Peer-graded Assignment: Regression Models Course Project

MPG difference between automatic and manual transmissions

Eddy Delta

09/01/2019

Introduction

Executive summary

You work 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:

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

Dataset description

The data was extracted from the 1974 Motor Trend US magazine, and comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles (1973–74 models).

The data frame contains 32 observations on 11 (numeric) variables. [, 1] `mpg` Miles/(US) gallon

[, 2] `cyl` Number of cylinders

[, 3] `disp` Displacement (cu.in.)

[, 4] `hp` Gross horsepower

[, 5] `drat` Rear axle ratio

[, 6] `wt` Weight (1000 lbs)

[, 7] `qsec` 1/4 mile time

[, 8] `vs` Engine (0 = V-shaped, 1 = straight)

[, 9] `am` Transmission (0 = automatic, 1 = manual)

[,10] `gear` Number of forward gears

[,11] `carb` Number of carburetors

Data exploratory analysis

Data observations (see figures in the appendix)

By having a look at the *Data overview* in the figures appendix, we can observe that te mean MPG consumption is lower for **automatic** transmission (17.1473684) than for **manual** transmission (24.3923077). Is it significantly lower ?

Hypothesis test

Let's consider $H_0 : mean(MPG_{automatic}) \geq mean(MPG_{manual})$ and $H_A : mean(MPG_{automatic}) < mean(MPG_{manual})$.

```
mpg_test <- t.test(mpg ~ am, data = mtcars, alternative = 'less', paired = FALSE)
mpg_test$p.value
```

```
## [1] 0.0006868192
```

The `p-value`, is lower than `0.05`, so the null hypothesis is rejected. The **automatic** transmission average MPG is significantly lower than the **manual** transmission average MPG.

Regression models

regression with transmission (am) predictor (see figures in the appendix)

```r
mtcars2 <- transform(mtcars,
                     vs = factor(vs, c(0, 1), c('V-shaped', 'Straight')),
                     am = factor(am, c(0, 1), c('Automatic', 'Manual')))
# mtcars2$color <- c('0' = 'blue', '1' = 'red')[as.character(mtcars2$am)]
fit_am <- lm(mpg ~ am, mtcars2)
```

This model shows a significant relationship between transmission and consumption based on the predictor p-value. It estimate to 7.2449393 the average MPG when we switch from automatic transmission to manual transmission. Although the R-square 0.3597989 is quite low.

It's then interresting to investigate if additional parameter can provide a better prediction.

ANOVA (see figures in the appendix)

```r
fit_multiple <- lm(mpg ~ am + cyl + hp + wt + qsec + vs + gear + carb, mtcars2)
library(MASS)
summary(stepAIC(fit_multiple, direction = 'both', trace = FALSE))$call
```

```
## lm(formula = mpg ~ am + wt + qsec, data = mtcars2)
```

The residuals shows homoscedasticity, don't follow a clear pattern, and don't have outliers.

This model reveal that the weight, with a 0.0297848 p-value, is a more reliable predictor than the transmission (am) with 95% of confidence. It also point our attention to the qsec 1/4 mile time to be considered as a predictor. The adjusted R-squared is also much higher even if a slight increase due to the addition of predictors is expected.

The analysis of variance between the simple predictor model and the multiple one gives us arguments in favour of the multiple predictor model as the F-test p.value is significantly low. For the purpose of this analysis, an interesting analysis to perform seems to be the regression with the weight and the acceleration (qsec) relative to the transmission.

regression with multiple predictors (see figures in the appendix)

```r
fit_wt_qsec <- lm(mpg ~ am:wt + am:qsec, mtcars2)
```

The prediction of the MPG between automatic and manual transmission using the weight (wt) and the acceleration as predictor highlight a smaller MPG decrease for automatic than manual transmission.

However the MPG increase is higher with the acceleration (qsec) increase for the manual transmission than the automatic.

Conclusion

The data exploratory analysis allows us to assume with 95% of confidence that the manual transmission is better for the MPG than the automatic transmission.

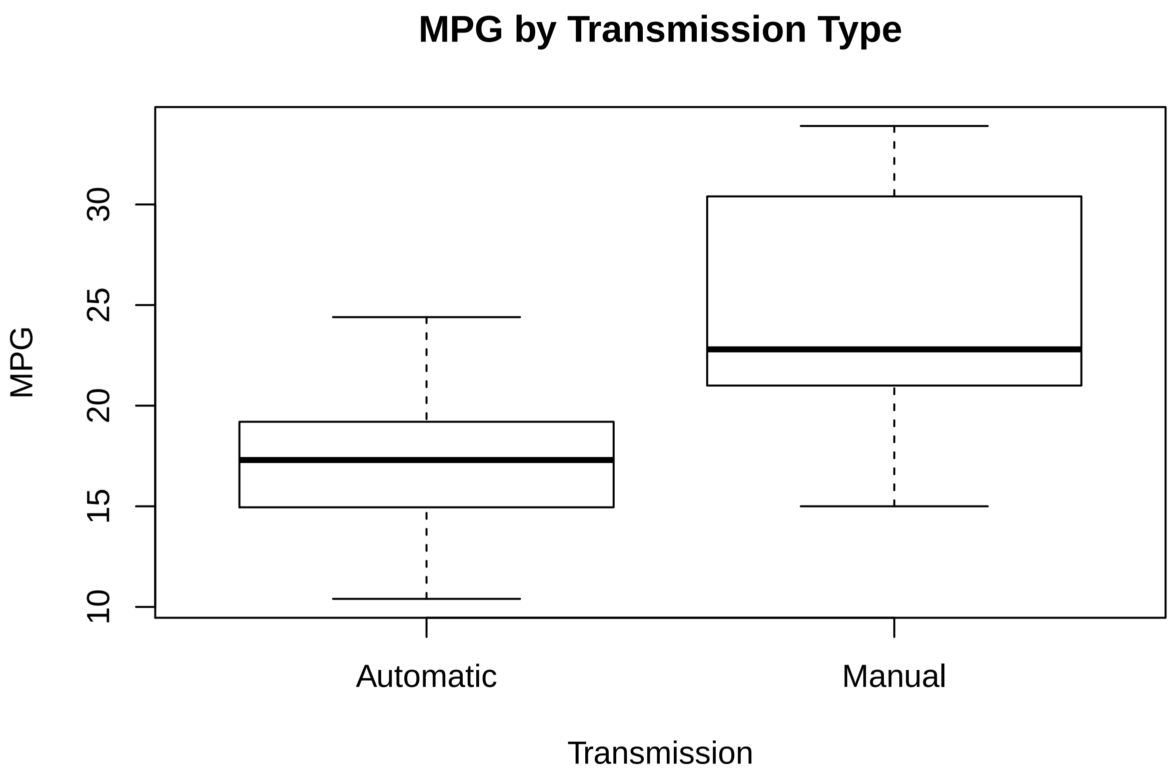
The regression analysis MPG overall average increase when you change from automatic to manual transmission. But looking deeper at the data reveal a strong but contrasted influence of the car weight and average acceleration on the MPG which requires a deeper look.

Appendix of figures

Data overview

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

Data distribution



regression with am predictor

```r
summary(fit_am)
```

```
##
## Call:
## lm(formula = mpg ~ am, data = mtcars2)
##
## 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
```

ANOVA

```
summary(fit_multiple)
```

```
##
## Call:
## lm(formula = mpg ~ am + cyl + hp + wt + qsec + vs + gear + carb,
##     data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9779 -1.3265 -0.2297  1.3211  4.6139
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.12752   16.50076   1.038   0.3101
## amManual     2.66026    1.98139   1.343   0.1925
## cyl         -0.05103    0.94991  -0.054   0.9576
## hp          -0.01328    0.01807  -0.735   0.4698
## wt          -2.71336    1.17127  -2.317   0.0298 *
## qsec         0.63603    0.68141   0.933   0.3603
## vsStraight   0.18232    2.04217   0.089   0.9296
## gear         0.81521    1.44686   0.563   0.5786
## carb        -0.55570    0.58763  -0.946   0.3541
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.585 on 23 degrees of freedom
## Multiple R-squared:  0.8635, Adjusted R-squared:  0.816
## F-statistic: 18.19 on 8 and 23 DF,  p-value: 3.089e-08
```

```
anova(
    fit_am,
    lm(mpg ~ am + wt + qsec, mtcars2),
    fit_multiple
)
```

```
## Analysis of Variance Table
##
## Model 1: mpg ~ am
## Model 2: mpg ~ am + wt + qsec
## Model 3: mpg ~ am + cyl + hp + wt + qsec + vs + gear + carb
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1     30 720.90
## 2     28 169.29  2    551.61 41.2703 2.458e-08 ***
## 3     23 153.71  5     15.58  0.4662    0.7973
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

regression with multiple predictors

```
summary(fit_wt_qsec)
```

```
##
## Call:
## lm(formula = mpg ~ am:wt + am:qsec, data = mtcars2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9361 -1.4017 -0.1551  1.2695  3.8862
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        13.9692     5.7756   2.419  0.02259 *
## amAutomatic:wt     -3.1759     0.6362  -4.992 3.11e-05 ***
## amManual:wt        -6.0992     0.9685  -6.297 9.70e-07 ***
## amAutomatic:qsec    0.8338     0.2602   3.205  0.00346 **
## amManual:qsec       1.4464     0.2692   5.373 1.12e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.097 on 27 degrees of freedom
## Multiple R-squared:  0.8946, Adjusted R-squared:  0.879
## F-statistic: 57.28 on 4 and 27 DF,  p-value: 8.424e-13
```

```
# pairs(mpg ~ ., data = mtcars)
par(mfrow=c(2,2))
plot(fit_wt_qsec)
```