STA302H1F: Mini Project 1

August 1st, 2021

Contents

Question 1

```
beta0 <- 0.5
beta1 <- 2
beta2 <- 1
sigma <- 2
n <- 100
set.seed(1005790495)
X_1 <- sample(seq(1,15, length = n))
X_2 <- sample(seq(1,15, length = n))
cor(X_1, X_2)
```

```
## [1] -0.2392319
```

The two predictor variables have a correlation of -0.239, which is quite low. Because the variables are simulated independently, the correlation makes sense.

Table 1: Table of coefficients

estimator	estimate
beta0	-0.6264
beta1	2.1120
beta2	1.0734

Question 2

```
set.seed(1005790495)
error <- rnorm(n, mean = 0, sd = sigma)
Y <- beta0 + beta1*X_1 + beta2*X_2 + error
sim_data <- data.frame(Y = Y, X_1 = X_1, X_2 = X_2)
multi.fit <- lm(Y ~ X_1 + X_2, data = sim_data)

coefficients <- data.frame(estimator = c("beta0", "beta1", "beta2"),
                           estimate = c(summary(multi.fit)$coefficients[1,1],
                                        summary(multi.fit)$coefficients[2,1],
                                        summary(multi.fit)$coefficients[3,1]))

kbl(coefficients, caption = "Table of coefficients", digits = 4) %>%
  kable_paper(full_width = F)
```

The regression coefficient for beta $\hat{\beta}_0$ from the multiple linear regression model is -0.6264, for $\hat{\beta}_1$ is 2.112, for $\hat{\beta}_2$ is 1.0734.

3. Matrix Algebra

$$X = \begin{pmatrix} 1 & x_1 \\ \vdots & \\ 1 & x_n \end{pmatrix}$$ # definition

$$X'X = \begin{pmatrix} 1 \cdots 1 \\ x_1 \cdots x_n \end{pmatrix} \cdot \begin{pmatrix} 1 & x_1 \\ \vdots & \\ 1 & x_n \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{pmatrix}$$

$$(X'X)^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{pmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{pmatrix}$$

$$= \begin{pmatrix} \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} & \frac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \\ \frac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} & \frac{n}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \end{pmatrix}$$ # hint

$$X'Y = \begin{pmatrix} 1 \cdots 1 \\ x_1 \cdots x_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}$$

$$= \begin{pmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{pmatrix}$$

$$(X'X)^{-1}X'Y$$

$$= \begin{pmatrix} \dfrac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} & \dfrac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ \dfrac{-\sum_{i=1}^{n} x_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} & \dfrac{n}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{pmatrix} \begin{pmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} x_i Y_i \end{pmatrix}$$

$$= \begin{pmatrix} \dfrac{(\sum_{i=1}^{n} x_i^2)(\sum_{i=1}^{n} Y_i) - \sum_{i=1}^{n} x_i (\sum_{i=1}^{n} x_i Y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \\ \dfrac{(-\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} Y_i) + n(\sum_{i=1}^{n} x_i Y_i)}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \end{pmatrix} \quad (1)$$

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i\bar{x} + \bar{x}^2)$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\sum_{i=1}^{n} x_i + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - 2\bar{x}\, n\bar{x} + n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

$$= \sum_{i=1}^{n} x_i^2 - n\left(\frac{\sum x_i}{n}\right)^2$$

$$= \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum x_i\right)^2}{n} \quad (2)$$

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y}) = \sum_{i=1}^{n}(x_i y_i - x_i\bar{y} - \bar{x}y_i + \bar{x}\bar{y})$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + n\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - \bar{y}\,n\bar{x} - \bar{x}\,n\bar{y} + n\bar{x}\bar{y}$$

$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \quad (3)$$

By (2) and (3),

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{n\sum_{i=1}^{n} x_i y_i - n\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Thus,

$$\bar{y} - \hat{\beta}_1\bar{x}$$

$$= \frac{\sum_{i=1}^{n} y_i}{n} - \frac{n\sum_{i=1}^{n} x_i y_i - n\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)$$

$$= \frac{\sum_{i=1}^{n} y_i\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right) - n\left(\sum_{i=1}^{n} x_i y_i\right)\left(\sum_{i=1}^{n} x_i\right) + \left(\sum_{i=1}^{n} x_i\right)^2 \sum_{i=1}^{n} y_i}{n^2\sum_{i=1}^{n} x_i^2 - n\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{n\left(\sum_{i=1}^{n} x_i^2\right)\sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} x_i\right)^2 \sum_{i=1}^{n} y_i - n\left(\sum_{i=1}^{n} x_i y_i\right)\sum_{i=1}^{n} x_i + \left(\sum_{i=1}^{n} x_i\right)^2 \sum_{i=1}^{n} y_i}{n^2\sum_{i=1}^{n} x_i^2 - n\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{n\left(\sum_{i=1}^{n} x_i^2\right)\sum_{i=1}^{n} y_i - n\left(\sum_{i=1}^{n} x_i y_i\right)\sum_{i=1}^{n} x_i}{n^2 \sum_{i=1}^{n} x_i^2 - n\left(\sum_{i=1}^{n} x_i\right)^2}$$

$$= \frac{\left(\sum_{i=1}^{n} x_i^2\right)\sum_{i=1}^{n} y_i - \left(\sum_{i=1}^{n} x_i y_i\right)\sum_{i=1}^{n} x_i}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

After adding ① to this, we conclude that

$$\hat{\beta} = \begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \begin{pmatrix} \bar{Y} - \hat{\beta}_1 \bar{x} \\ \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \end{pmatrix} = (X'X)^{-1} X'Y$$

Table 2: Table of beta

beta	value
beta0	-0.6264
beta1	2.1120
beta2	1.0734

Table 3: Standard error of beta

	beta	error
X_0	beta0	0.6648
X_1	beta1	0.0504
X_2	beta2	0.0504

Question 4

```
X_0=rep(1,n)
X=cbind(X_0, X_1, X_2)
XY=t(X)%*%Y
XX=t(X)%*%X
Inverse=solve(XX)
beta_Q4=Inverse%*%XY

error_Q4 <- Y-X %*% beta_Q4
S2 <- (t(error_Q4) %*% error_Q4) / (n-2-1)
S2 <- as.numeric(S2)

beta_value <- data.frame(beta = c("beta0", "beta1", "beta2"),
                         value = c(beta_Q4))

kbl(beta_value, caption = "Table of beta", digits = 4) %>%
  kable_paper(full_width = F)
```

```
standard_error_value <- data.frame(beta = c("beta0", "beta1", "beta2"),
                         error = c(sqrt(diag(Inverse * S2))))

kbl(standard_error_value, caption = "Standard error of beta", digits = 4) %>%
  kable_paper(full_width = F)
```

We could notice that the estimates from lm function in question 2 and the estimates we obtained in this quetsion are the same.

Table 4: Mean of each beta

	x
beta0	0.5164
beta1	1.9973
beta2	0.9995

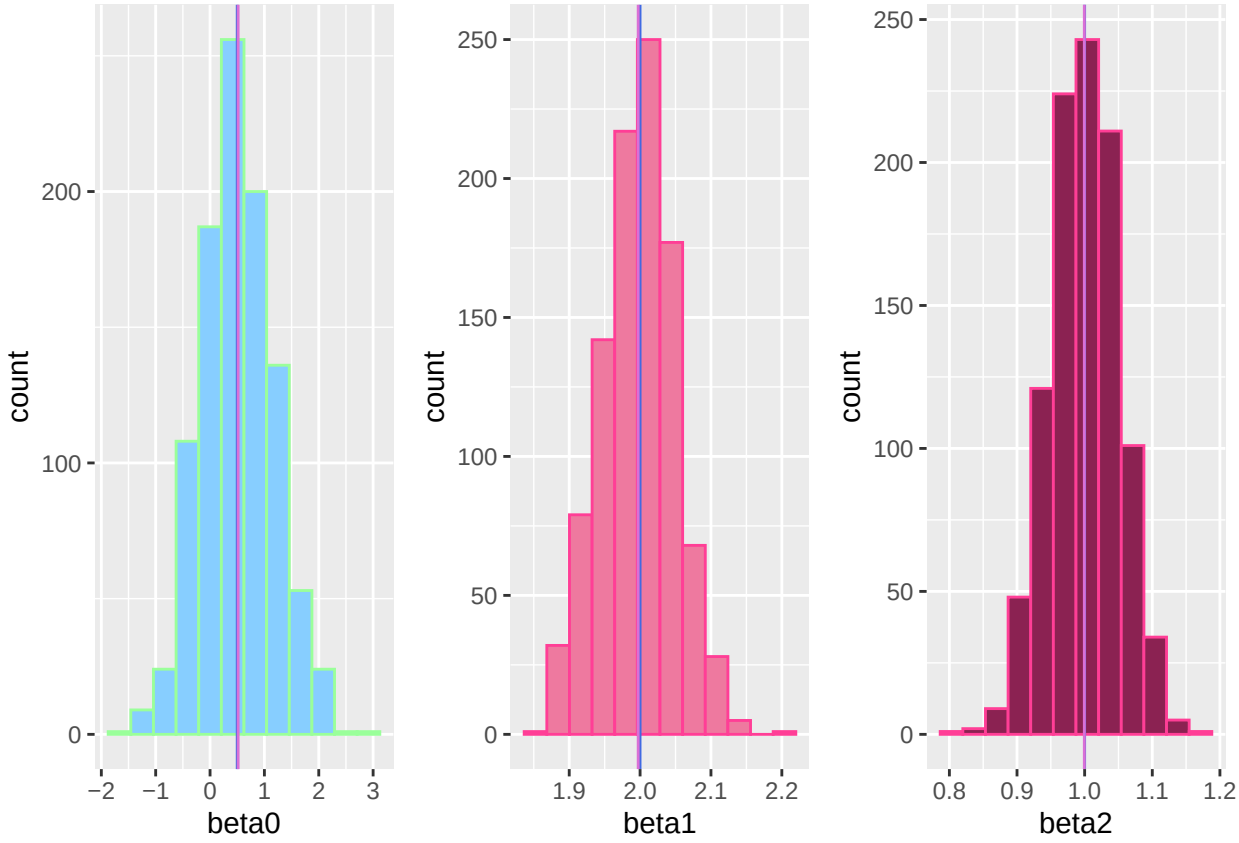
Question 5

```
set.seed(1005790495)
beta_0 <- c()
beta_1 <- c()
beta_2 <- c()
for (i in 1 : 1000) {
  error <- rnorm(n, mean = 0, sd = sigma)
  Y <- beta0 + beta1*X_1 + beta2*X_2 + error
  sim_data <- data.frame(Y = Y, X_1 = X_1, X_2 = X_2)
  multi.fit <- lm(Y ~ X_1 + X_2, data = sim_data)
  beta_0[i] <- coefficients(multi.fit)[1]
  beta_1[i] <- coefficients(multi.fit)[2]
  beta_2[i] <- coefficients(multi.fit)[3]
}
beta_Q5 <- data.frame(beta0 = beta_0,
                      beta1 = beta_1,
                      beta2 = beta_2)

kbl(colMeans(beta_Q5), caption = "Mean of each beta", digits = 4) %>%
  kable_paper(full_width = F)
```

Question 6

```
ggplot0 <- ggplot(beta_Q5, aes(x=beta0)) +
  geom_histogram(bins = 12, col="palegreen1", fill="skyblue1") +
              geom_vline(xintercept = beta0, col="royalblue") +
              geom_vline(xintercept = mean(beta_Q5$beta0), col="orchid")
ggplot1 <- ggplot(beta_Q5, aes(x=beta1)) +
  geom_histogram(bins = 12, col="violetred1", fill="palevioletred2") +
              geom_vline(xintercept = beta1, col="royalblue") +
              geom_vline(xintercept = mean(beta_Q5$beta1), col="orchid")
ggplot2 <- ggplot(beta_Q5, aes(x=beta2)) +
  geom_histogram(bins = 12, col="violetred1", fill="violetred4") +
              geom_vline(xintercept = beta2, col="royalblue") +
              geom_vline(xintercept = mean(beta_Q5$beta2), col="orchid")
grid.arrange(ggplot0, ggplot1, ggplot2, nrow = 1, ncol = 3)
```



According to the histograms, we could notice that the mean for each of the estimates is closely related to the true values of the parameters. The reasons can be explained as follows. In the multiple linear regression, the theoretical distribution of $\hat{\beta} = \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$ is the sampling distribution of $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$. By definition:

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$$
$$\hat{\boldsymbol{\beta}} = [\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2]$$

The experimental values from the variables we randomly stimulated from the distribution can be used to estimate the theoretical values of this distribution. The medium points of the histograms are the means of

the experimental values we simulated previously. The larger the N, the closer the histograms of experimental values and the theoretical sampling distribution will be, and thereby the closer the mean for each of the estimates and the true values of the parameters will be. According to Question 5, the number of simulations N is set at 1000, which is large enough. Thus, the results make sense.

Question 7

```
set.seed(1005790495)
alpha = 0.05
#number of parameters
p = 2
beta0_CI = 0
beta1_CI = 0
beta2_CI = 0
for (i in 1 : 1000) {
  error <- rnorm(n, mean = 0, sd = sigma)
  Y <- beta0 + beta1*X_1 + beta2*X_2 + error
  sim_data <- data.frame(Y = Y, X_1 = X_1, X_2 = X_2)
  multi.fit <- lm(Y ~ X_1 + X_2, data=sim_data)

  beta0_hat <- coefficients(multi.fit)[1]
  beta1_hat <- coefficients(multi.fit)[2]
  beta2_hat <- coefficients(multi.fit)[3]

  beta0_se <- summary(multi.fit)$coefficients[1,2]
  beta1_se <- summary(multi.fit)$coefficients[2,2]
  beta2_se <- summary(multi.fit)$coefficients[3,2]

  #t-quantile
  beta0_lower <- beta0_hat - qt(1-alpha/2, df = n-p-1)*beta0_se
  beta0_upper <- beta0_hat + qt(1-alpha/2, df = n-p-1)*beta0_se

  beta1_lower <- beta1_hat - qt(1-alpha/2, df = n-p-1)*beta1_se
  beta1_upper <- beta1_hat + qt(1-alpha/2, df = n-p-1)*beta1_se

  beta2_lower <- beta2_hat - qt(1-alpha/2, df = n-p-1)*beta2_se
  beta2_upper <- beta2_hat + qt(1-alpha/2, df = n-p-1)*beta2_se

  #count for any condition
  beta0_CI <- beta0_CI + sum(ifelse(beta0_lower <= beta0 & beta0 <= beta0_upper, 1, 0))
  beta1_CI <- beta1_CI + sum(ifelse(beta1_lower <= beta1 & beta1 <= beta1_upper, 1, 0))
  beta2_CI <- beta2_CI + sum(ifelse(beta2_lower <= beta2 & beta2 <= beta2_upper, 1, 0))

}


CI <- data.frame(Betas = c("beta0", "beta1", "beta2"),
                 Contain_true_parameter = c("YES", "YES", "YES"),
                 Lowerbound = c(beta0_lower, beta1_lower, beta2_lower),
                 Upperbound = c(beta0_upper, beta1_upper, beta2_upper))
```

Table 5: CI of betas

	Betas	Contain_true_parameter	Lowerbound	Upperbound
(Intercept)	beta0	**YES**	-0.3392092	2.258398
X_1	beta1	**YES**	1.8331914	2.029928
X_2	beta2	**YES**	0.9335051	1.130242

Table 6: Coverage probability

beta0_CI	beta1_CI	beta2_CI
0.949	0.937	0.946

```
kbl(CI, caption = "CI of betas") %>%
  kable_paper(full_width = F) %>%
  column_spec(1, bold = T, border_right = T) %>%
  column_spec(3, color = "white", bold = T, background = "orange")
```

```
coverage <- data.frame(beta0_CI = beta0_CI/1000,
                       beta1_CI = beta1_CI/1000,
                       beta2_CI = beta2_CI/1000)
kable(coverage, caption = "Coverage probability")
```