Chương 1:
MÔ HÌNH HỒI QUY TUYẾN TÍNH 2 BIẾN

1. MÔ HÌNH VÀ MỘT SỐ KHÁI NIỆM
2. PHƯƠNG PHÁP ƯỚC LƯỢNG OLS
3. TÍNH KHÔNG CHỆCH VÀ ĐỘ CHÍNH XÁC CỦA ƯỚC LƯỢNG OLS
4. ĐỘ PHÙ HỢP CỦA HÀM HỒI QUY MẪU – HỆ SỐ XÁC ĐỊNH R^2
5. MỘT SỐ VẤN ĐỀ BỔ SUNG

* Thuật ngữ “Hồi quy (Regression)” được Francis Galton sử dụng vào năm 1886 – “regression to mediocrity”.

Phân tích hồi quy nghiên cứu mối liên hệ phụ thuộc của một biến *(biến phụ thuộc, biến được giải thích–explained variable)* **với một hay nhiều biến khác** *(biến độc lập hay biến giải thích–explanatory variable)* **nhằm ước lượng hoặc dự báo *giá trị trung bình của biến phụ thuộc* với các giá trị đã biết của biến độc lập.**

- Biến phụ thuộc : biến ngẫu nhiên có quy luật phân phối xác suất.
- Biến độc lập : không ngẫu nhiên, giá trị được xác định trước.

Ví dụ:

Xu hướng về chiều cao của những đứa trẻ do cha mẹ cao không bình thường quy định. (Luật Galton).

* Karl Pearson nghiên cứu sự phụ thuộc chiều cao của các bé trai vào chiều cao của các ông bố: HS = f(HF) + U

Phân biệt quan hệ hồi quy với các quan hệ khác

* Với quan hệ hàm số.
* Với quan hệ nhân quả.
* Với quan hệ tương quan.

1.1 MÔ HÌNH HỒI QUY VÀ MỘT SỐ KHÁI NIỆM

1.1.1 Mô hình hồi quy

Ví dụ 1: Xét địa phương có 40 hộ gia đình.

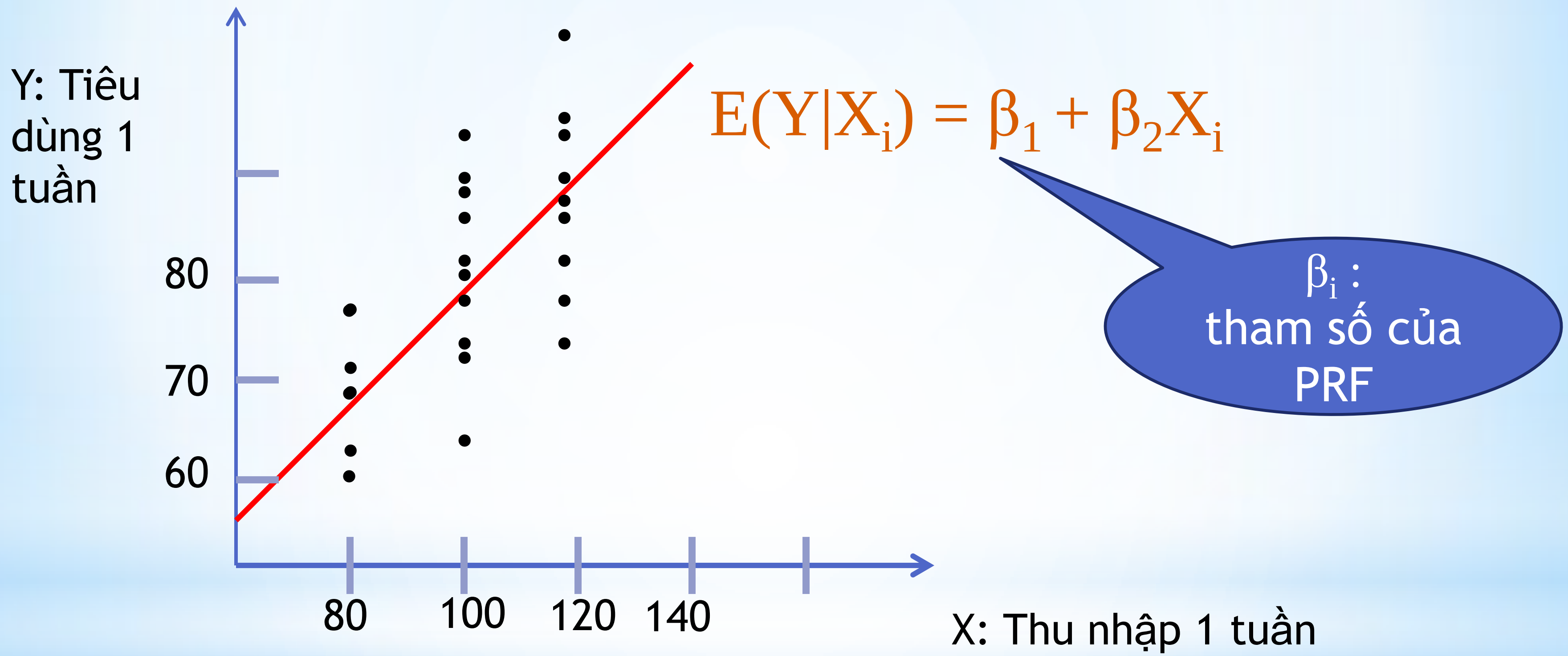
X: Thu nhập sau thuế của một gia đình trong một tuần (đôla)

Y: Chi tiêu của một gia đình trong một tuần (đôla)

X	80	100	120	140	160	180	200
Y	55	65	79	80	102	109	120
	60	70	84	93	107	117	136
	65	74	90	95	110	122	140
	70	80	93	103	116	125	144
	75	85	99	108	118	135	145
		88		113	125	142	
				115			
$E(Y/X_i)$	65	77	89	101	113	125	137

1.1.1 Mô hình hồi quy

Nhận thấy: Trung bình có điều kiện của mức chi tiêu trong tuần nằm trên 1 đường thẳng có hệ số góc dương



1.1.2 Mô hình hồi quy tổng thể
(PRF- Population Regression Function)

Mô hình hồi quy tổng thể

$E(Y|X_i) = f(X_i)$ với $f(X)$ là một hàm số của biến giải thích X.

- Hồi quy đơn (2 biến): nếu PRF có 1 biến độc lập.

- Hồi quy bội : nếu PRF có hơn 1 biến độc lập.

Mô hình PRF ngẫu nhiên: **$Y_i = f(X_i) + U_i$.**

Sai số ngẫu nhiên
$U_i = Y_i - E(Y|X_i)$

1.1.2 Mô hình hồi quy tổng thể

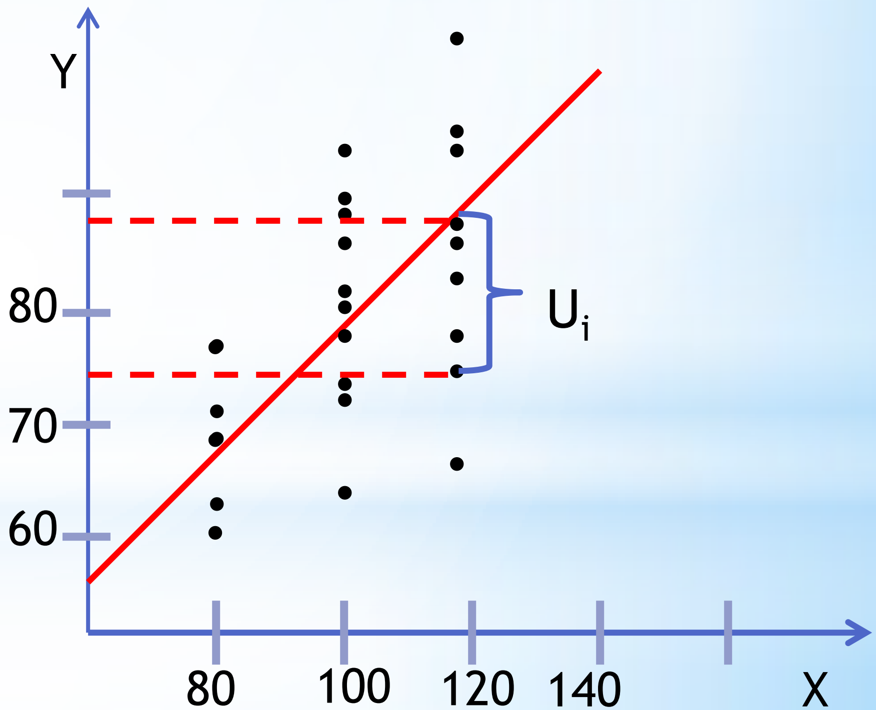
Sai số ngẫu nhiên (random error):

Giả thiết $E(U|X_i) = 0$.

U đại diện cho các yếu tố **không** có trong mô hình và có ảnh hưởng đến biến phụ thuộc Y.

***<u>Ví dụ 1.1</u>*:**

- Hãy tìm mô hình hồi quy tổng thể và xác định các giá trị sai số U.
- Cho biết ý nghĩa của các hệ số hồi quy trong mô hình PRF.



1.1.3 Mô hình hồi quy mẫu (SRF – Sample Regression Function)

* **Hàm hồi quy mẫu là** hàm hồi quy được xây dựng trên cơ sở một mẫu ngẫu nhiên.

$\hat{Y}_i = \hat{f}(X_i)$ hay $Y_i = \hat{f}(X_i) + e_i$

Trong đó:

SRF $\hat{f}(X_i)$ là ước lượng của PRF;

$\hat{Y}_i$ là giá trị tương ứng với X_i nhận được từ hàm SRF.

e_i là phần dư (residual), $e_i = Y_i - \hat{Y}_i$.

1.1.3 Mô hình hồi quy mẫu

Ví dụ 1.2 :

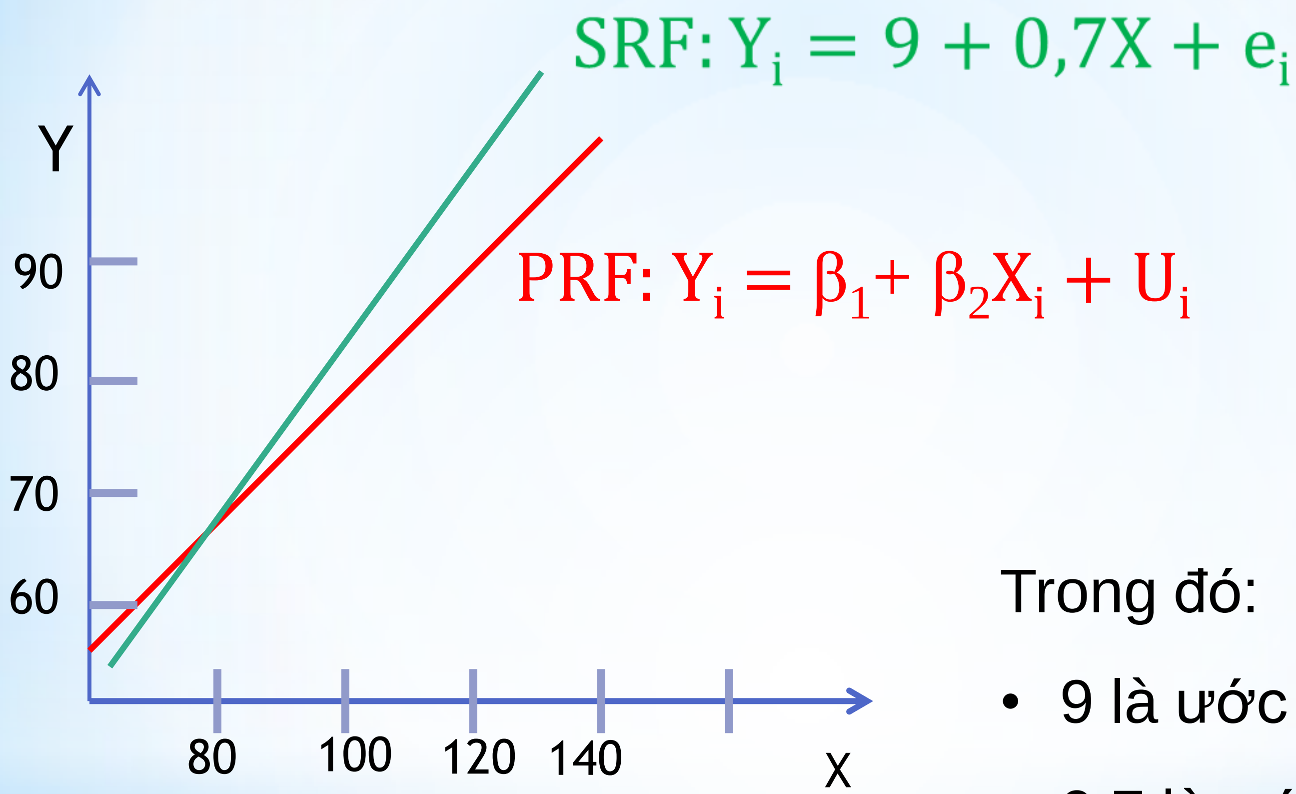
Hãy tìm hàm hồi quy mẫu cho mẫu ngẫu nhiên (gồm 5 quan sát) từ tổng thể sau:

X	80	100	120
Y	60 70	70 88	93
E(Y|X)	65	79	93

X	80	100	120
Y E(Y|X)	70	88	93

$$\widehat{Y}_i = 9 + 0{,}7X_i$$

Hãy nêu ý nghĩa các hệ số ước lượng trong mô hình SRF.



Hình: PRF và SRF

Trong đó:

- 9 là ước lượng của β_1
- 0,7 là ước lượng của β_2.

Hãy giải thích ý nghĩa của các giá trị 0,7 và 9.

1.1.4 Mô hình hồi quy tuyến tính

* Tính tuyến tính của hàm hồi quy được hiểu là **tuyến tính theo tham số**, nghĩa là theo các hệ số hồi quy.

Ví dụ:

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

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + U_i$$

$$Ln(Y_i) = \beta_1 + \beta_2 Ln(X_i) + U_i$$

$$1/Y_i = \beta_1 + \beta_2 X_i + U_i$$

1.1.4 Mô hình hồi quy tuyến tính

* Mô hình hồi quy tuyến tính 2 biến :

Hàm hồi quy tổng thể PRF: $Y_i = \beta_1 + \beta_2 X_i + U_i$

Hay $E(Y|X_i) = \beta_1 + \beta_2 X_i$

Trong đó: **Y:** biến phụ thuộc, **X** : biến độc lập.

Y_i, X_i là giá trị cụ thể của biến phụ thuộc và biến độc lập.

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

β_1 : tung độ gốc/ hệ số chặn (intercept), là giá trị trung bình của Y khi X=0.

β_2: độ dốc/ hệ số góc (slope) của hàm hồi quy, là lượng thay đổi của giá trị trung bình của Y khi X tăng 1 đơn vị.

1.1.4 Mô hình hồi quy tuyến tính

*Mô hình hồi quy tuyến tính 2 biến :

Hàm hồi quy mẫu SRF: $Y_i = \widehat{\beta_1} + \widehat{\beta_2} X_i + e_i$

Hay $\widehat{Y}_i = \widehat{\beta_1} + \widehat{\beta_2} X_i$

Trong đó:

$\widehat{\mathbf{Y}}_{\mathbf{i}}$ giá trị ước lượng của giá trị thực tế $E(Y_i/X_i)$

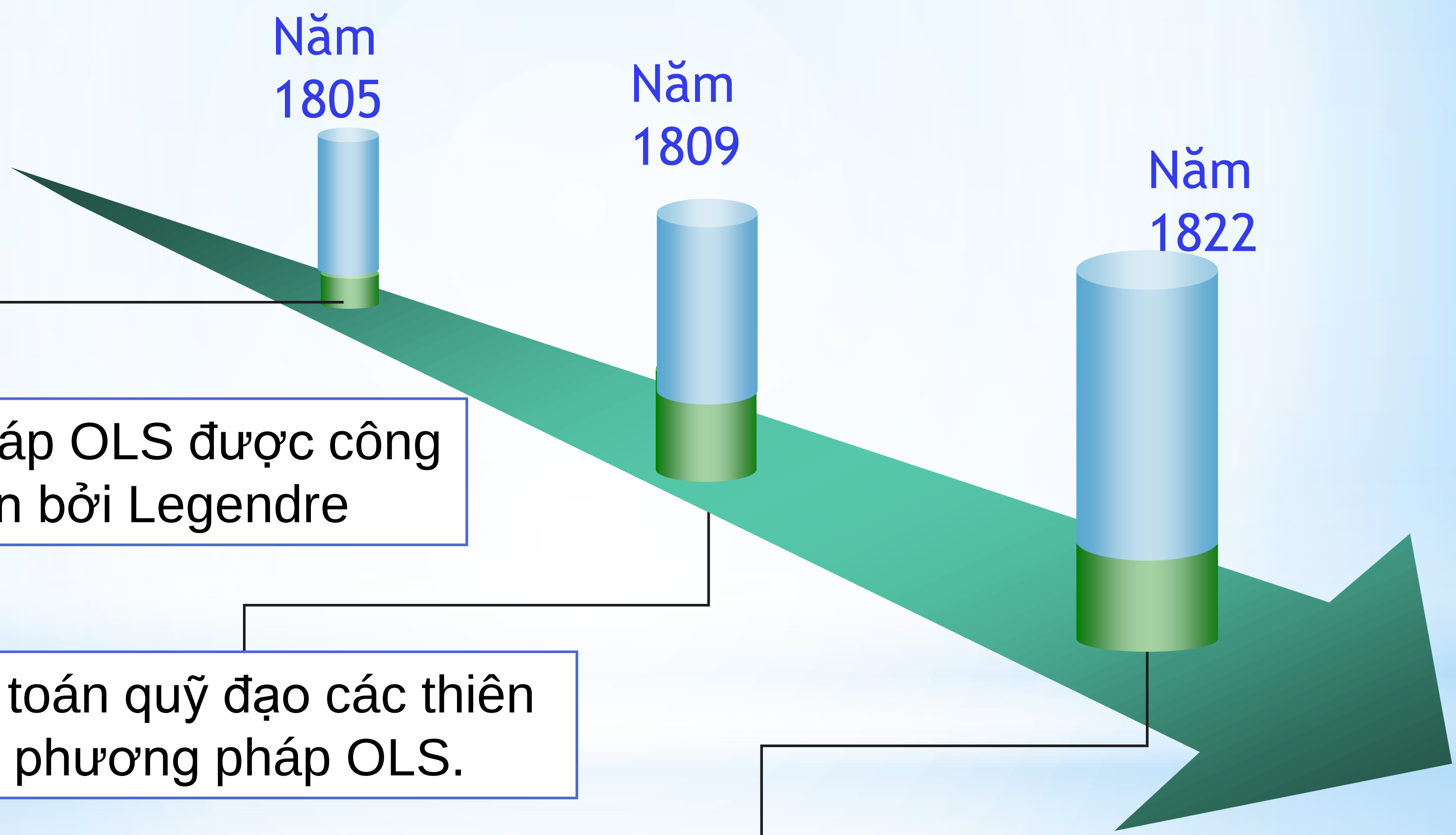
e_i: Phần dư, ước lượng điểm của U_i.

$\widehat{\boldsymbol{\beta}_1}$: ước lượng điểm của β_1 .

$\widehat{\boldsymbol{\beta}_2}$: ước lượng điểm của β_2 .

1.2 PHƯƠNG PHÁP ƯỚC LƯỢNG OLS

(OLS – Ordinary Least Squared)



Phương pháp OLS được công bố ngắn gọn bởi Legendre

Gauss tính toán quỹ đạo các thiên thạch bằng phương pháp OLS.

Định lý Gauss – Markov

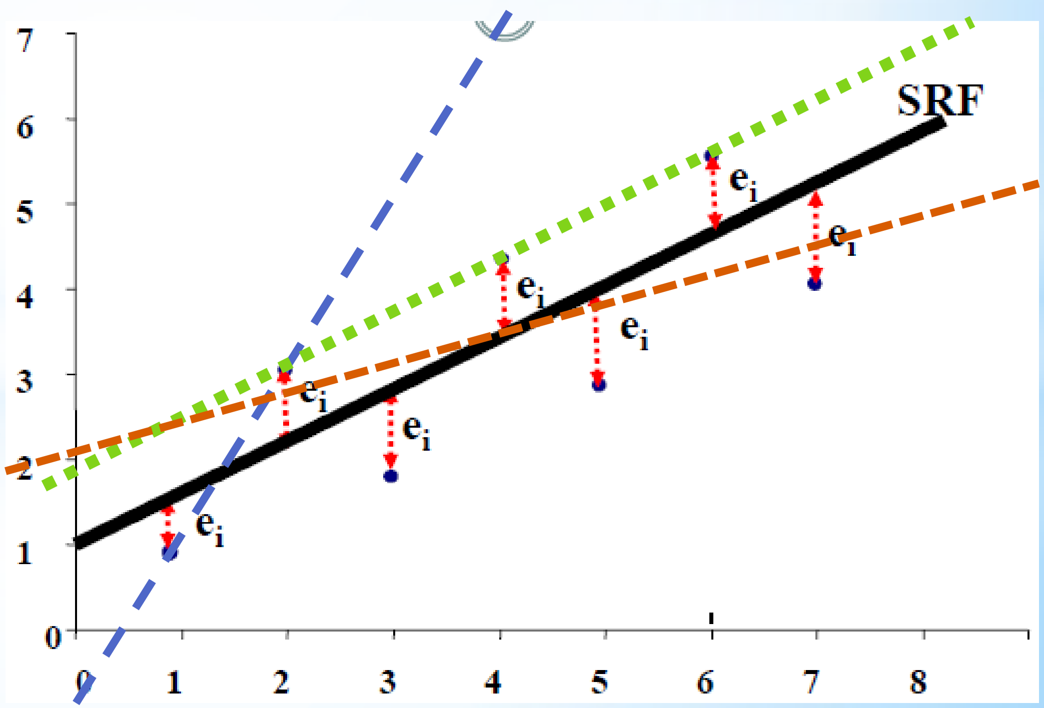
* Xét mô hình hồi quy tổng thể: $Y_i = \beta_1 + \beta_2 X_i + U_i$

Cần ước lượng các hệ số β_1 và β_2, bằng một mẫu cụ thể kích thước n $\{(X_i, Y_i)\ (i=1,..n)\}$.

Nội dung phương pháp OLS:

Tìm hàm hồi quy mẫu: $\hat{Y}_i = \widehat{\beta_1} + \widehat{\beta_2} X_i$ hay $Y_i = \widehat{\beta_1} + \widehat{\beta_2} X_i + e_i$

sao cho $\sum_{i=0}^{n} e_i^{\,2}$ **Min**



1.2 PHƯƠNG PHÁP ƯỚC LƯỢNG OLS

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \beta_1 - \beta_2 X_i\right)^2 \to Min$$

Giải bài toán cực trị hàm hai biến tìm $\widehat{\boldsymbol{\beta}_1}$ và $\widehat{\boldsymbol{\beta}_2}$, ta có:

$$\beta_2 = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

$$\beta_1 = \overline{Y} - \beta_2 \overline{X}$$

Trong đó:

- $\bar{X}$ và $\bar{Y}$ là trung bình mẫu của X và Y.
- $x_i = X_i - \bar{X}$;
- $y_i = Y_i - \bar{Y}$

1.2 PHƯƠNG PHÁP ƯỚC LƯỢNG OLS

Nhận xét:

- Hàm SRF thu được từ các mẫu khác nhau của cùng một tổng thể có thể rất khác nhau và rất khác với hàm PRF.
- $\widehat{\boldsymbol{\beta}_1}$ và $\widehat{\boldsymbol{\beta}_2}$ được xác định một cách duy nhất ứng với một mẫu cụ thể.
- SRF đi qua trung bình mẫu ($\bar{X}$; $\bar{Y}$).

<u>Ví dụ 1.3</u> :

Hãy tìm hàm hồi quy mẫu cho mẫu ngẫu nhiên (gồm 9 quan sát) sau:

X	80	100	120	140	160	180
Y	60 65	85	90	103 113	118 125	130

1.3 TÍNH KHÔNG CHỆCH VÀ ĐỘ CHÍNH XÁC CỦA ƯỚC LƯỢNG OLS

Các giả thiết đặt ra nhằm giúp phương pháp OLS thu được các ước lượng không chệch cho β_1, β_2.

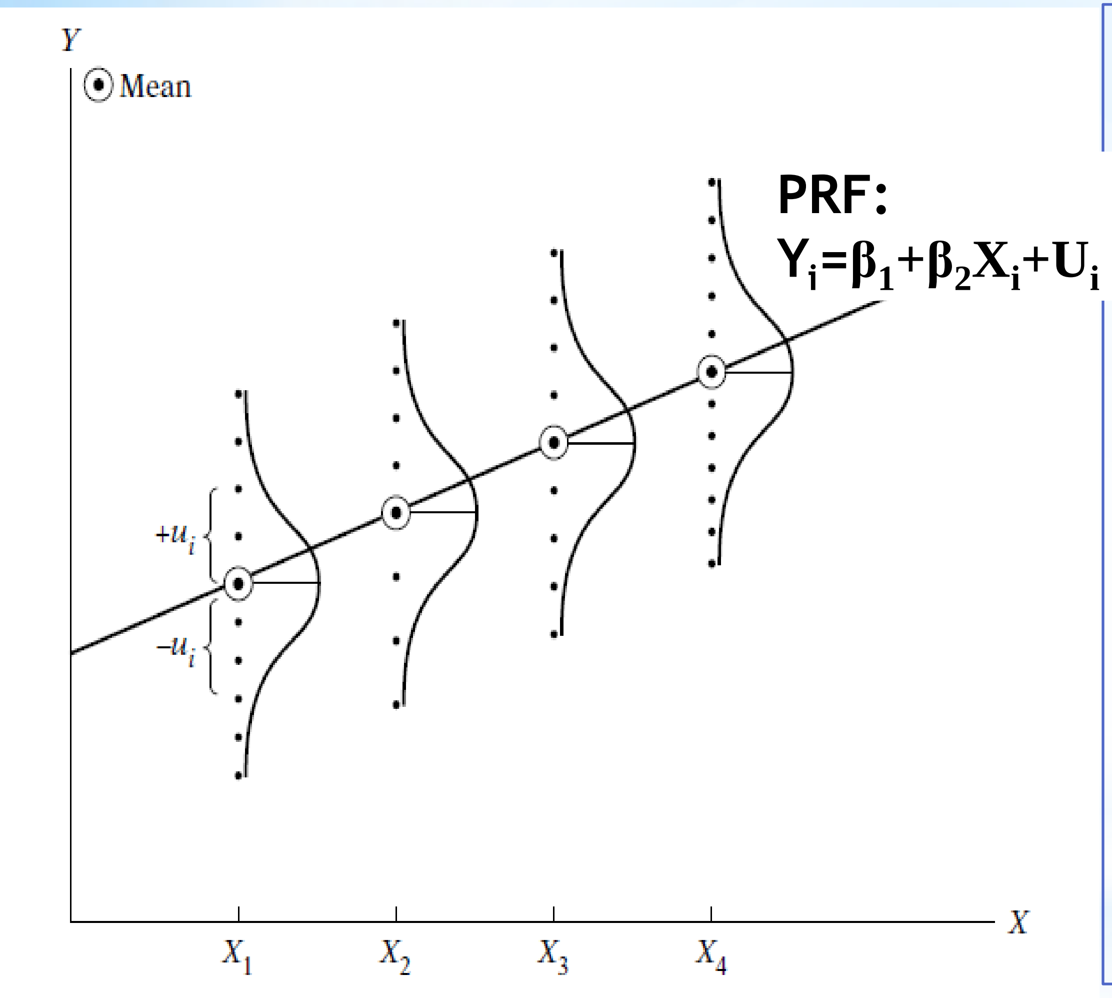
1.3.1 Các giả thiết của phương pháp OLS 2 biến

GT1: Mô hình được ước lượng trên cơ sở mẫu ngẫu nhiên.

GT2: Sai số U là biến ngẫu nhiên và $E(U|X_i) = 0$ với mọi i.

* Tính không chệch: Kỳ vọng của ước lượng bằng giá trị của tổng thể

1.3.1 Các giả thiết của phương pháp OLS 2 biến



Khi GT 2 thỏa mãn thì:

$$E(U) = 0.$$

$$Cov(X, U) = 0.$$

Ý nghĩa GT2:

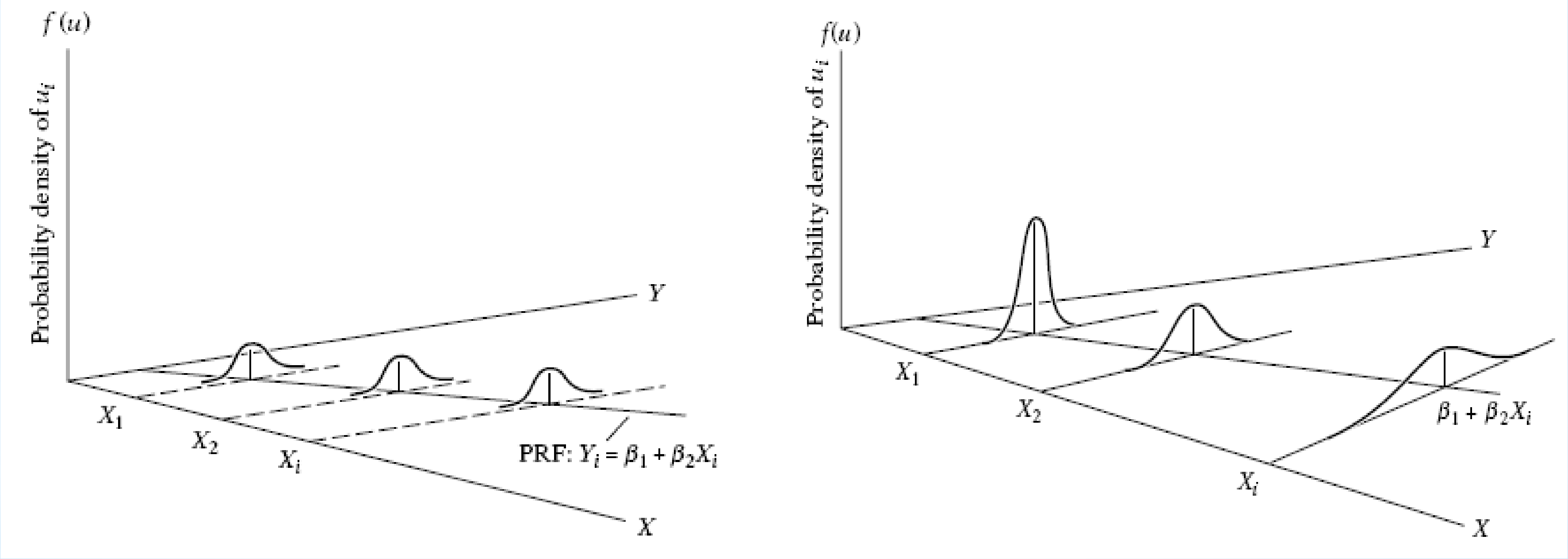
-Đảm bảo ý nghĩa hệ số β_2.

-Tại mỗi giá trị $X = X_i$, trung bình ảnh hưởng của các yếu tố khác X lên Y là bằng 0.

1.3.1 Các giả thiết của phương pháp OLS 2 biến

GT3: Phương sai của U_i là bằng nhau tại mọi giá trị X_i:

Var(U|X) = σ^2 const.



GT2, 3 thỏa mãn thì **Var(Y|X_i)= σ^2** với mọi i.

1.3.2 Tính không chệch và độ chính xác của các ước lượng OLS

* **Tính không chệch (unbiased):**

Khi GT2 thỏa mãn thì $E(\widehat{\beta_1}) = \beta_1$ và $E(\widehat{\beta_2}) = \beta_2$.

- **Độ chính xác:**

Khi các GT 1 – 3 được thỏa mãn thì phương sai của các hệ số ước lượng bằng:

$$var(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2} \; ; \quad var(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\sigma^2$$

1.3.2 Tính không chệch và độ chính xác của các ước lượng OLS

- Ước lượng của σ *là* $\hat{\sigma}$ - gọi là *sai số chuẩn của hàm hồi quy (standard error of regression)* với

$$\left(\hat{\sigma}\right)^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$$

- Khi các GT 1, 2, 3 thỏa mãn thì $E(\hat{\sigma}^2) = \sigma^2$.

- Sai số chuẩn (standard error) của hệ số ước lượng

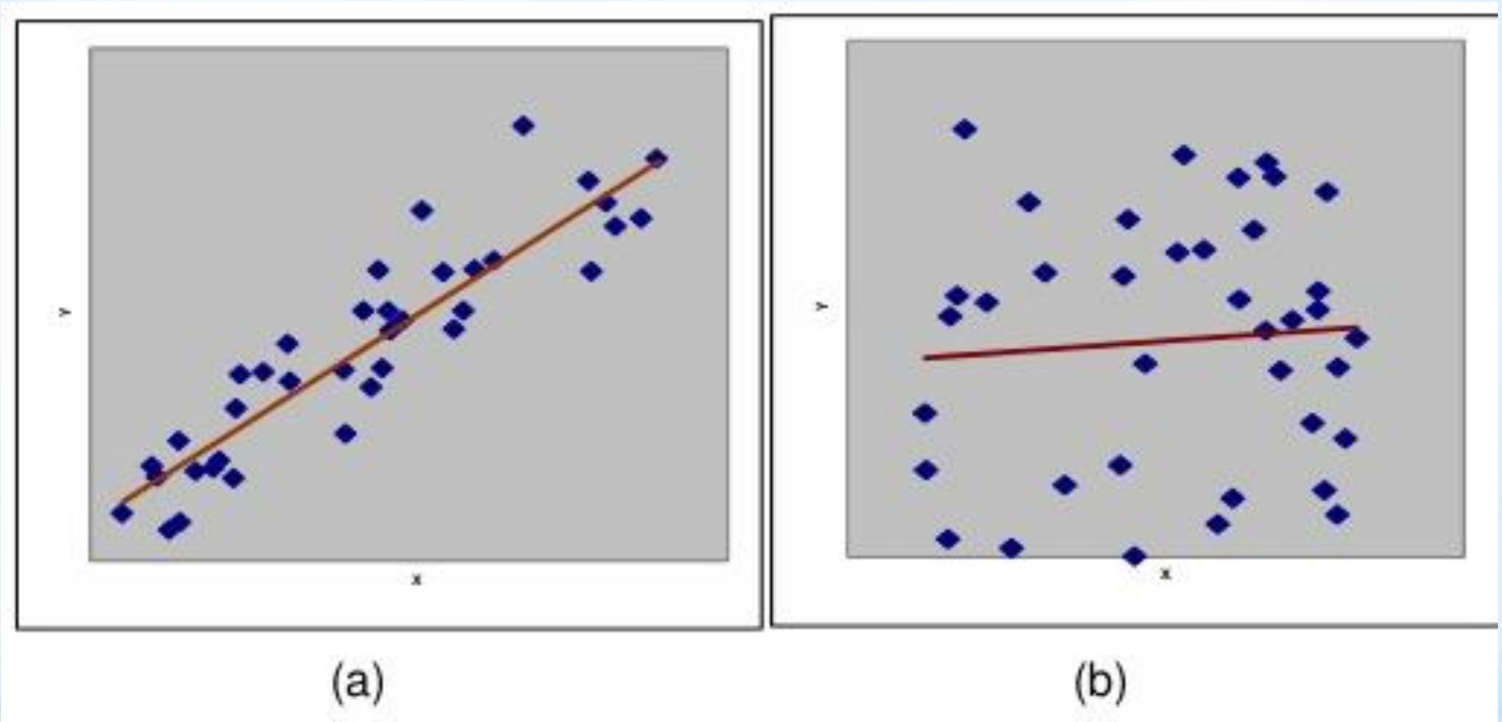
$$se(\hat{\beta}_i) = \sqrt{var(\hat{\beta}_i)}$$ với σ^2 được thay bởi ước lượng $\hat{\sigma}^2$

1.3.2 Tính không chệch và độ chính xác của các ước lượng OLS

* Khi mô hình hồi quy có hệ số chặn, các ước lượng OLS thỏa mãn các tính chất:

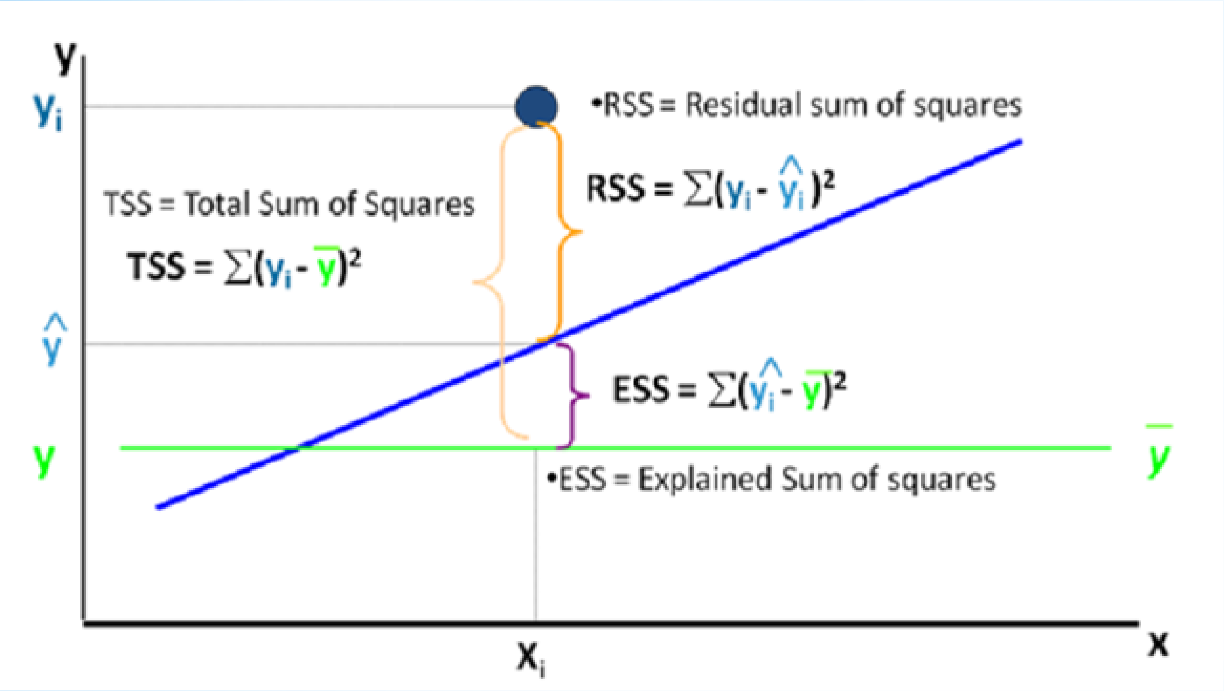
- $\sum_{i=0}^{n} e_i = 0$
- $\sum_{i=0}^{n} e_i X_i = 0$
- $\bar{\hat{Y}} = \bar{Y}$

1.4 ĐỘ PHÙ HỢP CỦA HÀM HỒI QUY MẪU – HỆ SỐ XÁC ĐỊNH R^2.



Hàm hồi quy mẫu nào phù hợp với số liệu mẫu hơn?

1.4 ĐỘ PHÙ HỢP CỦA HÀM HỒI QUY MẪU – HỆ SỐ XÁC ĐỊNH R².



$$TSS = \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$$

$$RSS = \sum_{i=1}^{n} e_i^2$$

$$ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$$

Nếu mô hình hồi quy có hệ số chặn thì **TSS = ESS + RSS**

* **Hệ số xác định của hàm hồi quy** (determination coefficient)

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

* $0 \le R^2 \le 1$

* **Ý nghĩa hệ số xác định R^2:**

1

2

3

Chú ý:

- Với mô hình hồi quy 2 biến <u>có chứa hệ số chặn</u>

- R^2 cũng chính là bình phương của hệ số tương quan mẫu giữa X và Y.

- R^2 = 0 khi và chỉ khi $\widehat{\beta_2}$ =0.

- Với mô hình hồi quy không có hệ nghĩa số chặn thì các phát biểu về R^2 nói trên có thể không đúng.

1.5 MỘT SỐ VẤN ĐỀ BỔ SUNG

1.5.1 Đơn vị đo lường trong phân tích hồi quy.

Xét hàm hồi quy mẫu : $\hat{Y}_i = \widehat{\beta_1} + \widehat{\beta_2} X_i$

1) Khi đổi đơn vị đo lường X bằng cách gia tăng 10 lần (ví dụ : từ met sang decimet), Y không đổi đơn vị thì các hệ số $\widehat{\beta_1}, \widehat{\beta_2}$ và R^2 thay đổi như thế nào?

2) Khi đổi đơn vị đo lường X và Y bằng cách gia tăng theo cùng tỷ lệ thì các hệ số $\widehat{\beta_1}, \widehat{\beta_2}$ và R^2 thay đổi như thế nào?

1.5 MỘT SỐ VẤN ĐỀ BỔ SUNG

* 1.5.2 Mô hình hồi quy không có hệ số chặn

Khi mô hình không có hệ số chặn thì $\sum_{\mathbf{i=0}}^{\mathbf{n}} \mathbf{e_i} \neq 0$.

Vì thế TSS ≠ ESS + RSS.

Suy ra …