# Homework Chapter 6

*Jinhong Du 15338039*

**6.23 (Calculus needed.) Consider the multiple regression model:**

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i \qquad i = 1, \cdots, n$$

**where the $\epsilon_i$ are uncorrelated, with $\mathbb{E}\{\epsilon\} = 0$ and $\sigma^2\{\epsilon\} = \sigma^2$**

**a. State the least squares criterion and derive the least squares estimators of $\beta_1$ and $\beta_2$.**

$$\begin{aligned} b_1, b_2 &= \arg\min_{b_1,b_2} \sum_{i=1}^n (Y_i - b_1 X_{i1} - b_2 X_{i2})^2 \\ &= \arg\min_{b_1,b_2} Q \end{aligned}$$

Let

$$\begin{cases} \dfrac{\partial Q}{\partial b_1} &= -2\sum\limits_{i=1}^n X_{i1}(Y_i - b_1 X_{i1} - b_2 X_{i2}) = 0 \\ \dfrac{\partial Q}{\partial b_2} &= -2\sum\limits_{i=1}^n X_{i2}(Y_i - b_1 X_{i1} - b_2 X_{i2}) = 0 \end{cases}$$

We have

$$\begin{cases} b_1 = \dfrac{\left(\sum\limits_{i=1}^n X_{i2}^2\right)\left(\sum\limits_{i=1}^n X_{i1}Y_i\right) - \left(\sum\limits_{i=1}^n X_{i2}Y_i\right)\left(\sum\limits_{i=1}^n X_{i1}X_{i2}\right)}{\left(\sum\limits_{i=1}^n X_{i1}^2\right)\left(\sum\limits_{i=1}^n X_{i2}^2\right) - \left(\sum\limits_{i=1}^n X_{i1}X_{i2}\right)^2} \\ b_2 = \dfrac{\left(\sum\limits_{i=1}^n X_{i1}^2\right)\left(\sum\limits_{i=1}^n X_{i2}Y_i\right) - \left(\sum\limits_{i=1}^n X_{i1}Y_i\right)\left(\sum\limits_{i=1}^n X_{i1}X_{i2}\right)}{\left(\sum\limits_{i=1}^n X_{i1}^2\right)\left(\sum\limits_{i=1}^n X_{i2}^2\right) - \left(\sum\limits_{i=1}^n X_{i1}X_{i2}\right)^2} \end{cases}$$

**b. Assuming that the $\epsilon_i$ are independent normal random variables, state the likelihood function and obtain the maximum likelihood estimators of $\beta_1$ and $\beta_2$. Are these the same as the least squares estimators?**

Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$X = \begin{pmatrix} b_1 X_{11} + b_2 X_{12} \\ \vdots \\ b_1 X_{n1} + b_2 X_{n2} \end{pmatrix}$$

$\because$

$$\epsilon_1, \cdots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

$\therefore$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim N(X, \sigma^2 I)$$

$$
\begin{aligned}
f_Y(Y) &= \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} e^{-\frac{1}{2\sigma^2}(Y-X)^T(Y-X)} \\
\ln f_Y(Y) &= -\frac{n}{2}\ln(2\pi\sigma) - \frac{1}{2\sigma^2}(Y-X)^T(Y-X) \\
&= -\frac{n}{2}\ln(2\pi\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - b_1X_{i1} - b_2X_{i2})^2
\end{aligned}
$$

Let

$$
\begin{cases}
\dfrac{\partial}{\partial b_1}\ln f_Y(Y) = \frac{1}{\sigma^2}\sum\limits_{i=1}^{n} X_{i1}(Y_i - b_1X_{i1} - b_2X_{i2}) = 0 \\
\dfrac{\partial}{\partial b_2}\ln f_Y(Y) = \frac{1}{\sigma^2}\sum\limits_{i=1}^{n} X_{i2}(Y_i - b_1X_{i1} - b_2X_{i2}) = 0
\end{cases}
$$

then get the same estimator

$$
\begin{cases}
b_1 = \dfrac{\left(\sum\limits_{i=1}^{n} X_{i2}^2\right)\left(\sum\limits_{i=1}^{n} X_{i1}Y_i\right) - \left(\sum\limits_{i=1}^{n} X_{i2}Y_i\right)\left(\sum\limits_{i=1}^{n} X_{i1}X_{i2}\right)}{\left(\sum\limits_{i=1}^{n} X_{i1}^2\right)\left(\sum\limits_{i=1}^{n} X_{i2}^2\right) - \left(\sum\limits_{i=1}^{n} X_{i1}X_{i2}\right)^2} \\
b_2 = \dfrac{\left(\sum\limits_{i=1}^{n} X_{i1}^2\right)\left(\sum\limits_{i=1}^{n} X_{i2}Y_i\right) - \left(\sum\limits_{i=1}^{n} X_{i1}Y_i\right)\left(\sum\limits_{i=1}^{n} X_{i1}X_{i2}\right)}{\left(\sum\limits_{i=1}^{n} X_{i1}^2\right)\left(\sum\limits_{i=1}^{n} X_{i2}^2\right) - \left(\sum\limits_{i=1}^{n} X_{i1}X_{i2}\right)^2}
\end{cases}
$$

**6.24 (Calculus needed.) Consider the multiple regression model:**

$$
Y_i = \beta_0 + \beta_1X_{i1} + \beta_2X_{i1}^2 + \beta_3X_{i2} + \epsilon_i \qquad i = 1, \cdots, n
$$

**where the $\epsilon_i$ are independent $N(0, \sigma^2)$.**

**a. State the least squares criterion and derive the least squares normal equations.**

$$
\begin{aligned}
b_1, b_2, b_3 &= \underset{b_1, b_2, b_3}{\arg\min} \sum_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i1}^2 - b_3X_{i2})^2 \\
&= \underset{b_1, b_2, b_3}{\arg\min}\, Q
\end{aligned}
$$

The least squares normal equations are

$$
\begin{cases}
\dfrac{\partial Q}{\partial b_0} & = -2\sum\limits_{i=1}^{n}(Y_i - b_0 - b_1X_{i1} - b_2X_{i1}^2 - b_3X_{i2}) = 0 \\
\dfrac{\partial Q}{\partial b_1} & = -2\sum\limits_{i=1}^{n} X_{i1}(Y_i - b_0 - b_1X_{i1} - b_2X_{i1}^2 - b_3X_{i2}) = 0 \\
\dfrac{\partial Q}{\partial b_2} & = -2\sum\limits_{i=1}^{n} X_{i1}^2(Y_i - b_0 - b_1X_{i1} - b_2X_{i1}^2 - b_3X_{i2}) = 0 \\
\dfrac{\partial Q}{\partial b_3} & = -2\sum\limits_{i=1}^{n} X_{i2}(Y_i - b_0 - b_1X_{i1} - b_2X_{i1}^2 - b_3X_{i2}) = 0
\end{cases}
$$

**b. State the likelihood function and explain why the maximum likelihood estimators will be the same as the least squares estimators.**

Let

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$$

$$X = \begin{pmatrix} b_0 + b_1 X_{11} + b_2 X_{11}^2 + b_3 X_{12} \\ \vdots \\ b_0 + b_1 X_{n1} + b_2 X_{n1}^2 + b_3 X_{n2} \end{pmatrix}$$

$\because$

$$\epsilon_1, \cdots, \epsilon_n \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

$\therefore$

$$Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \sim N(X, \sigma^2 I)$$

$$\begin{aligned} f_Y(Y) &= \frac{1}{\sqrt{(2\pi)^n \sigma^{2n}}} e^{-\frac{1}{2\sigma^2}(Y-X)^T(Y-X)} \\ \ln f_Y(Y) &= -\frac{n}{2}\ln(2\pi\sigma) - \frac{1}{2\sigma^2}(Y-X)^T(Y-X) \\ &= -\frac{n}{2}\ln(2\pi\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{i1} - b_2 X_{i1}^2 - b_3 X_{i2})^2 \end{aligned}$$

The least square normal equations are

$$\begin{cases} \frac{\partial}{\partial b_0} \ln f_Y(Y) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_{i1} - b_2 X_{i1}^2 - b_3 X_{i2}) = 0 \\ \frac{\partial}{\partial b_1} \ln f_Y(Y) = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_{i1}(Y_i - b_0 - b_1 X_{i1} - b_2 X_{i1}^2 - b_3 X_{i2}) = 0 \\ \frac{\partial}{\partial b_2} \ln f_Y(Y) = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_{i1}^2(Y_i - b_0 - b_1 X_{i1} - b_2 X_{i1}^2 - b_3 X_{i2}) = 0 \\ \frac{\partial}{\partial b_3} \ln f_Y(Y) = \frac{1}{\sigma^2}\sum_{i=1}^{n} X_{i2}(Y_i - b_0 - b_1 X_{i1} - b_2 X_{i1}^2 - b_3 X_{i2}) = 0 \end{cases}$$

**6.25 An analyst wanted to fit the regression model $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon, \quad i = 1, \ldots, n$, by the method of least squares when it is known that $\beta_2 = 4$. How can the analyst obtain the desired fit by using a multiple regression computer program?**

Let

$$\begin{aligned} Y_i' &= Y_i - \beta_2 X_{i2} \\ &= Y_i - 4X_{i2} \\ &= \beta_0 + \beta_1 X_{i1} + \beta_3 X_{i3} + \epsilon \end{aligned}$$

and fit the model

$$Y' \sim b_0 + b_1 X_1 + b_3 X_3$$

**6.26 For regression model (6.1), show that the coefficient of simple determination between $Y_i$ and $\hat{Y}_i$ equals the coefficient of multiple determination $R^2$.**

For (6.1),

$$R^2 = 1 - \frac{SSE}{SSTO}$$

Regress $Y_i$ on $\hat{Y}_i$,

$$Y_i = c_0 + c_1\hat{Y}_i$$

we have

$$\begin{aligned}
SSTO' &= Y^T\left[I - \frac{1}{n}J\right]Y \\
&= SSTO \\
\overline{\hat{Y}} &= \overline{Y} \\
c_1 &= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})(Y - \overline{Y})}{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2} \\
&= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})[(Y - \hat{Y}_i) + (\hat{Y}_i\overline{Y})]}{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2} \\
&= \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})[e_i + (\hat{Y}_i\overline{Y})]}{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2} \\
&= 1 \\
c_0 &= \overline{Y} - c_1\overline{\hat{Y}} \\
&= 0 \\
SSE' &= \sum_{i=1}^{n}(Y_i - c_0 - c_1\hat{Y})^2 \\
&= \sum_{i=1}^{n}(Y_i - \hat{Y})^2 \\
&= SSE
\end{aligned}$$

Therefore,

$$R'^2 = R^2$$

**6.27 In a small-scale regression study, the following data were obtained:**

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| $X_{i1}$ | 7 | 4 | 16 | 3 | 21 | 8 |
| $X_{i2}$ | 33 | 41 | 7 | 49 | 5 | 31 |
| $Y_i$ | 42 | 33 | 75 | 28 | 91 | 55 |

**Assume that regression model (6.1) with independent normal error terms is appropriate. Using matrix methods. obtain**

**(a)** $b$;

Let

$$
\begin{aligned}
Y &= \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\
X &= \begin{pmatrix} 1 & X_{11} & X_{12} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix} \\
\beta &= \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \\
\epsilon &= \begin{pmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{pmatrix}
\end{aligned}
$$

Then the regression model becomes

$$Y = X\beta + \epsilon$$

therefore

$$b = (X^T X)^{-1} X^T Y$$

```
data1 <- read.table("CH06PR27.txt",head=FALSE,col.names = c('Y','X1','X2'))
Y <- matrix(data1$Y)
n <- length(Y)
X <- cbind(rep(1,n),data1$X1,data1$X2)
b <- solve(crossprod(X))%*%crossprod(X,Y)
print(b)
```

```
##            [,1]
## [1,] 33.9321033
## [2,]  2.7847614
## [3,] -0.2644189
```

**(b)** $e$;

$$
\begin{aligned}
e &= Y - \hat{Y} \\
&= Y - Xb
\end{aligned}
$$

```
e <- Y-X%*%b
print(e)
```

```
##             [,1]
## [1,] -2.69960842
```

```
## [2,] -1.22997279
## [3,] -1.63735316
## [4,] -1.32985996
## [5,] -0.08999801
## [6,]  6.98679233
```

**(c)** $H$;

$$H = X(X^TX)^{-1}X^T$$

```
H <- X%*%solve(crossprod(X))%*%t(X)
print(H)
```

```
##             [,1]        [,2]        [,3]       [,4]        [,5]       [,6]
## [1,]  0.23143293  0.25167585  0.21178735  0.1488684 -0.05475543 0.21099091
## [2,]  0.25167585  0.31240459  0.09437844  0.2662773 -0.14787283 0.22313666
## [3,]  0.21178735  0.09437844  0.70442026 -0.3191744  0.10446672 0.20412159
## [4,]  0.14886839  0.26627729 -0.31917435  0.6142563  0.14143492 0.14833743
## [5,] -0.05475543 -0.14787283  0.10446672  0.1414349  0.94039955 0.01632707
## [6,]  0.21099091  0.22313666  0.20412159  0.1483374  0.01632707 0.19708635
```

**(d)** $SSR$;

$$\begin{aligned} SSR &= b^TX^TY - \frac{1}{n}Y^TJY \\ &= Y^T\left[H - \frac{1}{n}J\right]Y \end{aligned}$$

```
J <- matrix(rep(1,n*n),nrow=n,ncol=n)
SSR <- t(Y)%*%(H-J/n)%*%Y
print(SSR)
```

```
##          [,1]
## [1,] 3009.926
```

**(e)** $s^2\{b\}$;

$$\begin{aligned} s^2\{b\} &= MSE(X^TX)^{-1} \\ &= \frac{SSE}{n-p}(X^TX)^{-1} \end{aligned}$$

```
p <- 3
SSE <- crossprod(e)[1]
MSE <- SSE/(n-p)
s2b <- MSE * solve(crossprod(X))
print(s2b)
```

```
##           [,1]        [,2]        [,3]
## [1,] 715.47114 -34.1589166 -13.5949371
## [2,] -34.15892   1.6616664   0.6440674
## [3,] -13.59494   0.6440674   0.2624678
```

**(f)** $\hat{Y}_h$ **when** $X_{h1} = 10, X_{h2} = 30$

```
Xh <- matrix(c(1,10,30))
Yh <- crossprod(Xh,b)
print(Yh)
```

```
##          [,1]
## [1,] 53.84715
```

**(g)** $s^2\{\hat{Y}_h\}$**, when** $X_{h1} = 10, X_{h2} = 30$**.**

$$\begin{aligned} s^2\{\hat{Y}_h\} &= MSE(X_h^T(X^TX)^{-1}X_h) \\ &= X_h^T s^2\{b\}X_h \end{aligned}$$

```
s2Yh <- t(Xh)%*%s2b%*%Xh
print(s2Yh)
```

```
##         [,1]
## [1,] 5.42462
```

**6.31 Refer to the SENIC data set in Appendix C.1.**

**a. For each geographic region, regress infection risk ($Y$) against the predictor variables age ($X_1$), routine culturing ratio ($X_2$), average daily census ($X_3$), and available facilities and services ($X_4$). Use first—order regression model (6.5) with four predictor variables. State the estimated regression functions.**

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \epsilon$$

```
data2 <- read.table("APPENC01.txt",head=FALSE,col.names = c('Identification number',
'Length of stay','X1','Y','X2','Routine chest X-ray ratio','Number of beds',
'Medical school affiliation','Region','X3','Number of nurses','X4'))
fit <- list(1,2,3,4)
for (i in c(1:4)){
  fit[[i]] <- lm('Y~X1+X2+X3+X4',data2[which(data2['Region']==i),])
  print(sprintf('Region %d: Y=%f+%fX1+%fX2+%fX3+%fX4',i,
fit[[i]]$coefficients[1],fit[[i]]$coefficients[2],
fit[[i]]$coefficients[3],fit[[i]]$coefficients[4],
fit[[i]]$coefficients[5]))
}
```

```
## [1] "Region 1: Y=-3.349576+0.116954X1+0.058240X2+0.001508X3+0.006613X4"
## [1] "Region 2: Y=2.291536+0.004742X1+0.058030X2+0.001172X3+0.015018X4"
## [1] "Region 3: Y=-0.143858+0.030848X1+0.102281X2+0.004114X3+0.008039X4"
## [1] "Region 4: Y=1.566549+0.035241X1+0.040328X2+-0.000664X3+0.012792X4"
```

**b. Are the estimated regression functions similar for the four regions? Discuss.**

The regression functions for four regions are different.

**c. Calculate $MSE$ and $R^2$ for each region. Are these measures similar for the four regions? Discuss.**

```
print('              MSE         R2')
```

```
## [1] "              MSE         R2"
```

```
for (i in c(1:4)){
  Y <- data2[which(data2['Region']==i),'Y']
  n <- length(Y)
  Yh <-fit[[i]]$fitted.values
  MSE <- crossprod(Yh-Y)[1]/(n-5)
  print(sprintf('Region %d: %f    %f',i,MSE,summary.lm(fit[[i]])$r.squared))
}
```

```
## [1] "Region 1: 1.021771    0.461323"
## [1] "Region 2: 1.211917    0.411466"
## [1] "Region 3: 0.936720    0.608849"
## [1] "Region 4: 0.953804    0.089595"
```

These measures are not similar for the four regions.