ST300 Project Report

Candidate Numbers:
25983
26617
40088

1 Introduction

1.1 Objective

Our aim was to develop a multiple linear regression model that could represent a strong relationship between a car's price and the different features of a car. Using 3 variable selection processes, 3 models were developed and compared to see which linear regression model best fitted the real data.

1.2 Data Set

1.2.1 Finding the Data Set

Our data came from Kaggle.com in which we collected 205 rows of data with 20 variables. We chose to make car prices our respondent variable as it would be more desirable to understand how a car's properties would influence its price. This made the other 19 variables in our dataset candidates for the predictors to be included in our MLR model.

1.2.2 Cleaning the data

Before selecting variables, the data set had to be cleaned as it contained irrelevant columns such as 'Car Name', 'Symbolling' and 'Car ID'; these columns were removed in Excel as they intuitively had no contextual meaning regarding car features. In addition, several categorical variables were removed due to their irrelevance, such as 'Aspirations' and 'Engine Location' which only contained 2 levels each. In particular, 'Fuel Type' was removed as the level 'gas' covered 97% of fuel types on our data set, so it was unlikely we would see any significant dependence on changes in fuel type. Other variables like 'Peak rpm' and 'Fuel System lacked in variation so it would be difficult to see how changes in these variables affected pricing. Cleaning our dataset made our code more efficient and the best model showed a strong fit. In total, 7 variables were removed due to lack of variation in data or had very little relevance in pricing a car. The remaining 12 variables were used as our independent variables to analyse how price is associated with these factors.

1.2.3 Data Formats

Our dataset was comprised of a variety of continuous and discrete data. Since our categorical levels were not binary, we must create indicator functions for each level to correctly carry out a linear regression. We computed the linear model consisting of all the predictors to determine which levels were absorbed into the intercept. By including every level, we will get multicollinearity since there will be linear dependence in the levels. Take 'drivewheel' for example, each car must satisfy either fwd, rwd, or 4wd. Therefore, given that it's not two of them, then it must be the third implying dependency. Hence why we only need 2 of the levels for our model. For the remaining levels, we added columns for each level containing 0s and 1s. The datapoint takes value 1 if it satisfies the corresponding level and 0 otherwise. For example, for carbody = sedan

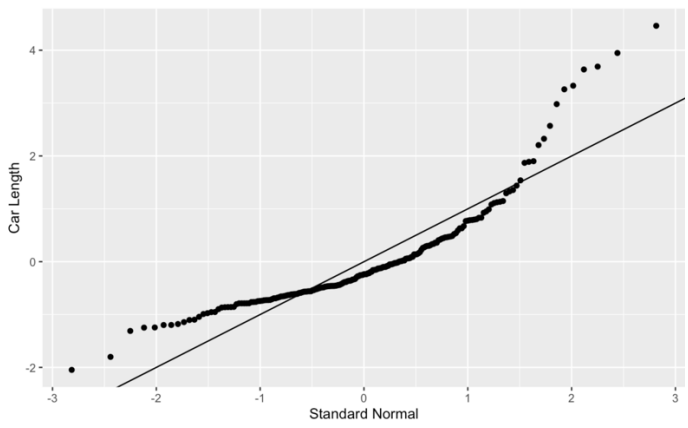
$$\underline{\boldsymbol{a}} = \begin{pmatrix} a_{11} \\ a_{12} \\ . \\ . \\ . \end{pmatrix} \qquad \boldsymbol{where} \qquad a_{i1} = \begin{cases} \mathbf{1}, & \boldsymbol{carbody = sedans} \\ \mathbf{0}, & \boldsymbol{otherwise} \end{cases}$$

These vectors were added to the data set as their own separate columns so we could analyse each level of a categorical variable individually.

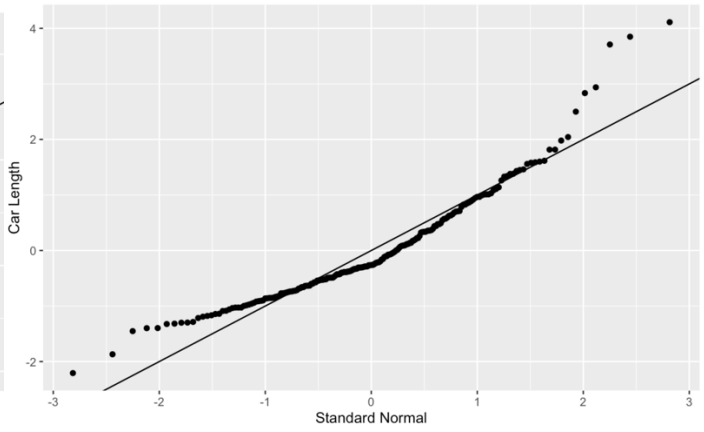
1.3 Transformations & Correlations

The first obstacle we encountered was the normality of residuals. It is important that our residuals are distributed normally with zero mean as this is an underlying assumption for MLR. To identify this problem, we individually regressed the predictors against price and then created a QQ-plot against the standard normal for each model. In doing so we discovered that some variables required transformations to behave more normally.

From the figure below you can see the variable car length's QQ Plot before and after applying a logarithmic transformation to price. As a group we concluded that the residuals in the transformed model are more normally distributed as they are closely fitted to the line. We then repeated this process for the other variables to make sure the residuals are as normal as possible.



QQ Plot Before Transformation



QQ Plot After Transformation

As a group, we decided that each proposed model should use the same transformations. Our reasoning behind this is we would like to accurately determine which model is best conditional on the same transformations which essentially implies using the same set of data. If we were to use different transformations, it would be difficult to compare at the end. For example, the effects on E(f(Y)) may not translate to the same effects on E(g(Y)) where E(.) are our fitted values.

Additionally, most of the transformations made led to a stronger correlation with price which can be seen in the table below.

Variables	Correlation before Transformation	Correlation after Transformation
Wheelbase	0.57781560	0.5778155983
Car length	0.68292002	0.6829200157
Car width	0.75932350	0.7593252997
Car height	0.11933623	0.1193362266
Curb weight	0.83530488	0.8353048793
Engine size	0.87414480	0.8741448025
Bore ratio	0.55317324	0.5531732368
Stroke	0.07944308	0.0926491987
Compression ratio	0.06798351	0.0007039066
Horsepower	0.80813882	0.8081388225
Highway MPG	-0.69759909	-0.6975990916

Despite compression ratio showing a drop in correlation, the QQ-plot suggests it is more normally distributed which is more important for the overall model.

2.0 Best Model (BIC)

After comparing all aspects of the proposed models, we decided that the following model is the best in showing the association between log(price) and the transformed predictors.

$$\log(price_i) = 3.91 + 0.0624 carwidth_i + 0.00101\sqrt{curbweight_i} + 0.00439\sqrt{horsepower_i} - \mathbb{1}_{\{carbody_i = wagon\}} 0.278 - \mathbb{1}_{\{carbody_i = hatchback\}} 0.265 - \mathbb{1}_{\{carbody_i = sedan\}} 0.226 + \mathbb{1}_{\{drivewheel_i = rwd\}} 0.00845$$

The rest of section 2 will go on to explain how we arrived at this model and section 3 will explore alternative models we could have selected and why we didn't.

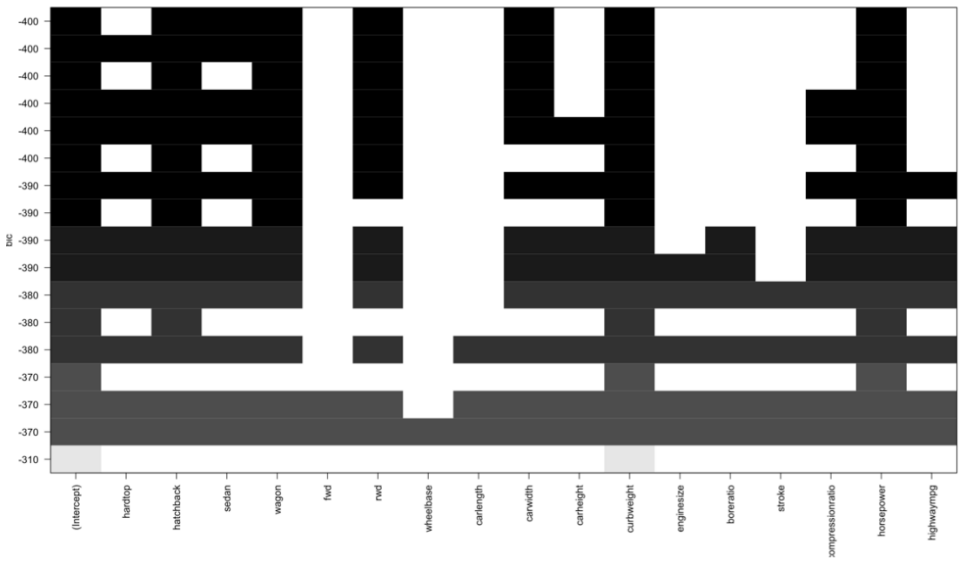
2.1 Variable Selector

The best model's variables were derived from using the Bayesian Information Criterion (BIC) in which we chose the set of variables that produced the lowest BIC. This metric heavily penalizes the addition of parameters as shown in the formula:

$$BIC = -2\ln(L) + p\,ln(n)$$

Where L is the likelihood of the linear regression model, p is the number of parameters in our model, and n is our sample size.

By using Best Subset Regression, we could find the lowest BIC values for every combination of variables. This process works by calculating the Residual Sum of Squares of every possible model given a subset size k, we then choose the model which has the lowest RSS in each subset size. As we are currently comparing models of the same subset size, we are not comparing model complexity, so we compare each model's RSS instead. With 18 models to choose from, we then calculate the BIC of each model to see which model offers the lowest BIC. Since our altered dataset only contained 18 parameters, we could compute 2^{18-1} combinations in R. For much larger number of parameters, we would have had to use stepwise regression, further justifying the cleaning of our data in 1.2.2.



BIC Values

From this table, we can select the best model from the top row of this table with a BIC value of -400, containing the following variables:

- Car body (levels: hatchback, sedan and wagon)
- Drive wheel (levels: rwd)
- Car width
- Curb weight
- Horsepower

2.2 Multicollinearity

We use OLS estimation to find the coefficients $\widehat{\boldsymbol{\beta}}$ for our chosen variables, hence we calculate $\widehat{\boldsymbol{\beta}}$ using the following equation:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\,\boldsymbol{X}'\,\boldsymbol{y}$$

Where $\boldsymbol{X}$ is a 205 x 13 matrix of the transformed values from the data set and $\boldsymbol{y}$ is a 205 x 1 vector containing the log-prices of every car.

Multicollinearity occurs when there is high dependence amongst the variables which can have negative effects on the estimation of our coefficients $\widehat{\boldsymbol{\beta}}$. High dependence means that the eigenvalues of the $\boldsymbol{X}'\boldsymbol{X}$ matrix would be close to 0, so by taking the inverse, we obtain very large estimations of $\widehat{\boldsymbol{\beta}}$ and its variance. To detect multicollinearity, we calculated the Variance Inflation Factor (VIF) for each variable in our model to determine if multicollinearity existed. VIF works by calculating the ratio between the variance of a coefficient $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ in a multiple linear regression and the variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}}$ if it were strictly in its own model with no other variables.

$$\boldsymbol{VIF_j} = \frac{\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}})}{\boldsymbol{Var}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{j}})_{\boldsymbol{min}}}$$

We aim to remove variables if their is VIF > 10 to reduce the effects of multicollinearity on our model. By running VIF calculations in R, we obtain the following values for each variable in our BIC model:

Variable	VIF
Car width	4.5
Curb weight	8.1
Horsepower	2.8
Hatchback	4.4
Sedan	4.8
Wagon	2.9
Rwd	2.1

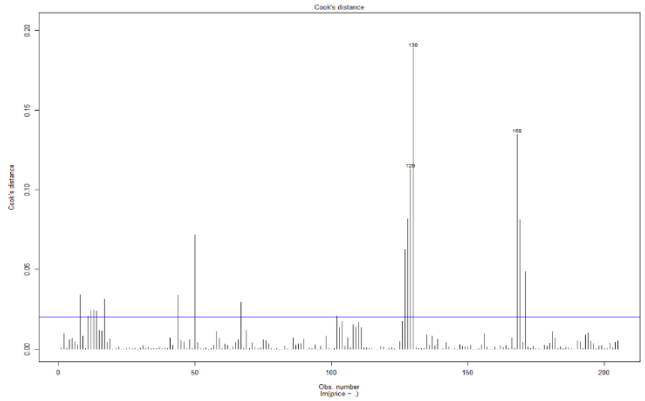
From this table, we can deduce that none of these variables need to be removed from our model. We can conduct further analysis (in appendix) to find the p-values of each variable and the overall model. Since the p-values of both are statistically significant, we can say this model does not suffer heavily from multicollinearity. Given how few variables this model has, if we chose to remove more to prevent linear dependence, it is likely we would end up underfitting our model.

2.3 Influential Points

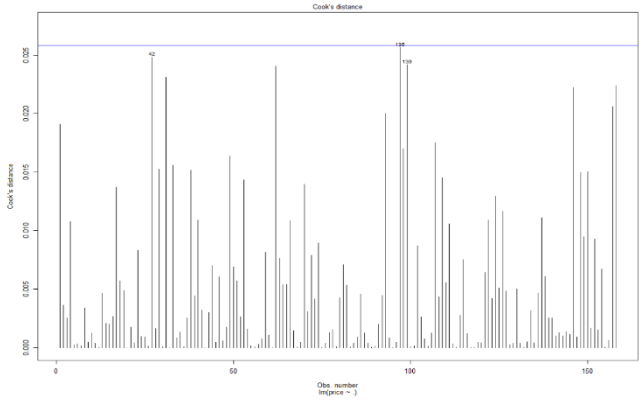
The next step in our model is finding data points of high influence that could be distorting our coefficient estimates, making our model less accurate. We use Cook's distance to identify the most influential point we wish to remove from the dataset then recreate the MLR model. Cook's distance measures how much a fitted value changes when we remove a point from the data set; highly influential points lead to a larger Cook's Distance. We say the Cook's Distance is large enough when it crosses a certain threshold. A commonly used threshold is given by $\boldsymbol{4/(n-p)}$. We repeat this process of removing the most influential point then recreating the model until no more reaches the threshold. We cannot remove all values above the threshold at once because removing the value with highest Cook's distance would change the model as well as the threshold which must be taken into account.

$$\boldsymbol{D_i = \frac{1}{p}\left(\frac{h_{ii}}{1-h_{ii}}\right) r_i^2}$$

Where p is the number of predictors, $\boldsymbol{r_i}$ being the residuals and $\boldsymbol{h_{ii}}$ being leverage points. The plots below show before and after removing all the influential points.



Before



After

The blue line represents the threshold at which we determine if a point is influential, on this graph, we can see there are no more influential points, so we recalculate the coefficient estimates and test the fit of the BIC model again.

2.4 Final Model

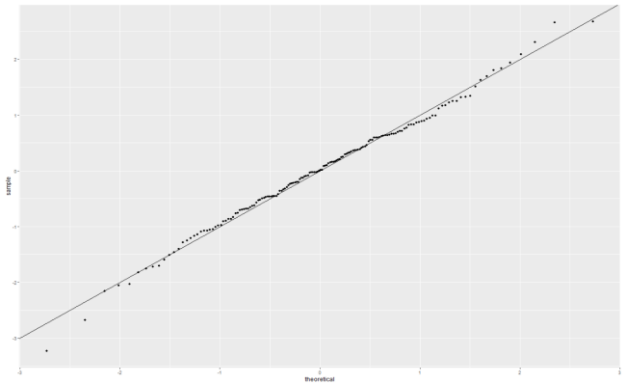
$$\begin{aligned}\log(\boldsymbol{price_i}) = \boldsymbol{3.91} &+ \boldsymbol{0.0624 carwidth_i} + \boldsymbol{0.00101}\sqrt{\boldsymbol{curbweight_i}} + \boldsymbol{0.00439}\sqrt{\boldsymbol{horsepower_i}} \\ &- \mathbb{1}_{\{\boldsymbol{carbody_i = wagon}\}}\boldsymbol{0.278} - \mathbb{1}_{\{\boldsymbol{carbody_i = hatchback}\}}\boldsymbol{0.265} - \mathbb{1}_{\{\boldsymbol{carbody_i = sedan}\}}\boldsymbol{0.226} \\ &+ \mathbb{1}_{\{\boldsymbol{drivewheel_i = rwd}\}}\boldsymbol{0.00845}\end{aligned}$$

The model formula above represents the linear model derived from the variables selected by the BIC. Note that the response variable is log(price) not price therefore we must take into consideration that the relationship seen may not be inherited by price itself. ***E(log(price))*** does not equal ***log(E(price))*** meaning we cannot just take the exponential of ***E(log(price))*** to calculate the price for unseen observations.

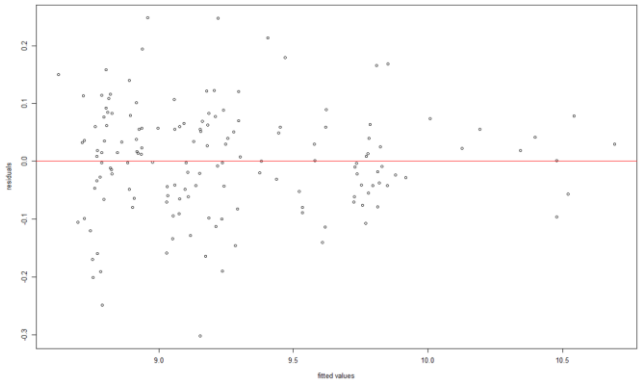
We can see all predictors have very small p-values < 5%, showing that they are all highly significant except for drive wheel rwd which has a p-value of 0.742. In addition, the coefficients of the levels for each categorical variable have a higher magnitude than the continuous variables. It is worth noting the continuous variables are much larger values so we cannot draw conclusions on which variables are the biggest influences in log(price). It is interesting to see curb weight have a positive coefficient as a heavier car would be less fuel efficient, but it could be due to manufacturing costs instead.

From this model, we obtain an $\bar{R}^2$ of 0.959, showing this model to have an extremely strong fit with the data set and the model's p-value < 5%, which indicates the overall model is statistically significant. A key issue with any data set is the curse of heteroscedasticity, we conducted the Breusch Pagan test to confirm if this was present in the model. With a 5% significance level, the p-value from this test was 0.71, showing we do not reject the null hypothesis that heteroscedasticity is not present.

In addition, we can create a QQ-plot and plot the residuals against fitted values of the final model to confirm if any structural changes are needed.



QQ-plot of BIC



Residual Plot

The QQ-plot shows that the residuals of the BIC model is extremely close to a normal distribution. Using a combination of both Breusch-Pagan test and a plot of the residuals, we conclude that we do not have heteroscedasticity.

3.0 Other Models

Using different variable selectors, we were able to create 2 other models which we will compare to the BIC model in section 3.

$$\boldsymbol{AIC}: \log(\boldsymbol{price_i}) = 3.10 + 0.0942\boldsymbol{carwidth_i} + 6.27 \times 10^{-3}\sqrt{\boldsymbol{horsepower_i}} - 0.0281\boldsymbol{highwaympg_i} + 6.02 \times 10^{-3}\boldsymbol{carheight_i} + 2.80 \times 10^{-11}exp\{\boldsymbol{compressionratio_i}\} - \mathbb{1}_{\{\boldsymbol{carbody_i=wagon}\}}0.0495 - \mathbb{1}_{\{\boldsymbol{carbody_i=hatchback}\}}0.0691 + \mathbb{1}_{\{\boldsymbol{carbody_i=hardtop}\}}0.163 + \mathbb{1}_{\{\boldsymbol{drivewheel_i=rwd}\}}0.132$$

$$\textbf{ADJR2}: \log(\boldsymbol{price_i}) = 2.21 + 0.119\boldsymbol{carwidth_i} + 3.35 \times 10^{-3}\sqrt{\boldsymbol{horsepower_i}} - 0.0278\boldsymbol{highwaympg_i} + 8.09 \times 10^{-3}\boldsymbol{carheight_i} + 2.66 \times 10^{-11}exp\{\boldsymbol{compressionratio_i}\} - 0.162\boldsymbol{boreratio_i} - \mathbb{1}_{\{\boldsymbol{carbody_i=wagon}\}}0.0345 - \mathbb{1}_{\{\boldsymbol{carbody_i=hatchback}\}}0.0464 + \mathbb{1}_{\{\boldsymbol{carbody_i=hardtop}\}}0.153 + \mathbb{1}_{\{\boldsymbol{drivewheel_i=rwd}\}}0.116$$

3.1 Alternative Variable Selectors

3.1.1 Akaike Information Criterion (AIC)

Similar to BIC, the Akaike Information Criterion (AIC) is another metric used to select the best model from a data set by finding which set of variables produces the smallest AIC.

$$\boldsymbol{AIC = -2\ln(L) + 2p}$$

With $\boldsymbol{L}$ and $\boldsymbol{p}$ having the same definitions as the BIC values, we can see how BIC penalizes additional parameters more than AIC by using a logarithmic function.

By using forward stepwise regression, we use this algorithm to test different values of AIC as the number of parameters in the model changes. Initially, we start the algorithm with the null model and add one variable at a time then test which variable leads to the largest fall in AIC. When found, the variable is permanently added to the model and the process is repeated with one less variable to choose from; the algorithm ends when adding a new variable leads to an increase in AIC. The resulting model is what we propose as our second option. However, stepwise regression is a greedy algorithm, implying that it will only consider the best option currently and not a better choice in the future. This means that stepwise regression can lead to useless models so further statistical analysis is needed to deduce the relevance of this proposed model.

```
Step:  AIC=-716.5
price ~ curbweight + horsepower + hatchback + wagon + rwd + carwidth +
    sedan + hardtop + compressionratio + carheight + highwaympg

             Df Sum of Sq    RSS     AIC
<none>                    5.5336 -716.50
+ boreratio   1  0.043692 5.4899 -716.12
+ enginesize  1  0.018799 5.5148 -715.19
+ stroke      1  0.009156 5.5244 -714.84
+ fwd         1  0.003203 5.5304 -714.61
+ carlength   1  0.002910 5.5307 -714.60
+ wheelbase   1  0.000110 5.5335 -714.50

Call:
lm(formula = price ~ curbweight + horsepower + hatchback + wagon +
    rwd + carwidth + sedan + hardtop + compressionratio + carheight +
    highwaympg, data = car_dat_trans)
```

Stepwise Regression Final Output in R

The final output from the algorithm in R shows AIC increasing from -716.50. If we were to add any more variable, the process stops to produce a model with the following variables:

- Curb weight
- Horsepower
- Car body (levels: hatchback, wagon, sedan, and hardtop)
- Drive wheel (levels: rwd)
- Compression Ratio
- Car height
- Highway mpg

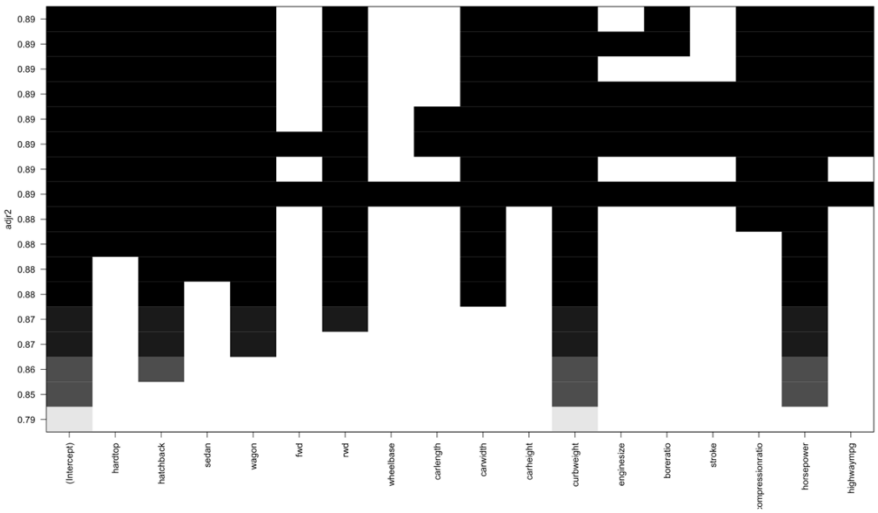
3.1.2 Adjusted R Squared

We choose to use Adjusted R Squared ($\bar{R}^2$) as the normal R Squared will always increase when we add variables to our model, defeating the purpose of a variable selector. Hence, we use $\bar{R}^2$ as it only increases when a significant variable is added. $\bar{R}^2$ is found by:

$$\bar{R}^2 = 1 - \frac{RSS(n-p)}{TSS(n-1)}$$

Where n is the sample size, p is the number of variables, RSS being the Residual Sum of Squares, and TSS being the Total Corrected Sum of Squares.

Using Best Subset Regression, we can find the display the highest $\bar{R}^2$ for each model of different subset sizes and plot our results in the table below.



$\bar{R}^2$ Values 1

The table shows the best model will have an $\bar{R}^2$ of 0.89 with the following variables:

- Car width
- Car height
- Curb weight
- Bore ratio
- Compression ratio
- Horsepower
- Car body (levels: hardtop, hatchback, sedan, and wagon)
- Drive wheel (level: rwd)

3.2 Model Assumptions for AIC and ADJ R-Squared

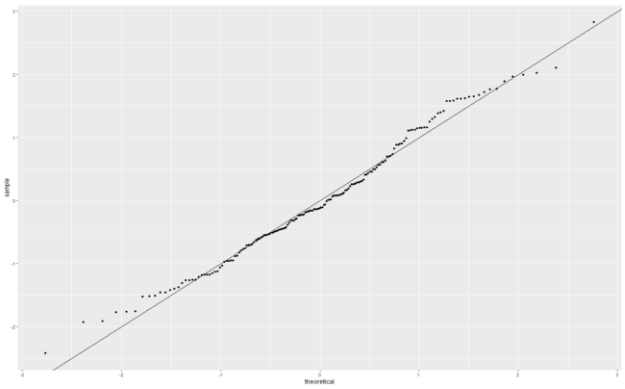
3.2.1 AIC

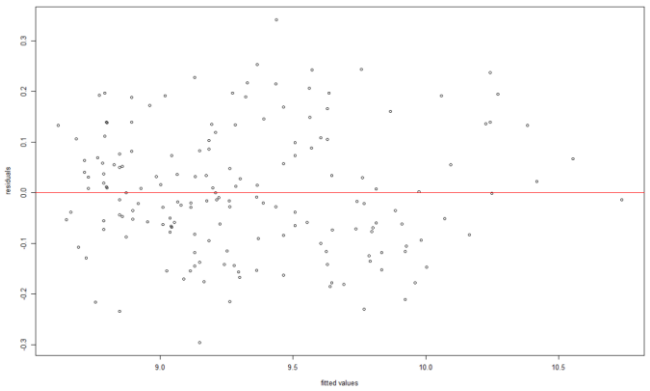
Using the same process as in section 2.2, we use the VIF to detect multicollinearity amongst predictors in our AIC model, from the following table:

Variable	VIF
Car width	4.6
Car height	2.1
Curb weight	10.6
Bore ratio	2.0
Compression ratio	1.5
Horsepower	4.0
Highway mpg	4.2
Hardtop	2.3
Hatchback	9.1
Sedan	10.3
Wagon	5.6
Rwd	2.2

We can deduce that we need to remove Curb weight and the Car body level 'Sedan' from our model to reduce the effects of multicollinearity since their VIFs > 10.

Below, we have added the QQ-plot of the residuals and residuals against fitted values plot respectively.



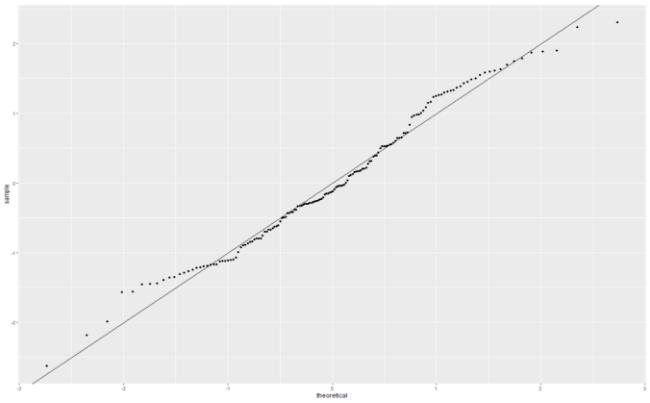
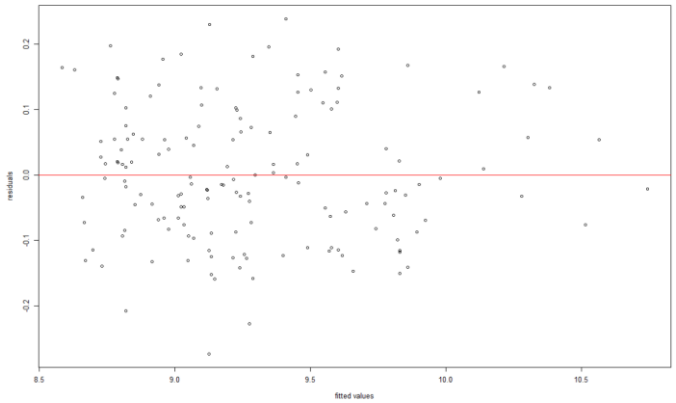


The QQ-plot shows that the residuals of the AIC model is very close to a normal distribution. Using a combination of both Breusch-Pagan test and a plot of the residuals, we conclude that we do not have heteroscedasticity. The p-value of the test is above 5% hence we do not have sufficient evidence to reject H0 and the plot of the residuals show no relationship between residuals and fitted values.

3.2.2 Adjusted R Squared

Using the same process as in section 3.2.1, we remove the following variables from our $\bar{R}^2$ model based on having a VIF > 10:

- Curb weight
- Car body level: Sedan

The QQ-plot shows that the residuals of the ADJR model is very close to a normal distribution. Using a combination of both Breusch-Pagan test and a plot of the residuals, we conclude that we do not have heteroscedasticity. The p-value of the test is above 5% hence we do not have sufficient evidence to reject H0 and the plot of the residuals show no relationship between residuals and fitted values.

3.3 Comparison of Three Models

3.3.1 Overview

To compare the 3 models, we can simply compare the statistics between each model in the table below:

Model	Number of Predictors	Adjusted R Squared	P-Value of Overall model	Is there heteroscedasticity?
BIC Model	5	0.959	< 2.2x10^-16	No
AIC Model	7	0.9305	< 2.2x10^-16	No
ADJR Model	8	0.9467	< 2.2x10^-16	No

From this table, we deduce that the BIC Model is the best model due to its higher adjusted R squared whilst having fewer predictors in its model. This is likely because the BIC Model captured the more important variables to consider such as curb weight. When transforming our variables in R, we noticed how curb weight had a correlation of 0.835 with log(price); higher than any other variable. Practically, curb weight affects other variables such as fuel consumption, which in this case is highway mpg. Theoretically, the values of highway mpg consider curb weight so we can remove curb weight with little consequence. However, the BIC model behaves opposite to AIC and ADJR2 models by including curb weight and not highway mpg, yet it obtains a higher $\overline{R}^2$. This could imply that our dataset is missing other variables that are affected by curb weight which could've been included in our other models. Of course, these 3 models do not only differ by curb weight, but the other variables included/excluded have a weaker correlation with log(price).

Homoscedasticity is an assumption for our MLR model. Heteroscedasticity means that the variance of our dependent variable (log-price) is constant which is equivalent to variance of the residuals being constant. We will be able to see any relationships by plotting the residuals against the fitted values. The BP-test essentially carries out a linear regression on the squared residuals of the original model then tests this. The null hypothesis is that we have homoscedasticity. Whilst we determined heteroscedasticity not to be present in any model through the BP-test, the BIC Model's p-value from the Breusch Pagan test was 0.71 compared to 0.05798 and 0.06188 for ADJR and AIC models respectively. Despite not rejecting the null for all models based on 5% significance, the ADJR and AIC models only barely passes this test for heteroscedasticity.

In addition, despite using VIF to reduce effects of multicollinearity, the AIC and ADJR models have more variables so it is possible those models will be influenced more by linear dependence than the BIC model.

3.3.2 Limitations & Conclusion

It is worth noting AIC's $\bar{R}^2$ could've been higher if we removed all influential points via Cook's Distance. However, the removal of these points in R for the AIC model would have led to the removal of too many data points to making the dataset far too small. Hence, we had added a limit to the number of iterations of removing influential points in the AIC model, unlike BIC or ADJR.

Furthermore, all 3 models have limitations due to the raw data set itself, with only 205 rows of data, it is likely our models would be less accurate if applied to a larger dataset. In addition, the dataset could be missing key variables which might influence car price; features like car colour and new technology might've made significant changes to all 3 models. Finally, we assume the data to be independent to carry out multiple linear regression. In reality, various components of a car affect each other, such as highway mpg and curb weight, so we cannot expect data points to be independent from each other.

Overall, we can conclude the best model is the proposed BIC model in section 2.4 due to its strong fit with fewer predictors; making it easier to produce estimated car prices. Despite its limitations, all 3 models suffer from the same issues mentioned above, so we achieve our objective with the BIC model.

Appendix