Regression Models: Course Project

Bryjólfur Gauti Jónsson

Synopsis

The purpose of this analysis is to determine whether the nature of a car's transmission (manual vs. automatic) has a significant effect on the car's MPG (mile's per gallon). To do this we will utilize the **'mtcars'** dataset that comes with R.

Getting and Cleaning the Data

We will start by reading the mtcars data that comes with r and having a first look at it.

```
data("mtcars")
str(mtcars)
```

```
## 'data.frame':    32 obs. of  11 variables:
##  $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
##  $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
##  $ disp: num  160 160 108 258 360 ...
##  $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
##  $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
##  $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
##  $ qsec: num  16.5 17 18.6 19.4 17 ...
##  $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
##  $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
##  $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
##  $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

```
head(mtcars)
```

```
##                   mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

We immediately spot a few variables that are numeric but should be factors.

```
mtcars$am <- as.factor(mtcars$am)
levels(mtcars$am) <- c("Auto", "Manual")
mtcars$cyl <- as.factor(mtcars$cyl)
mtcars$vs <- as.factor(mtcars$vs)
mtcars$gear <- as.factor(mtcars$gear)
mtcars$carb <- as.factor(mtcars$carb)
```

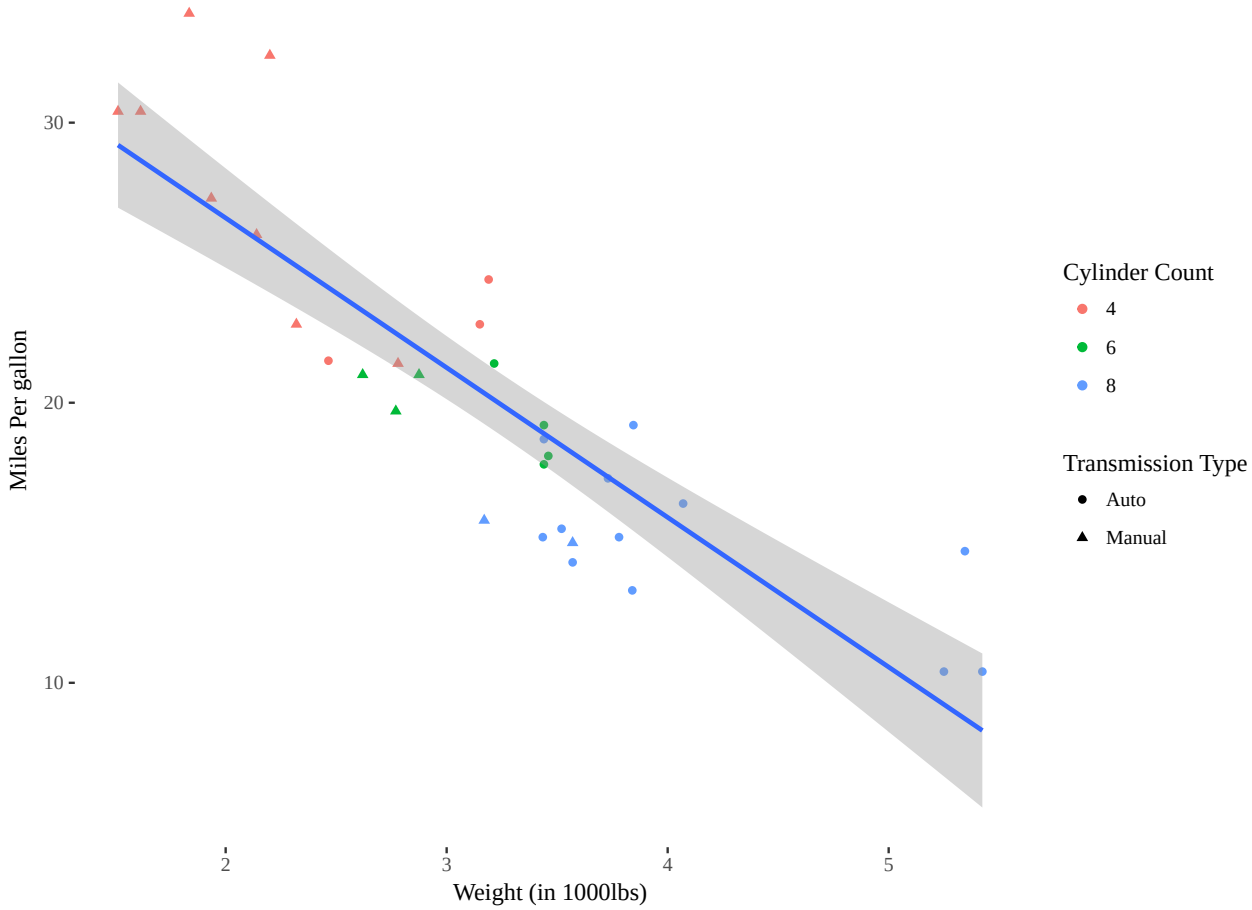
Exploratory Analysis

Let's start out by looking at some of the relationships between variables.

```
require(dplyr); require(ggplot2); require(ggthemes);require(knitr)
```

Transmission Type and Weight

```
# Points plot of relationship between mpg ~ am + wt
ggplot(mtcars) + geom_point(aes(x=wt, y=mpg, col=cyl, shape=am)) + theme_tufte() +
    geom_smooth(aes(x=wt, y=mpg), method='lm') +xlab('Weight (in 1000lbs)') +
    ylab('Miles Per gallon') + labs(col='Cylinder Count',
                                    shape = 'Transmission Type')
```



We see that manual transmission cars weigh less than cars with automatic transmissions and have higher MPG. We also see that most of the manual cars have 4 cylinders with only a few at 6 or 8 cylinders. This might introduce problems in our model later.

Data Analysis

Let's start out by making a linear model containing all variables to get a quick look at which ones have the greates effect on mpg. We remove the intercepts from these models since we have so many

```
kable(anova(lm(mpg ~ . , mtcars)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cyl	2	824.7845901	412.3922950	51.3766374	0.0000002
disp	1	57.6428040	57.6428040	7.1812531	0.0171391
hp	1	18.5022051	18.5022051	2.3050408	0.1497477
drat	1	11.9144757	11.9144757	1.4843287	0.2419149
wt	1	55.7868983	55.7868983	6.9500407	0.0186962
qsec	1	1.5245500	1.5245500	0.1899314	0.6691803
vs	1	0.3020868	0.3020868	0.0376346	0.8487815
am	1	16.5665336	16.5665336	2.0638911	0.1713540
gear	2	5.0215145	2.5107572	0.3127950	0.7360567
carb	5	13.5988573	2.7197715	0.3388344	0.8814442
Residuals	15	120.4026720	8.0268448	NA	NA

Lots of nonsignificant effects. Let's make a model including just the significant variables and transmission type.

```
kable(anova(lm(mpg ~ wt + disp + cyl + am, mtcars)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wt	1	847.7252500	847.7252500	120.5279009	0.0000000
disp	1	31.6394352	31.6394352	4.4984324	0.0436229
cyl	2	63.7334954	31.8667477	4.5307512	0.0205049
am	1	0.0796768	0.0796768	0.0113283	0.9160547
Residuals	26	182.8693302	7.0334358	NA	NA

The models fail to find a significant effect of transmission type on MPG. This result seems to agree with our exploratory analysis. Weight and number of cylinders in a car, among others, explain all the variance in MPG that we would otherwise get from the transmission type.

```
kable(anova(lm(mpg ~ wt + am, mtcars)))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
wt	1	847.7252500	847.7252500	88.3301918	0.0000000
am	1	0.0022403	0.0022403	0.0002334	0.9879146
Residuals	29	278.3196972	9.5972309	NA	NA

Again we fail to find a significant effect from transmission type beyond the effect of weight.

Results

In this analysis we aimed to discover a relationship between the transmission type of cars and their MPG. There is indeed a difference in MPG between manual and automatic cars:

```
model <- lm(mpg ~ am  -1, mtcars)
kable(summary(model)$coef)
```

	Estimate	Std. Error	t value	Pr(>|t|)
amAuto	17.14737	1.124602	15.24749	0
amManual	24.39231	1.359578	17.94109	0

Manual cars on average had an MPG of: **24.3923077**, while automatic cars had an MPG of: **17.1473684**.

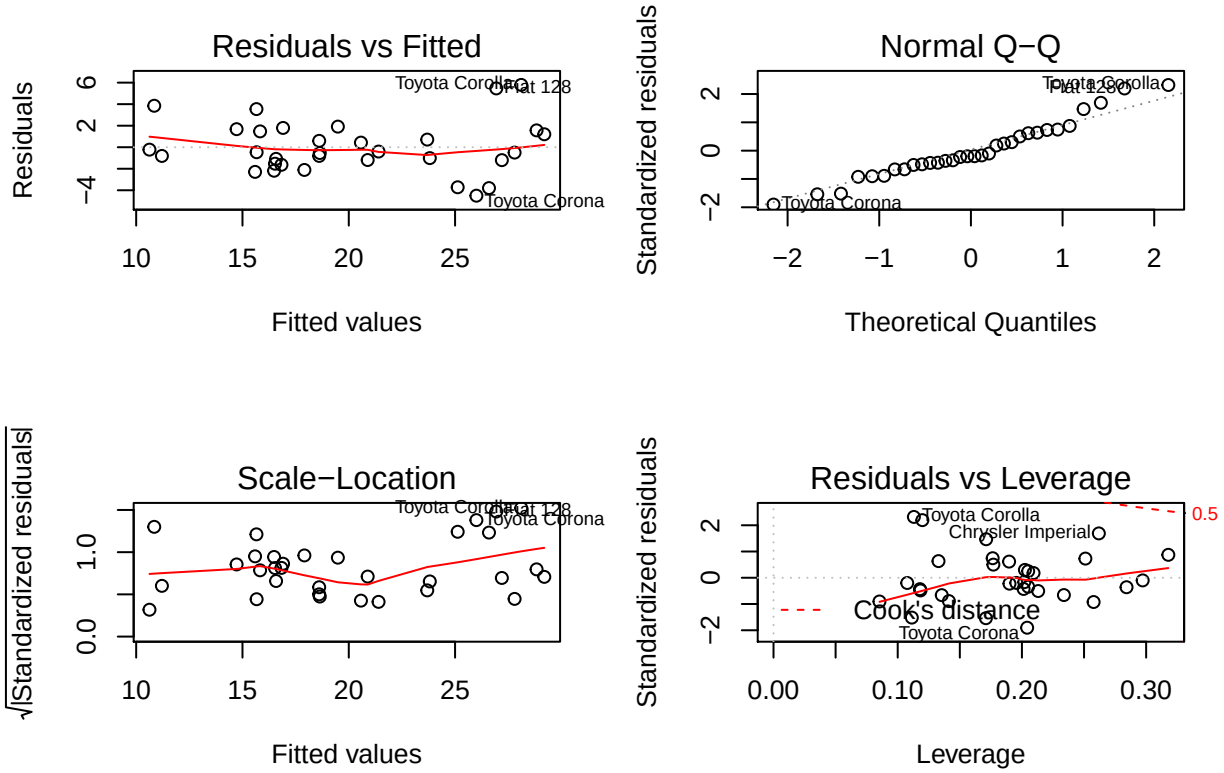
Unfortunately this effect of transmission type on MPG was not statistically significant once we corrected for other variables like weight and cylinder count. It seems that all the variance that would be explained by transmission type is instead explained solely by the weight and cylinder count of the car.

If we were to perform this analysis again, it would be a good idea to have a larger sample size with a larger spectrum of manual and automatic cars whose weights would cross in order to better compare the two groups correcting for weight. As we saw in the exploratory analysis we also need more manual cars with 8 cylinders and more automatic cars with 4 cylinders to perform a sound inference.

Appendix

Residual Plot of the last fitted model.

```
model <- lm(mpg ~ wt + disp + cyl + am, mtcars)
par(mfrow=c(2,2))
plot(model)
```



The error looks approximately normal with some skew by the outer bounds of the distribution. There is no hint of bias in the scale-location plot. There are a lot of high leverage points as the cars are not well distributed with almost all manual cars being way lighter than the automatic cars.