IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

MULTIPLE REGRESSION MODEL - I

Dr. A. Ramesh

DEPARTMENT OF MANAGEMENT STUDIES

Agenda

- Multiple regression model
- Least squares method
- Multiple coefficient of determination
- Model assumptions
- Testing for significance F-Test, t-Test

Multiple regression model

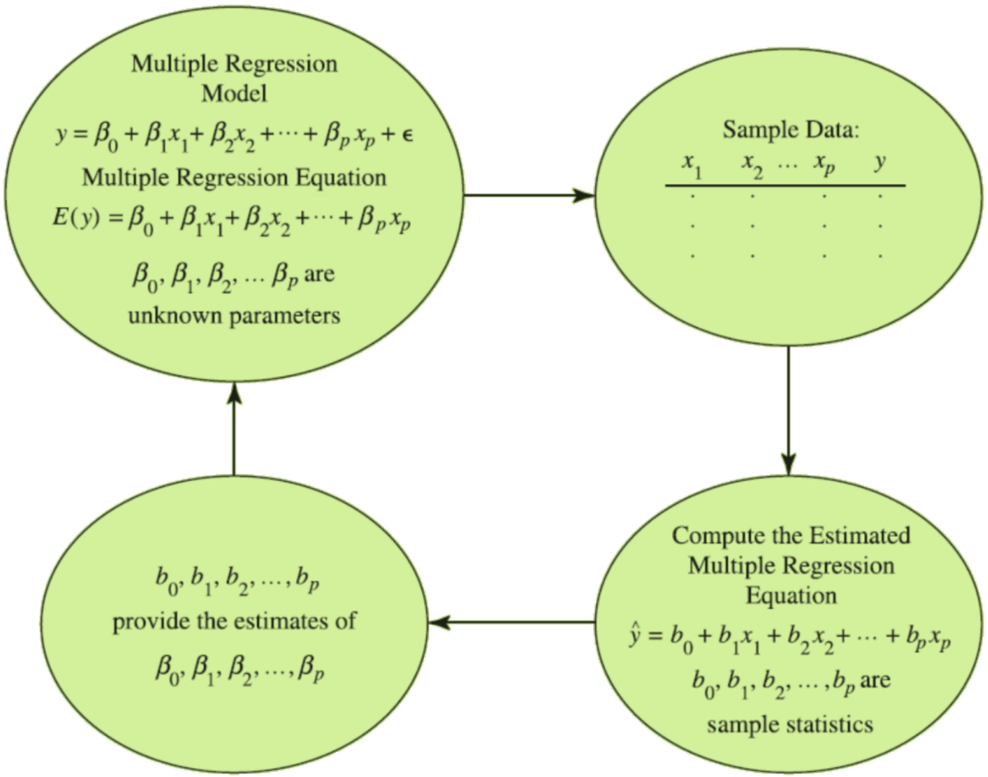
MULTIPLE REGRESSION MODEL

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

MULTIPLE REGRESSION EQUATION

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

The estimation process For multiple regression



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 $b_0, b_1, b_2, \ldots, b_p$ denoting the sample statistics used to estimate the parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$.

$b_0 \rightarrow \beta_0$

$b_1 \rightarrow \beta_1$

$b_2 \rightarrow \beta_2$

Least Squares Method

LEAST SQUARES CRITERION

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

b_0 b_1 $\hat{y}_i = b_0 + b_1 x_1$

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

Least Squares Method

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

b_1

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

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

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
```

Using python import data

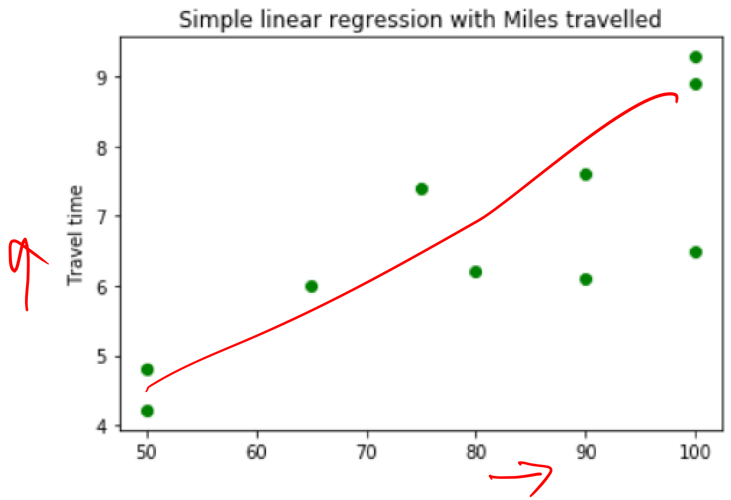
Out[2]:

	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

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 ')
```

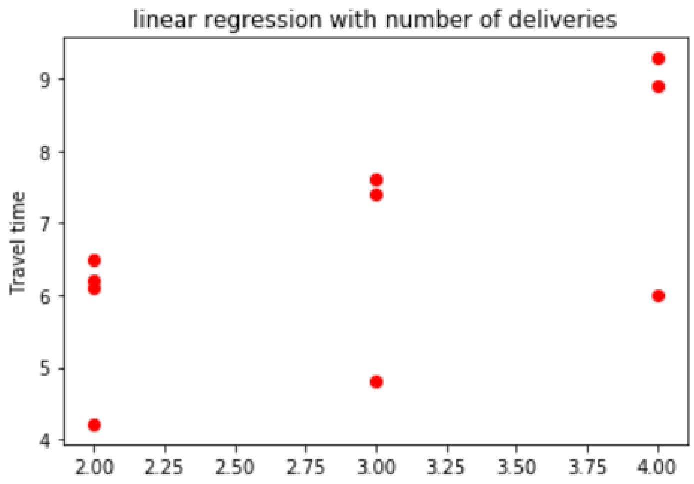
```
Out[3]: Text(0.5,1,' Simple linear regression with Miles travelled ')
```



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')
```

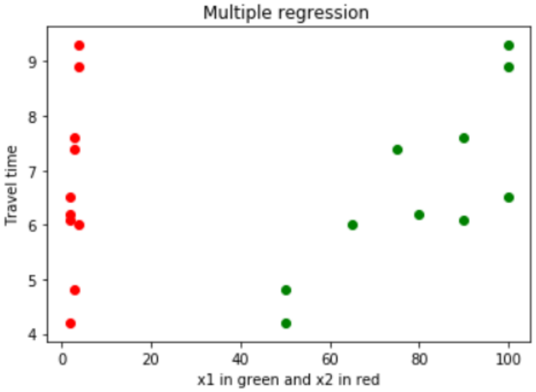
Out[11]: `Text(0.5,1,'linear regression with number of deliveries')`



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')



Linear regression Vs. multiple regression model

- Linear regression

$$\hat{y} = 1.27 + .0678x_1$$

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.
==============================================================================
```

$\hat{y} = 1.2734 + 0.067824$

Linear regression Vs. Multiple regression model

- Multiple regression

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

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: UserWa
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

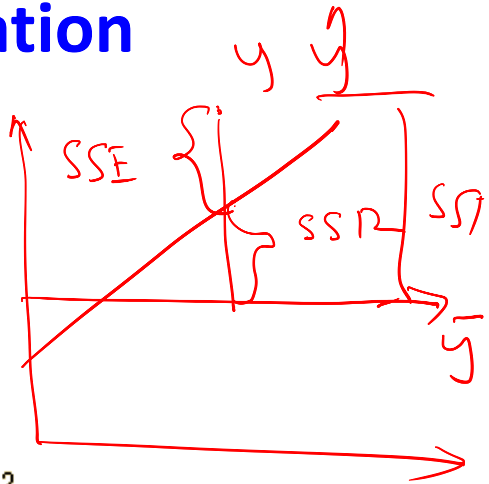
Omnibus:	0.039	**Durbin-Watson:**	2.515
Prob(Omnibus):	0.981	**Jarque-Bera (JB):**	0.151
Skew:	0.074	**Prob(JB):**	0.927
Kurtosis:	2.418	**Cond. No.**	435.

$H_0: \beta_1 = \beta_2 = 0$

$$\boxed{y = -0.8687 + 0.0611 x_1 + 0.9234 x_2}$$

x_2 = no. of deliveries

Multiple Coefficient of Determination



RELATIONSHIP AMONG SST, SSR, AND SSE

$$SST = SSR + SSE$$

where

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

Multiple Coefficient of Determination for linear model

In [9]:
```
print(anova_lm(Fit1))
```

	df	sum_sq	mean_sq	F	PR(>F)
x1	1.0	15.871304	15.871304	15.814578	0.00408
Residual	8.0	8.028696	1.003587	NaN	NaN

$SST = 15.87 + 8.02 = 23.89$

$SSE = 8.02$

$SSR = 15.871$

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

SST = 22
SSE = 2.29

SSR = 20.1

Multiple Coefficient of Determination

MULTIPLE COEFFICIENT OF DETERMINATION

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

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

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.

$SST = SSR + SS\underline{E}$

Adjusted Multiple Coefficient of Determination

n = number of observations
p = denoting the number of independent variables

$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST}$

$= 1 - \frac{SSE}{SST}$

$= 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$

$=$

ADJUSTED MULTIPLE COEFFICIENT OF DETERMINATION

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

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

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: UserWa
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.

R^2 adj R^2

$y \rightarrow (x_1, x_2, x_3, x_4)$ $\uparrow$ $\uparrow$

$y \rightarrow x_1$ $\uparrow$ $\uparrow$

$y \rightarrow x_1, x_2$ $\uparrow$ $\uparrow$

$y \rightarrow x_1, x_2, x_3$ $\uparrow$ $\downarrow$

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.

Adjusted Multiple Coefficient Vs Multiple Coefficient

- If the value of R^2 is small and the model contains a large number of independent variables, the adjusted coefficient of determination can take a negative value

Model Assumptions

MULTIPLE REGRESSION MODEL

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

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

Assumption about error term

1. The error term ε is a random variable with mean or expected value of zero;

   $E(\varepsilon) = 0.$

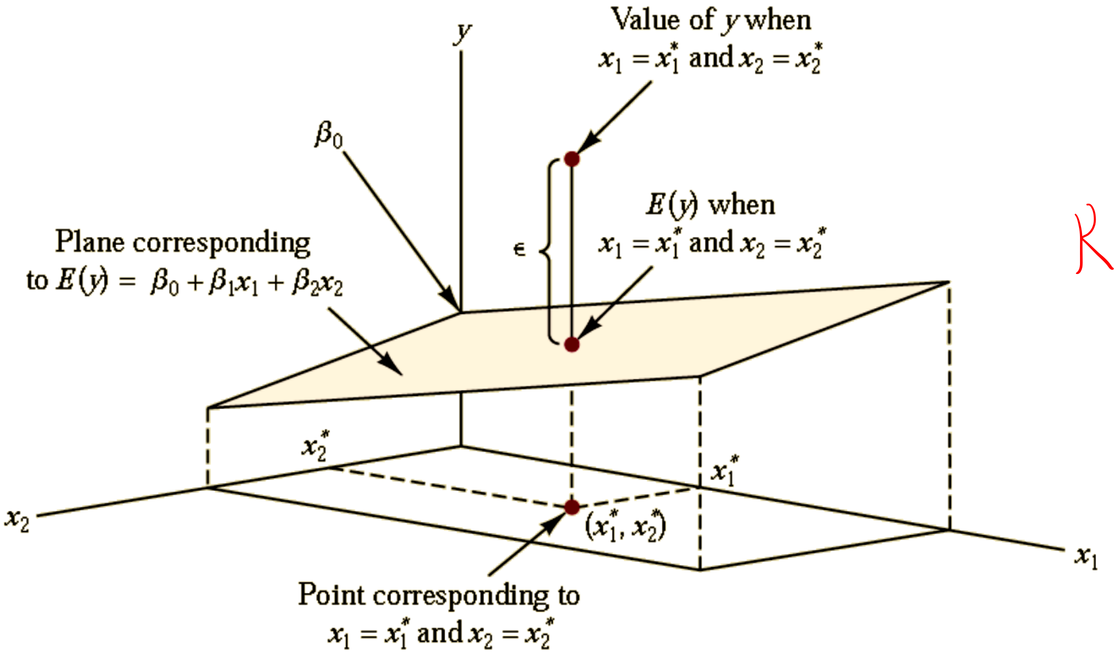
   Implication: For given values of $x_1, x_2, \ldots, x_p$ the expected , or average , value of y is given by $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots. + \beta_p x_p$

   – This equation represents the average of all possible values of y , that might occur for the given value of $x_1, x_2, \ldots, x_p$, by E(y).

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.

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



KSM

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.

Thank You