MPG analysis

Alex Fennell

Executive Summary

The goal of this report was to investigate whether an automatic or manual transmission is more fuel efficient (better MPG) using the mtcars dataset. Three models were compared. In the first model, mpg was predicted by transmission type. In the second model mpg was predicted by transmission type, number of cylinders, and engine style. The final model had mpg predicted by transmission type, number of cylinders, engine style, weight, and horsepower. A model selection analysis determined that the third model was the best model and provided additional information over and above the other two models. This model determined that tranmission type did not have a significant impact on mpg. The model states that manual cars get 2.7 more mpg than automatic cars but this is not significantly different from 0.

Exploratory Data Analysis summary

Boxplots of the data

I will begin my exploratory data analysis by first examining boxplots of various variable interactions. The plot is contained in the appendix, but I will summarize it here. The first boxplot seems to show that manual cars have a better mpg than automatic cars. The lower quartile of the manual boxplot is greater than the upper quartile of the automatic boxplot. The next boxplot shows the effect of transmission type on mpg depending on how many cylinders the cars have. It seems that regardless of transmission type, increasing numbers of cylinders result in worse mpg. When examining the influence of the number of gears a car has on the fuel efficiency of automatic versus manual cars, it seems that 3 gear cars are the least efficient. Then cars with 4 gears, with manual cars being slighlty more efficient than automatic. And 5 gear cars being similar to 4 gear cars if not having slighlty worse mpg. The last boxplot examines how engine type influences the mpg of automatic and manual cars. For both automatic and manual cars, v-shaped engines result in lower mpg than straight engines. For equivalent engine types, automatic cars have lower mpg than manual cars.

Correlation between numeric variables

The main insights from the correlation matrix is that all the numeric variables have a relationship with mpg. Variables such as horsepower, displacement, and weight all have strong negative correlations with mpg (>.7). Thus, as these variables increase in magnitude, mpg will decrease. Univariate distributions show that most variables are unimodal and roughly normally distributed with varying degrees of skew.

Modeling analysis

For the modelling analysis I compare 3 different models. The first examines the effect of transmission type on mpg. The second model examines how transmission type, number of cylinders, and engine type influence mpg. These additional variables were chosen because they seemed to account for additional variance in mpg based on the boxplots. The third model examines how these predictors in addition to weight and gross horsepower affect mpg. These numeric variables were selected based on the correlation information. Horsepower and weight were both related to mpg and were less related to each other than the other variables.

Model Residual Analysis

A plot of model diagnostics for the most basic model is presented in the appendix. To summarize, the distribution of residuals for automatic and manual cars appears to be normal and equivalent, satisfying one of the assumptions of linear regression. There also appears to be no observations that have high influence, increasing confidence in the unbiasdness of the model coefficients.

Model Comparison

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl + vs
## Model 3: mpg ~ am + cyl + vs + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     27 258.90  3    462.00 26.796 5.612e-08 ***
## 3     25 143.68  2    115.22 10.024 0.0006359 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This anova table shows model comparison results. It shows that the second model with cylinders and engine type improves upon the original model with just transmission type, since the p-value is less than .05. Furthermore, the third model improves upon the second one as the p-value for that comparison is also less than .05. Thus, including weight and horsepower in addition to the predictors from second model seems to explain more of the variance in mpg, than in the model that excludes these weight and horsepower.

Model Coefficient Comparison

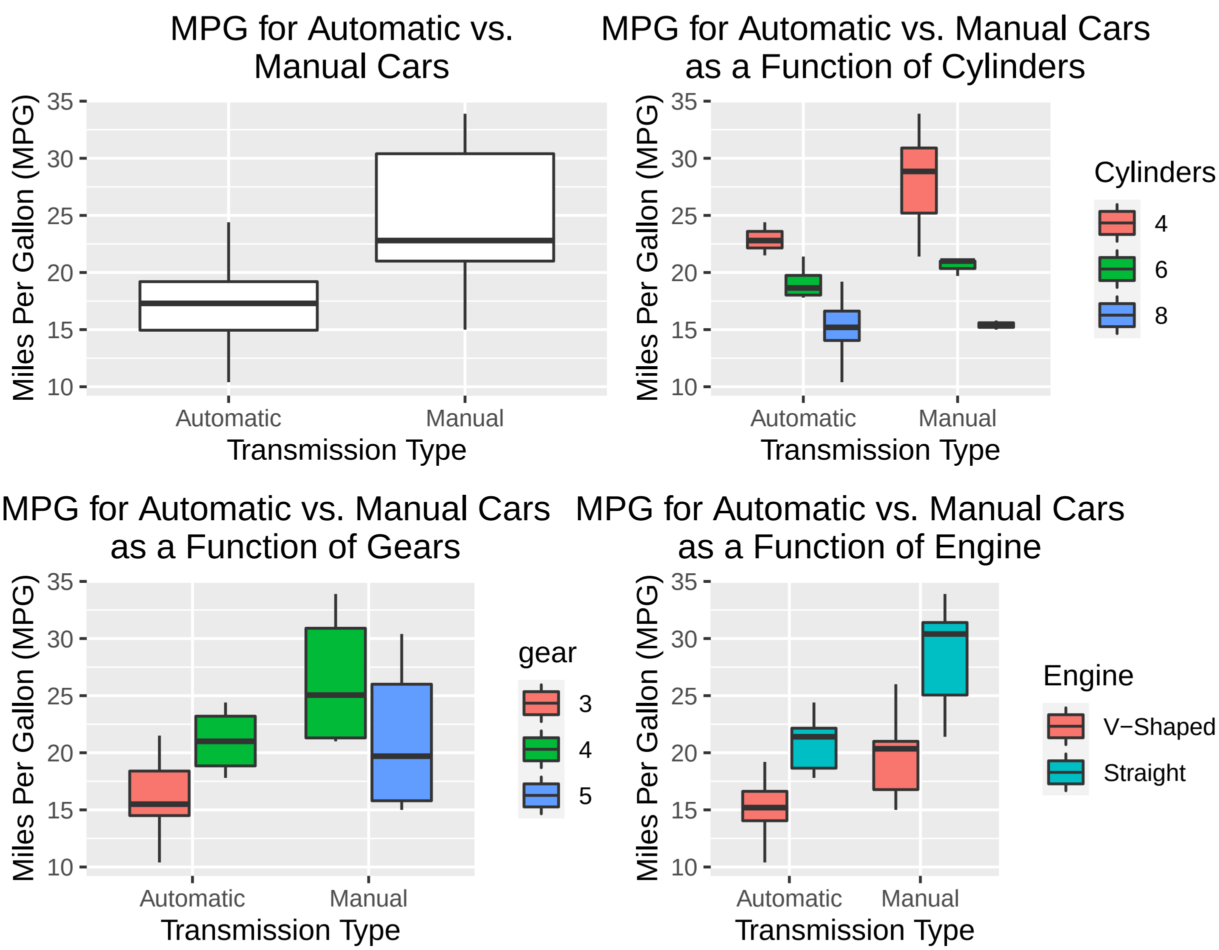
Table 1: MPG Regression Model Results

	Dependent variable:		
	mpg		
	(1)	(2)	(3)
amManual	7.245^{***} (3.787, 10.703)	3.165^{**} (0.170, 6.160)	2.704 (−0.429, 5.837)
cyl6		-5.399^{***} (−9.000, −1.797)	−2.090 (−5.282, 1.102)
cyl8		-8.161^{***} (−13.829, −2.494)	0.291 (−5.869, 6.451)
vsStraight		1.708 (−2.672, 6.089)	1.990 (−1.460, 5.440)
wt			-2.373^{**} (−4.113, −0.634)
hp			-0.035^{**} (−0.062, −0.008)
Constant	17.147^{***} (14.943, 19.352)	22.809^{***} (17.070, 28.548)	31.185^{***} (24.481, 37.888)
Observations	32	32	32
R^2	0.360	0.770	0.872
Adjusted R^2	0.338	0.736	0.842
Residual Std. Error	4.902 (df = 30)	3.097 (df = 27)	2.397 (df = 25)
F Statistic	16.860^{***} (df = 1; 30)	22.609^{***} (df = 4; 27)	28.488^{***} (df = 6; 25)
Note:			*p<0.1; **p<0.05; ***p<0.01

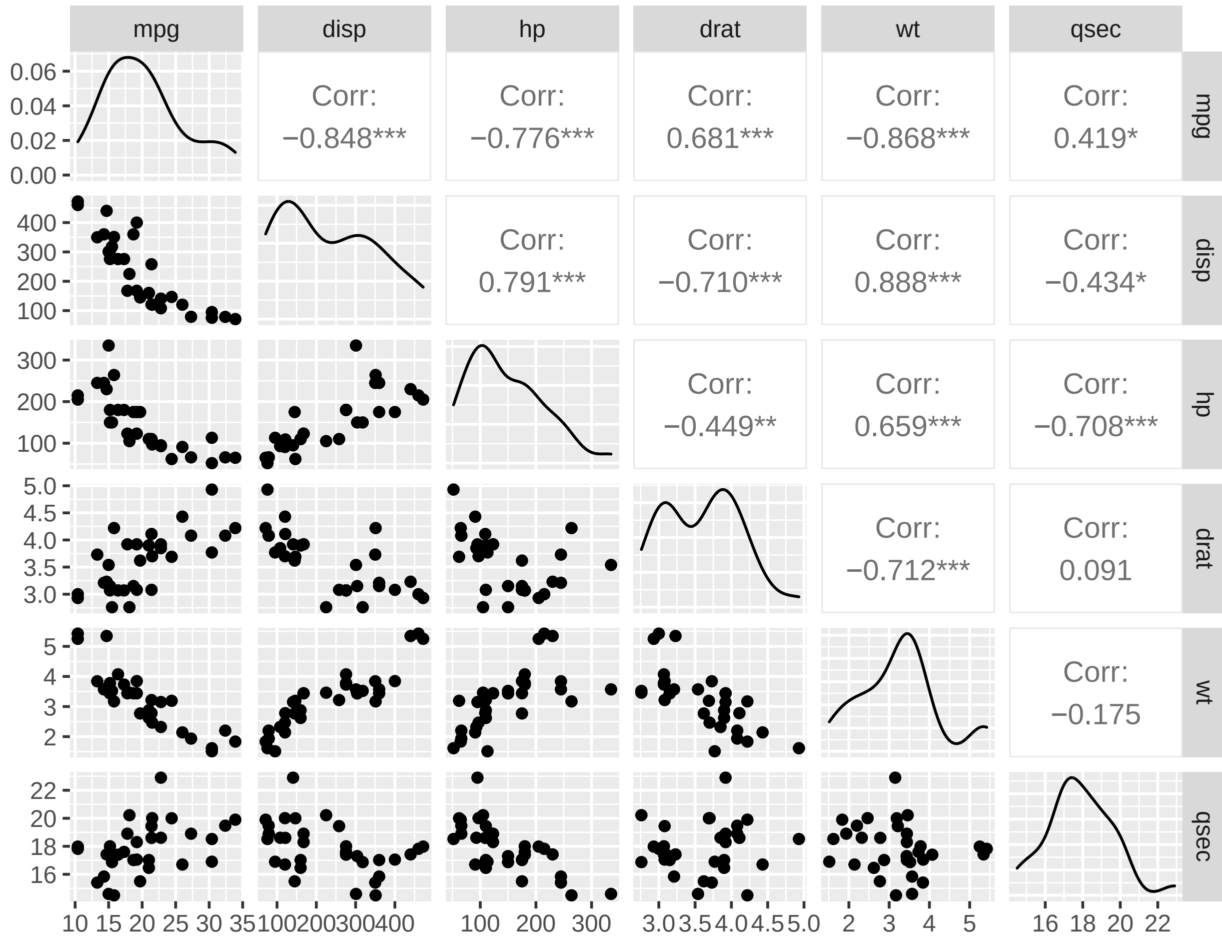
This table provides a comparison of all the model fits with model coefficients, and 95% confidence intervals in parentheses. The model in the leftmost column is the base model and shows a significant effect of transmission type on mpg, with a p-value less than .01. The model predicts mpg to be 17.15 for automatic cars, and the coefficient for amManual tells us that mpg for manual cars is 7.25 units greater and that this is a significant effect. Thus, manual cars are predicted to have 24.35 mpg on average. The 95% confidence interval for the amManual coefficient suggests that across repeated samples, 95% of the estimated coefficients will lie somewhere between 3.79 and 10.7. In the middle column, the coefficients for the second model are presented. The coefficient of 3.17 for amManual is statistically significant ($p<.05$) suggesting that manual cars with 4 cylinders, and a v-shaped engine have 3.17 more mpg than automatic cars with 4 cylinders and a v-shaped engine (22.8 vs. 25.97). The 95% confidence interval for the amManual coefficient again does not contain 0, as expected given the p-value is less than .05. The final column contains the last model with weight and horsepower added. In this model the amManual coefficient is not statistically signficant and the 95% confidence interval does contain 0. Thus, when weight and horsepower are taken into account, transmission type does not seem to be as important of a factor in influencing mpg. The amManual coefficient in this model says that manual cars of average weight and horsepower with 4 cylinders and a v-shaped engine have 2.7 more mpg on average than automatic cars. The only two significant coefficients in this model are weight and horsepower suggesting these are better at predicting mpg.

Appendix

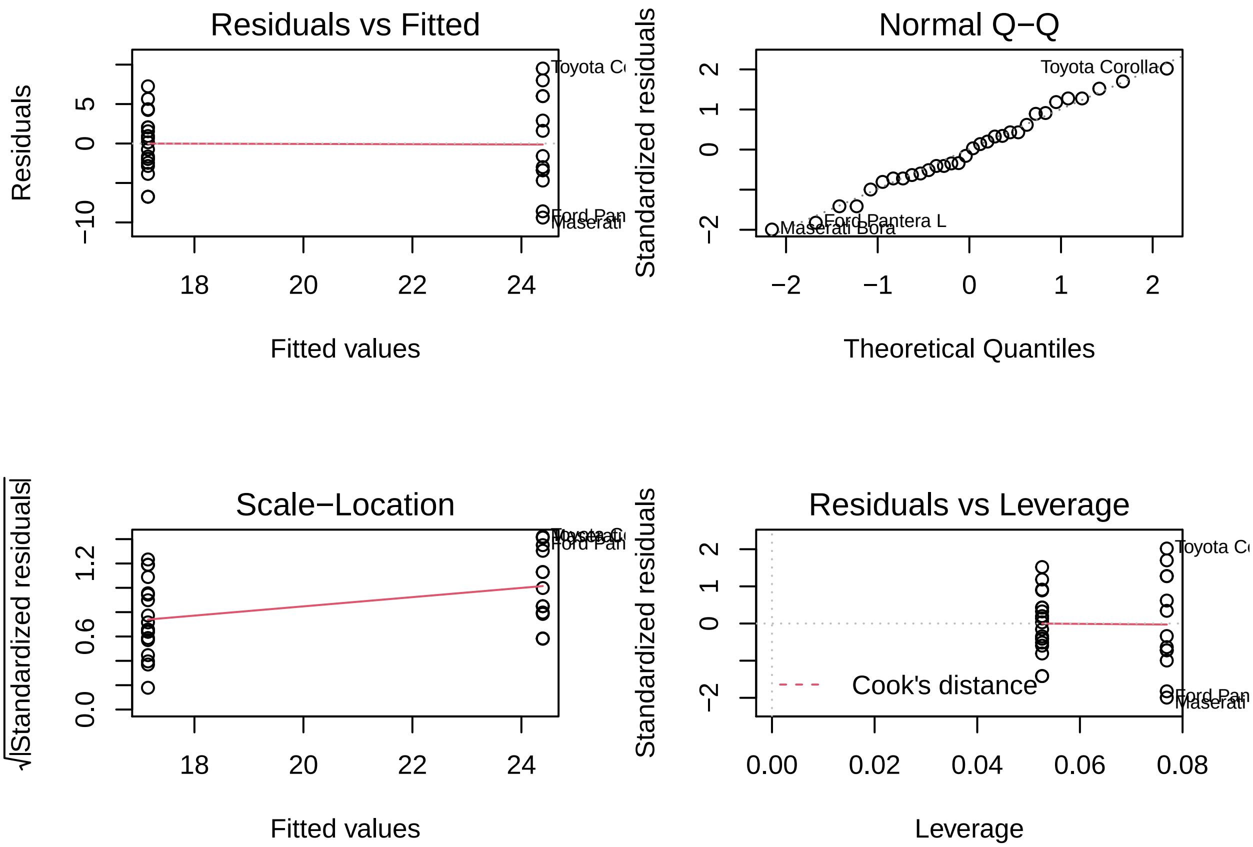
```
library(ggplot2);library(plyr);library(dplyr);library(gridExtra);library(GGally)
library(stargazer)
mtdat<-mtcars
mtdat$cyl<-factor(mtcars$cyl)
mtdat$vs<-factor(mtcars$vs,labels=c("V-Shaped","Straight"))
mtdat$am<-factor(mtcars$am,labels=c("Automatic","Manual"))
mtdat$gear<-factor(mtcars$gear)
mtdat$carb<-factor(mtcars$carb)
b1<-ggplot(mtdat,aes(y=mpg,x=am))+
    geom_boxplot()+
    xlab("Transmission Type")+
    ylab("Miles Per Gallon (MPG)")+
    ggtitle("MPG for Automatic vs. \n Manual Cars")+
    theme(plot.title=element_text(hjust=.5))

b2<-ggplot(data=mtdat,aes(y=mpg,x=am,fill=cyl))+
    geom_boxplot()+
    xlab("Transmission Type")+
    ylab("Miles Per Gallon (MPG)")+
    ggtitle("MPG for Automatic vs. Manual Cars \n as a Function of Cylinders")+
    theme(plot.title=element_text(hjust=.5))+
    guides(fill=guide_legend(title="Cylinders"))

b3<-ggplot(data=mtdat,aes(y=mpg,x=am,fill=gear))+
    geom_boxplot()+
    xlab("Transmission Type")+
    ylab("Miles Per Gallon (MPG)")+
    ggtitle("MPG for Automatic vs. Manual Cars \n as a Function of Gears")+
    theme(plot.title=element_text(hjust=.5))

b4<-ggplot(data=mtdat,aes(y=mpg,x=am,fill=vs))+
    geom_boxplot()+
    xlab("Transmission Type")+
    ylab("Miles Per Gallon (MPG)")+
    ggtitle("MPG for Automatic vs. Manual Cars \n as a Function of Engine")+
    theme(plot.title=element_text(hjust=.5))+
    guides(fill=guide_legend(title="Engine"))

grid.arrange(b1,b2,b3,b4,nrow=2)
```




```
mtdat%>%
    select_if(is.numeric)%>%
    ggpairs(.)
```




```
par(mfrow=c(2,2))
plot(fit1)
```



Residuals vs Fitted
Residuals
Fitted values
Toyota Co
Ford Pan
Maserati
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Toyota Corolla
Ford Pantera L
Maserati Bora
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
Toyota Co
Ford Pan
Maserati