Coursera: Linear Regression Course Project

YZ

October 25, 2017

Assignment Overview

Which car characteristics have a significant impact on miles per gallon (MPG)? In particular,

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

Data source: mtcars

Executive Summary

Based on our analysis, there are three car characteristics that determine MPG:

- Type of transmission: automotic is more taxing on MPG
- Weight of a car: heavier cars consume more gas per mile
- Acceleration time (as measured by time per 1/4 mile): faster cars have lower MPG
- Cars with maual transmission have around 3 MPG more compared to automatic transmission, keeping everything else fixed
- However, if we control for car's weight, the difference in MPG between manual and automatic varies significantly: the lighter the car, the bigger the difference in MPG between manual and automatic transmissions

Data Overview

Source dataset *mtcars* has 11 variables:

```
dim(mtcars)
```

```
## [1] 32 11
```

Our main variable of interest (dependent) is *mpg*.

Our main variable of interest (independent) is *am* (Transmission (0 = automatic, 1 = manual)).

```
table(mtcars$am)
```

```
##
##  0  1
## 19 13
```

We can quickly check if there is emperical difference between two types of transmission:

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

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

Indeed, manual cars on average have 7 MPG more. However, this doesn't account for any other factors. Is it possible that other car characteristics are more impactful on MPG? Let's take a look at MPG correlation to other available factors:

```
cor(mtcars)[1,]
```

```
##        mpg        cyl       disp         hp       drat         wt
##  1.0000000 -0.8521620 -0.8475514 -0.7761684  0.6811719 -0.8676594
##       qsec         vs         am       gear       carb
##  0.4186840  0.6640389  0.5998324  0.4802848 -0.5509251
```

Based on the above there are a lot of other factors that are highly correlated with MPG

What car characteristics are the strongest determinants of MPG?

After multiple itirations (see Appendix), we selected the model that contains three factors: Transmission, Weight, Time for 1/4 mile.

```
coefficients(lm(formula = mpg ~ am + wt + qsec, data = mtcars))
```

```
## (Intercept)          am          wt        qsec
##    9.617781    2.935837   -3.916504    1.225886
```

Based on the above, cars with manual Transmission have almost 3 MPG more than automatic.

However, the interaction between Transmission type and car Weight was also included due to its significance (full output in Appendix):

```
mpg_lm_final <- lm(formula = mpg ~ am + wt + qsec + am*wt, data = mtcars)
coefficients(mpg_lm_final)
```

```
## (Intercept)          am          wt        qsec       am:wt
##    9.723053   14.079428   -2.936531    1.016974   -4.141376
```

Based on this model, the difference in MPG between manual and automatic transmission is not constant. Instead, it is determined by the formula: 14.079 - 4.141 * Weight. This means that lighter cars have much higher difference at MPG compared to heavier cars.

All coefficients, except Intercept, are significantly significant at 95% level:

```
confint(mpg_lm_final)
```

```
##                 2.5 %    97.5 %
## (Intercept) -2.3807791 21.826884
## am           7.0308746 21.127981
## wt          -4.3031019 -1.569960
## qsec         0.4998811  1.534066
## am:wt       -6.5970316 -1.685721
```

We observe high multicollinearity (VIF > 2) among some predictors. But it is due to the fact that we have an interaction term.

```
library(car)
vif(mpg_lm_final)
```

```
##        am        wt      qsec     am:wt
## 20.970925  3.030963  1.447406 16.302453
```

The residual diagnostics (see Appendix) didn't identify any issues:

- Residuals don't have any distinct patterns (chart 1)
- For the most part they are normally distributed (with some deviations at higher quantiles) (char 2)
- Assumption of homoscedasticity does not seem violated (chart 3)
- There are no outliers that skew the model (chart 4)

Therefore, the final selected model provides reliable insights.

Appendix

Model Selection: stepwise itirations

We used the following method to select a model:

```
mpg_lm_base <- lm(data = mtcars,
                  mpg ~ am)
mpg_lm_full <- lm(data = mtcars,
                  mpg ~ .)
mpg_lm_step <- step(mpg_lm_base, scope=list(lower=mpg_lm_base, upper=mpg_lm_full), direction="both")
```

```
mpg_lm_step
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec, data = mtcars)
##
## Coefficients:
## (Intercept)           am           wt         qsec
##       9.618        2.936       -3.917        1.226
```

Final model:

```
summary(mpg_lm_final)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + qsec + am * wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5076 -1.3801 -0.5588  1.0630  4.3684
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    9.723      5.899   1.648 0.110893
## am            14.079      3.435   4.099 0.000341 ***
## wt            -2.937      0.666  -4.409 0.000149 ***
## qsec           1.017      0.252   4.035 0.000403 ***
## am:wt         -4.141      1.197  -3.460 0.001809 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.084 on 27 degrees of freedom
## Multiple R-squared:  0.8959, Adjusted R-squared:  0.8804
## F-statistic: 58.06 on 4 and 27 DF,  p-value: 7.168e-13
```

Residual Diagnostics

```
par(mfrow = c(2,2), oma = c(0, 0, 0, 0))
plot(mpg_lm_final)
```