Regression Models - Course Project

dillonchewwx

17/02/2021

Executive Summary

In this report, the `mtcars` dataset would be used to answer the following two questions:

1. Is an automatic or manual transmission better for MPG (miles per gallon)?
2. Quantify the MPG difference between automatic and manual transmissions.

Through a t-test, we have sufficient evidence (p=0.0014) to conclude that cars with manual transmissions provide better gas mileage than automatic transmission. Using regression models, a linear model shows that manual transmission provide 7.25 mpg more than automatic transmissions, but this single variable only accounts for 36% of the variation. By fitting additional variables to the model such as `disp`, `cyl`, `hp` and `wt`, the improvement was only 1.55 mpg, but this model accounts for 82.7% of the variation.

Exploratory Data Analysis

We will begin by loading the `mtcars` dataset and examining it.

```
library(datasets)
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 ...
```

```
summary(mtcars$mpg)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.40   15.43   19.20   20.09   22.80   33.90
```

It is noted that the transmission variable is stored in the column named `am` with 0 = automatic and 1 = manual - let's change the values for plotting.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
mtcars2<-mtcars %>%
    mutate(am=ifelse(am==0, "Automatic", "Manual"))
mtcars2$am
```

```
##  [1] "Manual"    "Manual"    "Manual"    "Automatic" "Automatic" "Automatic"
##  [7] "Automatic" "Automatic" "Automatic" "Automatic" "Automatic" "Automatic"
## [13] "Automatic" "Automatic" "Automatic" "Automatic" "Automatic" "Manual"
## [19] "Manual"    "Manual"    "Automatic" "Automatic" "Automatic" "Automatic"
## [25] "Automatic" "Manual"    "Manual"    "Manual"    "Manual"    "Manual"
## [31] "Manual"    "Manual"
```

Let's check the normality of the data.

```
library(rstatix)
```

```
##
## Attaching package: 'rstatix'

## The following object is masked from 'package:stats':
##
##     filter
```

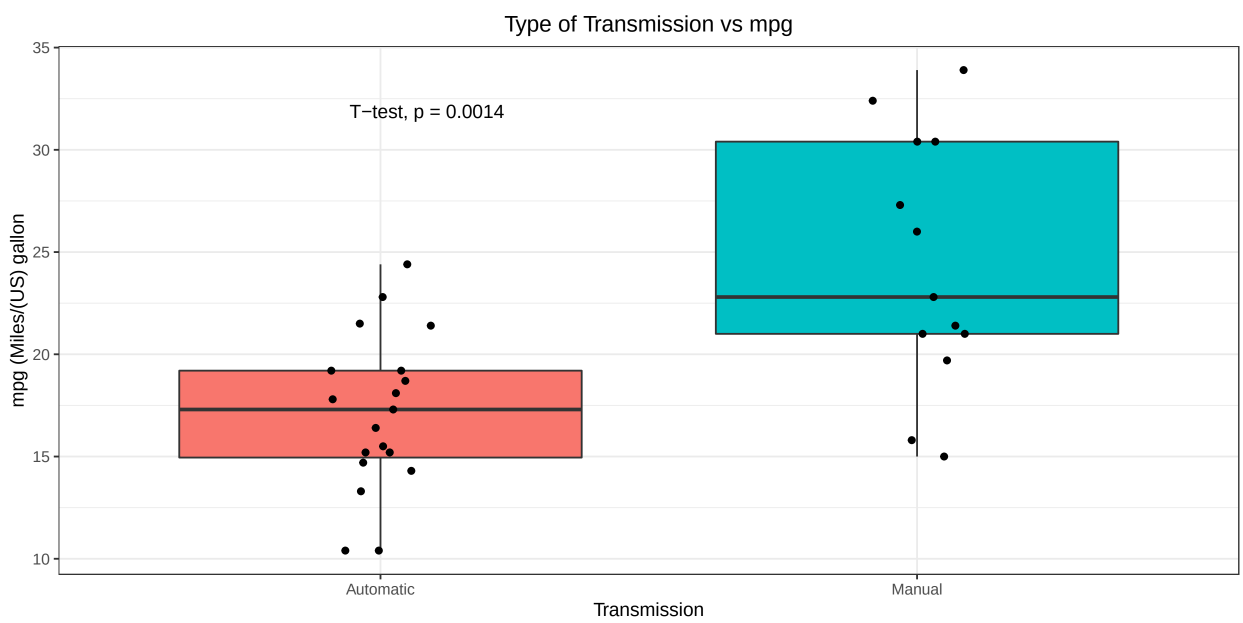
```r
mtcars2 %>%
    group_by(am) %>%
    shapiro_test(mpg)
```

```
## # A tibble: 2 x 4
##   am        variable statistic     p
##   <chr>     <chr>        <dbl> <dbl>
## 1 Automatic mpg          0.977 0.899
## 2 Manual    mpg          0.946 0.536
```

Since p > 0.05 for both Automatic and Manual transmission groups, we fail to reject the null hypothesis of the distributions being normal. We can proceed to use t-tests for the comparison of the mean MPG.

Let's see the boxplot of `mpg` with `am`.

```r
library(ggpubr)
ggplot(mtcars2, aes(x=am, y=mpg, fill=am)) +
    geom_boxplot(outlier.shape=NA) +
    geom_jitter(height=0, width=0.1) +
    theme_bw() +
    theme(legend.position="none", plot.title = element_text(hjust = 0.5)) +
    labs(x="Transmission", y="mpg (Miles/(US) gallon", title="Type of Transmission vs mpg") +
    stat_compare_means(method="t.test", label.x=1, label.y.npc=0.9)
```



From the boxplots and subsequent t-test (p=0.0014), we have sufficient evidence to conclude that cars with manual transmissions provide better gas mileage than automatic transmission.

Regression Models

In this section, we will attempt to use regression models to quantify the MPG difference between automatic and manual transmissions. To start, we can try a linear model to fit the variable `am` to the outcome `mpg`.

```r
fit1<-lm(mpg~am, mtcars)
summary(fit1)
```

```
##
## 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
```

Here, we see that the mean MPG for automatic cars is 17.1, while manual transmission provides 7.25 mpg more. However, the R^2 value suggests that transmission only accounts for 33.8% of the total variance, and thus a multivariate model might be a better fit to the data. Nonetheless, the p-value of < 0.05 for `am` suggests that there is a linear correlation with `mpg`.

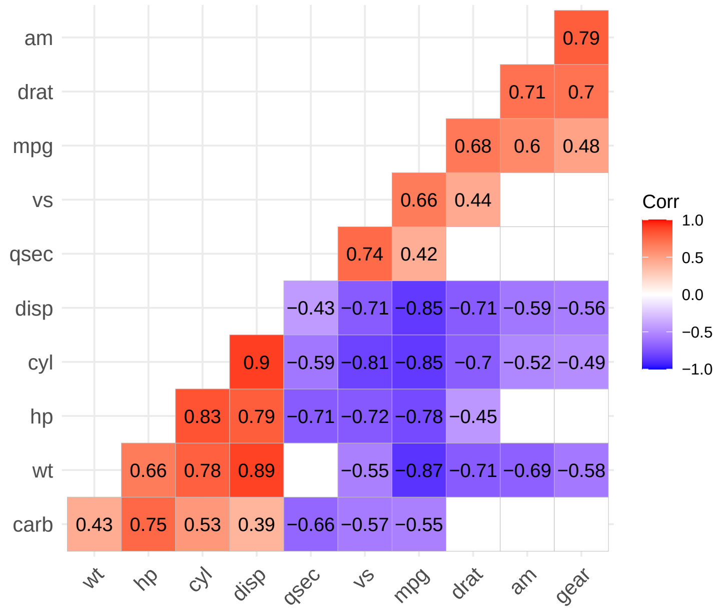
Lets check the correlation of the other variables to mpg.

```r
library(ggcorrplot)
```

```
##
## Attaching package: 'ggcorrplot'

## The following object is masked from 'package:rstatix':
##
##     cor_pmat
```

```r
corr<-cor(mtcars)
p.mat<-cor_pmat(mtcars)
ggcorrplot(corr, hc.order = TRUE, type = "lower", lab = TRUE, p.mat=p.mat, insig="blank")
```



From the correlation data, we observe that the variables drat and vs has moderately strong positive correlations with mpg (~0.6), while disp, cyl, hp and wt has strong negative correlations with mpg (~-0.8). Furthermore, we also note that all the abovementioned variables are significantly correlated with mpg. We shall move forward with multivariate models and begin by doing a nested model testing.

```
fit<-lm(mpg~., data=mtcars)
fit2<-update(fit, mpg~am+drat)
fit3<-update(fit, mpg~am+drat+vs)
fit4<-update(fit, mpg~am+disp)
fit5<-update(fit, mpg~am+disp+cyl)
fit6<-update(fit, mpg~am+disp+cyl+hp)
fit7<-update(fit, mpg~am+disp+cyl+hp+wt)
fit8<-update(fit, mpg~am+drat+vs+disp+cyl+hp+wt)
fit9<-update(fit, mpg~am+cyl)
fit10<-update(fit, mpg~am+cyl+hp)
fit11<-update(fit, mpg~am+cyl+hp+wt)
fit12<-update(fit, mpg~am+hp)
fit13<-update(fit, mpg~am+hp+wt)
fit14<-update(fit, mpg~am+wt)
anova(fit, fit1, fit2, fit3, fit4, fit5, fit6, fit7, fit8, fit9, fit10, fit11, fit12, fit13, fit14)
```

```
## Analysis of Variance Table
##
## Model  1: mpg ~ cyl + disp + hp + drat + wt + qsec + vs + am + gear + carb
## Model  2: mpg ~ am
## Model  3: mpg ~ am + drat
## Model  4: mpg ~ am + drat + vs
## Model  5: mpg ~ am + disp
## Model  6: mpg ~ am + disp + cyl
## Model  7: mpg ~ am + disp + cyl + hp
## Model  8: mpg ~ am + disp + cyl + hp + wt
## Model  9: mpg ~ am + drat + vs + disp + cyl + hp + wt
## Model 10: mpg ~ am + cyl
## Model 11: mpg ~ am + cyl + hp
```

```
## Model 12: mpg ~ am + cyl + hp + wt
## Model 13: mpg ~ am + hp
## Model 14: mpg ~ am + hp + wt
## Model 15: mpg ~ am + wt
##    Res.Df    RSS Df Sum of Sq       F    Pr(>F)    
## 1      21 147.49                                  
## 2      30 720.90 -9   -573.40  9.0711 1.779e-05 ***
## 3      29 573.64  1    147.26 20.9661 0.0001629 ***
## 4      28 339.99  1    233.65 33.2668 1.003e-05 ***
## 5      29 300.28 -1     39.71                      
## 6      28 252.08  1     48.20  6.8627 0.0160104 *  
## 7      27 216.37  1     35.71  5.0849 0.0349350 *  
## 8      26 163.12  1     53.25  7.5813 0.0119079 *  
## 9      24 158.65  2      4.47  0.3179 0.7311188    
## 10     29 271.36 -5   -112.71  3.2094 0.0262495 *  
## 11     28 220.55  1     50.81  7.2341 0.0137216 *  
## 12     27 170.00  1     50.56  7.1980 0.0139274 *  
## 13     29 245.44 -2    -75.44  5.3706 0.0130693 *  
## 14     28 180.29  1     65.15  9.2757 0.0061455 ** 
## 15     29 278.32 -1    -98.03 13.9571 0.0012194 ** 
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the nested model fit, we observe that the model with Df=1, lowest RSS, and significant p-value is model 8 with mpg ~ am + disp + cyl + hp + wt.

```r
summary(fit7)
```

```
## 
## Call:
## lm(formula = mpg ~ am + disp + cyl + hp + wt, 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    
## disp         0.01226    0.01171   1.047  0.30472    
## cyl         -1.10638    0.67636  -1.636  0.11393    
## hp          -0.02796    0.01392  -2.008  0.05510 .  
## wt          -3.30262    1.13364  -2.913  0.00726 ** 
## ---
## 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
```

From the summary, we see that the model explains 82.7% of the variation as given by the R^2 value, and manual transmissions result in a 1.55 mpg increase over automatic transmissions.

Appendix - Residual plots

```r
par(mfrow = c(2,2))
plot(fit7)
```