HW1

김경민(20210344)

14 4월, 2024

Question 1

Data

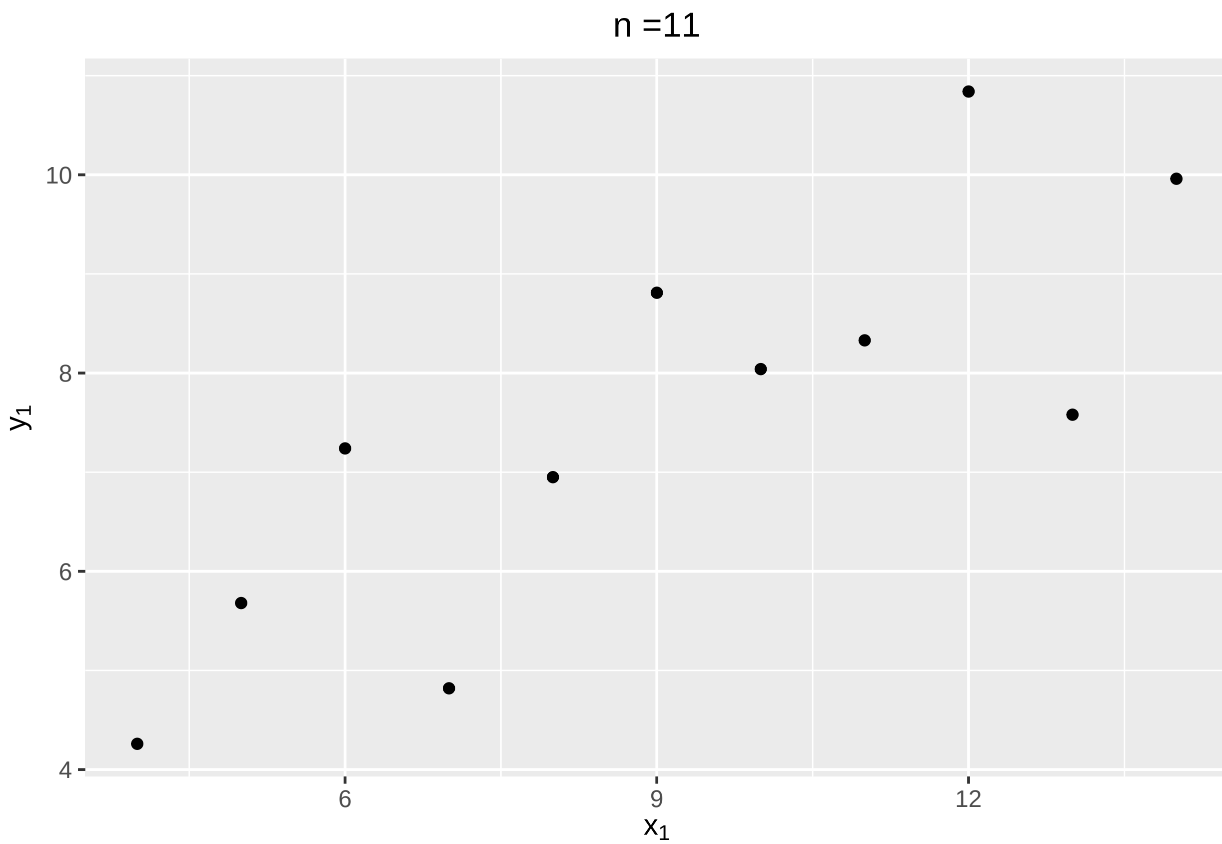
```r
library(ggplot2)
library(dplyr)

data("anscombe")
n = nrow(anscombe)
X1 = anscombe$x1
X2 = anscombe$x2
X3 = anscombe$x3
X4 = anscombe$x4
Y1 = anscombe$y1
Y2 = anscombe$y2
Y3 = anscombe$y3
Y4 = anscombe$y4
```

1.1 Plot the 4 data sets (x1, y1), (x2, y2), (x3, y3), (x4, y4) using ggplot2.

Plot (x1, y1)는 다음과 같다.

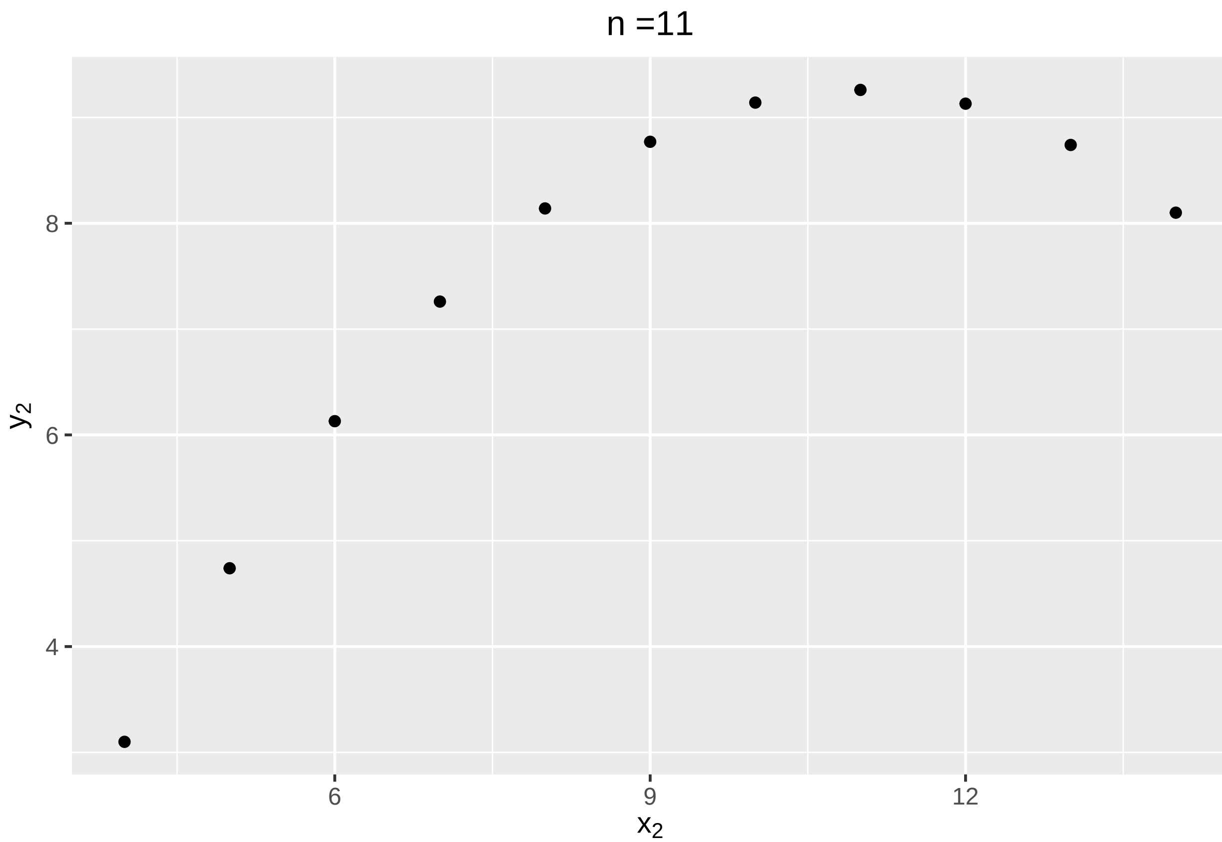
```
ggplot(data = anscombe) +
  geom_point(aes(x = x1, y = y1)) +
  xlab(bquote(x[1])) +
  ylab(bquote(y[1])) +
  ggtitle(paste0("n =", dim(anscombe %>% select(x1, y1))[1])) +
  theme(plot.title = element_text(hjust = 0.5))
```



선형식을 적합해도 괜찮아 보인다.

Plot (x2, y2)는 다음과 같다.

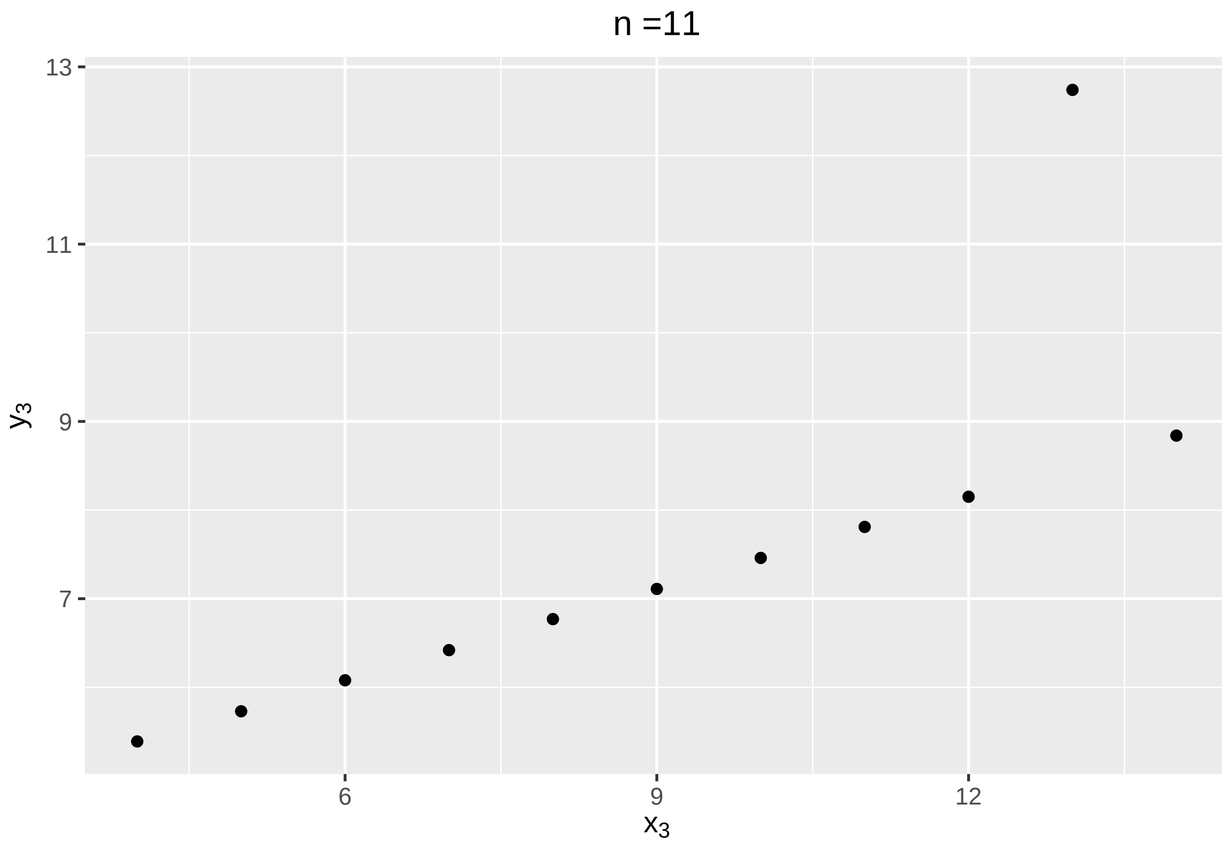
```
ggplot(data = anscombe) +
  geom_point(aes(x = x2, y = y2)) +
  xlab(bquote(x[2])) +
  ylab(bquote(y[2])) +
  ggtitle(paste0("n =", dim(anscombe %>% select(x2, y2))[1])) +
  theme(plot.title = element_text(hjust = 0.5))
```



포물선 형태라서 선형으로 적합하는 것은 적절해 보이지 않는다.

Plot (x3, y3)는 다음과 같다.

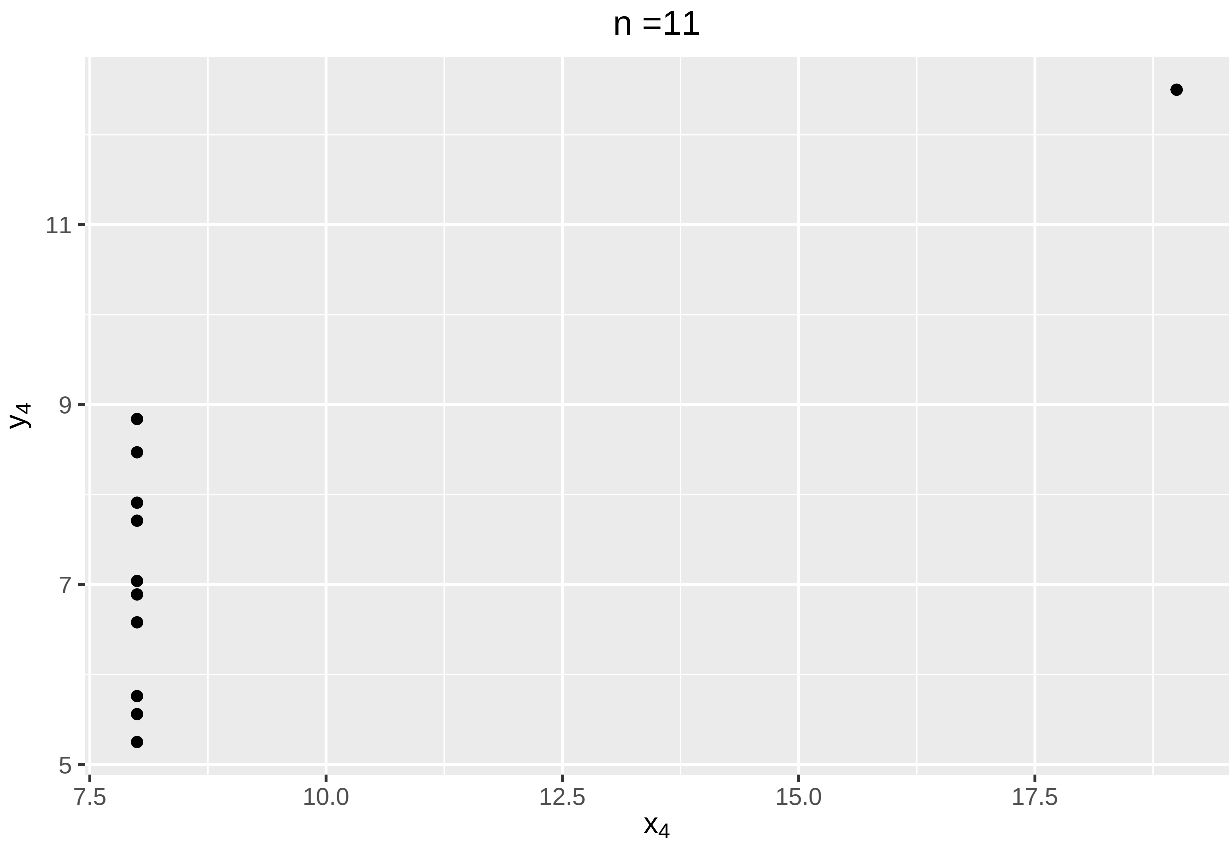
```
ggplot(data = anscombe) +
  geom_point(aes(x = x3, y = y3)) +
  xlab(bquote(x[3])) +
  ylab(bquote(y[3])) +
  ggtitle(paste0("n =", dim(anscombe %>% select(x1, y1))[1])) +
  theme(plot.title = element_text(hjust = 0.5))
```



이상치가 존재해서 주의해야 한다.

Plot (x4, y4)는 다음과 같다.

```
ggplot(data = anscombe) +
  geom_point(aes(x = x4, y = y4)) +
  xlab(bquote(x[4])) +
  ylab(bquote(y[4])) +
  ggtitle(paste0("n =", dim(anscombe %>% select(x1, y1))[1])) +
  theme(plot.title = element_text(hjust = 0.5))
```



데이터가 고르지 않다. X_4를 범주형 변수로 고려해볼 수 있을 것 같다.

1.2 Fit a regression model to the data sets

1.2.a y1 ~ x1

변수 설명

- Sxx1: $S_{x_1 x_1}$
- Syy1: $S_{y_1 y_1}$
- Sxy1: $S_{x_1 y_1}$
- beta1_hat1: y1 ~ x1에서의 $\hat{\beta}_1$
- beta0_hat1: y1 ~ x1에서의 $\hat{\beta}_0$

```
mean.X1 = sum( X1 ) / n
mean.Y1 = sum( Y1 ) / n

Sxx1 = sum( ( X1 - mean.X1 )^2 )
Syy1 = sum( ( Y1 - mean.Y1 )^2 )
Sxy1 = sum( ( X1 - mean.X1 ) * ( Y1 - mean.Y1 ) )

beta1_hat1 = Sxy1 / Sxx1
beta0_hat1 = mean.Y1 - beta1_hat1 * mean.X1


data.frame( beta0_hat1, beta1_hat1 )
```

```
##   beta0_hat1 beta1_hat1
## 1   3.000091  0.5000909
```

적합된 식은 $\hat{Y}_1 = \hat{\beta}_0 + \hat{\beta}_0 X_1 = 3.000091 + 0.5000909 X_1$ 이다.

1.2.b y2 ~ x2

변수 설명

- Sxx2: $S_{x_2x_2}$
- Syy2: $S_{y_2y_2}$
- Sxy2: $S_{x_2y_2}$
- beta1_hat2: y2 ~ x2에서의 $\hat{\beta}_1$
- beta0_hat2: y2 ~ x2에서의 $\hat{\beta}_0$

```
mean.X2 = sum ( X2 ) / n
mean.Y2 = sum ( Y2 ) / n

Sxx2 = sum( ( X2 - mean.X2 )^2 )
Syy2 = sum( ( Y2 - mean.Y2 )^2 )
Sxy2 = sum( ( X2 - mean.X2) * ( Y2 - mean.Y2) )

beta1_hat2 = Sxy2 / Sxx2
beta0_hat2 = mean.Y2 - beta1_hat2 * mean.X2


data.frame( beta0_hat2, beta1_hat2 )
```

```
##   beta0_hat2 beta1_hat2
## 1   3.000909        0.5
```

적합된 식은 $\hat{Y}_2 = \hat{\beta}_0 + \hat{\beta}_0 X_2 = 3.000909 + 0.5X_2$ 이다.

1.2.c y3 ~ x3

변수 설명

- Sxx3: $S_{x_3 x_3}$
- Syy3: $S_{y_3 y_3}$
- Sxy3: $S_{x_3 y_3}$
- beta1_hat3: y3 ~ x3에서의 $\hat{\beta}_1$
- beta0_hat3: y3 ~ x3에서의 $\hat{\beta}_0$

```
mean.X3 = sum ( X3 ) / n
mean.Y3 = sum ( Y3 ) / n

Sxx3 = sum( ( X3 - mean.X3 )^2 )
Syy3 = sum( ( Y3 - mean.Y3 )^2 )
Sxy3 = sum( ( X3 - mean.X3) * ( Y3 - mean.Y3) )

beta1_hat3 = Sxy3 / Sxx3
beta0_hat3 = mean.Y3 - beta1_hat3 * mean.X3


data.frame( beta0_hat3, beta1_hat3 )
```

```
##   beta0_hat3 beta1_hat3
## 1   3.002455  0.4997273
```

적합된 식은 $\hat{Y}_3 = \hat{\beta}_0 + \hat{\beta}_0 X_3 = 3.002455 + 0.4997273 X_3$ 이다.

1.2.d y4 ~ x4

변수 설명

- Sxx4: $S_{x_4x_4}$
- Syy4: $S_{y_4y_4}$
- Sxy4: $S_{x_4y_4}$
- beta1_hat4: y4 ~ x4에서의 $\hat{\beta}_1$
- beta0_hat4: y4 ~ x4에서의 $\hat{\beta}_0$

```
mean.X4 = sum ( X4 ) / n
mean.Y4 = sum ( Y4 ) / n

Sxx4 = sum( ( X4 - mean.X4 )^2 )
Syy4 = sum( ( Y4 - mean.Y4 )^2 )
Sxy4 = sum( ( X4 - mean.X4) * ( Y4 - mean.Y4) )

beta1_hat4 = Sxy4 / Sxx4
beta0_hat4 = mean.Y4 - beta1_hat4 * mean.X4


data.frame( beta0_hat4, beta1_hat4 )
```

```
##   beta0_hat4 beta1_hat4
## 1   3.001727  0.4999091
```

적합된 식은 $\hat{Y}_4 = \hat{\beta}_0 + \hat{\beta}_0 X_4 = 3.001727 + 0.4999091 X_4$ 이다.

1.3 Compute the sample correlation for each data set. Are they the same?

```r
correlation1 = Sxy1 / sqrt( Sxx1 * Syy1 ) # Corr(X1, Y1)
correlation2 = Sxy2 / sqrt( Sxx2 * Syy2 ) # Corr(X2, Y2)
correlation3 = Sxy3 / sqrt( Sxx3 * Syy3 ) # Corr(X3, Y3)
correlation4 = Sxy4 / sqrt( Sxx4 * Syy4 ) # Corr(X4, Y4)

data.frame(correlation1, correlation2, correlation3, correlation4)
```

```
##   correlation1 correlation2 correlation3 correlation4
## 1    0.8164205    0.8162365    0.8162867    0.8165214
```

4개의 데이터 셋에서 sample correlation은 각각 0.8164205, 0.8162365, 0.8162867, 0.8165214이고 이들은 거의 같다고 할 수 있다.

1.4 Compute the SSE, SST and R2 value for each data set. Are they the same?

```
# y1 ~x1에서의 SSE와 SST
SSE1 = sum( ( Y1 - beta0_hat1 - beta1_hat1 * X1)^2 ) # y1 ~x1에서의 SSE
SST1 = sum( ( Y1 - mean.Y1)^2 ) # y1 ~x1에서의 SST
R2_1 = 1 - SSE1 / SST1 # # y1 ~x1에서의 R^2

# y2 ~x2에서의 SSE와 SST
SSE2 = sum( ( Y2 - beta0_hat2 - beta1_hat2 * X2)^2 ) # y2 ~x2에서의 SSE
SST2 = sum( ( Y2 - mean.Y2)^2 ) # y2 ~x2에서의 SST
R2_2 = 1 - SSE2 / SST2 # y2 ~x2에서의 R^2

# y3 ~x3에서의 SSE와 SST
SSE3 = sum( ( Y3 - beta0_hat3 - beta1_hat3 * X3 )^2 ) # y3 ~x3에서의 SSE
SST3 = sum( ( Y3 - mean.Y3)^2 ) # y3 ~x3에서의 SST
R2_3 = 1 - SSE3 / SST3 # y3 ~x3에서의 R^2

# y4 ~x4에서의 SSE와 SST
SSE4 = sum( ( Y4 - beta0_hat4 - beta1_hat4 * X4)^2 ) # y4 ~x4에서의 SSE
SST4 = sum( ( Y4 - mean.Y4)^2 ) # y4 ~x4에서의 SST
R2_4 = 1 - SSE4 / SST4 # y4 ~x4에서의 R^2


data.frame(SSE1, SSE2, SSE3, SSE4)
```

```
##       SSE1     SSE2     SSE3     SSE4
## 1 13.76269 13.77629 13.75619 13.74249
```

```
data.frame(SST1, SST2, SST3, SST4)
```

```
##       SST1     SST2    SST3     SST4
## 1 41.27269 41.27629 41.2262 41.23249
```

```
data.frame(R2_1, R2_2, R2_3, R2_4)
```

```
##        R2_1     R2_2     R2_3      R2_4
## 1 0.6665425 0.666242 0.666324 0.6667073
```

4개의 데이터 셋에서 SSE, SSR, R^2 은 각각 모두 같다고 볼 수 있다.

Question 3

반응변수 Y는 1972년의 여성 노동 인구 참여율이고, 설명변수 X는 1968년의 여성 노동 인구 참여율이다. 데이터는 미국의 19개의 도시에서 수집되었다. 또한 $SSR = 0.0358, SSE = 0.0544$이다. 그리고 $SST = SSE + SSR$임을 이용해 다음과 같이 변수를 설정할 수 있다(n: 데이터 갯수, SSE: SSE, SSR: SSR, SST: SST, se_beta1: $s.e.(\hat{\beta}_1)$,se_beta0: $s.e.(\hat{\beta}_0)$, beta1_hat: $\hat{\beta}_1$, beta0_hat: $\hat{\beta}_0$).

```
n = 19
SSR = 0.0358
SSE = 0.0544
SST = SSE + SSR

beta1_hat = 0.656040
beta0_hat = 0.203311

se_beta1_hat = 0.1961
se_beta0_hat = 0.0976
```

적합된 모형은 $Y = 0.20311 + 0.65040X$이다.

1. Compute the sample variance of Y and the sample covariance betwen Y and X.

$S_{yy} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 = SST$이므로 다음과 같다(Syy: S_{yy}).

```
Syy = SST
```

그리고 Y의 표본분산은 $\dfrac{S_{yy}}{n-1}$이다. 따라서 다음과 같다.

```
sample_variance_Y = Syy / ( n - 1 ) # Y의 표본분산
sample_variance_Y
```

```
## [1] 0.005011111
```

그리고 $s.e.(\hat{\beta}_1) = \sqrt{\hat{\sigma^2} \times \dfrac{1}{S_{xx}}} = \sqrt{\dfrac{SSE}{n-2} \times \dfrac{1}{S_{xx}}}$이다. 따라서 $S_{xx} = \sqrt{\dfrac{SSE}{n-2} \times \dfrac{1}{s.e.(\hat{\beta}_1)^2}}$는 다음과 같다(Sxx: S_{xx}).

```
Sxx = sqrt( SSE / (( n - 1 )* ( se_beta1_hat^2 )) )
Sxx
```

```
## [1] 0.2803403
```

또한, $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$이므로 S_{xy}는 다음과 같다(Sxy: S_{xy}).

```
Sxy = Sxx * beta1_hat
Sxy
```

```
## [1] 0.1839145
```

표본공분산은 $\dfrac{S_{xy}}{n-1}$이므로 계산하면 다음과 같다.

```
sample_covariance_Y = Sxy / ( n - 1 )
sample_covariance_Y
```

```
## [1] 0.01021747
```

```
data.frame(sample_variance_Y, sample_covariance_Y)
```

```
##   sample_variance_Y sample_covariance_Y
## 1       0.005011111          0.01021747
```

결과적으로 Y의 표본 분산은 0.005011111이고, X와 Y 사이의 표본 공분산은 0.01021747이다.

2. Suppose participation rate of women in 1968 in a given city is 45%. What is the estimated participation rate of women in 1972 for the same city?

1968 년의 여성 노동 인구 참여율이 $x_0 = 0.45$ 일 경우 1972 년의 추정된 여성 노동 인구 참여율 $\hat{\mu_0}$ 은 다음과 같다. (`mu0_hat`: $\hat{\mu_0}$)

```
x0 = 0.45
mu0_hat = beta0_hat + beta1_hat * x0
mu0_hat
```

```
## [1] 0.498529
```

따라서 1972 년의 추정된 여성 노동 인구 참여율은 $\hat{\mu}_0 = 0.498529$ 이다.

3. Suppose further that the mean and variance of the participantion rate of women in 1968 are 0.5 and 0.005, respectively. Construct the 95% confidence interval for the estimate in (2).

$X = x_0$ 로 주어졌을 때, 1972년의 추정된 노동인구 참여율의 표준오차는 $s.e.(\hat{\mu}_0) = \sqrt{\hat{\sigma}^2(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}})}$ 인데 S_{xx} 자리에는 $0.005 \times (n-1)$ 을 넣고 $\bar{x}$ 자리에는 0.5를 넣어서 $\hat{\mu}_0$ 의 95% 신뢰구간을 구하면 다음과 같다.

```
se_mu0_hat = sqrt(( SSE / ( n - 1 )) * (1/n + (x0 - 0.5)^2 / (0.005*(n-1))) )

LB_mu0 = mu0_hat - qt(0.975, df = n-2) * se_mu0_hat
UB_mu0 = mu0_hat + qt(0.975, df = n-2) * se_mu0_hat

data.frame( LB_mu0, UB_mu0 )
```

```
##      LB_mu0    UB_mu0
## 1 0.4656392 0.5314188
```

$\hat{\mu}_0$ 의 95% 신뢰구간은 [0.4656392, 0.5314188] 이다.

4. Construct the 95% confidence interval for the slope of the true regression line.

```
LB_slope = beta1_hat - qt(0.975, df = n-2) * se_beta1_hat # 신뢰 하한
UB_slope = beta1_hat + qt(0.975, df = n-2) * se_beta1_hat # 신뢰 상한


data.frame( LB_slope, UB_slope )
```

```
##    LB_slope UB_slope
## 1 0.2423052 1.069775
```

$\hat{\beta}_1$ 의 95% 신뢰구간은 [0.2423052, 1.069775] 이다.

5. Test the hypothesis

\$ H_0: *1 = 1\$ versus \$ H*: _1 1\$ at the 5% significance level.

```
t_value = ( beta1_hat - 1 ) / se_beta1_hat
p_value = 2 *pt(t_value, df = n - 2 )

p_value < 0.05
```

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

p값이 0.05보다 크기 때문에 귀무가설을 기각할 수 없다.

6. Compute the R2 for this simple linear regression.

```
R_square = SSR / SST # R^2
R_square
```

```
## [1] 0.3968958
```

R^2 은 0.3968958이다.

7. If and were reversed in the above regression, what would you expect R2 to be?

X 와 Y 가 바뀌더라도 R^2 는 똑같을 것 같다. 단순선형회귀에서는 R^2 가 번응변수와 예측변수 사이의 상관계수의 제곱이기 때문에 X와 Y가 바뀌어도 같을 것이다다.

$$SSE = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x}))^2 = \hat{\beta}_1^2 S_{xx}$$

$$R^2 = \frac{SSE}{SST} = \frac{\hat{\beta}_1^2 S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}^2}\frac{S_{xx}}{S_{yy}} = \frac{S_{xy}^2}{S_{xx}S_{yy}} = Corr(X,Y)^2$$