Statistical Programming :Project

Isaiah Thompson Ocansey

Statistical Programming :Project

Introduction

In this paper, I present one of the parametric power transformations in regression models called Box-Cox Transformation. Box and Cox [1964] proposed the Box-Cox transformation in order to improve statistical models. It has been extensively studied on this subject with the most of the research concentrated on inferences about unknown parameters of interest [Box and Cox, 1964, Bickel and Doksum, 1981, Hinkley and Runger, 1984, Carroll and Ruppert, 1981]. The Box-Cox transformation is a linear transformation of the power transformation which attempts to overcome some violations of generalized linear model assumptions such as non-linearity, heteroscedascity and exogeneity.

Objectives

The objective of this project is to use a function to implement the Box-Cox transformation on the mtcars data set.

Data Description

The data set was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

Descrition of the mtcars Data Set	
Column	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement(cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	weight(1000 lbs)
qsec	1/4 mile time
vs	V/S
am	Transmission
gear	Number of forward gears

Firstly, I performed a simple linear regression on the data set and checked for some assumption violations of the linear model and attempt to overcome those violations using a function to implement the Box-Cox Transformation.

Data Cleaning

We would begin by cleaning the data for analysis

```
## [1] 0
```

It can be observed from the above that the data set has no missing values

Preliminary Analysis / Exploratory Data Analysis

I used exploratory data analysis to understand the dataset.

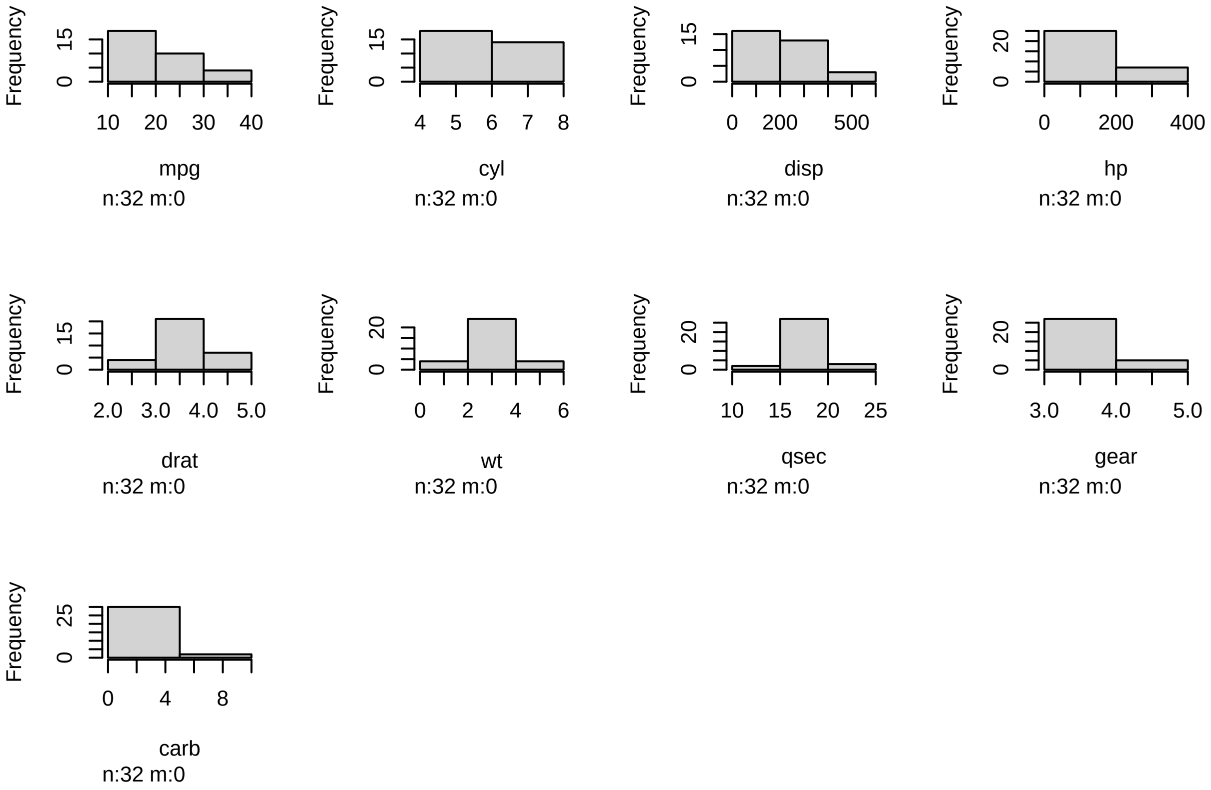
```
## # A tibble: 32 x 11
##      mpg   cyl  disp    hp  drat    wt  qsec    vs    am  gear  carb
##    <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1  21       6  160    110  3.9   2.62  16.5     0     1     4     4
##  2  21       6  160    110  3.9   2.88  17.0     0     1     4     4
##  3  22.8     4  108     93  3.85  2.32  18.6     1     1     4     1
##  4  21.4     6  258    110  3.08  3.22  19.4     1     0     3     1
##  5  18.7     8  360    175  3.15  3.44  17.0     0     0     3     2
##  6  18.1     6  225    105  2.76  3.46  20.2     1     0     3     1
##  7  14.3     8  360    245  3.21  3.57  15.8     0     0     3     4
##  8  24.4     4  147.    62  3.69  3.19  20       1     0     4     2
##  9  22.8     4  141.    95  3.92  3.15  22.9     1     0     4     2
## 10  19.2     6  168.   123  3.92  3.44  18.3     1     0     4     4
## # ... with 22 more rows
## # i Use ‘print(n = ...)‘ to see more rows
```

```
## [1] 32 11
```

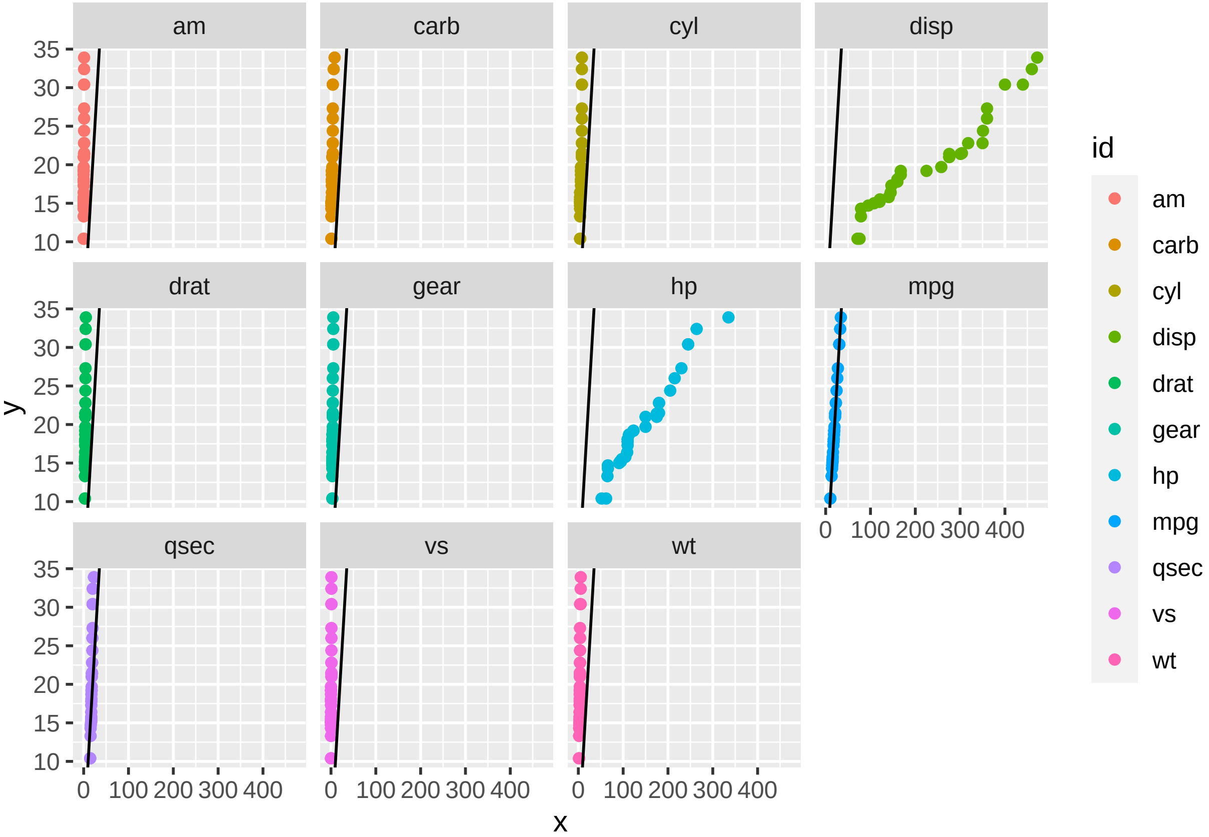
It can be observed that the data has 32 observations and 11 variables. Out of the 11 variables, the mpg variable is going to be the response variable whiles the rest of the remaining variables will be the independent variables.

We now explore the distribution of each variable in the data set

Histogram of all the variables



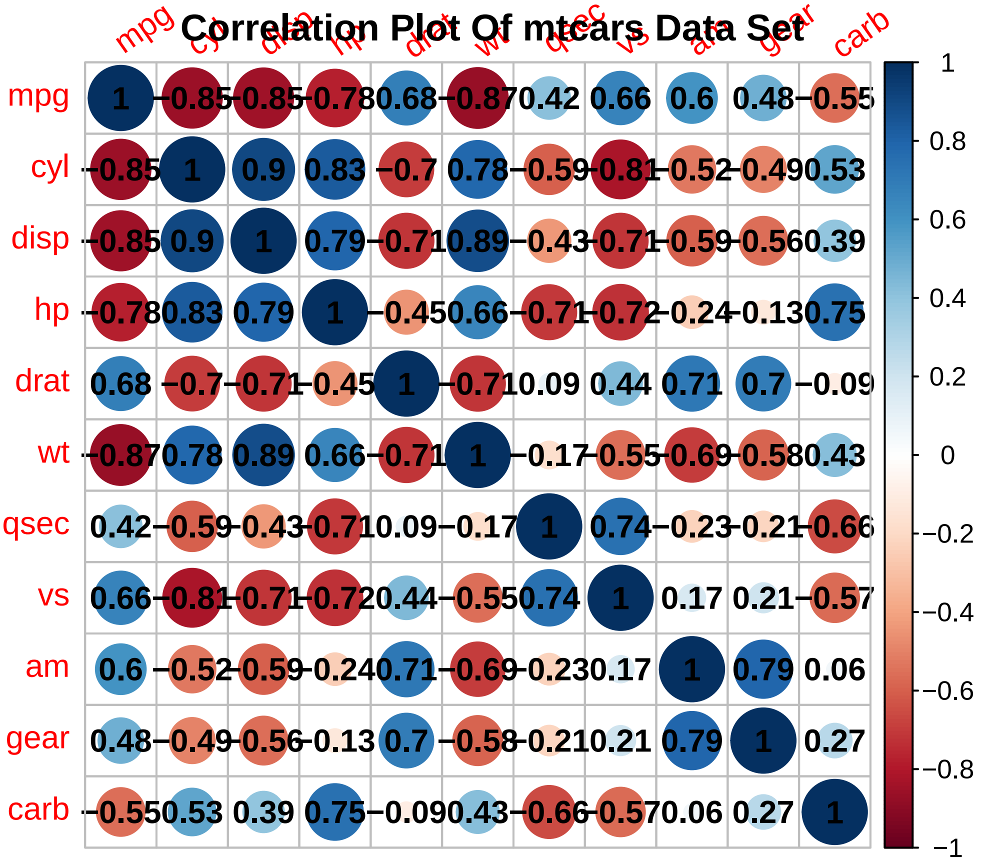
From the above histograms, it can be observed that almost all the histograms of the variables are skewed except drat, wt and qsec



From the plots above it can be observed that the distribution of most of the variables fails the normality test except the response variable,mpg which lies along the straight horizontal line. The code for checking the Shapiro-Wilks test is included in Apendix

I would like to check if there are multi-collinearity within the variables in the data set. multi-Collinearity refers to the situation in which two or more predictor variables are co-linear. The presence of multi-collinearity can pose problems because it can be difficult to separate out the individual effects of co-linear variables on the response variable.

Correlation Plot on mtcars data set


Correlation Plot Of mtcars Data Set

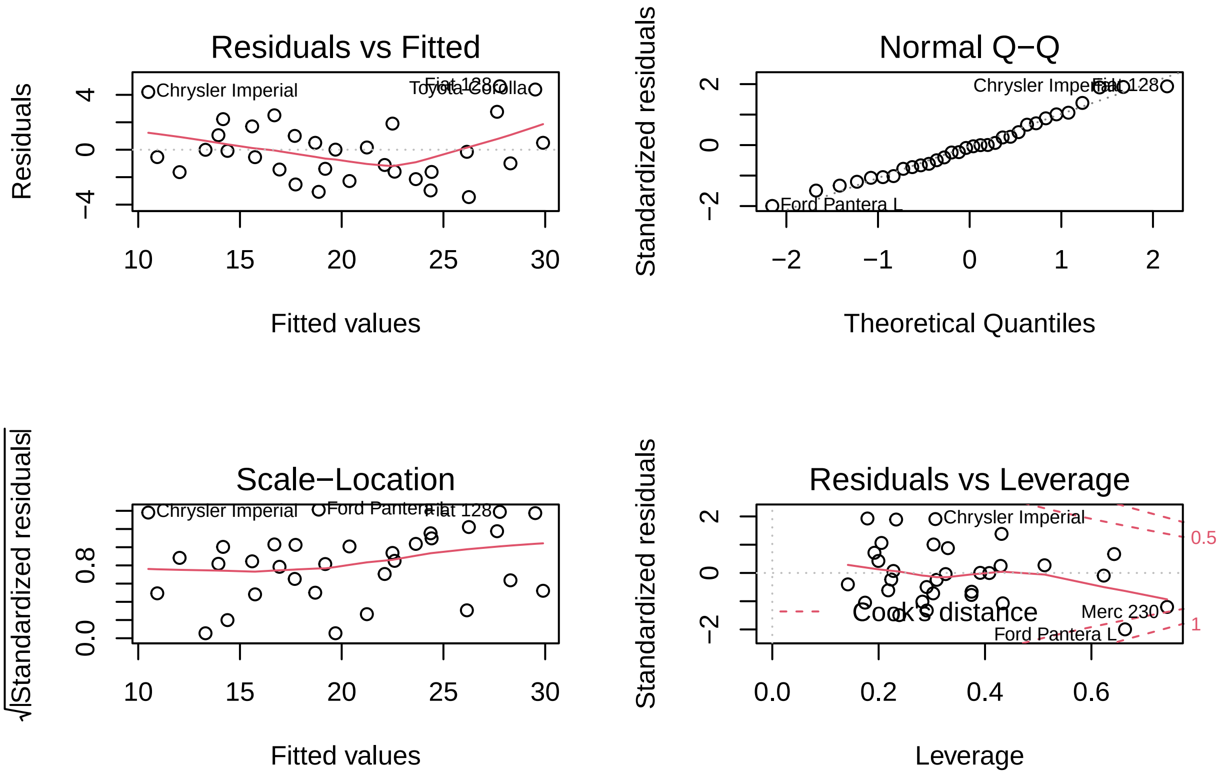
We can observe from the plot above that most of the variables are highly correlated which will pose a problem of individual effect on the response variable mpg.

Methods

We first begin by running a multiple linear regression on the mtcars dataset and check the model assumptions

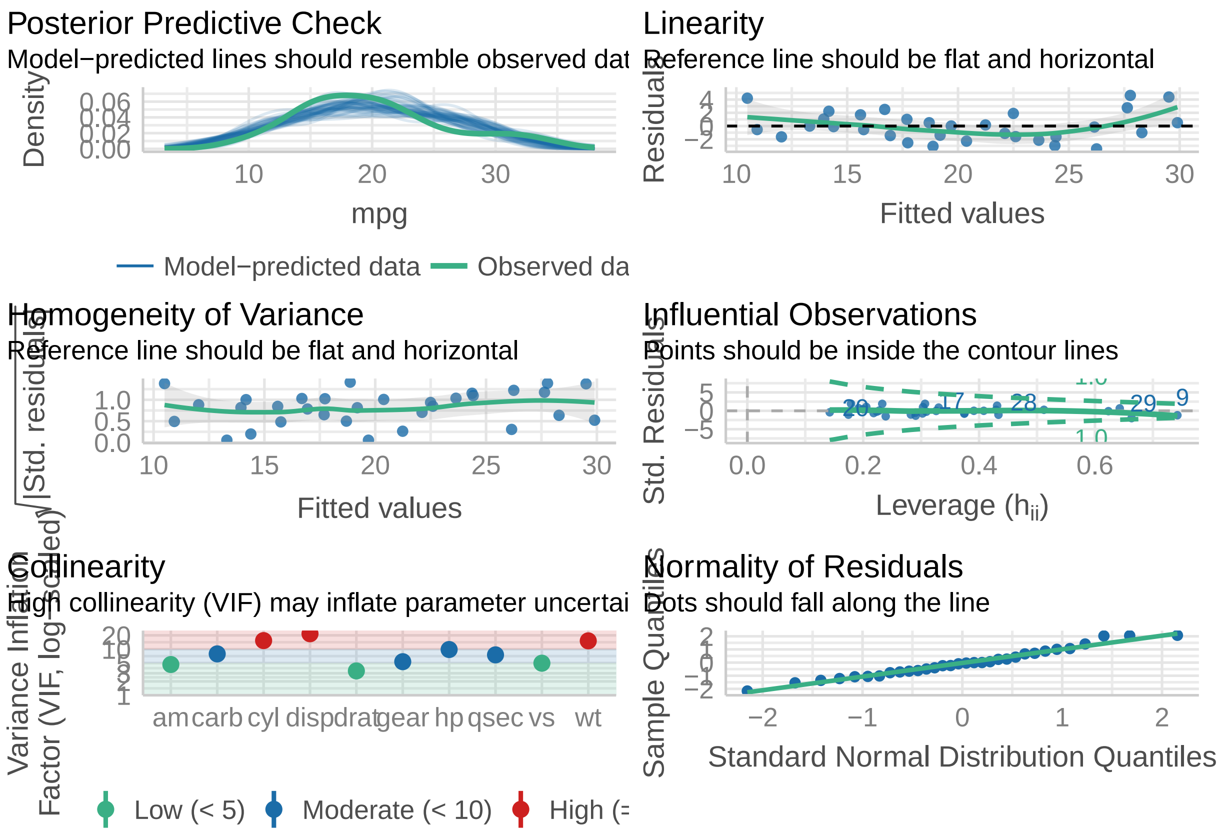
```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```



From the above,although we have an R-squared of 80.66% which shows that 80.66% of the variations in the data has been explained by the linear model, We will do better if we proceed to transforming the data using the Box Cox Transformation.

Checking Model Assumptions



We can observe from the plots above that the linearity and homogeneity assumptions are violated. Also,there are issues of multi-colinearity and a little issue with normality of the residuals

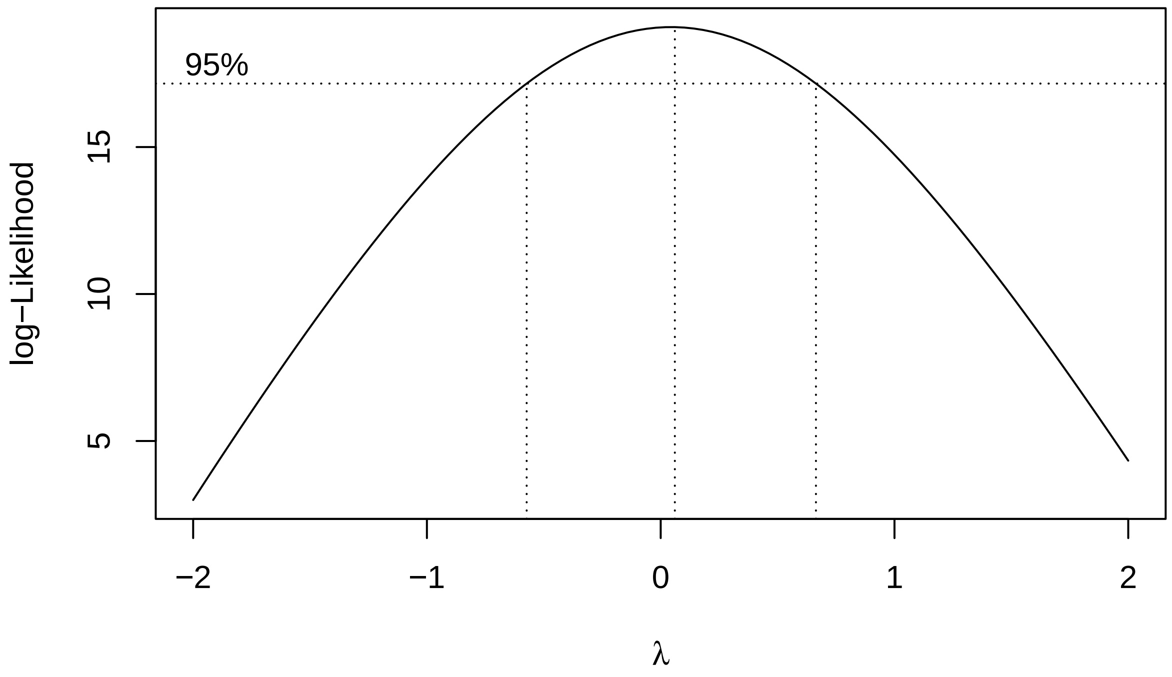
Box Cox Transformation

The ordinary least squares regression assumes normal distribution of residuals. When this is not the case, the Box-Cox Regression procedure may be useful (see Box, G. E. P. and Cox, D. R. 1964). It will transform the dependent variable using the Box-Cox Transformation function and employ maximum likelihood estimation to determine the optimal level of the power parameter lambda.

I would like to quickly remind the reader that because of issues with the pdf knitting, I did not include the mathematical definition of the Box Cox transform. However, I would like to refer the reader to "An Analysis of Transformations. Journal of the Royal Statistical

Society" cited in the reference page

We would attempt overcoming these model assumption violations using the Box Cox transformation.



```
## [1] 0.06060606
```

From the above diagram, it can be observed that the optimal λ is 0.06

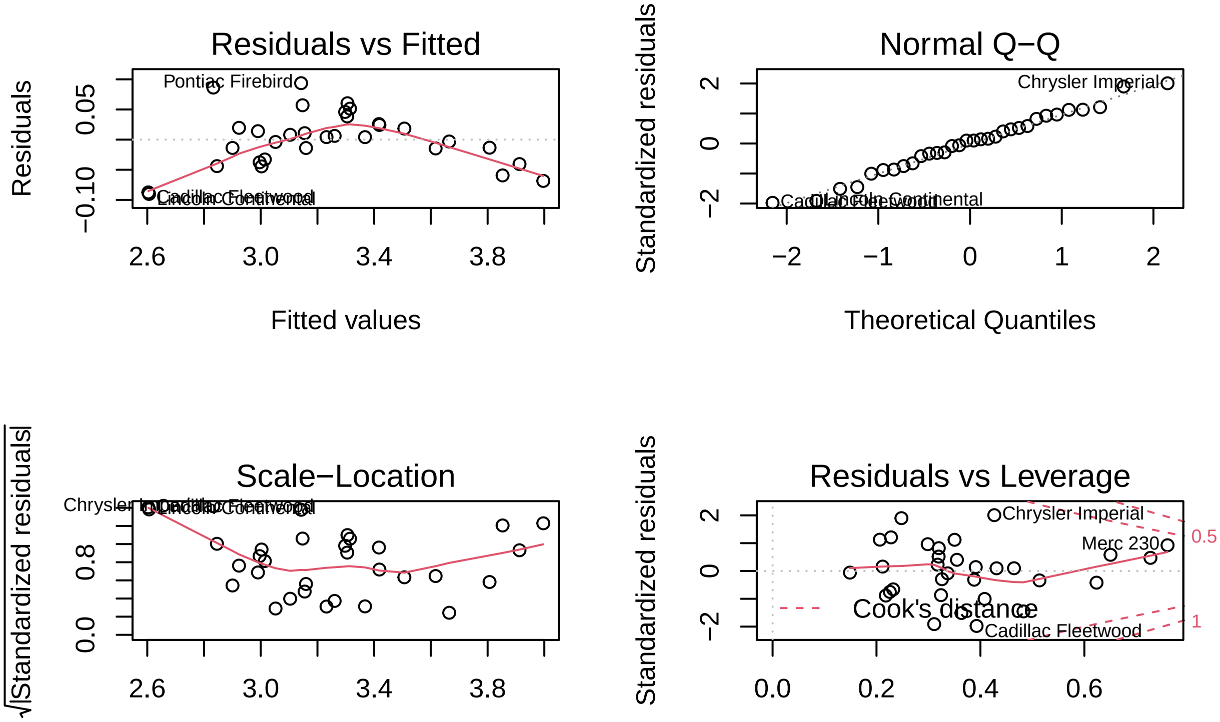
Now, we need to transform the response variable accordingly, and re-compute the linear model, with this transformed variable.

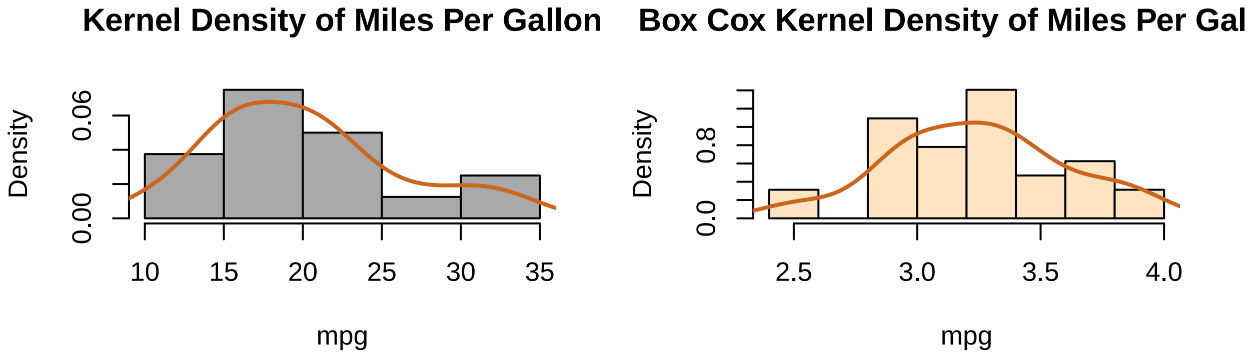
```
## [1] 0
```

Applying the Transformation

```
##
## Call:
## lm(formula = Box_Cox_mpg ~ ., data = mtcars)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.090131 -0.034089  0.004121  0.024638  0.093606
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.397e+00  4.055e-01   5.910 8.83e-06 ***
## mpg          5.006e-02  4.680e-03  10.698 1.01e-09 ***
## cyl          1.332e-02  2.242e-02   0.594    0.559
## disp        -5.569e-04  3.880e-04  -1.435    0.167
## hp          -1.595e-05  4.775e-04  -0.033    0.974
## drat        -1.179e-02  3.526e-02  -0.334    0.742
## wt          -2.114e-02  4.419e-02  -0.478    0.638
## qsec        -3.570e-03  1.614e-02  -0.221    0.827
## vs          -1.815e-02  4.516e-02  -0.402    0.692
## am          -6.366e-02  4.566e-02  -1.394    0.178
## gear         3.613e-02  3.217e-02   1.123    0.275
## carb        -1.318e-02  1.780e-02  -0.740    0.468
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.05683 on 20 degrees of freedom
## Multiple R-squared:  0.9836, Adjusted R-squared:  0.9745
## F-statistic: 108.9 on 11 and 20 DF,  p-value: 2.35e-15
```

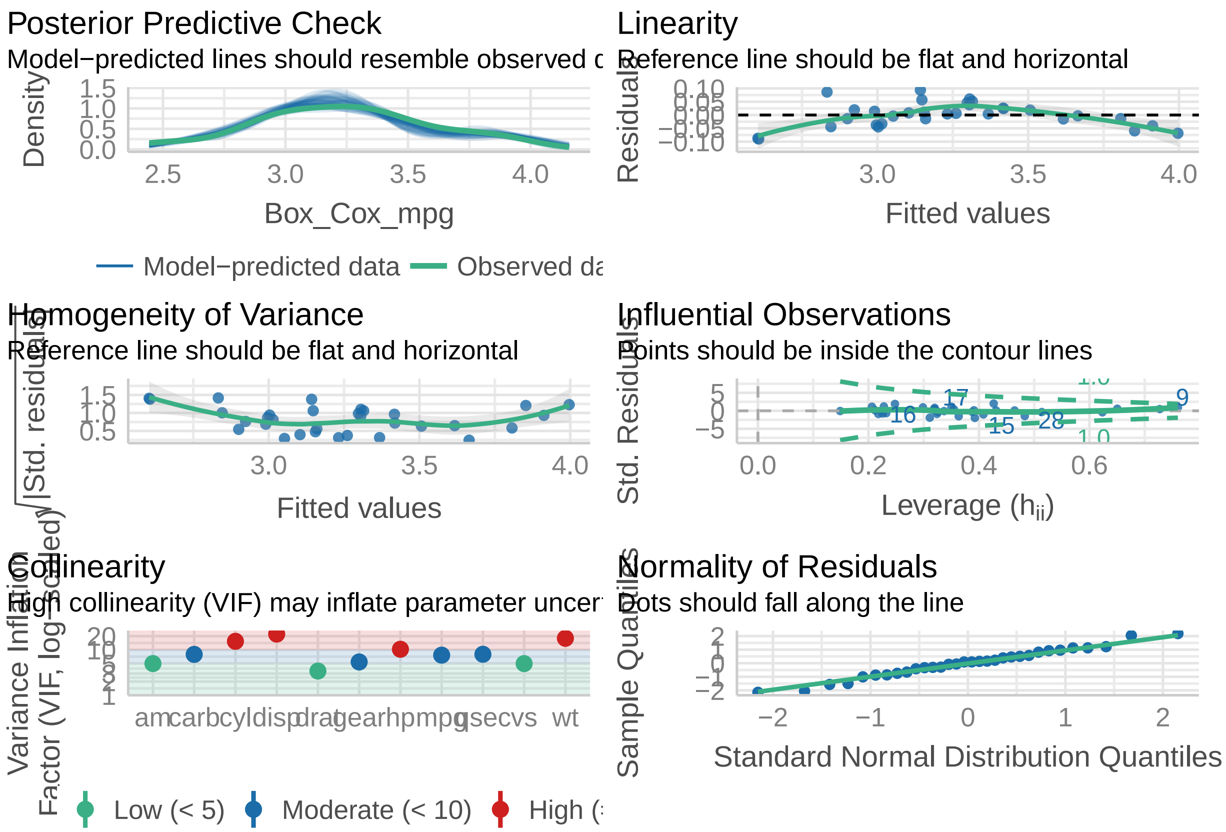
It can be observed from the above output that the adjusted R-squared has improved from 80.66% to 97.45%





It can be seen from the histogram of the response variable, mpg is fairly normal after the Box Cox transformation. Though fairly normal, we are interested in the normality of the residuals as well.

Checking for Model Assumptions after the Box Cox Transformation



Discussion

Overall, We can see a massive improvement in the adjusted R-squared from 80.66% to 97.45% after the Box Cox Transformation. This simply means that about 97.45% of the variations in the data has been explained by the linear model with the Box Cox transformation which is pretty descent. Although, the Box-Cox Transformation has helped in fitting the data well and improving the total variations explained by the model and the normality of the residuals, We still have issues with muti-colinearity and heteroscedascity.If the purpose of the model is for inference,these issues could be overcome using other transformations which are beyond the objective of this project however, If the purpose of the model was for predictions, we could ignore those assumption violations and proceed with the Box Cox transformed model as the best model because of the 97.45 R-squared.

Appendix

Checking for missing values: sum(is.na(mtcars))

histogram for each variable: hist.data.frame(mtcars)

QQ Plot for each variables: library(tidyverse) #set.seed(10) #dat <- data.frame(Observed = rnorm(20), sim1 = rnorm(20), sim2 = rnorm(20),sim3 = rnorm(20),sim4 = rnorm(20),sim5 = rnorm(20),sim6 = rnorm(20))

plot_data <- map_dfr(names(mtcars), ~as_tibble(qqplot(mtcars[[.x]], mtcars$mpg, plot.it = FALSE)) %>% mutate(id = .x))

ggplot(plot_data, aes(x, y, color = id)) + geom_point() + geom_abline() + facet_wrap(~id)

Shapiro Test: apply(mtcars,2,shapiro.test)

Correlation and correlation plot for each variable: y<-cor(mtcars)

library(corrplot) library(lattice) library(survival) library(Formula) library(ggplot2) library("Hmisc")

corrplot(y, tl.col = "red", bg = "White", tl.srt = 35, title = Correlation Plot Of mtcars Data set , addCoef.col = "black", type = "full")

Multiple Linear Regression Model and plot: fit<-glm(mpg~.,data=mtcars) summary(fit)

par(mfrow=c(2,2)) plot(fit)

Finding best lamda: library(MASS) Box_Cox <- boxcox(fit)

Function to implement Box Cox Transformation : Box_Cox_Transform <- function(y, lambda=0) { if (lambda == 0L) { log(y) } else { (y^lambda - 1) / lambda } }

Reversing the Box- Cox Transformation: Box_Cox_TransformInverse <- function(yt, lambda=0) { if (lambda == 0L) { exp(yt) } else { exp(log(1 + lambda * yt)/lambda) }

Histogram and density plot between mpg and Box Cox Transformed mpg:

```
par(mfrow=c(2,2)) hist(mtcars$mpg, col=“darkgray”, border=“black”, prob = TRUE, xlab = “mpg”, main = “Kernel Density of Miles Per Gallon”)

lines(density(mtcars$mpg), lwd = 2, col = “chocolate3”)

Box_Cox_mpg <- Box_Cox_Transform(mtcars$mpg, Box_Cox.power )

hist(Box_Cox_mpg, col=“bisque1”, border=“black”, prob = TRUE, xlab = “mpg”, main = “Box Cox Kernel Density of Miles Per Gallon”)

lines(density(Box_Cox_mpg), lwd = 2, col = “chocolate3”)
```

Fit of of the linear regression using the transformed response variable mpg:

```
Box_Cox_fit <- lm(Box_Cox_mpg ~., data=mtcars) summary(Box_Cox_fit)
```

Checking Model assumption before Box Cox Transformation:

```
library(performance) check_model(fit,lwd=3,cex=3)
```

Checking for Model Assumption after Transformation:

```
library(performance) check_model(Box_Cox_fit,lwd=3,cex=3)
```

Plot for the transformed fit:

```
par(mfrow=c(2,2)) plot(Box_Cox_fit)
```

References

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. Journal of the Royal Statistical Society. Series B (Methodological), 26(2), 211–252. http://www.jstor.org/stable/2984418

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations (with discussion). Journal of the Royal Statistical Society. Series B (Methodological), 26(2), 211–252