Effect of Automatic vs Manual transmission on MPG

Swathi Vijayakumar

December 24, 2015

Executive Summary

The purpose of this paper is to determine the relationship between a set of variables and miles per gallon. Specifical we want to determine if a manual vs an automatic transmision gives better gas milage. If so, we will quantify the MPG difference between the two engins. For this analysis we will be using simple and multi regression along with exploratory analysis to answer the required question. Analysis shows that over all manula transmission does offer a slight benifit in MPG by about 1.68.

Data Processing

```
library(ggplot2); library(leaps); library(corrplot); library(dplyr); library(lattice)
```

```
##
## Attaching package: 'dplyr'
##
## The following objects are masked from 'package:stats':
##
##     filter, lag
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(grid); library(gridExtra)
data("mtcars")
```

Data set summaries

```
summary(mtcars, n = 3)
```

```
##       mpg             cyl             disp             hp
##  Min.   :10.40   Min.   :4.000   Min.   : 71.1   Min.   : 52.0
##  1st Qu.:15.43   1st Qu.:4.000   1st Qu.:120.8   1st Qu.: 96.5
##  Median :19.20   Median :6.000   Median :196.3   Median :123.0
##  Mean   :20.09   Mean   :6.188   Mean   :230.7   Mean   :146.7
##  3rd Qu.:22.80   3rd Qu.:8.000   3rd Qu.:326.0   3rd Qu.:180.0
##  Max.   :33.90   Max.   :8.000   Max.   :472.0   Max.   :335.0
##       drat             wt             qsec             vs
##  Min.   :2.760   Min.   :1.513   Min.   :14.50   Min.   :0.0000
##  1st Qu.:3.080   1st Qu.:2.581   1st Qu.:16.89   1st Qu.:0.0000
##  Median :3.695   Median :3.325   Median :17.71   Median :0.0000
##  Mean   :3.597   Mean   :3.217   Mean   :17.85   Mean   :0.4375
##  3rd Qu.:3.920   3rd Qu.:3.610   3rd Qu.:18.90   3rd Qu.:1.0000
##  Max.   :4.930   Max.   :5.424   Max.   :22.90   Max.   :1.0000
##        am              gear            carb
##  Min.   :0.0000   Min.   :3.000   Min.   :1.000
##  1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:2.000
##  Median :0.0000   Median :4.000   Median :2.000
##  Mean   :0.4062   Mean   :3.688   Mean   :2.812
##  3rd Qu.:1.0000   3rd Qu.:4.000   3rd Qu.:4.000
##  Max.   :1.0000   Max.   :5.000   Max.   :8.000
```

```
str(mtcars)
```

```
## '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 ...
```

```
dim(mtcars)
```

```
## [1] 32 11
```

Exploratory Analysis

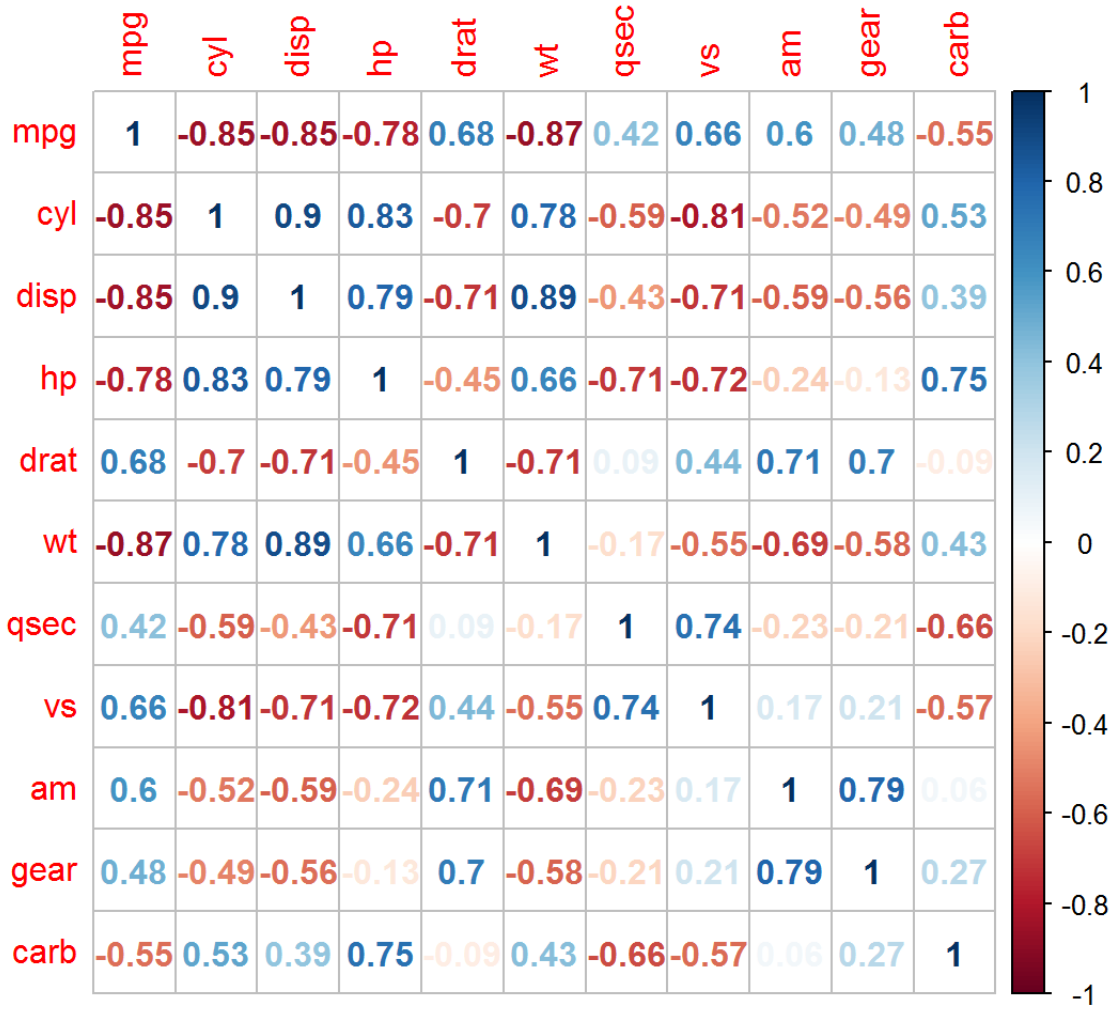
Correlation and variable selection

The variables to be used for the analysis in this paper were based on correlation to MPG. The variables that show the highest correlation to MPG are cyl, disp, hp, drat, wt and am.

```
cor = cor(mtcars)
cor
```

```
##             mpg        cyl       disp         hp        drat         wt
## mpg   1.0000000 -0.8521620 -0.8475514 -0.7761684  0.68117191 -0.8676594
## cyl  -0.8521620  1.0000000  0.9020329  0.8324475 -0.69993811  0.7824958
## disp -0.8475514  0.9020329  1.0000000  0.7909486 -0.71021393  0.8879799
## hp   -0.7761684  0.8324475  0.7909486  1.0000000 -0.44875912  0.6587479
## drat  0.6811719 -0.6999381 -0.7102139 -0.4487591  1.00000000 -0.7124406
## wt   -0.8676594  0.7824958  0.8879799  0.6587479 -0.71244065  1.0000000
## qsec  0.4186840 -0.5912421 -0.4336979 -0.7082234  0.09120476 -0.1747159
## vs    0.6640389 -0.8108118 -0.7104159 -0.7230967  0.44027846 -0.5549157
## am    0.5998324 -0.5226070 -0.5912270 -0.2432043  0.71271113 -0.6924953
## gear  0.4802848 -0.4926866 -0.5555692 -0.1257043  0.69961013 -0.5832870
## carb -0.5509251  0.5269883  0.3949769  0.7498125 -0.09078980  0.4276059
##             qsec         vs          am       gear        carb
## mpg   0.41868403  0.6640389  0.59983243  0.4802848 -0.55092507
## cyl  -0.59124207 -0.8108118 -0.52260705 -0.4926866  0.52698829
## disp -0.43369788 -0.7104159 -0.59122704 -0.5555692  0.39497686
## hp   -0.70822339 -0.7230967 -0.24320426 -0.1257043  0.74981247
## drat  0.09120476  0.4402785  0.71271113  0.6996101 -0.09078980
## wt   -0.17471588 -0.5549157 -0.69249526 -0.5832870  0.42760594
## qsec  1.00000000  0.7445354 -0.22986086 -0.2126822 -0.65624923
## vs    0.74453544  1.0000000  0.16834512  0.2060233 -0.56960714
## am   -0.22986086  0.1683451  1.00000000  0.7940588  0.05753435
## gear -0.21268223  0.2060233  0.79405876  1.0000000  0.27407284
## carb -0.65624923 -0.5696071  0.05753435  0.2740728  1.00000000
```

```
corrplot(cor, method = "number")
```




```
# Subsetting the desired variables
cars_mpg = select(mtcars, mpg:wt,am)

# changing cyl and am to the factor class.
cars_mpg$cyl = as.factor(cars_mpg$cyl)
cars_mpg$am = factor(cars_mpg$am,levels = c(0,1), labels = c("Automatic", "Manual"))

str(cars_mpg)
```

```
## 'data.frame':    32 obs. of  7 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
```

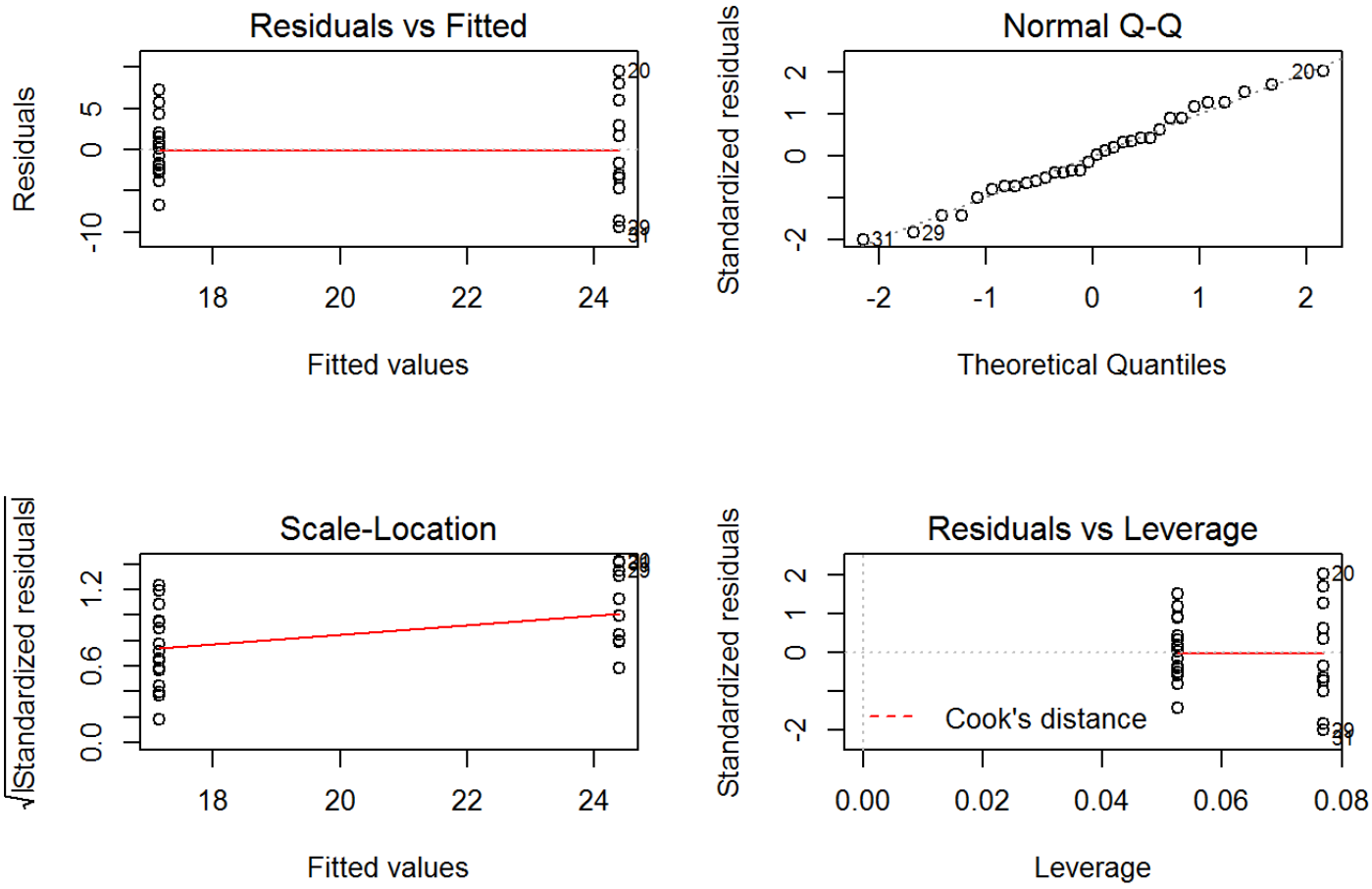
Regression models

Simple Linear Regression

```
y = cars_mpg$mpg; x = cars_mpg$am; n =length(y)
fit1 = lm(y ~ x)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## 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 ***
## xManual        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
```

```
par(mfrow = c(2,2))
plot(fit1)
```




```
# 95% confidence interval
sumCoef = summary(fit1)$coefficients
sumCoef[2,1] + c(-1, 1) * qt(.975, df = fit1$df) * sumCoef[2, 2]
```

```
## [1]  3.64151 10.84837
```

According to the linear fit coefficients, the average MPG for cars with manual transmission is 7.24 higher than for cars with automatic transmission with a 95% confidence interval of 3.64 and 10.84. The R-squared value of 0.3385 indicates that only 33.9% of the variation is explained by our model.

Multiple regression

For the multiple regression model only the variable that show a high correlation were chosen in addition to transmission(am). In this case, cyl, disp, hp and wt show the highest correlation. The correlation plot shows that cyl and disp also show a high correalation with each other indicating colinearity. For this reason, only one of the two should be included. cyl will be included in the multiple regression model.

```
fit2 = lm(mpg ~ wt + cyl + hp + am, data = cars_mpg)
summary(fit2)
```

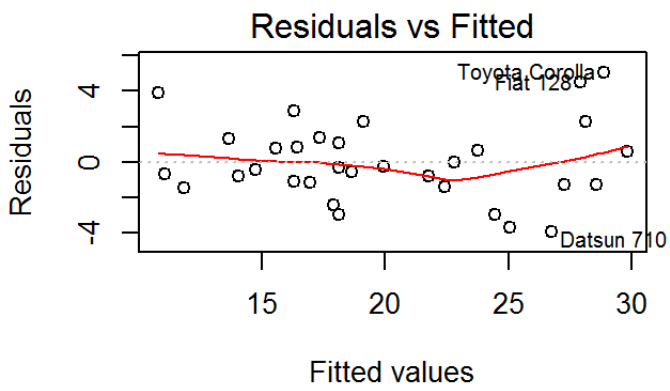
```
##
## Call:
## lm(formula = mpg ~ wt + cyl + hp + am, data = cars_mpg)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## wt          -2.49683    0.88559  -2.819  0.00908 **
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

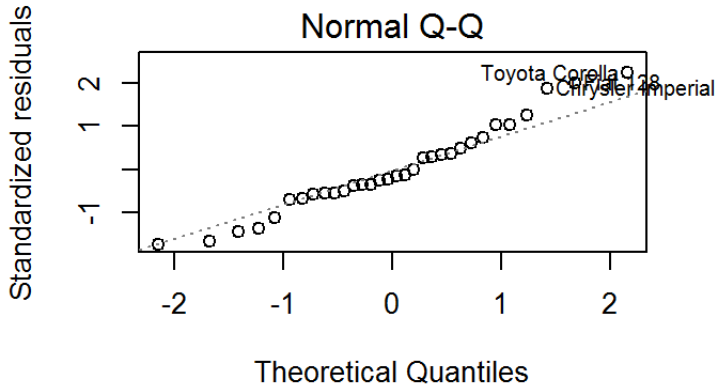
```
anova(fit1,fit2)
```

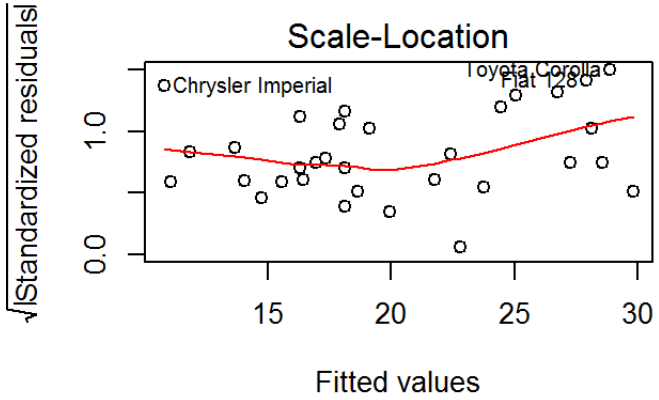
```
## Warning in anova.lmlist(object, ...): models with response '"mpg"' removed
## because response differs from model 1
```

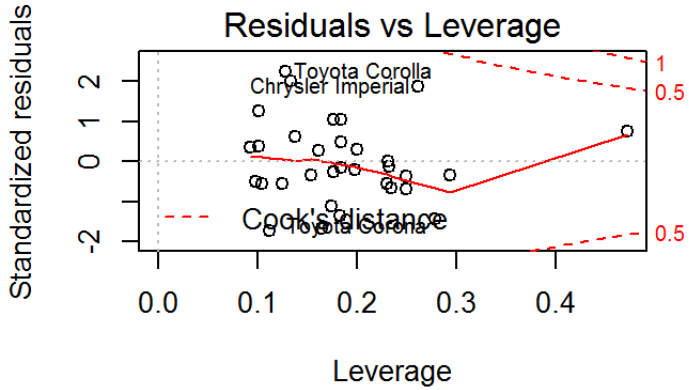
```
## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x          1 405.15  405.15   16.86 0.000285 ***
## Residuals 30 720.90   24.03
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
par(mfrow = c(2,2))
plot(fit2)
```





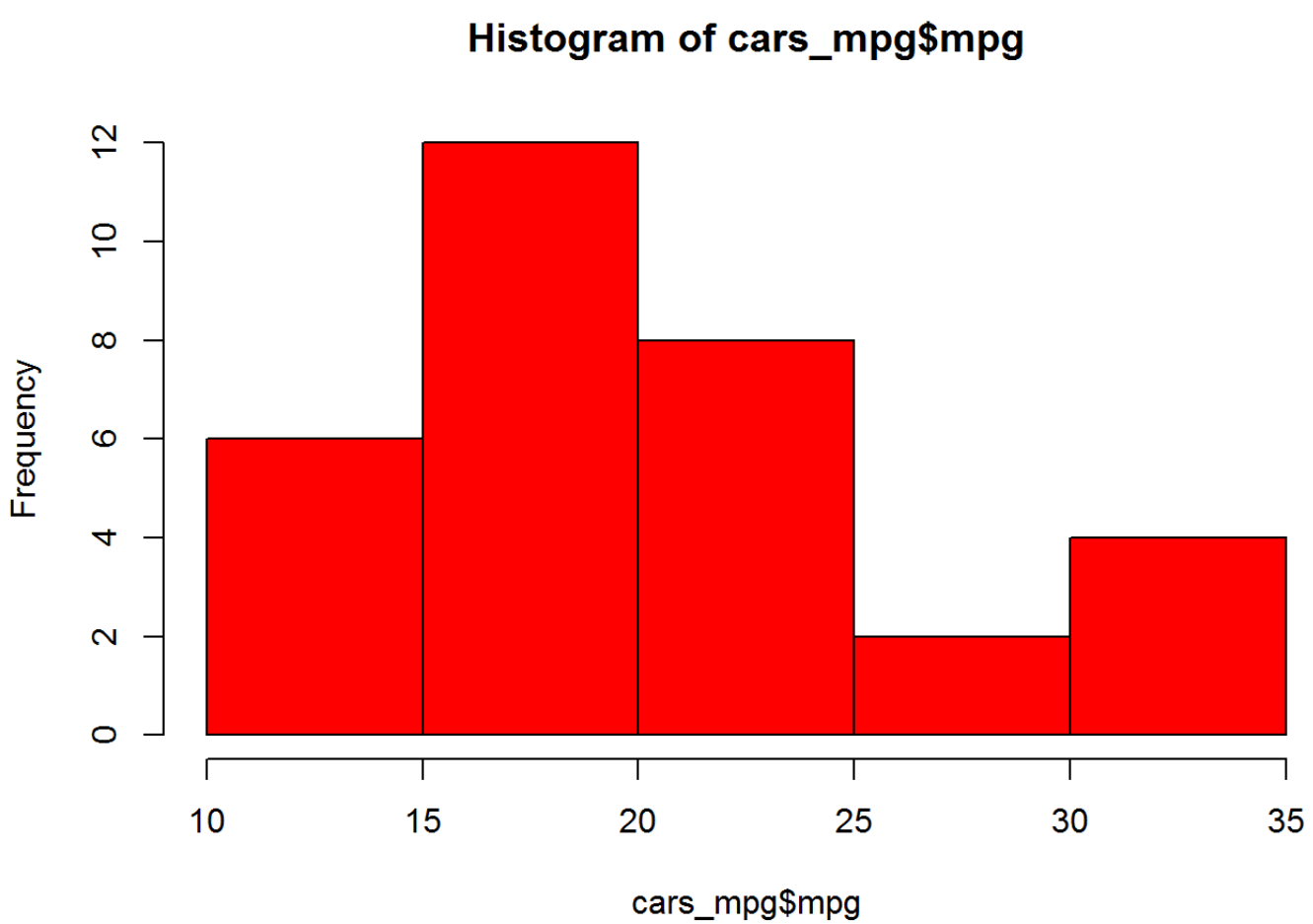


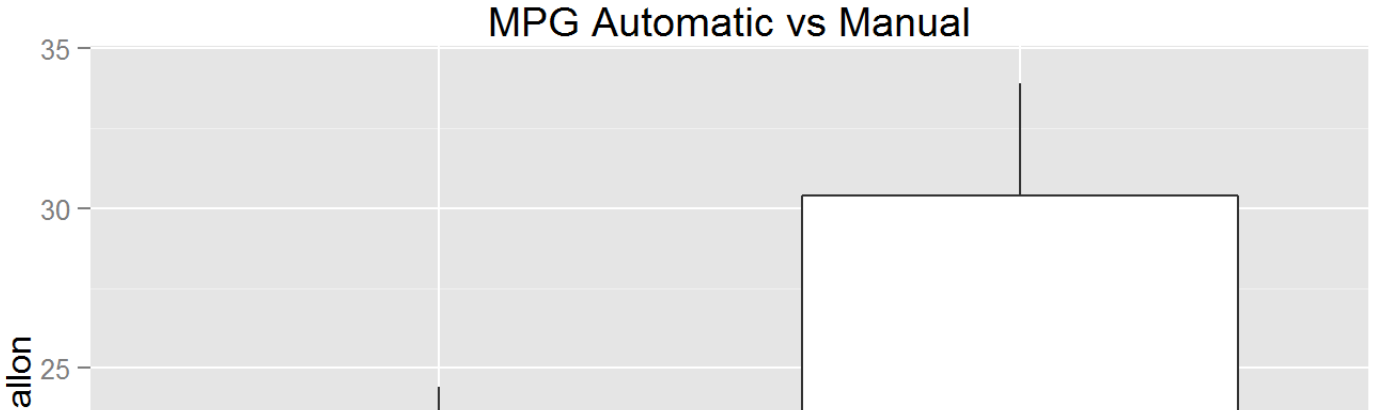


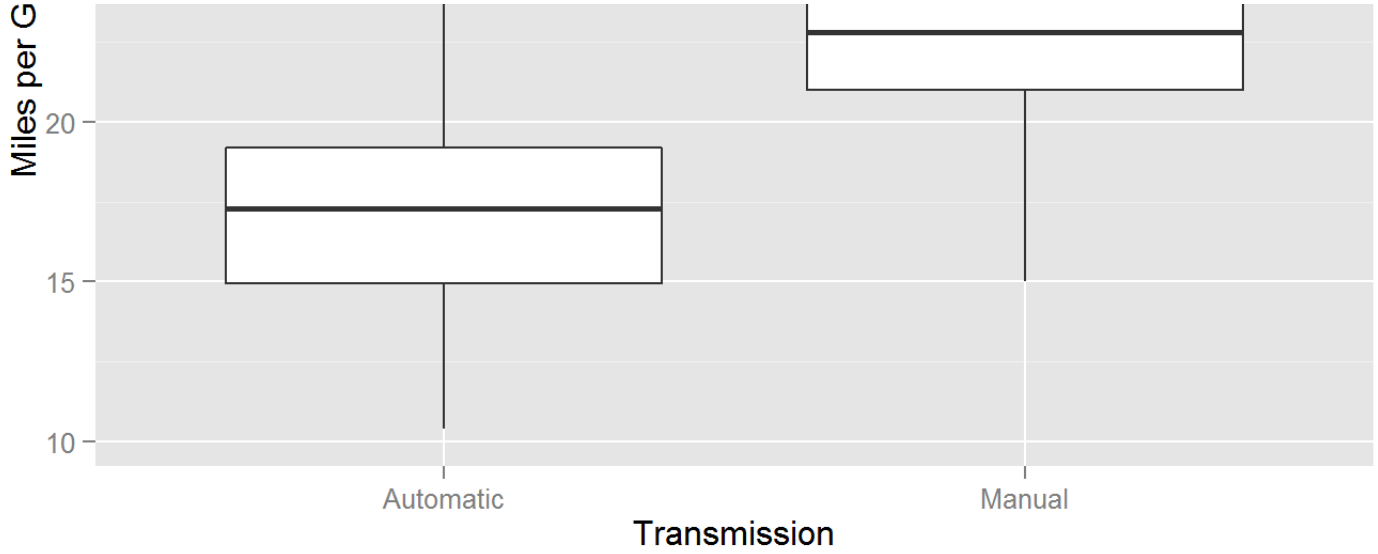
We performed an ANOVA on the two fits we've used. Anova showed a p-value of 0.000285 indicating that the simple linear regression is significantly different form the multiple regression model. The plot of the fit shows that the data is normally distributed and show no heteroskedasticity. According to the multiple regression model,cars with manual transmission have 1.68 MPGs more than automatic transmission. This model also explains 83% of the variation.

Apendix - Plots

The box plot below shows a higher over all MPG average for manual transmission compared to automatic transmission.

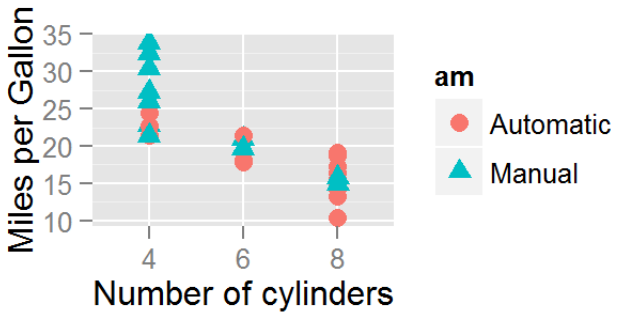

Histogram of cars_mpg$mpg
Frequency
0
2
4
6
8
10
12
10
15
20
25
30
35
cars_mpg$mpg

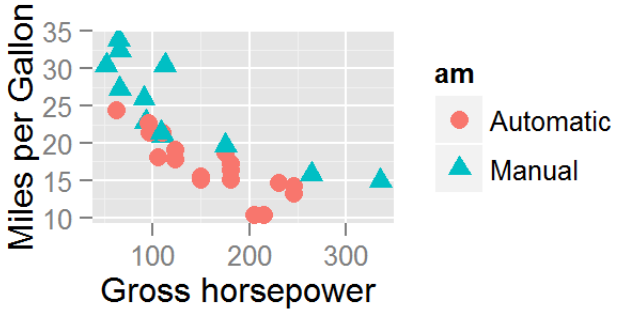

MPG Automatic vs Manual
35
30
25
allon

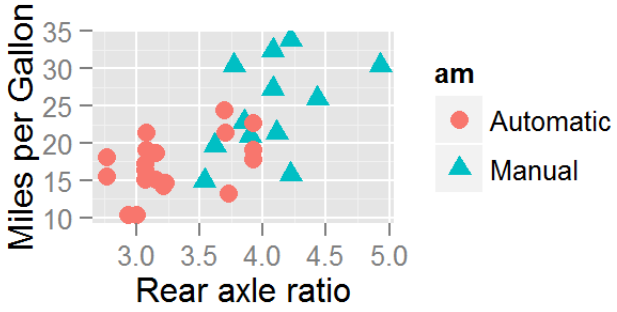


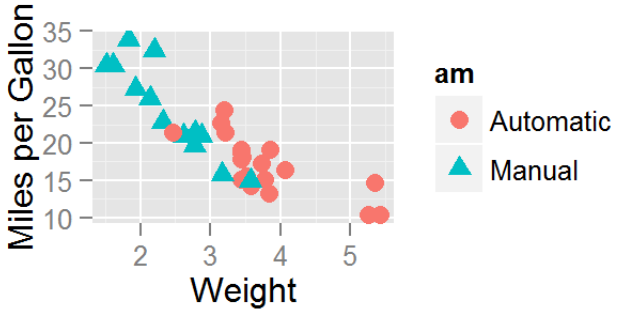
plot 1 clearly shows that with lower number of cylinders, cars with manual transmission have higher MPG. For a higher number of cylinders the transmission doesn't play much of a roll.
A similar relationship is seen for displacement, horsepower and weight. With lower horsepower, displacement and weight manual transmission has better MPG performance.
The opposit relationship is seen for rear axle ratio. When the rear axel ratio is higher, manual transmission hows better MPG performance.

```
p1 = qplot(cyl, mpg, data=cars_mpg, shape = am, color=am,
      size=I(3),xlab="Number of cylinders", ylab="Miles per Gallon")

p2 = qplot(disp, mpg, data=cars_mpg, shape = am, color=am,
      size=I(3),xlab="Displacement (cu.in.)", ylab="Miles per Gallon")

p3 = qplot(hp, mpg, data=cars_mpg, shape = am, color=am,
      size=I(3),xlab="Gross horsepower ", ylab="Miles per Gallon")

p4 = qplot(drat, mpg, data=cars_mpg, shape = am, color=am,
      size=I(3),xlab="Rear axle ratio", ylab="Miles per Gallon")

p5 = qplot(wt, mpg, data=cars_mpg, shape=am, color=am,
      size=I(3),xlab="Weight", ylab="Miles per Gallon")

grid.arrange(p1, p2, p3, p4, p5)
```










```
dev.off()
```

```
## null device
##           1
```