Lâm sàng thống kê

Chọn biến trong phân tích hồi qui logistic: một sai lầm phổ biến

Nguyễn Văn Tuấn

Hỏi*: "Trong một bài viết trước đây, Thầy viết rằng cách chọn biến cho một mô hình hồi qui logistic đa biến từ các phân tích đơn biến là sai lầm. Xin Thầy giải thích thêm tại sao?"*

Một nghiên cứu y học tiêu biểu thường đo lường nhiều yếu tố lâm sàng để nhằm tiên lượng một biến cố nào đó, chẳng hạn như tử vong, gãy xương, đái tháo đường, v.v... Lấy ví dụ một nghiên cứu về nguy cơ tử vong, nhà nghiên cứu có thể thu thập các thông tin như độ tuổi, chiều cao, cân nặng, tiền sử bệnh tật, lối sống, hay có thể đo lường các hormone, các chỉ số sinh hóa, v.v… (sẽ gọt tắt là "biến" hay variable) và câu hỏi đặt ra là trong những biến này, biến nào có liên quan đến tử vong. Đây là một vấn đề không đơn giản, và câu trả lời thường phải dựa vào kết quả phân tích thống kê và kiến thức sinh học. Một mô hình có thể tiên đoán rất chính xác, nhưng hoàn toàn vô dụng vì không có ý nghĩa lâm sàng hay sinh học; ngược lại, một mô hình có ý nghĩa lâm sàng nhưng không phù hợp với các giả định thống kê cũng chỉ là một trò chơi con số!

Một trong những khó khăn và có thể nói là vấn đề nan giải trong các nghiên cứu đa biến là các biến tiên lượng (predictor variables) thường có mối liên quan sinh học với nhau. Chẳng hạn như chiều cao và cân nặng có liên quan với nhau, hay các chỉ số sinh hóa biến chuyển theo từng độ tuổi. Và, những mối tương quan này làm cho vấn đề chọn mô hình thêm rắc rối, nhất là trong điều kiện nghiên cứu dựa vào một mẫu.

Vấn đề chọn mô hình

Để bạn đọc hiểu rõ vấn đề, tôi sẽ lấy một ví dụ đơn giản: một nghiên cứu lâm sàng nhằm mục đích phát triển một mô hình để tiên lượng nguy cơ tử vong (hay "khả năng sống sót" cho "tích cực" hơn) ở các bệnh nhân cấp cứu (ICU) dựa vào các chỉ số lâm sàng thu thập được từ lúc bệnh nhân nhập viện. Tiêu chí lâm sàng là tỉ lệ bệnh nhân sống sót sau 30 ngày xuất viện (và để tiết kiệm chữ nghĩa, gọi biến này là Y). Các biến thu thập lúc nhập viện gồm độ tuổi, cân nặng, và khoảng 8 chỉ số sinh hóa khác (gọi tắt là x_1, x_2, x_3, …., x_{10}). Để tiên lượng khả năng sống sót chúng ta có rất nhiều mô hình khả dĩ, chẳng hạn như:

$$Y = b_0 + b_1 \times x_1 + e$$

$$Y = b_0 + b_1 \times x_1 + b_2 \times x_2 + e$$
$$Y = b_0 + b_1 \times x_1 + b_2 \times x_2 + b_3 \times x_3 + e$$
$$Y = b_0 + b_1 \times x_1 + b_2 \times x_2 + b_6 \times x_6 + e$$
v.v…

trong đó, b_0, b_1, b_3, v.v… là những thông số liên quan đến từng biến cần ước tính, và e là phần ngẫu nhiên của mô hình. Thật ra, các mô hình trên đây còn đơn giản, vì chúng ta chưa xem xét đến các ảnh hưởng tương tác, ảnh hưởng phi tuyến tính, v.v… Có thể nói không ngoa rằng, với 10 biến số, con số mô hình khả dĩ có thể lên đến hàng trăm ngàn, thậm chí … bất tận. Nhưng trong những mô hình này, mô hình nào có thể tiên lượng chính xác nhất và đơn giản nhất?

Đây là một câu hỏi đã làm tốn biết bao công sức của nhiều nhà khoa học thống kê, nhà toán học và biết bao giấy mực để trả lời, nhưng cho đến nay vấn đề vẫn chưa ngã ngũ. Rất nhiều phương pháp đã được phát triển, nhưng chưa có một phương pháp nào hoàn chỉnh. Rất nhiều nhà thống kê học và toán học muốn giải quyết vấn đề, và đôi khi họ cũng phát triển một vài phương pháp, nhưng rất tiếc là các phương pháp này khi áp dụng vào môi trường y học thì rất vô nghĩa, vô duyên, và không thể sử dụng được. Tôi sẽ không bàn chi tiết tại sao vấn đề vẫn chưa ngã ngũ (tôi sẽ quay lại chủ đề này trong một bài viết khác), mà chỉ nhân cơ hội này để bàn về một sai lầm phổ biến trong việc đi tìm một mô hình tiên lượng.

Một sai lầm phổ biến

Đọc một bài báo khoa học trên một tập san y học trong nước trước đây, tôi các thấy tác giả viết: "*Các biến có liên quan với tử vong trong phân tích đơn biến với mức ý nghĩa p<0.05 sẽ được đưa vào phân tích hồi qui đa biến logistic*". Nói cách khác, các tác giả tiến hành phân tích hai giai đoạn:

- Giai đoạn 1, phân tích từng biến một và lưu ý các biến có ý nghĩa thống kê (tức $p < 0.05$);
- Giai đoạn 2, cho tất cả các biến có ý nghĩa thống kê trong giai đoạn 1 vào một mô hình đa biến.

Đây là một sai lầm rất "vô tư" và khá phổ biến trong y văn, không chỉ ở nước ta mà còn rất phổ biến ở các nước Tây phương. Thậm chí, theo kinh nghiệm của người viết bài này, các nhà thống kê chuyên nghiệp cũng sai! Sai lầm này không hẳn là do tác giả cố ý, nhưng do hiểu lầm (hay chưa thông hiểu) cơ chế của các mô hình thống kê.

Vấn đề chính của cách chọn mô hình theo hai giai đoạn trên là khi phân tích từng biến một (giai đoạn 1), mô hình hồi qui logistic không xem xét đến ảnh hưởng của các

biến khác cùng một lúc. Chẳng hạn như nếu biến x_1 và x_2 có tương quan với nhau, thì phân tích giai đoạn 1 có thể chọn cả hai biến, nhưng trong mô hình đa biến (giai đoạn 2), có thể chỉ có x_1 có ý nghĩa thống kê, còn x_2 thì không (hay ngược lại), bởi vì thông tin của biến này đã hàm chứa trong thông tin của biến kia (do hai biến có liên quan nhau).

Một vấn đế khác, tinh vi hơn và "tế nhị" hơn, là ảnh hưởng của một biến trung gian, rất khó hay không thể kiểm soát trong giai đoạn 1. (Tôi sẽ bàn qua về vấn đề ảnh hưởng của biến trung gian trong một bài khác). Trong trường hợp này, có thể hai biến có thể hai biến x_1 và x_5 (chẳng hạn) trong thực tế đều có ảnh hưởng đến Y, nhưng ảnh hưởng này chỉ tồn tại khi chúng xuất hiện bên nhau (cộng hưởng); do đó, khi phân tích riêng lẻ, chúng ta không phát hiện được ảnh hưởng của chúng, và do đó phân tích đơn giản trong giai đoạn 1 có thể bỏ qua cả hai biến!

Ví dụ 1: **Giới, thể dục, và tử vong**. Một nghiên cứu (mô phỏng) một thời điểm (cross-sectional study) nhằm đánh giá mối liên hệ của giới và nguy cơ tử vong vì bệnh nhồi máu cơ tim. Các nhà nghiên cứu còn thu thập thông tin liên quan đến thói quen tập thể dục và vận động cơ thể ở từng đối tượng. Kết quả nghiên cứu có thể tóm lược như sau:

Bảng 1. **Số đối tượng tử vong và còn sống chia theo giới và thói quen tập thể dục**

Biến	**Tử vong**	**Sống**	**Odds ratio và trị số P**
Giới			
• Nữ	113	2000	OR = 1.21
• Nam	94	2000	p = 0.176
Tập thể dục			
• Không	164	2000	OR = 4.06
• Có	43	2000	p = 0.0001

Trong nghiên cứu trên, nếu chúng ta áp dụng phương pháp phân tích hồi qui logistic cho từng biến riêng lẻ, chúng ta sẽ có:

- OR (odds ratio) cho nữ là 1.21 với trị số p = 0.176, tức không có ý nghĩa thống kê.

- OR cho nhóm không thường xuyên tập thể dục là 4.06 với p = 0.0001, tức có ý nghĩa thống kê.

Như vậy, nếu dựa vào phân tích này, chúng ta chỉ chọn biến tập thể dục vào mô hình đa biến. Nhưng kết quả này có thể sai. Quay lại với số liệu của nghiên cứu trên, chúng ta thử xác định tần số tử vong và sống sót theo cả hai biến *cùng một lúc* như sau:

Bảng 2. **Số đối tượng tử vong và còn sống chia theo thói quen tập thể dục cùng với giới**

Tập thể dục và giới	Tử vong	Sống	OR và trị số P
Không tập thể dục			
• Nữ	80	800	OR = 1.43
• Nam	84	1200	p = 0.028
Tập thể dục			
• Nữ	33	1200	OR = 2.20
• Nam	10	800	p = 0.026

Kết quả phân tích, như trình bày trong cột số 3 của bảng trên, rất khác với kết quả phân tích trong bảng 1. Ở đây, chúng ta thấy, giới có ảnh hưởng đến nguy cơ tử vong trong cả hai nhóm không tập thể dục và tập thể dục thường xuyên. Trong nhóm không tập thể dục thường xuyên, OR tử vong ở nữ là 1.43 với p = 0.028; trong nhóm tập thể dục thường xuyên, OR là 2.20 với p = 0.026.

Do đó, phương pháp phân tích đúng cho trường hợp này là chúng ta phải xem xét đến ảnh hưởng của hai biến cùng một lúc trong mô hình đa biến. Mô hình này có thể viết như sau:

$$Y = b_0 + b_1 \times x_1 + b_2 \times x_2 + e \quad [1]$$

Trong đó, Y là log của odd tử vong, x_1 là giới, x_2 là tập thể dục, và b_0, b_1, và b_2 là các thông số cần ước tính. Ước số của mô hình này có thể tóm lược như sau:

Biến	Hệ số của phương trình hồi qui logistic	OR và trị số P
Giới (Nữ)	$b_1 = 0.434$	OR = 1.54, p = 0.003
Tập thể dục (Không)	$b_2 = 1.425$	OR = 4.16, p < 0.0001

Kết quả phân tích đa biến trên cho chúng ta một "bức tranh" rất khác với phân tích đơn biến trong bảng 1. Đến đây, chúng ta có thể kết luận rằng ảnh hưởng của cả hai biến (giới và tập thể dục) đều có ý nghĩa thống kê, nhưng ảnh hưởng của tập thể dục có vẻ cao hơn ảnh hưởng của giới.

Một số nhà nghiên cứu cho rằng cách truy tầm biến có ý nghĩa thống kê cho phân tích đa biến có thể dựa vào kết quả của phân tích đơn biến bằng cách nâng trị số p lên 0.15 (thay vì 0.05). Nói cách khác, thay vì lưu giữ những biến có trị số $p < 0.05$ trong giai đoạn 1, có thể nâng cao tiêu chuẩn này thành $p < 0.15$ để lưu giữ những biến có thể bỏ sót vì tiêu chuẩn $p < 0.05$. Tuy nhiên, phương pháp này cũng … sai nốt! Để chứng minh cho sai lầm này, tôi sẽ lấy một ví dụ dưới đây.

Ví dụ 2: Vẫn với chủ đề của ví dụ 1, nhưng lần này, tôi thay đổi vài số liệu để chứng minh khiếm khuyết vừa nêu như sau:

Bảng 3. **Số đối tượng tử vong và còn sống chia theo thói quen tập thể dục cùng với giới**

Biến	Tử vong	Sống	Odds ratio và trị số P
Giới			
• Nữ	107	1935	OR = 1.18
• Nam	91	1935	p = 0.267
Tập thể dục			
• Không	107	1984	OR = 3.71
• Có	91	1886	p = 0.0001

Trong nghiên cứu trên, nếu phân tích từng biến riêng lẻ, một lần nữa, ảnh hưởng của yếu tố giới không có ý nghĩa thống kê ($p = 0.267$). Do đó, nếu dựa vào tiêu chuẩn $p < 0.15$, chúng ta phải loại bỏ yếu tố giới trong phân tích đa biến. Tuy nhiên, bảng số liệu dưới đây (Bảng 4) cho thấy nếu phân tích ảnh hưởng của giới trong từng nhóm tập thể dục, chúng ta thấy ảnh hưởng của giới có ý nghĩa thống kê.

Bảng 4. **Số lượng đối tượng tử vong và còn sống chia theo thói quen tập thể dục cùng với giới**

Tập thể dục và giới	Tử vong	Sống	OR và trị số P
Không tập thể dục			
• Nữ	75	774	OR = 1.39
• Nam	81	1161	p = 0.048
Tập thể dục			
• Nữ	32	1161	OR = 2.13
• Nam	10	774	p = 0.034

Bây giờ, chúng ta xem xét mô hình [1] (tức ước tính độ ảnh hưởng của hai biến số cùng một lúc trong một mô hình đa biến) cho số liệu trong Bảng 4, kết quả cho thấy cả hai biến đều có ý nghĩa thống kê:

Biến	Hệ số của phương trình hồi qui logistic	OR và trị số P
Giới (Nữ)	$b_1 = 0.4077$	OR = 1.50, p = 0.0064
Tập thể dục (Không)	$b_2 = 1.3938$	OR = 4.03, p < 0.0001

Tóm tắt

Xây dựng một mô hình hồi qui logistic đa biến là một vấn đề không đơn giản, nhất là trong trường hợp các biến tiên lượng có tương quan với nhau. Các ví dụ trên đây cho thấy phương pháp truy tầm biến có ý nghĩa thống kê trong mô hình đa biến dựa vào phân tích đơn biến có thể dẫn đến sai lầm quan trọng. Ngay cả nâng cao trị số p lên 0.15 cũng vẫn có thể phạm sai lầm.

Hiện nay, các phần mềm thống kê có sẵn một số thuật toán (algorithm) để truy tầm biến độc lập cho mô hình đa biến, như thuật toán stepwise, backward, và forward. Nhưng ngay cả các thuật toán này, nhất là thuật toán stepwise và forward, vẫn có nhiều khiếm khuyết và cho ra những kết quả "dương tính giả", tức là những biến chẳng có liên quan gì đến biến phụ thuộc. Rất nhiều người không hiểu các thuật toán này nên vẫn áp dụng chúng một cách vô tội vạ và hệ quả là có rất nhiều nghiên cứu với những kết quả sai trong y văn.

Xây dựng một mô hình đa biến là một khoa học, nhưng cũng là một nghệ thuật. Khoa học tính liên quan đến các tiêu chuẩn định lượng và thuật toán thích hợp. Nghệ thuật tính liên quan đến những yếu tố có thể nói là chủ quan, đòi hỏi nhà nghiên cứu phải vận dụng kiến thức chuyên ngành để đi đến một mô hình có ý nghĩa lâm sàng. Mộ mô hình đa biến nếu chỉ thỏa mản các tiêu chuẩn khoa học vẫn chưa thể là một mô hình có ích. Một mô hình có ý nghĩa lâm sàng nhưng không đáp ứng các tiêu chuẩn khoa học không thể là một mô hình có độ tin cậy cao. Do đó, phân tích đa biến, dù là mô hình logistic hay hồi qui tuyến tính, là một phương pháp phức tạp, đòi hỏi nhiều thời gian để suy nghĩ và tính toán. Không thể và không nên để cho máy tính suy nghĩ dùm cho chúng ta.

Chú thích kĩ thuật:

Phần dưới đây là các mã R sử dụng cho các ước tính trình bày trong bài viết.

```
# Phân tích số liệu ví dụ 1
# phân tích ảnh hưởng của giới
sex <- c("Nu", "Nam")
ntotal <- c(2000, 2000)
ndeaths <- c(113, 94)
pdeath <- ndeaths/ntotal
logistic <- glm(pdeath ~ sex, binomial, weight=ntotal)
logistic.display(logistic)

# phân tích ảnh hưởng của tập thể dục
pa <- c("No", "Yes")
ntotal <- c(2000, 2000)
ndeaths <- c(164, 43)
pdeath <- ndeaths/ntotal
logistic <- glm(pdeath ~ pa, binomial, weight=ntotal)
logistic.display(logistic)

# phân tích ảnh hưởng của tập thể dục và giới cùng một lúc, mô hình [1]
pa <- c("No", “No”, "Yes", “Yes”)
sex <- c(“Nu”, “Nam”, “Nu”, “Nam”)
ntotal <- c(880, 1284, 1233, 810)
ndeaths <- c(80, 84, 33, 10)
pdeath <- ndeaths/ntotal
logistic <- glm(pdeath ~ pa + sex, binomial, weight=ntotal)
summary(logistic)

# Ví dụ 2: phân tích đa biến – số liệu bảng 4
pa <- c("No", "No", "Yes", "Yes")
sex <- c("Nu", "Nam", "Nu", "Nam")
ntotal <- c(849, 1242, 1193, 784)
ndeaths <- c(75, 81, 32, 10)
pdeath <- ndeaths/ntotal
logistic <- glm(pdeath ~ pa + sex, binomial, weight=ntotal)
summary(logistic)
```