# Regression Models Assignment

*Thomas Sacchetti*

*August 15, 2015*

## Exevutive Summary

This project works with the mtcars data set. In this project we will examine the impact of transmission type on the miles per gallon. The dataset comes from a report from motortrend that contains 32 different models. The variable for transmission is reported as am and takes the value of 0 or 1, where 0 corresponds to automatic and 1 corresponds to manual. The t test confirms there is a signifigant relationship between between the two variables. From the linear model we can see that there is a gain of about 7 MPG to use a manual transmission on the 32 models tested. We can see a correlation between the weight of the car and the transmission type, which could potentially decrease the correlation.

## Load and Preprocess the Data

```
library(knitr)
library(rmarkdown)
data(mtcars)
dim(mtcars)
```

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

```
table(summary(is.na(mtcars)))
lapply(mtcars,class)
```

Now we know that there are no NA variables and that the data set has dimensions 32x11. Similarly we know that the column class is numeric for all values. We will now proceed to examine correlations between mpg and transmission.

## Statistical Insight

```
test = t.test(mpg~am,data= mtcars)
test$p.value
```

```
## [1] 0.001373638
```

```
test$estimate
```

```
## mean in group 0 mean in group 1
##        17.14737        24.39231
```

From the T test we get a P value of .00137, thus we are able to reject the null hypothesis and proceed to examine the effect of the transmission on MPG. We can see from the estimate the mean of automatic cars transmission is about 7 MPG worse than a manual.

## Linear Modeling

We will now proceed to develop and test accuracy of various linear models from the data provided.

The first model we will test will be the simple fit model

```
lm.trans= lm(mpg~am, data = mtcars)
summary(lm.trans)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
## 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 ***
## am             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
```

We can see that the model accounts for 34% of the change, and works at the 95% confidence interval. Now we look at several slightly improved models.

We will begin by looking at the basic model against all variables.

```
lm.gen = lm(mpg~., data = mtcars)
summary(lm.gen)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4506 -1.6044 -0.1196  1.2193  4.6271
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.30337   18.71788   0.657   0.5181
## cyl         -0.11144    1.04502  -0.107   0.9161
## disp         0.01334    0.01786   0.747   0.4635
## hp          -0.02148    0.02177  -0.987   0.3350
## drat         0.78711    1.63537   0.481   0.6353
## wt          -3.71530    1.89441  -1.961   0.0633 .
## qsec         0.82104    0.73084   1.123   0.2739
```

```
## vs           0.31776    2.10451   0.151   0.8814
## am           2.52023    2.05665   1.225   0.2340
## gear         0.65541    1.49326   0.439   0.6652
## carb        -0.19942    0.82875  -0.241   0.8122
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.65 on 21 degrees of freedom
## Multiple R-squared:  0.869,  Adjusted R-squared:  0.8066
## F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

From the above model we can see that it accounts for about 80% of the change of the MPG, but none of the variables are signifiganct at the 95% confidence interval. Now we will eliminate the variables that are covariant.

```
imp.gen <- step(lm.gen, k=log(nrow(mtcars)))
```

```
summary(imp.gen)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## 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 ***
## am            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
```

We can see the model has improved where the new model now accounts for the 83% of the change on MPG. Additionally we can see that see that all variables are siginifigant under the 95% confidence interval.

From the above above we will select the improved general model because each variable is signifigant and it explains the largest of change in MPG.

## Residual Analysis and Diagnostics

There are 4 standard types of error to check for in regression plots. See the Appendix for the graphics of the below phenomena.

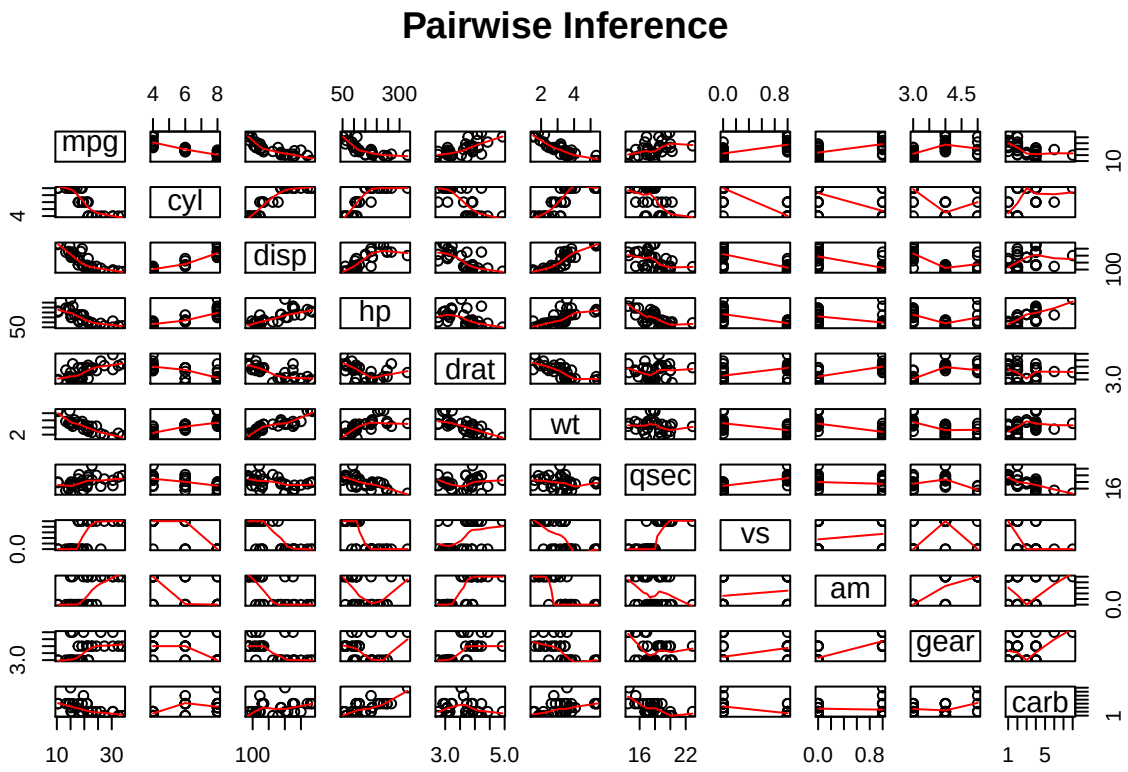
1. The Residual Plot shows no trends.

2. The normal Q-Q plot shows no negative trends.
3. The Scale Location plot shows no abnormal variation.
4. The Residuals and Leverage plot shows no outliers.

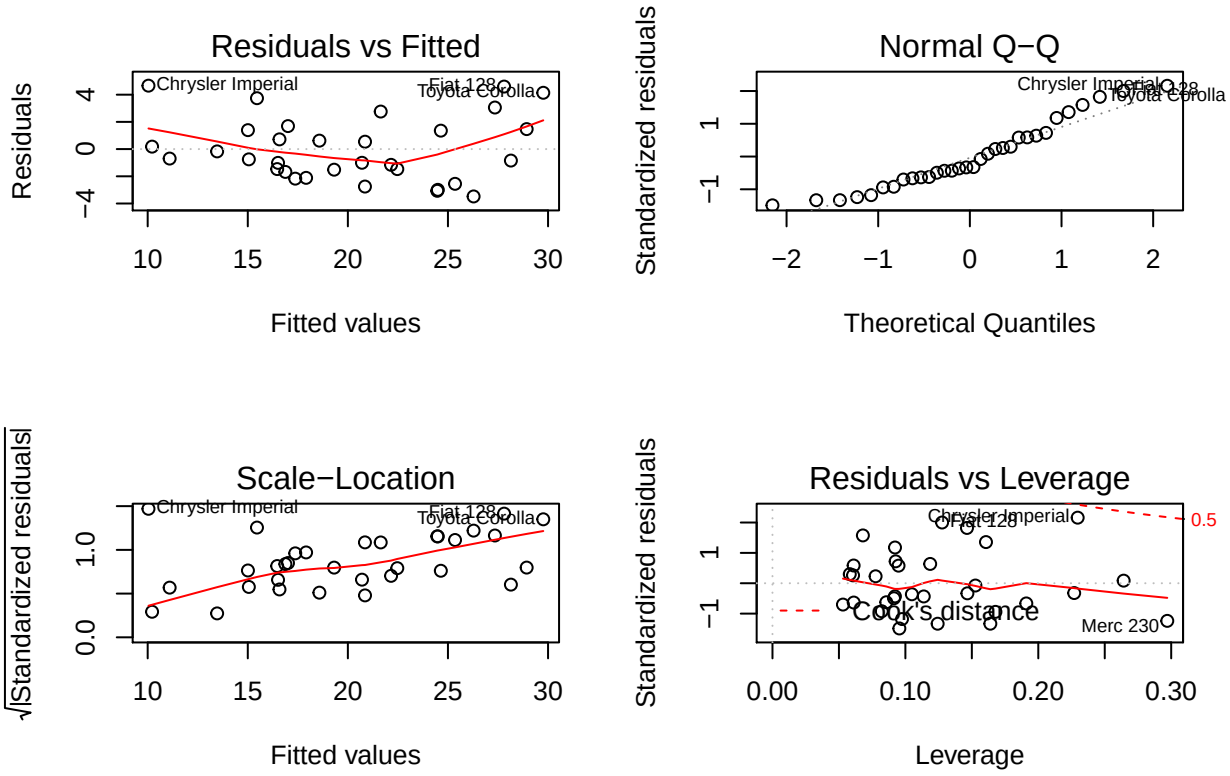
# Appendix for Graphics

1. Pairwise Graphic for visualization

```
pairs(mtcars,panel = panel.smooth,main="Pairwise Inference")
```




```
par(mfrow=c(2,2))
plot(imp.gen)
```


Residuals vs Fitted
Residuals
Fitted values
Chrysler Imperial
Fiat 128
Toyota Corolla
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Residuals vs Leverage
Cook's distance
0.5
Merc 230
Leverage