CHƯƠNG II. MÔ HÌNH HỒI QUY HAI BIẾN

TS. Đinh Thị Thanh Bình

Khoa Kinh Tế Quốc Tế- Đại học Ngoại Thương

1. Giới thiệu mô hình hồi qui

1.1. Khái niệm về phân tích hồi qui

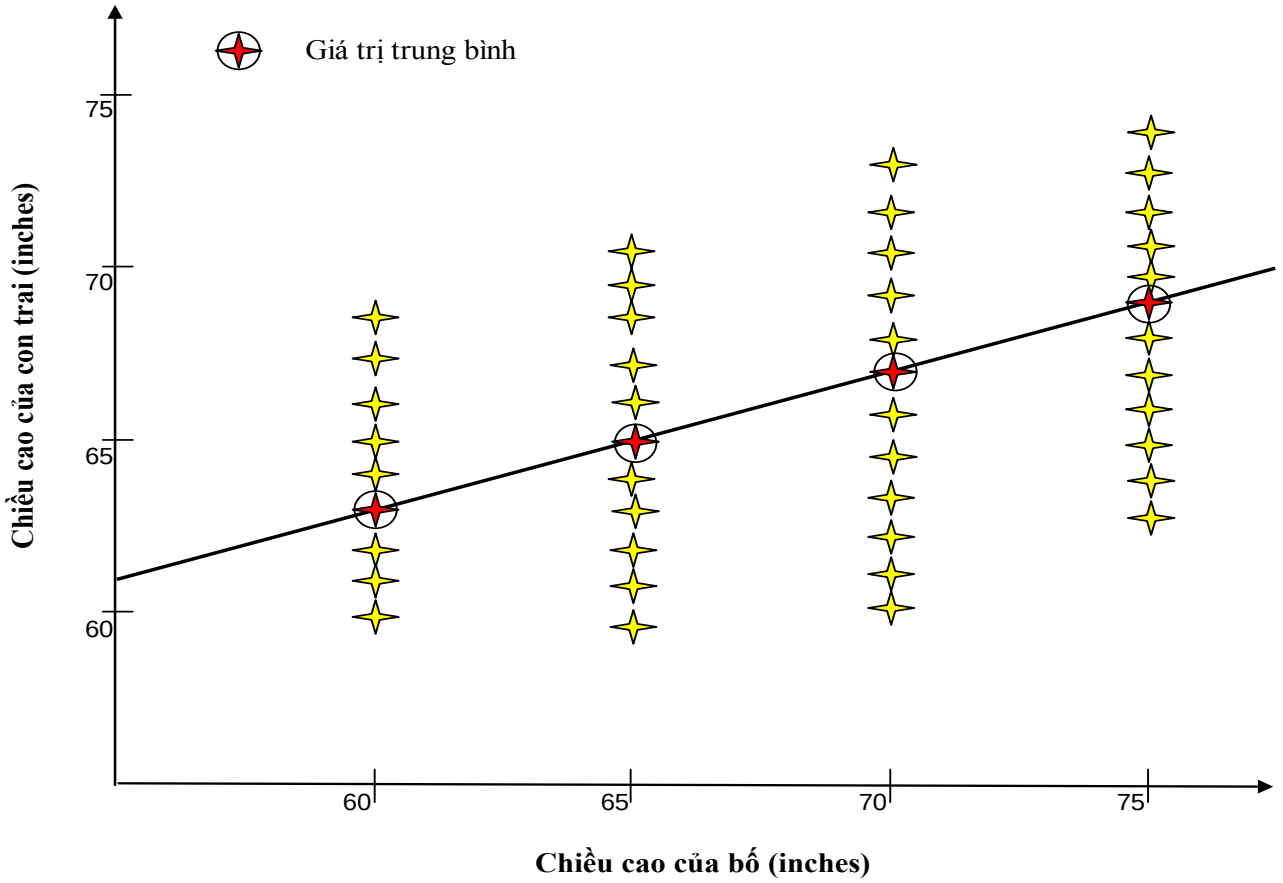
1.2. Sự khác nhau giữa các dạng quan hệ

1.1. Khái niệm về phân tích hồi qui

- Thuật ngữ hồi qui là «*regression to mediocrity*» nghĩa là « quy về giá trị trung bình »
- Thuật ngữ này ra đời khi Galton (1886) nghiên cứu sự phụ thuộc chiều cao của các con trai vào chiều cao của các ông bố.
- Ông đã xây dựng được đồ thị chỉ ra phân bố chiều cao của các con trai ứng với chiều cao của người cha.

1.1. Khái niệm về phân tích hồi qui

Hình 2.01. Đồ thị phân bố chiều cao của các cháu trai ứng với chiều cao của người cha



1.1. Khái niệm về phân tích hồi qui

Qua đồ thị phân bố, có thể thấy:

- Với chiều cao của người cha cho trước, thì chiều cao của con trai sẽ là một khoảng dao động quanh một giá trị trung bình.
- Chiều cao của cha tăng thì chiều cao của con trai cũng tăng.
- Các vòng tròn trên đồ thị chỉ ra giá trị TB của chiều cao con trai so với chiều cao của những ông bố.
- Nếu nối các điểm giá trị TB này, ta sẽ nhận được một đường thẳng như trong hình vẽ.
- Đường thẳng này được gọi là **đường hồi quy**- mô tả ***trung bình*** sự gia tăng chiều cao các con trai so với bố.

1.1. Khái niệm về phân tích hồi qui

- Như vậy, nghiên cứu giúp giải thích được câu hỏi: mặc dù có xu hướng bố cao đẻ con cao, bố thấp đẻ con thấp nhưng chiều cao trung bình của những người con có xu hướng **tiến tới (hồi quy) về chiều cao trung bình của toàn bộ dân số**, và xu hướng đó gọi là **hồi quy**.

- Từ đó, nghiên cứu giúp dự báo chiều cao trung bình của các con trai thông qua chiều cao cho trước của cha chúng.

1.1. Khái niệm về phân tích hồi qui

→ Bản chất của phân tích hồi quy là ***nghiên cứu mối liên hệ phụ thuộc của một biến (gọi là biến phụ thuộc hay biến được giải thích) với một hay nhiều biến khác (gọi là biến độc lập hay biến giải thích).***

→ Phân tích hồi quy tập trung giải quyết các vấn đề sau :

- Ước lượng **giá trị trung bình** của biến phụ thuộc với các giá trị đã cho của các biến độc lập.
- Kiểm định giả thuyết về bản chất của sự phụ thuộc đó.

1.2. Sự khác nhau giữa các dạng quan hệ

1.2.1. Hồi quy và quan hệ nhân quả

1.2.2. Hồi quy và tương quan

1.2.1. Hồi quy và quan hệ nhân quả

- Phân tích hồi quy nghiên cứu quan hệ giữa một biến phụ thuộc với một hoặc nhiều biến độc lập khác.

  → ***Điều này không đòi hỏi giữa biến độc lập và các biến phụ thuộc có mối quan hệ nhân quả***.

1.2.1. Hồi quy và quan hệ nhân quả

- **Ví dụ:** chúng ta có thể dự đoán sản lượng dựa vào lượng mưa và các biến khác nhưng không thể chấp nhận được việc dự báo lượng mưa dựa vào sự thay đổi của sản lượng.

  → ***Vì vậy, trước khi phân tích hồi quy, chúng ta phải nhận định chính xác mối quan hệ nhân quả.***

1.2.1. Hồi quy và quan hệ nhân quả

- Một sai lầm phổ biến nữa trong phân tích KTL là quy kết mối quan hệ nhân quả giữa hai biến số trong khi thực tế chúng đều là hệ quả của một nguyên nhân khác.
- **Ví dụ**: ta phân tích hồi quy số giáo viên với số phòng học trong toàn ngành giáo dục. Sự thực là cả số giáo viên và số phòng học đều phụ thuộc vào số học sinh.

1.2.2. Hồi quy và tương quan

- Hồi quy và tương quan khác nhau về : **mục đích** và **kỹ thuật**.
  - **Về mục đích**, phân tích tương quan đo mức độ kết hợp tuyến tính giữa hai biến. Ví dụ mức độ quan hệ giữa nghiện thuốc lá và ung thư phổi, giữa kết quả thi môn thống kê và môn toán. Nhưng phân tích hồi quy lại ước lượng hoặc dự báo một biến trên cơ sở giá trị đã cho của các biến khác.

1.2.2. Hồi quy và tương quan

- **Về kỹ thuật** trong phân tích hồi quy, các biến không có tính chất đối xứng. Biến phụ thuộc là đại lượng ngẫu nhiên còn giá trị của các biến giải thích đã được xác định. Trong phân tích tương quan, không có sự phân biệt giữa các biến, chúng có tính chất đối xứng.

2. Hàm hồi quy tổng thể và hàm hồi quy mẫu

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

2.2. Sai số ngẫu nhiên và bản chất của nó

2.3. Hàm hồi quy mẫu (SRF)

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

- ***Hàm hồi quy tổng thể là hàm hồi quy được xây dựng dựa trên kết quả nghiên cứu khảo sát tổng thể.***
- Ví dụ: Giả sử ở một địa phương chỉ có cả thảy 60 gia đình, 60 gia đình này được chia thành 10 nhóm, chênh lệch về thu nhập của các nhóm gia đình từ nhóm này sang nhóm tiếp theo đều bằng nhau.

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

Bảng 2.01. Số liệu về thu nhập và chi tiêu của 60 hộ gia đình

X	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
Y	55	65	79	80	102	110	120	135	137	150
Y	60	70	84	93	107	115	136	137	145	152
Y	65	74	90	95	110	120	140	140	155	175
Y	70	80	94	103	116	130	144	152	165	178
Y	75	85	98	108	118	135	145	157	175	180
Y	-	88	-	113	125	140	-	160	189	185
Y	-	-	-	115	-	-	-	162	-	191
Tổng	**325**	**462**	**445**	**707**	**678**	**750**	**685**	**1043**	**966**	**1211**

- **X= thu nhập sau thuế/hộ gia đình (USD)**
- **Y= Chi tiêu/hộ gia đình/tuần (USD)**

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

- Các số ở bảng trên có nghĩa là : với thu nhập trong một tuần chẳng hạn là X= 100$ thì có 6 gia đình mà chi tiêu trong tuần nằm giữa 65 và 88.
- Hay nói khác đi, ở mỗi cột của bảng cho ta phân bố xác suất của số chi tiêu trong tuần Y với mức thu nhập đã cho X, đó chính là ***phân bố xác suất có điều kiện của Y với giá trị X đã cho***.
- Vì bảng 2.01 là tổng thể nên ta dễ dàng tìm P(Y/X). Chẳng hạn, P(Y=85/X=100)= 1/6. Ta có bảng xác suất có điều kiện sau đây :

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

Bảng 2.02 Xác suất có điều kiện của chi tiêu/thu nhập của 60 hộ gia đình

X	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
P(Y/X)	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
P(Y/X)	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
P(Y/X)	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
P(Y/X)	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
P(Y/X)	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
P(Y/X)	-	1/6	-	1/7	1/6	1/6	-	1/7	1/6	1/7
P(Y/X)	-	-	-	1/7	-	-	-	1/7	-	1/7
$E(Y/X_i)$	**65**	**77**	**89**	**101**	**113**	**125**	**137**	**149**	**161**	**173**

$$E(Y / X_i) = \sum_j Y_j P(Y = Y_j / X = X_i)$$

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

- Chẳng hạn :

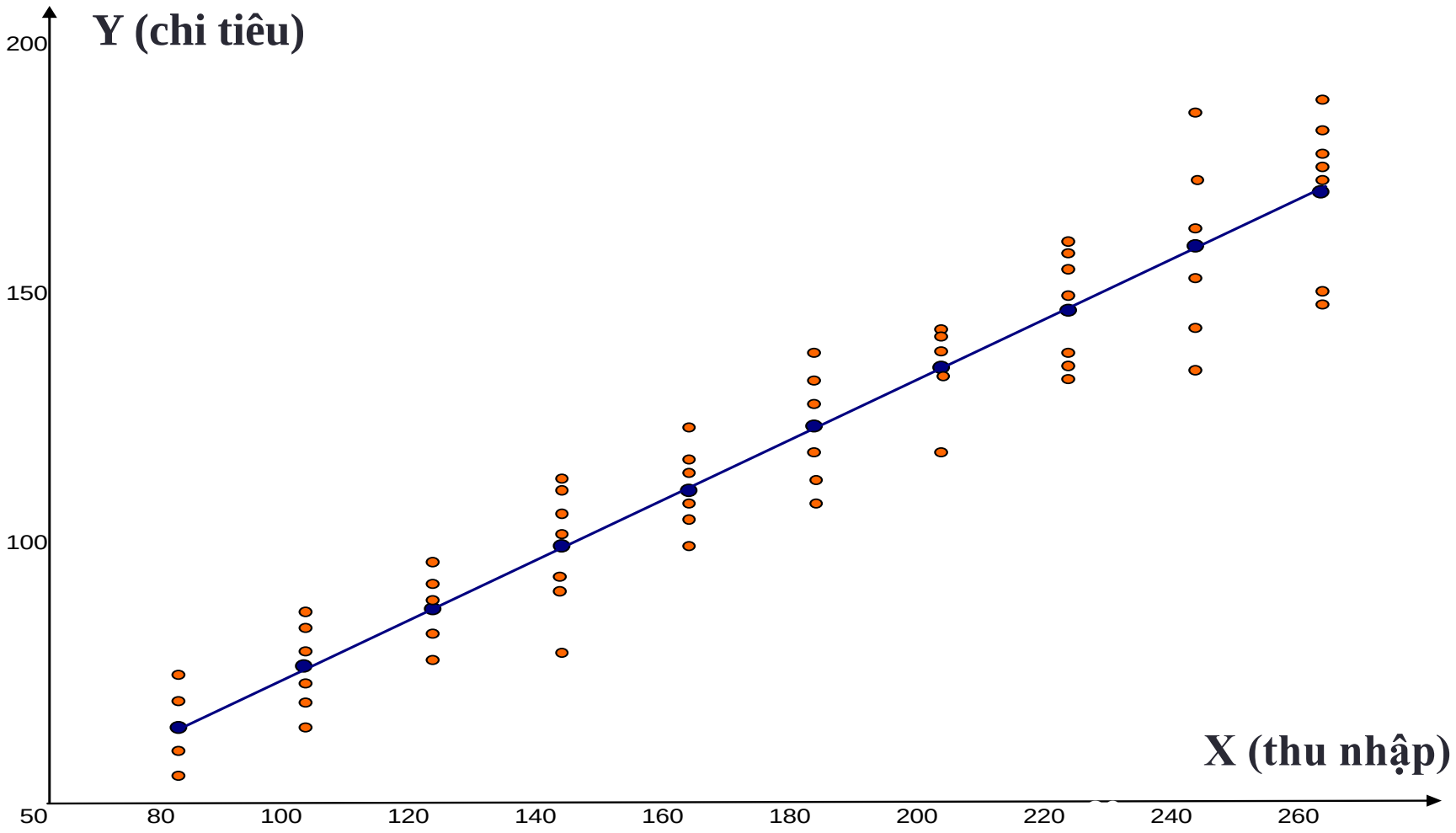
$$E(Y/100) = \sum_j Y_j P(Y = Y_j / X = 100)$$

= 65*1/6+ 70*1/6+ 74*1/6+ 80*1/6+ 85*1/6+ 88*1/6= 77

→ Biểu diễn các điểm của bảng 2.01 và các trung bình $E(Y/X_i)$ với i = 1,…,10 lên hệ tọa độ, ta được đồ thị sau đây :

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

Hình 2.02. Biểu đồ phân tán Y theo X và giá trị trung bình của Y theo X



2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

Biểu đồ 2 cho thấy:

- Mỗi »chấm » trên biểu đồ minh họa cho 1 quan sát thực tế, chính là tọa độ của cặp giá trị (X_i, Y_i)
- Nếu xét riêng từng hộ GĐ không thấy rõ xu hướng thay đổi của chi tiêu theo thu nhập.
- Nếu xét theo nhóm hộ gia đình, ta thấy:
  - ứng với cùng một mức thu nhập, có nhiều mức chi tiêu khác nhau
  - nếu chỉ quan tâm đến chi tiêu trung bình ($E(Y/X_i)$ thì thấy xu hướng tăng theo thu nhập.

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

→Vậy có thể xem $E(Y/X_i)$ là một hàm nào đó của biến giải thích X_i và biểu diễn như sau:

$$\mathbf{E(Y/X_i)= f(X_i) \quad [1]}$$

- Phương trình [1] gọi là hàm **hồi quy tổng thể- Population regression function (PRF).**
  - PRF cho biết giá trị trung bình của Y sẽ thay đổi như thế nào khi X nhận các giá trị khác nhau.
  - Nếu PRF có ***một biến độc lập*** thì gọi là ***hồi quy đơn (hồi quy hai biến)***, PRF có từ ***hai biến độc lập*** *trở lên* thì gọi là ***hồi quy bội (hồi quy nhiều biến)***.

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

- Giả sử PRF $E(Y/X_i)$ là hàm tuyến tính thì :

$$\mathbf{E(Y/X_i)= \beta_0+ \beta_1 X_i \quad [2]}$$

  - β_0, β_1= hệ số hồi quy
  - β_0= hệ số chặn
  - β_1= hệ số góc

- Phương trình [2] được gọi là **phương trình hồi quy tuyến tính đơn**.

2.1. Khái niệm về hàm hồi quy tổng thể (PRF)

- Thuật ngữ *"tuyến tính"* được hiểu theo hai nghĩa:
  - **Tuyến tính đối với tham số**.

  Ví dụ: $E(Y/X_i)= \beta_0+ \beta_1 X_i^2$ là hàm tuyến tính đối với tham số nhưng phi tuyến đối với biến.
  - **Tuyến tính đối với biến**.

  Ví dụ: $E(Y/X_i)= \beta_0+ \sqrt{\beta_1}\ X_i$ là hàm tuyến tính đối với biến nhưng phi tuyến với tham số.

→ *Trong phạm vi của môn học, hàm hồi quy tuyến tính được hiểu là hồi quy tuyến tính đối với các tham số*

2.2. Sai số ngẫu nhiên và bản chất của nó

- Giả sử ta có hàm hồi quy tổng thể $E(Y/X_i)$, vì $E(Y/X_i)$ là giá trị trung bình của biến Y với giá trị X_i đã biết, cho nên **các giá trị cá biệt Y_i không phải bao giờ cũng trùng với $E(Y/X_i)$, mà chúng xoay quanh $E(Y/X_i)$.**
- Kí hiệu u_i là chênh lệch giữa giá trị cá biệt Y_i và $E(Y/X_i)$, ta có :

$$\mathbf{u_i = Y_i - E(Y/X_i)} \qquad \mathbf{[3]}$$

- Hay :

$$\mathbf{Y_i = E(Y/X_i) + u_i} \qquad \mathbf{[4]}$$

→ u_i được gọi là biến ngẫu nhiên hay yếu tố ngẫu nhiên (hoặc nhiễu).

2.2. Sai số ngẫu nhiên và bản chất của nó

- Vậy các biến ngẫu nhiên ảnh hưởng đến mô hình là các biến nào và có thể đưa vào mô hình được không ?
- Câu trả lời là chúng ta có thể đưa nhiều biến ngẫu nhiên vào mô hình thông qua mô hình hồi quy bội, nhưng dù chúng ta có đưa vào bao nhiêu biến chăng nữa thì U_i vẫn tồn tại. (Vì sao?)

2.2. Sai số ngẫu nhiên và bản chất của nó

- Không thể biết rõ hết tất cả các yếu tố tác động đến biến phụ thuộc Y →U_i được sử dụng như yếu tố đại diện cho tất cả các biến tác động đến Y nhưng không có trong mô hình.
- Không phải lúc nào ta cũng tìm được số liệu của các biến tác động đến biến Y → phải loại các biến này khỏi mô hình.

2.2. Sai số ngẫu nhiên và bản chất của nó

- Có một số biến giải thích cho biến phụ thuộc Y nhưng những tác động của chúng tới biến Y là không đáng kể → không đưa các biến này vào mô hình. U_i sẽ đại diện cho chúng.
- Cần XĐ mục tiêu nghiên cứu→ sẽ có sự chọn lọc các biến đưa vào mô hình và làm nổi bật vai trò giải thích của các biến này đến biến phụ thuộc thay vì đưa vào mô hình một loạt các biến nhưng không tường minh.

2.3. Hàm hồi quy mẫu (SRF)

- Trong thực tế, ta không có điều kiện để khảo sát toàn bộ tổng thể → **ta không thể xây dựng được hàm hồi quy tổng thể (PRF).**
- Khi đó ta chỉ có thể ước lượng giá trị trung bình của biến phụ thuộc, hay nói cách khác, ước lượng hàm PRF từ một hoặc một số mẫu lấy ra từ tổng thể
- Tất nhiên, giá trị PRF mà ta ước lượng được khi đó không thể chính xác một cách tuyệt đối.
- **Hàm hồi quy được xây dựng trên cơ sở một mẫu được gọi là hàm hồi quy mẫu- SRF (Sample Regression Function).**

2.3. Hàm hồi quy mẫu (SRF)

- Ví dụ: Từ tổng thể 60 hộ gia đình, ta lấy ra ngẫu nhiên hai mẫu từ tổng thể này như sau :

Bảng 2.03. Mẫu thứ nhất- SRF1

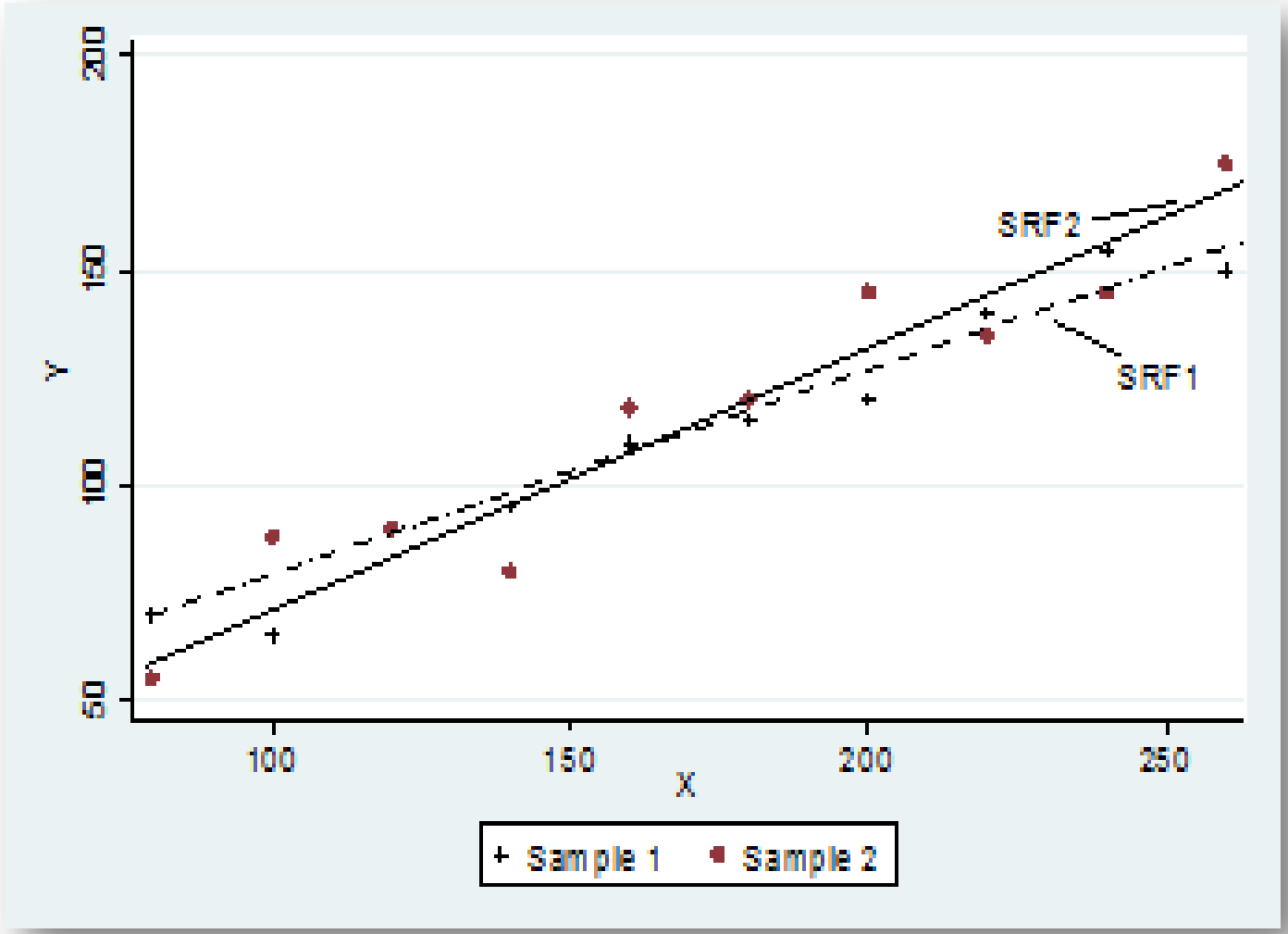
X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175

Bảng 2.04. Mẫu thứ hai- SRF2

X	80	100	120	140	160	180	200	220	240	260
Y	70	65	90	95	110	115	120	140	155	150

2.3. Hàm hồi quy mẫu (SRF)

Hình 2.03. Biểu đồ phân tán và đường hồi quy của hai mẫu SRF1 và SRF2



2.3. Hàm hồi quy mẫu (SRF)

- Mỗi dấu "chấm" trên hình 2.03 minh họa cho một quan sát thực tế, là tọa độ của một cặp giá trị (X_i, Y_i)
- Từ sự phân tán của các cặp giá trị, chúng ta phác họa được đường SRF.
- **Đường hồi quy của mẫu nào « gần » với đường hồi quy tổng thể hơn ?**
- Ta chỉ có thể biết đường nào tốt hơn khi có đường hồi quy tổng thể, tuy nhiên, trên thực tế, điều này **không có được do ta không thể khảo sát toàn bộ tổng thể.**

2.3. Hàm hồi quy mẫu (SRF)

- Mặc dù vậy, từ tổng thể, ta có thể rút ra được nhiều mẫu khác nhau và xây dựng được các đường hồi quy khác nhau.
- Những đường hồi quy mẫu này đều là ước lượng xấp xỉ cho đường hồi quy tổng thể
- Việc xem xét hàm hồi quy mẫu nào là xấp xỉ tốt cho hàm hồi quy tổng thể được xác định dựa theo một số **tiêu chuẩn** mà ta sẽ đề cập ở các phần sau.

2.3. Hàm hồi quy mẫu (SRF)

- Hàm hồi quy mẫu được biểu diễn theo hàm hồi quy tổng thể tương ứng.
- Ví dụ **PRF** có dạng :

$$\begin{cases} E(Y/X_i) = \beta_0 + \beta_1 X_i \\ Y_i = E(Y/X_i) + u_i = \beta_0 + \beta_1 X_i + u_i \end{cases}$$

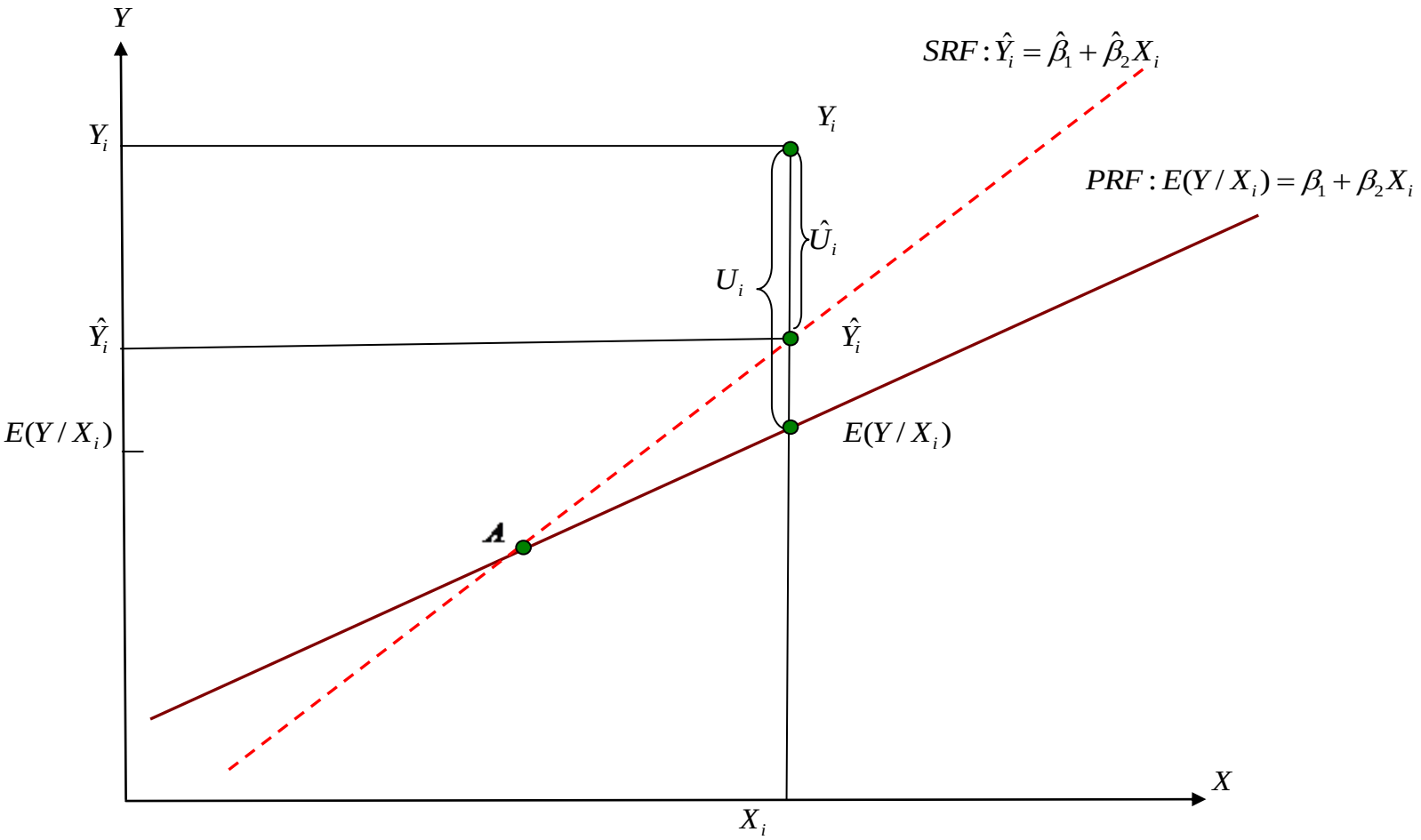
thì **SRF** được trình bày ở dạng tương ứng như sau :

$$\begin{cases} \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \\ Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i \end{cases}$$

với $\hat{Y}_i$ là ước lượng của E(Y/Xi) ;$\hat{\beta}_0$,$\hat{\beta}_1$ là ước lượng của β_0, β_1; $\hat{u}_i$ là ước lượng của u_i và được gọi là phần dư (residuals).

Mối liên hệ giữa SRF và PRF

Hình 2.04. Đường hồi quy tổng thể và đường hồi quy mẫu



Mối liên hệ giữa SRF và PRF

- Đồ thị 2.04 cho thấy mối liên hệ giữa SRF và PRF. Với $X = X_i$, ta có một mẫu quan sát là $Y = Y_i$.
- Dưới dạng hàm hồi quy mẫu SRF, giá trị quan sát Y_i được biểu diễn như sau :

$$Y_i = \hat{Y}_i + \hat{u}_i$$

- Dưới dạng hàm hồi quy tổng thể PRF, Y_i được viết như sau :

$$Y_i = E(Y/X_i) + u_i$$

Mối liên hệ giữa SRF và PRF

- Bây giờ, ta có thấy rằng, $\hat{Y}_i$ ước lượng « trên » giá trị thực của $E(Y/X_i)$ đối với những giá trị X_i nằm bên phải điểm A. Tương tự, $\hat{Y}_i$ ước lượng « dưới » giá trị thực của $E(Y/X_i)$ đối với những giá trị X_i nằm bên trái điểm A.

- Cần hiểu rằng việc ước lượng « trên » hay « dưới » giá trị thực là không thể tránh khỏi do có sự dao động (fluctuations) của việc lấy mẫu.

Mối liên hệ giữa SRF và PRF

- Vậy có quy tắc hay phương pháp nào để tìm ra hàm hồi quy mẫu « gần » với hàm hồi quy tổng thể nhất không ?
- Nói cách khác, làm thế nào để xác định được giá trị của các tham số $\hat{\beta}_1$, $\hat{\beta}_2$ gần với giá trị thực của β_1, β_2 nhất không, mặc dù trên thực tế, ta không bao giờ biết được các giá trị thực này.
- Phương pháp được áp dụng để ước lượng $\hat{\beta}_1$, $\hat{\beta}_2$ là phương pháp bình phương nhỏ nhất (Ordinary Least Square – OLS)

3. Phương pháp bình phương nhỏ nhất (OLS)

- Phương pháp OLS (Ordinary Least Square) do nhà toán học Đức Carl Friedrich Gauss đưa ra.
- Sử dụng phương pháp này kèm theo một vài giả thiết, các ước lượng thu được sẽ có một số tính chất đặc biệt, nhờ đó mà phương pháp này trở thành phương pháp mạnh nhất và phổ biến nhất trong phân tích hồi quy.

3.1. Nội dung phương pháp bình phương nhỏ nhất

- Giả sử hàm hồi quy tổng thể xác định hai biến có dạng như sau :

$$\textbf{PRF: } Y_i = \beta_0 + \beta_1 X_i + u_i \qquad [3.01]$$

- Do không thể trực tiếp ước lượng hàm PRF nên ta sẽ ước lượng nó thông qua hàm hồi quy mẫu có dạng :

$$\textbf{SRF: } \quad Y_i = \beta_0 + \beta_1 X_i + u_i = Y_i + u_i \qquad [3.02]$$

Trong đó $\hat{Y}_i$ là **giá trị dự đoán** của Y_i.

3.1. Nội dung phương pháp bình phương nhỏ nhất

Từ [3.02], ta có:

$$\hat{u}_i = Y_i - Y_i \qquad [3.03]$$

→[3.03] cho thấy ước lượng của biến ngẫu nhiên $\hat{u}_i$ là **chênh lệch** giữa giá trị thực và giá trị dự đoán của Y_i.

→ Nếu $\hat{u}_i$ càng nhỏ thì chênh lệch giữa Y_i và ước lượng$\hat{Y}_i$ càng nhỏ. Khi đó, giá trị của ước lượng$\hat{Y}_i$ càng gần với giá trị thực Y_i.

3.1. Nội dung phương pháp bình phương nhỏ nhất

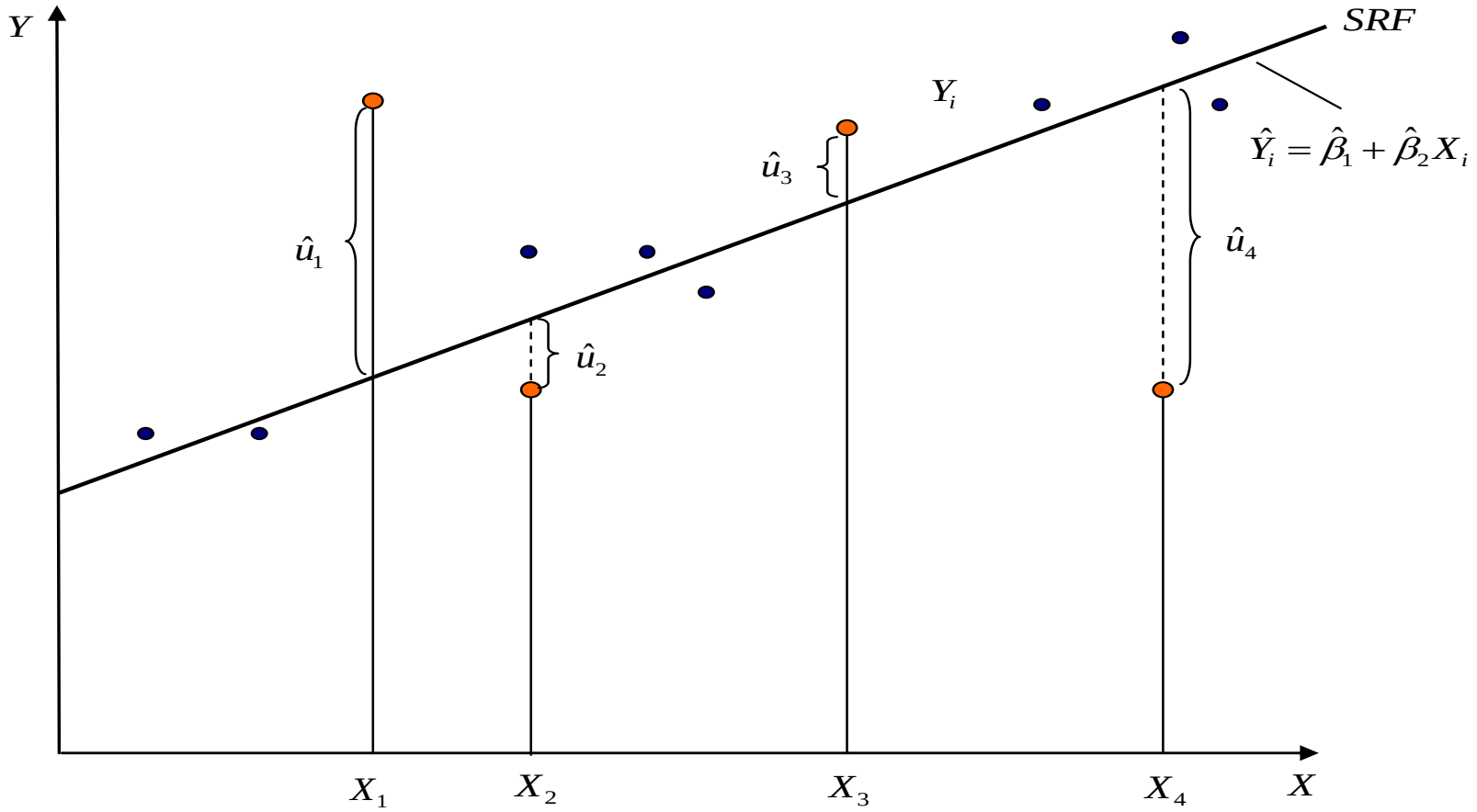
- Giả sử có n cặp quan sát giữa Y và X, ta sẽ thử đi tìm giá trị của hàm SRF sao cho Y_i gần với giá trị thực của Y nhất có thể.
- Để làm điều đó, ta sẽ áp dụng tiêu chuẩn: chọn hàm SRF nào có tổng các phần dư:

$$\sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)$$ đạt cực tiểu.

- Tuy nhiên, một cách trực quan, ta có thể thấy rằng đây không phải là phương pháp tối ưu vì các lý do sau đây.

3.1. Nội dung phương pháp bình phương nhỏ nhất

Hình 3.01. Tiêu chuẩn bình phương nhỏ nhất



3.1. Nội dung phương pháp bình phương nhỏ nhất

- Nếu áp dụng tiêu chuẩn cực tiểu hóa tổng các phần dư thì đồ thị 2.05 chỉ ra rằng các phần dư $\hat{u}_2$ và $\hat{u}_4$ tốt hơn các phần dư $\hat{u}_1$ và $\hat{u}_3$ vì chúng mang dấu âm (-).
- Vai trò của tất cả các phần dư mà ta nhận được bị đồng nhất hóa bất kể giá trị của chúng « gần » hay « xa » với các giá trị quan sát phân tán xung quanh đường SRF.
- Triệt tiêu ảnh hưởng của dấu

3.1. Nội dung phương pháp bình phương nhỏ nhất

Chúng ta có thể khắc phục được tình trạng này bằng cách tìm giá trị của SRF sao cho :

$$\sum_{i=1}^{n}\hat{u}_i^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \qquad [3.04]$$

đạt giá trị cực tiểu. Trong đó, $\sum_{i=1}^{n}\hat{u}_i^2$ là tổng bình phương các phần dư.

3.1. Nội dung phương pháp bình phương nhỏ nhất

- Phương pháp này cho phép vai trò của của $\hat{u}_1$; $\hat{u}_4$ và và $\hat{u}_2, \hat{u}_3$ ở trong ví dụ trên là như nhau.
- Với tiêu chuẩn cực tiểu tổng các phần dư thì tổng giá trị các phần dư có thể rất nhỏ mặc dù chúng phân tán xa SRF đến đâu. Nhưng điều này lại không thể xảy ra trong quy trình bình phương tối thiểu vì nếu $\hat{u}_i$ (giá trị tuyệt đối) càng lớn thì $\sum_{i=1}^{n} \hat{u}_i^2$ càng lớn.
- Các $\hat{u}_i$ có cùng độ lớn mà khác dấu sẽ không bị triệt tiêu nếu tính $\sum_{i=1}^{n} \hat{u}_i^2$

3.1. Nội dung phương pháp bình phương nhỏ nhất

Từ phương trình [3.03] ta có $\sum_{i=1}^{n}\hat{u}_i^2$ là một hàm của $\hat{\beta}_0$ và $\hat{\beta}_1$

$$\sum_{i=1}^{n}\hat{u}_i^2 = f(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

3.1. Nội dung phương pháp bình phương nhỏ nhất

- Ta biết rằng một hàm số f(X) đạt cực tiểu

$$\leftrightarrow \begin{cases} f'(X) = 0 \\ f''(X) > 0 \end{cases}$$

3.1. Nội dung phương pháp bình phương nhỏ nhất

- nên suy ra nếu coi $\sum_{i=1}^{n}\hat{u}_i^2$ là một hàm số thì $\sum_{i=1}^{n}\hat{u}_i^2$ đạt cực tiểu khi:

$$\begin{cases} f'(u) = 0 \\ f''(u) > 0 \end{cases}$$

3.1. Nội dung phương pháp bình phương nhỏ nhất

- Do đó, ta có $\hat{\beta}_0$ và $\hat{\beta}_1$ là nghiệm của hệ thống phương trình sau:

$$\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-1) = 0$$

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$$

3.1. Nội dung phương pháp bình phương nhỏ nhất

$$\frac{\partial f(\hat{\beta}_0, \hat{\beta}_1)}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)(-X_i) = 0$$

$$\hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i$$

3.1. Nội dung phương pháp bình phương nhỏ nhất

- Như vậy, $\hat{\beta}_0$ và $\hat{\beta}_1$ được tìm từ hệ phương trình:

$$\begin{cases} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} X_i + \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i \end{cases} \quad [3.05]$$

- Hệ phương trình [3.05] được gọi là hệ phương trình chuẩn trong đó n là kích thước mẫu (hay chính là số lượng các quan sát). Giải hệ phương trình trên ta được:

3.1. Nội dung phương pháp bình phương nhỏ nhất

$$\hat{\beta}_1 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \quad [3.06]$$

- Trong đó : $\bar{X}$ và $\bar{Y}$ là giá trị trung bình mẫu của X và Y;

3.1. Nội dung phương pháp bình phương nhỏ nhất

Thay $\hat{\beta}_1$ vào hệ phương trình [3.05] ta sẽ thu được $\hat{\beta}_0$ có giá trị là:

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad [3.07]$$

→ $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng của β_0 và β_1 được tính bằng phương pháp OLS và được gọi là **các ước lượng bình phương nhỏ nhất**.

Cho 1 mẫu ngẫu nhiên như sau:

X: Thu nhập của cá nhân trong 1 ngày, tính bằng 1000 đồng

Y: Chi tiêu của cá nhân trong 1 ngày, tính bằng 1000 đồng

X	**80**	**100**	**120**	**140**	**160**	**180**	**200**	**220**	**240**	**260**
Y	55	65	79	80	102	110	120	135	137	150

a. Tính các đặc trưng của X và Y

b. Ước lượng các tham số của mô hình hồi quy trên.

c. Viết phương trình hàm hồi quy mẫu.

3.2. Các tính chất của SRF theo OLS

Tính chất của các tham số ước lượng

- 1) $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng duy nhất ứng với 1 mẫu xác định gồm n quan sát (X_i, Y_i)
- 2) $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng điểm của β_1 và β_2
- 3) $\hat{\beta}_0$ và $\hat{\beta}_1$ là các đại lượng ngẫu nhiên. Với các mẫu khác nhau chúng sẽ có giá trị khác nhau.

3.2. Các tính chất của SRF theo OLS

Tính chất của đường SRF:

1. SRF đi qua điểm trung bình của dữ liệu mẫu $(\overline{X}, \overline{Y})$

2. Giá trị trung bình của $\hat{Y}_i$ bằng giá trị trung bình của các quan sát: $\overline{\hat{Y}} = \overline{Y}$

3. Tổng các phần dư bằng 0: $\sum_{i=1}^{n} \hat{u}_i = 0$

3.2. Các tính chất của SRF theo OLS

4. Các phần dư $\hat{u}_i$ không tương quan với giá trị ước lượng $\hat{Y}_i$: $\sum_{i=1}^{n} \hat{Y}_i \hat{u}_i = 0$

5. Các phần dư $\hat{u}_i$ không tương quan với X_i: $\sum_{i=1}^{n} \hat{u}_i X_i = 0$

3.3. CÁC TỔNG BÌNH PHƯƠNG ĐỘ LỆCH

- SST (Total Sum of Squares - Tổng bình phương sai số tổng cộng)

$$SST = \sum (Y_i - \bar{Y})^2$$

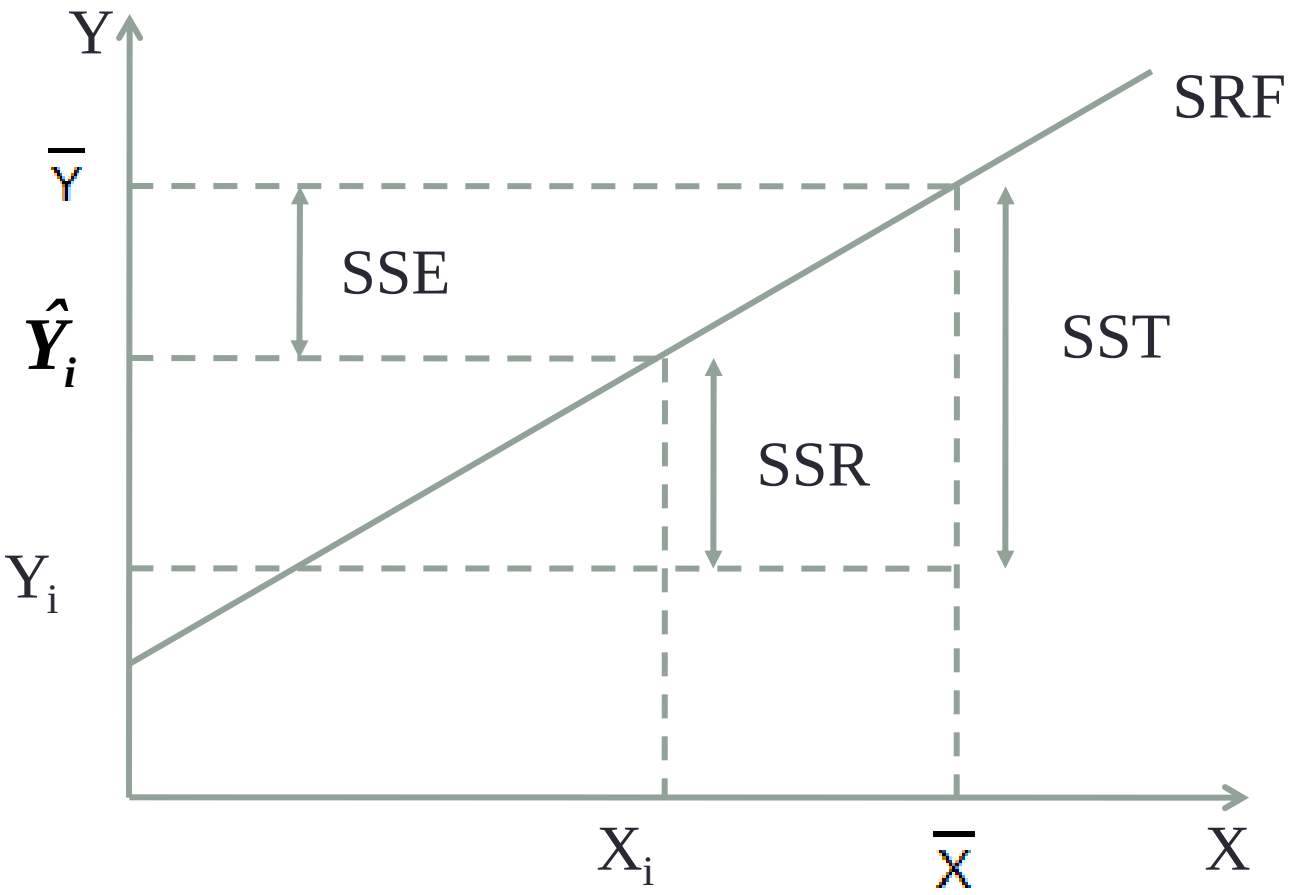
- SSE: (Explained Sum of Squares - Bình phương sai số được giải thích)

$$SSE = \sum (\hat{Y}_i - \bar{Y})^2$$

- SSR: (Residual Sum of Squares - Tổng bình phương các phần dư)

$$SSR = \sum_{i=1}^{n} \left(Y_i - Y_i \right)^2 = \sum u_i^2$$

3.3. CÁC TỔNG BÌNH PHƯƠNG ĐỘ LỆCH



Hình 2.3: Ý nghĩa hình học của SST, SSR và SSE

3.4. HỆ SỐ XÁC ĐỊNH R^2

Ta chứng minh được: SST = SSE + SSR

$$1 = \frac{SSE}{SST} + \frac{SSR}{SST}$$

3.4. HỆ SỐ XÁC ĐỊNH R^2

Hệ số xác định R^2: đo mức độ phù hợp của hàm hồi quy mẫu.

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Trong mô hình 2 biến:

$$R^2 = \frac{\hat{\beta}_1^2 \sum_{i=1}^{n} (X_i - \overline{X})^2}{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}$$

TÍNH CHẤT CỦA HỆ SỐ XÁC ĐỊNH R^2

$0 \leq R^2 \leq 1$

Cho biết % sự biến động của Y được giải thích bởi các biến số X trong mô hình.

$R^2 = 1$: đường hồi quy phù hợp hoàn hảo

$R^2 = 0$: X và Y không có quan hệ

Nhược điểm: R^2 tăng khi số biến X đưa vào mô hình tăng, dù biến đưa vào không có ý nghĩa.

=>Sử dụng R^2 điều chỉnh (adjusted R^2 , $\bar{R}^2$) để quyết định đưa thêm biến vào mô hình.

3.5. HỆ SỐ XÁC ĐỊNH ĐIỀU CHỈNH $\bar{R}2$

$$\bar{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k}$$

- Khi đưa thêm biến vào mô hình mà $\bar{R}^2$ tăng thì nên đưa biến vào và ngược lại.

3.6. HỆ SỐTƯƠNG QUAN r

Hệ số tương quan r: đo mức độ chặt chẽ của quan hệ tuyến tính giữa 2 đại lượng X và Y.

$$r_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{\sqrt{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2}\sqrt{n\sum_{i=1}^{n} Y_i^2 - (\sum_{i=1}^{n} Y_i)^2}}$$

Các tính chất của hệ số tương quan r

- r có thể âm hoặc dương, dấu của r phụ thuộc vào dấu của tử số, đó chính là dấu của cov(X,Y).
- r nằm giữa -1 và 1, tức là $-1 \leq r_{x,y} \leq 1$.
  - Nếu $r_{x,y}$ tiệm cận 1 → các biến tương quan cùng chiều
  - Nếu $r_{x,y}$ tiệm cận -1 → các biến tương quan ngược chiều
  - Nếu $r_{x,y}$ tiệm cận 0 → các biến không tương quan

- r có tính chất đối xứng : r(x,y)= r(y,x)
- r chỉ đo độ phụ thuộc tuyến tính giữa biến x và y, còn không có ý nghĩa trong các quan hệ phi tuyến. Đây là một hạn chế của hệ số tương quan r.
- Quan hệ tương quan mà r đo lường giữa x và y không nhất thiết phải là quan hệ nhân quả. Đây là hạn chế thứ hai của hệ số tương quan r.

Trong hồi qui đơn biến:

$$r = \pm\sqrt{R^2}$$

và r cùng dấu với $\hat{\beta}_1$

VD: $\hat{Y}_i = 6{,}25 + 0{,}75X_i$

Với $R^2 = 0{,}81 \Rightarrow r = 0{,}9$

3.7. Các giả thiết cơ bản của phương pháp OLS

Giả thiết 1: Trong mô hình tổng thể Y có mối quan hệ với X và u:

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

Giả thiết 2: Mẫu điều tra là mẫu ngẫu nhiên, kích cỡ n.

Giả thiết 3: X có các giá trị không đồng nhất.

Định lý 1: Ước lượng không chệch của các tham số

Với các giả thiết trên, ta có:

$$E(\beta_0) = \beta_0, \text{và } E(\beta_1) = \beta_1$$

Nghĩa là, $\beta_0 \text{và} \beta_1$ là ước lượng không chệch của $\beta_0 \text{và} \beta_1$

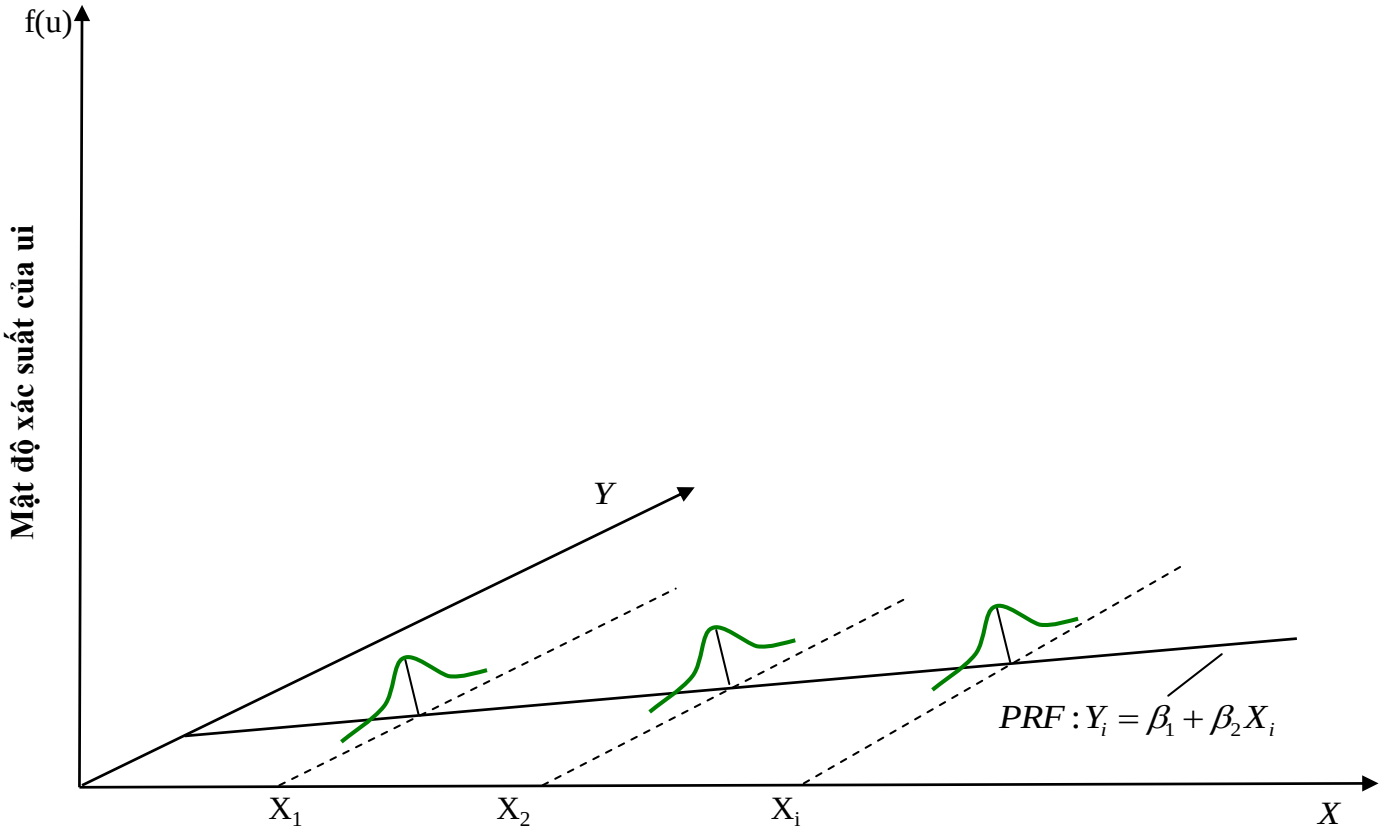
Giả thiết 4: Các u_i có phương sai thuần nhất (homoskedasticity), tức là các u_i có phương sai giống nhau với bất kỳ giá trị nào của X_i

$$var\ (u_i/X_i) = E[u_i - E(u_i/X_i)]^2 = E(u_i^2/X_i) = \sigma^2$$

→Phương sai của nhiễu thực chất phản ánh mức độ dao động hay phân tán của biến phụ thuộc Y quanh giá trị trung bình có điều kiện.

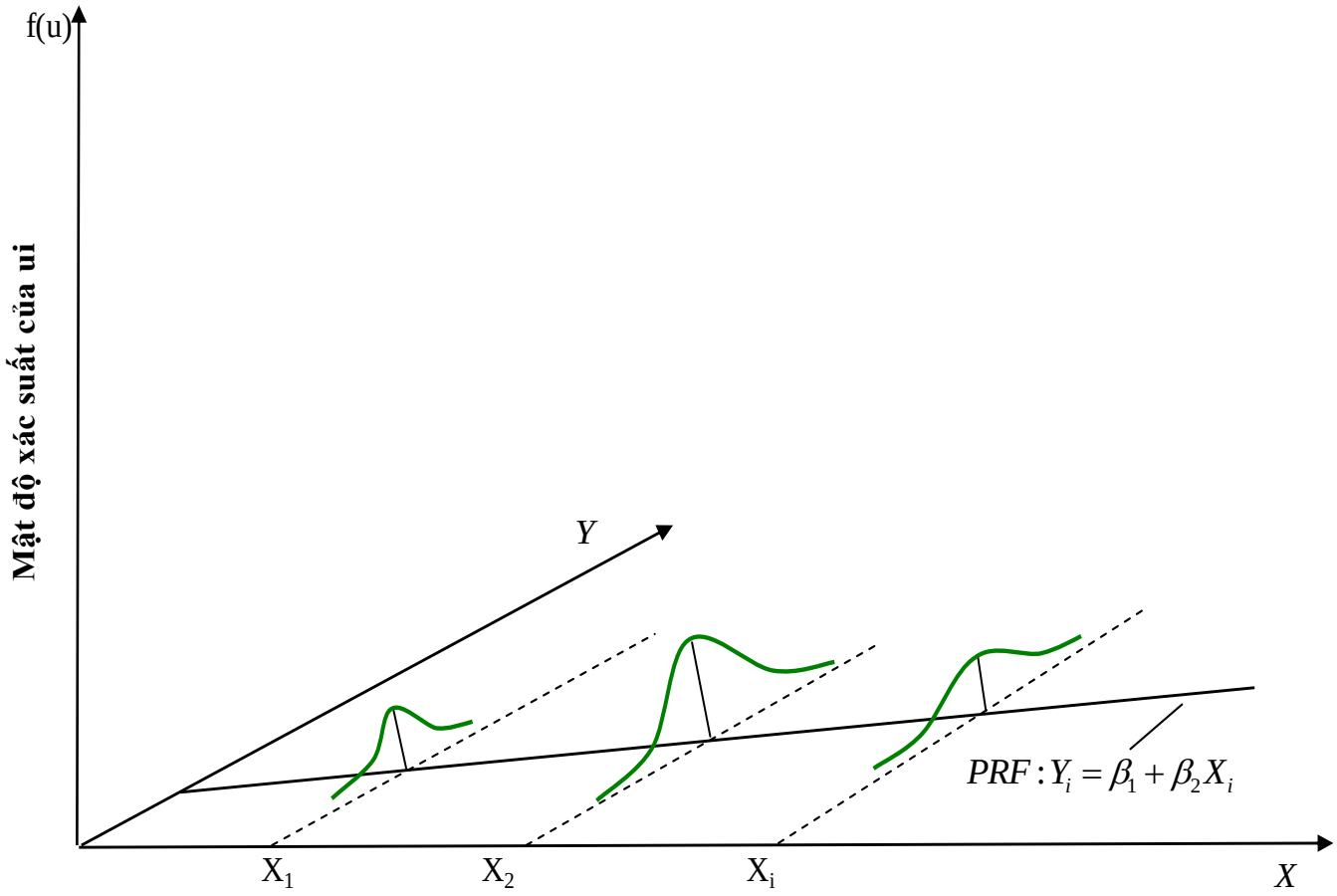
→ Giả thiết 5 có nghĩa là Y dao động quanh giá trị trung bình $E(Y/X_i)$ ứng với một giá trị của biến độc lập X nào đó với biên độ bằng nhau và không đổi. Tức là **giá trị phương sai có điều kiện của Y không thay đổi theo giá trị của X**.

Hình 3.04. Phương sai thuần nhất của nhiễu



- Trong thực tế, giả thiết 5 không phải lúc nào cũng thỏa mãn.
- Ví dụ, chi tiêu của những nhóm người có thu nhập thấp và thu nhập cao thường có khuynh hướng khác nhau.
  - Đối với nhóm thu nhập thấp, chi tiêu thường tập trung vào những hàng hóa thiết yếu.
  - Đối với nhóm thu nhập cao, ngoài các mặt hàng thiết yếu, còn có khoản chi cho các mặt hàng xa xỉ hoặc giải trí…
- ➔ có sự không đồng đều về chi tiêu giữa các nhóm thu nhập khác nhau ➔ giá trị phương sai có điều kiện của Y thay đổi theo giá trị của X ➔ hiện tượng phương sai không thuần nhất hoặc phương sai sai số thay đổi (heteroscedasticity).

Hình 3.05. Phương sai không thuần nhất của nhiễu



Định lý 2: Phương sai của các ước lượng

Với các giả thiết trên, ta có:

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{\sigma^2}{SST_x}$$

$$\text{var}(\hat{\beta}_0) = \frac{\sigma^2 n^{-1}\sum_{i=1}^{n} X_i^2}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}$$

σ^2 = var(u/x)= phương sai sai số

Định lý 3: Ước lượng không chệch của phương sai sai số của tổng thể:

Với các giả thiết 1-5, ta có:

$$E(\sigma^2) = \sigma^2$$

3.8. Độ chính xác của các ước lượng OLS

- Vì **phương sai** hay **độ lệch chuẩn** đặc trưng cho độ phân tán của đại lượng ngẫu nhiên so với giá trị trung bình của chúng, nên ta dùng chúng làm thước đo cho **chất lượng của ước lượng**.

3.8. Độ chính xác của các ước lượng OLS

Phương sai (var) và **độ lệch chuẩn của các ước lượng (sd)** được cho bởi các công thức sau :

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \quad \textbf{[3.08]}$$

$$\text{var}(\hat{\beta}_0) = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}\sigma^2 \quad \textbf{[3.10]}$$

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} \quad \textbf{[3.09]}$$

$$sd(\hat{\beta}_0) = \sigma\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n}(X_i - \overline{X})^2}} \quad \textbf{[3.11]}$$

σ^2 = var(u_i)= phương sai sai số, sd: độ lệch chuẩn

3.8. Độ chính xác của các ước lượng OLS

- Vì σ^2 khó biết được giá trị → σ^2 được ước lượng không chệch bằng công thức sau đây:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2} \quad \textbf{[3.12]}$$

  - $\hat{\sigma}^2$ = **ước lượng không chệch của σ^2**
  - n- 2 = số bậc tự do (number of degrees of freedom- df)
  - $\sum_{i=1}^{n} \hat{u}_i^2$ = tổng bình phương các phần dư (residual sum of squares- RSS)

3.8. Độ chính xác của các ước lượng OLS

- Lắp giá trị của $\hat{\sigma}^2$ vào 3.08 và 3.09, ta có ước lượng không chệch của $Var\beta_0$ và $Var\beta_1$

- Để có được ước lượng không chệch của $sd(\beta_0)$ và $sd(\beta_1)$, ta cần tính ước lượng của σ

$$\hat{\sigma} = \sqrt{\sigma^2} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}} = \sqrt{\frac{SSR}{n-2}} \qquad \textbf{[3.13]}$$

- σ là **sai số chuẩn của hồi quy**

3.8. Độ chính xác của các ước lượng OLS

$$sd(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}} = \frac{\sigma}{\sqrt{SST_x}}$$

- $$se(\beta_1) = \frac{\sigma}{\sqrt{SST_x}} \qquad \textbf{[3.14]}$$

- $se(\beta_1)$ **là sai số chuẩn của** β_1

- σ^2 = var(u/x) = var(y/x): phương sai sai số (error variance)
- σ : độ lệch chuẩn của sai số. σ càng lớn → sự phân tán của các giá trị ko quan sát được mà ảnh hưởng đến y càng lớn.
- $var(\beta_0)$ và $var(\beta_1)$ là phương sai của ước lượng
- $sd(\beta_0)$ và $sd(\beta_1)$ là độ lệch chuẩn của ước lượng
- σ^2 là ước lượng của σ^2
- $\sigma = \sqrt{\sigma^2}$ là ước lượng của σ , sai số chuẩn của hồi quy.
- $se(\beta_0)$ và $se(\beta_1)$ là sai số chuẩn của ước lượng

3.9. Đơn vị đo

- Ảnh hưởng của việc thay đổi đơn vị đo của biến phụ thuộc và biến độc lập đến giá trị ước lượng OLS
- Ví dụ với bộ số liệu "CEO Salary and Return on Equity"

Salary: lương hàng năm theo ngàn usd của CEO

Roe (average return on equity): lợi nhuận trung bình từ đầu tư của công ty trong 3 năm trước, %

$$salary = 963,191 + 18,501 roe \qquad [1]$$

Khi roe tăng 1%, lương được dự đoán là tăng 18501usd

3.9. Đơn vị đo

- Khi lương được tính theo usd → salarydol = 1000salary
- Đơn vị đo của roe không đổi
- $$salarydol = 963191 + 18501 roe \quad [2]$$

❖ **Khi đơn vị đo của biến độc lập ko đổi, đơn vị đo của biến phụ thuộc nhân hay chia một hằng số c khác 0 → giá trị của các hệ số ước lượng cũng nhân hoặc chia cho c.**

3.9. Đơn vị đo

- Khi đơn vị đo của salary không đổi
- Đơn vị đo của roedec = roe/100
- $$salary = 963,191 + 1850,1 roedec \quad [3]$$
- Hệ số của roedec gấp 100 lần hệ số của roe ở [1]

❖ **Khi đơn vị đo của biến phụ thuộc giữ nguyên, đơn vị đo của biến độc lập nhân hay chia với hằng số → β_1 sẽ chia hay nhân với c; nhưng β_0 không đổi.**

❖ **Đơn vị đo của Y và X thay đổi ko ảnh hưởng đến R^2**

3.10. Dạng hàm

Mô hình	Biến phụ thuộc	Biến độc lập	Cách giải thích β_1
Lin - lin	y	x	$\Delta y = \beta_1 \Delta x$
Lin-log	y	log(x)	$\Delta y = (\beta_1 / 100)\%\Delta x$
Log-lin	log(y)	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	log(y)	log(x)	$\%\Delta y = \beta_1 \%\Delta x$

3.10. Dạng hàm

Xem số liệu về "Wage and Education"
Wage: lương được đo bằng usd/1 giờ vào năm 1976 tại Mỹ
Educ: số năm học tại trường

1. Lin-lin:

$$wage = -0,90 + 0,54educ$$

- Mỗi năm học tăng thêm được dự đoán làm tăng mức lương theo giờ là 54 cent.
- Vì wage và educ có mối quan hệ tuyến tính → mức ảnh hưởng đến lương của mỗi năm học lên cao đều bằng 54 cent → mức ảnh hưởng của năm học lên cao thứ nhất = năm học lên cao thứ 20.

3.10. Dạng hàm

2. Log-lin:

$$\log wage = 0,584 + 0,083 educ$$

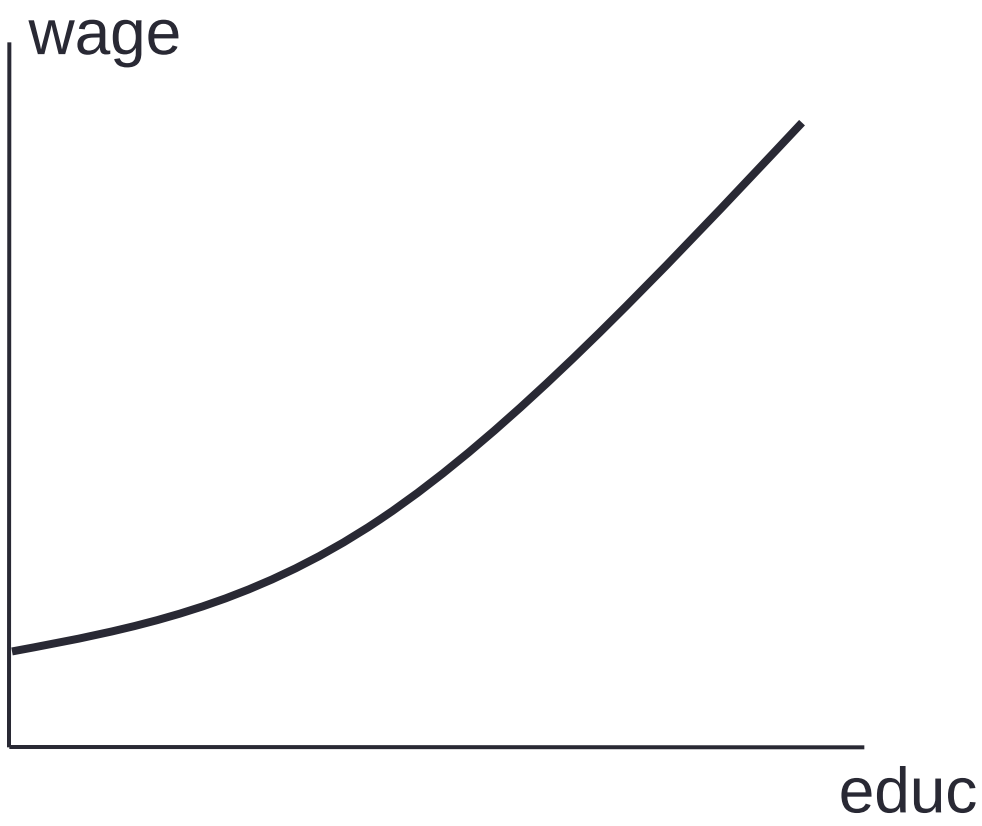
- Cách giải thích: $\%\Delta y = (100\beta_1)\Delta x$

- Mỗi năm học tăng thêm sẽ làm tăng lương ở một mức % cố định → sự thay đổi về lương tăng khi số năm theo học tăng → lợi ích tăng dần của việc học (increasing return to education)
- Mỗi một năm học lên cao sẽ làm tăng lương 8.3%.
- Học càng lên cao, giá trị càng lớn hơn

3.10. Dạng hàm

$$\log(\text{wage}) = \beta_0 + \beta_1 \text{educ} + u \rightarrow \text{wage} = \exp(\beta_0 + \beta_1 educ + u)$$

$$u = 0, \beta_1 > 0$$



3.10. Dạng hàm

3. Lin-log:

$$demand = 0,584 - 94,3\log(price)$$

- Cách giải thích $\Delta y = (\beta_1 / 100)\%\Delta x$
- Khi giá hàng hóa X tăng 1% thì lượng cầu của loại hàng này giảm 0,94 ngàn chiếc.
- (don vi: nghin chiec)

3.10. Dạng hàm

4. Log-log:

$$\log(\text{demand}) = 0,584 - 0,253\log(\textit{price})$$

- Cách giải thích $\%\Delta y = (\beta_1)\%\Delta x$
- Khi giá hàng hóa X tăng 1% thì lượng cầu của loại hàng này giảm 0.25%

- **Giả thiết 4:** Đại lượng sai số ngẫu nhiên (nhiễu) có kỳ vọng bằng 0, tức là: **E(u/X)=0**.

→Giả thiết này có nghĩa là các yếu tố không có trong mô hình mà U_i đại diện cho chúng không có ảnh hưởng hệ thống đến giá trị trung bình của Y. Về mặt hình học, giả thiết này được mô tả bằng đồ thị (hình 3.03)

→Đồ thị chỉ ra rằng với mỗi giá trị của X, các giá trị có thể có của Y xoay quanh giá trị trung bình. Phân bố của phần lớn hơn hay nhỏ hơn giá trị trung bình chính là các nhiễu u_i mà theo giả thiết này trung bình của các chênh lệch này phải bằng 0.

Hình 3.03. Phân phối có điệu kiện của các nhiễu u_i