PROJECT REPORT

Group 8: Car Price Prediction

Introduction

In the automotive sector, pricing analytics play an essential role for both companies and individuals to assess the market price of a vehicle before putting it on sale or buying it. There are many automobile companies aspires to enter the US market by setting up their manufacturing unit in US. Predicting car prices involves determining a vehicle's market worth based on various attributes like brand, model, year of manufacture, mileage, and overall state. This prediction holds significant value for the auto industry, aiding both prospective buyers and sellers in understanding pricing and making educated decisions regarding vehicle transactions.

Objective:

This project aims to predict car prices based on various car features in the dataset, exploring relationships between car specifications and their market prices. The attributes used in the dataset provide valuable insights into the factors influencing car prices which help the management to understand how exactly the prices vary with the independent variables and can be used to develop predictive models for estimating the selling price of cars. They can accordingly manipulate the design of the cars, the business strategy etc. to meet certain price levels.

Research Questions:

- What features significantly impact the price of a car?
- Can car price differ with number of cylinders, engine type, car body?
- Can car body be dependent on number of cylinders in the car?

Motivation:

Understanding price determinants can aid manufacturers in setting prices and help consumers make informed purchasing decisions.

Hypothesis:

We hypothesize that factors such as Higher engine size, horsepower, and specific body types will positively impact car prices. Features like fuel economy may negatively correlate with price due to a trade-off with performance.

Exploratory Data Analysis:

Initially there were 205 observations, 26 attributes with some missing values in some columns (peakrpm, price, highwaympg).

1. Missing values in numerical variables peakrpm, highwaympg replaced with median value.

2. Removed unnecessary ID column from the dataset.

3. Missing value in Price column (target variable) whole row (around 4 entries) has been deleted.

4. Qualitative variables like fueltype, aspiration, doornumber, enginelocation has only two values so its replaced with 0,1.

Symboling, carname are not considered for analysis. All other variables are used for analysis. One hot encoding applied for remaining categorical variables (ie., carbody, enginetype, enginesize etc.)

Verified correlation of numerical & few categorical variables against target and each other. (*Refer appendix for code output 1*). There is more multicollinearity observed in the variables citympg & highwaympg, curbweight & carlength, carwidth, carlength & wheelbase, drivewheel & enginelocation, aspiration & enginetype. Variables correlated with target variable. Negative correlation non linear relationship is noticed in citympg, highwaympg with target variable.

Some interaction terms, polynomial non linear termswere selected for model based on domain knowledge like wheelbase, carlength - The relationship between a car's wheelbase and its overall length can influence interior space, comfort, and handling characteristics, thereby affecting the price. Fueltype(X10), highwaympg(X9)- The impact of fuel efficiency on price might differ based on the type of fuel used. For instance, diesel cars might have different pricing dynamics compared to gasoline cars concerning their mpg ratings.

Regression Analysis

Linear regression is a popular choice, but it often faces the challenge of overfitting, especially with a high number of parameters. Regularization techniques are used to address overfitting and enhance model generalizability.

Ridge regression is one of the methods for regularizing a model to address multicollinearity. Since there is more multicollinearity removing few multicollinear variables didn't had much difference (*refer appendix code output 5)*, So used Ridge regression where it reduces the impact of multicollinearity and effectively shrinks the coefficients of these correlated variables, preventing large fluctuations in predictions while still including all features in the model, making it more stable and reliable compared to a standard linear regression.

Interest findings found from the cleaned dataset after handling outliers is significant predictors influencing car prices which includes enginesize, horsepower, curbweight, drivewheel, and enginelocation. Few of the interaction terms like Interaction terms (e.g., enginesize with horsepower) and polynomial terms (e.g., square of citympg) also significantly impact predictions.

Model assumptions & model fit analysis are important in statistical modelling. During Outlier diagnostics Breusch pagan test is performed to check if variance of error terms are constant, but Heteroscedasticity is noticed and handled with log transformations (*Refer Appendix code output 9(a)*). Multicollinearity was mitigated using Ridge regression. (*Refer Appendix code* *output 10)*. The R-squared value of 0.938 indicates that 93.8% of the variance in car prices is explained by the model. The ratio of mean squared prediction error on validation data (0.89) confirms the model's generalizability and reduced overfitting.

Discussion & Limitations

Upon applying ridge regression all the variable coefficients reached to near 0 and the model has nice predictions with test dataset. The final model with cleaned dataset successfully predicts car prices with high accuracy, performing better on the validation dataset than the training dataset. Conclusions highlight the critical impact of factors such as enginesize, horsepower, and drivewheel on pricing decisions.

Several methods were tried to handle multicollinearity and reduce heteroscadestiity, so effective handling of multicollinearity, robust variable selection using stepwise methods, and validation splits ensured model reliability which is positive point. After transformations still few outliers were noticed which is common but if there are more robust methods that could be considered which improves model. Model building is moved forward since no curve linear relationship noticed in residual plot which can give idea to go for polynomial regression model or handle nonlinear relationships from data.

Exploring advanced models like Elastic Net to improve feature selection further and mode deeper analysis of Lasso regression. Incorporating additional features like market conditions or brand reputation for a more holistic model.

The dataset is moderately sized (201 observations) with diverse variables. However, influential outliers could affect reliability. Validity is supported by EDA and transformations, but potential biases are in qualitative variables like carbody which needs consideration.

Regression analysis is appropriate given the linear relationships observed during EDA. Ridge regression addresses multicollinearity, and the addition of interaction and polynomial terms extends the model's flexibility.

Conclusion

This project aimed to predict car prices based on a dataset of 25 variables, identifying key predictors like enginesize, curbweight, and horsepower. Ridge regression was used to mitigate multicollinearity, resulting in a highly accurate model with good generalizability and incorporating interaction terms enhanced the model's predictive accuracy. By refining the regression model through feature selection and validation, we provide a reliable tool for predicting car prices. These insights are valuable for both manufacturers and consumers, aiding in strategic decision-making and aligning with market trends.

Ridge regression effectively handled multicollinearity and improved predictive performance. Outliers and heteroscedasticity were addressed to ensure robust results. Key predictors and their interactions provide valuable insights for pricing strategies.

Additional work

1. Applied Boxcox transformations while handling heteroscedasticity (*refer Appendix code output 10(c)*) apart from log transformations.

2. Effects of multicollinearity coefficients verified, Standardization is applied to check any correlation is reduced but only small difference is noticed. (*refer Appendix: Code attached Project_standard.R , Project_multicollinearity.R files for code*).

3. L1 regularization Lasso regression is performed on test set in order to minimize regression coefficients with optimal λ value. *(refer Appendix: code output 12)*

4. Have tried PCA model which transforms correlated variables into a smaller number of uncorrelated components and used those components in the regression model during model building, and there is a improvement noticed (*refer Appendix: Code attached Project_PCA.R, file for code, Appendix: code output 12)*

Robust regression could handle outliers better but that is complex instead Cook's distance was sufficient for detecting and addressing influential outliers. Ridge regression was preferred because it retains all variables, allowing the model to capture nuanced relationships that might be lost with LASSO even though lasso performs feature selection by shrinking some coefficients to zero. Moreover, Ridge regresion effectively mitigated multicollinearity, explained a large proportion of variance (93.8% R-squared), generalized well across train and validation datasets, also achieved a good balance between simplicity and predictive performance.

The main objective aligned with the project's focus on understanding key factors influencing car prices and ensuring the findings were actionable for stakeholders.

Appendix (files attached in canvas as well)

Dataset:

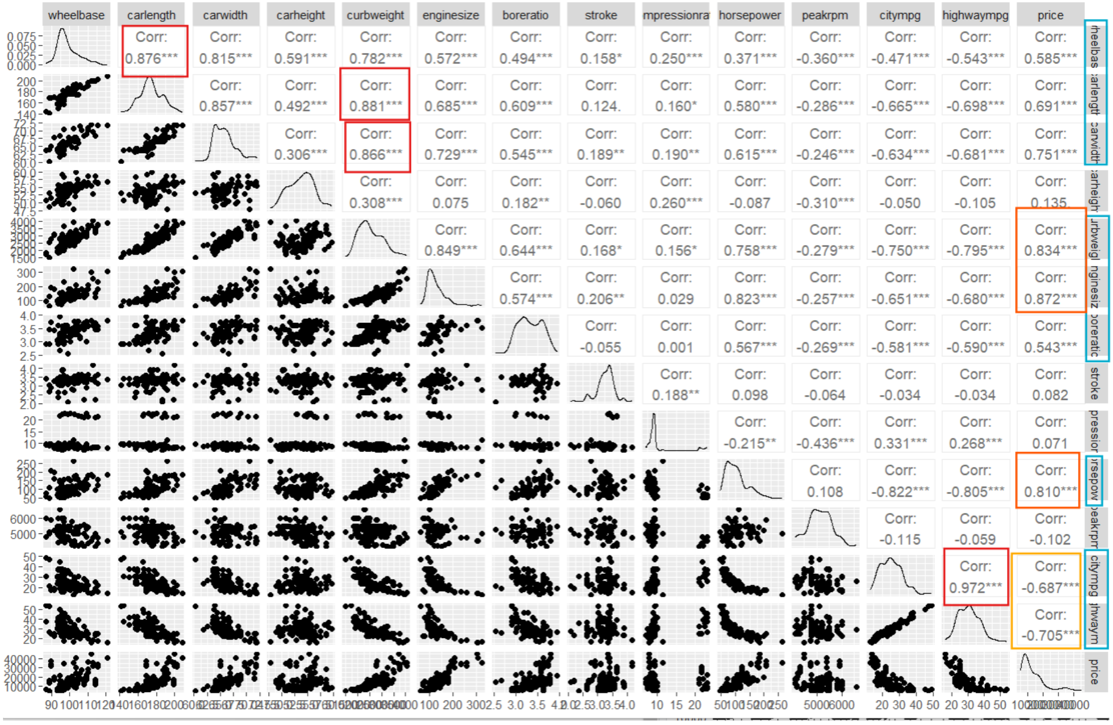
CarPrice.csv

Code: (4 files)

Codes.zip

Code Output:

1. Correlation between numerical variables



2. Qualitative variables correlation verified by Chisquare test.

Independent variables correlation

```
> c_t <- table(Cp$fueltype, Cp$fuelsystem)
> chisq.test(c_t)

        Pearson's Chi-squared test

data:  c_t
X-squared = 201, df = 7, p-value < 2.2e-16
```

```
> c_t1 <- table(Cp$drivewheel, Cp$enginelocation)
> chisq.test(c_t1)

        Pearson's Chi-squared test

data:  c_t1
X-squared = 5.1164, df = 2, p-value = 0.07745
```

```
> c_t2 <- table(Cp$aspiration, Cp$enginetype)
> chisq.test(c_t2)

        Pearson's Chi-squared test

data:  c_t2
X-squared = 10.471, df = 5, p-value = 0.06294
```

```
> c_t5 <- table(Cp$fuelsystem, Cp$enginetype)
> chisq.test(c_t5)

        Pearson's Chi-squared test

data:  c_t5
X-squared = 207.14, df = 35, p-value < 2.2e-16
```

Category variables correlation with target

```
> c_t3 <- table(Cp$carbody, Cp$price)
> chisq.test(c_t3)

        Pearson's Chi-squared test

data:  c_t3
X-squared = 770.5, df = 740, p-value = 0.212
```

```
> c_t4 <- table(Cp$fuelsystem, Cp$price)
> chisq.test(c_t4)

        Pearson's Chi-squared test

data:  c_t4
X-squared = 1368.2, df = 1295, p-value = 0.07711
```

Correlation between category variables. X13(carbody), X14(drivewheel), X18(cylindernumber) has more VIF > 10, which is critical multicollinearity.

```
> vif(catmodel)
       X10         X11         X12   X13hardtop X13hatchback    X13sedan    X13wagon      X14fwd
    3.4397      1.9235      2.6194      2.3918      10.6820     13.0660      6.7524      8.8682
    X14rwd         X15        X16l       X16ohc      X16ohcf     X16ohcv    X16rotor     X17five
   10.3150      2.0556      2.8148      5.1659       3.3787      3.0341      6.1617      6.1678
   X17four      X17six    X17three   X17twelve      X182bbl     X184bbl      X18mfi     X18mpfi
   18.7720      8.7131      1.8163      1.3188       4.8128      4.3122      1.1538      6.2787
   X18spdi     X18spfi
    2.1733      1.1379
>
```

3. Regression analysis coefficients ANOVA

```
> rownames(Reg_ana) <- cbind("Regression Analysis")
> print(Reg_ana)
                            SSR df_R       MSR       F*      p-value
Regression Analysis 10897178591   17 641010505 67.65013 3.643865e-68

                SSE df_E     MSE
Error 1733994097  183 9475378

               SST  df
Total 12631172689 200
```

4. General model with all variables with interaction terms.

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4 + X5 + X6 + X7 + X8 + X9 + 
    I(X8^2) + I(X9^2) + X10 + X11 + X12 + X13 + X14 + X15 + X16 + 
    X17 + X18 + X1X2 + X5X7 + X8X9 + X10X8 + X10X9 + X11X7 + 
    X15X5 + X1X6 + X1X8)

Residual standard error: 2032 on 155 degrees of freedom
Multiple R-squared:  0.9493,	Adjusted R-squared:  0.9346 
F-statistic: 64.53 on 45 and 155 DF,  p-value: < 2.2e-16
```

5. Model with few multicollinearity variables removed (highwaympg,carlength, carheight and the related interaction terms)

```
Call:
lm(formula = Y ~ X1 + X4 + X5 + X6 + X7 + X8 + I(X8^2) + X10 + 
    X11 + X12 + X13 + X14 + X15 + X16 + X17 + X18 + X5X7 + X10X8 + 
    X11X7 + X15X5 + X1X6 + X1X8)

Residual standard error: 2100 on 162 degrees of freedom
Multiple R-squared:  0.9434,	Adjusted R-squared:  0.9302 
F-statistic:  71.1 on 38 and 162 DF,  p-value: < 2.2e-16
```

6. Automatic regression procedure used, selected forward model based on lower AIC model.

Forward model

```
Step:  AIC=3096.07
Y ~ X5 + X17 + X16 + X5X7 + X4 + X13 + X18 + X14 + X7 + X3 +
    X1X8 + X10X9 + X15 + X1X6 + I(X8^2) + X8

          Df Sum of Sq       RSS    AIC
<none>                 694292738 3096.1
+ X1       1   4751348 689541390 3096.7
+ X12      1   3593121 690699616 3097.0
+ X10X8    1   2491948 691800790 3097.3
+ X6       1   1436780 692855958 3097.7
+ X2       1   1144033 693148705 3097.7
+ X9       1    264088 694028650 3098.0
+ X11X7    1    126836 694165902 3098.0
+ X11      1     25201 694267537 3098.1
+ X1X2     1     13280 694279458 3098.1
+ X8X9     1     12977 694279761 3098.1
+ I(X9^2)  1      2150 694290588 3098.1
```

Backward model

```
Step:  AIC=3192.81
Y ~ X3 + X4 + X5 + X7 + I(X8^2) + I(X9^2) + X13 + X14 + X16 +
    X17 + X5X7 + X8X9

          Df Sum of Sq        RSS    AIC
<none>                  822875850 3192.8
- X3       1  39062346  861938196 3196.8
- X14      2  93408163  916284014 3203.8
- I(X8^2)  1  81494615  904370466 3206.5
- X7       1  83412719  906288569 3206.9
- X8X9     1  83601321  906477171 3207.0
- I(X9^2)  1  86134869  909010720 3207.5
- X13      4 165088075  987963925 3208.3
- X4       1 115441296  938317146 3213.9
- X5       1 136069991  958945841 3218.3
- X16      4 320221758 1143097609 3237.7
- X5X7     1 252354597 1075230447 3241.3
- X17      5 684477247 1507353097 3288.0
```

Model built with forward selected variables , below are Rsquare & p value coefficient.

```
Residual standard error: 2043 on 163 degrees of freedom
Multiple R-squared:  0.9462,	Adjusted R-squared:  0.9339
F-statistic: 77.42 on 37 and 163 DF,  p-value: < 2.2e-16
```

7. Model validation on forward model.

Train

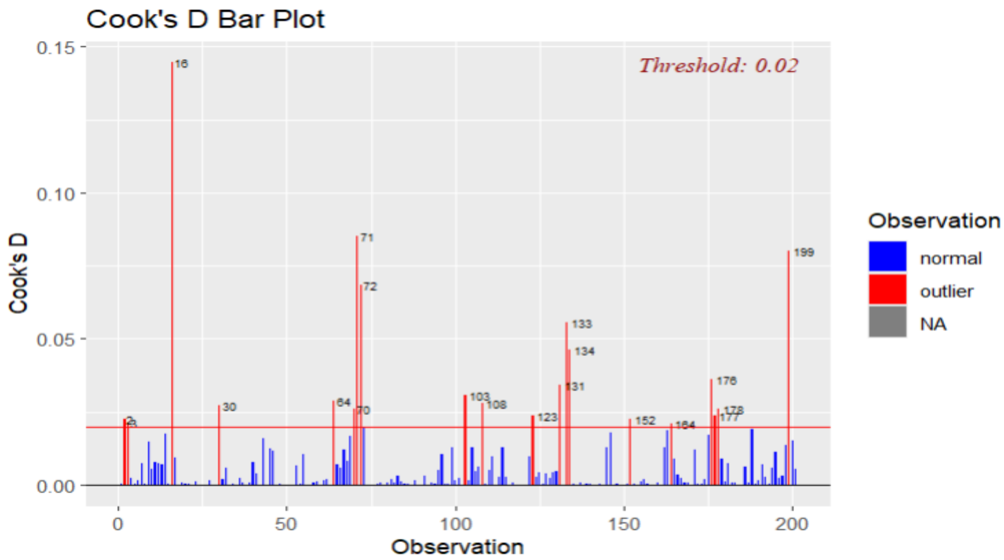
```
Residual standard error: 2468 on 125 degrees of freedom
Multiple R-squared:  0.9293,	Adjusted R-squared:  0.9152
F-statistic: 65.72 on 25 and 125 DF,  p-value: < 2.2e-16
```

Validation

```
Residual standard error: 1706 on 27 degrees of freedom
Multiple R-squared:  0.9496,	Adjusted R-squared:  0.9084
F-statistic:  23.1 on 22 and 27 DF,  p-value: 2.863e-12
```

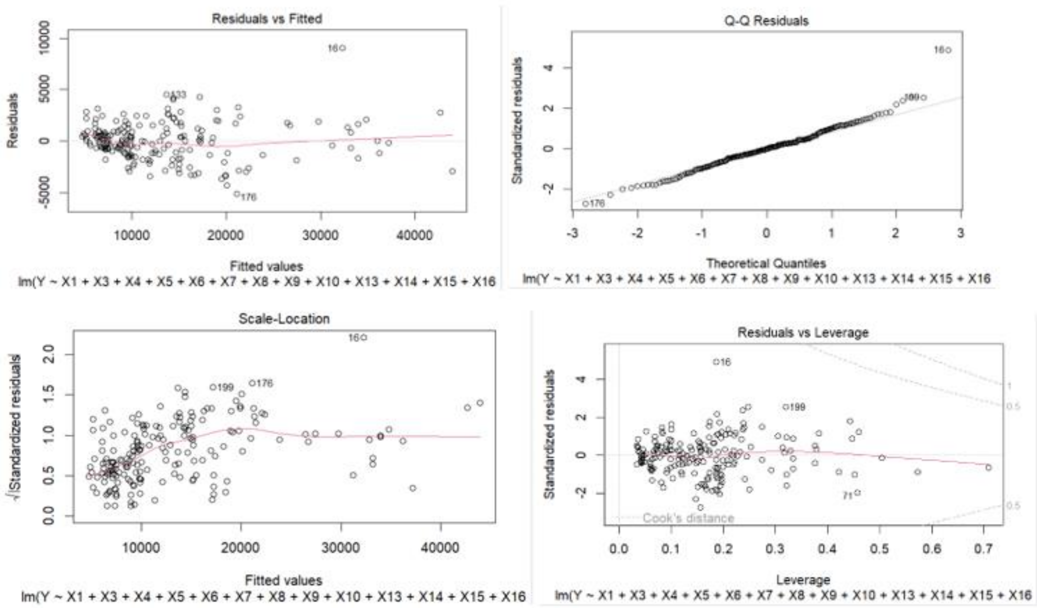
8. Outlier diagnostics.

a. 20 Influential cases found more than cooks distance threshold (0.02)



2	3.47	2.68	9.0	111	5000	21	27	16500
3	2.68	3.47	9.0	154	5000	19	26	16500
16	3.62	3.39	8.0	182	5400	16	22	41315
30	2.91	3.41	9.6	58	4800	49	54	6479
64	3.43	3.64	22.0	72	4200	31	39	18344
70	3.46	3.10	8.3	155	4750	16	18	35056
71	3.80	3.35	8.0	184	4500	14	16	40960
72	3.80	3.35	8.0	184	4500	14	16	45400
103	3.43	3.27	7.8	200	5200	17	23	19699
108	3.70	3.52	21.0	95	4150	25	25	13860
123	3.94	3.11	9.5	143	5500	19	27	22018
131	2.54	2.07	9.3	110	5250	21	28	15040
133	3.54	3.07	9.0	160	5500	19	26	18150
134	3.54	3.07	9.0	160	5500	19	26	18620
152	3.05	3.03	9.0	62	4800	27	32	8778
164	3.62	3.50	9.3	116	4800	24	30	8449
176	3.27	3.35	9.3	161	5200	19	24	15998
177	3.27	3.35	9.2	156	5200	20	24	15690
178	3.27	3.35	9.2	156	5200	19	24	15750
199	3.58	2.87	8.8	134	5500	18	23	21485

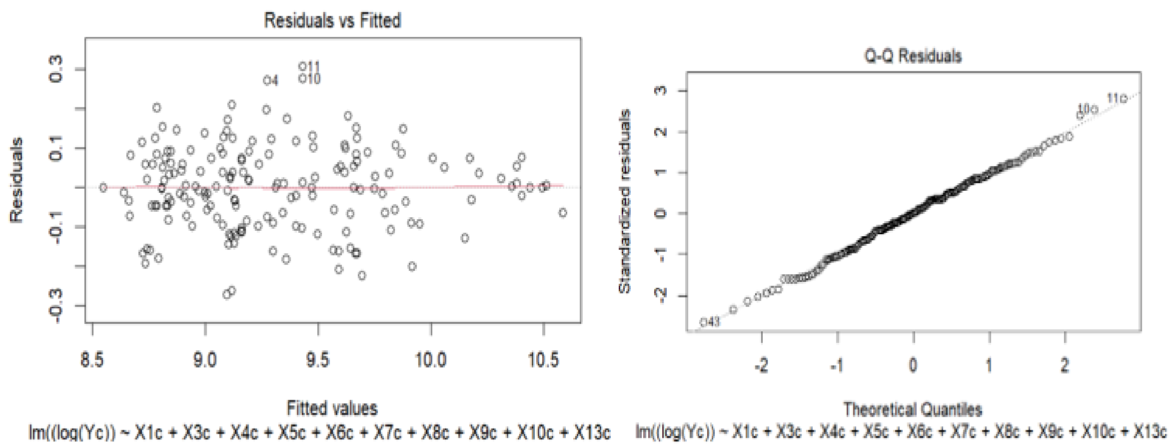
9. Residual plots



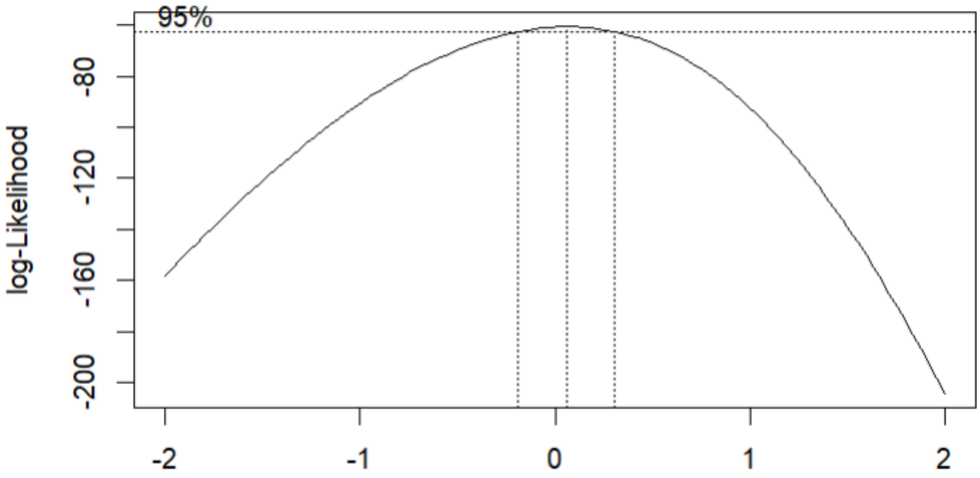
a. Breusch pagan test applied to check for homoscedasticity assumption

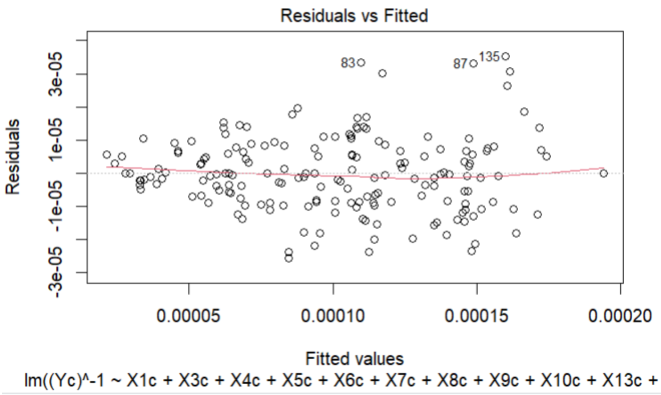
```
> bptest(model_c, studentize=FALSE)

        Breusch-Pagan test

data:  model_c
BP = 88.194, df = 37, p-value = 4.575e-06
```

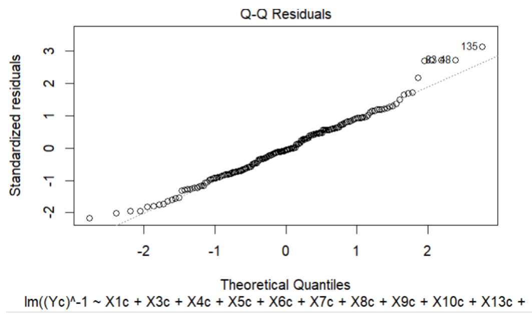
b. After handling heteroscedasticity with log transformation



c. Handling heteroscedasticity with boxcox transformation







From above figure it can be seen Heteroscedasticity is redued and the residuals can be seen splits all over.

10. Multicollinearity diagnostics

```
> vif(fmv2e)
 wheelbase    carwidth curbweight enginesize  boreratio horsepower    citympg highwaympg
    6.1009      5.5270    16.5910     7.0906     2.0622     7.6670    21.4480    21.8860
```

Applied Ridge regression technique to handle multicollinearity to have precise model with accurate predictions, below are ridge coefficients which are near to 0.

```
                       s0
(Intercept)    9.2990941221
X1c            0.0369353366
X3c            0.0534501055
X4c            0.0595818561
X5c            0.0255708669
X6c            0.0045755199
X7c            0.0428342846
X8c           -0.0298723732
X9c           -0.0171960659
X10c          -0.0166390939
X13chardtop   -0.0113218636
X13chatchback -0.0359503418
X13csedan     -0.0029505098
X13cwagon     -0.0175034002
X14cfwd       -0.0272188921
X14crwd        0.0296754836
X15c          -0.0619607903
X16cl         -0.0107506405
X16cohc        0.0360514901
X16cohcf      -0.0018387317
X16cohcv      -0.0280169549
X16crotor      0.0054909946
X17cfive       0.0065126825
X17cfour      -0.0541256260
X17csix        0.0230595419
X17cthree      0.0203517664
X17ctwelve    -0.0009805843
X17ctwo        0.0054750030
X18c2bbl      -0.0272976272
X18c4bbl      -0.0085435700
X18cidi        0.0169598678
X18cmfi        0.0002773727
X18cmpfi       0.0245124498
X18cspdi      -0.0057045645
X18cspfi      -0.0002361428
X1X6c          0.0224802560
X1X8c         -0.0260404914
X5X7c          0.0372553052
X8X8c         -0.0201544022
X10X9c        -0.0213590283
>
```

Predictions on ridge model below are the regression statistics.

```
> sst
[1] 40.67356
> sse
[1] 2.51111
> r_squared
[1] 0.9382619
>
```

Model validation with 75% train and 25% test on ridge regression model, below are statistical measures.

```
> cat("Total sum of squares (train):", sst_train, "\n")
Total sum of squares (train): 7113242191
> cat("Error sum of squares (train):", sse_train, "\n")
Error sum of squares (train): 467843921
> cat("Mean Absolute Error (train):", mae_train, "\n")
Mean Absolute Error (train): 1433.858
> cat("R-squared (train):", R_squaredtrain, "\n")
R-squared (train): 0.9342292
> cat("RMSE (train):", rmse_train, "\n")
RMSE (train): 1861.588
> cat("Total sum of squares (test):", sst_test, "\n")
Total sum of squares (test): 1822079535
> cat("Error sum of squares (test):", sse_test, "\n")
Error sum of squares (test): 120478697
> cat("Mean Absolute Error (test):", mae_test, "\n")
Mean Absolute Error (test): 1142.22
> cat("R-squared (test):", R_squaredtest, "\n")
R-squared (test): 0.9338785
> cat("RMSE (test):", rmse_test, "\n")
RMSE (test): 1618.364
```

11. Lasso regression technique is used to reduce multicollinearity and for feature selection.

Predicted vs actual values from lasso regression model

```
> head(final)
   y_test        s1
1   13495  9998.029
4   13950 10916.328
5   17450 18526.553
12  20970 19833.249
18   5151  4181.313
21   5572  5851.761
```

Good R^2 value on test set predictions.

```
> R_squaredtestl <- 1 - (sse_testl / sst_testl )
> cat("R-squared (test) with Lasso:", R_squaredtestl, "\n")
R-squared (test) with Lasso: 0.9531789
```

12. PCA model building is implemented, pca components were used in the place of highly correlated variables, below are the pca components.

```
> summary(pca_result)
Importance of components:
                         PC1    PC2     PC3     PC4    PC5     PC6
Standard deviation     2.181 0.7852 0.65594 0.33707 0.2400 0.15577
Proportion of Variance 0.793 0.1028 0.07171 0.01894 0.0096 0.00404
Cumulative Proportion  0.793 0.8957 0.96742 0.98636 0.9960 1.00000
```

Model summary

```
+                (X1c*pca8)+ (X10c*pca9) + X15c + (X1c*X6c) + X6c
> summary(modelpc) #R2=95,94

Call:
lm(formula = log(Yc) ~ X1c + pca5 + pca9 + X10c + X17c + X16c + 
    (pca5 * pca7) + pca4 + X13c + X18c + X14c + pca7 + pca3 + 
    (X1c * pca8) + (X10c * pca9) + X15c + (X1c * X6c) + X6c + 
    pca8 + (X10c * pca8), data = Cp_c)


Residual standard error: 0.1164 on 143 degrees of freedom
Multiple R-squared:  0.9523,    Adjusted R-squared:   0.94 
F-statistic: 77.24 on 37 and 143 DF,  p-value: < 2.2e-16
```