**15**

**Mô hình ảnh hưởng hỗn hợp**

**(mixed-effects model)**

Một trong những mô hình nghiên cứu phổ biến có tên là *prospective study*, hay nghiên cứu theo thời gian. Trong mô hình nghiên cứu theo thời gian, các đối tượng nghiên cứu được theo dõi theo thời gian để đánh giá hiệu quả của một thuật can thiệp. Chẳng hạn như hai nhóm bệnh nhân được điều trị bằng hai loại thuốc, và mỗi bệnh nhân được đo nồng độ glucose mỗi 6 tháng. Vấn đề đặt ra là làm thế nào để đánh giá hiệu quả của can thiệp một cách có hệ thống. Chúng ta không thể dùng t-test cho 6 thời điểm, vì làm như thế là gặp vấn đề kiểm định nhiều giả thuyết. Vả lại, mỗi bệnh nhân được đo lường nhiều lần, và các giá trị phải có liên quan nhau; do đó, kiểm định nhiều giả thuyết mà không xem xét đến sự tương quan trong mỗi cá nhân là một sai lầm nghiêm trọng.

Chương này sẽ hướng dẫn cách phân tích số liệu để kiểm định giả thuyết về ảnh hưởng của thuốc bằng một mô hình thống kê được xem là “chuẩn vàng” hiện nay: đó là mô hình *mixed-effects* mà tôi tạm dịch là “*mô hình ảnh hưởng hỗn hợp*”. Phân tích bằng mô hình này cần có máy tính và chương trình phân tích thống kê. Tôi sẽ sử dụng phần mềm R để phân tích, và sẽ trình bày các mã trong bài để bạn đọc dễ dàng theo dõi.

**15.1 Phân biệt fixed effect và random effect**

Một trong những khó khăn của người mới học thống kê và thí nghiệm là khái niệm fixed effects (có lẽ dịch là *ảnh hưởng cố định*) và random effects (*ảnh hưởng ngẫu nhiên*). Sách giáo khoa cũng không giải thích rõ ràng, một phần vì tác giả không xem đó là quan trọng, một phần vì … khó hiểu. Tuy nhiên, để sử dụng R (đặc biệt là package lme4) thì chúng ta cần phải phân biệt hai khái niệm trên. Trong phần này, tôi sẽ giải thích fixed effects và random effects qua một số ví dụ cụ thể.

**Phân tích phương sai 1 chiều (one-way analysis of variance)**

Số liệu dưới đây là điểm cho 3 nhóm vận động viên. Nhóm 1 gồm 10 người, nhóm hai gồm 6 người, và nhóm 3 có 9 người. Mục tiêu là kiểm định giả thuyết ba nhóm này có cùng điểm. Phương pháp kiểm định mà chúng ta nghĩ đến là phân tích phương sai (ANOVA), với hàm lm(Score ~ Group) như đã giới thiệu trong Chương 11.

|  |  |  |
| --- | --- | --- |
| **ID** | **Group** | **Score** |
| 1 | 1 | 3 |
| 2 | 1 | 13 |
| 3 | 1 | 13 |
| 4 | 1 | 8 |
| 5 | 1 | 11 |
| 6 | 1 | 9 |
| 7 | 1 | 12 |
| 8 | 1 | 7 |
| 9 | 1 | 16 |
| 10 | 1 | 15 |
| 11 | 2 | 18 |
| 12 | 2 | 6 |
| 13 | 2 | 21 |
| 14 | 2 | 34 |
| 15 | 2 | 26 |
| 16 | 2 | 11 |
| 17 | 3 | 24 |
| 18 | 3 | 14 |
| 19 | 3 | 21 |
| 20 | 3 | 5 |
| 21 | 3 | 17 |
| 22 | 3 | 17 |
| 23 | 3 | 23 |
| 24 | 3 | 19 |
| 25 | 3 | 7 |

**Mô hình trung bình**

Trong mô hình phân tích phương sai, bằng ngôn ngữ của mô hình tuyến tính, chúng ta có thể phát biểu rằng: Gọi Score của đối tượng *i* (*i*=1, 2, 3, …, 25) và nhóm *j*  là *Y*ij. Chúng ta nghĩ rằng những dao động về *Y*ij được xác định bởi giá trị trung bình của mỗi nhóm (tạm gọi là βj), với *j =* 1, 2, 3, cộng với một sai số ngẫu nhiên kí hiệu là *ε*ij. Nói cách khác, chúng ta có:

*Y*ij = βj + *ε*ij [1]

Trong đó, *ε*ij ~ N(0, σ2). Nói cách khác *ε*ij giả định tuân theo luật phân bố chuẩn với trung bình 0 và phương sai σ2. Có thể hiểu mô hình này bằng cách trình bày lại bảng số liệu trên như sau:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **Quan sát** | | **Mô hình hoá** | |
| **Group** | **Score** | βj | *ε*ij |
| 1 | 1 | 3 | β1 | *ε*11 |
| 2 | 1 | 13 | β1 | *ε*21 |
| 3 | 1 | 13 | β1 | *ε*31 |
| … | … | … | … | … |
| 8 | 1 | 7 | β1 | *ε*81 |
| 9 | 1 | 16 | β1 | *ε*91 |
| 10 | 1 | 15 | β1 | *ε*101 |
| 11 | 2 | 18 | β2 | *ε*12 |
| 12 | 2 | 6 | β2 | *ε*22 |
| 13 | 2 | 21 | β2 | *ε*32 |
| 14 | 2 | 34 | β2 | *ε*42 |
| 15 | 2 | 26 | β2 | *ε*52 |
| 16 | 2 | 11 | β2 | *ε*62 |
| 17 | 3 | 24 | β3 | *ε*13 |
| 18 | 3 | 14 | β3 | *ε*23 |
| 19 | 3 | 21 | β3 | *ε*33 |
| 20 | 3 | 5 | β3 | *ε*43 |
| … | … | … | … | … |
| 25 | 3 | 7 | β3 | *ε*93 |

Ở mức độ nhóm, chúng ta mô hình rằng những khác biệt giữa các cá nhân trong nhóm bằng giá trị trung bình μ0, cộng với các khác biệt ngẫu nhiên (kí hiệu là *u*oj) chung quanh giá trị trung bình:

βj = μ0 + *u*oj [2]

Trong đó, *u*oj ~ N(0, τ00). Thay thế [2] vào [1]:

*Y*ij = μ0 + (*u*oj + *ε*ij) [3]

Theo mô hình trên, μ0 là một *fixed effect* vì là một thông số. Còn (*u*oj + *ε*ij) được gọi là *random effects* vì phản ảnh phương sai.

Hàm **lmer** (linear mixed effect regression) trong package **lme4** rất có ích trong việc ước tính các thông số của mô hình trên. Hàm lmer có hai phần chính: fixed effects và random effects (trong ngoặc). Do đó, để hiểu và sử dụng lmer, chúng ta cần phải viết xuống mô hình phân tích. Chúng ta sẽ tạo ra một dataframe có tên là test như sau:

Score = c(3, 13, 13, 8, 11, 9, 12, 7, 16, 15, 18, 6, 21, 34, 26, 11, 24, 14, 21, 5, 17, 17, 23, 19, 7)

Group = c(rep("1", 10), rep("2", 6), rep("3", 9))

test = data.frame(Group, Score)

Để ước tính các thông số của mô hình [3], chúng ta dùng lmer như sau:

library(lme4)

fit = lmer(Score ~ 1 + (1 | Group), data=test, REML=0)

summary(fit)

Trong hàm trên số “1” chỉ intercept, tức mô hình đơn giản nhất, chỉ có số trung bình cho toàn bộ quần thể. Phần random effects, (1 | Group) phản ảnh *u*oj trong mỗi nhóm.

|  |
| --- |
| > summary(fit)  Linear mixed model fit by maximum likelihood ['lmerMod']  Formula: Score ~ 1 + (1 | Group)  Data: test  AIC BIC logLik deviance  175.0157 178.6723 -84.5079 **169.0157**  **Random effects:**  Groups Name Variance Std.Dev.  Group (Intercept) 7.082 2.661  Residual 45.797 6.767  Number of obs: 25, groups: Group, 3  **Fixed effects:**  Estimate Std. Error t value  (Intercept) 15.150 2.059 7.358 |

Đối chiếu với mô hình *Y*ij = μ0 + (*u*oj + *ε*ij), chúng ta thấy ước số của μ0 =15.2. Phương sai giữa các nhóm là τ00 = 7.08, và phương sai giữa các cá nhân là σ2 = 45.8.

**Mô hình ảnh hưởng nhóm**

Với mô hình mới, chúng ta cho rằng điểm của mỗi cá nhân *Y*ij là hàm số của intercept (μ0) + trung bình của mỗi nhóm (βj) + sai số ngẫu nhiên. Nói cách khác, mô hình mới là:

*Y*ij = μ0 + βj + *ε*ij [4]

Với *ε*ij ~ N(0, σ2). Trong đó, trung bình của mỗi nhóm βj là hàm số của trung bình thật của nhóm và sai số ngẫu nhiên.

βj = αj + *u*oj [5]

Với *u*oj ~ N(0, τ00). Thay [5] vào [4] chúng ta có mô hình:

*Y*ij = μ0 + αj +(*u*oj + *ε*ij)

Chúng ta dùng hàm lmer cho mô hình trên như sau:

fit1 = lmer(Score ~ 1 + Group + (1 | Group), data=test, REML=0)

summary(fit1)

|  |
| --- |
| > summary(fit1)  Linear mixed model fit by maximum likelihood ['lmerMod']  Formula: Score ~ 1 + Group + (1 | Group)  Data: test  AIC BIC logLik deviance  173.3541 179.4484 -81.6770 **163.3541**  **Random effects:**  Groups Name Variance Std.Dev.  Group (Intercept) 0.0 0.000  Residual 40.3 6.348  Number of obs: 25, groups: Group, 3  **Fixed effects:**  Estimate Std. Error t value  (Intercept) 10.700 2.007 5.330  Group2 8.633 3.278 2.634  Group3 5.633 2.917 1.931 |

Mô hình trên cho thấy so sánh với nhóm 1 điểm của nhóm 2 cao hơn 8.6 điểm, và nhóm 3 cao hơn 5.6 điểm, nhưng sự khác biệt đều không có ý nghĩa thống kê. Chú ý residual (tức σ2) giảm từ 45.8 xuống còn 40.3.

**15.2 Mô hình ảnh hưởng hỗn hợp cho nghiên cứu theo thời gian**

Bảng số liệu dưới đây trình bày kết quả của một thí nghiệm trên chuột. Nghiên cứu có 19 chuột, được chia thành 2 nhóm: nhóm thứ nhất được cho uống thuốc (n = 9 chuột) và nhóm thứ hai là nhóm chứng (không uống thuốc, gồm 10 chuột). Ở mỗi chuột, nồng độ đường trong máu (glucose) được đo 4 thời điểm: trước khi uống thuốc (T0), 2 giờ, 3 giờ, và 4 giờ sau khi uống thuốc (tạm kí hiệu T2, T3 và T4). Mô hình nghiên cứu này có khi được đề cập đến dưới cái tên *repeated measure design*. Kết quả của thí nghiệm như sau:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Bảng 1. Nồng độ glucose trong máu của chuột qua 4 thời điểm** | | | | | |
| **Treatment** | **Id** | **T0** | **T2** | **T3** | **T4** |
| Test | 1 | 5.9 | 3.9 | 3.9 | 3.6 |
| Test | 2 | 5.3 | 4.7 | 3.5 | 3.2 |
| Test | 3 | 4.6 | 3.7 | 3.3 | 3.2 |
| Test | 4 | 6.2 | 4.6 | 4.3 | 3.9 |
| Test | 5 | 6.0 | 5.4 | 5.2 | 4.8 |
| Test | 6 | 6.4 | 4.7 | 4.8 | 4.3 |
| Test | 7 | 7.6 | 4.1 | 3.8 | 4.1 |
| Test | 8 | 5.9 | 3.1 | 3.6 | 3.3 |
| Test | 9 | 7.5 | 6.1 | 5.4 | 4.6 |
| Control | 10 | 6.2 | 5.3 | 4.9 | 4.5 |
| Control | 11 | 6.9 | 5.6 | 5.9 | 5.9 |
| Control | 12 | 5.6 | 4.7 | 4.6 | 4.0 |
| Control | 13 | 5.1 | 3.9 | 2.9 | 2.9 |
| Control | 14 | 5.7 | 4.7 | 4.3 | 4.6 |
| Control | 15 | 5.0 | 4.0 | 3.5 | 3.3 |
| Control | 16 | 5.2 | 4.2 | 4.0 | 3.8 |
| Control | 17 | 7.7 | 6.2 | 6.1 | 5.7 |
| Control | 18 | 8.0 | 5.8 | 6.5 | 6.0 |
| Control | 19 | 7.7 | 5.0 | 6.3 | 6.2 |
| **Chú thích**: *id* là cột chỉ mã số của chuột. | | | | | |

Chúng ta thấy nồng độ đường (sẽ gọi là glucose) ở cả hai nhóm có xu hướng giảm theo thời gian. Câu hỏi chính là thuốc có ảnh hưởng đến sự thuyên giảm glucose hay không? Cụm từ “ảnh hưởng” ở đây có thể hiểu rõ hơn: nó chính là sự khác biệt (difference) về mức độ giảm glucose giữa hai nhóm chuột. Vì thế, câu hỏi trên có thể diễn giải một cách định lượng như sau: mức độ giảm glucose ở nhóm uống thuốc cao (hay thấp) hơn nhóm chứng hay không?

**Thẩm định số liệu**

Chúng ta cần một mô hình để mô tả mức độ giảm *cho mỗi chuột*. Hãy xem xét chuột số 1 (id = 1) với các số liệu sau đây (gọi T là thời điểm):

* Trước khi uống thuốc (T=0) thì glucose = 5.9 mmol/L
* Sau khi uống thuốc 2 giờ (T=2), glucose giảm xuống 3.9 (tức giảm 33%)
* Sau khi uống thuốc 3 giờ (T=3), glucose vẫn ở 3.9 (tức giảm 33%)
* Sau khi uống thuốc 4 giờ (T=4), glucose giảm thêm một chút 3.5 (tức giảm 41%)

Nhưng đối với chuột có id = 9 thì nồng độ glucose lúc đầu cao hơn (9 mmol//L), và giảm liên tục đến thời điểm 4 giờ sau khi uống thuốc còn 4.6 mmol/L, tức giảm 49%. Chúng ta có thể xem xét qua sự thay đổi nồng độ glucose cho từng chuột bằng biểu đồ 1 sau đây:

**# nhập số liệu glucose và gọi tên biến của là y:**

y = c(5.9, 3.9, 3.9, 3.6, 5.3, 4.7, 3.5, 3.2, 4.6, 3.7,

3.3, 3.2, 6.2, 4.6, 4.3, 3.9, 6.0, 5.4, 5.2, 4.8,

6.4, 4.7, 4.8, 4.3, 7.6, 4.1, 3.8, 4.1, 5.9, 3.1,

3.6, 3.3, 7.5, 6.1, 5.4, 4.6, 6.2, 5.3, 4.9, 4.5,

6.9, 5.6, 5.9, 5.9, 5.6, 4.7, 4.6, 4.0, 5.1, 3.9,

2.9, 2.9, 5.7, 4.7, 4.3, 4.6, 5.0, 4.0, 3.5, 3.3,

5.2, 4.2, 4.0, 3.8, 7.7, 6.2, 6.1, 5.7, 8.0, 5.8,

6.5, 6.0, 7.7, 5.0, 6.3, 6.2)

# 3 lệnh sau đây thực hiện 3 việc:

# tạo biến số id

# tạo biến số time có giá trị 0,2,3,4 lặp lại 19 lần

# tạo biến số treatment

id = rep(1:19, each=4)

time = rep(c(0,2,3,4), 19)

treatment = rep(1:2, c(9\*4, 10\*4))

glucose = data.frame(id,treatment, time ,y)

library(rms)

library(lattice)

library(lme4)

**# Vẽ biểu đồ cho từng chuột**

xyplot(y ~ time | id, type=c("p","r"), as.table=time, xlab="Time", ylab="Glucose")

|  |
| --- |
|  |
| **Biểu đồ 1**. Nồng độ glucose cho từng chuột qua từng thời điểm. |

Xem xét qua hai trường hợp trên và biểu đồ 1 cho chúng ta thấy hai điểm đáng chú ý: thứ nhất *nồng độ glucose lúc* *ban đầu* (baseline) khác nhau giữa các chuột; và thứ hai, *tốc độ* (rate) giảm glucose cũng khác nhau giữa các chuột. Chúng ta cần một số kí hiệu để bàn luận về mô hình:

* *i* là mã số định danh của chuột 1 đên 19 (i = 1, 2, 3, …, 19);
* *yi* là nồng độ glucose đo lường được cho chuột *i*;
* *ai* là nồng độ glucose trước khi uống thuốc của chuột *i;*
* *bi* là tốc độ giảm glucose của chuột *i*.
* Tốc độ giảm glucose tùy thuộc vào thời gian và thời gian **time** có thể tạm kí hiệu bằng *T*. Ở đây *T* = 0, 2, 3, và 4.

Chúng ta có thể hình dung ra một mô hình để mô tả sự thay đổi nồng độ glucose ở từng chuột như sau:

Mô hình trên phát biểu rằng giá trị glucose đo lường của từng chuột được xác định bằng giá trị lúc ban đầu (chưa can thiệp) và tốc độ thay đổi theo thời gian *T*. Nhưng ở đây, chúng ta chỉ đo lường nồng độ glucose qua 4 thời điểm, và mỗi lần đo lường đều có độ nhiễu (tức không hoàn toàn chính xác, do kĩ thuật đo lường hay do sự dao động sinh học ở mỗi chuột mà chúng ta chưa biết), cho nên chúng ta cần thêm một thông số khác để cho mô hình hoàn chỉnh hơn. Thông số đó là độ nhiễu và tạm kí hiệu bằng *ei*. Bây giờ thì mô hình trên trở thành:

[1]

Đây là mô hình hồi qui tuyến tính (linear regression model) mà có lẽ các bạn đã từng biết qua. Đối với chuột id=1, chúng ta có thể ước tính thông số *a* và *b* trên bằng lệnh R như sau:

time = c(0, 2, 3, 4)

y = c(5.9, 3.9, 3.9, 3.6)

summary(lm(y ~ time))

**> summary(lm(y ~ time))**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.6171 0.4467 12.574 0.00627 \*\*

time -0.5743 0.1659 -3.461 0.07428 .

Do đó, *a*1 = 5.6 và *b*1 = -0.57. Nói cách khác ước số nồng độ glucose lúc ban đầu của chuột 1 là 5.6 mmol/L, và giảm khoảng 0.57 mmol/L mỗi giờ (60 phút) sau khi uống thuốc. Phân tích tương tự chúng ta sẽ có ước số cho chuột số 9 là: *a*1 = 7.52 và *b*1 = -0.72.

time = c(0, 2, 3, 4)

y = c(7.5, 6.1, 5.4, 4.6)

summary(lm(y ~ time))

**> summary(lm(y ~ time))**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.52000 0.04071 184.73 2.93e-05 \*\*\*

time -0.72000 0.01512 -47.62 0.000441 \*\*\*

Biểu đồ 2 sau đây cho chúng ta thấy sự dao động của nồng đồ glucose lúc ban đầu cũng như tốc độ thay đổi glucose theo thời gian giữa các chuột:

fit = by(glucose, id,

function(data) fitted.values(lm(y ~ time, data=data)))

fit = unlist(fit)

names(fit) = NULL

interaction.plot(time, id, fit, xlab="Time", ylab="Glucose")

|  |
| --- |
|  |
| **Biểu đồ 2**. Sự dao về động nồng độ glucose cho từng chuột giữa các thời điểm. |

Chúng ta thấy rằng thông số *ai*, *bi* là hai thông số chúng ta quan tâm. Nên nhớ rằng hai thông số này dao động giữa các chuột, cho nên chúng ta cần kí hiệu *i* để nhắc nhở. Nếu chuột trong nghiên cứu được phân nhóm một cách ngẫu nhiên, chúng ta kì vọng rằng nồng độ trung bình lúc ban đầu (tức trung bình *ai*) giữa hai nhóm sẽ không khác nhau, và ảnh hưởng của thuốc chủ yếu được phản ảnh qua tốc độ tăng glucose giữa hai nhóm(tức trung bình *bi*).

Chúng ta thử kiểm định sự khác biệt giữa hai nhóm bằng phương pháp kiểm định t (qua R) như sau

y = c(5.9, 5.3, 4.6, 6.2, 6.0, 6.4, 7.6, 5.9, 7.5,

6.2, 6.9, 5.6, 5.1, 5.7, 5.0, 5.2, 7.7, 8.0, 7.7)

treatment = rep(c(1,0), c(9,10))

# lệnh trên tạo ra biến số treatment với 2 giá trị 0 để chỉ nhóm chứng gồm 10 chuột, và 1 để chỉ nhóm thuốc gồm 9 chuột

t.test(y ~ treatment)

|  |
| --- |
| t = 0.3164, df = 16.849, **p-value = 0.7556**  alternative hypothesis: true difference in means is not equal to 0  95 percent confidence interval:  -0.876183 1.185072  sample estimates:  mean in group 0 mean in group 1  6.310000 6.155556 |

Chúng ta thấy qua kết quả trên rằng nồng độ glucose lúc ban đầu trong nhóm chứng là 6.31 và nhóm điều trị là 6.16, và độ khác biệt này không có ý nghĩa thống kê (P = 0.7556).

**15.3 Mô hình tăng trưởng không điều kiện (unconditional linear growth)**

Mô hình đơn giản nhất là mô tả tăng trưởng của từng cá nhân, còn gọi là *unconditional growth model*. Xin nhắc lại, qua mô hình [1] chúng ta phát biểu rằng nồng độ glucose ở mỗi chuột *i* ở một thời điểm *T* là kết quả của 3 giá trị: (a) nồng độ glucose lúc ban đầu (baseline glucose); (b) tỉ lệ thay đổi tùy theo thời gian *T*; và (c) giá trị nhiễu do đo lường và do dao động sinh học mà chúng ta không giải thích được (*ei*):

Theo mô hình này: nồng độ glucose của chuột 1, 2, 3, …, 19 và số liệu ước tính (quan sát) được là:

|  |  |
| --- | --- |
| **Mô hình** | **Quan sát** |
| *y*1 = *a*1 + *b*1*T* + *e*1  *y*2 = *a*2 + *b*2*T* + *e*2  *y*3 = *a*3 + *b*3*T* + *e*3  …  *y*19 = *a*19 + *b*19*T* + *e*19 | *y*1 = 5.62 – 0.57*T* + *e*1  *y*2 = 5.42 – 0.55 *T* + *e*2  *y*3 = 4.52 – 0.37*T* + *e*3  …  *y*19 = 7.07 – 0.34*T* + *e*19 |

Quan sát mô hình trên chúng ta thấy nồng độ glucose lúc ban đầu dao động trong khoảng 5 đến 8 mmol/L, nhưng chúng ta không biết số thật là bao nhiêu. Do đó, chúng ta có thể giả định bằng một mô hình đơn giản rằng nồng độ glucose lúc ban đầu giữa các chuột bằng một số trung bình cộng (hay trừ) độ khác biệt giữa các các chuột với số trung bình đó. Gọi số trung bình là *A,* và *u* là độ khác biệt của *ai* so với *A*, chúng ta có:

*ai* = *A + u* [2]

Bởi vì tốc độ biến đổi (giảm) glucose cũng dao động giữa các chuột, do đó chúng ta cũng có thể giả định rằng tốc độ trung bình là *B*, và độ khác biệt giữa *bi* là B là *v* :

*bi* = *B + v* [3]

Thay thế [2] và [3] vào phương trình [1] chúng ta có (tôi viết lại phương trình 1 để dễ theo dõi):

 [1]

*yi* = (*A + u*) + (*B + v*)*T* + *ei*

Sắp xếp lại phương trình trên cho gọn, chúng ta có:

*yi* = (*A + BT*) +(*u + vT* + *ei*) [4]

Thấy gì qua phương trình trên? Phương trình có 2 phần: phần đầu, (*A + BT*), là phần mà giới thống kê học gọi là “fixed effects” (ảnh hưởng cố định – thuật ngữ không mấy chính xác nhưng chúng ta tạm hiểu là ảnh hưởng không ngẫu nhiên); phần hai, (*u + vT* + *ei*), là phần “random-effects” (ảnh hưởng ngẫu nhiên). Sở dĩ gọi ảnh hưởng ngẫu nhiên là vì *u, v* và *e* đều là các thông số phản ảnh độ nhiễu (noise) của mô hình. Các thông số trong phương trình này có ý nghĩa như sau:

* *A* là nồng độ glucose trung bình lúc ban đầu (trước khi uống thuốc) của quần thể các chuột được nghiên cứu;
* *B* là tốc độ trung bình (tỉ lệ giảm nồng độ glucose) tính trên mỗi giờ của các chuột trong nghiên cứu;
* *u* phản ảnh độ dao động về nồng độ glucose lúc ban đầu giữa các chuột;
* *v* cho chúng ta biết độ dao động về tốc độ giảm glucose giữa các chuột; và
* *ei* là độ dao động ở mỗi chuột.

Mô hình sinh học nào cũng có giả định. Trong mô hình [4] trên, chúng ta giả định rằng:

* *u* tuân theo luật phân phối chuẩn (normal distribution) với số trung bình bằng 0 và phương sai *u*2. Viết theo kí hiệu thống kê học là: *u* ~ *N*(0, *u*2).
* *v* tuân theo luật phân phối chuẩn (normal distribution) với số trung bình bằng 0 và phương sai *v*2, hay *v* ~ *N*(0, *v*2).
* *e­i* cũng theo luật phân phối tương tự: *e­i* ~ *N*(0, *e*2)

Tại sao 0? Tại vì chúng ta giả định rằng tính trung bình, các độ nhiễu bằng 0 để giá trị trung bình (expected value hay giá trị kì vọng) của mô hình [4] là (kí hiệu E có nghĩa là *expected*, hay nói nôm na là “tính trung bình”):

*E*(*yi*) = (*A + BT*) + (0 *+* 0*T* + 0)

*E*(*yi*) = *A + BT*

Chúng ta sẽ triển khai và ước tính các thông số trong mô hình [4] bằng R. Nhắc lại, mô hình [4]. Trước hết, chúng ta đọc dữ liệu glucose và tạo ra một dataframe tạm gọi là glucose:

y = c(5.9, 3.9, 3.9, 3.6, 5.3, 4.7, 3.5, 3.2, 4.6, 3.7,

3.3, 3.2, 6.2, 4.6, 4.3, 3.9, 6.0, 5.4, 5.2, 4.8,

6.4, 4.7, 4.8, 4.3, 7.6, 4.1, 3.8, 4.1, 5.9, 3.1,

3.6, 3.3, 7.5, 6.1, 5.4, 4.6, 6.2, 5.3, 4.9, 4.5,

6.9, 5.6, 5.9, 5.9, 5.6, 4.7, 4.6, 4.0, 5.1, 3.9,

2.9, 2.9, 5.7, 4.7, 4.3, 4.6, 5.0, 4.0, 3.5, 3.3,

5.2, 4.2, 4.0, 3.8, 7.7, 6.2, 6.1, 5.7, 8.0, 5.8,

6.5, 6.0, 7.7, 5.0, 6.3, 6.2)

id = rep(1:19, each=4)

time = rep(c(0,2,3,4), 19)

treatment = rep(1:2, c(9\*4, 10\*4))

glucose = data.frame(id,treatment, time ,y)

Đối với mô hình [4],

*yi* = (*A + BT*) +(*u + vT* + *ei*)

các mã R sau đây có thể sử dụng để phân tích:

**library(lme4)**

**fit1 = lmer(y ~ 1 + time + (1 + time | id), data=glucose, REML=0)**

**summary(fit1)**

Chú ý hàm lmer được mô phỏng theo cách viết 2 phần: phần đầu là fixed effects, và phần 2 (trong ngoặc) là random effects. Khi viết “1 + time” chúng ta có thể hiểu đó là “intercept + time” (hay A + B\*time). Kết quả của mô hình này như sau:

|  |
| --- |
| **Random effects:**  Groups Name Variance Std.Dev. Corr  id (Intercept) 0.6581762 0.8113  time 0.0007453 0.0273 1.00  Residual 0.2481143 0.4981  Number of obs: 76, groups: id, 19  **Fixed effects:**  Estimate Std. Error t value  (Intercept) 6.05038 0.21322 28.38  time -0.48496 0.03914 -12.39  **Correlation of Fixed Effects:**  (Intr)  time -0.263 |

Chú ý kết quả được chia thành 3 phần: phần random effects, phần fixed effects, và phần tương quan giữa các thông số fixed effects.

Chúng ta tập trung vào phần fixed effects trước. Ước số time = -0.485 chính là *B* trong phần *BT* của mô hình trên. Ngoài ra, intercept = 6.05 chính là ước số của A trong mô hình trên. Nói cách khác, mô hình [4] bây giờ là (phần fixed effects):

*yi* = 6.05 – 0.485*T*

Mô hình này cho biết nồng độ glucose *trung bình* của quần thể là 6.05 mmol/L và tốc độ giảm 0.485 mmol/L sau mỗi giờ theo dõi.

Tuy nhiên, đó chỉ là phần fixed effects; chúng ta còn một phần random effects, và R cung cấp các kết quả:

**Random effects:**

Groups Name Variance Std.Dev. Corr

id (Intercept) 0.6581762 0.8113

time 0.0007453 0.0273 1.00

Residual 0.2481143 0.4981

Các kết quả trên đây có thể diễn giải như sau (theo mô hình [4]):

* Phương sai *v*2 = 0.658 (tức var(intercept) = 0.658) phản ảnh độ dao động về nồng độ glucose lúc ban đầu (trước khi can thiệp) giữa các chuột. Như chúng ta có thể đoán được, có sự khác biệt giữa các chuột về nồng độ glucose, nên chúng ta không ngạc nhiên khi thấy khoảng tin cậy 95% có ý nghĩa thống kê;
* Phương sai *u*2 = 0.00075 (tức var(time) = 0.00075) phản ảnh độ dao động về tốc độ giảm glucose giữa các chuột. Như có thể thấy qua ước số này, độ dao động về tốc độ không khác nhau đáng kể giữa chuột;

Dòng Residual = 0.248 (chính là ước số của *e*2 = 0.248), phản ảnh độ dao động nồng độ glucose *ở mỗi* chuột.

**15.4 Mô hình ảnh hưởng của điều trị (conditional growth model)**

Chúng ta để ý thấy trong mô hình [4] không có biến số nào liên quan đến điều trị! Đó là vì mô hình căn bản. Bây giờ chúng ta thử xem qua lí giải trong phương trình [2] và [3]:

*ai* = *A + u*

*bi* = *B + v*

Chúng ta muốn có khác biệt đáng kể về nồng độ glucose lúc ban đầu (*ai*) giữa hai nhóm chuột, do đó phương trình [2] trở thành:

*ai* = *A*0 *+ A1 x* treatment *+ u*

Ngoài ra, có thể có sự khác biệt về tốc độ giảm glucose giữa hai nhóm (*bi*), và chúng ta cần tìm hiểu mức độ khác biệt hay ảnh hưởng này. Một cách để biết ảnh hưởng của điều trị là viết lại phương trình [3] thành:

*bi* = *B0 + B1 x* treatment *+ v*

Do đó, mô hình [1] trở thành:

 [1]

*yi* = (*A*0 *+ A1 x* treatment *+ u*) + (*B*0 *+ B*1 *x* treatment *+ v*)*T* + *ei*

*yi* = (*A*0 *+ A*1*x*treatment *+ B*0*T + B*1*x*treatment*xT*) *+* (*u + vT* + *ei*) [5]

Đây là mô hình chính của nghiên cứu. Trong mô hình trên, ngoài các thông số khác đã giải thích, chúng ta thấy có hai thông số mới và ý nghĩa của 2 thông số này như sau:

* ***A*1 *x* treatment**phản ảnh sự khác biệt nồng độ glucose lúc ban đầu giữa hai nhóm chuột;
* ***B*1 *x* treatment *x T***phản ảnh mức độ khác biệt về tốc độ giảm glucose giữa hai nhóm chuột; do đó, chính là kiểm định thống kê cho giả thuyết của nghiên cứu.

Bây giờ, chúng ta sẽ dùng lme4 để ước tính các thông số trong mô hình [5].

**fit2 = lmer(y ~ 1 + treatment + time + time:treatment + (1+time|id), data=glucose, REML=0)**

**summary(fit2)**

Kết quả như sau:

|  |
| --- |
| **Random effects:**  Groups Name Variance Std.Dev. Corr  id (Intercept) 6.878e-01 0.829313  time 5.157e-06 0.002271 1.00  Residual 2.299e-01 0.479518  Number of obs: 76, groups: id, 19  **Fixed effects:**  Estimate Std. Error t value  (Intercept) 5.81683 0.69150 8.412  treatment 0.15302 0.43060 0.355  time -0.74032 0.11962 -6.189  treatment:time 0.16730 0.07449 2.246  **Correlation of Fixed Effects:**  (Intr) trtmnt time  treatment -0.950  time -0.377 0.358  treatmnt:tm 0.358 -0.377 -0.950 |

Đối chiếu với mô hình [5], các thông số trên như sau:

|  |  |  |
| --- | --- | --- |
| *yi* = (*A*0 *+ A*1*x*treatment *+ B*0*T + B*1*x*treatment*xT*) *+* (*u + vT* + *ei*) | | |
| **Thông số** | **Ý nghĩa** | **Ước số** |
| *A0* | Nồng độ trung bình glucose lúc ban đầu của quần thể | 5.82 |
| *A*1 | Khác biệt về nồng độ glucose lúc ban đầu giữa hai nhóm | 0.15 |
| *B*0 | Ảnh hưởng của thời gian (time) đến sự giảm glucose | -0.74 |
| *B*1 | Khác biệt về tốc độ giảm glucose giữa hai nhóm | 0.17 |

Do đó, mô hình [5] trở thành:

*yi* = 5.82 + 0.15 x treatment – 0.74 x time + 0.17 x treatment x time

Nên nhớ rằng chúng ta gọi **treatment = 1** chỉ thuốc và **treatment = 2** chỉ nhóm chứng, và **time** có4 giá trị: 0, 2, 3, và 4. Cho nên, mô hình mô tả sự dao động của glucose trong nhóm được điều trị bằng thuốc là:

*y1* = 5.82 + 0.15 x 1 – 0.74 x time + 0.17 x 1 x time

***y1* = 5.97 – 0.57 x time**

và cho nhóm chứng là:

*y2* = 5.82 + 0.15 x 2 – 0.74 x time + 0.17 x 2 x time

***y2* = 6.12 – 0.40 x time**

Như vậy, hai nhóm có nồng độ glucose khởi đầu giống nhau (5.97 và 6.12 và P = 0.74), nhưng tốc độ giảm glucose của nhóm thuốc là 0.57 mmol/L/giờ, cao hơn nhóm chứng 0.40 mmol/L/giờ, và sự khác biệt này có ý nghĩa thống kê (P = 0.027). Do đó, qua kết quả này, chúng ta có thể kết luận rằng thuốc có hiệu quả giảm glucose cao hơn nhóm không được điều trị.

Thông số *B*0 (tức tốc độ trung bình giảm glucose của quần thể) không có ý nghĩa ở đây, bởi vì có sự tương tác giữa hai nhóm và thời gian.

Bây giờ chúng ta thử xét qua phần random effects:

|  |
| --- |
| **Random effects:**  Groups Name Variance Std.Dev. Corr  id (Intercept) 6.878e-01 0.829313  time 5.157e-06 0.002271 1.00  Residual 2.299e-01 0.479518  Number of obs: 76, groups: id, 19 |

Chúng ta có thể so sánh các thông số trong mô hình này với mô hình trước như sau:

|  |  |  |
| --- | --- | --- |
| **Phương sai (variance)** | **Mô hình [4]:** | **Mô hình [5]:** |
| *u*2: độ dao động về tốc độ giảm glucose giữa các chuột | 0.00075 | 0.0000051 |
| *v*2: độ dao động về nồng độ glucose lúc ban đầu | 0.658 | 0.687 |
| *e*2: độ dao động nồng độ glucose *ở mỗi* chuột | 0.248 | 0.230 |

Chúng ta chú ý trong so sánh trên, mô hình [4] có 2 thông số cố định, nhưng mô hình [5] có 4 thông số cố định. Do đó, chúng ta không ngạc nhiên khi mô hình [5] tốt hơn mô hình [4]. Sự “tốt hơn” đó có thể thấy qua *e*2, giảm từ 0.248 xuống 0.230 hay giảm 6%. Một sự giảm thiểu khiêm tốn nhưng vẫn có ý nghĩa.

**Tóm lại**

Các phân tích trên đây cho thấy thuốc có hiệu quả giảm nồng độ glucose trong máu. Mức độ giảm glucose tăng theo thời gian và có ý nghĩa thống kê, dù mức độ ảnh hưởng có phần khiêm tốn.

Trong quá khứ, chúng ta thường sử dụng phương pháp “repeated ANOVA” để phân tích các số liệu được đo nhiều lần ở một đối tượng nghiên cứu. Nhưng phương pháp đó có một số vấn đề về kĩ thuật tính toán và giả định, cho nên không còn được xem là phương pháp chuẩn nữa.

Ngày nay, với R, mô hình ảnh hưởng hỗn hợp được xem là một trong những mô hình phổ biến nhất và thích hợp nhất cho phân tích các thí nghiệm như vừa mô tả. Như trình bày ở phần trên vấn đề chủ yếu trong việc phân tích số liệu không phải là tính toán (vì đã có phần mềm máy tính), nhưng là những suy nghĩ và mô hình cho dữ liệu. Những suy nghĩ này phải xuất phát từ tình hình thực tế và quan trọng hơn hết là kiến thức sinh học. Ngoài ra, cũng cần phải xem xét các thông số cẩn thận để thẩm định tính phù hợp của các thông số cho từng trường hợp cụ thể.

Có lẽ một số bạn cảm thấy nao núng trước những lí giải và công thức trong mô hình phân tích trên, nhưng quả thật mô hình rất đơn giản. Chúng ta muốn có câu trả lời cho câu hỏi: sự giảm nồng độ glucose có khác biệt giữa hai nhóm chuột hay không? Từ câu hỏi đó, chúng ta đặt ra một số mô hình để mô tả dữ liệu. Mô hình mà không có ý nghĩa sinh học chỉ là những mô hình toán học, và theo tôi, những mô hình như thế chẳng có lợi ích gì cho khoa học hay y học lâm sàng.

Khoa học là hình trình đi tìm câu hỏi và câu trả lời. Các phân tích trên đây thật ra còn khá đơn giản. Còn một số câu hỏi khác cũng cần được đặt ra. Chẳng hạn như có nên hoán chuyển số liệu sang log để phân tích hay phân tích theo số liệu gốc (mmol/L)? Nếu có thông tin khác về đặc điểm sinh học của từng chuột, các đặc điểm này có ảnh hưởng gì đến ảnh hưởng của thuốc? Tuy nhiên, hiện nay với một thí nghiệm đơn giản như thế tôi nghĩ rằng mô hình trên cũng thích hợp và đầy đủ. Hi vọng lần sau chúng ta sẽ quay lại với một mô hình ảnh hưởng hỗn hợp khác.