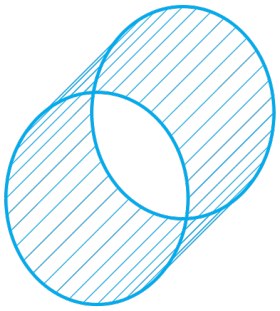
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP.HCM
CAO HỌC KHÓA 30



Khoa Toán - Tin học
Fac. of Math. & Computer Science

Bài tập lần 2

MÔ HÌNH HÓA THỐNG KÊ

Giảng viên hướng dẫn: **TS. Nguyễn Thị Mộng Ngọc**
Nhóm thực hiện: **Nhóm 4**

TP. Hồ Chí Minh – Tháng 01, 2021

BẢNG PHÂN CÔNG CÔNG VIỆC

Thành viên	**Công việc**	**MSHV**
Đặng Khánh Thi	– Bài 1: làm các ý 1, 2, 3, 4; kiểm tra các ý 5, 6, 7 – Bài 2: Làm các ý 4, 5, 6; kiểm tra các ý 1, 2, 3 – Bài 3: Làm các ý 1, 2, 3, 4, 5; kiểm tra các ý 6, 7, 8, 9, 10 – Kiểm tra và tổng hợp code R của bài 2 – Trình bày file bài tập	20C29038
Đinh Thị Nữ	– Bài 1: làm các ý 5, 6, 7; kiểm tra các ý 1, 2, 3, 4 – Bài 2: Làm các ý 1, 2, 3; kiểm tra các ý 4, 5, 6 – Bài 3: Làm các ý 6, 7, 8, 9, 10; kiểm tra các ý 1, 2, 3, 4, 5	20C29013
Lý Phi Long	– Bài 1: làm các ý 1, 2, 3, 4; kiểm tra các ý 5, 6, 7 – Bài 2: Làm các ý 4, 5, 6; kiểm tra các ý 1, 2, 3 – Bài 3: Làm các ý 1, 2, 3, 4, 5; kiểm tra các ý 6, 7, 8, 9, 10 – Kiểm tra và tổng hợp code R của bài 3	20C29028
Phan Thị Thùy An (Nhóm trưởng)	– Bài 1: làm các ý 5, 6, 7; kiểm tra các ý 1, 2, 3, 4 – Bài 2: Làm các ý 1, 2, 3; kiểm tra các ý 4, 5, 6 – Bài 3: Làm các ý 6, 7, 8, 9, 10; kiểm tra các ý 1, 2, 3, 4, 5 – Kiểm tra và tổng hợp code R của bài 1 – Trình bày file bài tập	20C29002

BÀI 1

- X_1: áp lực công việc
- X_2: kỹ năng quản lý
- X_3: mức độ hài lòng với chức vụ của mình
- Y: mức độ lo lắng (biến phụ thuộc)

Nguồn gốc của sự biến thiên	Tổng bình phương	Bậc tự do
Hồi quy trên X_1	981.326	1
Hồi quy trên $X_2 \mid X_1$	190.232	1
Hồi quy trên $X_3 \mid X_1, X_2$	129.431	1
Sai số	442.292	18
Tổng quát	1743.281	21

Hình 1: Bảng ANOVA bài 1

1. Tính tổng bình phương hồi quy trên X_1, X_2 và X_3?

$$SSR = SSR_{X_1} + SSR_{X_2|X_1} + SSR_{X_3|X_1,X_2} = 981.326 + 190.232 + 129.431 = 1300.989$$

2. Xác định tỷ lệ phần trăm sự biến thiên của mức độ lo lắng được giải thích bởi các biến độc lập.

$$R^2 = \frac{SSR}{SST} = \frac{1299.989}{1743.281} = 0.7462876$$

Sự biến thiên của mức độ lo lắng được giải thích bởi các biến độc lập có tỉ lệ phần trăm là 74.63%.

3. Có thể kết luận rằng trong tất cả ba biến giải thích đều có ảnh hưởng đáng kể đến mức độ lo lắng hay không? Chỉ rõ kiểm định nào được dùng.

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \text{tồn tại ít nhất một } \beta_j \neq 0, \text{ với } j = 1, 2, 3 \end{cases}$$

Ta tính được giá trị thống kê Fisher:

$$F_{obs} = \frac{SSR/(p-1)}{SSE/(n-p)} = 17.64882$$

Với mức ý nghĩa $\alpha = 0.05$, tra bảng Fisher ta được:

$$F_{1-\alpha}(p-1, n-p) = F_{0.95}(3, 18) = 3.16$$

Vì $F_{obs} > F_{0.95}(3, 18)$ nên ta bác bỏ H_0 với mức ý nghĩa 5%.
Vậy tồn tại **ít nhất** một trong ba biến áp lực công việc, kỹ năng quản lý, mức độ hài lòng với chức vụ của mình có ảnh hưởng đến mức độ lo lắng.

4. Nếu chúng ta chỉ xét biến giải thích X_1, hãy lập bảng ANOVA ?

Khi chỉ xét X_1, mô hình hồi quy trở thành:

$$Y = \beta_0 + \beta_1 X_1 + \epsilon$$

Vậy tổng sai số của biến giải thích X_1 là:

$$SSE_{X_1} = SST - SSR_{X_1} = 761.955$$

Với số mẫu $n = 22$, ta lập được bảng ANOVA với biến giải thích X_1 như sau:

Biến thiên	SS	DF	MS	Fisher
R_{X_1}	$SSR_{X_1} = 981.326$	1	$SSR_{X_1}/1 = 981.326$	
E_{X_1}	$SSE_{X_1} = 761.955$	$n - 2 = 20$	$SSE_{X_1}/20 = 38.09775$	$MSR_{X_1}/MSE_{X_1} = 25.75811$
Total	1743.281	$n - 1 = 21$		

5. Kiểm định giả thuyết sau với mức ý nghĩa 5%

(a) $H_0 : \beta_1 = 0$ **cho mô hình** $Y = \beta_0 + \beta_1 X_1 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Thống kê của kiểm định:

$$F = \frac{MSR}{MSE} \sim F_{(1,20)} \text{ khi } H_0 \text{ đúng,}$$

với $F_{(1,20)}$ là phân phối Fisher có bậc tự do 1 và 20.

Dựa vào bảng ANOVA ở câu 4, ta tính được giá trị thống kê:

$$F_{obs} = \frac{981.326}{38.09775} = 25.7581$$

Với mức ý nghĩa $\alpha = 0.05$, tra bảng Fisher ta được:

$$F_{1-\alpha}(1, n-2) = F_{0.95}(1, 20) = 4.3512$$

Vì $F_{obs} > F_{0.95}(1, 20)$ nên ta bác bỏ H_0 với mức ý nghĩa 5%.

Vậy Y được giải thích bởi một biến giải thích X_1.

(b) $H_0 : \beta_2 = 0$ **cho mô hình** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_2 = 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \epsilon \\ H_1 : \beta_2 \neq 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{cases}$$

Thực hiện kiểm định Fisher từng phần, ta có thống kê của kiểm định:

$$F = \frac{\left[SSE(H_0) - SSE(H_1)\right]/r}{SSE(H_1)/(n-p)} \sim F_{(r,n-p)} \text{ khi } H_0 \text{ đúng,}$$

trong đó $r = 1, n = 22, p = 3$.

Trước tiên, ta cần tính $SSE(H_0)$ và $SSE(H_1)$:

$$SSE(H_0) = SST - SSR_{X_1} = 761.955, \text{ (đặt là } SSE_{X_1})$$

$$SSE(H_1) = SST - SSR_{X_1} - SSR_{X_2|X_1} = 571.723, \text{ (đặt là } SSE_{X_1,X_2})$$

Giá trị thống kê

$$F_{obs} = 6.3219$$

Với mức ý nghĩa $\alpha = 0.05$, tra bảng Fisher ta được:

$$F_{1-\alpha}(r, n-p) = F_{0.95}(1, 19) = 4.3807$$

Vì $F_{obs} > F_{0.95}(1, 19)$ nên ta bác bỏ H_0 với mức ý nghĩa 5%.

Vậy Y được giải thích bởi hai biến giải thích X_1, X_2.

(c) $H_0 : \beta_3 = 0$ **cho mô hình** $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_3 = 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ H_1 : \beta_3 \neq 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \end{cases}$$

Thực hiện kiểm định Fisher từng phần, ta có thống kê của kiểm định:

$$F = \frac{\left[SSE(H_0) - SSE(H_1)\right]/r}{SSE(H_1)/(n-p)} \sim F_{(r,n-p)} \text{ khi } H_0 \text{ đúng,}$$

trong đó $r = 1, n = 22, p = 4$.

Trước tiên, ta cần tính $SSE(H_0)$ và $SSE(H_1)$:

$$SSE(H_0) = SST - SSR_{X_1} - SSR_{X_2|X_1} = 571.723, \text{ (đặt là } SSE_{X_1,X_2})$$

$$SSE(H_1) = SST - SSR_{X_1} - SSR_{X_2|X_1} - SSR_{X_3|X_1,X_2} = 442.292, \text{ (đặt là } SSE_{X_1,X_2,X_3})$$

Giá trị thống kê

$$F_{obs} = 5.2675$$

Với mức ý nghĩa $\alpha = 0.05$, tra bảng Fisher ta được:

$$F_{1-\alpha}(r, n-p) = F_{0.95}(1, 18) = 4.4138$$

Vì $F_{obs} > F_{0.95}(1, 18)$ nên ta bác bỏ H_0 với mức ý nghĩa 5%.

Vậy Y được giải thích bởi cả ba biến giải thích X_1, X_2 và X_3.

Kết luận cho câu 3

Vậy ta có thể kết luận rằng tất cả ba biến áp lực công việc, kỹ năng quản lý, mức độ hài lòng với chức vụ của mình **đều** có ảnh hưởng đến mức độ lo lắng của giám đốc.

6. Xác định hệ số xác định cho mỗi mô hình trong câu 5.

Mô hình 1: $Y = \beta_0 + \beta_1 X_1 + \epsilon$ có hệ số xác định là

$$R_1^2 = 1 - \frac{SSE_{X_1}}{SST} = 0.5629$$

Mô hình 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$ có hệ số xác định là

$$R_2^2 = 1 - \frac{SSE_{X_1,X_2}}{SST} = 0.6720$$

Mô hình 3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$ có hệ số xác định là

$$R_3^2 = 1 - \frac{SSE_{X_1,X_2,X_3}}{SST} = 0.7463$$

7. Trong các mô hình trên, mô hình nào thích hợp nhất để giải thích sự biến động mức độ lo lắng của các giám đốc ?

Để so sánh độ thích hợp của các mô hình có số lượng biến độc lập khác nhau, ta cần so sánh các hệ số xác định hiệu chỉnh theo công thức:

$$R_{adj}^2 = 1 - \frac{SSE/(n-p)}{SST/(n-1)} \text{ với } n = 22$$

Mô hình 1: với $p = 2$ ta có

$$R_{1adj}^2 = 1 - \frac{SSE_{X_1}/20}{SST/21} = 0.5411$$

Mô hình 2: với $p = 3$ ta có

$$R_{2adj}^2 = 1 - \frac{SSE_{X_1,X_2}/19}{SST/21} = 0.6375$$

Mô hình 3: với $p = 4$ ta có

$$R_{3adj}^2 = 1 - \frac{SSE_{X_1,X_2,X_3}/18}{SST/21} = 0.7040$$

Dựa vào các giá trị R^2 hiệu chỉnh vừa tính, có thể kết luận **mô hình 3** là mô hình thích hợp nhất để giải thích sự biến động mức độ lo lắng của các giám đốc.

BÀI 2

Essai numéro	Résistance à la rupture Y_i	Épaisseur du matériau X_{i_1}	Densité X_{i_2}
1	37,8	4	4,0
2	22,5	4	3,6
3	17,1	3	3,1
4	10,8	2	3,2
5	7,2	1	3,0
6	42,3	6	3,8
7	30,2	4	3,8
8	19,4	4	2,9
9	14,8	1	3,8
10	9,5	1	2,8
11	32,4	3	3,4
12	21,6	4	2,8

Hình 2: Bảng số liệu bài 2

- Y: mức độ bền dẻo của nhựa
- X_1: độ dày của vật liệu
- X_2: mật độ của vật liệu

1. Tìm 2 phương trình đường thẳng hồi quy và 1 phương trình siêu phẳng (nếu có) ?

Ta xây dựng các mô hình hồi quy như sau:

Mô hình 1: $Y = \beta_0 + \beta_1 X_1 + \epsilon$

Mô hình đường thẳng hồi quy tương ứng: $Y = \hat{\beta_0} + \hat{\beta_1} X_1$

Dựa vào kết quả của phần mềm R ở hình 3, ta có $\hat{\beta_0} = 3.523$ và $\hat{\beta_1} = 6.036$, do đó ta có phương trình đường thẳng hồi quy theo độ dày của vật liệu (X_1) là:

$$Y = \hat{\beta_0} + \hat{\beta_1} X_1 = 3.523 + 6.036 X_1$$

Mô hình 2: $Y = \beta_0 + \beta_2 X_2 + \epsilon$

Mô hình đường thẳng hồi quy tương ứng: $Y = \hat{\beta_0} + \hat{\beta_2} X_2$

```
> Y1<-lm(Y~X1)
> summary(Y1)

Call:
lm(formula = Y ~ X1)

Residuals:
   Min     1Q Median     3Q    Max
-8.266 -4.887 -1.208  3.232 10.770

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    3.523      4.383   0.804 0.440237
X1             6.036      1.279   4.721 0.000816 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.633 on 10 degrees of freedom
Multiple R-squared:  0.6903,    Adjusted R-squared:  0.6593
F-statistic: 22.29 on 1 and 10 DF,  p-value: 0.0008155
```

Hình 3: Tham số mô hình 1

```
> Y2<-lm(Y~X2)
> summary(Y2)

Call:
lm(formula = Y ~ X2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.1923  -5.1780  -0.2298   6.1123  12.3077

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -36.373     20.489  -1.775   0.1062
X2            17.464      6.069   2.878   0.0164 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.815 on 10 degrees of freedom
Multiple R-squared:  0.453,     Adjusted R-squared:  0.3983
F-statistic: 8.282 on 1 and 10 DF,  p-value: 0.01645
```

Hình 4: Tham số mô hình 2

Dựa vào kết quả của phần mềm R ở hình 4, ta có $\hat{\beta_0} = -36.373$ và $\hat{\beta_2} = 17.464$, do đó ta có phương trình đường thẳng hồi quy theo mật độ của vật liệu (X_2) là:

$$Y = \hat{\beta_0} + \hat{\beta_2} X_2 = -36.373 + 17.464 X_2$$

Mô hình 3: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Mô hình mặt phẳng hồi quy tương ứng: $Y = \hat{\beta_0} + \hat{\beta_1} X_1 + \hat{\beta_2} X_2$

Dựa vào kết quả của phần mềm R ở hình 5, ta có $\hat{\beta_0} = -30.081$, $\hat{\beta_1} = 4.905$ và $\hat{\beta_2} = 11.072$, do đó ta có phương trình mặt phẳng hồi quy theo độ dày của vật liệu (X_1) và mật độ

```
> #phuong trinh sieu phang hoi quy
> Y3<-lm(Y~X1+X2)
> summary(Y3)

Call:
lm(formula = Y ~ X1 + X2)

Residuals:
   Min     1Q Median     3Q    Max
-6.897 -2.135 -1.126  1.714 10.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -30.081     11.455  -2.626 0.027542 *
X1             4.905      1.014   4.838 0.000923 ***
X2            11.072      3.621   3.058 0.013617 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.897 on 9 degrees of freedom
Multiple R-squared:  0.8481,    Adjusted R-squared:  0.8143
F-statistic: 25.12 on 2 and 9 DF,  p-value: 0.0002075
```

Hình 5: Tham số mô hình 3

của vật liệu (X_2) là:

$$Y = \hat{\beta_0} + \hat{\beta_1} X_1 + \hat{\beta_2} X_2 = -30.081 + 4.905 X_1 + 11.072 X_2$$

2. Xác định tỷ lệ phần trăm sự biến thiên của biến phụ thuộc cho từng mô hình có thể có trên.

Mô hình 1: Dựa vào kết quả mô hình 1 (hình 3), hệ số xác định $R^2 = 0.6903$ cho biết có 69.03% sự thay đổi của mức độ bền dẻo của nhựa được giải thích bởi độ dày của vật liệu (X_1).

Mô hình 2: Dựa vào kết quả mô hình 2 (hình 4), hệ số xác định $R^2 = 0.453$ cho biết có 45.3% sự thay đổi của mức độ bền dẻo của nhựa được giải thích bởi mật độ của vật liệu (X_2).

Mô hình 3: Dựa vào kết quả mô hình 3 (hình 5), hệ số xác định $R^2 = 0,8481$ cho biết có 84.81% sự thay đổi của mức độ bền dẻo của nhựa được giải thích bởi hai yếu tố là độ dày vật liệu (X_1) và mật độ của vật liệu (X_2).

3. Nếu chúng ta chỉ quan tâm đến cả 2 biến giải thích, hãy lập bảng ANOVA?

Bảng ANOVA cho cả hai biến giải thích:

```
> anova(lm(Y~X1+X2))
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value   Pr(>F)
X1         1 980.63  980.63 40.8959 0.000126 ***
X2         1 224.22  224.22  9.3509 0.013617 *
Residuals  9 215.81   23.98
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình 6: Bảng ANOVA cho hai biến X_1, X_2

Biến thiên	SS	DF	MS	Fisher
R_{X_1,X_2}	$SSR = 980.63 + 224.22 = 1204.85$	2	$SSR/2 = 602.425$	$F = MSR/MSE$
E	$SSE = 215.81$	$n - 3 = 9$	$SSE/9 = 23.9789$	$= 25.123$
Total	980.63+224.22+215.81 = 1420.66	$n - 1 = 11$		

4. Kiểm định giả thuyết sau với mức ý nghĩa 5%

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_1 : \beta_j \neq 0 \text{ với } j = 1, 2 \end{cases}$$

```
Call:
lm(formula = Y ~ X1 + X2)

Residuals:
   Min     1Q Median     3Q    Max
-6.897 -2.135 -1.126  1.714 10.122

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -30.081     11.455  -2.626 0.027542 *
X1             4.905      1.014   4.838 0.000923 ***
X2            11.072      3.621   3.058 0.013617 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.897 on 9 degrees of freedom
Multiple R-squared:  0.8481,    Adjusted R-squared:  0.8143
F-statistic: 25.12 on 2 and 9 DF,  p-value: 0.0002075
```

Từ hình trên, ta có giá trị thống kê Fisher $F_{obs} = 25.12 \sim F_{0,95}(2, 9)$, tính được $p\_value = 0.0002075386$ qua hàm **pf**:

Vì $p\_value = 0.0002075386 < \alpha = 0.05$, ta bác bỏ giả thuyết H_0 với mức ý nghĩa 5%.

```
> F_obs = lm_sum$fstatistic["value"]
> p_value = pf(F_obs, p, n - p - 1, lower.tail = FALSE); p_value
       value
0.0002075386
```

5. Xác định khoảng tin cậy với mức ý nghĩa 5% cho β_1 trong trường hợp mô hình chỉ có biến độc lập là độ dày của vật liệu (X_1).

```
> lmX1_fit = lm(Y ~ X1)
> confintB1 = confint(lmX1_fit)[2,]; confintB1
   2.5 %   97.5 %
3.187036 8.884790
```

Với độ tin cậy 95%, khoảng tin cậy β_1 của mô hình có một biến độc lập X_1 là $[3.187036, 8.88479]$.

6. Với khoảng tin cậy vừa tìm được ở câu 5, chúng ta có thể khẳng định rằng hồi quy tuyến tính là có ý nghĩa giữa mức độ bền dẻo của nhựa và độ dày của vật liệu và mật độ của vật liệu không? Chứng minh điều khẳng định của bạn.

Nếu chỉ dựa vào khoảng tin cậy của β_1 ở câu 5, chúng ta **không** thể khẳng định được điều trên, vì kết quả ở câu 5 chỉ cho ta biết mối quan hệ tuyến tính giữa mức độ bền dẻo (Y) với độ dày của vật liệu (X_1).
Để chứng minh, ta lần lượt thực hiện kiểm định các giả thuyết sau:

i) $H_0 : \beta_1 = \beta_2 = 0$ cho mô hình $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = 0 \\ H_1 : \beta_j \neq 0 \text{ với } j = 1, 2 \end{cases}$$

Với kết quả từ câu 4, H_0 bị bác bỏ do chứng minh trên, nghĩa là tồn tại tham số β_1 hoặc β_2 trong mô hình hồi quy hai biến X_1, X_2.

ii) $H_0 : \beta_1 = 0$ cho mô hình $Y = \beta_0 + \beta_1 X_1 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Với kết quả từ câu 5, H_0 bị bác bỏ do $\beta_1 = 0$ không thuộc khoảng tin cậy $[3.187036, 8.88479]$, nghĩa là tồn tại tham số β_1 trong mô hình hồi quy hai biến X_1, X_2.

iii) $H_0 : \beta_2 = 0$ cho mô hình $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Ta thực hiện kiểm định sự tồn tại tham số β_2 trong mô hình hồi quy hai biến X_1, X_2 khi đã có X_1.

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_2 = 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \epsilon \\ H_1 : \beta_2 \neq 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{cases}$$

Thực hiện kiểm định Fisher từng phần, ta có thống kê của kiểm định:

$$F_{obs} = \frac{\left[SSE(H_0) - SSE(H_1)\right]/r}{SSE(H_1)/(n-p)} \sim F_{(r,n-p)} \text{ khi } H_0 \text{ đúng,}$$

trong đó $r = 1, n = 12, p = 3$.

Từ bảng ANOVA ở câu 3, ta có được các giá trị sau:

$$SST = 1420.667$$

$$SSR_{X_1} = 980.63$$

$$SSR_{X_2|X_1} = 224.22$$

Trước tiên, ta cần tính $SSE(H_0)$ và $SSE(H_1)$:

$$SSE(H_0) = SSE_{X_1} = SST - SSR_{X_1} = 440.032$$

$$SSE(H_1) = SSE_{X_2|X_1} = SST - SSR_{X_1} - SSR_{X_2|X_1} = 215.81$$

Giá trị thống kê

$$F_{obs} = 9.350885$$

Với mức ý nghĩa $\alpha = 0.05$, tra bảng Fisher ta được:

$$F_{1-\alpha}(r, n-p) = F_{0.95}(1, 9) = 5.1174$$

Vì $F_{obs} > F_{0.95}(1, 9)$ nên ta bác bỏ H_0 với mức ý nghĩa 5%, vậy tồn tại tham số β_2 trong mô hình hồi quy hai biến X_1, X_2.

Từ kết luận của các giả thuyết trên, ta suy ra được mức độ bền dẻo (Y) có quan hệ tuyến tính chặt chẽ với cả hai biến độc lập là độ dày của vật liệu (X_1) và mật độ của vật liệu (X_2).

BÀI 3

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

- Mô hình với hai biến x_1, x_2

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{1}$$

- Mô hình với hai biến x_1, x_3

$$y = \beta_0' + \beta_1' x_1 + \beta_3' x_3 + \epsilon' \tag{2}$$

- Mô hình với hai biến x_2, x_3

$$y = \beta_0'' + \beta_2'' x_2 + \beta_3'' x_3 + \epsilon'' \tag{3}$$

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

- Mô hình 1

```
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
```

Hình 7: Tham số mô hình 1

Hệ số hồi quy:

$$\hat{\beta_0} = 25.84214, \hat{\beta_1} = 0.7148959, \hat{\beta_2} = -0.3281129$$

- Mô hình 2

```
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
```

Hình 8: Tham số mô hình 2

Hệ số hồi quy:

$$\hat{\beta}'_0 = 8.609241, \hat{\beta}'_1 = 0.9272087, \hat{\beta}'_3 = 0.02323681$$

- Mô hình 3

```
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
```

Hình 9: Tham số mô hình 3

Hệ số hồi quy:

$$\hat{\beta}''_0 = 31.97642, \hat{\beta}''_2 = -0.4538954, \hat{\beta}''_3 = 0.01996295$$

3. Với độ tin cậy 95%, tìm khoảng tin cậy cho các tham số trong mô hình với 2 biến độc lập x_1 và x_2.

Viết lại mô hình: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

```
> ## 3
> confint(linearMod1, level=0.95)
                  2.5 %     97.5 %
(Intercept) 12.4938794 39.1903962
X1           0.1288532  1.3009387
X2          -0.6242802 -0.0319457
```

Dựa vào kết quả mô hình, với độ tin cậy 95%, khoảng tin cậy của:

- β_0 là $[12.4939, 39.1904]$
- β_1 là $[0.1289, 1.3009]$
- β_2 là $[-0.6243, -0.0319]$

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

- Mô hình 1 có bảng ANOVA:

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x1         1 117.659 117.659 18.2587 0.001314 **
x2         1  38.314  38.314  5.9458 0.032916 *
Residuals 11  70.884   6.444
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình 10: Bảng ANOVA của mô hình 1

Hệ số xác định:

$$R^2 = \frac{SSR}{SST} = 0.6875395$$

- Mô hình 2 có bảng ANOVA:

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value   Pr(>F)
x1         1 117.659 117.659 12.0443 0.005235 **
x3         1   1.741   1.741  0.1782 0.681057
Residuals 11 107.457   9.769
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình 11: Bảng ANOVA của mô hình 2

Hệ số xác định:

$$R'^2 = \frac{SSR}{SST} = 0.5263211$$

- Mô hình 3 có bảng ANOVA:

```
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value  Pr(>F)
x2         1 109.520 109.520 10.3725 0.00815 **
x3         1   1.192   1.192  0.1129 0.74319
Residuals 11 116.145  10.559
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Hình 12: Bảng ANOVA của mô hình 3

Hệ số xác định:

$$R''^2 = \frac{SSR}{SST} = 0.4880253$$

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

Do các mô hình đều có cùng số lượng biến độc lập, ta so sánh hệ số xác định (R^2) để xét xem mô hình nào thích hợp nhất. Với kết quả từ câu 4, ta có được thứ tự tăng dần các hệ số xác định từ các mô hình của Y là

$$R''^2 < R'^2 < R^2$$

Vậy với hệ số R^2 cao nhất thì mô hình hai biến độc lập x_1, x_2 là phù 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.

Giả thiết cho mô hình $\mathbf{Y} = \mathbf{X}\beta + \epsilon$:

- Tập dữ liệu trong mô hình không xảy ra hiện tượng đa cộng tuyến.
- $\epsilon \sim N(\vec{0}, \mathbf{I}_n\sigma^2)$.

Dựa vào dữ liệu đề bài, số biến độc lập nhiều nhất có thể là 3 biến (X_1, X_2, X_3):

$$Y_{14\times1} = \begin{bmatrix} y_1 \\ \vdots \\ y_{14} \end{bmatrix}; X_{14\times4} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & x_{1,3} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{14,1} & x_{14,2} & x_{14,3} \end{bmatrix}; \beta_{4\times1} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_3 \end{bmatrix}; \varepsilon_{14\times1} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{14} \end{bmatrix}$$

Với dữ liệu đề bài, một số dòng đầu và dòng cuối của ma trận **X** và vec-tơ **Y** là:

$$Y_{14\times1} = \begin{bmatrix} 12 \\ 14 \\ \vdots \\ 25 \\ 21 \end{bmatrix}; X_{14\times4} = \begin{bmatrix} 1 & 2 & 45 & 121 \\ 1 & 1 & 43 & 132 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 12 & 35 & 174 \\ 1 & 7 & 29 & 180 \end{bmatrix}$$

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

Viết lại mô hình: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

```
> Y3<-lm(Y~X1+X2+X3)
> summary(Y3)

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
```

Dựa vào kết quả mô hình, ta có các hệ số hồi quy:

$$\hat{\beta_0} = 32.89132, \hat{\beta_1} = 0.8019, \hat{\beta_2} = -0.38136, \hat{\beta_3} = -0.03713$$

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

(a) Ước lượng phương sai của sai số, $\mathbb{V}(\epsilon)$. Ta có $\mathbb{V}(\epsilon) = \mathbf{I}_{14}\sigma^2$

Với σ^2 được ước lượng bởi

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-4} = 6.745$$

Vậy ước lượng của $\mathbb{V}(\epsilon)$ là

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14
1	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
2	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
3	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
4	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
5	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
6	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
7	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
8	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
9	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000	0.000000
10	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000	0.000000
11	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000	0.000000
12	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000	0.000000
13	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767	0.000000
14	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	6.744767

(b) Ước lượng phương sai của $\hat{\beta}$:

$$\mathbb{V}(\hat{\beta}) = \begin{bmatrix} var(\hat{\beta}_0) \\ var(\hat{\beta}_1) \\ var(\hat{\beta}_2) \\ var(\hat{\beta}_3) \end{bmatrix} = \begin{bmatrix} \mathbf{Se}^2(\hat{\beta}_0) \\ \mathbf{Se}^2(\hat{\beta}_1) \\ \mathbf{Se}^2(\hat{\beta}_2) \\ \mathbf{Se}^2(\hat{\beta}_3) \end{bmatrix}$$

Cách khác:

	V1	V2	V3	V4
1	136.0328036	0.10161439	-1.561077677	-0.513781462
2	0.1016144	0.08906395	0.008055960	-0.006341360
3	-1.5610777	0.00805596	0.024517512	0.003881096
4	-0.5137815	-0.00634136	0.003881096	0.002706406

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

```
> ## 9
> alpha = 0.05
> (upper = (SD_hat ^ 2) * (n - (p + 1)) / qchisq(alpha / 2, n - (p + 1)))
[1] 20.77248
> (lower = (SD_hat ^ 2) * (n - (p + 1)) / qchisq(1 - alpha / 2, n - (p + 1)))
[1] 3.292832
```

Với độ tin cậy 95%, khoảng tin cậy của $\hat{\sigma}^2$ là $(3.292832; 20.77248)$

10. Khi thêm 2 biến độc lập x_3 và x_2 vào mô hình chỉ với 1 biến độc lập x_1 thì làm cho chất lượng ước lượng cao hơn không?

Để kiểm tra chất lượng ước lượng, trước tiên, ta lần lượt kiểm định các giả thuyết sau:

(i) $H_0 : \beta_1 = 0$ cho mô hình $Y = \beta_0 + \beta_1 X_1 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

Kết quả mô hình chỉ với 1 biến độc lập X_1

```
> Y1<-lm(Y~X1)
> summary(Y1)

Call:
lm(formula = Y ~ X1)

Residuals:
    Min      1Q  Median      3Q     Max
-4.6538 -1.6302  0.8048  2.1139  4.3698

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.5712     1.8891   6.125 5.14e-05 ***
X1            1.0118     0.2814   3.596  0.00367 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.017 on 12 degrees of freedom
Multiple R-squared:  0.5186,    Adjusted R-squared:  0.4785
F-statistic: 12.93 on 1 and 12 DF,  p-value: 0.003674
```

Từ đó, ta có $p\_value = 0.00367 < \alpha = 0.05$, suy ra bác bỏ H_0 với mức ý nghĩa 5%, nghĩa là Y được giải thích bởi X_1.

(ii) $H_0 : \beta_2 = 0$ cho mô hình $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_2 = 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \epsilon \\ H_1 : \beta_2 \neq 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \end{cases}$$

Thực hiện kiểm định Fisher từng phần, ta có thống kê của kiểm định:

$$F = \frac{\left[SSE(H_0) - SSE(H_1)\right]/r}{SSE(H_1)/(n-p)} \sim F_{(r,n-p)} \text{ khi } H_0 \text{ đúng,}$$

trong đó $r = 1, n = 14, p = 3$.

Trước tiên, ta cần tính $SSE(H_0)$ và $SSE(H_1)$, với $SST = 226.857$ từ câu 4:

$$SSE(H_0) = SST - SSR_{x_1} = 109.195, \text{ (đặt là } SSE_{x_1})$$

$$SSE(H_1) = SST - SSR_{x_1} - SSR_{x_2|x_1} = 70.881, \text{ (đặt là } SSE_{x_1,x_2})$$

Giá trị thống kê

$$F_{obs} = 5.946$$

Với mức ý nghĩa $\alpha = 0.05$, tra bảng Fisher ta được:

$$F_{1-\alpha}(r, n-p) = F_{0.95}(1, 11) = 4.8443$$

Vì $F_{obs} > F_{0.95}(1, 11)$ nên ta bác bỏ H_0 với mức ý nghĩa 5%, nghĩa là Y được giải thích bởi X_1, X_2.

(iii) $H_0 : \beta_3 = 0$ cho mô hình $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$

Đặt giả thuyết kiểm định:

$$\begin{cases} H_0 : \beta_3 = 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon \\ H_1 : \beta_3 \neq 0 \text{ hay } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \end{cases}$$

Thực hiện kiểm định Fisher từng phần, ta có thống kê của kiểm định:

$$F = \frac{\left[SSE(H_0) - SSE(H_1)\right]/r}{SSE(H_1)/(n-p)} \sim F_{(r,n-p)} \text{ khi } H_0 \text{ đúng,}$$

trong đó $r = 1, n = 14, p = 4$.

Trước tiên, ta cần tính $SSE(H_0)$ và $SSE(H_1)$:

$$SSE(H_0) = SST - SSR_{x_1} - SSR_{x_2} = 70.881, \text{ (đặt là } SSE_{x_1,x_2})$$

$$SSE(H_1) = SST - SSR_{x_1} - SSR_{x_2} - SSR_{x_3} = 69.14, \text{ (đặt là } SSE_{x_1,x_2,x_3})$$

Giá trị thống kê

$$F_{obs} = 0.2518$$

Với mức ý nghĩa $\alpha = 0.05$, tra bảng Fisher ta được:

$$F_{1-\alpha}(r, n-p) = F_{0.95}(1, 10) = 4.9646$$

Vì $F_{obs} < F_{0.95}(1, 10)$, suy ra chưa đủ cơ sở để bác bỏ H_0 với mức ý nghĩa 5%, nghĩa là Y chưa được giải thích bởi X_3.

Kết luận:

Qua kiểm định các giả thuyết trên, vì Y chưa được giải thích bởi X_3 mà chỉ có thể được giải thích bởi X_1, X_2, nên ta có kết luận như sau:

- Khi thêm biến độc lập X_2 vào mô hình X_1, do thực hiện trên cùng mẫu, số lượng biến độc lập khác nhau, ta so sánh giá trị R^2 hiệu chỉnh. Nhận thấy:

  $$R^2_{adj_1} = 0.4785 < R^2_{adj_2} = 0.6307$$

  Vậy khi thêm X_2 vào mô hình với một biến độc lập X_1 thì chất lượng ước lượng sẽ được cải thiện.

- Khi thêm biến độc lập X_3 vào mô hình X_1, X_2 sẽ không cải thiện được chất lượng ước lượng;