Regression Model Project: Transmission Type and MPG

Executive Summary

In this report we will investigate how miles per gallon (MGP) varies by the transmission type and other variables in mtcars data set. The following research questions will be answered using linear regression models: 1) Is an automatic or manual transmission better for MPG? 2) What's the MPG difference between automatic and manual transmissions? A linear model is built to predict mpg using the transmission type, number of cylinders, gross horsepower, and weight. It is found that, although there is a significant mpg difference between the automatic transmission (mean mpg=17.1473684) and manual cars (mean mpg=24.3923077), the transmission type only contributes about 1.81 mpg difference (p=0.20) while number of cylinders, gross horsepower and weight are significant predictors.

Data Analysis

t-test

```
t.test(mtcars$mpg[mtcars$am==0],mtcars$mpg[mtcars$am==1],
       paired = FALSE,var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  mtcars$mpg[mtcars$am == 0] and mtcars$mpg[mtcars$am == 1]
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean of x mean of y
##  17.14737  24.39231
```

The average mpg of automatic cars is 17.1473684, lower than that of manual cars 24.3923077. The independent t-test suggests that the difference is significant ($p<0.05$). (See [Box Plot])

From the Scatterplot , it is obvious that 4-cylinder cars have higher MPG while 8-cylinder cars have higher MPG; also, the heavier the car is, the lower the mpg is.In addition, the Correlation Matrix suggests that mpg is strongly correlated with weight, horsepower, and displacement

Linear Model From the summary results in Model Selection, we can see fit4 (with am, cyl,wt and hp as predictors) is the most optimized model; note hp and disp are strongly correlated (r=0.79) so it's not surprising that adding disp doesn't improve the model. We decided to select fit4 and below is the summary of the model fit:

```
fit4<-lm(data = mtcars,mpg~am+cyl+wt+hp)
summary(fit4)
```

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

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## am           1.80921    1.39630   1.296  0.20646
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## wt          -2.49683    0.88559  -2.819  0.00908 **
## hp          -0.03211    0.01369  -2.345  0.02693 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

Conclusion

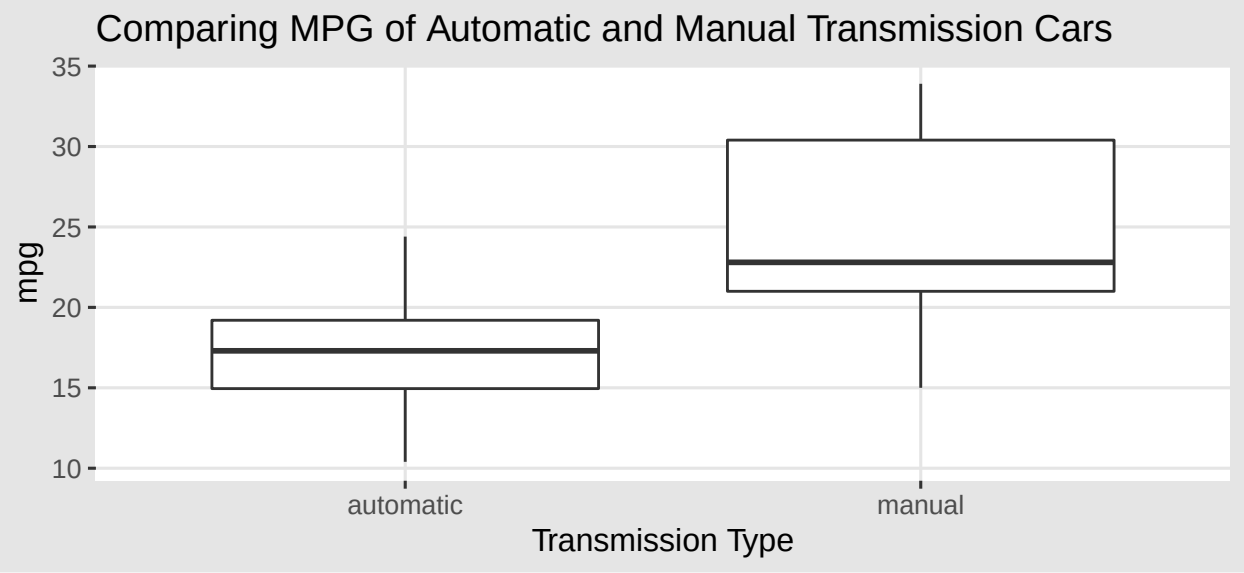
The model explains 84% of the variance. Although the difference between automatic and manual transmission cars is 1.81, when other factors (cylinder, horsepower, and weight) are controlled for, the difference is not significant (p=0.20)

Appendix

Exploratory Analysis The dataset only contaisn 32 observations - a relatively small sample size therefore the conclusion should be interpreted with causion.

Boxplot

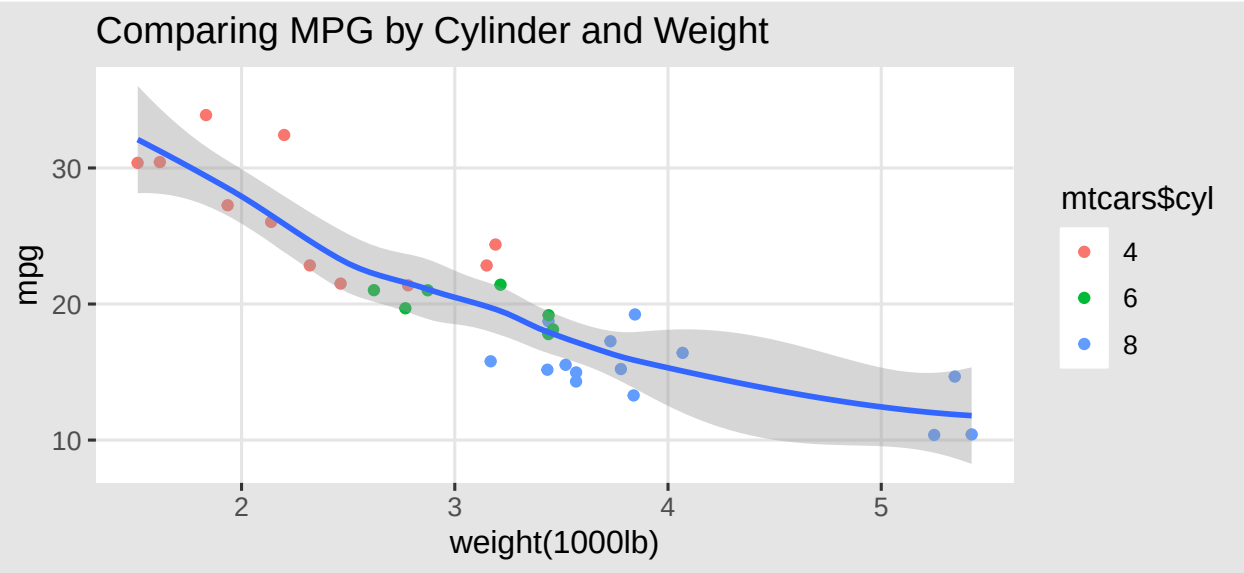
```
g<-ggplot(data = mtcars,aes(amname,mpg))
g+geom_boxplot()+
        ggtitle("Comparing MPG of Automatic and Manual Transmission Cars")+
        theme_igray()+xlab("Transmission Type")
```



Scatterplot

```
g<-ggplot(mtcars,aes(wt,mpg))
g+geom_jitter(aes(color=mtcars$cyl))+theme_igray()+
        geom_smooth()+
        ggtitle("Comparing MPG by Cylinder and Weight")+
        xlab("weight(1000lb)")
```

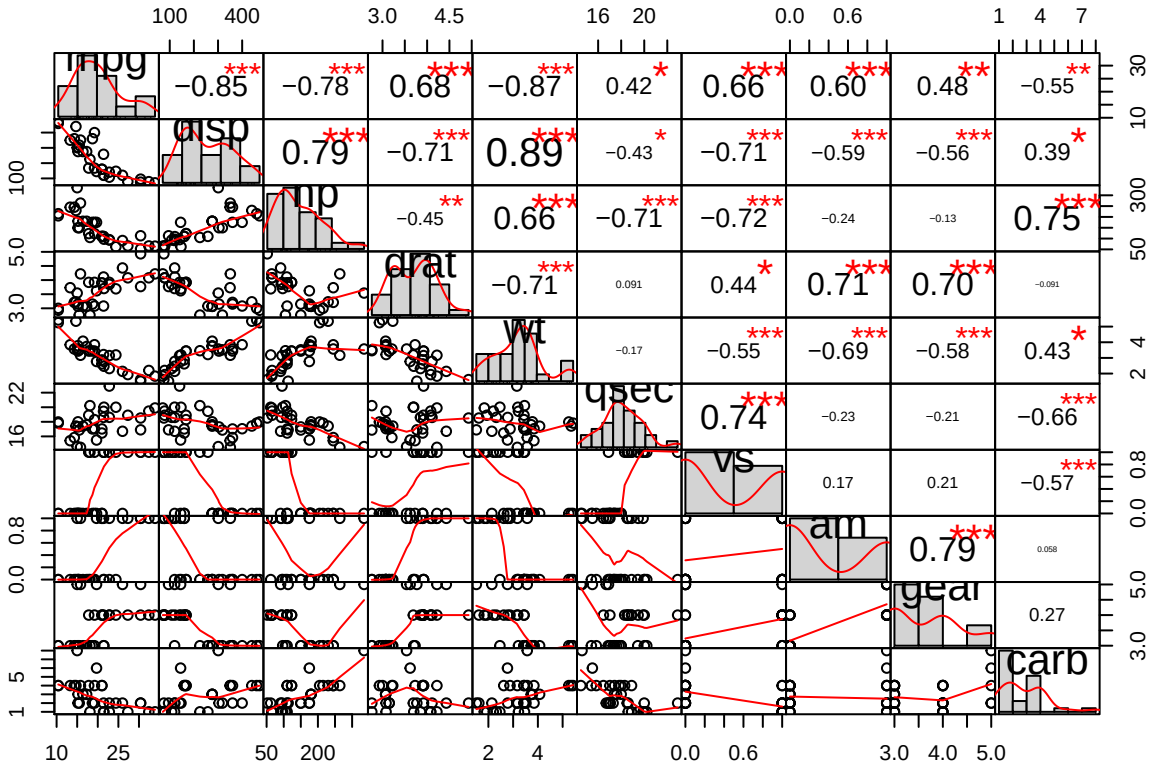
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Correlation Matrix

We now examine the correlation between mpg and other variables in the dataset.

```
mydata<-mtcars[,c(1,3:11)]
chart.Correlation(mydata,histogram = TRUE,pch=19)
```



Unfortunately, disp, hp, wt, all have a statistically significant and strong correlation ($p<0.05$) with mpg. Therefore, it is likely that they are confounding factors that affect mpg

Model Selection

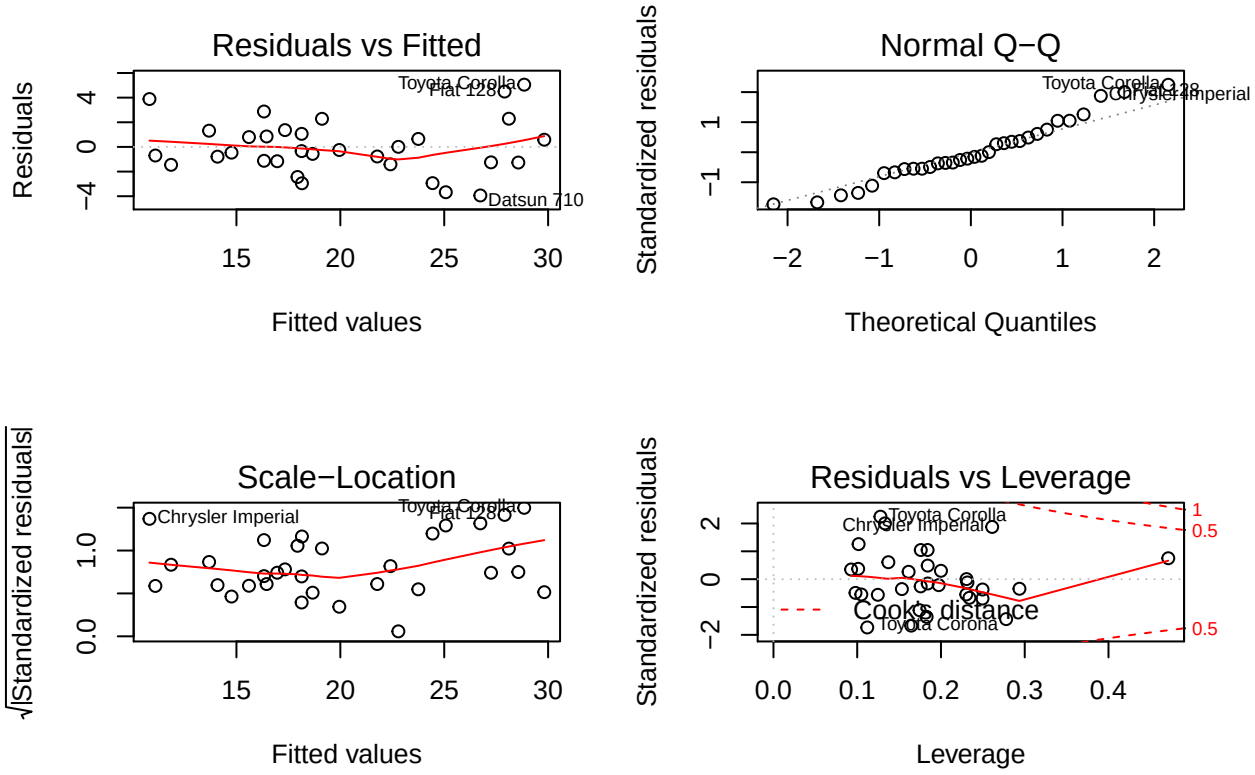
```
fit1<-lm(data=mtcars,mpg~am)
fit2<-lm(data = mtcars,mpg~am+cyl)
fit3<-lm(data = mtcars,mpg~am+cyl+wt)
fit4<-lm(data = mtcars,mpg~am+cyl+wt+hp)
fit5<-lm(data = mtcars,mpg~am+cyl+wt+hp+disp)
anova(fit1,fit2,fit3,fit4,fit5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + cyl
## Model 3: mpg ~ am + cyl + wt
## Model 4: mpg ~ am + cyl + wt + hp
## Model 5: mpg ~ am + cyl + wt + hp + disp
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
```

```
## 1     30 720.90
## 2     28 264.50  2    456.40 37.9300 2.678e-08 ***
## 3     27 182.97  1     81.53 13.5510  0.001118 **
## 4     26 151.03  1     31.94  5.3093  0.029801 *
## 5     25 150.41  1      0.62  0.1025  0.751489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual Plot

```
par(mfrow=c(2,2))
plot(fit4)
```



Interpretations of the Plots

*The residual vs. fitted values is more or less random
*The normal QQ plot should give a straight line
*Scale-location plot: should be random
* Residuals vs. Leverage: shows the Cook's distance
        -Toyota Corolla, Chrysler Imperial and Fiat seem to be outliers