Motor Trend Report

Tuesday, March 17, 2015

Executive Summary

This project is aimed at determining whether automatic or manual transmission provides better fuel economy for cars. For the analysis the data used is extracted from the 1974 Motor Trend magazine, which is stored in a data set called *mtcars*. In order to address the issue of fuel economy, this project first identifies any relationships between the variables. It then uses regression to try and fit a linear model to the data so that it becomes simple to see how miles per gallon (*mpg*) a car be expected to achieve, accounting for any other influencing factors.
The results of the regression indicate clearly that transmission type alone is not a good enough predictor of fuel economy. Other factors, such as number of cylinders and weight, are far more significant predictors. The final model produced can explain roughly 83.5% of the observed data, making it a fairly reliable model.

Exploratory Data Analysis

The aim of our exploratory analysis of the mtcars dataset will be to try and identify relationships between the variables so that we have all the information needed to accurately produce some regressions models.
To do this the first step will be to perform an analysis of variance to quickly assess how 10 other variables relate to our key variable, *mpg*. The results of this are shown here in the table below (see appendix marked "Analysis of Variance":

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cyl	1	817.71	817.71	116.42	0.0000
disp	1	37.59	37.59	5.35	0.0309
hp	1	9.37	9.37	1.33	0.2610
drat	1	16.47	16.47	2.34	0.1406
wt	1	77.48	77.48	11.03	0.0032
qsec	1	3.95	3.95	0.56	0.4617
vs	1	0.13	0.13	0.02	0.8932
am	1	14.47	14.47	2.06	0.1659
gear	1	0.97	0.97	0.14	0.7137
carb	1	0.41	0.41	0.06	0.8122
Residuals	21	147.49	7.02		

As we can see, three variables have been identified as significant (P-Value below 0.05). These variables are the number of cylinders (cyl), the displacement (disp) and the weight (wt). These three will therefore be important to consider when producing our regression models.
Another important aspect is to identify confounders in the data. To do this a simple correlation matrix will be sufficient. This is displayed in the appendix marked "Correlation Matrix". This shows some strong relationships between the variables, in particular between *mpg* and *wt*, *cyl* and *disp*, and also between *disp* and *wt*. This again backs up the earlier finding that it will be important to involve weight, cycliders and displacement when analysing fuel economy (mpg).

Regression Analysis and Diagnostics

Now, based on what we have observed, we will now produce some regression models (see appendix marked "Regression Models") to try and establish the effects that variabes have on *mpg*, with particular focus on the impacts of automatic or manual transmission.

First, see the appendix marked "Transmission Boxplot" for a simple box plot displaying the observed relationship between *mpg* and *am* (transmission). This clearly shows that when any confounding variables are ignored, manual transmission is better for fuel economy in the vast majority of cases. This conclusion can also be shown using the regression model `lm(mpg ~ am, data=mtcars)`:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	17.1474	1.1246	15.25	0.0000
am	7.2449	1.7644	4.11	0.0003

The coefficient in this linear model suggests a manual transmission will increase the miles per galon the car can achieve by 7.25. This supports what has been observed in the boxplot. However, the adjusted R-squared value for this model is very low at 0.3385, which implies only 33.85% of our observed data can be explained by the model. **It is clear therefore, the transmission type alone is not enough to accurately predict fuel economy.**

As observed in the exploratory analysis, other variables are having an influence on miles per galon. So we need to incorporate them into our model. This is shown in the next regression model `lm(mpg ~ am + wt + cyl + disp, data=mtcars)`, which gives the following:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	40.8983	3.6015	11.36	0.0000
am	0.1291	1.3215	0.10	0.9229
wt	-3.5834	1.1865	-3.02	0.0055
cyl	-1.7842	0.6182	-2.89	0.0076
disp	0.0074	0.0121	0.61	0.5451

This model is much better, giving an adjusted R-squared of 0.8079. It strongly suggests that both weight and the number of cylinders are important influencers, while displacement and transmission are not significant. But, again from the exploratory analysis, we know there is some confounding going on here, particularly between the number of cylcinders and the displacement. Therefore, this final model, `lm(mpg ~ am + wt + cyl + disp + cyl:disp, data=mtcars)` should be better still as it takes that confounding into account. We get the result:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	51.6605	5.6858	9.09	0.0000
am	-0.8164	1.2893	-0.63	0.5321
wt	-2.9840	1.1286	-2.64	0.0137
cyl	-3.1237	0.8101	-3.86	0.0007
disp	-0.0950	0.0452	-2.10	0.0455
cyl:disp	0.0121	0.0052	2.34	0.0274

Now we have an adjusted R-squared of 0.8351, indicating that around 83.5% of our observed data can be explained by our model. We can also see that all of the variables are within 95% significance, except for transmission type. **This is further evidence that although transmission type might have some impact on fuel economy, other factors are far more important in predicting it accurately.**

Finally, just to be sure that our final model is reliable, the residual plots (appendix marked "Residual Plots") confirm that our model has little bias. The Normal Q-Q plot in particular clearly shows this as the points lie very close to the line.

It is worth noting, however, that uncertainty will appear in any prediction model. While 83.5% seems like a high amount, it still means that 16.5% of the observed data remains unexplained. We also have 4 out of 5 of our prediction terms showing significance, but removing any one of these can drastically change that result. So in conclusion, this model is good, but I have no doubt that it could be better!

Appendix

Graphs

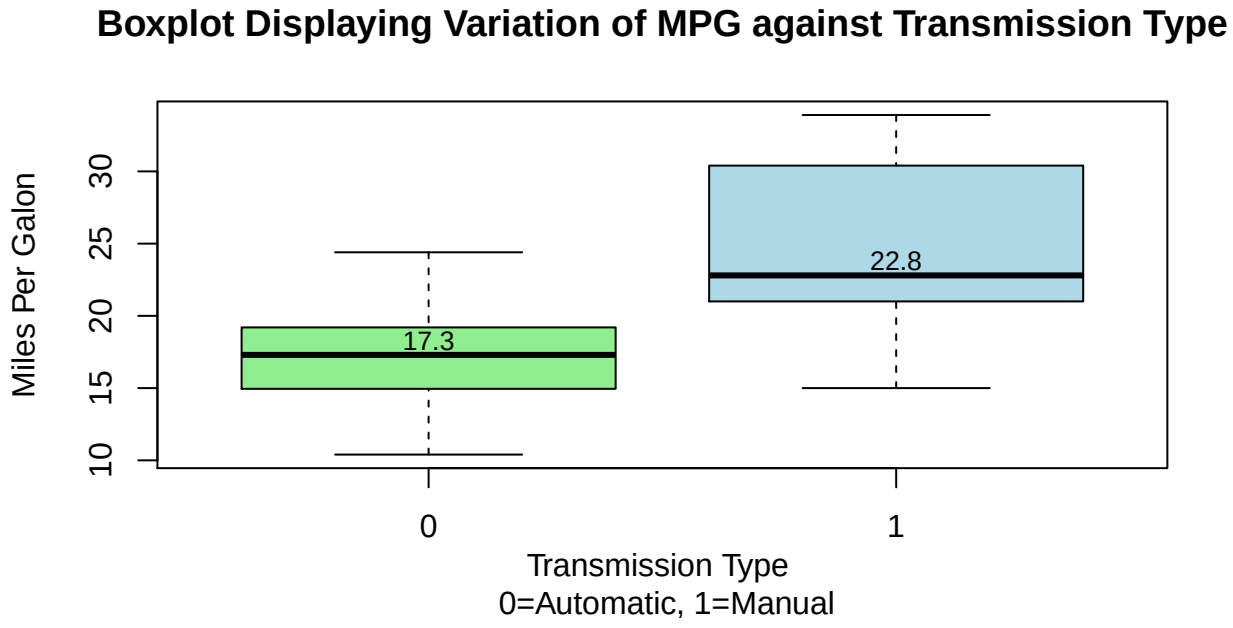
Correlation Matrix:

```
Corr <- cor(mtcars)
Corr[upper.tri(Corr)] <- NA
print(xtable(Corr), comment=FALSE)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
mpg	1.00										
cyl	-0.85	1.00									
disp	-0.85	0.90	1.00								
hp	-0.78	0.83	0.79	1.00							
drat	0.68	-0.70	-0.71	-0.45	1.00						
wt	-0.87	0.78	0.89	0.66	-0.71	1.00					
qsec	0.42	-0.59	-0.43	-0.71	0.09	-0.17	1.00				
vs	0.66	-0.81	-0.71	-0.72	0.44	-0.55	0.74	1.00			
am	0.60	-0.52	-0.59	-0.24	0.71	-0.69	-0.23	0.17	1.00		
gear	0.48	-0.49	-0.56	-0.13	0.70	-0.58	-0.21	0.21	0.79	1.00	
carb	-0.55	0.53	0.39	0.75	-0.09	0.43	-0.66	-0.57	0.06	0.27	1.00

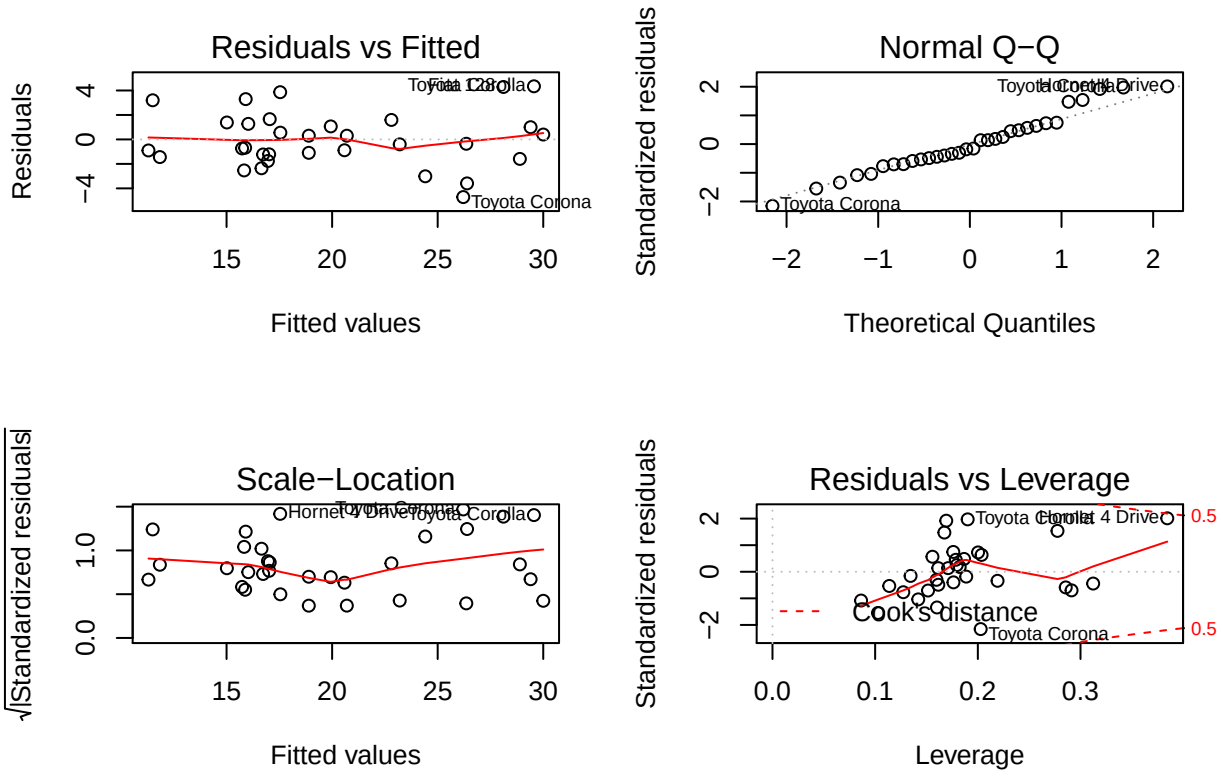
Transmission Boxplot:

```
m0 <- median(mtcars$mpg[mtcars$am==0]); m1 <- median(mtcars$mpg[mtcars$am==1])
boxplot(mpg~am, data=mtcars, xlab="Transmission Type \n 0=Automatic, 1=Manual"
        , ylab="Miles Per Galon", col=c("lightgreen","lightblue"))
title(main="Boxplot Displaying Variation of MPG against Transmission Type")
text(x=c(1,2), y=c(m0+1,m1+1),labels=c(m0,m1), cex=0.8)
```



Residual Plots:

```
par(mfrow=c(2,2))
plot(Fit3)
```



Code

Analysis of Variance:

```
#preliminary calcs
library(data.table)
library(xtable)
data(mtcars)

#ANOVA
AnVar <- aov(mpg ~ ., data=mtcars)
print(xtable(AnVar), comment=FALSE)
```

Regression Models:

```
#models
Fit1 <- lm(mpg ~ am, data=mtcars)
Fit2 <- lm(mpg ~ am + wt + cyl + disp, data=mtcars)
Fit3 <- lm(mpg ~ am + wt + cyl + disp + cyl:disp, data=mtcars)

#displaying results in table
print(xtable(Fit1), comment=FALSE)
print(xtable(Fit2), comment=FALSE)
print(xtable(Fit3), comment=FALSE)
```