Regression Models Final Project

Written by: Andrew Leonard

John Hopkins Data Science Specialization: Regression Models

Executive Summary

In analyzing data on 32 different car models, I look to answer two questions:

- Is an automatic or manual transmission better for MPG?
- Can we quantify the MPG difference between automatic and manual transmissions?

After statistical inference analysis, I have concluded that, at a 5% confidence level, manual transmission cars have higher MPG than automatic cars. The average difference in MPG for the data analyzed is 7.245 MPG, in favor of manual cars.

Through regression modeling, however, I cannot conclude that transmission type is a good predictor of MPG. While transmission type appears correlated with MPG, it does not appear as a causal variable. In the following report, I outline how the above conclusions have been reached and offer a better model, using weight and cylinders, to predict MPG. Additional figures and diagnostic testing can be found in the Appendix.

Exploratory Analysis

The mean MPG for auto versus manual transmission.

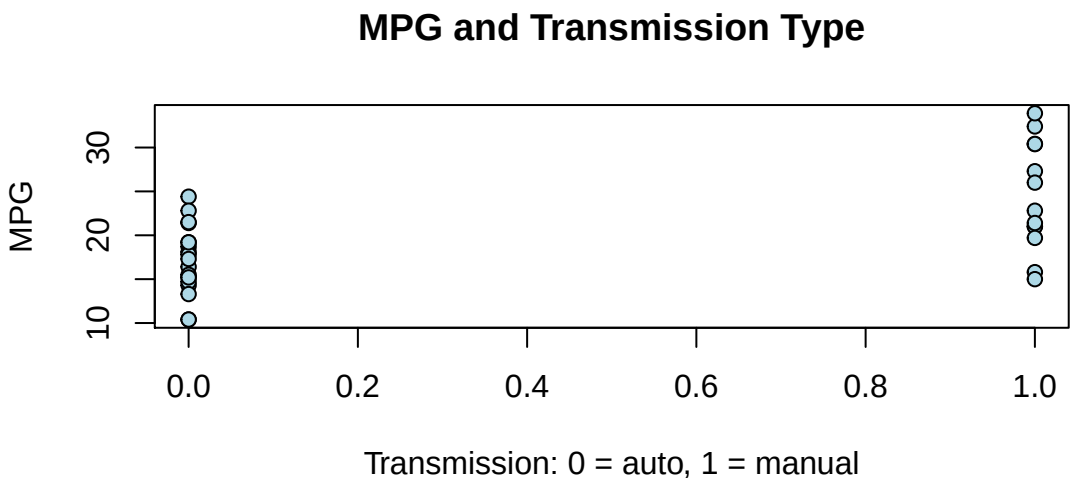
```
#mean of mpg for auto versus manual
mean(mtcars$mpg[which(mtcars$am == 1)])
```

```
## [1] 24.39231
```

```
mean(mtcars$mpg[which(mtcars$am == 0)])
```

```
## [1] 17.14737
```

```
#plot graph
with(mtcars, plot(am, mpg,pch = 21, bg = "lightblue", cex = 1, main = "MPG and Transmission Type", xlab
```




```
#test if difference in means of two types is statistically different
ttest <- t.test(mtcars$mpg[which(mtcars$am == 1)], mtcars$mpg[which(mtcars$am == 0)])
pvalue <- ttest$p.value
```

With a P-value of 0.0013736, we can infer that there is a difference in average MPG for automatic and manual transmission types.

Model Fitting

While it does appear that transmission type is correlated with MPG, it remains unclear if transmission type is a good predictor of MPG. Let's explore this by looking at multiple linear model fits.

First, let's fit a model with only transmission type as a predictor of MPG.

```
fit <- lm(mpg ~ factor(am), data = mtcars)

round(summary(fit)$coef,6)
```

```
##              Estimate Std. Error   t value Pr(>|t|)
## (Intercept) 17.147368   1.124603 15.247492 0.000000
## factor(am)1  7.244939   1.764422  4.106127 0.000285
```

From our first look, it seems transmission type may be a good predictor of MPG. However, it also appears both weight and number of cylinders have a linear relationship with MPG. See figure A in Appendix for graph. Let's fit two more models introducing weight and number of cylinders as variables.

```
fit2 <- lm(mpg ~ factor(am) + wt, data = mtcars)
fit3 <- lm(mpg ~ factor(am) + wt + factor(cyl), data = mtcars)
anova(fit, fit2, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ factor(am)
## Model 2: mpg ~ factor(am) + wt
## Model 3: mpg ~ factor(am) + wt + factor(cyl)
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     29 278.32  1    442.58 65.3095 1.107e-08 ***
## 3     27 182.97  2     95.35  7.0353  0.003473 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Through our anova function, we see that adding both weight and cylinders has a statistically significant effect on MPG at at 5% confidence interval.

Next, let's look at a model without transmission type as a variable, and then add it back in as a final variable to see if it has a significant effect on MPG.

```
fit4 <- lm(mpg~wt, data = mtcars)
fit5 <- lm(mpg ~ wt + factor(cyl), data = mtcars)
anova(fit4,fit5, fit3)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ wt
## Model 2: mpg ~ wt + factor(cyl)
## Model 3: mpg ~ factor(am) + wt + factor(cyl)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     30 278.32
## 2     28 183.06  2    95.263 7.0288 0.003488 **
## 3     27 182.97  1     0.090 0.0133 0.908947
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Our anova test shows that adding transmission type as a variable does not have a statistically significant effect on MPG, with a P-value over .9.

I added each additional variable in the dataset with wt and cyl to see if any other variable had a statistically significant effect on MPG outside of wt and cyl. All P-values of added variables were well above .05. Analysis of each not included in report. Here is an example trial with disp as a factor.

```
fit6 <- lm(mpg~wt+factor(cyl)+disp, data = mtcars)
round(summary(fit6)$coef,3)
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    34.042      1.963  17.341    0.000
## wt             -3.307      1.105  -2.992    0.006
## factor(cyl)6   -4.306      1.465  -2.939    0.007
## factor(cyl)8   -6.323      2.598  -2.433    0.022
## disp            0.002      0.013   0.127    0.900
```

Conclusion

While it does appear that manual cars tend to have higher MPG than automatic cars, we have found better predictors of MPG. Our model with wt and cyl as our predictors has the most statistical significance on MPG.

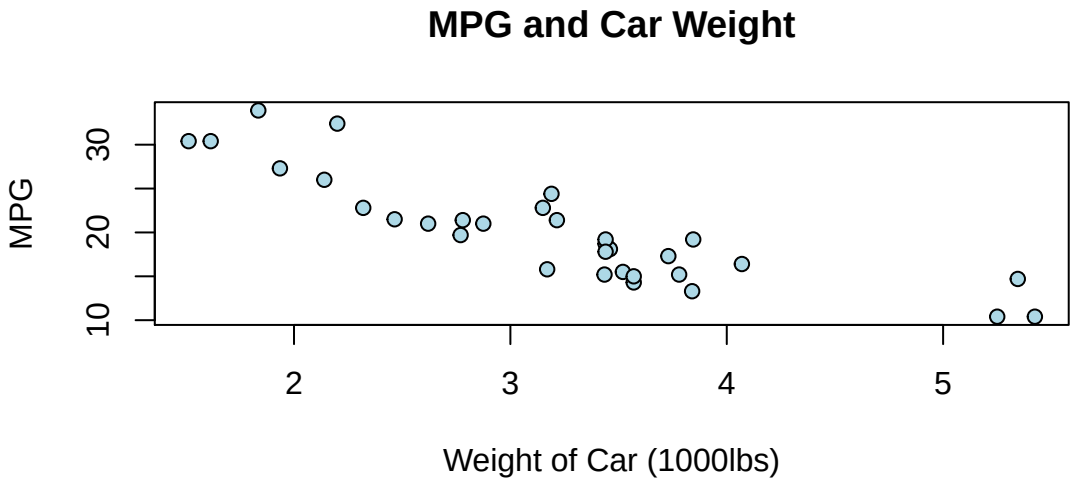
The appendix will look at the residuals of analyzed models to further prove this analysis. We will also interpret the coefficients of our selected model.

Appendix

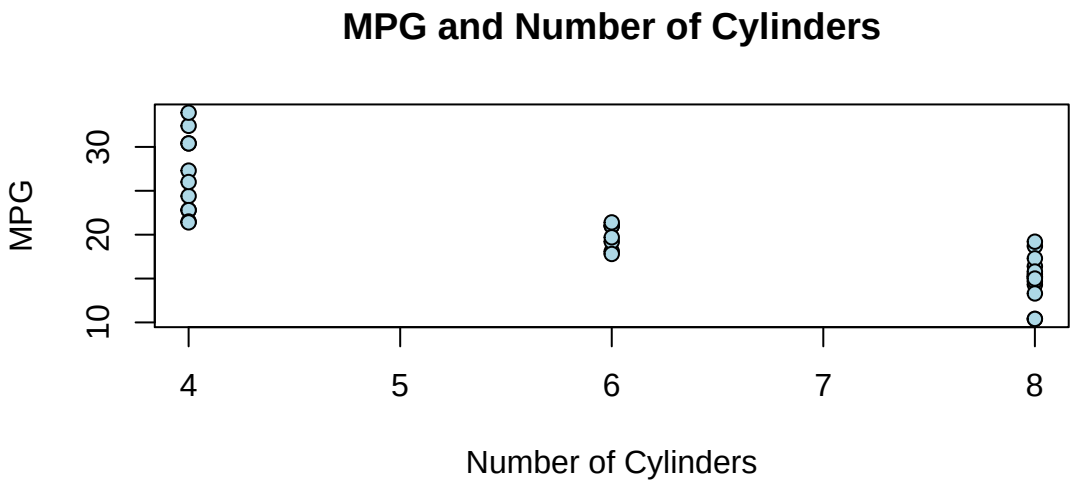
Figure A

Plots of weight and cylinders with MPG. Appears to be a linear relationship for both variables.

```
with(mtcars, plot(wt, mpg, pch = 21, bg = "lightblue", cex = 1, main = "MPG and Car Weight", xlab = "We
```




```
with(mtcars, plot(cyl, mpg, pch = 21, bg = "lightblue", cex = 1, main = "MPG and Number of Cylinders", x
```



Diagnostic Testing

Residuals

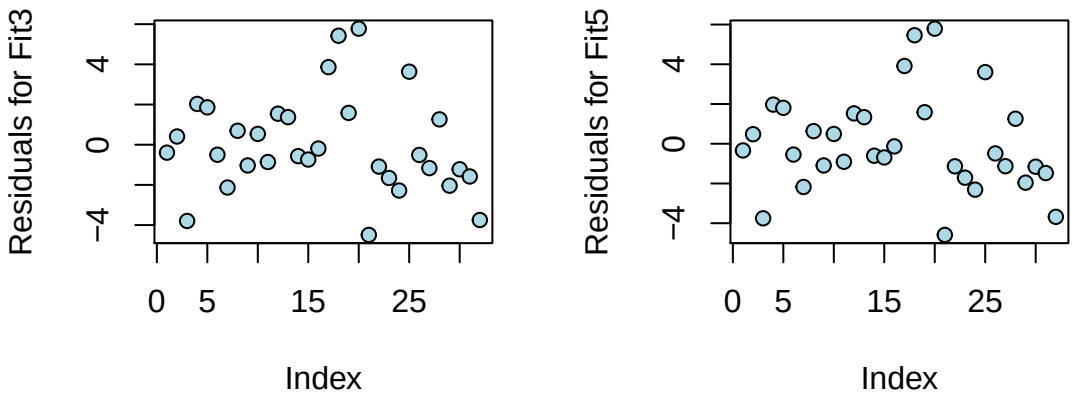
```
#sum of squared residuals after fitting first model (fit)
resfit <- round(sum(resid(fit)^2),3)

#after adding weight and cyl (fit3)
resfit3 <- round(sum(resid(fit3)^2),3)

#including only weight and cyl (fit5)
resfit5 <- round(sum(resid(fit5)^2),3)
```

By looking at our sums of the squared residuals for model fit, fit3 and fit5, it is clear including wt and cyl as predictors and excluding transmission as a predictor is a good choice.

Our sum of squared residuals for model one is 720.897. After adding wt and cyl, our sum of squared residuals for model three is 182.968. When we only include wt and cyl, and exclude transmission type, or sum of squared residuals barely changes at 183.059

Finally, we can graph the residuals to see that there is almost no change in a model that includes wt, cyl, and transmission (fit3) with a model that only includes wt and cyl (fit5).

```
par(mfrow=c(1,2))
plot(resid(fit3), pch = 21, bg = "lightblue", cex = 1, ylab = "Residuals for Fit3")
plot(resid(fit5), pch = 21, bg = "lightblue", cex = 1, ylab = "Residuals for Fit5")
```



Interpretation fo Coefficients

```
fit5[1]$coefficients
```

```
##  (Intercept)           wt factor(cyl)6 factor(cyl)8
##    33.990794    -3.205613    -4.255582    -6.070860
```

```
car <- round(predict(fit5, newdata = data.frame(wt = 1, cyl = 8)),3)
```

The intercept assumes a 4 cylinder car. For each 1000 lbs increase in weight, we will see a 3.2MPG reduction. If the car is 6 cylinders, we will see a 4.3MPG reduction. If the car is 8 cylinders, we will se a 6.1MPG reduction.

For example, a car weighing 1000 pounds with 8 cylinders will have a predicted MPG of 24.714.

Additional Diagnostics

While there are a few outliners, model reasonably resembles a normal distribution.

```
#qqplot
plot(fit5, which=2)
```