Least Squares & Linear Regression

Regression Models: Assignment
Coursera Data Science: Statistics & Machine Learning Specialization

Mircea Dumitru

02 March, 2023

Executive Summary

Looking at the data set of a collection of cars `mtcars`, we are interested in exploring the relationship between a set of variables and miles per gallon (MPG) (outcome). The main questions addressed are:

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

The main challenge in addressing the questions is to asses and quantify the influence of the other variables in (partially) explaining the response variable and quantify the transmission type influence over the MPG variable accounting for these possible cofounders. This implies deciding which is the "best" model (*model selection*) by performing *analysis of covariance* (ANCOVA) and by *adjusting* and considering the R^2 values as a measure to decide what how much of the variance is explained by the model.

Data Exploration

```
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)
```

A new dataset is created with the categorical variables transformed in factors:

```
mtcarsFact <- mtcars
mtcarsFact$cyl <- as.factor(mtcarsFact$cyl)
mtcarsFact$vs<- as.factor(mtcarsFact$vs); levels(mtcarsFact$vs) = c("V-shaped", 'straight')
mtcarsFact$am <- as.factor(mtcarsFact$am); levels(mtcarsFact$am) = c("Automatic", 'Manual')
mtcarsFact$gear <- as.factor(mtcarsFact$gear)
mtcarsFact$carb <- as.factor(mtcarsFact$carb)
```

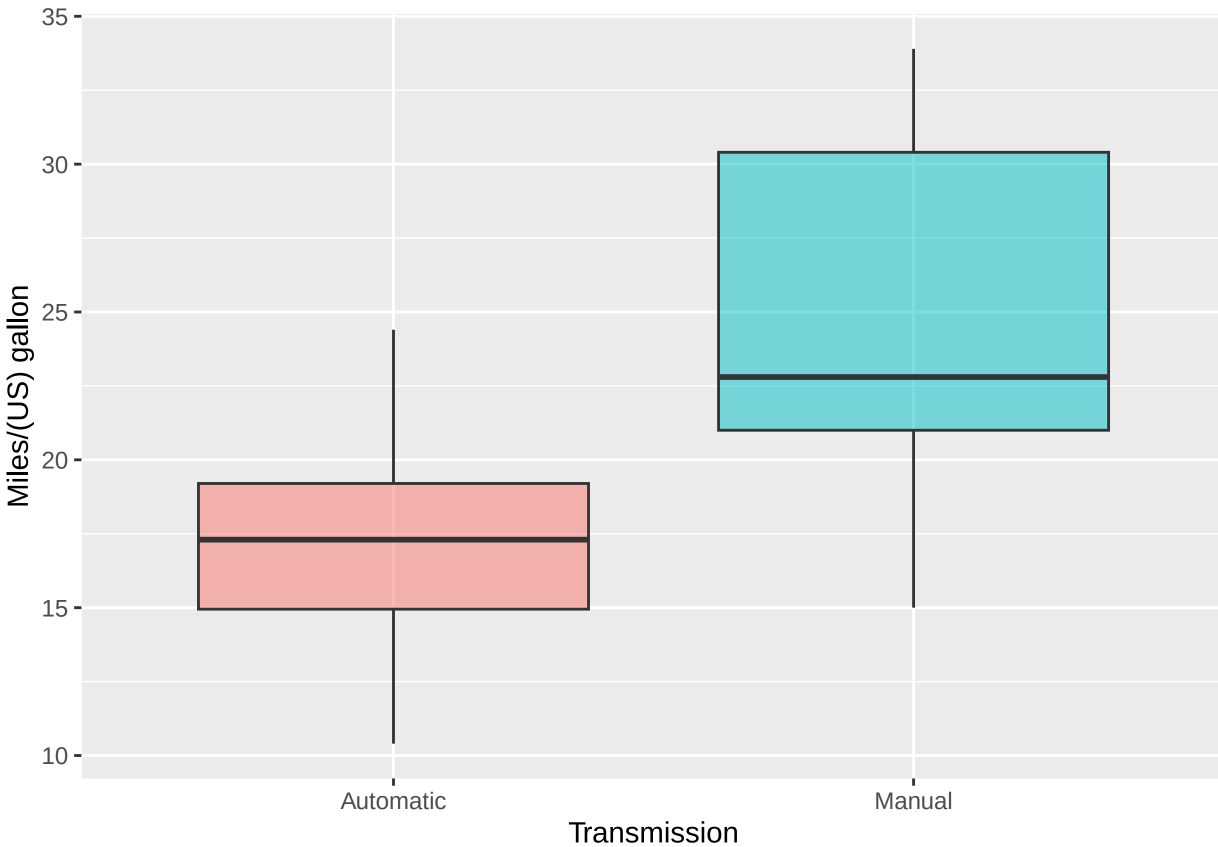
Exploratory analysis

The miles/gallon (response variable) versus the transmission type boxplot of the shows an *apparent* influence of the transmission type with an absolute mean value difference of 7.245 miles/gallon.

```r
aggregate(mtcarsFact$mpg, list(mtcarsFact$am), mean)
```

```
##     Group.1        x
## 1 Automatic 17.14737
## 2    Manual 24.39231
```

```r
library(ggplot2)
g <- ggplot(mtcarsFact, aes(x = am, y = mpg, fill = am)) + geom_boxplot(alpha=0.5) + theme(legend.posit
g <- g + xlab("Transmission") + ylab("Miles/(US) gallon")
g
```



This is precisely the expected change in response for a change in predictor from the automatic to manual transmission, since it is exactly the slope corresponding to a linear model that ignores all the other variables and assumes the transmission type as the only predictor for the response variable:

```r
lmPred1 <-lm(mpg ~ am, data = mtcarsFact)
lmPred1$coef
```

```
## (Intercept)    amManual
##   17.147368    7.244939
```

However, this model (i.e. the model accounting for the transmission type as the only predictor) can explain roughly 30% of the variation

```r
summary(lmPred1)$r.squared
```

```
## [1] 0.3597989
```

Model Selection via R^2

In order to decide how meaningful this influence is, the other variables available in the data set have to be accounted for. The variables will increase the R^2 are the ones that are have the smallest correlation with the transmission variable. The correlations between the transmission predictor variable and all the other potential predictor variables, excluding the response variable and the transmission variable itself, sorted according to their increasing magnitude are: :

```
corrs <- cor(mtcars$am, mtcars[ , !names(mtcars) %in% c('mpg', 'am')])
corrs <- corrs[, order(abs(corrs[1,]))]
corrs
```

```
##        carb          vs        qsec          hp         cyl        disp
##  0.05753435  0.16834512 -0.22986086 -0.24320426 -0.52260705 -0.59122704
##          wt        drat        gear
## -0.69249526  0.71271113  0.79405876
```

The model will improve in terms of R^2 score by sequentially adding the other variables, in increasing order with respect to their corresponding correlations w.r.t. the transmission variable:

```
# One additional predictor (ascending order of predictors w.r.t. their correlation with am)
# Fitting a linear model considering the carb variable beside the transmission variable as predictor
lmPred2 <-lm(mpg ~ am + carb, data = mtcarsFact)
# Extracting the corresponding R squared score
Rsq2 <- summary(lmPred2)$r.squared
# Extracting the correspnding p-value (for the F-test)
Pval2 <- pf(summary(lmPred2)$fstatistic[1], summary(lmPred2)$fstatistic[2], summary(lmPred2)$fstatistic
# Two additional predictors (ascending order of predictors w.r.t. their correlation with am)
lmPred3 <-lm(mpg ~ am + carb + vs, data = mtcarsFact)
Rsq3 <- summary(lmPred3)$r.squared
Pval3 <- pf(summary(lmPred3)$fstatistic[1], summary(lmPred3)$fstatistic[2], summary(lmPred3)$fstatistic
# Three additional predictors (ascending order of predictors w.r.t. their correlation with am)
lmPred4 <-lm(mpg ~ am + carb + vs + qsec, data = mtcarsFact)
Rsq4 <- summary(lmPred4)$r.squared
Pval4 <- pf(summary(lmPred4)$fstatistic[1], summary(lmPred4)$fstatistic[2], summary(lmPred4)$fstatistic
# Four additional predictors (ascending order of predictors w.r.t. their correlation with am)
lmPred5 <-lm(mpg ~ am + carb + vs + qsec + hp, data = mtcarsFact)
Rsq5 <- summary(lmPred5)$r.squared
Pval5 <- pf(summary(lmPred5)$fstatistic[1], summary(lmPred5)$fstatistic[2], summary(lmPred5)$fstatistic
# Five additional predictors (ascending order of predictors w.r.t. their correlation with am)
lmPred6 <-lm(mpg ~ am + carb + vs + qsec + hp + cyl, data = mtcarsFact)
Rsq6 <- summary(lmPred6)$r.squared
Pval6 <- pf(summary(lmPred6)$fstatistic[1], summary(lmPred6)$fstatistic[2], summary(lmPred6)$fstatistic
# Six additional predictors (ascending order of predictors w.r.t. their correlation with am)
lmPred7 <-lm(mpg ~ am + carb + vs + qsec + hp + cyl + disp, data = mtcarsFact)
Rsq7 <- summary(lmPred7)$r.squared
Pval7 <- pf(summary(lmPred7)$fstatistic[1], summary(lmPred7)$fstatistic[2], summary(lmPred7)$fstatistic
# Seven additional predictors (ascending order of predictors w.r.t. their correlation with am)
lmPred8 <-lm(mpg ~ am + carb + vs + qsec + hp + cyl + disp +  wt , data = mtcarsFact)
Rsq8 <- summary(lmPred8)$r.squared
Pval8 <- pf(summary(lmPred8)$fstatistic[1], summary(lmPred8)$fstatistic[2], summary(lmPred8)$fstatistic
# Eight additional predictors (ascending order of predictors w.r.t. their correlation with am)
lmPred9 <-lm(mpg ~ am + carb + vs + qsec + hp + cyl + disp +  wt + drat, data = mtcarsFact)
Rsq9 <- summary(lmPred9)$r.squared
Pval9 <- pf(summary(lmPred9)$fstatistic[1], summary(lmPred9)$fstatistic[2], summary(lmPred9)$fstatistic
# Nine additional predictors (ascending order of predictors w.r.t. their correlation with am)
```

```
lmPred10 <-lm(mpg ~ ., data = mtcarsFact)
Rsq10 <- summary(lmPred10)$r.squared
Pval10 <- pf(summary(lmPred10)$fstatistic[1], summary(lmPred10)$fstatistic[2], summary(lmPred10)$fstati

Rsq <- c(Rsq2, Rsq3, Rsq4, Rsq5, Rsq6, Rsq7, Rsq8, Rsq9, Rsq10)
Pval <- c(Pval2, Pval3, Pval4, Pval5, Pval6, Pval7, Pval8, Pval9, Pval10)

score <- data.frame(Rsq = Rsq, Pval = Pval, Correlations = corrs)
score
```

```
##            Rsq         Pval Correlations
## carb 0.7219336 6.084800e-06   0.05753435
## vs   0.8089823 3.057883e-07   0.16834512
## qsec 0.8119660 1.040108e-06  -0.22986086
## hp   0.8390198 7.766904e-07  -0.24320426
## cyl  0.8510404 4.934475e-06  -0.52260705
## disp 0.8549764 1.319714e-05  -0.59122704
## wt   0.8872699 5.240581e-06  -0.69249526
## drat 0.8895468 1.540471e-05   0.71271113
## gear 0.8930749 1.240147e-04   0.79405876
```

Including the variables with a corresponding correlation with the transmission variable having an absolute value less than 0.5, i.e. the model using as predictor variables `carb`, `vs`, `qsec`, `hp` beside the transmission variable `am` leads to a linear model with a corresponding R^2 equal to 0.8390198, i.e. a model that accounts for roughly 84% of the variation.

Including the variables with a corresponding correlation with the transmission variable having an absolute value less than 0.6, i.e. the model using as predictor variables `carb`, `vs`, `qsec`, `hp`, `cyl`, `disp`, beside the transmission variable `am` leads to a linear model with a corresponding R^2 equal to 0.8549764, i.e. a model that accounts for roughly 85.5% of the variation.

Any extra predictor will just marginally improve the R^2 (as expected, accounting for the fact that the extra variables included have significant correlations with the transmission variable): for the "full" linear model R^2 is 0.8930749 (i.e. an improvement of roughly 4.5%).

Hence any model that considers the variables with a corresponding correlation with the transmission variable having an absolute value less than 0.5 (`carb`, `vs`, `qsec`, `hp`) or less than 0.6 (`carb`, `vs`, `qsec`, `hp`, `cyl`, `disp`) explain at between 84% and 85.% of the variation.

For these models the quantification of the difference between the two transmission modes is given by the corresponding estimated regressor (the corresponding β):

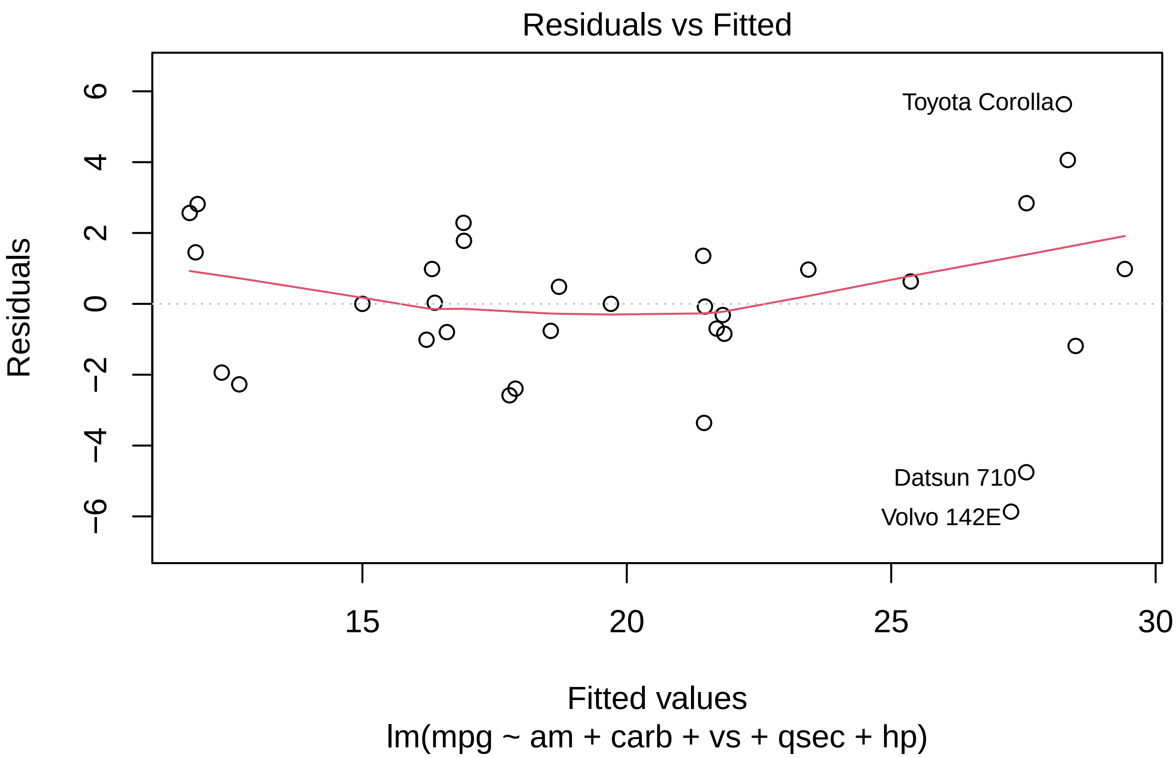
```
betas <- c(lmPred5$coef['amManual'], lmPred6$coef['amManual'], lmPred7$coef['amManual'])
betas
```

```
## amManual amManual amManual
## 5.225277 3.756240 3.575036
```

Influence measures

The model using as predictor variables `carb`, `vs`, `qsec`, `hp` beside the transmission variable `am` has a residuals vs. fitted that doesn't present a trend, hence it seems the variance unexplained by the model is due indeed to the noise.

```
plot(lmPred5, which=c(1,1))
```



MPG difference quantification

- A linear the model with predictos `carb`, `vs`, `qsec`, `hp` beside `am` has the R^2 equal to 0.8390198 and the expected change from switching from automatic to manual transmistion is 5.225277.
- A linear the model with predictos `carb`, `vs`, `qsec`, `hp`, `cyl` beside `am` has the R^2 equal to 0.8510404 and the expected change from switching from automatic to manual transmistion is 3.756240
- A linear the model with predictos `carb`, `vs`, `qsec`, `hp`, `cyl`, `disp` beside `am` has the R^2 equal to 0.8549764 and the expected change from switching from automatic to manual transmistion is 3.575036

Conclusions

1. It worth mentioning that the analysis makes sense in the context of the questions addressed, i.e. assessing the influence of the transmission `am` over the miles/gallon consumption `mpg`. This is why it makes sense to consider removing the variables that are highly correlated with `am` (namely `gear`, `drat` and `wt`). A straightforward way to select a model is via AIC, but in this case the high correlations between predictors are not accounted for, leading to an "optimal" model containing some of those variables. In this case it difficult to quantify the real influence of `am`.

```
#lmPredAIC <- step(lmPred10, direction = 'both', trace = FALSE)
#summary(lmPredAIC)
```

2. A manual tranmission apears to be marginally better for the number of miles per gallon. The expected change from switching from automatic to manual transmistion is estimated to be between 3.5 and 5.225 miles per galon.