Homework Chapter 8

Jinhong Du 15338039

8.31.

a. Derive the expressions for b'_0, b'_1 and b'_{11} in (8.12). $\because$

$$\begin{aligned}\hat{Y} &= b'_0 + b'_1 X' + b'_{11} X^2\\ &= b'_0 + b'_1 (X - \overline{X}) + b'_{11}(X - \overline{X})^2\\ &= b'_0 + b'_1 X - b_1 \overline{X} + b'_{11} X^2 - 2b'_{11} X \overline{X} + b'_{11} \overline{X}^2\\ &= (b'_0 - b'_1 \overline{X} + b'_{11} \overline{X}^2) + (b'_1 X - 2b'_{11} \overline{X} X) + b'_{11} X^2\end{aligned}$$

$\therefore$

$$\begin{cases} b'_0 - b'_1 \overline{X} + b'_{11} \overline{X}^2 = b_0\\ b'_1 X - 2b'_{11} \overline{X} X = b_1\\ b'_{11} = b_{11}\end{cases}$$

$\therefore$

$$\begin{cases} b'_0 = b_0 - b_1 \overline{X}\\ b'_1 = b_1 - 2b_{11} \overline{X}\\ b'_{11} = b_{11}\end{cases}$$

b. Using (5.46). obtain the variance-covariance matrix for the regression coefficients pertaining to the original X variable in terms of the variance-covariance matrix for the regresSion coefficients penaining to the transformed x variable.

$\because$ the transformed matrix is

$$A = \begin{bmatrix} 1 & -\overline{X} & \overline{X}^2\\ 0 & 1 & -2\overline{X}\\ 0 & 0 & 1\end{bmatrix}$$

$$\sigma^2\{b'\} = A\sigma^2\{b\}A^T$$

$\therefore$

$$\begin{cases} Var\{b'_0\} = Var\{b_0\} - 2\overline{X}Cov\{b_0, b_1\} + \overline{X}^2 Var\{b_1\} - 2\overline{X}^3 Cov\{b_1, b_{11}\} + \overline{X}^4 Var\{b_{11}\}\\ Var\{b'_1\} = Var\{b_1\} - 4\overline{X}Cov\{b_1, b_{11}\} + 4\overline{X}^2 Var\{b_{11}\}\\ Var\{b_{11}\} \qquad\qquad = Var\{b_{11}\}\\ Cov\{b'_0, b'_1\} = Cov\{b_0, b_1\} - 2\overline{X}Cov\{b_0, b_{11}\} + 3\overline{X}^2 Cov\{b_1, b_{11}\} - \overline{X}Var\{b_1\} - 2\overline{X}^3 Var\{b_{11}\}\\ Cov\{b'_0, b'_{11}\} = Cov\{b_0, b_{11}\} - \overline{X}Cov\{b_1, b_{11}\} + \overline{X}^2 Var\{b_{11}\}\\ Cov\{b'_1, b'_{11}\} = Cov\{b_1, b_{11}\} - 2\overline{X}Var\{b_{11}\}\end{cases}$$

8.34. In a regression study, three types of banks were involved, namely, commercial, mutual savings, and savings and loan. Consider the following system of indicator variables for type of bank:

Type of Bank	X_2	X_3
Commercial	1	0
Mutual savings	0	1
Savings and loan	-1	-1

a. Develop a first-order linear regression model for relating last year's profit or loss (Y) to size of bank (X_1) and type of bank (X_2, X_3).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

where $Y = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}$, $X_j = \begin{pmatrix} X_{1j} \\ X_{2j} \\ \vdots \\ X_{nj} \end{pmatrix}$ $\quad j = 1, 2, 3$

b. State the response functions for the three types of banks.

$$\mathbb{E}_{\text{Commercial}} Y = \beta_0 + \beta_1 X_1 + \beta_2$$
$$\mathbb{E}_{\text{Mutual savings}} Y = \beta_0 + \beta_1 X_1 + \beta_3$$
$$\mathbb{E}_{\text{Savings and loan}} Y = \beta_0 + \beta_1 X_1 - \beta_2 - \beta_3$$

c. Interpret each of the following quantities: (1) β_2, (2) β_3, (3) $-\beta_2 - \beta_3$

(1) β_2 shows how much higher (lower) the mean response line is for the type of bank is commercial than it is not, for any given level of X_1.
(2) β_3 shows how much higher (lower) the mean response line is for the type of bank is mutual savings than it is not, for any given level of X_1.
(3) $-\beta_2 - \beta_3$ shows how much higher (lower) the mean response line is for the type of bank is savings and loan than it is not, for any given level of X_1.

8.35 Refer to regression model (8.54) and exclude variable X_3.

a Obtain the $X^T X$ matrix for this special case of a single qualitative predictor variable, for $i = 1, \cdots, n$ when n_1 firms are not incorporated.

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} n & n - n_1 \\ n - n_1 & n - n_1 \end{bmatrix}$$

$$(X^T X)^{-1} = \begin{bmatrix} \frac{1}{n_1} & -\frac{1}{n_1} \\ -\frac{1}{n-n_1} & \frac{1}{n-n_1} + \frac{1}{n_1} \end{bmatrix}$$

b Using (6.25), find b.

$$\overline{Y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_i$$

$$\overline{Y}_2 = \frac{1}{n-n_1}\sum_{i=n_1+1}^{n} Y_i$$

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

$$\begin{aligned} b &= (X^T X)^{-1} X^T Y \\ &= \begin{bmatrix} \overline{Y}_1 \\ \overline{Y}_2 - \overline{Y}_1 \end{bmatrix} \end{aligned}$$

c Using (6.35) and (6.36), find SSE and SSR.

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

$$\begin{aligned} SSE &= Y^T(I-H)Y \\ &= \sum_{i=1}^{n} Y_i^2 - n_1\overline{Y}^2 - (n-n_1)\overline{Y}_2^2 \\ SSR &= Y^T(H - \frac{1}{n}J)Y \\ &= n_1\overline{Y}_1^2 + (n-n_1)\overline{Y}_2^2 - n\overline{Y}^2 \end{aligned}$$

8.38. Refer to the SENIC data set in Appendix C.1. Second-order regression model (8.2) is to be fitted for relating number of nurses (Y) to available facilities and services (X).

a. Fit the second-order regression model. Plot the residuals against the fitted values. How well does the second-order model appear to fit the data?

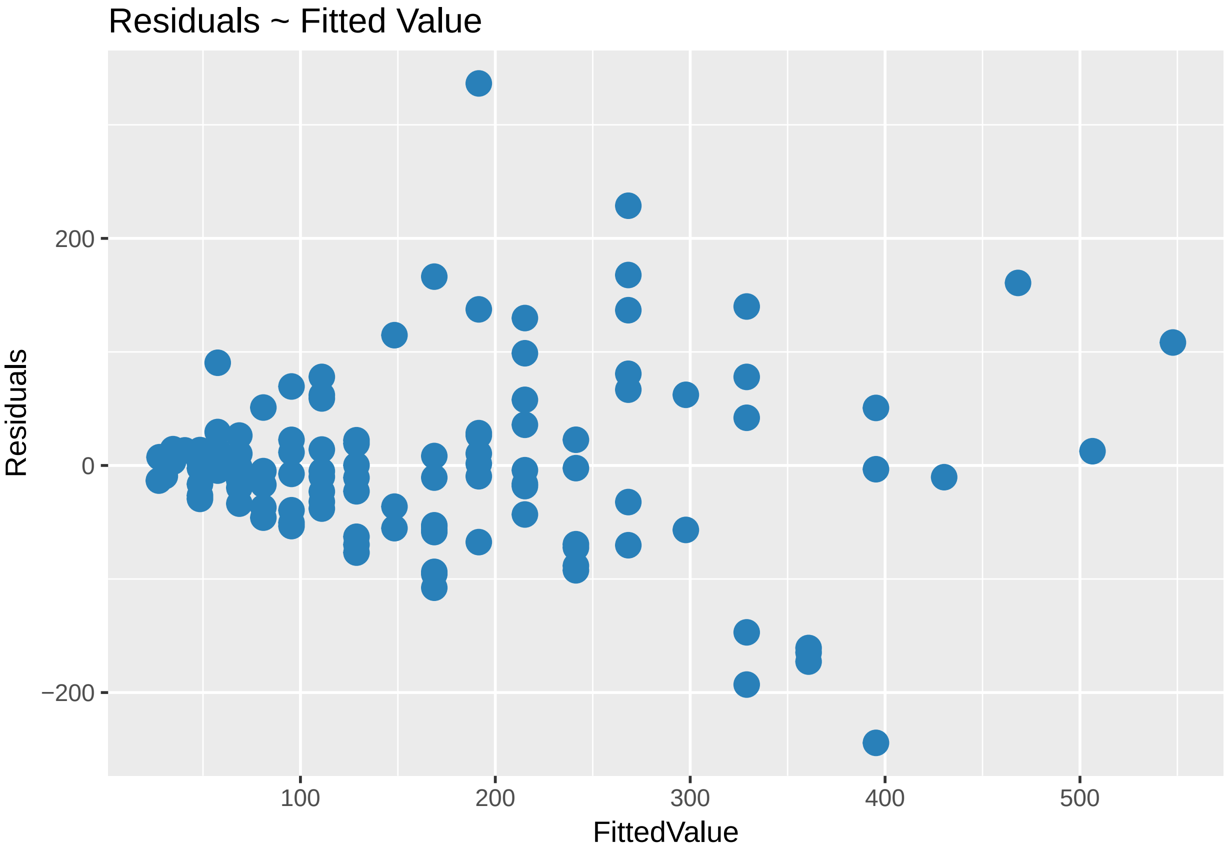
```r
data1 <- read.table("APPENC01.txt",head=FALSE,col.names = c('Identification number',
'Length of stay','Age','Infection risk',
'Routine culturing ratio','Routine chest X-ray ratio',
'Number of beds','Medical school affiliation',
'Region','Average daily census','Y','X'))
Y <- data1$Y
X <- data1$X
fit <- lm('Y~poly(X,2)')
summary(fit)
```

```
##
## Call:
## lm(formula = "Y~poly(X,2)")
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max 
## -244.32  -39.42   -4.55   26.48  336.48 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept)  173.248      7.743  22.375  < 2e-16 ***
## poly(X, 2)1 1154.767     82.308  14.030  < 2e-16 ***
## poly(X, 2)2  305.832     82.308   3.716  0.00032 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 82.31 on 110 degrees of freedom
## Multiple R-squared:  0.6569, Adjusted R-squared:  0.6507 
## F-statistic: 105.3 on 2 and 110 DF,  p-value: < 2.2e-16
```

```r
library(ggplot2)
library(gridExtra)

lm.scatter <- ggplot(data.frame(
  'FittedValue' = fit$fitted.values,
  'Residuals' = fit$residuals
  ), aes(x=FittedValue, y=Residuals)) +
  geom_point(color='#2980B9', size = 4)  +
  labs(title='Residuals ~ Fitted Value')
grid.arrange(lm.scatter)
```



It doesn't fit the data well since the absolute value of residuals tend to be larger as fitted value increases.

b. Obtain R^2 for the second-order regression model. Also obtain the coefficient of simple determination for the first-order regression model. Has the addition of the quadratic term in the regression model substantially increased the coefficient of determination?

```
fit.aov <- anova(fit)
R2 <- sum(fit.aov[1,2])/sum(fit.aov[, 2])
fit1 <- lm('Y~X')
fit1.aov <- anova(fit1)
R21 <- sum(fit1.aov[1,2])/sum(fit1.aov[, 2])
R2
```

```
## [1] 0.6569396
```

```
R21
```

```
## [1] 0.6138809
```

For the second-order regression model, $R^2 = 0.6569396$; For the first-order regression model, $R^2 = 0.6138809$.

c. Test whether the quadratic term can be dropped from the regression model; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

```
F <- ((sum(fit$residuals^2)-sum(fit1$residuals^2)
       )/(fit$df.residual-fit1$df.residual)
      )/(sum(fit$residuals^2)/fit$df.residual)
print(sprintf('F* is %f',F))
```

```
## [1] "F* is 13.806505"
```

```
print(sprintf('F(0.99,%d,%d) is %f',fit1$df.residual-fit$df.residual,
              fit$df.residual,
              qf(p=0.99,df1=fit1$df.residual-fit$df.residual,
                 df2=fit$df.residual)))
```

```
## [1] "F(0.99,1,110) is 6.871028"
```

Full model:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

Reduce model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Given α,

$$H_0 : \beta_2 = 0 \qquad H_a : \beta_2 \neq 0$$

$$F^* = \frac{\frac{SSE_F - SSE_R}{1}}{\frac{SSE_F}{n-2}} \overset{H_0}{\sim} F(1, n-3)$$

The decision rule is

If $F^* \leqslant F(1-\alpha, 1, n-3)$, then conclude H_0;

If $F^* > F(1-\alpha, 1, n-3)$, then conclude H_a;

Here, $F^* = 13.806505 > 6.871028$, therefore, conclude H_a.

8.41. Refer to the SENIC data set in Appendix C.1. Length of stay (Y) is to be regressed on age (X_1), routine culturing ratio (X_2), average daily census (X_3). available facilities and services (X_4), and region (X_5, X_6, X_7).

a. Fit a first-order regression model. Let $X_5 = 1$ if NE and 0 otherwise, $X_6 = 1$ if NC and 0 otherwise, and $X_7 = 1$ if S and 0 otherwise.

```
data1 <- read.table("APPENC01.txt",head=FALSE,col.names = c('Identification number',
'Y','X1','Infection risk','X2',
'Routine chest X-ray ratio','Number of beds',
'Medical school affiliation','Region','X3',
'Number of nurses','X4'))
Y <- data1$Y
X1 <- data1$X1
X2 <- data1$X2
X3 <- data1$X3
X4 <- data1$X4
Region <- data1$Region
fit <- lm('Y~X1+X2+X3+X4+I(Region==1)+I(Region==2)+I(Region==3)')
summary(fit)
```

```
##
## Call:
## lm(formula = "Y~X1+X2+X3+X4+I(Region==1)+I(Region==2)+I(Region==3)")
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7938 -0.7304  0.0037  0.5388  7.7231
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           2.047830   1.812955   1.130  0.26124
## X1                    0.103691   0.031459   3.296  0.00134 **
## X2                    0.040302   0.014303   2.818  0.00578 **
## X3                    0.006600   0.001404   4.700 7.92e-06 ***
## X4                   -0.020761   0.014369  -1.445  0.15148
## I(Region == 1)TRUE  2.149988   0.461517   4.659 9.37e-06 ***
## I(Region == 2)TRUE  1.190333   0.437058   2.724  0.00757 **
## I(Region == 3)TRUE  0.633478   0.427554   1.482  0.14143
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.399 on 105 degrees of freedom
## Multiple R-squared:  0.4981, Adjusted R-squared:  0.4647
## F-statistic: 14.89 on 7 and 105 DF,  p-value: 2.283e-13
```

$$\begin{aligned} Y = 2.047830 + 0.103691X_1 + 0.040302X^2 + 0.006600X_3 \\ - 0.020761X_4 + 2.149988X_5 + 1.190333X_6 + 0.633478X_7 \end{aligned}$$

b. Test whether the routine culturing ratio can be dropped from the model; use a level of significance of .05. State the alternatives, decision rule, and conclusion.

```r
fit1 <- lm('Y~X1+X3+X4+I(Region==1)+I(Region==2)+I(Region==3)')
F <- ((sum(fit$residuals^2)-sum(fit1$residuals^2)
       )/(fit$df.residual-fit1$df.residual)
      )/(sum(fit$residuals^2)/fit$df.residual)
print(sprintf('F* is %f',F))
```

```
## [1] "F* is 7.939198"
```

```r
print(sprintf('F(0.95,%d,%d) is %f',fit1$df.residual-fit$df.residual,
              fit$df.residual,
              qf(p=0.95,df1=fit1$df.residual-fit$df.residual,
                 df2=fit$df.residual)))
```

```
## [1] "F(0.95,1,105) is 3.931556"
```

Full model:

$$Y_i = \beta_0 + \sum_{i=1}^{7} \beta_i X_i + \epsilon_i$$

Reduce model:

$$Y_i = \beta_0 + \sum_{\substack{i=1 \\ i\neq 2}}^{7} \beta_i X_i + \epsilon_i$$

Given α,

$$H_0 : \beta_2 = 0 \qquad H_a : \beta_2 \neq 0$$

$$F^* = \frac{\frac{SSE_F - SSE_R}{1}}{\frac{SSE_F}{n-8}} \overset{H_0}{\sim} F(1, n-8)$$

The decision rule is

If $F^* \leqslant F(1-\alpha, 1, n-8)$, then conclude H_0;

If $F^* > F(1-\alpha, 1, n-8)$, then conclude H_a;

Here, $F^* = 7.939198 > 3.931556$, therefore, conclude H_a.

c. Examine whether the effect on length of stay for hospitals located in the western region differs from that for hospitals located in the other three regions by constructing an appropriate confidence interval for each pairwise comparison. Use the Bonferroni procedure with a 95 percent family confidence coefficient. Summarize your findings.

```r
n = length(Y)
b <- as.matrix(fit$coefficients)
X <- cbind(rep(1,n*1),X1,X2,X3,X4,I(Region==1),I(Region==2),I(Region==3))
res <- as.matrix(fit$residuals)
J = matrix(rep(1,n*n),nrow=n,ncol=n)
SSR <- t(b)%*%crossprod(X,Y) - t(Y)%*%J%*%Y/n
SSE <- crossprod(Y) - t(b)%*%crossprod(X,Y)
MSE <- SSE/fit$df.residual
s2b <- solve(crossprod(X)) * MSE[1,1]
print(sprintf('sb5=%f',sqrt(s2b[6,6])))
```

```
## [1] "sb5=0.461517"
```

```r
print(sprintf('sb6=%f',sqrt(s2b[7,7])))
```

```
## [1] "sb6=0.437058"
```

```r
print(sprintf('sb7=%f',sqrt(s2b[8,8])))
```

```
## [1] "sb7=0.427554"
```

```r
B <- qt(1-0.05/(2 * 3), fit$df.residual)
print(sprintf("B = %f",B))
```

```
## [1] "B = 2.432940"
```

```r
print(sprintf("The confidence interval for beta5 is (%f,%f)",
              b[6]-B*sqrt(s2b[6,6]),b[6]+B*sqrt(s2b[6,6])))
```

```
## [1] "The confidence interval for beta5 is (1.027146,3.272831)"
```

```r
print(sprintf("The confidence interval for beta6 is (%f,%f)",
              b[7]-B*sqrt(s2b[7,7]),b[7]+B*sqrt(s2b[7,7])))
```

```
## [1] "The confidence interval for beta6 is (0.126998,2.253667)"
```

```r
print(sprintf("The confidence interval for beta7 is (%f,%f)",
              b[8]-B*sqrt(s2b[8,8]),b[8]+B*sqrt(s2b[8,8])))
```

```
## [1] "The confidence interval for beta7 is (-0.406736,1.673692)"
```

β_5, β_6 and β_7 tends to be positive, which means that whether it is located in region NE or NC or S have effect to Y. $\beta_6 = \beta_6 = \beta_7 = 0$ indicates the region is W. Therefore, the effect on length of stay for hospitals located in the western region differs from that for hospitals located in the other three regions.