EXPERT SYSTEMS

Kỹ thuật hồi quy

PGS.TS. Hoàng Văn Dũng

dunghv@hcmute.edu.vn

Nội dung

Hồi quy tuyến tính hàm một biến

Hồi quy tuyến tính hàm nhiều biến

Hồi quy Logistic

Kỹ thuật hồi quy tuyến tính

- Phần này giới thiệu Linear Regression (hồi quy tuyến tính) thuộc nhóm Supervised learning. Hồi quy tuyến tính là một phương pháp đơn giản nhưng đã được chứng minh được tính hữu ích cho một số lượng lớn các tình huống.

Một số kiến thức về thống kê:

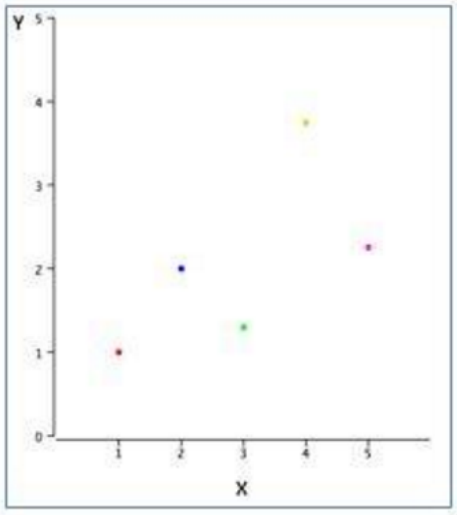
- Tương quan (r): Giải thích mối quan hệ giữa hai biến
- Giá trị trung bình (mean)
- Phương sai (σ^2), độ lệch chuẩn: Đánh giá độ phân tán trong dữ liệu
- Phân phối chuẩn
- Sai số (lỗi) - {giá trị thực tế - giá trị dự đoán}

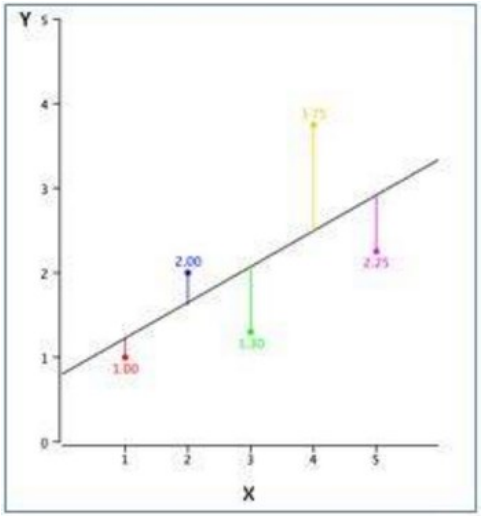
Kỹ thuật hồi quy tuyến tính

Ví dụ: Y = B0 + B1*X

- Y = Biến phụ thuộc
- X = biến độc lập
- B0 = Hằng số
- B1 = Hệ số mối quan hệ giữa X và Y

Đường hồi quy tuyến tính





Kỹ thuật hồi quy tuyến tính

- Xét trường hợp đơn giản nhất: $f(X_i)$ có dạng tuyến tính
  $\Rightarrow$ Hàm hồi quy tổng thể PRF :

  $$E(Y/X_i) = \beta_1 + \beta_2 X_i$$

  trong đó β_1, β_2 là các hệ số hồi quy
- Ý nghĩa các hệ số hồi quy:
  - $E(Y/_{Xi=0}) = \beta_1$: β_1 là hệ số tự do (hệ số chặn), cho biết trung bình của Y khi X = 0
  - $E(Y/_{Xi=Xi+1}) = \beta_1 + \beta_2 (X_i + 1)$

    $\Rightarrow E(Y/_{Xi=Xi+1}) - E(Y/X_i) = \beta_2$: β_2 là hệ số góc, cho biết khi X tăng 1 đơn vị thì trung bình Y tăng β_2 đơn vị

Kỹ thuật hồi quy tuyến tính

- **Tính chất của hồi quy tuyến tính**
  - Đường hồi quy luôn luôn đi qua trung bình của biến độc lập (x) cũng như trung bình của biến phụ thuộc (y)
  - Đường hồi quy tối thiểu hóa tổng của "Diện tích các sai số".

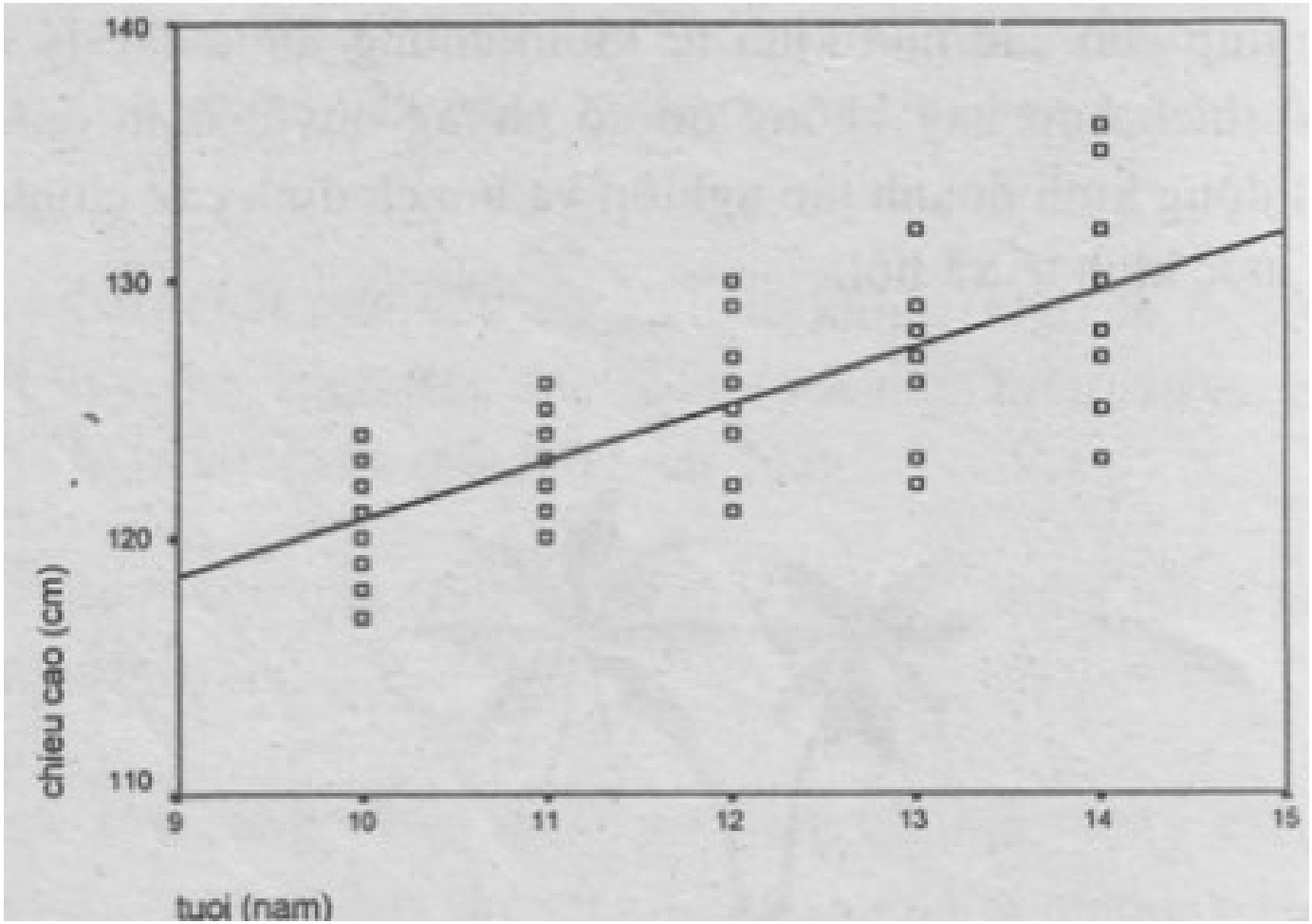
  => lý do hồi quy tuyến tính được gọi là "Ordinary Least Square (OLS)"
  - B1 giải thích sự thay đổi trong Y với sự thay đổi X bằng một đơn vị. Nếu chúng ta tăng giá trị của X bởi một đơn vị thì nó sẽ là sự thay đổi giá trị của Y

Hồi quy tuyến tính hàm một biến

- Phân tích hồi quy là nghiên cứu sự phụ thuộc của một biến (biến phụ thuộc) vào một hay nhiều biến khác (các biến giải thích) để ước lượng hay dự đoán giá trị trung bình của biến phụ thuộc trên cơ sở các giá trị biết trước của các biến giải thích.

Hồi quy tuyến tính hàm một biến

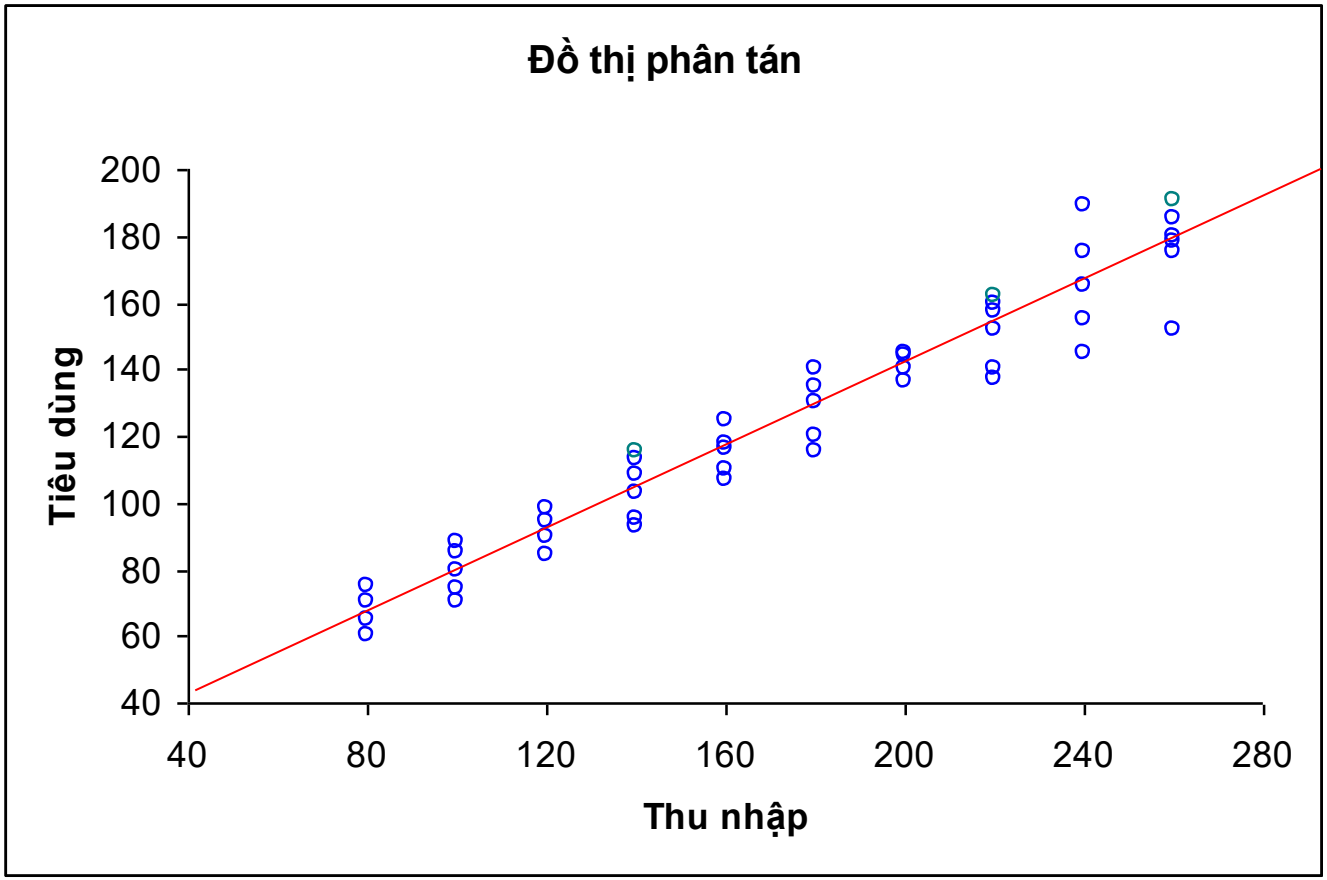
Ví dụ: Quan hệ giữa chiều cao của học sinh nam tính theo những độ tuổi cố định



Hồi quy tuyến tính hàm một biến

Ví dụ: Sự phụ thuộc của chi tiêu vào thu nhập thực tế

- Tỷ lệ thay đổi tiền lương trong mối quan hệ với tỷ lệ thất nghiệp



Mối quan hệ trong hồi quy

a. Quan hệ thống kê và quan hệ hàm số:

- Quan hệ thống kê thể hiện ở sự phụ thuộc thống kê của biến phụ thuộc vào các biến giải thích.
  - Biến phụ thuộc là đại lượng ngẫu nhiên có phân phối xác suất
  - Các biến giải thích có giá trị biết trước
  - Ứng với mỗi giá trị của biến giải thích có thể có nhiều giá trị khác nhau của biến giải thích
- Quan hệ hàm số:
  - Các biến không phải là ngẫu nhiên
  - Ứng với mỗi giá trị của biến giải thích có một giá trị của biến phụ thuộc
  - Phân tích hồi quy không nghiên cứu các quan hệ hàm số

Ví dụ: Sự phụ thuộc của năng suất lúa vào nhiệt độ, lượng mưa, lượng phân bón … là một quan hệ thống kê Tính chu vi hình vuông bằng 4 lần chiều dài $y = 4x$ là quan hệ hàm số.

Mối quan hệ trong hồi quy

b. Hồi quy và quan hệ nhân quả: Phân tích hồi quy nghiên cứu quan hệ phụ thuộc của Y vào X

Không đòi hỏi giữa Y và X phải có quan hệ 2 chiều (nhân quả)

c. Hồi quy và tương quan:

- Phân tích tương quan đo mức độ kết hợp tuyến tính giữa hai biến
- Phân tích hồi quy ước lượng, dự báo một biến trên cơ sở giá trị đã cho của các biến khác
- Trong phân tích hồi quy, khác với tương quan, các biến không có tính đối xứng

Ví dụ

Giả sử muốn dự đoán y từ x trong bảng sau và giả sử phương trình hồi quy có thể biểu diễn $y=B_0+B_1 * x$

x	y	Predict 'y'
1	2	B0+B1*1
2	1	B0+B1*2
3	3	B0+B1*3
4	6	B0+B1*4
5	9	B0+B1*5
6	11	B0+B1*6
7	13	B0+B1*7
8	15	B0+B1*8
9	17	B0+B1*9
10	20	B0+B1*10

Ví dụ

- Các giá trị thống kê tính toán được

Độ lệch chuẩn x	3.0277
Độ lệch chuẩn y	6.6173
Trung bình x	5.5
Trung bình y	9.7
Tương quan x và y	0.890944

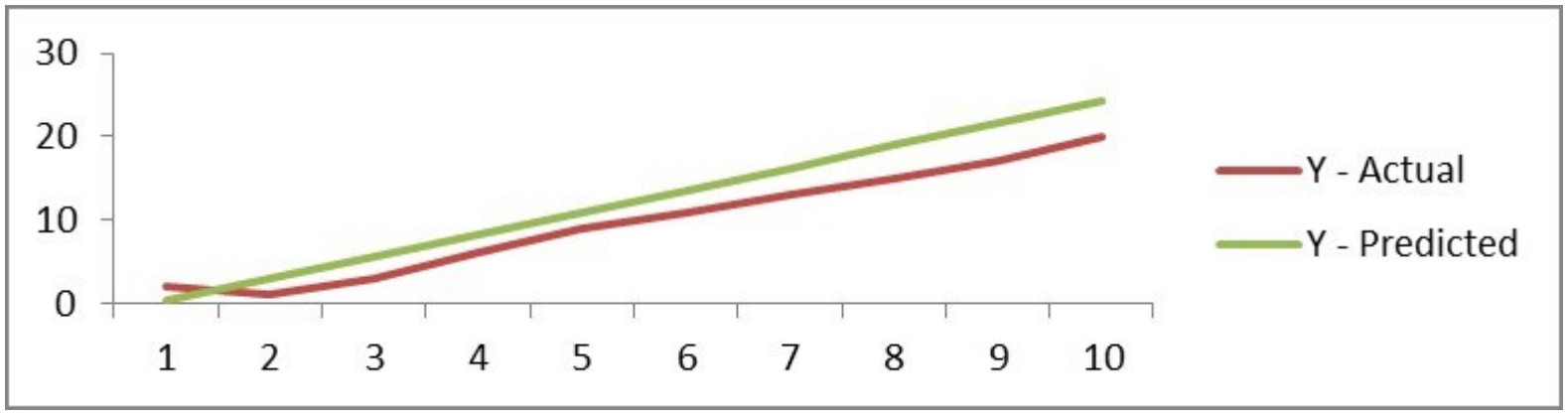
$$s = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N-1}}.$$

Hệ số tương quan $\rho_{X,\,Y}$

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E((X-\mu_X)(Y-\mu_Y))}{\sigma_X \sigma_Y}$$

Ví dụ

- Có thể tính các hệ số của phương trình mô hình
  - B1 = Tương quan*(Độ lệch chuẩn của y / Độ lệch chuẩn của x)
  - B0 = Trung bình (Y) - B1 * Trung bình (X)
- B1 = 1.947 và B0 = -1.01



2. Hồi quy tuyến tính hàm nhiều biến

- Về cơ bản không có sự khác biệt giữa hồi quy tuyến tính ‘đơn biến' và 'đa biến'.
- Cả hai đều làm việc tuân theo nguyên tắc OLS (Ordinary Least Square) và thuật toán để có được đường hồi quy tối ưu nhất cũng tương tự.

$$Y=B_0+B_1*X_1+B_2*X_2+B_3*X_3.....$$

B_i: Các hệ số khác nhau

X_i: Các biến độc lập khác nhau

2. Hồi quy tuyến tính hàm nhiều biến

- Hàm hồi quy tổng thể (PRF) của mô hình hồi quy hai biến:

$$PRF: Y_i = \beta_1 + \beta_2 X_i + U_i$$

Hay:

$$E(Y \mid X_i) = \beta_1 + \beta_2 X_i$$

Trong đó

Y : Biến phụ thuộc

Y_i : Giá trị cụ thể của biến phụ thuộc

X : Biến độc lập

X_i : Giá trị cụ thể của biến độc lập

U_i : Sai số ngẫu nhiên ứng với quan sát thứ i

3. Hồi quy Logistic

- Mục tiêu của hồi qui Logistic là nghiên cứu mối tương quan giữa một (hay nhiều) yếu tố nguy cơ (*risk factor*) và đối tượng phân tích (*outcome*).
- Ví dụ: Nghiên cứu mối tương quan giữa thói quen hút thuốc lá và nguy cơ mắc ung thư phổi thì yếu tố nguy cơ ở đây là thói quen hút thuốc lá và đối tượng phân tích ở đây là nguy cơ mắc ung thư phổi.

Hồi quy Logistic

- Trong hồi qui logistic thì các đối tượng nghiên cứu thường được thể hiện qua các biến số nhị phân (binary) như *xảy ra/ không xảy ra ; có/không,...*
- Các yếu tố nguy cơ có thể được thể hiện qua các biến số liên tục (huyết áp,…) hoặc các biến thứ bậc (thu nhập : Cao, trung bình, thấp).
- Để ước tính độ tương quan của các yếu tố nguy cơ và đối tượng phân tích.
- Các phương pháp phân tích như hồi qui tuyến tích không áp dụng được vì biến phụ thuộc không phải là biến liên tục mà là biến nhị phân.

Hồi quy Logistic

- *Ví dụ :* Bảng dữ liệu dưới đây thu thập để nghiên cứ mối tương quan giữa tình trạng phơi nhiễm chất độc gia cam (Agent Orange – AO) và ung thư tuyến tiền liệt.

	Ung thư (47)	Đối chứng (144)
Phơi nhiễm AO	11	17
Không phơi nhiễm AO	36	127

Hồi quy Logistic

- Bảng cho thấy 23.4% (11/47) người bị ung thư tuyến tiền liệt từng bị phơi nhiễm AO. Tỷ lệ này trong nhóm đối chứng là 11.8% (17/144).
- *Vấn đề đặt ra là có sự tương quan nào giữa tình trạng phơi nhiễm AO và ung thư tuyến tiền liệt hay không?*

Cần trả lời 2 vấn đề sau:

- Nguy cơ mắc bệnh ung thư tuyến tiền liệt của những người từng bị phơi nhiễm AO so với nguy cơ ở những người không từng bị phơi nhiễm là bao nhiêu?
- Sự khác biệt về nguy cơ ung thư tuyến tiền liệt giữa 2 nhóm phơi nhiễm và không phơi nhiễm AO có ý nghĩa thống kê không? (hay do ngẫu nhiên)

Tìm hiểu thêm ở đây http://bis.net.vn/forums/t/484.aspx

	Ung thư (47)	Đối chứng (144)
Phơi nhiễm AO	11	17
Không phơi nhiễm AO	36	127

Bài tập

1. Xác định hàm dự báo hồi quy tuyến tính dựa vào tập dữ liệu huấn luyện sau, tối thiểu hóa lỗi.
2. Hãy cho dùng hàm hồi quy tìm được để dự báo giá trị của Y

X	Y
7.3	1
8	2
8.7	3
7.3	1
9.3	4
10	5
10.7	6
11.3	7
12	8
12.7	9
13.3	10
14	11
10.3	5.5
11.2	6.8
9	3.5
10.3	5.4
11.7	7.5
12.3	8.5
20.7	21
26.7	30

X	Y
1	?
1.5	?
3	?
2	?
2.5	?
4	?
6	?
6.5	?
13	?
13.5	?
14	?
14.5	?
16	?
16.5	?
17	?
18	?
19	?
22	?
23	?
25	?

Q&A