Lab 4: Least Squares Reegression

Donggyun Kim
27008257

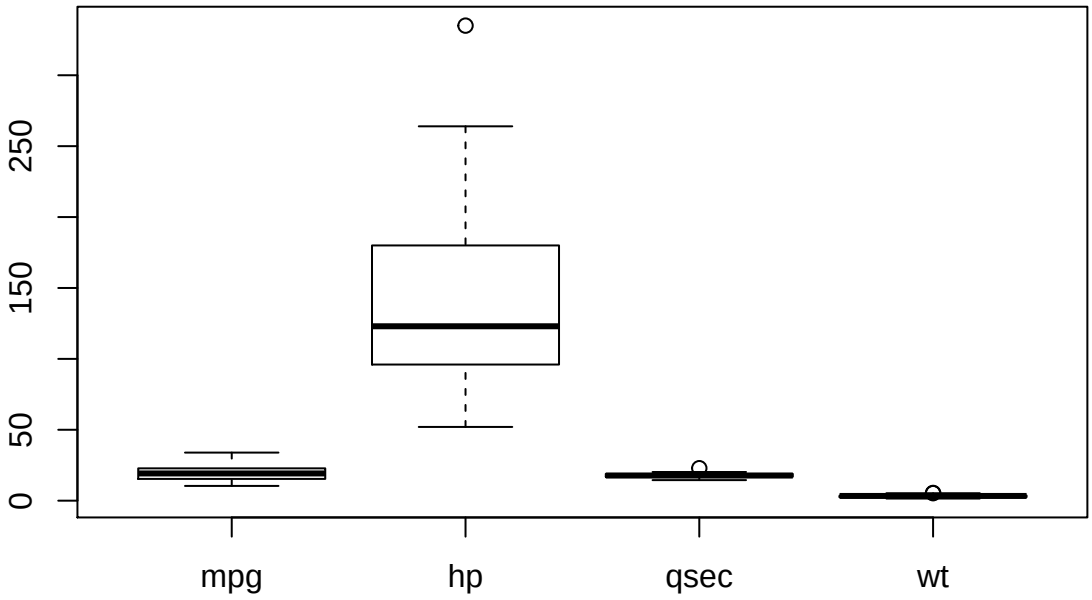
2/12/2018

Part 1: Exploratory Data Analysis (EDA)

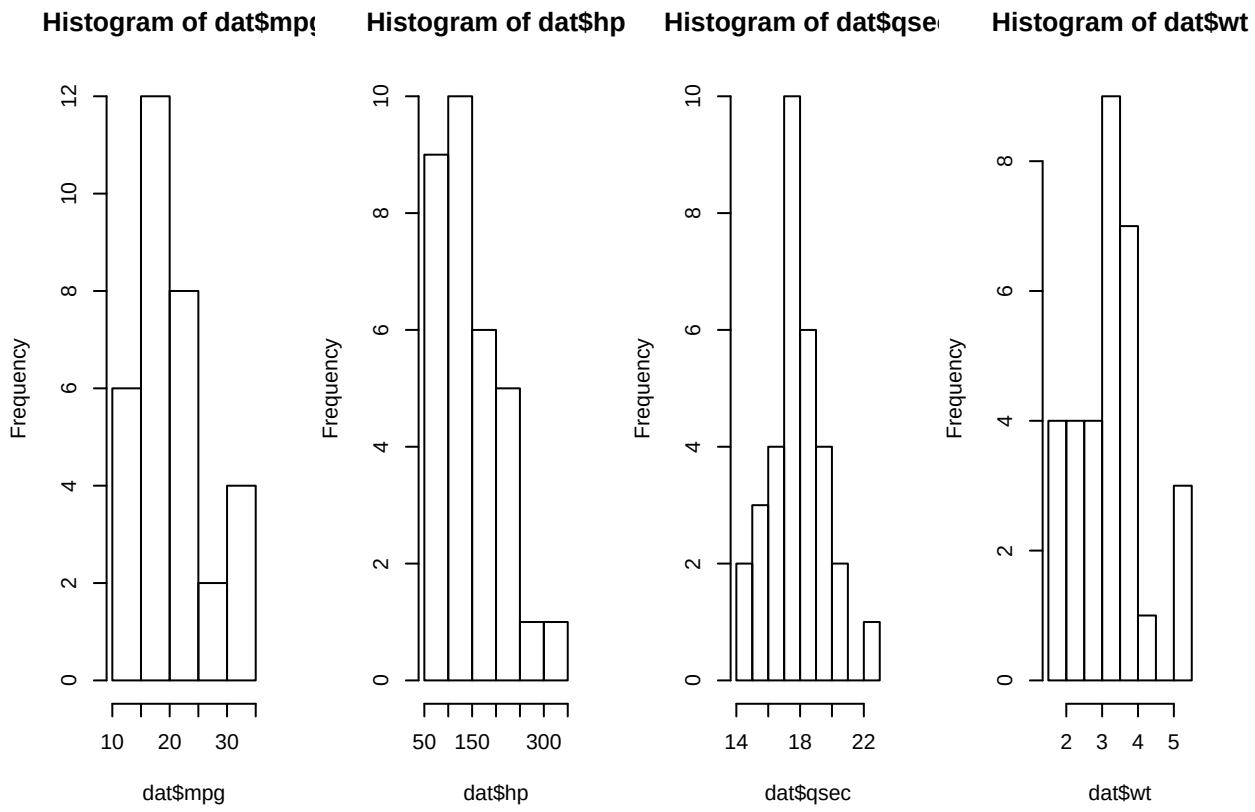
```
dat <- subset(mtcars, select = c(mpg, hp, qsec, wt))

summary(dat)
```

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

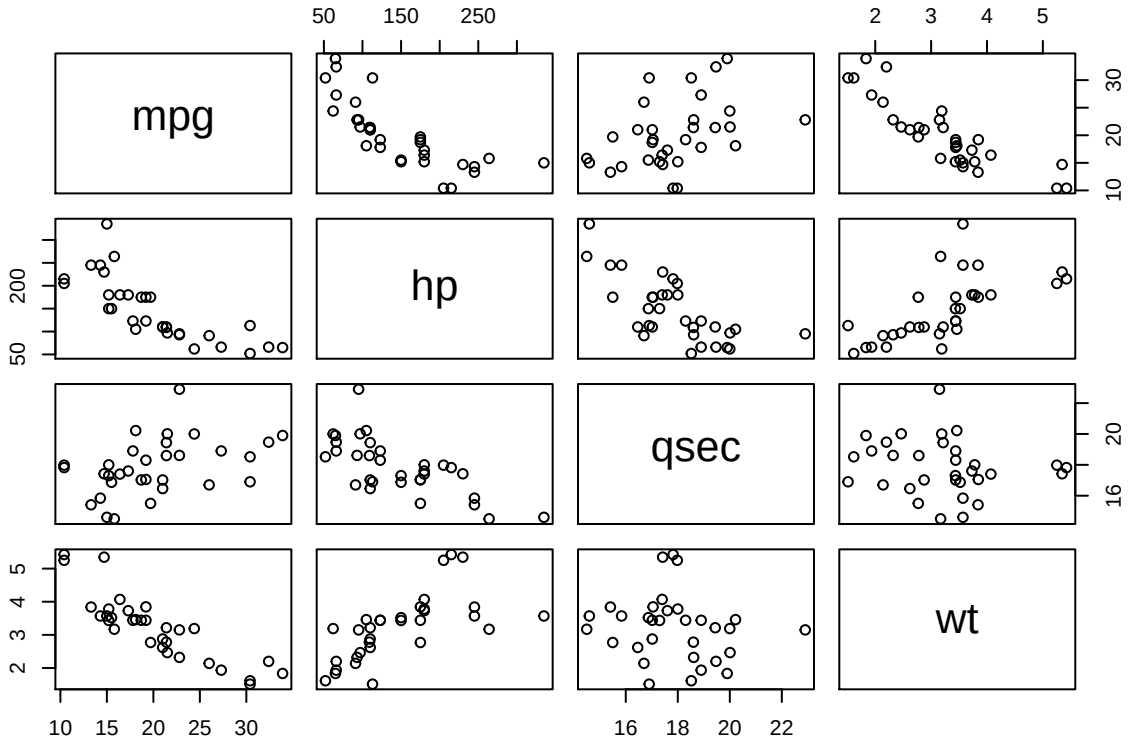
```
boxplot(dat)
```




```
par(mfrow = c(1, 4))
hist(dat$mpg)
hist(dat$hp)
hist(dat$qsec)
hist(dat$wt)
```




```
pairs(dat)
```




```
cor(dat)
```

```
##              mpg         hp       qsec         wt
```

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

```
prcomp(dat, scale. = TRUE)
```

```
## Standard deviations (1, .., p=4):
## [1] 1.6875568 0.9619487 0.3687684 0.3013580
##
## Rotation (n x k) = (4 x 4):
##             PC1        PC2         PC3        PC4
## mpg  -0.5508401  0.2685295 -0.52509223 -0.5905465
## hp    0.5539734  0.2098842 -0.76338848  0.2574880
## qsec -0.3836883 -0.7716617 -0.37558692  0.3409634
## wt    0.4924144 -0.5370091 -0.02122851 -0.6846157
```

Part 2: OLS Outputs

```
X <- as.matrix(cbind(1, dat[, 2:4]))  # explanatory variables
y <- as.matrix(subset(dat, select = mpg))  # response variable
coefficients <- solve(crossprod(X, X)) %*% crossprod(X, y)
fitted_values <- X %*% coefficients
residuals <- y - fitted_values
RSS <- sum(residuals^2)
sigma2 <- RSS / (nrow(X) - ncol(X))
TSS <- sum((y - mean(y))^2)
ESS <- sum((fitted_values - mean(y))^2)
R2 <- ESS/TSS
R2
```

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

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

```
##           mpg
## mpg 0.8347678
```

```
coefficients
```

```
##              mpg
## 1    27.61052686
## hp   -0.01782227
## qsec  0.51083369
## wt   -4.35879720
```

Part 3: QR Decomposition

$\hat{\beta} = (X^t X)^{-1} X^t y \implies X^t X \hat{\beta} = X^t y \implies R^t R \hat{\beta} = R^t Q^t y \implies R\hat{\beta} = Q^t y$

```
qr_ols <- function(M, v) {
  QR <- qr(M)
  Q <- qr.Q(QR)
```

```
  R <- qr.R(QR)
  V <- crossprod(Q, v)
  backsolve(R, V)
}

qr_ols(X, y)
```

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

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

```
##               mpg
## 1    27.61052686
## hp   -0.01782227
## qsec  0.51083369
## wt   -4.35879720
```

Part 4

```
Xc <- cbind(1, (sweep(X[, 2:4], 2, apply(X[, 2:4], 2, mean), "-")))  # mean-centered data

Xc_coef <- qr_ols(Xc, y)
Xc_coef
```

```
##              [,1]
## [1,] 20.09062500
## [2,] -0.01782227
## [3,]  0.51083369
## [4,] -4.35879720
```

```
X_colmean <- apply(X[, 2:4], 2, mean)
coef <- Xc_coef[2:4]
Xc_coef[1,1] <- Xc_coef[1,1] - sum(X_colmean * coef)
Xc_coef
```

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

```
qr_ols(X, y)
```

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

```
Xsd <- cbind(1, (sweep(Xc[, 2:4], 2, apply(X[, 2:4], 2, sd), "/")))  # standardized data
```

```r
Xsd_coef <- qr_ols(Xsd, y)
Xsd_coef
```

```
##             [,1]
## [1,] 20.0906250
## [2,] -1.2219461
## [3,]  0.9128308
## [4,] -4.2648976
```

```r
X_colsd <- apply(X[, 2:4], 2, sd)
coef <- Xsd_coef[2:4]
Xsd_coef[2:4, 1] <- coef / X_colsd
Xsd_coef[1, 1] <- Xsd_coef[1,1] - sum(X_colmean * Xsd_coef[2:4])
Xsd_coef
```

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

```r
qr_ols(X, y)
```

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

Part 5: Handling Categorical Variables

```r
gear4 <- rep(0, nrow(mtcars))
gear4[mtcars$gear == 4] <- 1
gear5 <- rep(0, nrow(mtcars))
gear5[mtcars$gear == 5] <- 1
X <- cbind(X, gear4, gear5)
qr_ols(X, y)
```

```
##             [,1]
## [1,] 24.1332699
## [2,] -0.0216439
## [3,]  0.5676117
## [4,] -3.6624819
## [5,]  1.1559309
## [6,]  2.2446800
```

```r
lm(mpg ~ hp + qsec + wt + factor(gear), data = mtcars)
```

```
##
## Call:
## lm(formula = mpg ~ hp + qsec + wt + factor(gear), data = mtcars)
##
## Coefficients:
##   (Intercept)             hp           qsec             wt  factor(gear)4
```

```
##      24.13327       -0.02164        0.56761       -3.66248        1.15593
## factor(gear)5
##       2.24468
```