Regression Models Course Final Project

Mike Holmquest

4/22/2021

Executive Summary

This project will be analyzing mileage data for Motor Trend magazine about the automobile industry. We will use a data set of a collection of cars, they are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). They are particularly interested in the following two questions:

Question 1 - "Is an automatic or manual transmission better for MPG"

Question 2 - "Quantify the MPG difference between automatic and manual transmissions"

After modeling the data it is clear that there is a statistically significant relationship between miles per gallon and the transmission type. When only reviewing the mpg and transmission variables there is a +7.245 mile per gallon improvement when using a manual transmission vs. an automatic transmission. When including other regressors in the model it is clear that the actual improvement is closer to + 2.08 miles per gallon. This model is the best because it takes into account the influence of weight, cylinder size, and horsepower.

To check these statistic I used linear models to find coefficients, and multiple tests to check significance. First, variance of inflation showed a low level of colinearity with values between 2.08 and 3.77. Next, I used a correlation test that showed a .59 relationship. While this is strong, I think it is important to note that other variables also have strong relationships with mpg. The confidence intervals showed that the predicted values are with an acceptable range. The ANOVA test showed how the significance levels were related between 2 models. The low p-value showed that statistically these results would not happen just by random chance. Lastly, I used residual plots to show that there isn't any trends or issues with the data.

Load data and programs

```
library(datasets)
data(mtcars)
datacor <- mtcars
library(ggplot2)
```

Exploratory Data Analysis

The first goal is going to be to take a high level look at the data to see its organization and structure.

```
## What do the variables mean
?mtcars
## Use str to see a compact summary of the data
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 ...
```

```
sapply(mtcars, class)
```

```
##       mpg       cyl      disp        hp      drat        wt      qsec        vs
## "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric" "numeric"
##        am      gear      carb
## "numeric" "numeric" "numeric"
```

From the information on the data set we see that the mtcars dataset is a data frame with 32 observations on 11 (numeric) variables.

mpg Miles/(US) gallon cyl Number of cylinders disp Displacement (cu.in.) hp Gross horsepower drat Rear axle ratio wt Weight (1000 lbs) qsec 1/4 mile time vs Engine (0 = V-shaped, 1 = straight) am Transmission (0 = automatic, 1 = manual) gear Number of forward gears carb Number of carburetors

From this information we know that all of the variables are numeric. We can also see that vs(Engine Type) and am(Transmission) are "0" , "1' factor variables. Let's convert these to factor variables for analysis purposes.

```
mtcars$vs = as.factor(mtcars$vs)
mtcars$am = as.factor(mtcars$am)
```

The most important variables for our analysis are mpg - miles per gallon, and am - transmission. We will also use some of the other variables like cyl - cylinders, and wt - weight, and hp -horsepower to see correlations and/or causations.

Lets take a quick look at the data to see what kind of trends it might show.

See Appendix -Exploratory Graph 1 for a visual of this data. For a quick sumamry let's look at the mean mpg for different kinds of transmissions.

```
aggregate(mpg ~ am, data=mtcars, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```

The mean comparison of the data shows a 7.245 mpg difference between Manual and Automatic Transmission. However, the data frame includes many variables that could also be related. My first analysis is going to be looking at the linear relationships of these variables using a regression model.

Regression Models

When selecting the model I will be comparing how different regressors influence the data, and also try to figure out which regressors are statistically insignificant and can be excluded to make the most parsimonious model possible.

Lets first start with a linear model relating the two variables of concern.

```
fit1 <- lm(mpg ~ am , data = mtcars)
summary(fit1)$coef
```

```
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 1.133983e-15
## am1          7.244939   1.764422  4.106127 2.850207e-04
```

From this model we see that there is 7.245 mpg increase on average for manual transmissions.

Let's check that by manually finding the coefficient

```
(coef(lm(mpg ~ am, mtcars))[2])
```

```
##      am1
## 7.244939
```

Now let's include more variables to see their effects on the model. This model will put all variables we think might be related.

```
fit2 <- lm(mpg ~ hp + wt + disp + am + cyl - 1, data = mtcars)
summary(fit2)$coef
```

```
##         Estimate Std. Error   t value     Pr(>|t|)
## hp   -0.02796002 0.01392172 -2.008374 5.509659e-02
## wt   -3.30262301 1.13364263 -2.913284 7.256888e-03
## disp  0.01225708 0.01170645  1.047036 3.047194e-01
## am0  38.20279869 3.66909647 10.412045 9.084987e-11
## am1  39.75929032 2.92165122 13.608500 2.457646e-13
## cyl  -1.10637984 0.67635506 -1.635797 1.139322e-01
```

We can see from this model that manual transmission is better than automatic for MPG by about 1.55, which is much less than the first model. We can also see from the p-values that the significance of cyl and disp are low. Let's create 1 more model to exclude these independent variables to keep our errors and variance as low as possible.

Model 3 will only include the significant variables.

```
fit3  <- lm(mpg ~ am + wt + hp, mtcars)
summary(fit3)$coef
```

```
##                Estimate  Std. Error   t value     Pr(>|t|)
## (Intercept) 34.00287512 2.642659337 12.866916 2.824030e-13
## am1          2.08371013 1.376420152  1.513862 1.412682e-01
## wt          -2.87857541 0.904970538 -3.180850 3.574031e-03
## hp          -0.03747873 0.009605422 -3.901830 5.464023e-04
```

As you can see when we include hp, cyl, and wt. This regression has used the am0, or "Automatic Transmission" as its reference. I left it this way to easily show the difference in coefficents for the transmissions. Therefore, the influence of manual transmissions (over automatic) is the coefficient shown for am1, which is 2.08.

Once again, let's check this by finding the coefficents.

```
(coef(lm(mpg ~ am + wt + hp, mtcars))[2])
```

```
##     am1
## 2.08371
```

So, including all of these variables changes the coefficient of the relationship of mpg to manual transmission to 2.08, compared to 7.245 if we just consider mpg and transmission type. So the influence of having a manual transmission to an automatic transmission is + 2.08.

Next, We will run a test to see how confident we are in these results.

```
confint(fit3)
```

```
##                   2.5 %      97.5 %
## (Intercept) 28.58963286 39.41611738
## am1         -0.73575874  4.90317900
## wt          -4.73232353 -1.02482730
## hp          -0.05715454 -0.01780291
```

Our confidence intervals show that we are 95% confident that the values of these coefficients will be in these ranges. The important statistic hear is that the confidence interval shows that the am1(manual transmission) could be negative in relation to am0(automatic transmission)

Correlation test: Let's take a look at the correlation of the first two variables

```
cor(datacor$am, datacor$mpg)
```

```
## [1] 0.5998324
```

This indicates a positive .599 relationship. This is a strong relationship between these variables. The previous data showed that there seemed to be relationships with other variables as well. Let's check how weight and mpg are related.

```
cor(datacor$wt, datacor$mpg)
```

```
## [1] -0.8676594
```

As this statistic shows, there is an even stronger negative correlation of -.87 between these two variables. This shows that as weight increases, the mpg will decrease.

Variance of Inflation Next, Lets use a variance of inflation to check correlation. A VIF describes the increase in the variance of a coefficient due to the correlation of its regressor with the other regressors. VIF is the square of standard error inflation. am wt hp 2.271082 3.774838 2.088124

These numbers all show an acceptable level of colinearity with these variables and the independent variable, mpg.

Next let's use an ANOVA test to see the significance between 2 linear models.

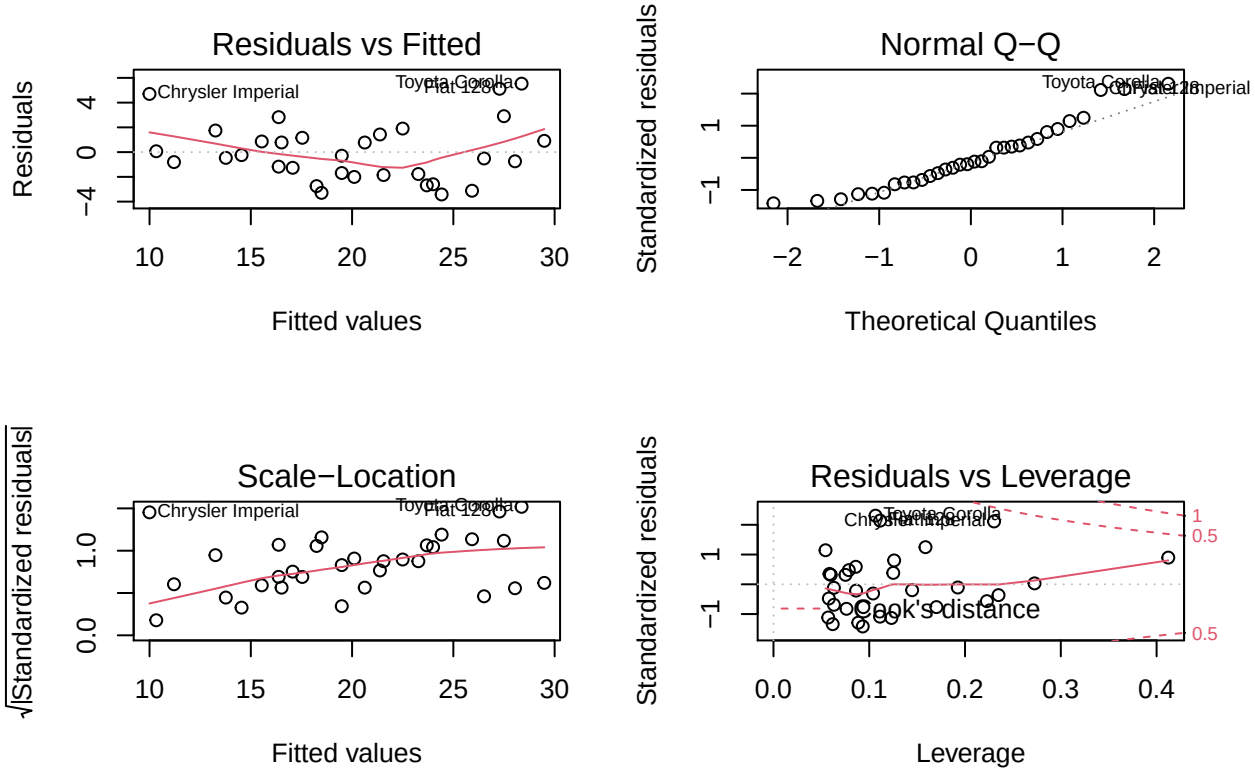
```
anova(fit1, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + hp
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     30 720.90
## 2     28 180.29  2    540.61 41.979 3.745e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The high f-statistic and low p-value show us that it highly unlikely that these results would be by pure chance the therefore would would reject the null-hypothesis and conclude that there is a relation between these variables.

Lastly I want to check how the residuals are in relation to the variables.

```
par(mfrow = c(2,2))
plot(fit3)
```

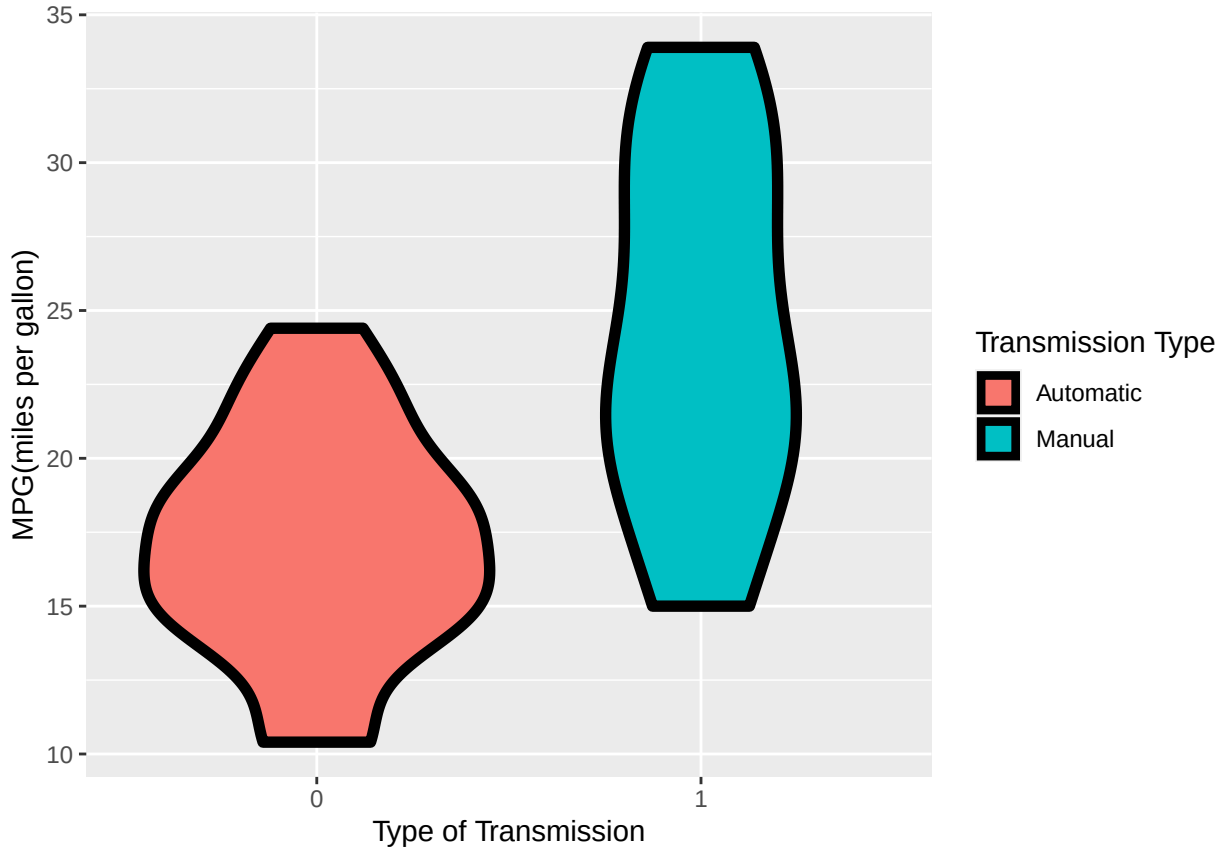


Their does not seem to be any out of place trends in the data.

Appendix

Exploratory Graph 1

```
plot1 <- ggplot(mtcars, aes(y = mpg, x = am, fill = am))
plot1 = plot1 + geom_violin(colour = "black", size = 2)
plot1 = plot1 + xlab("Type of Transmission") + ylab("MPG(miles per gallon)")
plot1 = plot1 + scale_fill_discrete(name = "Transmission Type", labels=c("Automatic", "Manual"))
plot1
```




```
## What are some useful summary statistics
aggregate(mpg ~ am, data=mtcars, mean)
```

```
##   am      mpg
## 1  0 17.14737
## 2  1 24.39231
```