LINEAR REGRESSION ANALYSIS

MODULE – XVI

Lecture - 44

Exercises

Dr. Shalabh
Department of Mathematics and Statistics
Indian Institute of Technology Kanpur

Exercise 1

The following data has been obtained on 26 patients on their systolic blood pressure and age. Now we illustrate the use of regression tools and the interpretation of the output from a statistical software through this example.

Patient no. (i)	Age (X_i)	Blood pressure (y_i)	Patient no. (i)	Age (X_i)	Blood pressure (y_i)
1	70	178	14	43	137
2	26	130	15	50	132
3	28	132	16	40	126
4	65	142	17	33	112
5	53	157	18	50	138
6	45	162	19	54	148
7	20	125	20	62	162
8	38	118	21	55	153
9	49	140	22	65	155
10	35	133	23	40	127
11	18	122	24	68	165
12	19	117	25	44	140
13	16	116	26	62	160

The first step involves the check whether a simple linear regression model can be fitted to this data or not. For this, we plot a scatter diagram as follows:

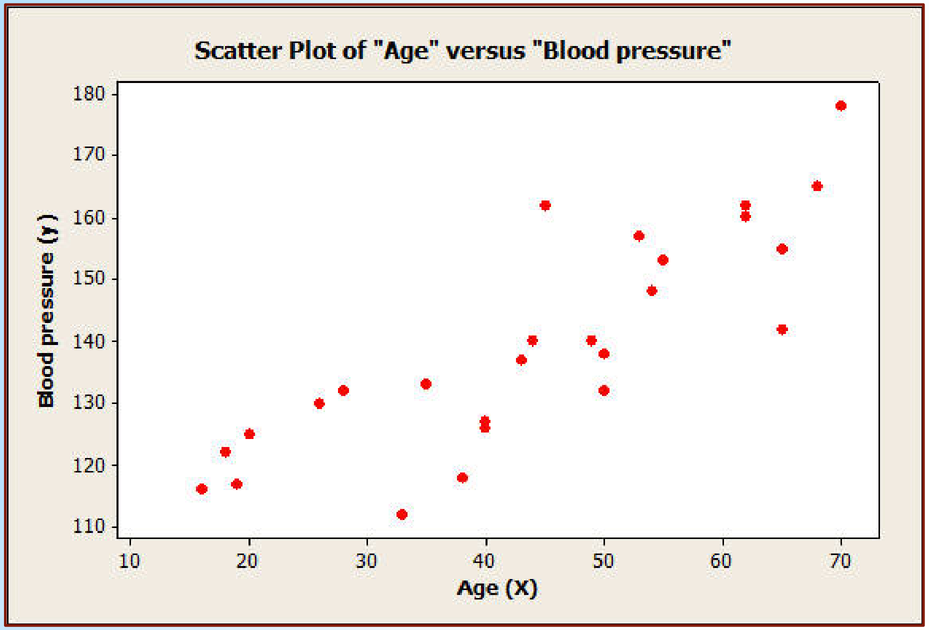


Figure 1

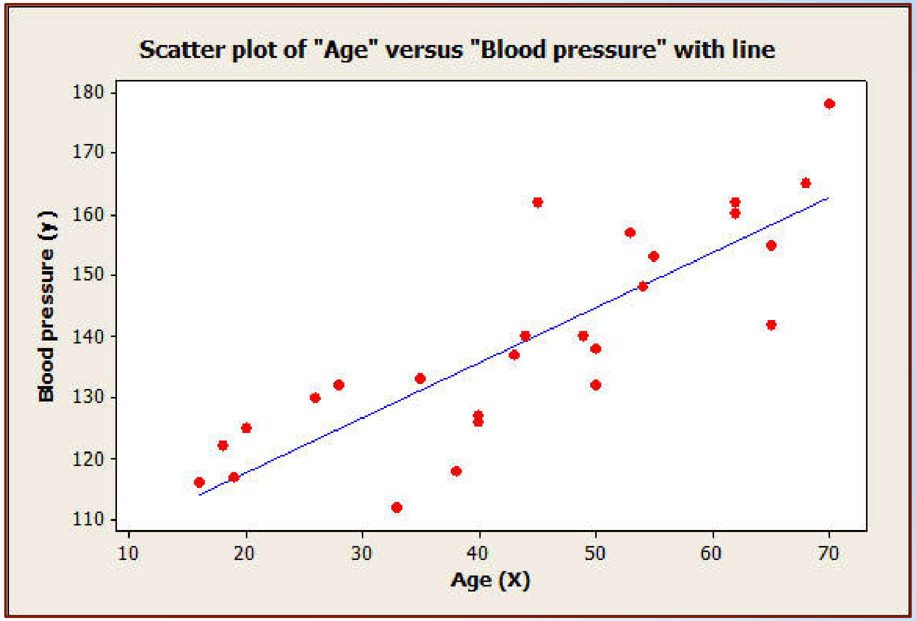


Figure 2

Looking at the scatter plots in Figures 1 and 2, an increasing linear trend in the relationship between blood pressure and age is clear. It can also be concluded from the figure 2 that it is appropriate to fit a linear regression model.

A typical output of a statistical software on regression of blood pressure (y) versus age (x) will look like as follows:

```
The regression equation is
Blood pressure (y) = 99.7 + 0.902 Age (X)

Predictor     Coef   SE Coef      T      P
Constant     99.7   5.636     17.69  0.000
Age (X)     0.902   0.1200     7.52  0.000


R-Sq = 70.2%   R-Sq(adj) = 69.0%

PRESS = 2673.61   R-Sq(pred) = 65.54%

Analysis of Variance

Source          DF    SS      MS      F      P
Regression       1  5446.8  5446.8  56.55  0.000
Residual Error  24  2311.7    96.3
  Lack of Fit   20  2206.7   110.3   4.20  0.086
  Pure Error     4   105.0    26.3
Total           25  7758.5


Unusual Observations
             Blood pressure
Obs  Age (X)      (y)     Fit  SE Fit   Residual  St Residual
  6     45.0    162.00  140.26    1.93     21.74      2.26R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.34
```

Now we discuss the interpretation of the results part-wise.

First look at the following section of the output.

```
The regression equation is
Blood pressure (y) = 99.7 + 0.902 Age (X)

Predictor     Coef  SE Coef       T      P
Constant     99.7   5.636     17.69  0.000
Age (X)     0.902   0.1200     7.52  0.000

R-Sq = 70.2%   R-Sq(adj) = 69.0%

PRESS = 2673.61   R-Sq(pred) = 65.54%
```

The fitted linear regression model is

`Blood pressure (y) = 99.7 + 0.902 Age (X)`

in which the regression coefficients are obtained as

$$b_0 = \bar{y} - b_1\bar{x} = 99.7$$

$$b_1 = \frac{s_{xy}}{s_{xx}} = 0.902$$

where

$$s_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \quad s_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2, \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n}x_i, \; \bar{y} = \frac{1}{n}\sum_{i=1}^{n}y_i.$$

The same is indicated by `Constant   99.7` and `Age (X)   0.902`.

The terms `SE Coef` denotes the standard errors of the estimates of intercept term and slope parameter. They are obtained as

$$\sqrt{\widehat{Var}(b_0)} = \sqrt{s^2\left(\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}}\right)} = 5.636$$

$$\sqrt{\widehat{Var}(b_1)} = \sqrt{\frac{s^2}{s_{xx}}} = 0.1200$$

where

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2}{n-2}, \quad n = 26.$$

The terms `T` the value of *t*-statistics for testing the significance of regression coefficient and are obtained as

$$H_0 : \beta_0 = \beta_{00} \ \text{ with } \ \beta_{00} = 0$$

$$t_0 = \frac{b_0 - \beta_{00}}{\sqrt{\frac{SS_{res}}{n-2}\left(\frac{1}{n} + \frac{\overline{x}^2}{s_{xx}}\right)}} = 17.69$$

$$H_0 : \beta_1 = \beta_{10} \ \text{ with } \ \beta_{10} = 0$$

$$t_0 = \frac{b_1 - \beta_{10}}{\sqrt{\frac{SS_{res}}{(n-2)s_{xx}}}} = 7.52$$

The corresponding `P` values are all 0.000 which are less than the level of significance $\alpha = 0.5.$ This indicates that the null hypothesis is rejected and the intercept term and slope parameter are significant.

Now we discuss the interpretation of the goodness of fit statistics.

The value of coefficient of determination is given by `R-Sq = 70.2%.`

This value is obtained from

$$R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n}(y_i - \overline{y})^2} = 0.702.$$

The value of adjusted coefficient of determination is given by `R-Sq(adj) = 69.0%.`

This value is obtained from

$$\overline{R}^2 = 1 - \left(\frac{n-1}{n-k}\right)(1 - R^2) = 0.69 \text{ with } k = 1.$$

This means that the fitted model can expalain about 70% of the variability in *y* through *X*.

This value of PRESS statistics is given by `PRESS = 2673.61`

This value is obtained from

$$PRESS = \sum_{i=1}^{n}\left[y_i - \hat{y}_{(i)}\right]^2 = \sum_{i=1}^{n}\left[\frac{e_i}{1-h_{ii}}\right]^2 = 2673.61.$$

This is also a measure of model quality. It gives an idea how well a regression model will perform in predicting new data. It is apparent that the value is quite high. A model with small value of PRESS is desirable.

The value of R^2 for prediction based on PRESS statistics is given by `R-Sq(pred) = 65.54%`

This value is obtained from

$$R^2_{\text{prediction}} = 1 - \frac{PRESS}{SS_T} = 0.6554.$$

This statistic gives an indication of the predictive capability of the fitted regression model.

Here $R^2_{\text{prediction}} = 0.6554$ indicates that the model is expected to explain about 66% of the variability in predicting new observations.

Now we look at the following output on analysis of variance.

```
Analysis of Variance

Source          DF      SS      MS      F      P
Regression       1  5446.8  5446.8  56.55  0.000
Residual Error  24  2311.7    96.3
  Lack of Fit   20  2206.7   110.3   4.20  0.086
  Pure Error     4   105.0    26.3
Total           25  7758.5
```

However, in this example with one explanatory variable, the analysis of variance does not make much sense and is equivalent to test the hypothesis

$H_0 : \beta_1 = \beta_{10}$ with $\beta_{10} = 0$ by

$$t_0 = \frac{b_1 - \beta_{10}}{\sqrt{\dfrac{SS_{res}}{(n-2)s_{xx}}}}.$$

We will discuss the details on analysis of variance in the next example.

Now we look into the following part of the output. The following outcome presents the analysis of residuals to find an unusual observation which can possibly be an outlier or an extreme observation. Its interpretation in the present case is that the 6th observation needs attention.

```
Unusual Observations
             Blood pressure
Obs  Age (X)       (y)     Fit  SE Fit   Residual   St Residal
  6     45.0    162.00  140.26    1.93      21.74      2.26R

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 1.34
```

The Durbin Watson statistics is obtained from `Durbin-Watson statistic = 1.34`

Its value is obtained from

$$d = \frac{\sum_{t=2}^{n} (e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = 1.34.$$

At d = 2, it is indicated that there is no first order autocorrelation in the data. As the value d = 1.34 which is less than 2, so it indicates the possible presence of first order positive autocorrelation in the data.

Next we find the residuals.

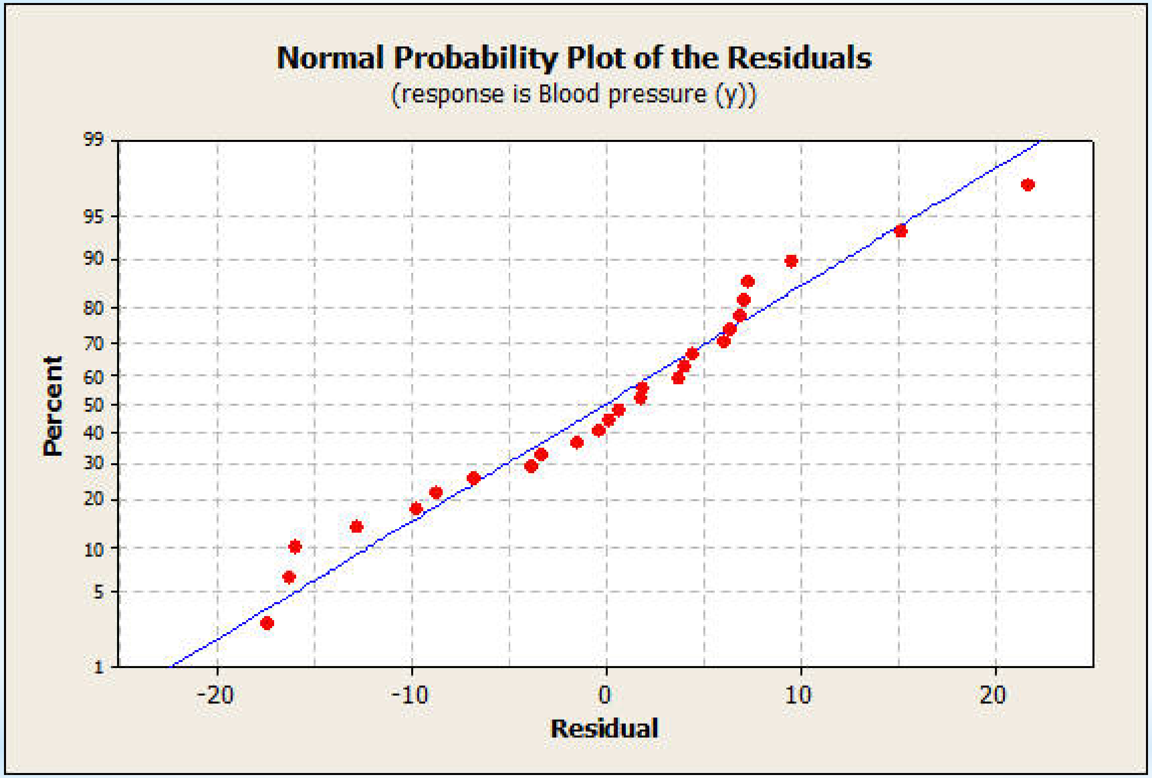
The residual are obtained as $e_i = y_i - \hat{y}_i$ where the observed values are denoted as y_i and the fitted values are obtained as

$$\text{Blood pressure } (\hat{y}_i) = 99.7 + 0.902 \text{ Age } (X_i)$$

Patient no. (i)	Observed values (y_i)	Fitted values $(\hat{y}_i)$	Residuals (e_i)	Patient no. (i)	Observed values (y_i)	Fitted values $(\hat{y}_i)$	Residuals (e_i)
1	178	162.84	-15.16	14	137	138.486	1.486
2	130	123.152	-6.848	15	132	144.8	12.8
3	132	124.956	-7.044	16	126	135.78	9.78
4	142	158.33	16.33	17	112	129.466	17.466
5	157	147.506	-9.494	18	138	144.8	6.8
6	162	140.29	-21.71	19	148	148.408	0.408
7	125	117.74	-7.26	20	162	155.624	-6.376
8	118	133.976	15.976	21	153	149.31	-3.69
9	140	143.898	3.898	22	155	158.33	3.33
10	133	131.27	-1.73	23	127	135.78	8.78
11	122	115.936	-6.064	24	165	161.036	-3.964
12	117	116.838	-0.162	25	140	139.388	-0.612
13	116	114.132	-1.868	26	160	155.624	-4.376

Next we consider the graphical analysis.

The normal probability plot is obtained like as follows:



The normal probability plots is a plot of the ordered standardized residuals versus the cumulative probability

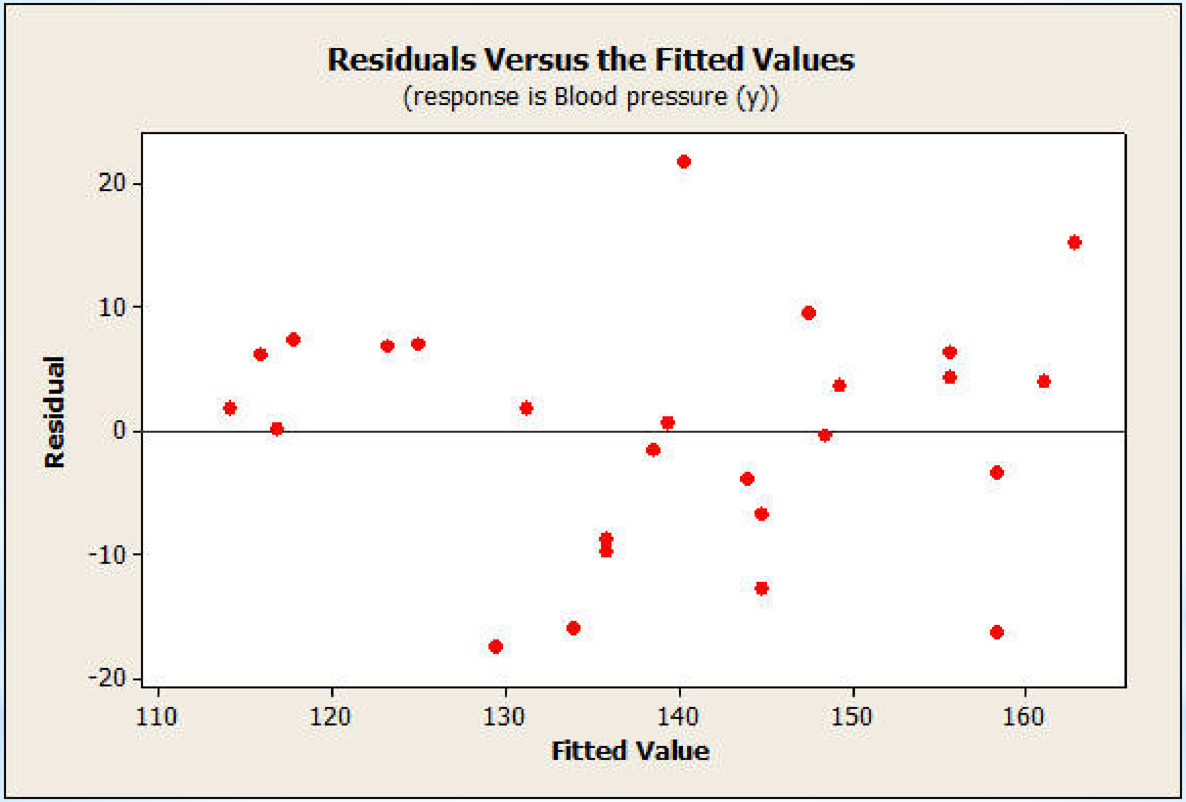
$$P_i = \frac{\left(i - \frac{1}{2}\right)}{n}, \ i = 1, 2, \ldots, n.$$

It can be seen from the plot that most of the points are lying close to the line. This clearly indicates that the assumption of normal distribution for random errors is satisfactory in the given data.

Next we consider the graph between the residuals and the fitted values.

Such a plot is helpful in detecting several common type of model inadequacies.

If plot is such that the residuals can be contained in a horizontal band (and residual fluctuates is more or less in a random fashion inside the band) then there are no obvious model defects.



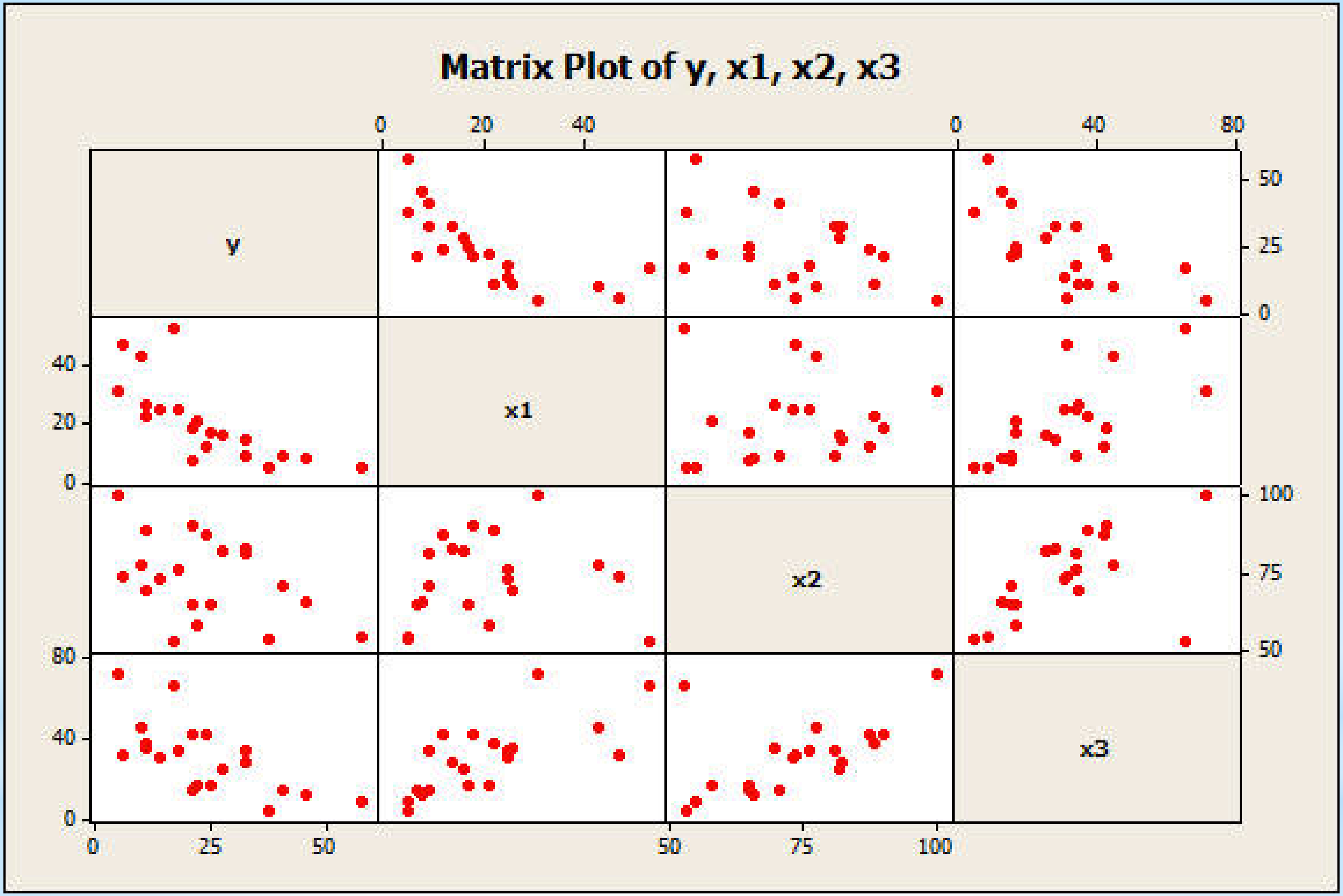
The points in this plot can more or less be contained in a horizontal band. So the model has no obvious defects.

Exercise 2

The following data has 20 observations on study variable with three independent variables- X_{1i}, X_{2i} and X_{3i}.
Now we illustrate the use of regression tools and the interpretation of the output from a statistical software through this example.

i	y_i	X_{1i}	X_{2i}	X_{3i}	i	y_i	X_{1i}	X_{2i}	X_{3i}
1	38	5	53.21	4.59	11	25	17	64.89	16.84
2	14	25	73.21	31.21	12	22	21	57.96	16.54
3	11	26	69.48	35.24	13	6	47	73.54	31.89
4	10	43	77.78	45.24	14	46	8	65.72	12.65
5	33	9	81.21	34.51	15	41	9	70.45	15.21
6	24	12	87.59	42.56	16	18	25	76.15	34.16
7	21	18	90.17	42.85	17	21	7	64.98	15.28
8	5	31	100.21	71.64	18	33	14	82.57	28.53
9	17	53	52.78	65.85	19	28	16	81.96	25.46
10	58	5	54.64	8.64	20	11	22	88.61	37.85

The first step involves the check whether a simple linear regression model can be fitted to this data or not. For this, we plot a matrix scatter diagram as follows:



Looking at the scatter plots in various blocks, a good degree of linear trend in the relationship between study variable and each of the explanatory variable is clear. The degree of linear relationships in various blocks are different. It can also be concluded that it is appropriate to fit a linear regression model.

A typical output of a statistical software on regression of y on X_1, X_2, X_3 will look like as follows:

```
The regression equation is
y = 72.6 - 0.797 x1 - 0.479 x2 + 0.099 x3


Predictor      Coef  SE Coef      T      P   VIF
Constant      72.64    13.52   5.37  0.000
x1          -0.7970   0.2271  -3.51  0.003   2.7
x2          -0.4789   0.2065  -2.32  0.034   2.0
x3           0.0988   0.2153   0.46  0.653   3.9


 R-Sq = 69.8%   R-Sq(adj) = 64.2%

PRESS = 2289.90   R-Sq(pred) = 38.61%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       3  2603.81  867.94  12.33  0.000
Residual Error  16  1125.99   70.37
Total           19  3729.80
```

Continued...

Obs	x1	y	Fit	SE Fit	Residual	St Resid
1	5.0	38.00	43.63	4.00	-5.63	-0.76
2	25.0	14.00	20.74	2.09	-6.74	-0.83
3	26.0	11.00	22.13	2.18	-11.13	-1.37
4	43.0	10.00	5.59	3.96	4.41	0.60
5	9.0	33.00	29.99	3.27	3.01	0.39
6	12.0	24.00	25.34	3.69	-1.34	-0.18
7	18.0	21.00	19.35	3.19	1.65	0.21
8	31.0	5.00	7.02	5.39	-2.02	-0.31
9	53.0	17.00	11.63	7.54	5.37	1.46 X
10	5.0	58.00	43.34	3.93	14.66	1.98
11	17.0	25.00	29.68	2.54	-4.68	-0.59
12	21.0	22.00	29.78	3.06	-7.78	-1.00
13	47.0	6.00	3.12	6.12	2.88	0.50
14	8.0	46.00	36.04	2.80	9.96	1.26
15	9.0	41.00	33.23	2.63	7.77	0.98
16	25.0	18.00	19.62	2.06	-1.62	-0.20
17	7.0	21.00	37.45	2.89	-16.45	-2.09R
18	14.0	33.00	24.76	2.60	8.24	1.03
19	16.0	28.00	23.15	2.85	4.85	0.61
20	22.0	11.00	16.41	3.06	-5.41	-0.69

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 1.80535

No evidence of lack of fit (P >= 0.1).

Now we discuss the interpretation of the results part-wise.

First look at the following section of the output.

```
The regression equation is
y = 72.6 - 0.797 x1 - 0.479 x2 + 0.099 x3

Predictor      Coef  SE Coef      T      P   VIF
Constant      72.64    13.52   5.37  0.000
x1          -0.7970   0.2271  -3.51  0.003  2.7
x2          -0.4789   0.2065  -2.32  0.034  2.0
x3           0.0988   0.2153   0.46  0.653  3.9

 R-Sq = 69.8%   R-Sq(adj) = 64.2%

PRESS = 2289.90   R-Sq(pred) = 38.61%
```

The fitted linear regression model is

`y = 72.6 - 0.797 x1 - 0.479 x2 + 0.099 x3`

in which the regression coefficients are obtained as

$$b = (X'X)^{-1} X'y$$

which is 4 X 1 vector

$$b = (b_1, b_2, b_3, b_4)' = (72.6,\ -0.797,\ -0.4789,\ 0.0988)'$$

where b_1 is the OLS estimator of intercept term and b_2, b_3, b_4 are the OLS estimators of regression coefficients.

The same is indicated by ‘`Constant  99.7`’, ‘`x1  -0.7970`’, ‘`x2  -0.4789`’, and ‘`x3  0.0988`’.

The terms `SE Coef` denotes the standard errors of the estimates of intercept term and slope parameter. They are obtained as positive square roots of the diagonal elements of the covariance matrix of

$$\hat{V}(b) = \hat{\sigma}^2 (X'X)^{-1}$$

where

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - b_1 - b_2 x_{2i} - b_3 x_{3i} - b_4 x_{4i})^2}{n-4}, \ n = 20.$$

The terms `T` the value of *t*-statistics for testing the significance of regression coefficient and are obtained for testing

$H_{0j} : \beta_j = \beta_{j0}$ with $\beta_{j0} = 0$ against $H_{1j} : \beta_j \neq \beta_{j0}$ using the statistic

$$t_j = \frac{\beta_j - \beta_{j0}}{\sqrt{Var(b_j)}}, \ j = 1, 2, 3, 4$$

$t_1 = 5.37$ with corresponding p-value 0.000

$t_2 = -3.51$ with corresponding p-value 0.003,

$t_3 = -2.32$ with corresponding p-value 0.034,

$t_4 = 0.46$ with corresponding p-value 0.653.

The null hypothesis is rejected at $\alpha = 0.5$ level of significance when the corresponding `P` value is less than the level of significance $\alpha = 0.5$ and the intercept term as well as slope parameter are significant.

Thus $H_{01} : \beta_1 = 0, H_{02} : \beta_2 = 0, H_{03} : \beta_3 = 0$ are rejected and $H_{04} : \beta_4 = 0$ is accepted. Thus the variables X_1, X_3, X_3 enter in the model and X_4 leaves the model.

Next we consider the values of variance inflation factor (VIF) given in the output as `VIF`. The variance inflation factor for the j^{th} explanatory variable is defined as

$$VIF_j = \frac{1}{1-R_j^2}.$$

This is the factor which is responsible for inflating the sampling variance. The combined effect of dependencies among the explanatory variables on the variance of a term is measured by the *VIF* of that term in the model. One or more large *VIFs* indicate the presence of multicollinearity in the data.

In practice, usually a *VIF* > 5 or 10 indicates that the associated regression coefficients are poorly estimated because of multicollinearity.

In our case, the values of *VIF*s due to first, second and third explanatory variables are 2.7, 2.0 and 3.9, respectively. Thus there is no indication of the presence of multicollinearity in the data.

Now we discuss the interpretation of the goodness of fit statistics.

The value of coefficient of determination is given by `R-Sq = 69.8%`

This value is obtained from

$$R^2 = 1 - \frac{e'e}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 0.698.$$

This means that the fitted model can expalain about 70% of the variability in *y* through *X*.

The value of adjusted coefficient of determination is given by `R-Sq(adj) = 64.2%`

This value is obtained from

$$\bar{R}^2 = 1 - \left(\frac{n-1}{n-k}\right)(1 - R^2) = 0.642 \text{ with } n = 20,\ k = 4.$$

This means that the fitted model can expalain about 64% of the variability in *y* through *X's*.

This value of PRESS statistics is given by `PRESS = 2289.90`

This value is obtained from

$$PRESS = \sum_{i=1}^{n} \left[y_i - \hat{y}_{(i)} \right]^2 = \sum_{i=1}^{n} \left[\frac{e_i}{1-h_{ii}} \right]^2 = 2289.90.$$

This is also a measure of model quality. It gives an idea how well a regression model will perform in predicting new data. It is apparent that the value is quite high. A model with small value of PRESS is desirable.

The value of R^2 for prediction based on PRESS statistics is given by `R-Sq(pred) = 38.61%`

This value is obtained from

$$R^2_{\text{prediction}} = 1 - \frac{PRESS}{SS_T} = 0.3861.$$

This statistic gives an indication of the predictive capability of the fitted regression model. Here $R^2_{\text{prediction}} = 0.3861$ indicates that the model is expected to explain about 39% of the variability in predicting new observations. This is also reflected in the value of PRESS statistic.

Now we look at the following output on analysis of variance. This output tests the null hypothesis $H_{01}: \beta_2 = \beta_3 = \beta_4 = 0$ against the alternative hypothesis that at least one of the regression coefficient is different from other. There are three explanatory variables.

```
Analysis of Variance

Source          DF      SS       MS       F       P
Regression       3   2603.81   867.94   12.33   0.000
Residual Error  16   1125.99    70.37
Total           19   3729.80
```

The sum of squares due to regression is obtained by

$$SS_{reg} = b'X'y - n\bar{y}^2 = y'Hy - n\bar{y}^2 = 2603.81.$$

The sum of squares due to total is obtained by

$$SS_T = y'y - n\bar{y}^2 = 3729.80.$$

The sum of squares due to error is obtained by

$$SS_{res} = SS_T - SS_{reg} = 1125.99.$$

The mean squares are obtained by dividing the sum of squares by the degrees of freedom.

The mean sqaure due to regression is obtained as $MS_{reg} = \dfrac{SS_{reg}}{3} = 867.94.$

The mean sqaure due to error is obtained as $MS_{res} = \dfrac{SS_{res}}{16} = 70.37.$

The F-statistic is obtained by $F = \dfrac{MS_{reg}}{MS_{res}} = 12.33.$

The null hypothesis is rejected at 5% level of significance because P value is less than then $\alpha = 0.5.$

Now we look into the following part of the output. The following outcome presents the analysis of residuals to find an unusual observation which can possibly be an outlier or an extreme observation.

```
Obs    x1      y    Fit    SE Fit  Residual  St Resid
  1   5.0  38.00  43.63    4.00     -5.63     -0.76
  2  25.0  14.00  20.74    2.09     -6.74     -0.83
  3  26.0  11.00  22.13    2.18    -11.13     -1.37
  4  43.0  10.00   5.59    3.96      4.41      0.60
  5   9.0  33.00  29.99    3.27      3.01      0.39
  6  12.0  24.00  25.34    3.69     -1.34     -0.18
  7  18.0  21.00  19.35    3.19      1.65      0.21
  8  31.0   5.00   7.02    5.39     -2.02     -0.31
  9  53.0  17.00  11.63    7.54      5.37      1.46 X
 10   5.0  58.00  43.34    3.93     14.66      1.98
 11  17.0  25.00  29.68    2.54     -4.68     -0.59
 12  21.0  22.00  29.78    3.06     -7.78     -1.00
 13  47.0   6.00   3.12    6.12      2.88      0.50
 14   8.0  46.00  36.04    2.80      9.96      1.26
 15   9.0  41.00  33.23    2.63      7.77      0.98
 16  25.0  18.00  19.62    2.06     -1.62     -0.20
 17   7.0  21.00  37.45    2.89    -16.45     -2.09R
 18  14.0  33.00  24.76    2.60      8.24      1.03
 19  16.0  28.00  23.15    2.85      4.85      0.61
 20  22.0  11.00  16.41    3.06     -5.41     -0.69

R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.


Durbin-Watson statistic = 1.80535

No evidence of lack of fit (P >= 0.1).
```

Now we look into the following part of the output. The following outcome presents the analysis of residuals to find an unusual observation which can possibly be an outlier or an extreme observation. Here `'y'` denotes the observed values, `'Fit'` denotes the fitted values $\hat{y}_i$, `'SE Fit'` denotes the standard error of fit, i.e., $\hat{y}_i$, `'Residual'` denotes the ordinary residuals obtained by

$$e_i = y_i - \hat{y}_i = y_i - (72.6 - 0.797X_1 - 0.479X_2 + 0.099X_3)$$

and `'St Resid'` denotes the standardized residuals.

The 9th observation has a large standardized residual.

The 17th observation whose X value gives it large influence.

There is no evidence of lack of fit.

The Durbin Watson statistics is obtained from `Durbin-Watson statistic = 1.80535`

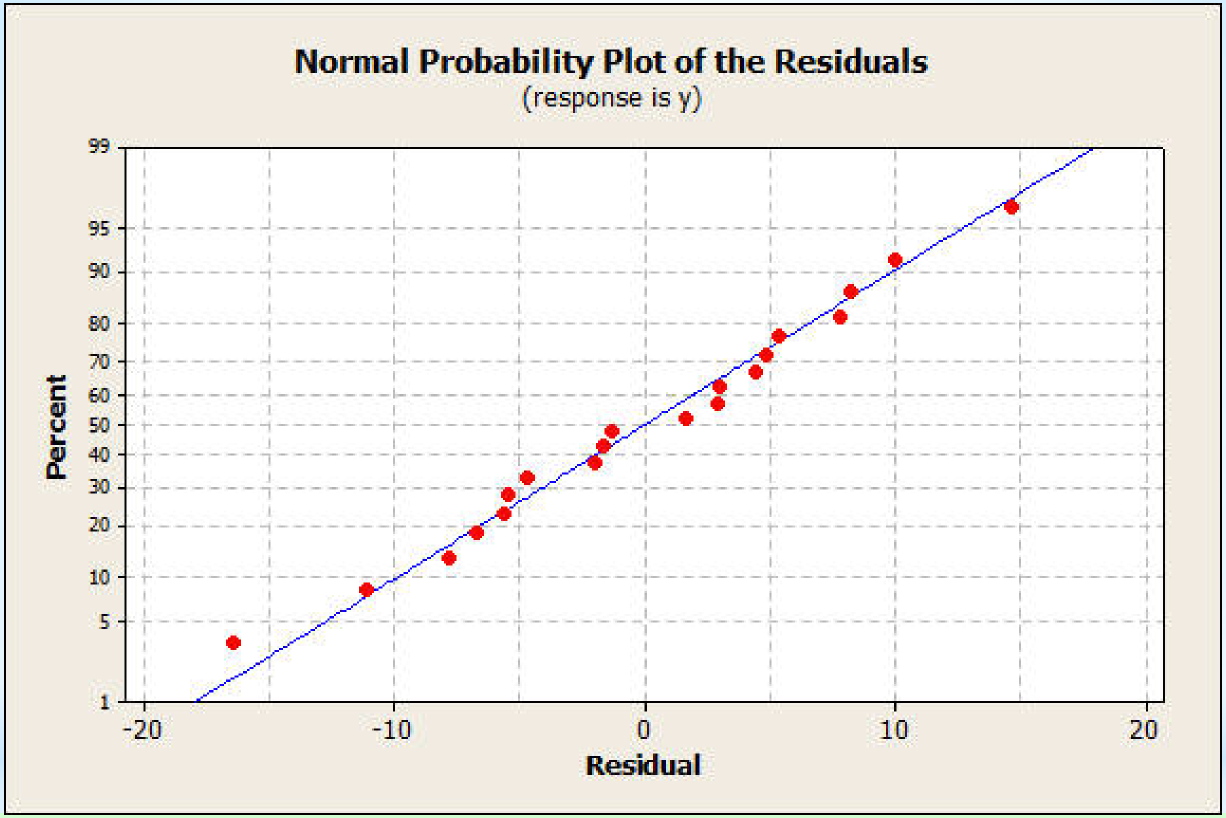
Its value is obtained from

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} = 1.80535.$$

At $d = 2$, it is indicated that there is no first order autocorrelation in the data. As the value $d = 1.80535$ is less than 2, so it indicates the possibility of presence of first order positive autocorrelation in the data.

Next we consider the graphical analysis.

The normal probability plot is obtained like as follows:



The normal probability plots is a plot of the ordered standardized residuals versus the cumulative probability

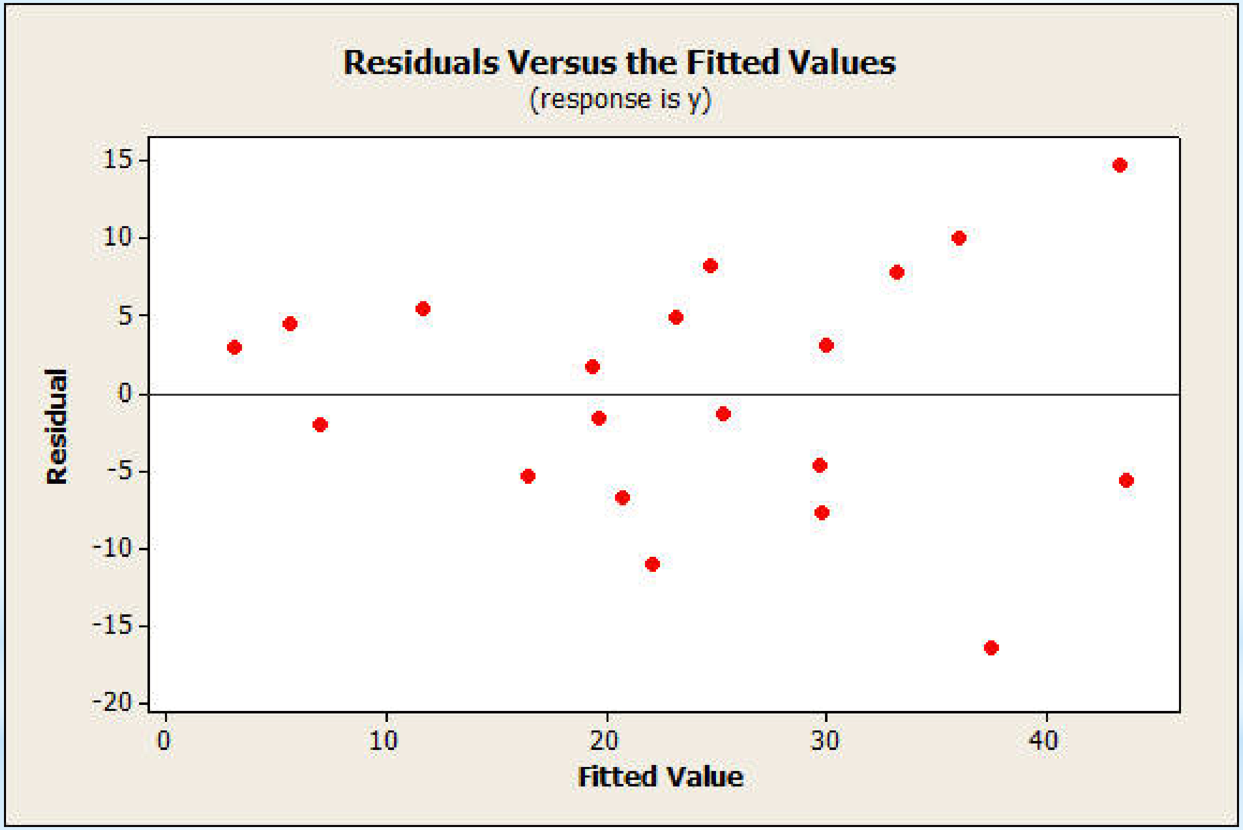
$$P_i = \frac{\left(i - \frac{1}{2}\right)}{n}, \ i = 1, 2, \ldots, n.$$

It can be seen from the plot that most of the points are lying close to the line. This clearly indicates that the assumption of normal distribution for random errors is satisfactory.

Next we consider the graph between the residuals and the fitted values.

Such a plot is helpful in detecting several common type of model inadequacies.

If plot is such that the residuals can be contained in a horizontal band (and residual fluctuates is more or less in a random fashion inside the band) then there are no obvious model defects.



The points in this plot can more or less be contained in an outward opening funnel. So the assumption of constant variance is violated and it indicates that possibly the variance increases with the increase in the values of study variable.