Lâm sàng thống kê

Cẩn thận với "confounder"

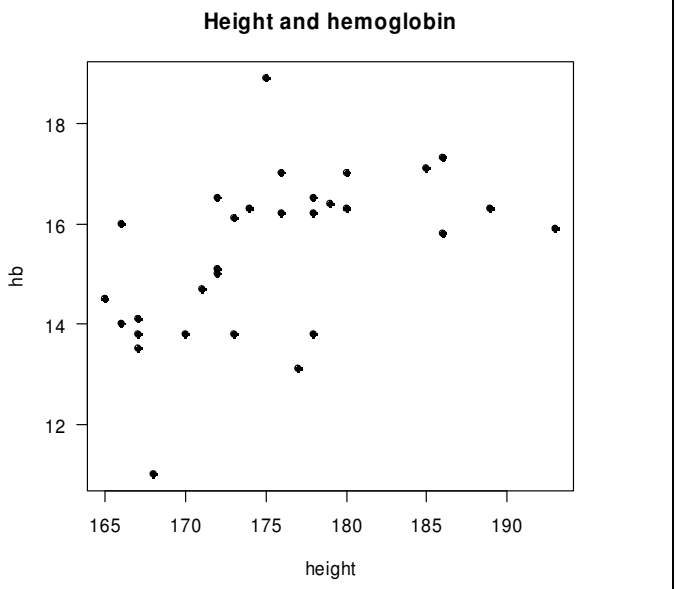
Một trong những đe dọa đến việc diễn giải các nghiên cứu dịch tễ học và lâm sàng là vấn đề *confounder* (đọc là con-phao-đơ). Thuật ngữ confounder rất khó dịch sang tiếng Việt, nhưng định nghĩa của nó có thể tóm tắt như sau: confounder là một biến trung gian có ảnh hưởng đến biến độc lập và biến phụ thuộc. Chẳng hạn như chúng ta biết rằng hormone testosterone có ảnh hưởng đến nguy cơ gãy xương ở đàn ông; ngoài ra, nguy cơ gãy xương cũng tăng dần theo độ tuổi; nhưng testosterone cũng suy giảm theo độ tuổi. Nếu một nghiên cứu báo cáo rằng mối liên hệ giữa testosterone và gãy xương mà không đề cập đến độ tuổi, chúng ta có thể nghi ngờ rằng mối liên hệ quan sát được giữa testosterone và gãy xương có phải bị ảnh hưởng bởi độ tuổi. Do đó, khi phân tích các nghiên cứu y học, chúng ta phải hết sức cẩn thận với ảnh hưởng của confounder.

Để minh họa cho ảnh hưởng của confounder, chúng ta hãy xem ví dụ sau đây: nồng độ hemoglobin (g/dL) và chiều cao (cm) được đo lường ở 31 sinh viên y khoa tuổi từ 19 đến 30, và kết quả như sau:

Sinh viên	Chiều cao (cm)	Hb (g/dL)
1	168	11.0
2	165	14.5
3	166	14.0
4	167	13.5
5	167	13.8
6	167	14.1
7	170	13.8
8	171	14.7
9	172	15.0
10	172	15.1
11	173	13.8
12	177	13.1
13	178	13.8
14	166	16.0
15	172	16.5
16	175	18.9
17	173	16.1
18	174	16.3
19	176	16.2
20	176	17.0
21	178	16.5
22	179	16.4
23	178	16.2
24	180	16.3

25	180	17.0
26	185	17.1
27	186	17.3
28	186	15.8
29	189	16.3
30	193	15.9

Giả sử chúng ta muốn tìm hiểu mối tương quan giữa chiều cao và Hb. Biểu đồ sau đây sẽ cho thấy mối tương quan đó:



Biểu đồ 1: Chiều cao và hemoglobin ở 30 sinh viên

Hệ số tương quan (coefficient of correlation) của mối tương quan này là: $r = 0.51$ với trị số $p = 0.0038$, tức có ý nghĩa thống kê. Nói cách khác, sinh viên với vóc người cao có độ hemoglobin cao hơn sinh viên với vóc người thấp.

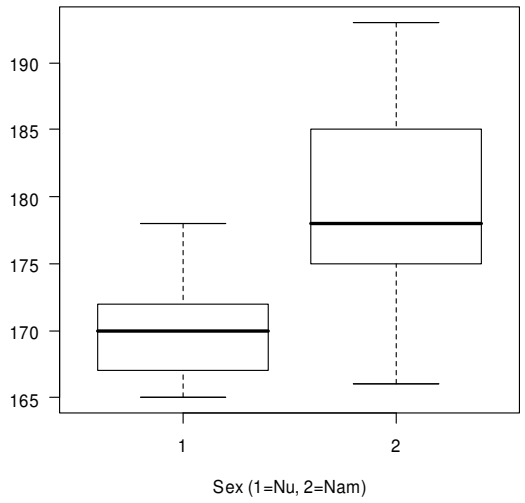
Tuy nhiên, kết luận trên có thể sai! Chúng ta biết rằng nam thường có vóc người cao hơn nữ, cho nên rất có thể hemoglobin chẳng có liên quan gì đến chiều cao, mà chỉ đơn thuần liên quan đến giới tính. Giả sử như chúng ta tiếp tục thu thập số liệu và ghi nhận giới tính của từng sinh viên, và biết rằng có 13 sinh viên nữ và 17 nam. Số liệu mới như sau:

Sinh viên	Giới [a]	Chiều cao (cm)	Hb (g/dL)
1	1	168	11.0
2	1	165	14.5
3	1	166	14.0
4	1	167	13.5

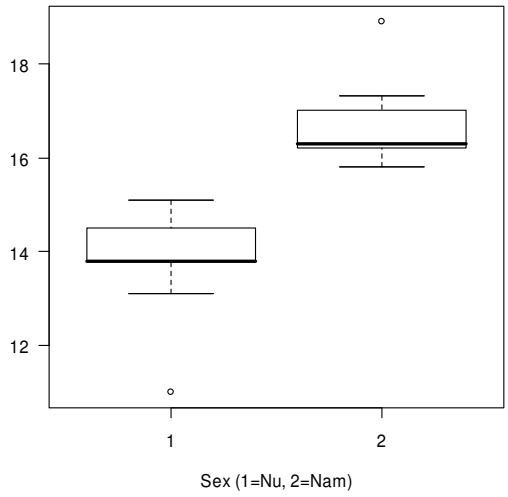
5	1	167	13.8
6	1	167	14.1
7	1	170	13.8
8	1	171	14.7
9	1	172	15.0
10	1	172	15.1
11	1	173	13.8
12	1	177	13.1
13	1	178	13.8
14	2	166	16.0
15	2	172	16.5
16	2	175	18.9
17	2	173	16.1
18	2	174	16.3
19	2	176	16.2
20	2	176	17.0
21	2	178	16.5
22	2	179	16.4
23	2	178	16.2
24	2	180	16.3
25	2	180	17.0
26	2	185	17.1
27	2	186	17.3
28	2	186	15.8
29	2	189	16.3
30	2	193	15.9

[a]Chú thích: Giới =1 có nghĩa là nữ và 2 có nghĩa là nam.

Bây giờ chúng ta thử làm một phân tích mô tả bằng biểu đồ hộp (box plot) theo giới như sau:

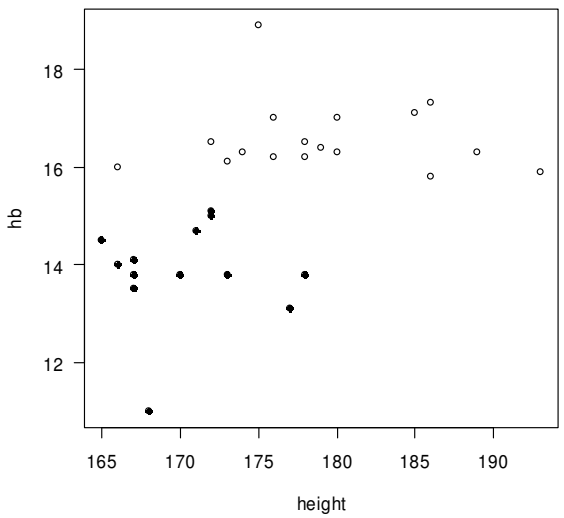


Biểu đồ 2a: Phân tích mô tả chiều cao theo từng giới. Chiều cao trung bình của nữ là 170.2 cm (SD: 4.1), và nam 179.2 cm (SD: 6.8). SD là độ lệch chuẩn.



Biểu đồ 2b: Phân tích mô tả hemoglobin theo từng giới. Hb trung bình của nữ là 13.9 cm (SD: 1.0), và nam 16.6 (SD: 0.7).

Rõ ràng qua phân tích đơn giản trên, chúng ta thấy nam có chiều cao cao hơn nữ, và cũng có độ hemoglobin cao hơn nữ. Như vậy, cách tốt nhất để biết thật sự có liên hệ giữa chiều cao và hemoglobin hay không, chúng ta phải phân tích theo từng giới. Biểu đồ 3 dưới đây cho thấy, trong từng giới, không có mối liên hệ nào giữa chiều cao và hemoglobin.



Biểu đồ 3: Mối liên hệ giữa chiều cao và hemoglobin theo giới.

Điểm chấm tròn đen thể hiện số liệu của nhóm nữ, điểm chấm tròn trắng thể hiện số liệu của nhóm nam.

Thật ra, chúng ta có thể áp dụng mô hình phân tích hiệp biến (analysis of covariance hay ANCOVA) để biết trong hai yếu tố (chiều cao và giới), yếu tố nào có ảnh hưởng đến Hb. Mô hình ANCOVA có thể viết đơn giản như sau:

$$Hb = a + b \times Height + c \times Sex$$

Trong đó, a, b và c là những thông số cần ước tính từ số liệu. Sử dụng ngôn ngữ R, ước số của a, b, và c là như sau:

Thông số	Hệ số ± SE	Trị số P
a (tức intercept)	11.81 ± 4.72	0.019
b (ảnh hưởng của chiều cao)	-0.004 ± 0.029	0.888
c (ảnh hưởng của giới)	2.75 ± 0.42	<0.0001

SE: sai số chuẩn (standard error)

Kết quả phân tích trên xác định chiều cao không có ảnh hưởng đến Hb, nhưng giới có ảnh hưởng. Phân tích trên cho thấy **sau khi điều chỉnh cho chiều cao**, Hb ở nam giới, tính trung bình, cao hơn nữ giới khoảng 2.75 g/dL với sai số chuẩn là 0.42 g/dL.

Kinh nghiệm quan trọng qua ví dụ này là nếu phân tích mà không xem xét đến ảnh hưởng của các yếu tố confounder rất dễ đi đến kết luận sai.

Chú thích kĩ thuật

Các mã R sau đây đã được sử dụng cho phân tích vừa trình bày.

```
# nhập số liệu chiều cao, hb và giới

height <- c(168,165,166,167,167,167,170,171,172,172,
            173,177,178,166,172,175,173,174,176,176,
            178,179,178,180,180,185,186,186,189,193)

hb    <- c(11.0, 14.5, 14.0, 13.5, 13.8, 14.1, 13.8, 14.7, 15.0, 15.1,
           13.8, 13.1, 13.8, 16.0, 16.5, 18.9, 16.1, 16.3, 16.2, 17.0,
           16.5, 16.4, 16.2, 16.3, 17.0, 17.1, 17.3, 15.8, 16.3, 15.9)

sex    <- c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
            2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
            2, 2, 2, 2)
```

```
# vẽ biểu đồ 1

plot(hb ~ height, main="Height and hemoglobin", pch=16)

# vẽ biểu đồ 2a và 2b

boxplot(height ~ sex, xlab="Sex (1=Nu, 2=Nam)")

boxplot(hb ~ sex, xlab="Sex (1=Nu, 2=Nam)")

# vẽ biểu đồ 3: chú ý 19 và 21 là mã số cho điểm chấm của biểu đồ

plot(hb ~ height, pch=ifelse(sex==1, 19, 21))

# Phân tích ancova

ancova <- lm(hb ~ height+sex)
summary(ancova)
```

 # kết quả sau khi chạy 2 lệnh trên:

```
Call:
lm(formula = hb ~ height + sex)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8707 -0.3797 -0.0781  0.4228  2.3063

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11.81196    4.71948    2.50    0.019 *
height      -0.00413    0.02907   -0.14    0.888
sex          2.75184    0.42005    6.55    5e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.895 on 27 degrees of freedom
Multiple R-Squared: 0.715,      Adjusted R-squared: 0.694
F-statistic: 33.9 on 2 and 27 DF,  p-value: 4.36e-08
```