Introduction to Machine Learning

(NPFL054)

Homework 1

François Leroy, PhD student at CZU

2021-04-17

Contents

1. Multiple linear regression

1.1

```r
library(ISLR)
library(tidyverse)
```

```r
# Perform the multiple linear regression
lm <-
  lm(mpg ~ ., data = subset(Auto, select = -name))
# Print the output
summary(lm)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = subset(Auto, select = -name))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707  0.00024 ***
## cylinders     -0.493376   0.323282  -1.526  0.12780
## displacement   0.019896   0.007515   2.647  0.00844 **
## horsepower    -0.016951   0.013787  -1.230  0.21963
## weight        -0.006474   0.000652  -9.929  < 2e-16 ***
## acceleration   0.080576   0.098845   0.815  0.41548
## year           0.750773   0.050973  14.729  < 2e-16 ***
## origin         1.426141   0.278136   5.127 4.67e-07 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 2.2e-16
```

First of all, the adjusted $R^2 = 0.82$, which means that 82% of the variance of the data is explained by this models. This is a very trustful model.

Here, I will talk about only about the covariates that have a significant influence (*i.e.* $p - value \leq 0.05$) on the `mpg` variable (*i.e.* rows with an asterix such as `displacement`, `weight`, `year` and `origin`):

- The miles per galon unit (*i.e.* `mpg`) expresses the fuel economy of a vehicle. Thus, when the coefficient of the `lm` is negative, it means that the vehicle will tend to go less further with a unit of fuel. Here, this is the case for the `weight` variable: a heavier vehicle will consume more fuel than a lighter one.
- The other significant relationships with the `displacement`, `year` and `origin` are positive which means that a more recent car, with a higher displacement volume and with a higher origin will tend to consume less fuel.

1.2

```r
## Perform the 5 polynomial simpple linear regression
for (i in 1:5){
  assign(paste0("fit", i),
         lm(mpg ~ poly(acceleration, i), data = subset(Auto, select = -name)))
}
## Plot them on a single plot
#### First merge the predicted values of mpg with the acceleration
Auto %>%
  select(acceleration) %>%
  cbind(poly1 = fit1$fitted.values,
        poly2 = fit2$fitted.values,
        poly3 = fit3$fitted.values,
        poly4 = fit4$fitted.values,
        poly5 = fit5$fitted.values) %>%
##### Then format the data for ggplot
  pivot_longer(cols = poly1:poly5,
               names_to = "poly",
               values_to = "mpg") %>%
  mutate(rsq = case_when(
    poly == "poly1" ~ round(summary(fit1)$adj.r.squared, digits = 2),
    poly == "poly2" ~ round(summary(fit2)$adj.r.squared, digits = 2),
    poly == "poly3" ~ round(summary(fit3)$adj.r.squared, digits = 2),
    poly == "poly4" ~ round(summary(fit4)$adj.r.squared, digits = 2),
    poly == "poly5" ~ round(summary(fit5)$adj.r.squared, digits = 2)
  )) %>%
  unite(poly, c("poly", "rsq"), sep = ", adj.R² = ") %>%
#### Now plot it
  ggplot()+
  geom_point(aes(acceleration, mpg), data = subset(Auto, select = -name))+
```

```
geom_line(aes(acceleration, mpg, color = poly), size = 1.2)+
theme_classic()
```