Does Choice of Transmission Affect MPG?

Executive Summary

In this analysis we explore the `mtcars` dataset in attempt to determine which type of transmission is better `mpg`, automatic or manual.

The quick take away is that our analysis showed, yes, there is a difference and that having a **manual** transmission will get you around **2.9** more miles per gallon than an automatic transmission.

Exploratory Data Analysis

In this section we will explore various relationship among the variables in the data. We are interested in the effects of transmission type on `mpg` and boxbplot of `mpg` to `am` (the transmission variable) ***figure 1*** shows an increase in `mpg` when the transmission is `Manual`.

However, there are other variables that could also impact mpg and they should also be considered. Plotting the data with a pairs plot, ***figure 2***, we see that the variables `cyl`, `disp`, `hp`, `wt`, `drat`, and `am` seem to have a strong correlation with `mpg`.

Regression Analysis

In this section we will build linear regression models and compare them using `anova`. After model selection we will perform some residual analysis.

Model Building and Selection

In this section we start building a linear model based on different variables. We first build a simple model with only `am` as the predictor, `lm1 <- lm(mpg ~ am, data=cars)`.

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.134e-15
## amManual       7.245      1.764   4.106 2.850e-04
```

```
## [1] 0.3385
```

Based on this simple model it appears the transmission type is significant with a p-value $< .05$. However, the adjusted R^2, **.34** for this model implies the model only explains about **34%** of the variation.

We can do better. We will build a model based on all the predictors, then use the step function to perform a forward and backwards stepwise selection.

```
lm2 <- lm(mpg ~ ., data=cars)
slm2 <- step(lm2, direction="both")
```

```
summary(slm2)$coef; summary(slm2)$adj.r.squared
```

```
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept)    9.618     6.9596   1.382 1.779e-01
## wt            -3.917     0.7112  -5.507 6.953e-06
## qsec           1.226     0.2887   4.247 2.162e-04
## amManual       2.936     1.4109   2.081 4.672e-02
## [1] 0.8336
```

From the above we see that we get an adjusted R^2 value of **.83** which us substantially better, and `wt`, `qsec`, and `amManual` all have **p-values** $<$ .05. Using `anova` to compare the our first simple model with the new model we get:

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ wt + qsec + am
##   Res.Df RSS Df Sum of Sq    F  Pr(>F)
## 1     30 721
## 2     28 169  2       552 45.6 1.6e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Looking at the **p-value** we see that results from the second model are significantly differernt and we reject the null hypothesis that `wt` and `qsec` do not contribute to the model accuracy.

Residuals and Diagnostics

Looking at the residual plots of the regression model, ***figure 3***, we can make the following observations:

- The points on the `Risduals vs Fitted` appear randomly scattered indicating supporting the independence condition.
- The `Normal Q-Q` plot points fall close to the line indicating that the risiduals are normally distributed.
- The `Scale-Location` plot has randomly scatted points in a constant band, indicating that the variance is constant.
- The `Risiduals vs Leverage` plot shows some points that could be exhibiting extra influence.

Following points show the three points with the most leverage and the three points with the greatest impact on the coefficients (using *Cooks Distance*)

```
##            Merc 230 Lincoln Continental   Chrysler Imperial
##              0.2970              0.2642              0.2296
```

```
## Chrysler Imperial          Merc 230          Fiat 128
##            0.3476            0.1621            0.1464
```

Conclusion

Our analysis has show that based upon our best fit modelwe can conclude that:

- For a given weight and quarter mile time, having a `Manual` transmission will results in higher gas milage, about **2.9**`mpg` than having and `Automatic` transmission.
- For every one second slower in the quarter mile `mpg` will increase by **1.2**`mpg`.
- For every 1000 lbs. increase in weight `mpg` will drop **3.9**`mpg`.
- Our model explains **83%** of the variation in `mpg`

Appendix

Figure 1: Boxplot of mpg and transmission

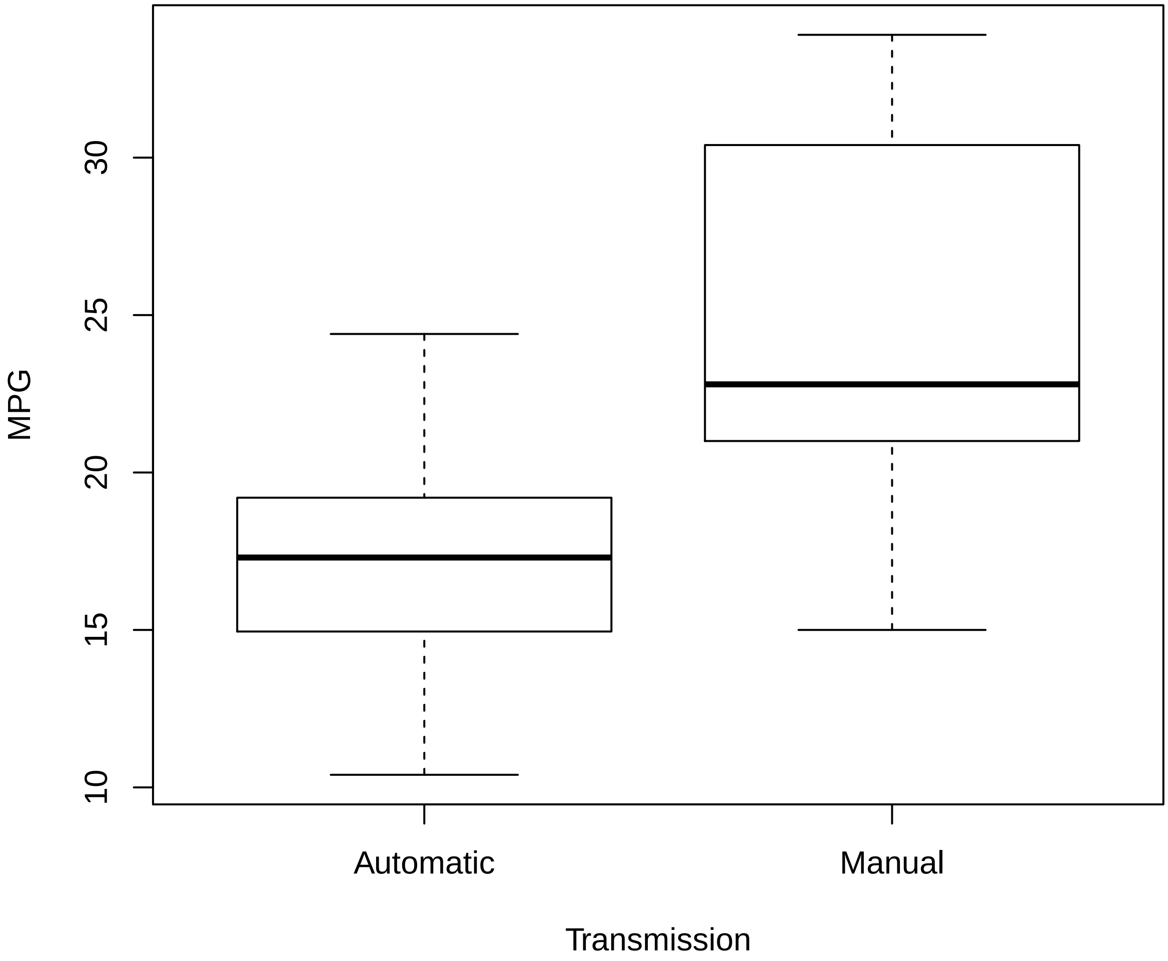


Figure 2: Pairs plot of mtcars variables

```r
## put (absolute) correlations on the panel,
## with size proportional to the correlations.
panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...)
{
    usr <- par("usr"); on.exit(par(usr))
    par(usr = c(0, 1, 0, 1))
    r <- abs(cor(x, y))
    txt <- format(c(r, 0.123456789), digits = digits)[1]
    txt <- paste0(prefix, txt)
    if(missing(cex.cor)) cex.cor <- 0.8/strwidth(txt)
    text(0.5, 0.5, txt, cex = cex.cor * r)
}
pairs(mtcars, upper.panel=panel.smooth, lower.panel=panel.cor, col=3 + (mtcars$am == 0))
```

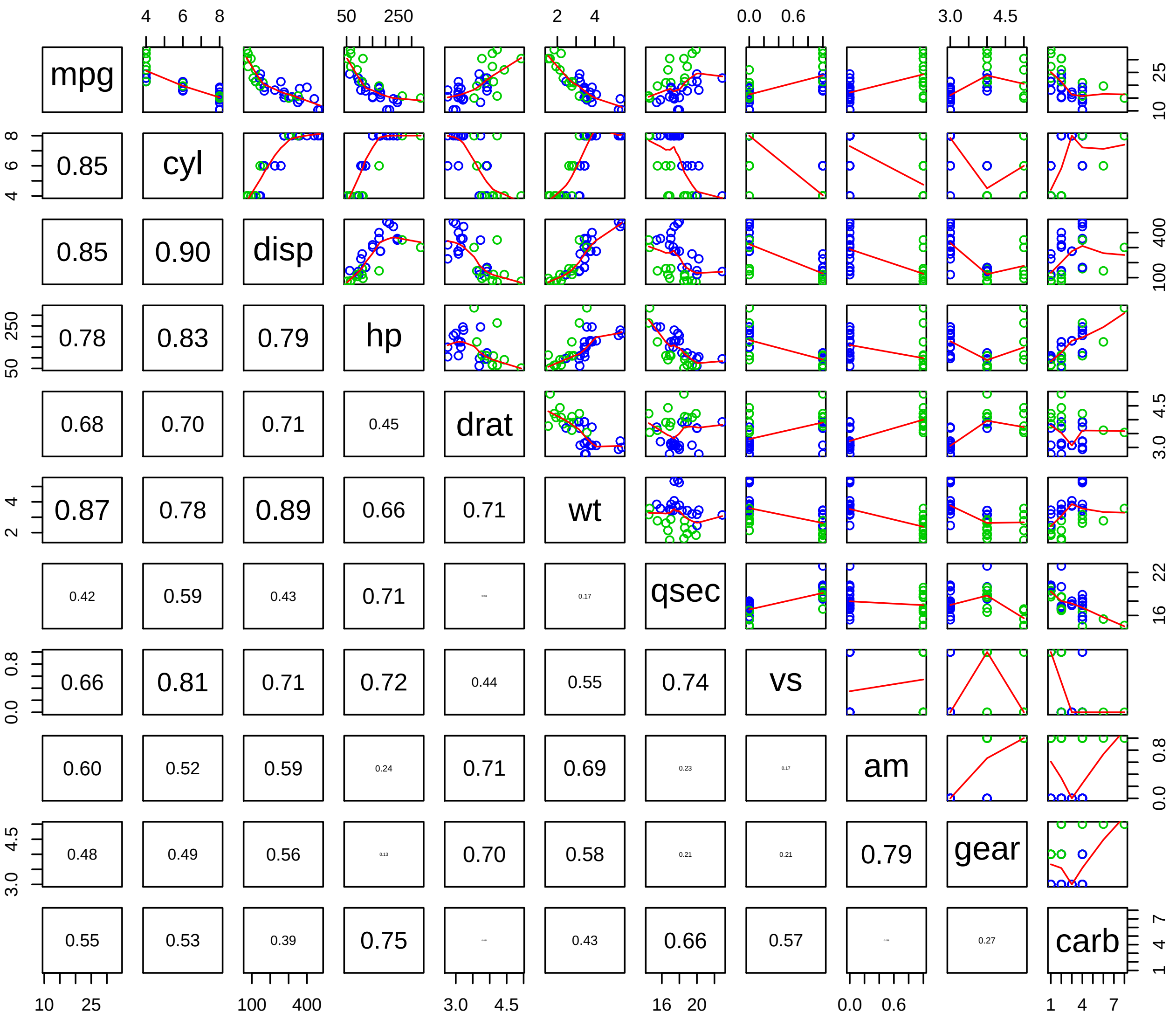
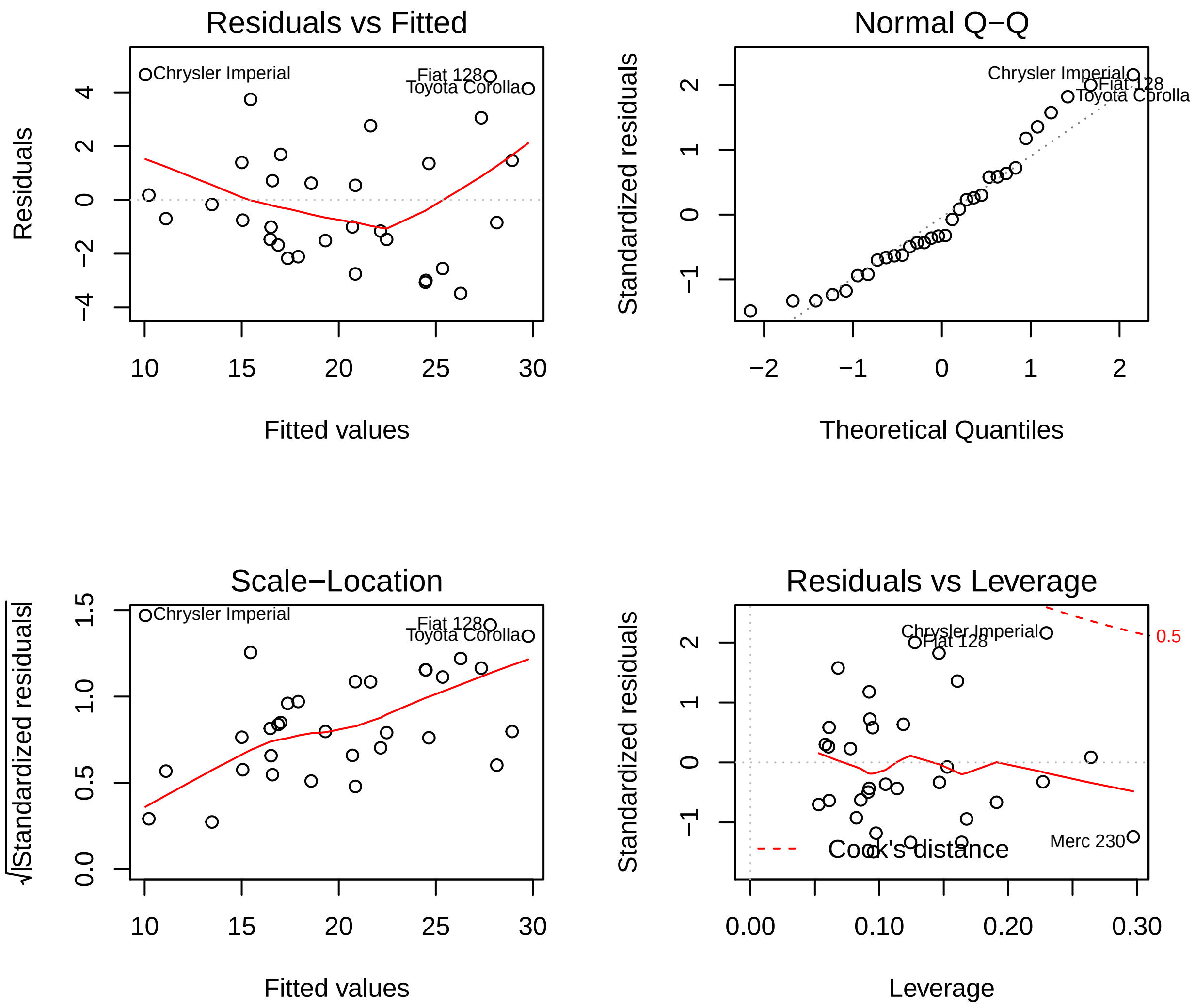


Figure 3: Residuals Plot

```
par(mfrow=c(2,2))
plot(slm2)
```



Data Preparation Code

```r
# Prepare data
library(datasets)
cars <- mtcars
cars$am <- factor(mtcars$am, labels=c("Automatic", "Manual"))
cars$cyl <- factor(mtcars$cyl)
cars$vs <- factor(mtcars$vs)
```