

Regression model:

$$y = f(x_1, \dots, x_p) + \text{"Sai số"}$$

Dự đoán (điều
prediction

logistic regression

Linear regression

$$f(x_1, \dots, x_p) = \sum_{i=1}^p \beta_i x_i \\ = \beta_1 x_1 + \dots + \beta_p x_p$$

$$Y \leftarrow \begin{cases} 0 \\ 1 \end{cases}$$

$$Y = f(x_1, x_2, \dots, x_p)$$

VĐ: $y = \theta^T x$ thi x STK măt hỷ.

x_1 : \sum số giờ ôn tập trong 1 tháng trước

x_2 : \sum số giờ làm việc $x \# \dots$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon. \quad \text{"Sai số"}$$

$y = \theta^T x$ thi x STK măt hỷ của 1 SV

$x = \sum$ số giờ thi học

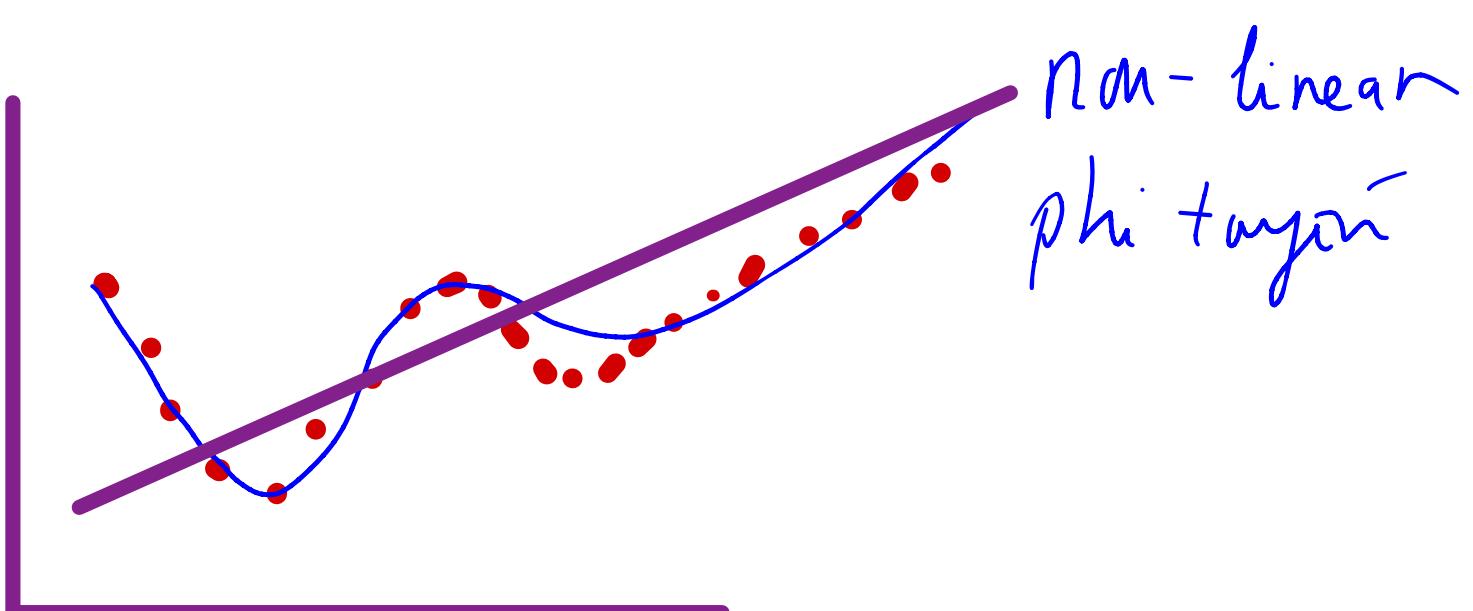
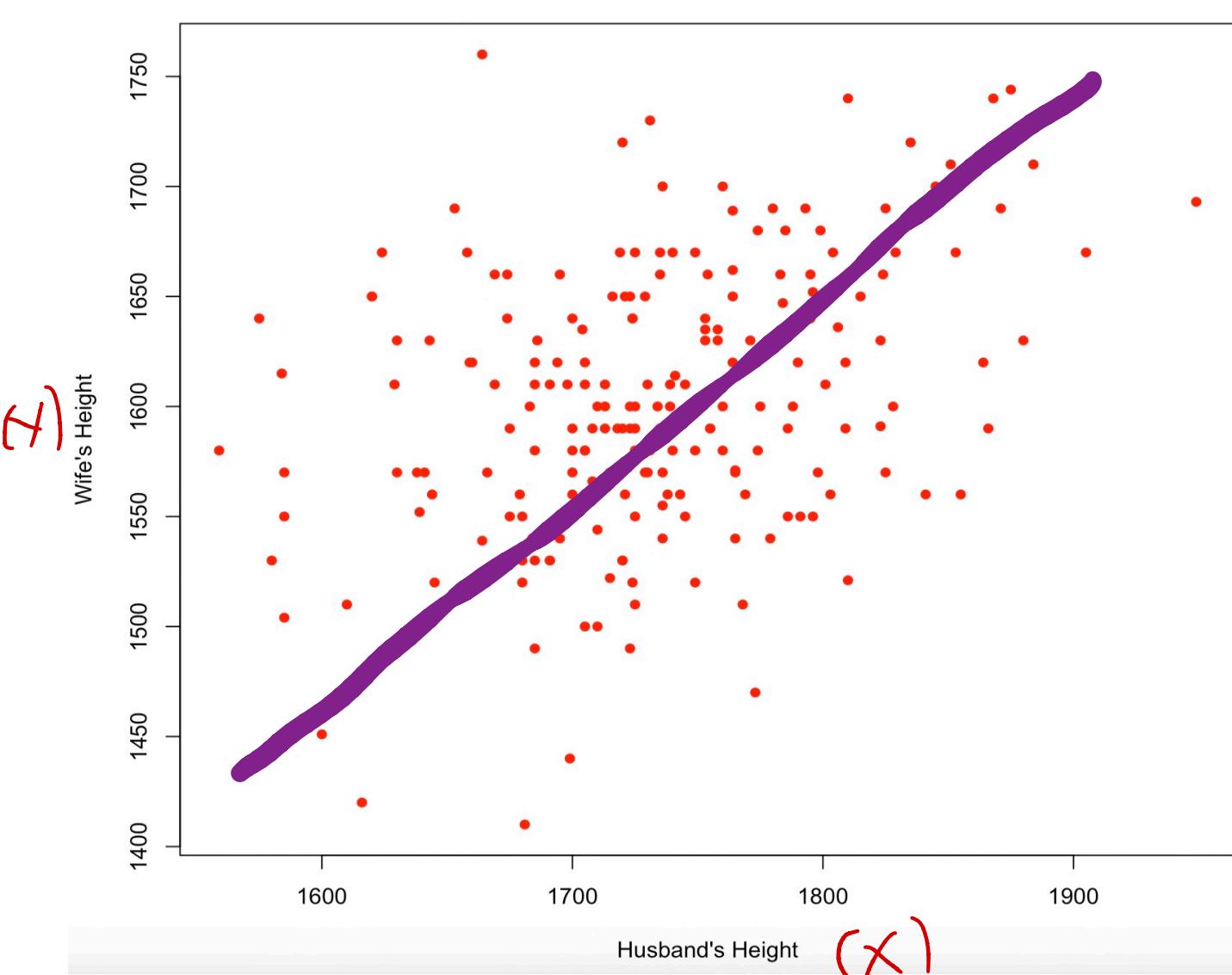
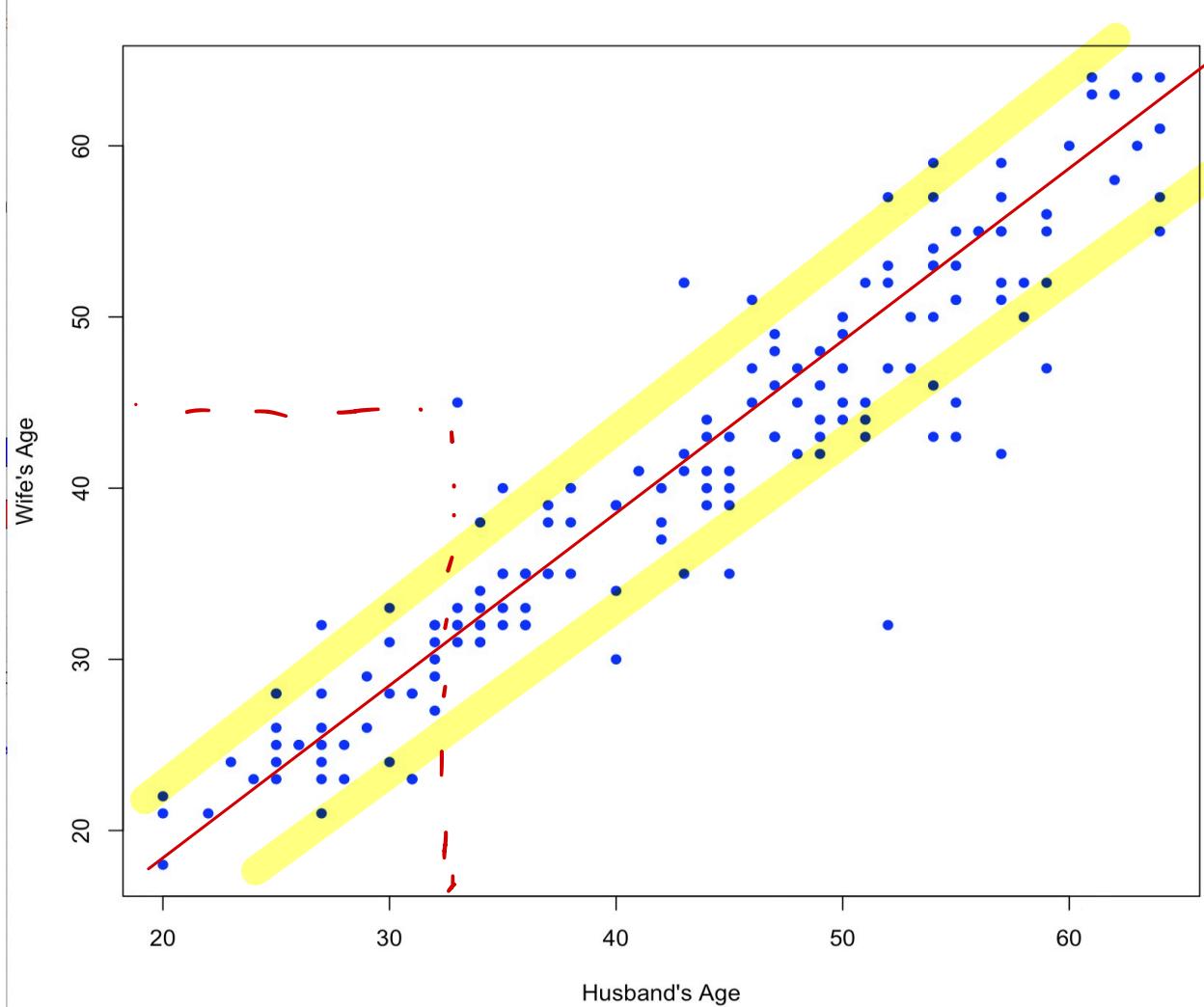
$$\varepsilon \sim N(0, \sigma^2)$$

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

Mối quan hệ tuyến tính
giữa x và y

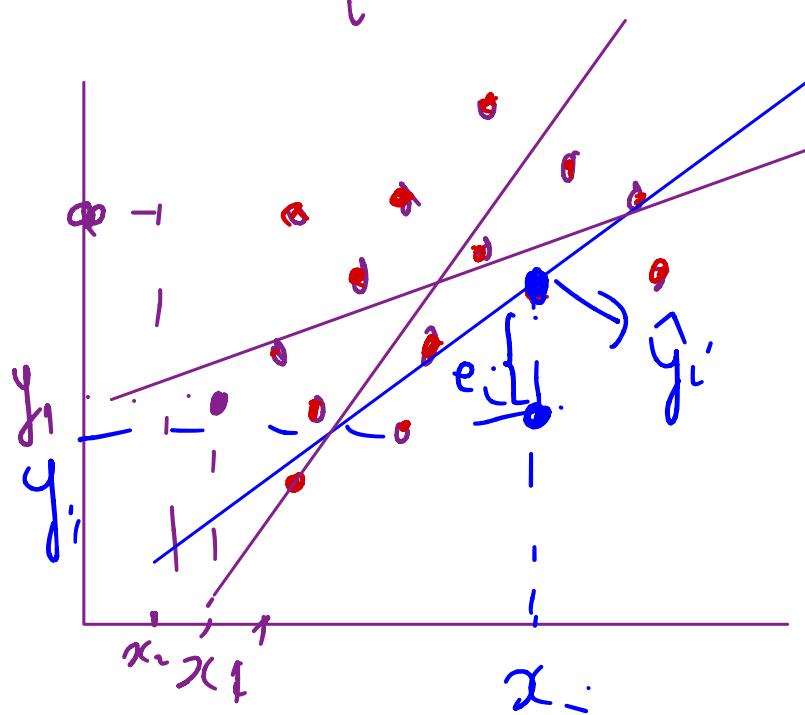
"Sai số"

σ^2 càng lớn
⇒ Mối quan
tập yếu



Xác định các hằng số hồi quy β_0, β_1 :

Dữ liệu quan sát: $(x_i, y_i), i = 1, \dots, n$



$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

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

Ước lường β_0, β_1 ?

↪ PP Bình phương bé nhất (Least-squares)

y_i : giá trị quan sát đ^tc (true value)

gọi: $\hat{\beta}_0$ và $\hat{\beta}_1$, ta 2 ước lường cho β_0

và $\hat{\beta}_1$

$$\hookrightarrow \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

↪ giá trị dự đoán (predicted value)

$$\epsilon_i = y_i - \hat{y}_i$$

SSE = Sum of Squares Error

$$= \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i} x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}$$

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum x_i^2 - \frac{(\sum x_i)^2}{n}$$

$$= (n-1) S_x^2 \quad \left| \begin{array}{l} \\ \\ S_x^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2 \end{array} \right.$$

$$S_{yy} = SST = \sum (y_i - \bar{y})^2$$

$$= \sum y_i^2 - \frac{(\sum y_i)^2}{n} = (n-1) S_y^2$$

$$\Rightarrow \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{và} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

(vì $\hat{\beta}_1$ là $k\hat{\beta}$ quy đổi qua (\bar{x}, \bar{y})).

x_i	y_i	x_i^2	y_i^2	$x_i y_i$
x_1	y_1			
x_2	y_2			
\vdots	\vdots			
x_n	y_n			
$\sum x_i$	$\sum y_i$	$\sum x_i^2$	$\sum y_i^2$	$\sum x_i y_i$

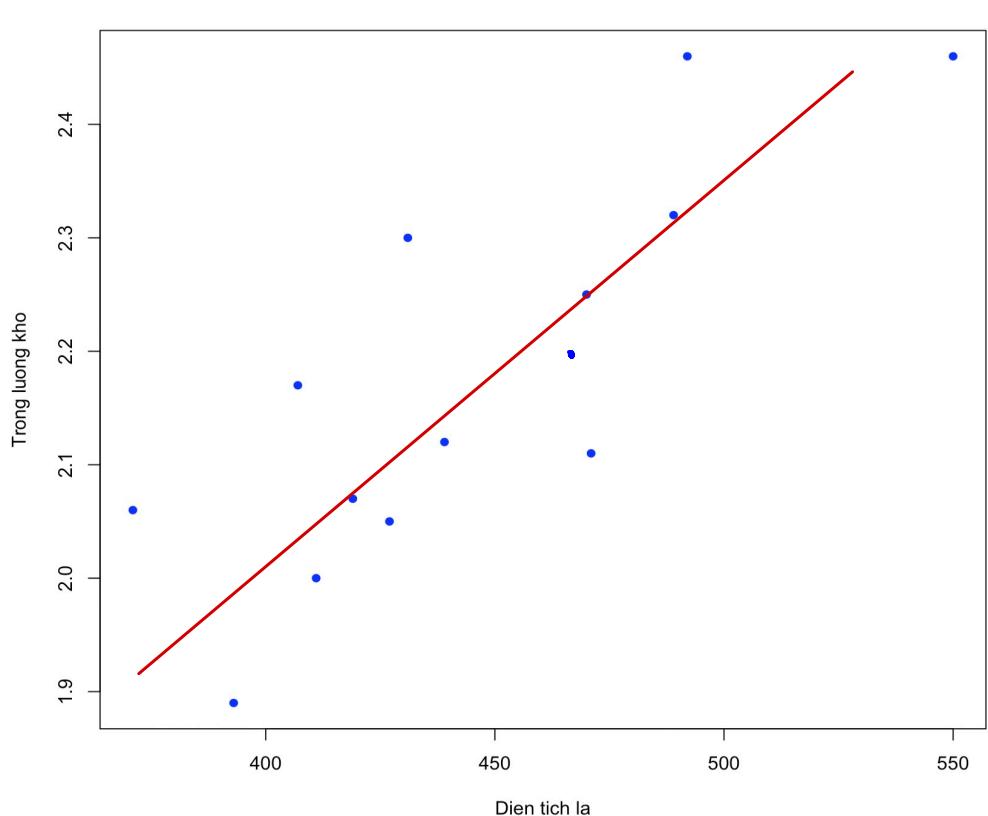
$$\Rightarrow \hat{\beta}_1 = \dots$$

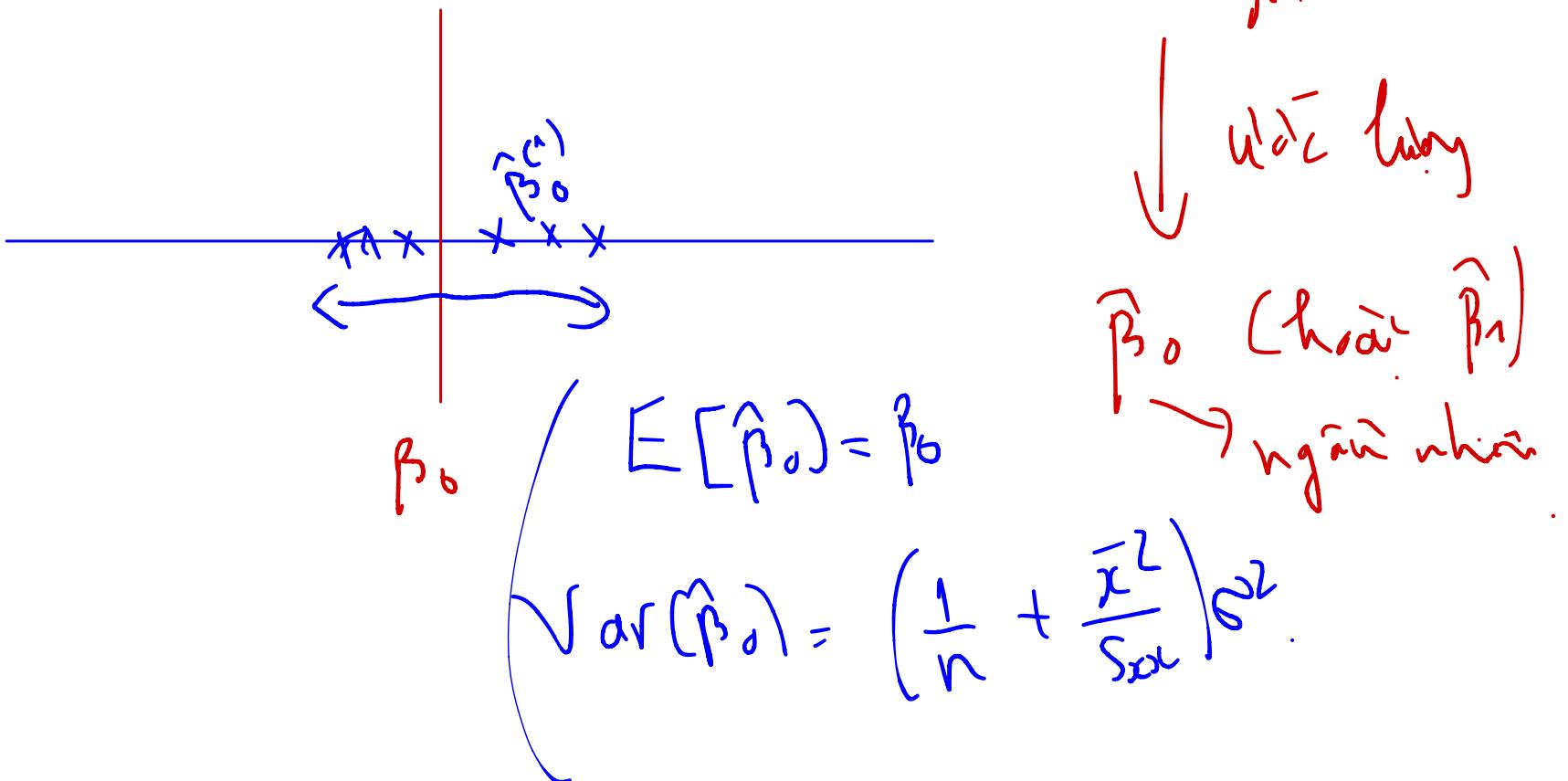
$$\hat{\beta}_0 = \dots$$

trong bảng white

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

dien tích
lô

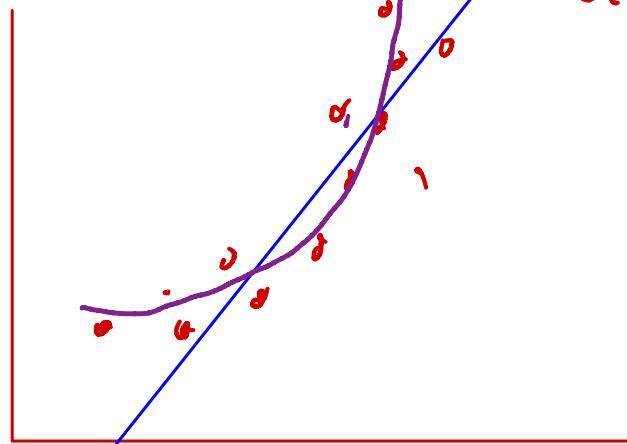
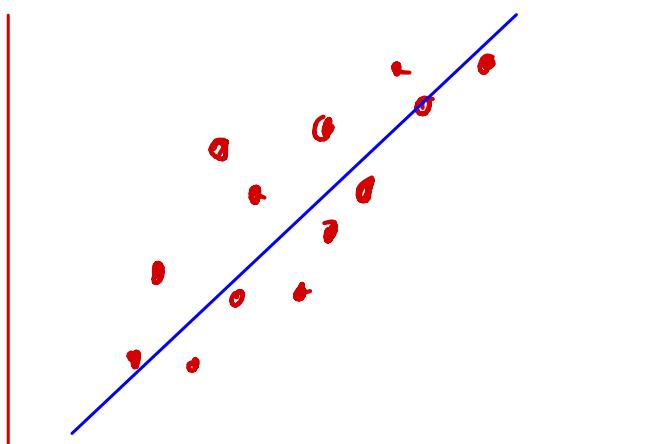




Cân hì: Xác định mô hình hồi quy tuyến tính giả[“] thích hợp cho mối quan hệ giữa X và Y .

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: best fitted line (đường thẳng "ulâm" tốt nhất)



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

doanh thu bán hàng

Chi phí quảng cáo

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$(Syy)$$

$$SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$$

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\begin{aligned}
 SST &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
 &= \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2
 \end{aligned}$$

SSR →
 SSE →

Định nghĩa 3

Hệ số xác định (Coefficient of Determination) là tỷ lệ của tổng sự biến thiên trong biến phụ thuộc gây ra bởi sự biến thiên của các biến độc lập (biến giải thích) so với tổng sự biến thiên toàn phần.

Hệ số xác định thường được gọi là R - bình phương (R -squared), ký hiệu là R^2 .

Công thức tính:

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}. \quad (20)$$

Chú ý: $0 \leq R^2 \leq 1$.

$$\begin{aligned}
 SST &= \overbrace{\text{SSR}} + \overbrace{\text{SSE}} \\
 y &= \beta_0 + \beta_1 x + \varepsilon
 \end{aligned}$$

$$\begin{aligned}
 \underbrace{y}_{\text{DVTB TL}} &= \beta_0 + \beta_1 \underbrace{x}_{\substack{\text{số liệu} \\ \text{tín hiệu}}} + \varepsilon \\
 R^2 &= 0..001 \quad \sum \text{thời gian tự học} \\
 &\quad R^2 = 0,5
 \end{aligned}$$

Tính toán:

$$R^2 = \frac{SSR}{SST} = \frac{\hat{\beta}_1 S_{xy}}{S_{yy}}$$

$$\begin{aligned}
 SST = S_{yy} &= \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{(\sum y_i)^2}{n} \\
 &= (n-1) S_y^2
 \end{aligned}$$

$$SSR = \hat{\beta}_1 \cdot S_{xy}$$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

gs $R^2 = 0.8 \Rightarrow$ Mối quan hệ mạnh
 80% sự thay đổi trong Y
 do do sự thay đổi của X .

$\begin{cases} Y = \text{doanh thu khi bán 1 sp} \Rightarrow R^2 = 0.7 \\ X = \text{chi phí q/c} \end{cases}$
 \Rightarrow 70% sự thay đổi trong
 doanh thu sẽ do bởi
 sự thay đổi chi phí quảng cáo.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

vD: $Y = \text{doanh thu (100 triệu)}$

$X = \text{chi phí q/c (chục triệu)}$

$\hat{Y} = 120 + 0.7X$ về mặt TB, nếu chi phí q/c tăng lên 1 đơn vị, doanh thu tăng thêm 0.7×100 triệu.

(*) về mặt truy bù, khi X thay đổi 1 đơn vị, thì Y tăng/giảm (tùy theo dấu của $\hat{\beta}_1$)

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$E[Y] = E[\underbrace{\beta_0 + \beta_1 X}_{\beta_0 + \beta_1 X} + \varepsilon] = \beta_0 + \beta_1 X + E[\varepsilon]$$

Giải ví dụ 1:

(c) Tính hệ số xác định R^2 : nhắc lại công thức tính hệ số xác định

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

với

$$\text{SST} = S_{yy} = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n} = 0.3637,$$

và

$$\text{SSR} = \hat{\beta}_1 S_{xy} = 0.002912 \times 82.8977 = 0.2414.$$

Vậy:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{0.2414}{0.3637} = 0.6637.$$

66.37% số thay đổi trong trọng lượng khô của các lá cây đã nén sẽ được giải thích bởi diện tích lá.

$\sim xx \quad \sim \dots, \quad \sim xy \quad \sim \dots$

Ta tính được

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{82.8977}{28465.69} = 0.002912,$$

$$\hat{\beta}_0 = y\bar{y} - \hat{\beta}_1 \bar{x} = 2.1738 - 0.002912 \times 443.8462 = 0.8813.$$

$\hat{y} = 0.8813 + 0.002912 \cdot x$ → nếu diện tích lá tăng lên 1 đơn vị (1 cm^2) thì trọng lượng khô thêm 0.002912 g và mặt trung bình

y = Điểm thi TOEFL,

x_1 = \sum giờ ôn tập

x_2 = Líthuật học (KHTN: 0; KAXV: 1; KT: 2)

x_3 = Túi tiền thi

x_4 = \sum tệp thi tháo

x_5 = Số tờ phi

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$

Mỗi qhtt giữa y và (x_1, \dots, x_p)

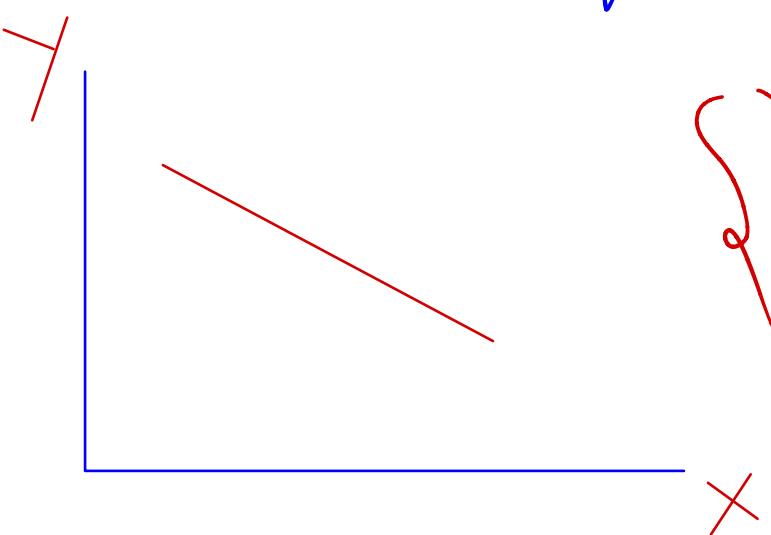
→ Mô hình hồi quy bì:

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon$$

$$\begin{cases} H_0: \beta_j = 0 & (\text{biến } X_j \text{ không có ý nghĩa giải thích cho } Y) \\ H_1: \beta_j \neq 0 & \rightarrow \text{biến } X_j \text{ có ý nghĩa.} \end{cases}$$

Hệ số tương quan: $r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}.$

$r_{XY} < 0$: tương quan nghịch

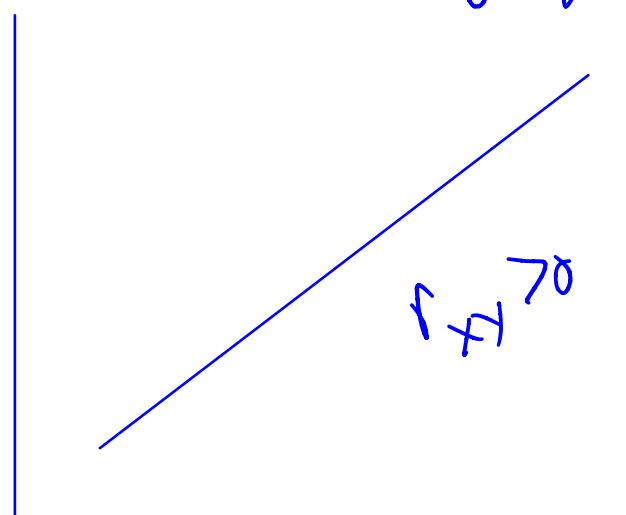


$$\begin{cases} Y = \beta_0 + \beta_1 X \\ X = T \text{ (giảm lín may X)} \end{cases}$$

$r_{XY} = 0$: K⁰ có mối quan hệ tuyến tính

$r_{XY} = 1$: mối quan hệ hoàn hảo

$r_{XY} > 0$: tương quan thuận

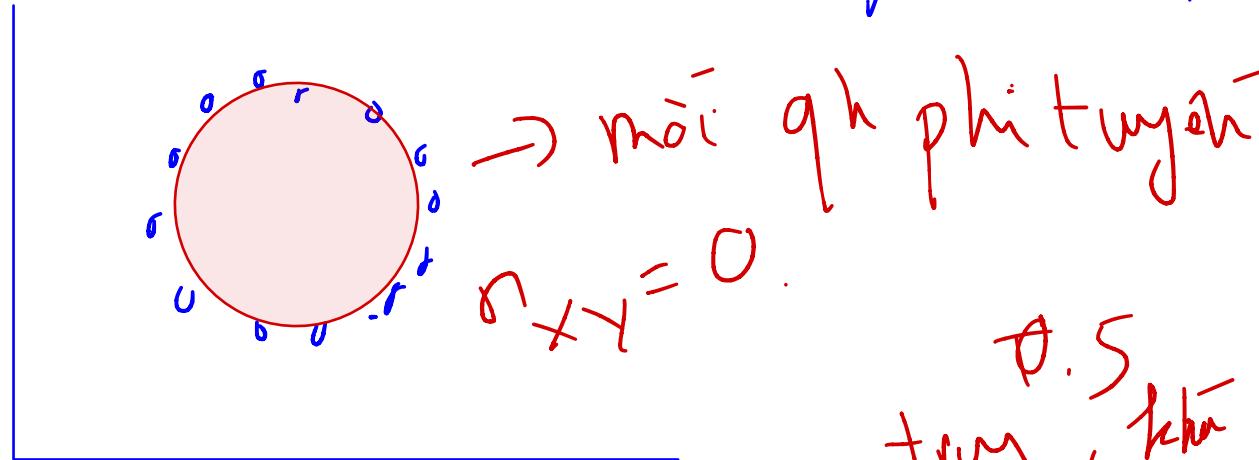


$$\begin{aligned} Y &= \text{chỉ cao 1 dùa tree} \\ X &= \text{hiện suất uống bia nh.} \\ Y &= \beta_0 + \beta_1 X + \epsilon \end{aligned}$$

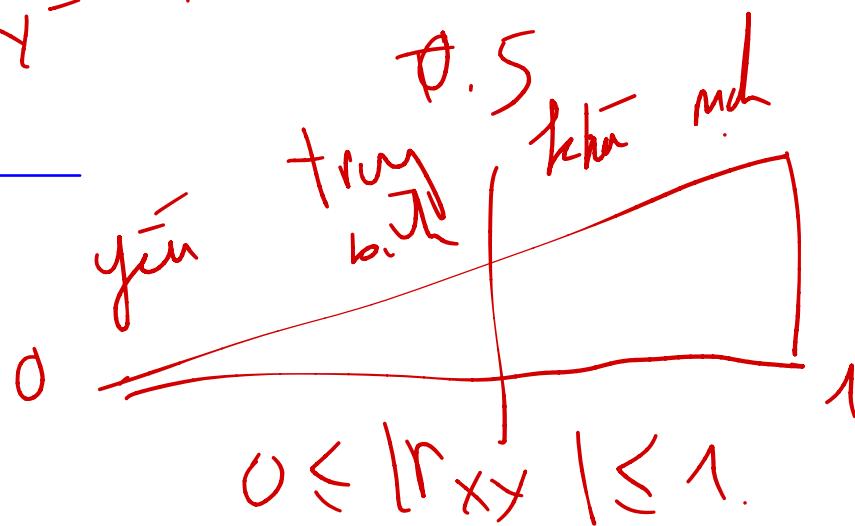
Chú ý: $r_{XY} = \sqrt{R^2} \rightarrow$ hệ số xđ.

$r_{XY} = 0$: K^o có mối quan hệ tuyêt tách.

(nhưng vẫn có thể tồn tại
mối quan hệ phi tuyến)



$$0 \leq |r_{XY}| \leq 1$$



Thi:

- 1) - Cho dữ liệu (x_i, y_i) : $i = 1, \dots, n$
hoặc dữ liệu tổng $(\sum x_i, \sum y_i, \sum x_i^2, \sum y_i^2, \sum x_i y_i)$

Tính $\hat{\beta}_0, \hat{\beta}_1$

$$\Rightarrow \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

Nhận xét mè hết ($\hat{\beta}_1$)

- 2) Tính hố số R^2 , nhận xét.

- 3) Tính hố số tuyêt quan r_{XY} ($r_{XY} = \sqrt{R^2}$)
nhận xét.