Statistics 206

Homework 1 (Solution)

Due : Oct. 6, 2022, 11:59PM

Instructions:

- You should upload homeworkX files on canvas (under "Assignments/hwX") before its due date.
- Your homework may be prepared by a word processor (e.g., Latex) or through handwriting.
- For handwritten homework, you should either scan or take photos of your homework: Please make sure the pages are clearly numbered and are in order and the scans/photos are complete and clear; Check before submitting.
- Please name the files following the format: "FirstName-LastName-HwX". If there are several files, you can use "-Questions1-5", "-Questions6", etc., to distinguish them. E.g., "Jie-Peng-Hw1-Questions1-5.pdf", "Jie-Peng-Hw1-Questions6.rmd".
- Your name should be clearly shown on the submitted files: By putting on your name, you also acknowledge that you are the person who did and prepared the submitted homework.
- **Optional Problems** are more advanced and are not counted towards the grade.
- Showing/sharing/uploading homework or solutions outside of this class is prohibited.

1. (a) $\mathbf{E}(A\mathbf{Z}) = A\mathbf{E}(\mathbf{Z})$.

Proof.

$$\begin{aligned}(A\mathbf{Z})_j &= \sum_k a_{jk}\mathbf{Z}_k, \quad j = 1, \cdots, s,\\ (E(A\mathbf{Z}))_j &= E((A\mathbf{Z})_j) = E(\sum_k a_{jk}\mathbf{Z}_k)\\ &= \sum_k a_{jk}E(\mathbf{Z}_k) = (AE(\mathbf{Z}))_j, \quad j = 1, \cdots, s.\end{aligned}$$

□

(b) $\mathbf{Cov}(A\mathbf{Z}, B\mathbf{Z}) = A\Sigma B^T$. So in particular, $\mathbf{Var}(A\mathbf{Z}) = A\Sigma A^T$.

Proof. Define

$$W = A\mathbf{Z}, U = B\mathbf{Z}, C = \mathrm{Cov}(W, U), D = A\Sigma B^T.$$

Then

$$\begin{aligned} C_{ij} &= \mathrm{Cov}(W_i, U_j) \\ &= \mathrm{Cov}(\sum_k a_{ik}\mathbf{Z}_k, \sum_k b_{jk}\mathbf{Z}_k) \\ &= \sum_k \sum_l a_{ik} b_{jl} \mathrm{Cov}(\mathbf{Z}_k, \mathbf{Z}_l) \\ &= \sum_k \sum_l a_{ik} b_{jl} \Sigma_{kl} = D_{ij}, \quad i = 1, \cdots, s, \quad j = 1, \cdots, t. \end{aligned}$$

□

2. (a) $\sum_{i=1}^n (X_i - \overline{X}) = 0, \quad \sum_{i=1}^n (X_i - \overline{X})^2 = \sum_{i=1}^n (X_i - \overline{X}) X_i = \sum_{i=1}^n X_i^2 - n(\overline{X})^2.$

Proof.

$$\begin{aligned} \sum_{i=1}^n (X_i - \overline{X}) &= \sum_{i=1}^n X_i - \sum_{i=1}^n \overline{X} = n\overline{X} - n\overline{X} = 0. \\ \sum_{i=1}^n (X_i - \overline{X})^2 &= \sum_{i=1}^n (X_i^2 - 2X_i\overline{X} + (\overline{X})^2) \\ &= \sum_{i=1}^n X_i^2 - 2\overline{X} \sum_{i=1}^n X_i + \sum_{i=1}^n (\overline{X})^2 \\ &= \sum_{i=1}^n X_i^2 - 2n(\overline{X})^2 + n(\overline{X})^2 \\ &= \sum_{i=1}^n X_i^2 - n(\overline{X})^2. \end{aligned}$$

□

(b) $\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^n (X_i - \overline{X}) Y_i = \sum_{i=1}^n X_i Y_i - n\overline{X}\ \overline{Y}.$

Proof.

$$\begin{aligned} \sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) &= \sum_{i=1}^n X_i Y_i - \sum_{i=1}^n X_i \overline{Y} - \sum_{i=1}^n \overline{X} Y_i + \sum_{i=1}^n \overline{X}\ \overline{Y} \\ &= \sum_{i=1}^n X_i Y_i - n\overline{X}\ \overline{Y} - n\overline{X}\ \overline{Y} + n\overline{X}\ \overline{Y} \\ &= \sum_{i=1}^n X_i Y_i - n\overline{X}\ \overline{Y}. \end{aligned}$$

□

3. (a) For a given line: $y = b_0 + b_1 x$, the *sum of squared vertical deviations* of the observations $\{(X_i, Y_i)\}_{i=1}^n$ from the corresponding points on the line is:

$$Q(b_0, b_1) = \sum_{i=1}^{n} \left(Y_i - (b_0 + b_1 X_i)\right)^2 .$$

The *least squares (LS) principle* is to fit the observed data by minimizing the sum of squared vertical deviations.

(b) *Proof.* From the lecture notes, the LS estimators can be derived by finding b_0 and b_1 which satisfy the normal equations:

$$nb_0 + b_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i, \tag{1}$$

$$b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i. \tag{2}$$

From equation (1),

$$b_0 = \overline{Y} - b_1 \overline{X}.$$

Using this in equation (2) we have

$$\begin{aligned} & b_0 n\overline{X} + b_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} X_i Y_i \\ \Rightarrow\ & n\overline{XY} + b_1 [\sum_{i=1}^{n} X_i^2 - n\overline{X}^2] = \sum_{i=1}^{n} X_i Y_i \\ \Rightarrow\ & b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \end{aligned}$$

Now from 2(a) and (b),

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}.$$

From equation (1),

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

□

(c)

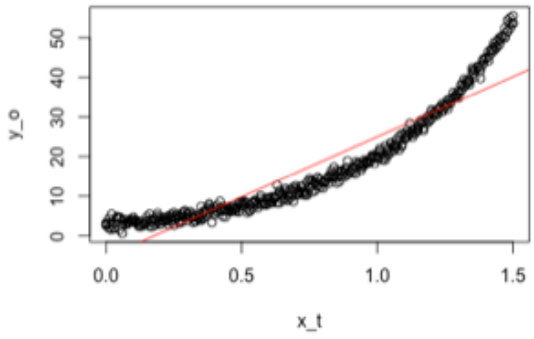
Solution. Plot implies linear regression model is not accurate for non-linear dataset.

$$Q(a,b) = \sum_{i=1}^{n} (Y_i - \exp(a + bX_i))^2 .$$

The *least squares (LS) principle* is to fit the observed data by minimizing the sum of squared vertical deviations:

$$(\hat{a}, \hat{b}) = \underset{a,b}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - \exp(a + bX_i))^2 .$$

□



4. (a) The least squares line always passes the center of the data $(\overline{X}, \overline{Y})$.

ANS. True. since $y = \overline{Y} + \hat{\beta}_1(x - \overline{X})$.

(b) If $\overline{X} = 0$, $\overline{Y} = 0$, then $\hat{\beta}_0 = 0$ no matter what is $\hat{\beta}_1$.

ANS. True since $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

(c) Given the sample size, the larger the sample variance of X_is, the smaller the standard errors of $\hat{\beta}_0, \hat{\beta}_1$ tend to be.

ANS. True. since $s\{\hat{\beta}_0\}$ and $s\{\hat{\beta}_1\}$ has $\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}$ in the denominator.

5. (a) $\sum_{i=1}^{n} e_i = 0$.

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} e_i &= \sum_{i=1}^{n}(Y_i - \overline{Y}) - \sum_{i=1}^{n} \hat{\beta}_1(X_i - \overline{X}) \\
&= (\sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} \overline{Y}) - \hat{\beta}_1(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \overline{X}) \\
&= (n\overline{Y} - n\overline{Y}) - \hat{\beta}_1(n\overline{X} - n\overline{X}) \\
&= 0.
\end{aligned}$$

□

(b) $\sum_{i=1}^{n} X_i e_i = 0$.

Proof.

$$\begin{aligned}
\sum_{i=1}^{n} X_i e_i &= \sum_{i=1}^{n} X_i((Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})) \\
&= (\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \overline{Y}) - \hat{\beta}_1(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} X_i \overline{X}) \\
&= (\sum_{i=1}^{n} X_i Y_i - n\overline{X}\ \overline{Y}) - \hat{\beta}_1(\sum_{i=1}^{n} X_i^2 - n(\overline{X})^2) \\
&= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) - \hat{\beta}_1 \sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= \sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) - \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \sum_{i=1}^{n}(X_i - \overline{X})^2 \\
&= 0.
\end{aligned}$$

□

(c) $\sum_{i=1}^{n} \widehat{Y}_i e_i = 0$.

Proof. By parts (a) and (b), and $\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$.

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \hat{\beta}_0 \sum_{i=1}^{n} e_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i e_i = 0 + 0 = 0.$$

□

6. *Proof.*

$$\begin{aligned} E(\hat{\beta}_0) &= E(\bar{Y}) - E(\hat{\beta}_1 \bar{X}) \\ &= \frac{1}{n}\sum(\beta_0 + \beta_1 X_i) - \bar{X}E(\hat{\beta}_1) \\ &= \beta_0 + \beta_1 \bar{X} - \bar{X}\beta_1 \qquad \text{(use the fact that } E(\hat{\beta}_1) = \beta_1) \\ &= \beta_0. \end{aligned}$$

□

HW1: Question 7

Wookyeong Song (mostly from Yan-Yu Chen)

2022/10/6

Simulation by R

You need to submit your codes alongside with the answers, plots, outputs, etc. For this homework, you are encouraged (though not required) to use R Markdown: Please submit a .rmd file and its corresponding .html file. Later in the quarter, you may be required to use R Markdown for some homework or quiz problems. *(Hint: Use the* ***help*** *function if needed)*

(a)

Create a sequence of consecutive integers ranging from 1 to 100. Record these in a vector x. *(Hint: use the* ***seq*** *function)*

```
x<-seq(1,100)
#x <- 1:100 #equivalent
```

(b)

Create a new vector w by the formula: $w = 2 + 0.5 * x$.

```
w<-2+0.5*x
```

(c)

Randomly sample 100 numbers from a Normal distribution with mean zero and standard deviation 5. Calculate the sample mean and sample variance and draw a histogram. What do you observe? *(Hint: use the* ***rnorm*** *function)*

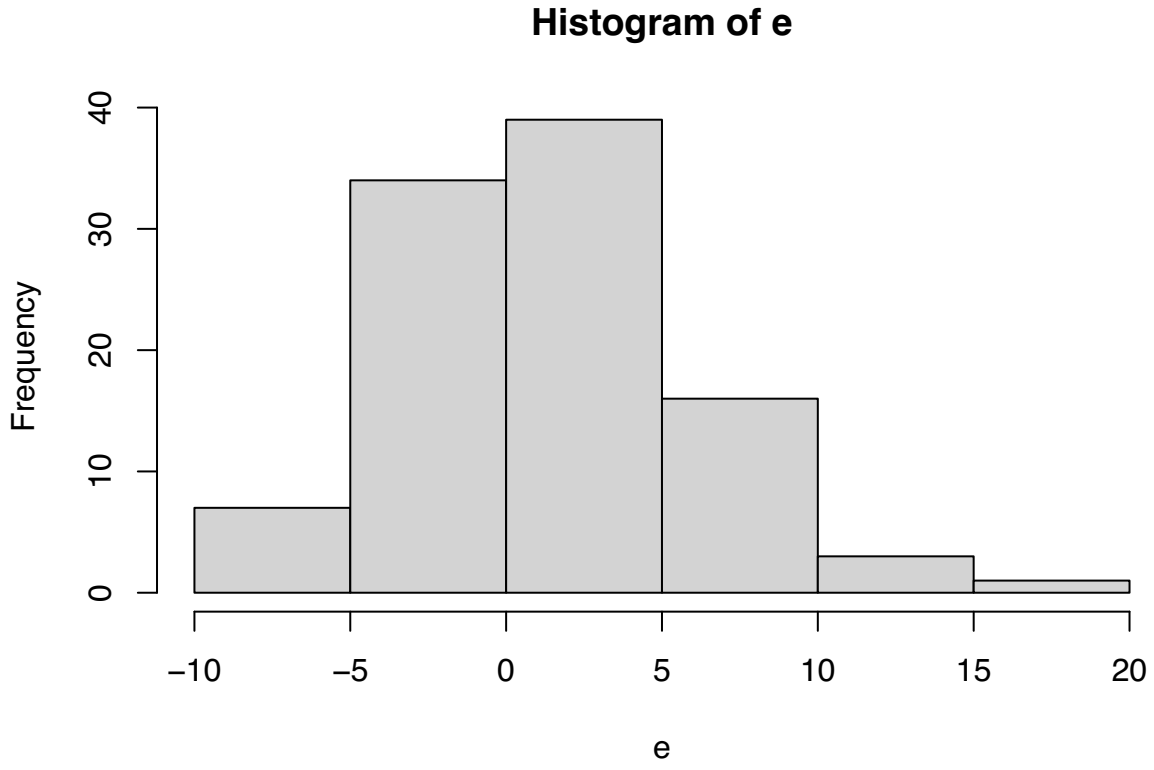
```
e <- rnorm(n = 100 , mean = 0 , sd = 5)
mean(e)
```

```
## [1] 1.036371
```

```
var(e)
```

```
## [1] 21.63843
```

```
hist(e)
```



We observe that the histogram is approximate to a normal distribution, and the sample mean and sample variance are not too far from their true values.

(d)

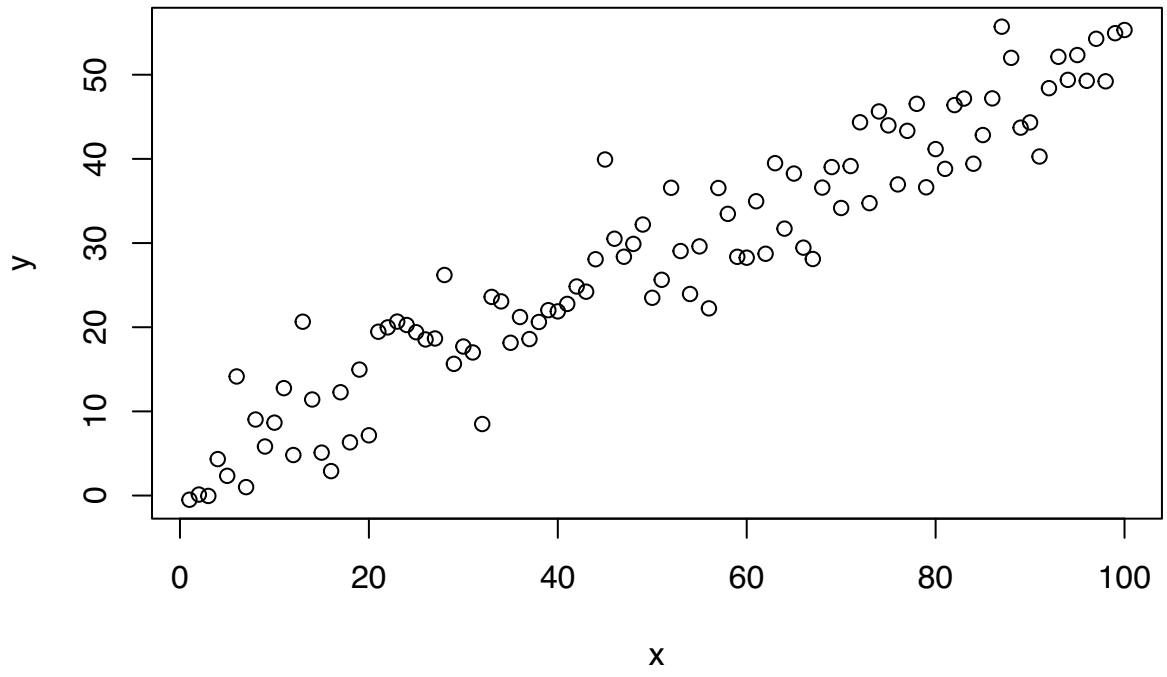
Add (element-wise) the numbers created in part (c) to the vector w. Record the new vector as y.

```
y<-w+e
```

(e)

Draw the scatter plot of y versus x. Make the axes labels 1.5 times as large as by default.

```
plot(x,y)
```



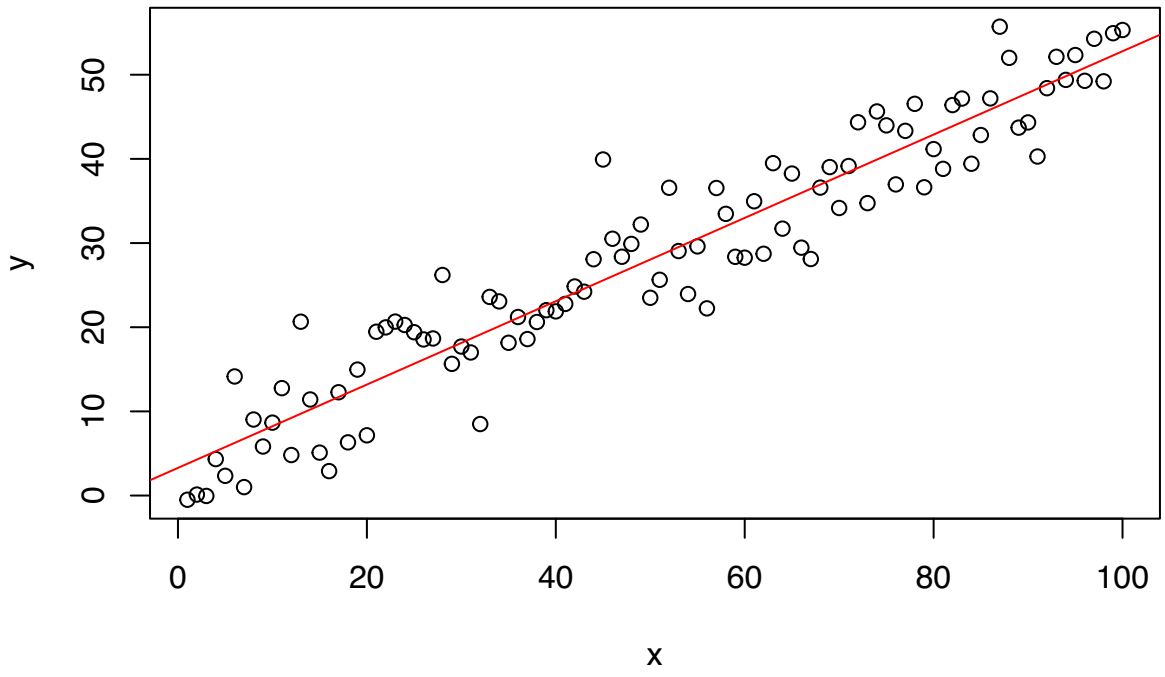
(f)

Estimate the regression coefficients of y on x. Add the fitted regression line to the scatter plot in part (e). What do you observe?

```
fit<- lm( y ~ x )
beta0 <- fit$coef[ 1 ]
beta1 <- fit$coef[2]
c(beta0,beta1)
```

```
## (Intercept)           x
##   3.2865876   0.4950452
```

```
plot (x, y)
abline ( a = beta0 , b = beta1 , col = "red")
```

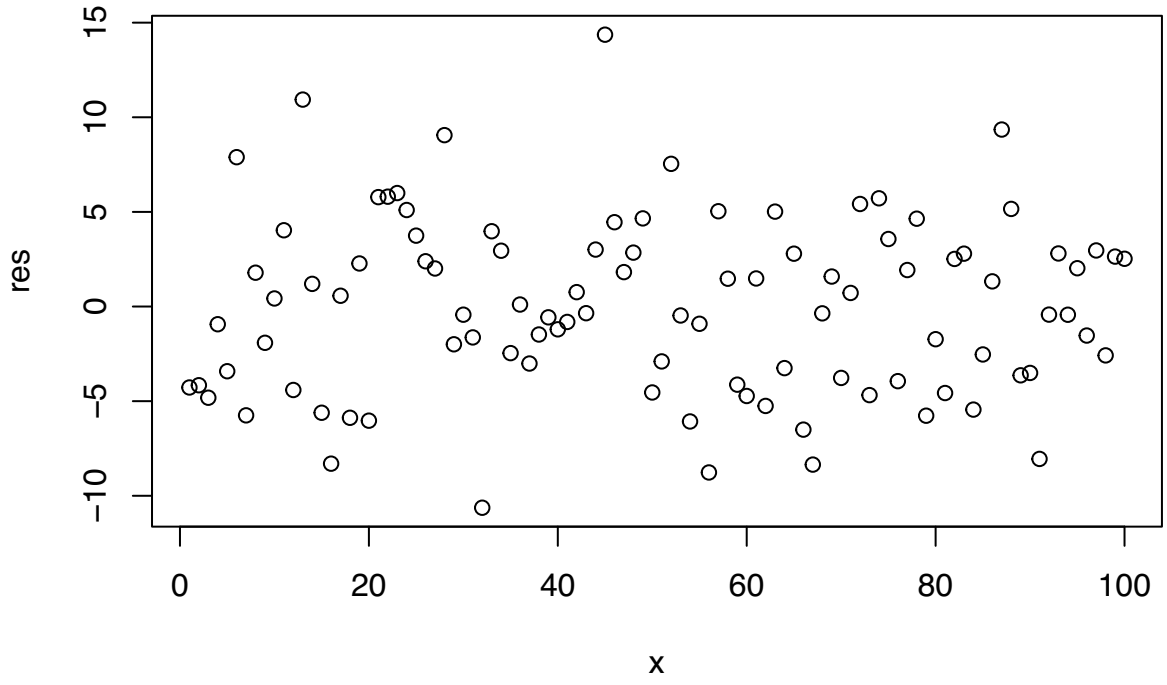


We observe the fitted regression line passing through the data cloud.

(g)

Calculate the residuals and draw a scatter plot of residuals versus x. What do you observe? Derive MSE.

```
res <- residuals (fit)
plot (x , res )
```




```
sse <- sum( res ^2)
mse <- sse /(100 - 2)
mse
```

```
## [1] 21.83835
```

We observe that residuals are randomly distributed, and the mean squared error is not too far from the true variance.

(h)

Repeat parts (c) – (g) a couple of times. What do you observe?

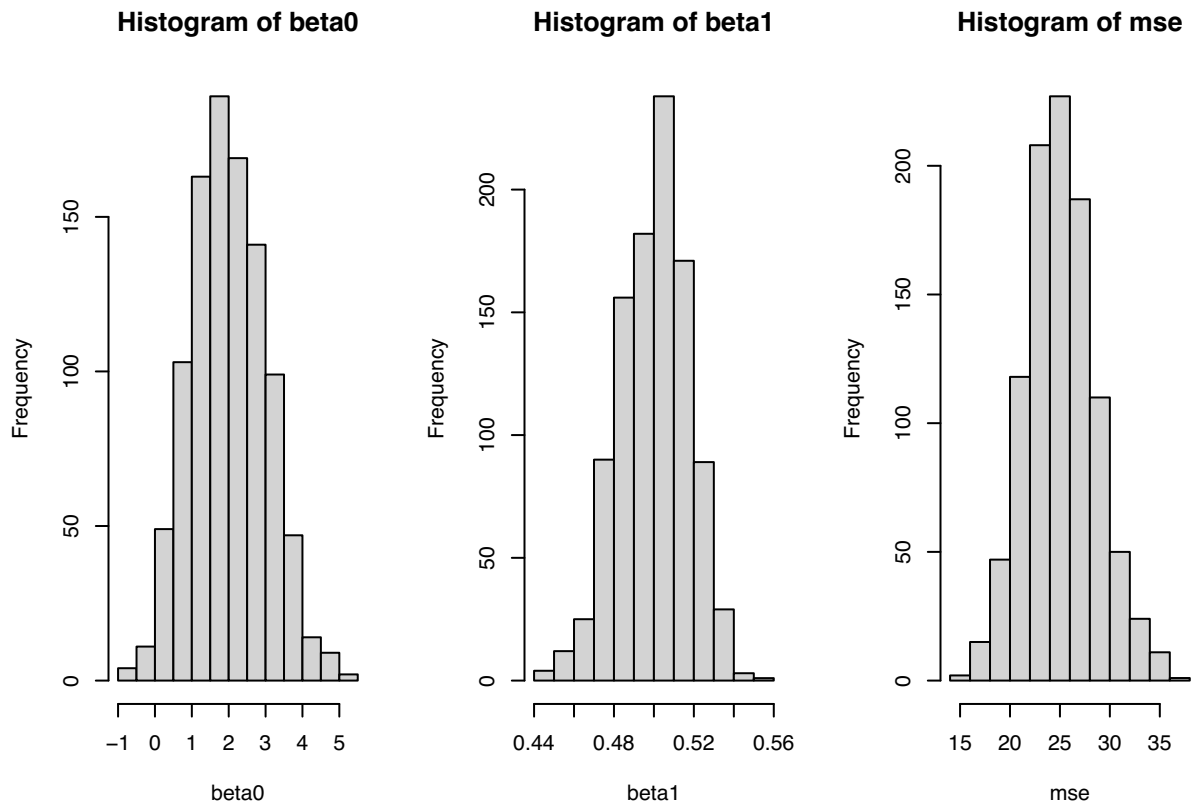
We observe the sampling variability.

(i)

(Optional Problem) Repeat parts (c) – (d) 1000 times. Each time, derive the fitted regression coefficients and MSE and record them. Draw histogram and calculate sample mean and sample variance for each of the three estimators. Summarize your observations. *(Hint: use the `for` loop)*

```
beta0 <- c ( )
beta1 <- c ( )
mse <- c ( )
for ( k in 1 : 1000 ) {
```

```
  e <- rnorm(n = 100 , mean = 0 , sd = 5)
  y <- w + e
  fit <- lm( y ~ x )
  beta0 [k] <- fit$coef[ 1 ]
  beta1 [k] <- fit$coef[ 2 ]
  res <- residuals (fit)
  mse[k]<-sum(res^2)/(100-2)
}
par(mfrow=c(1,3))
hist(beta0)
hist(beta1)
hist(mse)
```




```
mean(beta0)
```

```
## [1] 1.996409
```

```
mean(beta1)
```

```
## [1] 0.5000034
```

```
mean(mse)
```

```
## [1] 25.10441
```

```
var(beta0)
```

```
## [1] 1.037503
```

```
var(beta1)
```

```
## [1] 0.0002982367
```

```
var(mse)
```

```
## [1] 12.20248
```

The distribution of each of the three estimators is bell shaped, with sample mean close to true parameter.