(khái niệm hồi quy multilevel)

Bắt đầu với bộ dữ liệu khảo sát thời gian đọc sách của học sinh

|  |
| --- |
| stem <- read.csv(file="/bayesvl/STEM\_data.csv",header=T)  str(stem) |

|  |
| --- |
| data.frame': 4630 obs. of 5 variables:  $ school : int 4 4 4 4 4 4 4 4 4 4 ...  $ gradeid: int 8 8 8 8 8 8 8 8 8 8 ...  $ sex : int 2 2 1 2 1 2 1 2 2 2 ...  $ aps45id: int 3 2 2 3 2 3 2 2 3 5 ...  $ timesci: int 2 1 1 2 2 1 1 1 2 1 ... |

Bộ dữ liệu này bao gồm:

* School: mã trường học (có 16 trường trong khảo sát)
* Gradeid: mã khối học (6, 7, 8, 9)
* Sex: giới tính học sinh
* Asp45: điểm học lực trung bình của học sinh
* Timesci: thời gian đọc sách xã hội

Giả sử chúng ta đang định tìm mối liên hệ giữa học lực của học sinh và thời gian học sinh đọc sách. Với phương trình hồi quy tuyến tính ta có thể dựng mô hình tương quan giữa biến $$y\_{asp45id} ~ x\_{timesci}$$ (xem chương hồi quy tuyến tính) dạng:

$$y\_{asp45} = \alpha + \beta x\_{timesci}$$

Tuy nhiên ở đây dữ liệu khảo sát của chúng ta có khối lớp học (6, 7, 8, 9). Tất nhiên chúng ta có thể đưa biến khối lớp vào phương trình như bài toán hồi quy đa biến (xem chương hồi quy tuyến tính).

$$y\_{asp45} = \alpha + \beta\_{timesci} x\_{timesci} + \beta\_{grade} x\_{grade}$$

Tuy nhiên cũng có thể thấy việc đưa yếu tố khối học vào đây có thể chưa hoàn toàn hợp lý. Nếu tách riêng khối học ta có thể có các quan hệ tách biệt giữa thời gian đọc sách và học lực học sinh của từng khối. Như thế nếu phân tích theo hướng này ta có thể có sự so sánh tương quan giữa các khối học, mặt khác việc giả định yếu tố khối học sẽ ảnh hưởng tới tuyến tính với học lực học sinh là chưa hợp lý.

Như vậy, ta có thể tách dữ liệu ra thành 4 khối dữ liệu cho 4 khối lớp và tiến hành phân tích độc lập trên frequentist như sau:

|  |
| --- |
| stem6 = stem[stem$gradeid ==6,]  stem7 = stem[stem$gradeid ==7,]  stem8 = stem[stem$gradeid ==8,]  stem9 = stem[stem$gradeid ==9,]  fit6 <- lm(aps45 ~ timesci, data = stem6)  fit7 <- lm(aps45id ~ timesci, data = stem7)  fit8 <- lm(aps45id ~ timesci, data = stem8)  fit9 <- lm(aps45id ~ timesci, data = stem9)  summary(fit6) # Report the results  summary(fit7) # Report the results  summary(fit8) # Report the results  summary(fit9) # Report the results |



Như vậy, ta thấy cách hồi quy trên, ta phân hoạch dữ liệu ra 4 nhóm dữ liệu độc lập, kết quả sẽ cho ta 4 đường hồi quy khác nhau của 4 nhóm. Nếu viết lại công thức hồi quy cho p nhóm dữ liệu ta sẽ có dạng (mixed mode):

$$y\_{ki} = \alpha\_k + \beta\_k x\_{ki}$$

Trong đó k là số nhóm dữ liệu k = 1..p.

Nếu

3 loại mô hình tuyến tính multilevel

* Varying slope
* Varying intercept
* Mixed

Pool, non-pool và partial-pool

Mô hình bài STEM

Mô hình bài DKAP