JHU Regression Models - Motor Trends Analysis

Richard Allen

28 March 2022

Synopsis

Looking at a data set of a collection of cars, you are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). You are particularly interested in the following two questions:

- "Is an automatic or manual transmission better for MPG?" (H_0: $mpg_{auto} = mpg_{manual}$, H_A: $mpg_{auto} \neq mpg_{manual}$)
- "Quantify the MPG difference between automatic and manual transmissions"

Executive Summary

This report sets out to examine the relationship between fuel economy and transmission type in the `mtcars` data set.

With only 32 observations but 10 explanatory variables with high degrees of collinearity making 1,024 possible models, the best linear model suffers from a high degree of uncertainty.

Additionally, there is a high degree of bias in the data set with respect to sampled manual and automatic cars, with most manual cars having characteristics independent of transmission type that favour better fuel economy than the automatic cars included.

Using a model selected by backwards regression and VIF elimination, the model suggests that, ***for the cars in the `mtcar` data set, holding all other variables constant, changing from automatic to manual transmission will increase the fuel economy by 2.15 mpg***. However, with a residual standard error of 2.308, **the null hypothesis that transmission has no effect on fuel economy is failed to be rejected**.

Data

The `mtcars` dataset can be accessed from any standard R installation using `view(mtcars)`.

The data 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).

There are two categorical variables, `am` and `vs`, which are numeric in the data set. For the purposes of this analysis, these are converted to factor variables and renamed to `transmission` and `cyl_alignment` respectively.

There is no information available on sampling methodology, or even if this represents a random sample of vehicles. As such, it cannot be generalised beyond the scope of this data set. No randomised control group was used, no causal inference can be made from studies of this data set.

Additionally, the cars listed represent 50 year old technology, any model trained on this data set would be obsolete on modern cars.

With such a small sample set for modelling, no split for training and testing was made as this would have further degraded the viability of a model. The modelling was trained against the full dataset.

Exploratory Data Analysis Summary

See Exploratory Data Analysis for detailed analysis.

Most manual transmission cars in the data set lie in "lighter" vehicles - those with smaller engines, lighter weight, higher rear axle ratios. These are all factors that influence the fuel economy significantly and will bias the prediction of the effect of transmission type on fuel economy.

Another correlation brought out is the number forward gears which is a function of transmission type. No manual cars have only 3 gears while no automatic cars have 5 gears. This variable is not used in modelling analysis.

Each variable shows a strong relationship with fuel economy.

Displacement (engine size) and horsepower both show strongly exponential relationships, and carburettors mildly so. The log of each of these variables is used for model analysis.

All variables display significant collinearity presenting a challenge for meaningful linear modelling.

Regression Analysis Summary

Please refer to Regression Analysis in the Appendix for full details of each section below including model test graphs.

Choosing a Model

Because of the small sample size and high degree of collinearity, standard automated variable selection methods such as stepAIC, ANOVA and lasso failed to produce a meaningful model. In the end, a manual backwards regression was used, starting with all variables (other than `gear` - see above), then eliminating one variable at a time and examining the effect on the Adjusted R^2 and RMSE values. The model with the highest Adjusted R^2 was selected, the process repeated for that model recursively until no further improvement in the Adjusted R^2 could be seen.

Two models were found from this process:

The most parsimonious: `mpg ~ wt + transmission + log.disp + log.carb + log.hp`

The highest prediction rate (by RMSE): `mpg ~ cyl + drat + wt + qsec + transmission + log.disp + log.carb + log.hp`

Using VIF analysis on these two models showed the highest prediction model to be largely unworkable due to many variables showing high collinearity.

Removing displacement from the most parsimonious brought the VIF scores to an acceptable level, thus the final model used is `mpg ~ wt + transmission + log.carb + log.hp`

Testing the Model

- Because of the small sample size, no train/test split of the data set was made, the full data set was used for both training and testing.
- Residuals are mostly normally distributed, however there is a spike in the positive tail.
- Residuals are mostly scattered around the zero, however in both tails they tend to positive. There are no obvious structures in the residual distribution apart from this.
- Normal QQ plot shows a distribution mostly along the theoretical line, however the tails show some right-skew.
- Scale-Location is mostly horizontal until around 21mpg, after which it rises. Residuals are scattered evenly around the red line. This slope is a reflection of the right-skew mentioned previously and is less than ideal.
- No points lie outside or near the 0.5 line on the Cook's Distance plot, therefore no points are considered outliers for review. Leverage is well within acceptable range.

Evaluating the model

- 87.2% of variation can be explained by this model.
- `log.carb` is least significant, however removing this from the model reduces Adjusted R^2 and increases RMSE.
- 62.5% of observed values lie within their 95% confidence interval. This model would not be accepted under any normal conditions.
- `transmission` is not significant by P-Value, however it is kept in the model for the purposes of answering the research question.
- The summary indicates that ***for the cars in the data set, all other variables being held constant, changing from automatic to manual will increase the fuel economy by 2.15 mpg***. However, with a residual standard error of 2.308, **the null hypothesis that transmission has no effect on fuel economy is failed to be rejected**.

Conclusion

While the model predicted an improvement in fuel economy by using manual transmission over automatic (all other variables being held equal), the predicted change was well within possible ranges due to the high standard error and therefore no inference can made regarding the effect of transmission and the null hypothesis that transmission does not affect fuel economy is failed to be rejected.

This study highlights the need for meaningfully sized sample sets with greater distribution of properties and the complexities arising from highly co-dependent variables.

Appendix

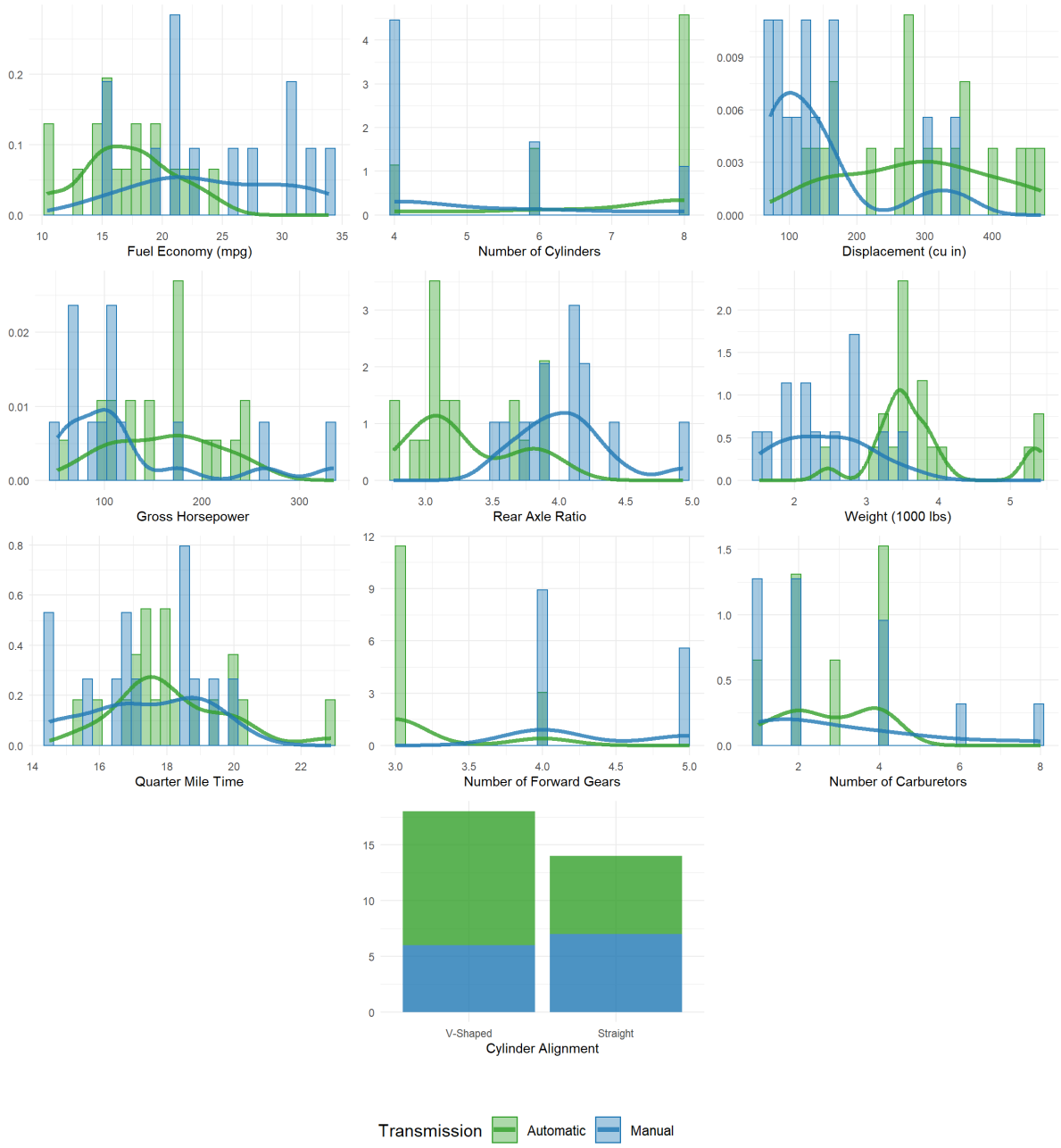
Exploratory Data Analysis

The `mtcars` data frame after transforming transmission and engine shape to factor variables:

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg          : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl          : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp         : num  160 160 108 258 360 ...
##  $ hp           : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat         : num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt           : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec         : num  16.5 17 18.6 19.4 17 ...
##  $ gear         : num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb         : num  4 4 1 1 2 1 4 2 2 4 ...
##  $ transmission : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ cyl_alignment: Factor w/ 2 levels "V-Shaped","Straight": 1 1 2 2 1 2 1 2 2 2 ...
```

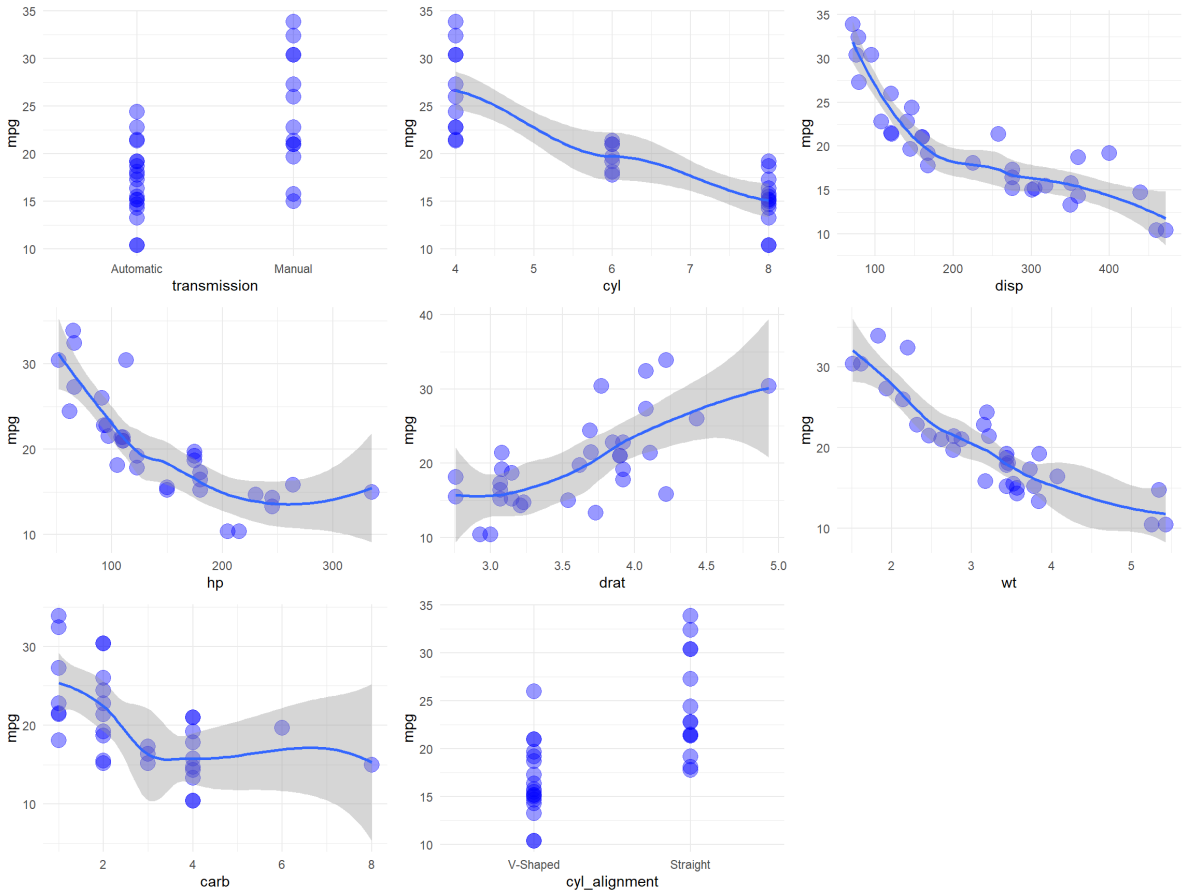
First examine the distribution of each variable against transmission type:

Density Function of Each Variable, Split by Transmission Type

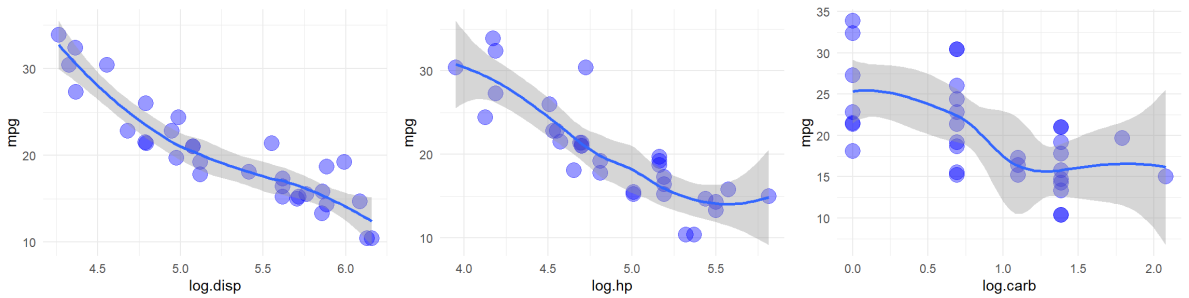


Next, prior to building regression models, check relationship type between mpg and each variable.

Relationship of Variables with Fuel Economy



Check log relationship with mpg for the possibly exponential variables:



- The log of Displacement and horsepower both show a much more linear relationship with fuel economy and will be used for modelling.
- Count of carburettors hasn't produced anything meaningful, it's more likely being pulled by the two data points at 6 and 8, though the relationship is more linear now. The log will be used for model analysis but only kept for final modelling with strong reasons.

Finally, check for collinearity between numeric variables (excluding fuel economy).

- all variables display significant collinearity which should be dealt with in model analysis

Regression Analysis

Choosing a Model

Variable Selection

Conduct backward elimination, selecting the model with highest Adjusted R^2 value. Repeat iteratively until suggested models no longer show improved Adjusted R^2.

```
lm_table <- tibble(model = character(), AR2 = numeric(), RMSE = numeric())
df_model <- df_cars %>% select(-gear)
lm1 <- lm(formula = mpg ~ . -cyl, data = df_model)
lm2 <- lm(formula = mpg ~ . -log.disp, data = df_model)
lm3 <- lm(formula = mpg ~ . -drat, data = df_model)
lm4 <- lm(formula = mpg ~ . -wt, data = df_model)
lm5 <- lm(formula = mpg ~ . -qsec, data = df_model)
lm6 <- lm(formula = mpg ~ . -cyl_alignment, data = df_model)
lm7 <- lm(formula = mpg ~ . -log.hp, data = df_model)
lm8 <- lm(formula = mpg ~ . -log.carb, data = df_model)
```

model	lm6	lm5	lm1	lm3	lm2	lm8	lm4	lm7
AR2	0.8390489	0.8389047	0.8377913	0.8375064	0.8323428	0.8317050	0.8311653	0.8286959
RMSE	2.049903	2.050821	2.057896	2.059702	2.092173	2.096148	2.099506	2.114805

Model 6 dropping `cyl_alignment` gives the best result. Drop from data frame and cycle through again:

```
df_model <- df_model %>% select(-cyl_alignment)
lm11 <- lm(formula = mpg ~ . -cyl, data = df_model)
lm12 <- lm(formula = mpg ~ . -log.disp, data = df_model)
lm13 <- lm(formula = mpg ~ . -drat, data = df_model)
lm14 <- lm(formula = mpg ~ . -wt, data = df_model)
lm15 <- lm(formula = mpg ~ . -qsec, data = df_model)
lm16 <- lm(formula = mpg ~ . -log.hp, data = df_model)
lm17 <- lm(formula = mpg ~ . -log.carb, data = df_model)
```

model	lm15	lm13	lm11	lm12	lm6	lm5	lm17	lm14
AR2	0.8455884	0.8442609	0.8441335	0.8392996	0.8390489	0.8389047	0.8386525	0.8381395
RMSE	2.051011	2.059809	2.060651	2.092361	2.049903	2.050821	2.096569	2.099900

Model 15: drop `qsec`

```
df_model <- df_model %>% select(-qsec)
lm21 <- lm(formula = mpg ~ . -cyl, data = df_model)
lm22 <- lm(formula = mpg ~ . -log.disp, data = df_model)
lm23 <- lm(formula = mpg ~ . -drat, data = df_model)
lm24 <- lm(formula = mpg ~ . -wt, data = df_model)
lm25 <- lm(formula = mpg ~ . -log.hp, data = df_model)
lm26 <- lm(formula = mpg ~ . -log.carb, data = df_model)
```

model	lm21	lm23	lm15	lm13	lm11	lm22	lm26	lm24
AR2	0.8503665	0.8503581	0.8455884	0.8442609	0.8441335	0.8438886	0.8425980	0.8424505
RMSE	2.060663	2.060720	2.051011	2.059809	2.060651	2.104795	2.113477	2.114467

Model 21: drop `cyl`

```
df_model <- df_model %>% select(-cyl)
lm31 <- lm(formula = mpg ~ . -log.disp, data = df_model)
lm32 <- lm(formula = mpg ~ . -drat, data = df_model)
lm33 <- lm(formula = mpg ~ . -wt, data = df_model)
lm34 <- lm(formula = mpg ~ . -log.hp, data = df_model)
lm35 <- lm(formula = mpg ~ . -log.carb, data = df_model)
```

model	lm32	lm21	lm23	lm31	lm35	lm15	lm34	lm13
AR2	0.8553433	0.8503665	0.8503581	0.8494520	0.8485923	0.8455884	0.8455575	0.8442609
RMSE	2.066229	2.060663	2.060720	2.107884	2.113894	2.051011	2.134974	2.059809

Model 32: drop `drat`

```
df_model <- df_model %>% select(-drat)
lm41 <- lm(formula = mpg ~ . -log.disp, data = df_model)
lm42 <- lm(formula = mpg ~ . -wt, data = df_model)
lm43 <- lm(formula = mpg ~ . -log.hp, data = df_model)
lm44 <- lm(formula = mpg ~ . -log.carb, data = df_model)
```

model	lm32	lm44	lm41	lm21	lm23	lm31	lm35	lm43
AR2	0.8553433	0.8540455	0.8533351	0.8503665	0.8503581	0.8494520	0.8485923	0.8484654

RMSE	2.066229	2.115014	2.120154	2.060663	2.060720	2.107884	2.113894	2.155064

No new models improve the R^2 value, model 32 is most parsimonious, however looking at the RMSE, model 6 has the best predictive rate.

```
fitR2 <- lm(mpg ~ wt + transmission + log.disp + log.carb + log.hp, df_cars)
fitRMSE <- lm(mpg ~ cyl + drat + wt + qsec + transmission + log.disp + log.carb + log.hp, df_cars)
```

Collinearity

```
vif(fitR2)
```

```
##           wt transmission     log.disp     log.carb       log.hp
##     6.014547     2.782875    12.287014     2.583308     7.753868
```

```
vif(fitRMSE)
```

```
##          cyl         drat           wt         qsec transmission     log.disp
##    12.342802     3.986027    10.296205     7.456551     4.616531    23.012337
##     log.carb       log.hp
##     4.290387     8.665107
```

```
vif(lm(mpg ~ wt + transmission + log.hp + log.carb, df_cars))
```

```
##           wt transmission       log.hp     log.carb
##     4.073262     2.451350     3.119289     2.371535
```

```
vif(lm(mpg ~ cyl + drat + wt + qsec + transmission + log.carb + log.hp, df_cars))
```

```
##          cyl         drat           wt         qsec transmission     log.carb
##    10.227642     3.983121     6.076469     6.545100     3.909631     3.537068
##       log.hp
##     7.481139
```

Displacement causes high VIF scores. Removing from the R^2 model reduces to an acceptable limit, however the RMSE model remains highly collinear.

```
fit <- lm(mpg ~ wt + transmission + log.carb + log.hp, df_cars)
```

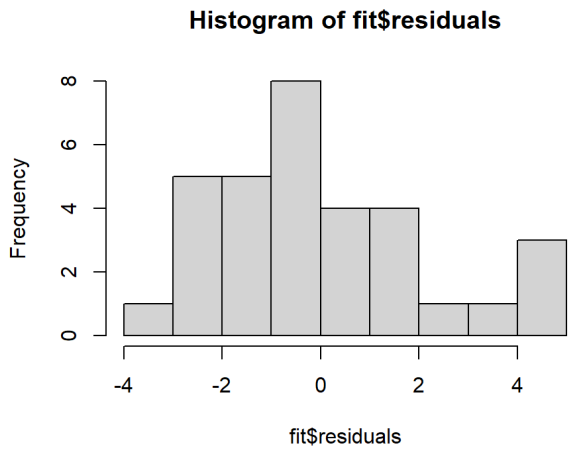
Final test for auto-collinearity of residuals using Watson test:

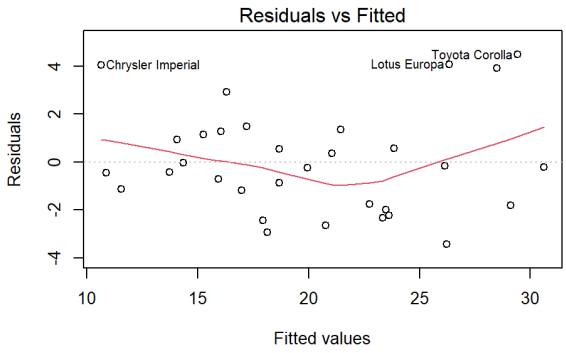
```
dwtest(fit)
```

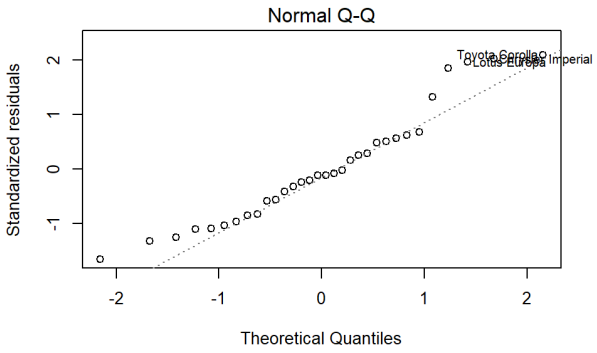
```
##
##  Durbin-Watson test
##
## data:  fit
## DW = 1.7854, p-value = 0.1496
## alternative hypothesis: true autocorrelation is greater than 0
```

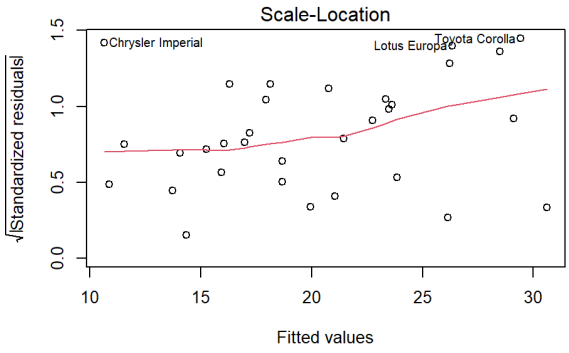
DW score is well below threshold of 2 with p-value of 0.15, the null-hypothesis is failed to be rejected, no evident auto-collinearity.

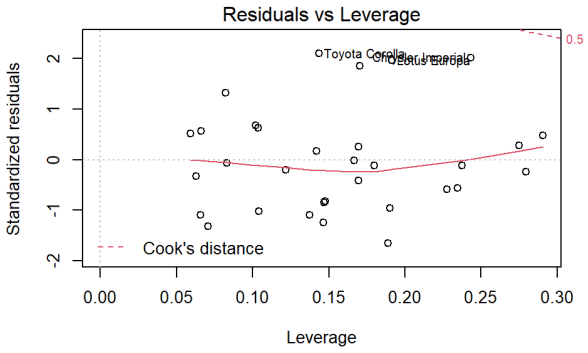
Testing the Model











Evaluating the model

```
##
## Call:
## lm(formula = mpg ~ wt + transmission + log.carb + log.hp, data = df_cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4252 -1.7724 -0.2315  1.1778  4.4742
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         55.8277     6.2994   8.862 1.77e-09 ***
## wt                  -2.2915     0.8551  -2.680 0.012395 *
## transmissionManual   2.1494     1.3007   1.652 0.110022
## log.carb            -0.8754     1.0843  -0.807 0.426554
## log.hp              -5.8323     1.5406  -3.786 0.000778 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.308 on 27 degrees of freedom
## Multiple R-squared:  0.8723, Adjusted R-squared:  0.8533
## F-statistic: 46.09 on 4 and 27 DF,  p-value: 1.101e-11
```

```
data.frame(predict(fit, interval="confidence")) %>%
    mutate(observed = mtcars$mpg) %>%
    mutate(in_ci = ifelse(observed >= lwr & observed <= upr,TRUE,FALSE)) %>%
    summarize(`% in CI` = round(mean(in_ci)*100,2))
```

```
##   % in CI
## 1    62.5
```