10
Phân tích hồi qui tuyến tính

Phân tích hồi qui tuyến tính (linear regression analysis) có lẽ là một trong những phương pháp phân tích số liệu thông dụng nhất trong thống kê học. Anon từng viết "Cho con người 3 vũ khí – hệ số tương quan, hồi qui tuyến tính và một cây bút, con người sẽ sử dụng cả ba"! Trong chương này, tôi sẽ giới thiệu cách sử dụng R để phân tích hồi qui tuyến tính và các phương pháp liên quan như hệ số tương quan và kiểm định giả thiết thống kê.

Ví dụ 1. Để minh họa cho vấn đề, chúng ta thử 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:

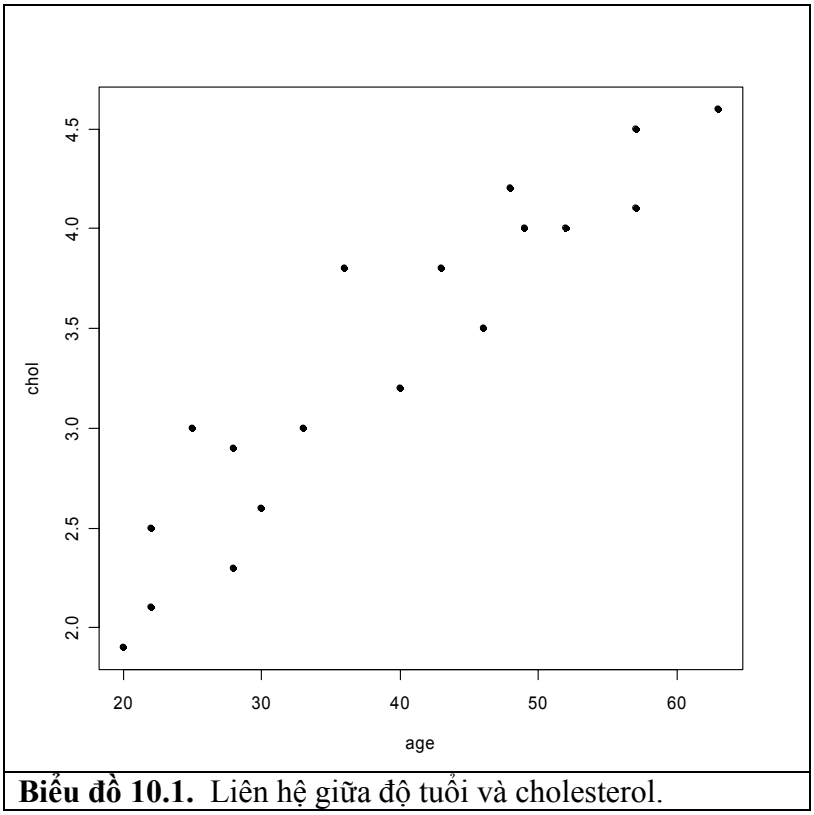
Bảng 1. Độ tuổi, tỉ trọng cơ thể và cholesterol

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

Nhìn sơ qua số liệu chúng ta thấy người có độ tuổi càng cao độ cholesterol cũng càng cao. Chúng ta thử nhập số liệu này vào R và vẽ một biểu đồ tán xạ như sau:

```
> age <- c(46,20,52,30,57,25,28,36,22,43,57,33,22,63,40,48,28,49)
```

```
> bmi <-c(25.4,20.6,26.2,22.6,25.4,23.1,22.7,24.9,19.8,25.3,23.2,
          21.8,20.9,26.7,26.4,21.2,21.2,22.8)

> chol <- c(3.5,1.9,4.0,2.6,4.5,3.0,2.9,3.8,2.1,3.8,4.1,3.0,
            2.5,4.6,3.2, 4.2,2.3,4.0)

> data <- data.frame(age, bmi, chol)
> plot(chol ~ age, pch=16)
```



Biểu đồ 10.1. Liên hệ giữa độ tuổi và cholesterol.

Biểu đồ 10.1 trên đây gợi ý cho thấy mối liên hệ giữa độ tuổi (`age`) và cholesterol là một đường thẳng (tuyến tính). Để "đo lường" mối liên hệ này, chúng ta có thể sử dụng hệ số tương quan (coefficient of correlation).

10.1 Hệ số tương quan

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số, như giữa độ tuổi (*x*) và cholesterol (*y*). Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm ($r < 0$) có nghĩa là khi *x* tăng cao thì *y* giảm (và ngược lại, khi *x* giảm thì *y* tăng); nếu giá trị hệ số tương quan là dương ($r > 0$) có nghĩa là khi *x* tăng cao thì *y* cũng tăng, và khi *x* tăng cao thì *y* cũng giảm theo.

Thực ra có nhiều hệ số tương quan trong thống kê, nhưng ở đây tôi sẽ trình bày 3 hệ số tương quan thông dụng nhất: hệ số tương quan Pearson r, Spearman ρ, và Kendall τ.

10.1.1 Hệ số tương quan Pearson

Cho hai biến số *x* và *y* từ *n* mẫu, hệ số tương quan Pearson được ước tính bằng công thức sau đây:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Trong đó, như định nghĩa phần trên, $\bar{x}$ và $\bar{y}$ là giá trị trung bình của biến số *x* và *y*. Để ước tính hệ số tương quan giữa độ tuổi `age` và cholesterol, chúng ta có thể sử dụng hàm `cor(x,y)` như sau:

```
> cor(age, chol)
[1] 0.936726
```

Chúng ta có thể kiểm định giả thiết hệ số tương quan bằng 0 (tức hai biến *x* và *y* không có liên hệ). Phương pháp kiểm định này thường dựa vào phép biến đổi Fisher mà R đã có sẵn một hàm `cor.test` để tiến hành việc tính toán.

```
> cor.test(age, chol)

        Pearson's product-moment correlation

data:  age and chol
t = 10.7035, df = 16, p-value = 1.058e-08
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8350463 0.9765306
sample estimates:
     cor
0.936726
```

Kết quả phân tích cho thấy kiểm định t = 10.70 với trị số p = 1.058e-08; do đó, chúng ta có bằng chứng để kết luận rằng mối liên hệ giữa độ tuổi và cholesterol có ý nghĩa thống kê. Kết luận này cũng chính là kết luận chúng ta đã đi đến trong phần phân tích hồi qui tuyến tính trên.

10.1.2 Hệ số tương quan Spearman ρ

Hệ số tương quan Pearson chỉ hợp lí nếu biến số *x* và *y* tuân theo luật phân phối chuẩn. Nếu *x* và *y* không tuân theo luật phân phối chuẩn, chúng ta phải sử dụng một hệ số tương quan khác tên là Spearman, một phương pháp phân tích phi tham số. Hệ số này

được ước tính bằng cách biến đổi hai biến số *x* và *y* thành thứ bậc (rank), và xem độ tương quan giữa hai dãy số bậc. Do đó, hệ số còn có tên tiếng Anh là Spearman's Rank correlation. R ước tính hệ số tương quan Spearman bằng hàm `cor.test` với thông số `method="spearman"` như sau:

```
> cor.test(age, chol, method="spearman")

        Spearman's rank correlation rho

data:  age and chol
S = 51.1584, p-value = 2.57e-09
alternative hypothesis: true rho is not equal to 0
sample estimates:
     rho
0.947205

Warning message:
Cannot compute exact p-values with ties in: cor.test.default(age,
chol, method = "spearman")
```

Kết quả phân tích cho thấy giá trị rho = 0.947, và trị số p = 2.57e-09. Kết quả từ phân tích này cũng không khác với phân tích hồi qui tuyến tính: mối liên hệ giữa độ tuổi và cholesterol rất cao và có ý nghĩa thống kê.

10.1.3 Hệ số tương quan Kendall τ

Hệ số tương quan Kendall (cũng là một phương pháp phân tích phi tham số) được ước tính bằng cách tìm các cặp số (*x, y*) "song hành" với nhau. Một cặp (*x, y*) song hành ở đây được định nghĩa là hiệu (độ khác biệt) trên trục hoành có cùng dấu hiệu (dương hay âm) với hiệu trên trục tung. Nếu hai biến số *x* và *y* không có liên hệ với nhau, thì số cặp song hành bằng hay tương đương với số cặp không song hành.

Bởi vì có nhiều cặp phải kiểm định, phương pháp tính toán hệ số tương quan Kendall đòi hỏi thời gian của máy tính khá cao. Tuy nhiên, nếu một dữ liệu dưới 5000 đối tượng thì một máy vi tính có thể tính toán khá dễ dàng. R dùng hàm `cor.test` với thông số `method="kendall"` để ước tính hệ số tương quan Kendall:

```
> cor.test(age, chol, method="kendall")

        Kendall's rank correlation tau

data:  age and chol
z = 4.755, p-value = 1.984e-06
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.8333333

Warning message:
```

```
Cannot compute exact p-value with ties in: cor.test.default(age,
chol, method = "kendall")
```

Kết quả phân tích hệ số tương quan Kendall một lần nữa khẳng định mối liên hệ giữa độ tuổi và cholesterol có ý nghĩa thống kê, vì hệ số tau = 0.833 và trị số p = 1.98e-06.

Các hệ số tương quan trên đây đo mức độ tương quan giữa hai biến số, nhưng không cho chúng ta một phương trình để nối hai biến số đó với nhau. Thành ra, vấn đề đặt ra là chúng ta tìm một phương trình tuyến tính để mô tả mối liên hệ này. Chúng ta sẽ ứng dụng mô hình hồi qui tuyến tính.

10.2 Mô hình của hồi qui tuyến tính đơn giản

10.2.1 vài dòng lí thuyết

Để tiện việc theo dõi và mô tả mô hình, gọi độ tuổi cho cá nhân i là x_i và cholesterol là y_i. Ở đây i = 1, 2, 3, …, 18. Mô hình hồi tuyến tính phát biểu rằng:

$$y_i = \alpha + \beta x_i + \varepsilon_i \qquad [1]$$

Nói cách khác, phương trình trên giả định rằng độ cholesterol của một cá nhân bằng một hằng số α cộng với một hệ số β liên quan đến độ tuổi, và một sai số ε_i. Trong phương trình trên, α là *chặn* (intercept, tức giá trị lúc x_i =0), và β là độ dốc (slope hay gradient). Trong thực tế, α và β là hai thông số (paramater, còn gọi là *regression coefficient* hay hệ số hồi qui), và ε_i là một biến số theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2.

Các thông số α, β và σ^2 phải được ước tính từ dữ liệu. Phương pháp để ước tính các thông số này là phương pháp *bình phương nhỏ nhất* (least squares method). Như tên gọi, phương pháp bình phương nhỏ nhất tìm giá trị α, β sao cho $\sum_{i=1}^{n}\left[y_i - (\alpha + \beta x_i)\right]^2$ nhỏ nhất. Sau vài thao tác toán, có thể chứng minh dễ dàng rằng, ước số cho α và β đáp ứng điều kiện đó là:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad [2]$$

và

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \qquad [3]$$

Ở đây, $\bar{x}$ và $\bar{y}$ là giá trị trung bình của biến số *x* và *y*. Chú ý, tôi viết $\hat{\alpha}$ và $\hat{\beta}$ (với dấu mũ phía trên) là để nhắc nhở rằng đây là hai ước số (estimates) của α và β, chứ không phải α và β (chúng ta không biết chính xác α và β, nhưng chỉ có thể ước tính mà thôi).

Sau khi đã có ước số $\hat{\alpha}$ và $\hat{\beta}$, chúng ta có thể ước tính độ cholesterol trung bình cho từng độ tuổi như sau:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta} x_i$$

Tất nhiên, $\hat{y}_i$ ở đây chỉ là số trung bình cho độ tuổi x_i, và phần còn lại (tức y_i - $\hat{y}_i$) gọi là *phần dư* (*residual*). Và phương sai của phần dư có thể ước tính như sau:

$$s^2 = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)}{n-2} \qquad [4]$$

s^2 chính là ước số của σ^2.

Trong phân tích hồi qui tuyến tính, thông thường chúng ta muốn biết hệ số β = 0 hay khác 0. Nếu β bằng 0, thì cũng có nghĩa là không có mối liên hệ gì giữa *x* và *y*; nếu β khác với 0, chúng ta có bằng chứng để phát biểu rằng *x* và *y* có liên quan nhau. Để kiểm định giả thiết β = 0 chúng ta dùng xét nghiệm t sau đây:

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \qquad [5]$$

$SE(\hat{\beta})$ có nghĩa là sai số chuẩn (standard error) của ước số $\hat{\beta}$. Trong phương trình trên, *t* tuân theo luật phân phối *t* với bậc tự do *n*-2 (nếu thật sự β = 0).

10.2.2 Phân tích hồi qui tuyến tính đơn giản bằng R

Hàm `lm` (viết tắt từ **l**inear **m**odel) trong R có thể tính toán các giá trị của $\hat{\alpha}$ và $\hat{\beta}$, cũng như s^2 một cách nhanh gọn. Chúng ta tiếp tục với ví dụ bằng R như sau:

```
> lm(chol ~ age)

Call:
lm(formula = chol ~ age)

Coefficients:
(Intercept)          age
    1.08922      0.05779
```

Trong lệnh trên, "`chol ~ age`" có nghĩa là mô tả `chol` là một hàm số của `age`. Kết quả tính toán của `lm` cho thấy $\hat{\alpha} = 1.0892$ và $\hat{\beta} = 0.05779$. Nói cách khác, với hai thông số này, chúng ta có thể ước tính độ cholesterol cho bất cứ độ tuổi nào trong khoảng tuổi của mẫu bằng phương trình tuyến tính:

$$\hat{y}_i = 1.08922 + 0.05779 \times \text{age}$$

Phương trình này có nghĩa là khi độ tuổi tăng 1 năm thì độ cholesterol tăng khoảng 0.058 mmol/L.

Thật ra, hàm `lm` còn cung cấp cho chúng ta nhiều thông tin khác, nhưng chúng ta phải đưa các thông tin này vào một object. Gọi object đó là `reg`, thì lệnh sẽ là:

```
> reg <- lm(chol ~ age)
> summary(reg)

Call:
lm(formula = chol ~ age)

Residuals:
     Min       1Q   Median       3Q      Max
-0.40729 -0.24133 -0.04522  0.17939  0.63040

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.089218   0.221466   4.918 0.000154 ***
age         0.057788   0.005399  10.704 1.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3027 on 16 degrees of freedom
Multiple R-Squared: 0.8775,     Adjusted R-squared: 0.8698
F-statistic: 114.6 on 1 and 16 DF,  p-value: 1.058e-08
```

Lệnh thứ hai, `summary(reg)`, yêu cầu R liệt kê các thông tin tính toán trong `reg`. Phần kết quả chia làm 3 phần:

(a) Phần 1 mô tả phần dư (residuals) của mô hình hồi qui:

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.40729 -0.24133 -0.04522  0.17939  0.63040
```

Chúng ta biết rằng trung bình phần dư phải là 0, và ở đây, số trung vị là -0.04, cũng không xa 0 bao nhiêu. Các số quantiles 25% (1Q) và 75% (3Q) cũng khá cân đối chung quan số trung vị, cho thấy phần dư của phương trình này tương đối cân đối.

(b) Phần hai trình bày ước số của $\hat{\alpha}$ và $\hat{\beta}$ cùng với sai số chuẩn và giá trị của kiểm định t. Giá trị kiểm định t cho $\hat{\beta}$ là 10.74 với trị số p = 1.06e-08, cho thấy β không phải bằng 0. Nói cách khác, chúng ta có bằng chứng để cho rằng có một mối liên hệ giữa cholesterol và độ tuổi, và mối liên hệ này có ý nghĩa thống kê.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.089218   0.221466   4.918 0.000154 ***
age         0.057788   0.005399  10.704 1.06e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) Phần ba của kết quả cho chúng ta thông tin về phương sai của phần dư (residual mean square). Ở đây, s^2 = 0.3027. Trong kết quả này còn có kiểm định F, cũng chỉ là một kiểm định xem có quả thật β bằng 0, tức có ý nghĩa tương tự như kiểm định t trong phần trên. Nói chung, trong trường hợp phân tích hồi qui tuyến tính đơn giản (với một yếu tố) chúng ta không cần phải quan tâm đến kiểm định F.

```
Residual standard error: 0.3027 on 16 degrees of freedom
Multiple R-Squared: 0.8775,     Adjusted R-squared: 0.8698
F-statistic: 114.6 on 1 and 16 DF,  p-value: 1.058e-08
```

Ngoài ra, phần 3 còn cho chúng ta một thông tin quan trọng, đó là trị số R^2 hay *hệ số xác định bội* (coefficient of determination). Hệ số này được ước tính bằng công thức:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{y}_i - \bar{y}\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2} \qquad [6]$$

Tức là bằng tổng bình phương giữa số ước tính và trung bình chia cho tổng bình phương số quan sát và trung bình. Trị số R^2 trong ví dụ này là 0.8775, có nghĩa là phương trình tuyến tính (với độ tuổi là một yếu tố) giải thích khoảng 88% các khác biệt về độ cholesterol giữa các cá nhân. Tất nhiên trị số R^2 có giá trị từ 0 đến 100% (hay 1). Giá trị R^2 càng cao là một dấu hiệu cho thấy mối liên hệ giữa hai biến số độ tuổi và cholesterol càng chặt chẽ.

Một hệ số cũng cần đề cập ở đây là *hệ số điều chỉnh xác định bội* (mà trong kết quả trên R gọi là "Adjusted R-squared"). Đây là hệ số cho chúng ta biết mức độ cải tiến của phương sai phần dư (residual variance) do yếu tố độ tuổi có mặt trong mô hình tuyến tính. Nói chung, hệ số này không khác mấy so với hệ số xác định bội, và chúng ta cũng không cần chú tâm quá mức.

10.2.3 Giả định của phân tích hồi qui tuyến tính

Tất cả các phân tích trên dựa vào một số giả định quan trọng như sau:

(a) x là một biến số cố định hay fixed, ("cố định" ở đây có nghĩa là không có sai sót ngẫu nhiên trong đo lường);

(b) ε_i phân phối theo luật phân phối chuẩn;

(c) ε_i có giá trị trung bình (mean) là 0;

(d) ε_i có phương sai σ^2 cố định cho tất cả x_i; và

(e) các giá trị liên tục của ε_i không có liên hệ tương quan với nhau (nói cách khác, ε_1 và ε_2 không có liên hệ với nhau).

Nếu các giả định này không được đáp ứng thì phương trình mà chúng ta ước tính có vấn đề hợp lí (validity). Do đó, trước khi trình bày và diễn dịch mô hình trên, chúng ta cần phải kiểm tra xem các giả định trên có đáp ứng được hay không. Trong trường hợp này, giả định (a) không phải là vấn đề, vì độ tuổi không phải là một biến số ngẫu nhiên, và không có sai số khi tính độ tuổi của một cá nhân.

Đối với các giả định (b) đến (e), cách kiểm tra đơn giản nhưng hữu hiệu nhất là bằng cách xem xét mối liên hệ giữa $\hat{y}_i$, x_i, và phần dư e_i ($e_i = y_i - \hat{y}_i$) bằng những đồ thị tán xạ.

Với lệnh `fitted()` chúng ta có thể tính toán $\hat{y}_i$ cho từng cá nhân như sau (ví dụ đối với cá nhân 1, 46 tuổi, độ cholestrol có thể tiên đoán như sau: `1.08922 + 0.05779 x 46 = 3.747`).

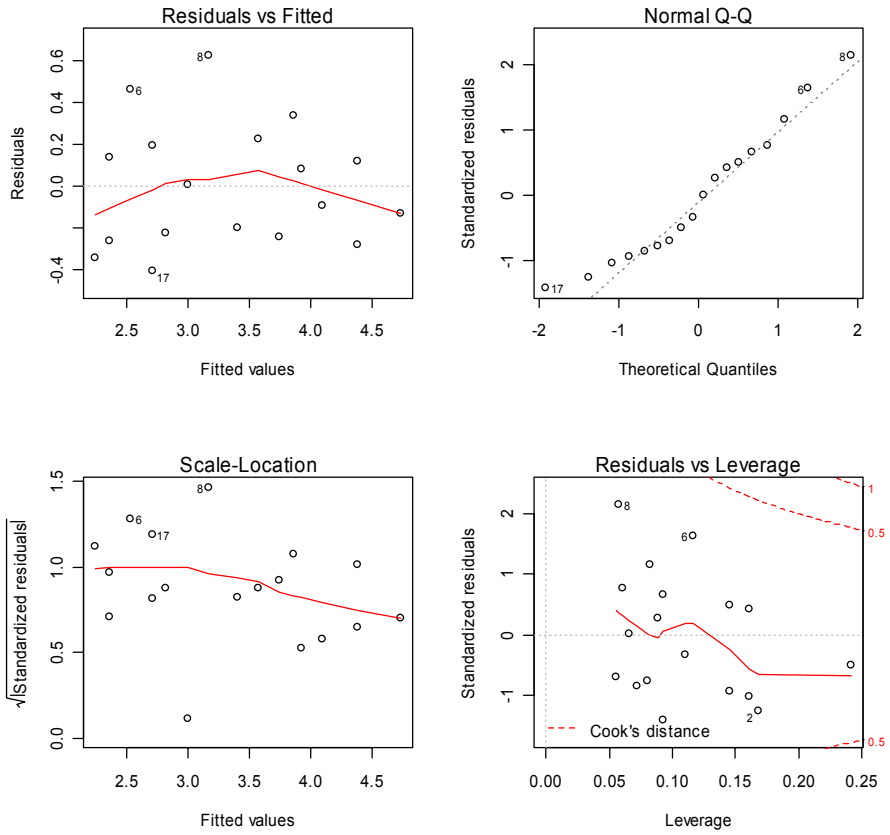
```
> fitted(reg)
       1        2        3        4        5        6        7        8
3.747483 2.244985 4.094214 2.822869 4.383156 2.533927 2.707292 3.169600
       9       10       11       12       13       14       15       16
2.360562 3.574118 4.383156 2.996234 2.360562 4.729886 3.400753 3.863060
      17       18
2.707292 3.920849
```

Với lệnh `resid()` chúng ta có thể tính toán phần dư e_i cho từng cá nhân như sau (với đối tượng 1, e_1 = `3.5 - 3.74748 = -0.24748`):

```
> resid(reg)
           1            2            3            4            5            6
-0.247483426 -0.344985415 -0.094213736 -0.222869265  0.116844338  0.466072660
           7            8            9           10           11           12
 0.192707505  0.630400424 -0.260562185  0.225881729 -0.283155662  0.003765579
          13           14           15           16           17           18
 0.139437815 -0.129885972 -0.200753116  0.336939804 -0.407292495  0.079151419
```

Để kiểm tra các giả định trên, chúng ta có thể vẽ một loạt 4 đồ thị mà tôi sẽ giải thích sau đây:

```
> op <- par(mfrow=c(2,2))          #yêu cầu R dành ra 4 cửa sổ
> plot(reg)                        #vẽ các đồ thị trong reg
```



Biểu đồ 10.2. Phân tích phần dư để kiểm tra các giả định trong phân tích hồi qui tuyến tính.

(a) Đồ thị bên trái dòng 1 vẽ phần dư e_i và giá trị tiên đoán cholesterol $\hat{y}_i$. Đồ thị này cho thấy các giá trị phần dư tập chung quanh đường y = 0, cho nên giả định (c), hay ε_i có giá trị trung bình 0, là có thể chấp nhận được.

(b) Đồ thị bên phải dòng 1 vẽ giá trị phần dư và giá trị kì vọng dựa vào phân phối chuẩn. Chúng ta thấy các số phần dư tập trung rất gần các giá trị trên đường chuẩn, và do đó, giả định (b), tức ε_i phân phối theo luật phân phối chuẩn, cũng có thể đáp ứng.

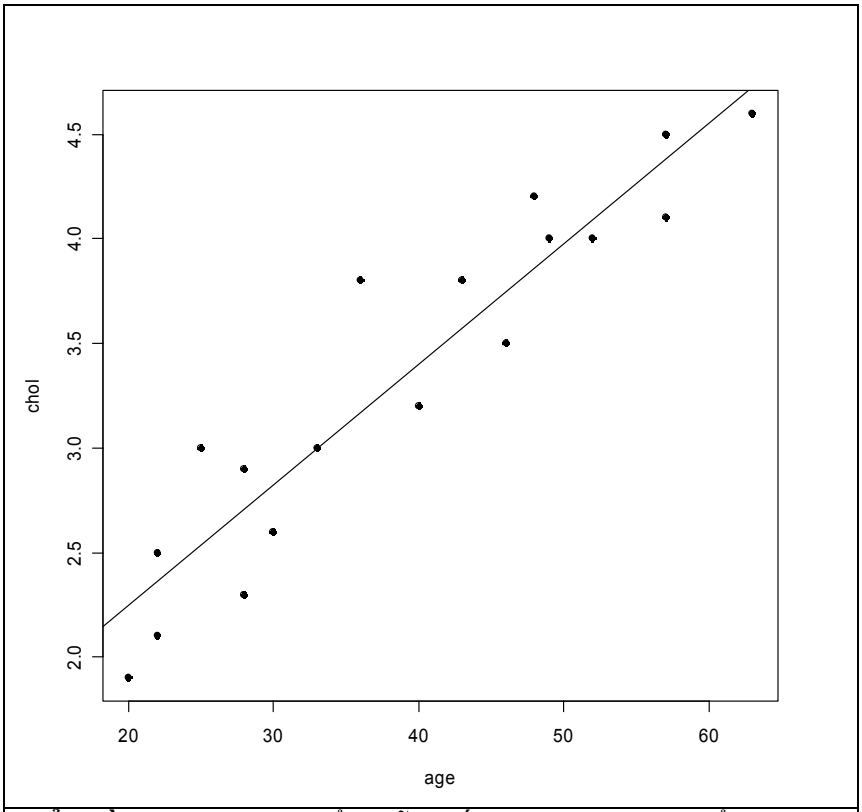
(c) Đồ thị bên trái dòng 2 vẽ căn số phần dư chuẩn (standardized residual) và giá trị của $\hat{y}_i$. Đồ thị này cho thấy không có gì khác nhau giữa các số phần dư chuẩn cho các giá trị của $\hat{y}_i$, và do đó, giả định (d), tức ε_i có phương sai σ^2 cố định cho tất cả x_i, cũng có thể đáp ứng.

Nói chung qua phân tích phần dư, chúng ta có thể kết luận rằng mô hình hồi qui tuyến tính mô tả mối liên hệ giữa độ tuổi và cholesterol một cách khá đầy đủ và hợp lí.

10.2.4 Mô hình tiên đoán

Sau khi mô hình tiên đoán cholesterol đã được kiểm tra và tính hợp lí đã được thiết lập, chúng ta có thể vẽ đường biểu diễn của mối liên hệ giữa độ tuổi và cholesterol bằng lệnh `abline` như sau (xin nhắc lại object của phân tích là `reg`):

```
> plot(chol ~ age, pch=16)
> abline(reg)
```

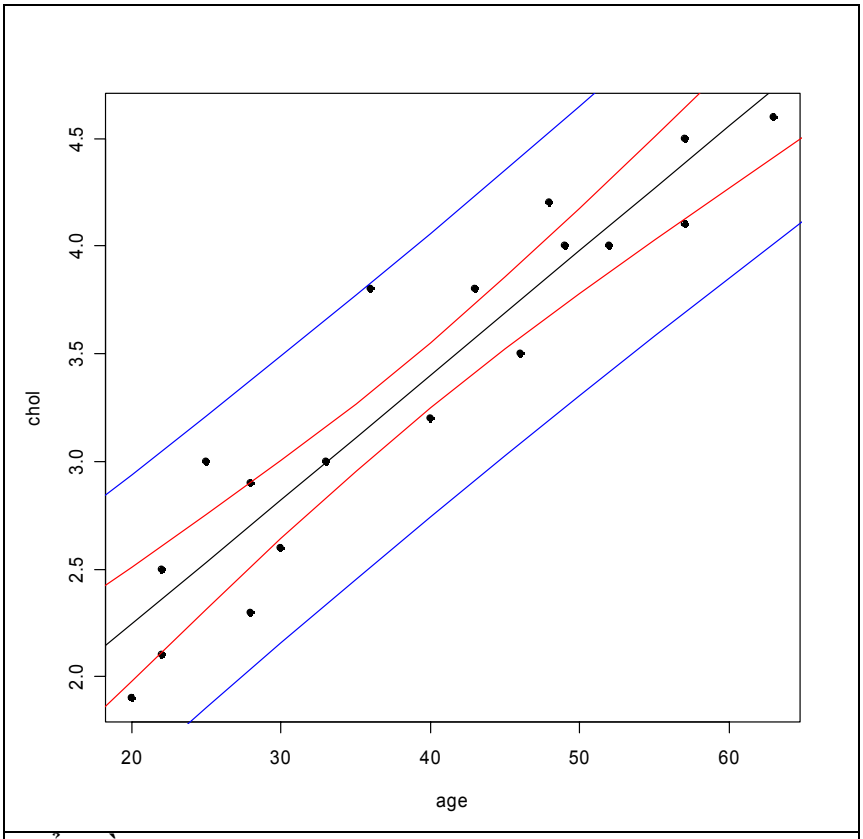


Biểu đồ 10.3. Đường biểu diễn mối liên hệ giữa độ tuổi (age) và cholesterol.

Nhưng mỗi giá trị $\hat{y}_i$ được tính từ ước số $\widehat{\alpha}$ và $\widehat{\beta}$, mà các ước số này đều có sai số chuẩn, cho nên giá trị tiên đoán $\hat{y}_i$ cũng có sai số. Nói cách khác, $\hat{y}_i$ chỉ là trung bình, nhưng trong thực tế có thể cao hơn hay thấp hơn tùy theo chọn mẫu. Khoảng tin cậy 95% này có thể ước tính qua R bằng các lệnh sau đây:

```
> reg <- lm(chol ~ age)
> new <- data.frame(age = seq(15, 70, 5))
```

```
> pred.w.plim <- predict.lm(reg, new, interval="prediction")
> pred.w.clim <- predict.lm(reg, new, interval="confidence")
> resc <- cbind(pred.w.clim, new)
> resp <- cbind(pred.w.plim, new)
> plot(chol ~ age, pch=16)
> lines(resc$fit ~ resc$age)
> lines(resc$lwr ~ resc$age, col=2)
> lines(resc$upr ~ resc$age, col=2)
> lines(resp$lwr ~ resp$age, col=4)
> lines(resp$upr ~ resp$age, col=4)
```



Biểu đồ 10.4. Giá trị tiên đoán và khoảng tin cậy 95%.

Biểu đồ trên vẽ giá trị tiên đoán trung bình $\hat{y}_i$ (đường thẳng màu đen), và khoảng tin cậy 95% của giá trị này là đường màu đỏ. Ngoài ra, đường màu xanh là khoảng tin cậy của giá trị tiên đoán cholesterol cho một độ tuổi mới trong quần thể.

10.3 Mô hình hồi qui tuyến tính đa biến (multiple linear regression)

Mô hình được diễn đạt qua phương trình [1] $y_i = \alpha + \beta x_i + \varepsilon_i$ có một yếu tố duy nhất (đó là x), và vì thế thường được gọi là mô hình hồi qui tuyến tính đơn giản (simple

linear regression model). Trong thực tế, chúng ta có thể phát triển mô hình này thành nhiều biến, chứ không chỉ giới hạn một biến như trên, chẳng hạn như:

$$y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_k x_{ki} + \varepsilon_i \quad [7]$$

nói cụ thể hơn:

$$\begin{aligned} y_1 &= \alpha + \beta_1 x_{11} + \beta_2 x_{21} + \ldots + \beta_k x_{k1} + \varepsilon_1 \\ y_2 &= \alpha + \beta_1 x_{12} + \beta_2 x_{22} + \ldots + \beta_k x_{k2} + \varepsilon_2 \\ y_3 &= \alpha + \beta_1 x_{13} + \beta_2 x_{23} + \ldots + \beta_k x_{k3} + \varepsilon_3 \\ &\ldots \\ y_n &= \alpha + \beta_1 x_{1n} + \beta_2 x_{2n} + \ldots + \beta_k x_{kn} + \varepsilon_n \end{aligned}$$

Chú ý trong phương trình trên, chúng ta có nhiều biến x (x_1, x_2, ... *đến* x_k), và mỗi biến có một thông số β_j (j = 1, 2, …, k) cần phải ước tính. Vì thế mô hình này còn được gọi là mô hình hồi qui tuyến tính đa biến.

Phương pháp ước tính β_j cũng chủ yếu dựa vào phương pháp bình phương nhỏ nhất. Gọi $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{1i} + ... + \hat{\beta}_k x_{ki}$ là ước tính của y_i , phương pháp bình phương nhỏ nhất tìm giá trị $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_k$ sao cho $\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ nhỏ nhất. Đối với mô hình hồi qui tuyến tính đa biến, cách viết và mô tả mô hình gọn nhất là dùng kí hiệu ma trận. Mô hình [7] có thể thể hiện bằng kí hiệu ma trận như sau:

$$\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$$

Trong đó: **Y** là một vector n x 1, **X** là một matrix n x k phần tử, β và một vector k x 1, và **ε** là vector gồm n x 1 phần tử:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \ldots \\ y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots x_{k1} \\ 1 & x_{12} & x_{22} & \ldots x_{k2} \\ \ldots & \ldots & \ldots & \ldots \\ 1 & x_{1n} & x_{2n} & x_{kn} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \ldots \\ \beta_k \end{bmatrix}, \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \ldots \\ \varepsilon_n \end{bmatrix}$$

Phương pháp bình phương nhỏ nhất giải vector β bằng phương trình sau đây:

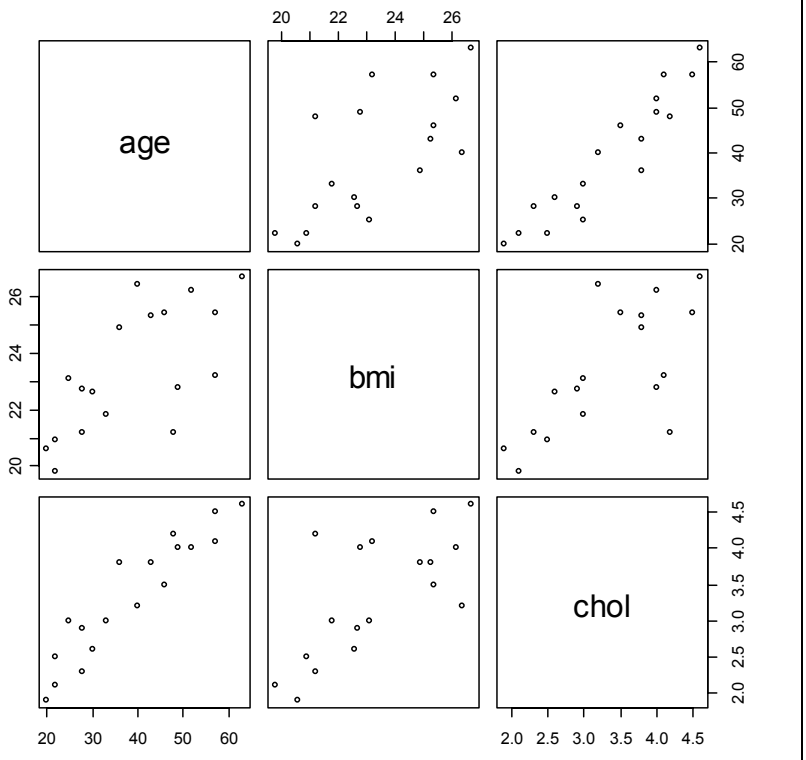
$$\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$$

và tổng bình phương phần dư:

$$\varepsilon^T \varepsilon = \left\| Y - \hat{Y} \right\|^2$$

Ví dụ 2. Chúng ta quay lại nghiên cứu về mối liên hệ giữa độ tuổi, `bmi` và cholesterol. Trong ví dụ, chúng ta chỉ mới xét mối liên hệ giữa độ tuổi 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)
```



Biểu đồ 10.5. Giá trị tiên đoán và khoảng tin cậy 95%.

Cũng như giữa độ tuổi và cholesterol, mối liên hệ giữa bmi và cholesterol cũng gần tuân theo một đường thẳng. Biểu đồ trên còn cho chúng ta thấy độ tuổi và bmi có liên hệ với nhau. Thật vậy, 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: $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ mà tôi vừa trình bày trên. Ở đây, **Y** là một vector 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:

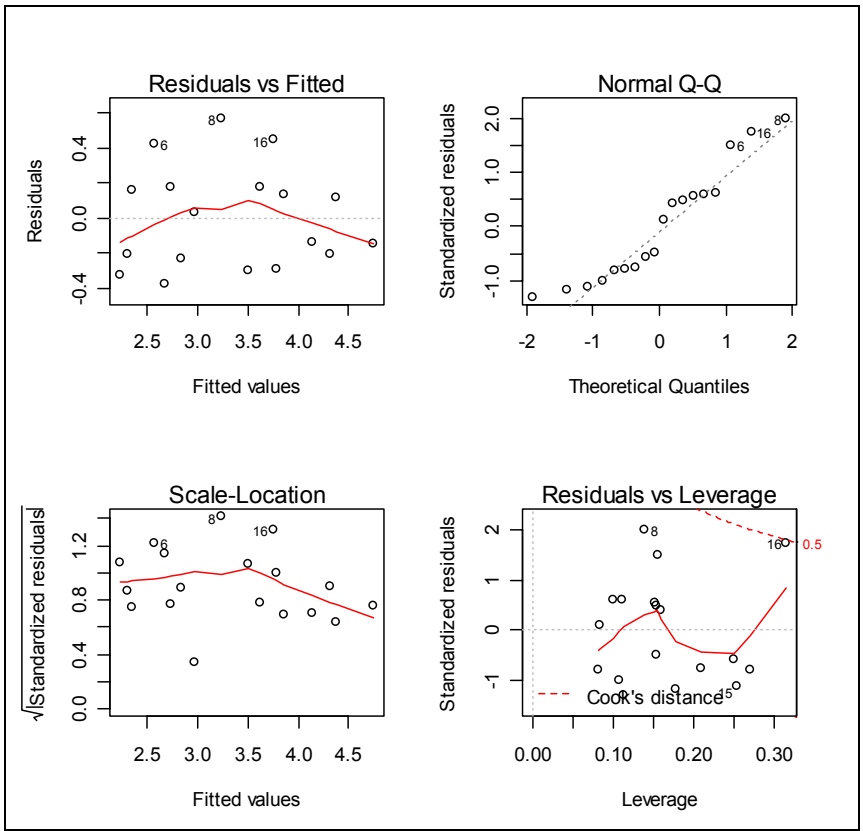
```
> mreg <- lm(chol ~ age + bmi)
> summary(mreg)

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. (Tuy nhiên, 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).



Biểu đồ 10.6. Phân tích phần dư để kiểm tra các giả định trong phân tích hồi qui tuyến tính đa biến.

Tuy BMI không có ý nghĩa thống kê trong trường hợp này, **Biểu đồ 10.6** cho thấy các giả định về mô hình hồi qui tuyến tính có thể đáp ứng.

10.4 Phân tích hồi qui đa thức (Polynomial regression analysis)

Một khai triển tất nhiên từ phân tích hồi qui đa biến độc lập là phân tích hồi qui đa thức. Mô hình hồi qui đa biến mô tả một biến phụ thuộc như là một *hàm số tuyến tính* (linear function) của nhiều biến độc lập, trong khi đó mô hình hồi qui đa thức mô tả một biến phụ thuộc là *hàm số phi tuyến tính* (non-linear function) của một biến độc lập.

Nói theo ngôn ngữ toán học, mô hình hồi qui đa thức tìm mối liên hệ giữa biến phụ thuộc y và biến độc lập x theo những hàm số sau đây:

$$y_i = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + .. + \beta_p x^p + \varepsilon_i.$$

Trong đó các thông số β_j (j = 1, 2, 3, … p) là hệ số đo lường mối liên hệ giữa y và x; và ε_i là phần dư của mô hình, với giả định ε_i tuân theo luật phân phối chuẩn với trung bình 0 và phương sai σ^2. Cho một dãy cặp số (y_1, x_1), (y_2, x_2), (y_3, x_3), …, (y_n, x_n), chúng ta có thể áp dụng phương pháp bình phương nhỏ nhất để ước tính β_j và σ^2.

Trong mô hình trên, chúng ta có thể dễ dàng thấy rằng mô hình hồi qui đa thức còn là một phát triển trực tiếp từ mô hình hồi qui tuyến tính đơn giản. Tức là nếu $\beta_2 = 0$, $\beta_3 = 0$, …, và $\beta_p = 0$, thì mô hình trên đơn giản thành mô hình hồi qui tuyến tính một biến mà chúng ta gặp trong phần đầu của chương này. Nếu $y_i = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon_i$ thì mô hình đơn giản là một phương trình bậc hai, v.v.

Ví dụ 3. Thí nghiệm sau đây tìm mối liên hệ giữa hàm lượng gỗ cứng (hardwoord concentration) và độ căng (tensile strength) của vật liệu. Mười chín vật liệu khác nhau với nhiều hàm lượng gỗ cứng được thử nghiệm để đo độ căng mạnh của vật liệu, và kết quả được tóm lược trong bảng số liệu sau đây:

Id	Hàm lượng gỗ cứng (x)	Độ căng mạnh (y)
1	1.0	6.3
2	1.5	11.1
3	2.0	20.0
4	3.0	24.0
5	4.0	26.1
6	4.5	30.0
7	5.0	33.8
8	5.5	34.0
9	6.0	38.1
10	6.5	39.9
11	7.0	42.0
12	8.0	46.1

13	9.0	53.1
14	10.0	52.0
15	11.0	52.5
16	12.0	48.0
17	13.0	42.8
18	14.0	27.8
19	15.0	21.9

Trước khi phân tích các số liệu này, chúng ta cần nhập số liệu vào R với những lệnh thông thường như sau:

```
> id <- 1:19
> conc <- c(1.0, 1.5, 2.0, 3.0, 4.0,
4.5,  5.0,  5.5,  6.0,
          6.5, 7.0, 8.0, 9.0, 10.0, 11.0, 12.0, 13.0, 14.0, 15.0)
> strength <- c(6.3, 11.1, 20.0, 24.0, 26.1, 30.0, 33.8, 34.0, 38.1,
          39.9, 42.0, 46.1, 53.1, 52.0, 52.5, 48.0, 42.8, 27.8, 21.9)
> data <- data.frame(id, conc, strength)
```

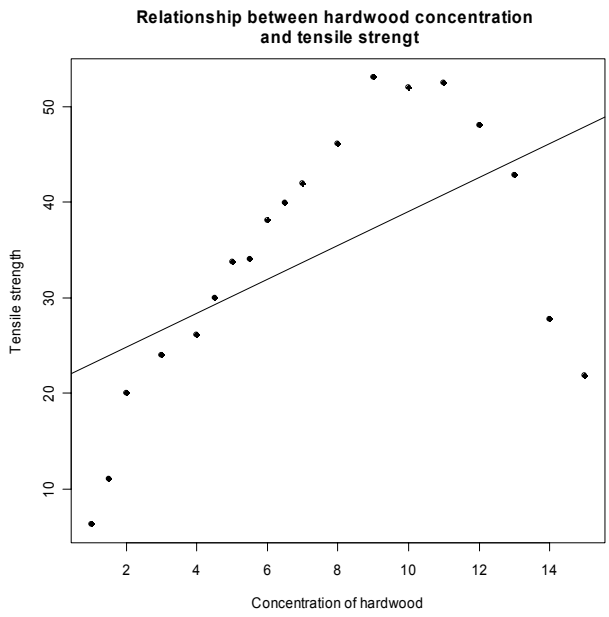
Chúng ta thử xem mô hình hồi qui tuyến tính đơn giản bằng lệnh:

```
> simple.model <- lm(strength ~ conc)
> summary(simple.model)

Call:
lm(formula = strength ~ conc)

Residuals:
    Min      1Q  Median      3Q     Max
-25.986  -3.749   2.938   7.675  15.840

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  21.3213     5.4302   3.926  0.00109 **
conc          1.7710     0.6478   2.734  0.01414 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 11.82 on 17 degrees of freedom
Multiple R-Squared: 0.3054,     Adjusted R-squared: 0.2645
F-statistic: 7.474 on 1 and 17 DF,  p-value: 0.01414
```

Kết quả trên cho thấy mô hình hồi qui tuyến tính đơn giản này (`strength = 21.32 + 1.77*conc`) giải thích khoảng 31% phương sai của `strength`. Ước số phương sai của mô hình này là: $s^2 = (11.82)^2 = 139.7$.

Bây giờ chúng ta xem qua biểu đồ và đường biểu diễn của mô hình trên:

```
> plot(strength ~ conc,
       xlab="Concentration of hardwood",
       ylab="Tensile strength",
       main="Relationship between hardwood concentration \n and tensile
strengt", pch=16)

> abline(simple.model)
```



Biểu đồ 10.7. Mối liên hệ giữa hàm lượng gỗ cứng và độ căng mạnh của vật liệu. Đường thẳng là đường biểu diễn của mô hình hồi qui tuyến tính đơn giản.

Qua biểu đồ này, chúng ta thấy rõ ràng mô hình hồi qui tuyến tính không thích hợp cho số liệu, bởi vì mối liên hệ giữa hai biến này không tuân theo một phương trình đường thẳng, mà là một đường cong. Nói cách khác, một mô hình phương trình bậc hai có lẽ thích hợp hơn. Gọi y là strength và x là conc, chúng ta có thể viết mô hình đó như sau:

$$y_i = \alpha + \beta_1 x + \beta_2 x^2$$

Bây giờ chúng ta sẽ sử dùng R để ước tính ba thông số trên.

```
> quadratic <- lm(strength ~
poly(conc, 2))
> summary(quadratic)

Call:
lm(formula = strength ~ poly(conc, 2))

Residuals:
    Min      1Q  Median      3Q     Max
-5.8503 -3.2482 -0.7267  4.1350  6.5506

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)      34.184      1.014  33.709 2.73e-16 ***
poly(conc, 2)1   32.302      4.420   7.308 1.76e-06 ***
poly(conc, 2)2  -45.396      4.420 -10.270 1.89e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.42 on 16 degrees of freedom
Multiple R-Squared: 0.9085,     Adjusted R-squared: 0.8971
F-statistic: 79.43 on 2 and 16 DF,  p-value: 4.912e-09
```

Như vậy, mô hình mới này `y = 34.18 + 32.30*x - 45.4*x^2` giải thích khoảng 91% phương sai của y. Phương sai của y bây giờ là $s^2 = (4.42)^2 = 19.5$. So với mô hình tuyến tính, mô hình này rõ ràng là tốt hơn rất nhiều.

Chúng ta thử xét một mô hình cubic (bậc ba) $y_i = \alpha + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ xem có mô tả y tốt hơn mô hình phương trình bậc hai hay không.

```
> cubic <- lm(strength ~ poly(conc, 3))
> summary(cubic)
```

```
Call:
lm(formula = strength ~ poly(conc, 3))

Residuals:
     Min       1Q   Median       3Q      Max
-4.62503 -1.61085  0.04125  1.58922  5.02159

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     34.1842     0.5931  57.641  < 2e-16 ***
poly(conc, 3)1  32.3021     2.5850  12.496 2.48e-09 ***
poly(conc, 3)2 -45.3963     2.5850 -17.561 2.06e-11 ***
poly(conc, 3)3 -14.5740     2.5850  -5.638 4.72e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.585 on 15 degrees of freedom
Multiple R-Squared: 0.9707,     Adjusted R-squared: 0.9648
F-statistic: 165.4 on 3 and 15 DF,  p-value: 1.025e-11
```

Mô hình cubic này thậm chí có khả năng mô tả *y* tốt hơn hai mô hình trước, với hệ số xác định bội (R^2) bằng 0.97, và tất cả các thông số trong mô hình đều có ý nghĩa thống kê. Biểu đồ sau đây so sánh 3 mô hình trên:

lặp lại các mô hình trên:

```
> linear <- lm(strength ~ conc)
> quadratic <- lm(strength ~ poly(conc, 2))
> cubic <- lm(strength ~ poly(conc, 3))
```

tạo nên một biến x với nhiều số gần nhau

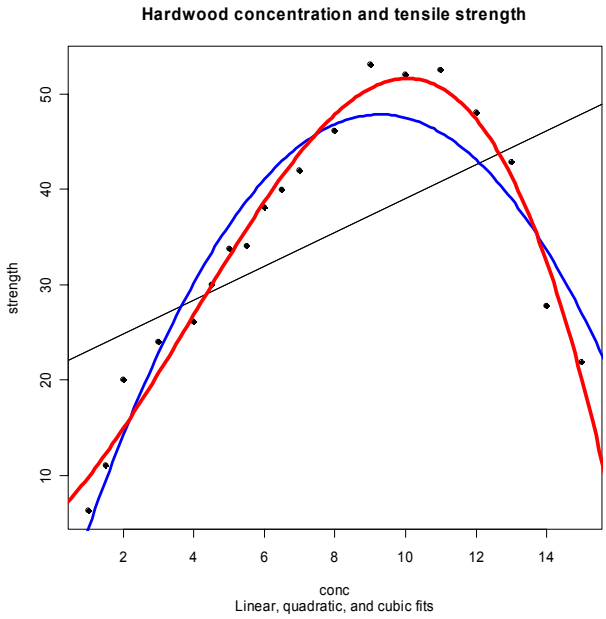
```
> xnew <- (0:160)/10
```

Tính giá trị tiên đoán (predictive values) của y

```
> y2 = predict(quadratic, data.frame(conc=xnew))
> y3 = predict(cubic, data.frame(conc=xnew))
```

Vẽ 3 đường thẳng, bậc hai và bậc 3

```
> plot(strength ~ conc, pch=16,
       main="Hardwood concentration and tensile strength",
       sub="Linear, quadratic, and cubic fits")
> abline(linear, col="black")
> lines(xnew, y2, col="blue", lwd=3)
> lines(xnew, y3, col="red", lwd=4)
```



10.5 Xây dựng mô hình tuyến tính từ nhiều biến

Trong một nghiên cứu thông thường với một biến số phụ thuộc, nhiều biến số độc lập $x_1, x_2, x_3, \ldots, x_k$, mà k có thể lên đến hàng chục, thậm chí hàng trăm. Các biến độc lập đó thường liên hệ với nhau. Có rất nhiều tổ hợp biến độc lập có khả năng tiên đoán biến phụ thuộc y. Ví dụ nếu chúng ta có 3 biến độc lập x_1, x_2, và x_3, để xây dựng mô hình tiên đoán y, chúng ta có thể phải xem xét các mô hình sau đây: $y = f_1(x_1)$, $y = f_2(x_2)$, $y = f_3(x_3)$, $y = f_4(x_1, x_2)$, $y = f_5(x_1, x_3)$, $y = f_6(x_3, x_3)$, $y = f_7(x_1, x_2, x_3)$, v.v… trong đó f_k là những hàm số được định nghĩa bởi hệ số liên quan đến các biến cụ thể. Khi k cao, số lượng mô hình cũng lên rất cao.

Vấn đề đặt ra là trong các mô hình đó, mô hình nào có thể tiên đoán y một cách đầy đủ, đơn giản và hợp lí. Tôi sẽ quay lại ba tiêu chuẩn này trong chương phân tích hồi qui logistic. Ở đây, tôi chỉ muốn bàn đến một tiêu chuẩn thống kê để xây dựng mộ mô hình hồi qui tuyến tính. Trong trường hợp có nhiều mô hình như thế, tiêu chuẩn thống kê để chọn một mô hình tối ưu thường dựa vào tiêu chuẩn thông tin Akaike (còn gọi là AIC hay Akaike Information Criterion).

Cho một mô hình hồi qui tuyến tính $\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$, chúng ta có k+1 thông số $\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$), và có thể tính tổng bình phương phần dư (residual sum of squares, *RSS*):

$$RSS = \sum_{i=1}^{n} \left(\hat{y}_i - y_i \right)^2$$

Trong đó, n là số lượng mẫu. Công thức trên cho thấy nếu mô hình mô tả y đầy đủ thì RSS sẽ thấp, vì độ khác biệt giữa giá trị tiên đoán $\hat{y}$ và giá trị quan sát y gần nhau. Một qui luật chung của phân tích hồi qui tuyến tính là một mô hình với k biến độc lập sẽ có *RSS* thấp hơn mô hình với k-1 biến; và tương tự mô hình với k-1 biến sẽ có RSS thấp hơn mô hình với k-2 biến, v.v… Nói cách khác, mô hình càng có nhiều biến độc lập sẽ "giải thích" y càng tốt hơn. Nhưng vì một số biến độc lập x liên hệ với nhau, cho nên có thêm nhiều biến không có nghĩa là *RSS* sẽ giảm một cách có ý nghĩa. Một phép tính để dung hòa *RSS* và số biến độc lập trong một mô hình là AIC, được định nghĩa như sau:

$$AIC = \log\left(\frac{RSS}{n}\right) + \frac{2k}{n}$$

Mô hình nào có giá trị AIC thấp nhất được xem là mô hình "tối ưu". Trong ví dụ sau đây, chúng ta sẽ dùng hàm `step` để tìm một mô hình tối ưu dựa vào giá trị AIC.

Ví dụ 4. Để nghiên cứu ảnh hưởng của các yếu tố như nhiệt độ, thời gian, và thành phần hóa học đến sản lượng CO_2. Số liệu của nghiên cứu này có thể tóm lược trong bảng số 2. Mục tiêu chính của nghiên cứu là tìm một mô hình hồi qui tuyến tính để tiên đoán sản lượng CO_2, cũng như đánh giá độ ảnh hưởng của các yếu tố này.

Bảng 2. Sản lượng CO_2 và một số yếu tố có thể ảnh hưởng đến CO_2

Id	y	X1	X2	X3	X4	X5	X6	X7
1	36.98	5.1	400	51.37	4.24	1484.83	2227.25	2.06
2	13.74	26.4	400	72.33	30.87	289.94	434.90	1.33
3	10.08	23.8	400	71.44	33.01	320.79	481.19	0.97
4	8.53	46.4	400	79.15	44.61	164.76	247.14	0.62
5	36.42	7.0	450	80.47	33.84	1097.26	1645.89	0.22
6	26.59	12.6	450	89.90	41.26	605.06	907.59	0.76
7	19.07	18.9	450	91.48	41.88	405.37	608.05	1.71
8	5.96	30.2	450	98.60	70.79	253.70	380.55	3.93
9	15.52	53.8	450	98.05	66.82	142.27	213.40	1.97
10	56.61	5.6	400	55.69	8.92	1362.24	2043.36	5.08
11	26.72	15.1	400	66.29	17.98	507.65	761.48	0.60
12	20.80	20.3	400	58.94	17.79	377.60	566.40	0.90
13	6.99	48.4	400	74.74	33.94	158.05	237.08	0.63
14	45.93	5.8	425	63.71	11.95	130.66	1961.49	2.04
15	43.09	11.2	425	67.14	14.73	682.59	1023.89	1.57
16	15.79	27.9	425	77.65	34.49	274.20	411.30	2.38
17	21.60	5.1	450	67.22	14.48	1496.51	2244.77	0.32
18	35.19	11.7	450	81.48	29.69	652.43	978.64	0.44
19	26.14	16.7	450	83.88	26.33	458.42	687.62	8.82
20	8.60	24.8	450	89.38	37.98	312.25	468.38	0.02
21	11.63	24.9	450	79.77	25.66	307.08	460.62	1.72
22	9.59	39.5	450	87.93	22.36	193.61	290.42	1.88
23	4.42	29.0	450	79.50	31.52	155.96	233.95	1.43
24	38.89	5.5	460	72.73	17.86	1392.08	2088.12	1.35
25	11.19	11.5	450	77.88	25.20	663.09	994.63	1.61
26	75.62	5.2	470	75.50	8.66	1464.11	2196.17	4.78

27	36.03	10.6	470	83.15	22.39	720.07	1080.11	5.88

Chú thích: y = sản lượng CO_2; X1 = thời gian (phút); X2 = nhiệt độ (C); X3 = phần trăm hòa tan; X4 = lượng dầu (g/100g); X5 = lượng than đá; X6 = tổng số lượng hòa tan; X7 = số hydrogen tiêu thụ.

Trước khi phân tích số liệu, chúng ta cần nhập số liệu vào R bằng các lệnh thông thường. Số liệu sẽ chứa trong đối tượng `REGdata`.

```
> y <- c(36.98,13.74,10.08, 8.53,36.42,26.59,19.07, 5.96,15.52,56.61,
       26.72,20.80, 6.99,45.93,43.09,15.79,21.60,35.19,26.14, 8.60,
       11.63, 9.59, 4.42,38.89,11.19,75.62,36.03)
> x1 <- c(5.1,26.4,23.8,46.4, 7.0,12.6,18.9,30.2,53.8,5.6,15.1,20.3,48.4,
        5.8,11.2,27.9,5.1,11.7,16.7,24.8,24.9,39.5,29.0, 5.5, 11.5,
        5.2,10.6)
> x2 <- c(400,400, 400, 400, 450, 450, 450, 450, 450, 400, 400, 400,
        400, 425, 425, 425, 450, 450, 450, 450, 450, 450, 450, 460,
        450, 470, 470)
> x3 <- c(51.37,72.33,71.44,79.15,80.47,89.90,91.48,98.60,98.05,55.69,
        66.29,58.94,74.74,63.71,67.14,77.65,67.22,81.48,83.88,89.38,
        79.77,87.93,79.50,72.73,77.88,75.50,83.15)
> x4 <- c(4.24,30.87,33.01,44.61,33.84,41.26,41.88,70.79,66.82,
        8.92,17.98,17.79,33.94,11.95,14.73,34.49,14.48,29.69,26.33,
        37.98,25.66,22.36,31.52,17.86,25.20, 8.66,22.39)
> x5 <- c(1484.83, 289.94, 320.79, 164.76, 1097.26, 605.06, 405.37,
        253.70, 142.27,1362.24, 507.65, 377.60,  158.05, 130.66,
        682.59, 274.20, 1496.51, 652.43,  458.42, 312.25, 307.08,
        193.61, 155.96,1392.08, 663.09,1464.11, 720.07)
> x6 <- c(2227.25, 434.90, 481.19, 247.14,1645.89, 907.59, 608.05,
        380.55, 213.40,2043.36, 761.48, 566.40, 237.08,1961.49,1023.89,
        411.30,2244.77, 978.64, 687.62, 468.38, 460.62, 290.42,
        233.95,2088.12, 994.63,2196.17,1080.11)
> x7 <- c(2.06,1.33,0.97,0.62,0.22,0.76,1.71,3.93,1.97,5.08,0.60,0.90,
        0.63,2.04,1.57,2.38,0.32,0.44,8.82,0.02,1.72,1.88,1.43,
        1.35,1.61,4.78,5.88)

> REGdata <- data.frame(y, x1,x2,x3,x4,x5,x6,x7)
```

Trước khi phân tích số liệu, chúng ta cần nhập số liệu vào R bằng các lệnh thông thường. Số liệu sẽ chứa trong đối tượng `REGdata`.

Bây giờ chúng ta bắt đầu phân tích. Mô hình đầu tiên là mô hình gồm tất cả 7 biến độc lập như sau:

```
> reg <- lm(y ~ x1+x2+x3+x4+x5+x6+x7, data=REGdata)
> summary(reg)

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = REGdata)

Residuals:
    Min      1Q  Median      3Q     Max
-20.035  -4.681  -1.144   4.072  21.214

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 53.937016  57.428952   0.939   0.3594
```

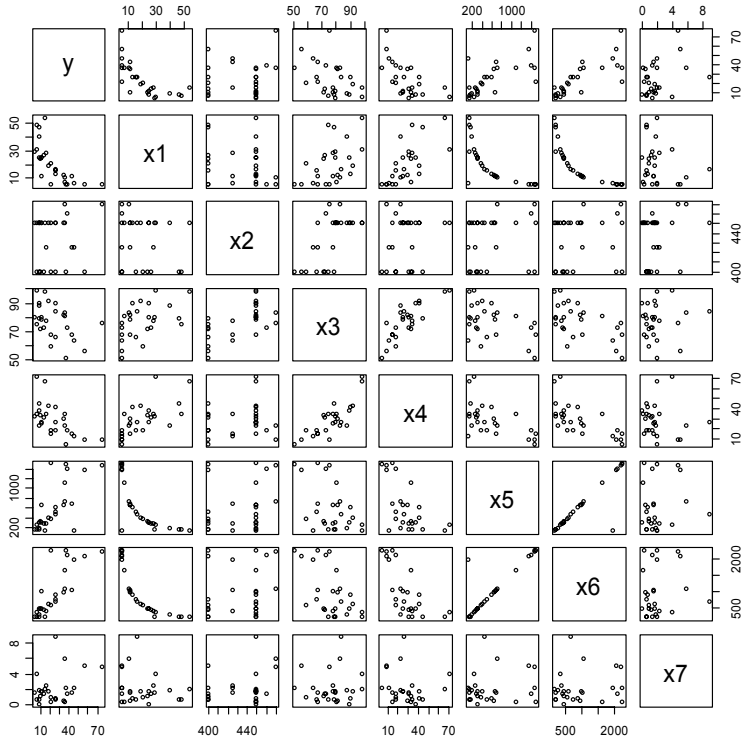
```
x1          -0.127653   0.281498  -0.453   0.6553
x2          -0.229179   0.232643  -0.985   0.3370
x3           0.824853   0.765271   1.078   0.2946
x4          -0.438222   0.358551  -1.222   0.2366
x5          -0.001937   0.009654  -0.201   0.8431
x6           0.019886   0.008088   2.459   0.0237 *
x7           1.993486   1.089701   1.829   0.0831 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.61 on 19 degrees of freedom
Multiple R-Squared: 0.728,      Adjusted R-squared: 0.6278
F-statistic: 7.264 on 7 and 19 DF,  p-value: 0.0002674
```

Kết quả trên cho thấy tất cả 7 biến số "giải thích" khoảng 73% phương sai của *y*. Nhưng trong 7 biến đó, chỉ có `x6` là có ý nghĩa thống kê (p = 0.024). Chúng ta thử giảm mô hình thành một mô hình hồi qui tuyến tính đơn giản với chỉ biến `x6`.

```
> summary(lm(y ~ x6, data=REGdata))

Call:
lm(formula = y ~ x6, data = REGdata)

Residuals:
    Min      1Q  Median      3Q     Max
-28.081  -5.829  -0.839   5.522  26.882

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.144181   3.483064   1.764     0.09 .
x6          0.019395   0.002932   6.616 6.24e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.7 on 25 degrees of freedom
Multiple R-Squared: 0.6365,     Adjusted R-squared: 0.6219
F-statistic: 43.77 on 1 and 25 DF,  p-value: 6.238e-07
```

Chỉ với một biến `x6` mà mô hình có thể giải thích khoảng 64% phương sai của *y*. Chúng ta chấp nhận mô hình này? Trước khi chấp nhận mô hình này, chúng ta phải xem xét độ tương quan giữa các biến độc lập:

```
> pairs(REGdata)
```



Kết quả trên cho thấy *y* có liên hệ với các biến như `x1`, `x5` và `x6`. Ngoài ra, biến `x5` và `x6` có một mối liên hệ rất mật thiết (gần như là một đường thẳng) với hệ số tương quan là 0.88. Ngoài ra, `x5` và `x1` hay `x6` và `x5` cũng có liên hệ với nhau nhưng theo một hàm số nghịch đảo. Điều này có nghĩa là biến `x5` và `x6` cung cấp một lượng thông tin như nhau để tiên đoán *y*, tức là chúng ta không cần cả hai trong mộ mô hình.

Để tìm một mô hình tối ưu trong bối cảnh có nhiều mối tương quan như thế, chúng ta ứng dụng `step` như sau. Chú ý cách cung cấp thông số `lm(y ~ .)`, dấu "." có nghĩa là yêu cầu R xem xét tất cả biến trong đối tượng `REGdata`.

```
> reg <- lm(y ~ ., data=REGdata)
> step(reg, direction="both")
```

```
Start:  AIC= 134.07
 y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7
       Df Sum of Sq     RSS     AIC
- x5    1      4.54 2145.37  132.13
- x1    1     23.17 2164.00  132.36
- x2    1    109.34 2250.18  133.42
- x3    1    130.90 2271.74  133.68
<none>               2140.83  134.07
- x4    1    168.31 2309.14  134.12
- x7    1    377.09 2517.92  136.45
- x6    1    681.09 2821.92  139.53
```

```
Step 1:  AIC= 132.13
 y ~ x1 + x2 + x3 + x4 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x1    1      22.7 2168.1  130.4
- x2    1     113.8 2259.1  131.5
- x3    1     133.5 2278.9  131.8
<none>               2145.4  132.1
- x4    1     170.8 2316.2  132.2
+ x5    1       4.5 2140.8  134.1
- x7    1     375.7 2521.1  134.5
- x6    1    1058.5 3203.8  141.0
```

```
Step 2:  AIC= 130.42
 y ~ x2 + x3 + x4 + x6 + x7
```

```
Step 3:  AIC= 129.59
 y ~ x3 + x4 + x6 + x7
```

```
       Df Sum of Sq    RSS    AIC
- x2    1      96.8 2264.9  129.6
- x3    1     122.0 2290.0  129.9
<none>              2168.1  130.4
- x4    1     187.4 2355.5  130.7
+ x1    1      22.7 2145.4  132.1
+ x5    1       4.1 2164.0  132.4
- x7    1     385.0 2553.1  132.8
- x6    1    1526.2 3694.3  142.8
```

```
       Df Sum of Sq    RSS    AIC
- x3    1      25.4 2290.3  127.9
- x4    1      90.9 2355.8  128.7
<none>              2264.9  129.6
+ x2    1      96.8 2168.1  130.4
+ x5    1       8.3 2256.5  131.5
+ x1    1       5.7 2259.1  131.5
- x7    1     384.9 2649.7  131.8
- x6    1    2015.6 4280.5  144.8
```

```
Step 4:  AIC= 127.9
 y ~ x4 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x4    1      73.5 2363.8  126.7
<none>              2290.3  127.9
+ x3    1      25.4 2264.9  129.6
+ x1    1      11.3 2279.0  129.8
+ x5    1       6.3 2284.0  129.8
+ x2    1       0.3 2290.0  129.9
- x7    1     486.6 2776.9  131.1
- x6    1    1993.8 4284.1  142.8
```

```
Step 5:  AIC= 126.75
 y ~ x6 + x7

       Df Sum of Sq    RSS    AIC
<none>              2363.8  126.7
+ x4    1      73.5 2290.3  127.9
+ x1    1      33.4 2330.4  128.4
+ x3    1       8.1 2355.8  128.7
+ x5    1       7.7 2356.1  128.7
+ x2    1       7.3 2356.6  128.7
- x7    1     497.3 2861.2  129.9
- x6    1    4477.0 6840.8  153.4
```

```
Call:
lm(formula = y ~ x6 + x7, data =
REGdata)

Coefficients:
(Intercept)           x6           x7
    2.52646      0.01852      2.18575
```

Quá trình tìm mô hình tối ưu dừng ở mô hình với hai biến x6 và x7, vì mô hình này có giá trị AIC thấp nhất. Phương trình tuyến tính tiên đoán *y* là: $y = 2.526 + 0.0185(x6) + 2.186(x7)$.

```
> summary(lm(y ~ x6+x7, data=REGdata))

Call:
lm(formula = y ~ x6 + x7, data = REGdata)

Residuals:
     Min       1Q   Median       3Q      Max
-23.2035  -4.3713   0.2513   4.9339  21.9682

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.526460   3.610055   0.700   0.4908
x6          0.018522   0.002747   6.742 5.66e-07 ***
x7          2.185753   0.972696   2.247   0.0341 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.924 on 24 degrees of freedom
Multiple R-Squared: 0.6996,     Adjusted R-squared: 0.6746
F-statistic: 27.95 on 2 and 24 DF,  p-value: 5.391e-07
```

Phân tích chi tiết (kết quả trên) cho thấy hai biến này giải thích khoảng 70% phương sai của *y*.

10.6 Xây dựng mô hình tuyến tính bằng Bayesian Model Average (BMA)

Một vấn đề trong cách xây dựng mô hình trên là mô hình với `x6` và `x7` được xem là mô hình sau cùng, trong khi đó chúng ta biết rằng một mô hình `x5` và `x7` cũng có thể là một mô hình khả dĩ, bởi vì `x5` và `x6` có mối tương quan rất gần nhau. Nếu nghiên cứu được tiến hành tiếp và với thêm số liệu mới, có lẽ một mô hình khác sẽ "ra đời".

Để đánh giá sự bất định trong việc xây dựng mô hình thống kê, một phép tính khác có triển vọng tốt hơn cách phép tính trên là BMA (Bayesian Model Average). Bạn đọc muốn tìm hiểu thêm về phép tính này có thể tham khảo vài bài báo khoa học dưới đây. Nói một cách ngắn gọn, phép tính BMA tìm tất cả các mô hình khả dĩ (với 7 biến độc lập, số mô hình khả dĩ là $2^7 = 128$, chưa tính đến các mô hình tương tác!) và trình bày kết quả của các mô hình được xem là "tối ưu" nhất về lâu về dài. Tiêu chuẩn tối ưu cũng dựa vào giá trị AIC.

Để tiến hành phép tính BMA, chúng ta phải dùng đến package BMA (có thể tải về từ trang web của R http://cran.R-project.org). Sau khi đã có cài đặt package BMA trong máy tính, chúng ta ra phải nhập BMA vào môi trường vận hành của R bằng lệnh:

```
> library(BMA)
```

Sau đó, tạo ra một ma trận chỉ gồm các biến độc lập. Trong data frame chúng ta biết `REGdata` có 8 biến, với biến số 1 là *y*. Do đó, lệnh `REGdata[, -1]` có nghĩa là tạo ra một data frame mới ngoại trừ cột thứ nhất (tức *y*).

```
> xvars <- REGdata[,-1]
```

Kế tiếp, chúng ta định nghĩa biến phụ thuộc tên `co2` từ `REGdata`:

```
> co2 <- REGdata[,1]
```

Bây giờ chúng ta đã sẵn sàng phân tích bằng phép tính BMA. Hàm `bicreg` được viết đặc biệt cho phân tích hồi qui tuyến tính. Cách áp dụng hàm `bicreg` như sau:

```
> bma <- bicreg(xvars, co2, strict=FALSE, OR=20)
```

Chúng ta sử dụng hàm `summary` để biết kết quả:

```
> summary(bma)
Call:
bicreg(x = xvars, y = co2, strict = FALSE, OR = 20)

  16  models were selected
```

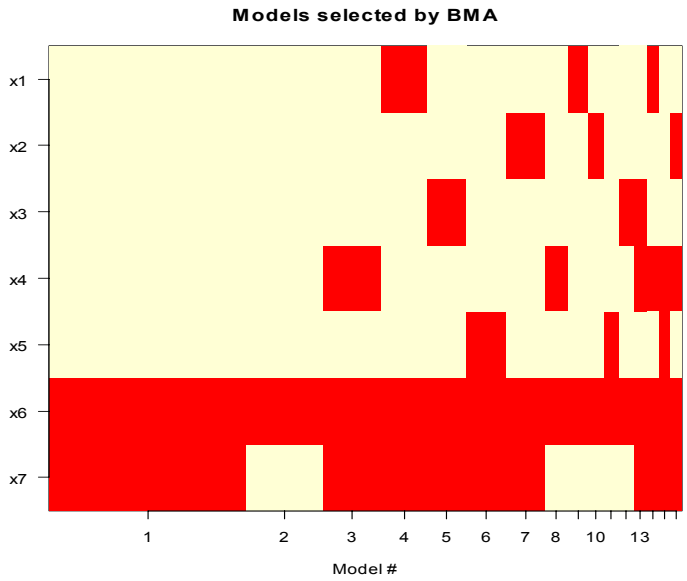
```
 Best  5  models (cumulative posterior probability =  0.6599 ):

           p!=0    EV       SD       model 1   model 2   model 3   model 4   model 5
Intercept  100.0   5.75672  14.6244    2.5264    6.1441    8.6120    7.5936    7.3537
x1          12.4  -0.01807   0.1008      .         .         .      -0.1393      .
x2          10.4  -0.00075   0.0282      .         .         .         .         .
x3          10.7   0.00011   0.0791      .         .         .         .      -0.0572
x4          20.2  -0.03059   0.1020      .         .      -0.1419      .         .
x5          10.5  -0.00023   0.0030      .         .         .         .         .
x6         100.0   0.01815   0.0040    0.0185    0.0193    0.0164    0.0162    0.0179
x7          73.7   1.60766   1.2821    2.1857      .       2.1628    2.1233    2.2382

nVar                                      2         1         3         3         3
r2                                      0.700     0.636     0.709     0.704     0.701
BIC                                   -25.8832  -24.0238  -23.4412  -22.9721  -22.6801
post prob                               0.311     0.123     0.092     0.072     0.063
```

BMA trình bày kết quả của 5 mô hình được đánh giá là tối ưu nhất cho tiên đoán *y* (`model 1, model 2, … model 5`).

- Cột thứ nhất liệt kê danh sách các biến số độc lập;
- Cột 2 trình bày xác suất giả thiết một biến độc lập có ảnh hưởng đến *y*. Chẳng hạn như xác suất là `x6` có ảnh hưởng đến *y* là 100%; trong khi đó xác suất mà `x7` có ảnh hưởng đến *y* là 73.7%. Tuy nhiên xác suất các biến khác thấp hơn hay chỉ bằng 20%. Do đó, chúng ta có thể nói rằng mô hình với `x6` và `x7` có lẽ là mô hình tối ưu nhất.
- Cột 3 (`EV`) và 4 (`SD`) trình bày trị số trung bình và độ lệch chuẩn của hệ số cho mỗi biến số độc lập.
- Cột 5 là ước tính hệ số ảnh hưởng (regression coefficient) của mô hình 1. Như thấy trong cột này, mô hình 1 gồm intercept (tức α), và hai biến `x6` và `x7`. Mô hình này giải thích (như chúng ta đã biết qua phân tích phần trên) 70% phương sai của *y*. Trị số BIC (Bayesian Information Criterion) thấp nhất. Trong số tất cả mô hình mà BMA tìm, mô hình này có xác suất xuất hiện là 31.1%.
- Cột 6 là ước tính hệ số ảnh hưởng của mô hình 2. Như thấy trong cột này, mô hình 2 gồm intercept (tức α), và biến `x6`. Mô hình này giải thích 64% phương sai của *y*. Trong số tất cả mô hình mà BMA tìm, mô hình này có xác suất xuất hiện chỉ là 12.3%.
- Các mô hình khác cũng có thể diễn dịch một cách tương tự.

Một cách thể hiện kết quả trên là qua một biểu đồ như sau:

```
> imageplot.bma(bma)
```



Biểu đồ này trình bày 13 mô hình. Trong 13 mô hình đó, biến `x6` xuất hiện một cách nhất quán. Kế đến là biến `x7` cũng có xuất hiện trong một số mô hình, nhưng như chúng ta biết xác suất là 74%.

Trong ví dụ này, cả hai phép tính đều cho ra một kết quả nhất quán, nhưng trong nhiều trường hợp, hai phép tính có thể cho ra kết quả khác nhau. Nhiều nghiên cứu lí thuyết gần đây cho thấy kết quả từ phép tính BMA rất đáng tin cậy, và trong tương lai, có lẽ là phương pháp chuẩn để xây dựng mô hình.

Tài liệu tham khảo cho BMA

Raftery, Adrian E. (1995). Bayesian model selection in social research (with Discussion). Sociological Methodology 1995 (Peter V. Marsden, ed.), pp. 111-196, Cambridge, Mass.: Blackwells.

Một số bài báo liên quan đến BMA có thể tải từ trang web sau đây:
www.stat.colostate.edu/~jah/papers.