MÔ HÌNH ẢNH HƯỞNG HỖN HỢP

Khương Quỳnh Long - Hà Nội, 2021

Trong phần này, chúng tôi giới thiệu mô hình ảnh hưởng hỗn hợp (mixed effects model), một phương pháp thống kê áp dụng cho những dữ liệu dạng đo lường lặp lại. Dữ liệu đo lường lặp lại thường gặp trong nghiên cứu theo dõi dọc và nghiên cứu thử nghiệm lâm sàng khi mỗi biến đầu ra được đo lường lặp lại trên cùng một cá nhân. Nội dung phần này gồm 3 mục chính. Mục đầu tiên nhắc lại về lý thuyết hồi quy tuyến tính và lý do cần áp dụng mô hình ảnh hưởng hỗn hợp. Mục thứ hai nêu lên lý thuyết và các định nghĩa cơ bản của mô hình ảnh hưởng hỗn hợp, và mục cuối cùng là ví dụ trên dữ liệu thực tế.

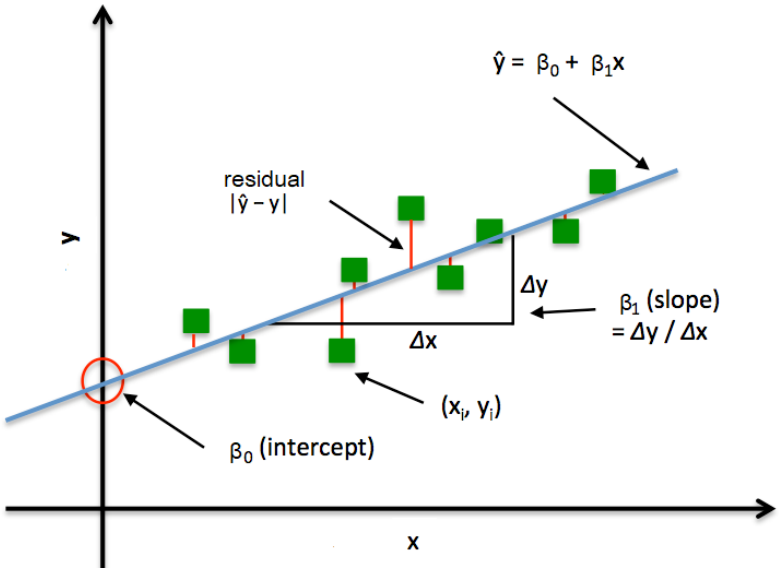
❖ Nhắc lại về hồi quy tuyến tính

- Mô hình hồi quy tuyến tính đơn giản

Hồi quy tuyến tính cho phép mô hình hóa mối quan hệ tuyến tính giữa biến phụ thuộc với một hay nhiều biến độc lập, nó cung cấp nền tảng cho những suy luận thống kê phức tạp hơn như mô hình ảnh hưởng hỗn hợp được đề cập trong chương này. Do đó, để thuận lợi hơn trong việc diễn giải mô hình say này, chúng tôi tóm tắt một số đặc điểm cơ bản của hồi quy tuyến tính sử dụng phương pháp bình phương nhỏ nhất (Ordinary Least Squares – OLS) cùng với các giả định ẩn phía sau mô hình và lý do cần đến mô hình ảnh hưởng hỗn hợp. Các suy luận thống kê có thể mở rộng ra cho các mô hình hồi quy khác, như các mô hình trong nhóm hồi quy tuyến tính tổng quát (Generalize linear model).

Một mô hình hồi quy tuyến tính đơn giản có thể được biểu thị bằng phương trình:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad [1]$$



Trong đó y_i là biến phụ thuộc của người thứ *i* trong bộ dữ liệu và x_i là biến độc lập của người thứ *i* (*i* = 1,…, N). Ký hiệu β_0 và β_1 lần lượt là điểm chặn (intercept) và độ dốc (slope) của mô hình. Trên biểu đồ, điểm chặn là điểm mà phương trình cắt tại trục y khi *x* = 0, nó còn mang ý nghĩa là trung bình của các giá trị *y* mà có *x* bằng 0. Độ dốc β_1 biểu thị mối liên hệ giữa *y* và *x*. Độ dốc nhận giá trị dương biểu thị mối liên hệ thuận chiều, tức là giá trị *x* càng tăng thì *y* càng tăng, và ngược lại độ dốc có giá trị âm chỉ mối liên hệ nghịch chiều, *x* càng tăng thì *y* càng giảm. Cuối cùng, e_i biểu thị cho sai số ngẫu nhiên (random error) hay phần dư (residual) của mô hình. Phần dư hàm ý rằng với mỗi cá nhân *i*, mô hình không thể tiên đoán chính xác hoàn toàn giá trị y_i, hay e đại diện cho tất cả các biến số không quan sát được (unobserved) mà có ảnh hưởng tới biến phụ thuộc *y*.

Các thông số của mô hình hồi quy (β_0 và β_1) được ước tính từ mẫu nghiên cứu (bao gồm *x* và *y*). Phương pháp phổ biến nhất để tính được các thông số này là phương pháp bình phương nhỏ nhất (hay OLS). Mục tiêu của phương pháp OLS là tối thiểu tổng bình phương sự khác biệt giữa giá trị y_i thực tế và giá trị $\hat{y}_i$ ước đoán từ mô hình ($\sum(y_i - \hat{y}_i)^2$) hay chính là tổng bình phương phần dư (sum of squared residuals - $\sum e_i^2$).

- **Giả định của hồi quy tuyến tính**

Mô hình hồi quy tuyến tính dựa trên một số giả định về phân phối của phần dư:

$$e \sim N(0, \sigma^2)$$

Giả định thứ nhất, phần dư có phân phối chuẩn với trung bình bằng 0. Phân phối của phần dư quanh giá trị trung bình 0 cũng chính là phân phối của biến phụ thuộc *y* quanh đường thẳng hồi quy, do đó giả định này hàm ý giá trị của biến phụ thuộc *y* phân bố đều hai bên đường thẳng hồi quy. Giả định thứ hai là phần dư có phương sai bất biến σ^2 với bất kì giá trị nào của *x*. Giả định thứ ba, phần dư của từng cá nhân *i* là độc lập với nhau. Giả định này có hàm ý phần dư (tức là phần nhân tố có ảnh hưởng tới biến phụ thuộc *y* nhưng không quan sát được) là độc lập giữa các cá nhân, tức là phần dư là ngẫu nhiên và không mô hình hóa được.

- **Vấn đề của hồi quy tuyến tính đối với dữ liệu nhiều bậc và đo lường lặp lại**

Dữ liệu nhiều bậc thường thấy nhiều ở thiết kế nghiên cứu cắt ngang, khi các cá nhân được "gộp" vào một cấu trúc lớn hơn, ví dụ như các học sinh được gộp vào trong lớp học, những lớp học trong trường học, hay những cộng đồng dân cư phân ra nhiều cấp xã, huyện, tỉnh. Trong nghiên cứu lâm sàng chẳng hạn như bệnh nhân được nhóm theo bác sĩ điều trị. Về mặt thực tế, những cá thể trong cùng một nhóm thường có xu hướng tương đồng nhau về một số đặc điểm hơn những cá thể trong nhóm khác, như cư dân trong cùng một huyện thường có một số đặc điểm chung về văn hóa, các bệnh nhân được điều trị chung một bác sĩ thường có xu hướng điều trị tương tự nhau…

Dữ liệu đo lường lặp lại thường thấy trong các nghiên cứu theo dõi dọc hay những nghiên cứu thử nghiệm lâm sàng khi các cá nhân được đo lường lặp lại vào nhiều thời điểm khác nhau. Ví dụ nghiên cứu thử nghiệm hiệu quả một loại thuốc X trong điều trị cao huyết áp, các bệnh nhân được khảo sát tại thời điểm t0, sau đó được phân nhóm điều trị (thuốc X/giả dược) và đo lặp lại huyết áp tại các thời điểm t2, t3, t4.

Trong cả hai loại cấu trúc dữ liệu trên các điểm dữ liệu đều có một sự tương quan nhất định với nhau, như các cá thể trong một cụm hay huyết áp qua các lần đo ở cùng một bệnh nhân, do đó vi phạm giả định của hồi quy tuyến tính (giả định phần dư của từng cá thể độc lập với nhau). Những trường hợp này sử dụng hồi quy tuyến tính sẽ cho ra kết quả thiếu chính xác. Mô hình ảnh hưởng hỗn hợp giới thiệu sau đây được sử dụng để giải quyết vấn đề này.

❖ **Mô hình ảnh hưởng hỗn hợp trong phân tích đo lường lặp lại**

Nội dung mục này giới thiệu về mô hình ảnh hưởng hỗn hợp ứng dụng trong phân tích dữ liệu đo lường lặp lại, đây là thiết kế nghiên cứu mà biến đầu ra (outcome) được đo lại nhiều lần trên cùng một đối tượng. Dữ liệu dạng đo lường lặp lại rất phổ biến trong nghiên cứu y khoa, như trong thử nghiệm lâm sàng (đánh giá hiệu quả của thuốc, hiệu quả của một phương pháp điều trị…), phục hồi chức năng, trong lĩnh vực tâm lý học và các ngành liên quan (đánh giá sự thay đổi nhận thức-hành vi theo thời gian)… Ở dạng dữ liệu này, kết quả các lần đo lường ở mỗi cá nhân thường có mối liên hệ với nhau do đó những mô hình hồi quy thông thường (vốn giả định các lần đo là độc lập) không thể tính toán chính xác. Mô hình ảnh hưởng hỗn hợp không những có thể mô hình hóa sự thay đổi về kết quả các lần đo trên từng bệnh nhân (tính toán tới sự tương quan giữa các lần đo trên cùng một bệnh nhân) mà còn có thể tính toán những thông số cho sự thay đổi của từng bệnh nhân dựa trên các biến tiên lượng (như đặc điểm điều trị, tuổi, giới…).

Mô hình ảnh hưởng hỗn hợp còn có nhiều tên gọi khác như mô hình đa bậc (multi-level models), mô hình thứ bậc (hierarchical models), mô hình ảnh hưởng ngẫu nhiên (random effect model), mô hình hệ số ngẫu nhiên (random coeffcients models)… Trong bài này chúng tôi sử dụng thuật ngữ mô hình ảnh hưởng hỗn hợp (mixed effects models) vì đây là thuật ngữ bao quát hơn cả, thuật ngữ này giúp truyền đạt một cách rõ ràng ý tưởng về hai loại ảnh hưởng trong mô hình, ảnh hưởng cố định (fixed effect) và ảnh hưởng ngẫu nhiên (random effect). Ảnh hưởng ngẫu nhiên cho phép mô hình hóa sự dao động (khác biệt) về kết quả giữa các cá nhân, trong khi ảnh hưởng cố định hàm ý rằng ảnh hưởng là như nhau đối với các cá nhân có cùng đặc điểm. Một ví dụ đơn giản, thử nghiệm lâm sàng đánh giá hiệu quả của thuốc X, ảnh hưởng cố định ở đây là phân nhóm điều trị (giả dược hay thuốc X), tức là ảnh hưởng của thuốc trên những bệnh nhân điều trị giả dược là như nhau và ảnh hưởng lên các bệnh nhân trong nhóm điều trị thuốc X là như nhau. Trong khi đó phần ảnh hưởng ngẫu nhiên có thể hiểu là cơ địa của bệnh nhân, do đó ảnh hưởng của thuốc lên các bệnh nhân này là dao động và không giống nhau.

- **Mô hình random intercept**

Mô hình hồi quy đơn giản mà chúng tôi giới thiệu ở phần trước (công thức [1]) chỉ bao gồm cố định (fixed effect) hay còn gọi là mô hình đơn bậc (single-level). Xin nhắc

lại, điểm chặn (intercept) chính là trung bình (có điều kiện) của *y* tại giá trị *x* = 0, trong mô hình này chỉ có 1 điểm chặn β_0 cố định cho mọi giá trị y_i, hay có nghĩa điểm chặn là như nhau cho toàn bộ các cá nhân. Tuy nhiên trong nhiều trường hợp các cá nhân có thể có đặc điểm khác nhau tại thời điểm ban đầu (baseline) dẫn tới giá trị trung bình của biến số quan tâm (giá trị trung bình của *y*) tại *x* = 0 là khác nhau. Do đó ngoài điểm chặn cố định (fixed intercept) cần phải có thêm một thông số để mô hình hóa những sự khác biệt này, và phần để diễn tả sự khác biệt về điểm chặn giữa các cá nhân được gọi là phần chặn ngẫu nhiên (random intercept). Khi đó công thức [1] được cộng thêm phần random intercept (kí hiệu là μ_0) và trở thành:

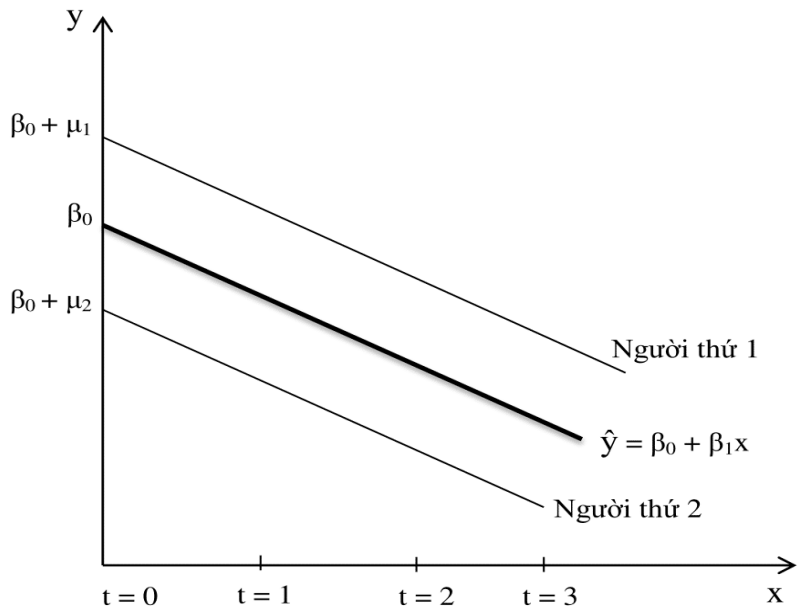
$$y_{ii} = (\beta_0 + \mu_{0j}) + \beta_1 x_{ij} + e_{ij} \quad [2]$$

Trong đó *i* chính là lần đo thứ *i* của người thứ *j*. Phương trình [2] cũng có thể viết lại theo phần cố định và phần ngẫu nhiên

$$y_{ii} = (\beta_0 + \beta_1 x_{ij}) + (\mu_{0j} + e_{ij}) \quad [3]$$

$$\text{Với } \mu_{0j} \sim N(0, \sigma_{\mu 0}^2) \text{ và } e_{ij} \sim N(0, \sigma_e^2)$$

Hình dưới đây minh họa rõ hơn về mô hình random intercept.



Giả sử một nghiên cứu tiến hành nhằm đánh giá hiệu quả giảm đường huyết của thuốc X, các bệnh nhân được đo đường huyết ban đầu tại thời điểm t = 0. Sau đó các

bệnh nhân tiếp tục được sử dụng thuốc và đánh giá lại vào các thời điểm 3 tháng, 6 tháng và 9 tháng. Giả định mối liên hệ giữa giảm đường huyết theo thời gian là mối liên hệ tuyến tính và tốc độ giảm đường huyết là như nhau (fixed slope – nội dung này sẽ được đề cập trong mục tiếp theo) thì đường huyết tại các thời điểm đo lường (tháng thứ 3, 6 và 9) sẽ khác nhau giữa các cá nhân phụ thuộc vào phần random intercept (ví dụ μ_1 và μ_2 ở hình minh họa trên). Một cách đơn giản hơn, random intercept trong đo lường lặp lại hàm ý rằng, với mỗi cá nhân có điểm bắt đầu khác nhau thì những đo lường tiếp theo cũng khác nhau phụ thuộc vào điểm ban đầu đó.

- **Mô hình random slope**

Phần độ dốc (slope) hay hệ số phương trình là thông số thể hiện mối liên hệ giữa biến phụ thuộc *y* và biến độc lập *x*. Khi giữ các yếu tố khác không thay đổi, độ dốc càng lớn (âm hoặc dương) càng thể hiện mối liên hệ mạnh giữa *y* và *x*. Trong mô hình hồi quy đơn bậc, phần độ dốc là cố định (fixed slope) nghĩa là mối liên hệ giữa biến phụ thuộc và biến độc lập là như nhau cho mọi cá nhân. Tuy nhiên, trong dữ liệu đo lường lặp lại, mỗi cá nhân được đo lường vào những thời điểm khác nhau hoặc đo lường sau mỗi liều lượng khác nhau, trong khi đó "cơ địa" mỗi cá nhân là khác nhau dẫn tới đáp ứng của mỗi cá nhân cũng khác nhau. Để mô hình hóa sự khác nhau giữa các cá nhân này, phần random slope được thêm vào mô hình. Như vậy mối liên hệ giữa *y* và *x* (hay hệ số phương trình) không chỉ có phần cố định (giống nhau giữa các cá nhân) mà còn có phần ngẫu nhiên (khác nhau giữa các cá nhân). Phần random slope được kí hiệu là μ_{1j}, khi đó phương trình [2] trở thành mô hình bao gồm cả radom intercept và random slope.

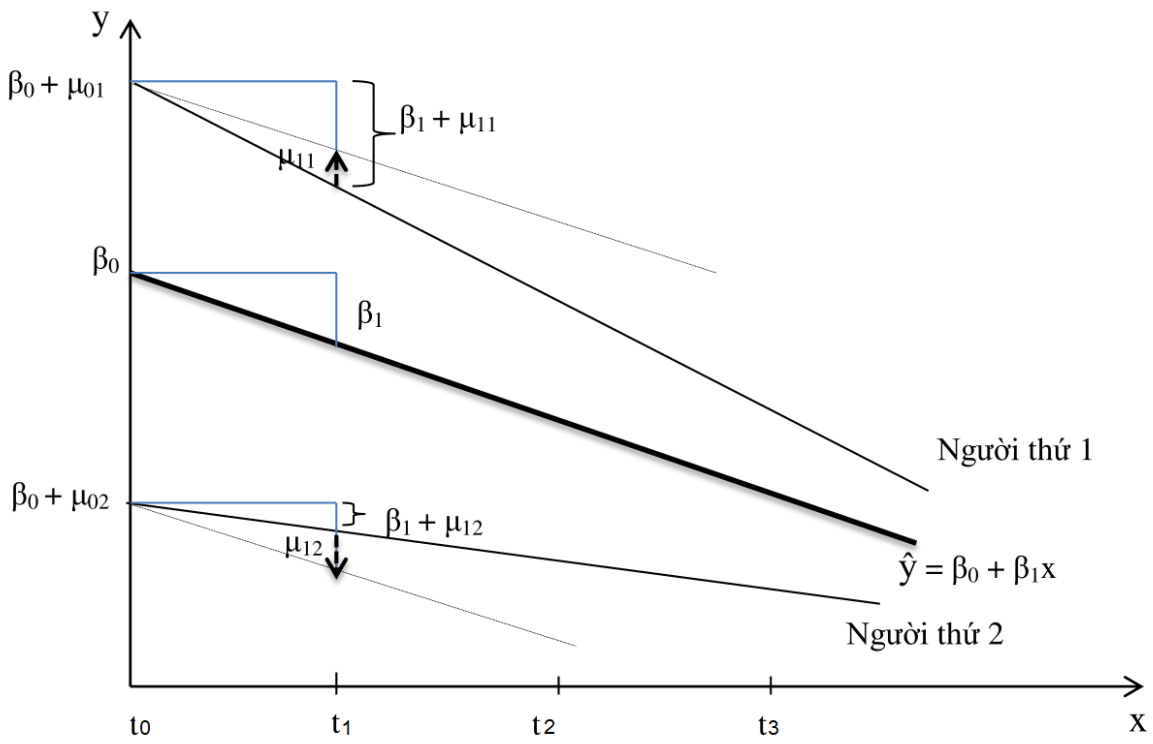
$$y_{ii} = (\beta_0 + \mu_{0j}) + (\beta_1 + \mu_{1j})\, x_{ij} + e_{ij} \quad [4]$$

Trong đó *i* chính là lần đo thứ *i* của người thứ *j*. Phương trình [4] cũng có thể viết lại theo phần cố định và phần ngẫu nhiên

$$y_{ii} = (\beta_0 + \beta_1\, x_{ij}) + (\mu_{1j}\, x_{ij} + \mu_{0j} + e_{ij}) \quad [5]$$

Với: $\mu_{0j} \sim N(0, \sigma_{\mu 0}^2)$, $\mu_{1j} \sim N(0, \sigma_{\mu 1}^2)$ và $e_{ij} \sim N(0, \sigma_e^2)$

Hình dưới đây minh họa rõ hơn về mô hình random slope.



Trở lại ví dụ thuốc điều trị đái tháo đường, mức độ giảm đường huyết chính là phần độ dốc. Trong ví dụ phần random intercept, giả định độ dốc là như nhau hay mức độ giảm đường huyết là như nhau giữa các cá nhân. Trong mô hình có thêm phần random slope thì mức độ giảm đường huyết giữa các cá nhân là khác nhau, như minh họa ở hình trên, người thứ nhất có mức độ giảm đường huyết nhanh hơn người thứ hai. Mô hình random slope còn quan tâm đến mối liên hệ giữa hai phần randome intercept và random slope, còn gọi là hiệp biến (covariance), hiệp biến có thể nhận ba loại giá trị: âm, dương và bằng 0. Trường hợp hiệp biến dương khi cá nhân có điểm chặn lớn và có độ dốc cũng lớn, hay trong trường hợp trên thì người có đường huyết cao tại thời điểm baseline sẽ có tốc độ giảm đường huyết nhanh hơn. Ngược lại, hiệp biến âm khi cá nhân có điểm chặn lớn và có độ dốc nhỏ, hay người có đường huyết cao tại thời điểm baseline sẽ có tốc độ giảm đường huyết chậm, và trường hợp cuối cùng là hiệp biến bằng 0, tức là mức độ giảm đường huyết không có mối liên hệ nào với mức độ đường huyết ban đầu.

- **Các bước xây dựng mô hình trong thực hành**

Việc xây dựng mô hình ảnh hưởng hỗn hợp trong thực hành bao gồm những bước chính sau

- Mô tả, thăm dò số liệu, vẽ biểu đồ
- Kiểm tra và thêm random intercept
- Kiểm tra và thêm random slope
- Kiểm tra và thêm các ảnh hưởng khác vào mô hình như biến gây nhiễu, ảnh hưởng tương tác…

Ví dụ ở mục sau đây sẽ giúp giải thích rõ hơn về quy trình từng bước của việc xây dựng mô hình ảnh hưởng hỗn hợp đối với dữ liệu đo lường lặp lại.

❖ **Ví dụ minh họa**

Phần ví dụ minh họa này sử dụng bộ dữ liệu "*Isoproterenol.dta*" được cung cấp bởi William D.Dupont (1). Bộ dữ liệu là một phần từ nghiên cứu của Lang và cộng sự năm 1995 (2). Nghiên cứu được mô tả:

Nghiên cứu đánh giá tác động của isoproterenol, một chất chủ vận β_2-adrenergic, đến lưu lượng máu cẳng tay trên 22 người đàn ông khỏe mạnh. Trong đó có 9 người da đen và 13 người da trắng. Lưu lượng máu cẳng tay (ml/min/dl) của mỗi người được đo lường tại thời điểm ban đầu và tại những thời điểm tăng liều isoproterenol (các liều 10, 20, 60, 150, 300 và 400 (ng/min)) (2).

Câu hỏi đặt ra là:

- Liều isoproterenol có ảnh hưởng tới lưu lượng máu ở cẳng tay hay không?
- Mức độ đáp ứng với isoproterenol có khác nhau giữa hai nhóm màu da (da đen và da trắng) hay không?

- **Mô tả dữ liệu**

```
list, nolabel
```

	id	race	fbf0	fbf10	fbf20	fbf60	fbf150	fbf300	fbf400
1.	1	0	1	1.4	6.4	19.1	25	24.6	28
2.	2	0	2.1	2.8	8.3	15.7	21.9	21.7	30.1
3.	3	0	1.1	2.2	5.7	8.2	9.3	12.5	21.6
4.	4	0	2.44	2.9	4.6	13.2	17.3	17.6	19.4
5.	5	0	2.9	3.5	5.7	11.5	14.9	19.7	19.3
6.	6	0	4.1	3.7	5.8	19.8	17.7	20.8	30.3
7.	7	0	1.24	1.2	3.3	5.3	5.4	10.1	10.6
8.	8	0	3.1	.	.	15.45	.	.	31.3
9.	9	0	5.8	8.8	13.2	33.3	38.5	39.8	43.3
10.	10	0	3.9	6.6	9.5	20.2	21.5	30.1	29.6
11.	11	0	1.91	1.7	6.3	9.9	12.6	12.7	15.4
12.	12	0	2	2.3	4	8.4	8.3	12.8	16.7
13.	13	0	3.7	3.9	4.7	10.5	14.6	20	21.7
14.	14	1	2.46	2.7	2.54	3.95	4.16	5.1	4.16
15.	15	1	2	1.8	4.22	5.76	7.08	10.92	7.08
16.	16	1	2.26	3	2.99	4.07	3.74	4.58	3.74
17.	17	1	1.8	2.9	3.41	4.84	7.05	7.48	7.05
18.	18	1	3.13	4	5.33	7.31	8.81	11.09	8.81
19.	19	1	1.36	2.7	3.05	4	4.1	6.95	4.1
20.	20	1	2.82	2.6	2.63	10.03	9.6	12.65	9.6
21.	21	1	1.7	1.6	1.73	2.96	4.17	6.04	4.17
22.	22	1	2.1	1.9	3	4.8	7.4	16.7	21.2

Dữ liệu đang ở dạng rộng/ngang (wide form) bao gồm 9 cột và 22 hàng

id	Mã số người tham gia nghiên cứu, gồm 22 người
race	Màu da (0 = da trắng, 1 = da đen)
fbf0	Lưu lượng máu cẳng tay (ml/min/dl) tại liều isoproterenol = 0 ng/min
fbf10	Lưu lượng máu cẳng tay (ml/min/dl) tại liều isoproterenol = 10 ng/min
…..	
fbf400	Lưu lượng máu cẳng tay (ml/min/dl) tại liều isoproterenol = 400 ng/min

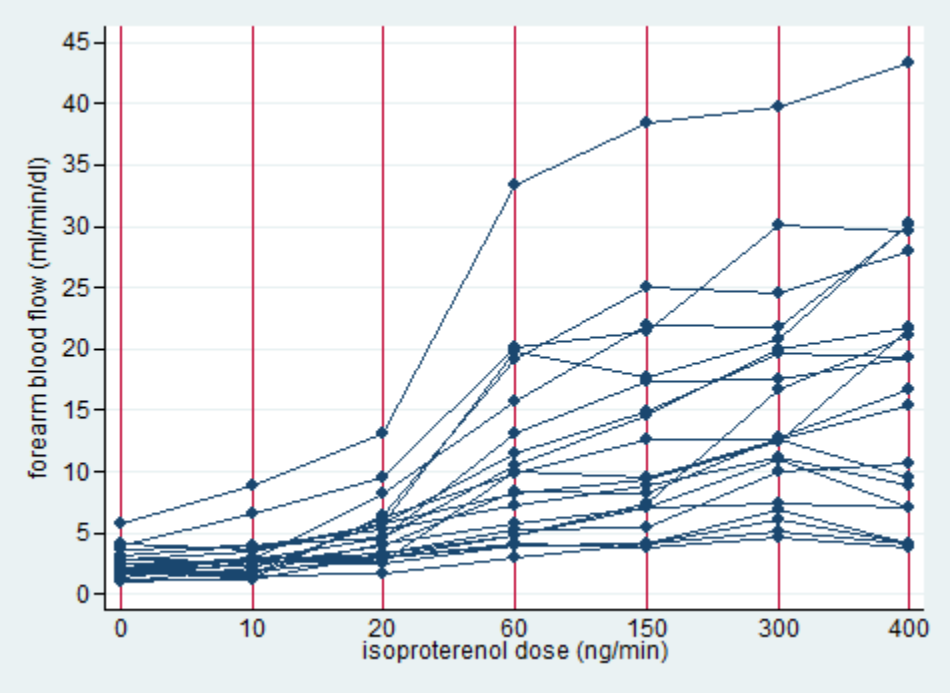
Để mô tả rõ hơn sự thay đổi của lưu lượng máu cẳng tay qua các liều isoproterenol, chúng ta sử dụng biểu đồ Parplot. Trước tiên cần cài đặt gói biểu đồ này (nếu sử dụng lần đầu tiên)

```
findit parplot
```

→ Chọn đường dẫn: parplot from http://fmwww.bc.edu/RePEc/bocode/p → (click here to install)

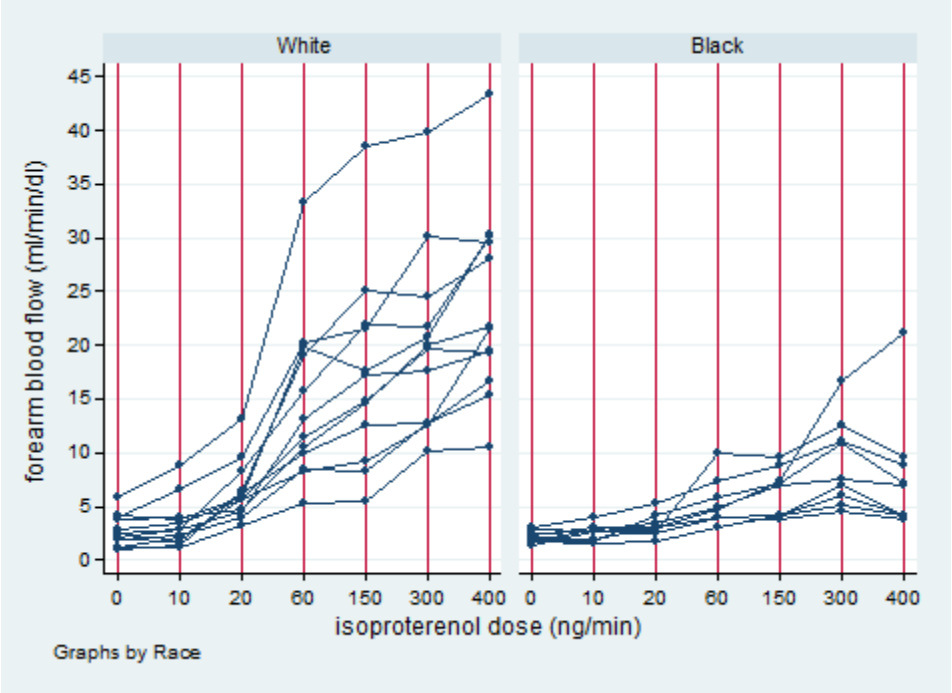
Vẽ biểu đồ: (lưu ý câu lệnh phải được lưu trong do-file để sử dụng)

```
#delimit ;
parplot fbf0-fbf400
, transform(raw)
xlabel(1 "0" 2 "10" 3 "20" 4 "60" 5 "150" 6 "300"
7 "400")
ylabel(0(5)45, angle(horizontal))
ytitle("forearm blood flow (ml/min/dl)")
xtitle("isoproterenol dose (ng/min)")
;
#delimit cr
```



Chúng ta cũng có thể so sánh mức độ đáp ứng với isoproterenol của hai nhóm, người da đen với người da trắng thông qua biểu đồ:

```
#delimit ;
parplot fbf0-fbf400
, transform(raw)
xlabel(1 "0" 2 "10" 3 "20" 4 "60" 5 "150" 6 "300"
7 "400")
ylabel(0(5)45, angle(horizontal))
ytitle("forearm blood flow (ml/min/dl)")
xtitle("isoproterenol dose (ng/min)")
by(race)
;
#delimit cr
```



Nhìn vào biểu đồ có thể thấy mức độ thay đổi lưu lượng máu cẳng tay của mỗi cá nhân khác nhau đáp ứng với mỗi liều isopritenol là khác nhau. Trong khi đó, xu hướng đáp ứng tại các liều isoproternol là tương tự cho từng đối tượng (nếu đáp ứng mạnh tại nồng độ 20 ng/min thì cũng đáp ứng mạnh tại nồng độ 300, 400 ng/min) tức là những lần đo này của từng cá nhân là không độc lập với nhau và vi phạm giả định của hồi quy tuyến tính, do đó sử dụng hồi quy tuyến tính để mô tả sự liên quan giữa liều isoproterenol và lưu lượng máu là không chính xác.

Một cách tiếp cận khác, nếu chỉ so sánh lưu lượng máu cẳng tay ở 2 liều isoproterenol khác nhau, hoặc so sánh lưu lượng máu cẳng tay ở hai nhóm màu da tại mỗi liều isoproterenol khác nhau thì chỉ cần sử dụng phép kiểm t-test. Tuy nhiên, nghiên cứu cần đánh giá một cách hệ thống sự thay đổi lưu lượng máu cẳng tay qua 7 liều isoproterenol, nếu sử dụng t-test sẽ phải cần đến 7 lần nếu so sánh lưu lượng máu cẳng tay giữa hai nhóm màu da (vì có 7 liều isoproterenol khác nhau) và nếu muốn so sánh mức độ thay đổi lưu lượng máu cẳng tay qua các liều isoproterenol khác nhau thì phải cần tới (7*6/2 = 21) phép kiểm t-test dạng bắt cặp!, như vậy sẽ gặp phải vấn đề kiểm định nhiều giả thuyết. Trong trường hợp này, chúng ta phải sử dụng mô hình ảnh hưởng hỗn hợp

- **Chuyển đổi dạng dữ liệu**

Như đã đề cập ở trên, dữ liệu đang ở dạng ngang (wide form), để phân tích được mô hình ảnh hưởng hỗn hợp, dữ liệu phải được chuyển sang dạng dọc (long form). Sở dĩ có tên như vậy vì ở dạng ngang, nếu càng đo lại nhiều lần thì dữ liệu sẽ càng phát triển theo chiều ngang (nhiều cột hơn) còn khi ở dạng dọc nếu đo lại càng nhiều lần thì dữ liệu phát triển theo chiều dọc (nhiều hàng hơn). Để chuyển sang dạng dọc, chúng ta sử dụng câu lệnh

```
reshape long fbf, i(id) j(dose)
list in 1/14, nolabel
```

	id	dose	race	fbf
1.	1	0	0	1
2.	1	10	0	1.4
3.	1	20	0	6.4
4.	1	60	0	19.1
5.	1	150	0	25
6.	1	300	0	24.6
7.	1	400	0	28
8.	2	0	0	2.1
9.	2	10	0	2.8
10.	2	20	0	8.3
11.	2	60	0	15.7
12.	2	150	0	21.9
13.	2	300	0	21.7
14.	2	400	0	30.1

Như vậy dữ liệu đã được chuyển sang dạng dọc. Ở dạng dữ liệu này mỗi lần đo sẽ ở một dòng (như vậy mỗi cá nhân có 7 dòng nếu có 7 lần đo), khác với dữ liệu dạng ngang (mỗi lần đo ở 1 cột), dữ liệu hiện tại còn 4 biến số: mã số bệnh nhân (id), liều isoproterenol (dose), màu da (race: 0 = da trắng, 1 = da đen), và lưu lượng máu cẳng tay (fbf). Việc chuyển đổi dữ liệu sang dạng dọc rất mạnh với dữ liệu trống, ví dụ một người chỉ đo lường được 5 lần (thiếu 2 lần), khi ở dữ liệu dạng ngang 2 ô sẽ bị trống, nhưng khi chuyển sang dữ liệu dạng dọc, cá nhân đó sẽ có 5 dòng dữ liệu và không có ô nào bị trống dữ liệu.

- **Xây dựng mô hình**

Trong phần xây dựng mô hình này, chúng ta sẽ giải đáp lần lượt 2 câu hỏi nghiên cứu:

(1) Liều isoproterenol có ảnh hưởng tới lưu lượng máu ở cẳng tay hay không?

(2) Mức độ đáp ứng với isoproterenol có khác nhau giữa hai nhóm màu da (da đen và da trắng) hay không?

- **Kiểm tra và thêm random intercept**

Trước khi thêm phần random intercept, cần thiết phải kiểm tra thử việc thêm random intercept có làm cho mô hình có "tốt" (phù hợp với số liệu hơn) hay không, nếu mô hình tốt hơn khi có phần random intercept thì chúng ta thêm vào mô hình. Quy trình có thể tóm tắt như sau:

(1) Tạo mô hình rỗng (chỉ có biến phụ thuộc) chỉ bao gồm phần cố định: $fbf = \beta_0 + e$

(2) Thêm vào mô hình rỗng phía trên phần random intercept μ: $fbf_{ij} = \beta_0 + \mu_{0j} + e_{ij}$

(3) Sử dụng phép kiểm Likelihood ratio test để kiểm tra nếu mô hình có random intercept có phù hợp hơn mô hình chỉ bao gồm phần cố định hay không

(4) Nếu mô hình có random intercept phù hợp hơn thì giữ mô hình và phân tích các bước tiếp theo

Để tạo mô hình rỗng chỉ bao gồm phần cố định, chúng ta dùng lệnh:

```
mixed fbf
* Lưu mô hình để so sánh với mô hình có random intercept
estimate store nullmodel0
```

Sau đó tạo mô hình rỗng bao gồm phần random intercept, ở đây, 7 lần đo "gộp" vào chung 1 người, hay chúng ta muốn mô hình sự dao động về lưu lượng máu tại thời điểm ban đầu, do đó random intercept sẽ được tính trên từng cá nhân (id)

```
mixed fbf || id:
* Lưu mô hình
estimate store nullmodel1
```

So sánh 2 mô hình bằng likelihood ratio test:

```
lrtest nullmodel0 nullmodel1
```

```
Likelihood-ratio test                                 LR chi2(1)  =     22.02
(Assumption: nullmodel0 nested in nullmodel1)         Prob > chi2 =    0.0000
```

Phép kiểm likelihood ratio test cho thấy có sự khác biệt có ý nghĩa thống kê ($\chi^2_{(1)}$ = 22,02; p <0,0001) giữa mô hình có và không random intercept (mô hình không có random intercept có log likelihood là -540.67 nhỏ hơn mô hình có random intercept -529.66). Do đó mô hình random intercept là phù hợp hơn.

- **Thêm biến tiên lượng (predictor)**

Sau khi kiểm tra và thêm phần random intercept, chúng ta thêm biến tiên lượng (ở đây là biến liều lượng isoproterenol (dose)) để kiểm tra mối liên hệ giữa lưu lượng máu cẳng tay và liều isoproterenol.

```
mixed fbf dose || id:
```

```
Computing standard errors:

Mixed-effects ML regression                     Number of obs     =        150
Group variable: id                              Number of groups  =         22

                                                Obs per group:
                                                              min =          3
                                                              avg =        6.8
                                                              max =          7

                                                Wald chi2(1)      =     163.42
Log likelihood = -476.95216                     Prob > chi2       =     0.0000
```

Phần cố định (fixed effect)

fbf	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dose	.0352574	.0027581	12.78	0.000	.0298517	.0406631
_cons	4.966772	1.24597	3.99	0.000	2.524717	7.408828

Phần ngẫu nhiên (random effect)

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Identity				
var(_cons)	27.38189	9.390553	13.98124	53.62668
var(Residual)	24.73093	3.091446	19.35697	31.59682

LR test vs. linear model: chibar2(01) = 64.59 Prob >= chibar2 = 0.0000

```
* Lưu mô hình để so sánh với mô hình random slope ở phần sau
estimate store modelA
```

Mô hình này có thể viết lại bằng phương trình

$$fbf = 4{,}967 + 0{,}035*dose + \mu_{0j} + e_{ij}.$$

Trong đó có hai phần, phần cố định bao gồm điểm chặn và độ dốc cố định (fixed slope). Kết quả cho thấy liều isoproterenol có ảnh hưởng thuận tới lưu lượng máu cẳng tay, cứ tăng liều isoproterenol lên 1 ng/min thì lưu lượng máu cẳng tay tăng lên 0,035 ml/min/dl (p < 0,001). Tuy nhiên mức độ thay đổi này là như nhau cho tất cả các cá nhân.

Mô hình còn có phần ngẫu nhiên, bao gồm điểm chặn ngẫu nhiên, là lưu lượng máu cẳng tay dao động giữa các cá nhân (biểu thị bằng var (_cons)) và phần sai số ngẫu nhiên của mô hình e_{ij} (biểu thị bằng var(Residual)). Trong mô hình này phương sai của e_{ij} bằng 24,73. Sai số càng nhỏ càng chứng tỏ mô hình càng phù hợp với số liệu, do đó phần sai số ngẫu nhiên (hay phần dư) rất quan trọng để đánh giá mô hình.

- **Kiểm tra và thêm random slope**

Ở phần trên kết quả đã cho thấy có mối liên quan giữa nồng độ isoproterenol với lưu lượng máu cẳng tay và mức độ thay đổi của lưu lượng máu cẳng tay như nhau cho mọi cá nhân. Tuy nhiên, như đã đề cập, mức độ đáp ứng với mỗi liều isoproterenol có thể khác nhau giữa các cá nhân, do đó cần kiểm tra việc thêm random slope vào mô hình theo mỗi liều isoproterenol (dose).

```
mixed fbf dose || id: dose, cov(unstructured)
```

```
Mixed-effects ML regression                     Number of obs     =        150
Group variable: id                              Number of groups  =         22

                                                Obs per group:
                                                              min =          3
                                                              avg =        6.8
                                                              max =          7

                                                Wald chi2(1)      =      50.84
Log likelihood = -437.38499                     Prob > chi2       =     0.0000
```

Phần cố định (fixed effect)

fbf	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dose	.0356267	.0049967	7.13	0.000	.0258334	.0454199
_cons	4.929402	.6681074	7.38	0.000	3.619935	6.238868

Phần ngẫu nhiên (random effect)

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Unstructured				
var(dose)	.0004634	.0001641	.0002314	.0009278
var(_cons)	6.388516	2.791335	2.713217	15.04235
cov(dose,_cons)	.0544077	.	.	.
var(Residual)	12.67931	1.585009	9.924052	16.19952

LR test vs. linear model: chi2(3) = 143.72 Prob > chi2 = 0.0000

```
* Lưu mô hình
estimate store modelB
```

So sánh 2 mô hình có random slope và không có random slope bằng likelihood ratio test.

```
lrtest modelA modelB
```

```
Likelihood-ratio test                                 LR chi2(1)  =     79.13
(Assumption: modelA nested in modelB)                 Prob > chi2 =    0.0000
```

Kết quả phép kiểm likelihood ratio test có sự khác biệt có ý nghĩa thống kê ($\chi^2_{(1)}$ = 79,13; p <0,0001) giữa mô hình có và không random slope, cho thấy mô hình có random slope là mô hình phù hợp hơn. Sai số ngẫu nhiên của mô hình var(e_{ij})= 12,68 cũng nhỏ hơn đáng kể so với mô hình không có random slope (var(e_{ij}) = 24,73).

Mô hình này có thể viết lại bằng phương trình

$$fbf = (4{,}929 + 0{,}036*dose) + (\mu_{1j}*dose + \mu_{0j} + e_{ij})$$

Sau khi thêm vào mô hình phần độ dốc ngẫu nhiên, liều lượng isoproterenol vẫn có ý nghĩa thống kê (giá trị thống kê z = 7,13; p < 0,001). Trong phần ảnh hưởng ngẫu nhiên đã có thêm phần random slope, tức là mức độ dao động về lưu lượng máu cẳng tay giữa các cá nhân (biểu thị bằng var(dose)). Ở đây có một chỉ số quan trọng là cov(dose,_cons), đây là hiệp biến giữa phần random intercept và random slope. Cov(dose,_cons) có giá trị dương, cho thấy người có lưu lượng máu cẳng tay cao hơn từ thời điểm ban đầu sẽ có mức độ thay đổi lưu lượng máu lớn hơn tại mỗi liều isoproterenol.

Như vậy chúng ta đã trả lời được câu hỏi nghiên cứu đầu tiên, đó là: liều lượng isoproterenol có ảnh hưởng thuận tới lưu lượng máu cẳng tay (cứ tăng liều isoproterenol lên 1 ng/min thì lưu lượng máu cẳng tay tăng lên 0,036 ml/min/dl). Hơn nữa, những người có lưu lượng máu cẳng tay cao hơn có xu hướng đáp ứng mạnh hơn với liều isoproterenol.

- **Mô hình tương tác giữa liều lượng và màu da**

Câu hỏi nghiên cứu thứ hai, chúng ta muốn biết mức độ đáp ứng với isoproterenol có khác nhau giữa hai nhóm màu da (da đen và da trắng) hay không**.** Để trả lời được câu hỏi này, chúng ta phải kiểm tra sự tương tác giữa biến số màu da (race) và biến liều lượng (dose). Mô hình có thể viết thành phương trình

$$fbf_{ii} = (\beta_0 + \beta_1{}^*dose_{ij} + \beta_2{}^*race + \beta_3{}^*race{}^*dose) + (\mu_{1j}{}^*dose_{ij} + \mu_{0j} + e_{ij})$$

Các câu lệnh để thực hiện mô hình tương tác:

```
* Tạo biến tương tác
```

```
gen race_dose = race*dose
* Thêm biến tương tác vào mô hình
mixed fbf dose race race_dose || id: dose, cov(unstructured)
```

```
                                                Wald chi2(3)      =     128.50
Log likelihood = -429.13311                     Prob > chi2       =     0.0000
```

fbf	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
dose	.0497674	.004543	10.95	0.000	.0408634	.0586715
race	-3.01543	1.203554	-2.51	0.012	-5.374353	-.6565077
race_dose	-.0345474	.0070824	-4.88	0.000	-.0484286	-.0206661
_cons	6.161752	.7749016	7.95	0.000	4.642973	7.680532

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id: Unstructured				
var(dose)	.0001824	.000077	.0000798	.0004171
var(_cons)	4.342916	2.109226	1.676401	11.25084
cov(dose,_cons)	.0281484	.0093922	.00974	.0465569
var(Residual)	12.51898	1.564468	9.799328	15.99343

```
LR test vs. linear model: chi2(3) = 86.84                 Prob > chi2 = 0.0000
```

Kết quả cho thấy sự tương tác giữa màu da và nồng độ isoproterenol có ý nghĩa thống kê ($p < 0,001$), tức là màu da có ảnh hưởng tới lưu lượng máu cẳng tay qua các liều isoproterenol khác nhau.

Phần cố định của mô hình có thể viết lại dưới dạng phương trình

$$fbf = 6,16 + 0,05*dose – 3,02*race – 0,03*race*dose$$

Vì màu da chỉ có 2 giá trị 0 = da trắng, 1 = da đen, do đó lưu lượng máu cẳng tay ở người da đen có thể viết là

$$fbf = 6,16 + 0,05*dose – 3,02*1 – 0,03*1*dose = 3,14 + 0,02*dose$$

Và người da trắng là:

$$fbf = 6,16 + 0,05*dose – 3,02*0 – 0,03*0*dose = 6,16 + 0,05*dose$$

Như vậy có thể kết luận rằng, mức độ đáp ứng với isoproterenol có khác nhau giữa hai nhóm màu da. Trong đó những người da trắng có mức độ đáp ứng với isoproterenol cao hơn những người da đen. Đến đây chúng ta đã trả lời được câu hỏi nghiên cứu thứ 2.

Trên đây là một ví dụ minh họa sử dụng mô hình ảnh hưởng hỗn hợp để phân tích nghiên cứu lâm sàng với thiết kế đo lường lặp lại. Chúng ta có thể mở rộng ra cho nhiều trường hợp với thiết kế tương tự, như đánh giá hiệu quả quả một loại thuốc, đánh giá hiệu

quả của một phương pháp can thiệp… và áp dụng với các hồi quy khác như hồi quy logistic, hồi quy poisson…

Tài liệu tham khảo

1. Dupont WD. Statistical Modeling for Biomedical Researchers. 2nd edition: Cambridge University Press; 2009.
2. Lang CC, Stein CM, Brown RM, Deegan R, Nelson R, He HB, et al. Attenuation of isoproterenol-mediated vasodilatation in blacks. N Engl J Med. 1995;333(3):155-60.