Regression Models - Course Project

Pavit Masson

February 10, 2019

Executive Summary

We will be performing an analysis for Motor Trend, a magazine about the automobile industry. Looking at 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:

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

Based on three different regression models, it is clear that manual transmission is better for MPG than automatic transmission. Our most reliable regression model (MPG dependent on transmission, weight, and gross horsepower) shows with reasonable confidence that manual transmission is better for MPG by about 2.084 (on average).

Analysis

First, we'll load the mtcars dataset and take a high-level look at it.

```
library(datasets)
data(mtcars)
data <- mtcars
str(data)
```

```
## '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 ...
```

We can see that the data contains 32 observations on 11 variables. The variables translate to the following:

- 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

We can factorize the variables vs (Engine) and am (Transmission) since they have 2 possible values.

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

Next, let's look at a summary of MPG for automatic (am = 0) vs. manual transmission (am = 1).

```
# Automatic
summary(subset(data, am == 0)$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   14.95   17.30   17.15   19.20   24.40
```

```
# Manual
summary(subset(data, am == 1)$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15.00   21.00   22.80   24.39   30.40   33.90
```

At a glance we can see that the mean MPG is higher for manual transmission. This can also be seen graphically in the boxplot **A-1** in the Appendix.

Next we'll look at correlations across all variables to see if other variables are correlated with MPG.

```
t(cor(mtcars[-1], mtcars$mpg))
```

```
##            cyl       disp         hp      drat         wt     qsec
## [1,] -0.852162 -0.8475514 -0.7761684 0.6811719 -0.8676594 0.418684
##             vs        am      gear       carb
## [1,] 0.6640389 0.5998324 0.4802848 -0.5509251
```

For our regression models, we'll examine MPG dependent on AM alone and MPG dependent on the next few highest correlated variables, which are cyl (number of cylinders), wt (weight), disp (displacement), and hp (gross horsepower).

Regression Models

We will do two regression models, as follows:

Reg 1 - MPG ~ AM (MPG dependent on AM) Reg 2 - MPG ~ AM + WT + CYL + DISP + HP (MPG dependent on WT, CYL, DISP, and HP)

```
reg1 <- lm(mpg ~ am, mtcars)
reg2 <- lm(mpg ~ am + wt + cyl + disp + hp, mtcars)
```

We can see from Appendix chart **A-2** the summary of reg 1. This shows that manual transmission is better than automatic for MPG by about 7.245, which is what we expected. When we include wt, cyl, disp, and hp (Appendix chart **A-3**) manual is still better, but only by 1.55.

By excluding the less significant variables from the multivariate regression, we might be able to better quantify how much better manual transmission is than automatic for MPG. We can see from **A-3** that cyl and disp have low significance, so we will create a new model without those two variables.

```
reg3  <- lm(mpg ~ am + wt + hp, mtcars)
```

Based on the reg3 summary in **A-4** of the Appendix, we see that this model has more reliable significance codes. This makes reg3 the most accurate model compared to reg1 and reg2. The model summary shows that manual transmission is better than automatic by an average of about 2.084.

Conclusions

To conclude, let's revisit the original two questions:

1. Is an automatic or manual transmission better for MPG?

As we can see above from the MPG mean and all three regression models that manual transmission is clearly better for MPG than automatic transmission, by varying degrees.

2. Quantify the difference between automatic and manual transmissions.

Our regression models have quantified how much better manual transmission is than automatic transmission by three different amounts:

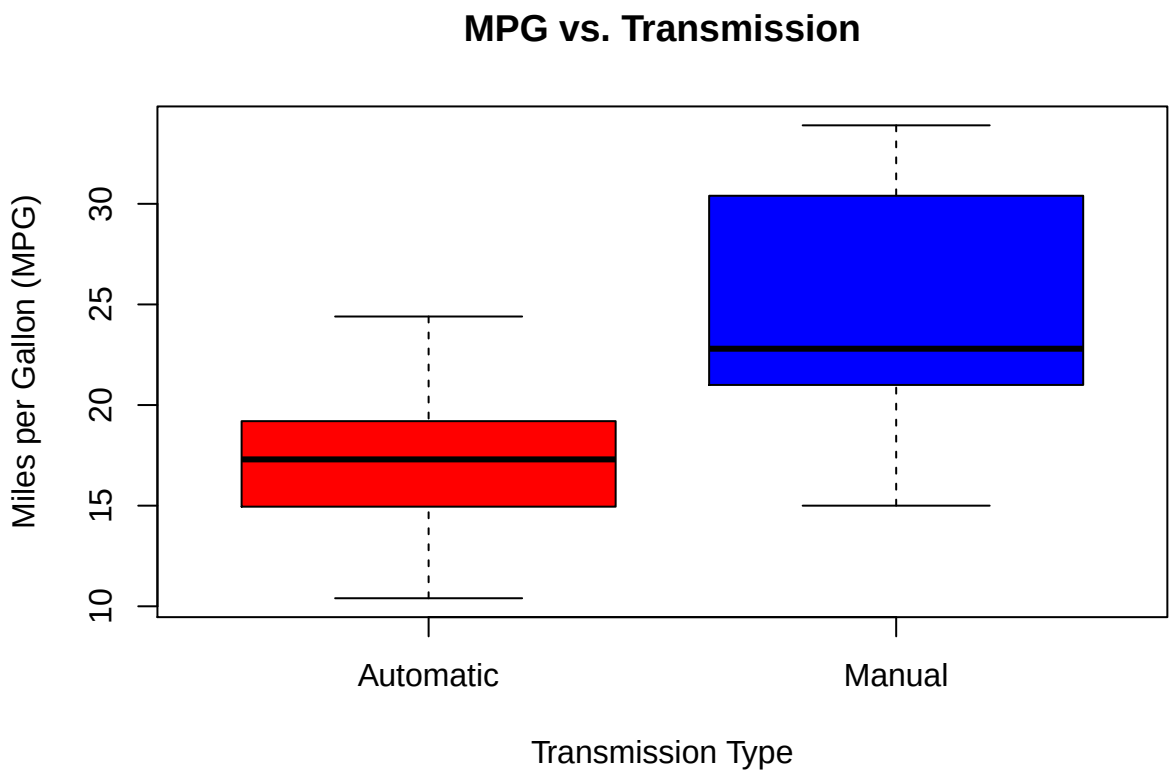
- 7.245 (reg 1)
- 1.55 (reg 2)
- 2.084 (reg 3)

Since reg 3 is the most reliable from the three models, we can say with confidence that manual transmission is better than automatic transmission by approximately 2.084 on average.

Appendix

A-1: MPG Boxplot per Transmission

```
boxplot(mpg ~ am, data = data, col = c("red", "blue"), names = c("Automatic", "Manual"), xlab = "Transmi
```



A-2 Summary of Regression Model 1 (Single Regression, am)

```
summary(reg1)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3923 -3.0923 -0.2974  3.2439  9.5077
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   17.147      1.125  15.247 1.13e-15 ***
## am             7.245      1.764   4.106 0.000285 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.902 on 30 degrees of freedom
## Multiple R-squared:  0.3598, Adjusted R-squared:  0.3385
## F-statistic: 16.86 on 1 and 30 DF,  p-value: 0.000285
```

A-3 Summary of Regression Model 2 (Multivar Regression, am + wt + hp + cyl + disp)

```
summary(reg2)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + cyl + disp + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5952 -1.5864 -0.7157  1.2821  5.5725
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.20280    3.66910  10.412 9.08e-11 ***
## am           1.55649    1.44054   1.080  0.28984
## wt          -3.30262    1.13364  -2.913  0.00726 **
## cyl         -1.10638    0.67636  -1.636  0.11393
## disp         0.01226    0.01171   1.047  0.30472
## hp          -0.02796    0.01392  -2.008  0.05510 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.505 on 26 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8273
## F-statistic:  30.7 on 5 and 26 DF,  p-value: 4.029e-10
```

A-4 Summary of Regression Model 3 (Multivar Regression, am + wt + hp)

```
summary(reg3)
```

```
##
## Call:
## lm(formula = mpg ~ am + wt + hp, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4221 -1.7924 -0.3788  1.2249  5.5317
##
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 34.002875   2.642659  12.867 2.82e-13 ***
## am           2.083710   1.376420   1.514 0.141268
## wt          -2.878575   0.904971  -3.181 0.003574 **
## hp          -0.037479   0.009605  -3.902 0.000546 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.538 on 28 degrees of freedom
## Multiple R-squared:  0.8399, Adjusted R-squared:  0.8227
## F-statistic: 48.96 on 3 and 28 DF,  p-value: 2.908e-11
```