10. Phân tích hồi qui tuyến tính

Ví dụ 15. 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:

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:

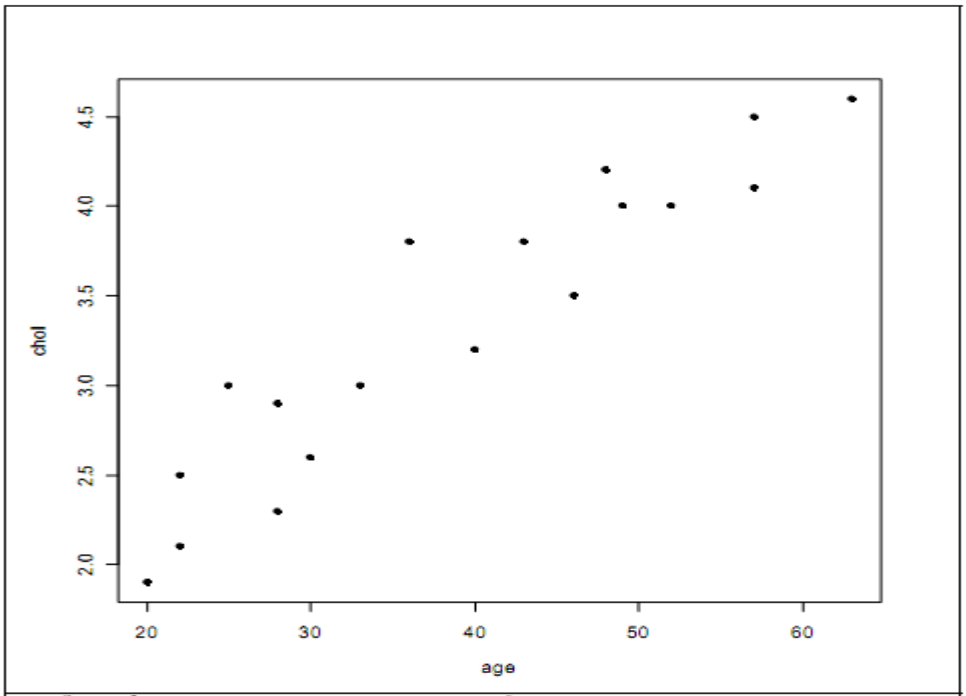
```
> age <- c(46,20,52,30,57,25,28,36,22,43,57,33,22,63,40,48,28,49)

> 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 18 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).



Biểu đồ 18. Liên hệ giữa độ tuổi và cholesterol.

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
```

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")
```

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")
```

10.2 Mô hình của hồi qui tuyến tính đơn giản

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:

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

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:

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \text{và} \quad \hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

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: $s^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{n-2}$. Ở đây, s^2 chính là ước số của σ^2.

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:

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

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). 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.

Các Giả thiết về hồi quy 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`).

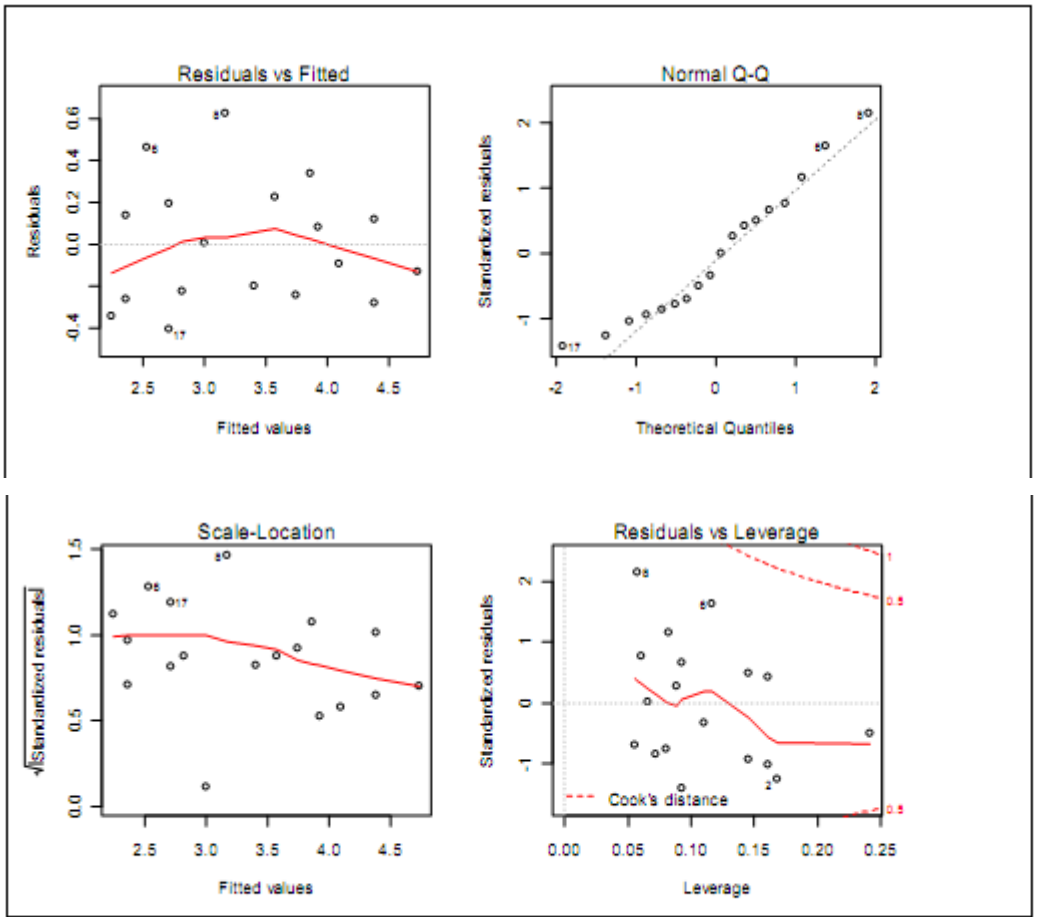
```
> fitted(reg)
       1        2        3        4        5        6        7        8
3.747483 2.244985 4.094214 2.822869 4.383156 2.533927 2.707292 3.169600
       9       10       11       12       13       14       15       16
2.360562 3.574118 4.383156 2.996234 2.360562 4.729886 3.400753 3.863060
      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 lại chúng ta sẽ vẽ các phần dư

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



Biểu đồ 19. 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

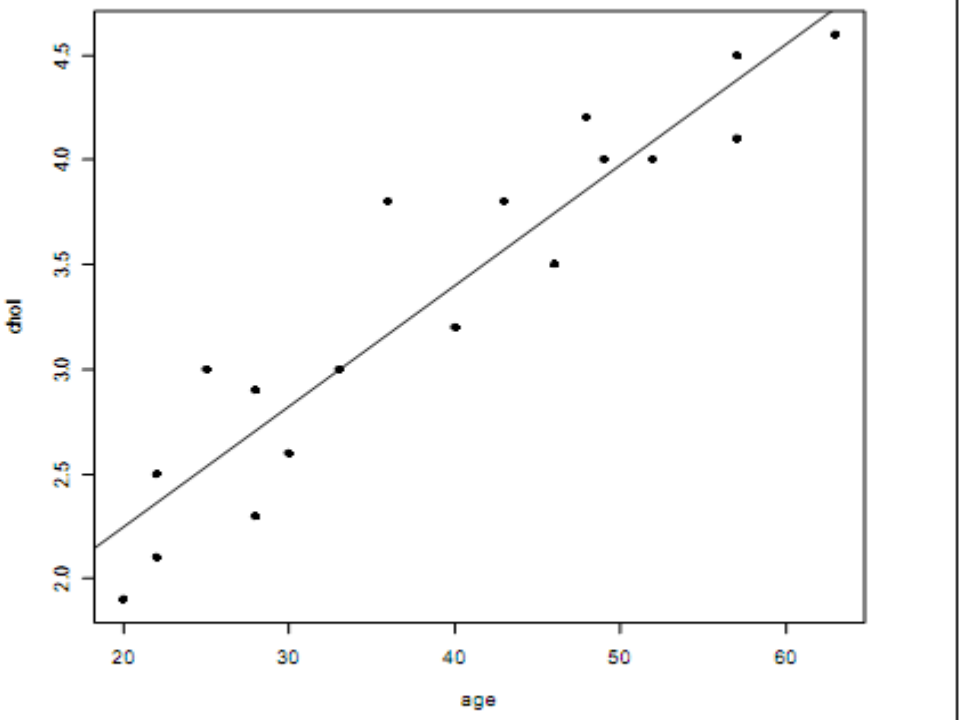
của $\hat{y}_i$, và do đó, giả định (d), tức ε_i có phương sai σ^2 cố định cho tất cả x_i, cũng có thể đáp ứng.

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.

Mô hình dự báo

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)
```

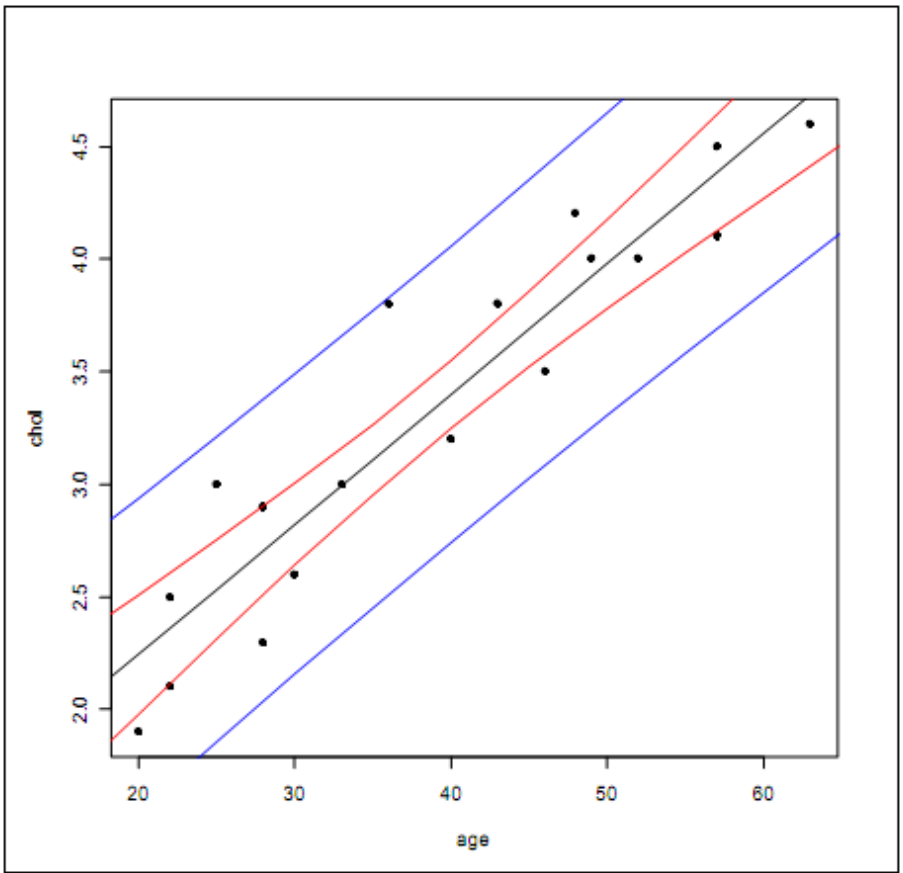


Biểu đồ 20. Đường biểu diễn mối liên hệ giữa độ tuổi (age) và cholesterol.

Nhưng mỗi giá trị $\hat{y}_i$ được tính từ ước số $\bar{\alpha}$ và $\bar{\beta}$, 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 21. 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)

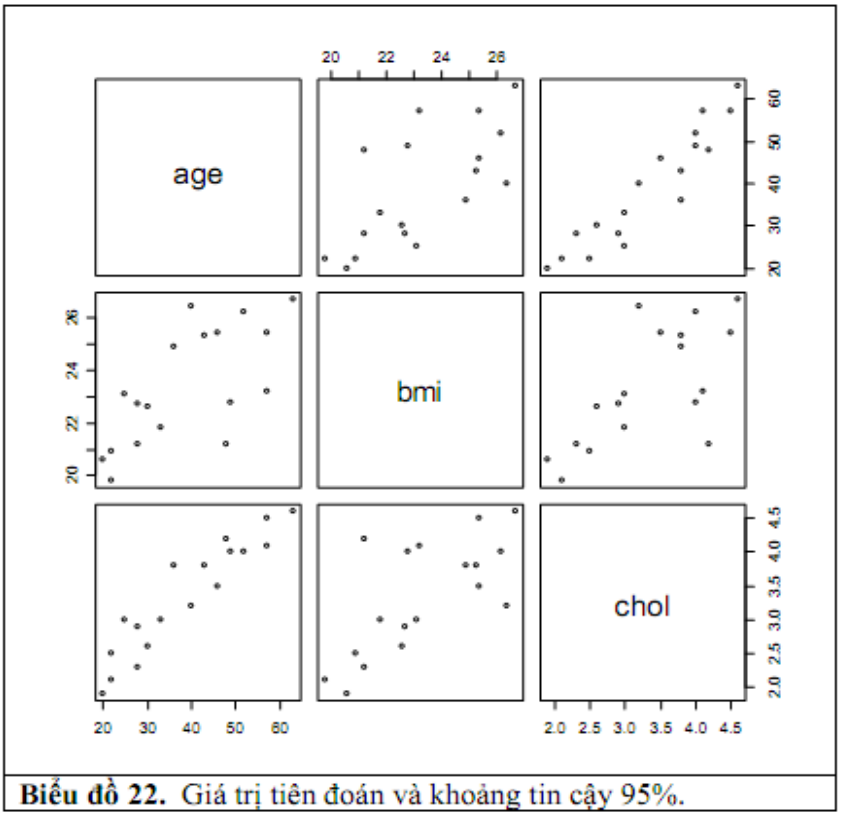
Mô hình được diễn đạt qua phương trình $y_i = \alpha + \beta x_i + \varepsilon_i$ có một yếu tố duy nhất (đó là x), và vì thế thường được gọi là mô hình hồi qui tuyến tính đơn giản (simple linear regression model). Trong thực tế, chúng ta có thể phát triển mô hình này thành nhiều biến, chứ không chỉ giới hạn một biến như trên, chẳng hạn như:

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

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.

Ví dụ 16. 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 22. 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

```
> 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
```

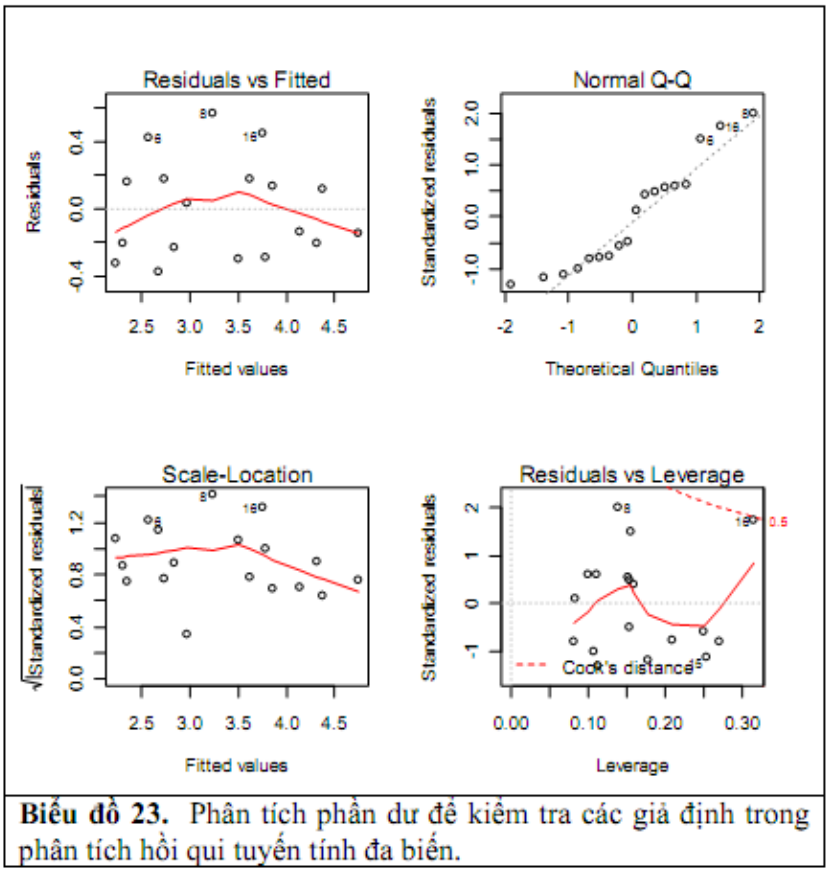
Kết quả phân tích trên cho thấy ước số $\hat{\alpha} = 0.455$, $\hat{\beta}_1 = 0.054$ và $\hat{\beta}_2 = 0.0333$. Nói cách khác, chúng ta có phương trình ước đoán độ cholesterol dựa vào hai biến số độ tuổi và bmi như sau:

```
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 23. 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.

12. Phân tích hồi qui logistic

Trong các phần trước về phân tích hồi qui tuyến tính và phân tích phương sai, chúng ta tìm mô hình và mối liên hệ giữa một biến phụ thuộc liên tục (continuous dependent variable) và một hay nhiều biến độc lập (independent variable) hoặc là liên tục hoặc là không liên tục. Nhưng trong nhiều trường hợp, biến phụ thuộc không phải là biến liên tục mà là biến mang tính đo lường nhị phân: có/không, mắc bệnh/không mắc bệnh, chết/sống, xảy ra/không xảy ra, v.v…, còn các biến độc lập có thể là liên tục hay không liên tục. Chúng ta cũng muốn tìm hiểu mối liên hệ giữa các biến độc lập và biến phụ thuộc.

Ví dụ 19. Trong một nghiên cứu do tôi tiến hành để tìm hiểu mối liên hệ giữa nguy cơ gãy xương (fracture, viết tắt là `fx`) và mật độ xương cùng một số chỉ số sinh hóa khác, 139 bệnh nhân nam (hay nói đúng hơn là đối tượng nghiên cứu) tuổi từ 60 trở lên. Năm 1990, các số liệu sau đây được thu thập cho mỗi đối tượng: độ tuổi (`age`), tỉ trọng cơ thể (body mass index hay `BMI`), mật độ chất khoáng trong xương (bone mineral density hay `BMD`), chỉ số hủy xương `ICTP`, chỉ số tạo xương `PINP`. Các đối tượng nghiên cứu được theo dõi trong vòng 15 năm. Trong thời gian theo dõi, các bệnh nhân bị gãy xương hay không gãy xương được ghi nhận. Câu hỏi đặt ra ban đầu là có một mối liên hệ gì giữa BMD và nguy cơ gãy xương hay không. Số liệu của nghiên cứu này được trình bày trong phần cuối của chương này, và sẽ trình bày một phần dưới đây để bạn đọc nắm được vấn đề.

Một phần số liệu nghiên cứu về các yếu tố nguy cơ cho gãy xương

id	fx	age	bmi	bmd	ictp	pinp
1	1	79	24.7252	0.818	9.170	37.383
2	1	89	25.9909	0.871	7.561	24.685
3	1	70	25.3934	1.358	5.347	40.620
4	1	88	23.2254	0.714	7.354	56.782
5	1	85	24.6097	0.748	6.760	58.358
6	0	68	25.0762	0.935	4.939	67.123
7	0	70	19.8839	1.040	4.321	26.399
8	0	69	25.0593	1.002	4.212	47.515
9	0	74	25.6544	0.987	5.605	26.132
10	0	79	19.9594	0.863	5.204	60.267
...						
137	0	64	38.0762	1.086	5.043	32.835
138	1	80	23.3887	0.875	4.086	23.837
139	0	67	25.9455	0.983	4.328	71.334

Ở đây, vì biến phụ thuộc (gãy xương) không được đo lường theo tính liên tục (mà chỉ là *có* hay *không*), cho nên phương pháp phân tích hồi qui tuyến tính để phân tích mối liên hệ giữa biến phụ thuộc và biến độc lập. Một phương pháp phân tích được phát triển tương đối gần đây (vào thập niên 1970s) có tên là logistic regression analysis (hay phân tích hồi qui logistic) có thể áp dụng cho trường hợp trên.

Trong nghiên cứu này, sau 15 năm theo dõi, có 38 bệnh nhân bị gãy xương. Tính theo phần trăm, tỉ lệ gãy xương là 38 / 139 = 0.273 (hay 27.3%).

12.1 Mô hình hồi qui logistic

Cho một tần số biến cố x ghi nhận từ n đối tượng, chúng ta có thể tính xác suất của biến cố đó là:

$$p = \frac{x}{n}$$

p có thể xem là một chỉ số đo lường nguy cơ của một biến cố. Một cách thể hiện nguy cơ khác là *odds* (một danh từ, nếu tôi không lầm, chỉ có trong tiếng Anh – ngay cả tiếng Pháp, Đức, Tây Ban Nha … cũng không có danh từ tương đương với *odds*). Tôi tạm dịch

odds là *khả năng*. Khả năng của một biến cố được định nghĩa đơn giản bằng tỉ số xác suất biến cố xảy ra trên xác suất biến cố không xảy ra:

$$odds = \frac{p}{1-p}$$

Hàm *logit* của *odds* được định nghĩa như sau:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Cho một biến độc lập x (x có thể là liên tục hay không liên tục), mô hình hồi qui logistic phát biểu rằng:

$$\text{logit}(p) = \alpha + \beta x$$

Tương tự như mô hình hồi qui tuyến tính, α và β là hai thông số tuyến tính cần phải ước tính từ dữ liệu nghiên cứu. Nhưng ý nghĩa của thông số này, đặc biệt là thông số β, rất khác với ý nghĩa mà ta đã quen với mô hình hồi qui tuyến tính. Để hiểu ý nghĩa của hai thông số này, tôi sẽ quay lại với ví dụ 19.

Vấn đề mà chúng ta muốn biết là mối liên hệ giữa mật độ xương `bmd` và nguy cơ gãy xương (`fx`). Để tiện cho việc minh họa, gọi `bmd` là x, vấn đề mà chúng ta cần biết có thể viết bằng ngôn ngữ mô hình như sau

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right)\alpha + \beta x$$

Nói cách khác:

$$odds(p) = \frac{p}{1-p} = e^{\alpha+\beta x}$$

Nói cách khác, mô hình hồi qui logistic vừa trình bày trên phát biểu rằng mối liên hệ giữa xác suất gãy xương (p) và mật độ xương `bmd` là một mối liên hệ theo hình chữ S. Mô hình trên còn cho thấy xác suất gãy xương p tùy thuộc vào giá trị của x. Thành ra, mô hình trên có thể viết một cách chính xác hơn rằng *khả năng* gãy xương với điều kiện x là:

$$odds(p \mid x) = e^{\alpha+\beta x}$$

Khi $x = x_0$, khả năng gãy xương là: $odds(p \mid x = x_0) = e^{\alpha+\beta x_0}$

Khi $x = x_0 + 1$ (tức tăng 1 đơn vị từ x_0), khả năng gãy xương là:

$$odds(p \mid x = x_0 + 1) = e^{\alpha+\beta(x_0+1)}$$

Khi $x = x_0$, khả năng gãy xương là: $odds(p \mid x = x_0) = e^{\alpha+\beta x_0}$

Khi $x = x_0 + 1$ (tức tăng 1 đơn vị từ x_0), khả năng gãy xương là:

$$odds(p \mid x = x_0 + 1) = e^{\alpha+\beta(x_0+1)}$$

Và, tỉ số của hai xác suất gãy xương:

$$\frac{odds\left(p \mid x = x_0 + 1\right)}{odds\left(p \mid x = x_0\right)} = \frac{e^{\alpha+\beta(x_0+1)}}{e^{\alpha+\beta x_0}} = e^{\beta}$$

Trong dịch tễ học, e^{β} được gọi là *odds ratio*. *Odds ratio*, như tên gọi là, *tỉ số khả năng* hay *tỉ số khả dĩ*. Nói cách khác, hệ số β trong mô hình hồi qui logistic chính là tỉ số khả dĩ.

Phương pháp để ước tính thông số trong mô hình [3] khá phức tạp (dùng phương pháp maximum likelihood – tức phương pháp *Hợp lí cực đại*) và không nằm trong phạm vi của cuốn sách này, nên tôi sẽ không trình bày ở đây (bạn đọc có thể tham khảo sách giáo khoa để biết thêm, nếu cần thiết). Tuy nhiên, tôi muốn đề cập ngắn gọn là phương pháp hợp lí cực đại cung cấp cho chúng ta một hệ phương trình như sau:

$$\begin{cases} \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \left(1 + e^{-\left(\hat{\alpha}+\hat{\beta}x_i\right)}\right)^{-1} \\ \sum_{i=1}^{n} x_i y_i = \sum_{i=1}^{n} x_i \left(1 + e^{-\left(\hat{\alpha}+\hat{\beta}x_i\right)}\right) \end{cases}$$

Trong đó, Trong đó, y_i là biến phụ thuộc (gãy xương với giá trị 0 hay 1), và x_i là biến độc lập (mật độ xương), và n là số mẫu. Để tìm ước số $\hat{\alpha}$ và $\hat{\beta}$, một trong những phép tính hay sử dụng là iterative weighted least square hay Newton-Raphson. R sử dụng phép tính Newton-Raphson để tìm hai ước số đó.

Sau khi đã có ước số $\hat{\alpha}$ và $\hat{\beta}$ chúng ta có thể ước tính xác suất p cho bất cứ giá trị nào của x như sau (sau vài thao tác đại số):

$$\hat{p} = \frac{e^{\hat{\alpha}+\hat{\beta}x}}{1 + e^{\hat{\alpha}+\hat{\beta}x}} = \frac{1}{1 + e^{-\left(\hat{\alpha}+\hat{\beta}x\right)}}$$

Chú ý tôi dùng dấu mũ $\hat{p}$ để chỉ số ước tính (predicted value), chứ không phải p là xác suất quan sát. Nếu mô hình mô tả dữ liệu tốt và đầy đủ, độ khác biệt giữa p và $\hat{p}$ nhỏ; nếu mô hình không thích hợp hay không tốt, độ khác biệt đó có thể sẽ cao. Độ khác biệt giữa p và $\hat{p}$ được gọi là *deviance*. Phương pháp tính deviance khá phức tạp, nhưng đó không phải là chủ đề ở đây, cho nên tôi chỉ nói qua khái niệm mà thôi. Khi chúng ta có nhiều mô hình để mô phỏng một hay nhiều mối liên hệ, deviance có thể được sử dụng để đánh giá sự thích hợp của một mô hình, hay để chọn một mô hình "tối ưu".

12.2 Phân tích hồi qui logistic bằng R

Bây giờ, chúng ta quay lại với ví dụ 1, dùng số liệu trong Bảng 12.1 để ước tính hai thông số α và β bằng R. Trước hết chúng ta phải nhập toàn bộ số liệu vào một data frame, và cho một cái tên, chẳng hạn như `fracture`. Trong trường hợp của tôi, dữ liệu được chứa trong directory c:\works\stats dưới tên `fracture.txt`, do đó, các lệnh sau đây cần thiết để nhập số liệu:

```
# báo cho R biết nơi chứa số liệu

> setwd("c:/works/stats")

# nhập số liệu và cho vào một data frame tên fracture

> fracture <- read.table("fracture.txt", header=TRUE, na.string=".")

# kiểm tra xem có bao nhiêu biến trong dữ liệu fracture

> names(fracture)
 [1] "id"   "fx"   "age"  "bmi"  "bmd"  "ictp" "pinp"

# Chọn những bệnh nhân có đầy đủ số liệu cho phân tích

> fulldata <- na.omit(fracture)
> attach(fulldata)
```

Hai biến mà chúng ta quan tâm trong ví dụ này là: `fx` (gãy xương) và `bmd` (mật độ xương). Chúng ta kiểm tra xem có bao nhiêu bệnh nhân gãy xương:

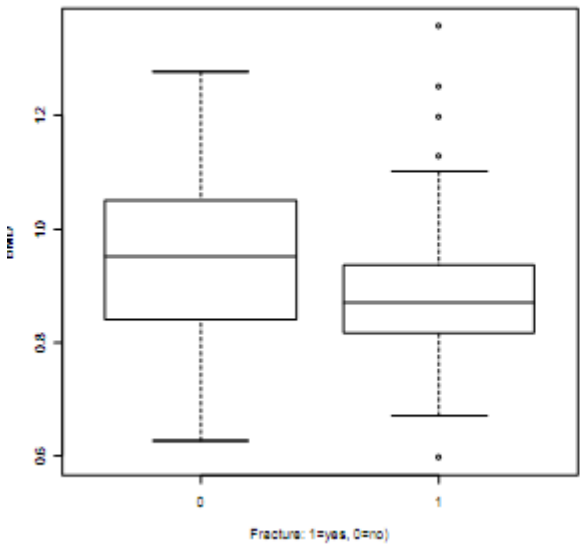
```
> table(fx)
fx
  0   1
101  38
```

Kế đến, xem mật độ xương trong nhóm gãy xương và không gãy xương ra sao:

Kế đến, xem mật độ xương trong nhóm gãy xương và không gãy xương ra sao:

```
> tapply(bmd, fx, mean)
        0         1
0.9444851 0.9016667

> boxplot(bmd ~ fx,
          xlab="Fracture: 1=yes, 0=no)",
          ylab="BMD")
```



Kết quả trên cho thấy, bmd trong nhóm bệnh nhân bị gãy xương thấp hơn so với nhóm không bị gãy xương (0.90 và 0.94). Và, kiểm định t sau đây cho thấy mức độ khác biệt này không có ý nghĩa thống kê (p = 0.15).

```
> t.test(bmd~fx)

        Welch Two Sample t-test

data:  bmd by fx
t = 1.4572, df = 53.952, p-value = 0.1508
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.01609226  0.10172922
sample estimates:
mean in group 0 mean in group 1
      0.9444851       0.9016667
```

Để ước tính thông số trong mô hình [4], hàm số glm (viết tắt từ *generalized linear model*) trong R có thể áp dụng, với "cú pháp" như sau:

```
> logistic <- glm(fx ~ bmd, family="binomial")
> summary(logistic)

Call:
glm(formula = fx ~ bmd, family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0287  -0.8242  -0.7020   1.3780   2.0709

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.063      1.342   0.792    0.428
bmd           -2.270      1.455  -1.560    0.119

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 157.81  on 136  degrees of freedom
Residual deviance: 155.27  on 135  degrees of freedom
AIC: 159.27

Number of Fisher Scoring iterations: 4
```

Tôi sẽ lần lượt giải thích các kết quả trên:

(a) Trong lệnh `logistic <- glm(fx ~ bmd, family="binomial")` chúng ta yêu cầu R phân tích theo mô hình fx là một hàm số với bmd như mô hình [4]. Trong glm có nhiều luật phân phối, mà trong đó phân phối nhị phân (binomial) là một luật phân phối chuẩn cho hồi qui logistic. Do đó, `family="binomial"` cần thiết cho R.

(b) Deviance: phần thứ nhất của kết quả cho biết qua về deviance.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0287  -0.8242  -0.7020   1.3780   2.0709
```

Deviance như giải thích trên phản ánh độ khác biệt giữa mô hình và dữ liệu (cũng tương tự như mean square residual trong phân tích hồi qui tuyến tính vậy). Đối với một mô hình đơn lẻ như ví dụ này thì giá trị của deviance không có ý nghĩa gì nhiều.

(c) Phần kế tiếp cung cấp ước số của $\hat{\alpha}$ (mà R đặt tên là `intercept`) và $\hat{\beta}$ (`bmd`) và sai số chuẩn (standard error).

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.063      1.342   0.792    0.428
bmd            -2.270      1.455  -1.560    0.119
```

Qua kết quả này, chúng ta có $\hat{\alpha}$ = 1.063 và $\hat{\beta}$ = -2.27. Ước số $\hat{\beta}$ là số âm cho thấy mối liên hệ giữa nguy cơ gãy xương và `bmd` là mối liên hệ nghịch đảo: xác suất gãy xương tăng khi giá trị của `bmd` giảm. Tuy nhiên, kiểm định z (tính bằng cách lấy ước số chia cho sai số chuẩn) cho chúng ta thấy ảnh hưởng của `bmd` không có ý nghĩa thống kê, vì trị số p = 0.119.

Nhớ rằng tỉ số khả dĩ (odds ratio hay viết tắt là OR) chính là $e^{-2.27}$ = 0.1033. Nói cách khác, khi `bmd` tăng 1 g/cm^2 (đơn vị đo lường của bmd là g/cm^2) thì tỉ số OR giảm 0.9067 hay 90.67%. Nhưng tăng 1 g/cm^2 là mật độ rất cao trong xương và không thực tế. Cho nên một cách tính khác là tính trên độ lệch chuẩn (standard deviation) của `bmd`. Chúng ta sẽ tìm hiểu độ lệch chuẩn của `bmd`:

```
> sd(bmd)
[1] 0.1406543
```

Do đó, OR sẽ tính trên mỗi 0.14 g/cm^2. Và OR cho mỗi độ lệch chuẩn, do đó, là:

$$e^{-2.27*0.1406} = 0.7267$$

Tức là, khi `bmd` tăng một độ lệch chuẩn thì tỉ số khả dĩ gãy xương giảm khoảng 28%. Cũng có thể nói cách khác, là khi `bmd` *giảm* một độ lệch chuẩn thì tỉ số khả dĩ tăng $e^{2.27*0.1406}$ = 1.376 hay khoảng 38%.

Một cách khác để biết ảnh hưởng của `bmd` là ước tính xác suất gãy xương qua phương trình:

$$\hat{p} = \frac{e^{1.063-2.27(bmd)}}{1+e^{1.063-2.27(bmd)}}$$

Theo đó, khi `bmd` = 1.00, p = 0.23. Khi `bmd` = 0.86 (tức giảm 1 độ lệch chuẩn), p = 0.291. Tức là, nếu BMD giảm 1 độ lệch chuẩn thì xác suất gãy xương tăng 0.291/0.23 = 1.265 hay 26%5.

(d) Phần cuối của kết quả cung cấp deviance cho hai mô hình: mô hình không có biến độc lập (`null deviance`), và mô hình với biến độc lập, tức là `bmd` trong ví dụ (`residual deviance`).

```
    Null deviance: 157.81  on 136  degrees of freedom
Residual deviance: 155.27  on 135  degrees of freedom
AIC: 159.27
```

Qua hai số này, chúng ta thấy `bmd` ảnh hưởng rất thấp đến việc tiên đoán gãy xương, chỉ làm giảm deviance từ 157.8 xuống còn 155.27, và mức độ giảm này không có ý nghĩa thống kê.

Ngoài ra, R còn cung cấp giá trị của AIC (Akaike Information Criterion) được tính từ deviance và bậc tự do. Tôi sẽ quay lại ý nghĩa của AIC trong phần sắp đến khi so sánh các mô hình.

12.3 Ước tính xác suất bằng R

Xin nhắc lại trong phân tích trên, chúng ta cho các kết quả vào đối tượng `logistic`. Trong đối tượng này có nhiều thông tin có ích, nhưng nếu muốn xem các thông tin này chúng ta phải dùng đến các lệnh như `summary` chẳng hạn. Trong phần này, tôi sẽ trình bày một vài hàm để xem xét các thông tin liên quan đến việc tiên đoán xác suất.

- `predict` dùng để liệt kê các giá trị ước tính (predicted values) của mô hình $\log\left(\frac{p}{1-p}\right) = \alpha + \beta x$ cho từng bệnh nhân.

```
> predict(logistic)
          1           2           3           4           5           6
 2.377576584  1.085694014 -2.141117756  1.492824115  0.965379946 -0.941253280
          7           8           9          10          11          12
-1.733686514 -1.675645430 -0.665282957 -0.507046129 -0.941854868 -0.648740461
...
```

Các số trên là log(p / (1 – p)), tức *log odds*, không có ý nghĩa hực tế bao nhiêu. Chúng ta muốn biết giá trị tiên đoán xác suất p tính từ phương trình $\hat{p} = \frac{e^{1.063-2.27(bmd)}}{1+e^{1.063-2.27(bmd)}}$. Để có giá trị này cho từng bệnh nhân, chúng ta cho thông số `type="response"` vào hàm `predict` như sau:

```
> predict(logistic, type="response")
        1         2         3         4         5         6         7
0.91510135 0.74757001 0.10516416 0.81650178 0.72419767 0.28064726 0.15011664
        8         9        10        11        12        13        14
0.15767295 0.33955387 0.37588624 0.28052582 0.34327343 0.44305196 0.23830776
...
```

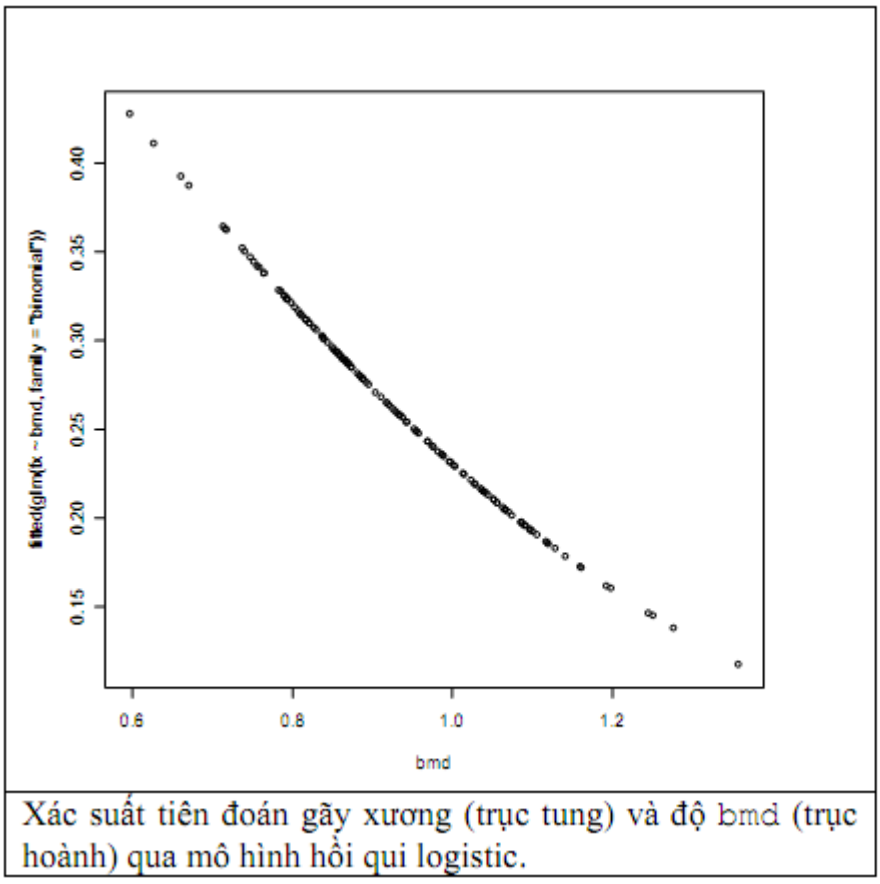
Các số trên là log(p / (1 – p)), tức *log odds*, không có ý nghĩa hực tế bao nhiêu. Chúng ta muốn biết giá trị tiên đoán xác suất p tính từ phương trình $\hat{p} = \frac{e^{1.063-2.27(bmd)}}{1+e^{1.063-2.27(bmd)}}$. Để có giá trị này cho từng bệnh nhân, chúng ta cho thông số `type="response"` vào hàm `predict` như sau:

```
> predict(logistic, type="response")
         1          2          3          4          5          6          7
0.91510135 0.74757001 0.10516416 0.81650178 0.72419767 0.28064726 0.15011664
         8          9         10         11         12         13         14
0.15767295 0.33955387 0.37588624 0.28052582 0.34327343 0.44305196 0.23830776
...
```

Trong kết quả trên (chỉ in một phần) ước tính xác suất gãy xương cho bệnh nhân 1 là 0.915, cho bệnh nhân 2 là 0.747, v.v…

- Chúng ta có thể xem xét các giá trị tiên đoán này với độ `bmd` bằng cách dùng hàm `plot` thông thường:

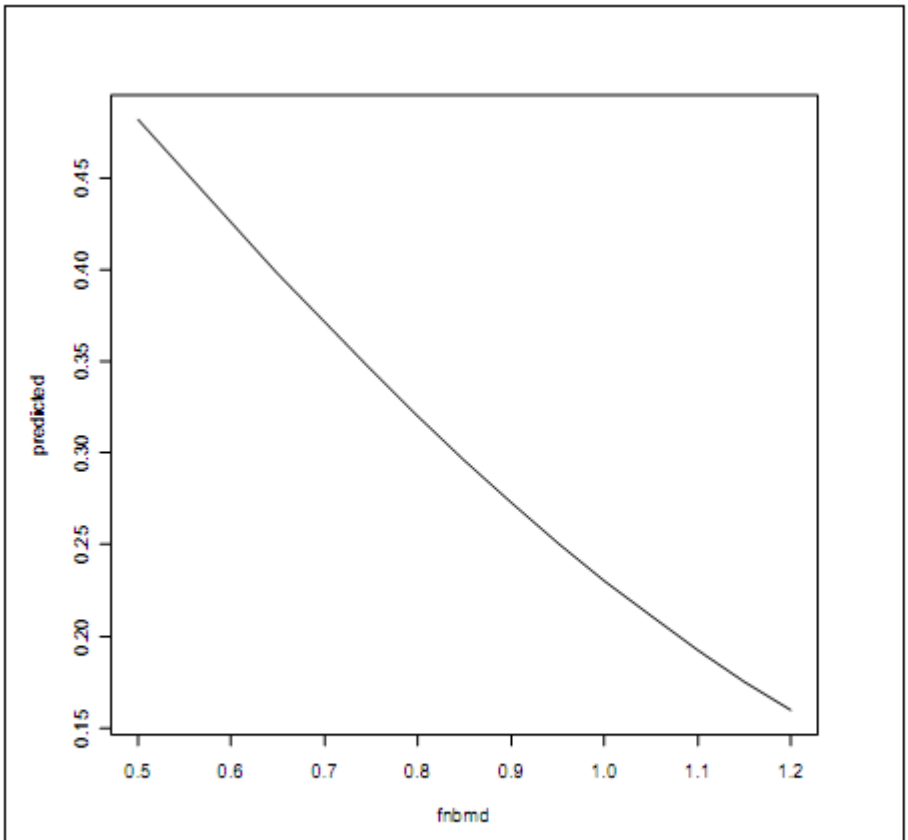
```
> plot(bmd, fitted(glm(fx ~ bmd, family="binomial")))
```



Xác suất tiên đoán gãy xương (trục tung) và độ `bmd` (trục hoành) qua mô hình hồi qui logistic.

Biểu đồ trên có thể cải tiến bằng cách cho các khoảng cách giá trị bmd gần nhau hơn (như 0.50, 0.55, 0.60, …, 1.20 chẳng hạn), và dùng đường biểu diễn thay vì dùng dấu chấm. Các lệnh sau đây sẽ cải tiến biểu đồ.

```
> logistic <- glm(fx ~ bmd, family="binomial")
> fnbmd <- seq(0.5, 1.2, 0.05)  #cho fnbmd từ > 0.50,0.55,0.6,...,1.2
> new.data <- data.frame(bmd = fnbmd) #cho vào một dataframe mới
> predicted <- predict(logistic, new.data, type="response")
> plot(predicted ~ fnbmd, type="l")
```



Xác suất tiên đoán gãy xương (trục tung) và độ bmd (trục hoành) qua mô hình hồi qui logistic.