Lab 9 # version 1

Yunting Chiu

2021-03-06

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 for b0, b1, b2.

```
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)
p = length(coef(mpg_model))
X = cbind(rep(1, n), autompg$wt, autompg$year)
y = autompg$mpg

beta_hat = solve(t(X) %*% X) %*% t(X) %*% y
beta_hat
```

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

Task 3

- https://stackoverflow.com/questions/43123462/how-to-obtain-rmse-out-of-lm-result

```
# Residual sum of squares
RSS <- c(crossprod(mpg_model$residuals))
RSS
```

```
## [1] 4556.646
```

```
# Mean squared error
MSE <- RSS / length(mpg_model$residuals)
MSE
```

```
## [1] 11.68371
```

```
# Root MSE
RMSE <- sqrt(MSE)
RMSE
```

```
## [1] 3.418144
```

Task 4

The Adjusted R-squared is 0.06476, meaning that after we adjusted the model, we have 80% of variability being explained by the model. The R-squared is how well the regression model fits the observed data.

```
summary(mpg_model)
```

```
##
## Call:
## lm(formula = mpg ~ wt + year, data = autompg)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.852 -2.292 -0.100  2.039 14.325
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.464e+01  4.023e+00  -3.638 0.000312 ***
## wt          -6.635e-03  2.149e-04 -30.881  < 2e-16 ***
## year         7.614e-01  4.973e-02  15.312  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.431 on 387 degrees of freedom
## Multiple R-squared:  0.8082, Adjusted R-squared:  0.8072
## F-statistic: 815.6 on 2 and 387 DF,  p-value: < 2.2e-16
```