

Phân tích hồi qui logistic (logistic regression analysis)

Nguyễn Văn Tuấn

Nhiều nghiên cứu y khoa (và khoa học thực nghiệm nói chung) có mục tiêu chính là phân tích mối tương quan giữa một (hay nhiều) *yếu tố nguy cơ* và *nguy cơ mắc bệnh*. Chẳng hạn như đối với một nghiên cứu về mối tương quan giữa thói quen hút thuốc lá và ung thư phổi, thì yếu tố nguy cơ ở đây là thói quen hút thuốc lá và đối tượng phân tích là nguy cơ mắc ung thư phổi. Nói theo thuật ngữ dịch tễ học, yếu tố nguy cơ chính là *risk factors*, và đối tượng phân tích là *outcome*. Trong các nghiên cứu này, đối tượng phân tích thường được thể hiện qua các biến số nhị phân, tức là *có/không*, *mắc bệnh/không mắc bệnh*, *chết/sống*, *xây ra/không xây ra*, v.v... Yếu tố nguy cơ có thể là các biến số liên tục (như độ tuổi, áp suất máu, mật độ xương, v.v...) hay các biến nhị phân (như giới tính) hay biến mang đặc tính thứ bậc (như tình trạng của bệnh dao động từ “nhẹ”, “trung bình” đến “nghiêm trọng”).

Vấn đề đặt ra cho các nghiên cứu dạng này là làm cách nào để ước tính độ tương quan (magnitude of association) giữa yếu tố nguy cơ và bệnh. Các phương pháp phân tích như mô hình hồi qui tuyến tính (linear regression model) không thể áp dụng được, bởi vì biến phụ thuộc (dependent variable) không phải là một biến liên tục, mà là biến nhị phân. Vào thập niên 1970s nhà thống kê học David R. Cox phát triển một mô hình có tên là “logistic regression model” (mà tôi tạm dịch là “mô hình hồi qui logistic”) để phân tích các biến nhị phân. Tôi sẽ giải thích cách ứng dụng mô hình này qua một số ví dụ từ đơn giản đến phức tạp. Tôi sẽ không bàn đến các chi tiết toán học của mô hình hồi qui logistic, mà chỉ tập trung vào các khía cạnh thực tế và diễn dịch kết quả phân tích.

I. Phân tích hồi qui logistic đơn giản cho nghiên cứu đối chứng

Ví dụ 1: Nghiên cứu mối tương quan giữa phơi nhiễm chất độc da cam và ung thư tuyến tiền liệt. Giri và đồng nghiệp (2004) tiến hành một nghiên cứu sơ bộ để thăm định mối liên hệ giữa phơi nhiễm chất độc màu da cam (Agent Orange – AO) và nguy cơ ung thư tuyến tiền liệt (prostate cancer risk) ở các cựu chiến binh Mỹ từng tham chiến ở Việt Nam trước đây. Các nhà nghiên cứu chẩn đoán 47 trường hợp ung thư tiền liệt tuyến từng tham chiến. Sau đó, họ ngẫu nhiên chọn 144 cựu chiến binh cũng từng tham chiến ở Việt Nam và nay nhập viện vì các lí do không liên quan đến ung thư. Gọi nhóm này là nhóm “Đối chứng” (control). Ở mỗi nhóm, các nhà nghiên cứu tìm trong hồ

sơ bệnh lí và phỏng vấn trực tiếp để biết ai là người đã từng phơi nhiễm AO trong thời chiến. Kết quả cho thấy trong số 47 trường hợp ung thư, có 11 người từng bị phơi nhiễm AO, 29 người không từng bị phơi nhiễm, và 7 người không rõ tiền sử; trong nhóm đối chứng có 17 người không từng bị phơi nhiễm, 106 người không từng bị phơi nhiễm, và 21 người không thể xác định phơi nhiễm. Kết quả có thể tóm lược trong bảng số liệu sau đây:

Bảng 1. Phơi nhiễm AO và ung thư tiền liệt tuyến

	Ung thư (n=47)	Đối chứng (n=142)
Phơi nhiễm AO	11	17
Không phơi nhiễm AO	29	106
Không rõ	7	21
Tổng số	47	144

Ghi chú: n là số bệnh nhân. Nguồn số liệu: *Giri VN, Cassidy AE, Beebe-Dimmer J, Ellis LR, Smith DC, Bock CH, Cooney KA. Association between Agent Orange and prostate cancer: a pilot case-control study. Urology. 2004 Apr;63(4):757-60; discussion 760-1. Correction in Urology. 2004 Jun;63(6):1213.*

Để minh họa cho phân tích hồi qui tuyến tính và đơn giản hóa vấn đề, tôi sẽ gộp chung hai nhóm “Không phơi nhiễm AO” và “Không rõ” thành một nhóm chung. (Cách làm này có thể là một đề tài phân tích khác!) Bảng số liệu trên, do đó, có thể rút gọn như sau:

	Ung thư	Đối chứng
Phơi nhiễm AO	11	17
Không phơi nhiễm AO và không rõ	36	127

Qua số liệu trên đây, có thể thấy 23.4% (hay 11/47) nhóm ung thư tiền liệt tuyến từng bị phơi nhiễm AO. Nhưng tỉ lệ này trong nhóm đối chứng là 11.8% (17/144). Vấn đề đặt ra là có sự tương quan nào giữa phơi nhiễm AO và ung thư tiền liệt tuyến hay không? Cụm từ “sự tương quan” có thể khai triển thành hai câu hỏi cụ thể:

- Nguy cơ mắc bệnh ung thư tiền liệt tuyến ở những người từng bị phơi nhiễm so với nguy cơ ở những người không từng bị phơi nhiễm là bao nhiêu?

- Độ khác biệt về nguy cơ ung thư giữa hai nhóm có ý nghĩa thống kê hay không?

Mô hình phân tích hồi qui logistic có thể trả lời hai câu hỏi này. Chỉ số thống kê quan trọng để phân tích số liệu từ các nghiên cứu bệnh – chứng (case-control study) như trên là *tỉ số nguy cơ (odds ratio hay OR)*. Để ước tính OR, tôi phải giải thích từng bước như sau:

Tiếng Anh có một danh từ để mô tả *nguy cơ* hay *khả năng* mà các ngôn ngữ Âu Á khác (như Pháp, Ý, Tây Ban Nha, Trung Quốc, Việt Nam, v.v...) không có: đó là danh từ *odd*. Do đó, tôi sẽ tạm thời không dịch chữ *odd* sang tiếng Việt. Nói một cách ngắn gọn, *odd là tỉ số của hai giá trị của một biến số nhị phân*. Do đó, *OR là tỉ số của hai odds*. Nói cách khác, OR là tỉ số của hai tỉ số! Trong ví dụ trên, chúng ta có:

- *odd* mắc ung thư trong nhóm từng bị phơi nhiễm AO là: $11/17 = 0.647$;
- *odd* mắc ung thư trong nhóm không từng bị phơi nhiễm AO là: $36/127 = 0.283$;
- và odds ratio mắc bệnh ung thư trong nhóm từng bị phơi nhiễm so với nhóm không từng bị phơi nhiễm là: $OR = 0.647 / 0.283 = 2.28$.

Thật ra, OR cũng có thể tính ngắn gọn bằng một công thức:

$$OR = \frac{11 \times 127}{17 \times 36} = 2.28$$

Nói cách khác, nguy cơ mắc bệnh ung thư tiền liệt tuyến trong các cựu chiến binh từng bị phơi nhiễm AO cao hơn các cựu chiến binh không từng bị phơi nhiễm AO khoảng 2.3 lần.

Nhưng vì đây là một nghiên cứu dựa vào một mẫu duy nhất, và ước tính trên đây có thể dao động từ mẫu này sang mẫu khác. Nên nhớ rằng, OR là một ước tính – estimate – của một *OR thật* – true OR – mà chúng ta không biết trong thực tế. Chỉ số nguy cơ thật này có thể dao động bất thường từ thấp hơn 1 đến cao hơn 1. Nếu OR *thật* thấp hơn 1, thì điều này có nghĩa là những người từng phơi nhiễm AO có nguy cơ ung

thấp hơn những người không từng phơi nhiễm AO; một chỉ số OR *thật* cao hơn 1 cho biết những người từng phơi nhiễm AO có nguy cơ ung thư cao hơn những người không từng phơi nhiễm AO; và nếu OR *thật* bằng 1 thì điều này có nghĩa là không có mối liên hệ nào giữa phơi nhiễm AO và ung thư tiền liệt tuyến.

Vì thế, câu hỏi thứ hai (và có lẽ quan trọng hơn) là mối tương quan như phản ánh qua OR có ý nghĩa thống kê hay không? Nói cách khác, nếu nghiên cứu trên được lặp lại rất nhiều lần, thì độ dao động của OR là bao nhiêu? Nếu nghiên cứu được lặp lại (chẳng hạn như) 100 lần, và 95 nghiên cứu cho ra ước số OR dao động từ 1.1 đến 3.8, và 5 nghiên cứu cho thấy OR thấp hơn 1.1 hay cao hơn 3.8, thì chúng ta có bằng chứng để phát biểu rằng mối liên hệ giữa phơi nhiễm AO và ung thư tiền liệt tuyến *có ý nghĩa thống kê – statistically significant*.

Nói cách khác, chúng ta cần phải ước tính *sai số chuẩn (standard error)* cho OR và khoảng tin cậy 95% của OR. Vì OR là một tỉ số, cho nên việc ước tính sai số chuẩn cho OR không thể tiến hành trực tiếp được (hay được nhưng rất phức tạp), mà phải ước tính bằng các phương pháp gián tiếp. Một trong những phương pháp gián tiếp đó là **phương pháp Woolf** và qui trình ước tính có thể mô tả từng bước như sau:

- Trước hết, chúng ta hoán chuyển OR sang đơn vị logarit (natural logarithm):

$$\log OR = \log(OR) = \log(2.28) = 0.824$$

- Bước thứ hai là ước tính sai số chuẩn (tạm cho kí hiệu SE) của logOR qua công thức sau đây:

$$SE = \sqrt{\frac{1}{11} + \frac{1}{17} + \frac{1}{36} + \frac{1}{127}} = 0.430$$

- Bước thứ ba, theo luật phân phối chuẩn, khoảng tin cậy 95% của logOR là: $\log OR \pm 1.96 \times SE$, và trong trường hợp trên, khoảng tin cậy 95% của logOR là:

$$0.824 - 1.96 \times 0.430 = -0.0188$$

$$0.824 + 1.96 \times 0.430 = +1.6668$$

- Vì đơn vị vừa tính là log, cho nên bước thứ tư là hoán chuyển khoảng tin cậy 95% sang đơn vị tỉ số như lúc ban đầu:

$$e^{-0.0188} = 0.98 \quad \text{đến} \quad e^{0.16668} = 5.30$$

Kết quả phân tích trên cho thấy tính trung bình, OR là 2.28, nhưng khoảng tin cậy 95% của OR dao động từ 0.98 đến 5.30. Nói cách khác, nếu nghiên cứu trên được lặp lại 100 lần, sẽ có 95 nghiên cứu cho thấy OR có thể thấp hơn 1 (0.98) hay thậm chí cao đến 5.30.

Đến đây, chúng ta có kết quả để phát biểu cho câu hỏi thứ hai. Bởi vì khoảng tin cậy 95% có thể thấp hơn 1 mà cũng có thể cao hơn 1, cho nên chúng ta phát biểu rằng mối liên hệ giữa phơi nhiễm AO và nguy cơ mắc ung thư tuyến tiền liệt không có ý nghĩa thống kê. Xin nhấn mạnh, đây chỉ mới là một kết luận thống kê, và tôi chưa bàn đến ý nghĩa của số liệu này trên quan điểm lâm sàng vì nó không nằm trong phạm vi của thảo luận.

II. Mô hình hồi qui logistic

Ví dụ trên minh họa cho phương pháp phân tích hồi qui logistic mang tính “thủ công”. Thật ra, mô hình hồi qui logistic có thể thể hiện bằng một mô hình chung. Gọi p là xác suất của một sự kiện (trong ví dụ trên, “sự kiện” ở đây là bệnh ung thư tuyến tiền liệt), thì odd có thể định nghĩa như sau:

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

Gọi tình trạng phơi nhiễm AO là x , và x có hai giá trị: 0 có nghĩa là không từng bị phơi nhiễm, và 1 biểu hiện cho tình trạng từng bị phơi nhiễm AO. Mô hình hồi qui logistic phát biểu rằng $\log(odd)$ tùy thuộc vào giá trị của x qua một hàm số tuyến tính gồm 2 thông số như sau:

$$\log(odd) = \alpha + \beta x + \varepsilon$$

hay,

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x + \varepsilon \quad [1]$$

Trong đó, $\log(\text{odd})$ hay $\log\left(\frac{p}{1-p}\right)$ còn được gọi là $\text{logit}(p)$ (và do đó, mới có tên logistic); α và β là hai thông số cần ước tính từ dữ liệu, và ε là phần dư (residual), tức là phần không thể giải thích bằng x . Lí do hoán chuyển từ p thành $\text{logit}(p)$ là vì p có giá trị trong khoảng 0 và 1, trong khi đó $\text{logit}(p)$ có giá trị vô giới hạn và do đó thích hợp cho việc phân tích theo mô hình hồi qui tuyến tính.

Mô hình trên giả định rằng ε tuân theo luật phân phối chuẩn (normal distribution) với trung bình bằng 0 và phương sai bất biến (constant variance). Với giả định này, giá trị kì vọng (expected value) hay giá trị trung bình của $\log\left(\frac{p}{1-p}\right)$ cho bất cứ giá trị nào của x là: $\alpha + \beta x$ (vì giá trị trung bình của ε là 0). Nói cách khác, odd bị ung thư, từ phương trình [1], là:

$$\text{odd} = \frac{p}{1-p} = e^{\alpha + \beta x + \varepsilon} \quad [2]$$

Như vậy mô hình hồi qui logistic phát biểu rằng odd của một sự kiện (ung thư tuyến tiền liệt) tùy thuộc vào x (tình trạng phơi nhiễm AO). Dựa vào phương trình [1], nhóm không bị phơi nhiễm ($x = 0$) có odd bị ung thư (gọi tắt odd_0) là:

$$\text{odd}_0 = e^{\alpha + \beta \times 0} = e^{\alpha} \quad [3]$$

và nhóm từng bị phơi nhiễm ($x = 1$) có odd bị ung thư (odd_1) là:

$$\text{odd}_1 = e^{\alpha + \beta \times 1} = e^{\alpha + \beta} \quad [4]$$

Tỉ số của hai odds chính là odds ratio (và đó chính là lí do tại sao tôi dịch *odds ratio* là *tỉ số nguy cơ*). Tỉ số nguy cơ – OR – có thể ước tính từ [3] và [4] như sau:

$$OR = \frac{\text{odd}_1}{\text{odd}_0} = \frac{e^{\alpha + \beta}}{e^{\alpha}} = e^{\beta} \quad [5]$$

Trong thực tế, chúng ta không biết giá trị thật của hai thông số α và β , và phải ước tính từ số liệu quan sát được. Theo qui ước thống kê, ước số (estimates) của hai

thông số này được kí hiệu hóa bằng dấu mũ: α và β . Như trong trường hợp ví dụ 1, ước số của thông số β là $\hat{\beta} = 0.824$. Do đó, OR phản ánh odd bị ung thư trong nhóm bị phơi nhiễm AO so với odd trong nhóm không từng bị phơi nhiễm AO. Trong ví dụ 1, $e^{\hat{\beta}} = e^{0.824} = 2.28$.

III. Ước tính thông số của mô hình hồi qui logistic bằng R

Như vừa trình bày, phương pháp ước tính OR và khoảng tin cậy 95% tuy đơn giản, nhưng khá dài dòng. Trong trường hợp có nhiều biến độc lập x , phương pháp tính toán phức tạp hơn và phân tích bằng phương pháp thủ công như trên sẽ tốn nhiều thì giờ. Ngày nay, máy tính và các phần mềm thống kê có thể cung cấp cho chúng ta một phương tiện phân tích rất hữu hiệu. Một trong những phần mềm chuyên phân tích thống kê có tên đơn giản là R mà tôi đã có dịp giới thiệu trong cuốn sách “*Phân tích số liệu và tạo biểu đồ bằng R*” (Nhà xuất bản Khoa học và Kỹ thuật, TPHCM 2007).

Ở đây, tôi sẽ hướng dẫn cách phân tích số liệu trên bằng R. Trước khi phân tích, cần phải nhập dữ liệu vào một khuôn khổ mà R có thể “đọc” được. Để tiện cho việc theo dõi, tôi trình bày bảng số liệu một lần nữa ở đây:

	Ung thư	Đối chứng
Phơi nhiễm AO	11	17
Không phơi nhiễm AO và không rõ	36	127

Ở đây, chúng ta có hai biến, gọi tắt là **ao** và **cancer**; mỗi biến có hai giá trị: 0 (không) và 1 (có). Trong nhóm **ao** = 1 (phơi nhiễm) có 28 đối tượng, và trong số này có 11 người bị ung thư; trong nhóm **ao** = 0 (không phơi nhiễm) có 143 đối tượng và trong số này có 36 người bị ung thư. Chúng ta sẽ “bổ trí” số liệu trên bằng R như sau:

```
ao <- c(1, 0)
ntotal <- c(28, 163)
cancer <- c(11, 36)
proportion <- cancer/ntotal
```

Chú thích:

- Dòng 1 định nghĩa biến **ao** có hai giá trị 1 và 0 (chú ý dấu <- có nghĩa tương đương như dấu bằng "=");
- Dòng 2 định nghĩa biến **ntotals**, và cho biết **ao=1** có 28 đối tượng, **ao=0** có 163 đối tượng;
- Dòng 3 định nghĩa biến **cancer**, và cho biết **ao=1** có 11 đối tượng, **ao=0** có 36 đối tượng;
- Dòng 4 định nghĩa biến **proportion** bằng **cancer** chia cho **ntotals**, có nghĩa là tỉ lệ ung thư cho từng nhóm **ao**.

Sau khi đã nhập số liệu, chúng ta đã sẵn sàng phân tích. Trong R có hàm **glm** chuyên dụng cho phân tích hồi qui logistic. Cách viết hàm này đã được mô tả trong sách của tôi. Ở đây, tôi chỉ giải thích ngắn gọn như sau:

```
logistic <- glm(proportion ~ ao, family="binomial",
               weight=ntotal)
```

Trong lệnh trên, chúng ta yêu cầu R sử dụng hàm **glm** để mô tả **proportion** như là một hàm số của **ao** (chú ý dấu ~ có nghĩa là mô hình), và phân phối của **proportion** là phân phối nhị phân (**binomial**) vì chỉ có 2 giá trị. Ngoài ra, trong lệnh trên, chúng ta còn cho một thông số `weight=ntotal`. Thông số `weight` yêu cầu R sử dụng **ntotal** là một số tóm lược (thay vì một bệnh nhân).

Kết quả phân tích được lưu trữ đối tượng có tên là **logistic** (tất nhiên, chúng ta có thể thay đổi với một tên nào khác mà mình thích). Bây giờ, chúng ta có thể xem qua kết quả phân tích bằng cách lệnh **summary** đối tượng **logistic** như sau:

```
summary(logistic)
```

```
Call:
glm(formula = proportion ~ ao, family = "binomial", weights = ntotal)

Deviance Residuals:
[1]  0  0
```



```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.2607      0.1888  -6.677 2.44e-11 ***
ao          0.8254      0.4306  1.917  0.0552 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 3.5022e+00 on 1 degrees of freedom
Residual deviance: -2.3093e-14 on 0 degrees of freedom
AIC: 12.933

Number of Fisher Scoring iterations: 3

```

Bảng 2. Kết quả phân tích hồi qui logistic bằng R.

Chú thích: Lệnh summary(logistic) cung cấp cho chúng ta các kết quả phân tích như trình bày trong Hình 1 trên.

- (a) Phần “**Call** :” báo cho chúng ta biết mô hình phân tích;
- (b) Deviance: phần thứ hai của kết quả cho biết qua về deviance, tức phần dư (hay residual trong mô hình [1]).

```

Deviance Residuals:
[1]  0  0

```

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 α (mà R đặt tên là **intercept**) và β (**ao**) và sai số chuẩn (standard error) cho từng ước số:

```

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.2607      0.1888  -6.677 2.44e-11 ***
ao          0.8254      0.4306   1.917  0.0552 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Qua kết quả này, chúng ta có $\hat{\alpha} = -1.2607$ và $\hat{\beta} = -0.8254$. Ước số $\hat{\beta}$ là số dương, cho thấy mối liên hệ giữa **cancer** và **ao** là mối liên hệ thuận: nguy cơ ung thư tăng khi giá trị của **ao** tăng. Tuy nhiên, kiểm định z (tính bằng cách lấy ước số chia cho sai số chuẩn) cho chúng ta thấy ảnh hưởng của **ao** không hẳn có ý nghĩa thống kê, vì trị số p = 0.055.

Xin nhắc lại, OR chính là $e^{0.8254} = 2.28$ (tức phương trình [5]) mà chúng ta vừa có được qua phân tích thủ công trong phần trên. Nói cách khác, khi **ao=1** thì nguy cơ ung thư tăng 2.28 lần so với nhóm **ao=0**.

(d) Các phần kế tiếp cung cấp một số chỉ số thống kê về mô hình, nhưng không có liên quan đến vấn đề chúng ta quan tâm, nên tôi sẽ không giải thích ở đây.

Như trình bày trên, không có khác biệt nào giữa kết quả phân tích bằng R và kết quả qua phân tích thủ công. Tuy nhiên, lợi thế khi phân tích bằng máy tính là thời gian. Sau khi nhập dữ liệu, tất cả các tính toán bằng R qua lệnh trên tốn không đầy 1 giây! Ngoài ra, R còn cung cấp cho chúng ta các sai số chuẩn thường rất khó tính trong trường hợp phân tích đa biến (mà tôi sẽ bàn qua trong một bài sau).

IV. Phân tích hồi qui logistic với một biến liên tục

Trong ví dụ 1, cả hai biến phụ thuộc (ung thư) và biến độc lập (phơi nhiễm AO) đều là biến nhị phân. Do đó, việc tính toán cũng đơn giản. Nhưng trong nhiều nghiên cứu, biến độc lập (hay yếu tố nguy cơ) là biến liên tục, và việc tìm hiểu mối tương quan giữa hai biến có phần phức tạp hơn. Trong phần này, tôi sẽ bàn qua một trường hợp như thế và sẽ sử dụng R để giải quyết vấn đề.

Ví dụ 2. Nghiên cứu mối tương quan giữa fibrinogen và EST. Erythrocyte sedimentation rate (ESR) là tỉ suất mà các hồng huyết cầu (erythrocytes) đọng lại trong huyết thanh. Bệnh nhân với ESR cao hơn 20 mm/giờ có nguy cơ cao bị bệnh thấp khớp, và các bệnh viêm mãn tính; và bệnh nhân với ESR thấp hơn 20 được xem là “bình thường”. Khi ESR tăng, một số protein trong máu cũng gia tăng. Một trong những protein đó là fibrinogen. Một nghiên cứu đo lường ESR và fibrinogen ở 29 đối tượng (Collett D, Jemain AA. Residuals, outliers and influential observations in regression analysis. Sains Malaysias 1985; 4:493-511), và các nhà nghiên cứu phát hiện trong nhóm này có 6 đối tượng với ESR cao hơn 20 mm/giờ. Các nhà nghiên cứu muốn biết có

mối tương quan nào giữa fibrinogen và ESR hay không. Số liệu của 29 đối tượng được trình bày trong **Bảng số 3** sau đây:

Bảng 3. Fibrinogen và ESR ở 29 đối tượng		
id	fibrinogen	ESR
1	2.52	0
2	2.56	0
3	2.19	0
4	2.18	0
5	3.41	0
6	2.46	0
7	3.22	0
8	2.21	0
9	3.15	0
10	2.60	0
11	2.29	0
12	2.35	0
16	3.15	0
18	2.68	0
19	2.60	0
20	2.23	0
21	2.88	0
22	2.65	0
24	2.28	0
25	2.67	0
26	2.29	0
27	2.15	0
28	2.54	0
30	3.34	0
31	2.99	0
32	3.32	0
13	5.06	1
14	3.34	1
15	2.38	1
17	3.53	1
23	2.09	1
29	3.93	1

Ghi chú: **id** là mã số của đối tượng nghiên cứu; **esr** được mã hóa 0 (nếu ESR thấp hơn 20) hay 1 (nếu ESR cao hơn 20).

Gọi p là xác suất $\mathbf{esr=1}$ và x là lượng protein fibrinogen trong máu, mô hình hồi qui logistic [1] có thể ứng dụng để trả lời câu hỏi trên:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x + \varepsilon \quad [6]$$

Chú ý rằng ở đây, x là một biến liên tục, chứ không phải biến nhị phân. Vì thế phương pháp ước tính thông số α và β cũng khác với ví dụ 1. Phương pháp chính để ước tính thông số trong mô hình [6] là phương pháp maximum likelihood – tức phương pháp *Hợp lí cực đại*, và không nằm trong phạm vi của bài viết 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:

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

Trong đó, Trong đó, y_i là biến phụ thuộc (**esr** với giá trị 0 hay 1), và x_i là biến độc lập (fibrinogen), và n là số mẫu. Để tìm ước số $\hat{\alpha}$ và $\hat{\beta}$ (ước số của α và β , 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.

Trước khi phân tích, chúng ta cần phải nhập số liệu vào R như sau (chúng ta không cần nhập biến **id**):

```
fibrinogen <- c(2.52, 2.56, 2.19, 2.18, 3.41, 2.46, 3.22, 2.21, 3.15,
               2.60, 2.29, 2.35, 3.15, 2.68, 2.60, 2.23, 2.88, 2.65,
               2.28, 2.67, 2.29, 2.15, 2.54, 3.34, 2.99, 3.32,
               5.06, 3.34, 2.38, 3.53, 2.09, 3.93)

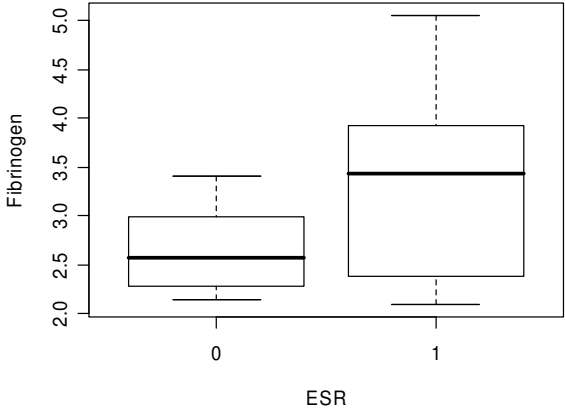
esr <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
         0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1)

data <- data.frame(fibrinogen, esr)
```

Chú ý lệnh thứ ba yêu cầu R nhập hai biến **fibrinogen** và **esr** vào một dữ liệu có tên là **data** để tiện cho việc phân tích sau này.

```
boxplot(fibrinogen ~ esr, xlab="ESR", ylab="Fibrinogen")
t.test(fibrinogen ~ esr)
```

Lệnh thứ nhất yêu cầu R vẽ biểu đồ hình hộp (box plot) về fibrinogen phân nhóm theo biến **esr**, và kết quả được trình bày trong biểu đồ 2 dưới đây. Lệnh thứ hai sử dụng kiểm định **t.test** trong R để xem sự khác biệt về fibrinogen giữa hai nhóm ESR có ý nghĩa thống kê hay không, và kết quả được trình bày trong Bảng 3 dưới đây:

	<pre>Welch Two Sample t-test data: fibrinogen by esr t = -1.6498, df = 5.331, p-value = 0.1563 alternative hypothesis: true difference in means is not equal to 0 95 percent confidence interval: -1.8666562 0.3907588 sample estimates: mean in group 0 mean in group 1 2.650385 3.388333</pre>
<p>Biểu đồ 2. Biểu đồ hình hộp độ phân phối của fibrinogen giữa hai nhóm ESR.</p>	<p>Bảng 4. Kiểm định t giữa hai nhóm cao và thấp ESR.</p>

Phân tích đơn giản trên đây cho thấy độ fibrinogen trung bình ở đối tượng có ESR cao (tức **esr = 1**) là 3.39 mm/giờ, có phần cao hơn so với nhóm ESR thấp với độ fibrinogen trung bình là 2.65 mm/giờ. Nhưng sự khác biệt này không có ý nghĩa thống kê ($p = 0.1563$).

Bây giờ chúng ta phân tích bằng phương pháp hồi qui logistic với hàm **glm** trong R như sau:

```
logit.esr <- glm(esr ~ fibrinogen, family="binomial")
summary(logit.esr)
```

Chú ý cách viết lệnh cũng không khác gì so với ví dụ 1, Kết quả của phân tích này được trình bày trong biểu đồ 3 sau đây:

Call:				
glm(formula = esr ~ fibrinogen, family = "binomial")				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-0.9298	-0.5399	-0.4382	-0.3356	2.4794
Coefficients:				
	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	-6.8451	2.7703	-2.471	0.0135 *
fibrinogen	1.8271	0.9009	2.028	0.0425 *

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 24.840  on 30  degrees of freedom
AIC: 28.840

Number of Fisher Scoring iterations: 5

```

Bảng 5. Kết quả phân tích sự tương quan giữa fibrinogen và ESR

Với kết quả trên, phương trình [6] có thể viết như sau:

$$\log\left(\frac{p}{1-p}\right) = -6.8451 + 1.8271x$$

Như vậy OR liên quan đến fibrinogen là: $OR = e^{1.827} = 6.21$ (như giải thích ở phương trình [5]). Nói cách khác, khi fibrinogen tăng 1 mmol/L, thì odd với **esr** cao tăng 6.21 lần. Chúng ta có thể tính khoảng tin cậy 95% của OR bằng lệnh sau đây:

```
exp(confint(logit.esr, parm="fibrinogen"))
```

Chú ý: lệnh trên yêu cầu tính số mũ (**exp**) của khoảng tin cậy 95% (**confint** – viết tắt từ confidence interval của thông số **fibrinogen** (**parm** – viết tắt của chữ parameter) trong đối tượng phân tích **logit.esr**. Kết quả là:

```

      2.5 %      97.5 %
1.403468 54.535954

```

Tức khoảng tin cậy 95% của OR liên quan đến fibrinogen dao động từ 1.40 đến 54.5. Bởi vì khoảng tin cậy 95% cao hơn 1, chúng ta có bằng chứng để phát biểu rằng mối liên hệ giữa fibrinogen và ESR có ý nghĩa thống kê. Thật ra, trị số p của mối liên hệ này là 0.0425 (xem **Bảng 5**).

V. Ảnh hưởng tương tác (interaction effect)

Hai ví dụ trên, tôi đã giới thiệu qua cách phân tích hồi qui logistic các nghiên cứu mà biến độc lập có thể là biến liên tục hay biến không liên tục, nhưng mô hình chỉ đơn

giản giới hạn một biến độc lập. Tuy nhiên, trong nhiều nghiên cứu khoa học, có rất nhiều biến độc lập mà nhà nghiên cứu muốn thăm định mối tương quan hay ảnh hưởng đến một biến phụ thuộc. Trong phần này, tôi sẽ bàn về một nghiên cứu với hai biến độc lập, và vấn đề tương tác giữa các biến độc lập.

Ví dụ 3. Nghiên cứu về vai trò của phụ nữ trong xã hội. Trong một điều tra xã hội thực hiện vào năm 1971-1972, các nhà nghiên cứu hỏi đối tượng – nam và nữ – đồng ý hay không đồng ý với câu hỏi sau đây: “*Phụ nữ nên lo việc nhà và để việc điều hành nhà nước cho đàn ông*” (Harberman SJ. The analysis of residuals in cross-classified tables. *Biometrics* 1973;29:205-220). Các nhà nghiên cứu ghi nhận trình độ học vấn và giới của mỗi đối tượng. Kết quả nghiên cứu có thể tóm lược bằng Bảng số liệu số 6 sau đây.

Bảng 6. Vai trò của phụ nữ trong xã hội

edu	sex	agree	disagree
0	Male	4	2
1	Male	2	0
2	Male	4	0
3	Male	6	3
4	Male	5	5
5	Male	13	7
6	Male	25	9
7	Male	27	15
8	Male	75	49
9	Male	29	29
10	Male	32	45
11	Male	36	59
12	Male	115	245
13	Male	31	70
14	Male	28	79
15	Male	9	23
16	Male	15	110
17	Male	3	29
18	Male	1	28
19	Male	2	13
20	Male	3	20
0	Female	4	2
1	Female	1	0
2	Female	0	0
3	Female	6	1
4	Female	10	0
5	Female	14	7
6	Female	17	5
7	Female	26	16
8	Female	91	36
9	Female	30	35
10	Female	55	67
11	Female	50	62
12	Female	190	403

13 Female	17	92
14 Female	18	81
15 Female	7	34
16 Female	13	115
17 Female	3	28
18 Female	0	21
19 Female	1	2
20 Female	2	4

Ghi chú: Trong bảng trên, biến **edu** là trình độ học vấn (đo bằng số năm theo học) của người trả lời, **agree** và **disagree** là số đối tượng đồng ý hay không đồng ý với câu hỏi. Chẳng hạn như trong dòng cuối của bảng số liệu có nghĩa là trong số phụ nữ với 20 năm học, 2 người đồng ý và 4 người không đồng ý với câu hỏi.

Các nhà nghiên cứu muốn ước lượng sự ảnh hưởng của giới tính và trình độ học vấn đến xu hướng trả lời câu hỏi trên.

Để tiện cho việc theo dõi, các số liệu trong bảng trên trước hết sẽ được nhập vào R. Các lệnh sau đây tạo ra 4 biến: **edu**, **sex**, **agree** và **disagree**. Ngoài ra, hai biến **ntotal** (tổng số đối tượng) và **proportion** (phần trăm đối tượng đồng ý với câu hỏi) cũng được tính toán từ hai biến **agree** và **disagree**. Các số liệu này sẽ được lưu trữ trong một dữ liệu có tên là **women**.

```
edu <- c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
        11, 12, 13, 14, 15, 16, 17, 18, 19, 20,
        0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
        11, 12, 13, 14, 15, 16, 17, 18, 19, 20)

sex <- c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
        1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)

agree <- c(4, 2, 4, 6, 5, 13, 25, 27, 75, 29, 32,
          36, 115, 31, 28, 9, 15, 3, 1, 2, 3, 4,
          1, 0, 6, 10, 14, 17, 26, 91, 30, 55, 50,
          190, 17, 18, 7, 13, 3, 0, 1, 2)

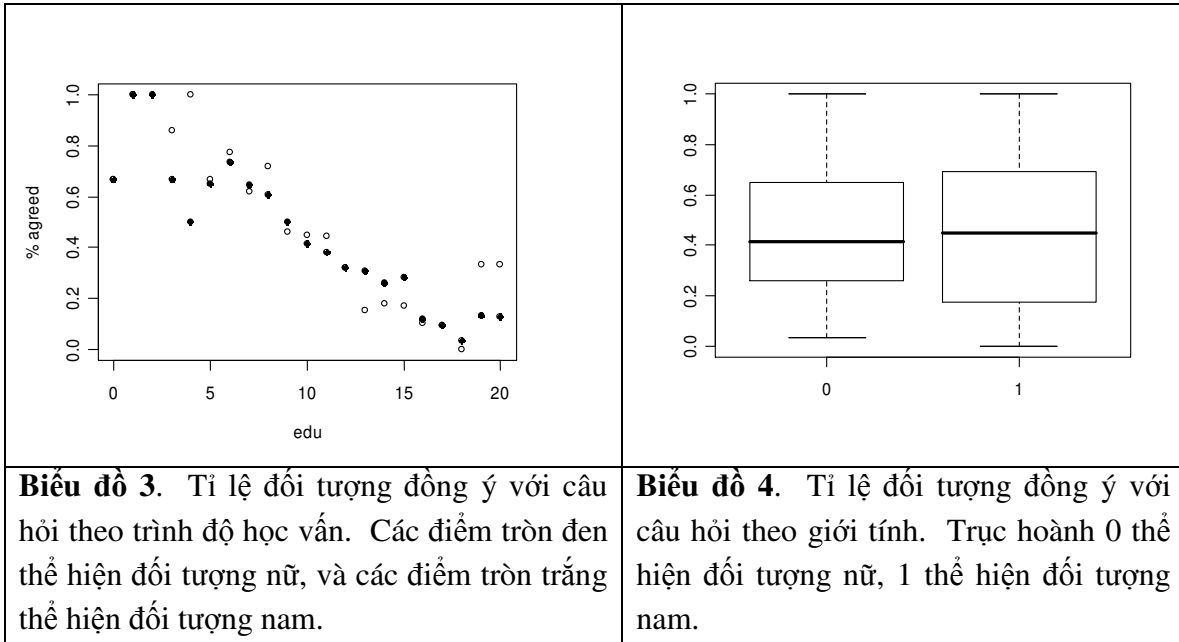
disagree <- c(2, 0, 0, 3, 5, 7, 9, 15, 49, 29, 45,
             59, 245, 70, 79, 23, 110, 29, 28, 13, 20, 2,
             0, 0, 1, 0, 7, 5, 16, 36, 35, 67, 62,
             403, 92, 81, 34, 115, 28, 21, 2, 4)

ntotal <- agree + disagree
proportion <- agree/ntotal
women <- data.frame(edu, sex, agree, disagree, ntotal, proportion)
```

Trước khi phân tích, chúng ta thử tìm hiểu tỉ lệ đồng ý (tức biến **proportion**) theo trình độ học vấn và giới tính, với hai lệnh sau đây: Lệnh thứ nhất thể hiện sự tương

quan giữa tỉ lệ đồng ý và trình độ học vấn, và kết quả trình bày trong **Biểu đồ 3**; lệnh thứ hai vẽ biểu đồ hình hộp về tỉ lệ đồng ý theo giới tính (**Biểu đồ 4**):

```
plot(proportion ~ edu, ylab="% agreed", pch=ifelse(sex==0,16,21))
boxplot(proportion ~ sex)
```



Biểu đồ 3 cho thấy rõ ràng có một mối tương quan nghịch đảo giữa tỉ lệ đồng ý và trình độ học vấn: đối tượng có trình độ văn hóa càng cao, tỉ lệ đồng ý càng thấp. Tuy nhiên, cả hai biểu đồ cho thấy ảnh hưởng của giới tính có vẻ không quan trọng, dù tỉ lệ nữ đồng ý có vẻ cao hơn so với nam giới.

Gọi p là xác suất đồng ý với câu hỏi, và với kết quả phân tích sơ bộ trên, chúng ta có thể xem xét một mô hình đơn giản mà theo đó tỉ xác suất đồng ý tùy thuộc vào trình độ học vấn và giới tính. Nói theo ngôn ngữ của mô hình hồi qui logistic:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \times \text{edu} + \gamma \times \text{sex} \quad [7]$$

Và, theo ngôn ngữ máy tính R (kết quả trình bày trong Bảng 7):

```
logistic <- glm(proportion ~ sex + edu, family="binomial",
                weight=ntotal)
summary(logistic)
```

```

Call:
glm(formula = proportion ~ sex + edu, family = "binomial", weights =
ntotal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.72544  -0.87168  -0.08448   0.88843   3.13315

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.50937    0.18389  13.646  <2e-16 ***
sex          -0.01145    0.08415  -0.136   0.892
edu          -0.27062    0.01541 -17.560  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 451.722  on 40  degrees of freedom
Residual deviance:  64.007  on 38  degrees of freedom
AIC: 208.07

Number of Fisher Scoring iterations: 4

```

Bảng 7. Kết quả phân tích hồi qui logistic của mô hình [7]

Kết quả trên cho thấy rõ ràng ảnh hưởng của trình độ học vấn đến xu hướng đồng ý với câu hỏi ($p < 0.0001$), nhưng giới tính không có ảnh hưởng đáng kể ($p = 0.892$).

Mô hình [7] còn có tên là *mô hình cộng hưởng (additive model hay main effect model)*, bởi vì mô hình này phát biểu rằng trình độ học vấn và giới tính ảnh hưởng *độc lập* đến tỉ lệ đồng ý. Cụm từ “*độc lập*” ở đây có nghĩa là ảnh hưởng của trình độ học vấn hoàn toàn không tùy thuộc vào ảnh hưởng của giới tính (và ngược lại, ảnh hưởng của giới tính – nếu có – hoàn toàn không phụ thuộc vào trình độ học vấn).

Trong thực tế, đó là một mô hình đơn giản, bởi vì thái độ và hành xử của nam và nữ có thể khác nhau dù họ có cùng một trình độ học vấn. Nếu điều đó xảy ra, thì mô hình cộng hưởng [7] không còn phù hợp trong thực tế nữa. Vì thế, trước khi chấp nhận mô hình cộng hưởng, chúng ta phải xem xét đến *mô hình tương tác (interaction model)* giữa giới tính và trình độ học vấn. Mô hình tương tác phát biểu rằng:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \times \text{edu} + \gamma \times \text{sex} + \zeta \times \text{edu} \times \text{sex} \quad [8]$$

Với R, mô hình trên được viết như sau:

```
interaction <- glm(proportion ~ sex + edu + sex:edu,
                  family="binomial", weight=ntotal)

summary(interaction)
```

```
Call:
glm(formula = proportion ~ sex * edu, family = "binomial", weights =
ntotal)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.39097  -0.94911   0.03065   0.75927   2.45262

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.09820    0.23550   8.910 < 2e-16 ***
sex           0.90474    0.36007   2.513  0.01198 *
edu          -0.23403    0.02019  -11.592 < 2e-16 ***
sex:edu      -0.08138    0.03109   -2.617  0.00886 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 451.722  on 40  degrees of freedom
Residual deviance:  57.103  on 37  degrees of freedom
AIC: 203.16

Number of Fisher Scoring iterations: 4
```

Bảng 8. Kết quả phân tích hồi qui logistic của mô hình tương tác [8]

Kết quả trên cho chúng ta một “bức tranh” hoàn toàn khác với mô hình cộng hưởng: tất cả ba thông số **sex**, **edu** và tương tác **sex:edu** (dấu “:” có nghĩa là tương tác trong R) đều có ý nghĩa thống kê. Để hiểu mô hình này, chúng ta cần phải viết lại mô hình [8] bằng các ước số trong Bảng 8:

$$\log\left(\frac{p}{1-p}\right) = 2.098 + 0.905 \times \text{sex} - 0.234 \times \text{edu} - 0.081 \times \text{edu} \times \text{sex}$$

Phương trình cho nữ (tức **sex = 0**) là:

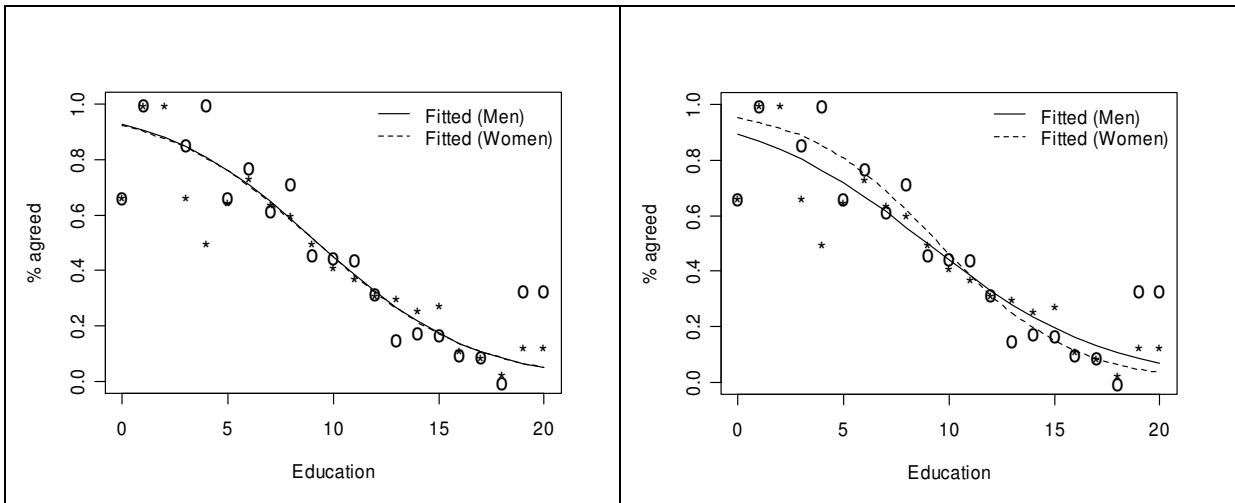
$$\log\left(\frac{P}{1-p}\right) = 2.098 - 0.234 \times edu$$

Phương trình cho nam (tức **sex = 1**) là:

$$\begin{aligned} \log\left(\frac{P}{1-p}\right) &= 2.098 + 0.905 - 0.234 \times edu - 0.081 \times edu \\ &= 3.003 - 0.315 \times edu \end{aligned}$$

Nói cách khác, chúng ta có 2 phương trình với cho hai giới tính. Ở nữ, mỗi năm tăng về học vấn, $OR = e^{-0.234} = 0.79$, nhưng ở nam, $OR = e^{-0.315} = 0.73$.

Một cách khác để cảm nhận sự khác biệt giữa hai nhóm là qua biểu đồ (xem mã để vẽ hai biểu đồ này trong phần **Chú thích**). Hai biểu đồ sau đây mô tả tỉ lệ đồng ý và trình độ học vấn cho nam và nữ dựa vào mô hình cộng hưởng [7] và mô hình tương tác [8]:



Biểu đồ 5. Tiên đoán tỉ lệ đối tượng đồng ý với câu hỏi theo giới tính dựa vào *mô hình cộng hưởng*: đường không đứt đoạn thể hiện nam, và đường đứt đoạn thể hiện nữ. Các điểm “o” thể hiện đối tượng nữ, và các điểm “*” thể hiện đối tượng nam.

Biểu đồ 6. Tiên đoán tỉ lệ đối tượng đồng ý với câu hỏi theo giới tính dựa vào *mô hình tương tác*: đường không đứt đoạn thể hiện nam, và đường đứt đoạn thể hiện nữ. Các điểm “o” thể hiện đối tượng nữ, và các điểm “*” thể hiện đối tượng nam.

Biểu đồ 5 cho thấy hai đường biểu diễn cho nam và nữ hầu như trùng nhau, nhưng **Biểu đồ 6** cho thấy xác suất đồng ý với câu hỏi khác nhau giữa nam và nữ và độ khác biệt còn tùy thuộc vào trình độ học vấn. Chẳng hạn như ở những đối tượng có trình độ học vấn thấp hơn 10 năm, nữ có xu hướng đồng ý cao hơn nam; nhưng ở những đối tượng có trình độ học vấn cao hơn 10 năm, nam có xu hướng đồng ý hơn nữ. Trong bối cảnh của câu hỏi, những đối tượng với trình độ học vấn thấp thường đồng ý với quan điểm rằng phụ nữ nên lo việc nhà và để việc “quốc gia đại sự” cho nam điều hành, nhưng với những đối tượng có trình độ học vấn cao, phần lớn đều không đồng ý với quan điểm này, và phản ứng của nữ khác với nam tùy vào trình độ học vấn. Đó chính là ý nghĩa của ảnh hưởng tương tác!

Qua ví dụ trên, chúng ta thấy nếu phân tích số liệu theo thói quen mà không xem xét đến khả năng ảnh hưởng tương tác, rất dễ đi đến kết luận sai hay bỏ qua những thông tin quan trọng. Xây dựng mô hình trong phân tích thống kê và khoa học nói chung là một vấn đề phức tạp, và tôi sẽ bàn đến trong phần sau.

Trong bài sau (hi vọng là có thì giờ) tôi sẽ bàn qua về mô hình hồi qui logistic đa biến, ảnh hưởng phi tuyến tính (non-linear effect) và các phương pháp cùng tiêu chuẩn để xây một mô hình logistic hoàn chỉnh.

Chú thích:

Thuật ngữ sử dụng trong bài

Tiếng Anh	Tiếng Việt
Logistic regression model	Mô hình hồi qui logistic
Control	Đối chứng
Variable	Biến
Continuous variable	Biến liên tục
Discrete variable	Biến không liên tục hay biến rời rạc
Dependent variable	Biến phụ thuộc
Independent variable	Biến độc lập
Maximum likelihood method	Phương pháp hợp lí cực đại
Additive model	Mô hình cộng hưởng
Interaction model	Mô hình tương tác

Mã R để vẽ biểu đồ 5 và 6

```
# tạo một hàm để vẽ, gọi hàm là myplot

myplot <- function(predicted)
{
  f <- data$sex == 1
  plot(data$edu, predicted, type="n",
        ylab="% agreed", xlab="Education", ylim=c(0,1))
  lines(data$edu[!f], predicted[!f], lty=1)
  lines(data$edu[f], predicted[f], lty=2)
  lgtxt <- c("Fitted (Men)", "Fitted (Women)")
  legend("topright", lgtxt, lty=1:2, bty="n")
  y <- data$agree/data$ntotal
  # text(data$edu, y, ifelse(f, "♂", "♀"), cex=1.25)
  text(data$edu, y, ifelse(f, "O", "*"), cex=1.25)
}

# vẽ biểu đồ 5 - mô hình cộng hưởng - additive model
additive <- glm(proportion ~ sex+edu,
                family="binomial", weight=ntotal, data=data)
p.additive <- predict(additive, type="response")
myplot(p.additive)

# vẽ biểu đồ 6 6 - mô hình tương tác - interactive model
```

```
interaction <- glm(proportion ~ sex+edu,  
                  family="binomial", weight=ntotal, data=data)  
p.predicted <- predict(interaction, type="response")  
myplot(p.predicted)
```