CHƯƠNG 1:
MÔ HÌNH HỒI QUY HAI BIẾN

ThS. Trương Bích Phương
ĐH Ngoại Thương CS II

1. Hàm hồi quy tổng thể và mẫu

Trong quan hệ hồi quy , *một* biến phụ thuộc có thể được giải thích bởi *nhiều* biến độc lập

Nếu chỉ nghiên cứu ***một*** biến phụ thuộc bị ảnh hưởng bởi ***một*** biến độc lập => ***Mô hình hồi quy hai biến***

Nếu mối quan hệ giữa hai biến này là tuyến tính => ***Mô hình hồi quy tuyến tính hai biến***

1. Hàm hồi quy tổng thể và mẫu

- Tuyến tính trong hồi quy
  - Mô hình hồi quy tuyến tính đối với tham số

    $Y = a + bx$

    $Y = a + bx^2$

    $Y = a + b(1/x)$
  - Mô hình hồi quy tuyến tính đối với biến số

    $Y = a + bx$

    $Y = a^2 + bx$

    $Y = a + b^3x$
  - Tuyến tính của các mô hình hồi quy là tuyến tính dựa vào tham số, không dựa vào biến số

1.1. Hàm hồi quy tổng thể (PRF)

- Mối liên hệ giữa biến phụ thuộc và biến giải thích dựa trên số liệu đã biết của toàn bộ tổng thể.

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

Trong đó

Y : Biến phụ thuộc, Y_i: Giá trị cụ thể của biến phụ thuộc

X : Biến độc lập, X_i : Giá trị cụ thể của biến độc lập

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

1.1. Hàm hồi quy tổng thể (PRF)

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

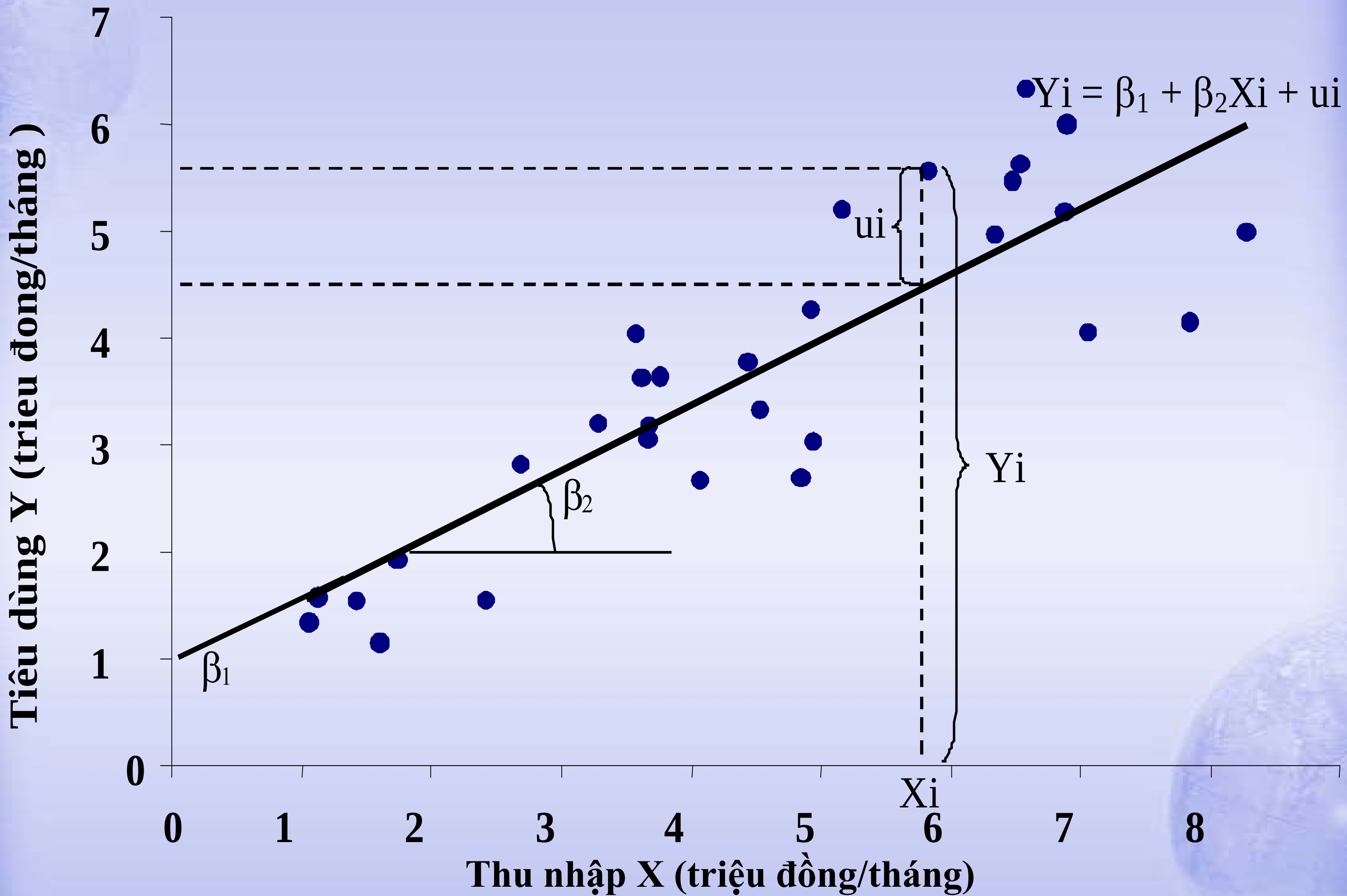
Trong đó

β_1,β_2 là các tham số của mô hình với ý nghĩa :

β_1 : Tung độ của hàm hồi quy tổng thể (intercept coefficient), là giá trị trung bình của biến phụ thuộc Y khi biến độc lập nhận giá trị bằng 0

β_2 : Hệ số gốc của hàm hồi quy (slope coefficient), phản ánh độ dốc của hàm hồi quy tổng thể , là lượng thay đổi trung bình của Y khi X thay đổi 1 đơn vị

1.1. Hàm hồi quy tổng thể



1.2. Hàm hồi quy mẫu

Sample Regression Function (SRF)

- Trong thực tế rất khó nghiên cứu trên tổng thể nên thông thường người ta nghiên cứu xây dựng hàm hồi quy trên một mẫu => Gọi là hàm hồi quy mẫu

- Biểu diễn mối quan hệ giữa biến phụ thuộc với biến độc lập dựa trên giá trị trung bình của tổng thể hay giá trị đã biết của mẫu

1.2. Hàm hồi quy mẫu

Sample Regression Function (SRF)

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

Trong đó

$\hat{\beta}_1$ Tung độ gốc của hàm hồi quy mẫu, là ước lượng điểm của β_1

$\hat{\beta}_2$ Độ dốc của hàm hồi quy mẫu, là ước lượng điểm của β_2

e_i Sai số ngẫu nhiên, là ước lượng điểm của U_i

1.2. Hàm hồi quy mẫu

Sample Regression Function (SRF)

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

Nếu bỏ qua sai số ngẫu nhiên e_i , thì giá trị thực tế Y_i sẽ trở thành giá trị ước lượng $\hat{Y}_i$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

2. Phương pháp bình phương bé nhất

(OLS-Ordinary Least Square)

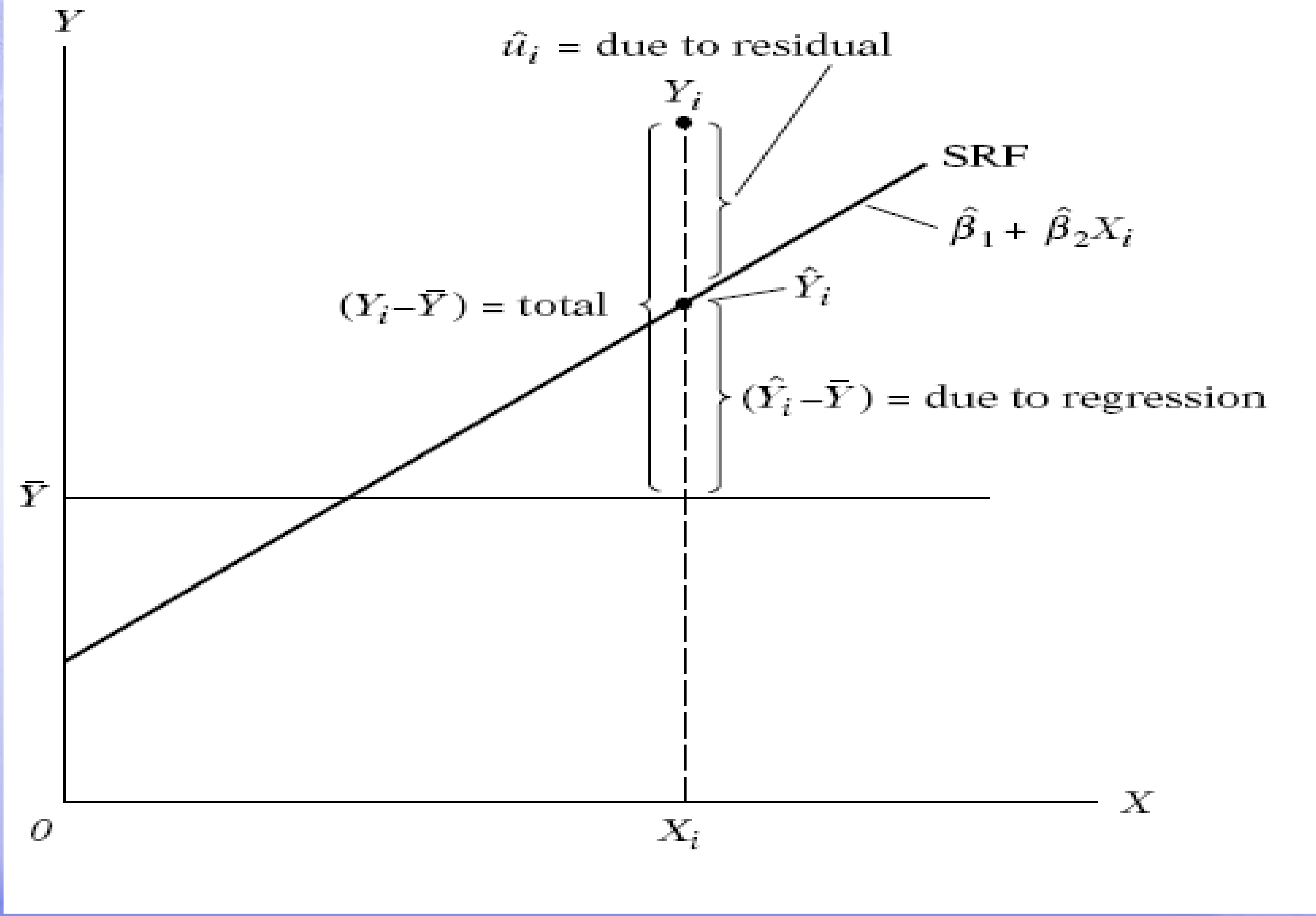
2.1. Ước lượng các tham số của mô hình

Sai số $e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$

Tìm $\hat{\beta}_1, \hat{\beta}_2$ sao cho tổng bình phương sai số là nhỏ nhất

Tức là $\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2 \rightarrow \min$

2.1. Ước lượng các tham số của mô hình



2.1. Ước lượng các tham số của mô hình

Giải bài toán cực trị hàm hai biến, ta được

Với $\bar{X} = \frac{\sum X_i}{n}$ là giá trị trung bình của X

$\bar{Y} = \frac{\sum Y_i}{n}$ là giá trị trung bình của Y

2.1. Ước lượng các tham số của mô hình

Ví dụ: Quan sát về thu nhập (X – triệu đồng/năm) và chi tiêu (Y – triệu đồng/năm) của 10 người, ta được các số liệu sau :

X_i	31	50	47	45	39	50	35	40	45	50
Y_i	29	42	38	30	29	41	23	36	42	48

Xây dựng hàm hồi quy mẫu $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$

2.1. Ước lượng các tham số của mô hình

Kết quả ví dụ :

Hàm hồi quy mẫu

2.2. Các giả thiết của mô hình

Giả thiết 1 : Các giá trị X_i cho trước và không ngẫu nhiên

Giả thiết 2 : Các sai số U_i là đại lượng ngẫu nhiên có giá trị trung bình bằng 0

Giả thiết 3 : Các sai số U_i là đại lượng ngẫu nhiên có phương sai không thay đổi

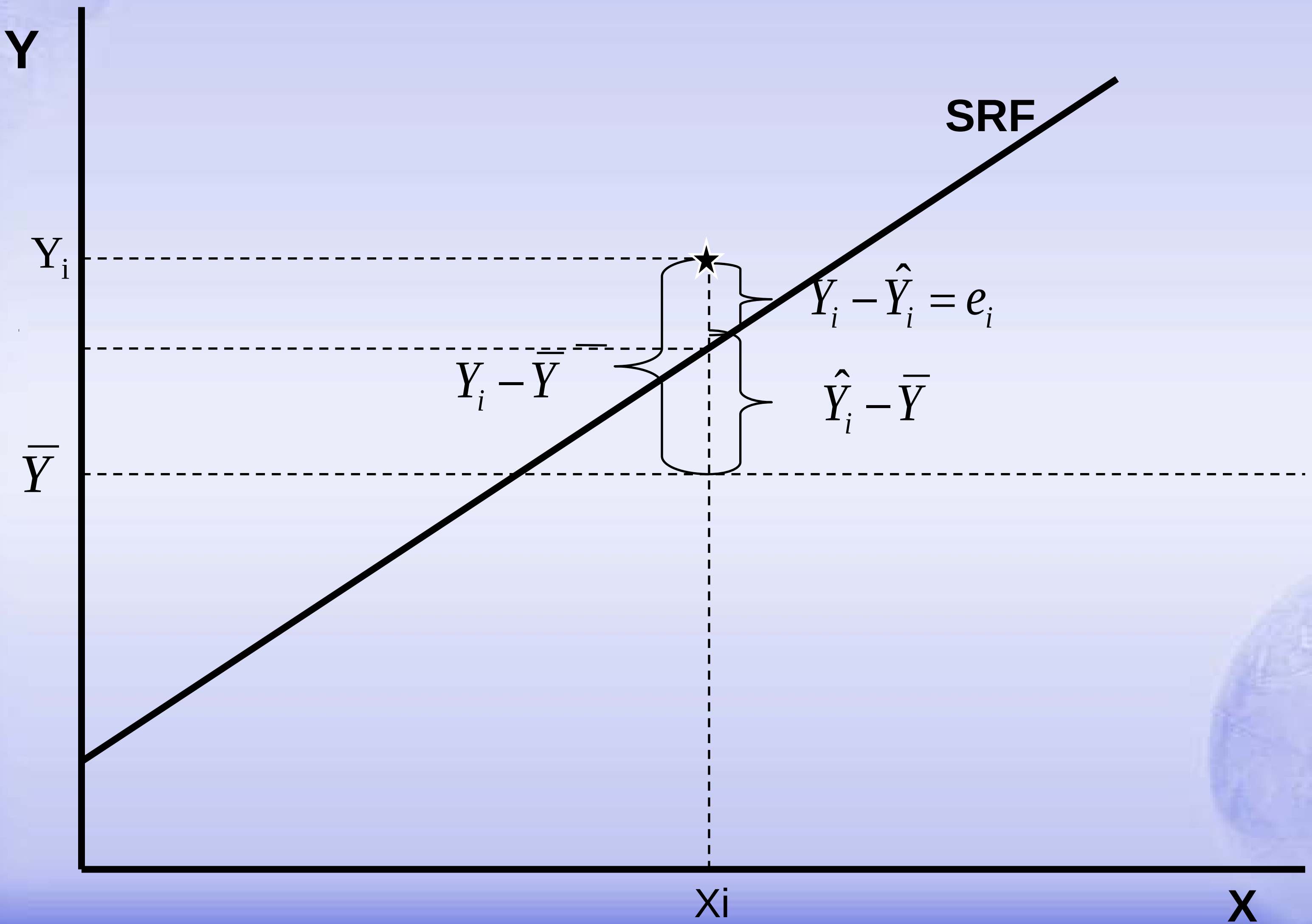
Giả thiết 4 : Không có sự tương quan giữa các U_i

Giả thiết 5 : Không có sự tương quan giữa U_i và X_i

Khi các giả thiết này được đảm bảo thì các ước lượng tính được bằng phương pháp OLS là các ước lượng tốt nhất và hiệu quả nhất của hàm hồi quy tổng thể

Ta nói, ước lượng OLS là ước lượng **BLUE** (*Best Linear Unbias Estimator*)

3. Hệ số xác định của mô hình



3. Hệ số xác định của mô hình

Dùng để đo mức độ phù hợp của hàm hồi qui

Tổng bình phương toàn phần

Total Sum of Squares-TSS

Tổng bình phương tất cả các sai lệch giữa giá trị thực tế của Y với giá trị trung bình của nó.

$$TSS =$$

3. Hệ số xác định của mô hình

Tổng bình phương hồi quy

Explained Sum of Squares-ESS

Tổng bình phương tất cả các sai lệch giữa giá trị của Y được tính theo mô hình với giá trị trung bình của nó.

$$ESS =$$

3. Hệ số xác định của mô hình

Tổng bình phương phần dư

Residual Sum of Squares-RSS

Tổng bình phương tất cả các sai lệch giữa giá trị thực tế với giá trị lý thuyết theo mô hình của Y.

$$RSS =$$

3. Hệ số xác định của mô hình

$$TSS = ESS + RSS$$

Hệ số xác định $R^2 =$

- $0 \leq R^2 \leq 1$
- $R^2 = 1$:
- $R^2 = 0$:

Ví dụ: với số liệu đã cho ở ví dụ trước, hãy tính hệ số xác định của mô hình

4. Hệ số tương quan (r)

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

$$\boldsymbol{r} =$$

4. Hệ số tương quan (r)

Tính chất của hệ số tương quan

1. Miền giá trị của r: $-1 \leq r \geq 1$

   |r| -> 1: X và Y có quan hệ tuyến tính càng chặt chẽ

2. r có tính chất đối xứng: $r_{XY} = r_{YX}$

3. Nếu X, Y độc lập thì r = 0. Điều ngược lại không đúng

5. Phương sai và sai số chuẩn của $\hat{\beta}_1, \hat{\beta}_2$

a. ***Phương sai***

$$Var(\hat{\beta}_1) = \sigma^2_{\hat{\beta}_1} = \frac{\sum X_i^2}{n(\sum X_i^2 - n\overline{X}^2)}\sigma^2 \approx \frac{\sum X_i^2}{n(\sum X_i^2 - n\overline{X}^2)}\hat{\sigma}^2$$

$$\mathrm{var}(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum X_i^2 - n\overline{X}^2} \approx \frac{\hat{\sigma}^2}{\sum X_i^2 - n\overline{X}^2}$$

5. Phương sai và sai số chuẩn của $\hat{\beta}_1, \hat{\beta}_2$

b. Sai số chuẩn

$$se(\hat{\beta}_1) = \sqrt{\sigma^2_{\hat{\beta}_1}} \qquad se(\hat{\beta}_2) = \sqrt{\sigma^2_{\hat{\beta}_2}}$$

Trong đó: σ^2 = Var(U_i)

Khi đó σ^2 được gọi là phương sai của tổng thể, rất khó tính được nên thường được ước lượng bằng phương sai mẫu

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2} = \frac{RSS}{n-2}$$

6. Khoảng tin cậy của các hệ số hồi quy

a. <u>Khoảng tin cậy của β_2</u>

Khoảng tin cậy của β_2 với độ tin cậy 1-α là

Với $t_{\frac{\alpha}{2}}$ có được khi tra bảng t-Student với bậc tự do (n-2), mức ý nghĩa α/2

6. Khoảng tin cậy của các hệ số hồi quy

b. ***<u>Khoảng tin cậy của β_1</u>***

Khoảng tin cậy của β_1 với độ tin cậy 1-α là

Với $t_{\frac{\alpha}{2}}$ có được khi tra bảng t-Student với bậc tự do (n-2), mức ý nghĩa α/2

Ví dụ áp dụng

- Từ số liệu trước, hãy tính khoảng tin cậy của các hệ số hồi quy và cho biết ý nghĩa của chúng với độ tin cậy 95%

7. Kiểm định giả thiết về hệ số hồi quy

a. Kiểm định giả thiết về β_2

H_o: $\beta_2 = \beta_o$
H_1: $\beta_2 \neq \beta_o$

Với độ tin cậy là *1-α*

Cách 1: Phương pháp khoảng tin cậy

Bước 1: Lập khoảng tin cậy của β_2

Bước 2: Nếu β_0 thuộc khoảng tin cậy thì chấp nhận H_0.
Nếu β_0 không thuộc khoảng tin cậy thì bác bỏ H_0

7. Kiểm định giả thiết về hệ số hồi quy

a. *Kiểm định giả thiết về β_2*

Cách 2: Phương pháp giá trị tới hạn (kiểm định t)

Bước 1 : tính giá trị tới hạn $t = \dfrac{\hat{\beta}_2 - \beta_0}{se(\hat{\beta}_2)}$

Bước 2 : tra bảng t-Student với bậc tự do (n-2) tìm $t_{\alpha/2}$

Bước 3 :

Nếu $-t_{\alpha/2} \le t \le t_{\alpha/2}$: chấp nhận giả thiết H_0

Nếu $t < -t_{\alpha/2}$ hoặc $t > t_{\alpha/2}$: bác bỏ giả thiết H_0

7. Kiểm định giả thiết về hệ số hồi quy

a. ***Kiểm định giả thiết về β_2***

Cách 3: Phương pháp p-value

7. Kiểm định giả thiết về hệ số hồi quy

b. *Kiểm định giả thiết về β_1*

$H_o: \beta_1 = \beta_o$
$H_1: \beta_1 \neq \beta_o$

Với độ tin cậy là *1-α*

Tương tự kiểm định giả thiết về β_2 nhưng giá trị tới hạn lúc này là

$$t = \frac{\hat{\beta}_1 - \beta_0}{se(\hat{\beta}_1)}$$

Ví dụ áp dụng

Từ số liệu đã cho của ví dụ trước , yêu cầu kiểm định các giả thiết sau

a) $H_o: \beta_2 = 0$
$H_1: \beta_2 \neq 0$

Với độ tin cậy là 99%

b) $H_o: \beta_1 = 0$
$H_1: \beta_1 \neq 0$

Với độ tin cậy là 99%

Kết quả ví dụ áp dụng

a. *<u>Kiểm định giả thiết β_2</u>*

Cách 1: Khoảng tin cậy của β_1 với độ tin cậy 1-α là

Kết quả ví dụ áp dụng

a. <u>Kiểm định giả thiết β_2</u>

Cách 2: Khoảng tin cậy của β_1 với độ tin cậy 1-α là

Kết quả ví dụ áp dụng

a. <u>Kiểm định giả thiết β_2</u>

Cách 3: Xác định P Value

Kết quả ví dụ áp dụng

b. *Kiểm định giả thiết β_1*

Cách 1: Khoảng tin cậy của β_1 với độ tin cậy 1-α là

$$\left(\hat{\beta}_1 - t_{\frac{\alpha}{2}} \times se(\hat{\beta}_1); \hat{\beta}_1 + t_{\frac{\alpha}{2}} \times se(\hat{\beta}_1) \right)$$

Kết quả ví dụ áp dụng

b. *Kiểm định giả thiết β_1*

Cách 2: Khoảng tin cậy của β_1 với độ tin cậy 1-α là

Kết quả ví dụ áp dụng

b. *Kiểm định giả thiết β_1*

Cách 3: Xác định P Value

8. Kiểm định sự phù hợp của mô hình

Kiểm định giả thiết

$H_o: R^2 = 0$

$H_1: R^2 \neq 0$

Với độ tin cậy là 1- α

Phương pháp kiểm định F

Bước 1 : tính $F = \dfrac{R^2(n-2)}{(1-R^2)}$

Bước 2 : Tra bảng tìm F(1,n-2), mức ý nghĩa là α

Bước 3 : Nếu F>F(1,n-2) , bác bỏ H_0

Nếu F≤F(1,n-2) , chấp nhận H_0

Ví dụ áp dụng

Từ số liệu đã cho của ví dụ trước , yêu cầu kiểm định sự phù hợp của mô hình

9. Dự báo

Cho trước giá trị X = X_0, hãy dự báo giá trị trung bình và giá trị cá biệt của Y với mức ý nghĩa α hay độ tin cậy 1 - α.

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

*** <u>Dự báo điểm</u>**

$$\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$$

9. Dự báo

• **Dự báo giá trị trung bình:**

Khoảng tin cậy của Y_0 với độ tin cậy (1-α) là

Với $$\sigma^2_{\hat{Y}_0} = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum X_i^2 - n(\bar{X})^2} \right]$$

$$se(\hat{Y}_0) = \sqrt{\sigma^2_{\hat{Y}_0}}$$

$$\boxed{\left(\hat{Y}_0 - t_{\frac{\alpha}{2}} \times se(\hat{Y}_0); \hat{Y}_0 + t_{\frac{\alpha}{2}} \times se(\hat{Y}_0) \right)}$$

Dự báo trung bình cá biệt

Ví dụ áp dụng

Từ số liệu đã cho của ví dụ trước , yêu cầu dự báo khoảng giá trị của Y khi X_0 = 60 (triệu đồng/năm) với độ tin cậy 95%

$$\hat{Y}_i = -5,4517 + 0,9549 X_i$$

* **Dự báo điểm**

10. Đổi đơn vị của biến

Trong hàm hồi quy hai biến, nếu đơn vị tính của X và Y thay đồi => áp dụng công thức đổi đơn vị

- Hàm hồi quy theo đơn vị cũ:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$$

- Hàm hồi quy theo đơn vị mới:

$$\hat{Y}_i^* = \hat{\beta}_1^* + \hat{\beta}_2^* X_i^*$$

Trong đó:

$$Y_i^* = k_1 Y_i$$

$$X_i^* = k_2 X_i$$

Khi đó:

$$\hat{\beta}_1^* = k_1 \hat{\beta}_1$$

$$\hat{\beta}_2^* = k_2 \hat{\beta}_2$$

10. Đổi đơn vị của biến

- Hàm Y (triệu đ/tháng) = 5 + 7X (tấn)
- Tính theo (triệu đ/tháng) và (Kg)

11. Mở rộng mô hình hồi quy 2 biến
11.1. Hồi quy qua gốc tọa độ

- Mô hình:

PRF:
$$Y_i = \beta_2 X_i + U_i$$

SRF:
$$Y_i = \hat{\beta}_2 X_i + e_i$$

Với
$$\boxed{\hat{\beta}_2 = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2}}$$
và
$$\operatorname{var}(\hat{\beta}_2) = \sigma^2_{\hat{\beta}_2} = \frac{\sigma^2}{\sum X_i^2}$$

σ^2 được ước lượng bằng
$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2} = \frac{RSS}{n-2}$$

11. Mở rộng mô hình hồi quy 2 biến
11.1. Hồi quy qua gốc tọa độ

Lưu ý:

- R^2 có thể âm đối với mô hình này, nên không dùng R^2 mà thay bởi $R^2_{thô}$

$$R^2_{thô} = \frac{(\sum X_i Y_i)^2}{\sum X_i^2 \sum Y_i^2}$$

- Không thể so sánh R^2 với $R^2_{thô}$
- Thường người ta dùng mô hình có tung độ gốc, trừ khi có một thực nghiệm cần phải dùng mô hình qua gốc toạ độ

11.2. Mô hình tuyến tính logarit

- *Mô hình log-log, log kép hay tuyến tính log*

❖ Mô hình hồi quy mũ $Y_i = \beta_1 X_i^{\beta_2} e^{u_i}$

Được biểu diễn dưới dạng sau:

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

Mô hình không tuyến tính theo tham số và biến số

$$\beta_1^* = \ln \beta_1 \qquad Y_i^* = \beta_1^* + \beta_2 X_i^* + U_i$$

$$Y^*_i = \ln Y_i$$

$$X_i^* = \ln X_i$$

- *Ý nghĩa của hệ số β_2:* khi X thay đổi 1% thì Y thay đổi *β_2%. (Đây chính là hệ số co giãn của Y đối với X)*

11.3. Mô hình log-lin

- Mô hình bán logarit nghĩa là chỉ có một biến thể hiện dưới dạng logarit:

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

- ***Ý nghĩa của hệ số β_2:*** khi X thay đổi 1 đơn vị thì Y thay đổi $100.\beta_2\%$.
- ***Ứng dụng:*** Nghiên cứu khảo sát tốc độ tăng trưởng (giảm sút) của các biến số kinh tế vĩ mô như GDP, dân số, lao động, năng suất.

11.4. Mô hình lin-log

- Mô hình bán logarit nghĩa là chỉ có một biến thể hiện dưới dạng logarit:

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

- ***Ý nghĩa của hệ số β_2:*** khi X thay đổi 1% thì Y thay đổi β_2 /100 đơn vị.
- ***Ứng dụng:*** Nghiên cứu khảo sát một số quan hệ: lượng cung tiền ảnh hưởng tới GNP, diện tích trồng trọt tác động tới sản lượng cây trồng, diện tích căn nhà tác động tới giá nhà,…

11.4. Mô hình nghịch đảo

$$Y_i = \beta_1 + \beta_2 \frac{1}{X} + u_i$$

- *Ứng dụng:*
  - Quan hệ giữa chi phí sản xuất cố định trung bình và sản lượng
  - Quan hệ giữa tỷ lệ thay đổi tiền lượng và tỷ lệ thất nghiệp (đường công Philips)
  - Đường chi tiêu Engel biểu diễn mối quan hệ giữa chi tiêu của người tiêu dùng cho một loại hàng hóa với tổng chi tiêu hay thu nhập của người đó

Bài tập

Bài 1: quan sát sự biến động của nhu cầu gạo Y (t/tháng) vào đơn giá X (1000đ/kg). Hãy lập mô hình hôi quy mẫu biễu diễn mối phụ thuộc về nhu cầu vào đơn giá gạo

Stt	Xi	Yi		
1	1	10		
2	4	6		
3	2	9		
4	5	5		
5	5	4		
6	7	2		
sum	24	36		

Bài tập

Bài 2: Với số liệu và kết quả ở Bài 1

a. Tìm khoảng tin cậy của $\beta 1$, $\beta 2$ với $\alpha=0,05$

b. Hãy xét xem nhu cầu của loại hàng trên có phụ thuộc vào đơn giá của nó không với $\alpha=0,05$.

c. Hãy dự báo nhu cầu trung bình và nhu cầu cá biệt của loại hàng trên khi đơn giá ở mức 6.000 đồng/kg với độ tin cậy 95%.

Bài tập

Bài 3: Với số liệu thống kê về 2 biến X và Y như sau

Yi	4	6	7	9	1O	11	13	15	16	19
Xi	5O	42	41	40	36	36	32	31	27	25

a. Tìm hàm hồi quy tuyến tính mẫu của Y theo X
b. Tính độ lệch chuẩn của hệ số gốc
c. Tính hệ số xác định
d. Kiểm định sự phù hợp của mô hình (β_1, β_2, δ^2) với mức ý nghĩa 1%
e. Xác định ước lượng khoảng của E(X/Y=12) với độ tin cậy 95%

Bài tập

Bài 4: Với số liệu thống kê về 2 biến X và Y như sau

Yi	23	19.5	24	21	25	22	26.5	23.1	25	28	29.5	26
Xi	3	2	4	2	5	4	7	6	9	8	10	8

Với Y là thu nhập (tr đ/năm), X là thâm niên (năm)

a. Tìm hàm hồi quy tuyến tính mẫu của Y theo X và phát biểu ý nghĩa các hệ số hồi quy

b. Tính hệ số tương quan tuyến tính r và đánh giá mức độ phụ thuộc tương quan tuyến tính

c. Dự báo thu nhập trung bình một người có thâm niên là 6 năm với độ tin cậy 95%