Bài giảng 19: Hồi qui tuyến tính đơn giản

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia

Đại học Tôn Đức Thắng, Việt Nam

Nội dung

- Mục tiêu của mô hình hồi qui tuyến tính
- Ý tưởng đằng sau mô hình hồi qui tuyến tính
- Ước tính tham số
- Ví dụ và R

Mục tiêu của mô hình hồi qui tuyến tính

Ba phân tích chính …

- Khác biệt (analysis of difference)
- Liên quan (association analysis)
- Tương quan (correlation analysis) và tiên lượng (prediction)

Phân tích khác biệt

- t-test, ANOVA
- z-test, Chi-square

Phân tích liên quan

- Odds ratio
- Risk ratio
- Prevalence ratio
- etc

Phân tích tương quan và tiên lượng

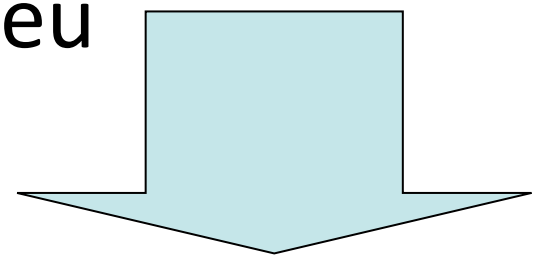
- Correlation analysis
- Linear regression analysis
- Logistic regression
- Cox's regression
- etc

Phân tích tương quan

- Đánh giá mối tương quan
- Hệ số tương quan (coefficient of correlation)
- Chúng ta cần biết thêm …
  - Mức độ ảnh hưởng của biến tiên lượng (predictor variable) trên biến phụ thuộc (dependent variable)
  - Tiên lượng

Mục tiêu của mô hình hồi qui tuyến tính

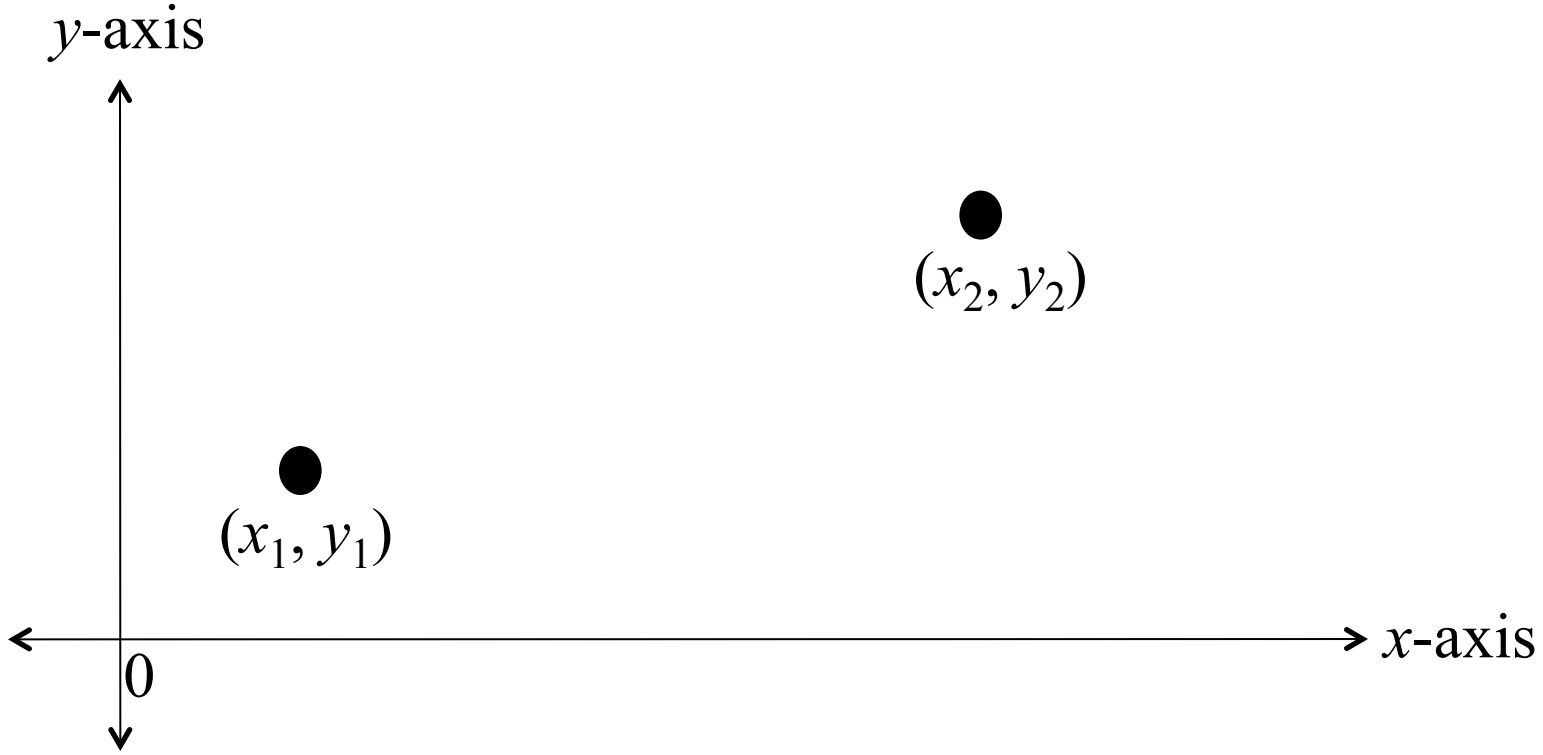
- Tìm một mô hình (phương trình) để mô tả một mối liên quan giữa X và Y
  - X có thể là độ tuổi, trọng lượng
  - Y có thể là BMD
- Điều chỉnh các yếu tố nhiễu
- Tiên lượng



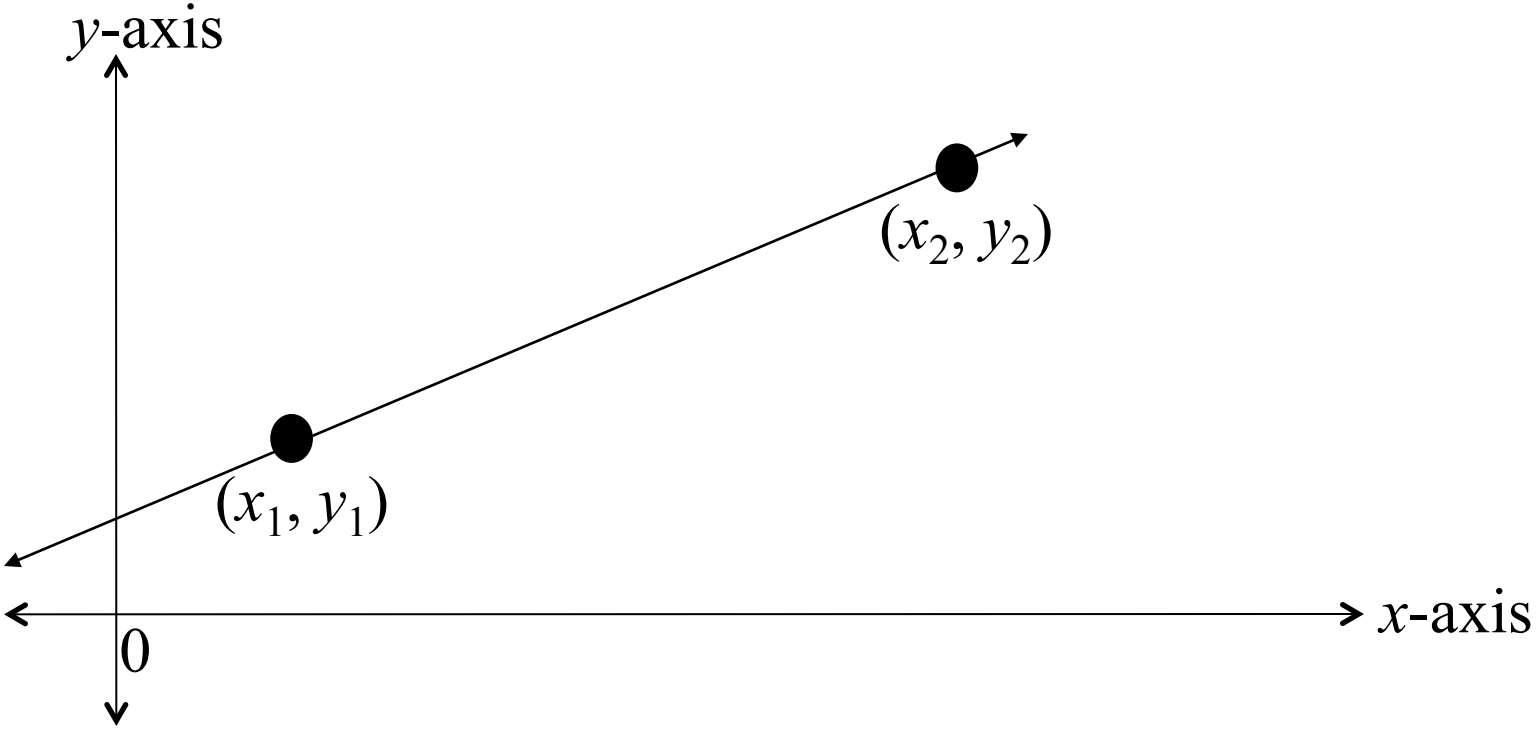
Linear regression model

Ý tưởng đằng sau mô hình hồi qui tuyến tính

Cho hai điểm (x_1, y_1) và (x_2, y_2)



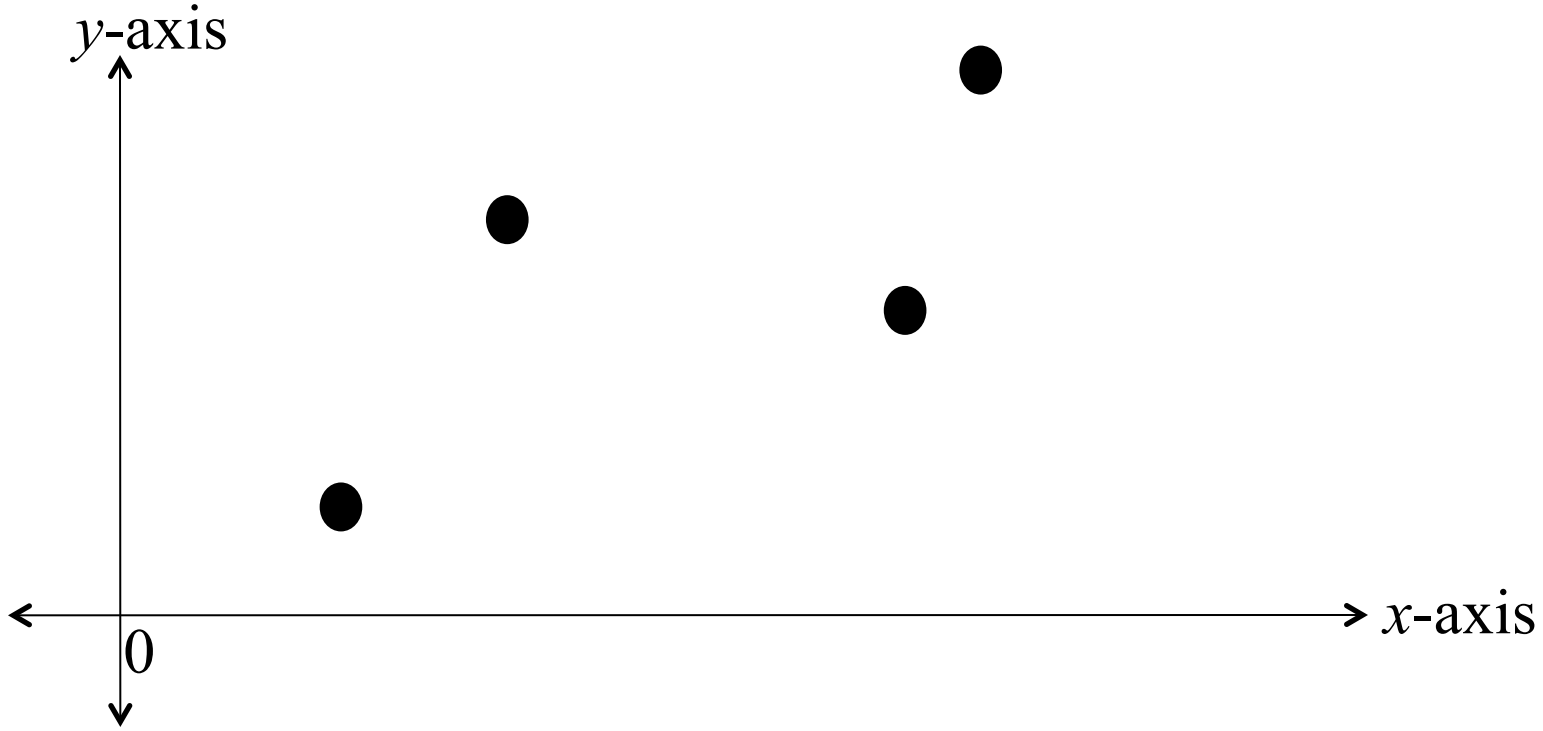
Làm sao để "phát triển" một phương trình nối 2 điểm này?



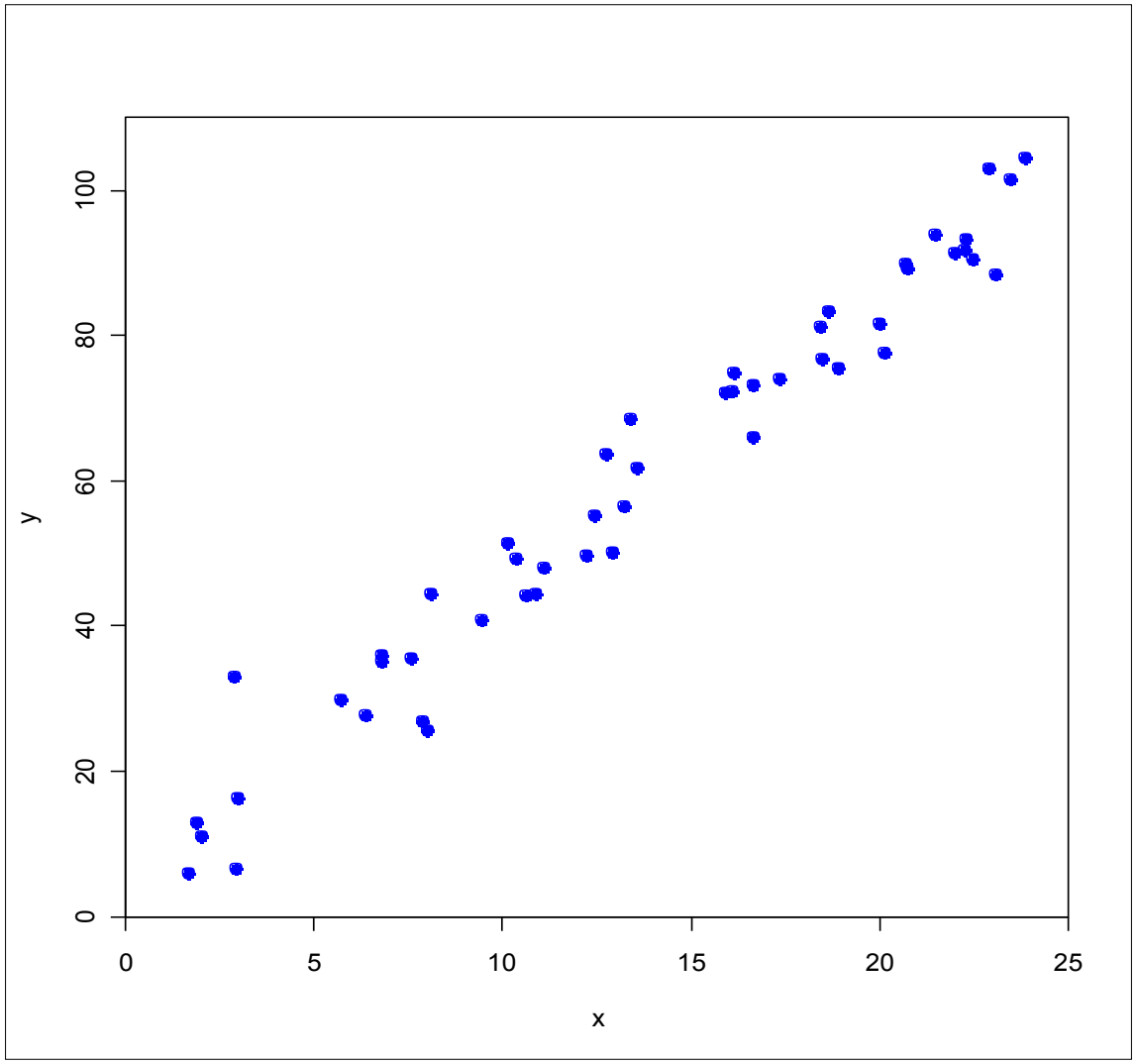
$$slope = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

- Tìm gradient (slope)
- Tìm giá trị khởi đầu (intercept) của y khi x=0

Nhưng nếu có nhiều điểm ...



và rất nhiều điểm



Mô hình hồi qui tuyến tính

- **Simple linear regression model**
- ***Y*** **-** biến phụ thuộc (response variable, dependent variable, v.v.)
  - *Y* là biến liên tục
- ***X*** - biến độc lập (predictor variable, independent variable, v.v.)
  - *X* là biến liên tục hay không liên tục

Mô hình hồi qui tuyến tính

- Mô hình:

$$Y = \alpha + \beta X + \varepsilon$$

α : intercept

β : slope / gradient

ε : sai số ngẫu nhiên (random error – những dao động về Y trong mỗi giá trị X)

Giả định

- Mối liên quan giữa X và Y là tuyến tính (linear) về *tham số*
- *X* không có sai số ngẫu nhiên
- Giá trị của *Y* độc lập với nhau (vd, Y_1 không liên quan với Y_2) ;
- Sai số ngẫu nhiên (e): phân bố chuẩn, trung bình 0, phương sai bất biến

$$\varepsilon \sim N(0, \sigma^2)$$

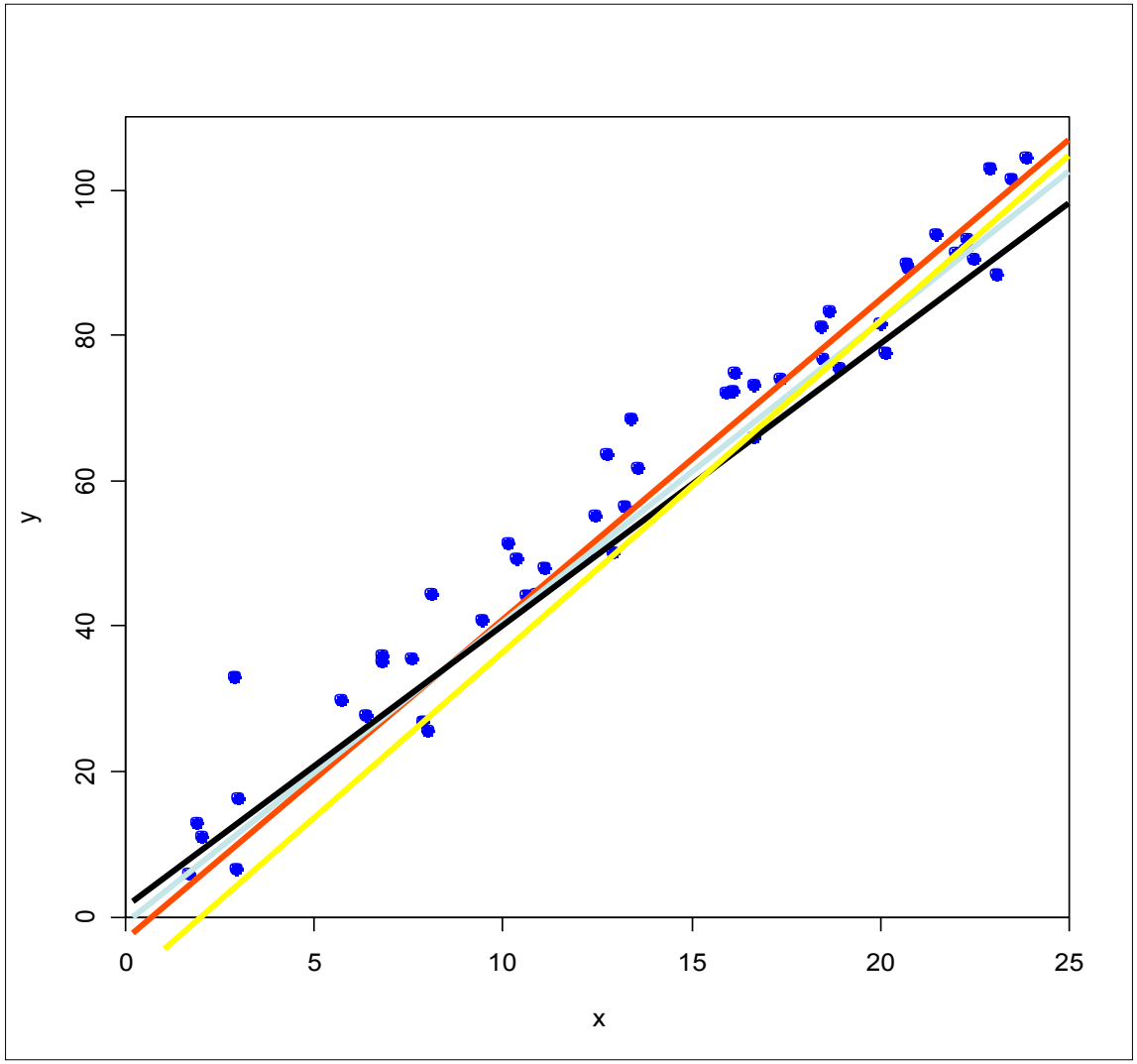
Ước tính tham số (parameters)

Mục tiêu

- Mô hình

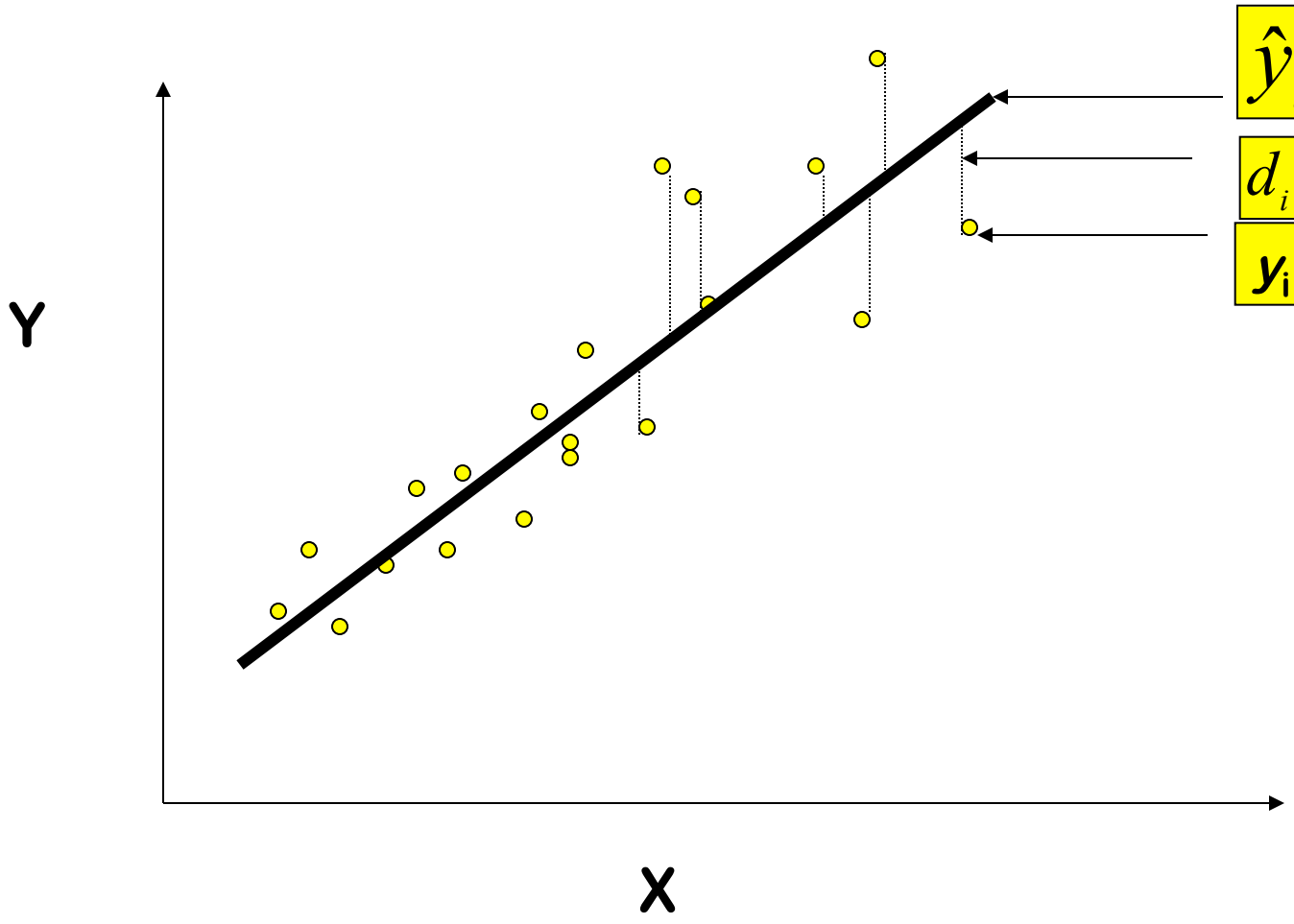
$$Y = \alpha + \beta X + \varepsilon$$

- Chúng ta không biết α và β
- Nhưng có thể dùng dữ liệu thí nghiệm / thực tế để ước tính 2 tham số đó
- Ước số (estimate) của α và β là *a* và *b*



- Có thể tính bằng mắt
- Nhưng không khách quan (biased)
- Chúng ta cần sự nhất quán -- consistency

Tiêu chuẩn để tìm tham số



Tìm công thức (estimator) để tính a và b sao cho tổng d^2 là nhỏ nhất →
Least square method = Bình phương nhỏ nhất

Ước tính bằng R

- Chúng ta muốn ước tính mối liên quan giữa BMD và trọng lượng
- Mô hình hồi qui tuyến tính:

$$BMD = \alpha + \beta * weight + \varepsilon$$

- R

```
lm(bmd ~ weight)
```

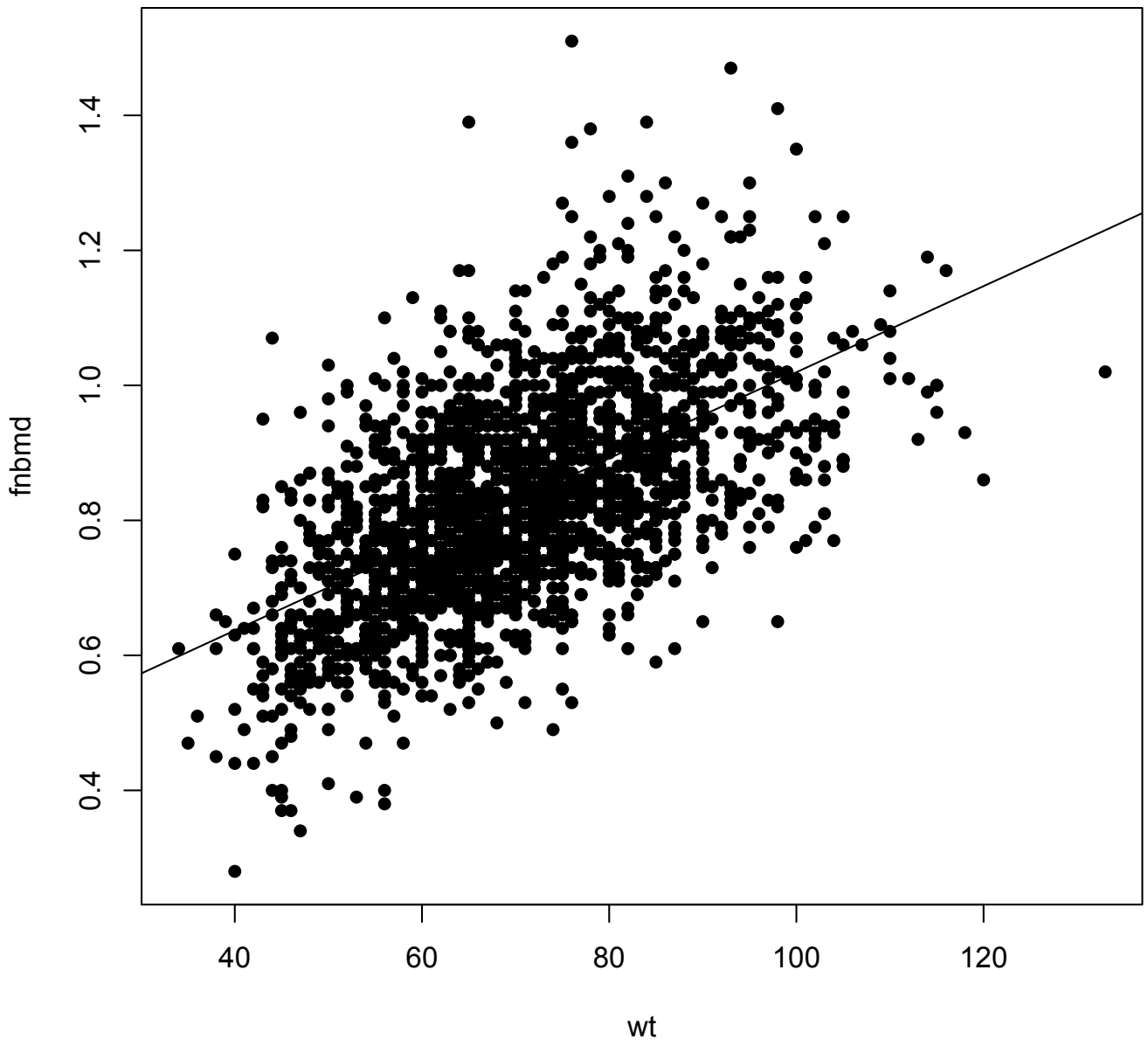
Phân tích bằng R

```
dat = read.csv("http://statistics.vn/data/
  does_vn07.csv",header=T)
attach(dat)
# Phân tích hồi qui tuyến tính
m1 = lm(fnbmd ~ wt)
summary(m1)
# vẽ biểu đồ
plot(fnbmd ~ wt, pch=16)
abline(m1)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.36371 -0.08817 -0.00733  0.07918  0.64354

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3818873  0.0138582   27.56   <2e-16 ***
wt          0.0063760  0.0001939   32.89   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
‘ ’ 1

Residual standard error: 0.1259 on 2120 degrees of
freedom (94 observations deleted due to missingness)
Multiple R-squared:  0.3379,  Adjusted R-squared:  0.3375
F-statistic:  1082 on 1 and 2120 DF,  p-value: < 2.2e-16
```


fnbmd
0.4
0.6
0.8
1.0
1.2
1.4
40
60
80
100
120
wt

Diễn giải kết quả

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.3818873  0.0138582   27.56   <2e-16 ***
wt          0.0063760  0.0001939   32.89   <2e-16 ***
```

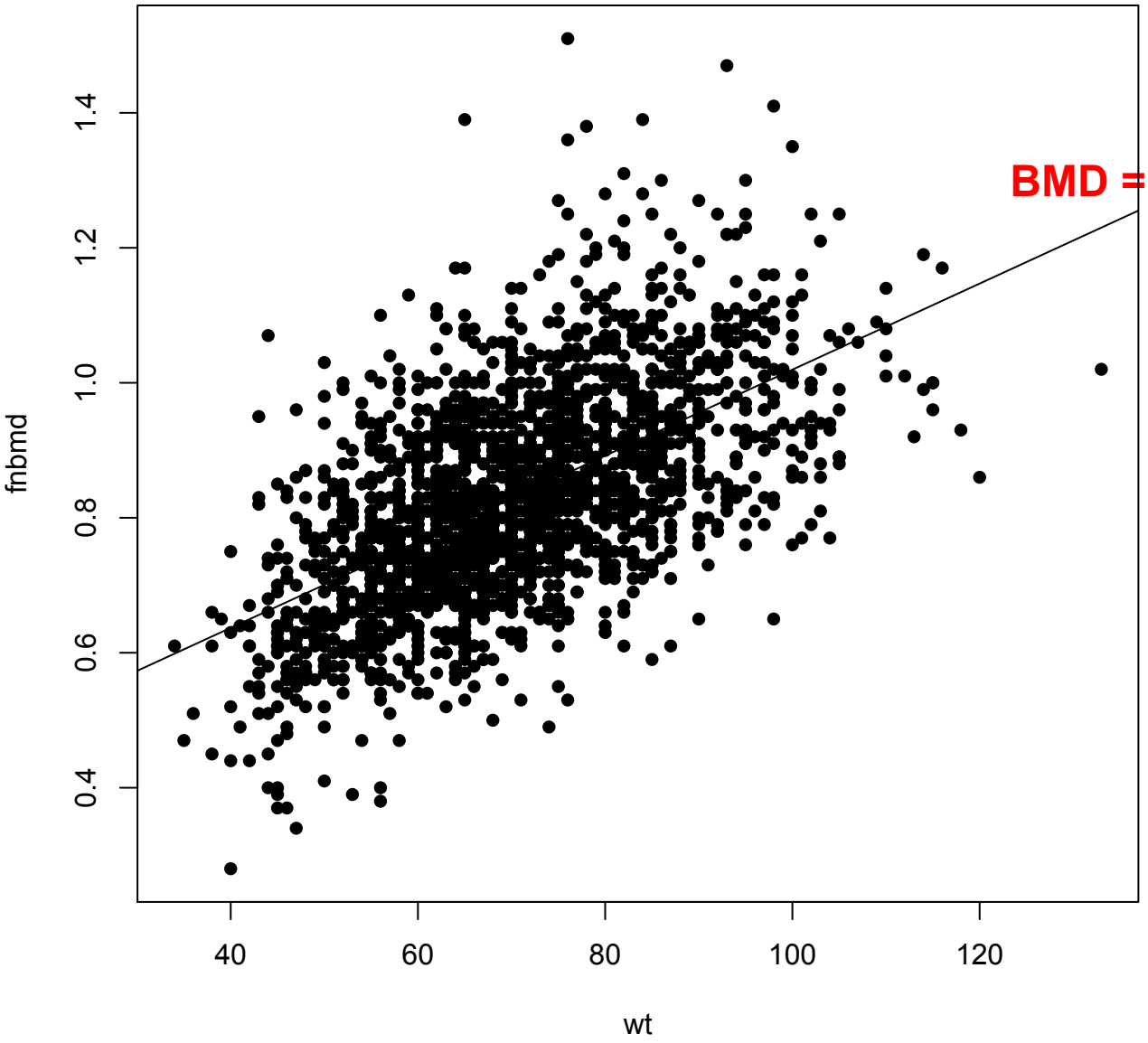
- Nhớ rằng mô hình là:

BMD = a + b*weight

- Phương trình:

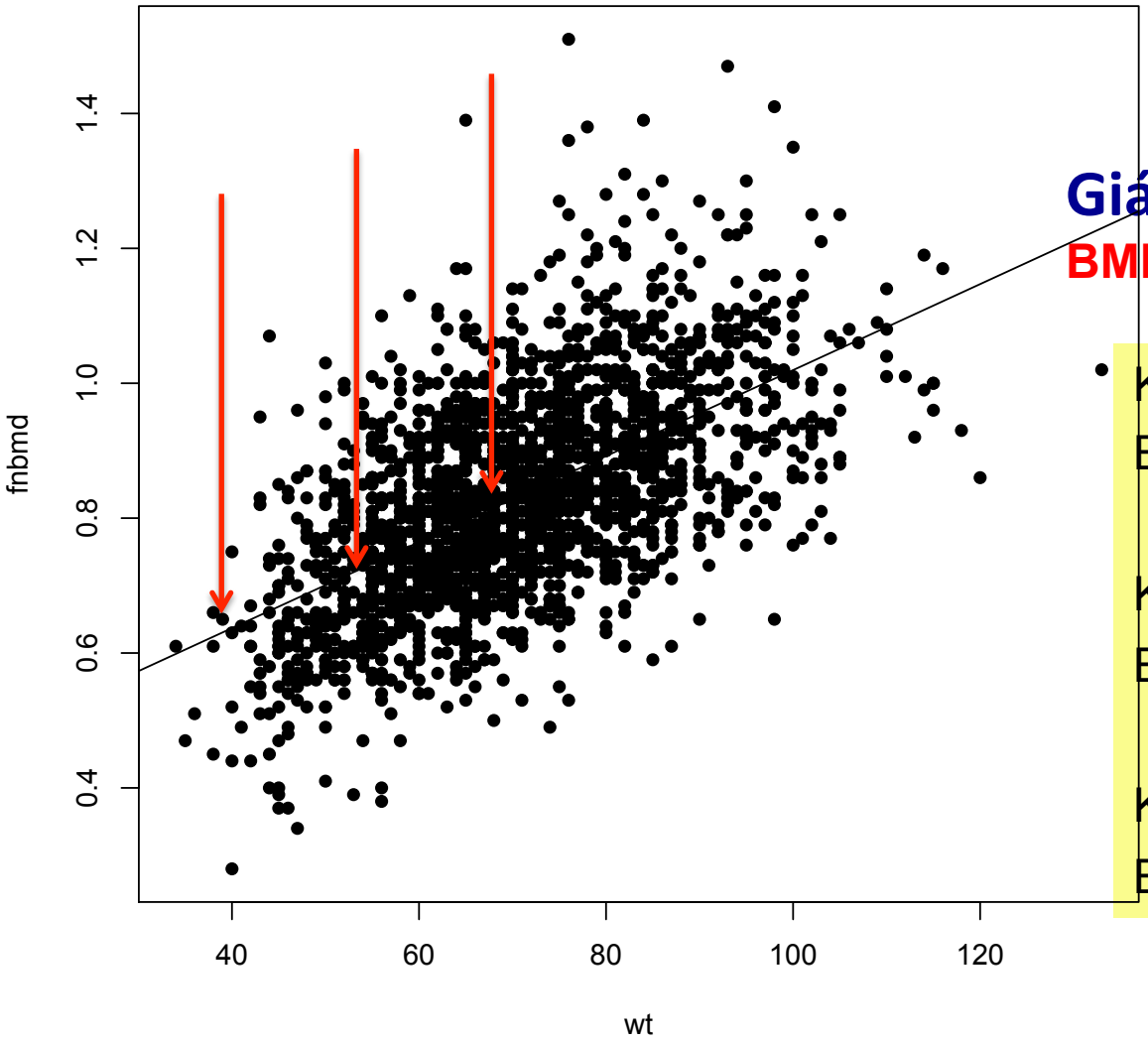
BMD = 0.38 + 0.0064*weight

- Ý nghĩa: người có trọng lượng tăng **1 kg thì mật độ xương tăng 0.0064 g/cm^2. Mối tương quan này có ý nghĩa thống kê (P < 0.0001)**



BMD = 0.38 + 0.0064*weight

Ý nghĩa của đường biểu diễn



Giá trị trung bình (kì vọng)

BMD = 0.38 + 0.0064*weight

Khi **weight = 40**,

BMD = 0.38 + 0.0064*40 = **0.64**

Khi **weight = 50**

BMD = 0.38 + 0.0064*50 = **0.70**

Khi **weight = 70**

BMD = 0.38 + 0.0064*70 = **0.83**

Tóm lược

- Mô hình hồi qui tuyến tính
  - Hiểu mức độ ảnh hưởng của biến tiên lượng
  - Tiên lượng
- Thích hợp khi mối liên quan là tuyến tính
- Hàm R:

lm(y ~ x)