BÀI TẬP THỰC HÀNH 2

PHÂN TÍCH HỒI QUY TUYẾN TÍNH

1. Nhập dữ liệu

Xem xét nghiên cứu sau đây, mà trong đó nhà nghiên cứu đo lường độ cholestrol trong máu của 18 đối tượng nam. Tỉ trọng cơ thể (body mass index) cũng được ước tính cho mỗi đối tượng bằng công thức tính BMI là lấy trọng lượng (tính bằng kg) chia cho chiều cao bình phương (m^2). Kết quả đo lường như sau:

Độ tuổi, tỉ trọng cơ thể và cholesterol

Mã số ID (id)	Độ tuổi (age)	BMI (bmi)	Cholesterol (chol)
1	46	25.4	3.5
2	20	20.6	1.9
3	52	26.2	4.0
4	30	22.6	2.6
5	57	25.4	4.5
6	25	23.1	3.0
7	28	22.7	2.9
8	36	24.9	3.8
9	22	19.8	2.1
10	43	25.3	3.8
11	57	23.2	4.1
12	33	21.8	3.0
13	22	20.9	2.5
14	63	26.7	4.6
15	40	26.4	3.2
16	48	21.2	4.2
17	28	21.2	2.3
18	49	22.8	4.0

- Nhập dữ liệu vào 1 file CSV tên "**data.csv**"

- Đọc dữ liệu từ file và đưa vào 1 dataframe (ví dụ tên của dataframe: **`data`**)

- Hiển thị dữ liệu theo mức độ **`chol`** theo **`age`** (Biểu đồ liên hệ giữa độ tuổi và cholesterol)

2. Hệ số tương quan

- Hiển thị hệ số tương quan giữa biến **`age`** và **`chol`**

- Kiểm định giả thiết hệ số tương quan của hai biến **age** và **chol** có bằng 0 hay không?

3. Mô hình hồi quy tuyến tính

- Tìm mô hình hồi quy tuyến tính mối tương quan giữ hai biến **chol** và **age**.

- Vẽ đường biểu diễn mối liên hệ giữa tuổi (**age**) và **chol**

4. Mô hình hồi quy tuyến tính đa biến – multiple linear regression (bổ sung)

Quay lại nghiên cứu về mối liên hệ giữa độ tuổi, **bmi** và **chol**. Trong ví dụ, chúng ta chỉ mới xét mối liên hệ giữa **age** và **cholesterol**, mà chưa xem đến mối liên hệ giữa cả hai yếu tố độ tuổi và **bmi** và **cholesterol**. Biểu đồ sau đây cho chúng ta thấy mối liên hệ giữa ba biến số này:

```
> pairs(data)
```

Phân tích hồi qui tuyến tính đơn giản giữa **bmi** và **cholesterol** cho thấy như mối liên hệ này có ý nghĩa thống kê:

```
> summary(lm(chol ~ bmi))
```

```
Call:
lm(formula = chol ~ bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9403 -0.3565 -0.1376  0.3040  1.4330

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.83187    1.60841  -1.761  0.09739 .
bmi          0.26410    0.06861   3.849  0.00142 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.623 on 16 degrees of freedom
Multiple R-Squared: 0.4808,     Adjusted R-squared: 0.4483
F-statistic: 14.82 on 1 and 16 DF,  p-value: 0.001418
```

BMI giải thích khoảng 48% độ dao động về cholesterol giữa các cá nhân. Nhưng vì BMI cũng có liên hệ với độ tuổi, chúng ta muốn biết nếu hai yếu tố này được phân tích cùng một lúc thì yếu tố nào quan trọng hơn. Để biết ảnh hưởng của cả hai yếu tố **age** (x_1) và **bmi** (tạm gọi là x_2) đến **cholesterol** (y) qua một mô hình hồi qui tuyến tính đa biến, và mô hình đó là:

$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

hay phương trình cũng có thể mô tả bằng kí hiệu ma trận: $\mathbf{Y} = \mathbf{X}\beta + \varepsilon$. Ở đây, **Y** là một vector 18 x 1, **X** là một matrix 18 x 2 phần tử, β và một vector 2 x 1, và ε là vector gồm 18 x 1 phần tử. Để ước tính hai hệ số hồi qui, β_1 và β_2 chúng ta cũng ứng dụng hàm **lm**() trong R như sau:

```
> mreg <- lm(chol ~ age + bmi)
> summary(mreg)
```

Kết quả

```
Call:
lm(formula = chol ~ age + bmi)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3762 -0.2259 -0.0534  0.1698  0.5679

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.455458   0.918230   0.496    0.627
```

```
age         0.054052   0.007591   7.120 3.50e-06 ***
bmi         0.033364   0.046866   0.712    0.487
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3074 on 15 degrees of freedom
Multiple R-Squared: 0.8815,     Adjusted R-squared: 0.8657
F-statistic: 55.77 on 2 and 15 DF,  p-value: 1.132e-07
```

Kết quả phân tích trên cho thấy ước số $\hat{\alpha} = 0.455$, $\hat{\beta}_1 = 0.054$ và $\hat{\beta}_2 = 0.0333$. Nói cách khác, chúng ta có phương trình ước đoán độ cholesterol dựa vào hai biến số **độ tuổi** và **bmi** như sau:

Cholesterol = 0.455 + 0.054(age) + 0.0333(bmi)

Phương trình cho biết khi **độ tuổi** tăng 1 năm thì **cholesterol** tăng 0.054 mg/L (ước số này không khác mấy so với 0.0578 trong phương trình chỉ có độ tuổi), và mỗi 1 kg/m^2 tăng BMI thì cholesterol tăng 0.0333 mg/L. Hai yếu tố này "giải thích" khoảng 88.2% (R^2 = 0.8815) độ dao động của cholesterol giữa các cá nhân.

Chúng ta chú ý phương trình với độ tuổi (trong phân tích phần trước) giải thích khoảng 87.7% độ dao động cholesterol giữa các cá nhân. Khi chúng ta thêm yếu tố BMI, hệ số này tăng lên 88.2%, tức chỉ 0.5%. Câu hỏi đặt ra là 0.5% tăng trưởng này có ý nghĩa thống kê hay không. Câu trả lời có thể xem qua kết quả kiểm định yếu tố bmi với trị số p = 0.487. Như vậy, **bmi** không cung cấp cho chúng thêm thông tin hay tiên đoán cholesterol hơn những gì chúng ta đã có từ độ tuổi. Nói cách khác, khi độ tuổi đã được xem xét, thì ảnh hưởng của **bmi** không còn ý nghĩa thống kê. Điều này có thể hiểu được, bởi vì qua Biểu đồ 10.5 chúng ta thấy độ tuổi và **bmi** có một mối liên hệ khá cao. Vì hai biến này có tương quan với nhau, chúng ta không cần cả hai trong phương trình. (Lưu ý: ví dụ này chỉ có tính cách minh họa cho việc tiến hành phân tích hồi qui tuyến tính đa biến bằng R, chứ không có ý định mô phỏng dữ liệu theo định hướng sinh học)