CHƯƠNG II. HỒI QUY TUYẾN TÍNH ĐƠN

ThS. Vũ Thị Phương Mai

Khoa Kinh Tế Quốc Tế- Đại học Ngoại Thương

CHƯƠNG II. MÔ HÌNH HỒI QUY HAI BIẾN

- Giới thiệu mô hình hồi qui
- Hàm hồi quy tổng thể và hàm hồi quy mẫu
- Phương pháp bình phương nhỏ nhất (OLS)
- Phương pháp hợp lý tối đa (MLE)
- Ước lượng khoảng và kiểm định giả thiết TK
- Phân tích phương sai và kiểm định sự phù hợp của mô hình hồi quy

1. Giới thiệu mô hình hồi qui

1.1. Khái niệm về phân tích hồi qui

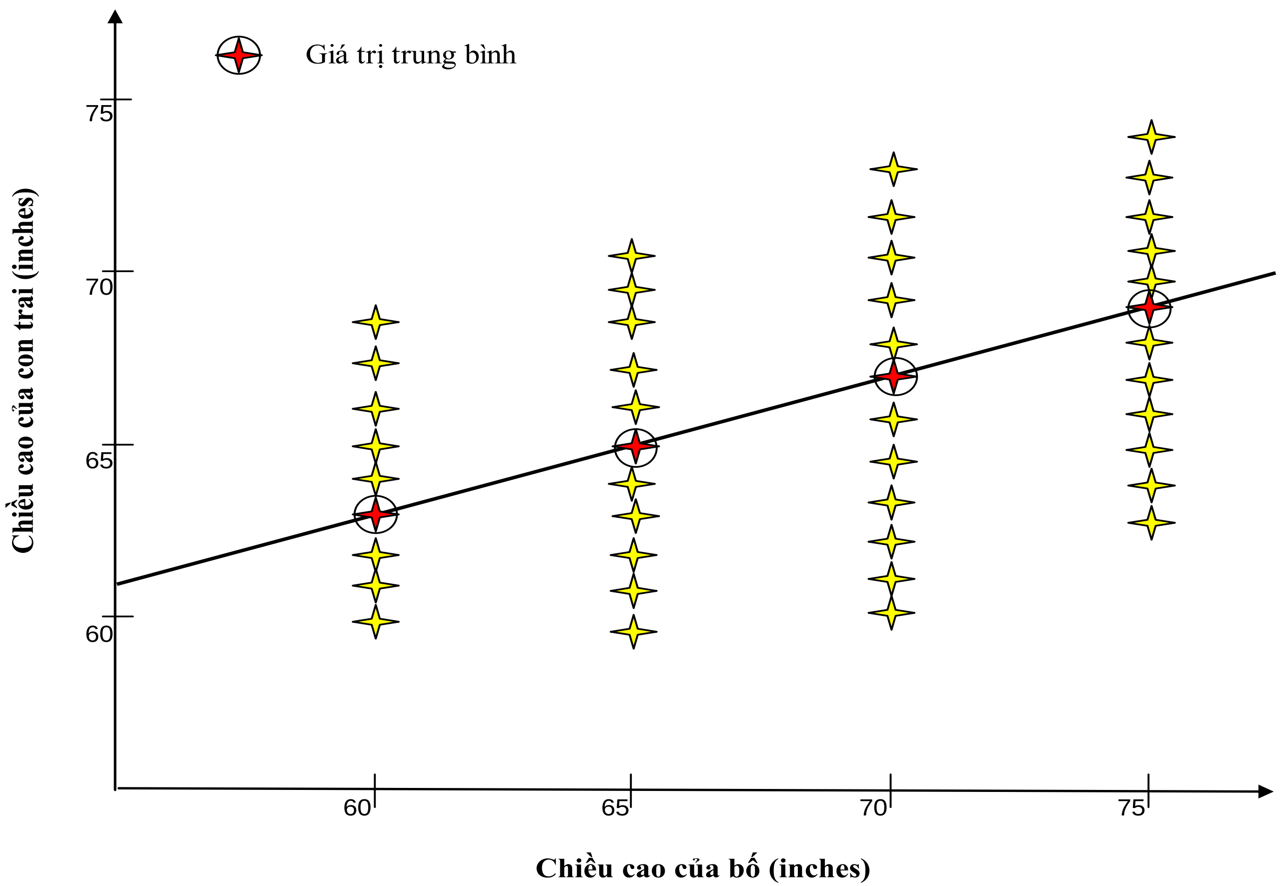
1.2. Sự khác nhau giữa các dạng quan hệ

1.1. Khái niệm về phân tích hồi qui

- Hồi qui là công cụ chủ yếu của KTL.
- *«regression to mediocrity»* nghĩa là « quy về giá trị trung bình »
- i khi Galton (1886) nghiên cứu sự phụ thuộc chiều cao của các cháu trai vào chiều cao của bố chúng.
- Ông đã xây dựng được đồ thị chỉ ra phân bố chiều cao của các cháu 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 các cháu trai sẽ là một khoảng dao động quanh một giá trị trung bình.
- Chiều cao của cha tăng thì chiều cao của các cháu trai cũng tăng.
- 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

i 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 thiết về bản chất của sự phụ thuộc đó.
- Dự báo giá trị trung bình của biến phụ thuộc khi biết giá trị của biến độc lập.
- Kết hợp cả ba vấn đề trên.

1.2.1. Quan hệ thống kê và quan hệ hàm số

- Trong quan hệ thống kê, biến phụ thuộc là đại lượng ngẫu nhiên, có phân bố xác suất.
- Ứng với mỗi giá trị đã biết của biến độc lập có thể có nhiều giá trị khác nhau của biến phụ thuộc. Phân tích hồi quy không xét đến các quan hệ hàm số.
- **Ví dụ**: sự phụ thuộc của năng suất một giống ngô vào nhiệt độ, lượng mưa, độ chiếu sáng, phân bón…là QH TK →không thể dự báo một cách chính xác năng suất của giống ngô này/ha (vì sao?)

- Trong quan hệ hàm số, các biến không phải là ngẫu nhiên
- ng với mỗi giá trị của biến độc lập chỉ có một giá trị của biến phụ thuộc.
- **Ví dụ**: trong vật lý, khi xét một động tử chuyển động đều, người ta có công thức :

S= v.t

  - S = độ dài quãng đường
  - i gian
  - t = thời gian

→ Đây sao?)

3. Phương pháp bình phương nhỏ nhất (OLS)

- 3.1. Nội dung phương pháp bình phương nhỏ nhất
- 3.2. Các tính chất thống kê của các ước lượng bình phương nhỏ nhất
- 3.3. Các giả thiết cơ bản của phương pháp bình phương nhỏ nhất
- 3.4. Độ chính xác của các ước lượng bình phương nhỏ nhất
- 3.5. Tiêu chuẩn của các ước lượng bình phương nhỏ nhất- Định lý Gauss- Markov
- 3.6. Phân bố xác suất của các ước lượng bình phương nhỏ nhất

3. Phương pháp bình phương nhỏ nhất (OLS)

- Phương pháp OLS (Ordinary Least Square) do nhà toàn 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 :

PRF: $Y_i = \beta_1 + \beta_2 X_i + u_i$ [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 :

SRF: $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i = \hat{Y}_i + \hat{u}_i$ [3.02]

- Trong đó $\hat{Y}_i$ là ước lượng của Y_i.
- Từ [3.02], ta có:

$$\hat{u}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \qquad [3.03]$$

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

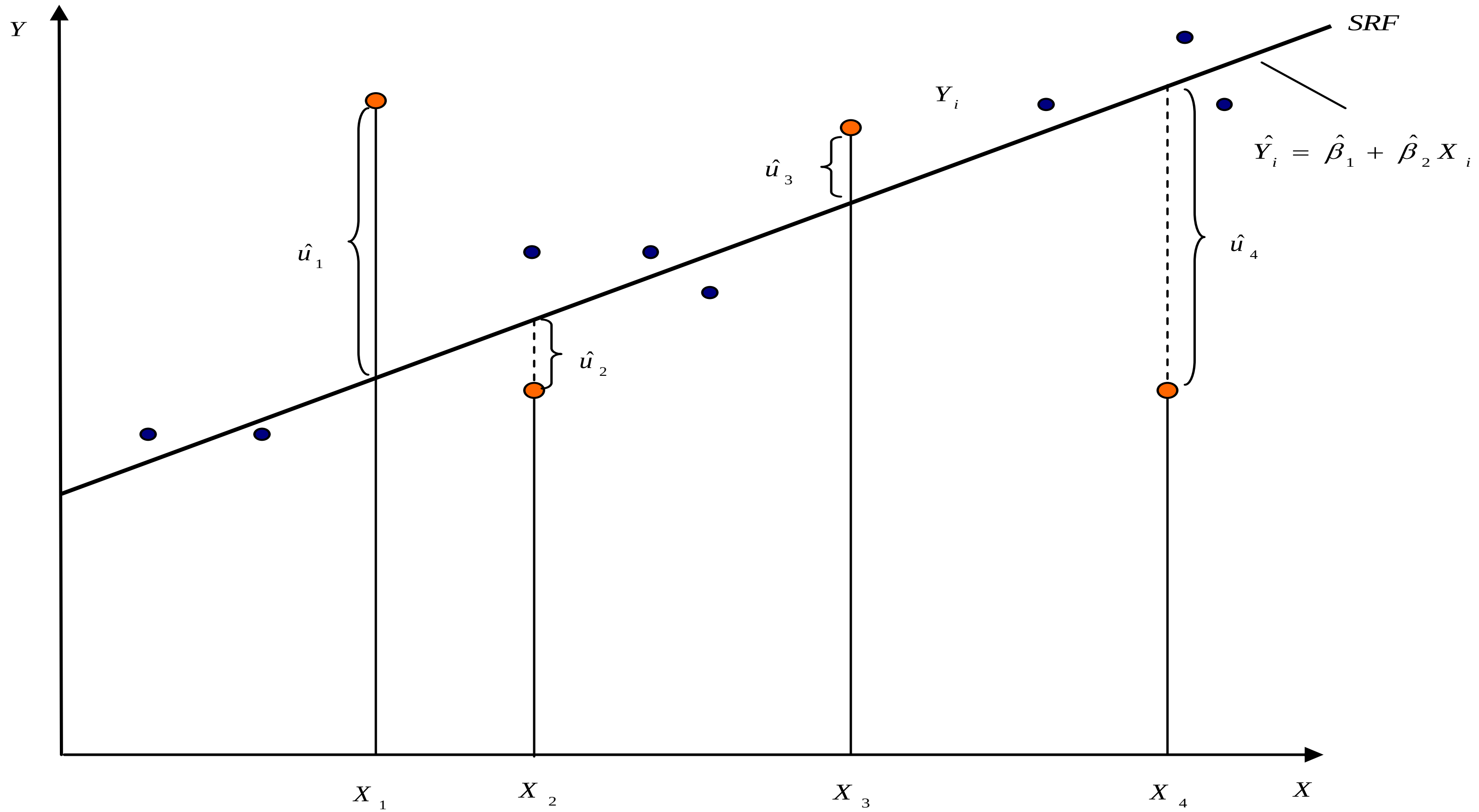
- Bây giờ, ta giả sử có n cặp quan sát giữa Y và X, ta sẽ thử đi tìm giá trị của hàm SRF sao cho nó gần với giá trị thực của Y nhất có thể. Để làm điều đó, ta sẽ áp dụng tiêu chuẩn: chọn hàm SRF nào có tổng các phần dư:

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

- Tuy nhiên, một cách trực quan, ta có thể thấy rằng đây không phải là phương pháp tối ưu vì lí 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 $\sum_{i=1}^{n} \hat{u}_i$ 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 (-). Mặc dù vậy khi cộng tổng các phần dư này lại ($\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \hat{u}_4$) thì vai trò của tất cả các phần dư này lại như nhau. Hay nói một cách khác, 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. Hậu quả của việc này là tổng đại số các phần dư $\hat{u}_i$ rất nhỏ (thậm chí bằng 0) mặc cho $\hat{u}_i$ phân tán xa SRF đến mấy.
- Để minh họa rõ hơn, ta hãy thử đặt giá trị của $\hat{u}_1, \hat{u}_2, \hat{u}_3, \hat{u}_4$ lần lượt là 10, -2, +2 và -10. Tổng đại số của các phần dư này bằng 0 mặc dù $\hat{u}_1$ và $\hat{u}_4$ phân tán xa hơn SRF so với $\hat{u}_2$ và $\hat{u}_3$.

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 :

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \quad [3.04]$$

đạt giá trị cực tiểu. Trong đó, $\sum_{i=1}^{n} \hat{u}_i^2$ là tổng bình phương các phần dư. Bằng việc bình phương $\hat{u}_i$, phương pháp này cho phép đề cao vai trò của của $\hat{u}_1$ và $\hat{u}_4$ hơn là $\hat{u}_2$ và $\hat{u}_3$ như trong ví dụ bên trên.

- 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.

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 $\hat{\beta}_1$ và $\hat{\beta}_2$:

$$\sum_{i=1}^{n} \hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2) = \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Nhắc lại về cực trị của hàm số

- 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 ↔

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

- Do đó, ta có $\hat{\beta}_1$ và $\hat{\beta}_2$ là nghiệm của hệ thống phương trình sau:

- $\dfrac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} 2(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-1) = 0$ hay $n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i$

- $\dfrac{\partial f(\hat{\beta}_1, \hat{\beta}_2)}{\partial \hat{\beta}_2} = \sum_{i=1}^{n} 2(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-X_i) = 0$ hay $\hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i$

3.1. Nội dung phương pháp bình phương nhỏ nhất

- Như vậy, $\hat{\beta}_1$ và $\hat{\beta}_2$ được tìm từ hệ phương trình:
- $$\begin{cases} n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i = \sum_{i=1}^{n} Y_i \\ \hat{\beta}_1 \sum_{i=1}^{n} X_i + \hat{\beta}_2 \sum_{i=1}^{n} X_i^2 = \sum_{i=1}^{n} Y_i X_i \end{cases} \qquad [3.05]$$
- Hệ phương trình [3.05] được gọi là hệ phương trình chuẩn (normal equations), 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

- $$\hat{\beta}_2 = \frac{n\sum_{i=1}^{n} X_i Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \quad [3.06]$$
- Trong đó : $\overline{X}$ và $\overline{Y}$ là giá trị trung bình mẫu của X và Y;

  $x_i = (X_i - \overline{X})$ và $y_i = (Y_i - \overline{Y})$
- Thay $\hat{\beta}_2$ vào hệ phương trình [3.06] ta sẽ thu được $\hat{\beta}_1$ có giá trị là:
- $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i^2 \sum_{i=1}^{n} Y_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} X_i Y_i}{n\sum_{i=1}^{n} X_i^2 - (\sum_{i=1}^{n} X_i)^2} = \overline{Y} - \hat{\beta}_2 \overline{X} \quad [3.07]$$

→ $\hat{\beta}_1$ và $\hat{\beta}_2$ là các ước lượng của β_1 và β_2 được tính bằng phương pháp OLS và được gọi là **các ước lượng bình phương nhỏ nhất**.

Ví dụ 1

- Ví dụ 1: Bảng 3.01 sau đây cho số liệu về tiêu dùng (Y) và thu nhập (X) trong 10 năm của một quốc gia.

Năm	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Y_i	**7389,99**	**8169,65**	**8831,71**	**8652,84**	**8788,08**	**9616,21**	**10593,45**	**11186,11**	**12758,09**	**13869,62**
X_i	**8000**	**9000**	**9500**	**9500**	**9800**	**11000**	**12000**	**13000**	**15000**	**16000**

- Giả sử rằng sự phụ thuộc E(Y/X) có dạng tuyến tính đối với cả biến số và tham số.

a. Viết phương trình hàm hồi quy mẫu.

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

Gợi ý: *Sử dụng Excel để tính toán!*

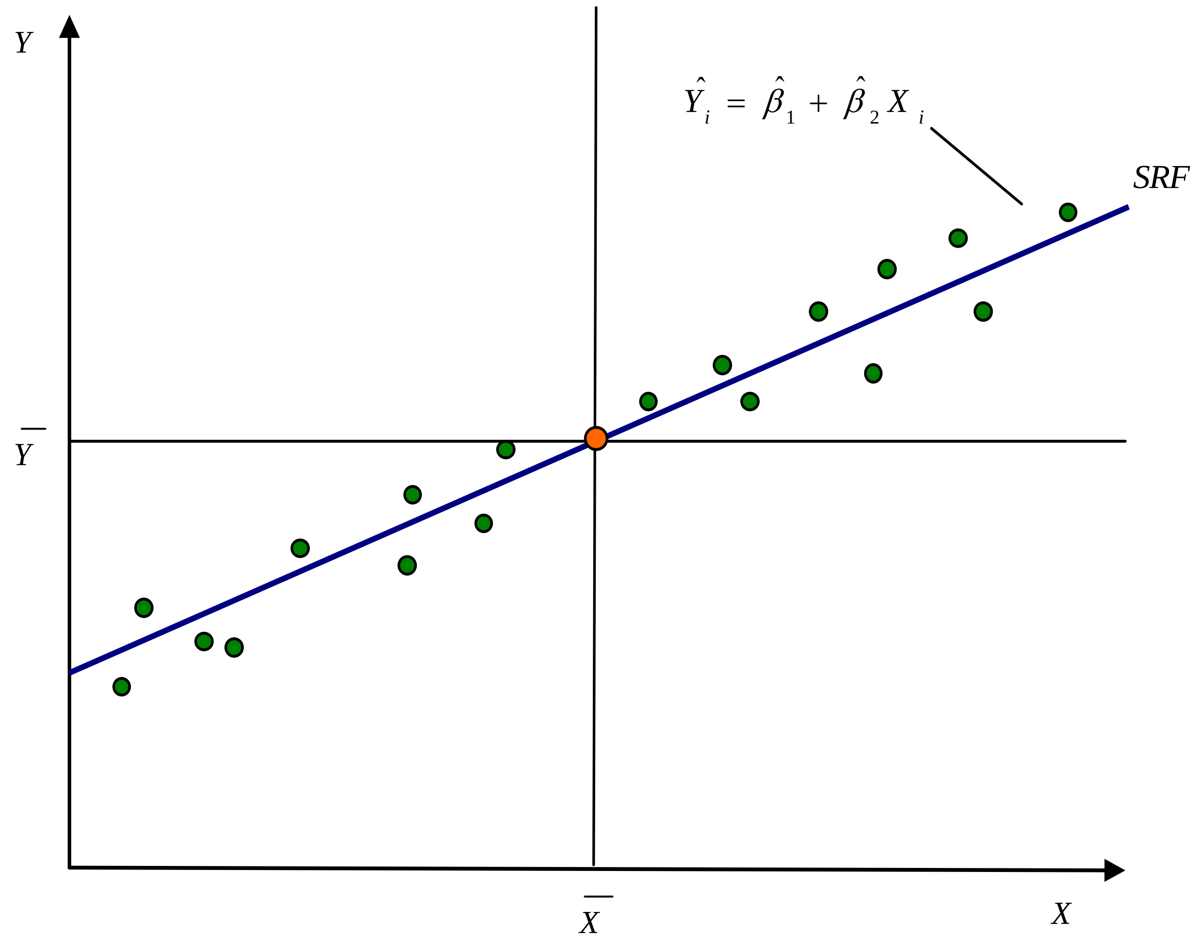
3.2. Các tính chất thống kê của các ước lượng OLS

- 1) $\hat{\beta}_1$ và $\hat{\beta}_2$ được xác định một cách duy nhất ứng với n cặp quan sát (X_i, Y_i)
- 2) $\hat{\beta}_1$ và $\hat{\beta}_2$ là các ước lượng điểm của β_1 và β_2 và là các đại lượng ngẫu nhiên, với các mẫu khác nhau chúng sẽ có giá trị khác nhau.
- 3) Đường hồi quy mẫu (**SRF**): $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ có các tính chất sau đây :
  - **a. SRF** đi qua trung bình mẫu $(\overline{X}, \overline{Y})$, nghĩa là : $\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X}$

→Tính chất này có thể được biểu diễn trên đồ thị như sau :

3.2. Các tính chất thống kê của các ước lượng OLS

Hình 3.02. Biểu đồ đường hồi quy đơn đi qua giá trị TB mẫu của X và Y



3.2. Các tính chất thống kê của các ước lượng OLS

- **b.** Giá trị trung bình của $\hat{Y}_i$ bằng giá trị trung bình của các quan sát: $\overline{\hat{Y}} = \overline{Y}$
- **c.** Giá trị trung bình của các phần dư bằng 0 : $\sum_{i=1}^{n} \hat{u}_i = 0$

→Từ tính chất này ta có thể suy ra được dạng hàm phương sai như sau:

Ta có : $Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \rightarrow \sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i + \sum_{i=1}^{n} \hat{u}_i$

$$\leftrightarrow \sum_{i=1}^{n} Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum_{i=1}^{n} X_i \quad \text{do} \quad \sum_{i=1}^{n} \hat{u}_i = 0$$

Chia cả hai vế của đẳng thức trên cho n ta được :

$$\overline{Y} = \hat{\beta}_1 + \hat{\beta}_2 \overline{X} \quad \rightarrow Y_i - \overline{Y} = \hat{\beta}_2 (X_i - \overline{X}) + \hat{u}_i$$

Hay $y_i = \hat{\beta}_2 x_i + \hat{u}_i$ → ***Đây được gọi là dạng hàm phương sai (deviation form) biểu thị độ lệch của giá trị quan sát so với giá trị trung bình của chúng.***

Từ đây, dễ dàng thấy đường hồi quy mẫu có dạng gốc $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ có thể được viết dưới dạng là : $\hat{y}_i = \hat{\beta}_2 x_i$

3.2. Các tính chất thống kê của các ước lượng OLS

- **d**. Các phần dư $\hat{u}_i$ không tương quan với giá trị ước lượng $\hat{Y}_i$:

$$\sum_{i=1}^{n} \hat{Y}_i \hat{u}_i = 0$$

- Hay ở dạng hàm phương sai, ta sẽ có : $\sum_{i=1}^{n} \hat{y}_i \hat{u}_i = 0$

- **e.** 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 giả thiết cơ bản của phương pháp OLS

- Trong phân tích hồi quy, mục đích của chúng ta là ước lượng, dự báo về tổng thể, tức là ước lượng $\mathbf{Y_i = \beta_1 + \beta_2 X_i + u_i}$.
- Các ước lượng $\hat{\beta}_1$ và $\hat{\beta}_2$ tìm được bằng phương pháp bình phương nhỏ nhất (OLS) là các ước lượng điểm của β_1 và β_2.
- Chúng ta không biết được chất lượng của các ước lượng điểm này như thế nào ngoại trừ việc biết rằng các ước lượng này phụ thuộc vào :
  - Dạng hàm của mô hình được lựa chọn
  - Phương pháp ước lượng được sử dụng
  - X_i và u_i
  - Kích thước mẫu

3.3. Các giả thiết cơ bản của phương pháp OLS

- **Giả thiết 1:** Các giá trị của X được xác định trước và không phải là đại lượng ngẫu nhiên (X_i không ngẫu nhiên).

→Giả thiết này không có gì mới vì phân tích hồi quy được đề cập là phân tích hồi quy có điều kiện, phụ thuộc và các giá trị X đã cho.

→Ví dụ như khi khảo sát mối quan hệ giữa thu nhập và chi tiêu thì các mức thu nhập mà chúng ta dự định sẽ tiến hành điều tra khảo sát đã được xác định trước.

3.3. Các giả thiết cơ bản của phương pháp OLS

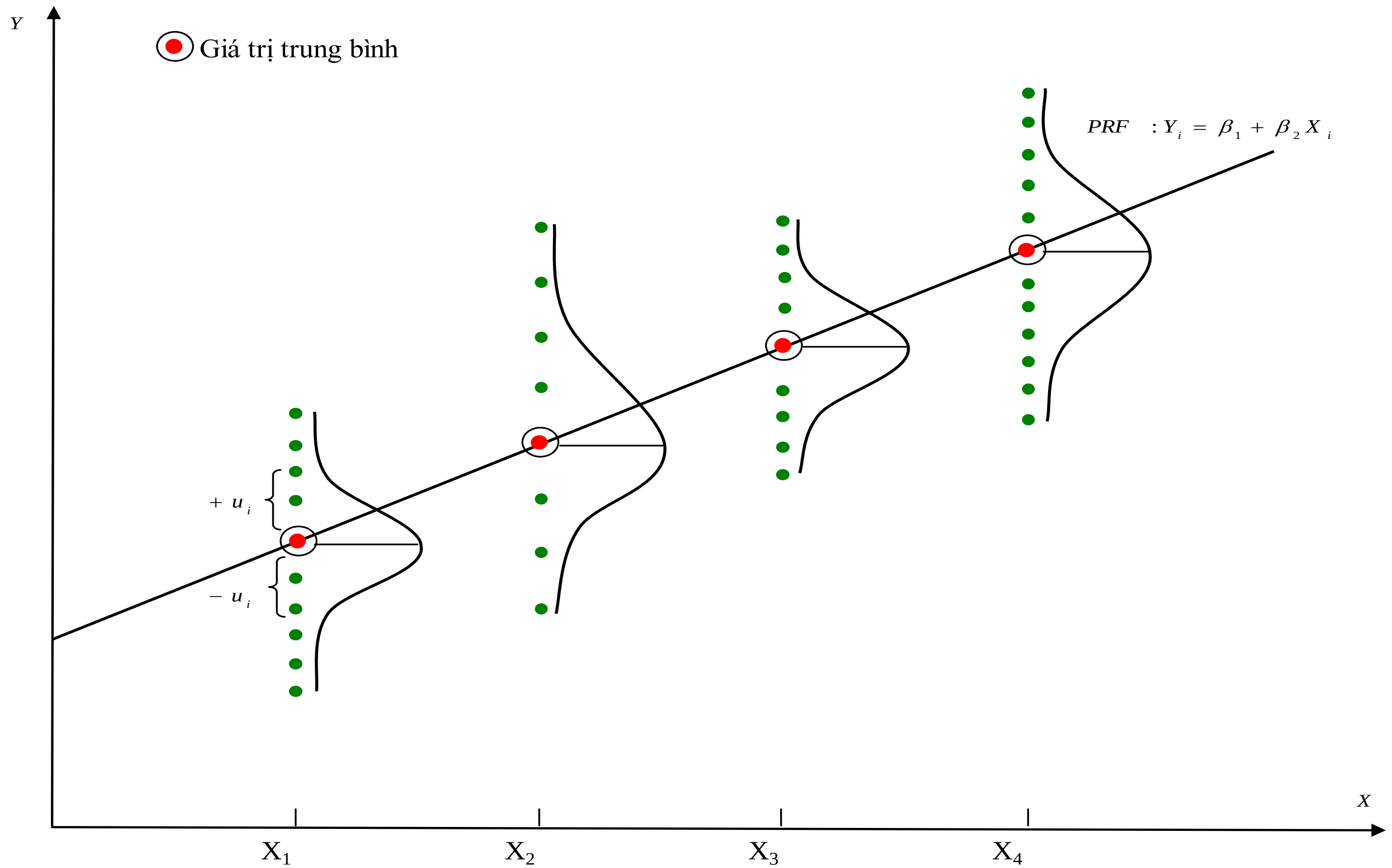
- **Giả thiết 2:** Đại lượng sai số ngẫu nhiên (nhiễu) có kỳ vọng bằng 0, tức là: $\mathbf{E(u_i/X_i)=0}$.

→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.

3.3. Các giả thiết cơ bản của phương pháp OLS

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



3.3. Các giả thiết cơ bản của phương pháp OLS

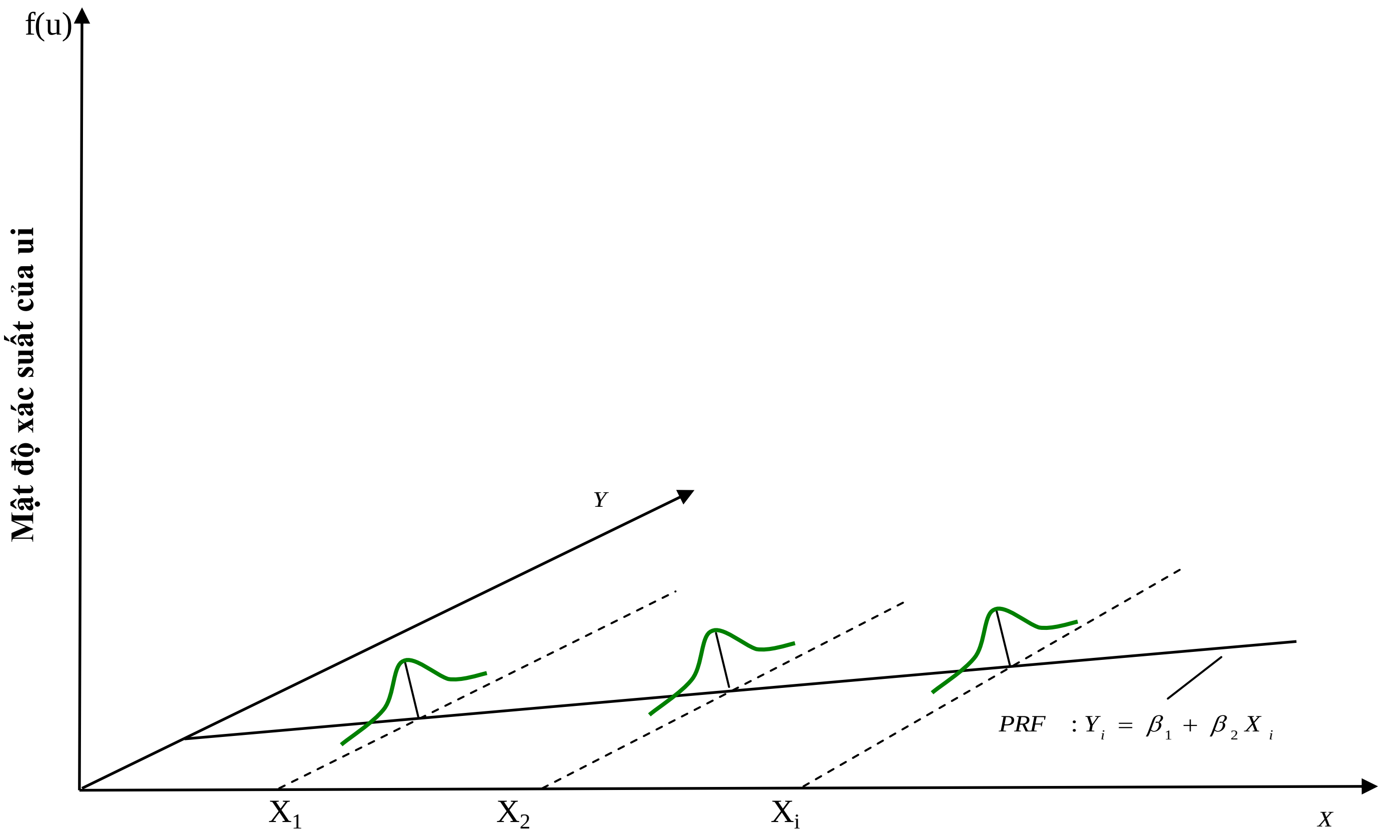
- **Giả thiết 3:** Các u_i có phương sai thuần nhất (Homoscedasticity of u_i), tức là các u_i có phương sai là một hằng số dương (không đổi) : **var (u_i/X_i)= $E[u_i- E(u_i/X_i)]^2$= $E(u_i^2/X_i)$= σ^2**.

→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.

→Như vậy, giả thiết 3 có nghĩa là biến phụ thuộc 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.

3.3. Các giả thiết cơ bản của phương pháp OLS

Hình 3.04. Phương sai thuần nhất của nhiễu

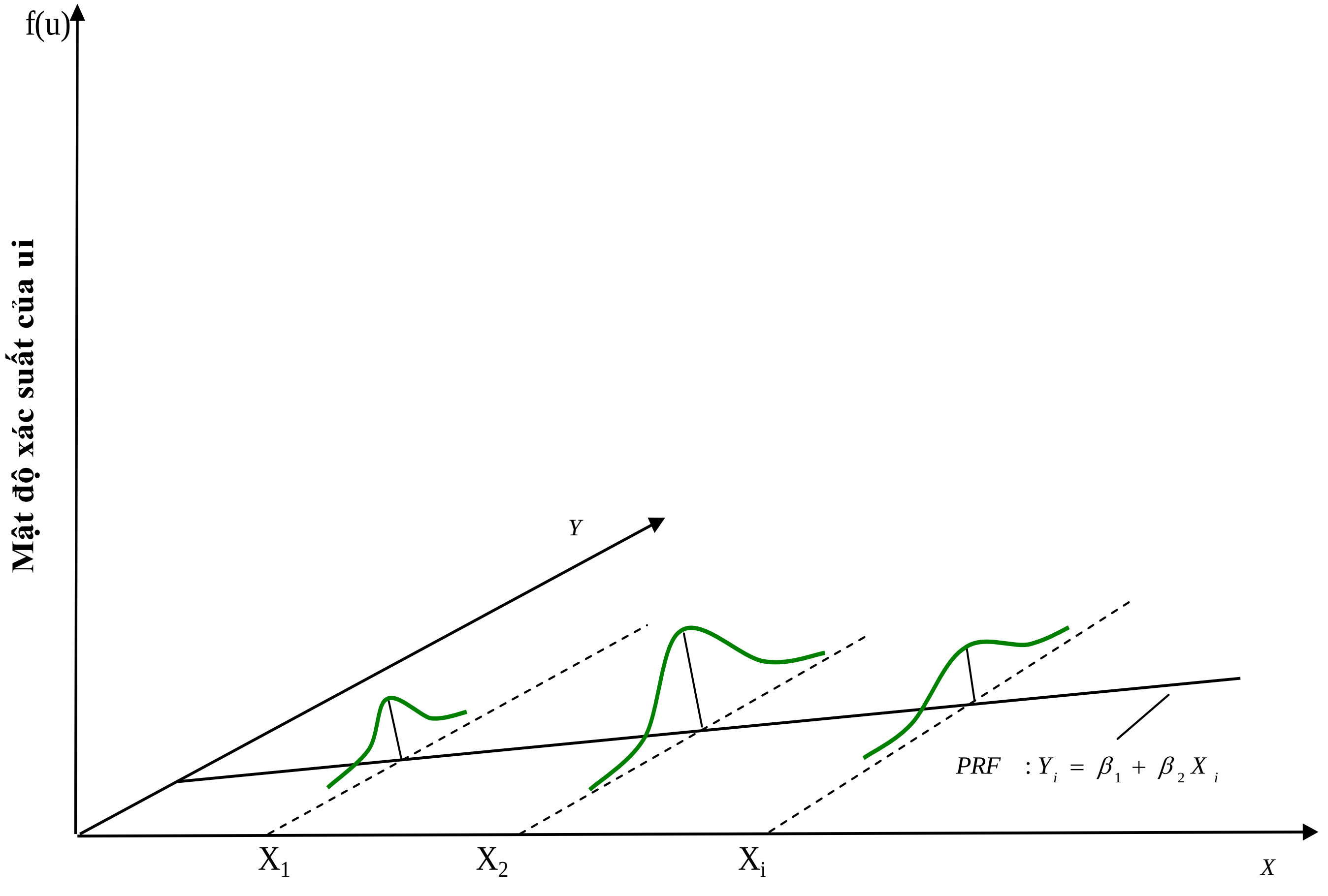


3.3. Các giả thiết cơ bản của phương pháp OLS

- Trong thực tế, giả thiết 3 không phải lúc nào cũng thỏa.
- Thí dụ như 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…
- Do vậy, có sự không đồng đều về chi tiêu giữa các nhóm thu nhập khác nhau. Trong trường hợp này, giá trị phương sai có điều kiện của Y thay đổi theo giá trị của X. Hiện tượng này được gọi là hiện tượng phương sai không thuần nhất (heteroscedasticity).

3.3. Các giả thiết cơ bản của phương pháp OLS

Hình 3.05. Phương sai không thuần nhất của nhiễu



3.3. Các giả thiết cơ bản của phương pháp OLS

- **Giả thiết 4** : Không có tự tương quan (no autocorrelation) giữa các sai số ngẫu nhiên (nhiễu) u_i. Điều này có nghĩa là : với hai giá trị bất kì X_i và X_j ($i \neq j$), hệ số tương quan giữa hai nhiễu bất kì của chúng ui và uj ($i \neq j$) là bằng 0 :

$$\mathbf{cov(u_i, u_j/X_i, X_j) = E\{[u_i - E(u_i)]/X_i\}\{[u_j - E(u_j)]/X_j\}}$$

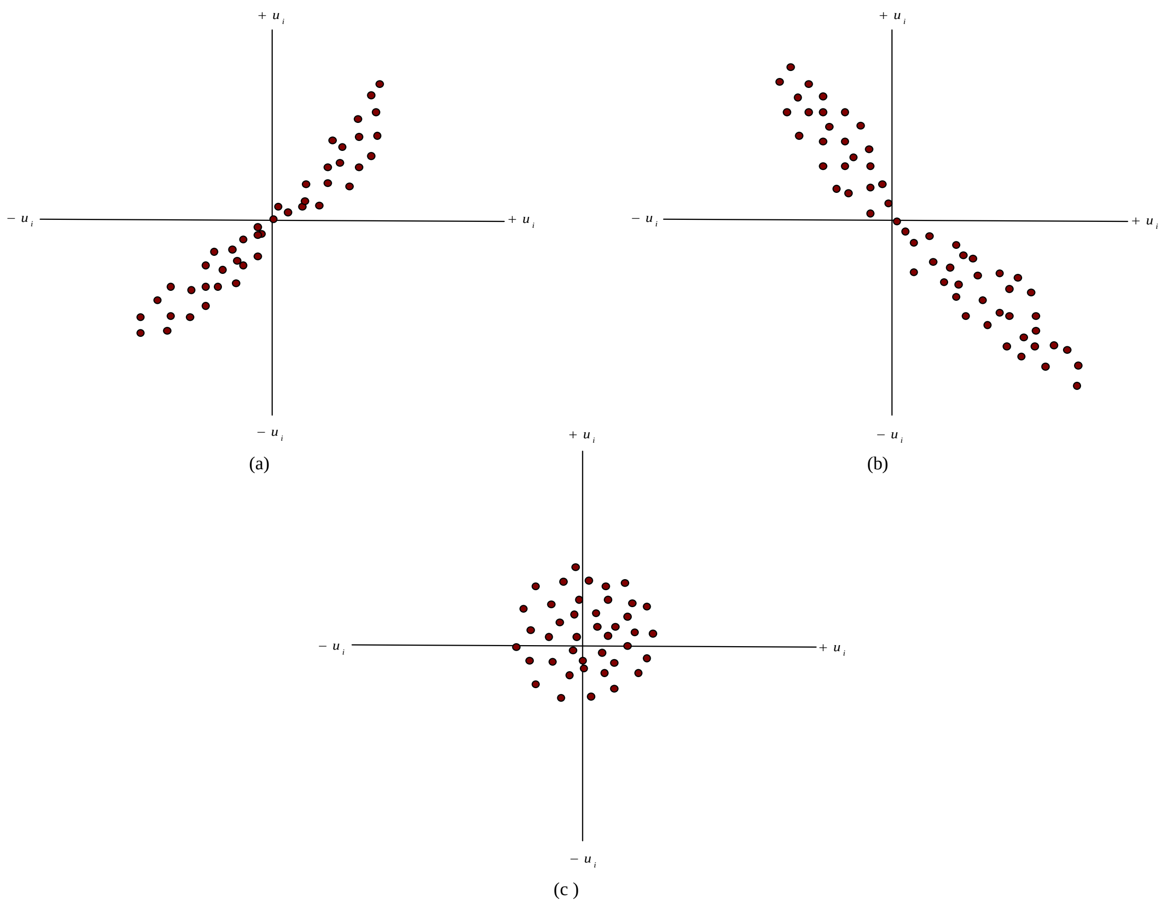
$$\mathbf{= E(u_i/X_i)(u_j/X_j) = 0}$$

→Về mặt ngôn ngữ, giả thiết 4 giả định rằng các nhiễu u_i và u_j không tương quan với nhau.

→Về mặt kĩ thuật, đây là giả thiết về sự không tồn tại của tự tương quan. Tức là với các giá trị X_i cho trước, sự sai lệch về giá trị của bất cứ hai giá trị Y nào so với giá trị trung bình của nó không dẫn tới các kết quả tự tương quan.

3.3. Các giả thiết cơ bản của phương pháp OLS

Hình 3.06. Tự tương quan giữa các nhiễu



3.3. Các giả thiết cơ bản của phương pháp OLS

- Ở hình 3.06 (a) ta thấy rằng các nhiễu tương quan dương : một giá trị dương của nhiễu u kết hợp với một giá trị dương của nhiễu u hoặc một giá trị âm của nhiễu u kết hợp với một giá trị âm của nhiễu u.
- Còn ở hình 3.06 (b), các nhiễu tương quan âm : một giá trị dương của nhiễu u kết hợp với một giá trị âm của nhiễu u và ngược lại.
- Nếu các nhiễu phân tán theo kiểu hình 3.06 (a) và 3.06 (b) thì ta nói rằng đó là hiện tượng là tự tượng quan. Còn nếu các nhiễu không tuân theo bất cứ nguyên tắc nào như mô phỏng trong hình 3.06 (c) thì ta nói rằng hệ số tự tương quan của nó bằng 0, tức là không tồn tại tự tương quan. Đây chính là yêu cầu đặt ra của giả thiết 4.

3.3. Các giả thiết cơ bản của phương pháp OLS

- **Giả thiết 5** : u_i và X_i không tương quan với nhau, hay $E(u_i/X_i)=0$. Một cách đầy đủ ta có :

$$\begin{aligned}
\operatorname{cov}(u_i, X_i) &= E[u_i - E(u_i)][X_i - E(X_i)] \\
&= E[u_i(X_i - E(X_i)] && \text{do } E(u_i)=0 \\
&= E(u_iX_i) - E(X_i)E(u_i) && \text{do } E(X_i) \text{ không phải là đại lượng ngẫu nhiên} \\
&= E(u_iX_i) && \text{do } E(u_i)=0 \\
&= 0
\end{aligned}$$

→ Khi xây dựng hàm hồi quy tổng thể PRF : $Y_i = \beta_1 + \beta_2 X_i + u_i$, ta giả định rằng biến X và u có tác động độc lập lên Y. Nhưng nếu X và u tương quan với nhau, thì ta không thể tách rời ảnh hưởng của chúng lên Y.

3.3. Các giả thiết cơ bản của phương pháp OLS

- **Giả thiết 6** : Đại lượng sai số ngẫu nhiên có phân phối chuẩn :

$$\mathbf{u_i \sim N\,(0,\ \sigma^2)}$$

với :

- Giá trị trung bình : $E\,(u_i) = 0$
- Phương sai : $E[u_i - E(u_i)]^2 = E(u_i^{\,2}) = \sigma^2$
- Tương quan $(u_i,\ u_j)$: $E\{[u_i - E(u_i)][u_j - E(u_j)]\} = E(u_i u_j) = 0$ với $i \neq j$

3.3. Các giả thiết cơ bản của phương pháp OLS

Lưu ý :

- Ta mặc nhiên thừa nhận cỡ mẫu n lớn hơn số tham số trong mô hình, nên điều này không được trình bày như là một giả thiết của mô hình.
- Nếu thỏa mãn giả thiết 1, tức là X là đại lượng đã xác định, thì X có các giá trị không đồng nhất.
- Giả thiết về quy luật chuẩn của nhiễu sẽ được ứng dụng trong phần ước lượng khoảng, kiểm định giả thiết và dự báo khoảng.

3.4. Độ chính xác của các ước lượng OLS

- Theo phương pháp OLS, các ước lượng $\hat{\beta}_1$ và $\hat{\beta}_2$ được xác định theo công thức:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

- Các ước lượng này là hàm của mẫu, là đại lượng ngẫu nhiên, với các mẫu khác nhau ta có các ước lượng khác nhau.
- Vì phương sai hay độ lệch chuẩn đặc trưng cho độ phân tán của đại lượng ngẫu nhiên, nên ta dùng chúng làm thước đo cho chất lượng của ước lượng.

3.4. Độ chính xác của các ước lượng OLS

- Với các giả thiết của phương pháp OLS, phương sai và độ lệch chuẩn của các ước lượng được cho bởi các công thức sau :

- $$\operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$$ **[3.08]**

$$\operatorname{var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}\sigma$$ **[3.10]**

- $$se(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} x_i^2}}$$ **[3.09]**

$$se(\hat{\beta}_1) = \sigma\sqrt{\frac{\sum_{i=1}^{n} X_i^2}{n\sum_{i=1}^{n} x_i^2}}$$ **[3.11]**

Với $\sigma^2 = \operatorname{var}(u_i)$= phương sai thuần nhất của u_i; $x_i = X_i - \bar{x}$

3.4. Độ chính xác của các ước lượng OLS

- Trong các công thức trên, σ^2 chưa biết. σ^2 được ước lượng bằng công thức sau đây:

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

  - $\hat{\sigma}^2$= ước lượng OLS 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)
- Từ công thức [3.12] → công thức tính sai số tiêu chuẩn của đường hồi quy (the standard error of the regression- se) như sau:
- $$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} \hat{u}_i^2}{n-2}} \quad [3.13]$$

$\hat{\sigma}$: độ lệch tiêu chuẩn (standard deviation) các giá trị Y quanh SRF

Ví dụ 2.

Dựa vào các số liệu đã cho trong ví dụ 5. Yêu cầu :

- a. Tính $\hat{Y}_i$ và $\hat{u}_i$
- b. Tính phương sai và sai số tiêu chuẩn của các hệ số hồi quy $(\hat{\beta}_1, \hat{\beta}_2)$ và thành phần nhiễu ($\hat{u}_i$) trong mô hình.

- ***Gợi ý:*** *Sử dụng Excel để tính toán!*

3.5. Tiêu chuẩn của các ước lượng OLS- Định lý Gauss- Markov

- Với các giả thiết từ 1-5 của phương pháp bình phương nhỏ nhất, các ước lượng bình phương nhỏ nhất thu được có tiêu chuẩn tốt nhất.
- Các tiêu chuẩn này được biết đến thông qua định lý nổi tiếng **Gauss- Markov**.
- Để hiểu được định lý này, trước hết chúng ta hãy làm quen với ***« tiêu chuẩn tuyến tính không chệch tốt nhất » (the best linear unbiasedness property)*** của một ước lượng.

3.5. Tiêu chuẩn của các ước lượng OLS- Định lý Gauss- Markov

- Một ước lượng, ví dụ như ước lượng theo phương pháp OLS, được gọi là ước lượng tuyến tính không chệch tốt nhất (Best Linear Unbiased Estimator- BLUE) của β_2 nếu nó thỏa mãn các tiêu chuẩn sau đây :
  - **Tuyến tính**: đó là hàm tuyến tính của một biến ngẫu nhiên, chẳng hạn như biến phụ thuộc Y trong mô hình hồi quy.
  - **Không chệch**: tức là giá trị trung bình của ước lượng hay chính là giá trị kỳ vọng của nó, E ($\hat{\beta}_2$), bằng với giá trị thực β_2.
  - **Có phương sai nhỏ nhất** trong lớp các ước lượng tuyến tính không chệch. Một ước lượng không chệch có phương sai nhỏ nhất được coi là một ước lượng hiệu quả.

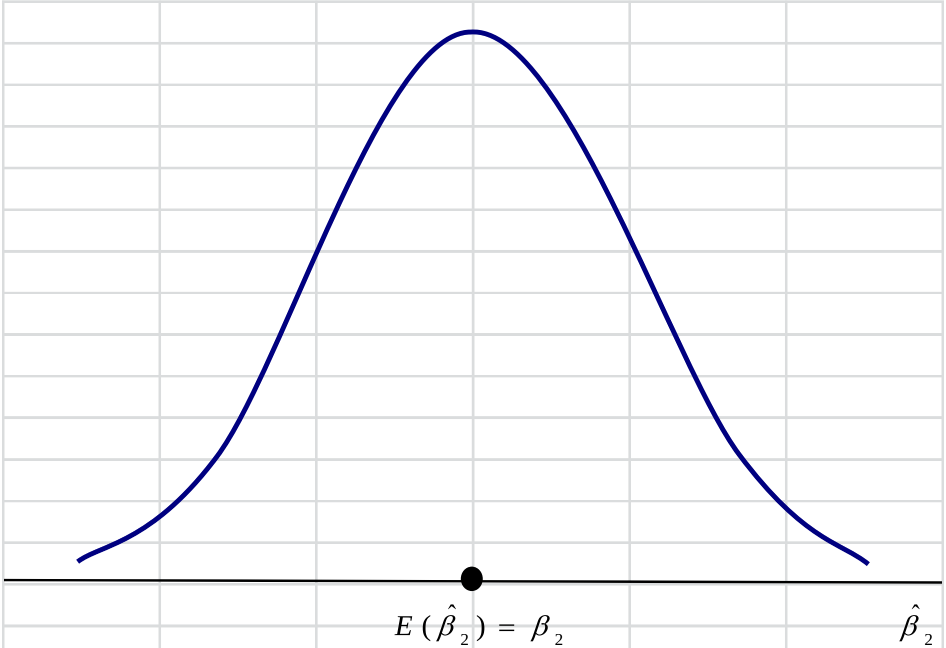
3.5. Tiêu chuẩn của các ước lượng OLS- Định lý Gauss- Markov

- Đối với mô hình hồi quy, thì các ước lượng theo phương pháp OLS được coi là các ước lượng BLUE. Đây chính là nội dung của định lý Gauss- Markov nổi tiếng, được phát biểu như sau:
- ***Định lý Gauss- Markov*: *Với các giả thiết 1-5 của phương pháp bình phương nhỏ nhất, các ước lượng bình phương thu được là các ước lượng tuyến tính, không chệch và có phương sai nhỏ nhất (BLUE) trong lớp các ước lượng tuyến tính không chệch.***
- Định lý Gauss- Markov có thể được giải thích thông qua các đồ thị phân bố xác suất trong hình [3.07].

3.5. Tiêu chuẩn của các ước lượng OLS- Định lý Gauss- Markov

Hình 3.07. Phân phối mẫu của ước lượng $\hat{\beta}_2$ (OLS) và β_2^* (phương pháp khác)

- Hình 3.07 (a) mô tả phân phối mẫu của ước lượng theo phương pháp OLS. Để thuận tiện, ta giả định rằng đồ thị phân bố xác suất của là đối xứng. Đồ thị này cho ta thấy trung bình các giá trị E($\hat{\beta}_2$) bằng với giá trị thực của β_2. Trong trường hợp này, ta nói rằng $\hat{\beta}_2$ là ước lượng không chệch của β_2.

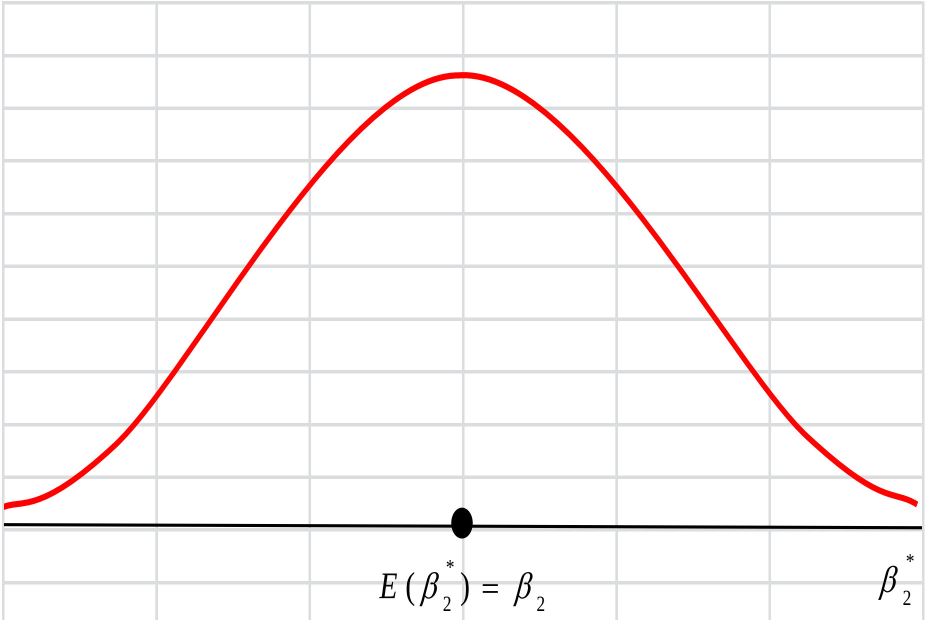


(a) Phân phối mẫu của β2

3.5. Tiêu chuẩn của các ước lượng OLS- Định lý Gauss- Markov

Hình 3.07. Phân phối mẫu của ước lượng $\hat{\beta}_2$ (OLS) và β_2^* (phương pháp khác)

- Hình 3.07 (b) biểu diễn phân phối mẫu của ƯL β_2^*, một giá trị ƯL của β_2 thu được bằng một phương pháp khác OLS. Giả định β_2^*, giống $\hat{\beta}_2$, là ƯL không chệch, nghĩa là giá trị TB của nó bằng giá trị của β_2. Ngoài ra, cũng giả định β_2^* và $\hat{\beta}_2$ đều là các ƯL tuyến tính, tức là chúng đều là hàm tuyến tính của biến phụ thuộc Y →giữa hai ƯL β_2^* và $\hat{\beta}_2$, ta chọn ước lượng nào?

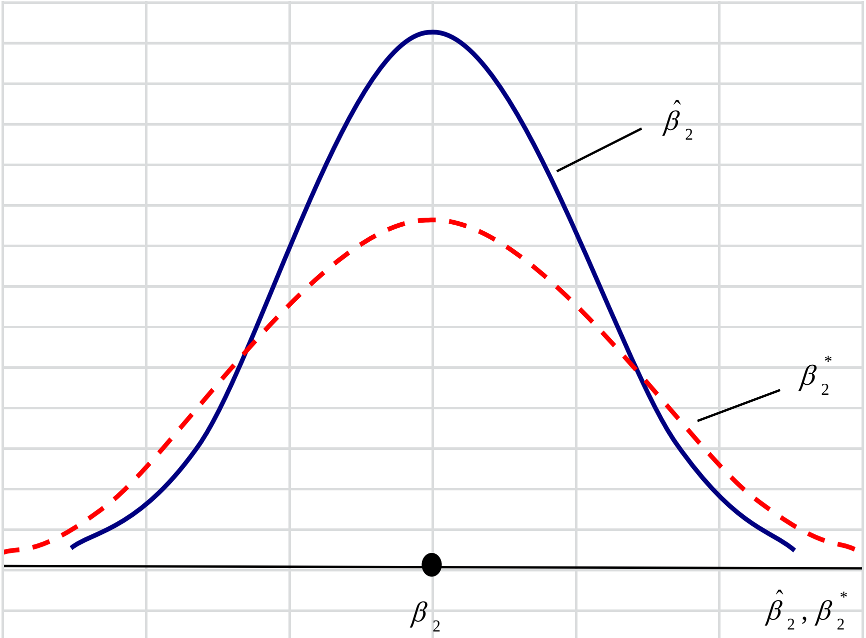


(b) Phân phối mẫu của β2*

3.5. Tiêu chuẩn của các ước lượng OLS- Định lý Gauss- Markov

Hình 3.07. Phân phối mẫu của ước lượng $\hat{\beta}_2$ (OLS) và β_2^* (phương pháp khác)

- hình 3.07(c): cả β_2^* và $\hat{\beta}_2$ đều là các ước lượng không chệch, tuy nhiên, β_2^* phân tán rộng quanh giá trị TB hơn $\hat{\beta}_2$ .Nói cách khác, phương sai của β_2^* lớn hơn phương sai của $\hat{\beta}_2$. Bây giờ, trong hai ƯL cùng là ƯL tuyến tính, không chệch, đương nhiên ta chọn ƯL nào có phương sai nhỏ hơn bởi đó là ƯL có giá trị gần với giá trị của β_2 hơn →đó chính là ước lượng $\hat{\beta}_2$ vì nó thỏa mãn tiêu chuẩn BLUE.



(c) Phân phối mẫu của β2 và β2*

3.6. Phân bố xác suất của các ước lượng OLS

- Với các giả thiết đưa ra trong phần 3.3, $\hat{\beta}_1$ và $\hat{\beta}_2$ là các ước lượng tuyến tính không chệch có phương sai nhỏ nhất của β_1 và β_2.
- Tuy vậy, mục đích của phân tích hồi quy không phải chỉ là suy đoán về β_1 và β_2 hay PRF mà còn phải kiểm tra bản chất của sự phụ thuộc, còn phải thực hiện các dự đoán khác.
- Do đó, cần phải biết phân bố xác suất của $\hat{\beta}_1$ và $\hat{\beta}_2$. Các phân bố này phụ thuộc vào phân bố của các u_i.
- Như ta đã biết, phân bố của u_i là phân bố chuẩn. Từ đó, ta có thể suy ra phân bố xác suất của $\hat{\beta}_1$ và $\hat{\beta}_2$ cũng như tổng kết các tính chất đặc trưng của chúng như sau:

3.6. Phân bố xác suất của các ước lượng OLS

- **1)** Chúng là các ước lượng không chệch.
- **2)** Có phương sai cực tiểu.
- **3)** Khi số quan sát đủ lớn thì các ước lượng này xấp xỉ với giá trị thực của phân bố.
- **4)** $\hat{\beta}_1$ tuân theo quy luật phân phối chuẩn với:
- a. Giá trị trung bình: $E(\hat{\beta}_1) = \beta_1$
- b. Phương sai: $\operatorname{var}(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \dfrac{\sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2} \sigma^2$

$$\rightarrow \quad \hat{\beta}_1 \sim N(\beta_1, \sigma^2_{\hat{\beta}_1})$$

3.6. Phân bố xác suất của các ước lượng OLS

→Từ tính chất 4, ta suy ra biến Z, với $Z = \dfrac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}}$, cũng tuân theo quy luật phân phối chuẩn, tức là: Z~ N (0, 1)

3.6. Phân bố xác suất của các ước lượng OLS

- **5)** $\hat{\beta}_2$ tuân theo quy luật phân phối chuẩn với:
- a. Giá trị trung bình: $E(\hat{\beta}_2) = \beta_2$
- b. Phương sai: $\mathrm{var}(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_1} = \dfrac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$

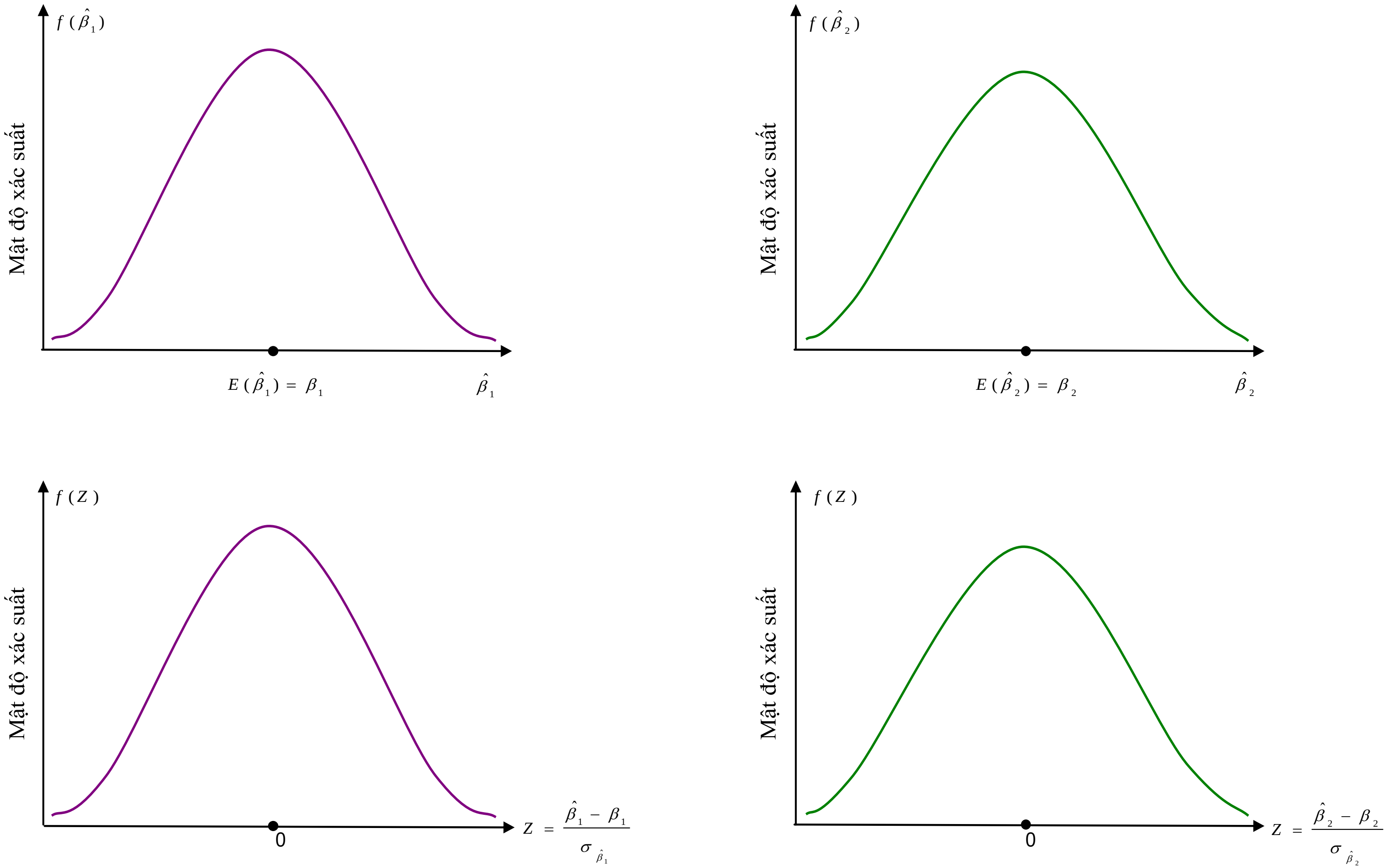
→ $\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2})$

→ $Z = \dfrac{\hat{\beta}_2 - \beta_2}{\sigma_{\hat{\beta}_2}} \sim N(0, 1)$

- Về mặt hình học, phân bố xác suất của $\hat{\beta}_1$ và $\hat{\beta}_2$ được minh họa như trong hình 3.08.

3.6. Phân bố xác suất của các ước lượng OLS

Hình 3.08. Phân bố xác suất của $\hat{\beta}_1$ và $\hat{\beta}_2$



3.6. Phân bố xác suất của các ước lượng OLS

- **6)** $\frac{(n-2)\hat{\sigma}^2}{\sigma^2}$ tuân theo quy luật phân phối Khi-đơ (χ^2) với (n-2)

bậc tự do:

$$\frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2$$

→ Biết được điều này sẽ giúp ta suy ra được giá trị thực của σ^2 từ việc ước lượng σ^2.

3.6. Phân bố xác suất của các ước lượng OLS

- **7)** Trong các ước lượng không chệch của β_1 và β_2 bất kể là tuyến tính hay phi tuyến thì $\hat{\beta}_1$ và $\hat{\beta}_2$ có phương sai nhỏ nhất.

- **8)** Từ giả thiết $u_i \sim N(0, \sigma^2)$, ta suy ra Y_i, vốn là hàm tuyến tính của u_i, cũng tuân theo quy luật phân phối chuẩn với:
- a. $E(Y_i) = \beta_1 + \beta_2 X_i$
- b. $Var(Y_i) = \sigma^2$

→ $Y_i \sim N(\beta_1 + \beta_2 X_i, \sigma^2)$