STA 4210 HW1

Yansheng Luo

1.(R) 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 the simple linear regression model is appropriate.

```
shipment <- c(1,0,2,0,3,1,0,1,2,0)
ampules <- c(16,9,17,12,22,13,8,15,19,11)
```

a. Obtain the estimated regression function and MSE "by hand" (i.e. without using the `lm` function).

Solution:

The estimated regression model is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Given that $n = 10$ and:

$$\sum x_i = 10, \quad \sum y_i = 142.$$

From the formula sheet,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

we get the sample mean of x and y :

$$\bar{x} = \frac{10}{10} = 1, \qquad \bar{y} = \frac{142}{10} = 14.2.$$

From the formula sheet,

$$\hat{\beta}_1 = b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = b_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

Compute the numerator:

$$\sum(x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \bar{x}\sum y_i - \bar{y}\sum x_i + n\bar{x}\bar{y}$$

$$= 182 - (1)(142) - (14.2)(10) + 10(1)(14.2) = 40.$$

Compute the denominator:

$$\sum(x_i - \bar{x})^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2$$

$$= 20 - 2(1)(10) + 10(1^2) = 10.$$

Therefore,

$$\hat{\beta}_1 = b_1 = \frac{40}{10} = 4, \qquad \hat{\beta}_0 = b_0 = 14.2 - 4(1) = 10.2.$$

Estimated regression function:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x = 10.2 + 4x.$$

From the formula sheet,

$$MSE = \frac{SSE}{n-2}.$$

Compute

$$SSE = \sum \big(y_i - (b_0 + b_1 x_i)\big)^2$$

$$= \sum y_i^2 + n b_0^2 + b_1^2 \sum x_i^2 + 2 b_0 b_1 \sum x_i - 2 b_0 \sum y_i - 2 b_1 \sum x_i y_i.$$

Substitute $b_0 = 10.2$, $b_1 = 4$:

$$SSE = 2194 + 10(10.2^2) + 16(20) + 2(10.2)(4)(10) - 2(10.2)(142) - 2(4)(182)$$

$$= 17.6.$$

Finally,

$$MSE = \frac{17.6}{10-2} = \frac{17.6}{8} = 2.2.$$

b. Obtain a point estimate of the expected number of broken ampules when $X = 1$ transfer is made using the regression function you calculated in part (a).

solution:

From part (a), the estimated regression function is

$$\hat{y} = 10.2 + 4x.$$

A point estimate of the expected number of broken ampules when $X = 1$ transfer is made is obtained by evaluating the regression function at $x = 1$:

$$\hat{y}(1) = 10.2 + 4(1) = 14.2.$$

Thus, the point estimate of the expected number of broken ampules when one transfer is made is 14.2.

c. Fit the simple linear regression model by the `lm` function. Plot the data and the estimated regression line. Based on your plot, do you think the linear regression model is appropriate for this data?

Solution:

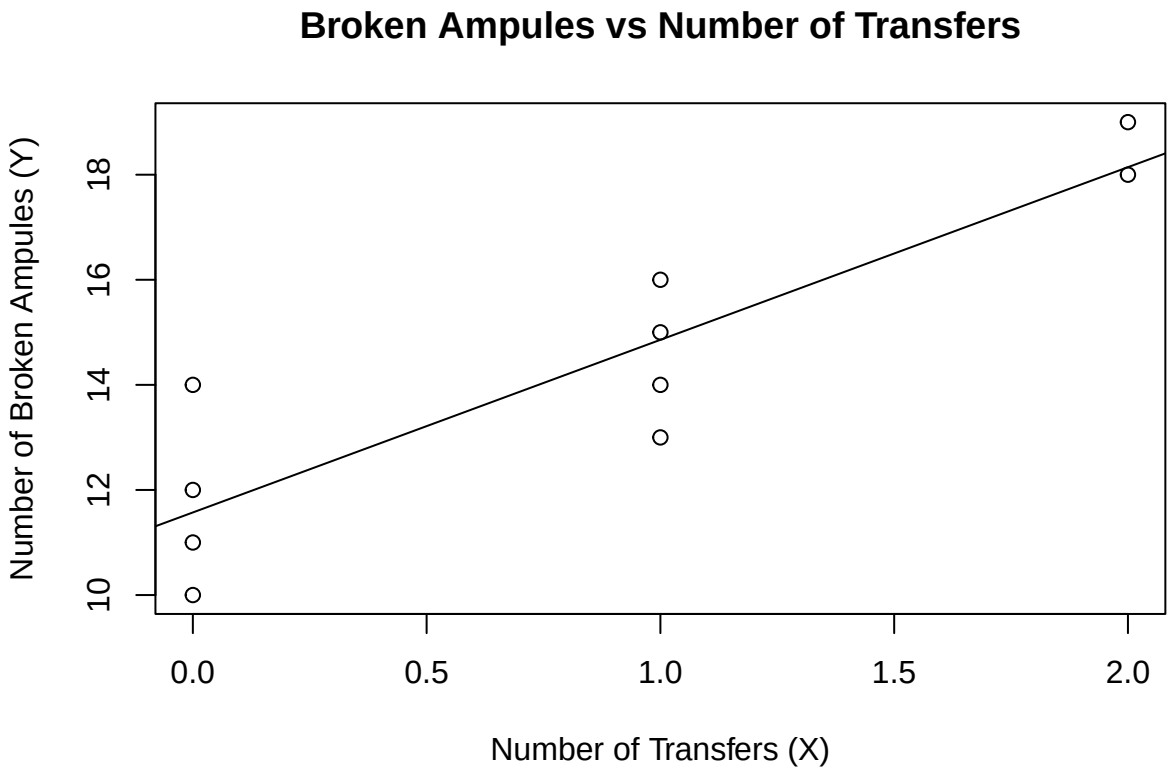
Fit the simple linear regression model using the `lm` function and plot the data with the estimated regression line.

```r
# observed data
x <- c(0, 0, 0, 0, 1, 1, 1, 1, 2, 2)
y <- c(10, 12, 14, 11, 14, 15, 16, 13, 18, 19)

# fit simple linear regression model
fit <- lm(y ~ x)

# plot data
plot(x, y,
     xlab = "Number of Transfers (X)",
     ylab = "Number of Broken Ampules (Y)",
     main = "Broken Ampules vs Number of Transfers")

# add regression line
abline(fit)
```



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

Solution:

From part (a), the fitted regression line is

$$\hat{y} = 10.2 + 4x.$$

From the data that we computed in part a,

$$\bar{X} = 1, \qquad \bar{Y} = 14.2.$$

Plug in 1 to valuate the fitted regression function at $\bar{X} = 1$:

$$\hat{y}(\bar{X}) = 10.2 + 4(1) = 14.2.$$

Since

$$\hat{y}(\bar{X}) = \bar{Y},$$

the fitted regression line does passes through the point $(\bar{X}, \bar{Y})$. :)s

2.(R) Consider the simple linear regression model through the origin: $Y_i = \beta_1 X_i + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$. Consider two estimators for β_1:

- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}$
- $\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$.

  We can show that $E[\hat{\beta}_1] = E[\tilde{\beta}_1] = \beta_1$. So these two estimators are both unbiased estimators for β_1. In general, an estimator with smaller variance is more efficient. For this problem, we will conduct simulation studies to analyze the Monte Carlo variance of the two estimators.

a. Run 1000 simulations. In each simulation, generate samples (X_i, Y_i, ϵ_i) with sample size $n = 30$ where $X_i \sim N(1, 9)$ (note that 9 is the variance), $\epsilon_i \sim N(0, 1)$ and $Y_i = 2X_i + \epsilon_i$ for $i = 1, .., n$, and obtain the $\hat{\beta}_1, \tilde{\beta}_1$ based on the sample. Use `set.seed(123)` at the beginning of the code to ensure that results are reproducible.

```r
set.seed(123)

# number of simulations and sample size
M <- 1000
n <- 30
beta1 <- 2

# storage for estimators
beta_hat <- numeric(M)
beta_tilde <- numeric(M)

for (m in 1:M) {

  # generate data
  X <- rnorm(n, mean = 1, sd = 3)       # X_i ~ N(1, 9)
  epsilon <- rnorm(n, mean = 0, sd = 1) # epsilon_i ~ N(0, 1)
  Y <- beta1 * X + epsilon              # Y_i = 2 X_i + epsilon_i

  # estimators
  beta_hat[m] <- sum(X * Y) / sum(X^2)
  beta_tilde[m] <- sum(Y) / sum(X)
}

# Monte Carlo means
mc_mean_beta_hat <- mean(beta_hat)
mc_mean_beta_tilde <- mean(beta_tilde)

# Monte Carlo variances
mc_var_beta_hat <- var(beta_hat)
```

```r
mc_var_beta_tilde <- var(beta_tilde)

# output
mc_mean_beta_hat
```

```
## [1] 2.002106
```

```r
mc_mean_beta_tilde
```

```
## [1] 2.001178
```

```r
mc_var_beta_hat
```

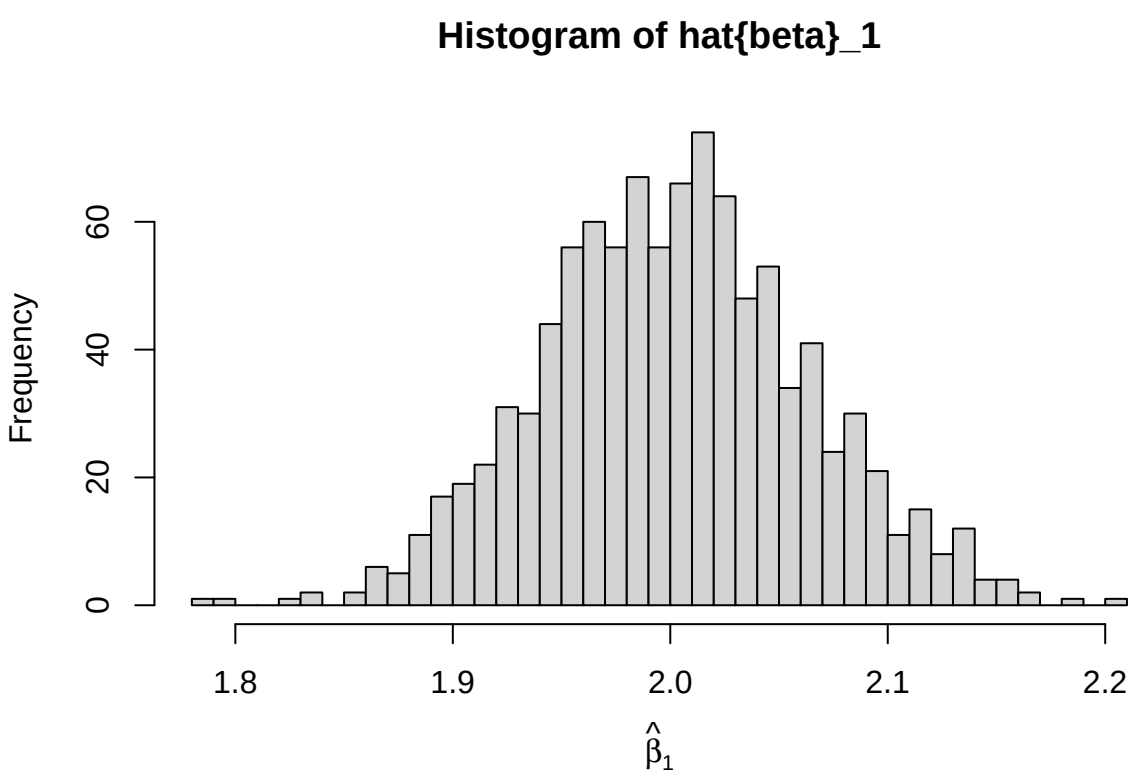
```
## [1] 0.00374595
```

```r
mc_var_beta_tilde
```

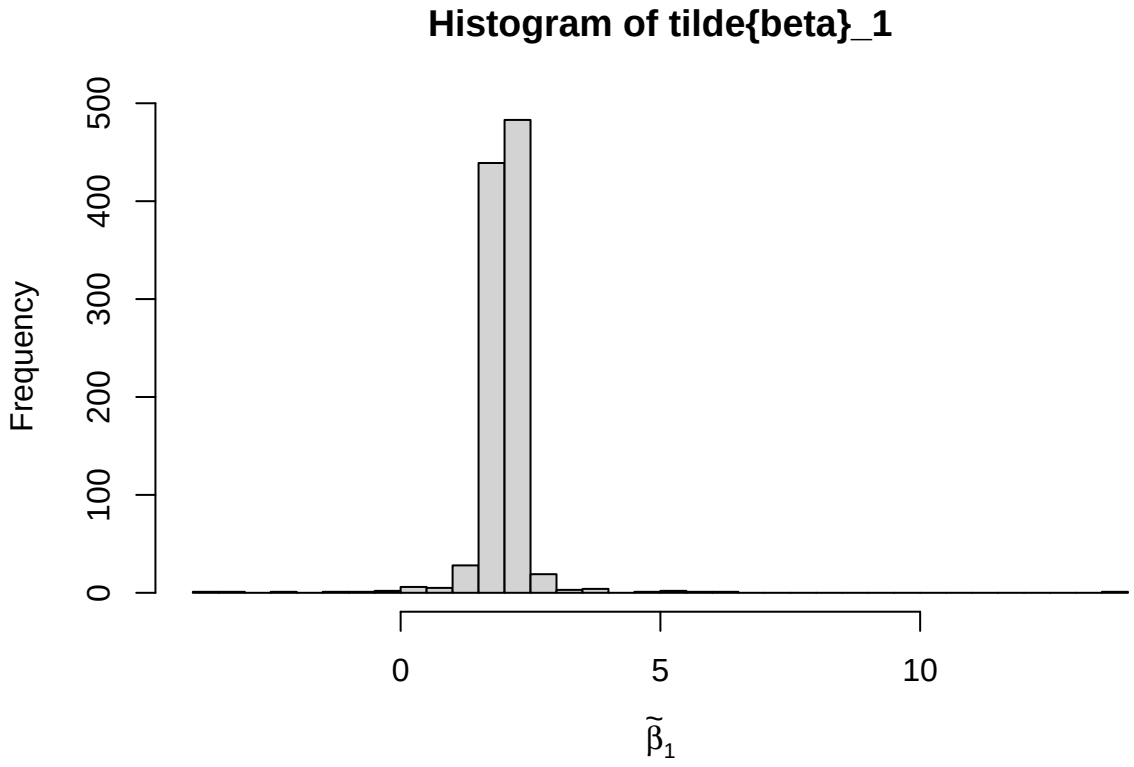
```
## [1] 0.4084336
```

b. Plot the histogram of $\hat{\beta}_1$ and $\tilde{\beta}$ across all simulations.

```r
# histogram of beta_hat
hist(beta_hat,
     breaks = 30,
     main = "Histogram of hat{beta}_1",
     xlab = expression(hat(beta)[1]))
```



```r
# histogram of beta_tilde
hist(beta_tilde,
     breaks = 30,
     main = "Histogram of tilde{beta}_1",
     xlab = expression(tilde(beta)[1]))
```



c. Compute the sample variance for $\hat{\beta}_1$ and $\tilde{\beta}_1$. Which estimator has smaller sample variance?

```r
var_beta_hat <- var(beta_hat)
var_beta_tilde <- var(beta_tilde)

var_beta_hat
```

```
## [1] 0.00374595
```

```r
var_beta_tilde
```

```
## [1] 0.4084336
```

The sample variance of $\hat{\beta}_1$ is smaller than the sample variance of $\tilde{\beta}_1$. Therefore, $\hat{\beta}_1$ is the more efficient estimator of β_1.

3. Consider the simple linear regression model through the origin and two estimators $\hat{\beta}_1$ and $\tilde{\beta}_1$ defined in problem 2. In the problem 2, we investigate the sampling distribution of these two estimators by simulations. For this problem, we will derive the **true** sampling distribution of the these two estimators.

a. Derive the $E(\hat{\beta}_1)$ and $V(\hat{\beta}_1)$. Write down the sampling distribution of $\hat{\beta}_1$. (Hint: $\hat{\beta}_1$ is a linear combination of Y_i's)
b. Derive the $E(\tilde{\beta}_1)$ and $V(\tilde{\beta}_1)$. Write down the sampling distribution of $\tilde{\beta}_1$.
c. Prove that $V(\hat{\beta}_1) \leq V(\tilde{\beta}_1)$. (Hint: there are different ways to show this. One option is to prove that $\frac{1}{V(\hat{\beta}_1)} - \frac{1}{V(\tilde{\beta}_1)} = \sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \geq 0$)