Bài 3: Sử dụng bảng số liệu cho như sau:

1. Viết các mô hình tuyến tính với 2 biến độc lập (có thể).

2. Ước lượng các hệ số hồi quy trong từng mô hình tuyến tính ở câu 1.

3. Với độ tin cậy 95%, tìm khoảng tin cậy cho các tham số trong mô hình với 2 biến độc lập x1 và x2.

4. Xác định hệ số xác định cho mỗi mô hình trong câu 1.

5. Trong các mô hình trên, mô hình nào thích hợp nhất để giải thích sự biến thiên của Y ?

6. Viết mô hình tuyến tính dưới dạng ma trận với số biến độc lập nhiều nhất có thể, và xác

định kích thước của ma trận.

7. Ước lượng các hệ số hồi quy trong mô hình tuyến tính ở câu 6.

8. Trong mô hình tuyến tính ở câu 6, tính ước lượng của $V(\epsilon)$ và $V(\hat{\beta})$.

9. Với độ tin cậy 95%, tìm khoảng tin cậy cho $V(\epsilon)$.

10. Khi thêm 2 biến độc lập x3 và x2 vào mô hình chỉ với 1 biến độc lập x1 thì làm cho chất lượng ước lượng cao hơn không?

y	x1	x2	x3
12	2	45	121
14	1	43	132
10	3	43	154
16	6	47	145
14	7	42	129
19	8	21	156
21	8	32	132
19	5	33	147
21	5	41	128
16	8	38	163
19	4	32	161
21	9	31	172
25	12	35	174
21	7	29	180

Tải dữ liệu vào R:

```
> library(readxl)

> bai3 <- read_excel("bai3.xlsx")
```

```
> View(bai3)
```

1. Viết mô hình tuyến tính với hai biến độc lập và ước lượng hệ số hồi quy

```
> y <- bai3$y

> x1 <- bai3$x1

> x2 <- bai3$x2

> x3 <- bai3$x3
```

- Mô hình hồi quy tuyến tính y theo biến độc lập x1, x2:

```
> M12 <- lm(y ~ x1 + x2)
> coef(M12)
(Intercept)          x1          x2
 25.8421378   0.7148959  -0.3281129
```

Kết quả fit: $y = 25.8421378 + 0.7148959 \times x1 - 0.3281129 \times x2 + \in$
Ước lượng hệ số hồi quy:
$\hat{\beta}_0 = 25.8421378$
$\hat{\beta}_1 = 0.7148959$
$\hat{\beta}_2 = -0.3281129$

- Mô hình hồi quy tuyến tính y theo biến độc lập x2, x3:

```
> M23 <- lm(y ~ x2 + x3)
> coef(M23)
(Intercept)          x2          x3
31.97642386 -0.45389541  0.01996295
```

Kết quả fit: $y = 31.97642386 - 0.45389541 \times x2 + 0.01996295 \times x3 + \in$
Ước lượng hệ số hồi quy:
$\hat{\beta}_0 = 31.97642386$
$\hat{\beta}_2 = -0.45389541$
$\hat{\beta}_3 = 0.01996295$

- Mô hình hồi quy tuyến tính y theo biến độc lập x1, x3:

```
> M13 <- lm(y ~ x1 + x3)
> coef(M13)
(Intercept)          x1          x3
 8.60924098  0.92720866  0.02323681
```

Kết quả fit: $y = 8.60924098 + 0.92720866 \times x1 + 0.02323681 \times x3 + \in$
Ước lượng hệ số hồi quy:
$\hat{\beta}_0 = 8.60924098$
$\hat{\beta}_1 = 0.92720866$
$\hat{\beta}_3 = 0.02323681$

3. Với độ tin cậy 95%, tìm khoảng tin cậy cho các tham số trong mô hình với 2 biến độc lập x1 và x2.

```
> confint(M12)
                2.5 %     97.5 %
(Intercept) 12.4938794 39.1903962
x1           0.1288532  1.3009387
x2          -0.6242802 -0.0319457
```

Dựa vào kết quả trên cho thấy:

- Với độ tin cậy 95%, x1 nằm trong khoảng giá trị từ 0.1288532 đến 1.3009387
- Với độ tin cậy 95%, x2 nằm trong khoảng giá trị từ 0-0.6242802 đến -0.0319457

4. Xác định hệ số xác định cho mỗi mô hình trong câu 1.

```
> summary(M12)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
   Min     1Q  Median     3Q    Max
-3.8780 -0.9552  0.1747  1.1902  5.0360

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  25.8421     6.0647   4.261  0.00134 **
x1            0.7149     0.2663   2.685  0.02122 *
x2           -0.3281     0.1346  -2.438  0.03292 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.539 on 11 degrees of freedom
Multiple R-squared:  0.6875,      Adjusted R-squared:  0.6307
F-statistic:  12.1 on 2 and 11 DF,  p-value: 0.001665
```

Hệ số xác định của mô hình hồi quy tuyến tính y theo x1, x2:

$R^2 = 68.75\%$

```
> summary(M23)


Call:
lm(formula = y ~ x2 + x3)

Residuals:
   Min     1Q Median     3Q    Max
-5.533 -1.621 -1.013  2.075  5.436

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

```
(Intercept) 31.97642   14.58671   2.192   0.0508 .
x2          -0.45390    0.19298  -2.352   0.0383 *
x3           0.01996    0.05941   0.336   0.7432
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.249 on 11 degrees of freedom
Multiple R-squared:  0.488,        Adjusted R-squared:  0.3949
F-statistic: 5.243 on 2 and 11 DF,  p-value: 0.02517
```

Hệ số xác định của mô hình hồi quy tuyến tính y theo x2, x3:

$R^2 = 48.8\%$

```
> summary(M13)

Call:
lm(formula = y ~ x1 + x3)

Residuals:
   Min      1Q  Median      3Q     Max
-4.9693 -1.4752  0.6351  1.8588  4.7804

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.60924    7.28437   1.182   0.2622
x1           0.92721    0.35378   2.621   0.0238 *
x3           0.02324    0.05505   0.422   0.6811
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.126 on 11 degrees of freedom
Multiple R-squared:  0.5263,      Adjusted R-squared:  0.4402
F-statistic: 6.111 on 2 and 11 DF,  p-value: 0.01641
```

Hệ số xác định của mô hình hồi quy tuyến tính y theo x1, x3:

$R^2 = 52.63\%$

5. Trong các mô hình trên, mô hình nào thích hợp nhất để giải thích sự biến thiên của Y ?

Mô hình tuyến tính của y theo hai biến X1 và X2 có p-value = 0.001665 < 0.05, do đó ta bác bỏ giả thuyết H0 là Y không được giải thích bởi X1 và X2. Tức là Y được giải thích với ít nhất một trong các biến X1 và X2 với mức ý nghĩa 5%, hệ số xác định R^2 = 68.75%, tức có 68.75% Y được giải thích với X1 và X2.

Mô hình tuyến tính của y theo hai biến X2 và X3 có p-value = 0.02517 < 0.05, do đó ta bác bỏ giả thuyết H0 là Y không được giải thích bởi X2 và X3. Tức là Y được giải thích với ít nhất một trong các biến X2 và X3 với mức ý nghĩa 5%, hệ số xác định R^2 = 0.448, tức có 44.8% Y được giải thích với X2 và X3.

Mô hình tuyến tính của y theo hai biến X1 và X3 có p-value = 0.01641 **< 0.05, do đó ta bác bỏ giả thuyết H0 là Y không được giải thích bởi X1 và X3. Tức là Y được giải thích với ít nhất một trong**

các biến X1 và X3 với mức ý nghĩa 5%, hệ số xác định R^2 = 0.5263, tức có 52.63% Y được giải thích với X1 và X3.

Như vậy, mô hình hồi quy tuyến tính theo hai biến độc lập x1, x2 phù hợp nhất để giải thích cho sự biến thiên của y.

6. Viết mô hình tuyến tính dưới dạng ma trận với số biến độc lập nhiều nhất có thể, và xác

định kích thước của ma trận.

Mô hình hồi quy tuyến tính cho ba biến độc lập x1, x2, x3

```
model <- lm(y ~ x1 + x2 + x3)

summary(model)

Call:

lm(formula = y ~ x1 + x2 + x3)


Residuals:

    Min      1Q  Median      3Q     Max

-3.6973 -1.1259  0.1907  1.4846  4.4880


Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept) 32.89132   11.66331   2.820   0.0182 *

x1           0.80190    0.29844   2.687   0.0228 *

x2          -0.38136    0.15658  -2.436   0.0351 *

x3          -0.03713    0.05202  -0.714   0.4917

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 2.597 on 10 degrees of freedom

Multiple R-squared: 0.7027, Adjusted R-squared: 0.6135

F-statistic: 7.878 on 3 and 10 DF, p-value: 0.005452

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_{14} \end{pmatrix} \text{ kích thước } (14 \times 1)$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \dots & x_{13} \\ 1 & x_{21} & \dots & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{141} & \dots & x_{143} \end{pmatrix} \text{ kích thước } (14 \times 4)$$

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \text{ kích thước } (4 \times 1)$$

$$\boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_{14} \end{pmatrix} \text{ kích thước } (14 \times 1)$$

7. Ước lượng các hệ số hồi quy trong mô hình tuyến tính cho 3 biến độc lập x1, x2, x3

> coef(M3)

(Intercept) x1 x2 x3

32.89132428 0.80190069 -0.38136236 -0.03713244

Kết quả fit: y = 32.89132428 + 0.80190069×x1 - 0.38136236×x2 - 0.03713244×x3 + ∈
Ước lượng hệ số hồi quy:
$\hat{\beta}_0$ = 32.89132428
$\hat{\beta}_1$ = 0.80190069
$\hat{\beta}_2$ = -0.38136236
$\hat{\beta}_3$ = -0.03713244

8. Trong mô hình tuyến tính ở câu 6, tính ước lượng của V(ϵ) và V($\hat{\beta}$).

out <- summary(model)

```
se_B0 <- out$coefficients[1,2]
#11.66331
V_B0 <- (se_B0)**2
#136.0328
```

$$\text{Var}(\widehat{\beta_0}) = s(\widehat{\beta_0})^2 = 136.0328$$

```
se_B1 <- out$coefficients[2,2]
#0.2984358
V_B1 <- (se_B1)**2
#0.0.08906395
```

$$\text{Var}(\widehat{\beta_1}) = s(\widehat{\beta_1})^2 = 0.0.08906395$$

```
se_B2 <- out$coefficients[3,2]
#0.1565807
V_B2 <- (se_B2)**2
#0.02451751
```

$$\text{Var}(\widehat{\beta_2}) = s(\widehat{\beta_2})^2 = 0.02451751$$

```
se_B3 <- out$coefficients[4,2]
#0.05202312
V_B3 <- (se_B3)**2
#0.002706406
```

$$\text{Var}(\widehat{\beta_3}) = s(\widehat{\beta_3})^2 = 0.002706406$$

```
s_square <- anova_model$`Mean Sq`[4]
s_square
#6.744767
```

$S^2 = \frac{SSE}{n-p}$ là ước lượng không chệch cho $\sigma^2 = Var(\varepsilon_i) = Var(Y_i)$

9. Với độ tin cậy 95%, tìm khoảng tin cậy cho V(ϵ).

10. Khi thêm 2 biến độc lập x3 và x2 vào mô hình chỉ với 1 biến độc lập x1 thì làm cho chất lượng ước lượng cao hơn không?

Mô hình hồi quy tuyến tính theo x1

M1 <- lm(y ~ x1)

Bảng ANOVA cho mô hình hồi quy tuyến tính M1

anova(M1)

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	117.66	117.66	12.93	0.003674	**
Residuals	12	109.20	9.10			

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Mô hình hồi quy tuyến tính theo ba biến độc lập x1, x2, x3:

model <- lm(y ~ x1 + x2 + x3)

Bảng ANOVA cho mô hình hồi quy tuyến tính theo ba biến

> anova(model)

Analysis of Variance Table

Response: y

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
x1	1	117.659	117.659	17.4445	0.001898	**

x2 1 38.314 38.314 5.6806 0.038389 *

x3 1 3.436 3.436 0.5095 0.491694

Residuals 10 67.448 6.745

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Giả thuyết:

H_0: Mô hình một biến x1

H_1: Mô hình theo ba biến độc lập x1, x2, x3

Ta tính giá trị F_{obs} và so sánh nó với F(0.05, r, n-p)

Tính giá trị thống kê F dựa trên công thức:

$$F = \frac{(SSE(H_0) - SSE(H_1))/r}{SSE(H_\cdot)/(n-p)}$$

```
SSE_H0 <- anova_M1$`Sum Sq`[2]

df_H0 <- anova_M1$Df[2]

SSE_H1 <- anova_model$`Sum Sq`[4]

df_H1 <- anova_model$Df[4]

r <- (df_H0 - df_H1)

F_obs <- ((SSE_H0 - SSE_H1)/r)/(SSE_H1/df_H1)

#3.095036

F_val <- qt(1-0.05,r,df_H1)

F_obs >= F_val

#FALSE
```

Sau khi so sánh kết quả và nhận thấy F_{obs} < F(0.05,2,10), do đó với mức nghĩa 5%, không đủ cơ sở để bác bỏ giả thuyết H_0, nghĩa là mô hình hồi quy tuyến tính với 1 biến x1 cho chất lượng ước lượng cao hơn.