Regression Project

Alim Ray

10/25/2015

Executive Summary

Using the mtcars data included in RStudio, I have been asked to answer the folowing questions. Is an automatic or manual transmission better for MPG? Quantify the MPG difference between automatic and manual transmissions?

Setting up data

Reading mtcars data and adding cylinder and transmission factors. Making transmission a factor variable will allow me to determine the marginal differences between manual and automatic transmissions when creating a linear regression.

```
##
## Attaching package: 'dplyr'
##
## The following object is masked from 'package:stats':
##
##     filter
##
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
##
##
## Attaching package: 'GGally'
##
## The following object is masked from 'package:dplyr':
##
##     nasa
```

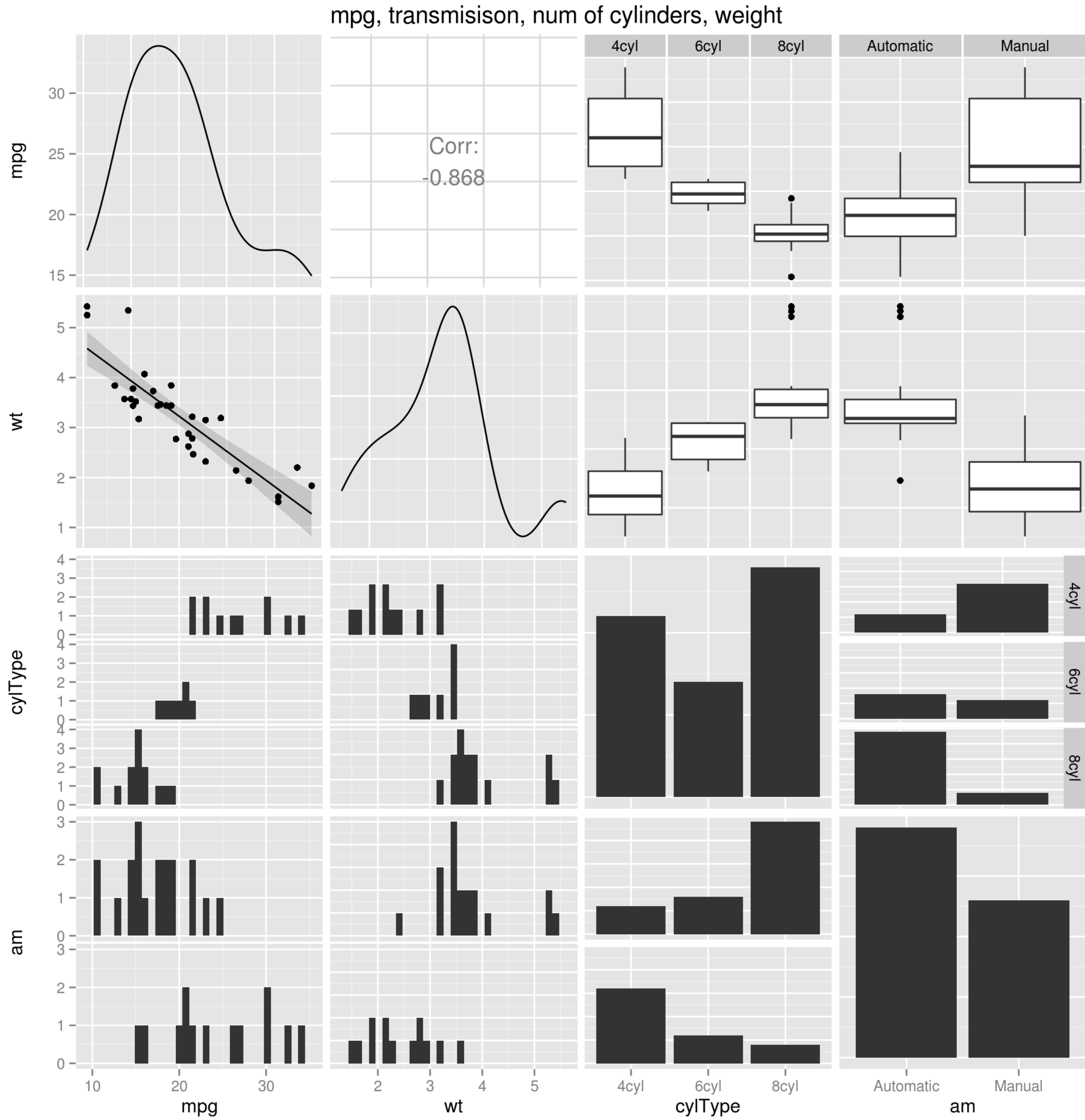
Expolaratory Analysis

Looking at the linear regression between mpg and transmissions only yields the following model.

Intercept = 17.1473684

Slope = 7.2449393

This shows a positive correlation before adjustments. I then looked at number of cylinders and weight because I believed they would effect the model. Below is a pairs plot of mpg, transmission, number of cylinders, and weight.



Cylinders and Weight

Cell 1,3 of this graph is a box and whisker plot which shows that as number of cylinders increase, mpg decreases. Cell 2,1 also shows that as weight increases, mpg decreases. I checked the correlation between weight and cylinders and got 0.7824958. Since these data are highly correlated, I decided to remove weight from the model.

Displacement and Horsepower

Engine displacement is the volume of an engine's cylinders. This is directly correlated to the number of cylinders (correlation = 0.9020329) and was removed from the model. Horsepower is also a function of cylinders (correlation = 0.8324475) and was excluded from the model.

Elimination of variables via ANOVA

Using the rest of the variables (and transmission + number of cylinders), I created five linear models that I tested using ANOVA.

```
everything <- lm( mpg ~ am + drat + cylType, cars)
everything2 <- lm( mpg ~ am + drat + cylType + qsec, cars)
everything3 <- lm( mpg ~ am + drat + cylType + qsec + vs, cars)
everything4 <- lm( mpg ~ am + drat + cylType + qsec + vs + gear, cars)
everything5 <- lm( mpg ~ am + drat + cylType + qsec + vs + gear + carb, cars)
anova(everything, everything2, everything3, everything4, everything5)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am + drat + cylType
## Model 2: mpg ~ am + drat + cylType + qsec
## Model 3: mpg ~ am + drat + cylType + qsec + vs
## Model 4: mpg ~ am + drat + cylType + qsec + vs + gear
## Model 5: mpg ~ am + drat + cylType + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 264.32
## 2     26 259.83  1     4.494 0.5205 0.4779
## 3     25 256.61  1     3.217 0.3726 0.5476
## 4     24 250.87  1     5.740 0.6647 0.4233
## 5     23 198.59  1    52.279 6.0547 0.0218 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The qsec, v/s, and number of gears are shown to have high P-values (all > 0.4). Removing these fields, the final model is transmission, rear axle ratio, number of cylinders, and number of carburetors.

Conclusion

```
finalM <- lm( mpg ~ am + drat + cylType + carb, cars)
confInt <- summary(finalM)$coefficients[2,1] + c(-1,1) * summary(finalM)$coeffi
cients[2,2] * qt(0.975, df = finalM$df)
```

Using the final model the coefficient for manual transmission is 3.5736741. Since this is positive, the effect of manual transmission on miles per gallon is positive. with a confidence interval of 0.4904621 to 6.656886 is also the average mpg increase when using a manual transmission.