Applied Regression Model Assignment 1

Overview:

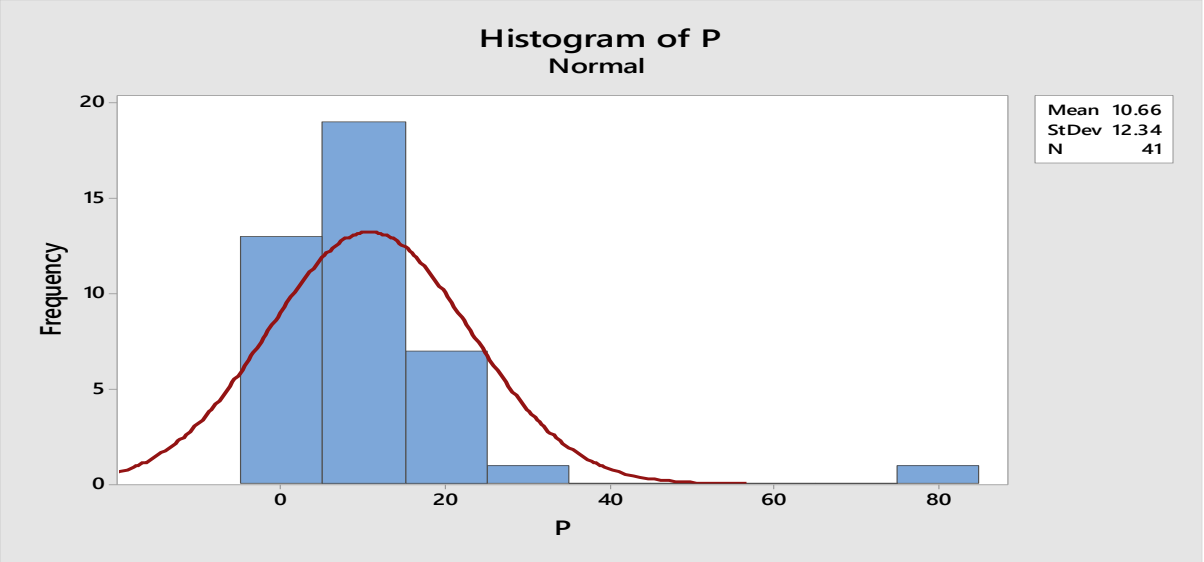
Student ID: 17230755

This example relates to the percentage of expenditure, P, a publishing house spends on advertising and the change in revenue, R (expressed as a percentage) at the end of the following year. We are interested in prediction R based on P.

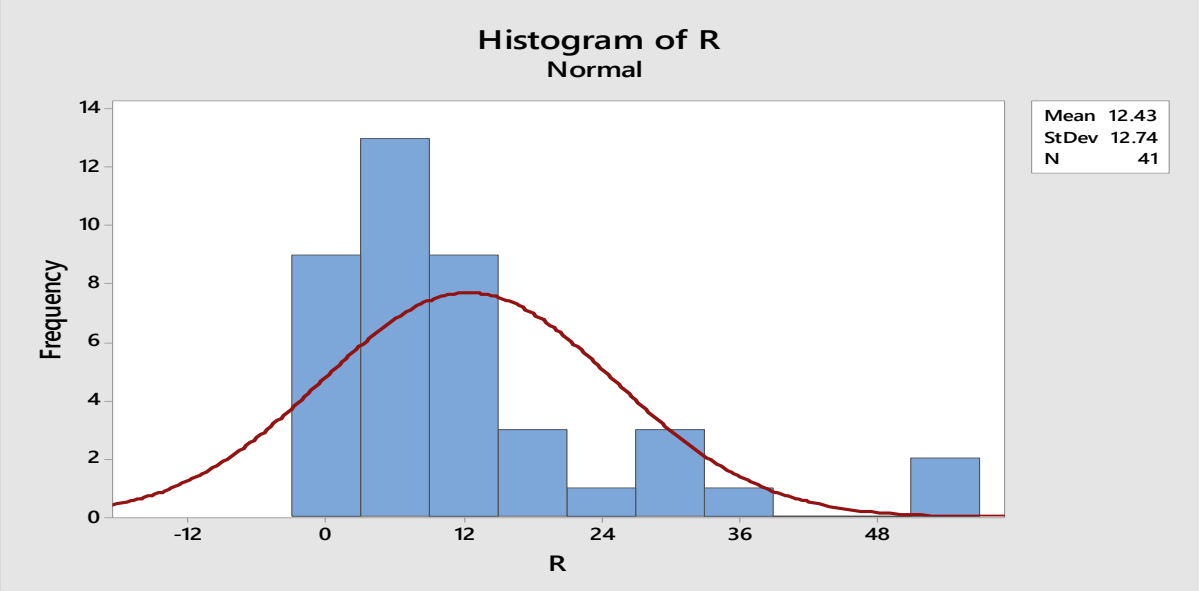
Step 1:

In this case Response Variable is R (dependent Variable) and the predictor is P (Independent Variable). First let us explore the given data by plotting the distribution of R, P and plotting a scatter plot.

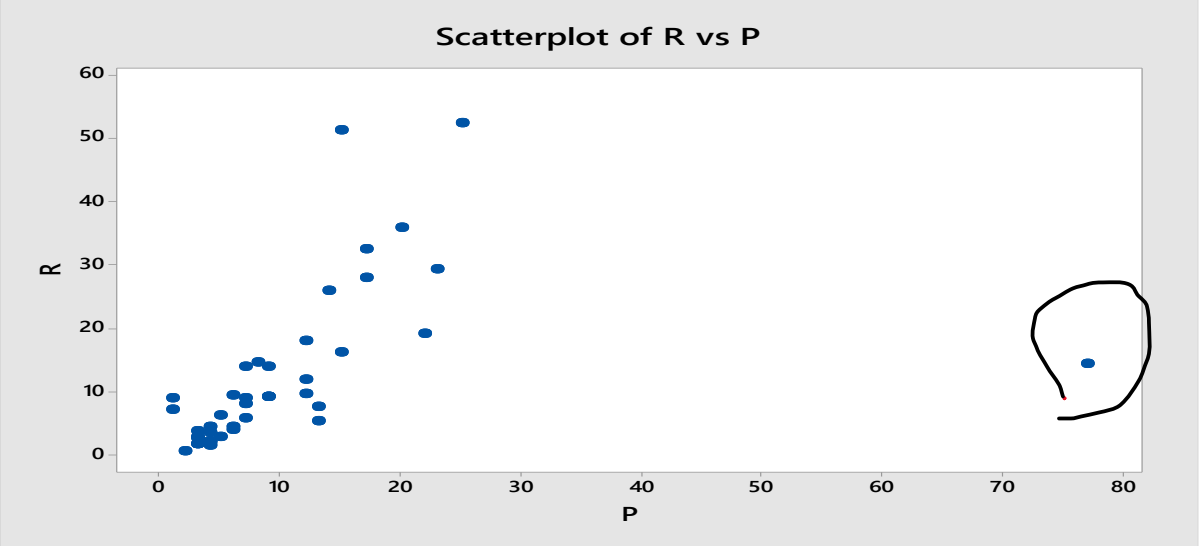
- **Distribution of P (Predictor):** From below plot we can say that the variable P is approximately normally distributed data with **Mean = 10.266** and **SD = 12.34.**



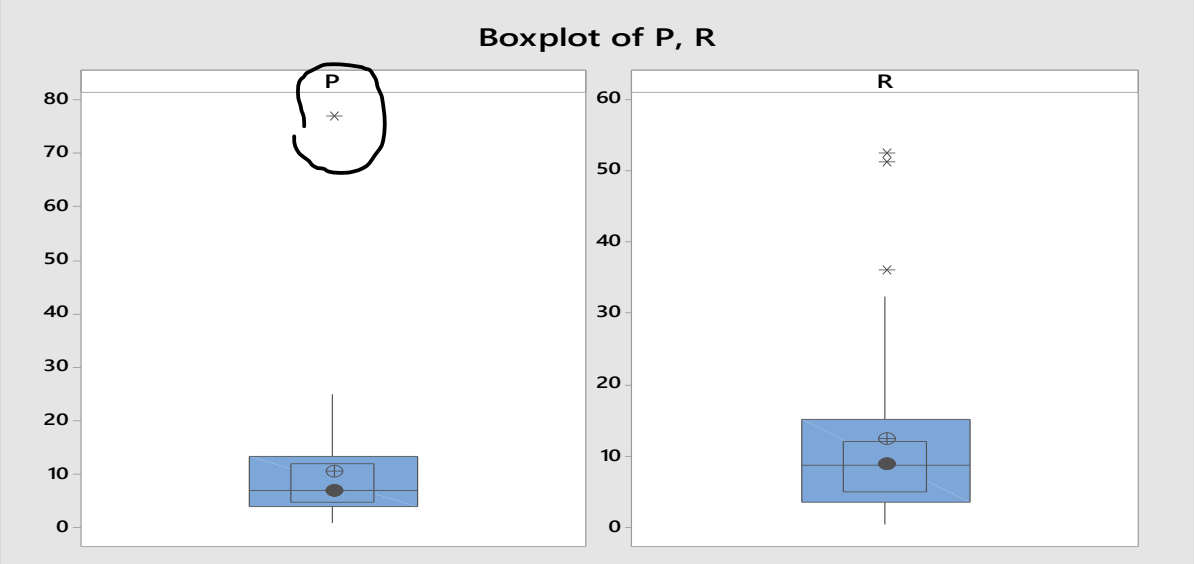
- **Distribution of Change in Revenue R (Response):** From below plot we can say that the Reponse R is approximately normally distributed data with **Mean = 12.43 and SD = 12.74.**



<u>Scatter Plot</u>: From below scatter plot we can say increase in P will result in increase in R. (However, some outliers can be observed from the below plot). We will again plot the box plot to see the outliers in P. Suspected Outlier is marked in below plot.



Box Plot: Below is the box plot representing Median (middle line), Maximum, Minimum, Mean and 95% Confidence Interval. Outlier in P is marked below. This point corresponds to **True Story, P = 77, R = 14.3 (Outlier in predictor i.e. P).**



Step 2: Fitting a Regression Line.

We know that regression in the form: **response = systematic + random**

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \qquad i = 1, \ldots, n$$

Regression Analysis: R versus P

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	1249.1	1249.14	9.28	0.004
P	1	1249.1	1249.14	9.28	0.004
Error	39	5247.1	134.54		
Lack-of-Fit	17	4497.8	264.58	7.77	0.000
Pure Error	22	749.3	34.06		
Total	40	6496.3			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
11.5992	19.23%	17.16%	0.00%

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	7.60	2.41	3.16	0.003	
P	0.453	0.149	3.05	0.004	1.00

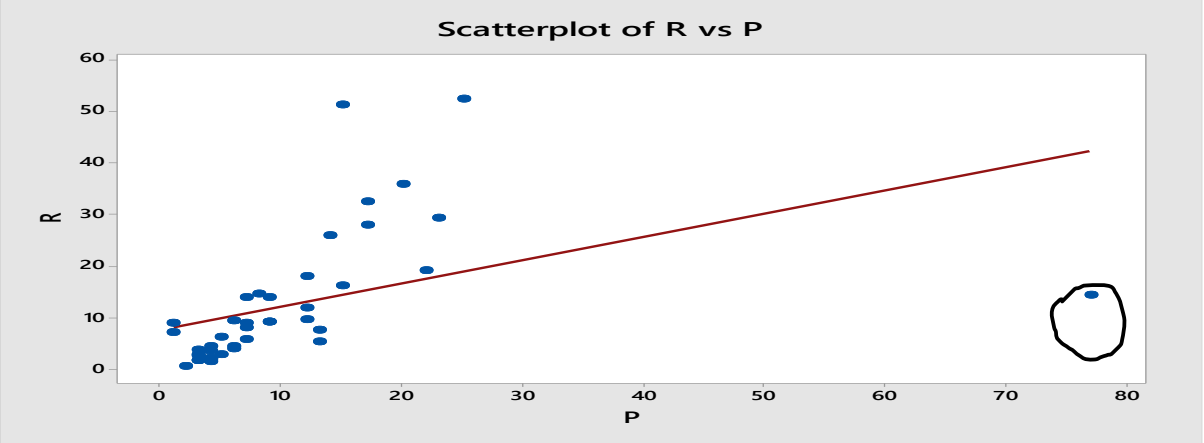
Regression Equation

R = 7.60 + 0.453 P

Observation 1:

From the output of minitab, **R-sq = 17.16%** value is too low to use this model for prediction. Furthermore, standard error is considerably high i.e. **s = 11.5992.** Hence this model is not suitable to use for prediction.

Fitted Regression Line is shown below.



Observation 2:

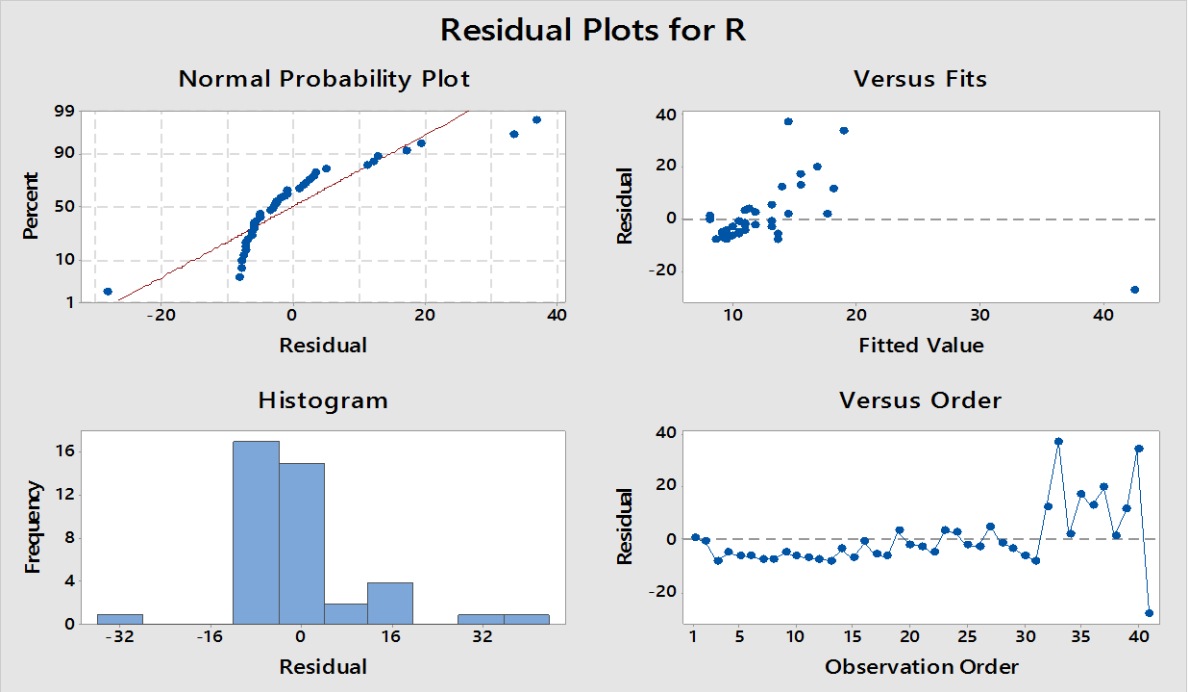
From above plot because of the **outlier** fitted line is away from densely populated approximately linear plots. Hence, the fitted regression line is not suitable for prediction. Because of this **R-sq** value is **less** and **Standard Error** is more.

Four in one graph is show below for the fitted regression line.

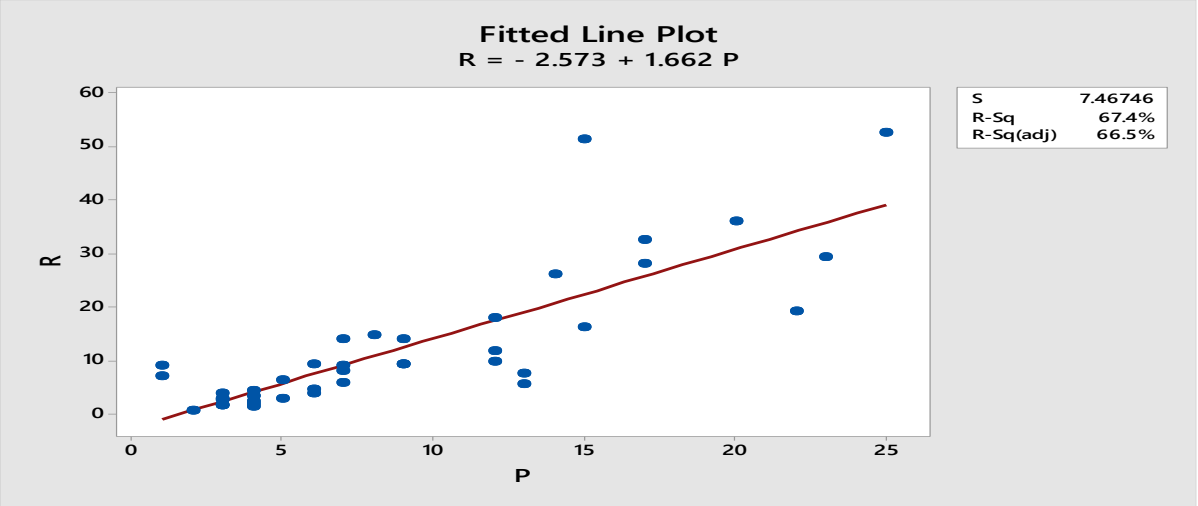
Observation 3*:*

From 1st graph (percent vs Residual of R) Residual is not exactly normal distribution **(if residual is in normal distribution then, points should be along straight red line). Normal distribution is one the key assumption in simple linear regression model**. In addition, in graph 2 points should be distributed approximately along the straight dotted line (Which ensures residual is minimum). Moreover, from graph 4 residual is more in 40th observation (possibly outlier).

Note: It is possible to make response Y approximately normal distributed by doing Box – Cox Transformation, which will be explained in further steps.



Step 3: **Removing Outlier which is Tue Story, P = 77 and R = 14.3 and fitting the regression line.**



Regression Analysis: R versus P

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	1	4373.7	4373.71	78.43	0.000
P	1	4373.7	4373.71	78.43	0.000
Error	**38**	2119.0	55.76		
Lack-of-Fit	16	1369.7	85.60	2.51	0.023
Pure Error	22	749.3	34.06		
Total	39	6492.7			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
7.46746	67.36%	66.50%	62.00%

Coefficients

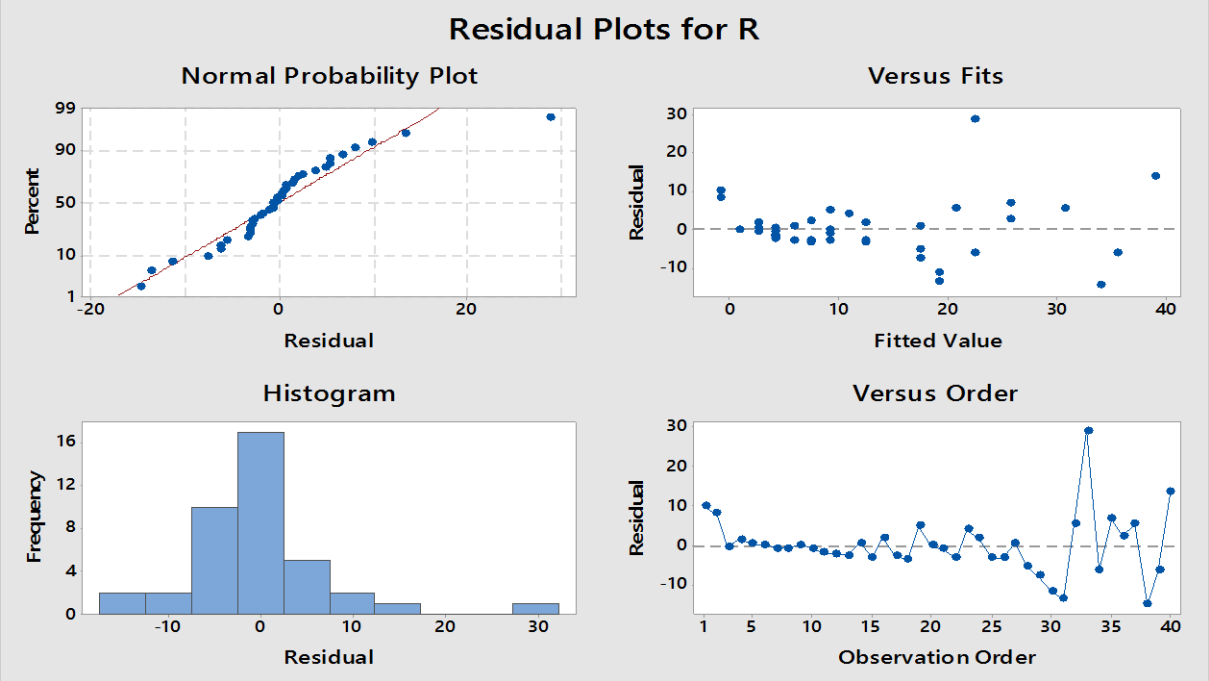
Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	-2.57	2.06	-1.25	0.219	
P	1.662	0.188	8.86	0.000	1.00

Regression Equation ------- New Regression Equation

R = -2.57 + 1.662 P

Observation 1: From the above **Regression Analysis: R versus P** we can observe that **R-sq value** increased to **67.36%** and **R-sq(adj) = 66.50%, R-sq(pred)= 62.00%. So better than previous model.**

Observation 2: From 2nd graph we can observe that residual is near to fitted value and from 1st graph residual is normally distributed (Approximately).



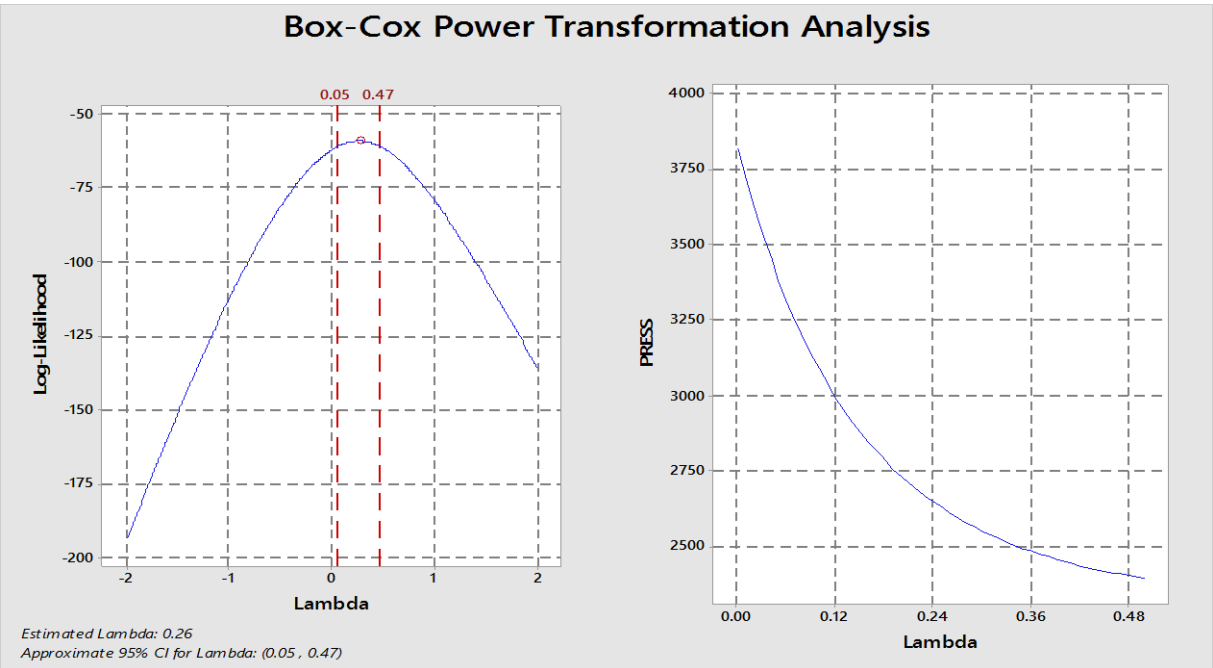
Observation 3: Error term does not have any pattern.

Step 4: Box Cox Transformation to determine best trans formation of Y response (R).

Macros used: %bctrans17 Y X

A Box Cox transformation is a way to transform non-normal dependent variables into a normal shape. Normality is an important assumption for many statistical techniques; if your data isn't normal, applying a Box-Cox means that you are able to run a broader number of tests. ***[1]***

To make distribution of response as normal as possible.



Lambda: Lambda value indicates the power to which all data should be raised to response normally distributed. In Box Cox Transformation Lambda searches from -5 to +5.

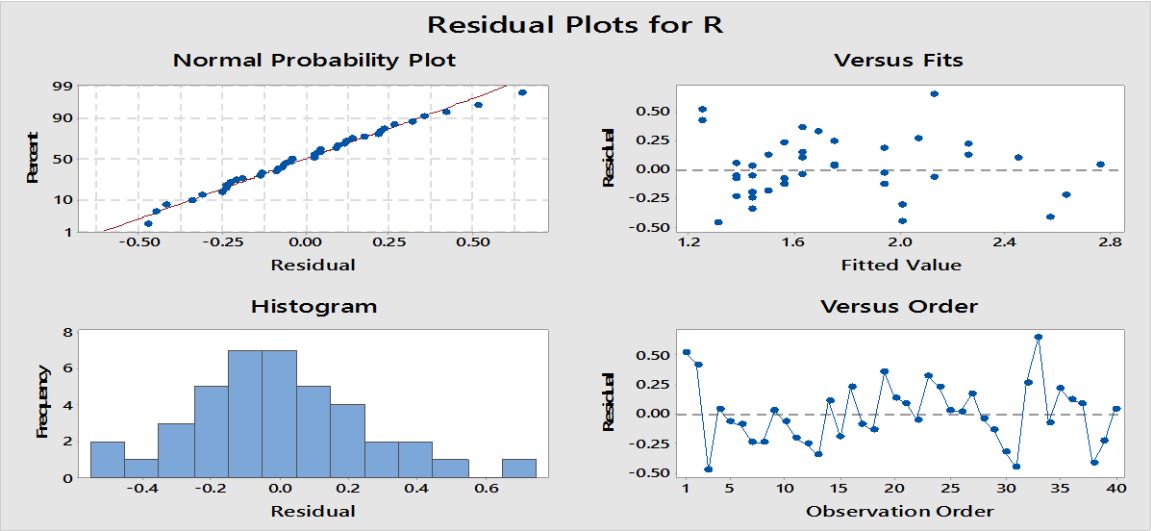
- **Peak Lambda = 0.26**
- **95% CI for Lambda is [0.06,0.47]**

Note: Lambda or Box Cox transformation is based on Maximum Likehood Estimation.

Step 5: Fitting Regression line with lambda = 0.26 (Obtained from step 4).

```
Method

Box-Cox transformation  λ = 0.26

Analysis of Variance for Transformed Response

Source         DF  Adj SS   Adj MS  F-Value  P-Value
Regression      1  6.3064  6.30643    91.10    0.000
  P             1  6.3064  6.30643    91.10    0.000
Error          38  2.6306  0.06923
  Lack-of-Fit  16  1.9146  0.11966     3.68    0.003
  Pure Error   22  0.7161  0.03255
Total          39  8.9371

Model Summary for Transformed Response

       S    R-sq  R-sq(adj)  R-sq(pred)
0.263111  70.56%     69.79%      67.02%

Coefficients for Transformed Response

Term         Coef  SE Coef  T-Value  P-Value   VIF
Constant   1.1834   0.0726    16.30    0.000
P         0.06310  0.00661     9.54    0.000  1.00

Regression Equation

R^0.26 = 1.1834 + 0.06310 P

Fits and Diagnostics for Unusual Observations

Original Response

Obs       R     Fit
  1   8.900   2.334
 33  51.200  18.320
 40  52.500  49.698

Transformed Response
                              Std
Obs      R'     Fit   Resid  Resid
  1  1.7654  1.2465  0.5189   2.04  R
 33  2.7823  2.1299  0.6524   2.54  R
 40  2.8005  2.7609  0.0397   0.17     X

R' = transformed response
R  Large residual
X  Unusual X
```

Four in plot of box cox transformation:



Observation 1: From above Graph 1 (Residual vs Percent) points are along straight line which indicates after box cox transformation residuals is more normal than before. Furthermore, there is no unusual patterns observed.

Step 6:

Observation 1: **R-sq. increased to 70.56%** after box cox transformation.

Observation 2: Pearson Correlation P value = 0.00 **(P),** which indicates there is a linear relation between P and R.

Step 7:

(a). **Two tailed T significance test:**
Given: alpha = 0.05. That means Confidence Interval = 100(1-alpha) % = 100(1-0.05) = 95%
DF = Degree of Freedom for T test = n-2 = 38
Observed from output **T_{obs}** for P = 9.54

(b). **F significance test:**
Given: **alpha** = 0.05. That means Confidence Interval = 100(1-alpha) % = 100(1-0.05) = 95%
DF Numerator = Degree of Freedom for F Test Numerator = 1
DF Denominator = n - 2 = 40 - 2 = 38
Observed from output **F_{obs}** for (P)= 91.10

(c). Verification = T_{obs}^2 = F_{obs}

= 9.54^2 = 91.10 = F_{obs}

i.e. **T_{obs}^2 = F_{obs}**

Step 8: To make prediction as to the average and actual levels of % Revenue when the level of Expenditure is 15%.

Prediction for R

```
Regression Equation

R^0.26 = 1.1834 + 0.06310 P


Variable  Setting
P              15


    Fit        95% CI              95% PI
18.3199  (14.7598, 22.4790)  (5.87528, 44.0171)
```

Conclusion: For a given P value the response Y (R) is 18.3199 . However, with 95% confidence we can say for given value P = 15 the response R may be anywhere between (14.758 to 22.479).

Reference:

[1] Box Cox Transformation: http://www.statisticshowto.com/box-cox-transformation/