Regression Analysis is one of the most popular statistical methods in Data Science to predict a continuous dependent variable from a number of independent variables.

Regression Analysis

Pressure Drop Data

Prajkta Kamat

Regression Analysis can be performed on a variety of software today. The ubiquitous Microsoft Excel is still by far the most popular tool. A variety of other free and paid tools are available to run regression analysis. Some of these include SPSS, SAS, R, Python and JMP, etc.

Each of these tools presents the regression analysis output data in different ways. However, all of these tools provide essentially the same data. We present the regression output from some of the tools mentioned above.

Regression analysis is a predictive modelling technique. Linear regression is one of the simplest and most common supervised machine learning algorithms that data scientists use for predictive modelling. In this report, I'll use linear regression to build a model that predicts the results of pitch carbon analysis tests.

The determination of explicit form of regression equation is the ultimate objective of regression analysis. It is finally a good and valid relationship between study variable and explanatory variables. Such regression equation can be used for several purposes. For example, to determine the role of any explanatory variable in the joint relationship in any policy formulation, to forecast the values of response variable for given set of values of explanatory variables. The regression equation helps is understanding the interrelationships of variables among them.

The following topics are covered in the report.

- Scatter Plots
- Simple Linear Regression
- Multiple Linear Regression
- Regression diagnostics
- Variable Selection and Model Building
- Multicollinearity
- Residual and Residual Diagnostics

Pages 35

DATASET:

Pressure Drop Data

The data shows the data for 62 observations. The Regressors are x_1, x_2, x_3, and x_4 .

x1	x2	x3	x4	y	x1	x2	x3	x4	y
2.14	10	0.34	1	28.9	4.3	2.63	0.34	0.472	15.8
4.14	10	0.34	1	31	4.3	2.63	0.34	0.398	15.4
8.15	10	0.34	1	26.4	5.6	10.1	0.25	0.789	19.2
2.14	10	0.34	0.246	27.2	5.6	10.1	0.25	0.677	8.4
4.14	10	0.34	0.379	26.1	5.6	10.1	0.25	0.59	15
8.15	10	0.34	0.474	23.2	5.6	10.1	0.25	0.523	12
2.14	10	0.34	0.141	19.7	5.6	10.1	0.34	0.789	21.9
4.14	10	0.34	0.234	22.1	5.6	10.1	0.34	0.677	21.3
8.15	10	0.34	0.311	22.8	5.6	10.1	0.34	0.59	21.6
2.14	10	0.34	0.076	29.2	5.6	10.1	0.34	0.523	19.8
4.14	10	0.34	0.132	23.6	4.3	10.1	0.34	0.741	21.6
8.15	10	0.34	0.184	23.6	4.3	10.1	0.34	0.617	17.3
2.14	2.63	0.34	0.679	24.2	4.3	10.1	0.34	0.524	20
4.14	2.63	0.34	0.804	22.1	4.3	10.1	0.34	0.457	18.6
8.15	2.63	0.34	0.89	20.9	2.4	10.1	0.34	0.615	22.1
2.14	2.63	0.34	0.514	17.6	2.4	10.1	0.34	0.473	14.7
4.14	2.63	0.34	0.672	15.7	2.4	10.1	0.34	0.381	15.8
8.15	2.63	0.34	0.801	15.8	2.4	10.1	0.34	0.32	13.2
2.14	2.63	0.34	0.346	14	5.6	10.1	0.55	0.789	30.8
4.14	2.63	0.34	0.506	17.1	5.6	10.1	0.55	0.677	27.5
8.15	2.63	0.34	0.669	18.3	5.6	10.1	0.55	0.59	25.2
2.14	2.63	0.34	1	33.8	5.6	10.1	0.55	0.523	22.8
4.14	2.63	0.34	1	31.7	2.14	112	0.34	0.68	41.7
8.15	2.63	0.34	1	28.1	4.14	112	0.34	0.803	33.7
5.6	1.25	0.34	0.848	18.1	8.15	112	0.34	0.889	29.7
5.6	1.25	0.34	0.737	16.5	2.14	112	0.34	0.514	41.8
5.6	1.25	0.34	0.651	15.4	4.14	112	0.34	0.672	37.1
5.6	1.25	0.34	0.554	15	8.15	112	0.34	0.801	40.1
4.3	2.63	0.34	0.748	19.1	2.14	112	0.34	0.306	42.7
4.3	2.63	0.34	0.682	16.2	4.14	112	0.34	0.506	48.6
4.3	2.63	0.34	0.524	16.3	8.15	112	0.34	0.668	42.4

Entering Data:

R-Commands:

```
data=read.table("E:\\Msc data sets\\htr_data.csv",header=T,sep=",");data
attach(data)
```

The file has been entered in Excel, saved as a .csv extension file and the next step is to read the csv file or import the data in R. Before that let us understand the data by exploring it in R.

Minitab Commands:

```
Command Line
MTB > WOPEN "C:\Users\Admin\OneDrive\Desktop\Manjushree.csv";
SUBC>   FTYPE;
SUBC>     CSV;
SUBC>   FIELD;
SUBC>     COMMA;
SUBC>   TDELIMITER;
SUBC>     DOUBLEQUOTE;
SUBC>   DECSEP;
SUBC>     PERIOD;
SUBC>   DATA;
SUBC>     IGNOREBLANKROWS;
SUBC>     EQUALCOLUMNS;
SUBC>   SHEET 1;
SUBC>     VNAMES 1;
SUBC>     FIRST 2;
SUBC>     NROWS 32.
Retrieving worksheet from file: 'C:\Users\Admin\OneDrive\Desktop\Manjushree.csv '
Worksheet was saved on 30-03-2020
```

Data Preprocessing:

After loading the data, it's a good practice to see if there are any missing values in the data.In R the missing values are coded by the symbol NA. To identify missings in the dataset the function is.na() is used. Here our dataset is called data .Now checking for missing values in our dataset.

```
>sum(is.na(data))
```

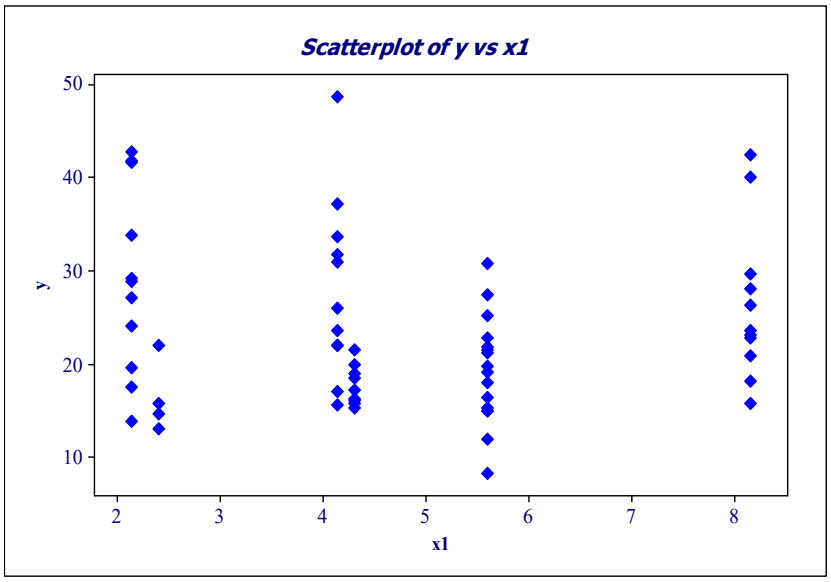
Since all the values are FALSE , and sum is 0 there are no missing values found in the dataset as seen from the R ouutput above. (However if missing values are found in the dataset they need to be handled with some statistical techniues or need to be predicted, but handling them in an intelligent way and giving rise to robust models is a challenging task)

Scatter Plots

Using a scatter plot for detection of a linear pattern:

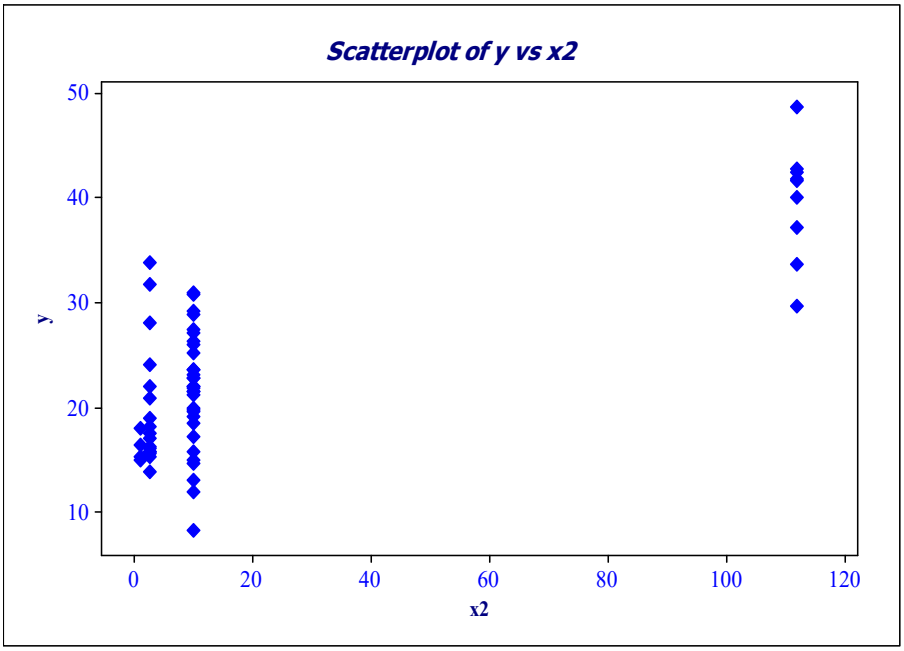
A scatter plot is used to analyze the data for directionality and correlation of data. If correlation exists or a linear pattern is observed then we assume that the data is fit to run a regression analysis. However seeing the scatter plot we can choose to fit the regression model that's best to our data.

1. Scatter plot Y v/s X_1



Interpretation :
From the above scatter plot we can observe that the relationship between Y and the observed results of x1 shows a very weak correlation.There exists multicollinearity as well. There is not a smooth pattern seen. Outliers can also be seen. But it can be also seen that replicated observations exists here. i.e. **for same values of x we get different values of y**. For further regression analysis we need to test lack of fit here using an appropriate test.

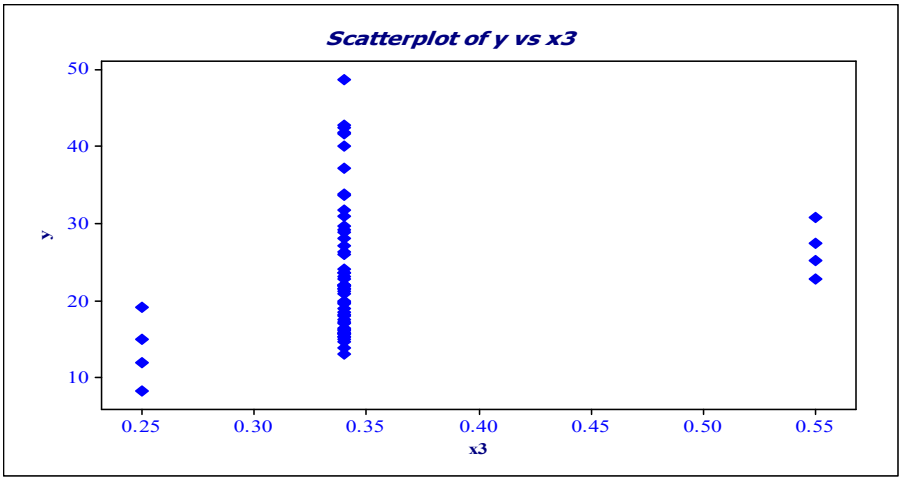
2.Scatter plot for Y V/s X_2



Interpretation :

The scatter plot shows very weak correlation between Y and x2. As there is an increasing trend we can say that the variable x2 increases, the values for y also increases. Some but low outliers are detected.It can be also seen that replicated observations exists here. i.e. for same values of x we get different values of y.

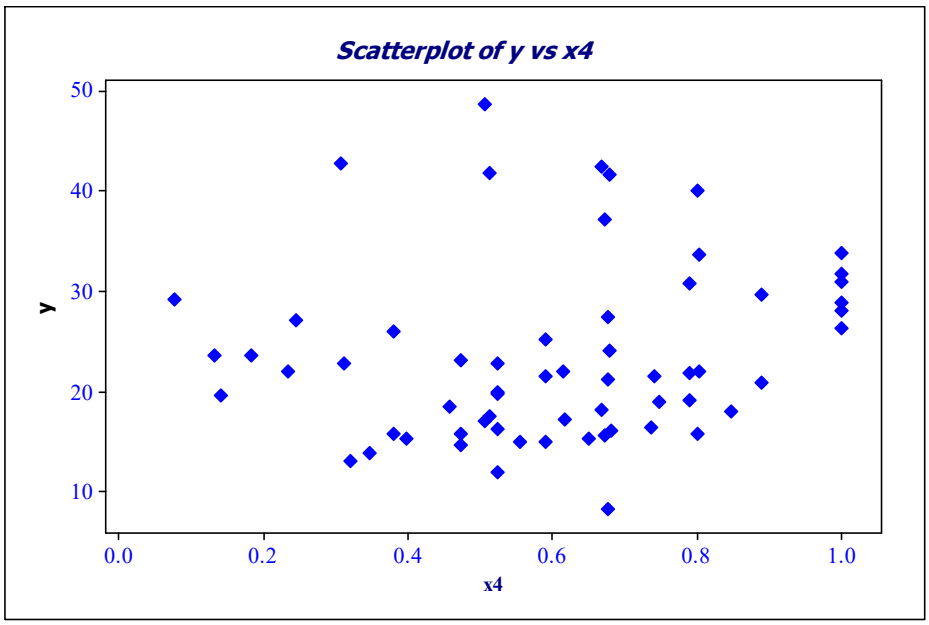
3.Scatter plot for Y V/s X_3



Interpretation :

From the scatter plot it can be seen that there exists a positive correlation between y and x3. A large number of replicated observations can be observed. From this scatter plot we can say that that quantity of x3 at the same point gives different results of the y. For less variation in x3 there exixts large variation in the values of y . But a distinct linear trend cannot be seen.

4.Scatter plot for Y V/s X_4



Interpretation :

The above scatter plot indicates that there is a positive correlation between the variables y and x4. A point that is seperated from the cluster of observations is is termed as an outlier. Some Replicated observations can also be seen here.

SIMPLE LINEAR REGRESSION

In simple linear regression, we predict scores on one variable from the scores on a second variable. The variable we are predicting is called the criterion variable or dependent variable and is referred to as Y. The variable we are basing our predictions on is called the predictor or independent or exploratory variable and is referred to as X. When there is only one predictor variable, the prediction method is called simple regression. In simple linear regression, the topic of this section, the predictions of Ywhen plotted as a function of X form a straight line(best-fitted line).

★ Regression between independent variable Y and dependent variable X_1 :

We wish to fit a linear model $Y = \beta_0 + \beta_1 X_1$.

Testing correlation is statistically significant or not:

USING R :

$H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

```
> cor.test(data$y,data$x1)

        Pearson's product-moment correlation

data:  data$y and data$x1
t = -0.35253, df = 60, p-value = 0.7257
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2919174  0.2066508
sample estimates:
        cor
-0.04546405
```

Interpretation

A higher p-value accepts the null hypothesis that $\rho = 0$

Decision :

Since our p-value > 0.05, we accept H_0 at 5 % l.o.s. and conclude that correlation is insignificant. This is also indicated from the scatter plot in the 1st part.

- **We estimate the model using least square method.**

Minitab output:

Regression Analysis: y versus x1

The regression equation is
y = 24.5 - 0.199 x1

Predictor	Coef	SE Coef	T	P
Constant	24.462	2.926	8.36	0.000
x1	-0.1993	0.5653	-0.35	0.726

S = 8.78722 R-Sq = 0.2% R-Sq(adj) = 0.0%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	1	9.60	9.60	0.12	0.726
Residual Error	60	4632.91	77.22		
Total	61	4642.51			

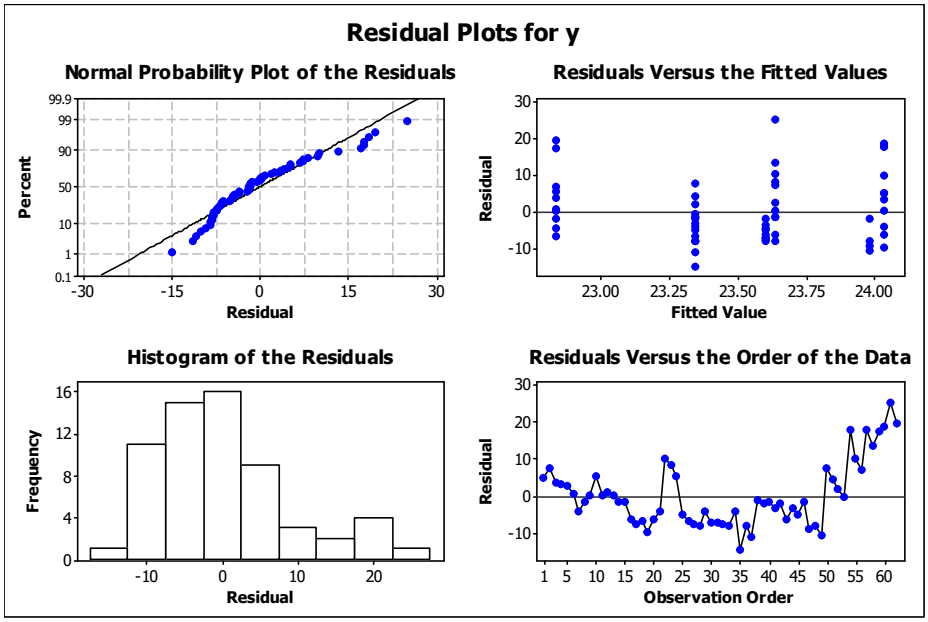
Obs	x1	y	Fit	SE Fit	Residual	St Resid
1	2.14	28.90	24.04	1.87	4.86	0.57
2	4.14	31.00	23.64	1.17	7.36	0.85
3	8.15	26.40	22.84	2.21	3.56	0.42
4	2.14	27.20	24.04	1.87	3.16	0.37
5	4.14	26.10	23.64	1.17	2.46	0.28
6	8.15	23.20	22.84	2.21	0.36	0.04
7	2.14	19.70	24.04	1.87	-4.34	-0.50
8	4.14	22.10	23.64	1.17	-1.54	-0.18
9	8.15	22.80	22.84	2.21	-0.04	-0.00
10	2.14	29.20	24.04	1.87	5.16	0.60
11	4.14	23.60	23.64	1.17	-0.04	-0.00
12	8.15	23.60	22.84	2.21	0.76	0.09
13	2.14	24.20	24.04	1.87	0.16	0.02
14	4.14	22.10	23.64	1.17	-1.54	-0.18
15	8.15	20.90	22.84	2.21	-1.94	-0.23
16	2.14	17.60	24.04	1.87	-6.44	-0.75
17	4.14	15.70	23.64	1.17	-7.94	-0.91
18	8.15	15.80	22.84	2.21	-7.04	-0.83
19	2.14	14.00	24.04	1.87	-10.04	-1.17
20	4.14	17.10	23.64	1.17	-6.54	-0.75
21	8.15	18.30	22.84	2.21	-4.54	-0.53
22	2.14	33.80	24.04	1.87	9.76	1.14
23	4.14	31.70	23.64	1.17	8.06	0.93
24	8.15	28.10	22.84	2.21	5.26	0.62
25	5.60	18.10	23.35	1.21	-5.25	-0.60
26	5.60	16.50	23.35	1.21	-6.85	-0.79
27	5.60	15.40	23.35	1.21	-7.95	-0.91
28	5.60	15.00	23.35	1.21	-8.35	-0.96
29	4.30	19.10	23.60	1.15	-4.50	-0.52
30	4.30	16.20	23.60	1.15	-7.40	-0.85
31	4.30	16.30	23.60	1.15	-7.30	-0.84
32	4.30	15.80	23.60	1.15	-7.80	-0.90
33	4.30	15.40	23.60	1.15	-8.20	-0.94
34	5.60	19.20	23.35	1.21	-4.15	-0.48
35	5.60	8.40	23.35	1.21	-14.95	-1.72
36	5.60	15.00	23.35	1.21	-8.35	-0.96
37	5.60	12.00	23.35	1.21	-11.35	-1.30
38	5.60	21.90	23.35	1.21	-1.45	-0.17
39	5.60	21.30	23.35	1.21	-2.05	-0.24
40	5.60	21.60	23.35	1.21	-1.75	-0.20

41	5.60	19.80	23.35	1.21	-3.55	-0.41
42	4.30	21.60	23.60	1.15	-2.00	-0.23
43	4.30	17.30	23.60	1.15	-6.30	-0.72
44	4.30	20.00	23.60	1.15	-3.60	-0.41
45	4.30	18.60	23.60	1.15	-5.00	-0.57
46	2.40	22.10	23.98	1.75	-1.88	-0.22
47	2.40	14.70	23.98	1.75	-9.28	-1.08
48	2.40	15.80	23.98	1.75	-8.18	-0.95
49	2.40	13.20	23.98	1.75	-10.78	-1.25
50	5.60	30.80	23.35	1.21	7.45	0.86
51	5.60	27.50	23.35	1.21	4.15	0.48
52	5.60	25.20	23.35	1.21	1.85	0.21
53	5.60	22.80	23.35	1.21	-0.55	-0.06
54	2.14	41.70	24.04	1.87	17.66	2.06R
55	4.14	33.70	23.64	1.17	10.06	1.16
56	8.15	29.70	22.84	2.21	6.86	0.81
57	2.14	41.80	24.04	1.87	17.76	2.07R
58	4.14	37.10	23.64	1.17	13.46	1.55
59	8.15	40.10	22.84	2.21	17.26	2.03R
60	2.14	42.70	24.04	1.87	18.66	2.17R
61	4.14	48.60	23.64	1.17	24.96	2.87R
62	8.15	42.40	22.84	2.21	19.56	2.30R

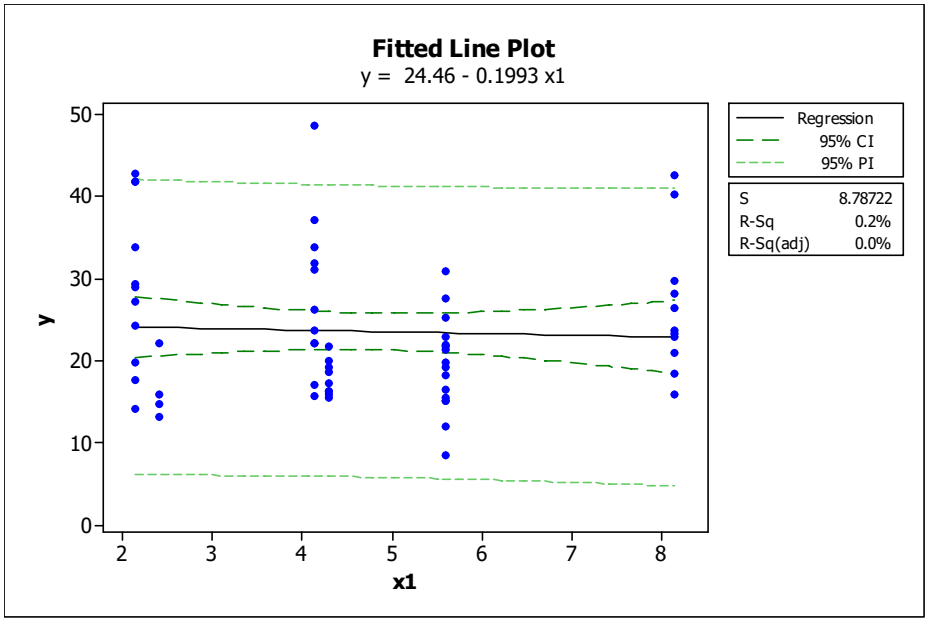
R denotes an observation with a large standardized residual.

Fitted Line: y versus x1

Residual Plots for y



Visualising confidence and prediction bands along with fitted regression line



Interpretation :

- The significance F is the probability that the null hypothesis in our regression model cannot be rejected, i.e. probability that all the coefficients in our regression output are zero.

  H_0 : Regression Model is insignificant Vs H_1 : Regression Model is significant.

- Here, in our output since p-value is significantly more, we accept null hypothesis and conclude that the model is insignificant.

Using the t test Here we test the hypothesis:

$H_0 : \beta_0 = 0$ Vs $H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0$ Vs $H_1 : \beta_1 \neq 0$

Conclusion:

Regression is insignificant. The sample estimates of β_0 and β_1 are 24.5 and -0.199 respectively. The corresponding test statistics are 8.36 and – 0.35 along with their displayed p-values indicates that , p-vales < 0.05 and p-value > 0.05, we reject H_0 and conclude that β_0 and contribute significantly to the model and accept H_0 and conclude that β_1 does not contribute significantly to the model.

Also we check the hypothesis as follows

H_0 : Residuals are Normal Vs H_1 : Residuals are not normal.

Conclusion

From the residual plots we can conclude that the assumption is not verified.

★ Regression between independent variable Y and dependent variable X_2 :

We wish to fit a linear model $Y = \beta_0 + \beta_2 X_2$.

Testing correlation is statistically significant or not:

USING R :

$H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

```
> cor.test(data$y,data$x2)

        Pearson's product-moment correlation

data:  data$y and data$x2
t = 9.7185, df = 60, p-value = 6.229e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6614206 0.8631747
sample estimates:
      cor
0.7820008
```

Interpretation

A lower p-value rejects the null hypothesis that $\rho = 0$

Decision :

Since our p-value < 0.05, we reject H_0 at 5 % l.o.s. and conclude that the sample estimates just did not come from the noise. There is a meaningful relationship between the two variables. This is also indicated from the scatter plot in the 1st part.

- **We estimate the model using least square method.**

Minitab output:

Regression Analysis: y versus x2

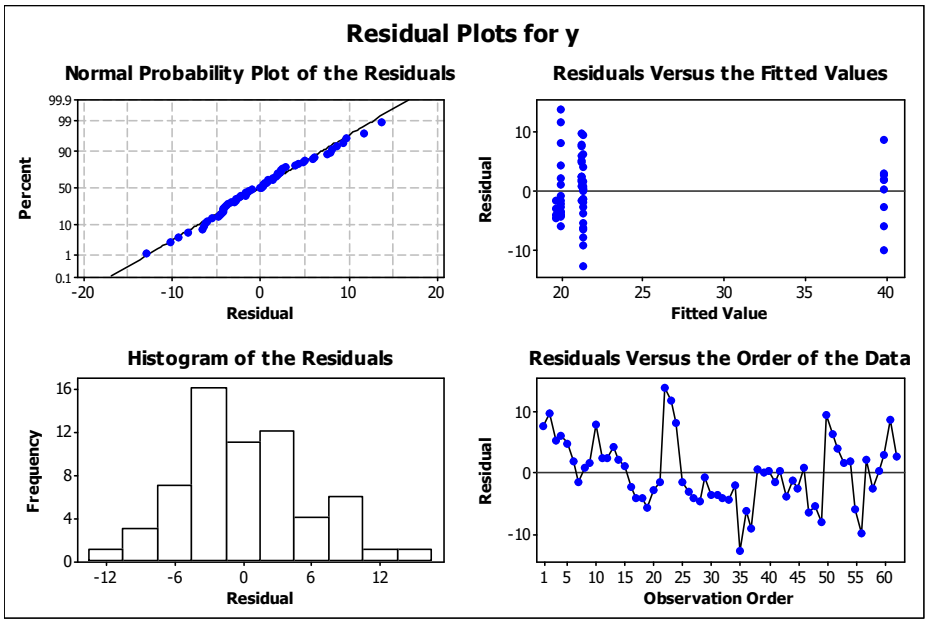
```
The regression equation is
y = 19.5 + 0.182 x2


Predictor      Coef  SE Coef      T      P
Constant    19.4544   0.8117  23.97  0.000
x2          0.18216  0.01874   9.72  0.000


S = 5.48254   R-Sq = 61.2%   R-Sq(adj) = 60.5%

PRESS = 1929.38   R-Sq(pred) = 58.44%


Analysis of Variance

Source           DF      SS      MS      F      P
Regression        1  2839.0  2839.0  94.45  0.000
Residual Error   60  1803.5    30.1
  Lack of Fit     3   312.5   104.2   3.98  0.012
  Pure Error     57  1491.0    26.2
Total            61  4642.5
```

Obs	x2	y	Fit	SE Fit	Residual	St Resid
1	10	28.900	21.276	0.733	7.624	1.40
2	10	31.000	21.276	0.733	9.724	1.79
3	10	26.400	21.276	0.733	5.124	0.94
4	10	27.200	21.276	0.733	5.924	1.09
5	10	26.100	21.276	0.733	4.824	0.89
6	10	23.200	21.276	0.733	1.924	0.35
7	10	19.700	21.276	0.733	-1.576	-0.29
8	10	22.100	21.276	0.733	0.824	0.15
9	10	22.800	21.276	0.733	1.524	0.28
10	10	29.200	21.276	0.733	7.924	1.46
11	10	23.600	21.276	0.733	2.324	0.43
12	10	23.600	21.276	0.733	2.324	0.43
13	3	24.200	19.933	0.787	4.267	0.79
14	3	22.100	19.933	0.787	2.167	0.40
15	3	20.900	19.933	0.787	0.967	0.18
16	3	17.600	19.933	0.787	-2.333	-0.43
17	3	15.700	19.933	0.787	-4.233	-0.78
18	3	15.800	19.933	0.787	-4.133	-0.76
19	3	14.000	19.933	0.787	-5.933	-1.09
20	3	17.100	19.933	0.787	-2.833	-0.52
21	3	18.300	19.933	0.787	-1.633	-0.30
22	3	33.800	19.933	0.787	13.867	2.56R
23	3	31.700	19.933	0.787	11.767	2.17R
24	3	28.100	19.933	0.787	8.167	1.51
25	1	18.100	19.682	0.800	-1.582	-0.29
26	1	16.500	19.682	0.800	-3.182	-0.59
27	1	15.400	19.682	0.800	-4.282	-0.79
28	1	15.000	19.682	0.800	-4.682	-0.86
29	3	19.100	19.933	0.787	-0.833	-0.15
30	3	16.200	19.933	0.787	-3.733	-0.69
31	3	16.300	19.933	0.787	-3.633	-0.67
32	3	15.800	19.933	0.787	-4.133	-0.76
33	3	15.400	19.933	0.787	-4.533	-0.84
34	10	19.200	21.294	0.733	-2.094	-0.39
35	10	8.400	21.294	0.733	-12.894	-2.37R
36	10	15.000	21.294	0.733	-6.294	-1.16
37	10	12.000	21.294	0.733	-9.294	-1.71
38	10	21.900	21.294	0.733	0.606	0.11
39	10	21.300	21.294	0.733	0.006	0.00
40	10	21.600	21.294	0.733	0.306	0.06
41	10	19.800	21.294	0.733	-1.494	-0.28
42	10	21.600	21.294	0.733	0.306	0.06
43	10	17.300	21.294	0.733	-3.994	-0.74
44	10	20.000	21.294	0.733	-1.294	-0.24
45	10	18.600	21.294	0.733	-2.694	-0.50
46	10	22.100	21.294	0.733	0.806	0.15
47	10	14.700	21.294	0.733	-6.594	-1.21
48	10	15.800	21.294	0.733	-5.494	-1.01
49	10	13.200	21.294	0.733	-8.094	-1.49
50	10	30.800	21.294	0.733	9.506	1.75
51	10	27.500	21.294	0.733	6.206	1.14
52	10	25.200	21.294	0.733	3.906	0.72
53	10	22.800	21.294	0.733	1.506	0.28
54	112	41.700	39.856	1.821	1.844	0.36 X
55	112	33.700	39.856	1.821	-6.156	-1.19 X
56	112	29.700	39.856	1.821	-10.156	-1.96 X
57	112	41.800	39.856	1.821	1.944	0.38 X
58	112	37.100	39.856	1.821	-2.756	-0.53 X
59	112	40.100	39.856	1.821	0.244	0.05 X
60	112	42.700	39.856	1.821	2.844	0.55 X
61	112	48.600	39.856	1.821	8.744	1.69 X
62	112	42.400	39.856	1.821	2.544	0.49 X

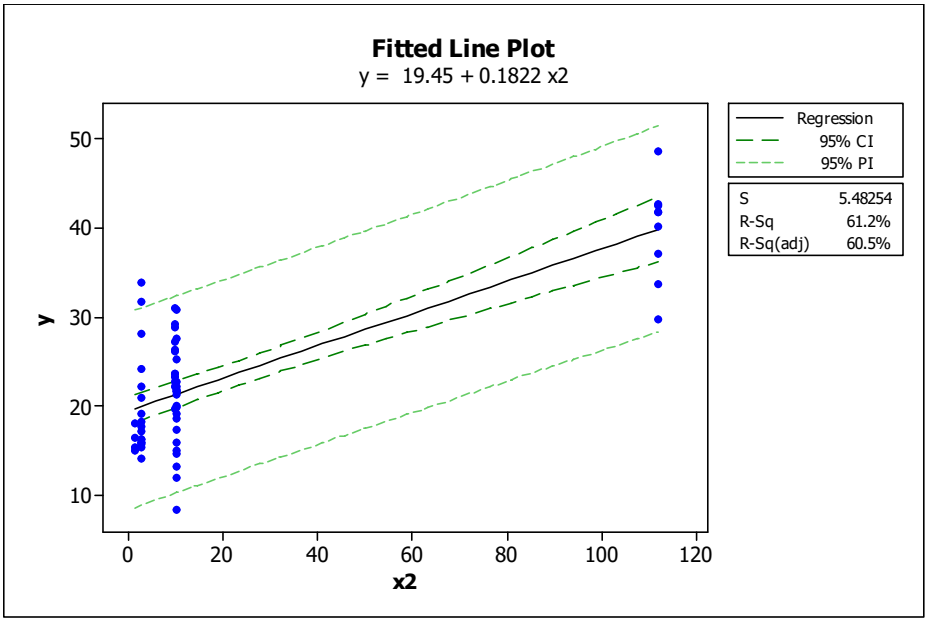
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

```
Durbin-Watson statistic = 0.885743
```

Residual Plots for y



Fitted Line: y versus x2



Visualising confidence and prediction bands along with fitted regression line

Interpretation :

- The significance F is the probability that the null hypothesis in our regression model cannot be rejected, i.e. probability that all the coefficients in our regression output are zero.

  H_0 : Regression Model is insignificant Vs H_1 : Regression Model is significant.

- Here, in our output since p-value is significantly less, we reject null hypothesis and conclude that the model is significant.

Using the t test Here we test the hypothesis:

$H_0 : \beta_0 = 0$ Vs $H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0$ Vs $H_1 : \beta_1 \neq 0$

Conclusion:

Regression is insignificant. The sample estimates of β_0 and β_1 are 19.5 and 0.182 respectively. The corresponding test statistics are 23.9 and 9.72 along with their displayed p-values indicates that , p-vales < 0.05, we reject H_0 and conclude that β_0 and β_1 contribute significantly to the model.

Also we check the hypothesis as follows

H_0 : Residuals are Normal Vs H_1 : Residuals are not normal.

Conclusion

Using the residual plots to help you determine whether the model is adequate and meets the assumptions of the analysis. If the assumptions are not met, the model may not fit the data well and you should use caution when you interpret the results.

We made the assumptions that the all the error terms are identically and independently normally distributed with mean 0 and common variance sigma –square.

From the residual plots we can conclude that the assumption is verified.

The regression equation is :

```
y = 19.5 + 0.182 x2
```

Interpretation :

$\beta_0 = 19.5$ $\beta_1 = 0.182$

The line intersects y axis at 19.5 with a slope of 0.182 i.e. at 0 x2 , the value of the y will be 19.5 and for each increase in x2, the y results will also increase on an average by 0.182.

★Regression between independent variable Y and dependent variable X_3 :

We wish to fit a linear model $Y = \beta_0 + \beta_3 X_3$.

Testing correlation is statistically significant or not:

USING R :

$H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

```
> cor.test(data$y,data$x3)

        Pearson's product-moment correlation

data:  data$y and data$x3
t = 1.5688, df = 60, p-value = 0.1219
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05393658  0.42709830
sample estimates:
      cor
0.1985058
```

Interpretation

A higher p-value accepts the null hypothesis that $\rho = 0$

Decision :

Since our p-value > 0.05, we accept H_0 at 5 % l.o.s. and conclude that correlation is insignificant.

- **We estimate the model using least square method.**

Minitab output:

Regression Analysis: y versus x3

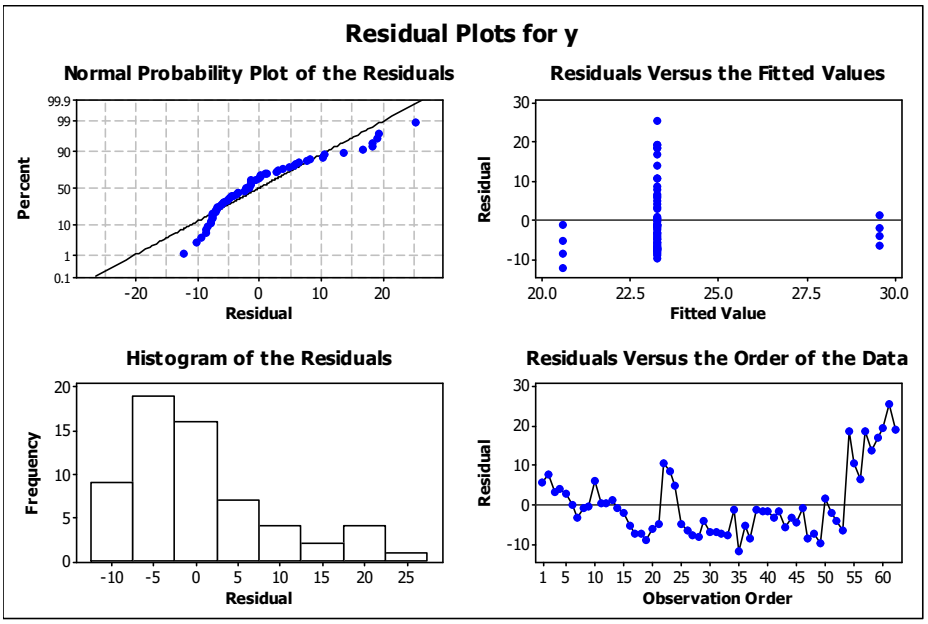
```
The regression equation is
y = 13.1 + 29.9 x3


Predictor     Coef  SE Coef     T      P
Constant   13.122    6.710  1.96  0.055
x3          29.87    19.04  1.57  0.122


S = 8.62126   R-Sq = 3.9%   R-Sq(adj) = 2.3%

PRESS = 4678.00   R-Sq(pred) = 0.00%


Analysis of Variance

Source          DF       SS      MS     F      P
Regression       1   182.94  182.94  2.46  0.122
Residual Error  60  4459.57   74.33
  Lack of Fit    1   257.08  257.08  3.61  0.062
  Pure Error    59  4202.49   71.23
Total           61  4642.51
```

Obs	x3	y	Fit	SE Fit	Residual	St Resid
1	0.340	28.90	23.28	1.10	5.62	0.66
2	0.340	31.00	23.28	1.10	7.72	0.90
3	0.340	26.40	23.28	1.10	3.12	0.37
4	0.340	27.20	23.28	1.10	3.92	0.46
5	0.340	26.10	23.28	1.10	2.82	0.33
6	0.340	23.20	23.28	1.10	-0.08	-0.01
7	0.340	19.70	23.28	1.10	-3.58	-0.42
8	0.340	22.10	23.28	1.10	-1.18	-0.14
9	0.340	22.80	23.28	1.10	-0.48	-0.06
10	0.340	29.20	23.28	1.10	5.92	0.69
11	0.340	23.60	23.28	1.10	0.32	0.04
12	0.340	23.60	23.28	1.10	0.32	0.04
13	0.340	24.20	23.28	1.10	0.92	0.11
14	0.340	22.10	23.28	1.10	-1.18	-0.14
15	0.340	20.90	23.28	1.10	-2.38	-0.28
16	0.340	17.60	23.28	1.10	-5.68	-0.66
17	0.340	15.70	23.28	1.10	-7.58	-0.89
18	0.340	15.80	23.28	1.10	-7.48	-0.87
19	0.340	14.00	23.28	1.10	-9.28	-1.08
20	0.340	17.10	23.28	1.10	-6.18	-0.72
21	0.340	18.30	23.28	1.10	-4.98	-0.58
22	0.340	33.80	23.28	1.10	10.52	1.23
23	0.340	31.70	23.28	1.10	8.42	0.99
24	0.340	28.10	23.28	1.10	4.82	0.56
25	0.340	18.10	23.28	1.10	-5.18	-0.61
26	0.340	16.50	23.28	1.10	-6.78	-0.79
27	0.340	15.40	23.28	1.10	-7.88	-0.92
28	0.340	15.00	23.28	1.10	-8.28	-0.97
29	0.340	19.10	23.28	1.10	-4.18	-0.49
30	0.340	16.20	23.28	1.10	-7.08	-0.83
31	0.340	16.30	23.28	1.10	-6.98	-0.82
32	0.340	15.80	23.28	1.10	-7.48	-0.87
33	0.340	15.40	23.28	1.10	-7.88	-0.92
34	0.250	19.20	20.59	2.16	-1.39	-0.17
35	0.250	8.40	20.59	2.16	-12.19	-1.46
36	0.250	15.00	20.59	2.16	-5.59	-0.67
37	0.250	12.00	20.59	2.16	-8.59	-1.03
38	0.340	21.90	23.28	1.10	-1.38	-0.16
39	0.340	21.30	23.28	1.10	-1.98	-0.23
40	0.340	21.60	23.28	1.10	-1.68	-0.20
41	0.340	19.80	23.28	1.10	-3.48	-0.41
42	0.340	21.60	23.28	1.10	-1.68	-0.20
43	0.340	17.30	23.28	1.10	-5.98	-0.70
44	0.340	20.00	23.28	1.10	-3.28	-0.38
45	0.340	18.60	23.28	1.10	-4.68	-0.55
46	0.340	22.10	23.28	1.10	-1.18	-0.14
47	0.340	14.70	23.28	1.10	-8.58	-1.00
48	0.340	15.80	23.28	1.10	-7.48	-0.87
49	0.340	13.20	23.28	1.10	-10.08	-1.18
50	0.550	30.80	29.55	4.00	1.25	0.16 X
51	0.550	27.50	29.55	4.00	-2.05	-0.27 X
52	0.550	25.20	29.55	4.00	-4.35	-0.57 X
53	0.550	22.80	29.55	4.00	-6.75	-0.88 X
54	0.340	41.70	23.28	1.10	18.42	2.15R
55	0.340	33.70	23.28	1.10	10.42	1.22
56	0.340	29.70	23.28	1.10	6.42	0.75
57	0.340	41.80	23.28	1.10	18.52	2.17R
58	0.340	37.10	23.28	1.10	13.82	1.62
59	0.340	40.10	23.28	1.10	16.82	1.97
60	0.340	42.70	23.28	1.10	19.42	2.27R
61	0.340	48.60	23.28	1.10	25.32	2.96R
62	0.340	42.40	23.28	1.10	19.12	2.24R

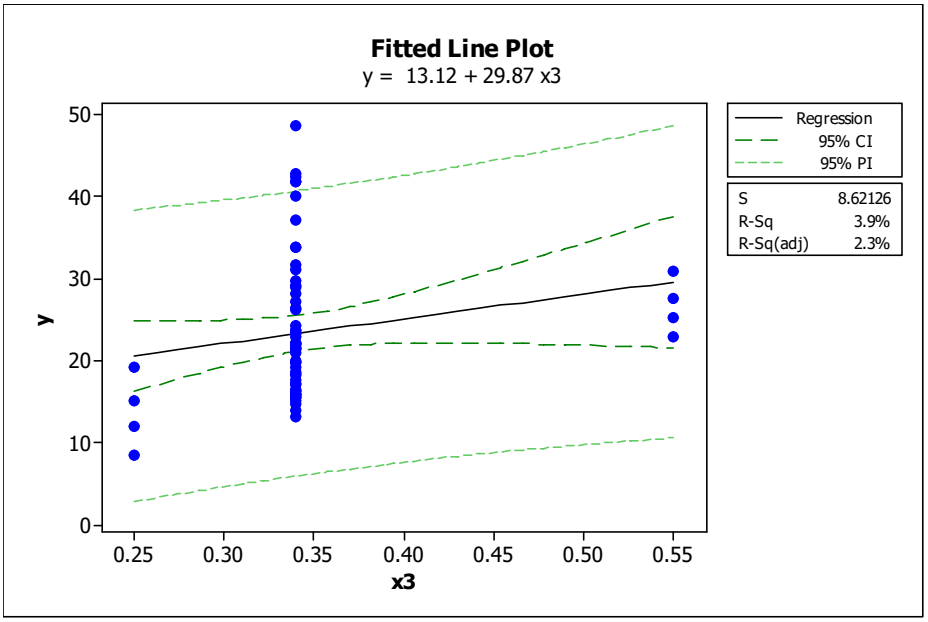
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

```
Durbin-Watson statistic = 0.458163
```

Residual Plots for y



Fitted Line: y versus x3



Visualising confidence and prediction bands along with fitted regression line.

Interpretation :

- The significance F is the probability that the null hypothesis in our regression model cannot be rejected, i.e. probability that all the coefficients in our regression output are zero.

  H_0 : Regression Model is insignificant Vs H_1 : Regression Model is significant.

- Here, in our output since p-value is not significantly less, we accept the null hypothesis and conclude that the model is insignificant.

Using the t test Here we test the hypothesis:

$H_0 : \beta_0 = 0$ Vs $H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0$ Vs $H_1 : \beta_1 \neq 0$

Conclusion:

Regression is insignificant. The sample estimates of β_0 and β_1 are 13.122 and 29.9 respectively. The corresponding test statistics are 1.96 and 1.57 along with their displayed p-values indicates that , p-vales < 0.05 and p-value > 0.05, we reject H_0 and conclude that β_0 and contribute significantly to the model and accept H_0 and conclude that β_1 does not contribute significantly to the model.

Also we check the hypothesis as follows

H_0 : Residuals are Normal Vs H_1 : Residuals are not normal.

Conclusion

Using the residual plots to help you determine whether the model is adequate and meets the assumptions of the analysis. If the assumptions are not met, the model may not fit the data well and you should use caution when you interpret the results.

We made the assumptions that the all the error terms are identically and independently normally distributed with mean 0 and common variance sigma –square.

From the residual plots we can conclude that the assumption is verified.

The regression equation is :

```
y = 13.1 + 29.9 x3
```

Interpretation :

$\beta_0 = 13.1$ $\beta_3 = 29.9$

The line intersects y axis at 13.1 with a slope of 29.9 i.e. at 0 x3 , the value of the y will be 13.1 and for each increase in x3, the y results will also increase on an average by 29.9.

★Regression between independent variable Y and dependent variable X_4 :

We wish to fit a linear model $Y = \beta_0 + \beta_4 X_4$.

Testing correlation is statistically significant or not:

USING R :

$H_0 : \rho = 0$ vs $H_1 : \rho \neq 0$

```
> cor.test(data$y,data$x4)

        Pearson's product-moment correlation

data:  data$y and data$x4
t = 1.3281, df = 60, p-value = 0.1892
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.0843338  0.4018025
sample estimates:
      cor
0.1689941
```

Interpretation

A higher p-value accepts the null hypothesis that $\rho = 0$

Decision :

Since our p-value > 0.05, we accept H_0 at 5 % l.o.s. and conclude that correlation is insignificant.

- **We estimate the model using least square method.**

Minitab output:

Regression Analysis: y versus x4

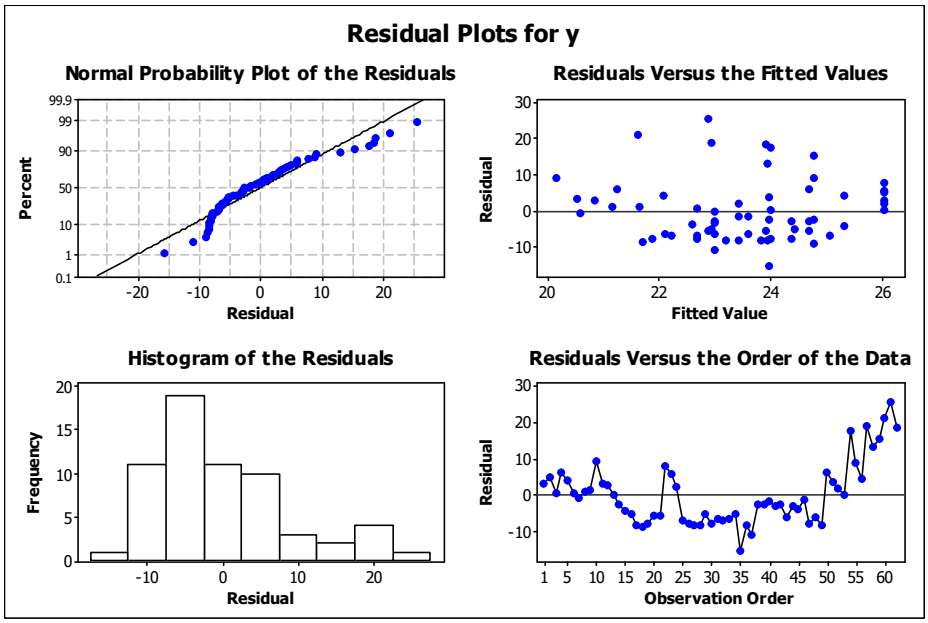
```
The regression equation is
y = 19.7 + 6.34 x4


Predictor     Coef  SE Coef     T      P
Constant    19.687    3.081  6.39  0.000
x4           6.338    4.772  1.33  0.189


S = 8.66979   R-Sq = 2.9%   R-Sq(adj) = 1.2%

PRESS = 4760.59   R-Sq(pred) = 0.00%


Analysis of Variance

Source           DF       SS      MS     F      P
Regression        1   132.59  132.59  1.76  0.189
Residual Error   60  4509.92   75.17
  Lack of Fit    42  2774.54   66.06  0.69  0.845
  Pure Error     18  1735.38   96.41
Total            61  4642.51


 34 rows with no replicates
```

Obs	x4	y	Fit	SE Fit	Residual	St Resid	
1	1.00	28.90	26.03	2.19	2.87	0.34	
2	1.00	31.00	26.03	2.19	4.97	0.59	
3	1.00	26.40	26.03	2.19	0.37	0.04	
4	0.25	27.20	21.25	2.03	5.95	0.71	
5	0.38	26.10	22.09	1.53	4.01	0.47	
6	0.47	23.20	22.69	1.26	0.51	0.06	
7	0.14	19.70	20.58	2.46	-0.88	-0.11	
8	0.23	22.10	21.17	2.08	0.93	0.11	
9	0.31	22.80	21.66	1.78	1.14	0.13	
10	0.08	29.20	20.17	2.74	9.03	1.10	X
11	0.13	23.60	20.52	2.50	3.08	0.37	
12	0.18	23.60	20.85	2.28	2.75	0.33	
13	0.68	24.20	23.99	1.16	0.21	0.02	
14	0.80	22.10	24.78	1.46	-2.68	-0.31	
15	0.89	20.90	25.33	1.76	-4.43	-0.52	
16	0.51	17.60	22.94	1.18	-5.34	-0.62	
17	0.67	15.70	23.95	1.15	-8.25	-0.96	
18	0.80	15.80	24.76	1.45	-8.96	-1.05	
19	0.35	14.00	21.88	1.65	-7.88	-0.93	
20	0.51	17.10	22.89	1.19	-5.79	-0.67	
21	0.67	18.30	23.93	1.15	-5.63	-0.65	
22	1.00	33.80	26.03	2.19	7.77	0.93	
23	1.00	31.70	26.03	2.19	5.67	0.68	
24	1.00	28.10	26.03	2.19	2.07	0.25	
25	0.85	18.10	25.06	1.61	-6.96	-0.82	
26	0.74	16.50	24.36	1.27	-7.86	-0.92	
27	0.65	15.40	23.81	1.12	-8.41	-0.98	
28	0.55	15.00	23.20	1.13	-8.20	-0.95	
29	0.75	19.10	24.43	1.30	-5.33	-0.62	
30	0.68	16.20	24.01	1.16	-7.81	-0.91	
31	0.52	16.30	23.01	1.16	-6.71	-0.78	
32	0.47	15.80	22.68	1.27	-6.88	-0.80	
33	0.40	15.40	22.21	1.47	-6.81	-0.80	
34	0.79	19.20	24.69	1.41	-5.49	-0.64	
35	0.68	8.40	23.98	1.16	-15.58	-1.81	
36	0.59	15.00	23.43	1.10	-8.43	-0.98	
37	0.52	12.00	23.00	1.17	-11.00	-1.28	
38	0.79	21.90	24.69	1.41	-2.79	-0.33	
39	0.68	21.30	23.98	1.16	-2.68	-0.31	
40	0.59	21.60	23.43	1.10	-1.83	-0.21	
41	0.52	19.80	23.00	1.17	-3.20	-0.37	
42	0.74	21.60	24.38	1.28	-2.78	-0.32	
43	0.62	17.30	23.60	1.10	-6.30	-0.73	
44	0.52	20.00	23.01	1.16	-3.01	-0.35	
45	0.46	18.60	22.58	1.30	-3.98	-0.46	
46	0.62	22.10	23.59	1.10	-1.49	-0.17	
47	0.47	14.70	22.69	1.26	-7.99	-0.93	
48	0.38	15.80	22.10	1.53	-6.30	-0.74	
49	0.32	13.20	21.72	1.74	-8.52	-1.00	
50	0.79	30.80	24.69	1.41	6.11	0.71	
51	0.68	27.50	23.98	1.16	3.52	0.41	
52	0.59	25.20	23.43	1.10	1.77	0.21	
53	0.52	22.80	23.00	1.17	-0.20	-0.02	
54	0.68	41.70	24.00	1.16	17.70	2.06	R
55	0.80	33.70	24.78	1.46	8.92	1.04	
56	0.89	29.70	25.32	1.75	4.38	0.52	
57	0.51	41.80	22.94	1.18	18.86	2.20	R
58	0.67	37.10	23.95	1.15	13.15	1.53	
59	0.80	40.10	24.76	1.45	15.34	1.79	
60	0.31	42.70	21.63	1.79	21.07	2.48	R
61	0.51	48.60	22.89	1.19	25.71	2.99	R
62	0.67	42.40	23.92	1.14	18.48	2.15	R

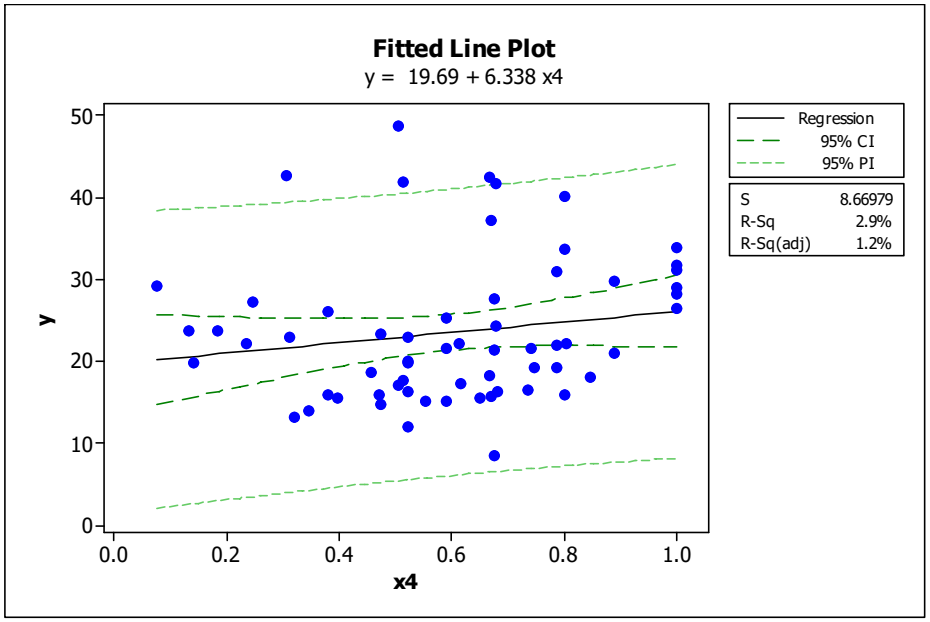
R denotes an observation with a large standardized residual.
X denotes an observation whose X value gives it large influence.

Durbin-Watson statistic = 0.402619

Residual Plots for y



Fitted Line: y versus x4



Visualising confidence and prediction bands along with fitted regression line

Interpretation :

- The significance F is the probability that the null hypothesis in our regression model cannot be rejected, i.e. probability that all the coefficients in our regression output are zero.

  H_0 : Regression Model is insignificant Vs H_1 : Regression Model is significant.

- Here, in our output since p-value is not significantly less, we accept null hypothesis and conclude that the model is insignificant.

Using the t test Here we test the hypothesis:

$H_0 : \beta_0 = 0$ Vs $H_1 : \beta_0 \neq 0$

$H_0 : \beta_1 = 0$ Vs $H_1 : \beta_1 \neq 0$

Conclusion:

Regression is insignificant. The sample estimates of β_0 and β_1 are 13.122 and 29.9 respectively. The corresponding test statistics are 6.39 and 1.33 along with their displayed p-values indicates that , p-vales < 0.05 and p-value > 0.05, we reject H_0 and conclude that β_0 and contribute significantly to the model and accept H_0 and conclude that β_1 does not contribute significantly to the model.

Also we check the hypothesis as follows

H_0 : Residuals are Normal Vs H_1 : Residuals are not normal.

Conclusion

Using the residual plots to help you determine whether the model is adequate and meets the assumptions of the analysis. If the assumptions are not met, the model may not fit the data well and you should use caution when you interpret the results.

We made the assumptions that the all the error terms are identically and independently normally distributed with mean 0 and common variance sigma –square.

From the residual plots we can conclude that the assumption is not verified.

The regression equation is :

```
y = 19.7 + 6.34 x4
```

Interpretation :

$\beta_0 = 19.7$ $\beta_4 = 6.34$

The line intersects y axis at 19.7 with a slope of 6.34 i.e. at 0 x4 , the value of the y will be 19.7 and for each increase in x4, the y results will also increase on an average by 6.34.

MULTIPLE LINEAR REGRESSION

First enter the data. To decide whether we can make a predictive model, our first step is to see if there appears to be a strong relationship between our predictor and response variables. Summary function gives the range,quartiles,median for numerical variables and table wth frequencies for categorial variables.

Studying correlation between variables:

Correlations: x1, x2, x3, x4, y

```
        x1       x2       x3       x4
x2   0.002
x3   0.056   -0.044
x4   0.312    0.053    0.024
y   -0.045    0.782    0.199    0.169

Cell Contents: Pearson correlation
```

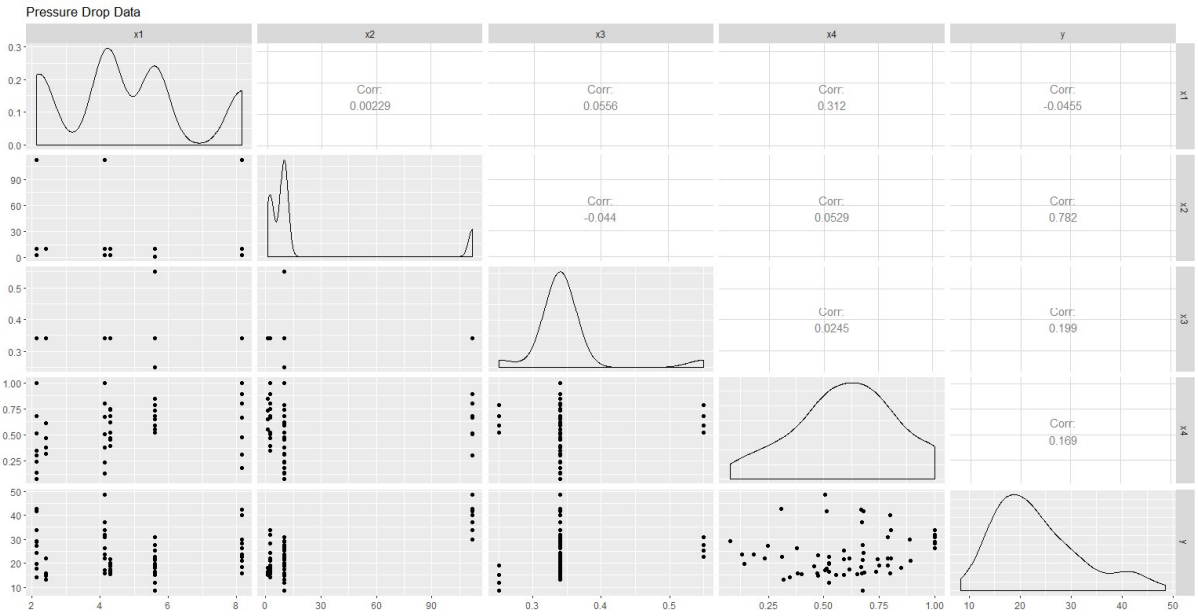
Visualising the correlation between the independent variables uding ggpairs.

Interpretation

ggpairs() function creates a plot matrix to see how the variables relate to one another and gives scatter plots for each variable combination, as well as density plots for each variable and the strength of correlations between variables.

The correlation coefficients provide information about relationships i.e. the closer the coefficient is closer to 1, the stronger the relationship is.

```
> ggpairs(data,title="Pressure Drop Data")
```



Interpretation

- From all the above plots we are getting to know that there exists correlation between variables.
- Scatter plots, linear pattern can be seen between variable combinations.
- From the density plots, the proportion of data points that are in accumulated in one variable is visualised
- Here assumption of non-multicollinearity is also verified. i.e. the independent variables are not highly correlated with each other(magnitude of correlation coefficients is less than .8 .

Fitting the model

```
> summary(lm(data$y~data$x1+data$x2+data$x3+data$x4,data=data))

Call:
lm(formula = data$y ~ data$x1 + data$x2 + data$x3 + data$x4,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.9958 -3.3092 -0.2419  3.3924 10.5668

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.89453    4.32508   1.363  0.17828
data$x1     -0.47790    0.34002  -1.406  0.16530
data$x2      0.18271    0.01718  10.633 3.78e-15 ***
data$x3     35.40284   11.09960   3.190  0.00232 **
data$x4      5.84391    2.90978   2.008  0.04935 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.014 on 57 degrees of freedom
Multiple R-squared:  0.6914,    Adjusted R-squared:  0.6697
F-statistic: 31.92 on 4 and 57 DF,  p-value: 5.818e-14
```

Interpretation :

H_0 : Regression is insignificant Vs H_1 : Regression is significant.

- From the above r output , since p-value < 0.05 , we reject H_0 at 5% l.o.s. and conclude that overall regression is significant .
- Here R-Sq = 69.14 % which indicates that whenever we observe variation in the value of the dependent variable Y, only 69.14 % of the variation is explained by our fitted model and only 30.86 % of the variation is due to some unexplained factor.
- Therefore from R-Sq we can say that the model is a good fit to the data.

Here we test :

$H_0 : \beta_0 = 0$ Vs $H_1 : \beta_0 \neq 0$

From ANOVA, Since p-value 0.17828 > 0.05 , we accept H_0 and conclude that $\beta_0 = 0$ and it is insignificant.

$H_0 : \beta_1 = 0$ Vs $H_1 : \beta_1 \neq 0$

From ANOVA, Since p-value 0.16530 > 0.05 , we accept H_0 and conclude that $\beta_1 = 0$ and the variable x1 does not contributes significantly to the model.

$H_0 : \beta_2 = 0$ Vs $H_1 : \beta_2 \neq 0$

From ANOVA, Since p-value 0.00 < 0.05 , we reject H_0 and conclude that $\beta_2 = 0$ and the variable x2 contributes significantly to the model.

$H_0 : \beta_3 = 0$ Vs $H_1 : \beta_3 \neq 0$

From ANOVA, Since p-value 0.00232 < 0.05 , we reject H_0 and conclude that $\beta_3 = 0$ and the variable x3 contributes significantly to the model.

$H_0 : \beta_4 = 0$ Vs $H_1 : \beta_4 \neq 0$

From ANOVA, Since p-value 0.04935< 0.05 , we reject H_0 and conclude that $\beta_4 \neq 0$ and the variable x4 contributes significantly to the model.

Hence variables x2, x3, and x4 are significant in predicting y values.

Regression Equation

Y = 5.89453 – 0.4779 x1 + 0.18271 x2 + 35.40284 x3 + 5.84391 x4

Outliers:

```
> outlier(data)
    x1      x2     x3     x4     y
 8.150 112.000  0.550  0.076 48.600
```

Interpretation

The observation vector shown above are outliers. To improve the model we have to either check the cause and review the situation or delete these observations and recompute the model.

Residual Analysis and Residual Diagnostics.

Residual Analysis: y versus x1, x2, x3, x4

The regression equation is
y = 5.89 - 0.478 x1 + 0.183 x2 + 35.4 x3 + 5.84 x4

S = 5.01364 R-Sq = 69.1% R-Sq(adj) = 67.0%

PRESS = 1762.86 R-Sq(pred) = 62.03%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	3209.72	802.43	31.92	0.000
Residual Error	57	1432.79	25.14		
Total	61	4642.51			

No replicates.
Cannot do pure error test.

Source	DF	Seq SS
x1	1	9.60
x2	1	2839.78
x3	1	258.95
x4	1	101.39

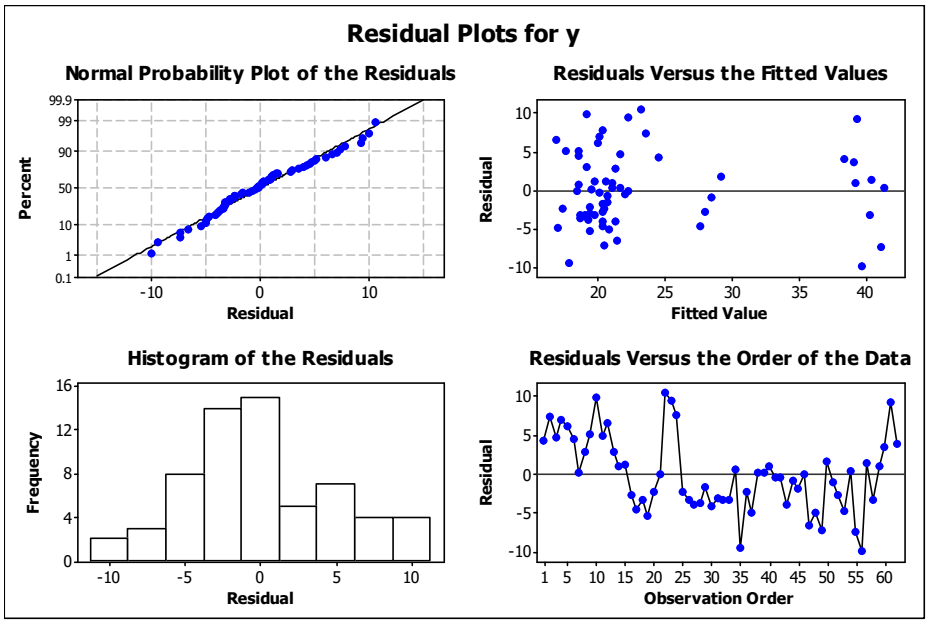
Obs	x1	y	Fit	SE Fit	Residual	St Resid
1	2.14	28.900	24.580	1.810	4.320	0.92
2	4.14	31.000	23.624	1.424	7.376	1.53
3	8.15	26.400	21.708	1.520	4.692	0.98
4	2.14	27.200	20.174	1.317	7.026	1.45
5	4.14	26.100	19.995	0.908	6.105	1.24
6	8.15	23.200	18.634	1.473	4.566	0.95
7	2.14	19.700	19.560	1.511	0.140	0.03
8	4.14	22.100	19.148	1.218	2.952	0.61
9	8.15	22.800	17.681	1.754	5.119	1.09
10	2.14	29.200	19.180	1.648	10.020	2.12R
11	4.14	23.600	18.552	1.470	5.048	1.05
12	8.15	23.600	16.939	2.024	6.661	1.45
13	2.14	24.200	21.357	1.233	2.843	0.58
14	4.14	22.100	21.132	1.011	0.968	0.20
15	8.15	20.900	19.718	1.405	1.182	0.25
16	2.14	17.600	20.393	1.118	-2.793	-0.57
17	4.14	15.700	20.361	0.808	-4.661	-0.94
18	8.15	15.800	19.198	1.335	-3.398	-0.70
19	2.14	14.000	19.411	1.203	-5.411	-1.11
20	4.14	17.100	19.391	0.779	-2.291	-0.46
21	8.15	18.300	18.427	1.321	-0.127	-0.03
22	2.14	33.800	23.233	1.835	10.567	2.26R
23	4.14	31.700	22.277	1.455	9.423	1.96
24	8.15	28.100	20.361	1.547	7.739	1.62
25	5.60	18.100	20.439	1.019	-2.339	-0.48
26	5.60	16.500	19.791	0.851	-3.291	-0.67
27	5.60	15.400	19.288	0.790	-3.888	-0.79
28	5.60	15.000	18.721	0.813	-3.721	-0.75
29	4.30	19.100	20.728	0.891	-1.628	-0.33
30	4.30	16.200	20.343	0.801	-4.143	-0.84

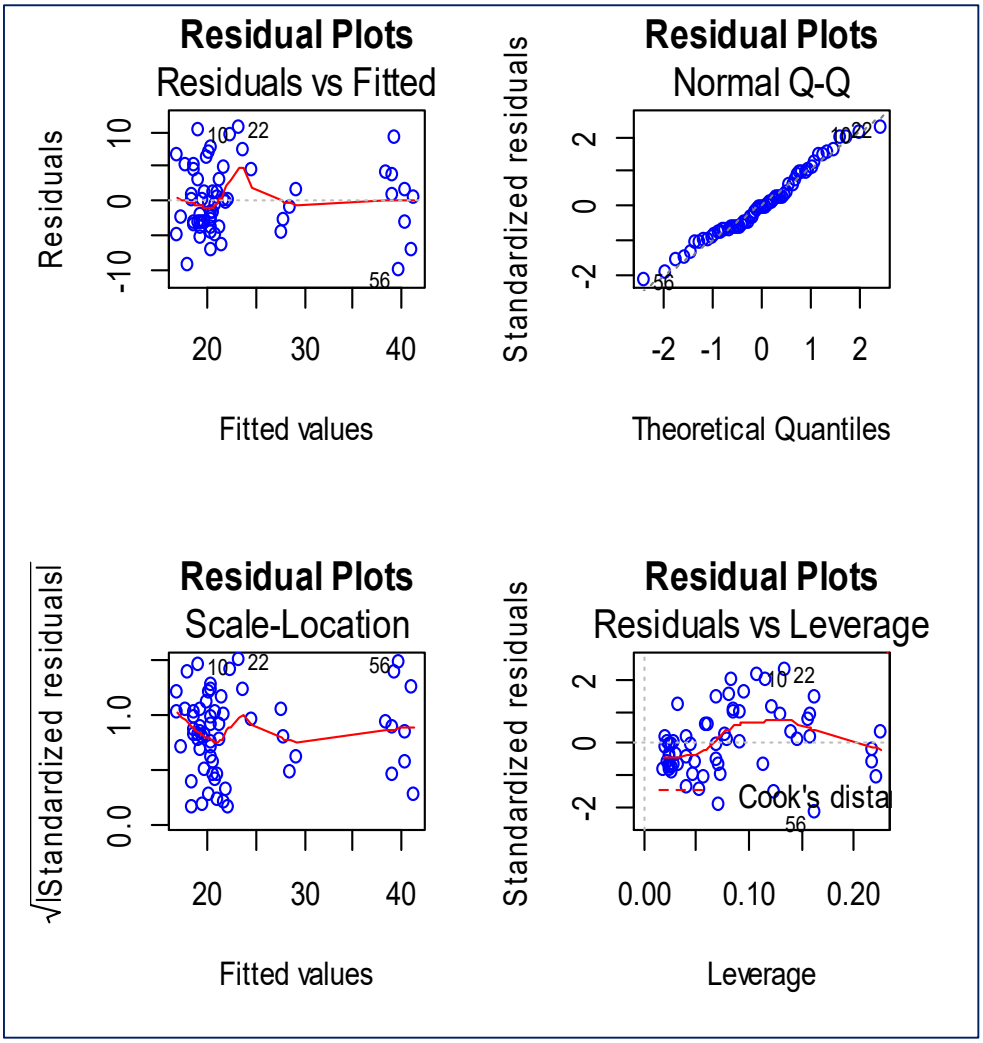
31	4.30	16.300	19.419	0.759	-3.119	-0.63
32	4.30	15.800	19.115	0.805	-3.315	-0.67
33	4.30	15.400	18.683	0.909	-3.283	-0.67
34	5.60	19.200	18.525	1.406	0.675	0.14
35	5.60	8.400	17.871	1.331	-9.471	-1.96
36	5.60	15.000	17.362	1.326	-2.362	-0.49
37	5.60	12.000	16.971	1.355	-4.971	-1.03
38	5.60	21.900	21.711	0.865	0.189	0.04
39	5.60	21.300	21.057	0.742	0.243	0.05
40	5.60	21.600	20.549	0.737	1.051	0.21
41	5.60	19.800	20.157	0.790	-0.357	-0.07
42	4.30	21.600	22.052	0.835	-0.452	-0.09
43	4.30	17.300	21.328	0.701	-4.028	-0.81
44	4.30	20.000	20.784	0.713	-0.784	-0.16
45	4.30	18.600	20.393	0.781	-1.793	-0.36
46	2.40	22.100	22.224	1.064	-0.124	-0.03
47	2.40	14.700	21.394	1.027	-6.694	-1.36
48	2.40	15.800	20.856	1.089	-5.056	-1.03
49	2.40	13.200	20.500	1.163	-7.300	-1.50
50	5.60	30.800	29.146	2.377	1.654	0.37
51	5.60	27.500	28.492	2.338	-0.992	-0.22
52	5.60	25.200	27.983	2.339	-2.783	-0.63
53	5.60	22.800	27.592	2.358	-4.792	-1.08
54	2.14	41.700	41.347	1.919	0.353	0.08
55	4.14	33.700	41.110	1.771	-7.410	-1.58
56	8.15	29.700	39.696	2.025	-9.996	-2.18R
57	2.14	41.800	40.376	1.874	1.424	0.31
58	4.14	37.100	40.344	1.688	-3.244	-0.69
59	8.15	40.100	39.182	1.992	0.918	0.20
60	2.14	42.700	39.161	1.988	3.539	0.77
61	4.14	48.600	39.374	1.704	9.226	1.96
62	8.15	42.400	38.404	2.003	3.996	0.87

R denotes an observation with a large standardized residual.

Durbin-Watson statistic = 0.772251

Residual Plots for Y





Interpretation

- The residuals versus fitted plot , it is observed that the residuals are randomly distributed.
- From the normality plot, the normality of the residuals are also verified.
- From the bottom left plot of histogram in the minitab output, the assumption of normality is verified here.
- From the order versus residuals plot we can say that the residuals may be independent.
- In the Scale-Location plot we can see that the residuals are linearly distributed and least affected by change in scale and location.
- In residual versus leverage plot the leverage points are seen.

VARIABLE SELECTION AND MODEL BUILDING

In the regression analysis that only correct and important explanatory variables appears in the model. While choosing a subset of explanatory variables, there are possible options: In order to make the model as realistic as possible, the analyst may include as many as possible explanatory variables. On the other hand, when small number of variables are included then the predictive variance of y decreases. Also, when the observations on more number are to be collected, then it involves more cost, time, etc. A compromise between these consequences is striked to select the "best regression equation".

1.Forward Selection

In this methodology it is assumed that there is no explanatory variable in the model except an intercept term. Here the variables are added one by one and the fitted model is tested at each step using a suitable criterion.

Stepwise Regression: y versus x1, x2, x3, x4

Forward selection. Alpha-to-Enter: 0.25

Response is y on 4 predictors, with N = 62

Step	1	2	3	4
Constant	19.454	7.190	4.641	5.895
x2	0.182	0.185	0.183	0.183
T-Value	9.72	10.52	10.56	10.63
P-Value	0.000	0.000	0.000	0.000
x3		35	35	35
T-Value		3.10	3.10	3.19
P-Value		0.003	0.003	0.002
x4			4.6	5.8
T-Value			1.64	2.01
P-Value			0.107	0.049
x1				-0.48
T-Value				-1.41
P-Value				0.165
S	5.48	5.13	5.06	5.01
R-Sq	61.15	66.59	68.07	69.14
R-Sq(adj)	60.51	65.46	66.42	66.97
Mallows C-p	13.7	5.7	5.0	5.0
PRESS	1929.38	1711.10	1750.52	1762.86
R-Sq(pred)	58.44	63.14	62.29	62.03

Interpretation

By using forward selection, the final model obtained is

$$Y = 4.641 + 0.183\,x2 + 35\,x3 + 4.6\,x4$$

It required 3 steps to obtain this model.

We can conclude that if we carry out our regression analysis using there 3 variables instead of using all of 4, we can get better results.

2. Backward Elimination

The backward elimination methodology begins with all the explanatory variables and keep on deleting on variable at a time until a suitable model is obtained.

Stepwise Regression: y versus x1, x2, x3, x4

Backward elimination. Alpha-to-Remove: 0.1

Response is y on 4 predictors, with N = 62

Step	1	2	3
Constant	5.895	4.641	7.190
x1	-0.48		
T-Value	-1.41		
P-Value	0.165		
x2	0.183	0.183	0.185
T-Value	10.63	10.56	10.52
P-Value	0.000	0.000	0.000
x3	35	35	35
T-Value	3.19	3.10	3.10
P-Value	0.002	0.003	0.003
x4	5.8	4.6	
T-Value	2.01	1.64	
P-Value	0.049	0.107	
S	5.01	5.06	5.13
R-Sq	69.14	68.07	66.59
R-Sq(adj)	66.97	66.42	65.46
Mallows C-p	5.0	5.0	5.7
PRESS	1762.86	1750.52	1711.10
R-Sq(pred)	62.03	62.29	63.14

Interpretation

By using backward elimination, the final model obtained is

$$Y = 4.641 + 0.183\,x2 + 35\,x3 + 4.6\,x4$$

It required 3 steps to obtain this model.

We can conclude that if we carry out our regression analysis using there 3 variables instead of using all of 4, we can get better results.

3.Stepwise Selection

A combination of forward selection and backward elimination procedure is the stepwise regression.

Stepwise Regression: y versus x1, x2, x3, x4

Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is y on 4 predictors, with N = 62

Step	1	2	3
Constant	19.454	7.190	4.641
x2	0.182	0.185	0.183
T-Value	9.72	10.52	10.56
P-Value	0.000	0.000	0.000
x3		35	35
T-Value		3.10	3.10
P-Value		0.003	0.003
x4			4.6
T-Value			1.64
P-Value			0.107
S	5.48	5.13	5.06
R-Sq	61.15	66.59	68.07
R-Sq(adj)	60.51	65.46	66.42
Mallows C-p	13.7	5.7	5.0
PRESS	1929.38	1711.10	1750.52
R-Sq(pred)	58.44	63.14	62.29

By using stepwise regression, the final model obtained is

$$Y = 4.641 + 0.183\,x2 + 35\,x3 + 4.6\,x4$$

It required 3 steps to obtain this model.
We can conclude that if we carry out our regression analysis using there 3 variables instead of using all of 4, we can get better results.

Best Subsets Regression:

Using best subsets regression we compare different regression models that contain subsets of the predictors of our dataset. Best Subsets regression is an efficient way to identify model that adequately fits our data with as few predictors as possible.

Best Subsets Regression: y versus x1, x2, x3, x4

Response is y

Vars	R-Sq	R-Sq(adj)	Mallows C-p	S	x 1	x 2	x 3	x 4
1	61.2	60.5	13.7	5.4825		X		
1	3.9	2.3	119.4	8.6213			X	
1	2.9	1.2	121.4	8.6698				X
1	0.2	0.0	126.3	8.7872	X			
2	66.6	65.5	5.7	5.1273		X	X	
2	62.8	61.5	12.7	5.4113		X		X
2	61.4	60.1	15.3	5.5129	X	X		
2	6.6	3.5	116.4	8.5712			X	X
2	4.3	1.0	120.8	8.6795	X		X	
3	68.1	66.4	5.0	5.0556		X	X	X
3	67.0	65.2	7.0	5.1431	X	X	X	
3	63.6	61.7	13.2	5.3956	X	X		X
3	7.9	3.2	116.1	8.5850	X		X	X
4	69.1	67.0	5.0	5.0136	X	X	X	X

Interpretation

- The best fitting models have the highest R^2 values.
- The most appropriate model has low Mallow's Cp
- By using best subsets regression, the final model obtained will have predictors : x2, x3, x4 with a mallows cp of 5.0 which is lowest and corresponding higher r square value.

Multicollinearity

- If r is close to 0, then multicollinearity does not harm and it is termed as non-harmful multicollinearity .
- If r is close to +1 or -1 then multicollinearity inflates the variance and it rises terribly. This is termed as harmful multicollinearity.

Multicollinearity Diagnostics

1. Using determinant of X'X :

If Rank(X' X) < k then |X'X| will be singular and so |X'X|=0. So as |X'X|→ 0, the degree of multicollinearity increases and it becomes exact or perfect at |X'X|=0.

We will determine it using R

```
> x=matrix(c(rep(1,n),d[,1],d[,2],d[,3],d[,4]),n,p)
> c=x%*%t(x);qr(c)$rank;det(c)
[1] 5
[1] 0
```

Thus |X'X| serves as a measure of multicollinearity and |X'X|=0 indicates that perfect multicollinearity exists.

2. Using Determinant of correlation matrix:

```
> c=cor(data);c
           x1          x2          x3          x4          y
x1  1.000000000  0.002288954  0.05562132 0.31237223 -0.04546405
x2  0.002288954  1.000000000 -0.04403982 0.05288934  0.78200081
x3  0.055621320 -0.044039817  1.00000000 0.02445348  0.19850582
x4  0.312372229  0.052889337  0.02445348 1.00000000  0.16899408
y  -0.045464054  0.782000806  0.19850582 0.16899408  1.00000000
> D=det(c);D
[1] 0.27624
```

Let D be the determinant of correlation matrix then $0 \leq D \leq 1$.
Since D is close to 0 this is an indication of high degree of multicollinearity.
Any value of D between 0 and 1 gives an idea of the degree of multicollinearity.

Diagnostic for Leverage

```
> n=62;p=5;d=as.matrix(data)
> x=matrix(c(rep(1,n),d[,1],d[,2],d[,3],d[,4]),n,p)
> h=x%*%solve(t(x)%*%x)%*%t(x)
> for(i in 1:n)
+ {
+ print(h[i,i])
+ }
[1] 0.1303549
[1] 0.08069374
[1] 0.0919687
[1] 0.06899521
[1] 0.03282139
[1] 0.08635833
[1] 0.09083095
[1] 0.05905534
[1] 0.1224483
[1] 0.1080703
```

```
[1] 0.08599569
[1] 0.162973
[1] 0.06045242
[1] 0.04067902
[1] 0.07849929
[1] 0.04970293
[1] 0.0259764
[1] 0.07090045
[1] 0.05760169
[1] 0.02414913
[1] 0.06945631
[1] 0.1339128
[1] 0.0841644
[1] 0.09526436
[1] 0.04128667
[1] 0.0287917
[1] 0.02481751
[1] 0.02631409
[1] 0.03160728
[1] 0.02553778
[1] 0.0229289
[1] 0.02574848
[1] 0.03290158
[1] 0.07859805
[1] 0.07044101
[1] 0.0699363
[1] 0.07302304
[1] 0.02979851
[1] 0.02189836
[1] 0.02159321
[1] 0.02483363
[1] 0.02774825
[1] 0.01955599
[1] 0.02020939
```

```
[1] 0.02429096
[1] 0.04506133
[1] 0.04197198
[1] 0.04722174
[1] 0.0538462
[1] 0.2247122
[1] 0.2174114
[1] 0.2175719
[1] 0.2211709
[1] 0.1465219
[1] 0.1248072
[1] 0.1631955
[1] 0.1396462
[1] 0.1134127
[1] 0.1578294
[1] 0.1572337
[1] 0.1155803
[1] 0.1596198
> cof=2*(p/n);cof
[1] 0.1612903
> for(i in 1:n)
+ {
+ if(h[i,i]>cof)
+ {
+ print(paste("Leverage Point:",i,h[i,i]))
+ }
+ }
[1] "Leverage Point: 12 0.162972958975964"
[1] "Leverage Point: 50 0.22471215540724"
[1] "Leverage Point: 51 0.217411431666014"
[1] "Leverage Point: 52 0.217571897388476"
[1] "Leverage Point: 53 0.221170901778092"
[1] "Leverage Point: 56 0.163195463965471"
```

Interpretation

- The location of observations in x -space can play an important role in determining the regression coefficients. Influential points are “bad” values, they should be eliminated from the sample .
- The hat matrix

$$\mathbf{H = X(X^IX)^{-1}X^I}$$

plays an important role is identifying the influential observations or the leverage points which is displayed in R Console.

The $\mathbf{i^{th}}$ diagonal element of **H** is

$$\mathbf{h_{ii} = x_i^I(X^IX)^{-1}x_i}$$

where x_i is the i^{th} row of X-matrix.

The **leverage**, h_{ii}, quantifies the influence that the observed response y_i has on its predicted value . That is, if h_{ii} is small, then the observed response y_i plays only a small role in the value of the predicted response .

On the other hand, if h_{ii} is large, then the observed response y_i plays a large role in the value of the predicted response. It's for this reason that the h_{ii} are called the "**leverages**."

A common rule is to flag any observation whose leverage value,

$$\mathbf{h_{ii} > 2(k+1)/n}$$

⇒ The point is leverage point.

The leverage depends only on the predictor values.

The leverage points are shown in the data above in the r console output .

REGRESSION ANALYSIS

- Out of the models fitted , multiple linear regression model fitted to the pressure drop data using x2, x3, and x4 was most appropriate.
- Multiple linear regression model fitted to the pressure drop data was given by

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$$Y = 5.89453 - 0.4779\ x1 + 0.18271\ x2 + 35.40284\ x3 + 5.84391\ x4$$

- From the quadratic and cubic polynomials , these models were best-fitted while carrying out the regression analysis with Y and .
- This can be checked by the prediction and confidence bands or the values of R-Sq.
- In different model building stratagies, i.e. forward selection, backward elimination , stepwise regression and best subsets regression, we obtained the same model which may be the most adequate to carry out the further regression analysis.
- Using the criterias specified in the methodologies above, variables selected and found to be important are X_2, X_3 and X_4.
- Observations 12, 50, 51, 52, 53, and 56 were found out to be leverage points
- All procedures leads to different models.
- But the model with better R-Sq is more adequate and appropriate generally.

References

- Regression Analysis with R 1st Edition
- Regression Analysis – Google Books
- Draper, N. R. and Smith H. (1998) Applied regression analysis 3rd edition(John Wiley)
- Applied Regression Analysis(Wiley Series in Probability and Statistics)
- Applied Regression Analysis: A Research Tool, Second Edition
- R Documentation : library
- Regression Analysis: Moving on with Minitab.
- Introduction to Linear Regression Analysis by Example
- Sanford Weisberg, Applied Linear Regression(Wiley Series in Probability and Statistics)
- John Fox and Sanford Weisberg(2010) An R Companion to applied Regression, 2nd Edition.