Bài giảng 1: Giới Thiệu Mô Hình Tuyến Tính Tổng Quát

TS. Tô Đức Khánh

Khoa Toán-Tin Học, Trường Đại Học Khoa Học Tự Nhiên
Đại Học Quốc Gia Tp. HCM

Mô Hình Thống Kê Tuyến Tính Nâng Cao

Mục lục

1 Một số thông tin của khóa học

2 Nhắc lại một số kiến thức cơ bản

3 Giới thiệu mô hình tuyến tính tổng quát
- Ví dụ
- Mô hình thống kê

Thông tin của môn học

Giảng viên:

TS. Tô Đức Khánh (tôi)
Bộ môn Xác suất-Thống kê, khoa Toán-Tin học
Trường Đại Học Khoa Học Tự Nhiên, Tp. HCM
e-mail: tdkhanh@hcmus.edu.vn
Phòng F.012, Trường Đại Học Khoa Học Tự Nhiên, 227 Nguyễn Văn Cừ, Quận 5, Tp. Hồ Chí Minh.

Thông tin của môn học

Nội dung của môn học trong 15 tuần bao gồm:

- Mô hình hồi quy cho biến nhị thức;
- Mô hình hồi quy cho biến cho biến phân loại;
- Mô hình hồi quy cho biến số đếm;
- Lý thuyết cơ sở cho mô hình tuyến tính tổng quát.

Thực hành trên phần mềm R/Rstudio, với các thư viện: `tidyverse`, `ggplot2`, `nnet`, `pROC`.

Thông tin của môn học

Số tín chỉ

	Tín chỉ	Số giờ học
Lý thuyết	3	45
Thực hành/Bài tập	1	30
Tổng cộng	4	75

Lịch học

- Tất cả các bài giảng lý thuyết, thực hành và bài tập được tổ chức tại cở sở 2 của trường Đại Học Khoa Học Tự Nhiên, Linh Trung, Thủ Đức.
- Bắt đầu từ 27/02/2024 tới 18/06/2024. Ngoại trừ ngày 23/04/2024 (thi giữa kỳ), 30/04/2024.
- Buổi lý thuyết: thứ 3 (hàng tuần) từ tiết 6 tới 8 (12:30 - 15h00) tại phòng D207
- Buổi thực hành R: thứ 3 (hàng tuần) từ tiết 7.5 tới 10 (15:00 - 17:00) tại phòng C205, bắt đầu từ tuần thứ 3.

Xem chi tiết lịch học tại trang quản lý môn học.

Thông tin của môn học

Cấu trúc điểm

Môn học có cấu trúc điểm như sau:

- Điểm thực hành/bài tập: 20%
- Điểm thi giữa kỳ: 20%
- Điểm thi cuối kỳ: 60%

Hình thức thi

- Điểm thi giữa kỳ: viết tự luận trong thời gian 90 phút.
- Điểm thi cuối kỳ: làm đề tài môn học theo nhóm.
  (Danh sách đề tài sẽ được cung cấp vào tuần thứ 12 của môn học.)

Chú ý: Sinh viên cần phải tham gia ít nhất 80% số tiết học.

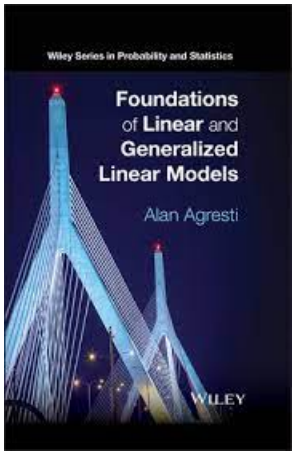
Yêu cầu kiến thức chuẩn bị

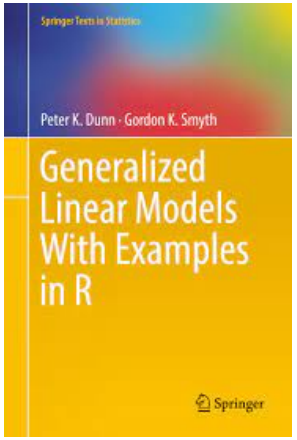
Trước khi tham gia khóa học này, sinh viên cần nắm được các khái niệm dưới đây:

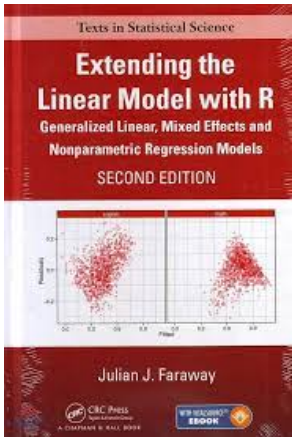
1. các kiến thức cơ bản về thống kê mô tả: đồ thị tần số, biểu đồ mật độ, biểu đồ hộp (boxplot), giá trị trung bình, phương sai, ...
2. các kiến thức cơ bản về hàm mật độ xác suất, hội tụ xác suất, phân phối của biến ngẫu nhiên, ...
3. các kiến thức cơ bản về kiểm định giả thuyết thống kê
4. ngoài ra, các kiến thức về giải tích hàm (khả vi, liên tục, khả tích) hay đại số tuyến tính cũng cần thiết.

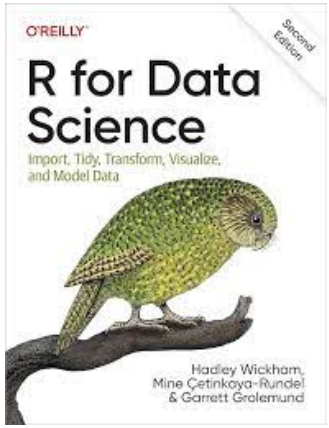
Tài liệu học tập và tài liệu tham khảo

Slide bài giảng được cập nhật tại trang của môn học:
`https://courses.hcmus.edu.vn/course/view.php?id=5578`









1 Một số thông tin của khóa học

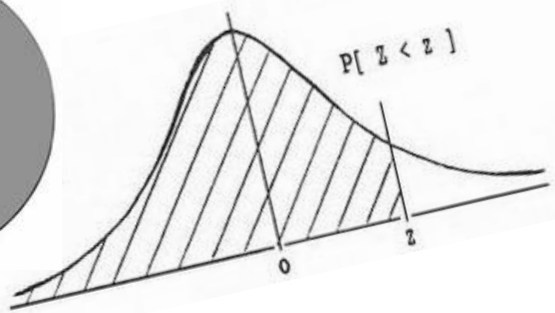
2 Nhắc lại một số kiến thức cơ bản

3 Giới thiệu mô hình tuyến tính tổng quát
- Ví dụ
- Mô hình thống kê

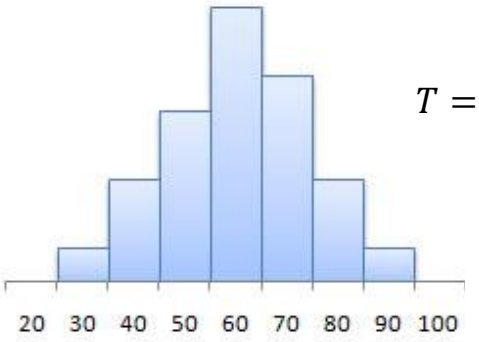
Thống kê?

$$Z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$$

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$



p-value $= \Pr(|T| \geq t)$



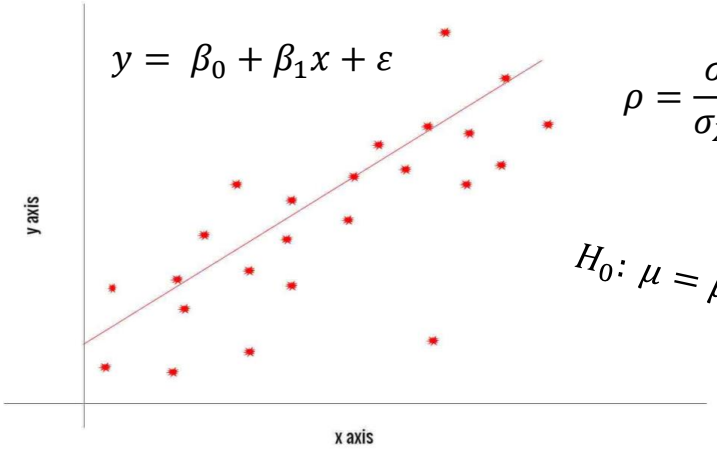
$$T = \frac{\bar{x} - \mu_0}{S_x/\sqrt{n}} \xrightarrow{d} t_{n-1}$$

$$\sqrt{np(1-p)}$$

$$\left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\alpha/2}\right)$$

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$y = \beta_0 + \beta_1 x + \varepsilon$$



$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

$$H_0: \mu = \mu_0$$

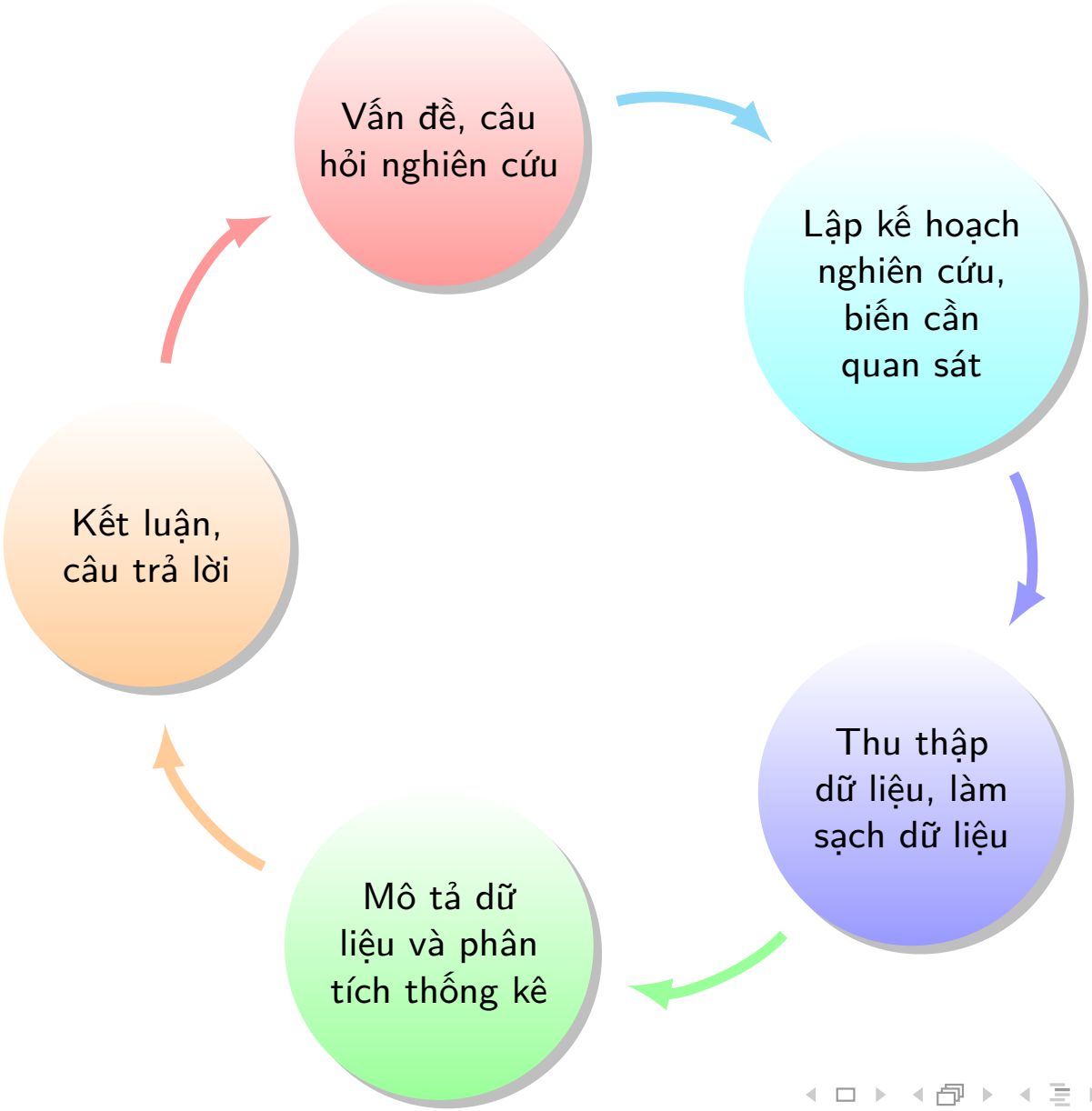
Thống kê?

Thống kê (**Statistics**) là một ngành khoa học của việc:

- **thu thập** dữ liệu
- **biểu diễn** dữ liệu
- **phân tích** dữ liệu
- **giải thích** kết quả

nhằm tìm ra các bằng chứng để cung cố hoặc bác bỏ các giả thuyết khoa học hoặc câu hỏi nghiên cứu.

Thống kê?


Vấn đề, câu hỏi nghiên cứu
Lập kế hoạch nghiên cứu, biến cần quan sát
Thu thập dữ liệu, làm sạch dữ liệu
Mô tả dữ liệu và phân tích thống kê
Kết luận, câu trả lời

Thế nào là một câu hỏi nghiên cứu?

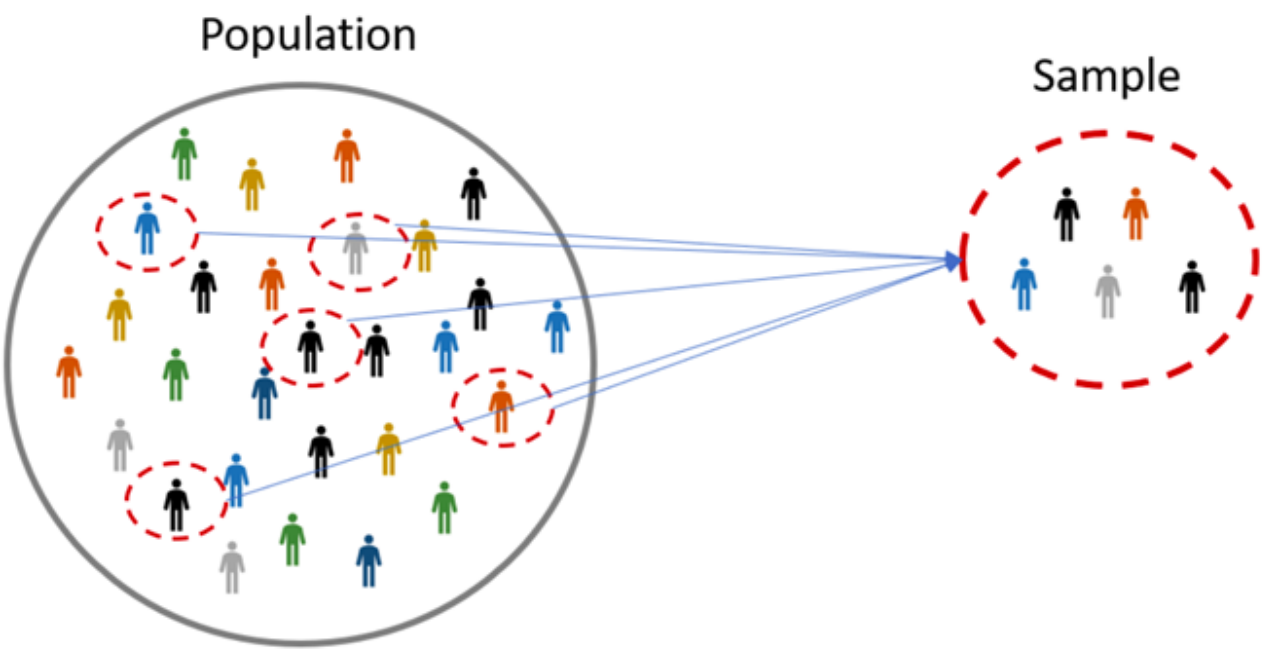
Một câu hỏi nghiên cứu là "một câu hỏi được đặt ra bởi một nhà nghiên cứu hoặc một dự án nghiên cứu nhằm tìm ra câu trả lời đề giải thích cho các quan sát hoặc các giả thuyết." Ví dụ như:

- Khả năng bị ung thư phổi có bị ảnh hưởng bởi thói quen thuốc lá?
- Phương pháp giảng dạy mới có hiệu quả hơn phương pháp dạy cũ?
- Tại sao mức độ hoạt động MAO (monoamine oxidase) khác nhau giữa các nhóm bệnh nhân tâm thần phân liệt khác nhau?
- Làm thế nào để phân loại chất lượng các loại rượu vang dựa trên các tính chất của chúng?
- Nhóm người nào có nguy cơ cao phải đối mặt với các cơn đau tim bất thường?

Dữ liệu thống kê là gì?

- Một dữ liệu thống kê là một bộ **mẫu ngẫu nhiên** của các **đối tượng nghiên cứu**, nó bao gồm các thông tin
- Một bộ dữ liệu thống kê có thể được thu thập thông qua việc quan sát và thu mẫu thực tế (**observations**), hoặc thu mẫu các đối tượng trong phòng thí nghiệm **experiments**.
- Một quần thể (**population**) là bộ dữ liệu lớn bao gồm tất cả đối tượng nghiên cứu có thể được bao hàm trong câu hỏi nghiên cứu.
- Chú ý: một đối tượng nghiên cứu được coi là một đơn vị thống kê (**statistical unit**).

Dữ liệu thống kê là gì?



Dữ liệu thống kê là gì?

Characteristics of a iris flower

	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa

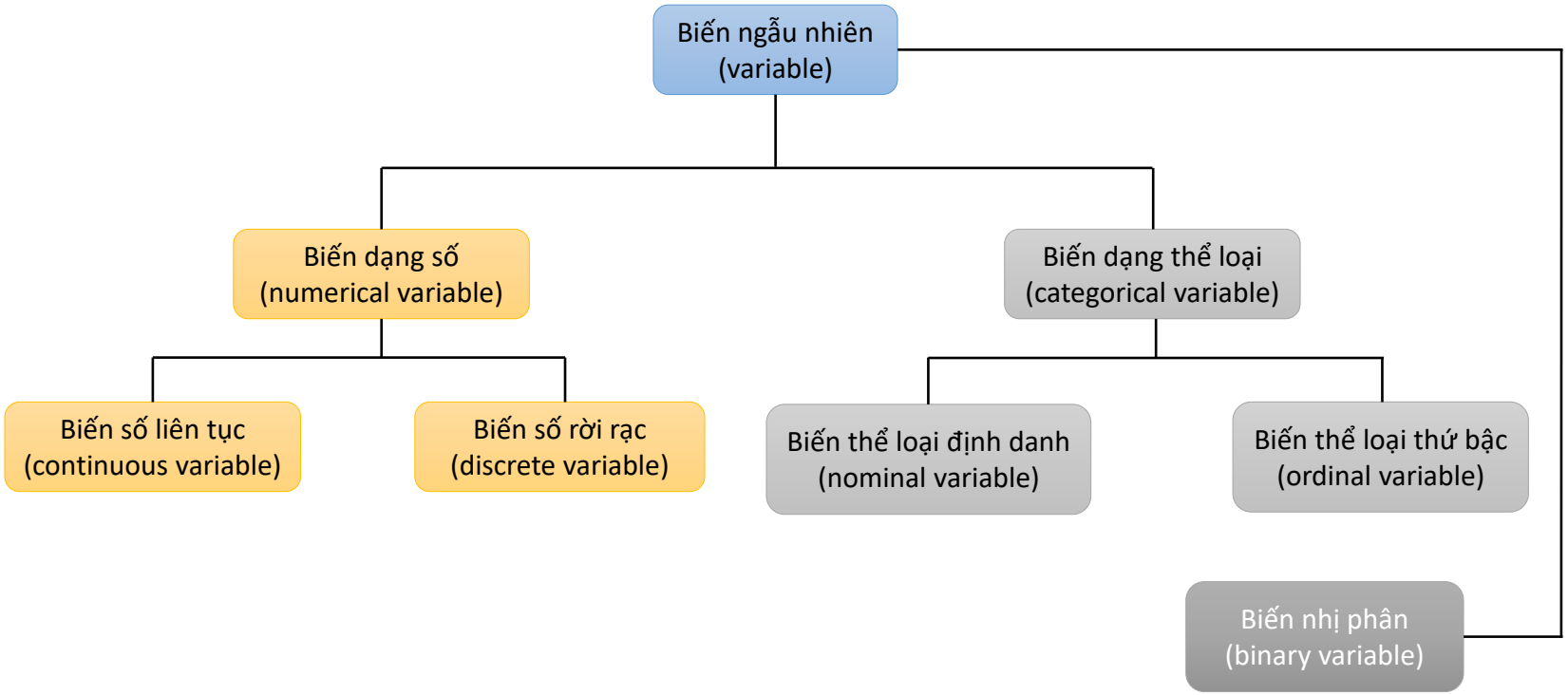
Number of iris flowers

- Câu hỏi nghiên cứu: Làm thế nào dự đoán được 3 loài hoa diên vĩ (iris): setosa, versicolor và virginica, dựa vào các số đo của chiều dài đài hoa (sepal), chiều rông đài hoa, chiều dài cánh hoa (petal) và chiều rộng cánh hoa?
- Đối tượng nghiên cứu: hoa diên vĩ thuộc 3 loài: setosa, versicolor, và virginica.
- Quần thể: Tất cả các hoa diên vĩ trên thế giới, với khoảng 310 loài đã được ghi nhận.
- Dữ liệu: một mẫu gồm 150 hoa diên vĩ của 3 loài: setosa, versicolor, and virginica.

Biến ngẫu nhiên là gì?

Biến ngẫu nhiên

Trong thống kê, một tính chất của đối tượng nghiên cứu thường được gán bởi một biến ngẫu nhiên (*random variable*).



Biến ngẫu nhiên là gì?

Biến ngẫu nhiên dạng số (*numerical variable* hay *quantitative variable*) là biến ngẫu nhiên với giá trị là dạng:

- liên tục (*continuous variable*);
- rời rạc hay số đếm (*discrete variable*),

ghi lại số lượng của một đối đối tượng nào đó.

Biến ngẫu nhiên dạng phân loại (*categorical variable* hay *qualitative variable*) là biến ngẫu nhiên có giá trị là các thể loại khác nhau của người hoặc vật. Thông thường ta có hai dạng phổ biến sau:

- biến định danh (*nominal variable*)
- biến thứ bậc (*ordinal variable*)

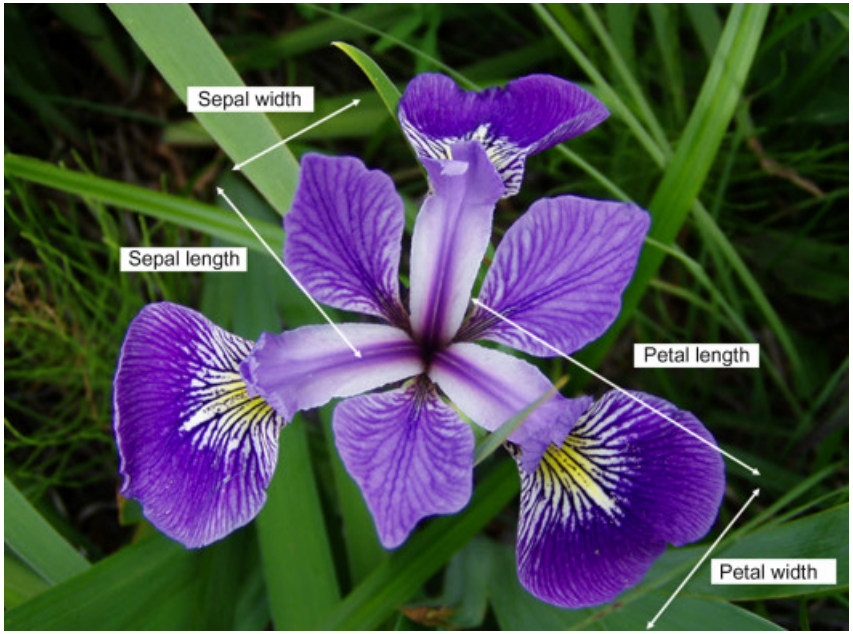
Đặc biệt, đối với biến định danh, ta có:

- biến nhị định danh hay biến nhị phân (*binary variable*), khi biến có 2 giá trị;
- biến đa định danh hay biến đa thức (*nominal multinomial variable* hoặc *multinomial variable*).

Biến ngẫu nhiên là gì?

Ví dụ, đối với dữ liệu hoa diên vĩ, ta có:

- 04 biến dạng số liên tục: chiều dài đài hoa, chiều rộng đài hoa, chiều dài cánh hoa, chiều rộng cánh hoa;
- 01 biến dạng phân loại định danh: loài hoa.



Biến ngẫu nhiên là gì?

Câu hỏi: phân loại dạng biến ngẫu nhiên

- Trạng thái hút thuốc của một người.
- Giới tính sinh học của một người.
- Chiều cao của một người.
- Xếp loại tốt nghiệp của một học sinh.
- Các nhóm tính trạng bệnh của bệnh ung thư phổi.

Kỹ thuật phân tích thống kê

Thống kê mô tả (*descriptive statistics*) là tổ hợp các kỹ thuật để tổ chức, mô tả, tổng hợp và trình bày dữ liệu hoặc thông tin của dữ liệu theo cách thuận tiện cho việc hiểu và sử dụng.

Thống kê suy luận (*inferential statistics*) là tổ hợp các phương pháp thống kê dùng để đưa ra các kết luận và suy luận về tính chất của quần thể đối tượng dựa trên thông tin của dữ liệu.

Mô hình thống kê (*statistical model*) là một dạng phân tích thống kê dựa trên một mô hình toán cụ thể cho dữ liệu. Mô hình thống kê thường được dùng để mô tả sự biến động của một (hoặc nhiều) biến dựa trên giá trị của một hay nhiều biến khác. Khi phân tích mô hình thống kê, ta thường phải sửa dụng một số phương pháp thống kê mô tả và thống kê suy luận.

1 Một số thông tin của khóa học

2 Nhắc lại một số kiến thức cơ bản

3 Giới thiệu mô hình tuyến tính tổng quát
- Ví dụ
- Mô hình thống kê

Ví dụ 1 - Dung tích phổi

Ta xét bộ dữ liệu về dung tích phổi được thu thập trên 654 người trẻ tuổi tại East Boston, Mỹ, như sau:

id	Age	FEV	Ht	Gender	Smoke
1	3	1.072	46	F	0
2	4	0.839	48	F	0
3	4	1.102	48	F	0
4	4	1.389	48	F	0
5	4	1.577	49	F	0
6	4	1.418	49	F	0
⋮	⋮	⋮	⋮	⋮	⋮
654	18	4.404	70.5	M	1

trong đó:

- Age, Ht, Gender và Smoke lần lượt là các biến đo độ tuổi, chiều cao (inch), giới tính và tình trạng hút thuốc;
- FEV (forced expiratory volume) là thể tích (đơn vị: lít) thở ra cưỡng bức được dùng để chỉ số đo dung tích phổi.

Ví dụ 1 - Dung tích phổi

id	Age	FEV	Ht	Gender	Smoke
1	3	1.072	46	F	0
2	4	0.839	48	F	0
3	4	1.102	48	F	0
4	4	1.389	48	F	0
5	4	1.577	49	F	0
6	4	1.418	49	F	0
⋮	⋮	⋮	⋮	⋮	⋮
654	18	4.404	70.5	M	1

Trong các biến trên:

- Age, FEV và Ht là các biến ngẫu nhiên liên tục;
- Gender và Smoke là các biến ngẫu nhiên định danh,

tại sao?

Ví dụ 1 - Dung tích phổi

Câu hỏi nghiên cứu

Liệu rằng dung tích phổi có thay đổi theo độ tuổi, chiều cao, giới tính hay trạng thái hút thuốc?

Ví dụ 1 - Dung tích phổi

Câu hỏi nghiên cứu

Liệu rằng dung tích phổi có thay đổi theo độ tuổi, chiều cao, giới tính hay trạng thái hút thuốc?

Để trả lời câu hỏi nghiên cứu này, trước hết ta sẽ dùng biểu đồ để biểu diễn mối liên hệ giữa khả năng dung tích phổi (FEV) và các biến: Age, Ht, Gender, Smoke.

Ví dụ 1 - Dung tích phổi

Câu hỏi nghiên cứu

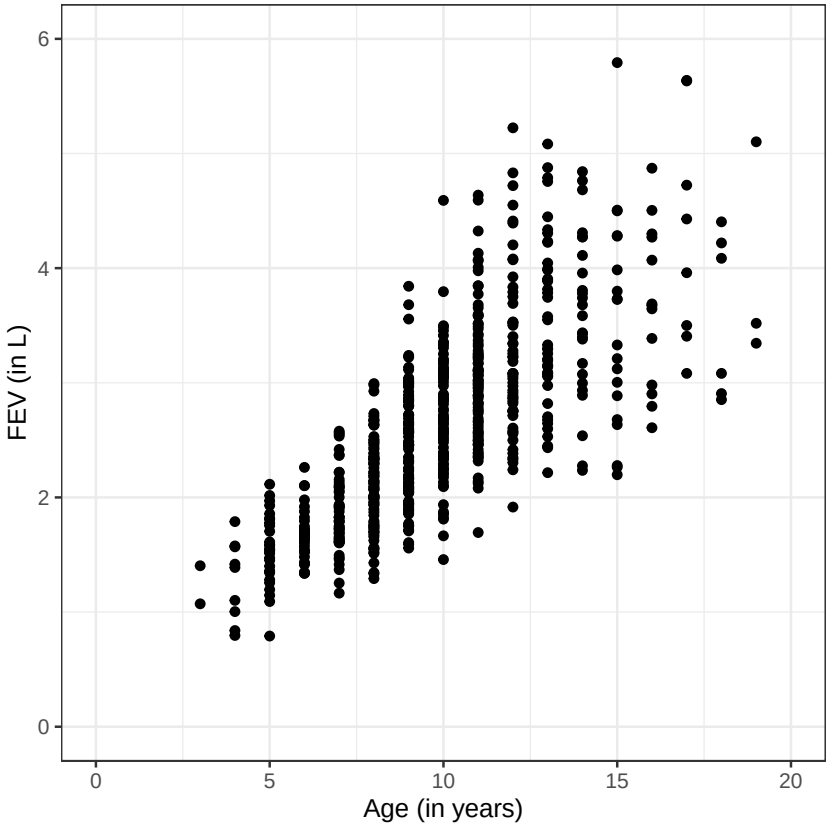
Liệu rằng dung tích phổi có thay đổi theo độ tuổi, chiều cao, giới tính hay trạng thái hút thuốc?

Để trả lời câu hỏi nghiên cứu này, trước hết ta sẽ dùng biểu đồ để biểu diễn mối liên hệ giữa khả năng dung tích phổi (FEV) và các biến: Age, Ht, Gender, Smoke.

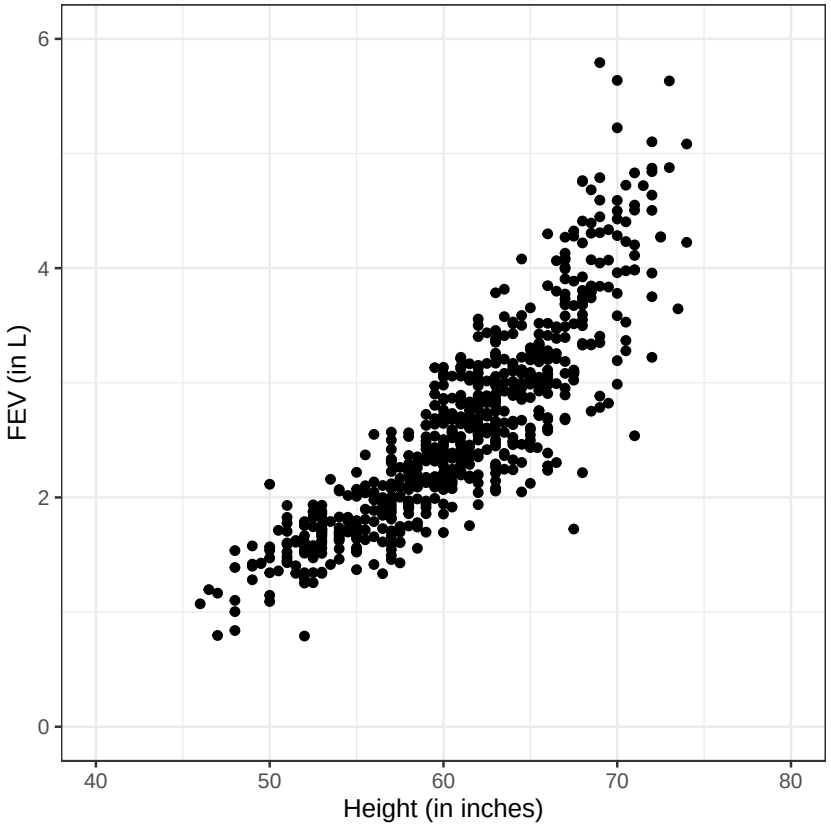
Đồ thị nào phù hợp để biểu diễn mối liên hệ của:

- **Age vs. FEV**
- **Ht vs. FEV**
- **Gender vs. FEV**
- **Smoke vs. FEV**

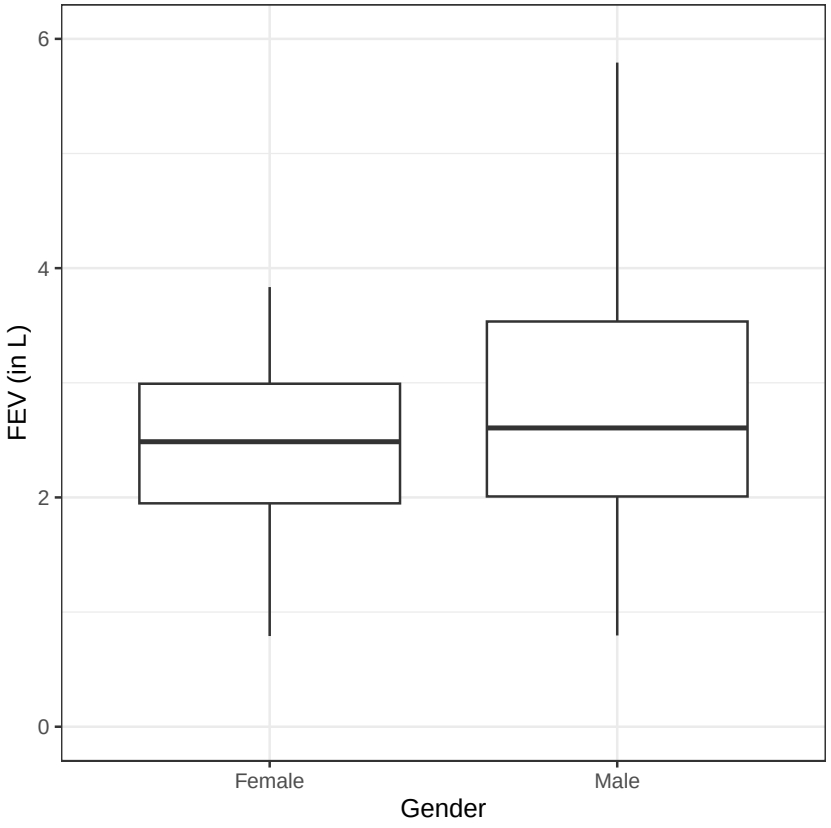
Ví dụ 1 - Dung tích phổi



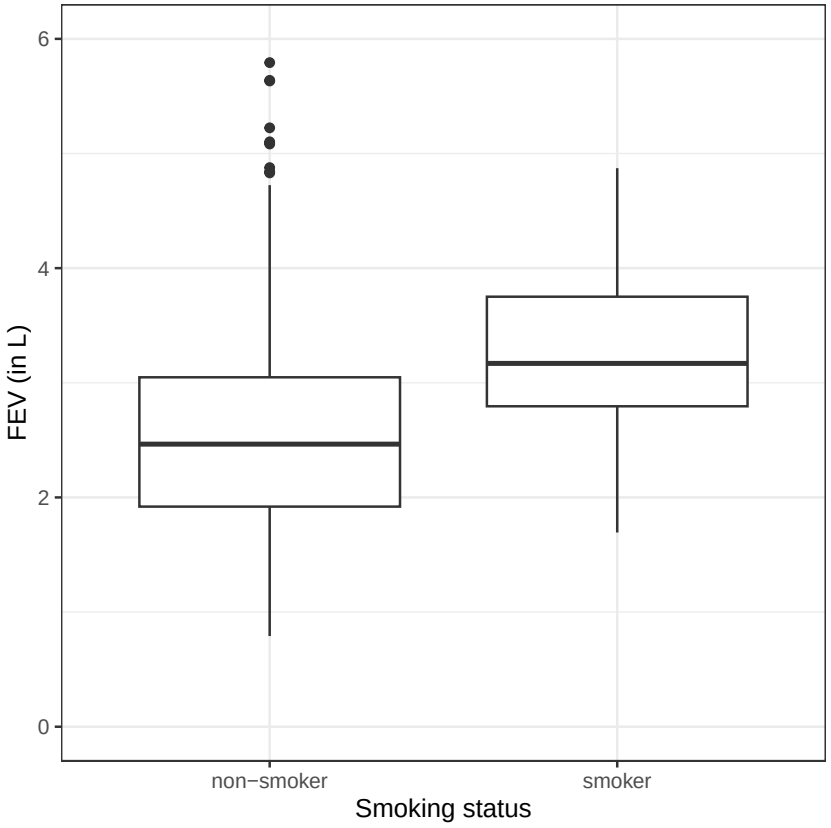
Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi

Tổng kết lại,

- độ tuổi (Age) và dung tích phổi (FEV) có sự tương quan tuyến tính (một cách tương đối, đặc biệt từ 1 tới 15 tuổi);
- sự tương quan giữa chiều cao (Ht) và dung tích phổi (FEV) là rất rõ ràng, tuy nhiên, sự tương quan có thể là phi tuyến;
- dung tích phổi (FEV) ở nam giới, nhìn chung, là cao hơn nữ giới, tuy nhiên, sự cao hơn là không thực sự rõ ràng;
- giá trị FEV ở những người hút thuốc (Smoke $= 1$) có xu hướng cao hơn ở những người không hút thuốc (Smoke $= 0$).

Ví dụ 1 - Dung tích phổi

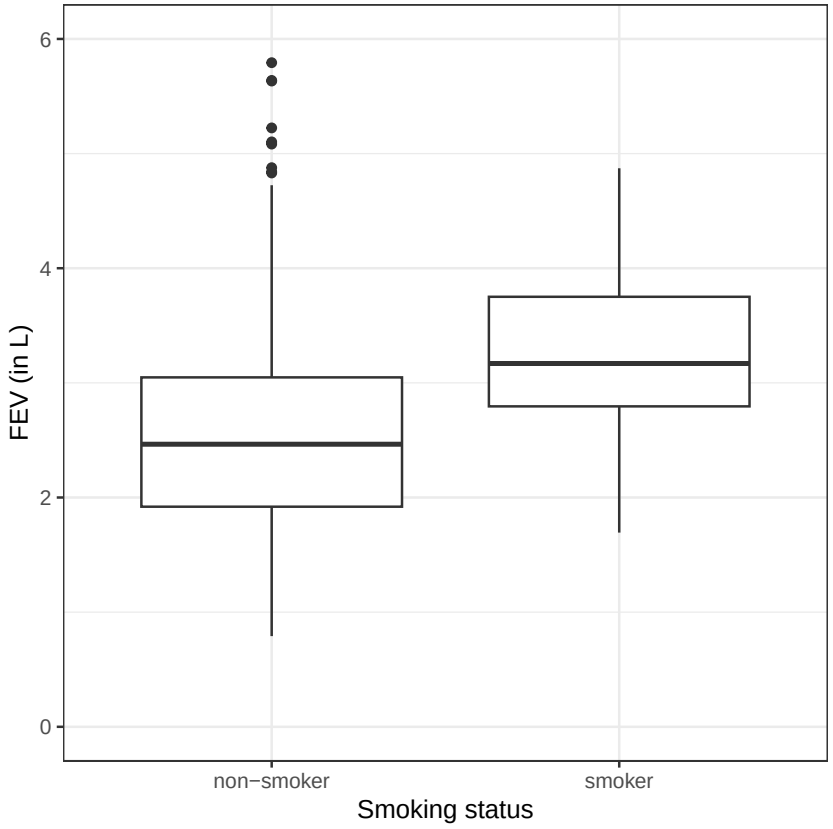
Tổng kết lại,

- độ tuổi (Age) và dung tích phổi (FEV) có sự tương quan tuyến tính (một cách tương đối, đặc biệt từ 1 tới 15 tuổi);
- sự tương quan giữa chiều cao (Ht) và dung tích phổi (FEV) là rất rõ ràng, tuy nhiên, sự tương quan có thể là phi tuyến;
- dung tích phổi (FEV) ở nam giới, nhìn chung, là cao hơn nữ giới, tuy nhiên, sự cao hơn là không thực sự rõ ràng;
- giá trị FEV ở những người hút thuốc (Smoke $= 1$) có xu hướng cao hơn ở những người không hút thuốc (Smoke $= 0$).

Tuy nhiên, những nhận xét này mới cho chúng ta góc nhìn về sự tương quan của một biến tới FEV.

Ví dụ 1 - Dung tích phổi

Giá trị FEV ở những người hút thuốc (Smoke = 1) có xu hướng cao hơn ở những người không hút thuốc (Smoke = 0). **Tại sao? Liệu rằng việc hút thuốc giúp tăng dung tích phổi?**



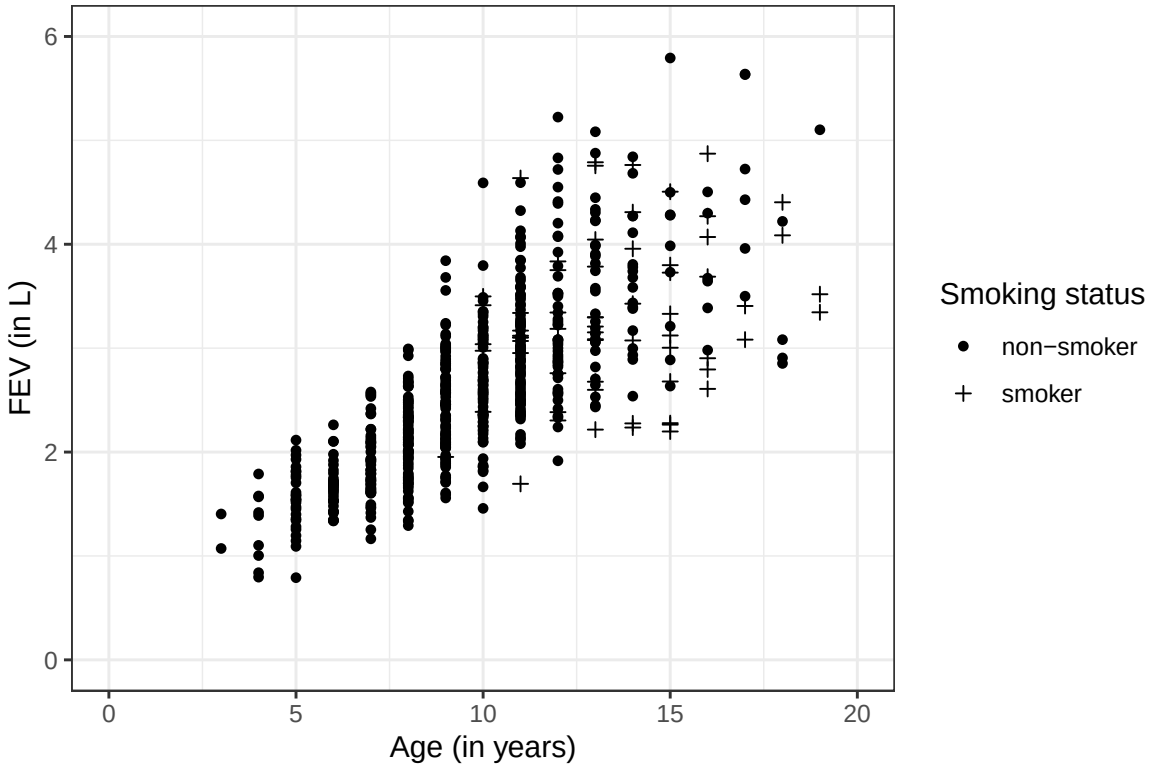
Ví dụ 1 - Dung tích phổi

Để tìm hiểu nguyên nhân tại sao giá trị FEV ở những người hút thuốc có xu hướng cao hơn ở những người không hút thuốc, ta sẽ tìm hiểu sự tương quan của các biến tới FEV dựa trên hai mẫu con:

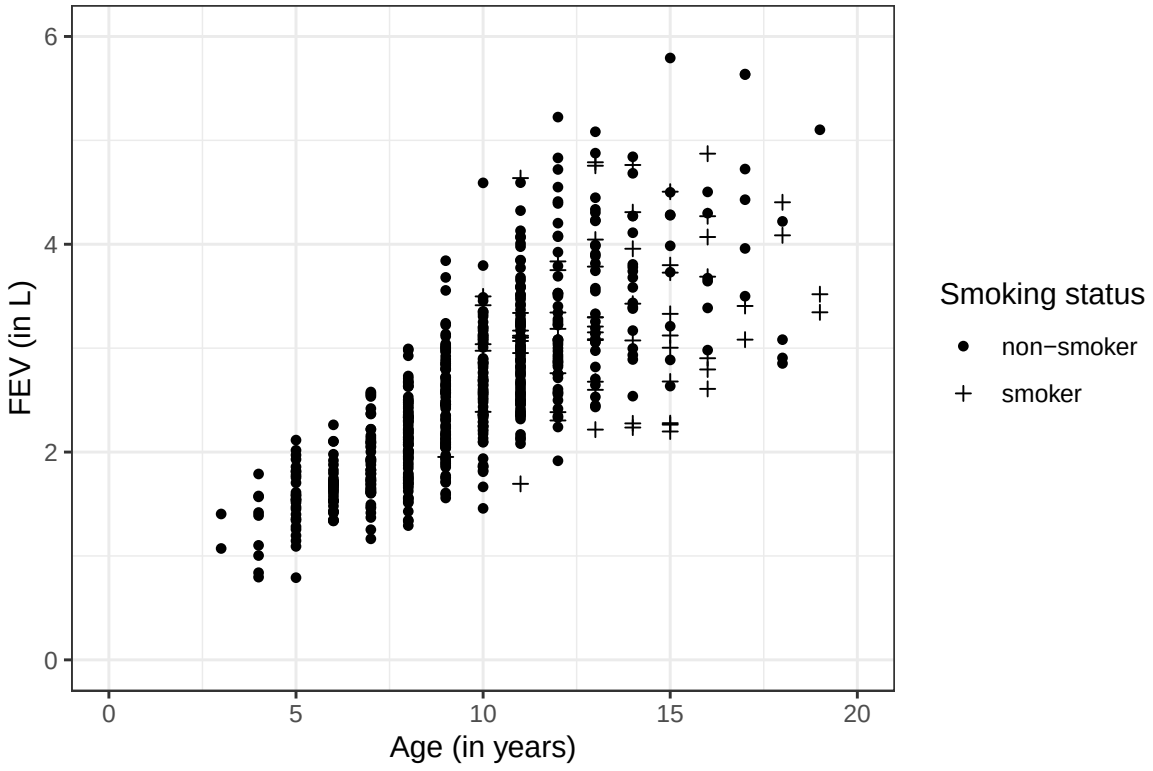
- không hút thuốc (Smoke $= 0$);
- hút thuốc (Smoke $= 1$).

Ta sẽ tập trung vào hai biến Age và Ht, bởi sự tương quan của chúng tới FEV đã được kiểm chứng.

Ví dụ 1 - Dung tích phổi

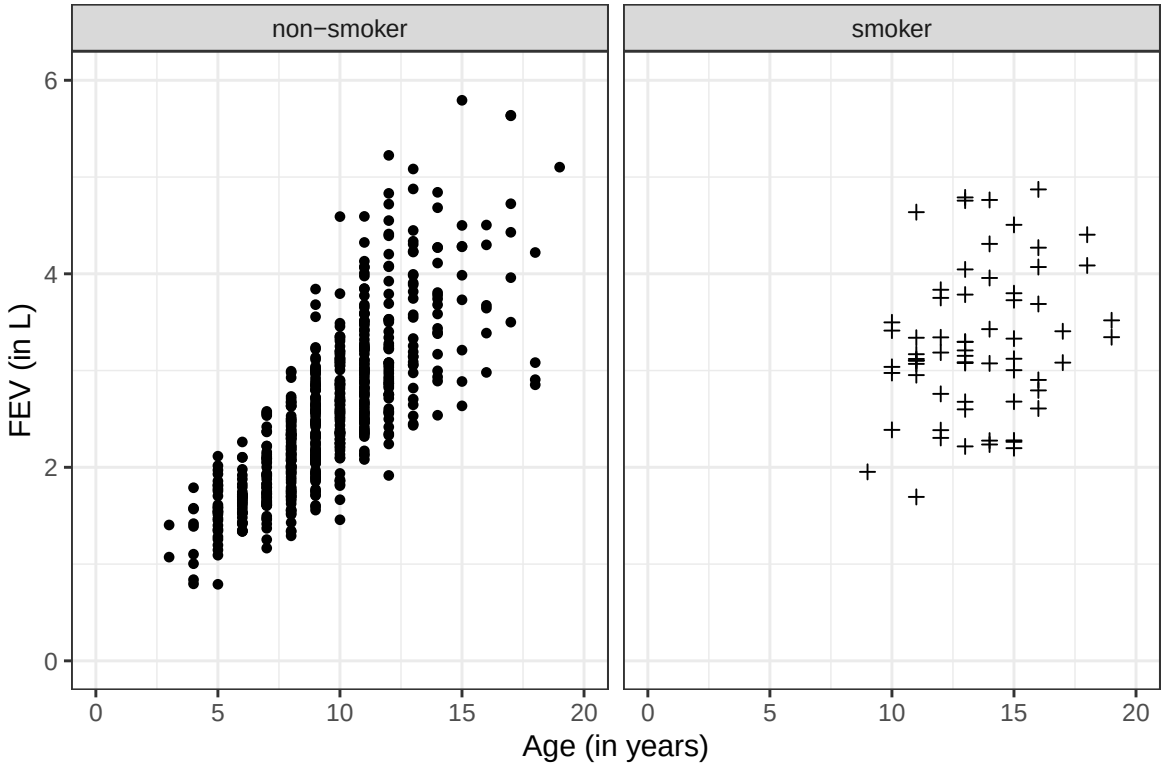


Ví dụ 1 - Dung tích phổi

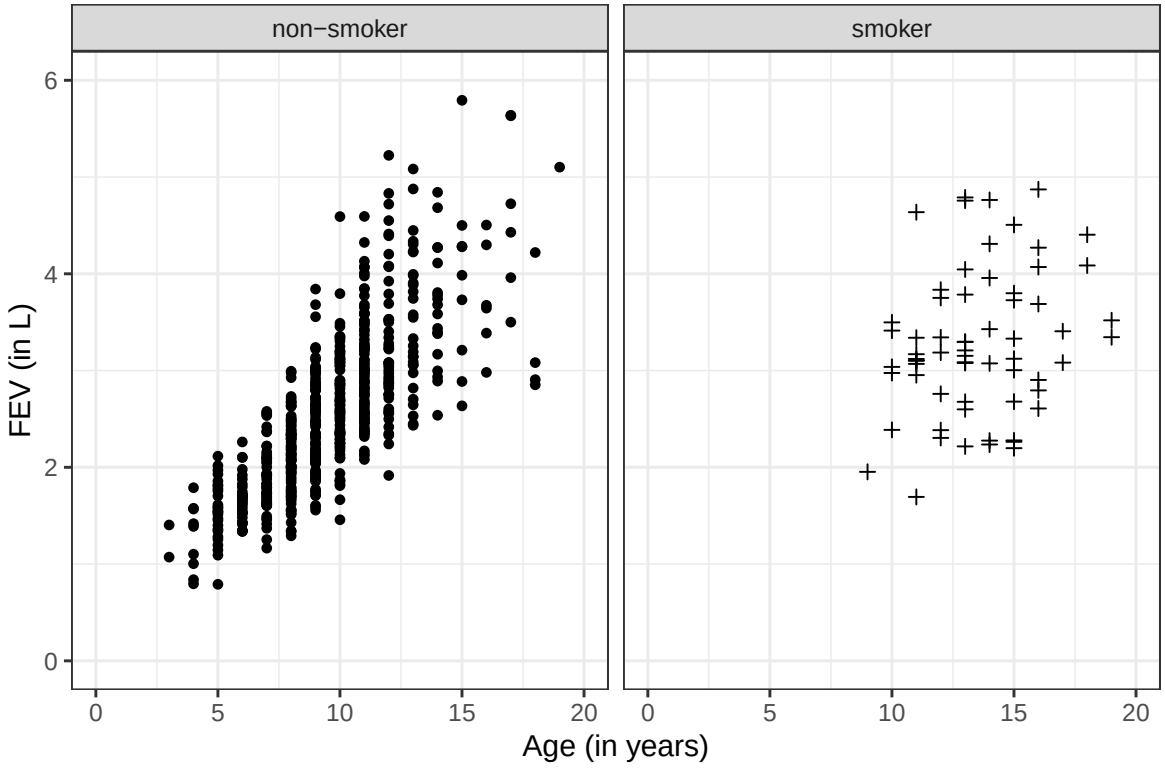


- những người hút thuốc có độ tuổi lớn (ít nhất 10);
- quan sát vùng dữ liệu từ 10 tuổi trở lên, ta thấy rằng giá trị FEV là không thực sự khác biệt giữa những người hút thuốc hoặc không hút thuốc.

Ví dụ 1 - Dung tích phổi

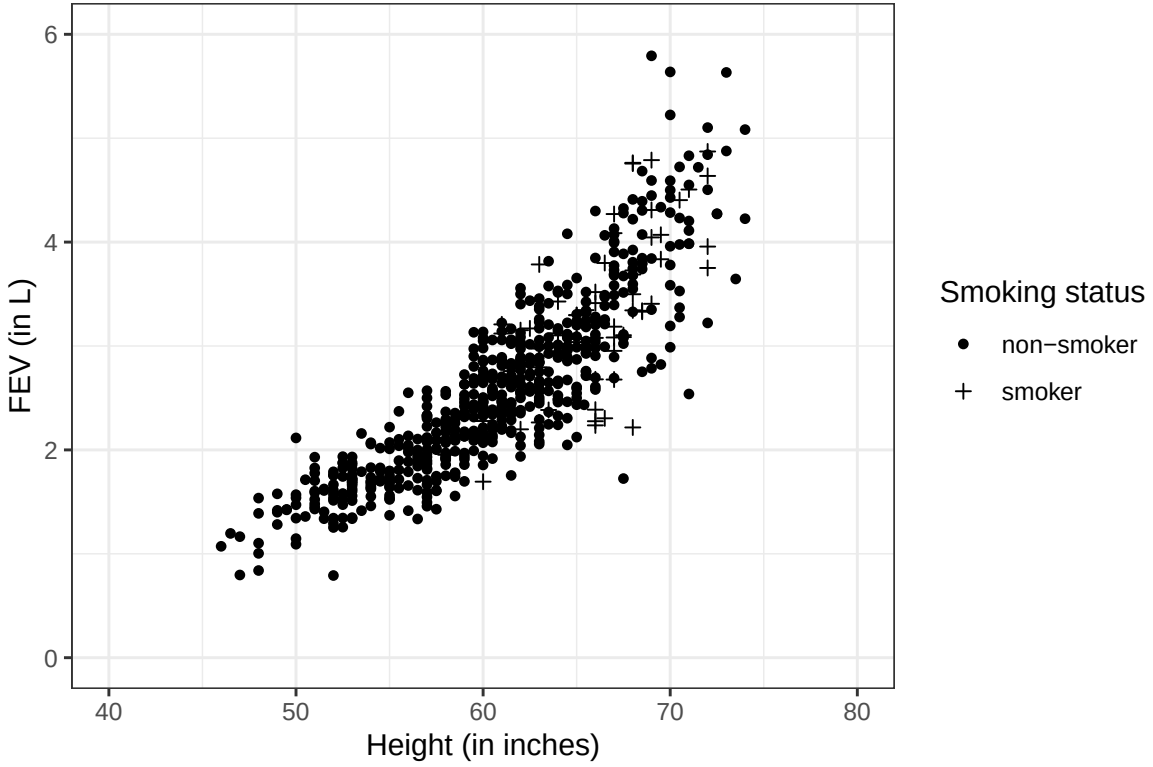


Ví dụ 1 - Dung tích phổi

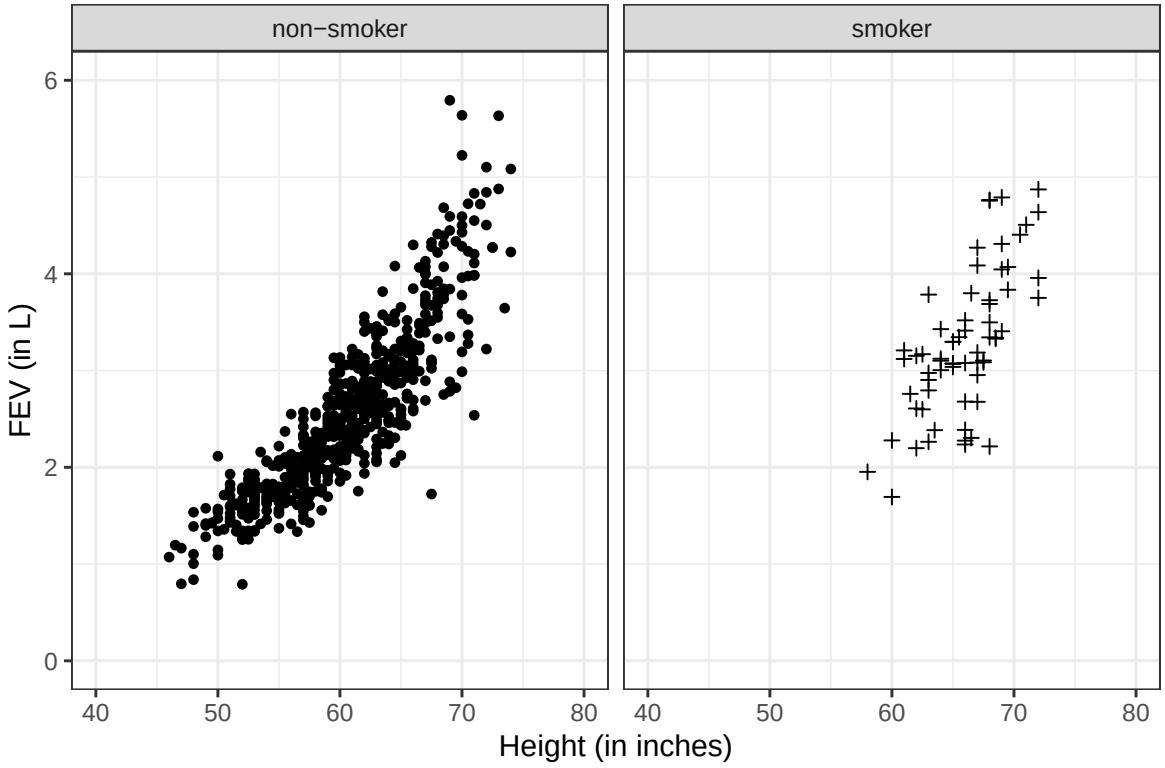


Hai đồ thị cho thấy việc giá trị FEV của những người hút thuốc cao hơn so với những người không hút thuốc là do độ tuổi!

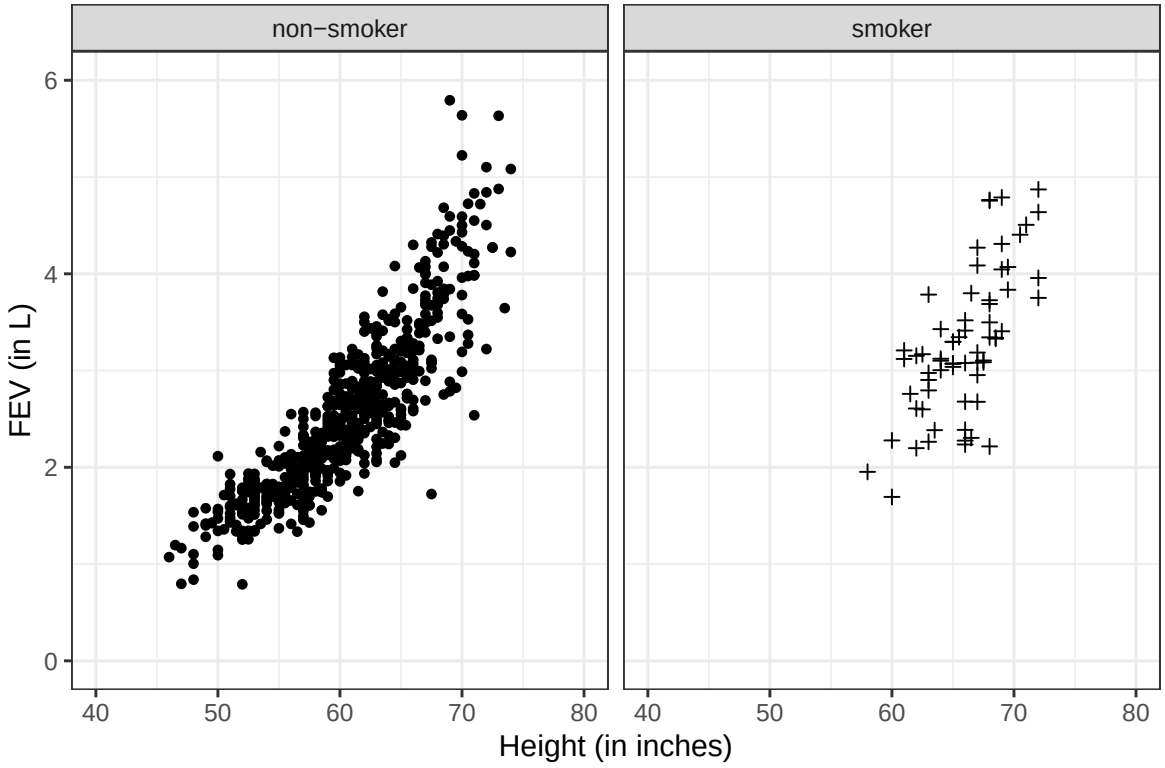
Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Ví dụ 1 - Dung tích phổi



Các đồ thị cho thấy việc giá trị FEV của những người hút thuốc cao hơn so với những người không hút thuốc là do chiều cao!

Ví dụ 1 - Dung tích phổi

Như vậy, về mặt tổng quan, các biến Age, Ht và Smoke có tương quan hay nói cách khác, là có ảnh hưởng hay tác động tới FEV.

Tuy nhiên, để có thể định lượng được sự tác động đó, về mặt toán học, ta cần thiết lập một *mô hình thống kê*.

1 Một số thông tin của khóa học

2 Nhắc lại một số kiến thức cơ bản

3 Giới thiệu mô hình tuyến tính tổng quát
- Ví dụ
- Mô hình thống kê

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

biến giải thích (explanatory variable) biến được dùng để giải thích sự thay đổi của biến đáp ứng bởi mô hình;

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

biến giải thích (explanatory variable) biến được dùng để giải thích sự thay đổi của biến đáp ứng bởi mô hình;

biến nhân tố (factor) là biến giải thích nhưng với dạng phân loại (định danh hoặc thứ bậc), biến này có tên khác là giả biến (*dummy variable*);

Biến ngẫu nhiên trong mô hình

Trong một mô hình thống kê, ta thường có những định danh như sau cho các biến ngẫu nhiên (bất kể dạng biến ngẫu nhiên):

biến đáp ứng (response variable) biến ngẫu nhiên được quan tâm trong nghiên cứu, giá trị và sự biến động được diễn tả bởi mô hình thông qua những biến khác;

biến giải thích (explanatory variable) biến được dùng để giải thích sự thay đổi của biến đáp ứng bởi mô hình;

biến nhân tố (factor) là biến giải thích nhưng với dạng phân loại (định danh hoặc thứ bậc), biến này có tên khác là giả biến (*dummy variable*);

biến gây nhiễu (compounding variable) là một dạng đặc biệt của biến giải thích, một biến giải thích được coi là gây nhiễu nếu sự xuất hiện của nó làm thay đổi sự tác động của một hoặc nhiều biến giải thích khác.

Biến ngẫu nhiên trong mô hình

Giả sử rằng ta muốn xây dựng một mô hình thống kê để giải thích sự thay đổi của FEV bởi sự tách động của Age (độ tuổi), Ht (chiều cao), Gender (giới tính) và Smoke (trạng thái hút thuốc). Khi đó,

- FEV là biến đáp ứng;
- Age, Ht, Gender và Smoke là các biến giải thích;
- Gender và Smoke là biến nhân tố.

Mô hình thống kê

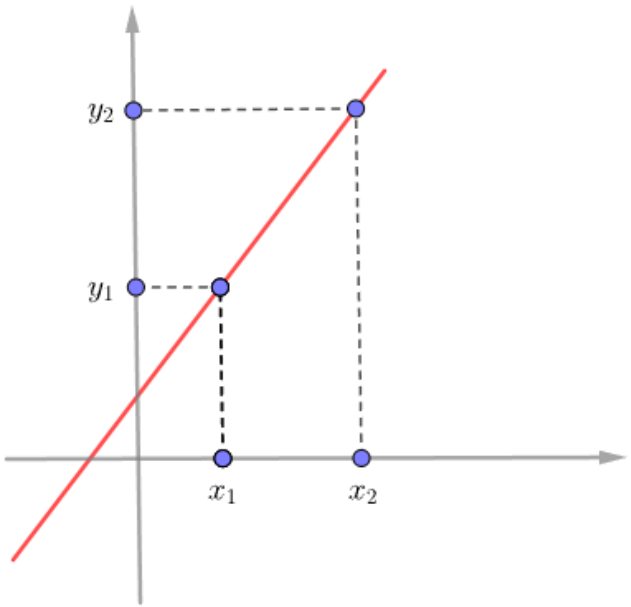
Xét phương trình đường thẳng:

$$y = a + bx$$

với a và b là các hằng số được biết trước.

- Cho trước một giá trị x_1 ta dễ dàng tính được tương ứng một giá trị y_1.

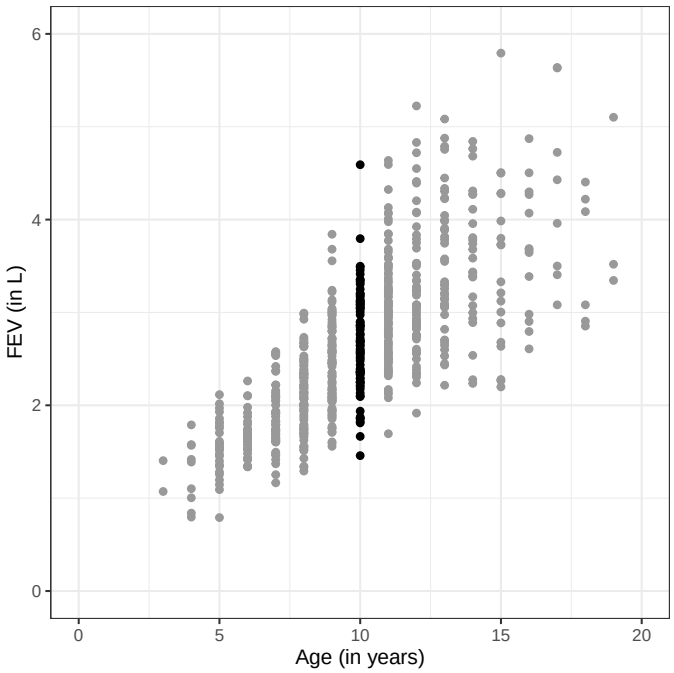
$\hookrightarrow$ Phương trình $y = a + bx$ là một dạng **mô hình toán** mô tả sự thay đổi giá trị y bởi giá trị của x.



Về mặt mô hình toán, với một giá trị của x ta sẽ quan sát được một giá trị của y, tuy nhiên, đối với dữ liệu quan sát thực tế thì không như vậy.

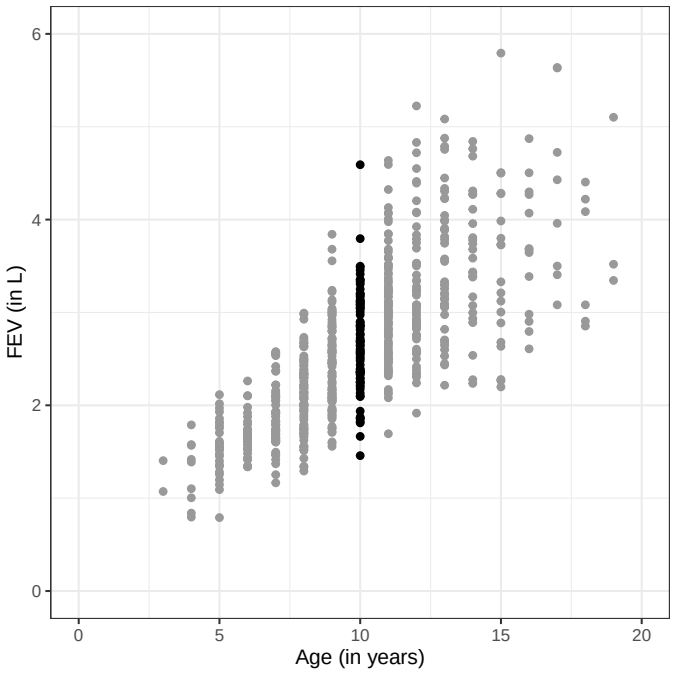
Mô hình thống kê

Một giá trị của Age (x) có thể ghi nhận nhiều giá trị khác nhau của FEV (y). Ví dụ Age $= 10$ (các điểm màu đen).

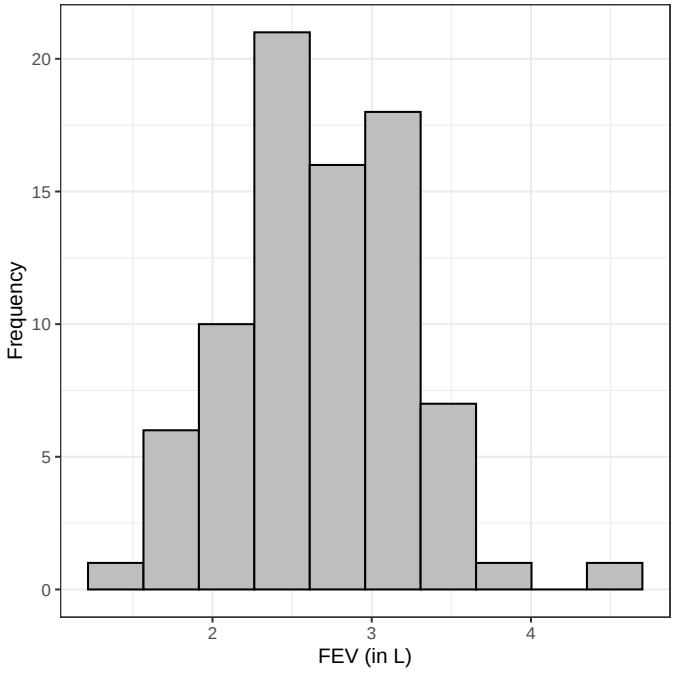


Mô hình thống kê

Một giá trị của Age (x) có thể ghi nhận nhiều giá trị khác nhau của FEV (y). Ví dụ Age $= 10$ (các điểm màu đen).



Những quan sát FEV của các đối tượng 10 tuổi là ngẫu nhiên, và do đó, chúng có phân phối (*phân phối có điều kiện*).



Mô hình thống kê

Do đó, một mô hình toán cho FEV dựa vào Age, chẳng hạn:

$$\text{FEV} = a + b \times \text{Age},$$

sẽ chỉ có thể mô tả được giá trị trung bình của FEV tương ứng của một giá trị Age, mà không thể mô tả được phân phối của FEV.

$\hookrightarrow$ Điều này đòi hỏi phải có thêm một thành phần trong mô hình để mô tả cho phân phối của FEV.

Mô hình thống kê

Mô hình thống kê

Một mô hình thống kê (*statistical model*) là một mô hình gồm hai thành phần:

- thành phần hệ thống (*systematic component*);
- thành phần ngẫu nhiên (*random component*),

mô tả lần lượt hai đặc trưng của biến đáp ứng: trung bình và phân phối, dựa vào giá trị của một hoặc nhiều biến giải thích.

Mô hình thống kê

Ví dụ 1: Mô hình thống kê cho FEV dựa vào các biến Age, Ht, Gender và Smoke có thể là:

- thành phần hệ thống:

$$\mu = \mathbb{E}(\text{FEV}) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Ht} + \beta_3 \text{Gender} + \beta_4 \text{Smoke},$$

- thành phần ngẫu nhiên: $\text{FEV} \sim \mathcal{N}(\mu, \sigma^2)$;

trong đó, các tham số $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ và σ là chưa biết, cần phải ước lượng từ dữ liệu.

Mô hình thống kê

Ví dụ 1: Mô hình thống kê cho FEV dựa vào các biến Age, Ht, Gender và Smoke có thể là:

- thành phần hệ thống:

$$\mu = \mathbb{E}(\text{FEV}) = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Ht} + \beta_3 \text{Gender} + \beta_4 \text{Smoke},$$

- thành phần ngẫu nhiên: $\text{FEV} \sim \mathcal{N}(\mu, \sigma^2)$;

trong đó, các tham số $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ và σ là chưa biết, cần phải ước lượng từ dữ liệu.

Cách viết mô hình như trên là cách viết dạng tổng quát. Khi ta có bộ dữ liệu với n quan sát độc lập, mô hình thống kê sẽ được biểu diễn cho quan sát thứ i, ví dụ:

- thành phần hệ thống:

$$\mu_i = \mathbb{E}(\text{FEV}_i) = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Ht}_i + \beta_3 \text{Gender}_i + \beta_4 \text{Smoke}_i,$$

- thành phần ngẫu nhiên: $\text{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$ (chúng ta đang giả định các quan sát FEV_i có phương sai đồng nhất).

Mô hình thống kê

Ví dụ 2: Một mô hình thống kê cho FEV có thể là,

- thành phần hệ thống:
$$\mu_i = \mathbb{E}(\text{FEV}_i) = \beta_0,$$
- thành phần ngẫu nhiên: $\text{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

Mô hình thống kê

Ví dụ 2: Một mô hình thống kê cho FEV có thể là,

- thành phần hệ thống:
$$\mu_i = \mathbb{E}(\text{FEV}_i) = \beta_0,$$
- thành phần ngẫu nhiên: $\text{FEV}_i \sim \mathcal{N}(\mu_i, \sigma^2)$.

Thông thường sẽ có nhiều dạng khác nhau cho thành phần hệ thống cũng như thành phần ngẫu nhiên.

Mô hình hồi quy

Mô hình hồi quy

Nếu ta giả định rằng thành phần hệ thống, hay trung bình μ_i là một hàm f của p biến giải thích với các tham số chưa biết, tức là

$$\mu_i = \mathbb{E}(y_i) = f(x_{1i}, \ldots, x_{pi}; \beta_0, \beta_1, \ldots, \beta_p),$$

khi đó mô hình thống kê sẽ được gọi là một mô hình hồi quy (*regression model*).

Mô hình hồi quy

Mô hình hồi quy

Nếu ta giả định rằng thành phần hệ thống, hay trung bình μ_i là một hàm f của p biến giải thích với các tham số chưa biết, tức là

$$\mu_i = \mathbb{E}(y_i) = f(x_{1i}, \ldots, x_{pi}; \beta_0, \beta_1, \ldots, \beta_p),$$

khi đó mô hình thống kê sẽ được gọi là một mô hình hồi quy (*regression model*).

Về mặt toán học, sẽ có rất nhiều sự kết hợp khác nhau của $x_{1i}, \ldots, x_{pi}$ và $\beta_0, \beta_1, \ldots, \beta_p$. Thông thường, ta giả sử rằng sự kết hợp này là tuyến tính, tức là

$$\mu_i = \mathbb{E}(y_i) = f(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}),$$

khi này, ta có mô hình hồi quy tuyến tính theo tham số.

Mô hình hồi quy

Ta có hai dạng mô hình tuyến tính như sau:

Mô hình hồi tuyến tính - Linear regression model: là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi},$$

- thành phần ngẫu nhiên $\mathbb{V}\mathrm{ar}(y_i) = \sigma^2$. Chú ý, chúng ta không cần đưa ra giả định cụ thể nào về phân phối.

Mô hình hồi quy

Ta có hai dạng mô hình tuyến tính như sau:

Mô hình hồi tuyến tính - Linear regression model: là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi},$$

- thành phần ngẫu nhiên $\mathbb{V}\mathrm{ar}(y_i) = \sigma^2$. Chú ý, chúng ta không cần đưa ra giả định cụ thể nào về phân phối.

Mô hình hồi tuyến tính tổng quát - Generalized linear model: là một mô hình thống kê với

- thành phần hệ thống

$$\mu_i = \mathbb{E}(y_i) = g^{-1}(\beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}),$$

  trong đó, $g(\cdot)$ là một hàm được xác định sao cho đồng biến và khả vi, và được gọi là hàm liên kết (*link function*),
- thành phần ngẫu nhiên: y_i tuân theo một phân phối xác định F với trung bình μ_i.

Mô hình hồi quy

Ví dụ 3: *Những tác nhân nào có thể làm gia tăng nguy cơ bị bệnh tim mạch vành (coronary heart disease)?* Để tìm ra câu trả lời, một nhóm các nhà nghiên cứu đã ghi lại trạng thái bệnh tim mạch vành của 3154 người đàn ông (với độ tuổi 39 tới 59, từ khu vực San Francisco) trong vòng 8 năm và 6 tháng. Đồng thời, họ cũng ghi lại những biến có khả năng ảnh hưởng tới khả năng bị bệnh tim mạch vành.

id	age	chd	weight	chol	cigs
1	49	no	150	225	25
2	42	no	160	177	20
3	42	no	160	181	0
4	41	no	152	132	20
5	59	yes	150	255	20
6	44	no	204	182	0
⋮	⋮	⋮	⋮	⋮	⋮
3154	39	no	155	264	40

Mô hình hồi quy

- age: độ tuổi
- chd: trạng thái bị bệnh tim mạch vành (có - yes hoặc không - no)
- weight: cân nặng
- chol: hàm lượng cholesterol
- cigs: số lượng thuốc lá hút trong một ngày

Mô hình hồi quy

- age: độ tuổi
- chd: trạng thái bị bệnh tim mạch vành (có - yes hoặc không - no)
- weight: cân nặng
- chol: hàm lượng cholesterol
- cigs: số lượng thuốc lá hút trong một ngày

Bởi vì trọng tâm của câu hỏi là về khả năng bị bệnh tim mạch vành, do đó, biến chd sẽ là biến đáp ứng.

Một số nghi vấn cho rằng, hút thuốc là nhiều, béo phì, và hàm lượng cholesterol cao, là những nguyên nhân tăng khả năng bị bệnh tim mạch vành.

Để làm rõ nghi vấn này, ta xem xét một mô hình hồi quy tuyến tính tổng quát như sau.

Mô hình hồi quy

Với mỗi quan sát thứ i, ta thấy rằng, chd_i chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd_i có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

Mô hình hồi quy

Với mỗi quan sát thứ i, ta thấy rằng, chd$_i$ chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd$_i$ có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

$\hookrightarrow$ thành phần ngẫu nhiên của mô hình sẽ là: chd$_i \sim \mathcal{B}$er(p_i).

Mô hình hồi quy

Với mỗi quan sát thứ i, ta thấy rằng, chd_i chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd_i có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

$\hookrightarrow$ thành phần ngẫu nhiên của mô hình sẽ là: $\text{chd}_i \sim \mathcal{B}\text{er}(p_i)$.

thành phần hệ thống có thể là một tổ hợp tuyến tính bao gồm các biến: age, weight, chol và cigs, được xác định như sau:

$$\mu_i = \text{expit}\left(\beta_0 + \beta_1 \text{age}_i + \beta_2 \text{weight}_i + \beta_3 \text{chol}_i + \beta_4 \text{cigs}_i\right),$$

với $\text{expit}(u) = \dfrac{\exp(u)}{1+\exp(u)}$; chú ý, $\mu_i = \mathbb{E}(\text{chd}_i) = p_i$.

Mô hình hồi quy

Với mỗi quan sát thứ i, ta thấy rằng, chd_i chỉ có hai giá trị:

- no - không bị bệnh, ký hiệu là 0;
- yes - bị bệnh, ký hiệu là 1;

do đó, chd_i có thể là một biến ngẫu nhiên Bernoulli với xác suất bị bệnh p_i.

$\hookrightarrow$ thành phần ngẫu nhiên của mô hình sẽ là: $\text{chd}_i \sim \mathcal{B}\text{er}(p_i)$.

thành phần hệ thống có thể là một tổ hợp tuyến tính bao gồm các biến: age, weight, chol và cigs, được xác định như sau:

$$\mu_i = \text{expit}\left(\beta_0 + \beta_1 \text{age}_i + \beta_2 \text{weight}_i + \beta_3 \text{chol}_i + \beta_4 \text{cigs}_i\right),$$

với $\text{expit}(u) = \dfrac{\exp(u)}{1+\exp(u)}$; chú ý, $\mu_i = \mathbb{E}(\text{chd}_i) = p_i$.

$\hookrightarrow$ mô hình dạng này được gọi là mô hình hồi quy logistic (*logistic regression model*).

Một số dạng có thể của thành phần hệ thống

Giả sử ta có biến đáp ứng là y với trung bình μ, các biến giải thích lần lượt là x_1, x_2, x_3 và x_4. Các dạng có thể của thành phần hệ thống là:

$$\begin{aligned}
\mu &= \beta_0 + \beta_1 x_1 + \beta_4 x_4 && (1)\\
\mu &= \beta_0 + \beta_2 x_2 + \beta_3 x_2^2 + \beta_4 x_4 && (2)\\
\mu &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 && (3)\\
\mu &= \beta_0 + \beta_1 \log(x_1) + \beta_2 x_2 + \beta_4 x_4 && (4)\\
\mu &= \beta_0 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_4 && (5)\\
1/\mu &= \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 && (6)\\
\log(\mu) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 && (7)\\
\mu &= \beta_0 + \exp(\beta_1 x_1) - \exp(\beta_2 x_2) + \beta_4 x_4^2 && (8)
\end{aligned}$$

- Các phương trình từ (1) - (7) đều có dạng tuyến tính theo tham số.
- Các phương trình từ (1) - (5) có thể được sử dụng để chỉ định một mô hình quy tuyến tính.

Diễn giải mô hình

Các mô hình hữu ích nhất khi chúng có những diễn giải hợp lý.

So sánh hai thành phần hệ thống sau:

$$\mu = \beta_0 + \beta_1 x \tag{9}$$

$$\log(\mu) = \beta_0 + \beta_1 x \tag{10}$$

- mô hình (9): x tăng 1 đơn vị thì μ tăng β_1 đơn vị;
- mô hình (10): x tăng 1 đơn vị thì $\log(\mu)$ tăng β_1 đơn vị;

  $\hookrightarrow$ x tăng 1 đơn vị thì μ tăng $\exp(\beta_1)$ lần. Tại sao?

Trong ứng dụng, ta cần xem xét lựa chọn mô hình (thành phần hệ thống) sao cho phù hợp với thực tế, hơn là tập trung quá nhiều vào công thức toán.

Diễn giải mô hình

Hai mục tiêu chính của mô hình hồi quy là

Dự đoán: để tạo ra những dự đoán chính xác từ dữ liệu mới hoặc dữ liệu trong tương lai.

Hiểu và diễn giải: để hiểu các biến có liên quan với nhau như thế nào.

Diễn giải mô hình

Hai mục tiêu chính của mô hình hồi quy là

Dự đoán: để tạo ra những dự đoán chính xác từ dữ liệu mới hoặc dữ liệu trong tương lai.

Hiểu và diễn giải: để hiểu các biến có liên quan với nhau như thế nào.

Ví dụ, hãy xem xét nghiên cứu về dung tích phổi.

- Ta bắt đầu với một câu hỏi về xác định mối liên hệ giữa trạng thái hút thuốc, giới tính, độ tuổi, chiều cao với dung tích phổi (FEV).
  $\hookrightarrow$ các hệ số trong mô hình sẽ miêu tả mối liên hệ (tuyến tính), và ta cần diễn giải triệt để các hệ số này.
- Mặt khác, trong tương lai, nếu ta thu thập được dữ liệu về trạng thái hút thuốc, giới tính, độ tuổi và chiều cao của một nhóm người, ta có thể dự đoán dung tích phổi (FEV) dựa vào mô hình, mà không cần yêu cầu thực hiện thí nghiệm.

Độ chính xác vs Tính đơn giản

Một mô hình thống kê phù hợp (adequate) cần cân bằng hai tiêu chí:

Độ chính xác: mô hình phải mô tả chính xác cả thành phần hệ thống và thành phần ngẫu nhiên của dữ liệu.

Tính đơn giản: mô hình nên đơn giản nhất có thể.

Độ chính xác vs Tính đơn giản

Một mô hình thống kê phù hợp (adequate) cần cân bằng hai tiêu chí:

Độ chính xác: mô hình phải mô tả chính xác cả thành phần hệ thống và thành phần ngẫu nhiên của dữ liệu.

Tính đơn giản: mô hình nên đơn giản nhất có thể.

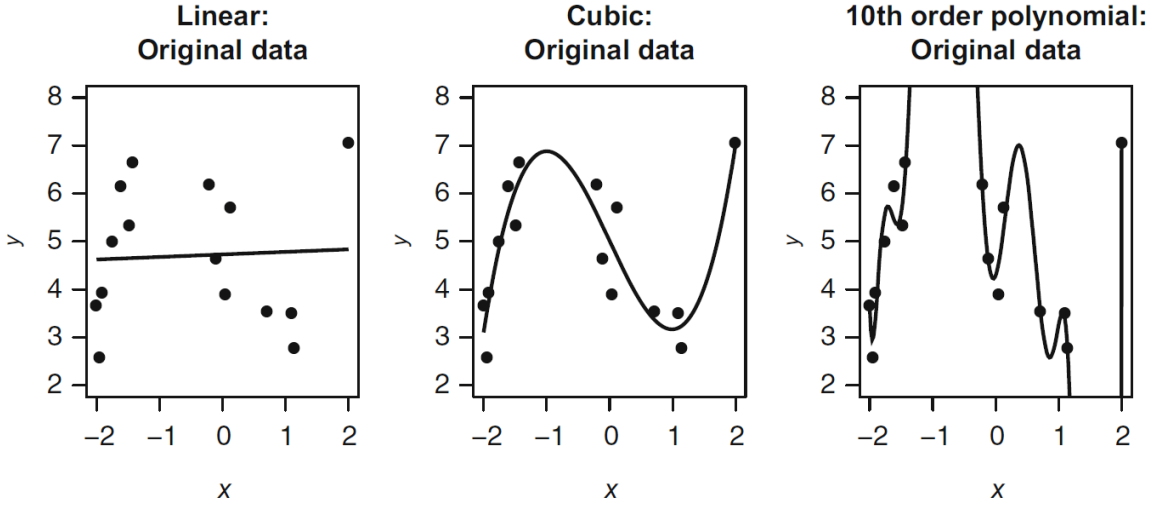
Ví dụ: Xét bộ dữ liệu mô phỏng từ mô hình hồi quy sau:

$$\begin{cases} y \sim \mathcal{N}(\mu, 0.35), \\ \mu = x^3 - 2x + 5 \end{cases}$$

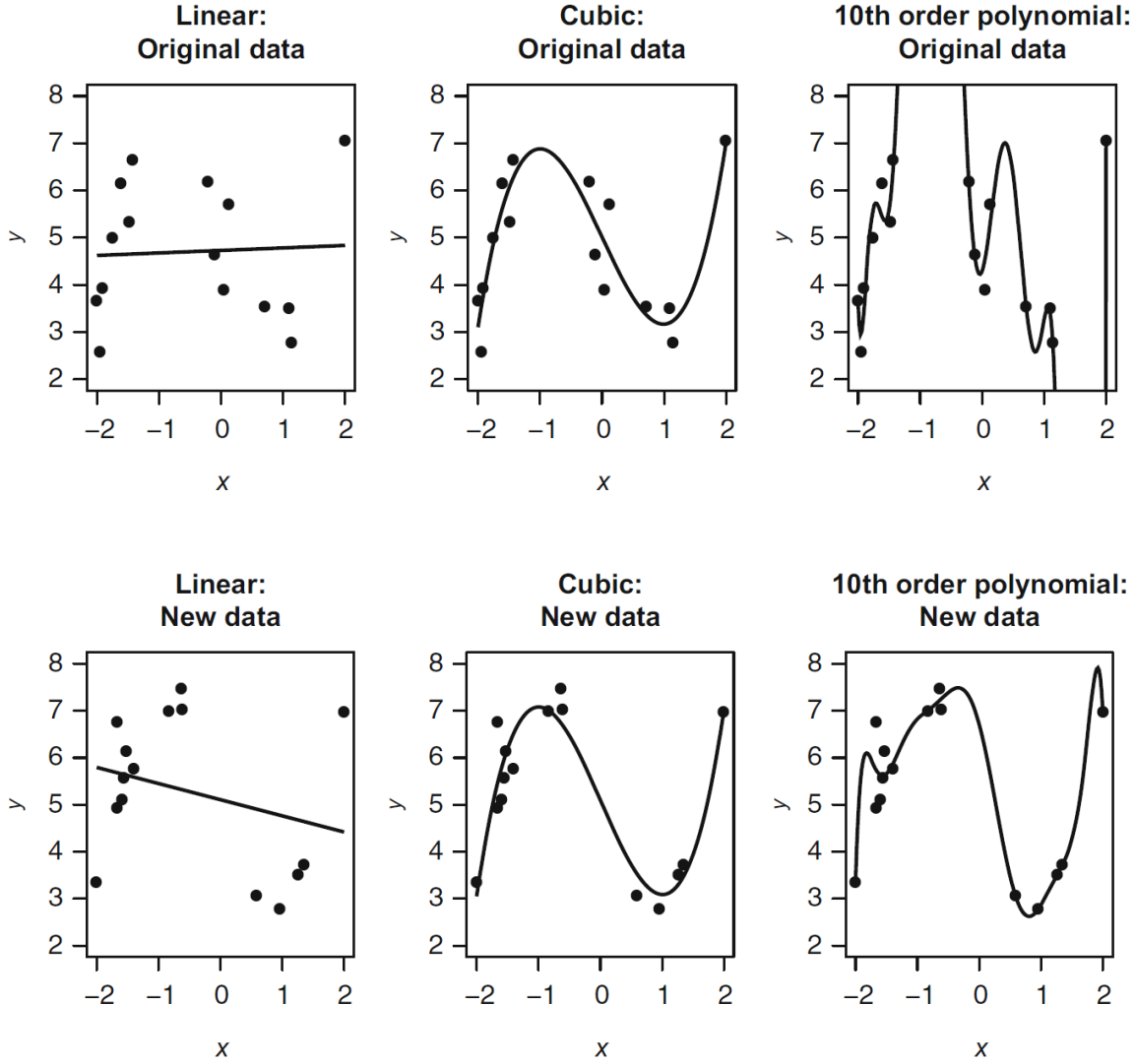
Ta xem xét ba mô hình tiềm năng sau

- tuyến tính đơn: $\mu = \beta_0 + \beta_1 x$;
- tuyến tính bậc ba: $\mu = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$;
- tuyến tính bậc mười: $\mu = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots + \beta_{10} x^{10}$

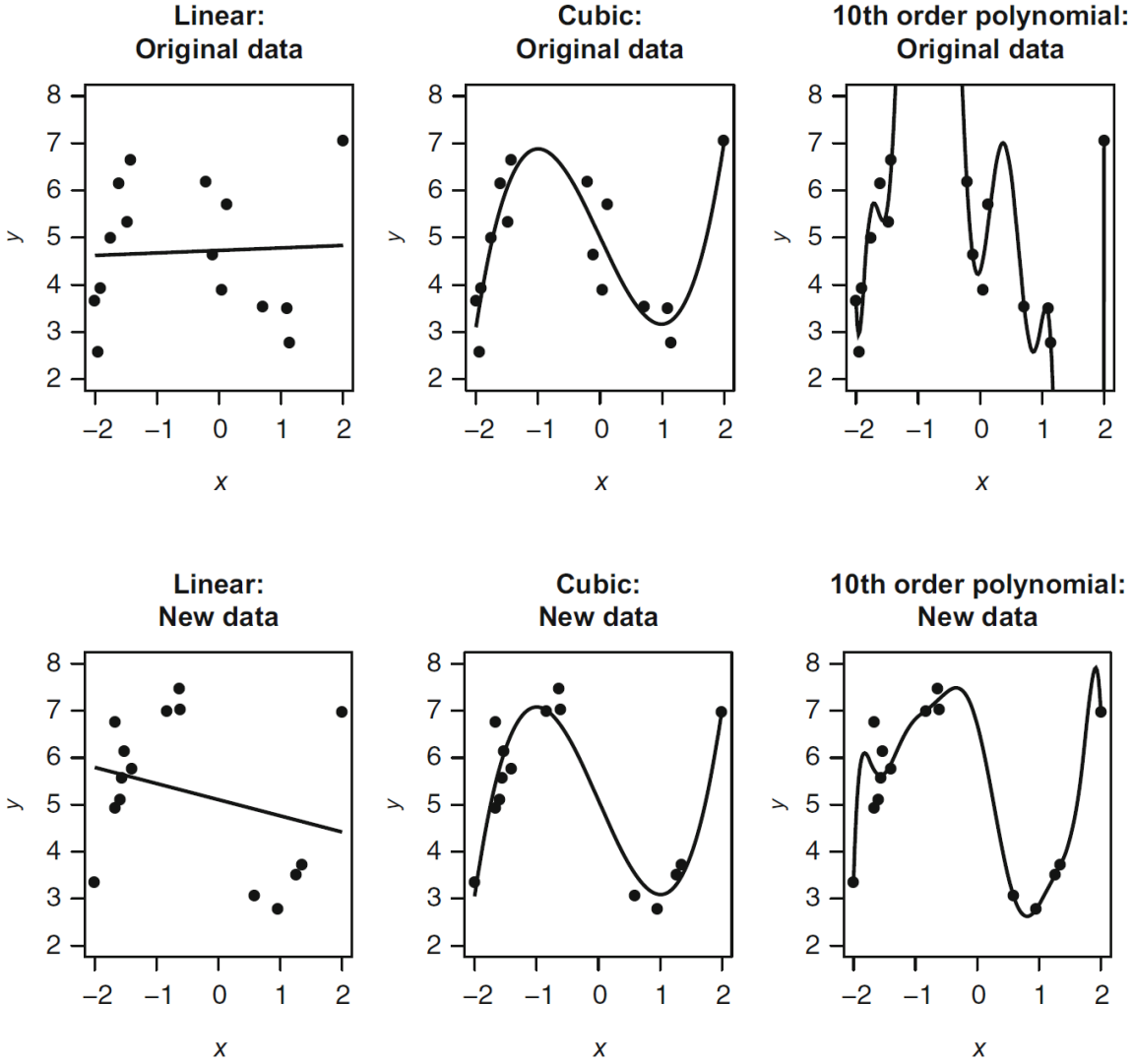
Độ chính xác vs Tính đơn giản



Độ chính xác vs Tính đơn giản



Độ chính xác vs Tính đơn giản



Một mô hình tốt sẽ giống nhau cho cả hai bộ dữ liệu.