Motor Trend Cars Miles per Gallon Analysis

Jon Ting

20/08/2020

Executive Summary

This report investigates the relationship between the type of transmission and miles per gallon (MPG) based on the `mtcars` dataset. A T-test between manual and automatic transmission vehicles shows that vehicles with the former transmission have a 7.245 greater MPG than the latter. Multiple linear regressions revealed that the former contributed less significantly to MPG, only an improvement of 1.81 MPG. The main contribution to the overall vehicle MPG originates from the other variables, including weight, horsepower, and number of cylinders.

Setup, Load and Preprocess Data

The data is loaded and preprocessed as below:

```
# Setup
knitr::opts_chunk$set(warning=FALSE)
# Load required libraries
library(ggplot2)
# Load data and examine data structure (see Appendix 1)
data(mtcars)
# Convert the categorical variable to appropriate class
categorical_vars <- c("cyl", "vs", "am", "gear", "carb")
mtcars[categorical_vars] <- lapply(mtcars[categorical_vars], factor)
# Rename binary variables
levels(mtcars$am) <- c("Automatic", "Manual")
levels(mtcars$vs) <- c("V-shaped", "Straight")
```

Exploratory Analysis

The relationships between all variables are first explored through a scatter plot matrix (see **Appendix 2**), where some strong correlations were observed between certain variables and MPG.

A box plot that compares the MPG of the 2 transmission types indicates that manual transmission vehicle allows longer MPG in general (see **Appendix 3**).

Regression Analysis

Simple Model

A simple model is first fitted to the data, including only the main explanatory and response variables. Albeit the p-value returned is small, the R-squared value indicates that only around 34% of the variance in MPG could be explained, hinting at the need for inclusion of more variables (see **Appendix 4**).

```
simpleModel <- lm(formula=mpg~am, data=mtcars)
```

Full Model

Therefore, all variables are included in the next analysis, resulting in the full model. This model returns opposite result compared to the previous model (see **Appendix 5**). Although about 89% of the MPG variance could be explained, all p-values are greater than 0.05, thus no result is significant. This implies that only the most significant variables should be included.

```
fullModel <- lm(formula=mpg~., data=mtcars)
```

Step Model

The determination of most statistically significant variables is achieved by both forward selection and backward elimination methods by AIC algorithm.

```
stepModel <- step(fullModel, direction="both", trace=0)
```

The resulting model includes cylinders, horsepower, and weight as confounder variables, and transmission being the response variable. It explains about 87% of the variance in MPG. The p-values are statistically significantly for all 3 confounder variables at the significance level of 0.05 (see **Appendix 6**).

The coefficients indicate that holding the other variables constant, the increment in the number of cylinders from 4 to 6 leads to lower MPG by 3.03, while further increment to 8 cylinders reduce the MPG by 2.16. Every unit increase in the horsepower corresponds to reduction in MPG by 0.0321 while every 1000 lbs increase in the vehicle weight decreases the MPG by 2.5 if other variables are held constant. A manual transmission corresponds to 1.81 higher MPG compared to automatic transmission.

Statistical Inference

The 95% confidence interval of each coefficient is computed and attached to **Appendix 7**. A Welch 2 sample T-test on transmission type and MPG indeed results in a p-value much smaller than 0.05 and a confidence interval that does not include 0, giving evidence to reject the null hypothesis that transmission type has no impact on MPG (see **Appendix 8**).

Residuals and Diagnostics

The residual plots (see **Appendix 9**) leads to the following conclusions:

- The assumption of independence is supported by the randomness in the residuals vs fitted plot
- The normality assumption is supported by the rather straight diagonal fit of the residuals in the normal Q-Q plot.
- The constant variance assumption is supported by the random distribution in the scale-location plot.
- No outlier is found as all points fall within the 0.5 boundaries in the residuals vs leverage plot.

Some regression diagnostics of the model are computed to find the leverage points:

```
leverage <- hatvalues(stepModel)
tail(sort(leverage), 3)
```

```
##       Toyota Corona Lincoln Continental       Maserati Bora
##          0.2777872           0.2936819          0.4713671
```

```
influential <- dfbetas(stepModel)
tail(sort(influential[, 6]), 3)
```

```
## Chrysler Imperial          Fiat 128     Toyota Corona
##         0.3507458         0.4292043         0.7305402
```

The top three points in each case of influence measures are found in the residual plots, thus indicating that the analysis was correct.

Conclusion

The inclusion of confounding variables like weight, horsepower, and number of cylinders in the analysis revealed that the difference in MPG based on transmission type is not found to be significant. The difference could largely be explained by the other variables instead. Thus the question of whether an automatic or manual transmission is better for MPG could not be answered.

Based on the best fit model, it could be said that holding the other confounder variables constant, a manual transmission vehicle could travel anywhere between 4.7 MPG longer or 1.1 MPG shorter than its counterpart. Other variables would be much better variables to be tuned for obtaining optimal MPG.

Appendix

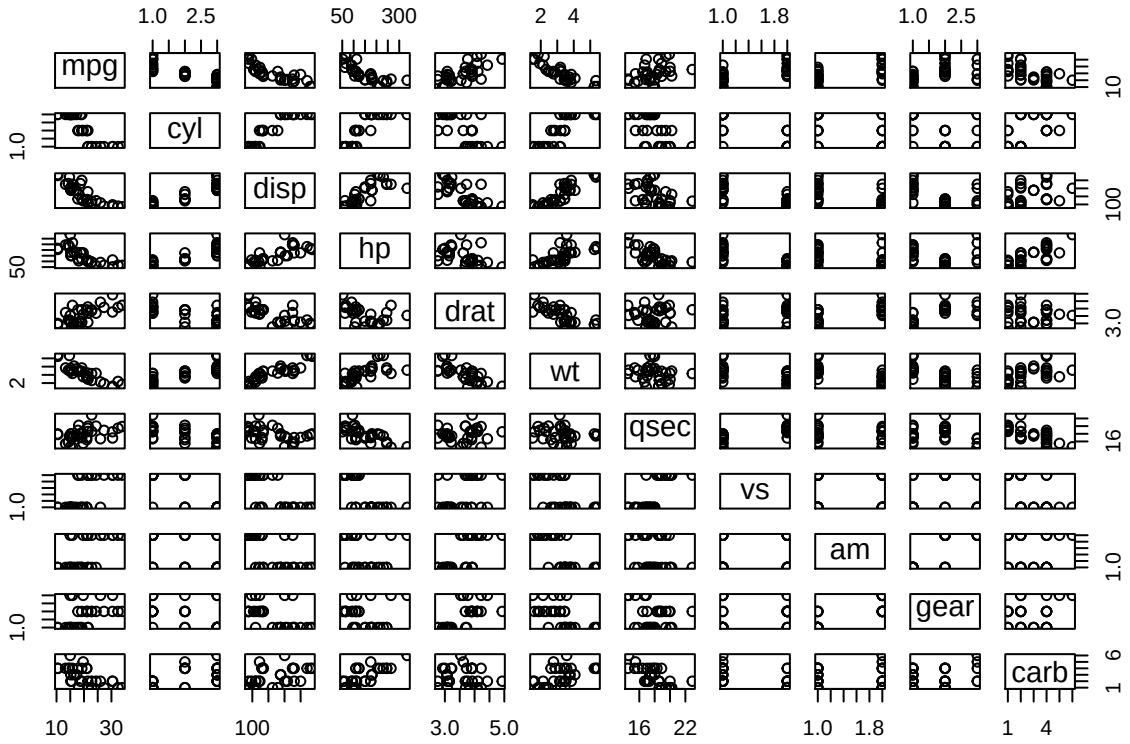
1. Data structure

```
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 : Factor w/ 3 levels "4","6","8": 2 2 1 2 3 2 3 1 1 2 ...
##  $ 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  : Factor w/ 2 levels "V-shaped","Straight": 1 1 2 2 1 2 1 2 2 2 ...
##  $ am  : Factor w/ 2 levels "Automatic","Manual": 2 2 2 1 1 1 1 1 1 1 ...
##  $ gear: Factor w/ 3 levels "3","4","5": 2 2 2 1 1 1 1 2 2 2 ...
##  $ carb: Factor w/ 6 levels "1","2","3","4",..: 4 4 1 1 2 1 4 2 2 4 ...
```

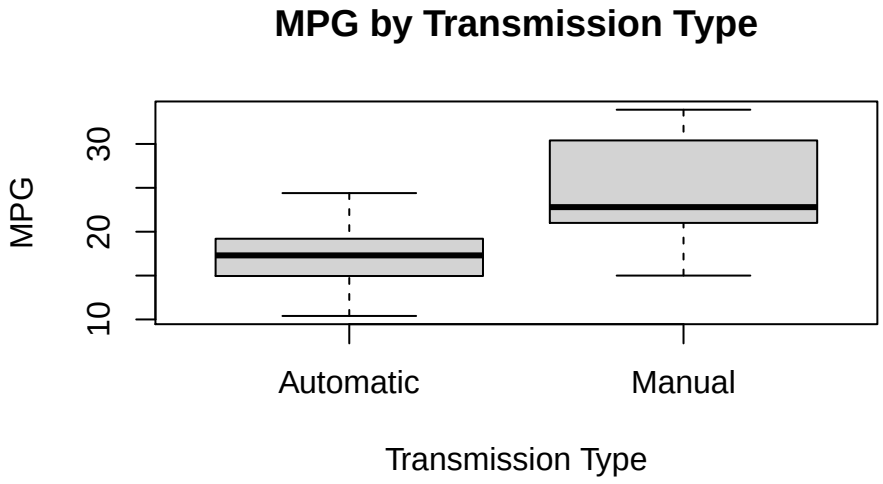
2. Scatter plot matrix

```
pairs(mpg~., data=mtcars)
```



3. Boxplot between transmission type and MPG

```
boxplot(formula=mpg~am, data=mtcars, xlab="Transmission Type", ylab="MPG", main="MPG by Transmission Ty
```



4. Simple model summary

```r
summary(simpleModel)
```

```
##
## 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 ***
## amManual       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
```

5. Full model summary

```r
summary(fullModel)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5087 -1.3584 -0.0948  0.7745  4.6251
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.87913   20.06582   1.190   0.2525
## cyl6        -2.64870    3.04089  -0.871   0.3975
## cyl8        -0.33616    7.15954  -0.047   0.9632
## disp         0.03555    0.03190   1.114   0.2827
## hp          -0.07051    0.03943  -1.788   0.0939 .
## drat         1.18283    2.48348   0.476   0.6407
## wt          -4.52978    2.53875  -1.784   0.0946 .
## qsec         0.36784    0.93540   0.393   0.6997
## vsStraight   1.93085    2.87126   0.672   0.5115
## amManual     1.21212    3.21355   0.377   0.7113
## gear4        1.11435    3.79952   0.293   0.7733
## gear5        2.52840    3.73636   0.677   0.5089
## carb2       -0.97935    2.31797  -0.423   0.6787
## carb3        2.99964    4.29355   0.699   0.4955
## carb4        1.09142    4.44962   0.245   0.8096
## carb6        4.47757    6.38406   0.701   0.4938
## carb8        7.25041    8.36057   0.867   0.3995
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 15 degrees of freedom
## Multiple R-squared:  0.8931, Adjusted R-squared:  0.779
## F-statistic:  7.83 on 16 and 15 DF,  p-value: 0.000124
```

6. Step model summary

```
summary(stepModel)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt + am, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9387 -1.2560 -0.4013  1.1253  5.0513
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 33.70832    2.60489  12.940 7.73e-13 ***
## cyl6        -3.03134    1.40728  -2.154  0.04068 *
## cyl8        -2.16368    2.28425  -0.947  0.35225
## hp          -0.03211    0.01369  -2.345  0.02693 *
## wt          -2.49683    0.88559  -2.819  0.00908 **
## amManual     1.80921    1.39630   1.296  0.20646
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.41 on 26 degrees of freedom
## Multiple R-squared:  0.8659, Adjusted R-squared:  0.8401
## F-statistic: 33.57 on 5 and 26 DF,  p-value: 1.506e-10
```

7. Confidence intervals

```
confint(stepModel)
```

```
##                   2.5 %       97.5 %
## (Intercept) 28.35390366 39.062744138
## cyl6        -5.92405718 -0.138631806
## cyl8        -6.85902199  2.531671342
## hp          -0.06025492 -0.003963941
## wt          -4.31718120 -0.676477640
## amManual    -1.06093363  4.679356394
```

8. Welch 2 sample T-test

```
t.test(formula=mpg~am, data=mtcars)
```

```
##
##  Welch Two Sample t-test
##
## data:  mpg by am
## t = -3.7671, df = 18.332, p-value = 0.001374
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -11.280194  -3.209684
## sample estimates:
## mean in group Automatic    mean in group Manual
##                17.14737                24.39231
```

9. Residual plot

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