204: Regression Analysis

Instructor: Prof. Sugata Sen Roy

Report Card
Data Analysis with Regression

Transformation of Variable

Submitted by: Soumarya Basak

Submitted on: June 22, 2022

University of Calcutta

Department of Statistics

Contents

Problem 1

1.1 About the Practical

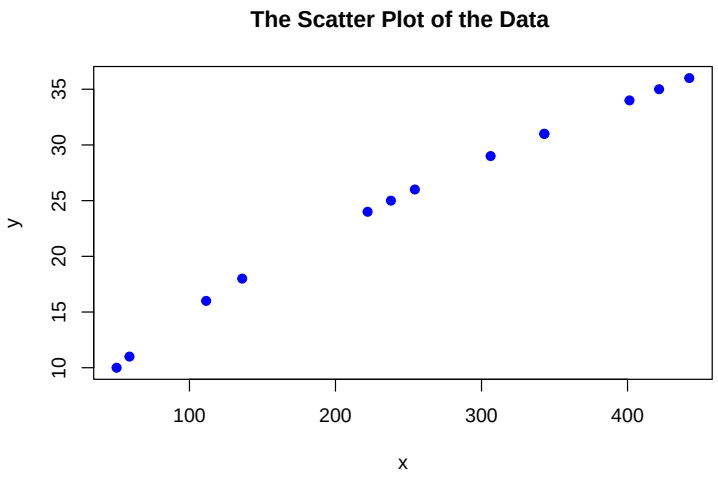
We have a data consist of 2 variables, firstly we need to plot the data and fit a linear regression. Then if you are not satisfied with the fit we need to use a Box-Tidwell transformation to see whether the fit improve or not.

1.2 Analysis

Note that, y variable and X variable is given in the data set so we will use them accordingly to fit models on the data set.

Scatter Plot

From the scatter plot it is clear that, the y variable and the x variable is very nearly linearly related. So a linear regression plot will be a good choice here.
If we notice more carefully the plot will be look like a parabolic arc. So a *square root transformation* maybe give a better fit but it is very risky to say at this early stage without fitting the models.



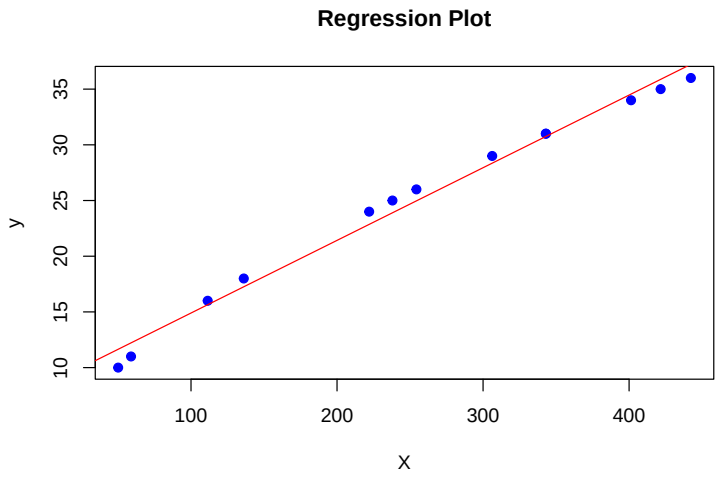
1.2.1 Fitting Regression Model

It is better to see linear regression first. So considering y as our response variable and X as our regressor variable, we fit a linear regression on the data.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	8.376417	0.629721	13.30	4.01e-08
$log(L)$	0.065227	0.002194	29.73	7.35e-12

Residual standard error: 1.026 on 11 degrees of freedom
Multiple R-squared: 0.9877
Adjusted R-squared: 0.9866
F-statistic: 884 on 1 and 11 DF
p-value: 7.35e-12



Observation and Conclusion

The above fit is significant with all the significant coefficient and intercept and explains nearly 98.6% of variation of the response variable.
So we can consider it as a good model.

Now as we assume the plot looks like parabolic arc, next we'll fit regression model on the square root transformed regressor variable.

1.2.2 Box-Tidwell Transformation

Box Tidwell transformation is necessarily used to make a nonlinear relationship to a linear relationship.
Former scatter plot we assume that the scatter plot exhibits parabolic arc. So a **square root** transformation on X variable maybe work here.
Consider the following transformation,

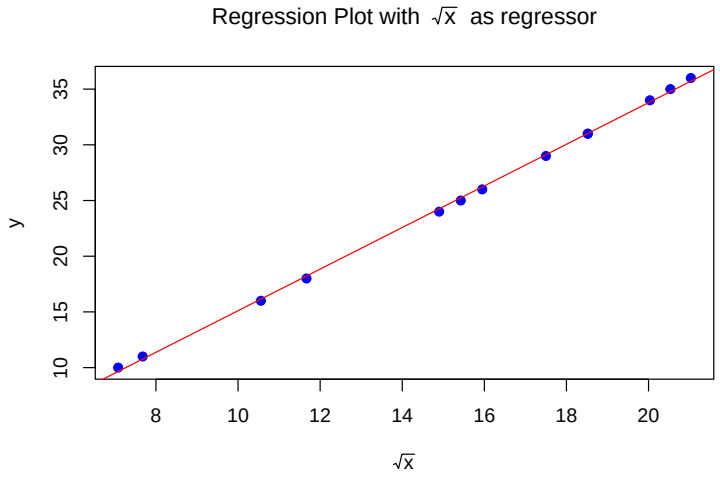
$$x \to x^* = x^{(1/2)}$$

Now regressing y with respect to $x^{1/2}$, we come to the following result.

Output:

Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
Intercept	-3.56727	0.22497	-15.86	6.34e-09
$log(L)$	1.86754	0.01406	132.83	< 2e-16

Residual standard error: 0.2311 on 11 degrees of freedom
Multiple R-squared: 0.9994
Adjusted R-squared: 0.9993
F-statistic: 1.764e+04 on 1 and 11 DF
p-value: < 2.2e-16



Observation and Conclusion

Now this model is also significant with all the significant coefficient and intercept, and in edition the Adjusted R^2 increases to 0.99 which indicates the model explains almost 99% variation of the response variable and also the residual standard error decreases as compared to the previous model with same degrees of freedom. Also note that the scatter plot for the transformed variable shows more linearity than the original one.
So we prefer to work with the transformed regressor.

1.2.3 Final Model

$$y = -3.56727 + 1.86754 \times (\sqrt{X})$$

with *Adjusted R square: 0.99* and *Residual Standard Error:0.2311*

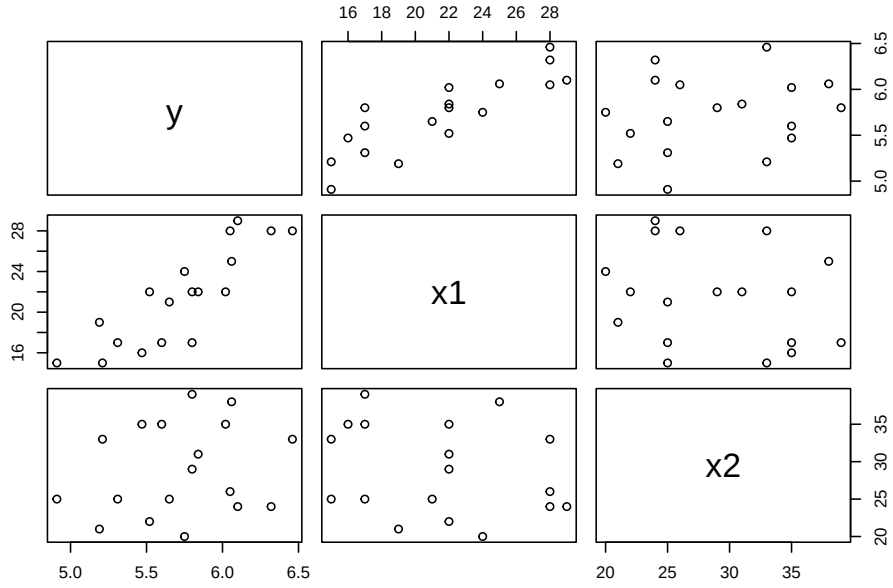
Problem 2

2.1 About the Practical

Our data consist of one response variable and 2 regressor variables X_1 and X_2, next we need to fit a linear regression for the data and suggest a better model after suitable transformation.

2.2 Analysis

As per the given data we use Y as response variable and X_1 , X_2 as 2 regressor variables. As we have 2 regressors, it's better to visualise their relationship with the response variable



The plot shows that though X_1 and y is quite linearly related but X_2 and Y is hardly linearly related, which may affect the model.
Firstly, let fit regression model with these regressors.

Output:

Parameters	***Estimate***	***Std. Error***	$t-value$	$Pr(>\|t\|)$
Intercept	3.044288	0.223767	13.605	7.64e-10
X_1	0.083508	0.006377	13.096	1.30e-09
X_2	0.030664	0.004999	6.135	1.91e-05

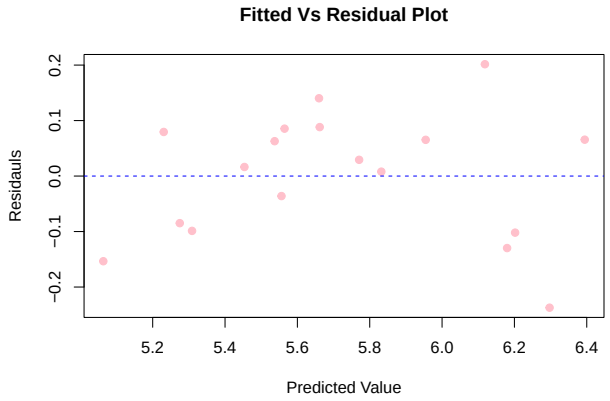
Residual standard error: 0.1213 on 15 degrees of freedom
Multiple R-squared:0.9234
Adjusted R-squared: 0.9132
F-statistic: 90.4 on 2 and 15 DF
p-value: 4.288e-09

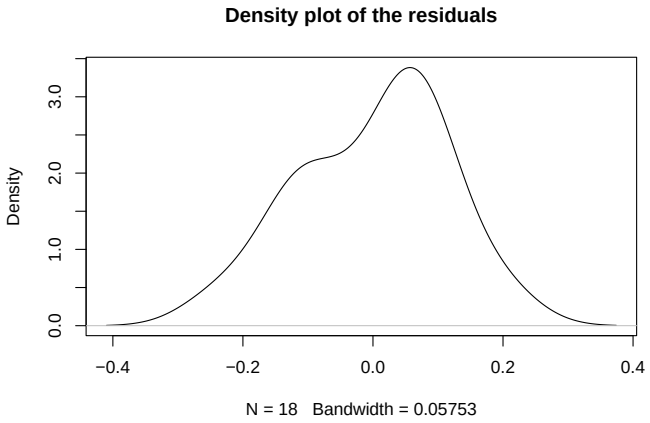
Observation and Conclusion

The above model is significant with all the significant coefficient and intercept. Note that, the adjusted R^2 for the model is 0.91 which indicates the model explains 91% variability of the response variable so we can say that the model is a good model for work purpose.

Next we will study the residuals for the model, weather they follows the assumptions of linear model or not.

2.2.1 Residual Analysis





From the residual plot of the data it is clear that variability of the residuals are not stable and the density plot shows that the residuals are not like *Normally distributed*.

One way to recover this is to take a transformation on response variable as the distributional assumption depends on the distribution of response variable.

2.2.2 Transformation

As we need a transformation on our response variable we consider the following transformation,

$$y \to y^*$$

where,

$$Y^* = \begin{cases} \frac{Y^\lambda}{Y_g} & if \quad \lambda \neq 0 \\ lnY & if \quad \lambda = 0 \end{cases}, \quad Y_g = \text{GM of Y}$$

For different values of λ we will get different transformation and for each of the transformation we will get different models. i.e. for different λ's we have different models.
We will try to find a better model based on our residual sum of squares.

Below we tabulate the residual sum of squares for different models for different λ's

λ	RSS	Adj R^2
-2.0	0.0000012	0.8812271
-1.0	0.0000080	0.8956760
-0.5	0.0000106	0.9015071
0.0	0.0074078	0.9063771
0.5	0.0003080	0.9102693
2.0	0.8507640	0.9159984
3.0	63.7191012	0.9148843

Conclusion

From the tabular it is clear that for λ = -0.5 the model is better then the other, in terms of residual sum of square although for this model the Adjusted R^2 decreases the least and indicates the model explains nearly 90.1% of the response variability.

Working with other values of λ which minimises the RSS, they also decreases the Adjusted R^2 which indicates that their explanation power decreases. So it is better to not to use them.

Now we will fit the model with taking the $\lambda = -0.5$ for the transformation on response.

The Model with $\lambda = -0.5$

Output:

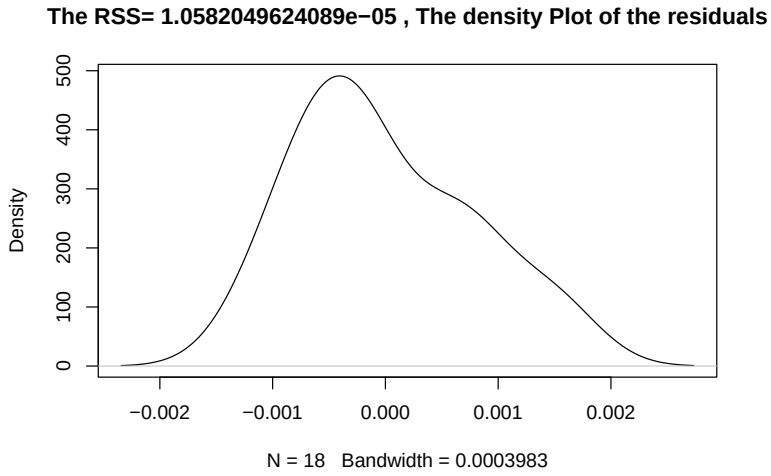
Parameters	***Estimate***	***Std. Error***	$t - value$	$Pr(>\|t\|)$
Intercept	9.072e-02	1.549e-03	58.553	< 2e-16
X_1	-5.387e-04	4.415e-05	-12.200	3.45e-09
X_2	-2.018e-04	3.461e-05	-5.832	3.30e-05

Residual standard error: 0.0008399 on 15 degrees of freedom
Multiple R-squared: 0.9131
Adjusted R-squared: 0.9015
F-statistic: 78.8 on 2 and 15 DF
p-value: 1.104e-08

Conclusion

For the above model the model is significant with all the significant coefficient and intercept at 5% level.
And mainly the residual sum of squared decreases from 0.1213 to 0.0008399 with same

degrees of freedom, without hampering the Adjusted R^2 much. A low RSS indicates that the variability of the residuals decreases of stabilized.



By the above transformation the density of the residuals also approaches to the normal density. So this transformation is good enough to work with, where as the other transformations do not change the density much.

2.2.3 Final Model

$$y^* = (9.072e-02) + (-5.387e-04) \times x_1 + (-2.018e-04) \times x_2$$

where,

$$y^* = \frac{y^{-1/2}}{Y_g}$$

Appendix

The overall analysis is performed in R software, and the necessary output is discussed here. The data set and the codes can be found in the following Github repository.

Source code : https://github.com/SoumaryaBasak/Regression_Analysis_1.git