Lab 9 # version 2

Yunting Chiu

2021-03-12

R Lab 9

```
# read the data from the web
autompg = read.table(
  "http://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data",
  quote = "\"",
  comment.char = "",
  stringsAsFactors = FALSE)
# give the dataframe headers
colnames(autompg) = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year", "origin", "name") # remove missi
autompg = subset(autompg, autompg$hp != "?")
# remove the plymouth reliant, as it causes some issues
autompg = subset(autompg, autompg$name != "plymouth reliant")
# give the dataset row names, based on the engine, year and name
rownames(autompg) = paste(autompg$cyl, "cylinder", autompg$year, autompg$name)
# remove the variable for name, as well as origin
autompg = subset(autompg, select = c("mpg", "cyl", "disp", "hp", "wt", "acc", "year")) # change horsepo
autompg$hp = as.numeric(autompg$hp)
# check final structure of data
str(autompg)
```

```
## 'data.frame':    390 obs. of  7 variables:
##  $ mpg : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cyl : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ disp: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ hp  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ wt  : num  3504 3693 3436 3433 3449 ...
##  $ acc : num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year: int  70 70 70 70 70 70 70 70 70 70 ...
```

```
head(autompg)
```

```
##                                      mpg cyl disp  hp   wt  acc year
## 8 cylinder 70 chevrolet chevelle malibu  18   8  307 130 3504 12.0   70
## 8 cylinder 70 buick skylark 320          15   8  350 165 3693 11.5   70
## 8 cylinder 70 plymouth satellite         18   8  318 150 3436 11.0   70
## 8 cylinder 70 amc rebel sst              16   8  304 150 3433 12.0   70
## 8 cylinder 70 ford torino                17   8  302 140 3449 10.5   70
## 8 cylinder 70 ford galaxie 500           15   8  429 198 4341 10.0   70
```

Task 1

Use the lm function and provide estimates of $\beta 0$, $\beta 1$, and $\beta 2$.

```
mpg_model <- lm(mpg ~ wt+year, data = autompg)
coef(mpg_model) # b0, b1, b2
```

```
##   (Intercept)            wt          year
## -14.637641945  -0.006634876   0.761401955
```

Task 2

```
n = nrow(autompg) # 390 observations
p = length(coef(mpg_model)) # b0, b1, b2
X = cbind(rep(1, n), autompg$wt, autompg$year) # x as defined above
y = autompg$mpg # column vector

# solve: a %*% x = b for x, where b can be either a vector or a matrix.
beta_hat = solve(t(X) %*% X) %*% t(X) %*% y # equation
# transport  = t(X), solve : find the inverse
beta_hat
```

```
##               [,1]
## [1,] -14.637641945
## [2,]  -0.006634876
## [3,]   0.761401955
```

Task 3

- In statistics, the residual sum of squares (RSS), also known as the sum of squared residuals (SSR) or the sum of squared estimate of errors (SSE), is the sum of the squares of residuals (deviations predicted from actual empirical values of data).

```
MSE <- sum(residuals(mpg_model)^2)/(n-p) # sum of residual square
sqrt(MSE)
```

```
## [1] 3.431367
```

- Second method

```
yhat = X %*% solve(t(X) %*% X) %*% t(X) %*% y
e = y - yhat
# e
sqrt(t(e) %*% e / (n-p))
```

```
##          [,1]
## [1,] 3.431367
```

Task 4

The Adjusted R-squared is 0.8082355, meaning that the model has 81% of variability being explained. The observed variation `in miles per gallon` is explained of 81% by the linear relationship with other two predictors (`weight` and `year`.)

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

```
## [1] 0.8082355
```