Efficiency comparisons between car transmission types

Kevin Tham

May 12, 2018

Overview

In this report we will explore the relationship between a set of variables and miles per gallon (MPG) from a data set of a collection of cars. Specifically, we would like to answer the following two questions:

1. How different is the MPG between automatic and manual transmissions?
2. Is an automatic of manual transmission better for MPG?

Using the dataset `mtcars` we shall embark on a statistical study to address the above two questions.

Exploratary Data Analysis

We begin the study by conducting some exploratory data analysis. First we load in required libraries:

```
if (!require("pacman"))
  install.packages("pacman", repos = "http://cran.us.r-project.org")
pacman::p_load(knitr, dplyr, ggplot2, GGally, tidyr, grid, gridExtra, car, broom, tibble)
```

Next we import and examine the dataset:

```
data(mtcars)
head(mtcars)
```

```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
str(mtcars)
```

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

Some of the variables are in the wrong data type and require coercion to the correct data type:

```
mtcars$am <- factor(mtcars$am, labels = c('automatic',' manual'))
mtcars$vs <- factor(mtcars$vs, labels = c('V-shaped', 'straight'))
```

```r
mtcars$cyl <- ordered(mtcars$cyl)
mtcars$gear <- ordered(mtcars$gear)
```

We can make a direct comparison between the transmission type and MPG with a boxplot:

From the boxplot in Figure 1 we can conclude that from the dataset, cars with a manual transmission have a larger median MPG than cars with an automatic transmission. The MPG for cars with a manual transmission also appear to have a larger spread between the first and third quartiles.

In order to visualise the relationship of MPG and transmission type with the other variables we can utilise a pairplot, shown in Figure **??**. From the pairplot we can observe that many of the variables are fairly correlated with each other.

In particular, we can see how the nominal variables clearly separate some of the numerical variables. For example, the variable `cyl`, the number of cylinders in the car engine, splits the variables `disp`, `hp` and `drat` into distinct groups. The transmission type `am` also splits `disp`, `hp` and `drat` into two groups. Now if `am` is correlated with some of the other variables, which one actually is the variable responsible for the effect on MPG? Or are they all equally responsible?

This suggests that it will be difficult to intepret linear regression results to answer question 2 due to confounding variables.

Statistical Analysis

Linear Model with a single variable

If we are only concerned with the bulk effect of transmission type on MPG disregarding other values, we can simply regress `mpg` on `am` and examine the regression coefficients.

```r
fit1 <- lm(mpg ~ am, data=mtcars)
tidy(fit1)
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept) 17.147368  1.124603 15.247492 1.133983e-15
## 2   am manual  7.244939  1.764422  4.106127 2.850207e-04
```

```r
glance(fit1)
```

```
##   r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.3597989     0.3384589 4.902029  16.86028 0.0002850207  2 -95.24219
##        AIC      BIC deviance df.residual
## 1 196.4844 200.8816 720.8966          30
```

The coefficient of the linear model with only 1 variable is 7.25, with a p-value of 0.0002, indicating significance and that we should reject the null hypothesis. Therefore is a difference of 7.25 MPG between automatic and manual transmission types, neglecting adjustment for other variables. We note here that the R-squared value for this model is fairly low. This is expected as other variables that can explain the variance in MPG have not been included.

Linear Model with multiple variables

```r
fit2 <- lm(mpg ~ ., data=mtcars)
tidy(fit2)
```

```
##           term    estimate   std.error  statistic    p.value
## 1  (Intercept) 15.73289830 16.55441672  0.9503747 0.35385548
```

```
## 2        cyl.L  2.16015247  3.41523225  0.6325053 0.53459525
## 3        cyl.Q  2.22646814  1.43686806  1.5495286 0.13775130
## 4         disp  0.01256810  0.01774024  0.7084518 0.48726645
## 5           hp -0.05711722  0.03174603 -1.7991927 0.08789210
## 6         drat  0.73576811  1.98461241  0.3707364 0.71493502
## 7           wt -3.54511861  1.90895437 -1.8570997 0.07886857
## 8         qsec  0.76801287  0.75221895  1.0209964 0.32008122
## 9   vsstraight  2.48849171  2.54014636  0.9796647 0.33956206
## 10   am manual  3.34735713  2.28948094  1.4620594 0.16006890
## 11      gear.L  0.75274795  2.14062152  0.3516492 0.72897110
## 12      gear.Q  1.25045717  1.80854870  0.6914147 0.49766706
## 13        carb  0.78702815  1.03599487  0.7596834 0.45676696
```

```r
glance(fit2)
```

```
##   r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.8845064      0.811563 2.616258  12.12594 1.764049e-06 13 -67.84112
##        AIC      BIC deviance df.residual
## 1 163.6822 184.2025 130.0513          19
```

```r
vif(fit2)
```

```
##           GVIF Df GVIF^(1/(2*Df))
## cyl  44.446614  2        2.582020
## disp 21.894422  1        4.679148
## hp   21.456428  1        4.632108
## drat  5.099622  1        2.258234
## wt   15.800677  1        3.975007
## qsec  8.182966  1        2.860588
## vs    7.423472  1        2.724605
## am    5.910988  1        2.431252
## gear 25.668180  2        2.250861
## carb 12.681439  1        3.561101
```

```r
fit21 <- lm(mpg ~ . - disp, mtcars)
vif21 <- as.data.frame(vif(fit21))
vif21 %>% rownames_to_column('var') %>% filter(GVIF^(1/(2*Df)) > 3) %>%
  column_to_rownames('var')
```

```
##           GVIF Df GVIF^(1/(2*Df))
## hp   19.389094  1        4.403305
## carb  9.541618  1        3.088951
```

```r
fit22 <- lm(mpg ~ . - wt, mtcars)
vif22 <- as.data.frame(vif(fit22))
vif22 %>% rownames_to_column('var') %>% filter(GVIF^(1/(2*Df)) > 3) %>%
  column_to_rownames('var')
```

```
##           GVIF Df GVIF^(1/(2*Df))
## hp   21.325033  1        4.617904
## carb  9.798614  1        3.130274
```

```r
fit23 <- lm(mpg ~ . - carb, mtcars)
vif23 <- as.data.frame(vif(fit23))
vif23 %>% rownames_to_column('var') %>% filter(GVIF^(1/(2*Df)) > 3) %>%
  column_to_rownames('var')
```

```
##           GVIF Df GVIF^(1/(2*Df))
```

```
## disp 16.473542  1        4.058761
## hp    9.955196  1        3.155185
## wt   12.208767  1        3.494105
```

```
fit24 <- lm(mpg ~ . - hp, mtcars)
vif24 <- as.data.frame(vif(fit24))
vif24 %>% rownames_to_column('var') %>% filter(GVIF^(1/(2*Df)) > 3) %>%
  column_to_rownames('var')
```

```
##          GVIF Df GVIF^(1/(2*Df))
## disp 19.78489  1        4.448021
## wt   15.70392  1        3.962817
```

```
fit3 <- step(fit2, trace=0)
tidy(fit3)
```

```
##          term  estimate std.error statistic      p.value
## 1 (Intercept)  9.617781 6.9595930  1.381946 1.779152e-01
## 2          wt -3.916504 0.7112016 -5.506882 6.952711e-06
## 3        qsec  1.225886 0.2886696  4.246676 2.161737e-04
## 4   am manual  2.935837 1.4109045  2.080819 4.671551e-02
```

```
glance(fit3)
```

```
##   r.squared adj.r.squared    sigma statistic      p.value df    logLik
## 1 0.8496636     0.8335561 2.458846  52.74964 1.210446e-11  4 -72.05969
##        AIC      BIC deviance df.residual
## 1 154.1194 161.4481 169.2859          28
```

```
sqrt(vif(fit3))
```

```
##       wt     qsec       am
## 1.575738 1.168049 1.594189
```

Appendix

```
ggplot(mtcars, aes(x=am,y=mpg)) +
  geom_boxplot()
```

```
#ggpairs(mtcars, lower=list(combo=wrap('facethist',binwidth=0.8)))
```

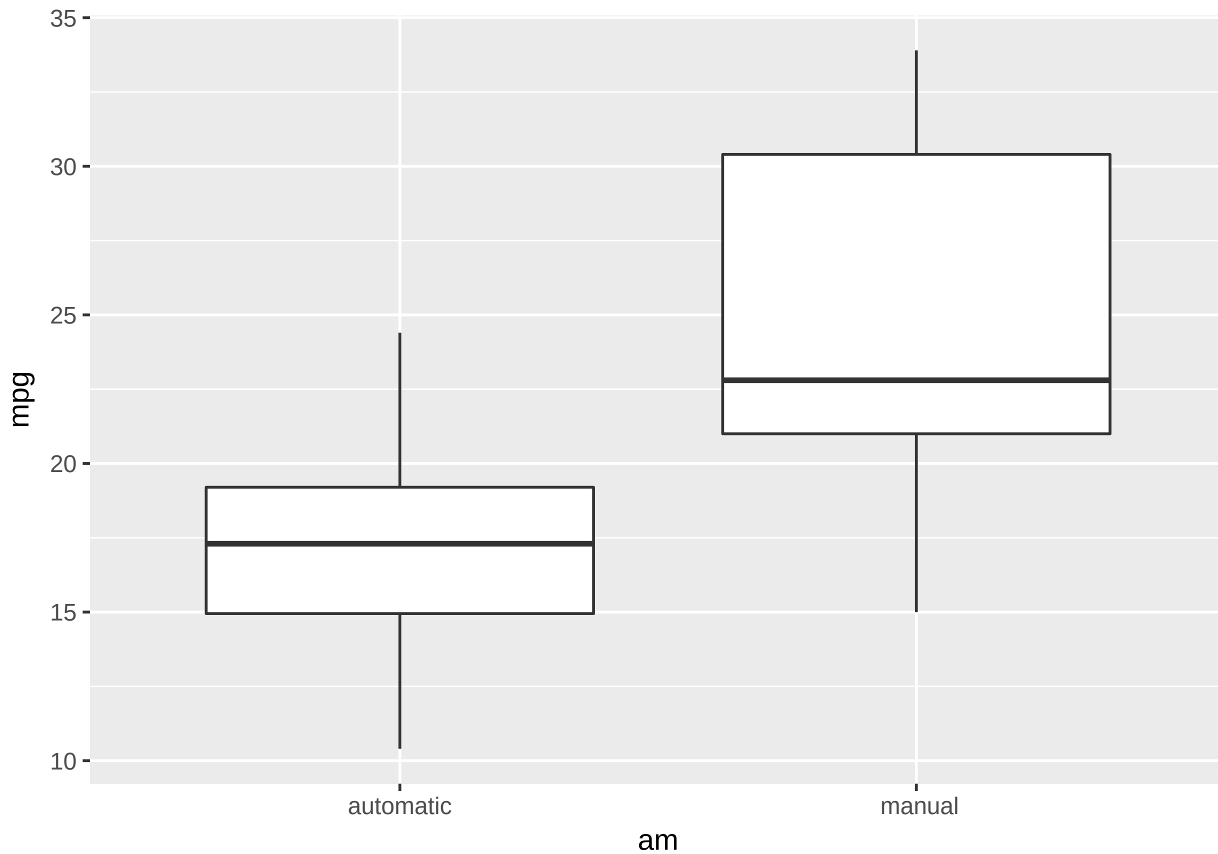

Figure 1: Box plot of MPG against transmission type.