Project: Regression Model - Coursera project

N. Lakhani

21 January 2018

Report on: Miles Per Gallon - Automatic vs Manual Transmission

Executive Summary

This project uses linear regression to explore relationships between Miles per gallon (mpg) and other variables in the mtcars dataset. The mtcars dataset has data on fuel consumption and 11 aspects of automobile design and performance for 32 automobiles (1973-74 models). The goal of the analysis is to answer:

- Is an automatic or manual transmission better for MPG?
- Quantify the MPG difference between the 2 different transmissions

In additon, this document provides analysis on several other questions of interest. Based on the analysis, I conclude that:

- mpg for manual transmission is better than automatic by 7.25. This was validated by various statistical tests at a 95% confidence level
- the best model for predicting mpg is: mpg = 37.22 - 3.87 * wt - 0.03 * hp. This model has adj R-sq of 0.81 with a very small p value. The model accounts for 81% of the mpg variance, the rest being explained by residuals
- The residual distribution is random, which indicates the soundness of the model, even given that the mpg data is not completely normal

1. Data

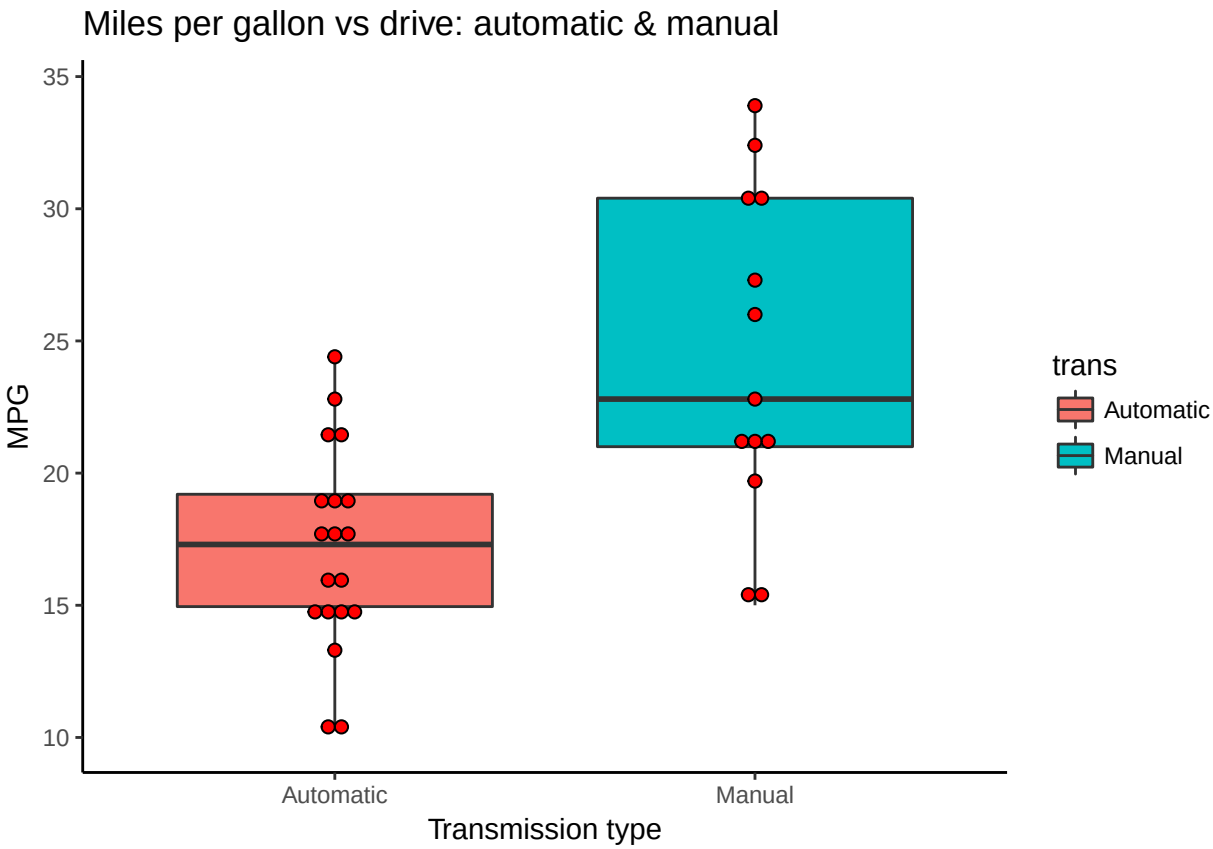
Details of the dataset **mtcars** is provided in the Appendix - section 1

2. Exploratory Data Analysis

A summary of the basic analysis is provided below:

- Average MPG for manual drive is 24.39 with sd of 6.16; for automatic drive is 17.14 with sd of 3.83. So the variance (& sd) in mpg for each transmission is large, perhaps due to size of the sample size, outliers or other factors such as # of cylinders. etc. More analysis follows in Appendix section 2 on this.
- A histogram and qq-norm plot in appendix section 3, shows the data distribution is not normal.

We further explore the data visually with plots:



The exploratory plots (box and density) indicate that:

- mpg for manual transmission has a larger variance as compared to automatic. There are some data points outside the quantiles for both transmissions as indicated by the box plot (outliers). This could indicate the influence of other variables such cyl on mpg
- median mpg of manual is higher vs automatic. This is analyzed further.

We further examine if the mean mpg for the 2 transmission types are statistically different using a t-test at a 95% confidence level.

A null hypothesis (H0) stating that **nothing is going on and the mean mpg's are not statistically different is tested using a t-test.**

The t-test results show that $p < 0.05$, hence the null hypothesis is rejected implying that *there is something going on and the mean mpgs for the 2 transmissions are different*, as also indicated by the difference in the 2 mean's (24.39-17.14 = 7.25), which lies between the CI [3.2,11.28]. Details of the t-test are in the appendix, section 4

3. Exploring relationships to identify key influencing variables:

We do this initially using the R corr function to eliminate some variables and make the model simpler. The pairs plot also indicate the relationships visually:

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am
##  0.4186840  0.6640389  0.5998324
```

The correlation coefficients and pairs plot indicate that:

- variable qsec has the lowest absolute coefficient (0.41), so we will eliminate this from further analysis, keeping the rest (absolute value > 0.5) for further detailed investigation.
- Negative coefficients of cyl,disp,hp,wt indicate reduction in mpg as these variable take on higher values

4. Model evaluation and selection

The approach taken to arrive at a simpler relationship using model selection is:

- Start with the full model (mpg ~., excluding qsec) using function **lm**. Examine adj R-squared and p values to determine significance of each variable
- Use the function stepAIC to rank the variables based on their influence and eliminate variables with low p values (<0.05) resulting in a reduced model.
- Run the regression function lm on the reduced model and evaluate it for suitability based on adj R-squared and p values.
- Further evaluate the reduced model vs other models for best fit evaluation using Anova. Also look at residuals and other fit parameters for validation of best fit model

The step wise model selection and regression summaries provides the following insights:

- for the simplest model mpg ~ am, adjusted R-squared is 0.36 indicating that **am** explains only 36% of the variance in mpg, this model ignores many other inluencing variables
- for the full model (mpg ~.), R-squared is 0.81 indicating that this model explains 81% of the variation in mpg. The model though is complex with all variables included and yet none of the p values are < 0.05 for any predictor. Hence simplification is needed.
- The stepAIC function helps arrive at at simpler model mpg ~ wt + hp.

```
##
## Call:
## lm(formula = mpg ~ wt + hp + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## am1          2.083710   1.376420   1.514 0.141268
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + hp + am
## Model 3: mpg ~ cyl + disp + hp + drat + wt + vs + am + gear + trans
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 42.5248 2.762e-08 ***
## 3     22 139.84  6     40.45  1.0607    0.4151
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We next examine the final model (mpg ~ wt + hp + am) for residuals and influence of each selected variable using function stepAIC.

- the final model is: mpg = 37 - 3.87 * wt - 0.03 * hp. This model has adj R-sq of 0.81 with small p values. The model accounts for 81% of the mpg variance
- The residual plots shown in the appendix 5 brings out the random distribution of the residuals for this model. Also shown in the Residuals vs Leverage plot are outliers like Maserati, Chrysler perhaps due to the number of cylinders and hence larger wt

5. Conclusion

- The analysis helps determine that transmission types influences mpg significantly with manual mpg being higher than automatic.

- A model selection was explored, which helped arrive at final model of mpg ~ wt + hp. This model has adj R-squared of 0.82. While other variables also influence mpg to various degrees, they are also very likely a factor in influencing wt and hp.

Appendix

1. Description of mtcars dataset

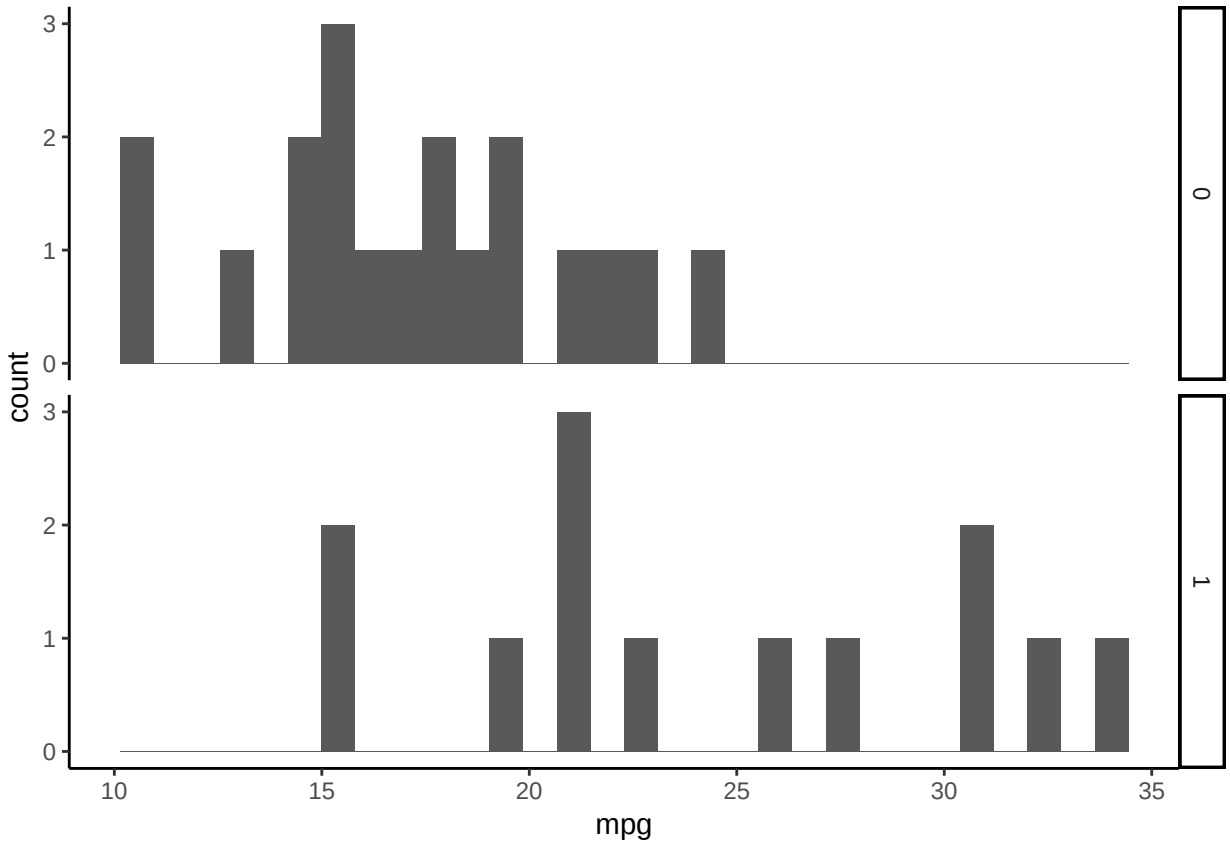
mtcars has 32 observations on 11 variables:
1 mpg: Miles/(US) gallon
2 cyl: Number of cylinders
3 disp: Displacement (cu.in.)
4 hp: Gross horsepower
5 drat: Rear axle ratio
6 wt: Weight (1000 lbs)
7 qsec: 1/4 mile time
8 vs: V/S
9 am: Transmission (0 = automatic, 1 = manual)
10 gear: Number of forward gears
11 carb: Number of carburetors

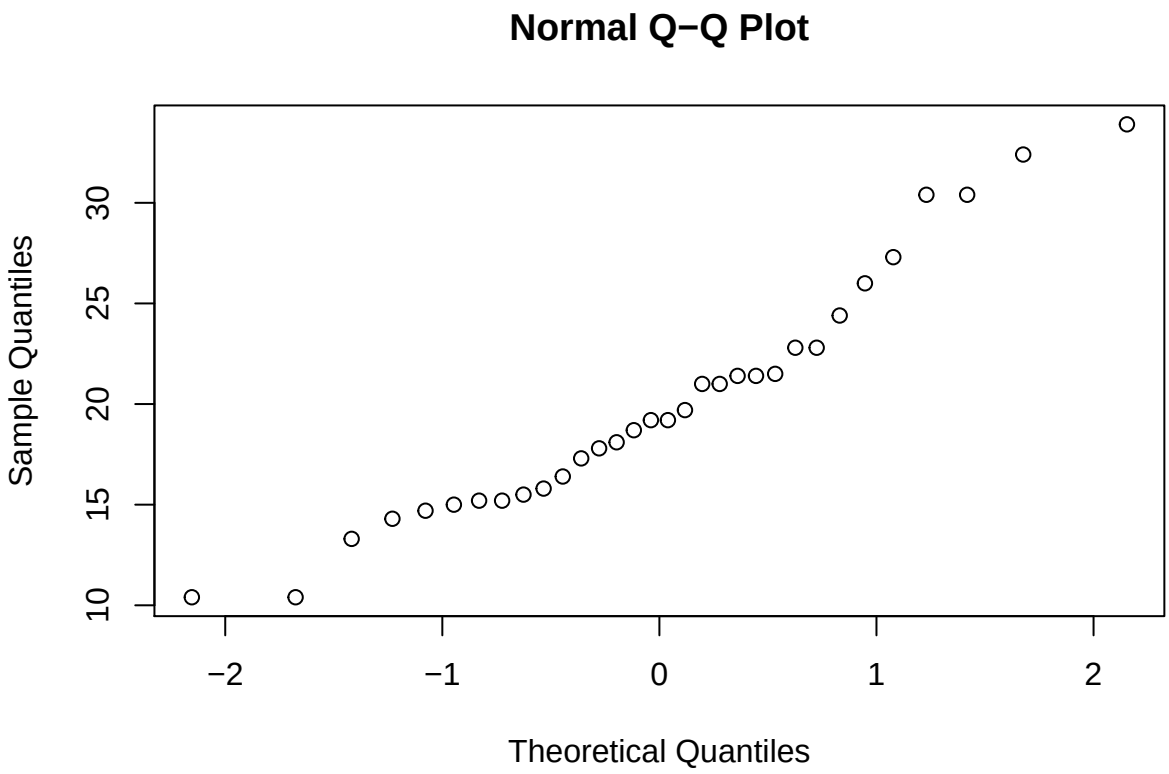
2. Summary of mtcars dataset

```
##       mpg        cyl         disp             hp             drat
##  Min.   :10.40   4:11   Min.   : 71.1   Min.   : 52.0   Min.   :2.760
##  1st Qu.:15.43   6: 7   1st Qu.:120.8   1st Qu.: 96.5   1st Qu.:3.080
##  Median :19.20   8:14   Median :196.3   Median :123.0   Median :3.695
##  Mean   :20.09          Mean   :230.7   Mean   :146.7   Mean   :3.597
##  3rd Qu.:22.80          3rd Qu.:326.0   3rd Qu.:180.0   3rd Qu.:3.920
##  Max.   :33.90          Max.   :472.0   Max.   :335.0   Max.   :4.930
##        wt        vs     am          gear            trans
##  Min.   :1.513   0:18   0:19   Min.   :3.000   Automatic:19
##  1st Qu.:2.581   1:14   1:13   1st Qu.:3.000   Manual   :13
##  Median :3.325                 Median :4.000
##  Mean   :3.217                 Mean   :3.688
##  3rd Qu.:3.610                 3rd Qu.:4.000
##  Max.   :5.424                 Max.   :5.000
```

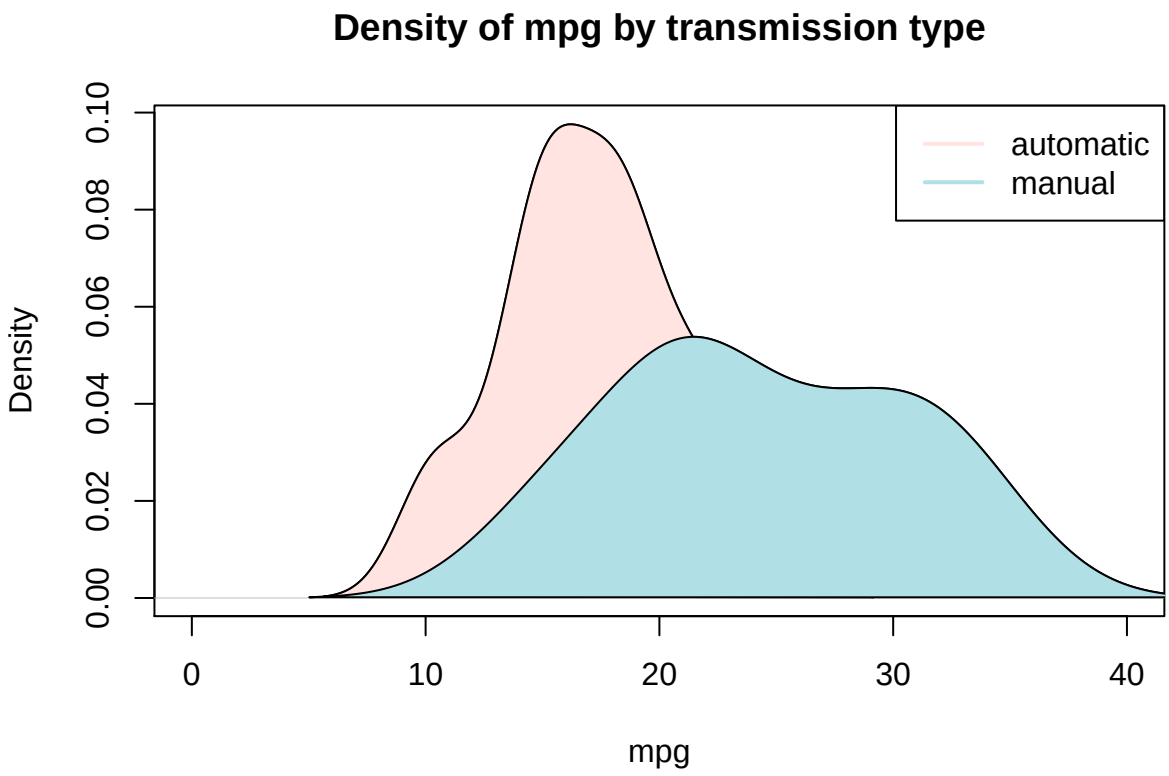
```
## 'data.frame':    32 obs. of  10 variables:
##  $ mpg  : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl  : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ disp : num  160 160 108 258 360 ...
##  $ hp   : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat : 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 ...
##  $ vs   : Factor w/ 2 levels "0","1": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am   : Factor w/ 2 levels "0","1": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear : num  4 4 4 3 3 3 3 4 4 4 ...
##  $ trans: Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
```

3. Histogram of mpg data by transmission type

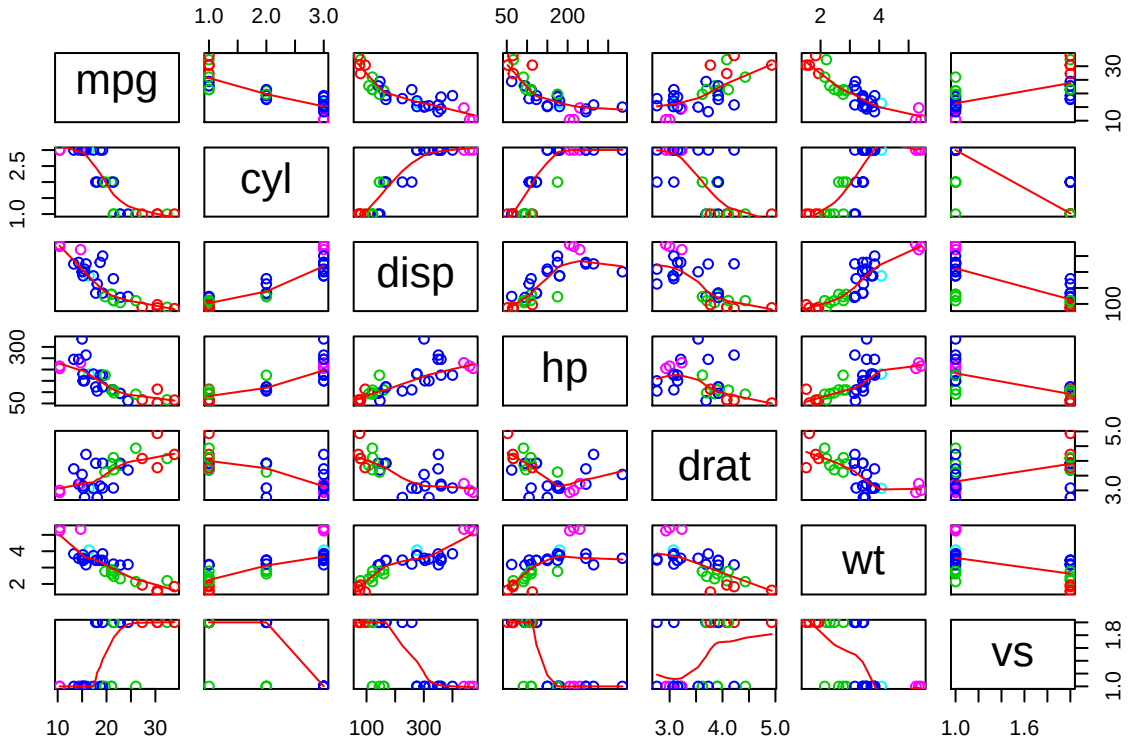




4. Density plots for mpg



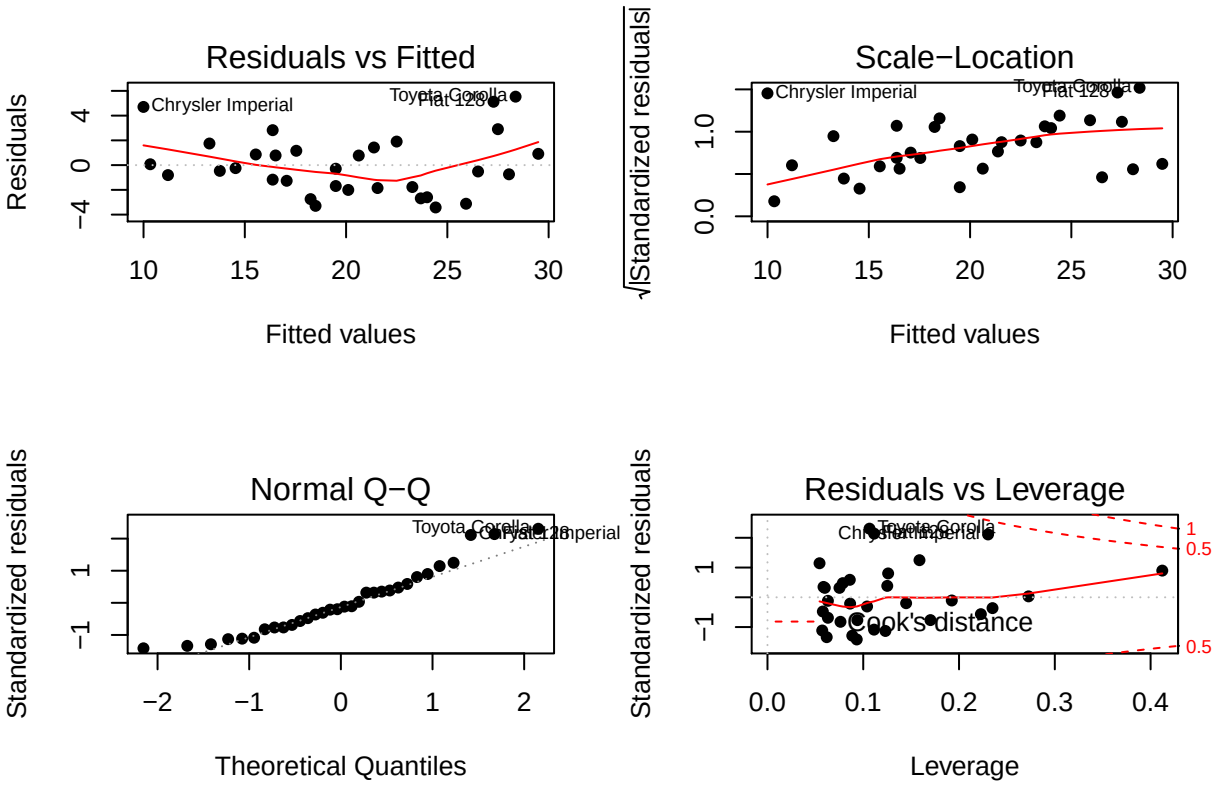
5. Plot - relationship between variables



6. T-test results of mpg vs transmission type

```
##
##  Welch Two Sample t-test
##
## data:  auto$mpg and man$mpg
## t = 3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   3.209684 11.280194
## sample estimates:
## mean of x mean of y
##  24.39231  17.14737
```

7. Residual Analysis - plots




```
##                   2.5 %      97.5 %
## (Intercept) 28.58963286 39.41611738
## wt          -4.73232353 -1.02482730
## hp          -0.05715454 -0.01780291
## am1         -0.73575874  4.90317900
```