Regression Model - Assessment

Executive SummaryExecutive Summary

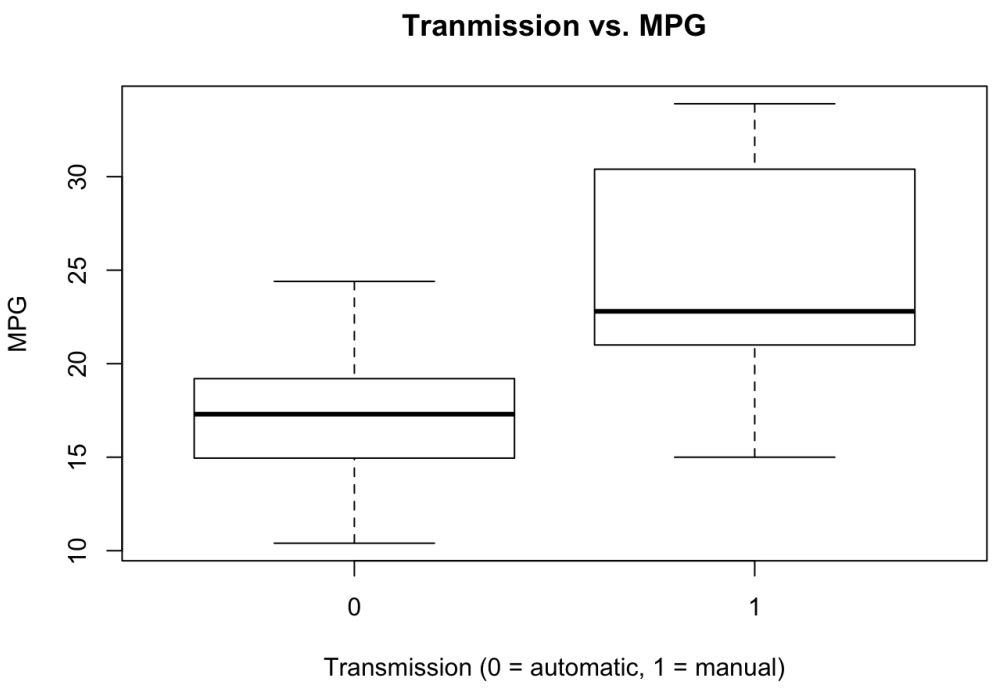
The questions posed are:

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

Throught analysis I have found that whilst manual transmission did appear to have a positive impact on MPG it wasn't the full story as weight and clyinders were the most statistically significant variables and the sample (appendix 2) showed that the manual cars tended to be lighter with less cylinders and there appear more fuel efficient but transmission would need to be further testing.

Regression model to test transmission against MPG

The use of a boxplot is a great visual tool to understand the relationships in the data and from this one we can see that there is relationship between tranmission and MPG; with manual transmissions getting more fuel efficiency.




```
# The library is one of the demo one's from R-Studio, first thing I do is boxplot the tr ansmission against MPG
boxplot(mpg ~ am, data = mtcars, main = "Tranmission vs. MPG", xlab = "Transmission (0 = automatic, 1 = manual)", ylab = "MPG")MPG")
```

I want to better quantify this position so I construct a simple regression model with transmission as predictor (1 as manual and 0 as automatic) and MPG as the outcome.

```
mtcarsmodel1 <- lm(mpg ~ am, data = mtcars) summary(mtcarsmodel1)
```

Transmission accounts for 36% of the variance (so still a lot of factors in play) in fuel consumption, which is statistically significantly better than chance ($p < .0003$). The intercept shows the baseline MPG of 17.15 miles per gallon and the coefficients shows a 7.24 increase where the transmission is manual, however we do need to look at the other data to understand the overall effects.

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars) ##
## Residuals:
##    Min     1Q Median     3Q    Max ## -9.392 -3.092 -0.297  3.244  9.508 ##
## Coefficients:

##                Estimate Std. Error t value Pr(>|t|)
## (Intercept) ##17.15               1.12   15.25  1.1e-15 ***
am
                    7.24          1.76      4.11  0.00029 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## Residual standard error: 4.9 on 30 degrees of freedom
## Multiple R-squared:  0.36,   Adjusted R-squared:  0.338
## F-statistic: 16.9 on 1 and 30 DF,  p-value: 0.000285
```

Regression model to test the other variables

The first thing that I am checking is the analysis of variance across all the variables.

```
analysis <- aov(mpg ~ ., data = mtcars)
summary(analysis)
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## cyl            1    818     818  116.42  5e-10 ***
## disp           1     38      38    5.35 0.0309 *
## hp ##          1      9       9    1.33 0.2610
drat ##          1     16      16    2.34 0.1406
wt ##            1     77      77   11.03 0.0032 **
qsec ##          1      4       4    0.56 0.4617
vs ## am         1      0       0    0.02   0.8932
                 1     14      14    2.06   0.1659
 ## gear          1      1       1    0.14   0.7137
 ## carb          1      0       0    0.06 0.8122
## Residuals   21   147               7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

So if the p-value is very low, then there is a higher probability that we are seeing data that is counter-indicative of zero effect i.e. having an effect. With this in mind I will look more into cyl (no. cylinders), disp (displacement) and wt (weight) - still retaining am given its our focus.

Appendix 3 also shows us that we have little fear of heteroscedascity, we see normality of residuals and also the residuals increase with the fitted values (which is normal).

```
lm <- lm(mpg ~ cyl + disp + wt + am, data = mtcars)
summary(lm)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + disp + wt + am, data = mtcars) ##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.318 -1.362 -0.479  1.354  6.058
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
 ## (Intercept)  40.8983 ##     3.6015   11.36  8.7e-12 ***
    cyl          -1.7842    0.6182   -2.89    0.0076 **
## disp             0.0074       0.0121 0.61          0.5451
## wt              -3.5834       1.1865   -3.02   0.0055 **
## am               0.1291       1.3215      0.10     0.9229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## Residual standard error: 2.64 on 27 degrees of freedom
## Multiple R-squared:  0.833,  Adjusted R-squared:  0.808
## F-statistic: 33.6 on 4 and 27 DF,  p-value: 4.04e-10
```

At this point we can see the high p-value indicates that transmission (in this model) has a high probability of 0 effect, now I want to see the effects with transmission removed - I also remove disp given its high p-value too.

```
lm <- lm(mpg ~ cyl + wt, data = mtcars)
summary(lm)
```

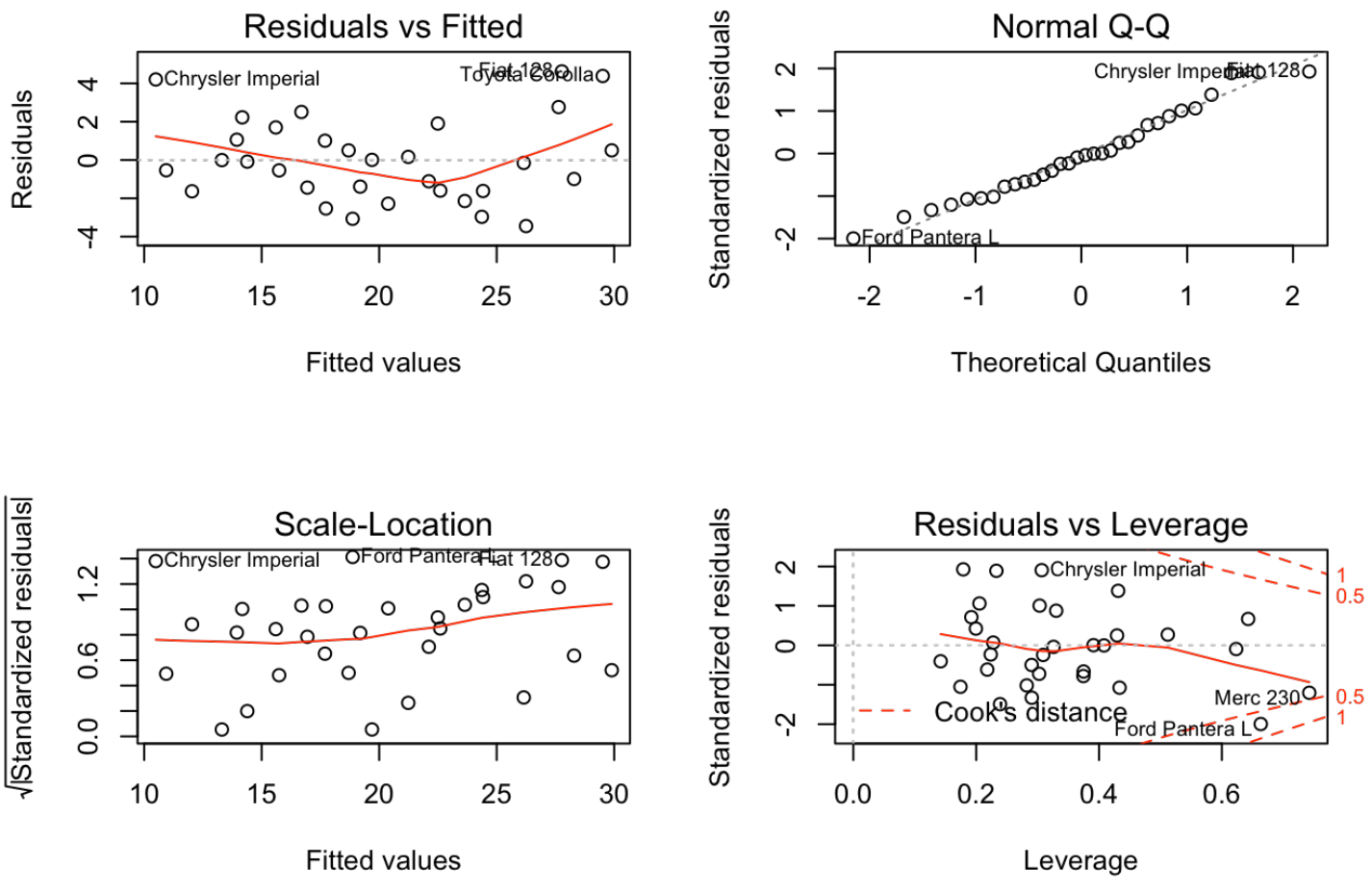
```
##
## Call:
## lm(formula = mpg ~ cyl + wt, data = mtcars) ##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.289 -1.551 -0.468  1.574  6.100
##
## Coefficients:

##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.686           1.715   23.14  < 2e-16 *** 0.415
## cyl              -1.508 -3.64  0.00106 ** 0.757   -4.22
## wt               -3.191 0.00022 ***
## ---
## 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
```

So in conclusion, cylinders and weight seem to be the biggest influece on MPG but how come we saw such a dramatic impact of transmission at the start? The final check I have done is in Appendix 2 to better understand the types of car we are looking at. What we see is that, in our sample, the automatic cards tend to be heavier and have more cylinders and the manuals are lighter and less cylinders, therefore, given our last model we would expect manuals to see better MPG so there is more at play here than simply transmission and would need to test equitable samples to discover more.

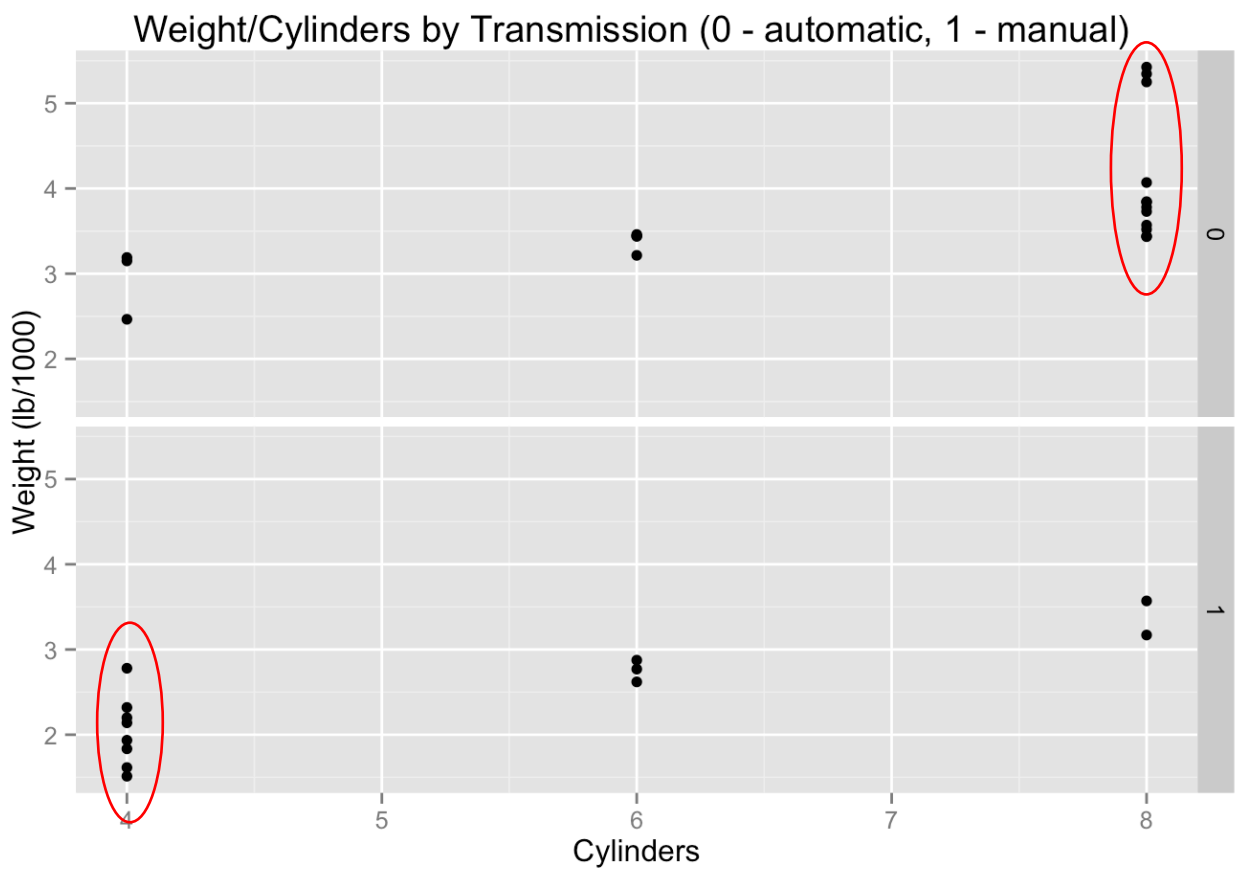
Appendix 1 - Model diagnostics

```
library(ggplot2)
par(mfrow = c(2, 2))
plot(analysis)
```



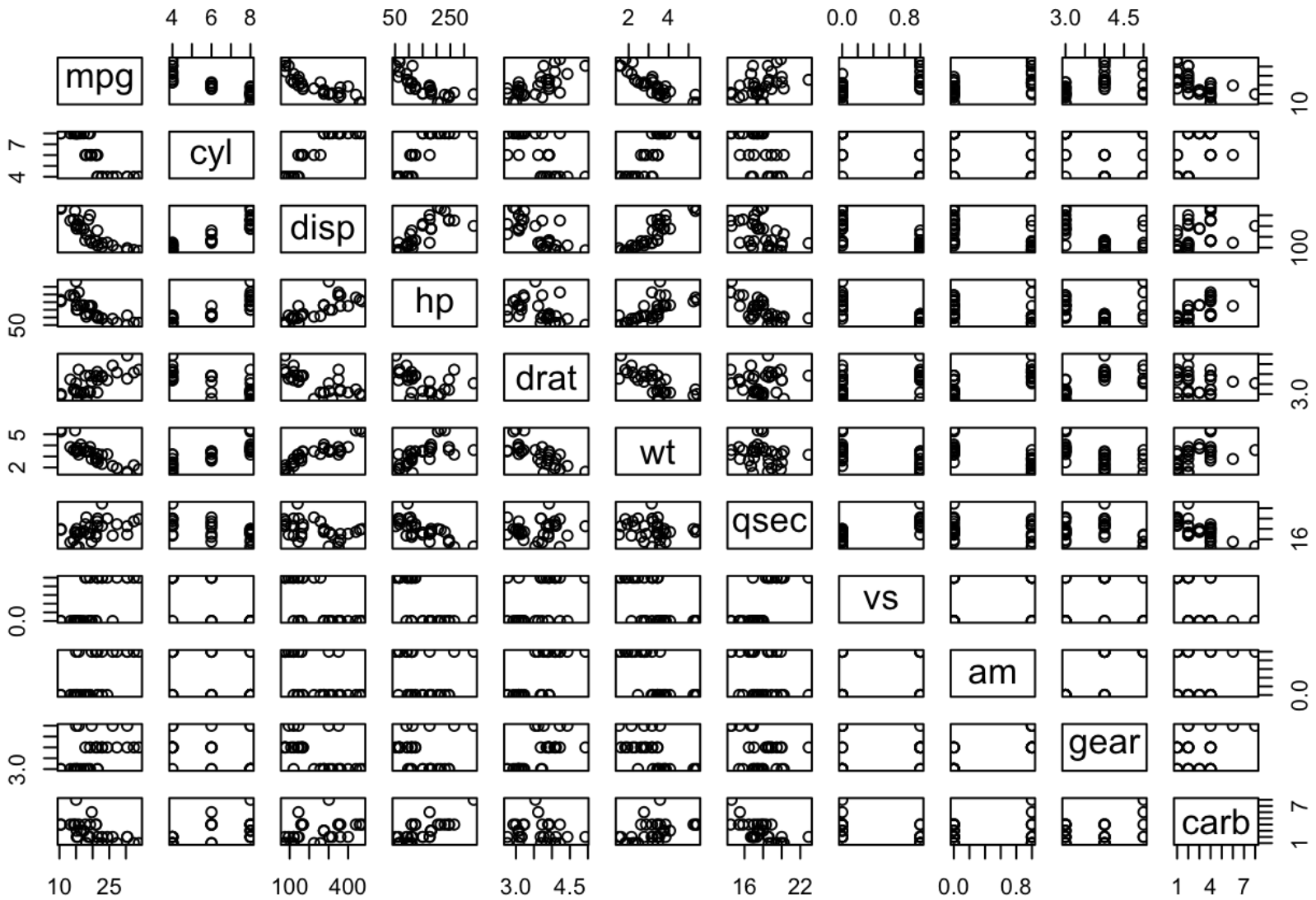
Appendix 2 - Understanding the carsAppendix 2 - Understanding the cars

```
qplot(cyl, wt, data=mtcars, main="Weight/Cylinders by Transmission (0 - automatic, 1 - m anual) ", xlab =
"Cylinders", ylab = "Weight (lb/1000)") +
  facet_grid(am ~ .)
```



Appendix 3 – Bivariate plots of variables

```
pairs(mtcars)
```



ScubaSteve22. August 2014