Regression Models - Report (OF)

Question

Using the mtcars data set from R, answer the following questions:

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

Executive Summary

It is not possible to answer the question with the provided data.

At first sight it looks as if automatic cars get less millage per gallon, however in the data set automatic vehicles are also heavier (692 lb). But the data does not allow to know if this is an intrinsic characteristic of automatic vehicles or if the sample selection was falwed.

Adding the transmission type to a model of miles per gallon versus weight results in a model with a negative coefficient for the amManual feature, suggesting that when controlling for weight automatic cars are more efficient. But the model is notstatistically significantly better model (P-value: 0.9879) when compared with the simple model ussing annova.

The best model includes the following variables

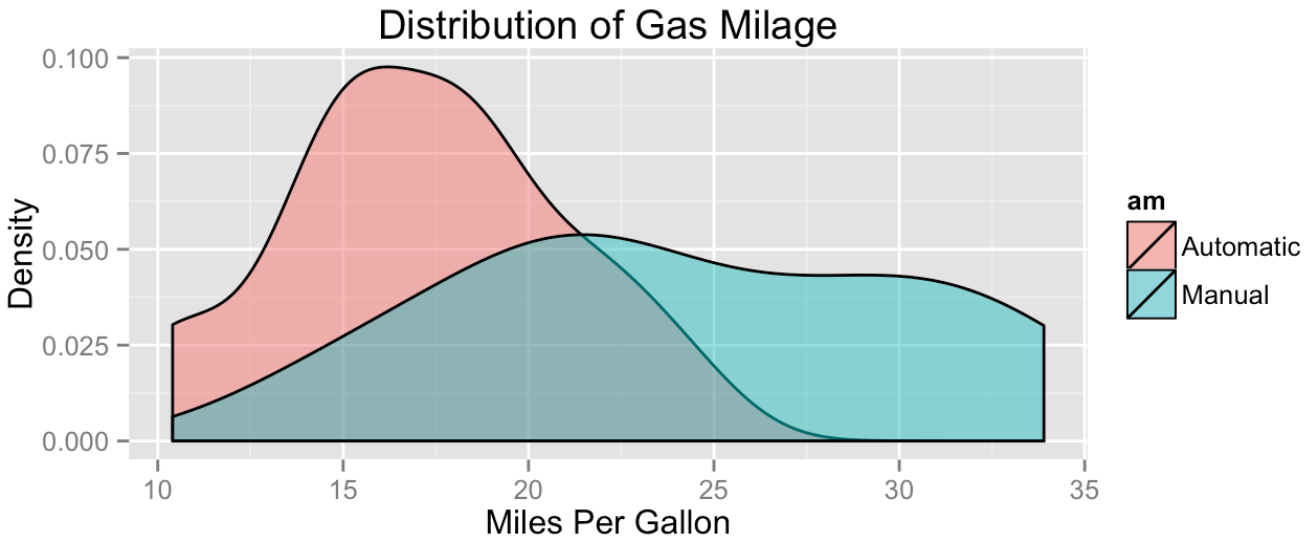
In order to conclude if automatic cars are less efficient or not, we will need to research if automatic cars are intrinsically heavier than manual ones.

Data cleaning

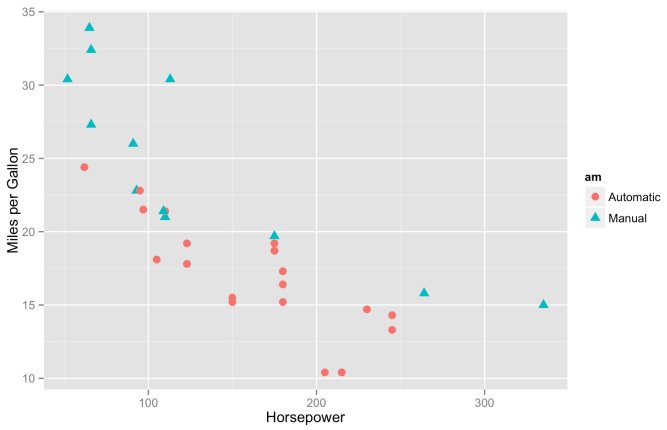
I created a second data set ‘mtcarsX’ with the column ‘am’ converted to a factor variable. I used the original data set to calculate the correlations and the modified one to fit the models.

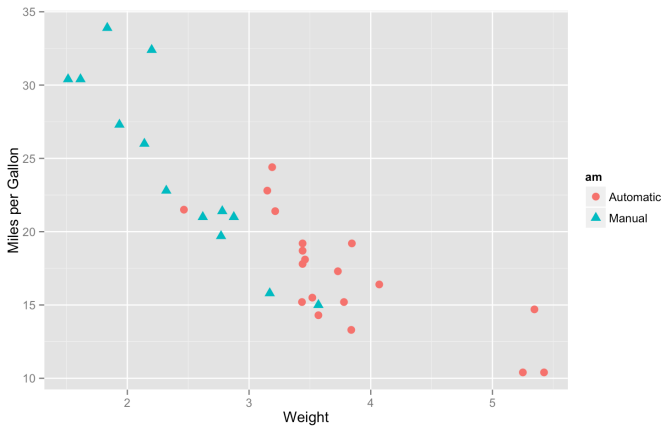
Explorartory Data Analysis

First I compared the transmission type versus the fuel efficiency using a density plot:



The results suggest that Manual transmission cars are better than automatic ones, but it is not sufficient for a conlcussion because other factors as the horse power, cylinders, or weight could influence the fuel consumption.





This shows that in this data set the automatic cars are both heavier and have more horse power than the manual ones, which could at least partially explain the higher full consumption of automatic vehicles observed in the density graph.

Building a linear model

Model Selection Strategy

To select between the possible models I decided to guide the search by the correlation between the variables, choose the variable with the highest correlation to it and then add other variables to the model then choosing the one with the lowest P-Value.

Then repeat the steps with the two predictors model to obtain a three predictors model, then repeat the process until adding variables do not result in better models.

To compare the models I used annova.

```
##          mpg      wt       am
## mpg   1.0000 -0.8677  0.59983
## cyl  -0.8522  0.7825 -0.52261
## disp -0.8476  0.8880 -0.59123
## hp   -0.7762  0.6587 -0.24320
## drat  0.6812 -0.7124  0.71271
## wt   -0.8677  1.0000 -0.69250
## qsec  0.4187 -0.1747 -0.22986
## vs    0.6640 -0.5549  0.16835
## am    0.5998 -0.6925  1.00000
## gear  0.4803 -0.5833  0.79406
## carb -0.5509  0.4276  0.05753
```

Selected model

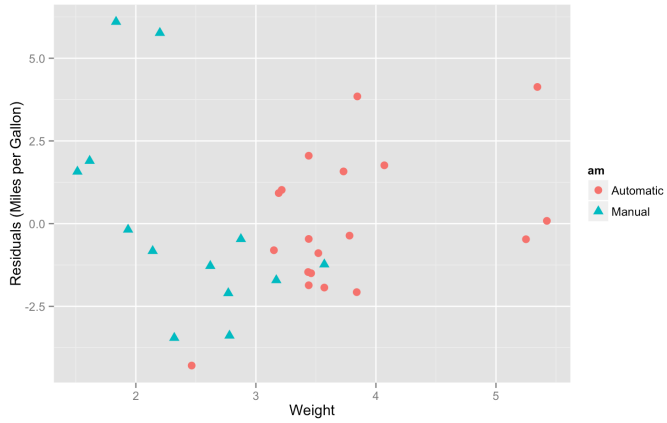
The selected model is `mpg ~ wt + cyl`, adding `am` to this model does not result in a significantly better model.

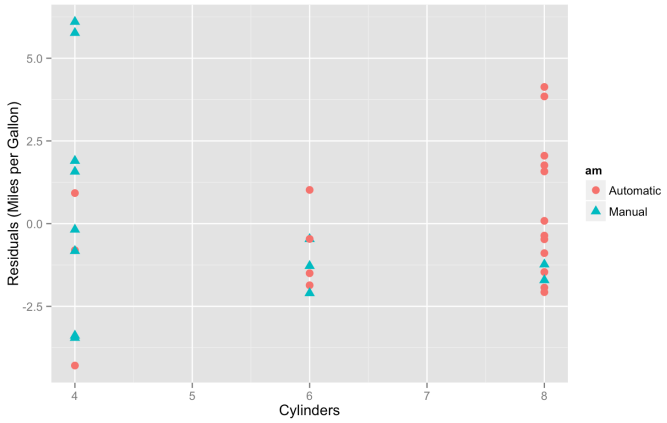
```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcarsX)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.289 -1.551 -0.468  1.574  6.100
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.686      1.715   23.14  < 2e-16 ***
## wt            -3.191      0.757   -4.22  0.00022 ***
## cyl           -1.508      0.415   -3.64  0.00106 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.57 on 29 degrees of freedom
## Multiple R-squared:  0.83,   Adjusted R-squared:  0.819
## F-statistic: 70.9 on 2 and 29 DF,  p-value: 6.81e-12
```

Diagnostigs & Confidence

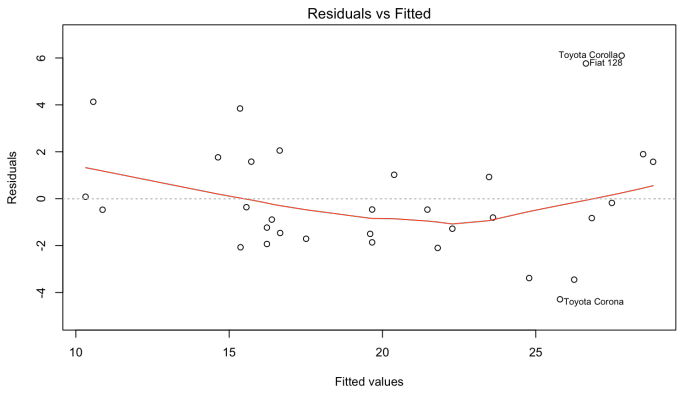
The following is the 95% confidence interval for the model coeficients.

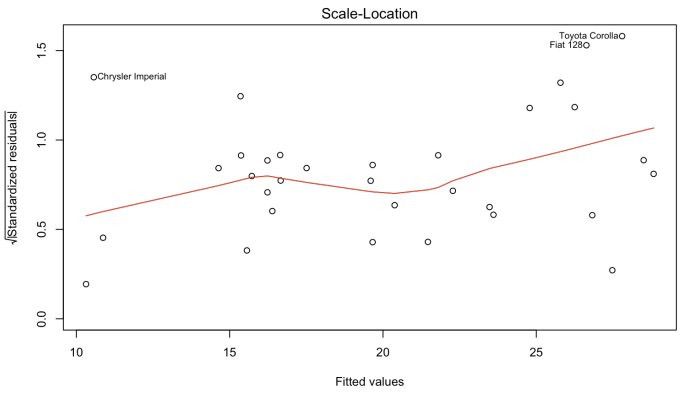
```
##                2.5 %  97.5 %
## (Intercept) 36.179 43.1938
## wt          -4.739 -1.6429
## cyl         -2.356 -0.6597
```

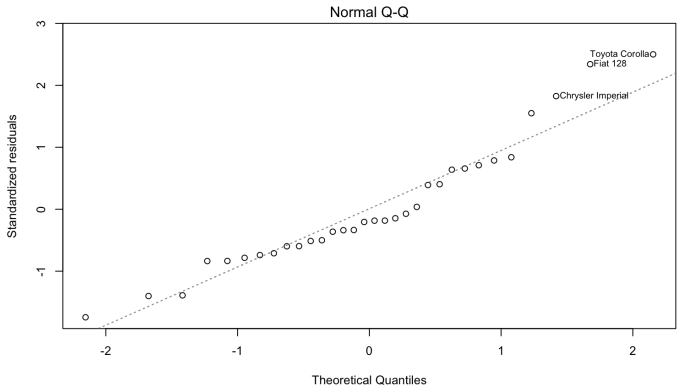


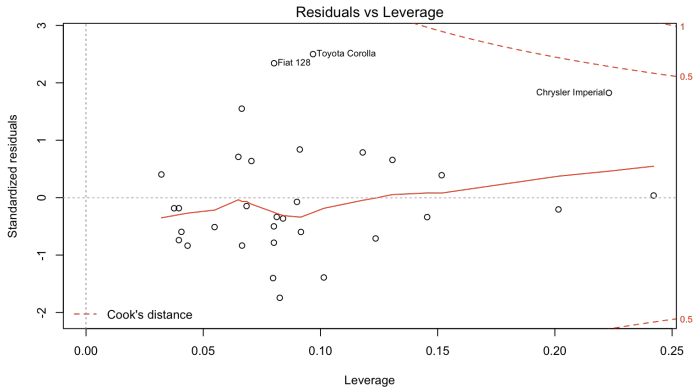


The selected model seems to be a good fit for the data according to the diagnostig plots.









Appendices

Data Description

Source (http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html)

A data frame with 32 observations on 11 variables.

- 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

Sample of Fitted Models

```
##
## Call:
## lm(formula = mpg ~ wt, data = mtcarsX)
##
## Coefficients:
## (Intercept)           wt
##       37.29        -5.34
```

```
##
## Call:
## lm(formula = mpg ~ wt + cyl, data = mtcarsX)
##
## Coefficients:
## (Intercept)           wt          cyl
##       39.69        -3.19        -1.51
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt + cyl
## Model 2: mpg ~ wt + cyl + disp
##   Res.Df RSS Df Sum of Sq   F Pr(>F)
## 1     29 191
## 2     28 188  1      2.68 0.4   0.53
```