Effects of Transmission Type on Fuel Consumption

Woldetsadick Selam Getachew

Friday, October 24, 2014

Executive Summary

The following analysis was made using the mtcars data for R datasets. It was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. After some initial change in data set, and summary of variables an exploratory data analysis howed that there was in fact a statistically significant difference between the mileage of automatic transmission cars and manual transmission cars. Some variables were selected using stepwise regression and AIC and fitted data using the best model generated by inclusion or exclusion of said variables and their interactions. Results showed that manual transmissions are better for mileage (or MPG). Actually the gain makes for about 9.14 mpg ceteris paribus when shfiting from automatic to Manual transmission. The following analysis was made using the R statistic software version R version 3.1.1 (2014-07-10) on a computer using Windows OS version x86_64-w64-mingw32/x64 (64-bit) and completed on Fri, Oct 24 2014, 8:47:44 PM.

1. Loading, Processing and Summarizing Data

The dataset used in this analysis was extracted from the 1974 *Motor Trend* US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles. The names of the 11 variables present in the dataset, their descriptions and corresponding units are given in Table I in Annex.

Looking at variable description one concludes that the variables `cyl, vs, am, gera` and `carb` should be categorical (or factor class in R) while the rest are continuous quantitative (or class numeric in R). Let's transform each variables to corresponding class. Additionally let values of variables `vs` and `am` reflect real values making variables more descriptive.

```
data(mtcars); data <- mtcars # Loading data
data$vs <- ifelse(data$vs == 0, "veng", "seng"); data$am <- ifelse(data$am == 0, "auto", "manu") #
 Descriptive values
for(i in 1:11){if(i %in% c(2,8,9,10,11)) {data[,i] <- as.factor(data[,i])} else{data[,i] <- as.num
eric(data[,i])}} # Class Change
nas <- sum(is.na(data)) # Total NAs in Data
len <- nrow(data)# Dataset dimensions
library(psych) ; a <- describe(data$mpg)
```

The dataset has exactly `0` missing values throughout, and includes `32` observations. Let's now conduct a quick summary of the variables. First, the re-partition of observation along values of variable `cyl`, or values *4*, *6*, *8* are 11, 7 and 14 respectively. Similarly along values of variable `am`, or values *auto* and *manu*, are 19 and 13 respectively, along values of variable `gear`, or values *3*, *4* and *5*, distribution is 15, 12 and 5, and finally along variable `carb` of values *1*, *2*, *3*, *4*, *6* and *8* re-partition is 7, 10, 3, 10, 1 and 1. The re-partition of data points along values of variable `vs`, which are *seng* and *veng*, are 14 and 18 respectively.(Data Description Paragraph in Annex)

The research on 1/4 Mile Time shows that values for this variable can be deduced from weight and horsepower values. This link between these three variables makes information from one of them redundant. Here one can choose to exclude variable `qsec` from analysis. The minimums and maximums for variables `disp, hp, drat` and `wt` are respectively, in order, 71.1 - 472.0, 52.0 - 335.0, 2.76 - 49.93, 1.51 - 5.42 for range `400.9`, `283`, `47.17` and `3.91` for each variable respectively. The mean - median pairs for the same variables in the same order are 230.7 - 196.3, 146.7 - 123.0, 3.60 - 3.69, 3.22 - 3.33. The proximity of mean and median for all variables considered tells us that mass of data on upper and lower side of mean value is nearly equal. The mpg variable has a mean of `20.0906` and standard deviation `6.0269` for a median of `19.2`. Again the near equality of mean and median suggest the equal mass of data upper and lower mean value. Maximum and minimum values are `33.9` and `10.4` for range `23.5`. A skewness of `0.6107` and a kurtosis of `-0.3728` describes in general a flat and positively skewed distribution.

2. Exploratory Data Analysis

First a pairwise scatterplot is made in order to visualize the correlation between pairs of variables present in the data, except the `qsec` variable which is excluded. One can see in Figure I in the Annex not only that mpg variable is correlated strongly positively with `cyl`, `drat` and `am` variables, strongly negatively with `hp`, `disp`, `wt`, `disp`, `vs`, and `carb`, and weak positively with `drat` and `carb` variable, but also that correlation between the 10 other variables is strong. Note a strong positive correlation between `wt` and `cyl`, `wt` and `hp`, `hp` and `cyl`, a strong negative correlation between `wt` and `am`, and absence of correlation between `hp` and `am` and `cyl` and `am`.
The two variables of interest here are `mpg` and `am` because the study is about the impact of transmission on consumption. There seem to be a positive correlation between the two variables. However observing mean and variance within groups formed along the line of `am` variable could be of interest here. The boxplot in Figure II in Annex suggests that not only that cars with manual transmission have in average a higher consumption of fuel compared to automatic transmission cars, but also that within group variability in fuel consumption for the latter group is much narrower than the earlier.
However one should think of correcting for other variables before setting this conclusion in stone.

3. Building our Model

One way of looking at the relation of interest correcting for additional counfounders is to make a multivariate regression model to fit the data. In this case, one only considers linear models. However, linear regressions depend on model selection heavily, in our case in inclusion and exclusion of variables. To lighten the decision, a step regression in both ways is conduct with the Akaike Information Criterion from the different models as comparator statistic informing the best model. The use of this algorithm informs then the choice of inclusion and exclusion of variables.

```
full.model <- lm( mpg ~ ., data = data[, -7])
best.model <- step(full.model, direction = "both")
```

```
summary(best.model)$call
```

```
## lm(formula = mpg ~ cyl + hp + wt + am, data = data[, -7])
```

The best model fit in the dataset is one including variable `cyl`, `hp`, `wt` and `am` according to AIC. That will be then the model of choice hereinafter.Now let's consider interactions between these different selected variables. Let's create a pairwise interaction variable for each four variables, then use the stepwise regression method to judge inclusion of interaction variables.

```
full.model1 <- lm(mpg ~ cyl+hp+wt+am+(cyl:am)+(cyl:hp)+(cyl:wt)+hp*wt+(hp:am)+(wt:am), data = data
)
best.model1 <- step(full.model1, direction = "both")
```

```
summary(best.model1)$call
```

```
## lm(formula = data$mpg ~ data$cyl + data$hp + data$wt + data$am +
##     data$cyl:data$hp + data$wt:data$am)
```

Considering all pairwise interactions between variables in the dataset as regressors, the best model fit is given above. This will be then the end model used to make statistical inference of the model.

---

4. Statistical Inference and Coefficient Interpretation

First one can conduct a two sided t-test to look at the significance at 95% confidence level of difference in means of fuel consumption between a car with manual and automatic transmission, assuming unequal variances like observed in boxplot.

```
t <- t.test(mpg ~ am, data = mtcars)
```

A p-value equal to `0.0014` > .05 of the t-test conducted permits to reject the null hypothesis of equal means. Also observe that the confidence interval [`-11.2802, -3.2097`] for this test does not include 0.
Then data is fitted to model choose in the previous section, the results are displayed in *Chosen Model Fit* section of the Annex. Notice here that the intercept cannot have an interpretation as baseline values for weight and horse power are zero. A car with 0 weight and horse power is a *dead car*.
First observe that for a confidence level at 95%, only estimates of the betas for variables hp wt am and of shift in mileage when passing from 4 to 8 cylinders and interaction variable of 8 cylinders level with hp are significant, they will be the only one interpreted.
The mean loss in mileage for 1000 pounds gain in vehicle weight is estimated at `-2.31` mpg and the mean loss in mileage for a horse power gain of one is estimated at `-0.08`. The two observations are of course made ceteris paribus. Ceteris paribus, the shift from automatic to manual transmission in a car will correspond to a gain of `9.14` mpg in mean mileage. In the same manner a shift from a 4 cylinders car to 8 cylinders cars will correspond to a gain of `-10.82` mpg in mean mileage. Also the mean change in mileage due to horse power change when shifting from 4 cylinders cars to 8 cylinders car is estimated at `-3.05` mpg.
The R^2 gives a value of `0.9036` which means that our data explains `90.3616`% of the variability in the data.

---

5. Residuals and Diagnostics

In this section, residual plots are made for the choose regression model: This will help of course point out to any problems in the residual of the fitted model and study the quality of fitness.
The Residuals vs Fitted and Scale-Location graphs in residuals analysis plots in Figure III in the Annex shows no particular pattern. However the difference between the two red lines show the specially influential outliers namely, The Pontiac Firebird, The FIAT 128 and The Toyota Corolla. While the QQplot shows pretty normal residuals, the three points mentioned before are outstanding outliers. Three leverage points are pointed out to here, namely The Lotus Europa, The Maserati Bora and The Honda Civic. Let's then compute some regression diagnostics of the model to find out interesting leverage and influential points, and confirm visual observations.

```
influential <- dfbetas(best.model1); leverage <- hatvalues(best.model1)
no <- head(sort(influential[,6], decreasing = TRUE),3); yes <- head(sort(leverage, decreasing = TR
UE),3)
```

The results are significant departure of visual analysis of influential and leverage points in the dataset. In top 3 leverage data points `Ferrari Dino, Lotus Europa, Maserati Bora` all of them represent race cars with low mpg. In top 3 influential data points one finds `Lotus Europa, Chrysler Imperial, Toyota Corolla`An interesting thing to do would be to exclude these points to study the relationship of interest anew.

6. Conclusion and Answers

In summary, after some initial change in data set, and summary of variables present in the data set, we conducted a exploratory data analysis that showed that there was in fact a statistically significant difference between the mileage of automatic transmission cars and manual transmission cars. We then selected some variables using stepwise regression and AIC and fitted data to the best model. This model makes use of variables `cyl` , `hp` , `wt` and `am` and some of their interactions. Then, correcting for the selected variables and some of their interactions, one finds that :

- Manual transmissions are better for mileage (or MPG).
- Actually the gain makes for about 9.14 mpg ceteris paribus when shfiting from automatic to Manual transmission.

7. Annex

Table I : Variables Description

Variable Name	Description	Unit
mpg	Fuel Consumption	US Miles per Gallon
cyl	Number of Cylinders	No Unit
disp	Engine Displacement	Cubic Inches
hp	Motor’s Gross Horsepower	hp
drat	Rear Axle Ratio	No Units
wt	Vehicule weight	Thousand Tons
qsec	1/4 Mile Time	Seconds
vs	Engine Type	0 for V-Engine - 1 for Straight Engine
am	Transmission Type Dummy	0 for Automatic - 1 for Manual
gear	Number of Forward Gears	No Unit
carb	Number of Carburetors	No Unit

Data Summary

```
c(summary(data$cyl), summary(data$vs), summary(data$am), summary(data$gear), summary(data$carb), s
ummary(data$vs))
```

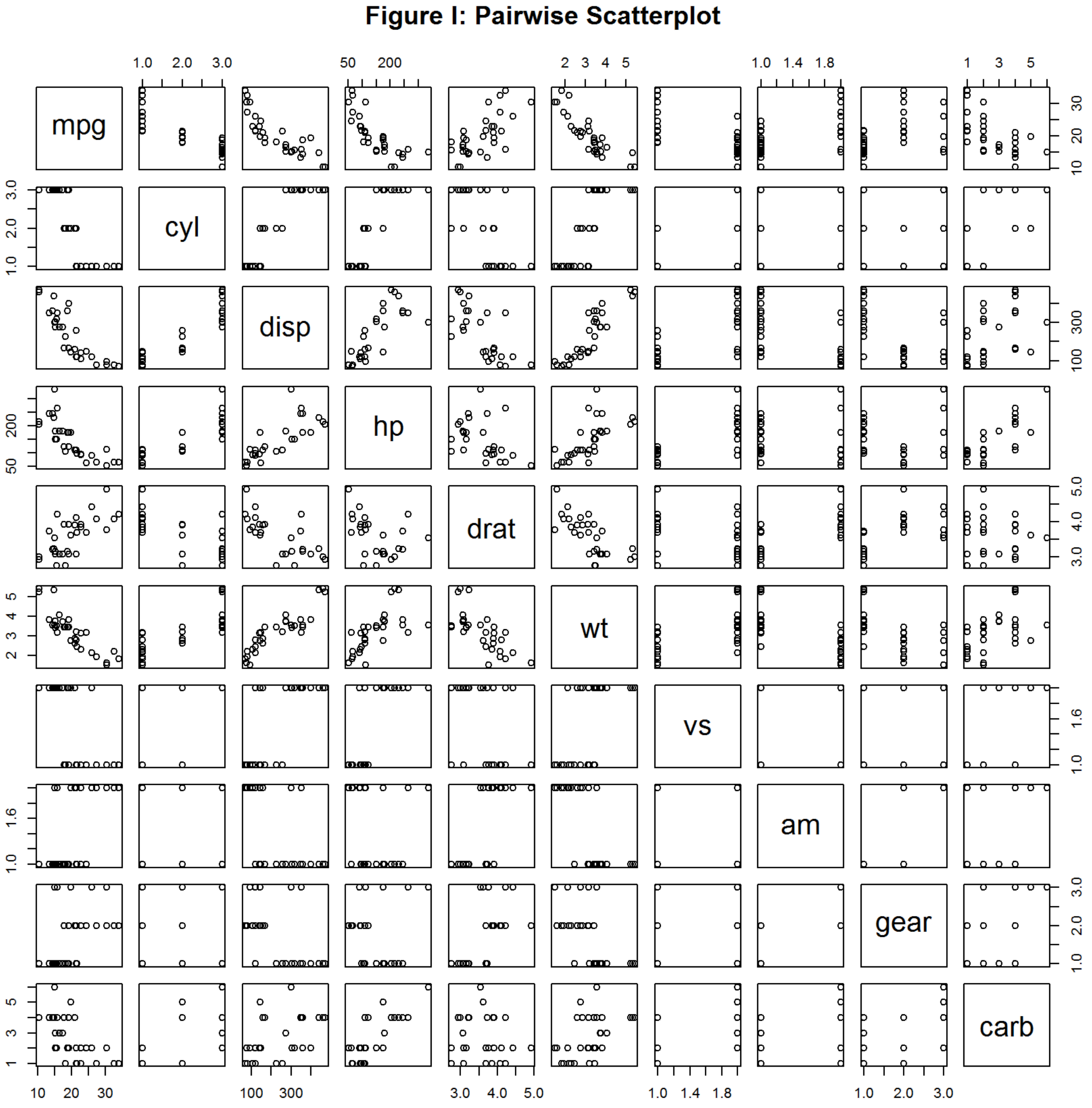
```
##    4    6    8 seng veng auto manu    3    4    5    1    2    3    4    6
##   11    7   14   14   18   19   13   15   12    5    7   10    3   10    1
##    8 seng veng
##    1   14   18
```

```
summary(data[,-c(1, 2, 7, 8, 9, 10, 11)])
```

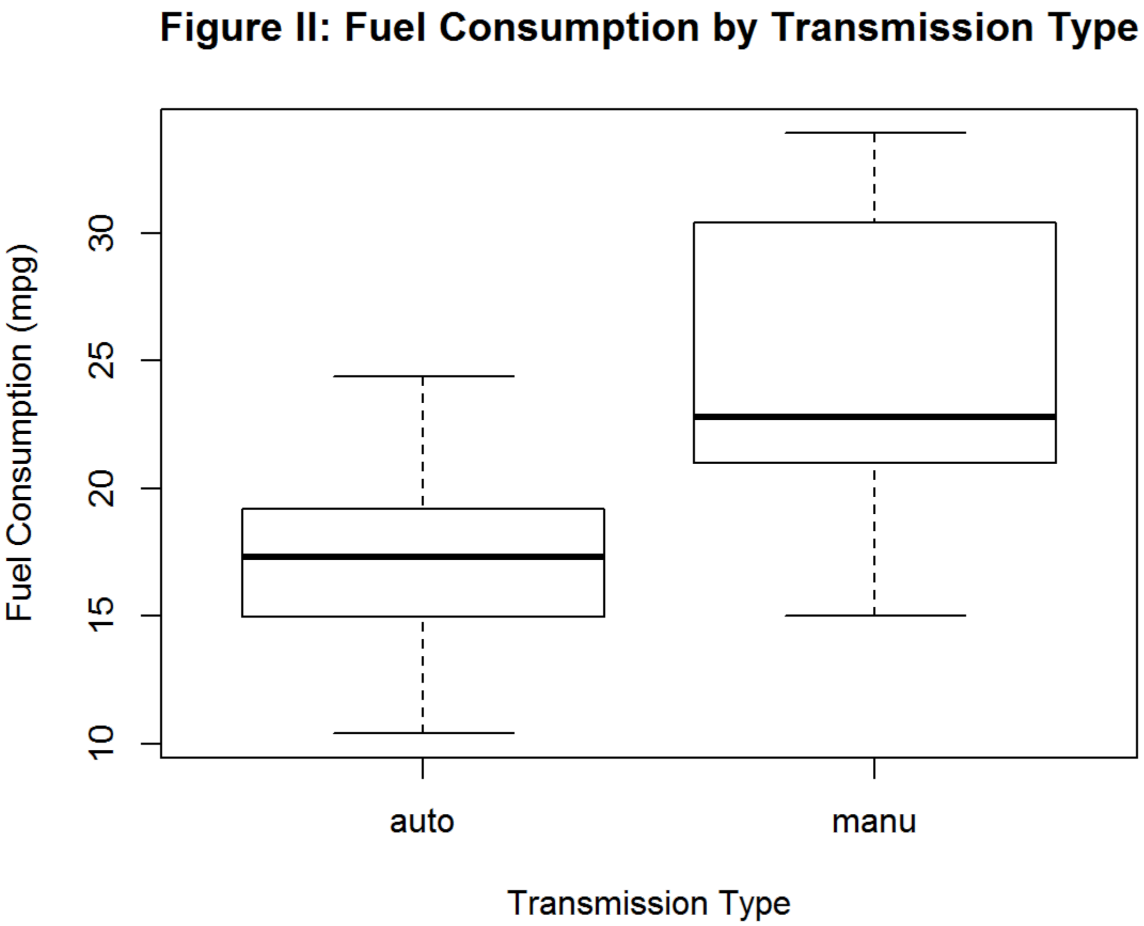
```
##       disp             hp             drat            wt
##  Min.   : 71.1   Min.   : 52.0   Min.   :2.76   Min.   :1.51
##  1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.08   1st Qu.:2.58
##  Median :196.3   Median :123.0   Median :3.69   Median :3.33
##  Mean   :230.7   Mean   :146.7   Mean   :3.60   Mean   :3.22
##  3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.92   3rd Qu.:3.61
##  Max.   :472.0   Max.   :335.0   Max.   :4.93   Max.   :5.42
```

Correlations

```
par(mfrow=c(1,1)); title <- "Figure II: Fuel Consumption by Transmission Type"
pairs(data[,-7], main="Figure I: Pairwise Scatterplot")
```



```
boxplot(mpg ~ am, data = data,main = title, xlab = "Transmission Type",ylab = "Fuel Consumption (m
pg)")
```



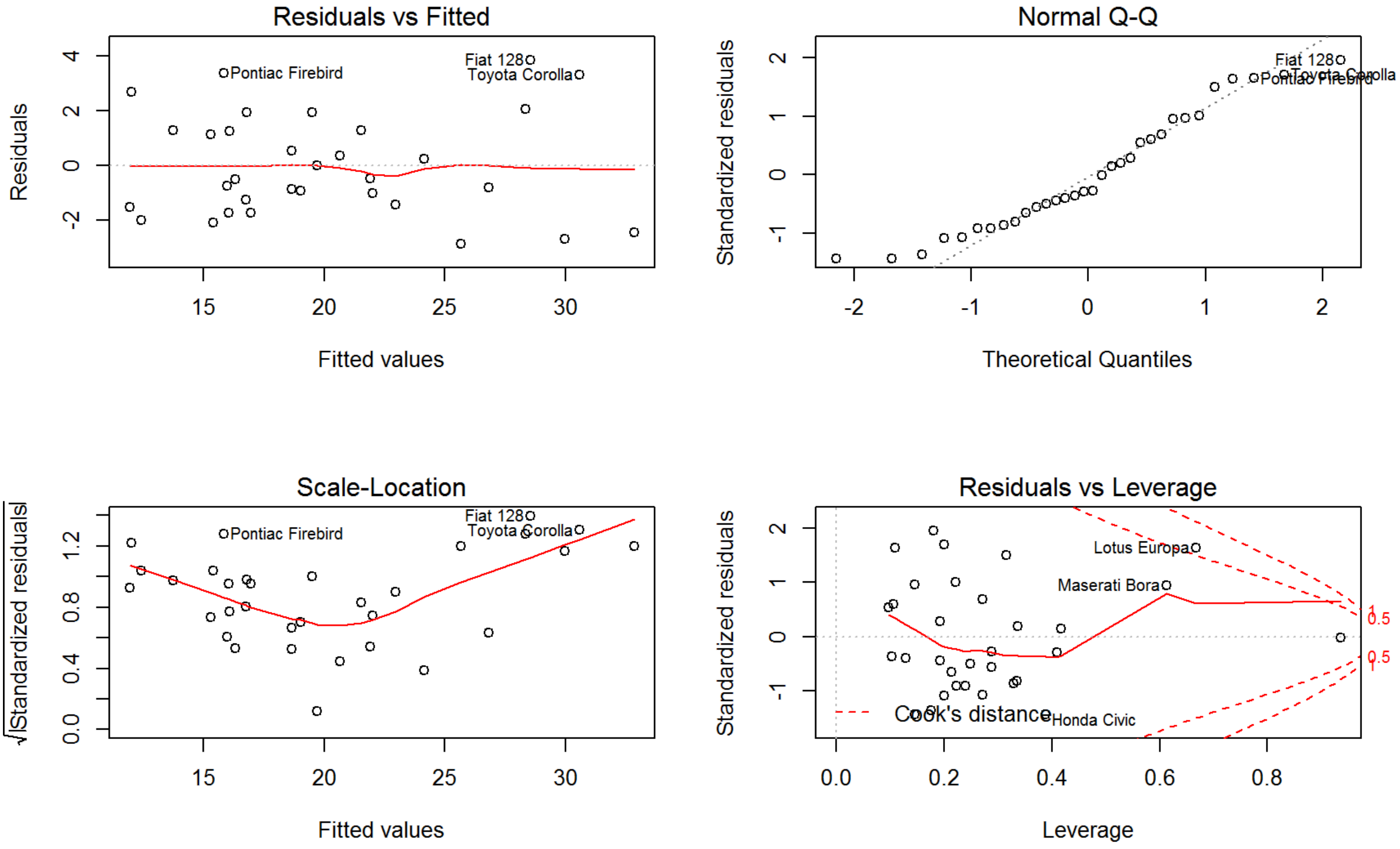
Chosen Model Fit

```
chosen <- lm(mpg ~ cyl + hp + wt + am + (cyl:hp) + (wt:am), data = data)
summary(chosen)$coefficients
```

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  36.65881    3.98711   9.194 3.639e-09
## cyl6         -7.19711    5.60784  -1.283 2.121e-01
## cyl8        -10.82118    4.22762  -2.560 1.752e-02
## hp           -0.08268    0.03401  -2.431 2.326e-02
## wt           -2.31293    0.81181  -2.849 9.082e-03
## ammanu        9.14282    4.12170   2.218 3.669e-02
## cyl6:hp       0.05954    0.05035   1.182 2.491e-01
## cyl8:hp       0.07634    0.03565   2.142 4.305e-02
## wt:ammanu    -3.04685    1.51646  -2.009 5.639e-02
```

Figure III: Residuals Plots

```
par(mfrow=c(2, 2))
plot(best.model1)
```


Residuals vs Fitted
Residuals
Fitted values
4
2
0
-2
15
20
25
30
Pontiac Firebird
Fiat 128
Toyota Corolla
Normal Q-Q
Standardized residuals
Theoretical Quantiles
2
1
0
-1
-2
-1
0
1
2
Fiat 128
Toyota Corolla
Pontiac Firebird
Scale-Location
√|Standardized residuals|
Fitted values
0.0
0.4
0.8
1.2
15
20
25
30
Pontiac Firebird
Fiat 128
Toyota Corolla
Residuals vs Leverage
Standardized residuals
Leverage
2
1
0
-1
0.0
0.2
0.4
0.6
0.8
Lotus Europa
Maserati Bora
Honda Civic
Cook's distance
1
0.5
0.5
1