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 we learned is

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

```
# load sample size
n <- length(shipment)

# compute sample means for xbar and ybar
xbar <- mean(shipment)
ybar <- mean(ampules)

#print
xbar
```

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

```
ybar
```

```
## [1] 14.2
```

From the formula sheet,

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

```
# slope and intercept (by hand formulas)
numerator   <- sum((shipment - xbar) * (ampules - ybar))
denominator <- sum((shipment - xbar)^2)

b1 <- numerator / denominator
b0 <- ybar - b1 * xbar

b1
```

```
## [1] 4
```

```r
b0
```

```
## [1] 10.2
```

Thus, the estimated regression function is

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

From the formula sheet,

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

```r
# fitted values
yhat <- b0 + b1 * shipment

# SSE and MSE
SSE <- sum((ampules - yhat)^2)
MSE <- SSE / (n - 2)

SSE
```

```
## [1] 17.6
```

```r
MSE
```

```
## [1] 2.2
```

$$MSE = 2.2.$$

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?

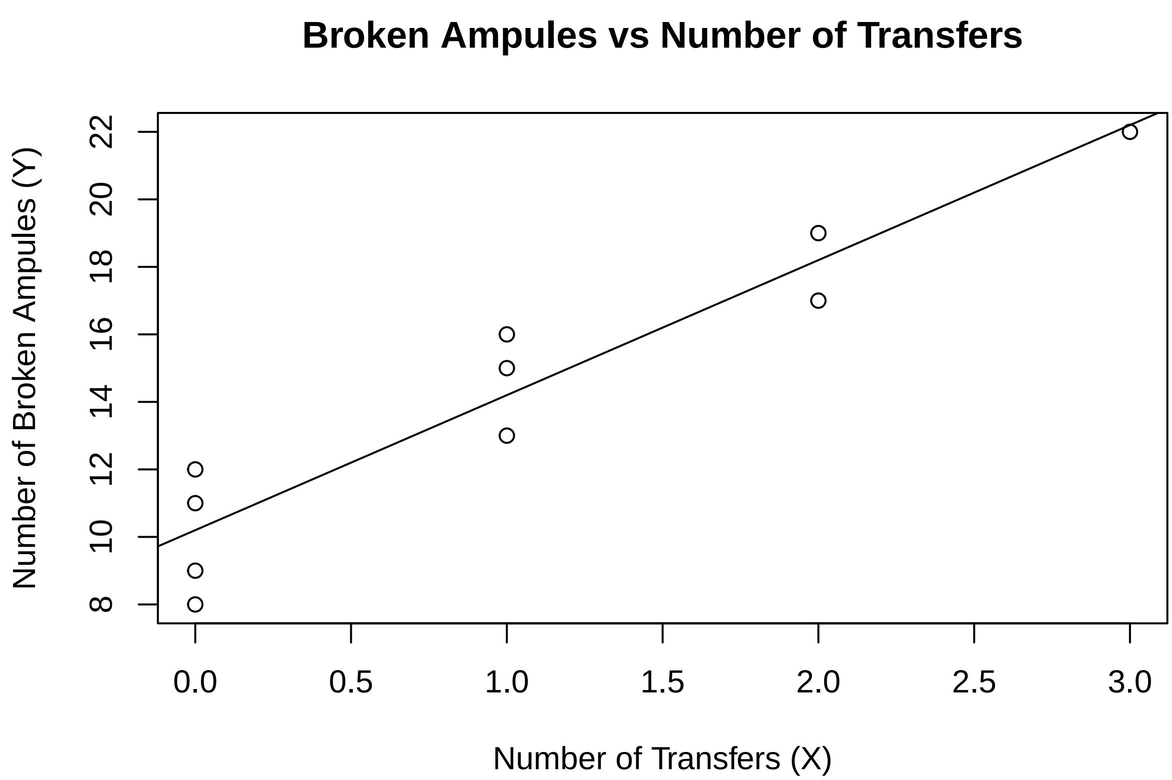
Solution:

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

```r
# fit simple linear regression model using the given data
fit <- lm(ampules ~ shipment)

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

# add regression line
abline(fit)
```



Based on the plot, the relationship between the number of transfers and the number of broken ampules appears approximately linear, with an increasing trend and no strong curvature. Therefore, the simple linear regression model appears appropriate for this data.

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.

```
set.seed(123)

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

# storage for estimators
beta1_hat_values <- numeric(num_sim)     # sum(X_i Y_i) / sum(X_i^2)
beta1_tilde_values <- numeric(num_sim)   # sum(Y_i) / sum(X_i)

for (m in 1:num_sim) {

  # 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
  beta1_hat_values[m] <- sum(X * Y) / sum(X^2)
  beta1_tilde_values[m] <- sum(Y) / sum(X)
}

# empirical means across simulations
mean_hat <- mean(beta1_hat_values)
mean_tilde <- mean(beta1_tilde_values)
```

```r
# sample variances across simulations
var_hat <- var(beta1_hat_values)
var_tilde <- var(beta1_tilde_values)

# output
mean_hat
```

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

```r
mean_tilde
```

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

```r
var_hat
```

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

```r
var_tilde
```

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

Based on the 1000 simulated samples, the empirical mean of

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

is

$$2.002106,$$

and the empirical mean of

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

is

$$2.001178.$$

The sample variance of

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

across the 1000 simulations is

$$0.00374595,$$

while the sample variance of

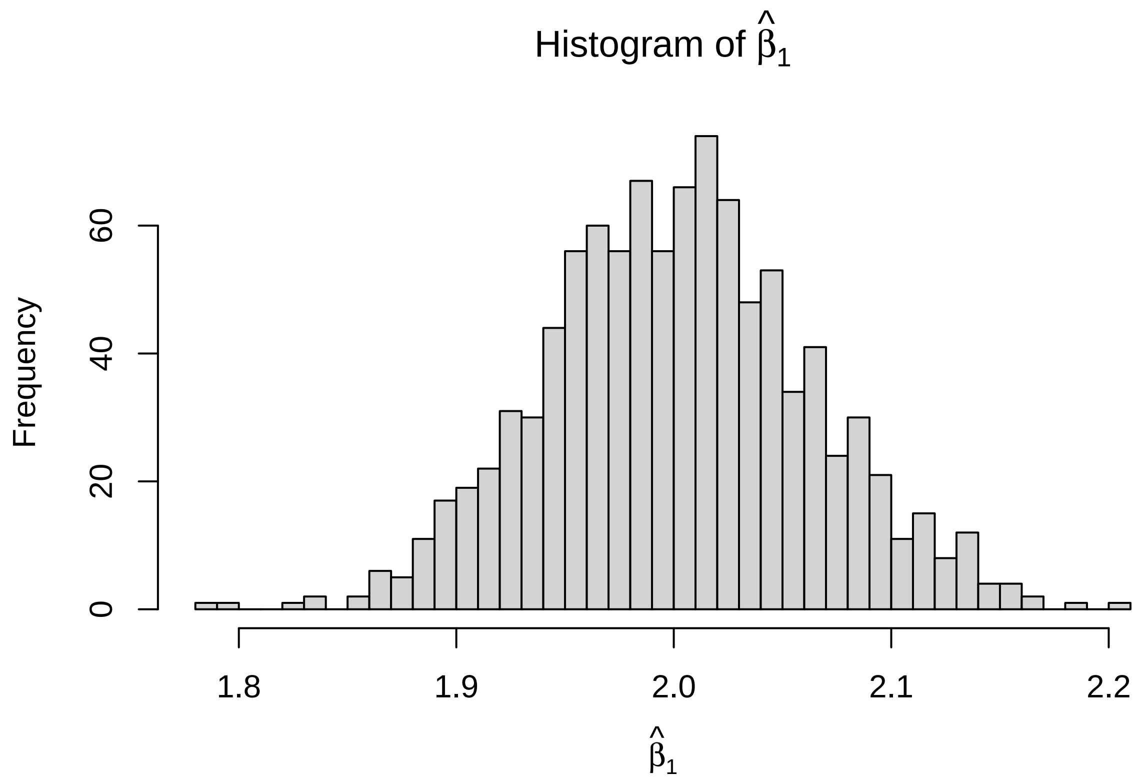
$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} X_i}$$

across the 1000 simulations is

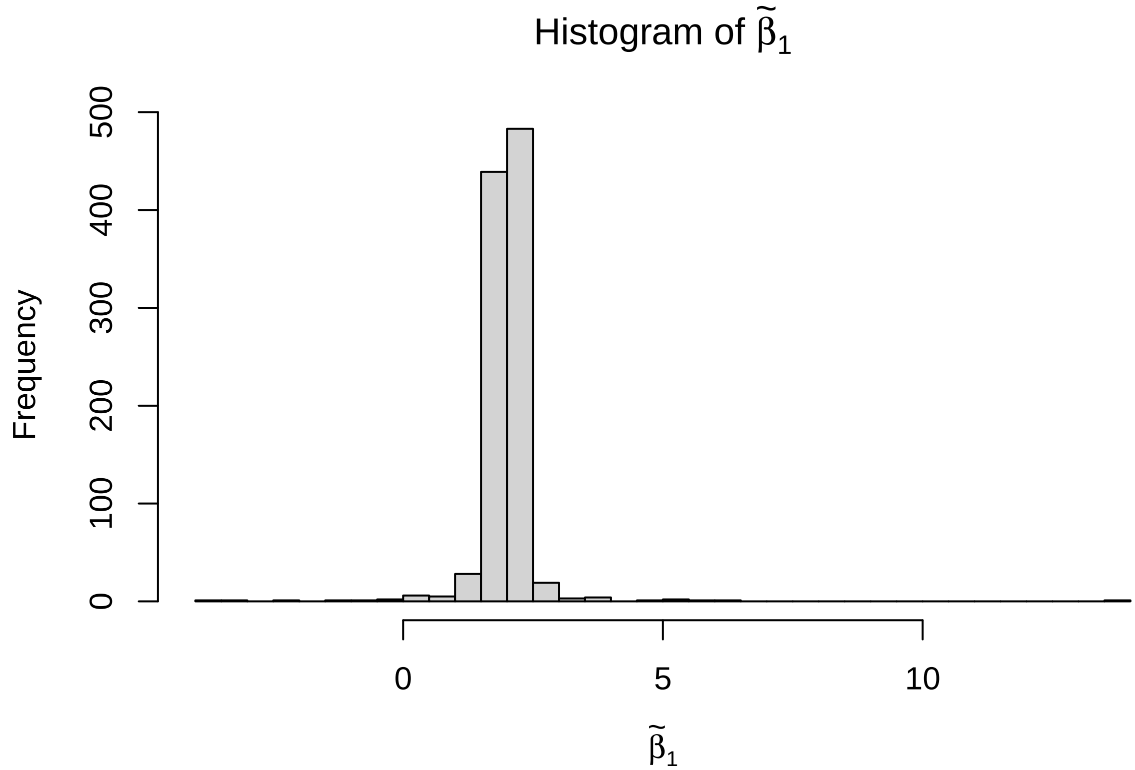
$$0.4084336.$$

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

```r
# histogram of hat{beta}_1
hist(beta1_hat_values,
     breaks = 30,
     main = expression(paste("Histogram of ", hat(beta)[1])),
     xlab = expression(hat(beta)[1]))
```




```r
# histogram of tilde{beta}_1
hist(beta1_tilde_values,
     breaks = 30,
     main = expression(paste("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_hat <- var(beta1_hat_values)
var_tilde <- var(beta1_tilde_values)

var_hat
```

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

```r
var_tilde
```

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

Since both $\hat{\beta}_1$ and $\tilde{\beta}_1$ are unbiased estimators of β_1, efficiency is determined by their variances. The estimator with smaller variance is more efficient because its values are more tightly concentrated around the true parameter. Since

$$\mathrm{Var}(\hat{\beta}_1) = 0.00374595 < \mathrm{Var}(\tilde{\beta}_1) = 0.4084336,$$

$\hat{\beta}_1$ 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)

Solution:

The simple linear regression model is:

$$Y_i = \beta_1 X_i + \varepsilon_i, \qquad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

The estimator is

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2} = \sum_{i=1}^n a_i Y_i, \qquad a_i = \frac{X_i}{\sum_{j=1}^n X_j^2}.$$

From the formula sheet for linear combinations,

$$E\left(\sum_{i=1}^n a_i Y_i\right) = \sum_{i=1}^n a_i E(Y_i).$$

Since $E(Y_i) = \beta_1 X_i$,

$$E(\hat{\beta}_1) = \sum_{i=1}^n a_i \beta_1 X_i = \beta_1 \frac{\sum_{i=1}^n X_i^2}{\sum_{i=1}^n X_i^2} = \beta_1.$$

Using the variance formula for linear combinations and $Var(Y_i) = \sigma^2$,

$$V(\hat{\beta}_1) = \sum_{i=1}^n a_i^2 \sigma^2 = \sigma^2 \sum_{i=1}^n \frac{X_i^2}{\left(\sum_{j=1}^n X_j^2\right)^2} = \frac{\sigma^2}{\sum_{i=1}^n X_i^2}.$$

Since $\hat{\beta}_1$ is a linear combination of normally distributed Y_i's,

$$\hat{\beta}_1 \sim N\left(\beta_1,\ \frac{\sigma^2}{\sum_{i=1}^n X_i^2}\right).$$

b. Derive the $E(\tilde{\beta}_1)$ and $V(\tilde{\beta}_1)$. Write down the sampling distribution of $\tilde{\beta}_1$.

The estimator is

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i} = \sum_{i=1}^n c_i Y_i, \qquad c_i = \frac{1}{\sum_{j=1}^n X_j}.$$

Using the expectation formula for linear combinations,

$$E(\tilde{\beta}_1) = \sum_{i=1}^n c_i E(Y_i) = \frac{1}{\sum_{j=1}^n X_j} \sum_{i=1}^n \beta_1 X_i = \beta_1.$$

Using the variance formula from the formula sheet,

$$V(\tilde{\beta}_1) = \sum_{i=1}^n c_i^2 Var(Y_i) = \sum_{i=1}^n \left(\frac{1}{\sum_{j=1}^n X_j}\right)^2 \sigma^2 = \frac{n\sigma^2}{\left(\sum_{i=1}^n X_i\right)^2}.$$

Thus,

$$\tilde{\beta}_1 \sim N\left(\beta_1,\ \frac{n\sigma^2}{\left(\sum_{i=1}^n X_i\right)^2}\right).$$

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 $\frac{1}{V(\hat{\beta}_1)} - \frac{1}{V(\tilde{\beta}_1)} = \sum_{i=1}^n (X_i - \bar{X})^2/\sigma^2 \geq 0$)

From parts (a) and (b),

$$\frac{1}{V(\hat{\beta}_1)} = \frac{\sum_{i=1}^n X_i^2}{\sigma^2}, \qquad \frac{1}{V(\tilde{\beta}_1)} = \frac{\left(\sum_{i=1}^n X_i\right)^2}{n\sigma^2}.$$

Then,

$$\frac{1}{V(\hat{\beta}_1)} - \frac{1}{V(\tilde{\beta}_1)} = \frac{1}{\sigma^2}\left[\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}\right].$$

Using the identity from the formula sheet,

$$\sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n},$$

we obtain

$$\frac{1}{V(\hat{\beta}_1)} - \frac{1}{V(\tilde{\beta}_1)} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \geq 0.$$

Therefore,

$$V(\hat{\beta}_1) \leq V(\tilde{\beta}_1).$$