Course Project - Regression Models

This document is created for *"Course Project"* assignment in the framework of Regression Models course (part of Data Science Specialization by Johns Hopkins University Bloomberg School of Public Health on Coursera).

The Challenge

Working for *Motor Trend*, a magazine about the automobile industry. Looking at a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon. They are particularly interested in the following two questions:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between automatic and manual transmissions?

Executive Summary

It is statistically proved that transmission type affects fuel efficiency in cars. Cars with manual transmission can travel on average of 7.2 miles more per gallon. (With 95% chance, manual transmission cars travel 3.6 - 10.8 miles more per gallon.) Although transmission type can differentiate cars in respect of fuel efficiency but in order to predict mpg I found that cylinder number and weight are better predictors.

Data Overview

Data to be analysed for this exercise reside as a built-in object, called `mtcars`, comes along with any R distribution. This is an extract from the 1974 Motor Trend US magazine that comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models) Let's take a first look at it:

```
## '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 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

First few rows of the dataset:

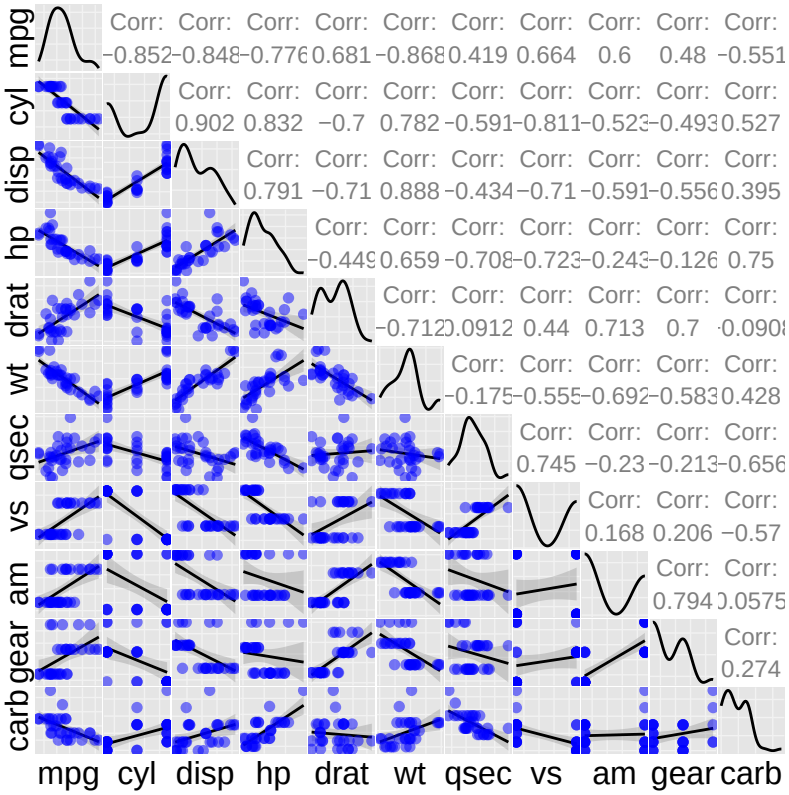
```
##                mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4     21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag 21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710    22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
```

Each row represents a model that can be seen in the row names and each column is an attribute of a particular model. Further description is available in R documentation (`?mtcars`), that is presented below:

Variable	Description
mpg	Miles/(US) gallon
cyl	Number of cylinders
disp	Displacement (cu.in.)
hp	Gross horsepower
drat	Rear axle ratio
wt	Weight (lb/1000)
qsec	1/4 mile time
vs	V/S
am	Transmission (0 = automatic, 1 = manual)
gear	Number of forward gears
carb	Number of carburetors

Exploratory Data Analyses

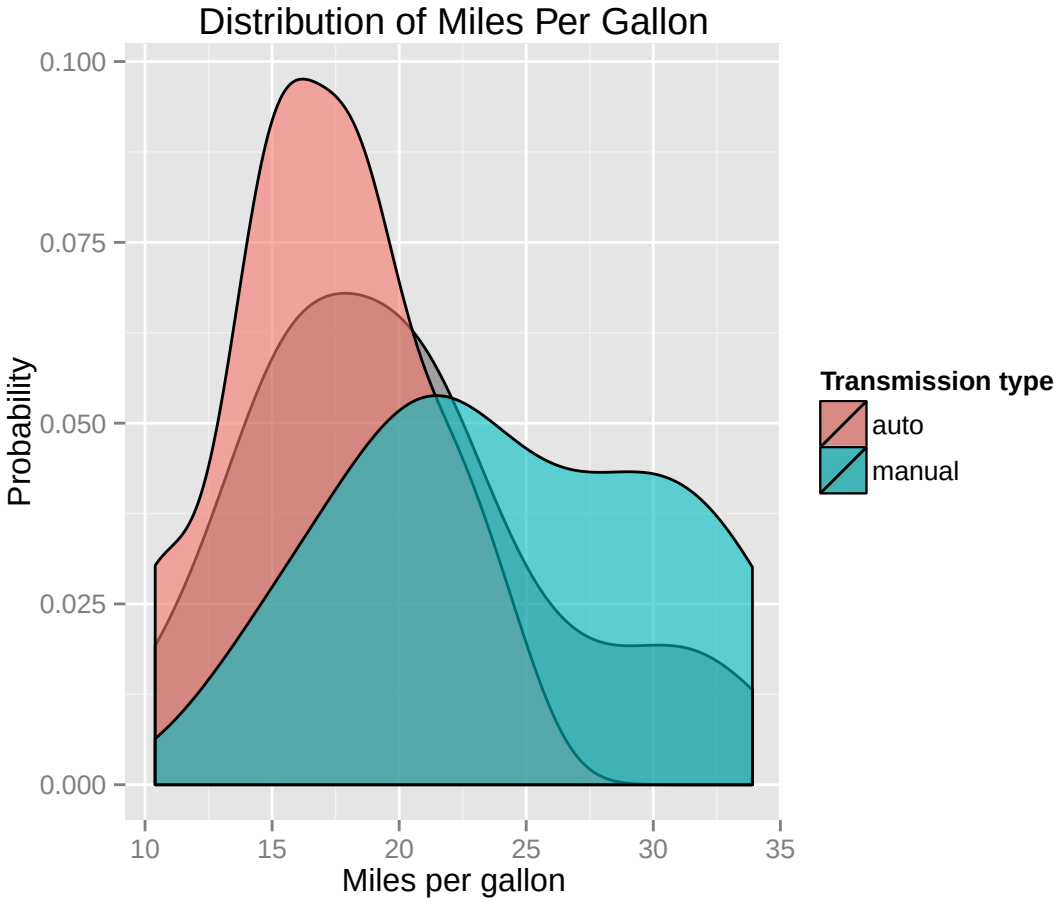
I start with some initial explaratory data analysis to get a better insight of the potential patterns in the data set. Below I create pairwise scatterplots for every variable.



This presents that `mpg` shows high correlation (>0.8) with number of cylinders, displacement and weight. Other notable (corr>0.5) variables might be the horsepower, rear axle ratio, cylinder alignment (V/S), transmission type or the number of carburetors. Additionally, one can point out in this table that there are

several variable combinations that show significant correlation, suggesting that including all of them will raise the threat of multicollinearity.

Let's also take a look at the outcome variable (miles per gallon) and see how it varies by automatic vs manual transmission.



At this point I will form a hypothesis that cars with automatic transmission can travel less distance per gallon then their counterparties with manual transmission. To check whether this pattern happened by random, I performed the necessary statistical analyses in the subsequent chapters.

Model fitting and inference

First I test if an automatic or manual transmission better for MPG. My initial hypothesis is that cars with automatic transmission consume more fuel, thus have lower range, than cars with manual transmission.

Population	Observed Mean	Standard Deviation
Entire population	20.09	6.03
Cars with automatic transmission	17.15	3.83
Cars with manual transmission	24.39	6.17

Simple linear regression

I fit a linear regression model on the type of transmission (as categorical predictor) having miles per gallon as outcome. The estimations in the table below are in comparison to automatic transmission.

```
##
## Call:
## lm(formula = mpg ~ factor(am), data = mtc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## factor(am)1    7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

This says that **in the group of cars with manual transmission we expect a 7.2449 miles per gallon efficiency improvement (in the confidence interval of 3.6415096, 10.848369) comparing to group of cars with automatic transmission.** Intercept (17.1474) represents the mean of "mpg" for automatic transmission group. Since the p value for this difference is so close to zero the mean difference between the two groups is statistically significance.

Although the difference is proved, the model above has low power to explain variance in `mpg`: R Squared equals 0.3598 that implies that this simple linear regression explains only 35.98% of outcome variable variation.

Multivariable model

Using a single categorical variable to estimate expected `mpg` is not very precise. In order to create a better prediction model I examined additional variables that are available in `mtcars` dataset. My goal is to determine the optimal set of variables that predicts `mpg` outcome for any given model by knowing its certain attributes.

Going forward I will treat `am` (transmission type) and `vs` (cylinder alignment) as factors since the numbers they represent do not correspond to anything in the physical world rather encode categories.

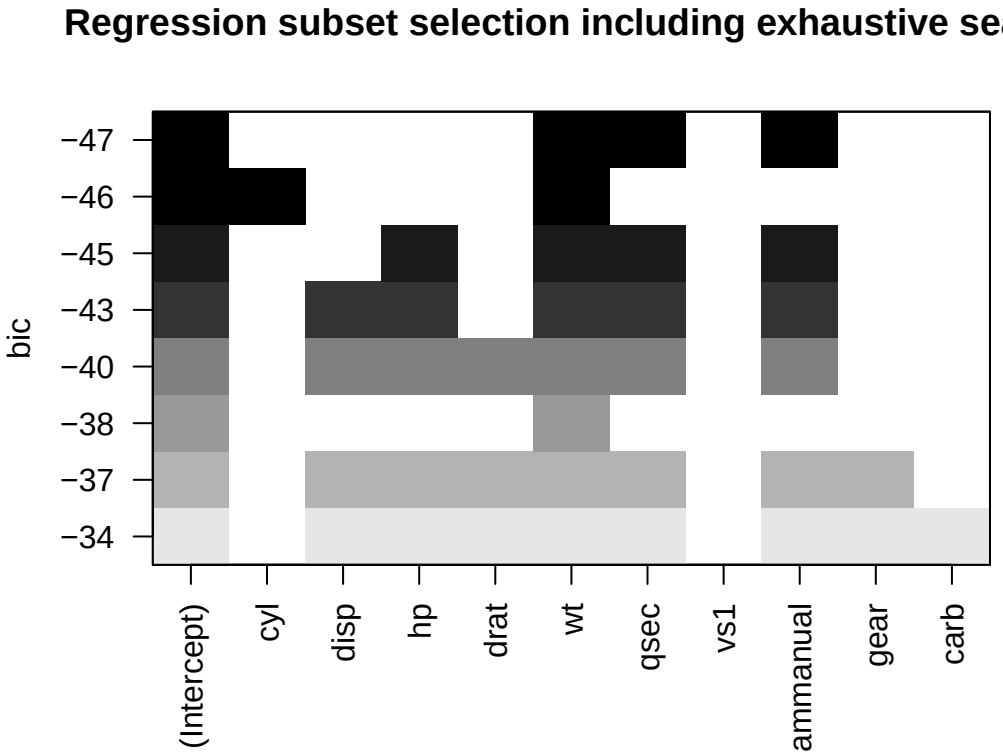
Stepwise variable selection

The first, most obvious try for variable selection is the stepwise algorythm, provided by `step()` function. This function executes stepwise feature selection by fetching through all the options between empty (interception only) and full (including all variables) models. This process can be performed in two directions; empty to full (forward) or full to empty (backward). The decision criterion, where iteration stops, is based on Akaike Information Criterion (AIC); selection process stops when the next variable exclusion/inclusion would not decrease the value of AIC anymore.

The following model has been selected in a *backward* stepwise process as the best fit that incorporates the following variables: *(Intercept), wt, qsec, ammanual* with the corresponding coefficients of *9.6177805, -3.9165037, 1.225886, 2.9358372*.

Regression subset selection including exhaustive search

To verify the selected variables, I performed a different method for variable selection provided by `leaps` R library. This result proofed the previously selected three variables. Exhaustive search method use Bayesian Information Criteria (BIC) to value model performance. (Model with the lowest value is considered as "best" fit.)



Model valuation

So both model selection algorythm returned with the same set of features that resulted the following model:

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## ammanual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
```

```
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

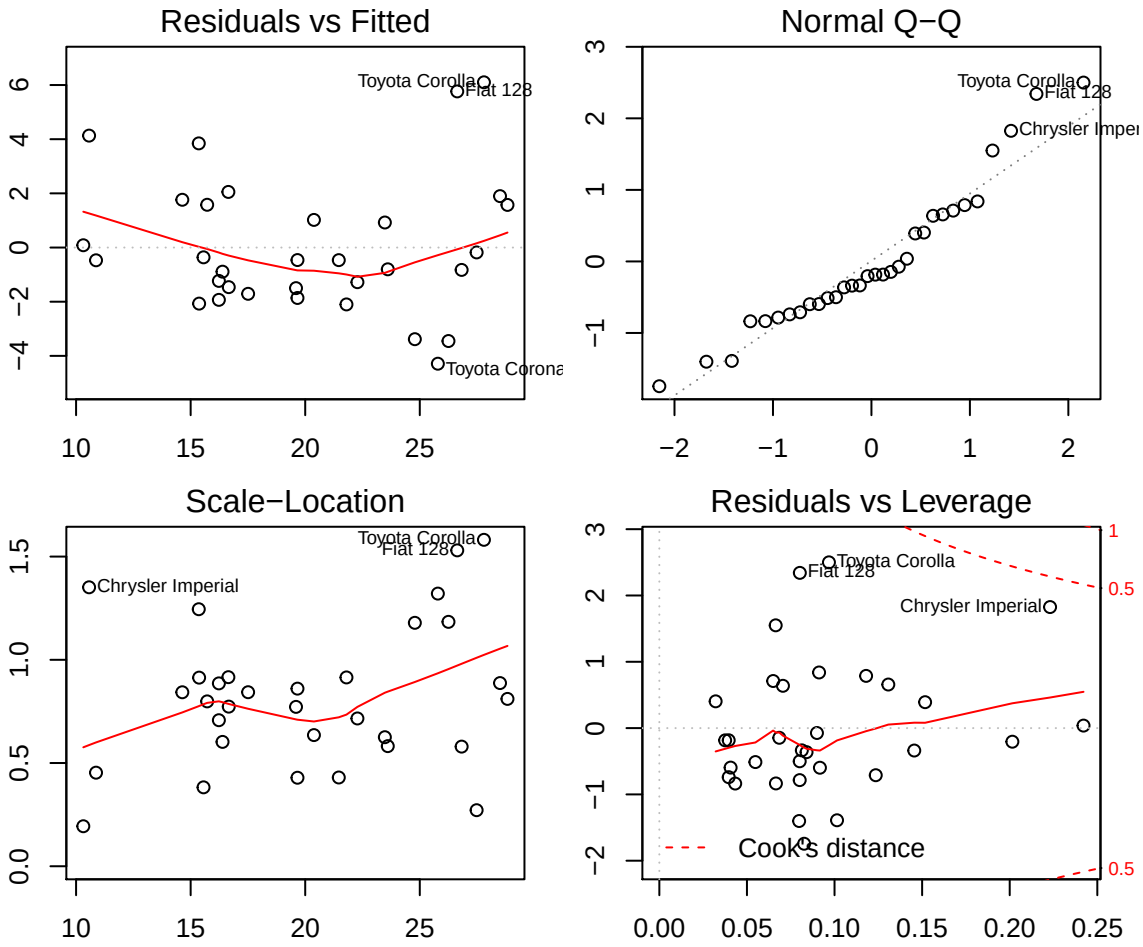
Looking at P values of coefficients, I see that intercept became insignificant thus we can not be sure if it is different from zero or not. This situation hinders the interpretation capability of the model so I went forward to see what other options I had. From exhaustive feature search figure I pointed out that a model including number of cylinders & weight as predictors proved to be the second best according to Bayesian Information Criteria calculation.

```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtc)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2893 -1.5512 -0.4684  1.5743  6.1004
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.6863     1.7150  23.141  < 2e-16 ***
## cyl          -1.5078     0.4147  -3.636 0.001064 **
## wt           -3.1910     0.7569  -4.216 0.000222 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.568 on 29 degrees of freedom
## Multiple R-squared:  0.8302, Adjusted R-squared:  0.8185
## F-statistic: 70.91 on 2 and 29 DF,  p-value: 6.809e-12
```

Although this model is not considered as the best for prediction (higher AIC, BIC and lower R-Squared) but it is way better interpretable, considering that it is simplier and even the estimated coefficients are statistically significant. Therefore I chose this model to proceed with.

Regression diagnostics

In this section, I performed some diagnostics on the final model.



Residuals vs fitted values By plotting residuals versus the fitted values, we're looking for any sign of a particular pattern. Scale-Localtion figure basically presents the same thing but plots fitted values versus standardized residuals. In my model they do not show any notable pattern.

Normal Q-Q plot - Residual Normality The normal Q-Q plot serves as an indicator to detect deviance from normal distribution in residuals. Due to some items on the right tail indicates that residuals do not follow exactly normal but a slightly left skewed distribution.

Residual vs Leverage Plot shows that there is no high leverage data points in the population, however there are some high residual points. But these do not alter the regression, so they reside below the critical 0.5 Cook's distance.