Lâm sàng thống kê 20

Mô hình ảnh hưởng hỗn hợp (mixed-effects model)

Nguyễn Văn Tuấn

Trong bài trước ("Kiểm định outlier) tôi sử dụng số liệu thu thập từ một thí nghiệm trên chuột liên quan đến nồng độ đường trong máu. Bài 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.

Xin nhắc lại thiết kế của thí nghiệm. 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 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). Kết quả của thí nghiệm như sau:

Bảng 1. Nồng độ glucose của nhóm chuột được điều trị và nhóm chứng

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.

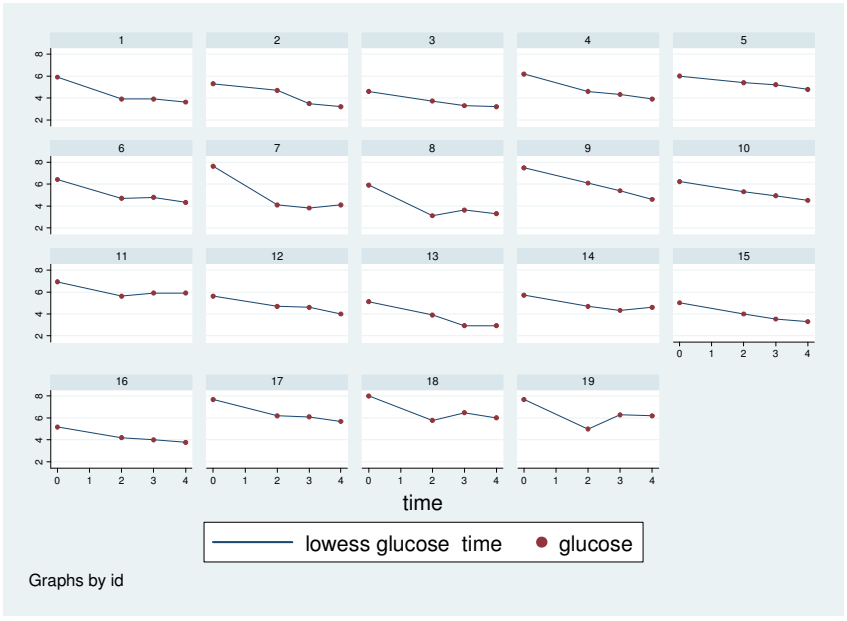
Chúng ta thấy nồng độ đường (sẽ gọi là glucose) ở cả hai nhóm có xu hướng giảm theo thời gian. Câu hỏi chính là thuốc có ảnh hưởng đến sự thuyên giảm glucose hay không? Cụm từ "ảnh hưởng" ở đây có thể hiểu rõ hơn: nó chính là sự khác biệt (difference) về mức độ giảm glucose giữa hai nhóm chuột. Vì thế, 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 (xem chú thích về lệnh Stata):



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 đến 19 (i = 1, 2, 3, …, 19);
- y_i là nồng độ glucose đo lường được cho chuột i;
- a_i là nồng độ glucose trước khi uống thuốc của chuột i;
- b_i là tốc độ giảm glucose của chuột i.
- Tốc độ giảm glucose tùy thuộc vào thời gian và thời gian 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:

$$y_i = a_i + b_i T$$

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 e_i. Bây giờ thì mô hình trên trở thành:

$$y_i = a_i + b_i T + e_i \qquad [1]$$

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:

```
T = c(0, 2, 3, 4)
y = c(5.9, 3.9, 3.9, 3.6)
lm(y ~ T)
```

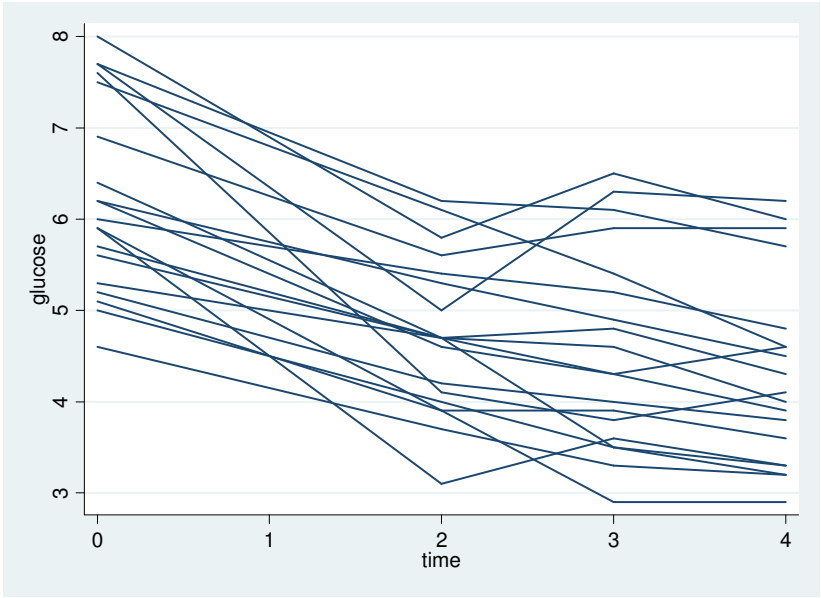
```
Coefficients:
(Intercept)            T
     5.6171      -0.5743
```

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.

```
T = c(0, 2, 3, 4)
y = c(7.5, 6.1, 5.4, 4.6)
lm(y ~ T)
```

```
Coefficients:
(Intercept)            T
     7.52          -0.72
```

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 (xem phần chú thích để biết cách vẽ biểu đồ này bằng Stata, nếu bạn đọc muốn biết):



Biểu đồ 2. Dao động nồng độ glucose cho từng chuột giữa các thời điểm.

Chúng ta thấy rằng thông số a_i, b_i 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 a_i) 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 b_i).

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).

Mô hình ảnh hưởng hỗn hợp (mixed-effects model)

Đây là mô hình "state-of-the-art" trong thống kê học ngày nay. Nói một cách ngắn gọn, mô hình này cho phép chúng ta đánh giá các yếu tố ngẫu nhiên (random effects) và không ngẫu nhiên (fixed effects). Rất khó giải thích hai khái niệm này (và ngay cả sách giáo khoa còn làm cho bạn đọc rối rắm thêm), nhưng tôi sẽ cố gắng giải thích một cách đơn giản nhất và hi vọng rằng bạn đọc sẽ hiểu được. (Nếu không hiểu, các bạn có thể viết cho tôi và nói như thế, để tôi tự nhận mình thất bại trong giải thích và sẽ tìm cách giải thích tiếp). Xin nói trước: chúng ta đang nói chuyện lí thuyết (mô hình), nên cần một chút … tưởng tượng!

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 (e_i):

$$y_i = a_i + b_i T + e_i$$

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
$y_1 = a_1 + b_1 T + e_1$	$y_1 = 5.62 - 0.57T + e_1$
$y_2 = a_2 + b_2 T + e_2$	$y_2 = 5.42 - 0.55\ T + e_2$
$y_3 = a_3 + b_3 T + e_3$	$y_3 = 4.52 - 0.37T + e_3$
…	…
$y_{19} = a_{19} + b_{19} T + e_{19}$	$y_{19} = 7.07 - 0.34T + e_{19}$

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 a_i so với A, chúng ta có:

$$a_i = A + u \qquad [2]$$

Bởi vì tốc độ thay 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 b_i và B là v :

$$b_i = B + v \qquad [3]$$

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

$$y_i = a_i + b_i T + e_i \qquad [1]$$

$$y_i = (A + u) + (B + v)T + e_i$$

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

$$y_i = (A + BT) + (u + vT + e_i) \qquad [4]$$

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, $(u + vT + e_i)$, 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
- e_i 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: $u \sim N(0, u^2)$.

- v tuân theo luật phân phối chuẩn (normal distribution) với số trung bình bằng 0 và phương sai v^2, hay $v \sim N(0, v^2)$.

- e_i cũng theo luật phân phối tương tự: $e_i \sim N(0, e^2)$

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"):

$$E(y_i) = (A + BT) + (0 + 0T + 0)$$

$$E(y_i) = A + BT$$

Mô hình ảnh hưởng của điều trị

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]:

$$a_i = A + u$$

$$b_i = B + v$$

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

$$a_i = A_0 + A_1 \times \text{treatment} + u$$

Ngoài ra, có thể có sự khác biệt về tốc độ giảm glucose giữa hai nhóm (b_i), 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:

$$b_i = B_0 + B_1 \times \text{treatment} + v$$

Do đó, mô hình [1] trở thành:

$$y_i = a_i + b_i T + e_i \qquad [1]$$

$$y_i = (A_0 + A_1 \times \text{treatment} + u) + (B_0 + B_1 \times \text{treatment} + v)T + e_i$$

$$y_i = (A_0 + A_1 \times \text{treatment} + B_0 T + B_1 \times \text{treatment} \times T) + (u + vT + e_i) \quad [5]$$

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.

Ước tính thông số mô hình ảnh hưởng hỗn hợp

Như vậy, trong mô hình trên chúng ta có các thông số *A, B*, u^2, v^2, và e^2. Chúng ta cần máy tính và chương trình **`nlme`** trong R để ước tính các thông số trên. Nhưng trước hết, chúng ta cần sắp xếp số liệu trong **Bảng 1** sao cho có 3 yếu tố (hay 3 cột): cột **treatment** để chỉ chuộc thuộc nhóm nào, cột **T** để chỉ giá trị glucose đo vào thời điểm nào, và cột **id** để nhận dạng chuột. Số liệu đó như sau:

Bảng 2. Kết quả thí nghiệm được sắp xếp theo dòng

treatment	id	T	Y
Test	1	0	3.9
Test	1	2	4.7
Test	1	3	3.7
Test	1	4	4.6
Test	2	0	5.3
Test	2	2	4.7
Test	2	3	3.5
Test	2	4	3.2
vân vân …	…	…	…
Control	10	0	6.2
Control	10	2	5.3
Control	10	3	4.9
Control	10	4	4.5
…	…	…	…
Control	19	0	7.7
Control	19	2	5.0
Control	19	3	6.3
Control	19	4	6.2

Nói cách khác, chúng ta sắp xếp lại mỗi chuột có 4 dòng, nhưng **mỗi** dòng phải được nhận dạng rõ ràng thuộc nhóm nào và thời điểm nào. Chúng ta có 19 x 4 = 76 dòng như trên.

Tôi sẽ sử dụng phần mềm Stata để ước tính các thông số trong mô hình [4] và [5] vừa trình bày trên. Trước hết, tôi cho số liệu vào Stata như sau:

```
id treatment time glucose
1     1      0     5.9
1     1      2     3.9
1     1      3     3.9
1     1      4     3.6
2     1      0     5.3
2     1      2     4.7
2     1      3     3.5
2     1      4     3.2
...
19    2      0     7.7
19    2      2     5.0
19    2      3     6.3
19    2      4     6.2
```

Đối với mô hình [4], các mã Stata sau đây có thể sử dụng để phân tích:

```
xtmixed glucose time || id: time, variance cov(un)
```

Đối chiếu với mô hình [4]:

$$y_i = (A + BT) + (u + vT + e_i)$$

Chú ý rằng mã `glucose time` có nghĩa là ước tính thông số *BT*, tức thông số phản ảnh ảnh hưởng cố định (fixed effects) trong mô hình trên. Các mã phía sau `||` (tức `id: time`) có mục đích phản ảnh sự ảnh hưởng ngẫu nhiên *vT* (random effects). Kết quả của mô hình này như sau:

```
Mixed-effects REML regression                   Number of obs      =        76
Group variable: id                              Number of groups   =        19

                                                Wald chi2(1)       =    150.71
Log restricted-likelihood = -82.425783          Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        time |  -.4849624   .0395033   -12.28   0.000    -.5623874   -.4075374
       _cons |   6.050376   .2183835    27.71   0.000     5.622352    6.4784
------------------------------------------------------------------------------
```

Ước số `time = -0.485` chính là *B* trong phần *BT* của mô hình trên. Ngoài ra, `_cons = 6.05` (`_cons` viết tắt từ chữ constant) chính là ước số của A trong mô hình trên. Nói cách khác, mô hình [4] bây giờ là:

$$y_i = 6.05 - 0.485T$$

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 ước tính ảnh hưởng cố định; chúng ta còn một phần ước tính ngẫu nhiên, và Stata cung cấp các kết quả sau đây:

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                   var(time) |   .0007889   .0023444      2.33e-06     .267106
                  var(_cons) |   .6968951   .2975647      .3017977    1.609233
             cov(time,_cons) |   .0234463   .0329776     -.0411887    .0880813
-----------------------------+------------------------------------------------
               var(Residual) |   .2525322    .047723      .1743646    .3657422
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =    57.41   Prob > chi2 = 0.0000
```

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

- Phương sai $u^2 = 0.00078$ (tức `var(time) = 0.0079`) 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;

- Phương sai $v^2 = 0.697$ (tức `var(_cons) = 0.697`) 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;

- Dòng `var(Residual) = 0.253` (chính là ước số của $e^2 = 0.253$), phản ảnh độ dao động nồng độ glucose *ở mỗi* chuột.

Mô hình trên mang tính mô tả. Chúng ta cần biết tỉ lệ giảm glucose có khác nhau giữa hai nhóm hay không. Để trả lời câu hỏi này, chúng ta cần đến mô hình 5. Các mã Stata sau đây sẽ giúp chúng ta ước tính thông số của mô hình đó:

```
gen treat_time = treatment*time
xtmixed glucose time treatment treat_time || id: time, variance cov(un)
```

Kết quả của mã trên là như sau:

```
Log restricted-likelihood = -81.103522          Prob > chi2       =    0.0000

------------------------------------------------------------------------------
     glucose |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
```

```
-------------+----------------------------------------------------------------
        time |  -.7403175   .1217784    -6.08   0.000    -.9789987   -.5016363
   treatment |   .1530158   .4527252     0.34   0.735    -.7343092    1.040341
  treat_time |   .1673016   .0758314     2.21   0.027     .0186749    .3159283
       _cons |   5.816825   .7270359     8.00   0.000     4.391861     7.24179
-------------------------------------------------------------------------------
```

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

$$y_i = (A_0 + A_1 x \text{treatment} + B_0 T + B_1 x \text{treatment} x T) + (u + vT + e_i)$$

Thông số	Ý nghĩa	Ước số
A_0	Nồng độ trung bình glucose lúc ban đầu của toàn 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:

$$y_i = 5.82 + 0.15 \text{ x treatment} - 0.74 \text{ x time} + 0.17 \text{ x treatment x time}$$

Nên nhớ rằng chúng ta gọi **treatment = 1** chỉ thuốc và **treatment = 2** chỉ nhóm chứng, và **time** có 4 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:

$$y_i = 5.82 + 0.15 \text{ x } 1 - 0.74 \text{ x time} + 0.17 \text{ x } 1 \text{ x time}$$

$$\mathbf{y_i = 5.97 - 0.57 \text{ x time}}$$

và cho nhóm chứng là:

$$y_i = 5.82 + 0.15 \text{ x } 2 - 0.74 \text{ x time} + 0.17 \text{ x } 2 \text{ x time}$$

$$\mathbf{y_i = 6.12 - 0.40 \text{ x time}}$$

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 ảnh hưởng ngẫu nhiên, và phần 2 của Stata cung cấp cho chúng ta kết quả sau đây:

```
------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
id: Unstructured             |
                   var(time) |   5.80e-06   .0001984      4.45e-35    7.57e+23
                  var(_cons) |   .7734249   .3290391      .3359622    1.780516
             cov(time,_cons) |   .0021156   .0359408     -.0683272    .0725583
-----------------------------+------------------------------------------------
               var(Residual) |    .238288   .0454387       .163979     .346271
------------------------------------------------------------------------------
```

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	0.00078	0.0000058
v^2: độ dao động về nồng độ glucose lúc ban đầu	0.697	0.773
e^2: độ dao động nồng độ glucose *ở mỗi* chuột	0.253	0.238

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 0.253 xuống 0.238 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, người 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 sức mạnh của máy tính và phát triển nhanh chóng các hệ thống phần mềm phân tích thống kê, 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.

Chú thích:

Mã Stata sử dụng trong phân tích

Trước hết, chúng ta cần sắp xếp số liệu từ cột sang dòng qua lệnh sau đây. Số liệu trong Stata có hình thức như sau:

treatment	id	glucose0	glucose2	glucose3	glucose4
1	1	5.9	3.9	3.9	3.6
1	2	5.3	4.7	3.5	3.2
1	3	4.6	3.7	3.3	3.2
1	4	6.2	4.6	4.3	3.9
1	5	6.0	5.4	5.2	4.8
1	6	6.4	4.7	4.8	4.3
1	7	7.6	4.1	3.8	4.1
1	8	5.9	3.1	3.6	3.3
1	9	7.5	6.1	5.4	4.6
2	10	6.2	5.3	4.9	4.5
2	11	6.9	5.6	5.9	5.9
2	12	5.6	4.7	4.6	4.0
2	13	5.1	3.9	2.9	2.9
2	14	5.7	4.7	4.3	4.6
2	15	5.0	4.0	3.5	3.3
2	16	5.2	4.2	4.0	3.8
2	17	7.7	6.2	6.1	5.7
2	18	8.0	5.8	6.5	6.0
2	19	7.7	5.0	6.3	6.2

Sử dụng lệnh:

```
reshape long glucose, i(id) j(time)
```

sẽ chuyển số liệu sang dạng:

treatment	id	time	glucose
Test	1	0	3.9
Test	1	2	4.7
Test	1	3	3.7
Test	1	4	4.6
Test	2	0	5.3

Test	2	2	4.7
Test	2	3	3.5
Test	2	4	3.2
…	…	…	…

Lệnh để vẽ biểu đồ 1:

```
graph twoway (lowess glucose time) (scatter glucose time) by(id)
```

và biểu đồ 2:

```
graph twoway (scatter glucose time, msym(i) connect(L))
```

Phân tích bằng 3 mô hình chính:

- Mô hình cơ bản: không có ảnh hưởng của bất cứ yếu tố nào:

  xtmixed glucose time || id: time, variance cov(un)

  Phương sai $e^2 = 1.3874$

- Mô hình 4: ảnh hưởng của time nhưng không có treatment:

  $$y_i = (A + BT) + (u + vT + e_i)$$

  xtmixed glucose time || id: time, variance cov(un)

  Phương sai $e^2 = 0.252$

- Mô hình 5: ảnh hưởng của time và treatment:

  $$y_i = (A_0 + A_1 x \text{treatment} + B_0 T + B_1 x \text{treatment} x T) + (u + vT + e_i)$$

  ```
  gen treat_time = treatment*time
  xtmixed glucose time treatment treat_time || id: time,
          variance cov(un)
  ```

  Phương sai $e^2 = 0.238$

Các phân tích trên cũng có thể thực hiện qua ngôn ngữ R:

```
library(Design)
library(lattice)
library(nlme)

# 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)
T = rep(c(0,2,3,4), 19)
treatment = rep(1:2, c(9*4, 10*4))
glucose = data.frame(id,treatment,T,y)
```

 # vẽ biểu đồ 1

```
xyplot(y ~ T | id, as.table=T, xlab="Time", ylab="Glucose")
```

 # vẽ biểu đồ 2

```
fit <- by(glucose, id,
       function(data) fitted.values(lm(y ~ T, data=data)))
fit <- unlist(fit)
names(fit) <- NULL

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

xyplot(y ~ T| id, glucose,
  panel = function(x, y){
    panel.xyplot(x, y)
    panel.lmline(x, y)
  }, ylim=c(0, 8), as.table=T)

# Phân tích mô hình [4]

fit1 = lme(y ~ T, data=glucose, random=~1+T | id,
control=lmeControl(opt="optim"))
summary(fit1)
```

Nếu bạn nào sử dụng SAS, các lệnh sau đây sẽ có ích:

```
data glucose;
input id treatment T y;
logy = log(y);
cards;
 1          1 0 5.9
 1          1 2 3.9
 1          1 3 3.9
 1          1 4 3.6
 2          1 0 5.3
 2          1 2 4.7
 2          1 3 3.5
 2          1 4 3.2
 3          1 0 4.6
 3          1 2 3.7
 3          1 3 3.3
 3          1 4 3.2
 4          1 0 6.2
 4          1 2 4.6
 4          1 3 4.3
 4          1 4 3.9
 5          1 0 6.0
 5          1 2 5.4
 5          1 3 5.2
 5          1 4 4.8
 6          1 0 6.4
 6          1 2 4.7
 6          1 3 4.8
 6          1 4 4.3
 7          1 0 7.6
 7          1 2 4.1
 7          1 3 3.8
 7          1 4 4.1
 8          1 0 5.9
 8          1 2 3.1
 8          1 3 3.6
 8          1 4 3.3
 9          1 0 7.5
 9          1 2 6.1
 9          1 3 5.4
 9          1 4 4.6
10          2 0 6.2
10          2 2 5.3
10          2 3 4.9
10          2 4 4.5
11          2 0 6.9
11          2 2 5.6
11          2 3 5.9
11          2 4 5.9
12          2 0 5.6
12          2 2 4.7
12          2 3 4.6
12          2 4 4.0
13          2 0 5.1
13          2 2 3.9
```

```
 13          2 3 2.9
 13          2 4 2.9
 14          2 0 5.7
 14          2 2 4.7
 14          2 3 4.3
 14          2 4 4.6
 15          2 0 5.0
 15          2 2 4.0
 15          2 3 3.5
 15          2 4 3.3
 16          2 0 5.2
 16          2 2 4.2
 16          2 3 4.0
 16          2 4 3.8
 17          2 0 7.7
 17          2 2 6.2
 17          2 3 6.1
 17          2 4 5.7
 18          2 0 8.0
 18          2 2 5.8
 18          2 3 6.5
 18          2 4 6.0
 19          2 0 7.7
 19          2 2 5.0
 19          2 3 6.3
 19          2 4 6.2
 ;
 run;

proc mixed noclprint covtest;
 class id;
 model y = T / solution ddfm=bw notest;
 random intercept T / subject=id type=un;
title "Model 4";
run;

proc mixed noclprint covtest;
 class id treatment;
 model y = T treatment T*treatment / solution ddfm=bw notest;
 random intercept T / subject=id type=un;
title "Model 5";
run;
```