Homework Chapter 1

Jinhong Du 15338039

1 Under the linear regression model (1.1) with error distribution unspecified (in which the errors have expectation zero and are uncorrelated and have equal variances σ^2), calculate

(1) the expectations of random variables SS_{YY} and SS_{XY}

$\because$ $\epsilon_1, \cdots, \epsilon_n$ i.i.d., $E\epsilon_i = 0$, $Var\epsilon_i = \sigma^2$ and

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

$\therefore$

$$\begin{aligned} EY_i &= \beta_0 + \beta_1 X_i \\ VarY_i &= \sigma^2 \end{aligned}$$

$\therefore$

$$\begin{aligned} EY_i^2 &= VarY_i + (EY_i)^2 \\ &= (\beta_0 + \beta_1 X_i)^2 + \sigma^2 \\ &= \beta_0^2 + 2\beta_0\beta_1 X_i + \beta_1^2 X_i^2 + \sigma^2 \\ E\overline{Y} &= \beta_0 + \beta_1 \overline{X} \\ Var\overline{Y} &= \frac{\sigma^2}{n} \\ E\overline{Y}^2 &= Var\overline{Y} + (E\overline{Y})^2 \\ &= \beta_0^2 + 2\beta_0\beta_1\overline{X} + \beta_1^2\overline{X}^2 + \frac{\sigma^2}{n} \end{aligned}$$

$\because$

$$\begin{aligned} SS_{YY} &= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \\ SS_{XY} &= \sum_{i=1}^{n} X_i Y_i - n\overline{XY} \end{aligned}$$

$\therefore$

$$
\begin{aligned}
ESS_{YY} &= \sum_{i=1}^{n} EY_i^2 - nE\overline{Y}^2 \\
&= \sum_{i=1}^{n}(\beta_0^2 + 2\beta_0\beta_1 X_i + \beta_1^2 X_i^2 + \sigma^2) - n(\beta_0^2 + 2\beta_0\beta_1\overline{X} + \beta_1^2\overline{X}^2 + \frac{\sigma^2}{n}) \\
&= \sum_{i=1}^{n}[2\beta_0\beta_1(X_i - \overline{X}) + \beta_1^2(X_i^2 - \overline{X}^2)] + (n-1)\sigma^2 \\
&= \beta_1^2(\sum_{i=1}^{n} X_i^2 - n\overline{X}) + (n-1)\sigma^2 \\
ESS_{XY} &= \sum_{i=1}^{n} X_i EY_i - n\overline{X}E\overline{Y} \\
&= \sum_{i=1}^{n}(\beta_0 + \beta_1 X_i)X_i - n\overline{X}(\beta_0 + \beta_1\overline{X}) \\
&= \sum_{i=1}^{n}[\beta_0(X_i - \overline{X}) + \beta_1(X_i^2 - \overline{X}^2)] \\
&= \beta_1(\sum_{i=1}^{n} X_i^2 - n\overline{X})
\end{aligned}
$$

(2) $cov(e_i, e_j), i \neq j.$

$$
\begin{aligned}
cov(e_i, e_j) &= cov(Y_i - \overline{Y}, Y_i - \hat{Y}) \\
&= cov(\beta_0 + \beta_1 X_i + \epsilon_i - \hat{Y}, \beta_0 + \beta_1 X_j + \epsilon_j - \hat{Y}) \\
&= cov(\epsilon_i, \epsilon_j) \\
&= 0
\end{aligned}
$$

1.21 Airfreight breakage. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1,000 ampules. The data below, involving 10 shipments, were collected on the number of times the carton was transferred from one aircraft to another over the shipment route (X) and the number of ampules found to be broken upon arrival (Y). Assume that first-order regression model (1.1) is appropriate.

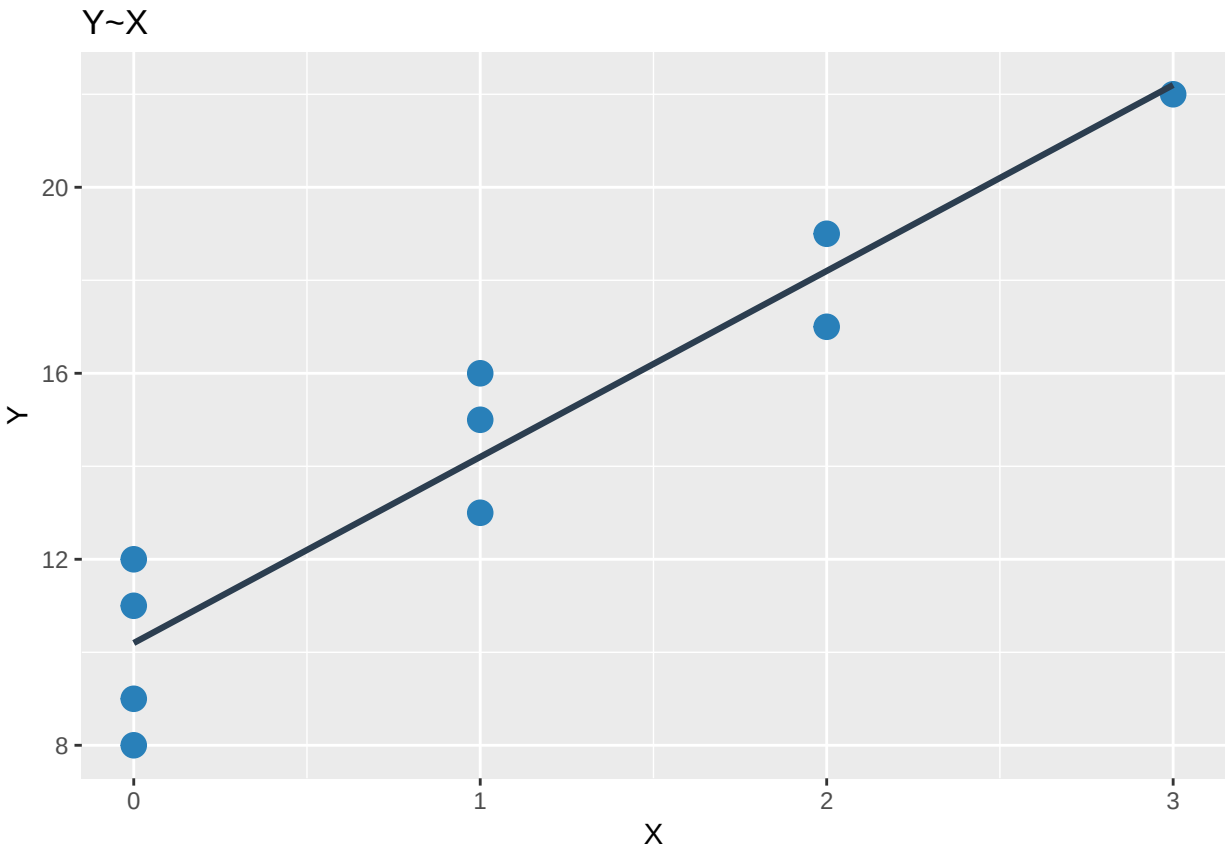
(a) Obtain the estimated regression function. Plot the estimated regression function and the data. Does a linear regression function appear to give a good fit here?

```
library(ggplot2)
library(gridExtra)
X <- c(1, 0, 2, 0, 3, 1, 0, 1, 2, 0)
Y <- c(16, 9, 17, 12, 22, 13, 8, 15, 19, 11)
data1 <- data.frame(X,Y)
fit <- lm('Y~X',data1)
summary(fit)
```

```
##
## Call:
## lm(formula = "Y~X", data = data1)
```

```
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -2.2   -1.2    0.3    0.8    1.8
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  10.2000     0.6633  15.377 3.18e-07 ***
## X             4.0000     0.4690   8.528 2.75e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.483 on 8 degrees of freedom
## Multiple R-squared:  0.9009, Adjusted R-squared:  0.8885
## F-statistic: 72.73 on 1 and 8 DF,  p-value: 2.749e-05
```

```
lm.scatter <- ggplot(data1, aes(x=X, y=Y)) +
  geom_point(color='#2980B9', size = 4) +
  geom_smooth(method = lm, se=FALSE, fullrange=TRUE, color='#2C3E50', size=1.1) +
  labs(title='Y~X')
grid.arrange(lm.scatter)
```



The linear regression function is

$$\hat{Y} = 10.20 + 4X$$

It seems like a good fit.

(b) Obtain a point estimate of the expected number of broken ampules when $X = 1$ transfer is made.

When $X = 1$, $Y_1 = 10.2 + 4 \times 1 = 14.2$

(c) Estimate the increase in the expected number of ampules broken when there are 2 transfers as compared to 1 transfer.

When $X = 2$, $Y_2 = 10.2 + 4 \times 2 = 18.2$

Therefore $Y_2 - Y_1 = 4$

(d) Verify that your fitted regression line goes through the point $(\overline{X}, \overline{Y})$.

$\overline{X} = 1$, $\overline{Y} = 14.2$ lies in the regression line

1.33 (Calculus needed.) Refer to the regression model, $Y_i = \beta_0 + \epsilon_i$ in Exercise 1.30. Derive the least squares estimator of β_0 for this model.

When $\beta_1 = 0$,

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0)^2$$

Let

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0) = 0$$

We get

$$\hat{\beta}_0 = \overline{Y}$$

1.34 Prove that the least squares estimator of β_0 obtained in Exercise 1.33 is unbiased.

$\because$

$$\begin{aligned} E\hat{\beta}_0 &= E\overline{Y} \\ &= \frac{1}{n} \sum_{i=1}^{n} E(\beta_0 + \epsilon_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \beta_0 \\ &= \beta_0 \end{aligned}$$

$\therefore$ $\hat{\beta}_0$ is the UE of β

1.39 Two observations on Y were obtained at each of three X levels, namely, at $X = 5$, $X = 10$, and $X = 15$.

(a) Show that the least squares regression line fitted to the three points $(5, \overline{Y}_1)$, $(10, \overline{Y}_2)$, and $(15, \overline{Y}_3)$, where $\overline{Y}_1, \overline{Y}_2$ and $\overline{Y}_3$ denote the means of the Y observations at the three X levels is identical to the least squares regression line fitted to the original six cases.

For

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

let

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = -2 \sum\limits_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = -2 \sum\limits_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i) = 0 \end{cases}$$

we have

$$\begin{cases} \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \\ \hat{\beta}_1 = \dfrac{SS_{XY}}{SS_{XX}} \end{cases}$$

For

$$Q' = \sum_{i=1}^{3} (\overline{Y_i} - \beta_0 - \beta_1 \overline{X_i})^2$$

let

$$\begin{cases} \dfrac{\partial Q'}{\partial \beta_0} = -2 \sum\limits_{i=1}^{3} (\overline{Y_i} - \beta_0 - \beta_1 \overline{X_i}) = 0 \\ \dfrac{\partial Q'}{\partial \beta_1} = -2 \sum\limits_{i=1}^{3} \overline{X_i} (\overline{Y_i} - \beta_0 - \beta_1 \overline{X_i}) = 0 \end{cases}$$

we have

$$\begin{cases} \hat{\beta}_0{}' = \overline{Y} - \hat{\beta}_1{}' \overline{X} \\ \hat{\beta}_1{}' = \dfrac{SS_{XY}}{SS_{XX}} \end{cases}$$

Therefore the two fits are the same.

(b) In this study, could the error term variance σ^2 be estimated without fitting a regression line? Explain.

$$\begin{aligned} MSE &= \frac{SSE}{3-2} \\ &= \sum_{i=1}^{3} \left[\frac{SS_{XY}}{SS_{XX}} (\overline{X}_i - \overline{X}) + \overline{Y} - \overline{Y}_i \right]^2 \\ &= \left(-5 \frac{SS_{XY}}{SS_{XX}} + \overline{Y} - \overline{Y}_1 \right)^2 \\ &\quad + \left(\overline{Y} - \overline{Y}_1 \right)^2 \\ &\quad + \left(5 \frac{SS_{XY}}{SS_{XX}} + \overline{Y} - \overline{Y}_3 \right)^2 \end{aligned}$$

It is only relevant to 6 sample points. Therefore we can estimate σ^2 without fitting a regression line.

From vairance analysis, we divide the data set into 3 groups such that in every group X_i have the same level. So the variance can be estimate unbiasedly by

$$\underbrace{SSTO}_{df=5} = \underbrace{SSW}_{3} + \underbrace{SSB}_{2}$$

$$\begin{aligned}\hat{\sigma^2} &= \frac{SSW}{3} \\ &= \frac{1}{3}\sum_{i=1}^{3}\sum_{j=1}^{2}(Y_{ij} - \overline{Y}_i)^2\end{aligned}$$

1.41 (Calculus needed.) Refer to the regression model $Y_i = \beta_1 X_i + \epsilon_i,\ i = 1, \cdots, n$, in Exercise 1.29.

(a) Find the least Squares estimator of β_1.

For

$$Q = \sum_{i=1}^{n}(Y_i - \beta_1 X_i)^2$$

let

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i(Y_i - \beta_1 X_i) = 0$$

we have

$$\hat{\beta}_{1LS} = \frac{\sum\limits_{i=1}^{n} X_i Y_i}{\sum\limits_{i=1}^{n} X_i^2}$$

(b) Assume that the error terms ϵ_i are independent $N(0, \sigma^2)$ and that σ^2 is known. State the likelihood function for the n sample observations on Y and obtain the maximum likelihood estimator of β_1. Is it the same as the least squares estimator?

$$\begin{aligned}L(\beta_1; x_0, x_2, \cdots, x_n) &= \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{1}{2\sigma^2}(y_i - \beta_1 x_i)^2} \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}e^{-\frac{1}{2\sigma^2}\sum\limits_{i=1}^{n}(y_i - \beta_1 x_i)^2} \\ \ln L &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta_1 x_i)^2\end{aligned}$$

Let

$$\begin{cases}\dfrac{\partial \ln L}{\partial \sigma^2} = -\dfrac{n}{2\sigma^2} + \dfrac{1}{\sigma^4}\sum\limits_{i=1}^{n} x_i^2 = 0 \\ \dfrac{\partial \ln L}{\partial \beta_1} = -\dfrac{1}{\sigma^2}\sum\limits_{i=1}^{n} x_i(y_i - \beta_1 x_i) = 0\end{cases}$$

we get

$$\begin{cases}\hat{\sigma^2}_{MLE} = \dfrac{1}{n}\sum\limits_{i=1}^{n} X_i^2 \\ \hat{\beta}_{1MLE} = \dfrac{\sum\limits_{i=1}^{n} X_i Y_i}{\sum\limits_{i=1}^{n} X_i^2}\end{cases}$$

(c) Show that the maximum likelihood estimator of β, is unbiased.

$\because$

$$\begin{aligned} E\hat{\beta}_{1MLE} &= E\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= E\left[\frac{\sum_{i=1}^{n} X_i(\beta_1 X_i + \epsilon_i)}{\sum_{i=1}^{n} X_i^2}\right] \\ &= \beta_1 + \frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} X_i^2} E\epsilon_i \\ &= \beta_1 \end{aligned}$$

$\therefore$ $\hat{\beta}_{1MLE}$ is the unbiased estimator of β_1

(Optional) Show least square estimator b_0 is BLUE of β_0 in model (1.1) with error distribution unspecified.

$\because$

$$b_0 = \overline{Y} - b_1\overline{X} = \sum_{i=1}^{n}\left(\frac{1}{n} - k_i\overline{X}\right)Y_i = \sum_{i=1}^{n} l_i Y_i$$

where $k_i = \frac{X_i - \overline{X}}{SS_{XX}}$, $\sum_{i=1}^{n} k_i = 0$, $\sum_{i=1}^{n} k_i^2 = \frac{1}{SS_{XX}}$, $\sum_{i=1}^{n} l_i = 1$, $\sum_{i=1}^{n} l_i^2 = \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}}$

Have proved that

$$Eb_0 = \beta_0$$

$$Varb_0 = \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}}\sigma^2$$

For any linear unbiased estimator of β_0,

$$b = \sum_{i=1}^{n} c_i Y_i$$

$\because$

$$\begin{aligned} \mathbb{E}b &= \sum_{i=1}^{n} c_i \mathbb{E}Y_i \\ &= \sum_{i=1}^{n} c_i(\beta_0 + \beta_1 X_i) \\ &= \sum_{i=1}^{n} c_i\beta_0 + \sum_{i=1}^{n} c_i X_i \beta_1 \\ &= \beta_0 \end{aligned}$$

$\therefore$

$$\begin{cases} \sum_{i=1}^{n} c_i = 1 \\ \sum_{i=1}^{n} c_i X_i = 0 \end{cases}$$

Let $d_i = c_i - l_i$, we have

$$\begin{aligned}
Varb &= \sum_{i=1}^{n} c_i^2 VarY_i \\
&= \sigma^2 \sum_{i=1}^{n} (d_i + l_i)^2 \\
&= \sigma^2 \sum_{i=1}^{n} (d_i^2 + l_i^2 + 2d_i l_i) \\
&= \sigma^2 \sum_{i=1}^{n} l_i^2 + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \sum_{i=1}^{n} d_i l_i \\
&= Var\beta_0 + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \sum_{i=1}^{n} (c_i - l_i) l_i \\
&= Var\beta_0 + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \sum_{i=1}^{n} c_i l_i - 2\sigma^2 \sum_{i=1}^{n} l_i^2 \\
&= Var\beta_0 + \sigma^2 \sum_{i=1}^{n} d_i^2 + 2\sigma^2 \sum_{i=1}^{n} c_i (\frac{1}{n} - k_i \overline{X}) - 2\sigma^2 \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}} \\
&= Var\beta_0 + \sigma^2 \sum_{i=1}^{n} d_i^2 + \frac{2\sigma^2}{n} - 2\sigma^2 \sum_{i=1}^{n} \frac{c_i X_i \overline{X} - c_i \overline{X}^2}{SS_{XX}} - 2\sigma^2 \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}} \\
&= Var\beta_0 + \sigma^2 \sum_{i=1}^{n} d_i^2 + \frac{2\sigma^2}{n} + 2\sigma^2 \sum_{i=1}^{n} \frac{\overline{X}^2}{SS_{XX}} - 2\sigma^2 \frac{\sum_{i=1}^{n} X_i^2}{nSS_{XX}} \\
&= Var\beta_0 + \sigma^2 \sum_{i=1}^{n} d_i^2 \\
&\geqslant Var\beta_0
\end{aligned}$$

the equation holds iff $d_1 = d_2 = \cdots = d_n = 0$

$\therefore$ b_0 is the BLUE of β_0