

# Giới thiệu mô hình hồi qui tuyển tính

**Tuan V. Nguyen**

Garvan Institute of Medical Research

University of New South Wales (UNSW Sydney), Australia

University of Technology, Sydney (UTS), Australia

Ton Duc Thang University, Vietnam



# Một thoáng lịch sử ...

- Hệ số tương quan là một ý tưởng của Francis Galton
- Karl Pearson chỉ ra cách tính hệ số tương quan
- Francis Galton dùng chữ “Regression” đầu tiên vào thế kỉ 19 (1875)
- Regression: mô tả hiện tượng chiều cao của con [của những cha mẹ có chiều cao cao] có xu hướng quay về số trung bình quần thể
- (Trước đó, Carl F Gauss khám phá regression, nhưng ông xem là ‘chuyện nhỏ’)

# Ý tưởng của mô hình hồi qui

- Tìm một phương trình để mô tả mối liên quan tuyến tính giữa X và Y
  - X là biến độc lập (vd predictor, *independent* variable)
  - Y là biến phụ thuộc (vd outcome variable, *dependent* variable)
- Mô hình có thể có nhiều biến độc lập → hiệu chỉnh cho yếu tố nhiễu (*confounding factors*)
- Tiên lượng

# Ý tưởng mô hình hồi qui tuyến tính

- Biến phụ thuộc (Y) phải là biến liên tục (vd: pcfat)
- Biến tiên lượng (X) hay predictor variables: không giới hạn (vd: giới tính, tuổi)
- Hồi qui tuyến tính đơn giản (simple linear regression model)
  - có một biến tiên lượng

# Mô hình hồi qui tuyến tính

Gọi  $Y_i$  là giá trị quan sát của cá nhân  $i$ , và  $X_i$  là biến độc lập, mô hình hồi qui tuyến tính phát biểu:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

$\alpha$  : intercept (giá trị của  $Y$  khi  $X = 0$ )

$\beta$  : slope / gradient

$\epsilon$  : sai số ngẫu nhiên (random error)

# Giả định mô hình hồi qui tuyến tính

- 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 ( $\varepsilon$ ): phân bố chuẩn, trung bình 0, phương sai bất biến

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

# Tham số của mô hình hồi qui tuyến tính

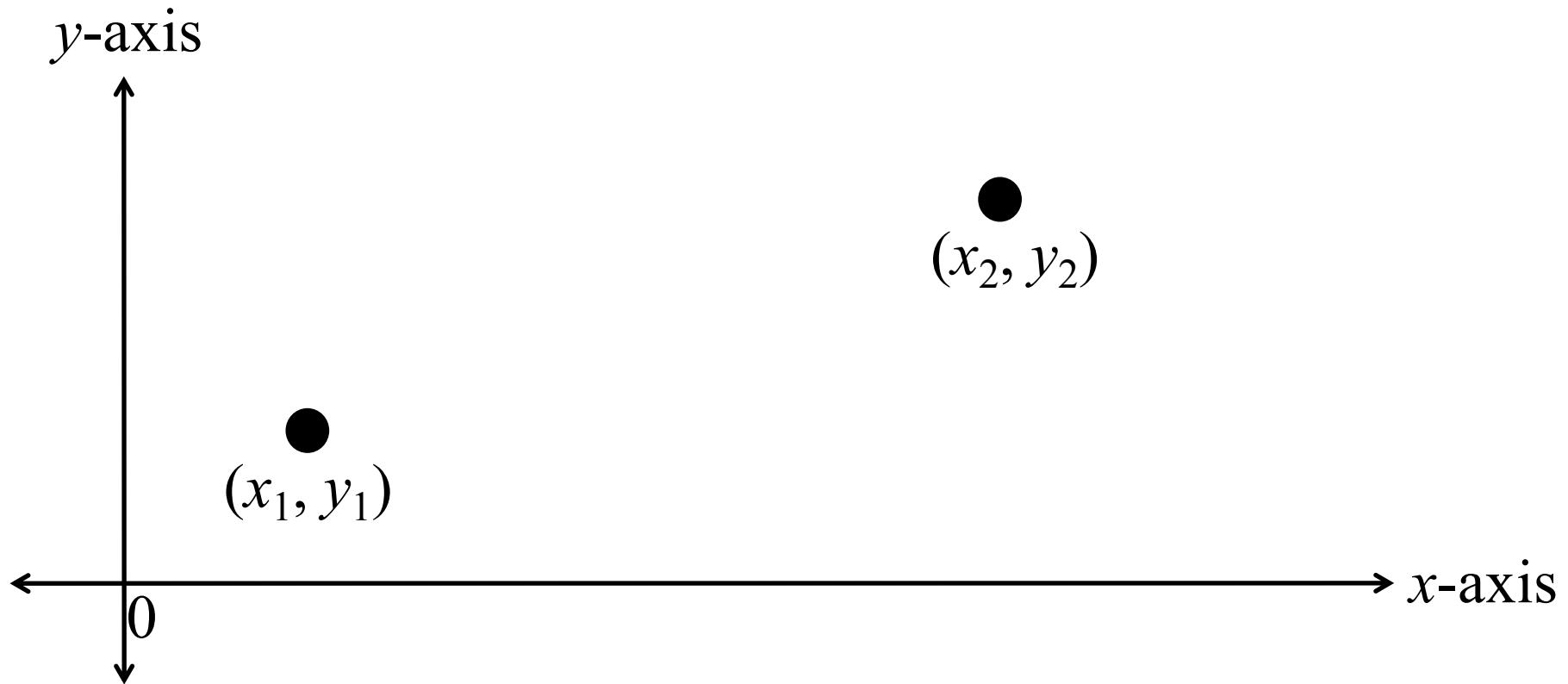
- Mô hình

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

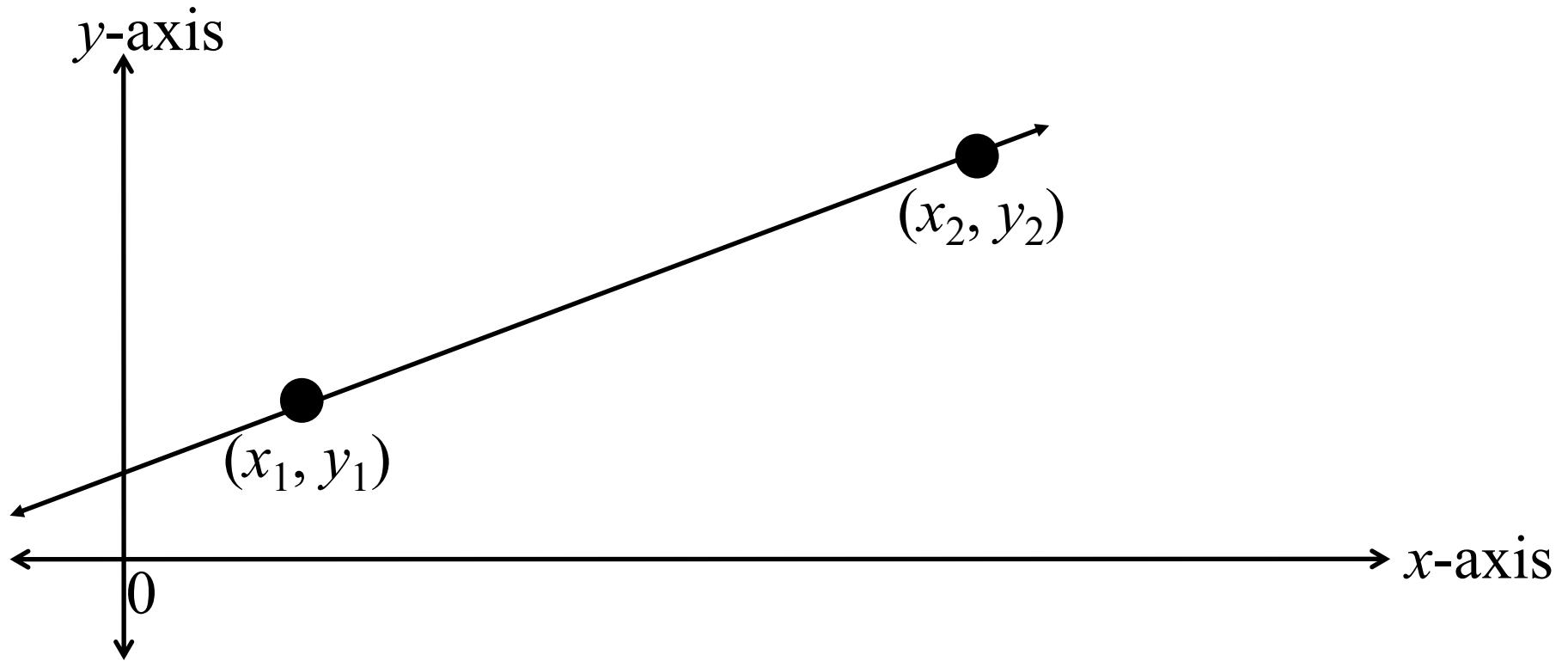
- Chúng ta không biết tham số  $\alpha$  và  $\beta$
- Nhưng có thể dùng dữ liệu thí nghiệm / thực tế để ước tính 2 tham số đó
- **Ước số (estimate)** của  $\alpha$  và  $\beta$  là  $a$  và  $b$

$$Y_i = a + bX_i + e_i$$

Với 2 điểm trong không gian  $(x_1, y_1)$  và  $(x_2, y_2)$



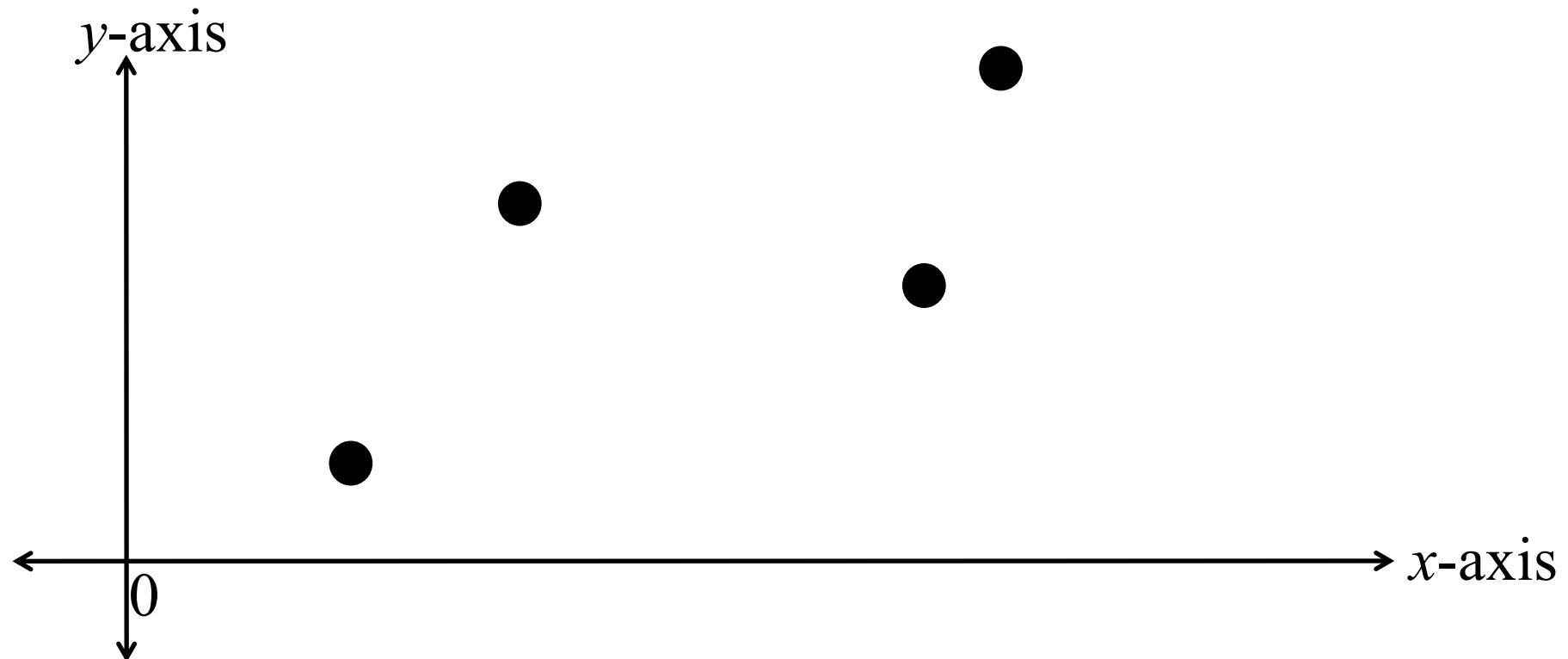
Làm sao tìm một phương trình để nối 2 điểm?



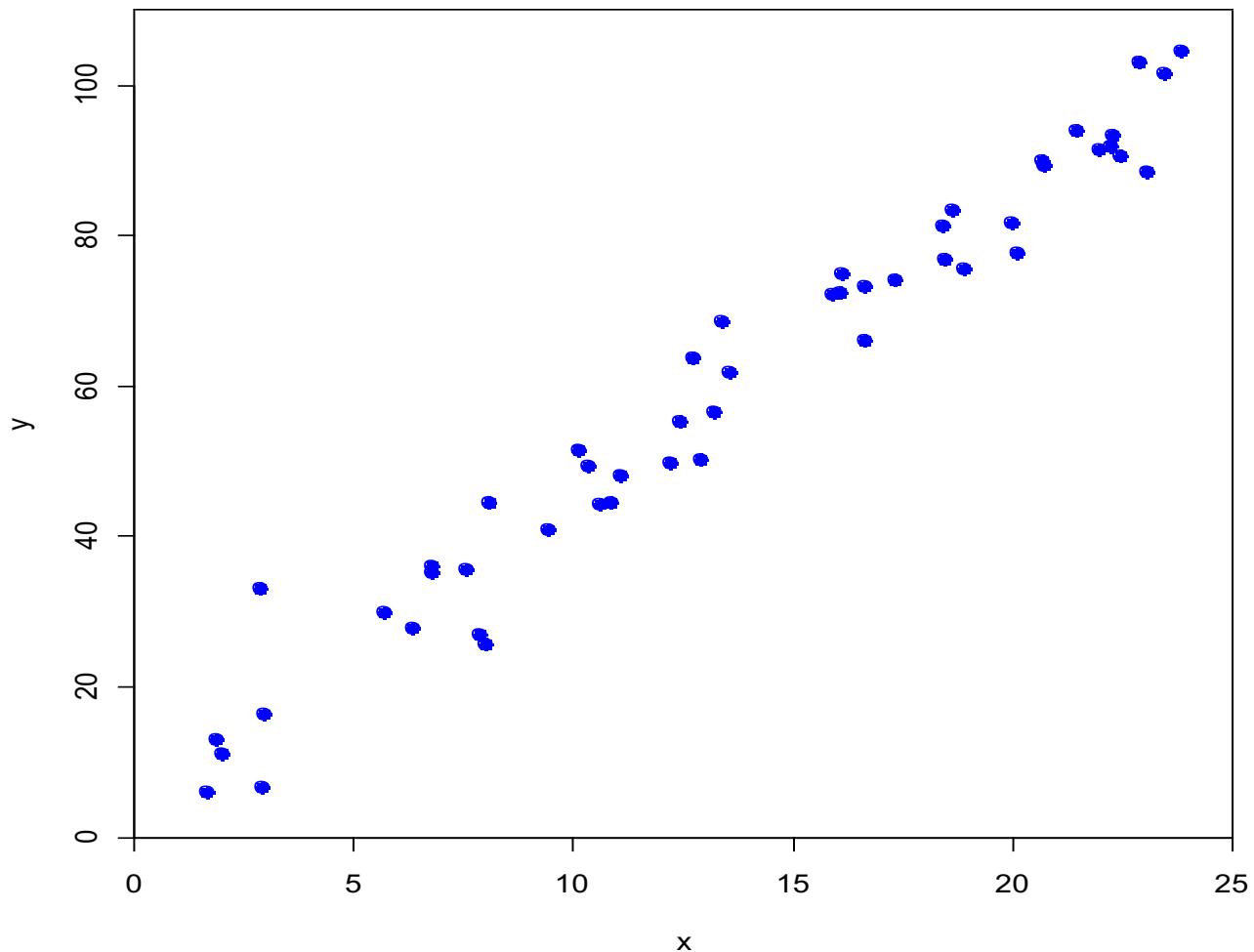
$$slope = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$$

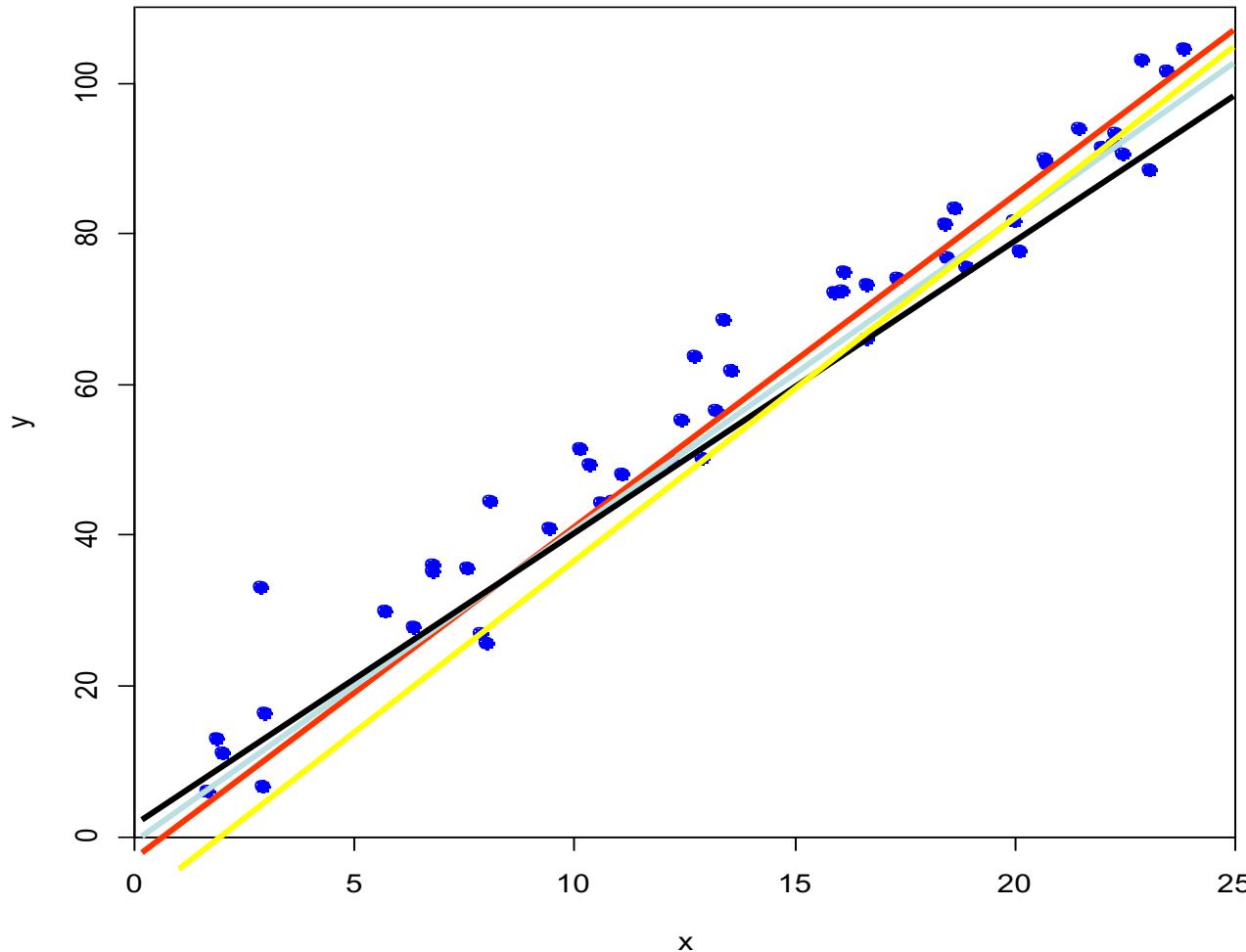
- Tìm **slope (gradient)**
- Tìm điểm khởi đầu **intercept** (giá trị của y khi  $x = 0$ )

# Nhưng chúng ta có *nhiều* điểm ...



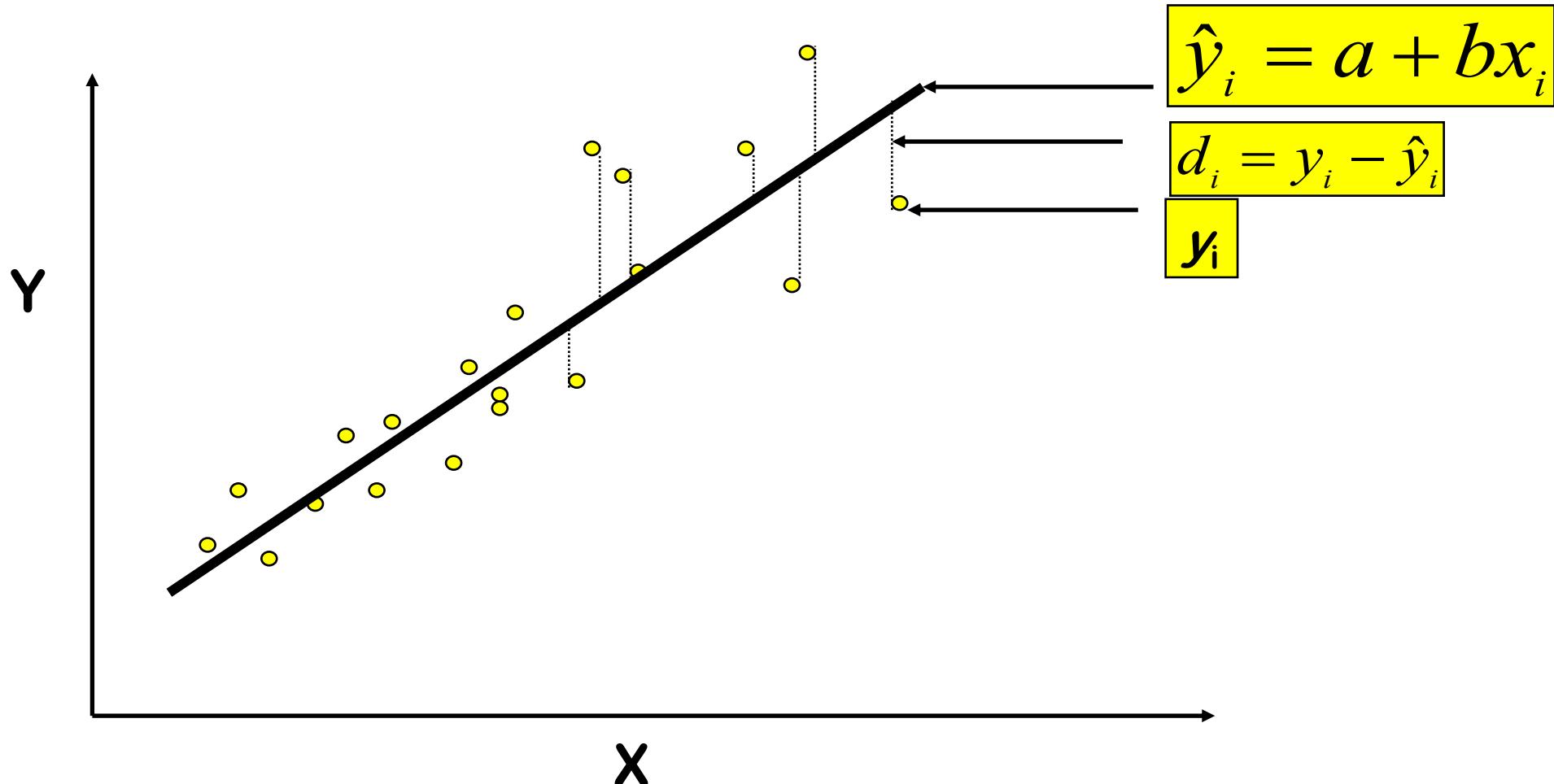
# và rất nhiều điểm ...





- Có thể dùng mắt vẽ nhiều đường thẳng qua các điểm
- Nhưng có thể biased và inconsistent
- Chúng ta cần một phương pháp tránh tình trạng bias và inconsistent

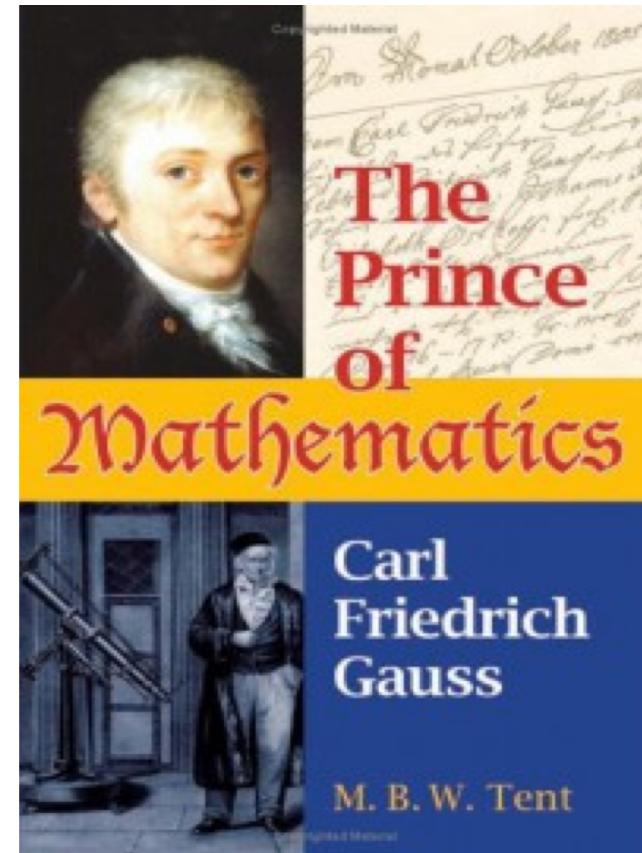
# Tiêu chuẩn ước tính tham số



Tìm một công thức (estimator) để ước tính  $a$  và  $b$  sao cho tổng bình phương  $d^2$  là thấp nhất → **Least square method**

# Carl Friedrich Gauss (1777 – 1855)

- Sanh Brunswick, Đức
- Thiên tài toán học vĩ đại nhất, "*the greatest mathematician since antiquity*"
- Sáng chế phương pháp "least square"



# Estimators

- Cho một dãy số  $X_i, Y_i$

$$b = \frac{s_{XY}}{s_X^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})}$$

$$a = \bar{Y} - b\bar{X}$$

- Và, mô hình / phương trình tuyến tính là:

$$\hat{Y}_i = a + bX_i$$

# Dùng hàm 'lm' trong R ước tính tham số

- Mô hình hồi qui tuyến tính:

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

- Triển khai bằng R qua hàm "lm" (linear model)

**lm(y ~ x)**

# Dữ liệu lịch sử của Francis Galton

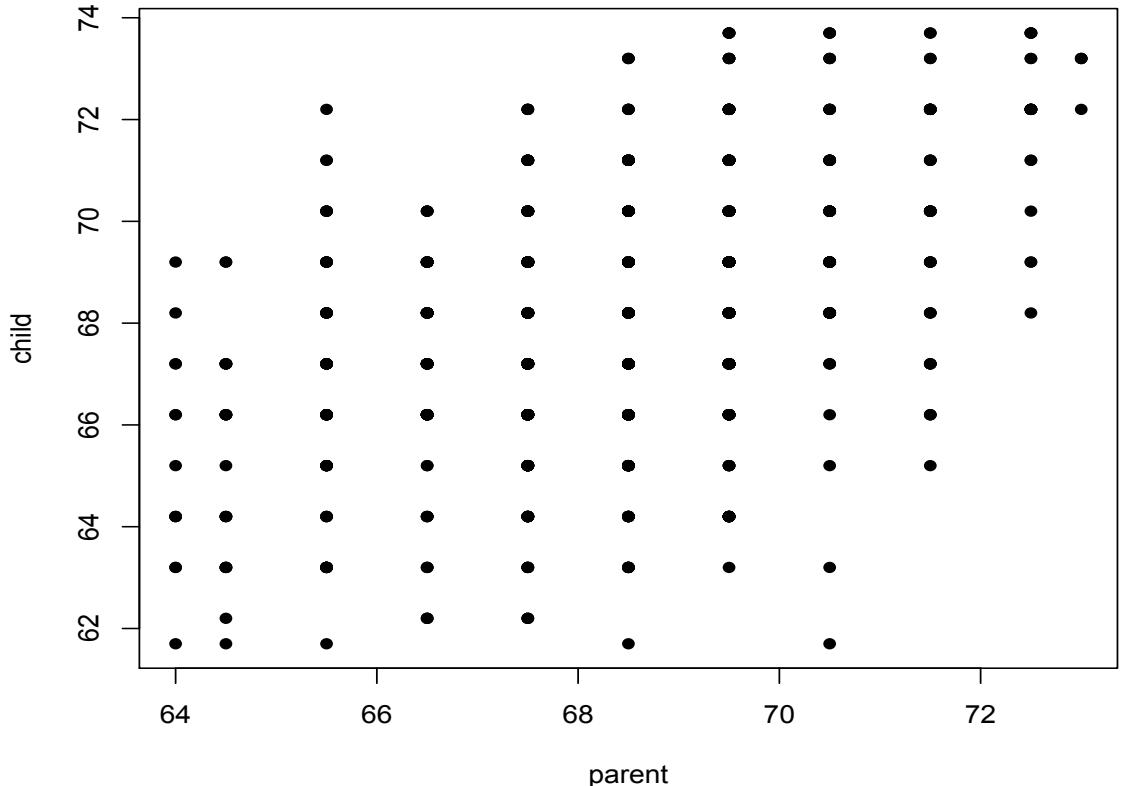
- Galton F (1869). *Hereditary Genius: An Inquiry into its Laws and Consequences*. London: Macmillan
- Dữ liệu thu thập từ 928 người con của 205 cặp cha mẹ
- Dữ liệu
  - mid-parent's height (chiều cao trung bình của cha và mẹ)
  - child's height (chiều cao của con)

# Dữ liệu lịch sử của Francis Galton

```
galton = read.csv("Galton  
data.csv", header=T)
```

```
head(galton)
```

	<code>id</code>	<code>parent</code>	<code>child</code>
1	1	70.5	61.7
2	2	68.5	61.7
3	3	65.5	61.7
4	4	64.5	61.7
5	5	64.0	61.7
6	6	67.5	62.2



# Mô hình mô tả mối liên quan giữa chiều cao cha mẹ và con

- Câu hỏi nghiên cứu: có mối liên quan giữa chiều cao của cha mẹ và con?
- Phát biểu thống kê:

$$\text{Child} = \alpha + \beta \cdot \text{Parent} + \varepsilon$$

- Phát biểu bằng R:

```
m = lm(child ~ parent, data=galton)
```

```
> m = lm(child ~ parent, data=galton)
> summary(m)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	23.94153	2.81088	8.517	<2e-16 ***
parent	0.64629	0.04114	15.711	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.239 on 926 degrees of freedom  
Multiple R-squared: 0.2105, Adjusted R-squared: 0.2096  
F-statistic: 246.8 on 1 and 926 DF, p-value: < 2.2e-16

# Ý nghĩa của kết quả phân tích

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	23.94153	2.81088	8.517	<2e-16 ***	intercept
parent	0.64629	0.04114	15.711	<2e-16 ***	slope

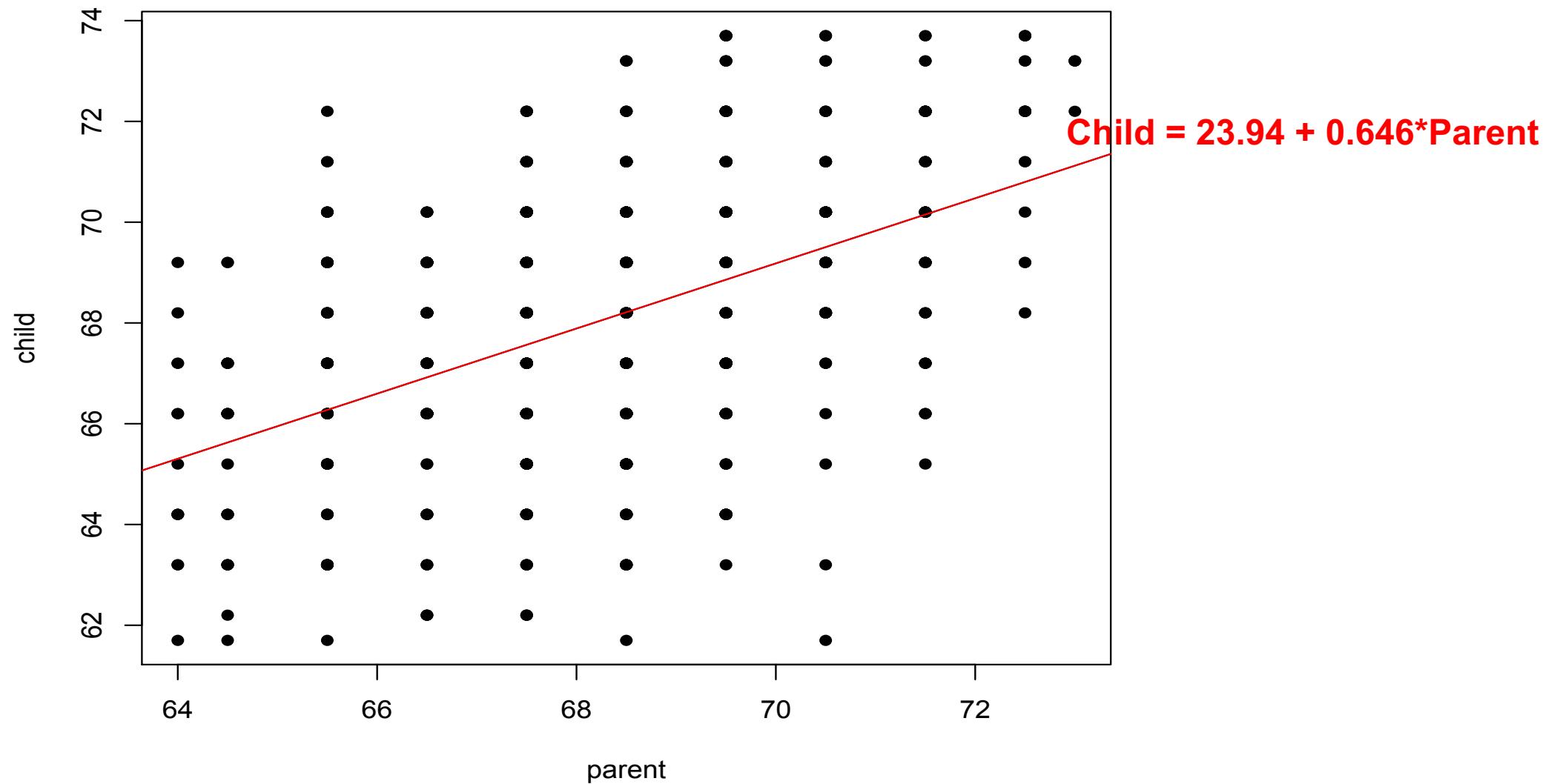
- Mô hình của chúng ta là:

$$\text{Child} = a + b * \text{Parent}$$

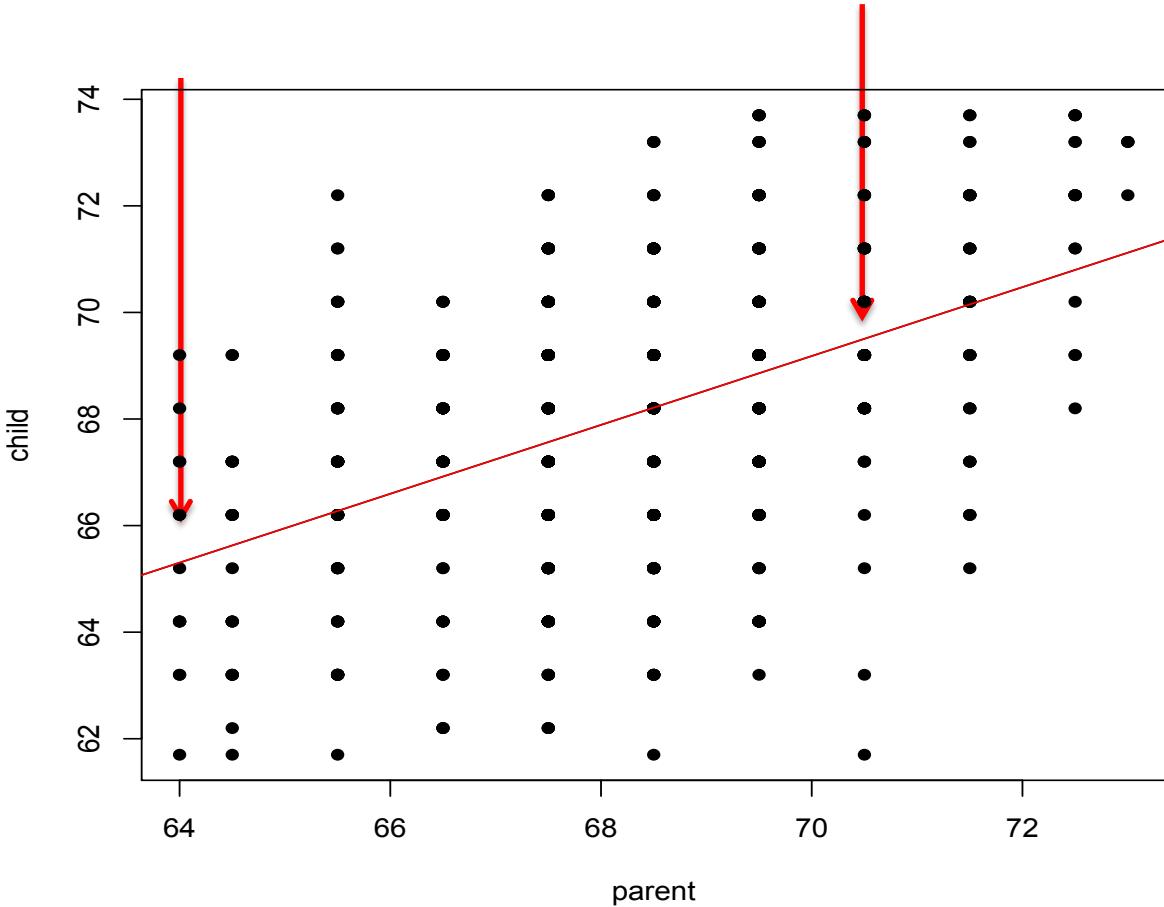
- Phương trình bây giờ là:

$$\text{Child} = 23.94 + 0.646 * \text{Parent}$$

- Điễn giải: *mỗi cm tăng chiều cao của cha/mẹ có liên quan đến 0.646 cm tăng chiều cao của con.*



# Ý nghĩa của mô hình tuyến tính



Expected value (trung bình)

$$\text{Child} = 23.94 + 0.646 * \text{Parent}$$

When parental height = 64 in

$$\text{Child} = 23.94 + 0.646 * 64 = 65.3$$

When parental height = 70

$$\text{Child} = 23.94 + 0.646 * 70 = 69.2$$

# Kiểm tra giả định

# 4 giả định của mô hình hồi qui tuyến tính

- Mỗi 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 ( $\epsilon$ ): phân bố chuẩn, trung bình 0, phương sai bất biến

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

# Đao động dư -- residual (e)

**Giá trị quan sát = Giá trị tiên lượng + Phần dư**

- Residual = phần dư của mô hình tuyến tính
- Giá trị tiên lượng =  $23.94 + 0.646 * \text{Parent}$
- $e = \text{Giá trị quan sát} - \text{Giá trị tiên lượng}$

tính cho mỗi đối tượng

# Residual và predicted values dùng R

```
m = lm(child ~ parent, data=galton)
res = resid(m)
pred = predict(m)

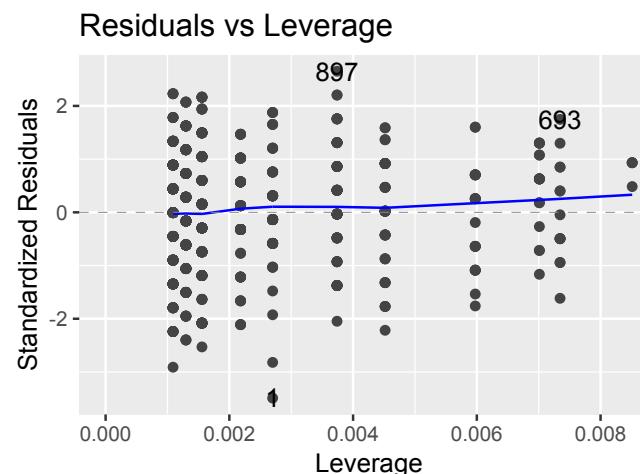
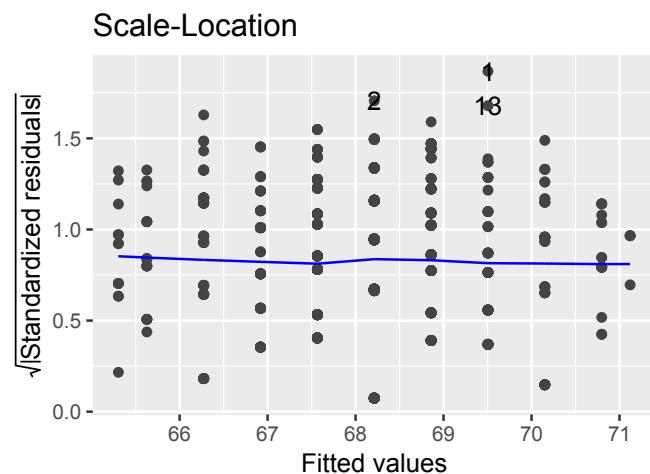
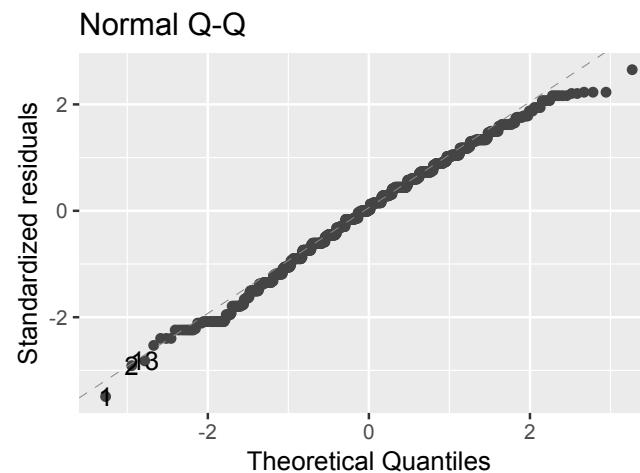
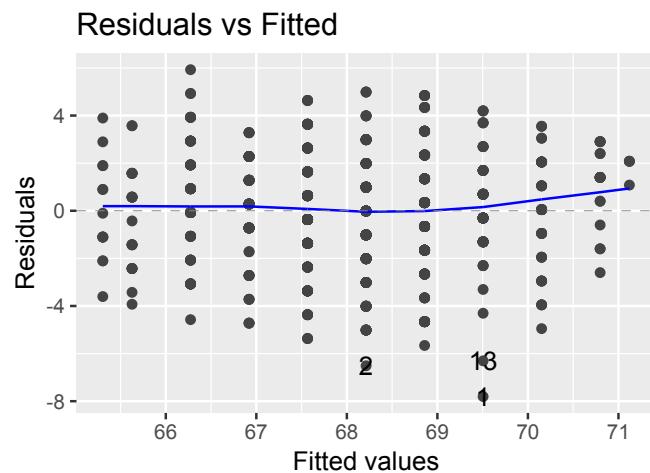
> cbind(parent, child, pred, res)
  parent child     pred      res
1    70.5  61.7 69.50502 -7.80501621
2    68.5  61.7 68.21244 -6.51243505
3    65.5  61.7 66.27356 -4.57356330
4    64.5  61.7 65.62727 -3.92727272
5    64.0  61.7 65.30413 -3.60412743
6    67.5  62.2 67.56614 -5.36614446
7    67.5  62.2 67.56614 -5.36614446
8    67.5  62.2 67.56614 -5.36614446
9    66.5  62.2 66.91985 -4.71985388
10   66.5  62.2 66.91985 -4.71985388
11   66.5  62.2 66.91985 -4.71985388
```

# Kiểm tra giả định mô hình hồi qui tuyến tính

- Có thể dùng hàm `autoplot` trong package **ggfortify**
  - normal Q-Q plot of residuals: kiểm tra giả định phân bố)
  - residual vs predicted values: kiểm tra giả định đồng dạng
  - residual vs leverage: kiểm tra giá trị ngoại vi

# Kiểm tra giả định mô hình bằng R

```
library(ggfortify)  
m = lm(child ~ parent, data=galton)  
autoplot(m)
```



# Ví dụ về mô hình có vấn đề (không đáp ứng giả định)

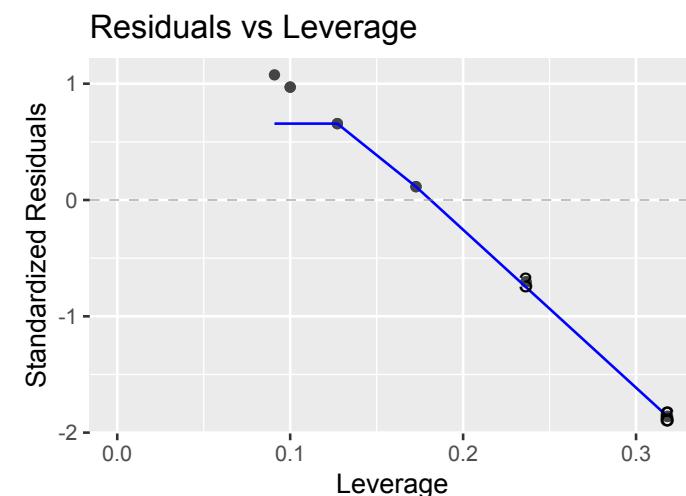
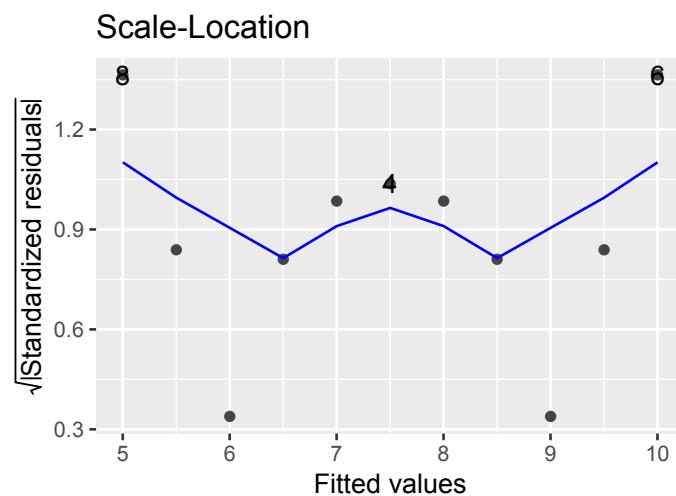
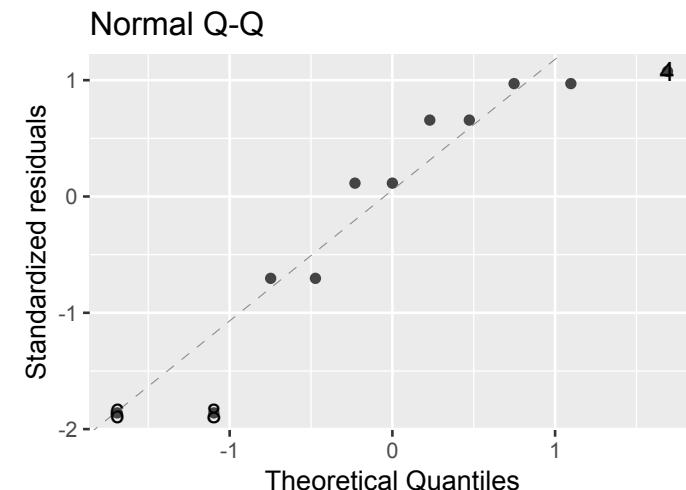
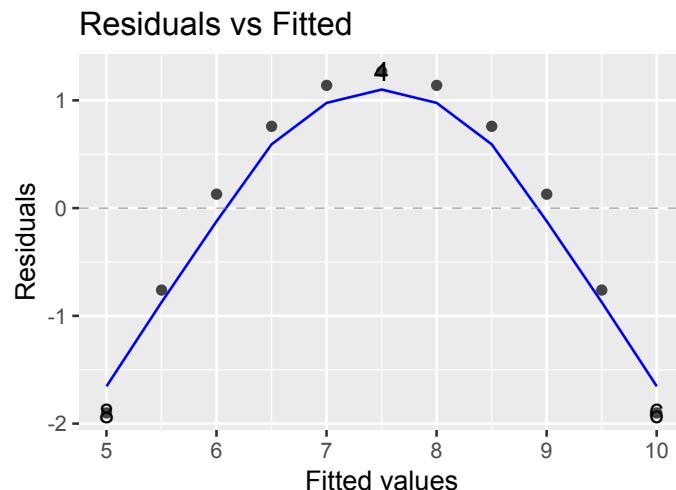
```
# Lấy dữ liệu anscombe từ dplyr
```

```
library(dplyr)
data(anscombe)
anscombe
```

	x1	x2	x3	x4	y1	y2	y3	y4
1	10	10	10	8	8.04	9.14	7.46	6.58
2	8	8	8	8	6.95	8.14	6.77	5.76
3	13	13	13	8	7.58	8.74	12.74	7.71
4	9	9	9	8	8.81	8.77	7.11	8.84
5	11	11	11	8	8.33	9.26	7.81	8.47
6	14	14	14	8	9.96	8.10	8.84	7.04
7	6	6	6	8	7.24	6.13	6.08	5.25
8	4	4	4	19	4.26	3.10	5.39	12.50
9	12	12	12	8	10.84	9.13	8.15	5.56
10	7	7	7	8	4.82	7.26	6.42	7.91
11	5	5	5	8	5.68	4.74	5.73	6.89

# Ví dụ về mô hình có vấn đề (không đáp ứng giả định)

```
library(ggfortify)  
m = lm(y2 ~ x2, data=anscombe)  
autoplot(m)
```



# Ví dụ về mô hình có vấn đề (không đáp ứng giả định)

```
m = lm(y2 ~ x2, data=anscombe)
summary(m)
plot(y2 ~ x2, pch=16, data=anscombe)
abline(m, col="red")
```

-----

```
> summary(m)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	3.001	1.125	2.667	0.02576 *
x2	0.500	0.118	4.239	0.00218 **

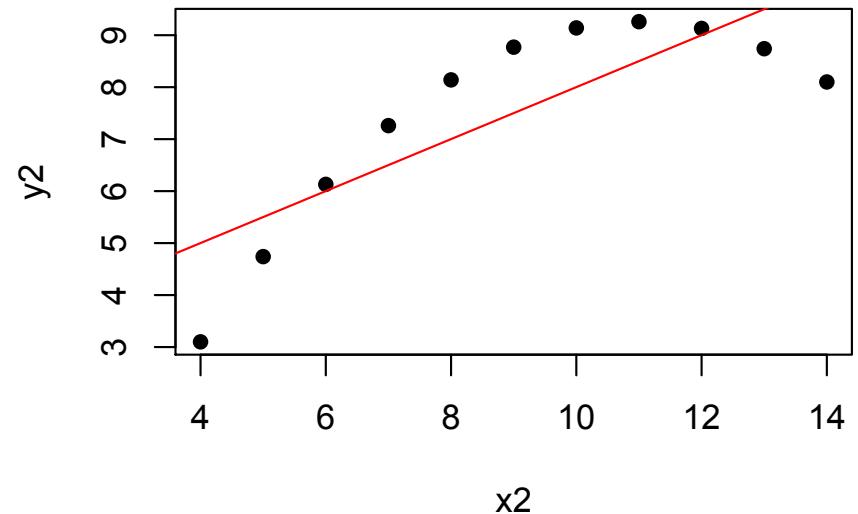
---

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 1.237 on 9 degrees of freedom

Multiple R-squared: 0.6662, Adjusted R-squared: 0.6292

F-statistic: 17.97 on 1 and 9 DF, p-value: 0.002179



# Mô hình hồi qui tuyến tính đơn giản: tóm lược

- Regression - mở rộng từ khái niệm correlation
- Mô hình hồi qui tuyến tính đơn giản chỉ có 1 biến  $X$ :  $Y = \alpha + \beta X + \varepsilon$
- Giả định mô hình:  $\varepsilon \sim N(0, \sigma^2)$
- Ước tính tham số qua hàm ‘lm’ trong R