plane. So, multiple regression equation is explained with the help of here a surface otherwise this is called a surface the reference model is a plane now the equation is not the line it is the plane.

Otherwise they will call it is RSM also response surface model another name for regression is response surface model because now this is the surface.

(Refer Slide Time: 28:26)



Response variable and response surface

- In regression analysis, the term response variable is often used in place of the term dependent variable.
- Furthermore, since the multiple regression equation generates a plane or surface, its graph is called a response surface.

A response variable and response surface in regression analysis the term response variable is often used in place of the term dependent variable instead of saying dependent variable we will say the response variable. Furthermore since the multiple regression equation generates a plane or surface the graph is called response surface. In this lecture I have explained what is a multiple regression model? Then I have explained what is the connection between simple linear regression model and multiple regression model.

Then I explained the least square model then I have explained what is the meaning of R square and adjusted R square? Then I have explained various model assumptions. The next lecture I am going to test the significance of beta 1 beta 2 and beta 3 with the help of F test and t test and also we will see a demo on Python programming to do a multiple regression, thank you very much.