CHƯƠNG II. MÔ HÌNH HỒI QUY HAI BIẾN (P. I)

- 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

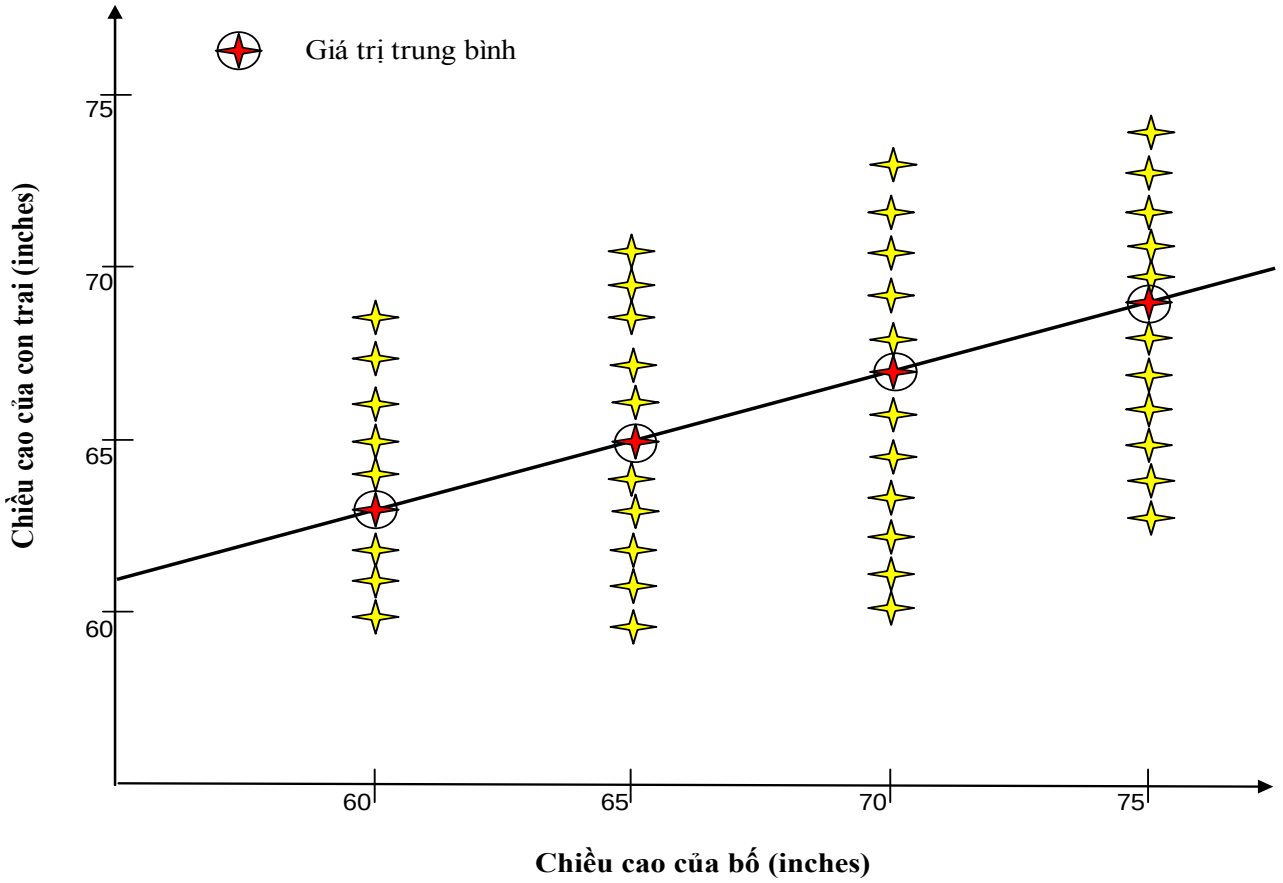
1.2. Sự khác nhau giữa các dạng quan hệ

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. Sự khác nhau giữa các dạng quan hệ

1.2.1. Quan hệ thống kê và quan hệ hàm số

1.2.2. Hồi quy và quan hệ nhân quả

1.2.3. Hồi quy và tương quan

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?)

1.2.2. 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***.

→***Nếu như quan hệ nhân quả tồn tại thì nó phải được xác lập dựa trên các lý thuyết kinh tế khác***.

- **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.***

1.2.2. 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.

→ Như vậy phân tích mối quan hệ nhân quả dựa vào kiến thức và phương pháp luận của môn khác chứ không từ phân tích hồi quy.

1.2.3. 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.

  - **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.***
- : 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)

ng 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)

ng 2.02 **p 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 :

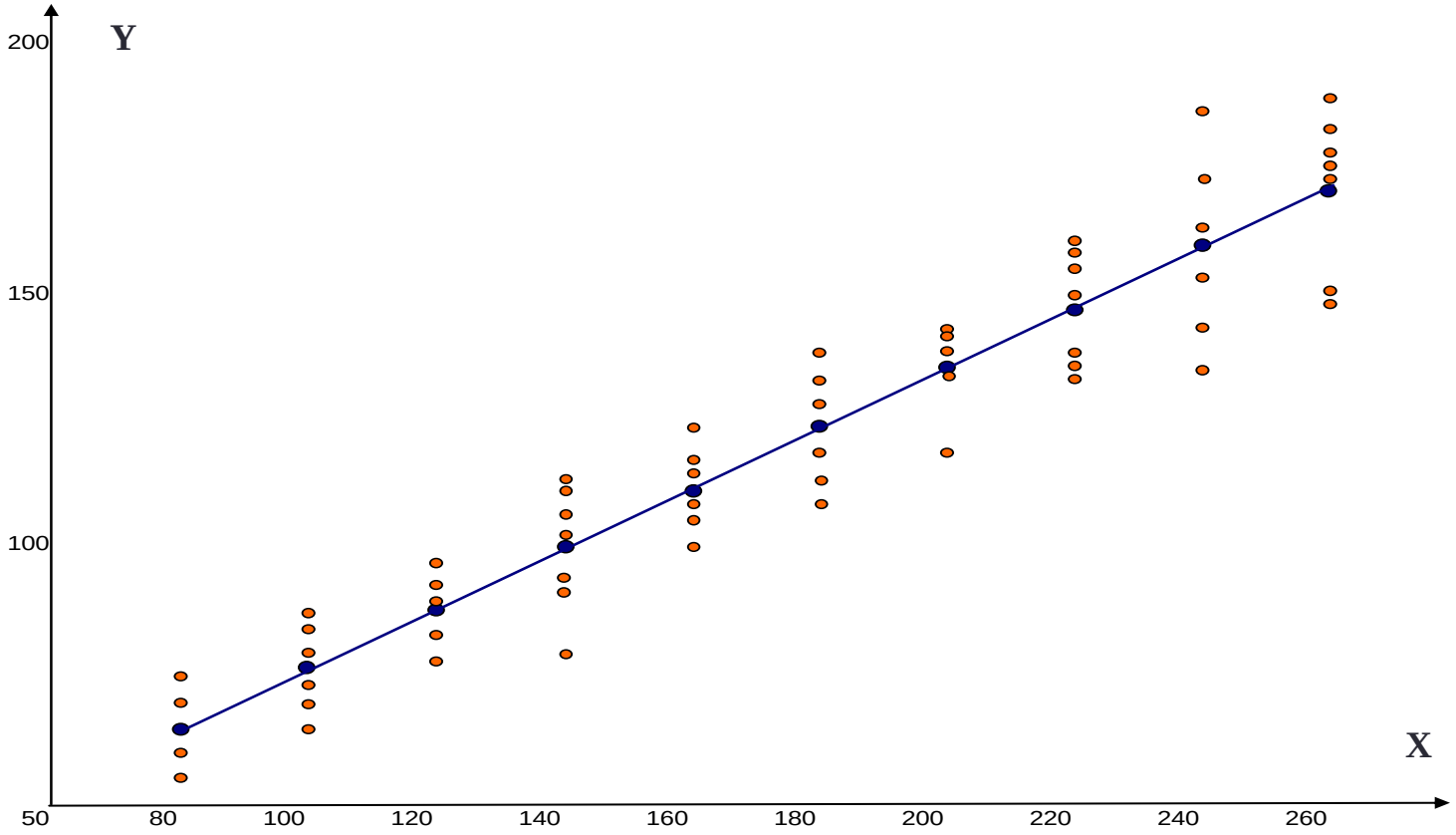
$$E(Y/100) = \sum_j Y_j P(Y = Y_j / X = 100)$$

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

→ 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:

- Nếu xét riêng từng hộ GĐ thì mức độ biến động của chi tiêu lớn và không thấy rõ xu hướng thay đổi của chi tiêu theo thu nhập.
- Nếu xét theo nhóm hộ gia đình có cùng thu nhập và quan tâm đến chi tiêu trung bình ($E(X/Y_i)$ thì mức độ biến động của chi tiêu trung bình ít và có xu hướng tăng theo thu nhập.

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

→Vậy có thể xem $E(X/Y_i)$ là một hàm nào đó của biến giải thích X_i n như sau:

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

- 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 :

$$\mathbf{E(Y/X_i)= \beta_1+ \beta_2 X_i \quad [2]}$$

  - β_1, β_2= hệ số hồi quy
  - β_1= hệ số chặn
  - β_2= 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: $E(Y/X_i)= \beta_1+ \beta_2 X_i^2$ là hàm tuyến tính đối với tham số nhưng phi tuyến đối với biến.
  - **Tuyến tính đối với biến**. Ví dụ: $E(Y/X_i)= \beta_1+ \sqrt{\beta_2} X_i$ là hàm tuyến tính đối với biến nhưng phi tuyến với tham số.

→ ***Hàm hồi quy tuyến tính luôn luôn được hiểu là hồi quy tuyến tính đối với các tham số, nó có thể hoặc không phải là tuyến tính đối với biến.***

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 :

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

- Hay :

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

→ 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

- Nếu $E(Y/X_i)$ là tuyến tính đối với X_i thì phương trình [4] có thể được trình bày dưới dạng như sau :

$$\mathbf{Y_i = \beta_1 + \beta_2 X_i + u_i \quad [5]}$$

- Từ phương trình [4] ta có :

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

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

$$\mathbf{\leftrightarrow E(Y_i/X_i) = E(Y_i/X_i) + E(u_i/X_i) \quad [5]}$$

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

→ Như vậy, ngoài các biến giải thích trong mô hình, giá trị trung bình của tất cả các yếu tố tác động đến biến phụ thuộc Y (đại diện bởi U_i) bằng 0.

2.2. Sai số ngẫu nhiên và bản chất của nó

Ví dụ với X = 100 $ (bảng 2.01), hãy tính $E(u_i/100)$.

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?)

2.3. 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

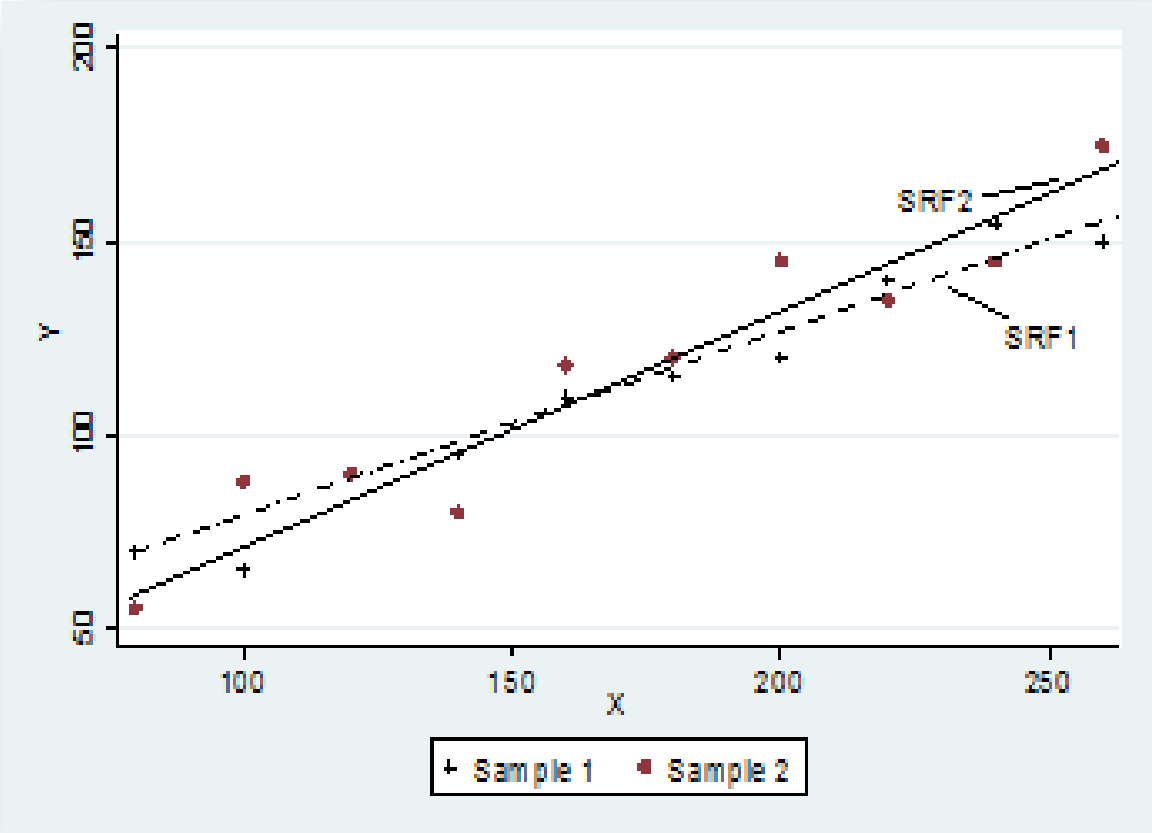
X	80	100	120	140	160	180	200	220	240	260
Y	55	88	90	80	118	120	145	135	145	175

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)

- Hình 2.03 trình bày biểu đồ phân tán và hai đường hồi quy tương ứng với hai mẫu trên. Vậy 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.
- 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 và 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 :

$$\begin{cases} E(Y/X_i) = \beta_1 + \beta_2 X_i \\ Y_i = E(Y/X_i) + u_i = \beta_1 + \beta_2 X_i + u_i \end{cases}$$

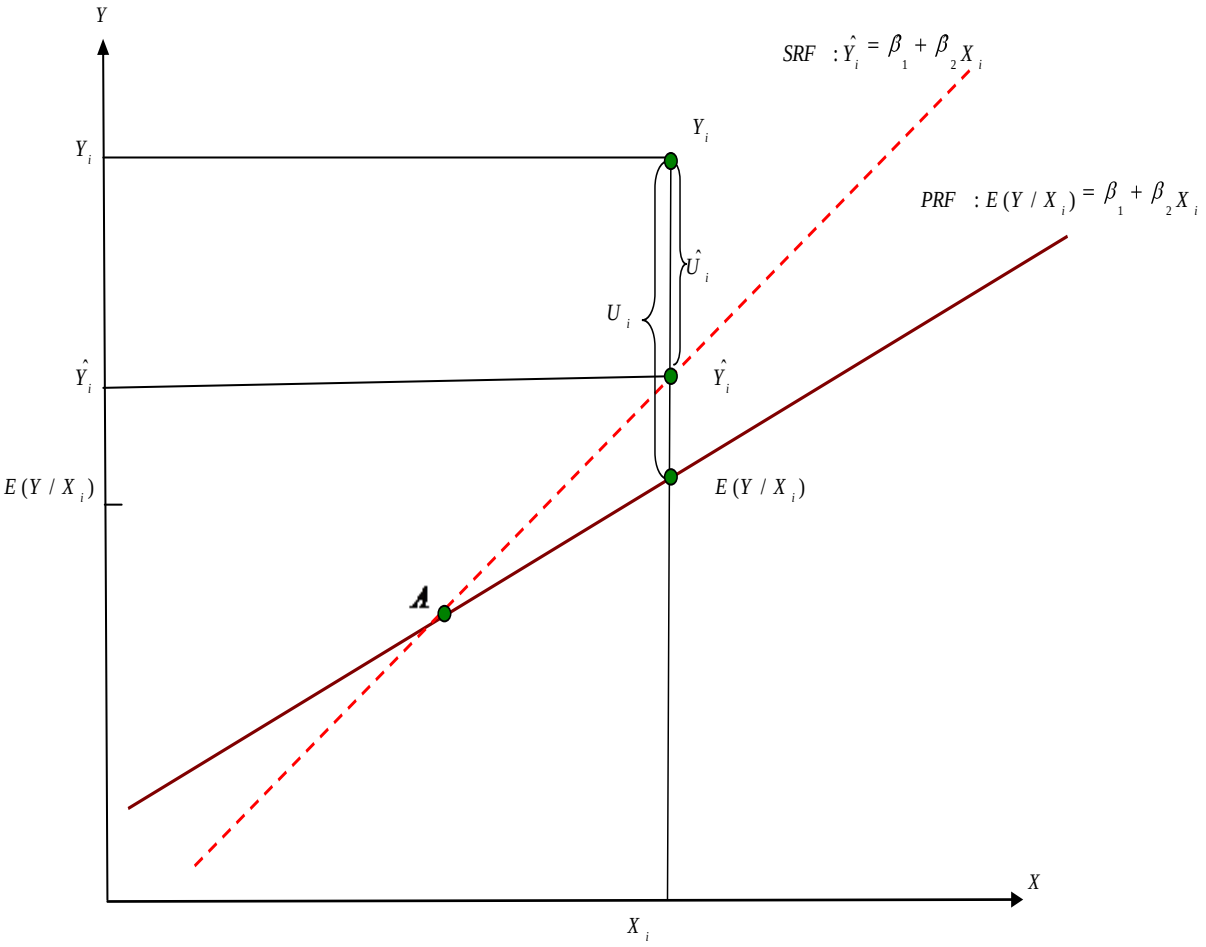
thì **SRF** được trình bày ở dạng tương ứng như sau :

$$\begin{cases} \hat{Y}_i = \beta_1 + \beta_2 X_i \\ Y_i = \hat{Y}_i + \hat{u}_i = \beta_1 + \beta_2 X_i + \hat{u}_i \end{cases}$$

với $\hat{Y}_i$ là ước lượng của E(Y/Xi) ;β_1 ,β_2 là ước lượng của β_1, β_2; $\hat{u}_i$ 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.
- 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 β_1, β_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.