lab04-ols-regression-Katia-Williams

Katia Williams

2/12/2018

Part 1:EDA

```
names(mtcars)
```

```
##  [1] "mpg"  "cyl"  "disp" "hp"   "drat" "wt"   "qsec" "vs"   "am"   "gear"
## [11] "carb"
```

```
cars <- mtcars[c(4, 7, 6, 1)]
head(cars)
```

```
##                    hp  qsec    wt  mpg
## Mazda RX4         110 16.46 2.620 21.0
## Mazda RX4 Wag     110 17.02 2.875 21.0
## Datsun 710         93 18.61 2.320 22.8
## Hornet 4 Drive    110 19.44 3.215 21.4
## Hornet Sportabout 175 17.02 3.440 18.7
## Valiant           105 20.22 3.460 18.1
```

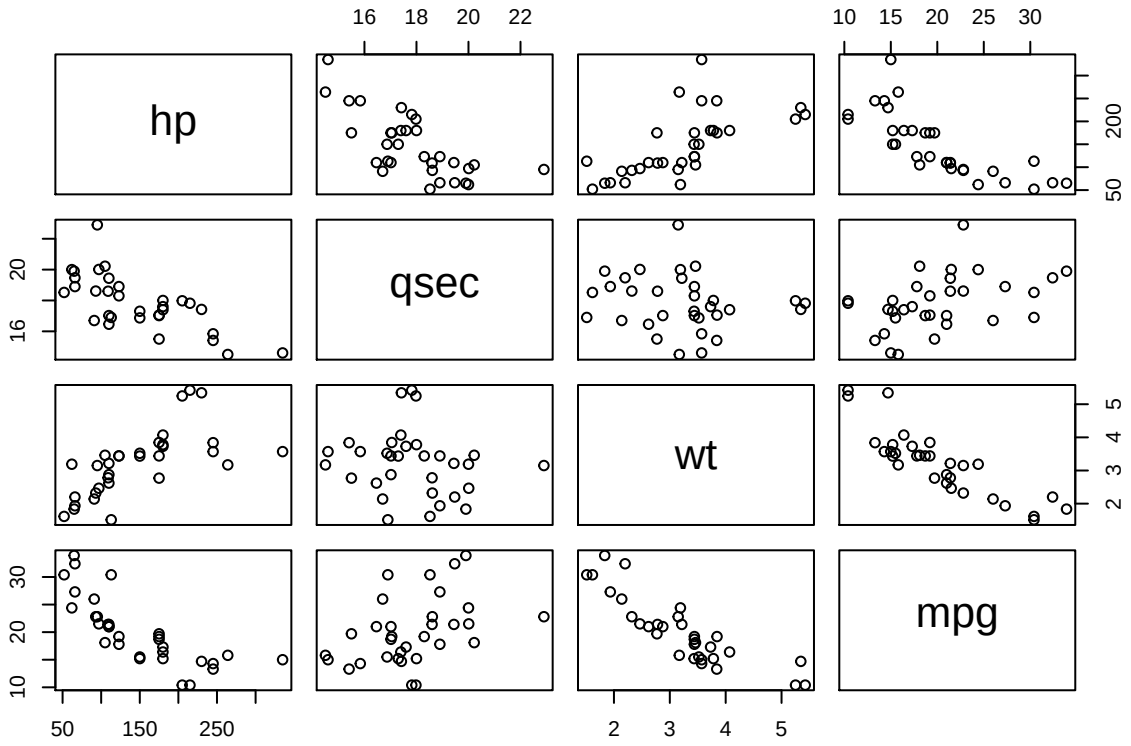
```
summary(cars)
```

```
##        hp             qsec             wt             mpg
##  Min.   : 52.0   Min.   :14.50   Min.   :1.513   Min.   :10.40
##  1st Qu.: 96.5   1st Qu.:16.89   1st Qu.:2.581   1st Qu.:15.43
##  Median :123.0   Median :17.71   Median :3.325   Median :19.20
##  Mean   :146.7   Mean   :17.85   Mean   :3.217   Mean   :20.09
##  3rd Qu.:180.0   3rd Qu.:18.90   3rd Qu.:3.610   3rd Qu.:22.80
##  Max.   :335.0   Max.   :22.90   Max.   :5.424   Max.   :33.90
```

```
cor(cars)
```

```
##              hp       qsec         wt        mpg
## hp    1.0000000 -0.7082234  0.6587479 -0.7761684
## qsec -0.7082234  1.0000000 -0.1747159  0.4186840
## wt    0.6587479 -0.1747159  1.0000000 -0.8676594
## mpg  -0.7761684  0.4186840 -0.8676594  1.0000000
```

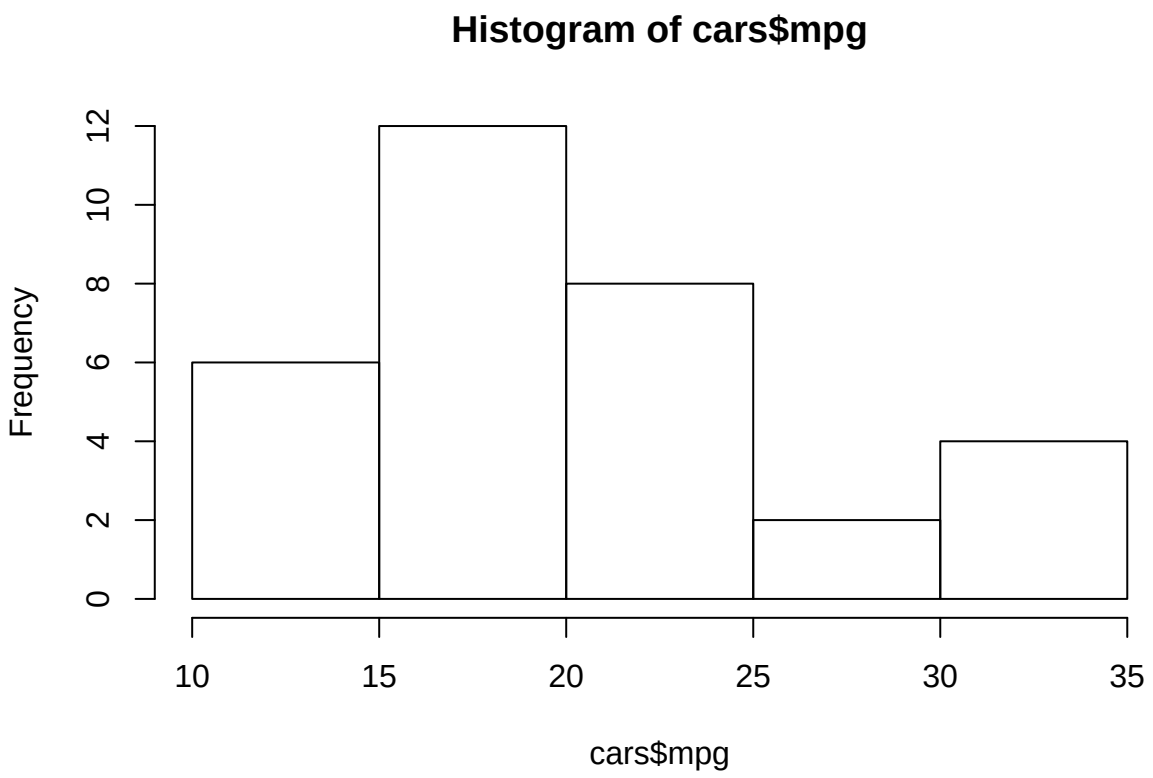
```
pairs(cars)
```



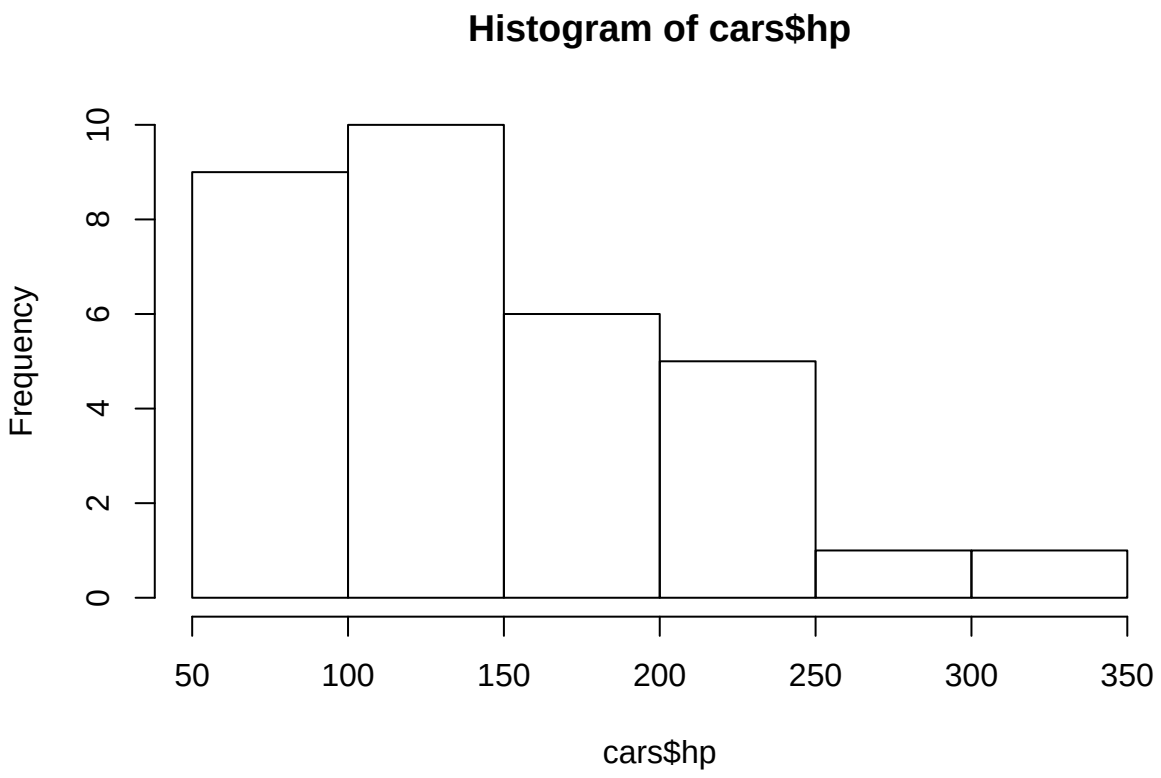

```
prcomp(cars)
```

```
## Standard deviations (1, .., p=4):
## [1] 68.7374255  3.8533877  1.2237768  0.4021126
##
## Rotation (n x k) = (4 x 4):
##                PC1         PC2        PC3          PC4
## hp    0.997453468  0.06643932 0.02542544 -0.005093549
## qsec -0.018396105 -0.11140955 0.96740653 -0.226658563
## wt    0.009406474 -0.14735595 0.20966894  0.966559195
## mpg  -0.068261860  0.98054064 0.13970358  0.119846874
```

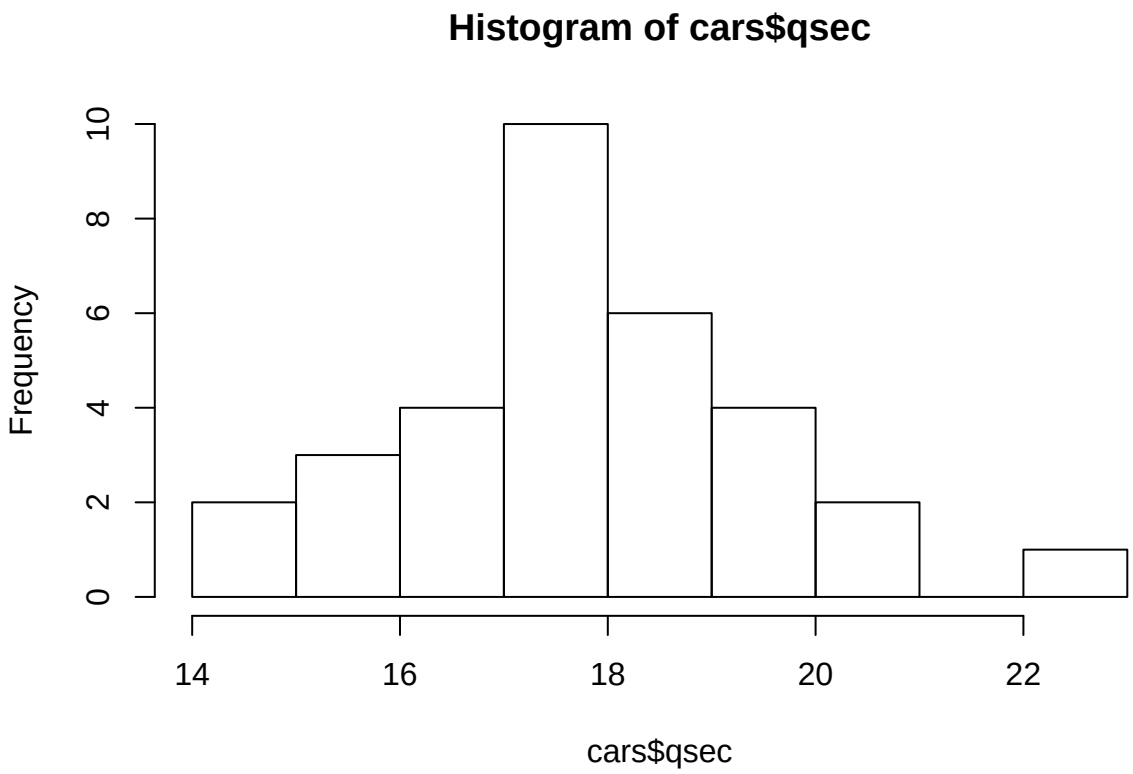
```
hist(cars$mpg)
```



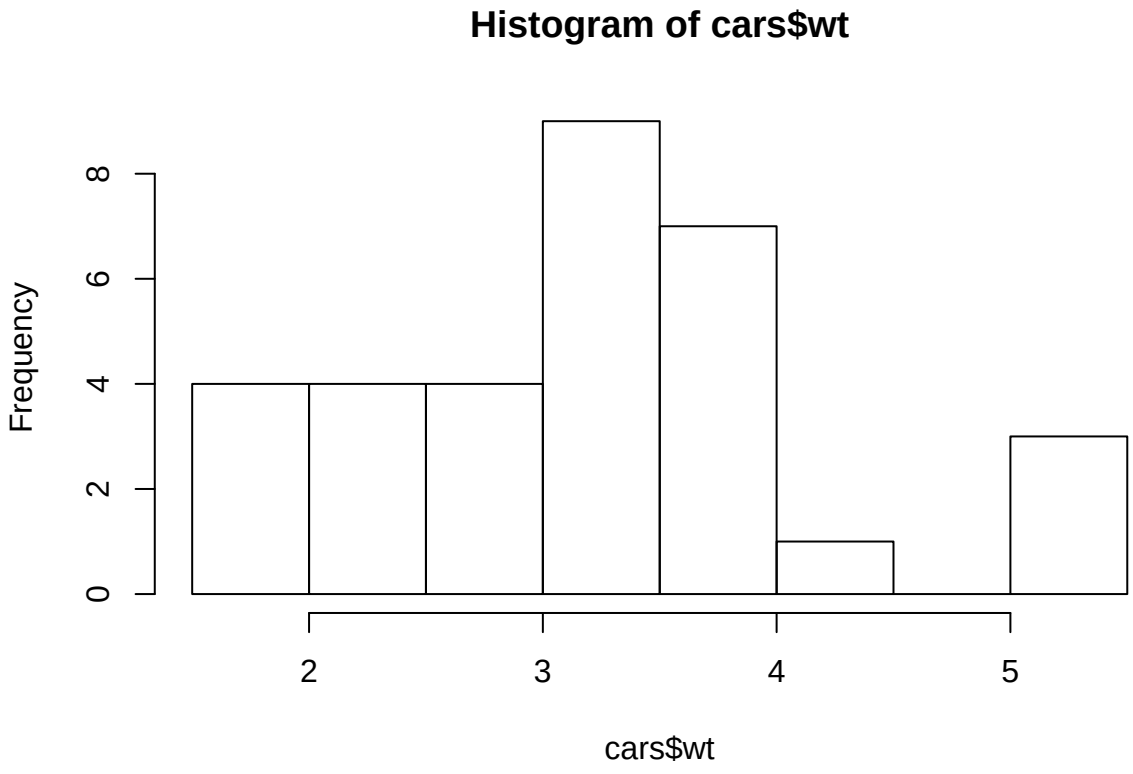

```
hist(cars$hp)
```




```
hist(cars$qsec)
```


Histogram of cars$qsec
Frequency
0
2
4
6
8
10
14
16
18
20
22
cars$qsec

```
hist(cars$wt)
```


Histogram of cars$wt
Frequency
0
2
4
6
8
2
3
4
5
cars$wt

Part 2: OLS Outputs

X

```
get_X <- function(df, n, p) {
  rawX <- df[c(1:p)]
  rawX['ones'] = rep(1, n)
  return (as.matrix(rawX[c(p+1, c(1:p))]))
}
X <- get_X(cars,length(cars$hp), 3)
head(X)
```

```
##                   ones  hp  qsec    wt
## Mazda RX4            1 110 16.46 2.620
## Mazda RX4 Wag        1 110 17.02 2.875
## Datsun 710           1  93 18.61 2.320
## Hornet 4 Drive       1 110 19.44 3.215
## Hornet Sportabout    1 175 17.02 3.440
## Valiant              1 105 20.22 3.460
```

Y

```
get_y <- function(df) {
  return(as.matrix(df[,ncol(df)]))
}
y <- get_y(cars)
head(y)
```

```
##      [,1]
## [1,] 21.0
## [2,] 21.0
## [3,] 22.8
## [4,] 21.4
## [5,] 18.7
## [6,] 18.1
```

Coefficients

Assumption: X has full rank

```
get_betahats <- function(X, y) {
  return (solve( t(X) %*% X ) %*% t(X) %*% y)
}
Coefficients <- get_betahats(X, y)
Coefficients
```

```
##             [,1]
## ones 27.61052686
## hp   -0.01782227
## qsec  0.51083369
## wt   -4.35879720
```

Fitted values

```
fitted_values <- X %*% Coefficients
fitted_values
```

```
##                           [,1]
## Mazda RX4            22.638351
## Mazda RX4 Wag        21.812925
## Datsun 710           25.347261
## Hornet 4 Drive       21.567151
## Hornet Sportabout    18.191756
## Valiant              20.986807
## Duster 360           15.774770
## Merc 240D            22.817657
## Merc 230             23.885291
## Merc 280             19.772382
## Merc 280C            20.078882
## Merc 450SE           15.550720
## Merc 450SL           17.134877
## Merc 450SLC          17.121271
## Cadillac Fleetwood   10.258066
## Lincoln Continental   9.239679
## Chrysler Imperial     9.112356
## Fiat 128             26.790835
## Honda Civic          29.104951
## Toyota Corolla       28.619277
## Toyota Corona        25.359114
## Dodge Challenger     18.211984
## AMC Javelin          18.802141
## Camaro Z28           14.378236
## Pontiac Firebird     16.441769
## Fiat X1-9            27.654741
## Porsche 914-2        25.191797
## Lotus Europa         27.634839
## Ford Pantera L       16.495149
## Ferrari Dino         20.335683
## Maserati Bora        13.537332
## Volvo 142E           23.051950
```

Residuals

```
residuals <- fitted_values - y
head(residuals)
```

```
##                           [,1]
## Mazda RX4           1.6383509
## Mazda RX4 Wag       0.8129245
## Datsun 710          2.5472611
## Hornet 4 Drive      0.1671510
## Hornet Sportabout -0.5082436
## Valiant             2.8868073
```

RSS

```
RSS <- t(residuals) %*% residuals
RSS
```

```
##           [,1]
## [1,] 186.0593
```

Sigma2

```
n <- length(X[,1])
p <- 3
sigma2 <- RSS/(n - p - 1)
sigma2
```

```
##           [,1]
## [1,] 6.644975
```

TSS

```
ybar <- mean(y)
TSS <- (1/n)*sum((y - ybar)**2)
TSS
```

```
## [1] 35.18897
```

ESS

```
ESS <- (1/n)*sum((fitted_values - ybar)**2)
ESS
```

```
## [1] 29.37462
```

R2

```
R2 <- ESS/TSS
R2
```

```
## [1] 0.8347678
```

```
(cor(y, fitted_values))**2
```

```
##           [,1]
## [1,] 0.8347678
```

They are the same!

Checking with lm()

```
reg <- lm(mpg ~ hp + qsec + wt, data = mtcars)
reg
```

```
##
## Call:
## lm(formula = mpg ~ hp + qsec + wt, data = mtcars)
##
## Coefficients:
## (Intercept)           hp         qsec           wt
##    27.61053     -0.01782      0.51083     -4.35880
```

```
Coefficients
```

```
##             [,1]
## ones 27.61052686
## hp   -0.01782227
## qsec  0.51083369
## wt   -4.35879720
```

Part 3: QR Decomposition

```
qr_ols <- function(X, y) {
  QRD <- qr(X)
  Q <- qr.Q(QRD)
  R <- qr.R(QRD)
  QTy <- t(Q) %*% y
  return (backsolve(r=R, QTy, upper.tri=TRUE))
}

b <- qr_ols(X, y)
b
```

```
##             [,1]
## [1,] 27.61052686
## [2,] -0.01782227
## [3,]  0.51083369
## [4,] -4.35879720
```

```
qr.solve(X, y)
```

```
##             [,1]
## ones 27.61052686
## hp   -0.01782227
## qsec  0.51083369
## wt   -4.35879720
```

Part 4

```
centcars <- data.frame(scale(cars))
centX <- get_X(centcars, n, p)
```

```
centy <- get_y(centcars)
centbetas <- qr_ols(centX, centy)
centbetas
```

```
##               [,1]
## [1,]  2.245792e-18
## [2,] -2.027471e-01
## [3,]  1.514582e-01
## [4,] -7.076380e-01
```