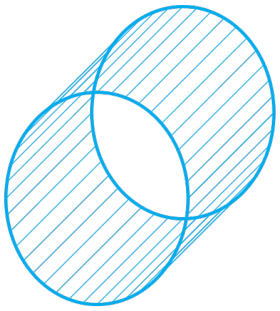
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN TP.HCM
CAO HỌC KHÓA 30



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**	**Mã số học viên**
1. Đặng Khánh Thi	– – –	20C29038
2. Đinh Thị Nữ	– – –	20C29013
3. Lý Phi Long	– – –	20C29028
4. Phan Thị Thùy An (Nhóm trưởng)	– –	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)

Bảng ANOVA:

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

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

$SSR = SSR_{X_1} + SSR_{X_2|X_1} + SSR_{X_3|X_1,X_2} = 981.326 + 190.232 + 129.431 = 1299.989$

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%.

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:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \beta_3 = 0 \\ H_1 : \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0 \end{cases}$$

Với giả thuyết trên, không thể kết luận cả ba biến đều ảnh hưởng đáng kể đến mức độ lo lắng, mà chỉ có thể kết luận rằng tồn tại ít nhất một biến có ảnh hưởng đến mức độ lo lắng nếu giả thiết H_0 sai.

Ta tính được kiểm định Fisher cho quan trắc:

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

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

$$F_{1-\alpha}(p, n-p-1) = F_{0.95}(3, 18) = 3.609$$

Vì $F_{obs} > F_{0.95}(3, 18)$ nên ta bác bỏ H_0.

Vậy tồn tại ít nhất một biến 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:

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

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 thiết sau với mức ý nghĩa 5%

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

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

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

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

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 ?

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 1: 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) ?

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

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

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

$$H_0 : \beta_1 = \beta_2 = 0$$

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).

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.

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

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

- Mô hình với hai biến x_1, x_3

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

- Mô hình với hai biến x_2, x_3

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

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

- Mô hình 1

```
Call:
lm(formula = y ~ 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 2: Thông số mô hình 1

Hệ số hồi quy:

$$\beta_0 = 25.84214, \beta_1 = 0.7148959, \beta_2 = -0.3281129$$

```
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 3: Thông số mô hình 2

- Mô hình 2

  Hệ số hồi quy:

$$\beta_0' = 8.609241, \beta_1' = 0.9272087, \beta_3' = 0.02323681$$

- 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 4: Thông số mô hình 3

Hệ số hồi quy:

$$\beta_0'' = 31.97642, \beta_2'' = -0.4538954, \beta_3'' = 0.01996295$$

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.

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 5: 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 6: Bảng ANOVA của mô hình 2

Hệ số xác định:

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

- Mô hình 3 có bảng ANOVA:

Hệ số xác định:

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

```
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 7: Bảng ANOVA của mô hình 3

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

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.

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 $\mathbb{V}(\epsilon)$ và $\mathbb{V}(\hat{\beta})$.

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

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?