Regression Analysis Assignment

Minki Jo

December 3rd, 2017

Introduction

This report explores the relationship between a set of variables and outcome, miles per gallon (MPG). This answers to the following two questions :

- "Is an automatic or manual transmission better for MPG"
- "Quantify the MPG difference between automatic and manual transmissions"

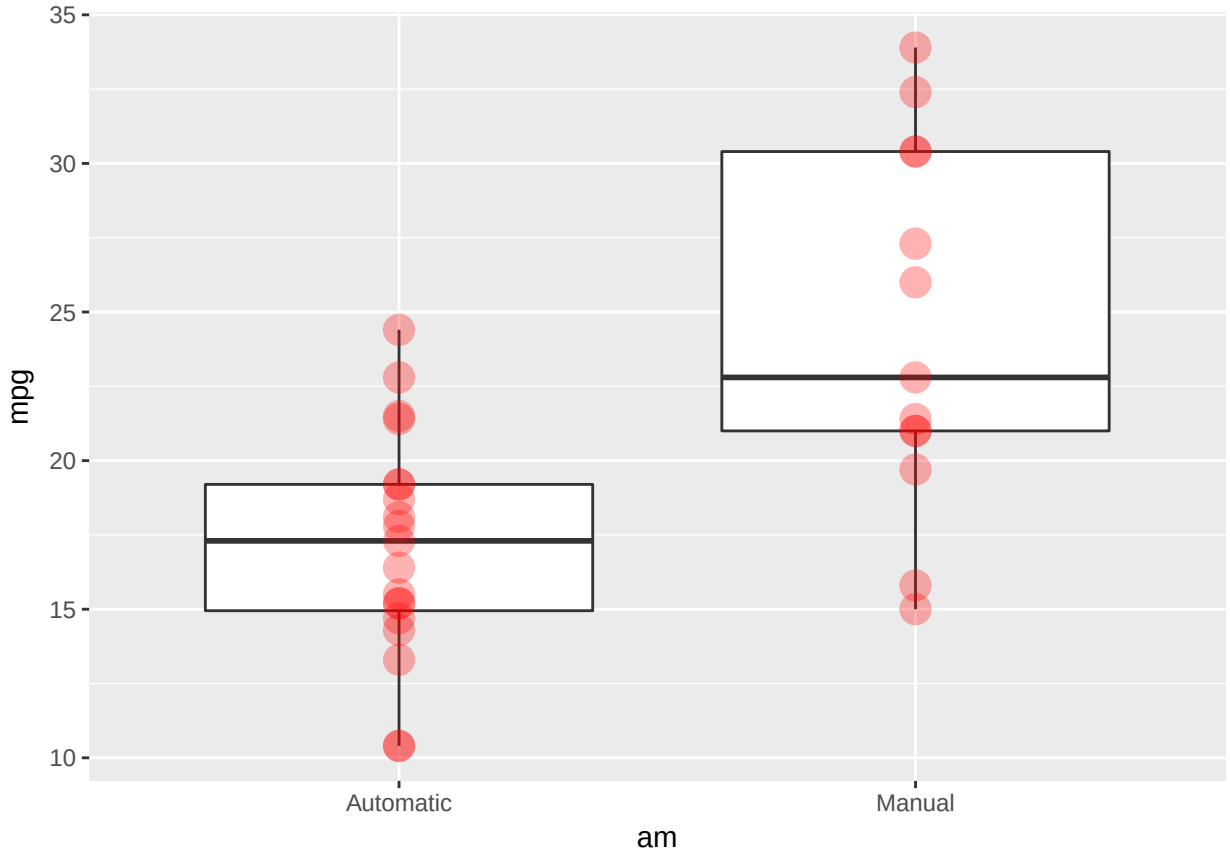
Executive summary

This reports shows that manual transmission increases efficiency in mpg compared with cars with automatic transmission. From the multiple regression analysis, the model that has the least AIC for explanation of mpg has three variables such as wt, qsec, and transmission type. However qsec variable remains controversial. This might come from size of the data set. Therefore, further investigation should be executed for better accuracy and interpretation.

Explorating data analysis

Data consists of 32 observations with 12 variables. Let's take a look at how outcome (MPG) data looks like in terms of transmission type.

```
g <- ggplot(data = mtcars, aes(x = am, y = mpg))
g <-  g + geom_boxplot() + geom_point(col = "red", alpha = 0.3, size =5)
g
```



It looks a lot like like cars with manual trasmission has better mpg than those with automatic transmission. But since some observations in manual transmission have worse mpg than those of automatic transmission, one might say it is not obvious to say manual cars has better in mpg than automatic transmission cars. In order to verify that, independent 2 sample t-test is required. As we know, simple linear regression analysis provides p-value and it results from 2 sample t-test when one variable has only two factors. Here is the result with the assumption that each observation has equal variable.

```
fit <- lm(mpg ~ am, data = mtcars)
summary(fit)$coefficients
```

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

Based on p-value for factor variable, **it is proved that cars with manual transmission has better in mpg than automatic transmission cars.** Now, let's quantifiy how different it is between transmission types. The coefficient of amManual shows that cars with manual transmission has **7.245** higher mpg in average than those with automatic transmission. It is also said that mpg in average for automatic cars **17.147** because coefficient of intercepts indicates automatic factor.

Variable selection for multiple regresion model

There are several methods to select variable for multi-variable linear regression model. This reports uses AIC which trades off between squared error and number of variable. That is, when AIC is the least, number of variables included and error are overally minimum. In search for minimum AIC, step function is used which applies backward step.

```r
full <- lm(mpg ~., data = mtcars)
final <- step(full)
```

```r
summary(final)
```

```
##
## Call:
## lm(formula = mpg ~ wt + qsec + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4811 -1.5555 -0.7257  1.4110  4.6610
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   9.6178     6.9596   1.382 0.177915
## wt           -3.9165     0.7112  -5.507 6.95e-06 ***
## qsec          1.2259     0.2887   4.247 0.000216 ***
## amManual      2.9358     1.4109   2.081 0.046716 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.459 on 28 degrees of freedom
## Multiple R-squared:  0.8497, Adjusted R-squared:  0.8336
## F-statistic: 52.75 on 3 and 28 DF,  p-value: 1.21e-11
```

Results indicates that model with the least AIC have three variables such as vehicle weight, time to reach 1/4 miles, and transmission type. Since R-squared is high enough like 84.97%, mpg can be explained by linear model.

If selected variables are highly correlated each other, calculated coefficients can be distorted and its variabion can be inflated Here are variation inflation factors for the selected variables

```r
vif(final)
```

```
##       wt     qsec       am
## 2.482952 1.364339 2.541437
```

vif with 5 is usually used for baseline to check whether or not the variable makes variation inflated. Frome the results above, **selected variables are not highly correlated each other.**

Conclusion

Looking into residual and diagnostic plot (Appendix 1), 4 assumptions for regression model such as linearity, normality, heteroscedasticity, leverage with discrepancy are met. Therefore, our final model is justifiable.

Final model shows that cars' mpg with manual transmission is increased by 2.93 in average compared with cars with automatic transmission when other variables (qsec and wt) are fixed. Also, when a car weights 1,000 lbs more, mpg is reduced by 3.92. This is also the case that other variables (transmission type and qsec) are fixed. These interpretation can be accepted because of the common sense

However, the interpretation that reducing car acceleration(qsec) increases mpg remain controverial. This requires either to investigate further for the qsec variation(e.g. get more data) or to omit the variable.

Appendix1: Residual plot and diagnostics for multiple linear regression model

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