

CHƯƠNG 2

PHÂN TÍCH HỒI QUY HAI BIẾN: MỘT SỐ Ý TƯỞNG CƠ BẢN

Trong chương 1 chúng ta đã thảo luận về khái niệm hồi quy một cách tổng quát. Trong chương này chúng ta sẽ tiếp cận vấn đề một cách tương đối hệ thống hơn. Đặc biệt, chương này và ba chương tiếp theo sẽ giúp bạn đọc làm quen với lý thuyết làm nền tảng cho một phân tích hồi quy đơn giản nhất có thể có được, gọi là hồi quy hai biến. Chúng ta xem xét trường hợp này trước, không nhất thiết bởi vì khả năng thực tế của nó, mà bởi vì nó trình bày cho chúng ta những ý tưởng cơ bản của phân tích hồi quy một cách đơn giản nhất có thể được và một số trong những ý tưởng này có thể được minh họa bằng các biểu đồ hai chiều. Hơn nữa, như chúng ta sẽ thấy, đứng về nhiều phương diện trường hợp phân tích hồi quy bội tổng quát là sự mở rộng hợp lý của trường hợp hồi quy hai biến.

2.1 MỘT VÍ DỤ GIẢ THIẾT

Như đã chỉ ra ở Phần 1.2, phân tích hồi quy chủ yếu là để ước lượng và/hay dự đoán trung bình (tổng thể) hoặc giá trị trung bình của biến độc lập trên cơ sở các giá trị đã biết hoặc đã xác định của (các) biến giải thích. Để hiểu điều này được thực hiện như thế nào, hãy xem xét ví dụ sau.

Giả thiết có một quốc gia với một **tổng thể**¹ là 60 gia đình. Giả sử chúng ta quan tâm đến việc nghiên cứu mối quan hệ giữa Y chỉ tiêu tiêu dùng hàng tuần của gia đình và X thu nhập khả dụng hàng tuần của gia đình hay thu nhập sau khi đã đóng thuế. Nói một cách cụ thể hơn là giả định rằng chúng ta muốn dự đoán mức trung bình (tổng thể) của chi tiêu tiêu dùng hàng tuần khi biết thu nhập hàng tuần của gia đình. Để thực hiện điều này, giả sử chúng ta chia 60 gia đình thành 10 nhóm có thu nhập tương đối như nhau và xem xét chỉ tiêu tiêu dùng của các gia đình trong từng mỗi nhóm thu nhập này. Các dữ liệu giả thiết nằm ở Bảng 2.1. (Với mục đích để thảo luận, giả định rằng chỉ những mức thu nhập đưa ra ở bảng 2.1 là thật sự được quan sát.)

Bảng 2.1 sẽ được giải thích như sau: Ví dụ như, tương ứng với thu nhập hàng tuần là 80 đôla, có năm gia đình có mức chi tiêu tiêu dùng hàng tuần trong khoảng 55 đến 75 đôla. Tương tự, với $X = 240\$$, có sáu gia đình có mức chi tiêu tiêu dùng hàng tuần nằm trong khoảng 137\$ và 189\$. Nói một cách khác, mỗi cột dọc (dãy đứng) của Bảng 2.1 cho thấy sự phân phối của chi tiêu tiêu dùng Y tương ứng với một mức thu nhập X cố định: có nghĩa là, nó cho thấy **phân phối có điều kiện** của Y phụ thuộc vào các giá trị nhất định của X .

Lưu ý rằng các dữ liệu trong Bảng 2.1 tiêu biểu cho tổng thể, chúng ta có thể dễ dàng tính toán các **xác suất có điều kiện** của Y , $p(Y|X)$, xác suất của Y với điều kiện X sẽ như sau.² Ví dụ, với $X = 80\$$, có 5 giá trị của Y : 55\$, 60\$, 65\$, 70\$, và 75\$. Do đó, với $X=80$, xác suất để

¹ Ý nghĩa thống kê của thuật ngữ tổng thể được giải thích ở phần phụ lục A. Nói đơn giản, nó là tập hợp của tất cả các kết quả có thể xảy ra của một thí nghiệm hay một đo đạc, ví dụ: tung một đồng tiền nhiều lần hay ghi chép lại giá cả của tất cả các chứng khoán trên Thị trường Trao đổi Chứng khoán New York vào cuối một ngày kinh doanh.

² Giải thích về ký hiệu: biểu thức $p(Y|X)$ hay $p(Y|X_i)$ là viết tắt cho $p(Y=Y_j|X=X_i)$, có nghĩa là, xác suất để biến ngẫu nhiên (rời rạc) Y có giá trị bằng số là Y_j với điều kiện biến ngẫu nhiên (rời rạc) X có giá trị bằng số là X_i . Tuy nhiên để tránh làm lộn xộn các ký hiệu, chúng tôi sẽ dùng chỉ số ở dưới i (chỉ số của quan sát) cho cả hai biến. Như vậy, $p(Y|X)$ hay $p(Y|X_i)$ sẽ thay thế cho $p(Y=Y_i|X=X_i)$, có nghĩa là, xác suất để Y có giá trị Y_i khi X lấy giá trị X_i , vấn đề gặp phải ở đây là làm sáng tỏ phạm vi giá trị của Y và X . Trong Bảng 2.1, khi $X=220$, Y sẽ nhận 7 giá trị khác nhau, nhưng khi $X = 120$, Y chỉ nhận 5 giá trị.

có được bất kỳ một trong số những chỉ tiêu tiêu dùng này là 1/5. Biểu thị bằng các ký hiệu toán học là $p(Y=55 | X=80) = 1/5$. Tương tự, $p(Y=150 | X=260) = 1/7$, v.v. Xác suất có điều kiện của các dữ liệu trong Bảng 2.1 được trình bày trong Bảng 2.2.

Bây giờ đối với mỗi phân phối xác suất có điều kiện của của Y chúng ta có thể tính được số trung bình hoặc giá trị trung bình của nó, được gọi là **trung bình có điều kiện** hay **kỳ vọng có điều kiện**, được thể hiện bằng $E(Y | X = X_i)$ và được diễn giải là "giá trị kỳ vọng của Y khi X nhận một giá trị cụ thể X_i ," để đơn giản hóa về mặt ký hiệu chúng ta viết lại thành như sau: $E(Y | X_i)$. (Lưu ý: một giá trị kỳ vọng chỉ đơn thuần là trung bình tổng thể hay giá trị trung bình). Đối với các dữ liệu giả thiết của chúng ta, những kỳ vọng có điều kiện này có thể được tính toán một cách dễ dàng bằng cách nhân các giá trị Y tương ứng trong Bảng 2.1 với các xác suất có điều kiện của chúng trong Bảng 2.2 và cộng các kết quả này lại. Để minh họa, trung bình có điều kiện tức kỳ vọng có điều kiện của Y với $X = 80$ là $55(1/5) + 60(1/5) + 65(1/5) + 70(1/5) + 75(1/5) = 65$. Như vậy kết quả các trung bình có điều kiện được đặt trong hàng cuối cùng của Bảng 2.2.

BẢNG 2.1

Thu nhập gia đình hàng tuần X , \$

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

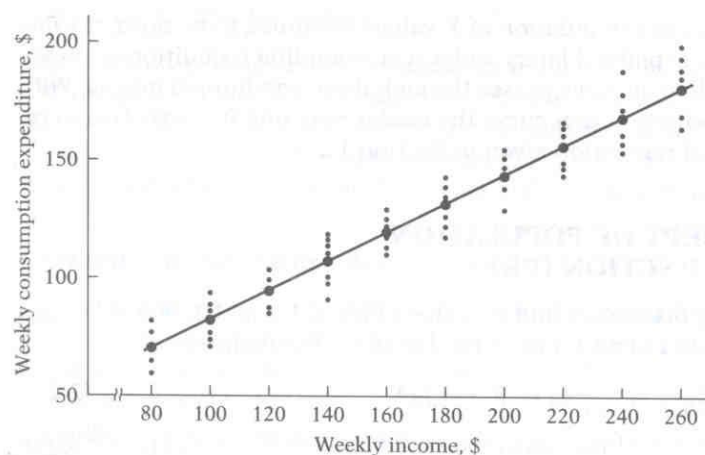
Trước khi tiếp tục, việc xem xét các dữ liệu của Bảng 2.1 trên một đồ thị phân tán sẽ giúp cho ta nhiều điều bổ ích, như trong hình 2.1. Đồ thị phân tán cho thấy phân phối có điều kiện của Y ứng với các giá trị khác nhau của X . Mặc dù có sự biến đổi trong chỉ tiêu tiêu dùng của từng gia đình, Hình 2.1 cho thấy một cách rất rõ ràng là chỉ tiêu tiêu dùng về mặt *trung bình* sẽ tăng khi thu nhập tăng. Nói một cách khác, đồ thị phân tán cho thấy rằng các giá trị trung bình (có điều kiện) của Y tăng khi X tăng. Có thể nhận thấy quan sát này một cách sinh động hơn nếu chúng ta tập trung vào các điểm có kích thước lớn thể hiện các trung bình có điều kiện khác nhau của Y . Đồ thị phân tán cho thấy rằng các trung bình có điều kiện này nằm trên một hàng thẳng với một độ dốc đồng biến.³ Đường thẳng này được gọi là **đường hồi qui tổng thể**, hoặc gọi một cách khái quát, là **đường cong hồi qui tổng thể**. Đơn giản hơn, **đường thẳng đó chính là hồi qui của Y trên X** .

³Các bạn đọc cần nhớ các dữ liệu của ta là giả thiết. Ở đây chúng tôi không gợi ý rằng trung bình có điều kiện sẽ luôn nằm trên một đường thẳng; chúng có thể nằm trên một đường cong.

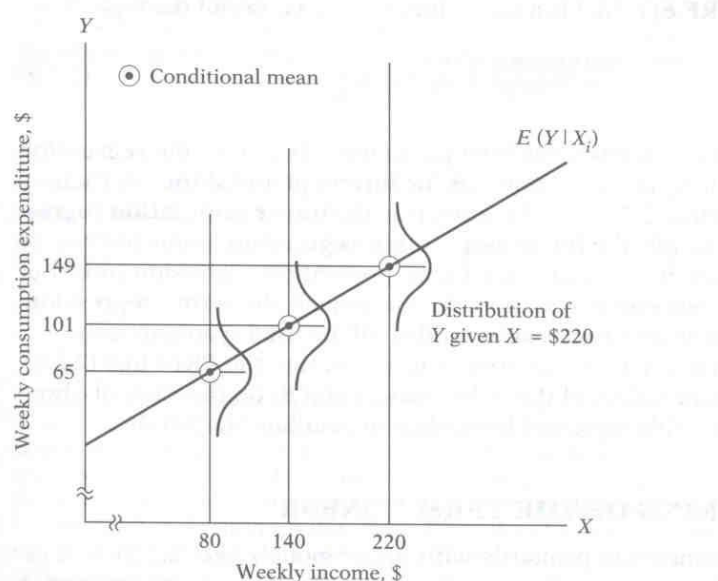
BẢNG 2.2**Xác suất có Điều kiện $p(Y | X_i)$ của dữ liệu trong Bảng 2.1**

$p(Y X_i)$ $X \rightarrow$ ↓	80	100	120	140	160	180	200	220	240	260
Xác suất có điều kiện $p(Y X_i)$	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	—	1/6	—	1/7	1/6	1/6	—	1/7	1/6	1/7
	—	—	—	1/7	—	—	—	1/7	—	1/7
Trung bình có điều kiện của Y	65	77	89	101	113	125	137	149	161	173

Như vậy về mặt hình học, một đường cong hồi qui tổng thể đơn giản là quỹ tích của các trung bình có điều kiện hay các kỳ vọng có điều kiện của biến số phụ thuộc đối với các giá trị xác định của (các) biến giải thích. Có thể vẽ đường này như trong hình 2.2, cho thấy đối với mỗi X_i có một tổng thể các giá trị Y (được giả định là có phân phối chuẩn vì những lý do chúng tôi sẽ giải thích sau) và một trung bình (có điều kiện) tương ứng. Và đường thẳng hay đường cong hồi qui đi ngang qua những giá trị trung bình có điều kiện này. Với cách giải thích này về đường cong hồi qui các bạn có lẽ cảm thấy sẽ bổ ích hơn nếu đọc lại định nghĩa của hồi qui đã cho trong phần 1.2.

**Hình 2.1**

Phân phối có điều kiện của chi tiêu đối với những mức độ thu nhập khác nhau (dữ liệu ở Bảng 2.1)

**Hình 2.2**

Đường hồi quy tổng thể (dữ liệu của Bảng 2.10)

2.2 KHÁI NIỆM HÀM HỒI QUI TỔNG THỂ (PRF)

Từ phần thảo luận trước và đặc biệt là từ hai hình 2.1 và 2.2, rõ ràng là mỗi trung bình có điều kiện $E(Y | X_i)$ là một hàm của X_i . Thể hiện bằng các ký hiệu:

$$E(Y | X_i) = f(X_i) \quad (2.2.1)$$

trong đó $f(X_i)$ là hàm của biến giải thích X_i . [Trong ví dụ giả thiết của chúng ta, $E(Y | X_i)$ là hàm tuyến tính của X_i .] Phương trình (2.2.1) được gọi là **hàm hồi qui tổng thể** (hai biến) (PRF), hay một cách ngắn gọn là **hồi qui tổng thể** (PR). Phát biểu một cách đơn giản là, trung bình (tổng thể) của phân phối của Y với điều kiện X_i là có quan hệ hàm số với X_i . Nói một cách khác, nó cho biết giá trị trung bình của Y biến đổi như thế nào so với X .

Hàm $f(X_i)$ có dạng như thế nào? Câu hỏi này quan trọng bởi vì trong những tình huống thực tế chúng ta không có sẵn toàn bộ tổng thể để xem xét. Do đó, dạng hàm của PRF là một vấn đề thực nghiệm, mặc dù trong các trường hợp cụ thể lý thuyết có thể giúp cho ta một vài điều. Ví dụ, một nhà kinh tế học có thể giả thiết rằng chi tiêu tiêu dùng là có quan hệ tuyến tính với thu nhập. Như vậy, giả thiết gần đúng hay có thể đúng đầu tiên của chúng ta là giả định rằng PRF $E(Y | X_i)$ là một hàm tuyến tính của X_i , giả dụ thuộc loại

$$E(Y | X_i) = \beta_1 + \beta_2 X_i \quad (2.2.2)$$

trong đó β_1 và β_2 là những thông số không biết nhưng không thay đổi được gọi là các **hệ số hồi qui**; β_1 và β_2 còn được tuần tự gọi là **hệ số tung độ gốc** và **hệ số độ dốc**. Phương trình (2.2.2) được gọi là **hàm hồi qui tổng thể tuyến tính**. Một số biểu thức thay thế được dùng trong các tài liệu là mô hình hồi qui tổng thể tuyến tính hay phương trình hồi qui tổng thể tuyến tính. Trong các phần tiếp theo sau, các thuật ngữ hồi qui, phương trình hồi qui, và mô hình hồi qui sẽ được dùng với nghĩa như nhau.

Khi phân tích hồi qui mối quan tâm của chúng ta là để dự đoán các PRF như (2.2.2), có nghĩa là, dự đoán các giá trị không biết β_1 và β_2 trên cơ sở quan sát trên Y và X . Vấn đề này sẽ được nghiên cứu chi tiết ở Chương 3.

2.3 Ý NGHĨA CỦA THUẬT NGỮ "TUYẾN TÍNH"

Bởi vì tài liệu này quan tâm chủ yếu đến các mô hình tuyến tính như (2.2.2), do đó điều cần thiết là phải biết thuật ngữ "tuyến tính" thật sự có ý nghĩa gì, bởi vì có thể hiểu từ này theo hai cách khác nhau.

Sự tuyến tính theo các Biến số

Ý nghĩa đầu tiên và có lẽ "tự nhiên" hơn của sự tuyến tính đó là kỳ vọng có điều kiện của Y là một hàm tuyến tính của X_i , ví dụ như là (2.2.2).⁴ Về mặt hình học, đường cong tuyến tính trong trường hợp này là một đường thẳng. Theo cách giải thích này, một hàm tuyến tính như $E(Y|X_i) = \beta_1 + \beta_2 X_i^2$ không phải là một hàm tuyến tính bởi vì biến số X xuất hiện với số mũ hay lũy thừa 2.

Sự tuyến tính theo các Thông số

Cách giải thích thứ hai của sự tuyến tính là kỳ vọng có điều kiện của Y , $E(Y|X_i)$, là một hàm tuyến tính theo các thông số, các β ; nó có thể tuyến tính hoặc có thể không tuyến tính theo biến X .⁵ Theo cách giải thích này, $E(Y|X_i) = \beta_1 + \beta_2 X_i^2$ là một mô hình tuyến tính nhưng $E(Y|X_i) = \beta_1 + \sqrt{\beta_2} X_i$ thì không phải. Biểu thức thứ hai là một ví dụ của mô hình hồi qui không tuyến tính (theo các thông số); chúng ta sẽ không bàn tới những mô hình như vậy trong tài liệu này.

Trong hai cách giải thích về sự tuyến tính, tuyến tính theo các thông số là có liên quan đến sự phát triển của lý thuyết hồi qui dưới đây. Do đó, từ đây trở đi, thuật ngữ hồi qui "tuyến tính" sẽ luôn có nghĩa là một hồi qui tuyến tính theo các thông số, các β , (có nghĩa là, các thông số chỉ có lũy thừa bằng 1 mà thôi); nó có thể có tuyến tính hoặc có thể không tuyến tính theo các biến giải thích, tức các giá trị X . Điều này được trình bày một cách sơ đồ hóa trong Bảng 2.3. Như vậy, $E(Y|X_i) = \beta_1 + \beta_2 X_i$ sẽ tuyến tính theo thông số và theo biến số, là một LRM, và $E(Y|X_i) = \beta_1 + \beta_2 X_i^2$ cũng vậy, sẽ tuyến tính theo các thông số nhưng không tuyến tính theo biến số X .

BẢNG 2.3

Các Mô hình Hồi qui Tuyến tính

Mô hình tuyến tính theo các thông số ?	Mô hình tuyến tính theo các biến số ?	
	Phải	Không phải
Phải	LRM	LRM
Không phải	NLRM	NLRM

Chú ý: LRM = mô hình hồi qui tuyến tính

⁴ Hàm $Y = f(x)$ được coi là tuyến tính theo X nếu X xuất hiện với lũy thừa hay chỉ số chỉ bằng 1 mà thôi (có nghĩa là những số hạng như X^2 , $\sqrt{X}$ v.v. được loại bỏ) và không được nhân hay chia với bất cứ một biến nào khác (ví dụ, $X \cdot Z$ hay X/Z , trong đó Z là một biến khác). Nếu Y chỉ phụ thuộc vào một biến X , một cách khác để nói rằng Y có quan hệ tuyến tính với X là tỉ lệ thay đổi của Y so với X (có nghĩa là độ dốc, hay đạo hàm, của Y so với X , dY/dX) là không phụ thuộc vào giá trị của X . Như vậy, nếu $Y=4X$, $dY/dX=4$, tức kết quả này không phụ thuộc vào giá trị của X . Nhưng nếu $Y=4X^2$, $dY/dX=8X$, tức có phụ thuộc vào giá trị của X . Do đó hàm này không tuyến tính theo X .

⁵ Một hàm được gọi là tuyến tính theo thông số, ví dụ như β_1 , nếu β_1 xuất hiện với lũy thừa bằng 1 và không nhân hay chia bất cứ một thông số nào khác (ví dụ $\beta_1\beta_2$, β_2/β_1 , v.v.)

NLRM = mô hình hồi qui không tuyến tính

2.4 ĐẶC TRƯNG NGẪU NHIÊN CỦA PRF

Từ hình 2.1 ta thấy rõ rằng khi thu nhập gia đình tăng, chi tiêu tiêu dùng của gia đình về mặt trung bình cũng tăng theo. Nhưng còn chi tiêu tiêu dùng của từng gia đình so với mức thu nhập (không đổi) của mình thì sao? Từ hình 2.1 và Bảng 2.1 ta thấy rõ chi tiêu tiêu dùng của từng gia đình không nhất thiết phải tăng khi mức thu nhập tăng. Ví dụ, trong Bảng 2.1 chúng ta quan sát thấy tương ứng với mức thu nhập 100 đôla có một gia đình với mức chi tiêu tiêu dùng là 65 đôla thấp hơn mức chi tiêu tiêu dùng của hai gia đình mà mức thu nhập hàng tuần chỉ có 80 đôla. Nhưng lưu ý rằng mức chi tiêu tiêu dùng *trung bình* của các gia đình với thu nhập hàng tuần là 100 đôla là lớn hơn mức chi tiêu tiêu dùng trung bình của những gia đình có mức thu nhập hàng tuần là 80 đôla (77 đôla so với 65 đôla).

Như vậy, chúng ta có thể nói gì về mối tương quan giữa mức chi tiêu tiêu dùng của một gia đình cá thể và một mức thu nhập nhất định? Từ hình 2.1 chúng ta thấy rằng với mức thu nhập là X_i , mức chi tiêu tiêu dùng của một gia đình cá thể nằm xung quanh chi tiêu trung bình của tất cả các gia đình ở tại X_i , có nghĩa là xung quanh kỳ vọng có điều kiện của nó. Do đó, chúng ta có thể diễn đạt *độ lệch* của một Y_i xung quanh giá trị kỳ vọng của nó như sau:

$$u_i = Y_i - E(Y | X_i)$$

hay

$$Y_i = E(Y | X_i) + u_i \quad (2.4.1)$$

trong đó độ lệch u_i là một biến số ngẫu nhiên không thể quan sát có các giá trị âm và dương. Diễn đạt bằng thuật ngữ chuyên môn, u_i được gọi là **số hạng nhiễu ngẫu nhiên** hay **số hạng sai số ngẫu nhiên**.

Chúng ta giải thích (2.4.1) như thế nào? Chúng ta có thể nói rằng chi tiêu của một gia đình cá thể, khi biết mức thu nhập của nó, có thể được thể hiện như là tổng của hai thành tố, (1) $E(Y | X_i)$, đơn giản là chi tiêu tiêu dùng trung bình của tất cả các gia đình có cùng mức thu nhập. Thành tố này được gọi là thành tố **tất định** hay **hệ thống**, và (2) u_i , là thành tố **ngẫu nhiên** hay **không hệ thống**. Chúng ta sẽ nhanh chóng xem xét bản chất của số hạng nhiễu ngẫu nhiên, nhưng tạm thời giả định rằng nó là một số hạng **thay thế** hay **đại diện** cho tất cả các biến số ta bỏ ra ngoài hay bỏ sót mà có thể ảnh hưởng đến Y nhưng không được (hay không thể) đưa vào trong mô hình hồi qui.

Nếu $E(Y | X_i)$ được giả định là tuyến tính theo X_i , như trong (2.2.2), phương trình (2.4.1) có thể được biểu thị như sau:

$$\begin{aligned} Y_i &= E(Y | X_i) + u_i \\ &= \beta_1 + \beta_2 X_i + u_i \end{aligned} \quad (2.4.2)$$

Phương trình (2.4.2) giả định rằng chi tiêu tiêu dùng của một gia đình có quan hệ tuyến tính đối với thu nhập cộng với số hạng nhiễu. Như vậy, chi tiêu tiêu dùng của một gia đình, với $X = 80$ (xem Bảng 2.1), có thể được biểu thị như sau

$$\begin{aligned} Y_1 &= 55 = \beta_1 + \beta_2(80) + u_1 \\ Y_2 &= 60 = \beta_1 + \beta_2(80) + u_2 \\ Y_3 &= 65 = \beta_1 + \beta_2(80) + u_3 \\ Y_4 &= 70 = \beta_1 + \beta_2(80) + u_4 \end{aligned} \quad (2.4.3)$$

$$Y_5 = 75 = \beta_1 + \beta_2(80) + u_5$$

Bây giờ nếu chúng ta lấy giá trị kỳ vọng của (2.4.2) ở cả hai vế, chúng ta được

$$\begin{aligned} E(Y_i | X_i) &= E[E(Y | X_i)] + E(u_i | X_i) \\ &= E(Y | X_i) + E(u_i | X_i) \end{aligned} \quad (2.4.4)$$

trong đó ta vận dụng một đặc tính là giá trị kỳ vọng của một hằng số chính là hằng số đó.⁶ Lưu ý cẩn thận rằng trong (2.4.4) chúng ta đã lấy giá trị kỳ vọng có điều kiện, phụ thuộc vào giá trị của X đã cho.

Bởi vì $E(Y_i | X_i)$ cũng chính là $E(Y | X_i)$, phương trình (2.4.4) cho thấy rằng

$$E(u_i | X_i) = 0 \quad (2.4.5)$$

Như vậy, giả định cho rằng đường hồi qui đi ngang qua các giá trị trung bình có điều kiện của Y (xem hình 2.2) có nghĩa là các giá trị trung bình có điều kiện của u_i (phụ thuộc vào các giá trị của X) là bằng zero.

Từ lý luận ở trên chúng ta thấy rõ ràng là (2.2.2) và (2.4.2) và các hình thức tương đương nếu $E(u_i | X_i) = 0$.⁷ Nhưng đặc trưng ngẫu nhiên của (2.4.2) có ưu điểm ở chỗ nó cho thấy một cách rõ ràng là có những biến số khác ngoài thu nhập ra có thể ảnh hưởng đến chi tiêu tiêu dùng và không thể giải thích một cách đầy đủ chi tiêu tiêu dùng của một gia đình chỉ bằng (những) biến số nằm trong mô hình hồi qui.

2.5 Ý NGHĨA CỦA SỐ HẠNG NHIỀU NGẪU NHIÊN

Như đã được lưu ý trong Phần 2.4, số hạng nhiễu u_i là số hạng thay thế cho tất cả những biến số bị bỏ ra khỏi mô hình nhưng tất cả những biến số này tập hợp lại có ảnh hưởng đến Y . Câu hỏi đặt ra là: Tại sao không đưa thẳng những biến này vào trong mô hình một cách công khai? Nói một cách khác, tại sao không phát triển một mô hình hồi qui bội với càng nhiều biến càng tốt? Có rất nhiều lý do.

1. *Sự mơ hồ của lý thuyết*: Lý thuyết quyết định hành vi của Y , có thể, và thường là, không hoàn chỉnh. Chúng ta có thể biết chắc chắn rằng thu nhập hàng tuần X ảnh hưởng đến chi tiêu tiêu dùng hàng tuần Y , nhưng chúng ta có thể không biết hoặc không biết chắc về những biến khác ảnh hưởng đến Y . Do đó, u_i có thể được sử dụng làm một biến thay thế cho tất cả những biến bị loại bỏ hay bỏ ra khỏi mô hình.
2. *Dữ liệu không có sẵn*: Ngay cả nếu chúng ta biết một số trong những biến bị loại bỏ là những biến gì và do đó có thể xem xét đến một hồi qui bội thay vào hồi qui đơn, chúng ta chưa chắc có thể có được những thông tin định lượng về những biến này. Một kinh nghiệm thường gặp trong phân tích thực nghiệm là những dữ liệu lý tưởng mà chúng ta muốn có thông thường lại là không có được. Ví dụ, trên nguyên tắc chúng ta có thể đưa sự giàu có của gia đình vào làm biến giải thích thêm với biến thu nhập để giải thích chi tiêu tiêu dùng của gia đình. Nhưng không may là thông tin về sự giàu có của gia đình thông thường là không có. Do đó chúng ta buộc phải loại bỏ biến giàu có ra khỏi mô hình của mình mặc dù nó có tầm quan trọng lý thuyết rất lớn và cần thiết để giải thích chi tiêu tiêu dùng.

⁶ Xem Phụ lục A về phần thảo luận về các đặc tính của toán tử kỳ vọng E . Chú ý rằng $E(Y | X_i)$, một khi giá trị của X_i là không đổi, sẽ là một hằng số.

⁷ Sự thật là, trong phương pháp bình phương tối thiểu sẽ được phát triển ở chương 3, chúng ta giả định một cách rõ ràng là $E(u_i | X_i) = 0$. Xem Phần 2.3.

3. *Các biến cốt lõi (core) và biến ngoại vi (peripheral)*: Giả định rằng trong ví dụ về thu nhập-chi tiêu của chúng ta, ngoài thu nhập X_1 ra, số con trong mỗi gia đình X_2 , giới tính X_3 , tôn giáo X_4 , giáo dục X_5 , và khu vực địa lý X_6 cũng ảnh hưởng đến chi tiêu tiêu dùng. Nhưng hoàn toàn có thể là ảnh hưởng chung của tất cả hay của một vài biến này có thể rất nhỏ và thậm chí là rất không hệ thống hoặc ngẫu nhiên đến mức xét về phương diện thực tế và vì những lý do về chi phí việc đưa chúng vào trong mô hình một cách rõ ràng là không có ích lợi. Chúng ta hy vọng rằng ảnh hưởng kết hợp chung của chúng có thể được xử lý như là biến ngẫu nhiên u_i .⁸

4. *Bản chất ngẫu nhiên trong hành vi của con người*: Ngay cả khi chúng ta thành công trong việc đưa tất cả các biến liên quan vào trong mô hình, chắc chắn vẫn còn một số "ngẫu nhiên" thuộc bản chất trong cá thể Y mà không thể giải thích được dù cho chúng ta có cố gắng đến mấy. Các biến nhiễu, các biến số u , rất có thể đã thể hiện được bản chất ngẫu nhiên này.

5. *Các biến thay thế kém*: Mặc dù mô hình hồi qui cổ điển (sẽ được phát triển ở chương 5) giả định rằng các biến Y và X được tính toán một cách chính xác, trên thực tế các dữ liệu có thể không chính xác vì những sai số về tính toán. Ví dụ như xem lý thuyết nổi tiếng của Milton Friedman về hàm chi tiêu.⁹ Ông xem *tiêu thụ thường xuyên* (Y^p) là một hàm của *thu nhập thường xuyên* (X^p). Nhưng bởi vì dữ liệu về những biến số này không thể trực tiếp quan sát được, trên thực tế chúng ta dùng các biến thay thế, ví dụ như chi tiêu hiện thời (Y) và thu nhập hiện thời (X), là những biến mà chúng ta có thể quan sát được. Bởi vì Y và X quan sát được có thể không tương đương với Y^p và X^p , ta gặp phải vấn đề về sai sót trong tính toán. Như vậy số hạng nhiễu u trong trường hợp này có thể còn tượng trưng cho sai sót trong tính toán. Như chúng ta sẽ thấy trong chương sau, nếu có những sai sót như vậy trong tính toán, chúng có thể có những tác động nghiêm trọng đối với việc tính toán các hệ số hồi qui β .

6. *Nguyên tắc chi li*: Tuân theo nguyên tắc Lưỡi dao Occam,¹⁰ chúng tôi muốn giữ cho mô hình hồi qui của mình càng đơn giản càng tốt. Nếu chúng ta có thể giải thích hành vi của Y "một cách đầy đủ" bằng hai hay ba biến giải thích và nếu lý thuyết của chúng ta không đủ mạnh để cho ta thấy có thể đưa những biến nào khác vào, tại sao còn đưa thêm biến vào? Hãy để u_i biểu thị tất cả những biến khác. Dĩ nhiên, chúng ta không nên loại bỏ những biến quan trọng và liên quan chỉ nhằm để giữ cho mô hình đơn giản.

7. *Dạng hàm sai*: Ngay cả khi về mặt lý thuyết chúng ta có được những biến đúng để giải thích cho một hiện tượng và ngay cả khi chúng ta có thể thu được dữ liệu về những biến này, thông thường chúng ta không biết dạng quan hệ hàm số giữa các biến hồi qui phụ thuộc và biến hồi qui độc lập. Có phải chi tiêu tiêu dùng là một hàm (theo biến số) tuyến tính của thu nhập hay là hàm không tuyến tính (theo biến số)? Nếu là trường hợp đầu, $Y_i = \beta_1 + \beta_2 X_i + u_i$ là quan hệ hàm số thích hợp giữa Y và X , nhưng nếu là trường hợp sau, $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i$ có thể là dạng hàm đúng. Trong các mô hình hai biến có thể suy xét dạng hàm của mối quan hệ từ đồ thị phân tán. Nhưng trong một mô hình hồi qui bội, không dễ dàng xác định dạng hàm thích hợp, bởi vì chúng ta không thể tưởng tượng ra được đồ thị phân tán trong không gian đa chiều.

Vì tất cả những lý do này, các số hạng nhiễu u_i đóng một vai trò vô cùng quan trọng trong phân tích hồi qui, chúng ta sẽ thấy điều này khi chúng ta tiếp tục.

⁸ Một khó khăn nữa là các biến như giới tính, giáo dục, tôn giáo v.v. là rất khó định lượng.

⁹ Milton Friedman, *A Theory of the Consumption Function* (Một lý thuyết về hàm tiêu dùng), Princeton University Press, Princeton, N.J., 1957.

¹⁰ "Nên giữ cho sự diễn tả càng đơn giản càng tốt cho đến khi nào tỏ ra không thỏa đáng thì thôi," *The World of Mathematics* (Thế giới toán học), tập 2, J. R. Newman, Simon & Schuster, New York, 1956, trang 1247, hay "Không nên nhân các đối tượng vượt quá mức cần thiết," Donald F. Morrison, *Applied Linear Statistical Methods*, Prentice Hall, Englewood Cliffs, N.J., 1983, trang 58.

2.6 HÀM HỒI QUI MẪU (SRF)

Cho tới giờ bằng cách giới hạn sự thảo luận của chúng ta vào tổng thể các giá trị Y tương ứng với các giá trị không đổi của X , chúng ta đã cố tình tránh không xem xét đến việc lấy mẫu (lưu ý rằng các dữ liệu trong Bảng 2.1 là tiêu biểu cho tổng thể, không phải là một mẫu). Nhưng giờ đây đã đến lúc phải đối diện với những vấn đề về lấy mẫu, bởi vì trong hầu hết các tình huống thực tế những gì chúng ta có chỉ là một mẫu những giá trị của Y tương ứng với một số X không đổi. Do đó, nhiệm vụ của chúng ta bây giờ là phải tính toán PRF trên cơ sở thông tin mẫu.

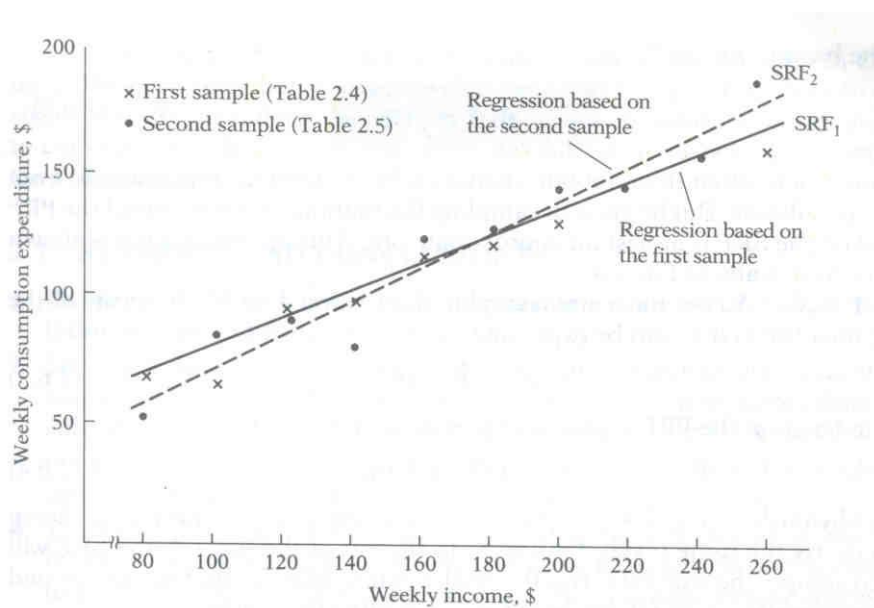
Bảng 2.4**Một mẫu ngẫu nhiên từ tổng thể của Bảng 2.1**

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

Để minh họa, giả vờ rằng chúng ta chưa biết được tổng thể của Bảng 2.1 và thông tin duy nhất chúng ta có là một mẫu lựa chọn ngẫu nhiên các giá trị Y tương ứng với X không đổi đã cho trong Bảng 2.4. Không giống như trong Bảng 2.1, ở đây chúng ta có chỉ một giá trị Y tương ứng với giá trị X đã biết; mỗi Y (đã biết X_i) trong Bảng 2.4 được chọn một cách ngẫu nhiên từ những Y tương tự nhau tương ứng với cùng một X_i từ tổng thể ở Bảng 2.1.

Vấn đề là: Từ mẫu Bảng 2.4 liệu chúng ta có thể tiên đoán được chi tiêu tiêu dùng hàng tuần trung bình Y trong tổng thể tương ứng với X được chọn? Nói một cách khác, liệu chúng ta có thể tính được PRF từ dữ liệu mẫu không? Như các bạn đọc chắc chắn đã nghi vấn, chúng ta có thể sẽ không thể tính được PRF "một cách chính xác" bởi vì những giao động của việc lấy mẫu. Để thấy được điều này, giả sử chúng ta lấy một mẫu ngẫu nhiên khác từ tổng thể ở Bảng 2.1, như được trình bày trong Bảng 2.5.

Về đồ thị các dữ liệu của Bảng 2.4 và 2.5, chúng ta đặt được đồ thị phân tán như trong hình 2.3. Trong đồ thị phân tán hai đường hồi qui mẫu được vẽ sao cho tương đối "thích hợp" với các điểm rời rạc: SRF_1 được vẽ trên cơ sở mẫu thứ nhất, và SRF_2 trên cơ sở mẫu thứ hai. Đường nào trong hai đường hồi qui này thể hiện đường hồi qui tổng thể "thực"? Nếu chúng ta không xem hình 2.1, được cho là thể hiện PR, không có cách nào chúng ta có thể hoàn toàn chắc chắn rằng một trong hai đường hồi qui trong hình 2.3 thể hiện đường (đường cong) hồi qui tổng thể thực. Đường hồi qui trong hình 2.3 được gọi là các **đường hồi qui mẫu**. Chúng được xem là thể hiện đường hồi qui tổng thể, nhưng bởi vì các giao động của việc lấy mẫu chúng chỉ có thể là sự gần bằng của đường PR thật. Nhìn chung, chúng ta sẽ thu được N lần các SRF khác nhau cho N các mẫu khác nhau, và những SRF này ít có khả năng sẽ giống nhau.



Hình 2.3. Đường hồi quy dựa trên hai mẫu khác nhau

Bảng 2.5

Một mẫu ngẫu nhiên khác từ tổng thể của Bảng 2.1

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

Giờ đây, tương tự như đường PRF nằm dưới đường hồi qui tổng thể, chúng ta có thể phát triển khái niệm **hàm hồi qui mẫu** (SRF) để thể hiện đường hồi qui mẫu. Biểu thức mẫu tương ứng với (2.2.2) có thể được viết thành

$$Y_i = \beta_1 + \beta_2 X_i \quad (2.6.1)$$

trong đó Y được đọc là "Y mũ"

Y_i = hàm ước lượng của $E(Y | X_i)$

trong đó β_1 = hàm ước lượng của β_1

β_2 = hàm ước lượng của β_2

Lưu ý rằng **hàm ước lượng**, còn được biết như là một trị thống kê (mẫu), đơn giản chỉ là một quy tắc hay công thức hay phương pháp cho chúng ta biết làm cách nào để tính toán thông số của

tổng thể từ các thông tin được cung cấp từ mẫu đang xem xét. Một giá trị bằng số nhất định thu được bằng cách áp dụng hàm ước lượng được gọi là một **giá trị ước lượng**.¹¹

Cũng giống như chúng ta đã biểu diễn PRF qua hai biểu thức tương đương (2.2.2) và (2.4.2), chúng ta có thể biểu diễn SRF (2.6.1) dưới dạng ngẫu nhiên của nó như sau:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (2.6.2)$$

trong đó, ngoài những ký hiệu mà chúng ta đã định nghĩa, u_i là số hạng **phần dư** (mẫu). Về mặt khái niệm u_i cũng tương tự như u_i và có thể được xem như một *ước lượng* của u_i . Nó được đưa vào trong SRF cũng cùng với một lý do như u_i được đưa vào trong PRF.

Nói tóm lại, mục tiêu chính của chúng ta trong phân tích hồi quy là để tính PRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (2.4.2)$$

trên cơ sở của SRF

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (2.6.2)$$

bởi vì thông thường phương pháp phân tích của chúng ta được dựa trên một mẫu duy nhất lấy từ một tổng thể. Nhưng bởi vì những giao động của việc lấy mẫu ước lượng của chúng ta về PRF trên cơ sở SRF chỉ có thể là một sự gần đúng tốt nhất. Sự gần đúng này được đưa thể hiện bằng biểu đồ thông qua hình 2.4.

Đối với $X = X_i$, chúng ta có một quan sát (mẫu) $Y = Y_i$. Theo SRF, có thể thể hiện Y_i quan sát được như sau

$$Y_i = Y_1 + u_i \quad (2.6.3)$$

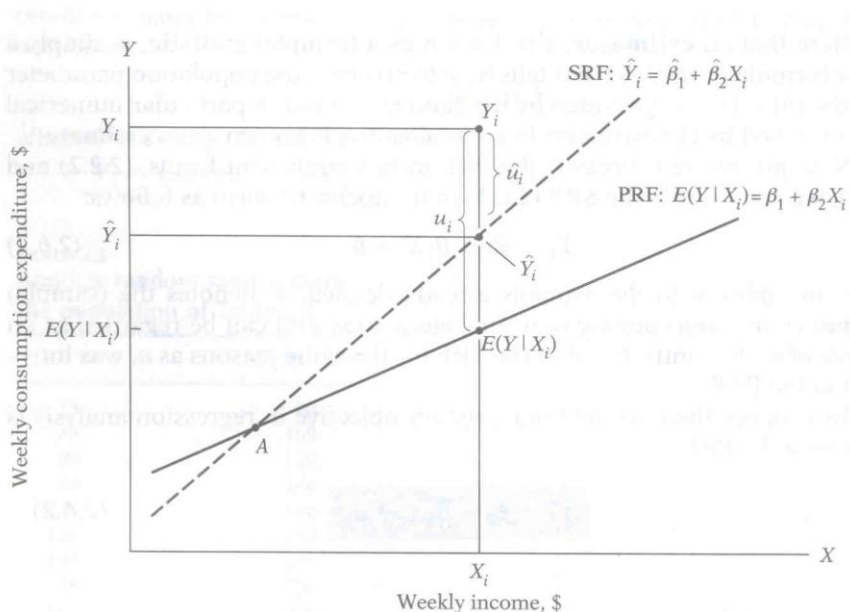
và theo PRF nó có thể được thể hiện như sau

$$Y_i = E(Y | X_i) + u_i \quad (2.6.4)$$

Rõ ràng là trong hình 2.4 Y_i *ước lượng quá cao* $E(Y | X_i)$ thực đối với X_i trong hình 2.4. Cũng tương tự như vậy, đối với bất cứ một X_i nằm bên trái của điểm A, SRF sẽ *ước lượng quá thấp* PRF thực. Nhưng các bạn có thể dễ dàng thấy rằng những ước lượng quá cao và quá thấp này là điều không thể tránh khỏi bởi vì những giao động của việc lấy mẫu.

Bây giờ câu hỏi quan trọng là: Giả sử rằng SRF chỉ là một sự gần đúng của PRF, liệu chúng ta có thể đặt ra một quy luật hay một phương pháp để đưa ước lượng này càng "gần" đúng hơn được không? Nói một cách khác, làm cách nào để thiết lập SRF sao cho β_1 càng "gần" với β_1 thực và β_2 càng "gần" với β_2 thực ngay cả khi chúng ta không thể biết được β_1 và β_2 thực?

¹¹ Như đã lưu ý trong phần Giới thiệu, dấu mũ ở trên một biến số tượng trưng cho hàm ước lượng của giá trị tổng thể liên quan.

**Hình 2.4.** Mẫu và đường hồi quy dân số

Câu trả lời cho vấn đề này sẽ chiếm nhiều công sức giải thích trong chương 3. Ở đây chúng ta lưu ý rằng chúng ta có thể phát triển những phương pháp có thể chỉ cho chúng ta làm cách nào để thiết lập SRF để thể hiện PRF một cách trung thực nhất. Quan niệm rằng có thể làm điều này được ngay cả khi chúng ta không thật sự có thể xác định được PRF là một điều lý thú.

2.7 TÓM TẮT VÀ KẾT LUẬN

1. Khái niệm chính làm nền tảng cho phân tích hồi qui là khái niệm **hàm hồi qui tổng thể** (PRF).
2. Tập sách này đề cập đến PRF tuyến tính, có nghĩa là, những hồi qui tuyến tính theo các tham số chưa biết. Chúng có thể tuyến tính hay có thể không tuyến tính theo các biến phụ thuộc hay biến hồi qui phụ thuộc Y và các biến độc lập hay (các) biến hồi qui độc lập X .
3. Vì mục đích thực nghiệm, PRF ngẫu nhiên mới chính là điều quan trọng. Số hạng nhiễu ngẫu nhiên u_i đóng một vai trò quyết định trong việc ước lượng PRF.
4. Đường PRF là một khái niệm lý tưởng hóa, bởi vì trên thực tế chúng ta ít khi có thể được toàn bộ một tổng thể mà chúng ta cần. Thông thường, chúng ta có được một mẫu những quan sát từ tổng thể. Do đó, chúng ta dùng hàm hồi qui mẫu ngẫu nhiên (SRF) để ước lượng PRF. Chúng ta sẽ thấy điều này được thực hiện như thế nào ở chương 3.

BÀI TẬP

2.1 Bảng dưới đây cho ta các suất sinh lời dự đoán trong một năm của một dự án đầu tư và các xác suất liên quan của chúng.

Suất sinh lời X , %	Xác suất p_i
-20	0.10
-10	0.15
10	0.45

25	0.25
30	0.05

Sử dụng các định nghĩa đã cho trong bảng phụ lục A, hãy thực hiện các yêu cầu sau:

- Tính suất sinh lời kỳ vọng, $E(X)$.
- Tính phương sai (σ^2) và độ lệch chuẩn (σ) của các suất sinh lời.
- Hãy tính hệ số của độ biến thiên, V , được định nghĩa là $V = \sigma / E(X)$. Chú ý: V thường được nhân với 100 để biểu thị nó dưới dạng phần trăm.
- Dùng định nghĩa của độ lệch (skewness), hãy tính độ lệch của phân phối các suất sinh lời cho trong bảng. Phân phối suất sinh lời trong ví dụ này là lệch dương hay lệch âm?
- Dùng định nghĩa về độ nhọn (kurtosis), hãy tính độ nhọn trong ví dụ này. Phân phối suất sinh lời cho trong bảng này có độ nhọn vượt chuẩn (dạng đuôi hẹp) hay dưới chuẩn (đuôi dài)?

2.2 Bảng dưới đây cho ta phân phối xác suất liên kết, $p(X,Y)$, của các biến X và Y .

Y	X		
	1	2	3
1	0.03	0.06	0.06
2	0.02	0.04	0.04
3	0.09	0.18	0.18
4	0.06	0.12	0.12

Sử dụng các định nghĩa đã cho trong bảng phụ lục A, hãy tính các yêu cầu sau:

- Phân phối xác suất không điều kiện hay xác suất biên của X và Y .
- Tính các phân phối xác suất có điều kiện $p(X/Y_i)$ và $p(Y/X_i)$.
- Các kỳ vọng có điều kiện $E(X/Y_i)$ và $E(Y/X_i)$.

2.3 Bảng dưới đây cho ta phân phối xác suất liên kết, $p(X,Y)$, của các biến ngẫu nhiên X và Y trong đó X = suất sinh lời trong năm đầu tiên (%) kỳ vọng sẽ đạt được từ dự án A và Y = suất sinh lời trong năm đầu tiên (%) kỳ vọng sẽ đạt được từ dự án B

Y	X			
	-10	0	20	30
20	0.27	0.08	0.16	0.00
50	0.00	0.04	0.10	0.35

- Tính suất sinh lời kỳ vọng của dự án A, $E(X)$.
- Tính suất sinh lời kỳ vọng của dự án B, $E(Y)$.
- Các suất sinh lời của hai dự án có độc lập không? (Gọi ý: $E(XY) = E(X)E(Y)$?) Lưu ý rằng

$$E(XY) = \sum_{i=1}^4 \sum_{j=1}^2 X_i Y_j p(X_i Y_j)$$

2.4 Có 50 cặp vợ chồng, tuổi (tính bằng năm) của những người vợ X và chồng Y được xếp thành nhóm trong bảng sau với khoảng của các nhóm là 10 năm, tần số của các nhóm khác nhau được trình bày trong phần giữa của Bảng. Các giá trị của X và Y là các giá trị ở giữa trong các nhóm.

Y \ X	20	30	40	50	60	70	Tổng
20	1						1
30	2	11	1				14
40		4	10	1			15
50			3	6	1		10
60				2	3	2	7
70					1	2	3
Tổng	3	15	14	9	5	4	50

Như vậy, đối với nhóm trong đó tuổi của người chồng nằm giữa 35 và 45 và tuổi của người vợ là giữa 25 và 35, các giá trị của Y và X lần lượt (được tập trung vào) là 40 và 30, và tần số là 4.

- Xác định trung bình của mỗi dãy, có nghĩa là, mỗi hàng ngang và mỗi cột dọc.
- Đặt biến X trên hoành độ và biến Y trên tung độ, vẽ đồ thị cho các trung bình dãy (hay có điều kiện) đã tính được ở câu trên. Các Anh/Chị có thể sử dụng ký hiệu $+$ cho trung bình cột dọc và $\oplus$ cho trung bình hàng ngang.
- Chúng ta có thể đưa ra nhận xét gì về quan hệ giữa X và Y ?
- Các trung bình cột dọc và hàng ngang có điều kiện có nằm trên một đường tương đối thẳng không? Vẽ các đường hồi qui.

2.5 Bảng dưới đây cung cấp kết quả định mức (X) và lãi suất hoàn vốn (yield to maturity) Y (%) của 50 trái phiếu, trong đó việc định mức được đánh giá theo 3 cấp: $X=1$ (Bbb), và $X=2$ (Bb), và $X=3$ (B). Theo định mức của Công ty Per Standard & Poor, Bbb, Bb và B tất cả đều là trái phiếu chất lượng trung bình, Bb được đánh giá cao hơn B một ít và Bbb lại được đánh giá cao hơn Bb một ít.

Y \ X	1	2	3	Tổng
	Bbb	Bb	B	cộng
8.5	13	5	0	18
11.5	2	14	2	18
17.5	0	1	13	14
Tổng cộng	15	20	15	50

- Chuyển Bảng ở trên thành một bảng cung cấp phân phối xác suất liên kết, $p(X,Y)$, ví dụ, $p(X=1, Y=8.5) = 13/50 = .26$.
- Tính $p(Y|X=1)$, $p(Y|X=2)$, và $p(Y|X=3)$.
- Tính $E(Y|X=1)$, $E(Y|X=2)$, và $E(Y|X=3)$.
- Các kết quả suất sinh lợi trong câu (c) có phù hợp với những kỳ vọng tiên nghiệm về mối quan hệ giữa định mức trái phiếu và lãi suất hoàn vốn không?

2.6* Hàm mật độ (density) liên kết của hai biến ngẫu nhiên liên tục X và Y là như sau

$$f(X,Y) = \begin{cases} 4 - X - Y & \text{nếu } 0 \leq X \leq 1; \quad 0 \leq Y \leq 1 \\ 0 & \text{những trường hợp khác} \end{cases}$$

- Tính các hàm mật độ biên, $f(X)$ và $f(Y)$.
- Tính các hàm mật độ có điều kiện $f(X|Y)$ và $f(Y|X)$.

* Tùy ý

- c) Tính $E(X)$ và $E(Y)$.
d) Tính $E(X | Y = 0.4)$

2.7 Xem xét các dữ liệu dưới đây.

Lương trung vị của các nhà kinh tế trong theo các nhóm kinh nghiệm và tuổi tác chọn lọc, sổ sách quốc gia, 1966 (ngàn đôla)

Số năm kinh nghiệm chuyên môn										
Tuổi	0-2	2-4	5-9	10-14	15-19	20-24	25-29	30-34	35-39	40-44*
20-24	7.5									
25-29	9.0	9.1	10.0							
30-34	9.0	9.5	11.0	12.6						
35-39		10.0	11.7	13.2	15.0					
40-44		9.6	11.0	13.0	15.5	17.0				
45-49				12.0	15.0	17.0	20.0			
50-54				11.3	13.3	15.0	18.2	20.0		
55-59						13.8	16.0	18.0	19.0	
60-64							13.1	16.0	17.2	18.8
65-69									13.8	17.0
70-74†										12.5
#										

Ghi chú: Các nhóm được chọn bao gồm tất cả những người do 25 người đại diện trả lời hoặc hơn, họ báo cho biết sự kết hợp giữa tuổi tác và kinh nghiệm như trên.

* Nhóm thực gồm có 40 hoặc hơn.

Nhóm thực gồm có 70 hoặc hơn.

Nguồn: N. Arnold Tolles and Emanuel Melichar, "Studies of the Structure of Economists' Salaries and Income" (Các nghiên cứu về Cấu trúc lương và Thu nhập của các Nhà kinh tế), American Economic Review, vol.57, no. 5, pt.2, Suppl., December 1968, bảng H, trang 119

- a) Các dữ liệu này cho ta thấy gì?
b) Tuổi tác hay kinh nghiệm có quan hệ gần hơn đối với mức lương hay không? Làm sao Anh /Chị biết?
c) Hãy vẽ hai hình riêng biệt, một trình bày mức lương trung vị quan hệ với tuổi tác và một trình bày mức lương trung vị quan hệ với kinh nghiệm nghề nghiệp (tính bằng năm).

2.8 Xem xét các dữ liệu dưới đây.

- a) Dùng trục Y để biểu thị thu nhập bằng tiền trung bình và trục X để tượng trưng cho các trình độ học vấn - 8 năm trở xuống, 1-3 năm học trung học, 4 năm trung học, 1-3 năm đại học, 4 năm đại học và 5 năm đại học trở lên - vẽ đồ thị cho dữ liệu của nam và nữ riêng biệt cho từng nhóm tuổi.
b) Anh / Chị có thể rút ra được kết luận tổng quát gì?

Tuổi và giới tính	Tổng cộng	Tiểu học, 8 năm hay ít hơn	Trung học			Đại học			
			Tổng cộng	1-3 năm	4 năm	Tổng cộng	1-3 năm	4 năm	5 năm hay hơn
Nam, tổng cộng	34,886	19,188	27,131	22,564	28,043	43,217	34,188	44,554	55,831
25 đến 34 tuổi	27,743	15,887	23,255	19,453	24,038	33,003	28,298	35,534	39,833
35 đến 44 tuổi	37,958	18,379	28,205	23,621	28,927	45,819	36,180	47,401	58,542
45 đến 54 tuổi	40,231	19,686	31,235	24,133	32,862	50,545	39,953	50,718	62,902
55 đến 64 tuổi	37,469	22,379	29,460	25,280	30,779	50,585	36,954	55,518	61,647

65 tuổi trở lên	33,145	17,028	24,003	19,530	25,516	44,424	34,323	43,092	52,149
Nữ, tổng cộng	22,768	13,322	18,469	15,381	18,954	27,493	22,654	28,911	35,827
25 đến 34 tuổi	21,337	11,832	16,673	13,385	17,076	25,194	20,872	27,210	32,563
35 đến 44 tuổi	24,453	13,714	19,344	15,695	19,886	29,287	23,307	31,631	37,599
45 đến 54 tuổi	23,429	13,490	19,500	16,651	19,986	29,334	24,608	29,242	38,307
55 đến 64 tuổi	21,388	13,941	18,607	15,202	19,382	26,930	23,364	27,975	33,383
65 tuổi trở lên	19,194	*	18,281	*	18,285	23,277	*	*	*

*Các giá trị cơ sở quá nhỏ để thỏa mãn các tiêu chuẩn thống kê đối với độ tin cậy của các con số tính được.

Nguồn: Statistical Abstract of United States (Tóm Lược Thống Kê của Mỹ), 1992, Bộ thương mại Mỹ, Bảng 713, trang 454.

2.9 Xem xét bảng ở trang bên cạnh:

- Vẽ đồ thị các mức lương trung vị của ba nhóm so với giá trị ở giữa của các khoảng theo số lượng năm kinh nghiệm khác nhau và vẽ các đường hồi qui.
- Những yếu tố nào giải thích cho sự khác biệt trong mức lương của ba nhóm kinh tế gia? Đặc biệt là tại sao các nhà kinh tế có bằng cử nhân kiếm được nhiều tiền hơn các đồng nghiệp của họ có bằng tiến sĩ có 15 năm kinh nghiệm trở lên? Quan sát này có ngụ ý cho thấy rằng có bằng tiến sĩ là không có ích lợi gì hay không?

Các mức lương trung vị của các nhà kinh tế học (ngàn đôla) theo bằng cấp đại học, 1966

Năm kinh nghiệm	Tiến sĩ	Thạc sĩ	Cử nhân
Dưới 2	9.8	8.0	9.0
2 -	10.0	8.8	8.9
5-9	11.5	10.5	10.6
10-14	13.0	12.3	13.0
15-19	15.0	15.0	15.6
20-24	16.2	15.6	17.0
25-29	18.0	17.0	20.0
30-34	17.9	17.7	20.0
35-39	16.9	16.2	20.5
40-14*	17.5	14.2	22.0

*Số nhóm thực là 40 hoặc hơn

Nguồn: N. Arnold Tolles and Emanuel Melichar, "Studies of the Structure of Economists' Salaries and Income," *America Economic Review*, vol. 57, no. 5, pt. 2, Suppl., December 1968, bảng III-B-3, trang 92.

2.10 Xem xét Bảng ở dưới đây:

Số lượng các nhà kinh tế học theo năm kinh nghiệm và tuổi tác (chỉ các nhà kinh tế học làm việc toàn thời gian chuyên nghiệp)

Nhóm tuổi (năm)	Số năm kinh nghiệm						Tổng cộng
	0-2	2 -	5-9	10-14	15-19	20-24*	
20-24	24	13	1	-	-	-	38
25-29	121	405	184	-	-	-	710
30-34	77	497	825	197	3	-	1599
35-39	18	125	535	780	194	1	1653
40-44	6	36	161	652	761	235	1851
45-49	1	15	48	183	433	751	1431
50-54	1	5	19	52	119	784	980
55-59	1	2	10	18	27	612	670

60-64	1	-	3	6	8	382	400
65-69	-	1	1	2	4	206	214
70-74	-	-	-	-	1	27	28
Tổng cộng	250	1099	1787	1890	1550	2998	9574

*Số nhóm thực là 20 hay nhiều hơn.

Ả Số nhóm thực là 70 hay cao hơn.

Source: Adapted from "The Structure of Economists' Employment and Salaries, 1964," *American Economic Review*, vol. 55, no. 4, December 1965, table VII, p. 40.

Bảng ở trên cho thấy tần số tuyệt đối liên kết của các biến tuổi tác và năm kinh nghiệm. Dùng các tần số tương đối (chia tần số tuyệt đối cho tổng số) làm các số đo của xác suất, thực hiện các yêu cầu sau:

- Tính phân phối xác suất liên kết của tuổi tác và các năm kinh nghiệm.
- Tính các phân phối xác suất có điều kiện của tuổi tác cho các năm kinh nghiệm khác nhau.
- Tính phân phối xác suất có điều kiện của các năm kinh nghiệm cho các mức tuổi tác khác nhau.
- Dùng các điểm giữa của các khoảng mức tuổi tác và khoảng năm kinh nghiệm, tính các trung bình có điều kiện của các kết quả phân phối ở các câu (b) và (c) trên.
- Về các đồ thị phân tán thích hợp thể hiện các trung bình có điều kiện khác nhau.
- Nếu liên kết các trung bình có điều kiện trong câu (e), Anh / Chị thu được gì?
- Anh / Chị có nhận xét gì về mối quan hệ giữa năm kinh nghiệm và tuổi tác?

2.11 Xem xét xem các mô hình sau đây có tuyến tính theo các thông số hay các biến hay không, hay có cả hai. Mô hình nào trong số những mô hình sau là mô hình hồi qui tuyến tính?

Mô hình

Từ mô tả

- | | |
|---|-------------------------|
| a) $Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i} \right) + u_i$ | Nghịch đảo |
| b) $Y_i = \beta_1 + \beta_2 \ln X_i + u_i$ | Nửa logarit |
| c) $\ln Y_i = \beta_1 + \beta_2 X_i + u_i$ | Nửa logarit nghịch |
| d) $\ln Y_i = \ln \beta_1 + \beta_2 \ln X_i + u_i$ | Logarit hay logarit bội |
| e) $\ln Y_i = \beta_1 - \beta_2 \left(\frac{1}{X_i} \right) + u_i$ | Logarit nghịch đảo |

Chú ý: $\ln$ = logarit tự nhiên (có nghĩa là, log với cơ số e); u_i là số hạng nhiễu ngẫu nhiên. Chúng ta sẽ nghiên cứu những mô hình này ở chương 6.

2.12 Những mô hình sau đây có phải là những mô hình hồi qui tuyến tính không? Tại sao?

- $Y_i = e^{\beta_1 + \beta_2 X_i + u_i}$
- $Y_i = \frac{1}{1 + e^{\beta_1 + \beta_2 X_i + u_i}}$
- $\ln Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i} \right) + u_i$

- d) $Y_i = \beta_1 + (0.75 - \beta_1)e^{-\beta_2(X_i - 2)} + u_i$
 e) $Y_i = \beta_1 + \beta_2^3 X_i + u_i$

2.13 Nếu $\beta_2 = 0.8$ trong (d) của bài 2.12, vậy mô hình có trở thành một mô hình hồi quy tuyến tính không? Tại sao?

2.14 Xem xét những mô hình không ngẫu nhiên. Chúng có phải là mô hình tuyến tính không, có nghĩa là, những mô hình có tuyến tính theo thông số hay không? Nếu không, bằng các phép toán đại số thích hợp có thể chuyển chúng thành những mô hình tuyến tính hay không?

a) $Y_i = \frac{1}{\beta_1 + \beta_2 X_i}$

b) $Y_i = \frac{X_i}{\beta_1 + \beta_2 X_i}$

c) $Y_i = \frac{1}{1 + \exp(-\beta_1 - \beta_2 X_i)}$

2.15 Một biến ngẫu nhiên rời rạc X có phân phối đều hoặc tam giác (rời rạc) nếu PDF của nó có dạng sau:

$$f(X) = 1/k \text{ với } X = X_1, X_2, \dots, X_k \text{ } [X_i \neq X_j \text{ khi } i \neq j]$$

a) Chứng minh rằng đối với phân phối này $E(X) = \sum X_i (1/k)$ và phương sai

$$\sigma_X^2 = \sum [X_i - E(X)]^2 \cdot (1/k) \text{ trong đó } E(X) \text{ là giống ở trên.}$$

b) Nếu $X = 1, 2, \dots, k$ thì các giá trị của $E(X)$ và σ_X^2 bằng bao nhiêu?

2.16 Bảng dưới đây cung cấp dữ liệu về điểm Kiểm tra Năng khiếu Học đường (SAT) trung bình của những học sinh năm cuối sắp lên đại học trong 1967-1990.

- Dùng trục hoành cho năm và trục tung cho điểm SAT để vẽ hai đồ thị riêng biệt điểm toán và điểm văn đáp cho nam và nữ.
- Chúng ta có thể rút ra được những kết luận gì?
- Khi đã biết điểm văn đáp của nam và nữ, làm cách nào bạn có thể tiên đoán được điểm toán của họ?
- Vẽ đồ thị điểm SAT tổng cộng của nữ so với điểm SAT tổng cộng của nam. Vẽ đường hồi qui đi qua những điểm rời rạc này. Các Anh/Chị quan sát được gì?

Điểm Kiểm Tra Năng Khiếu Học Đường (SAT) Trung Bình Của Những Học Sinh Năm Cuối Sắp Lên Đại Học, 1967-1990*

Năm	Văn đáp			Toán		
	Nam	Nữ	Tổng cộng	Nam	Nữ	Tổng cộng
1967	463	468	466	514	467	492
1968	464	466	466	512	470	492
1969	459	466	463	513	470	492
1970	459	461	460	509	465	488
1971	454	457	455	507	466	488
1972	454	452	453	505	461	484
1973	446	443	445	502	460	431
1974	447	442	444	501	459	480
1975	437	431	434	495	449	472
1976	433	430	431	497	446	472
1977	431	427	429	497	445	470
1978	433	425	429	494	444	468
1979	431	423	427	493	443	467
1980	428	420	424	491	443	466
1981	430	418	424	492	443	466
1982	431	421	426	493	443	467
1983	430	420	425	493	445	468
1984	433	420	426	495	449	471
1985	437	425	431	499	452	475
1986	437	426	431	501	451	475
1987	435	425	430	500	453	476
1988	435	422	428	498	455	476
1989	434	421	427	500	454	476
1990	429	419	424	499	455	476

* Dữ liệu cho 1967-1971 là những số ước lượng

Source: The College Board. The New York Times, Aug. 28, 1990, p.B-5.

2.17 Đường hồi quy trong hình 1.3 của Phần Giới thiệu có là đường PRF hay SRF? Tại sao? Các Anh/Chị giải thích các điểm rời rạc nằm quanh đường hồi quy như thế nào? Ngoài GDP, còn có các yếu tố nào, hay các biến nào, có thể quyết định đến chi tiêu tiêu dùng của cá nhân?

CHƯƠNG 3

MÔ HÌNH HỒI QUY HAI BIẾN: VẤN ĐỀ ƯỚC LƯỢNG

Như đã lưu ý ở Chương 2, nhiệm vụ đầu tiên của chúng ta là ước lượng chính xác tối đa hàm hồi quy tổng thể (PRF) trên cơ sở hàm hồi quy mẫu (SRF). Có nhiều phương pháp xây dựng hàm SRF, nhưng cho đến nay, liên quan tới quá trình phân tích hồi quy, **phương pháp bình phương tối thiểu thông thường (OLS)**¹² là phương pháp được sử dụng nhiều và phổ biến nhất. Trong chương này, ta sẽ thảo luận về phương pháp này cho mô hình hồi quy hai biến. Sau đó, ở Chương 7, ta sẽ xem xét sự tổng quát hoá của phương pháp này cho các mô hình hồi quy đa biến.

3.1. PHƯƠNG PHÁP BÌNH PHƯƠNG TỐI THIỂU THÔNG THƯỜNG:

Phương pháp bình phương tối thiểu thông thường do Carl Friedrich Gauss, nhà toán học người Đức đưa ra. Dựa trên các giả thiết nhất định (được thảo luận ở Phần 3.2), phương pháp bình phương tối thiểu có một số tính chất thống kê rất hấp dẫn đã làm cho nó trở thành phương pháp phân tích hồi quy mạnh nhất và phổ biến nhất. Để hiểu phương pháp này, trước tiên ta phải giải thích nguyên tắc bình phương tối thiểu.

Ta nhắc lại hàm PRF hai biến:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (2.4.2)$$

Tuy nhiên như đã lưu ý trong Chương 2, hàm PRF không thể quan sát trực tiếp được. Ta ước lượng nó từ hàm SRF:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \quad (2.6.2)$$

$$= \hat{Y}_i + \hat{u}_i \quad (2.6.3)$$

trong đó $\hat{Y}_i$ là giá trị ước lượng (giá trị trung bình có điều kiện) của Y_i .

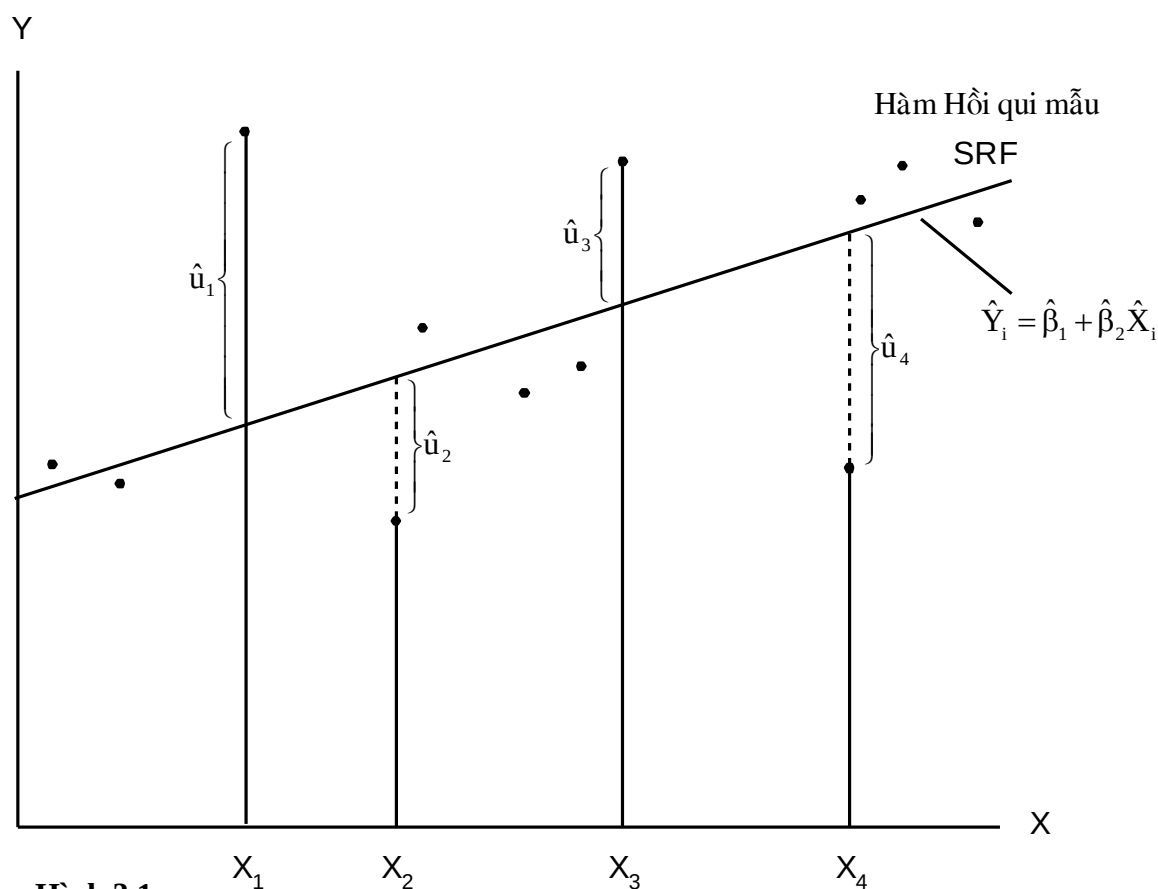
Nhưng ta sẽ xác định chính hàm SRF như thế nào? Để thấy được điều này, ta hãy tiến hành như sau. Đầu tiên, ta biểu thị (2.6.3) thành :

$$\begin{aligned} \hat{u}_i &= Y_i - \hat{Y}_i \\ &= Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \end{aligned} \quad (3.1.1)$$

biểu thức đó chỉ rằng, $\hat{u}_i$ (các phần dư) chỉ đơn giản là chênh lệch giữa các giá trị thực và giá trị ước lượng của Y .

Bây giờ, cho n cặp quan sát của X và Y , ta muốn xác định hàm SRF bằng cách nào đó để nó gần nhất với giá trị thực của Y . Để đạt được đích này, ta có thể chọn tiêu chuẩn sau đây: chọn hàm SRF sao cho tổng các phần dư $\sum \hat{u}_i = \sum (Y_i - \hat{Y}_i)$ là càng nhỏ càng tốt. Tuy nhiên, mặc dù hấp dẫn về trực giác, đây không phải là tiêu chuẩn tốt lắm, như có thể thấy trên đồ thị phân tán giả thiết (hình 3.1).

¹² Một phương pháp khác, được biết gọi là “Phương pháp thích hợp tối đa” sẽ được xem xét ngắn gọn trong Chương 4.



Hình 3.1
Tiêu chuẩn bình

phương tối thiểu

Nếu ta chấp nhận điều kiện cực tiểu của tổng $\sum \hat{u}_i$, hình 3.1 cho thấy rằng các phần dư $\hat{u}_2$ và $\hat{u}_3$ cũng như các phần dư $\hat{u}_1$ và $\hat{u}_4$ có cùng trọng số trong tổng $(\hat{u}_1 + \hat{u}_2 + \hat{u}_3 + \hat{u}_4)$, mặc dầu hai phần dư đầu gần hàm SRF hơn nhiều so với hai phần dư sau. Nói cách khác, tất cả các phần dư đều có vai trò quan trọng như nhau, bất kể các quan sát riêng biệt có gần hay phân tán rộng tới đâu so với hàm SRF. Hậu quả của điều này là hoàn toàn có khả năng là tổng đại số của $\hat{u}_i$ rất nhỏ (thậm chí bằng 0) mặc dù các $\hat{u}_i$ được phân tán rộng xung quanh hàm SRF. Để thấy được điều này, ta hãy cho rằng $\hat{u}_1, \hat{u}_2, \hat{u}_3, \hat{u}_4$ trên hình 3.1 có các giá trị tương ứng bằng 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 rộng hơn xung quanh hàm SRF so với $\hat{u}_2$ và $\hat{u}_3$. Chúng ta có thể tránh được vấn đề này nếu ta chấp nhận *tiêu chuẩn bình phương tối thiểu*, nó khẳng định rằng hàm SRF có thể được cố định theo cách để

$$\begin{aligned} \sum \hat{u}_i^2 &= \sum (Y_i - \hat{Y}_i)^2 \\ &= \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \end{aligned} \quad (3.1.2)$$

càng nhỏ càng tốt, trong đó $\hat{u}_i^2$ là bình phương của các phần dư. Bằng cách bình phương $\hat{u}_i$, phương pháp này sẽ cho các phần dư $\hat{u}_1$ và $\hat{u}_4$ trên hình 3.1 một trọng số lớn hơn phần dư $\hat{u}_2$ và $\hat{u}_3$. Như đã lưu ý trước đây, với tiêu chuẩn giá trị cực tiểu của $\sum \hat{u}_i$, tổng này có thể nhỏ ngay khi $\hat{u}_i$ phân tán rộng xung quanh hàm SRF. Tuy nhiên điều này không thể xảy ra với quy trình

bình phương tối thiểu, vì $\hat{u}_i$ càng lớn (về giá trị tuyệt đối) thì $\sum \hat{u}_i^2$ càng lớn. Một minh chứng tiếp theo cho phương pháp bình phương tối thiểu nằm trong thực tế là các hàm ước lượng thu được từ phương pháp này có một số tính chất thống kê rất đúng như mong muốn, như ta sẽ thấy ngay sau đây.

Rõ ràng từ (3.1.2) ta có

$$\sum \hat{u}_i^2 = f(\hat{\beta}_1, \hat{\beta}_2) \quad (3.1.3)$$

nghĩa là tổng các bình phương phần dư là một hàm nào đó của các hàm ước lượng $\hat{\beta}_1$ và $\hat{\beta}_2$. Với một bộ dữ liệu cho trước bất kỳ, việc chọn các giá trị khác nhau cho $\hat{\beta}_1$ và $\hat{\beta}_2$ sẽ cho các giá trị khác nhau của $\sum \hat{u}_i^2$ và do đó dẫn tới các giá trị khác nhau của $\sum \hat{u}_i^2$. Để thấy rõ điều này, hãy xét các dữ liệu giả thiết của Y và X cho trong 2 cột đầu của Bảng 3.1. Ta hãy thực hiện hai thử nghiệm. Trong thử nghiệm 1, cho $\hat{\beta}_1 = 1.572$ và $\hat{\beta}_2 = 1.357$ (ngay lúc này đừng lo lắng về việc làm thế nào ta thu được các giá trị này, coi như chỉ là dự đoán)¹³. Sử dụng các giá trị này của $\hat{\beta}$ và các giá trị của X cho trong cột (2) của Bảng 3.1, ta có thể dễ dàng tính ra giá trị ước lượng $\hat{Y}_i$ của $\hat{Y}_{1i}$ như là các giá trị Y_i đã cho trong cột (3) của bảng này (chỉ số 1 ký hiệu cho thử nghiệm 1). Bây giờ, chúng ta hãy thực hiện thử nghiệm 2, nhưng lần này, ta sử dụng giá trị $\hat{\beta}_1 = 3$ và $\hat{\beta}_2 = 1$. Các giá trị ước lượng của Y_i từ thử nghiệm này được cho như $\hat{Y}_{2i}$ trong cột (6) của Bảng 3.1. Vì các giá trị $\hat{\beta}$ trong hai thử nghiệm là khác nhau, ta thu được các giá trị khác nhau cho các phần dư ước lượng, như trong bảng; $\hat{u}_{1i}$ là các phần dư từ thử nghiệm đầu và $\hat{u}_{2i}$ là các phần dư từ thử nghiệm thứ 2. Các bình phương của các phần dư này được cho trong cột (5) và (8). Rõ ràng, như đã kỳ vọng từ (3.1.3), các tổng phần dư bình phương này sẽ khác nhau vì chúng dựa trên các giá trị $\hat{\beta}$ khác nhau.

Bảng 3.1
Thông số thử nghiệm của hàm SRF

	Y_i	X_i	$\hat{Y}_{1i}$	$\hat{u}_{1i}$	$\hat{u}_{1i}^2$	$\hat{Y}_{2i}$	$\hat{u}_{2i}$	$\hat{u}_{2i}^2$
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	4	1	2,929	1,071	1,147	4	0	0
	5	4	7,000	-2,000	4,000	7	-2	4
	7	5	8,357	-1,357	1,841	8	-1	1
	12	6	9,714	2,286	5,226	9	3	9
Cộng:	28	16		0,0	12,214		0	14
Chú ý	$\hat{Y}_{1i} = 1.572 + 1.357 X_i$ (với $\beta_1=1.572$ và $\beta_2 = 1.357$)							
	$\hat{Y}_{2i} = 3.0 + 1.0 X_i$ (với $\beta_1=3$ và $\beta_2 = 1.0$)							
	$\hat{u}_{1i} = (Y_i - \hat{Y}_{1i})$							
	$\hat{u}_{2i} = (Y_i - \hat{Y}_{2i})$							

¹³ Để thoả mãn tính toán, các giá trị này thu được từ phương pháp bình phương tối thiểu, được nói đến một cách ngắn gọn. Xem các phương trình (3.1.6) và (3.1.7)

Bây giờ, ta nên chọn bộ giá trị $\hat{\beta}$ nào đây? Vì các giá trị $\hat{\beta}$ của thử nghiệm thứ 1 cho ta $\sum \hat{u}_i^2$ (=12,214) thấp hơn là ở thử nghiệm thứ 2 (=14), ta có thể nói rằng các $\hat{\beta}$ của thử nghiệm thứ 1 là các giá trị “tốt nhất”. Nhưng làm thế nào ta biết? Bởi vì, nếu có được thời gian và lòng kiên nhẫn vô hạn, ta đã có thể làm thêm nhiều thử nghiệm như thế, bằng cách chọn các bộ $\hat{\beta}$ khác nhau mỗi lần và so sánh kết quả $\sum \hat{u}_i^2$, rồi cuối cùng lọc ra bộ giá trị $\hat{\beta}$ cho ta giá trị $\sum \hat{u}_i^2$ nhỏ nhất có thể, giả định rằng ta đã xem xét tất cả các giá trị có thể tính tới được của β_1 và β_2 . Tuy nhiên, vì thời gian và cả lòng kiên nhẫn của con người nói chung đều hiếm hoi, ta cần xem xét một số đường tắt đi tới quá trình thử-và-sai này. May mắn là phương pháp bình phương tối thiểu cho ta cách làm tắt này. Nguyên tắc này hay là phương pháp bình phương tối thiểu chọn $\hat{\beta}_1$ và $\hat{\beta}_2$ theo cách để với một mẫu hoặc bộ dữ liệu đã cho $\sum \hat{u}_i^2$ càng nhỏ càng tốt. Nói cách khác, đối với một mẫu cho trước, phương pháp bình phương tối thiểu cho ta các giá trị ước lượng duy nhất của β_1 và β_2 , các giá trị này cho giá trị nhỏ nhất có thể có được của $\sum \hat{u}_i^2$. Công việc này được thực hiện như thế nào? Đây chỉ là một bài tập đơn giản trong toán giải tích. Như đã nói ở Phụ lục 3A, Phần 3A.1, quá trình vi phân cho các phương trình sau để ước lượng β_1 và β_2 :

$$\sum Y_i = n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \quad (3.1.4)$$

$$\sum Y_i X_i = \hat{\beta}_1 \sum X_i + \hat{\beta}_2 \sum X_i^2 \quad (3.1.5)$$

trong đó n là cỡ mẫu. Phương trình này được gọi là các **phương trình chuẩn**.

Giải hệ phương trình chuẩn này, ta thu được:

$$\begin{aligned} \hat{\beta}_2 &= \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\ &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum x_i y_i}{\sum x_i^2} \end{aligned} \quad (3.1.6)$$

trong đó $\bar{X}$ và $\bar{Y}$ là các trung bình mẫu của X và Y và trong đó ta định nghĩa $x_i = X_i - \bar{X}$ và $y_i = Y_i - \bar{Y}$. Từ bây giờ trở về sau, ta chọn quy ước đặt chữ cái viết thường để biểu thị độ lệch khỏi các giá trị trung bình.

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum X_i^2 \sum Y_i - \sum X_i \sum X_i Y_i}{n \sum X_i^2 - (\sum X_i)^2} \\ &= \bar{Y} - \hat{\beta}_2 \bar{X} \end{aligned} \quad (3.1.7)$$

Bước cuối cùng trong (3.1.7) có thể thu được trực tiếp từ (3.1.4) bằng vài biến đổi đại số đơn giản.

Nhân đây, lưu ý rằng, bằng cách dùng các đồng nhất thức đại số đơn giản, công thức (3.1.6) để ước lượng β_2 có thể biểu thị theo cách khác như là:

$$\begin{aligned}\hat{\beta}_2 &= \frac{\sum x_i y_i}{\sum x_i^2} \\ &= \frac{\sum x_i Y_i}{\sum X_i^2 - n\bar{X}^2} \\ &= \frac{\sum X_i y_i}{\sum X_i^2 - n\bar{X}^2}\end{aligned}\quad (3.1.8)^{14}$$

nó có thể giảm gánh nặng tính toán cho những ai sử dụng máy tính tay để giải quyết một bài toán hồi quy với một bộ dữ liệu nhỏ.

Hàm ước lượng thu được trên đây gọi là **các hàm ước lượng bình phương tối thiểu**, vì chúng được xác định từ các nguyên tắc bình phương tối thiểu. Lưu ý rằng **các tính chất bằng số** sau đây của các hàm ước lượng thu được từ phương pháp bình phương tối thiểu thông thường: “Các tính chất bằng số là các tính chất thể hiện như là hệ quả của việc dùng bình phương tối thiểu thông thường, bất kể dữ liệu được tạo ra như thế nào.”¹⁵ Nói ngắn gọn, ta cũng sẽ xem xét **các tính chất thống kê** của các hàm ước lượng bình phương tối thiểu thông thường, tức là, các tính chất “có được khi có các giả định nào đó về các dữ liệu đã được tạo nên.”¹⁶ (Xem mô hình hồi quy tuyến tính cổ điển ở Phần 3.2).

- I. Các hàm ước lượng *bình phương tối thiểu thông thường* OLS được biểu thị duy nhất dưới dạng các số lượng (nghĩa là X và Y) có thể quan sát được (nghĩa là mẫu). Do đó chúng có thể tính được dễ dàng.
- II. Chúng là các **hàm ước lượng điểm**, nghĩa là nếu cho trước một mẫu mỗi hàm ước lượng sẽ chỉ cho một giá trị đơn lẻ (điểm) của thông số tổng thể phù hợp. (Trong Chương 5, ta sẽ xét cái gọi là các **hàm ước lượng khoảng**, chúng cung cấp một khoảng các giá trị có thể có đối với các thông số tổng thể chưa biết).

III. Một khi đã thu được các ước lượng *bình phương tối thiểu thông thường* OLS từ dữ liệu mẫu, ta có thể dễ dàng vẽ được *đường hồi quy mẫu*. Đường hồi quy thu được như vậy có các tính chất sau:

1. Nó đi qua các giá trị *trung bình mẫu* của Y và X . Thực tế này có thể được thấy rõ từ (3.1.7), đối với dòng sau có thể viết thành $\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X}$, biểu thức này được mô tả bằng đồ thị trong hình 3.2.
2. Giá trị trung bình của ước lượng $Y = \hat{Y}_i$ bằng giá trị trung bình của Y thực đối với

¹⁴ Lưu ý 1: $\sum x_i^2 = \sum (X_i - \bar{X})^2 = \sum X_i^2 - 2\sum X_i \bar{X} + \sum \bar{X}^2 = \sum X_i^2 - 2\bar{X} \sum X_i + \sum \bar{X}^2$, vì $\bar{X}$ là hằng số. Sau đó lưu ý rằng $\sum X_i = n\bar{X}$ và $\sum \bar{X}^2 = n\bar{X}^2$ với $\bar{X}$ là một hằng số, chúng ta thu được $\sum x_i^2 = \sum X_i^2 - n\bar{X}^2$.

Lưu ý 2: $\sum x_i y_i = \sum x_i (Y_i - \bar{Y}) = \sum x_i Y_i - \bar{Y} \sum x_i = \sum x_i Y_i - \bar{Y} \sum (X_i - \bar{X}) = \sum x_i Y_i$ vì $\bar{Y}$ là một hằng số và vì tổng các độ lệch của các biến so với các giá trị trung bình [ví dụ $\sum (X_i - \bar{X})$] luôn luôn bằng 0. Nghĩa là, $\sum y_i = \sum (Y_i - \bar{Y}) = 0$.

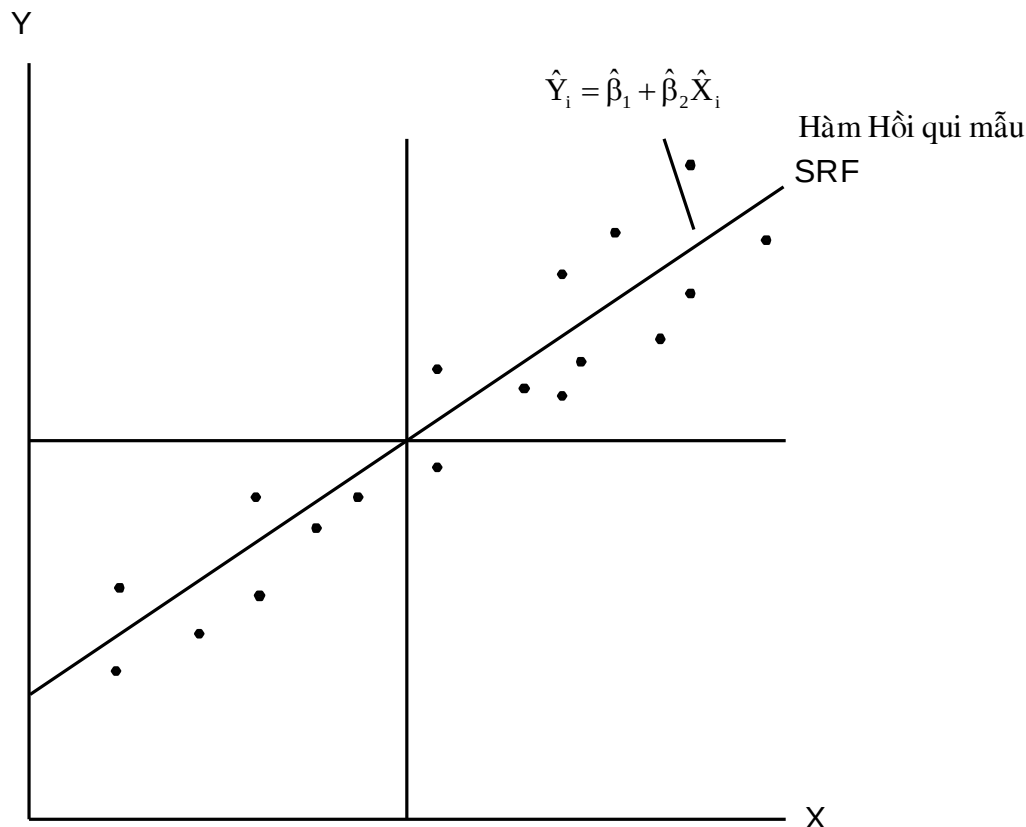
¹⁵ Cuốn *Estimation and Inference in Econometrics* của Russell Davidson và James G. MacKinnon, nhà xuất bản Oxford University Press, New York, 1993, trang 3.

¹⁶ Như sách trên

$$\begin{aligned}
 \hat{Y}_i &= \hat{\beta}_1 + \hat{\beta}_2 X_i \\
 &= (\bar{Y} - \hat{\beta}_2 \bar{X}) + \hat{\beta}_2 X_i \\
 &= \bar{Y} + \hat{\beta}_2 (X_i - \bar{X})
 \end{aligned}
 \tag{3.1.9}$$

Lấy tổng hai vế của đẳng thức cuối cùng đối với các giá trị mẫu rồi chia cho cỡ mẫu n , cho ta:

$$\bar{\hat{Y}} = \bar{Y} \tag{3.1.10}^{17}$$



trong đó ứng dụng được lập ra bởi thực tế: $\sum (X_i - \bar{X}) = 0$ (Tại sao?)

Hình 3.2

Đồ thị cho thấy đường hồi qui mẫu xuyên qua các giá trị trung bình mẫu của X và Y

3. Giá trị trung bình của các phần dư $\hat{u}_i$ bằng 0. Từ phụ lục 3A, Phần 3A.1, phương trình đầu tiên là:

$$-2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = 0$$

¹⁷ Lưu ý: Kết quả này chỉ đúng khi mô hình hồi quy có số hạng tung độ gốc β_1 trong đó. Như phụ lục 6A, Phần 6A.1, kết quả này không áp dụng khi thiếu β_1 trong mô hình

Nhưng vì $\hat{u}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$, phương trình trên giảm xuống còn $-2 \sum \hat{u}_i = 0$, khi $\bar{\hat{u}} = 0$.¹⁸

Do tính chất trên, hồi quy mẫu:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{u}_i \quad (2.6.2)$$

có thể biểu diễn theo một dạng khác thay thế trong đó cả Y và X đều được biểu thị như là các độ lệch từ các giá trị trung bình của chúng. Để thấy điều này, ta lấy tổng (2.6.2) cho cả 2 vế để có:

$$\begin{aligned} \sum Y_i &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i + \sum \hat{u}_i \\ &= n\hat{\beta}_1 + \hat{\beta}_2 \sum X_i \quad \text{vì} \quad \sum \hat{u}_i = 0 \end{aligned} \quad (3.1.11)$$

Chia phương trình (3.1.11) cho n, ta có:

$$\bar{Y} = \hat{\beta}_1 + \hat{\beta}_2 \bar{X} \quad (3.1.12)$$

biểu thức này cũng giống như (3.1.7). Lấy phương trình (2.6.2) trừ đi (3.1.12), ta có:

$$Y_i - \bar{Y} = \hat{\beta}_2 (X_i - \bar{X}) + \hat{u}_i$$

hoặc

$$y_i = \hat{\beta}_2 x_i + \hat{u}_i \quad (3.1.13)$$

trong đó y_i và x_i , theo quy ước của chúng ta, là độ lệch từ các giá trị trung bình tương ứng (mẫu) của chúng.

Phương trình (3.1.13) được biết như là **dạng độ lệch**. Lưu ý rằng số hạng tung độ gốc $\hat{\beta}_1$ không còn có mặt trong phương trình đó. Nhưng số hạng tung độ gốc luôn có thể được ước lượng bởi (3.1.7), nghĩa là, từ thực tế rằng đường hồi quy mẫu đi qua các trung bình mẫu của Y và X. Một ưu điểm của dạng độ lệch là nó luôn đơn giản hoá các phép tính số học khi phải làm việc trên máy tính bàn. Tuy nhiên trong kỷ nguyên thông tin này, lợi điểm này trở nên thứ yếu.

Nhân đây, xin lưu ý rằng trong dạng độ lệch, hàm SRF có thể được viết như là:

$$\hat{y}_i = \hat{\beta}_2 x_i \quad (3.1.14)$$

trong khi nó chính là $\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i$ trong các đơn vị đo lường chính gốc, như thấy ở (2.6.1).

4. Các phần dư $\hat{u}_i$ là không tương quan với giá trị dự báo $\hat{Y}_i$. Có thể kiểm chứng điều này như sau, sử dụng bằng cách dạng độ lệch, ta có thể viết:

$$\begin{aligned} \sum y_i u_i &= \hat{\beta}_2 \sum x_i u_i \\ &= \hat{\beta}_2 \sum x_i (y_i - \hat{\beta}_2 x_i) \\ &= \hat{\beta}_2 \sum x_i y_i - \hat{\beta}_2^2 \sum x_i^2 \\ &= \hat{\beta}_2^2 \sum x_i^2 - \hat{\beta}_2^2 \sum x_i^2 \end{aligned}$$

¹⁸ Kết quả này cũng đòi hỏi số hạng tung độ gốc β_1 phải có mặt trong mô hình (xem phụ lục 6A, Phần 6A.1)

$$= 0 \quad (3.1.15)$$

trong đó ứng dụng được lập ra bởi thực tế $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$.

5. Các phần dư $\hat{u}_i$ là không tương quan với X_i , nghĩa là $\sum \hat{u}_i X_i = 0$. Điều này tiếp theo từ phương trình (2) trong phụ lục 3A, Phần 3A.1.

3.2. MÔ HÌNH HỒI QUY TUYẾN TÍNH CỔ ĐIỂN: GIẢ THIẾT CƠ SỞ CỦA PHƯƠNG PHÁP BÌNH PHƯƠNG TỐI THIỂU

Nếu như mục đích của chúng ta chỉ là ước lượng β_1 và β_2 thì *phương pháp bình phương tối thiểu* OLS đã thảo luận ở phần trên là quá đủ. Nhưng xin được nhắc lại Chương 2, rằng trong phân tích hồi quy, mục đích của chúng ta không chỉ dừng ở việc tính được $\hat{\beta}_1$ và $\hat{\beta}_2$ mà còn phải rút ra kết luận giá trị thực của β_1 và β_2 . Ví dụ, ta muốn biết $\hat{\beta}_1$ và $\hat{\beta}_2$ gần như thế nào đối với thành phần tương ứng của chúng trong tổng thể hoặc là $\hat{Y}_i$ gần như thế nào tới giá trị thực $E(Y|X_i)$. Để trả lời các câu hỏi đó, chúng ta không chỉ phải định được dạng hàm số của phương trình, như trong (2.4.2), mà còn phải đưa ra các giả thiết chắc chắn về cách thức Y_i được sinh ra. Để hiểu vì sao đòi hỏi này là cần thiết, hãy nhìn vào hàm PRF: $Y_i = \beta_1 + \beta_2 X_i + \hat{u}_i$. Nó cho thấy rằng Y_i phụ thuộc vào cả X_i và u_i . Do đó, trừ phi chỉ rõ được X_i và u_i được tạo ra như thế nào, ta không có cách nào để suy diễn thống kê về Y_i , và như ta sẽ thấy, cũng không thể làm được điều đó về β_1 và β_2 . Do đó, giả thiết đưa ra về các biến X_i và số hạng sai số là tối hạn trong cách giải thích hiệu lực của phép ước lượng hồi quy.

Mô hình hồi quy tuyến tính cổ điển hay mô hình chuẩn, mô hình Gauss (CLRM) được coi là nền tảng của hầu hết lý thuyết kinh tế lượng, nó đưa ra 10 giả thiết¹⁹. Đầu tiên, ta hãy thảo luận các giả thiết này cho trường hợp mô hình hồi quy hai biến, và trong Chương 7 ta sẽ mở rộng chúng ra mô hình hồi quy đa biến, nghĩa là mô hình có nhiều hơn một biến hồi qui độc lập:

Giả thiết 1: Mô hình hồi quy tuyến tính. Mô hình hồi quy là **tuyến tính theo các thông số**, như được thấy ở (2.4.2)

$$Y_i = \beta_1 + \beta_2 X_i + \hat{u}_i \quad (2.4.2)$$

Ta đã thảo luận mô hình (2.4.2) trong Chương 2. Vì các mô hình hồi quy tuyến tính trong các thông số là khởi điểm cho CLRM, chúng ta sẽ duy trì giả thiết này trong suốt quyển sách. Hãy nhớ rằng biến hồi qui phụ thuộc Y và biến hồi qui độc lập X tự chúng có thể không tuyến tính, như đã đề cập ở Chương 2.²⁰

¹⁹ Nó được coi là cổ điển theo cảm giác vì được phát triển lần đầu tiên bởi Gauss vào năm 1821 và từ đó được coi là một khuôn mẫu hay tiêu chuẩn mà có thể được so sánh với các mô hình hồi quy không thỏa mãn các giả thiết Gauss.

²⁰ Nói vậy không có nghĩa là mô hình hồi quy không tuyến tính theo các thông số là không quan trọng hay ít được sử dụng. Tuy nhiên, việc đề cập đến các mô hình như thế đòi hỏi một số kiến thức toán học và thống kê nằm ngoài phạm vi của cuốn sách. Để thảo luận sâu sắc về các mô hình phi tuyến tính theo thông số, có thể tham khảo cuốn *Estimation and Inference in Econometrics* (Ước lượng và suy diễn thống kê trong kinh tế lượng) của tác giả Russell

Giả thiết 2: Các giá trị X được cố định trong việc lấy mẫu lặp lại. Các giá trị rút ra bởi biến hồi qui độc lập X được coi là cố định trong các mẫu lặp lại. Nói rõ hơn, X được giả thiết là *không ngẫu nhiên*.

Giả thiết này đã ngụ ý trong phần thảo luận của ta về hàm PRF ở Chương 2. Nhưng điều rất quan trọng đối với ta là hiểu được khái niệm về “các giá trị cố định trong việc lấy mẫu lặp lại”, nó được giải thích dưới dạng ví dụ đã cho ở Bảng 2.1. Xét các tổng thể Y khác nhau tương ứng với mức thu nhập được trình bày trong bảng đó. Giữ cho giá trị thu nhập X cố định và giả sử bằng \$80, ta rút ra một cách ngẫu nhiên một gia đình ngẫu nhiên nào đó và quan sát chi tiêu hàng tuần Y của gia đình đó, giả sử là \$60. Vẫn giữ X ở mức \$80, ta lại rút một cách ngẫu nhiên một gia đình khác và thấy giá trị quan sát Y của nó là \$75. Trong mỗi lần rút ra một gia đình để xem xét (nghĩa là lấy mẫu lặp lại), giá trị X được cố định ở mức \$80. Ta có thể lặp lại quá trình này cho tất cả các giá trị X đã ghi trong Bảng 2.1. Thực ra, dữ liệu mẫu ghi trên bảng 2.4 và 2.5 đều được rút ra theo cách này.

Tất cả những điều này có nghĩa là sự phân tích hồi quy của ta là **phân tích hồi quy có điều kiện**, nghĩa là có điều kiện với các giá trị đã cho của (các) biến hồi qui độc lập X .

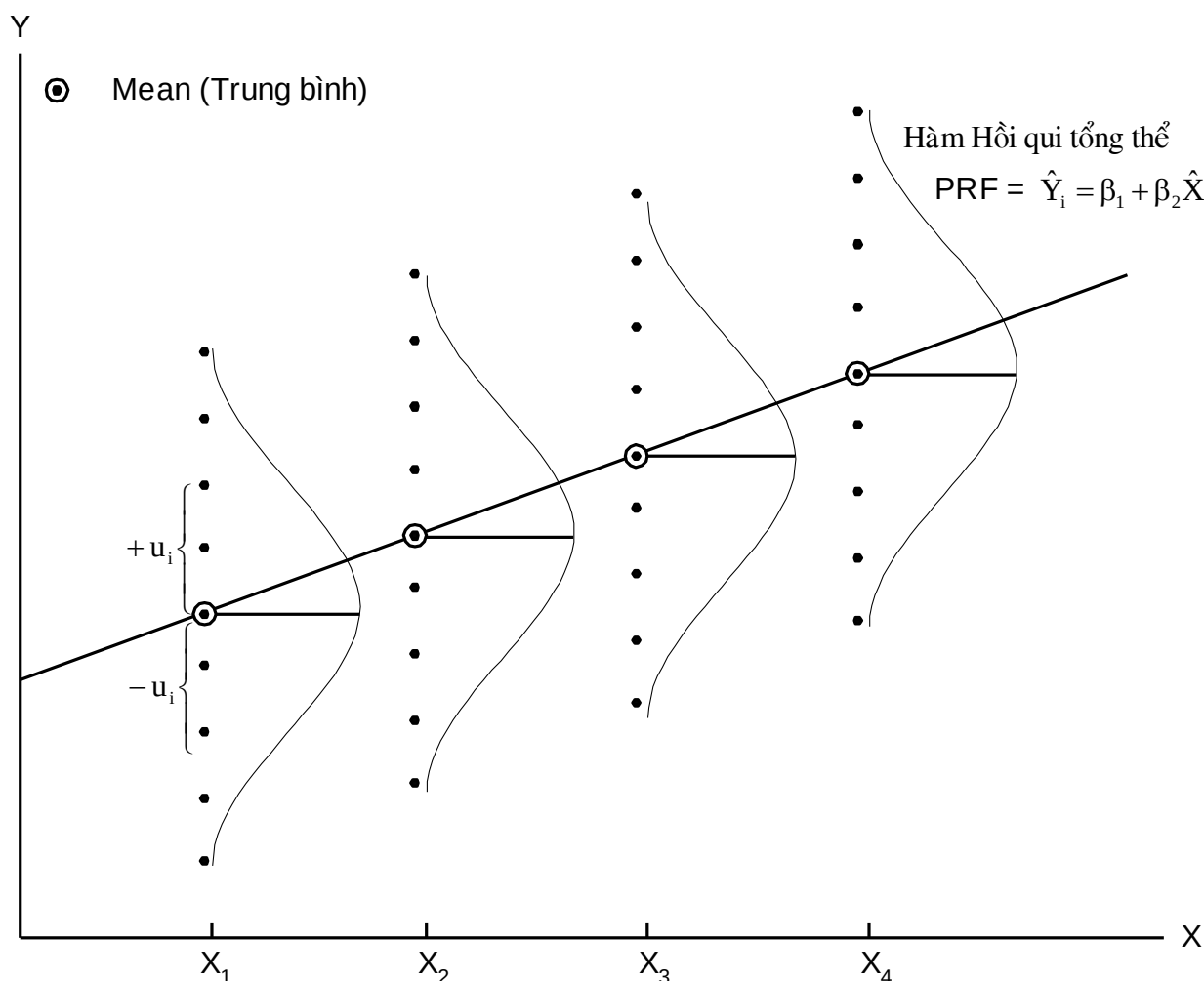
Giả thiết 3: Giá trị trung bình bằng không của các nhiễu u_i . Cho trước giá trị của X , giá trị trung bình hay kỳ vọng của các số hạng nhiễu u_i bằng 0. Nói rõ hơn, giá trị trung bình có điều kiện của u_i là 0. Về mặt ký hiệu, ta có:

$$E(u_i | X_i) = 0 \quad (3.2.1)$$

Giả thiết 3 cho rằng, giá trị trung bình của u_i có điều kiện theo với X_i đã cho, là bằng 0. Bằng hình học, giả thiết này có thể được vẽ trên hình 3.3, nó chỉ ra một vài giá trị của biến X và tổng thể Y liên kết với chúng. Như đã thấy, mỗi một tổng thể Y tương ứng với một X cho trước được phân phối xung quanh giá trị trung bình của nó (có thể thấy được nhờ những chấm được khoanh tròn trên PRF) cùng với một vài giá trị Y ở phía trên và dưới nó. Khoảng cách phía trên và dưới đối với giá trị trung bình không là gì nhưng u_i và cái mà (3.2.1) đòi hỏi là giá trị trung bình của các độ lệch này tương ứng với bất kỳ X đã cho phải bằng 0²¹.

Davidson và James MacKinnon, NXB Oxford University Press, New York, 1993. Cuốn sách không dành cho người mới bắt đầu.

²¹ Để minh họa, ta chỉ coi rằng các u được phân bố đối xứng như đã chỉ trên hình 3.3. Nhưng trong Chương 4 ta sẽ coi rằng các u được phân phối chuẩn.



Hình 3.3
 Phân bố có điều kiện của nhiều u_i

Từ cách nhìn nhận của những gì đã thảo luận Phần 2.4 (xem phương trình 2.4.5), giả thiết này không có gì là khó hiểu. Tất cả những gì mà giả thiết này khẳng định là các yếu tố không bao gồm rõ rệt trong mô hình và do đó sẽ được kể vào trong u_i , không ảnh hưởng một cách có hệ thống đến giá trị trung bình của Y ; cho nên, có thể nói, các giá trị u_i dương triệt tiêu các giá trị u_i âm sao cho trung bình của chúng ảnh hưởng lên Y bằng 0.²²

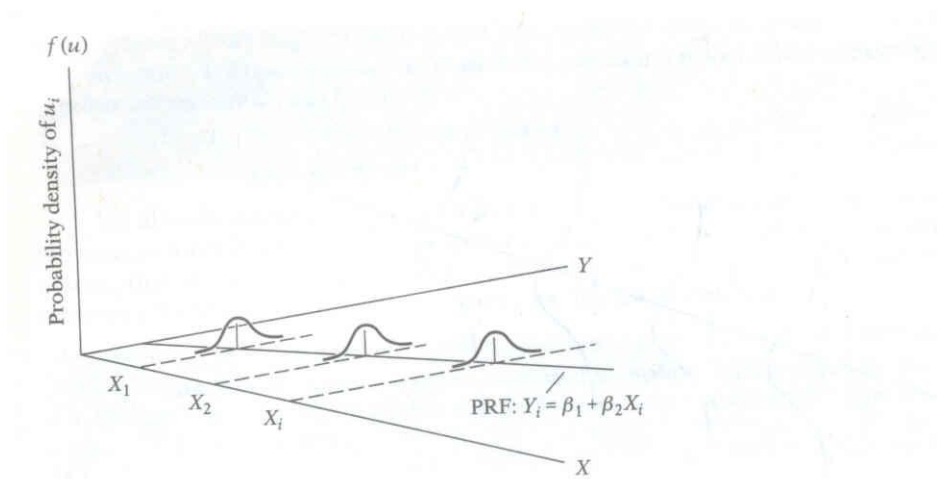
Nhân đây, lưu ý rằng giả thiết $E(u_i|X_i) = 0$ ngụ ý rằng $E(Y_i|X_i) = \beta_1 + \beta_2 X_i$. (Tại sao?). Do đó, hai giả thiết này là tương đương nhau.

Giả thiết 4: Phương sai có điều kiện không đổi hay phương sai bằng nhau của u_i . Cho các giá trị của X , phương sai của u_i sẽ như nhau đối với tất cả mọi quan sát. Nghĩa là, các phương sai có điều kiện của u_i đều đồng nhất. Về mặt ký hiệu, ta có:

²² Để hiểu thêm vì sao mà giả thiết 3 là cần thiết có thể đọc *Statistical Methods of Econometrics* (Phương pháp thống kê của kinh tế lượng của E.Malinvaud, NXB Rand McNally, 1996, trang 75. Xem thêm bài tập 3.3

$$\begin{aligned}
 \text{var}(u_i|X_i) &= E[u_i - E(u_i)|X_i]^2 \\
 &= E(u_i^2|X_i) \text{ (do giả thiết 3)} \\
 &= \sigma^2
 \end{aligned}
 \tag{3.2.2}$$

trong đó **var** là phương sai.



Hình 3.4

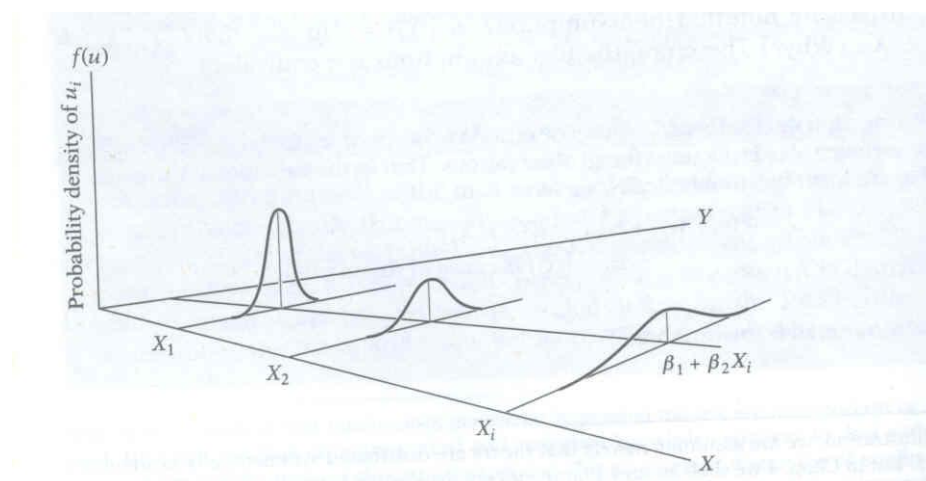
Phương sai có điều kiện không đổi

Phương trình (3.2.2) khẳng định phương sai của u_i cho mỗi X_i (nghĩa là, phương sai điều kiện của u_i) là một hằng số dương nào đó bằng σ^2 . Một cách kỹ thuật, phương trình (3.2.2) thể hiện giả thiết về **phương sai có điều kiện không đổi**, hay là *đẳng truyền*, hay là *phương sai bằng nhau*. Nói cách khác, (3.2.2) có nghĩa là các tổng thể Y tương ứng với các giá trị X khác nhau sẽ có phương sai như nhau. Về mặt đồ thị, điều này được mô tả trên hình 3.4.

Ngược lại, hãy xét hình 3.5, trong đó phương sai điều kiện của các tổng thể Y biến thiên đối với X . Người ta gọi hiện tượng này một cách gần đúng là **phương sai của sai số thay đổi** hay là *sự truyền bất đẳng*, hoặc là *phương sai*. Về mặt ký hiệu, trong trường hợp này, (3.2.2) có thể viết thành

$$\text{var}(u_i|X_i) = \sigma_i^2 \tag{3.2.3}$$

Lưu ý chỉ số của σ^2 trong phương trình (3.2.3), nó chỉ rõ rằng phương sai của tổng thể Y đã không còn là một hằng số.



Hình 3.5
Phương sai của sai số thay đổi

Để làm rõ hơn sự khác biệt giữa hai trường hợp trên, hãy gọi Y là mức chi tiêu tiêu dùng hàng tuần và X là thu nhập hàng tuần. Hình 3.4 và 3.5 cho thấy khi thu nhập tăng thì chi tiêu tiêu dùng trung bình cũng tăng. Nhưng trên hình 3.4, phương sai của mức chi tiêu tiêu dùng giữ nguyên tại tất cả các mức thu nhập, trong khi đó ở hình 3.5 phương sai lại tăng khi mức thu nhập tăng. Nói cách khác, mức chi phí trung bình của các gia đình giàu hơn thì lớn hơn là mức chi phí của các gia đình nghèo hơn, nhưng cũng có biến thiên lớn hơn trong mức chi tiêu tiêu dùng của gia đình giàu.

Để hiểu được lý do căn bản đằng sau giả thiết này, ta hãy tham khảo hình 3.5 theo đó $\text{var}(u | X_1) < \text{var}(u | X_2), \dots, < \text{var}(u | X_i)$. Do đó, có thể đúng là các quan sát Y từ tổng thể với $X = X_1$ có thể gần tới hàm hồi quy tổng thể PRF hơn là những quan sát đó từ các tổng thể tương ứng với $X = X_2, X = X_3, \dots$. Nói gọn hơn, không phải tất cả các giá trị Y tương ứng với các X khác nhau sẽ đều đáng tin cậy như nhau. Độ tin cậy được đánh giá bởi các giá trị Y phân phối gần hay xa thế nào xung quanh vị trí trung bình của chúng, nghĩa là các điểm trên hàm PRF. Nếu đúng là có trường hợp đó, ta có nên coi trọng các mẫu lấy từ các tổng thể Y nào gần giá trị trung bình hơn là các mẫu với các giá trị phân phối rộng hay không? Nhưng làm như thế cũng có nghĩa là giới hạn những biến đổi ta có được thông qua các giá trị X .

Bằng cách dẫn ra giả thiết 4, ta nói rằng tại giai đoạn này, tất cả các giá trị Y tương ứng với các X khác nhau đều quan trọng như nhau. Trong Chương 11 ta sẽ thấy điều gì sẽ xảy ra nếu đây không phải là trường hợp có phương sai của sai số thay đổi.

Nhân đây, xin lưu ý, giả thiết 4 ngụ ý rằng các phương sai điều kiện của Y_i cũng là phương sai có điều kiện không đổi. Nghĩa là:

$$\text{var}(Y_i | X_i) = \sigma^2 \quad (3.2.4)$$

Tất nhiên, *phương sai vô điều kiện* của Y là σ_Y^2 . Sau này, ta sẽ thấy tầm quan trọng của sự phân biệt giữa các phương sai điều kiện và vô điều kiện của Y (xem phụ lục A để biết thêm chi tiết về các phương sai không điều kiện và có điều kiện).

Giả thiết 5: Không có tự tương quan giữa các nhiễu. Cho trước hai giá trị X bất kỳ, X_i và X_j ($i \neq j$), tương quan giữa u_i và u_j bất kỳ ($i \neq j$) bằng 0. Về mặt ký hiệu:

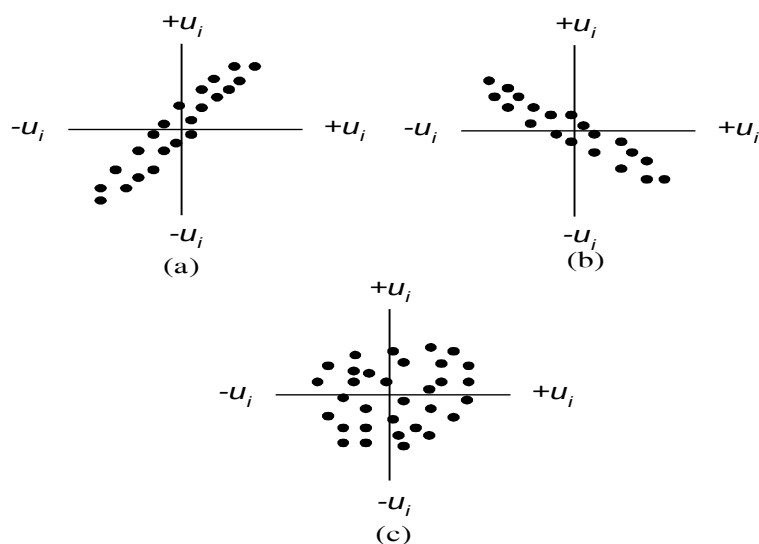
$$\begin{aligned}\text{cov}(u_i, u_j | X_i, X_j) &= E[u_i - E(u_i) | X_i][u_j - E(u_j) | X_j] \\ &= E(u_i | X_i)(u_j | X_j) \text{ (tại sao?)} \\ &= 0\end{aligned}$$

(3.2.5)

trong đó i và j là hai quan sát khác nhau và **cov** nghĩa là **đồng phương sai**.

Nói đúng hơn, (3.2.5) định ra rằng các nhiễu u_i và u_j là không tương quan. Nói bằng thuật ngữ, đây là giả thiết về không có **tương quan chuỗi**, hay là **không có tự tương quan**. Điều này có nghĩa là với các X_i đã cho, các độ lệch của bất kỳ hai giá trị Y từ giá trị trung bình của chúng đều không biểu hiện kiểu như đã mô tả ở trên hình 3.6a và 3.6b. Trên hình 3.6a ta thấy các u **tương quan đồng biến**, một giá trị u dương được có bởi một giá trị u dương hay là một u âm sẽ có từ một giá trị u âm. Trên hình 3.6b, các u lại **tương quan nghịch**, một giá trị u dương sẽ tiếp theo bởi một u âm và ngược lại.

Nếu các nhiễu (các độ lệch) tuân theo các kiểu hệ thống, như là các kiểu trên hình 3.6a và b, đó là tương quan chuỗi hay là tự tương quan, và cái mà giả thiết 5 đòi hỏi là sự vắng mặt của các kiểu tương quan này. Hình 3.6c chỉ rằng không có kiểu hệ thống đối với các u , do đó nó chỉ tương quan zero (không tương quan).

**Hình 3.6**

Các kiểu tương quan giữa các nhiễu. (a) tương quan chuỗi đồng biến; (b) tương quan chuỗi nghịch biến; (c) tương quan zero.

Tầm quan trọng toàn diện của giả thiết này sẽ được giải thích kỹ càng trong Chương 12. Nhưng ta có thể giải thích nó bằng trực giác như sau. Trong hàm PRF của chúng ta ($Y_t = \beta_1 + \beta_2 X_t + u_t$) ta cho rằng u_t và u_{t-1} là tương quan đồng biến. Thì Y_t không chỉ phụ thuộc vào X_t mà còn phụ thuộc vào u_{t-1} , u_{t-1} cùng với một vài sự mở rộng sẽ định ra u_t . Tại giai đoạn phát triển này của đối tượng nghiên cứu, bằng cách dẫn chứng giả thiết 5, ta nói rằng ta sẽ xét ảnh hưởng có tính hệ thống, nếu có, của X_t và Y_t và không quan tâm đến các ảnh hưởng khác có thể tác động đến Y như là kết quả của các tự tương quan có thể có giữa các u . Thế nhưng, như đã lưu ý ở Chương 12, ta sẽ thấy các tương quan giữa các nhiễu sẽ được đưa vào phép phân tích như thế nào, và cùng với kết quả nào.

Giả thiết 6: Đồng phương sai zero giữa u_i và X_i , hay là $E(u_i, X_i) = 0$. Nói chung,

$$\begin{aligned} \text{cov}(u_i, X_i) &= E[u_i - E(u_i)][X_i - E(X_i)] \\ &= E[u_i(X_i - E(X_i))], \quad \text{vì } E(u_i) = 0 \\ &= E(u_i X_i) - E(X_i)E(u_i), \quad \text{vì } E(X_i) \text{ là không ngẫu nhiên} \\ &= E(u_i X_i), \quad \text{vì } E(u_i) = 0 \\ &= 0, \quad \text{bởi giả thiết} \end{aligned} \quad (3.2.6)$$

Giả thiết 6 phát biểu rằng nhiễu u và các biến giải thích X là không tương quan. Lý do căn bản cho giả thiết này như sau: Khi biểu thị hàm PRF trong (2.4.2), ta cho rằng X và u (đại diện cho ảnh hưởng của tất cả các biến bị bỏ qua) có ảnh hưởng riêng (và bổ sung) tới Y . Thế nhưng, nếu X và u có tương quan, ta không thể nào đánh giá các ảnh hưởng của mỗi biến tới Y . Do đó, nếu X và u là tương quan dương, X tăng khi u tăng và X giảm khi u giảm. Tương tự, nếu X và u là tương quan âm, X tăng khi u giảm và X giảm khi u tăng. Trong mỗi trường hợp, sẽ rất khó khăn để tách rời ảnh hưởng của X và u lên Y .

Giả thiết 6 được đáp ứng một cách tự động nếu biến X là không ngẫu nhiên và khi giả thiết 3 được áp dụng, trong trường hợp đó, $\text{cov}(u_i, X_i) = [X_i - E(X_i)]E[u_i - E(u_i)] = 0$ (tại sao?) Nhưng bởi vì ta cho rằng biến X của ta không chỉ là không ngẫu nhiên, mà còn giả thiết là các giá trị cố định trong các mẫu lặp lại²³, giả thiết 6 không phải là giới hạn đối với chúng ta, nó được nêu ra ở đây chỉ để cho thấy rằng lý thuyết hồi quy đã được trình bày trong kết quả suy diễn logic sẽ vẫn đúng thậm chí nếu các X là ngẫu nhiên, miễn là chúng là độc lập, hay ít ra là không tương quan với các nhiễu u_i ²⁴. (Ta sẽ kiểm tra hệ quả này khi kéo nối lỏng thiết 6 trong Phần II).

Giả thiết 7: Số lượng các quan sát n phải lớn hơn số lượng các thông số được ước lượng. Một cách khác, số lượng các quan sát n phải lớn hơn số lượng các biến giải thích.

²³ Nhắc lại rằng khi thu được mẫu như được trình bày trên Bảng 2.4 và 2.5, ta đã giữ cho các giá trị X là như nhau.

²⁴ Như ta sẽ thảo luận ở Phần II, nếu các X là ngẫu nhiên nhưng phân bố độc lập với u_i , các tính chất của hàm ước lượng nhỏ nhất đã thảo luận ngắn gọn việc tiếp tục được áp dụng, nhưng nếu các biến ngẫu nhiên X chỉ là không tương quan với u_i , các tính chất của hàm ước lượng bình phương tối thiểu thông thường OLS chỉ đúng khi nếu kích cỡ mẫu thật lớn. Tuy nhiên, tại giai đoạn này, không cần thiết phải sa lầy vào điểm lý thuyết này.

Giả thiết này không hề là vô thường vô phạt như ta có thể thoáng nghĩ. Trong ví dụ giả định của Bảng 3.1, hãy tưởng tượng ta chỉ có cặp quan sát đầu tiên cho Y và X (4 và 1). Từ quan sát đơn này, không có cách nào để ước lượng hai đại lượng chưa biết β_1 và β_2 . Ta cần ít nhất là hai cặp các quan sát để ước lượng hai đại lượng chưa biết. Trong Chương sau ta sẽ thấy tầm quan trọng cực kỳ của giả thiết này.

Giả thiết 8: Sự biến thiên trong các giá trị X . Các giá trị X trong một mẫu cho trước không thể tất cả đều bằng nhau. Nói theo từ ngữ kỹ thuật, $\text{var}(X)$ phải là một số dương hữu hạn²⁵.

Giả thiết này cũng không phải là vô thường vô phạt. Hãy nhìn vào phương trình (3.1.6). Nếu tất cả các giá trị X đều là đồng nhất, thì $X_i = \bar{X}$ (tại sao?) và mẫu số của phương trình này sẽ bằng 0, nên không thể tính được β_2 và do đó cả β_1 . Một cách trực giác, ta sẵn sàng thấy vì sao giả thiết này lại quan trọng. Nhìn vào ví dụ chi tiêu tiêu dùng trong gia đình ở Chương 2, nếu như có sự biến thiên rất nhỏ trong thu nhập gia đình, ta sẽ không thể giải thích nhiều về sự biến thiên trong chi tiêu tiêu dùng. Độc giả nên nhớ rằng sự biến thiên trong cả Y và X là điều thiết yếu để sử dụng phép phân tích hồi quy như là một công cụ nghiên cứu. Nói ngắn gọn: các biến phải biến đổi!

Giả thiết 9: Mô hình hồi quy được xác định một cách đúng đắn. Nói cách khác, trong các mô hình được sử dụng trong phép phân tích thực nghiệm không có **độ thiên lệch hoặc sai số đặc trưng**.

Như đã đề cập ở Phần Giới thiệu, phương pháp luận kinh tế lượng cổ điển giả thiết điều ả ý, nếu không phải là lộ rõ, rằng mô hình được sử dụng để kiểm định một lý thuyết kinh tế là “được xác định một cách đúng đắn”. Giả thiết này có thể được giải thích một cách không chính thức như sau. Một sự điều tra kinh tế lượng bắt đầu với việc định rõ một mô hình kinh tế lượng trên cơ sở hiện tượng cần quan tâm. Một số câu hỏi quan trọng phát sinh trong việc xác định một mô hình có thể là: (1) Những biến nào nên được bao gồm trong mô hình? (2) dạng hàm số của mô hình như thế nào? Nó tuyến tính theo các thông số, các biến hay là cả hai? (3) Ta sẽ đặt các giả thiết có tính xác suất nào về Y_i , X_i , và u_i khi đưa chúng vào mô hình?

Đó là những câu hỏi cực kỳ quan trọng vì như ta sẽ chỉ ra ở Chương 13, bằng cách bỏ qua các biến quan trọng ra khỏi mô hình, hay bằng cách chọn dạng hàm số sai, hay là bằng cách đặt các giả thiết ngẫu nhiên sai cho các biến của mô hình, tính hiệu lực đúng đắn trong cách giải thích hồi quy ước lượng sẽ mang độ nghi vấn cao. Để có cảm giác thực về điều này, ta hãy tham khảo đường cong Philips trên hình 1.3, (cho là ta chọn hai mô hình sau đây để mô tả mối liên quan cơ sở giữa tỉ lệ thay đổi tiền lương và tỉ lệ thất nghiệp:

$$Y_i = \alpha_1 + \alpha_2 X_i + u_i \quad (3.2.7)$$

²⁵ Phương sai mẫu của X là

$$\text{var}(X) = \frac{\sum (X_i - \bar{X})^2}{n-1} \text{ trong đó } n \text{ là cỡ mẫu.}$$

$$Y_i = \beta_1 + \beta_2 \left(\frac{1}{X_i} \right) + u_i \quad (3.2.8)$$

trong đó Y_i là tỉ lệ thay đổi tiền lương và X_i là tỉ lệ thất nghiệp.

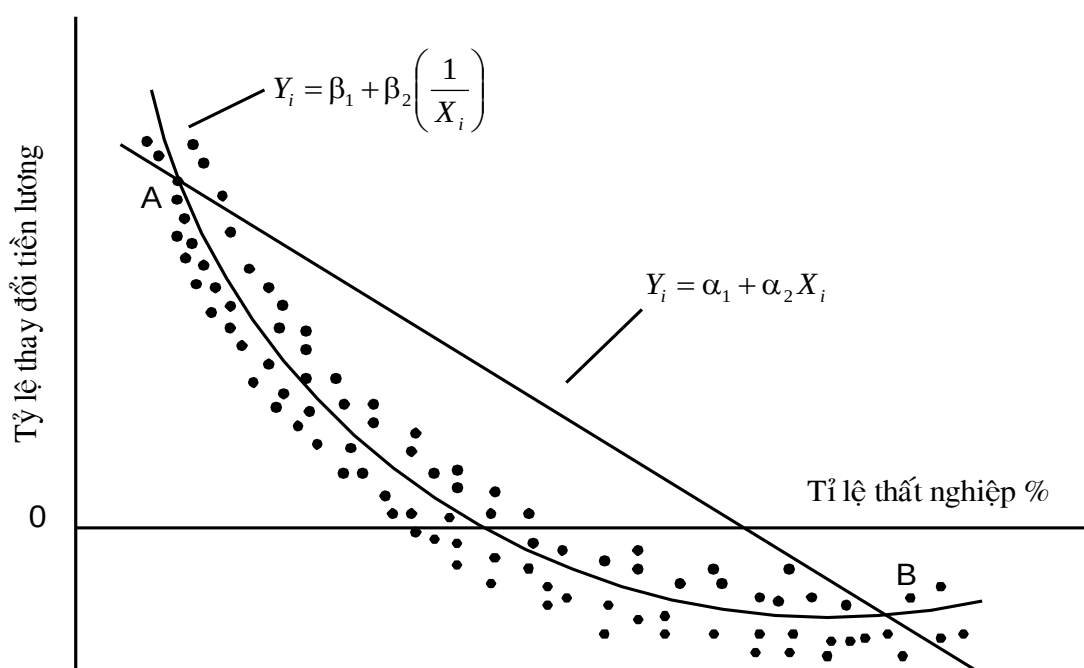
Mô hình hồi quy (3.2.7) là tuyến tính theo các thông số và các biến số trong khi (3.2.8) là tuyến tính theo thông số (do đó, là mô hình hồi quy tuyến tính đúng với định nghĩa của ta) nhưng phi tuyến tính trong biến số X . Bây giờ ta xét hình 3.7 ở cuối trang.

Nếu mô hình (3.2.8) là mô hình “đúng” hay là mô hình “thực” thì sự làm thích hợp mô hình (3.2.7) vào các điểm phân tán trên hình 3.7 sẽ cho ta các dự báo sai: Giữa hai điểm A, B đối với giá trị bất kỳ X_i cho trước, mô hình (3.2.7) sẽ ước lượng quá cao giá trị trung bình thực của Y , trong khi ở phía trái của A (hay phía phải của B) nó sẽ ước lượng thấp (hay là ước lượng cao, khi nói về trị tuyệt đối) giá trị trung bình thực của Y .

Ví dụ trên chính là minh họa cho cái gọi là **độ thiên lệch đặc trưng** hay là **sai số đặc trưng**; độ thiên lệch ở đây có là do chọn dạng hàm số sai. Ta sẽ thấy các loại sai số đặc trưng khác trong Chương 13.

Thật không may là trong thực tế, người ta hiếm khi biết các biến đúng để đặt vào mô hình hay là các hàm đúng của mô hình hay là các giả thiết xác suất đúng về các biến nhập vào mô hình đối với lý thuyết nền tảng kiểm tra cụ thể, (ví dụ như sự đánh đổi giữa tỉ lệ thay đổi tiền lương và tỉ lệ thay đổi thất nghiệp kiểu Phillips) có thể không đủ mạnh hay vững chắc để trả lời các câu hỏi trên. Do đó, trong thực hành, các nhà kinh tế lượng phải sử dụng một sự phán quyết nào đó khi chọn số lượng các biến nhập vào mô hình và dạng hàm của mô hình và phải đặt ra vài giả thiết về bản chất ngẫu nhiên của các biến trong mô hình. Để mở rộng, có vài cách thử và sai nào đó liên quan đến việc chọn mô hình “đúng” cho phép phân tích thực nghiệm.²⁶

²⁶ Người ta có thể tránh cái gọi là “kiểm dữ liệu” nghĩa là, thử mỗi mô hình có thể với hy vọng rằng ít nhất sẽ có một dữ liệu tốt. Đó là vì sao mà nó là thiết yếu mà có vài lý do kinh tế làm cơ sở cho mô hình được chọn và bất cứ sự biến tấu nào của mô hình cũng cần có sự phân xét về kinh tế nào đó. Mô hình đặc biệt thuần túy có thể rất khó mà phân xét trên nền cơ sở lý thuyết hay là priori. Ngắn gọn hơn, lý thuyết nên là cơ sở cho phép ước lượng.

**Hình 3.7**

Các đường cong tuyến tính và phi tuyến tính Phillips

Nếu điều phán xét được đòi hỏi trong việc chọn mô hình thì điều gì cần thiết đối với giả thiết 9? Không cần đi vào chi tiết ở đây (xem Chương 13), giả thiết này có mặt ở đó để nhắc nhở ta rằng phép phân tích hồi quy của ta, và do đó, kết quả dựa trên phép phân tích này là có điều kiện kèm theo với mô hình được chọn và để báo trước cho ta rằng ta nên suy nghĩ thật cẩn thận khi thiết lập các mô hình kinh tế lượng, đặc biệt là khi mà có thể có nhiều học thuyết cạnh tranh cùng cố muốn giải thích một hiện tượng kinh tế, như tỷ lệ lạm phát, hay là nhu cầu về tiền, hay việc xác định giá trị cân bằng hay giá trị cân bằng thích hợp của cổ phiếu hay trái phiếu. Vì vậy, như sau này ta sẽ thấy, việc xây dựng mô hình kinh tế lượng thường nghiêng về phần nghệ thuật hơn là khoa học.

Việc thảo luận về các giả thiết cơ sở của mô hình hồi quy tuyến tính cổ điển của chúng ta đến đây là hoàn tất. Rất quan trọng để lưu ý rằng tất cả các giả thiết này chỉ gắn liền với hàm PRF chứ không gắn với hàm SRF. Nhưng cũng thật thú vị khi quan sát thấy rằng phương pháp bình phương tối thiểu đã đề cập ở trên lại có vài tính chất tương tự như các giả thiết mà ta phải đặt về hàm PRF. Ví dụ, việc tìm ra rằng $\sum \hat{u}_i = 0$, và vì vậy $\bar{\hat{u}} = 0$ là giống với giả thiết $E(u_i | X_i) = 0$. Cũng giống như vậy, việc tìm ra $\sum \hat{u}_i X_i = 0$ cũng tương tự với giả thiết $cov(u_i, X_i) = 0$. Cũng có thể lưu ý rằng phương pháp bình phương tối thiểu do vậy cố gắng là “phó bản” nào đó của các giả thiết mà ta phải đặt cho hàm PRF.

Đương nhiên, hàm SRF không làm phó bản cho tất cả các giả thiết của mô hình hồi quy tuyến tính cổ điển. Như ta sẽ chỉ ra sau này, mặc dù $cov(u_j, u_j) = 0$ do giả thiết, sẽ không đúng sự thực rằng $cov(u_j, u_j)$ của mẫu $= 0$ ($i \neq j$). Thực ra, ta sẽ chỉ ra sau này rằng mặc dù phần dư không chỉ là tự tương quan mà chúng còn có phương sai của sai số thay đổi (xem Chương 12).

Khi bước ra ngoài mô hình hai biến và xem xét các mô hình hồi quy đa biến, nghĩa là, mô hình chứa nhiều biến hồi quy độc lập, ta phải bổ sung các giả thiết sau.

Giả thiết 10: Không có tính đa cộng tuyến hoàn toàn. Nghĩa là *không có các mối tương quan tuyến tính hoàn toàn trong các biến để giải thích.*

Ta sẽ thảo luận về giả thiết này ở Chương 7, khi nói về các mô hình hồi quy đa biến.

Các mô hình giả thiết này thực tế đến mức nào?

Câu hỏi đáng giá cả triệu đô la là: Tất cả các giả thiết này có tính thực tiễn như thế nào? “Tính thực tiễn của các giả thiết” là câu hỏi rất xưa của triết lý trong khoa học. Có người lập luận rằng không cần thiết phải để ý xem các giả thiết có tính thực tiễn hay không. Sự việc nào sẽ là các dự báo dựa trên các giả thiết này. Milton Friedman là người được chú ý, trong luận đề về “tính không thích hợp của các giả thiết”. Theo ông, tính phi thực tiễn của các giả thiết chính là các lợi thế tích cực: “Để trở thành quan trọng ... Mỗi giả thiết phải là điều giả dối trong cách mô tả trong chính các giả thiết của nó.”²⁷

Người ta không thể tán thành hoàn toàn với quan điểm này, nhưng cũng nên nhắc lại rằng trong mỗi nghiên cứu khoa học bất kỳ, ta đưa ra các giả thiết nhất định bởi vì chúng hỗ trợ sự phát triển của các chủ thể trong các bước xa hơn, chứ không phải vì chúng cần có tính thực tiễn trong cảm giác rằng chúng lập lại thực tế một cách chính xác. Như một tác giả đã viết “... Nếu như tính đơn giản là tiêu chuẩn mong muốn của một lý thuyết tốt, tất cả các lý thuyết tốt đều lý tưởng hoá và đơn giản hoá một cách mãnh liệt.”²⁸

Một phép tương tự có thể có ích ở đây. Các sinh viên kinh tế được giới thiệu chung về mô hình của sự cạnh tranh hoàn hảo trước khi họ được giới thiệu về các mô hình cạnh tranh không hoàn hảo như là cạnh tranh độc quyền và cạnh tranh nhóm, bởi vì các ý tiềm ẩn xuất phát từ mô hình này sẽ làm cho ta đánh giá tốt hơn các mô hình cạnh tranh không hoàn hảo, không phải vì mô hình cạnh tranh hoàn hảo mang tính thực tiễn cần thiết. Mô hình hồi quy tuyến tính cổ điển trong kinh tế lượng là tương đương với mô hình cạnh tranh hoàn hảo trong lý thuyết về giá!

Trong kế hoạch của ta, điều cần làm đầu tiên là tìm hiểu các tính chất của mô hình hồi quy tuyến tính cổ điển một cách lý tưởng, và sau đó, trong các chương sau sẽ xem xét thật sâu rằng điều gì sẽ xảy ra nếu như một hay vài giả thiết trong mô hình hồi quy tuyến tính cổ điển không được thực hiện. Ở cuối chương này, trong Bảng 3.5 chúng tôi cung cấp một chỉ dẫn để mỗi người quan tâm có thể tìm điều gì xảy ra với mô hình hồi quy tuyến tính cổ điển khi một giả thiết riêng nào đó không được thoả mãn.

Một đồng nghiệp đã chỉ cho tôi rằng khi ta xem lại một công trình nghiên cứu của người khác nào đó, ta cần phải xem xét các giả thiết nhà nghiên cứu đặt ra có thích hợp với dữ liệu và vấn đề không. Rất thường xảy ra trường hợp khi công trình đã phát hành dựa vào các giả thiết ẩn tàng về vấn đề và dữ liệu, mà vấn đề và dữ liệu này chưa chắc là đúng và nó sinh ra các ước lượng dựa trên các giả thiết đó. Rõ hơn, người đọc có kiến thức nên nhận thức đúng về vấn đề, và lựa chọn cách tiếp cận nghiêm khắc các công trình nghiên cứu. Các giả thiết liệt kê trong Bảng 3.5 sẽ cung cấp một danh sách kiểm tra để hướng dẫn các nghiên cứu của chúng ta và để đánh giá nghiên cứu của người khác.

²⁷ Milton Friedman, *Essay in Positive Economics* (Luận văn về Kinh tế học Thực chứng), University of Chicago Press, Chicago, 1953, trang 14

²⁸ Mark Blaug, cuốn *The Methodology of Economics: Or How Economists Explain* (Phương pháp luận của kinh tế lượng: hay các nhà kinh tế lượng giải thích như thế nào, 2nd Edition, NXB Cambridge University Press, New York, 1992, trang 92.

Với một bước nhỏ quay lại, bây giờ ta sẵn sàng nghiên cứu mô hình hồi quy tuyến tính cổ điển. Nói riêng là ta muốn tìm ra các **tính chất thống kê** của các bình phương tối thiểu thông thường OLS được so sánh với các **tính chất bằng số** mà ta đã thảo luận trước đây. Các tính chất thống kê của bình phương tối thiểu thông thường dựa trên các giả thiết của mô hình hồi quy tuyến tính cổ điển đã được thảo luận và giữ gìn trong **định lý Gauss-Markov** nổi tiếng. Nhưng trước khi quay về với định lý này, định lý cung cấp sự công nhận lý thuyết về tính phổ biến của các bình phương tối thiểu thông thường, đầu tiên ta cần xét **tính chính xác** hay là **các sai số chuẩn** của các phép ước lượng bình phương tối thiểu.

3.3 TÍNH CHÍNH XÁC HAY LÀ CÁC SAI SỐ CHUẨN CỦA CÁCH ƯỚC LƯỢNG BÌNH PHƯƠNG TỐI THIỂU

Từ phương trình (3.1.6) và (3.1.7) ta thấy rõ các ước lượng bình phương tối thiểu là hàm của các dữ liệu mẫu. Nhưng vì dữ liệu có khả năng sẽ thay đổi từ mẫu này sang mẫu khác nên các ước lượng cũng thay đổi từ việc đó. Vì vậy, cần thiết có đại lượng đo “độ tin cậy” nào đó hay là **tính chính xác** của các hàm ước lượng $\hat{\beta}_1$ và $\hat{\beta}_2$. Trong môn thống kê, tính chính xác của một ước lượng nào đó được đo bởi sai số chuẩn của nó²⁹. Cho các giả thiết Gauss như trong phụ lục 3A, Phần 3A.3 chỉ rõ các sai số chuẩn của các ước lượng bình phương tối thiểu thông thường OLS, ta thu được như sau:

$$\text{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_i^2} \quad (3.3.1)$$

$$\text{se}(\hat{\beta}_2) = \frac{\sigma}{\sqrt{\sum x_i^2}} \quad (3.3.2)$$

$$\text{var}(\hat{\beta}_1) = \frac{\sum X_i^2}{n \sum x_i^2} \sigma^2 \quad (3.3.3)$$

$$\text{se}(\hat{\beta}_1) = \sqrt{\frac{\sum X_i^2}{n \sum x_i^2}} \sigma \quad (3.3.4)$$

trong đó **var** là phương sai và **se** là sai số chuẩn và trong đó s^2 là phương sai có điều kiện không đổi hay phương sai hằng số của u_i , trong giả thiết 4.

Trừ đại lượng s^2 , tất cả các số lượng nhập vào phương trình trên đều có thể tính từ dữ liệu. Như đã chỉ ra ở mục 3A, Phần 3A.5, s^2 tự nó được tính bằng công thức sau:

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-2} \quad (3.3.5)$$

²⁹ Sai số chuẩn không là gì nhưng độ lệch chuẩn của sự phân phối mẫu của hàm ước lượng, và sự phân phối mẫu của các hàm ước lượng đơn giản là xác suất hay phân phối tần số của các hàm ước lượng, nghĩa là, phân phối của một bộ các giá trị hàm ước lượng thu được từ các mẫu cùng cỡ từ một tổng thể cho trước. Các phân phối mẫu được sử dụng để rút ra kết luận về các giá trị của các thông số tổng thể trên cơ sở các giá trị hàm ước lượng được tính từ một hay nhiều mẫu (Chi tiết hơn, xem Phụ lục A).

trong đó $\hat{\sigma}^2$ là hàm ước lượng bình phương tối thiểu thông thường OLS của giá trị thực nhưng chưa biết s^2 và trong đó $n-2$ là **số bậc tự do (df)**, $\sum \hat{u}_i^2$ là tổng của bình phương phần dư hay là **tổng bình phương của các phần (RSS)**³⁰.

Nếu đã biết $\sum \hat{u}_i^2$, $\hat{\sigma}^2$ có thể tính được dễ dàng. $\sum \hat{u}_i^2$ tự nó có thể được tính từ (3.1.2) hoặc từ biểu thức sau (xem chứng minh ở phần 3.5)

$$\sum \hat{u}_i^2 = \sum y_i^2 - \hat{\beta}_2^2 \sum x_i^2 \quad (3.3.6)$$

So sánh với phương trình (3.1.2), phương trình (3.3.6) rất dễ sử dụng, vì nó không đòi hỏi phải tính toán $\hat{u}_i$ cho mỗi quan sát mặc dù các tính toán này sẽ rất có ích trong vẽ phải của chính nó (như ta sẽ xem trong Chương 11 và 12).

$$\text{Vì} \quad \hat{\beta}_2 = \frac{\sum x_i y_i}{\sum x_i^2}$$

dạng biểu hiện thay thế cho việc tính $\sum \hat{u}_i^2$ là:

$$\sum \hat{u}_i^2 = \sum y_i^2 - \frac{(\sum x_i y_i)^2}{\sum x_i^2} \quad (3.3.7)$$

Nhân thể, lưu ý rằng căn bậc hai dương của $\hat{\sigma}^2$:

$$\hat{\sigma} = \sqrt{\frac{\sum \hat{u}_i^2}{n-2}} \quad (3.3.8)$$

được biết như **sai số chuẩn của ước lượng**. Nó đơn giản là độ lệch chuẩn của các giá trị $\hat{Y}$ so với đường hồi quy ước lượng và nó thường được sử dụng như là đại lượng đo “độ thích hợp” của đường hồi quy ước lượng, nội dung đó sẽ được thảo luận trong Phần 3.5.

Trước đây, ta lưu ý rằng, cho trước các X_i , s^2 đại diện cho phương sai (điều kiện) của cả hai đại lượng u_i và Y_i . Vì vậy, sai số chuẩn của sự ước lượng có thể gọi là độ lệch chuẩn (điều kiện) của u_i và Y_i . Đương nhiên, như thông thường, s_Y^2 và s_Y sẽ đại diện tương ứng cho phương sai không điều kiện và độ lệch chuẩn không điều kiện của Y .

Lưu ý các đặc tính sau đây của phương sai (vì vậy, của cả sai số chuẩn) của $\hat{\beta}_1$ và $\hat{\beta}_2$

1. Phương sai của $\hat{\beta}_2$ tỷ lệ thuận với s^2 nhưng tỷ lệ nghịch với $\sum x_i^2$. Nghĩa là, nếu cho trước s^2 , sự biến thiên của X càng lớn thì phương sai của $\hat{\beta}_2$ càng nhỏ, do đó, tính chính xác của b_2 ước lượng được cũng sẽ tăng. Ngắn gọn hơn, nếu cho trước s^2 , nếu có sự biến thiên thực sự đối với các giá trị X (coi lại giả thiết 8), b_2 có thể được xác định chính xác hơn là khi X_i không biến thiên thực sự. Cũng như vậy, cho trước $\sum x_i^2$, phương sai của s^2 càng lớn thì phương sai của b_2 càng lớn. Lưu ý rằng, khi cỡ mẫu n tăng thì số lượng các số hạng trong

³⁰ Thuật ngữ **số bậc tự do** nghĩa là số lượng tổng cộng các quan sát trong mẫu ($= n$) trừ đi số ràng buộc hay giới hạn độc lập (tuyến tính) đặt vào các quan sát đó. Nói cách khác, nó là số lượng các quan sát độc lập ở bên ngoài tổng số n lần quan sát. Ví dụ, trước khi tính được RSS (tổng bình phương của các phần dư) theo (3.1.2), phải tính được $\hat{\beta}_1$ và $\hat{\beta}_2$. Hai giá trị ước lượng này đặt hai giới hạn vào RSS. Cho nên, có $n-2$ chứ không phải n quan sát tự do để tính RSS. Theo lộ gíc này, trong RSS hồi quy 3 biến ta sẽ có $n-3$ bậc tự do, và cho mô hình k biến mô hình sẽ có $n-k$ bậc tự do. Quy tắc chung là bậc tự do $= n - \text{số lượng các thông số được ước lượng}$.

tổng $\sum x_i^2$ sẽ tăng. Vì khi n tăng thì tính chính xác mà với nó β_2 ước lượng được cũng sẽ tăng (Tại sao?).

2. Phương sai của $\hat{\beta}_1$ tỷ lệ thuận với s^2 và $\sum X_i^2$ nhưng tỷ lệ nghịch với $\sum x_i^2$ và kích thước n của mẫu.
3. Vì $\hat{\beta}_1$ và $\hat{\beta}_2$ là các hàm ước lượng, chúng sẽ không chỉ biến đổi từ mẫu này đến mẫu khác, mà ngay trong một mẫu cho trước, chắc chắn chúng sẽ phụ thuộc lẫn nhau, sự phụ thuộc này được đo bởi đồng phương sai giữa chúng. Điều này được đề cập đến trong phụ lục 3A, Phần 3A.4, như sau:

$$\begin{aligned}\text{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -\bar{X} \text{var}(\hat{\beta}_2) \\ &= -\bar{X} \left(\frac{\sigma^2}{\sum x_i^2} \right)\end{aligned}\quad (3.3.9)$$

Vì là phương sai của một biến bất kỳ $\text{var}(\hat{\beta}_2)$ luôn luôn dương, bản chất của đồng phương sai giữa $\hat{\beta}_1$ và $\hat{\beta}_2$ phụ thuộc vào dấu của $\bar{X}$. Nếu $\bar{X}$ dương, thì theo công thức, đồng phương sai sẽ âm. Vì vậy, nếu hệ số góc b_2 được *ước lượng cao* (nghĩa là độ dốc rất dốc), hệ số tung độ β_1 sẽ được *ước lượng thấp* (nghĩa là tung độ góc sẽ rất nhỏ). Sau này, (đặc biệt trong Chương 10 về tính đa cộng tuyến), ta sẽ thấy rõ lợi ích của việc nghiên cứu đồng phương sai giữa các hệ số hồi quy được ước lượng.

Các phương sai và các sai số chuẩn của các hệ số hồi quy ước lượng sẽ đưa người ta đến việc phán xét về tính thực tiễn của các ước lượng này như thế nào? Đây là vấn đề trong suy diễn thống kê và nó sẽ được đưa vào Chương 4 và 5.

3.4 CÁC TÍNH CHẤT CỦA HÀM ƯỚC LƯỢNG: ĐỊNH LÝ GAUSS-MARKOV³¹

Như đã đề cập trước đây, khi cho trước các giả thiết của mô hình hồi quy tuyến tính cổ điển, các phép ước lượng bình phương tối thiểu đặt vài tính chất tối ưu hoặc là lý tưởng nào đó. Các tính chất này được chứa đựng trong **định lý Gauss-Markov** nổi tiếng. Để hiểu lý thuyết này, ta cần xét tính chất **không thiên lệch tuyến tính tốt nhất** của hàm ước lượng.³² Như đã giải thích ở phụ lục A, một hàm ước lượng, gọi là hàm ước lượng bình phương tối thiểu thông thường của $\hat{\beta}_2$, được cho là hàm ước lượng không thiên lệch tuyến tính tốt nhất (BLUE) của b_2 nếu như theo đúng các điều sau:

1. Nó là **tuyến tính**, nghĩa là hàm tuyến tính của biến ngẫu nhiên, như là biến phụ thuộc Y trong mô hình hồi quy.
2. Nó là **không thiên lệch**, nghĩa là giá trị trung bình của nó hay là giá trị kỳ vọng, $E(\hat{\beta}_2)$, bằng giá trị thực β_2 .
3. Nó có phương sai nhỏ nhất trong nhóm tất cả các hàm ước lượng không thiên lệch tuyến tính; hàm ước lượng không thiên lệch với phương sai tối thiểu được gọi là **hàm ước lượng hiệu quả**.

³¹ Tuy gọi là lý thuyết Gauss-Markov nhưng phép tính gần đúng các bình phương tối thiểu của Gauss xảy ra trước (1821) phép tính gần đúng phương sai cực tiểu của Markov (1900).

³² Bạn đọc nên tham khảo phụ lục A để biết tầm quan trọng của các hàm ước lượng tuyến tính cũng như cách thảo luận chung về các tính chất mong muốn của các hàm ước lượng thống kê.

Trong nội dung hồi quy, nó có thể được chứng minh rằng các hàm ước lượng bình phương tối thiểu thông thường là hàm ước lượng không thiên lệch tuyến tính tốt nhất. Đây là thực chất của định lý Gauss-Markov nổi tiếng, lý thuyết đó có thể được phát biểu như sau:

Định lý Gauss-Markov: Cho trước các giả thiết của mô hình hồi quy tuyến tính cổ điển, các hàm ước lượng bình phương tối thiểu, trong nhóm các hàm ước lượng tuyến tính không thiên lệch, có phương sai nhỏ nhất, nghĩa là chúng là các hàm ước lượng không thiên lệch tuyến tính tốt nhất.(BLUE)

Bằng chứng của định lý này đã được phác họa trong Phụ lục 3A, Phần 3A.6. Sự xâm nhập rộng rãi của định lý Gauss-Markov sẽ trở nên rõ ràng hơn khi ta đi xa hơn. Sẽ không thừa khi muốn lưu ý rằng định lý này có tầm quan trọng về lý thuyết cũng như về thực hành.³³

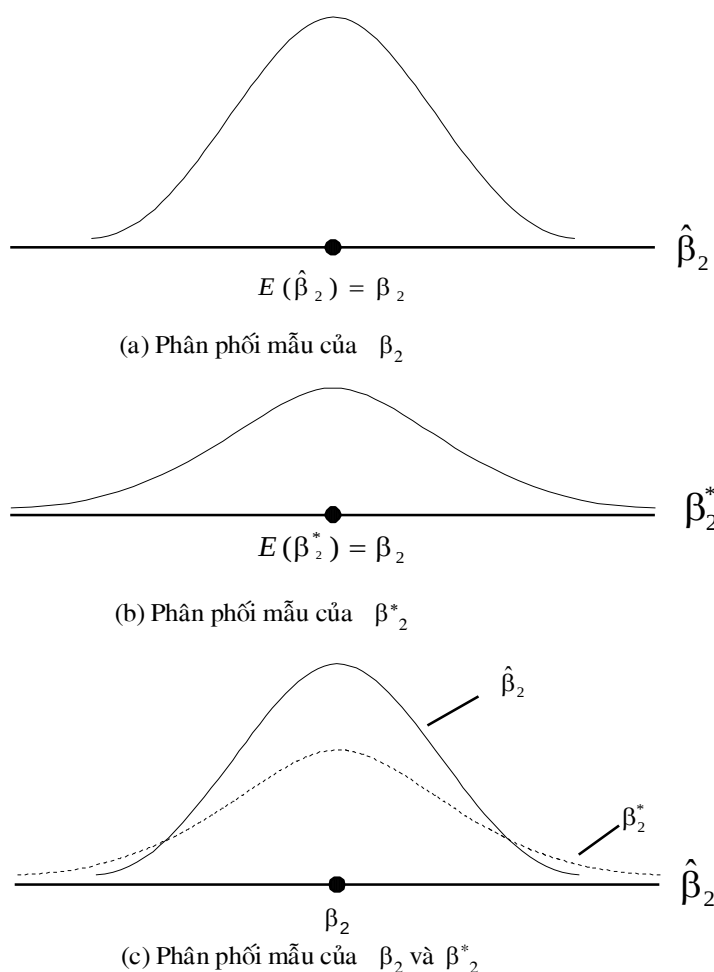
Tất cả những ý nghĩa này có thể được giải thích bằng hình 3.8.

Trong hình 3.8(a) ta đã chỉ ra **phân phối mẫu** của hàm ước lượng bình phương tối thiểu thông thường $\hat{\beta}_2$, nghĩa là phân phối các giá trị lấy bởi $\hat{\beta}_2$ trong các thử nghiệm lấy mẫu lặp lại (nhắc lại Bảng 3.1). Để thuận lợi, ta giả thiết rằng $\hat{\beta}_2$ phân phối đối xứng (nhiều hơn có thể xem Chương 4). Như các hình đưa ra, trung bình của các giá trị $\hat{\beta}_2$, $E(\hat{\beta}_2)$ bằng giá trị thực b_2 . Ở đây, ta nói $\hat{\beta}_2$ là *hàm ước lượng không thiên lệch* của b_2 . Trong hình 3.8(b) ta thấy phân phối mẫu của b_2^* , hàm ước lượng thay thế b_2 thu được bằng phương pháp khác (nghĩa là không phải phương pháp bình phương tối thiểu thông thường). Để tiện lợi, ta giả sử b_2^* giống như $\hat{\beta}_2$ là không thiên lệch, nghĩa là trung bình hay là giá trị kỳ vọng của chúng bằng b_2 . Tiếp theo, ta giả sử rằng $\hat{\beta}_2$ và b_2^* là các hàm ước lượng tuyến tính, nghĩa là chúng là các hàm tuyến tính của Y .

Ta sẽ chọn hàm ước lượng nào: $\hat{\beta}_2$ hay b_2^* ?

Để trả lời câu hỏi này, ta chồng hai hình này lên nhau như trên hình (3.8)(c). Rõ ràng là mặc dù $\hat{\beta}_2$ và b_2^* đều là không thiên lệch, phân phối của b_2^* phân tán hơn hay là trải rộng hơn so với phân phối của $\hat{\beta}_2$ xung quanh giá trị trung bình. Nói cách khác, phương sai của b_2^* rộng hơn phương sai của $\hat{\beta}_2$. Bây giờ, cho trước hai hàm ước lượng, đều là không thiên lệch và tuyến tính, ta cần chọn hàm ước lượng với phương sai nhỏ hơn vì nó sẽ gần với b_2 hơn là hàm thay thế. Nói gọn hơn, ta nên chọn hàm ước lượng không thiên lệch tuyến tính tốt nhất.(BLUE).

³³ Ví dụ: có thể cho rằng một kết hợp tuyến tính bất kỳ của b như $(b_1 - 2b_2)$, có thể được ước lượng bởi $(\hat{\beta}_1 - 2\hat{\beta}_2)$, và hàm ước lượng này là BLUE. Chi tiết hơn có thể xem cuốn *Introduction to Econometrics* (Nhập môn Kinh tế lượng) của Henri Theil, Prentice Hall, Englewood Cliffs, New Jersey, 1978, trang 401-402.

**Hình 3.8**

Phân phối mẫu của hàm ước lượng bình phương tối thiểu
thông thường $\hat{\beta}_2$ và hàm ước lượng thay thế β_2^*

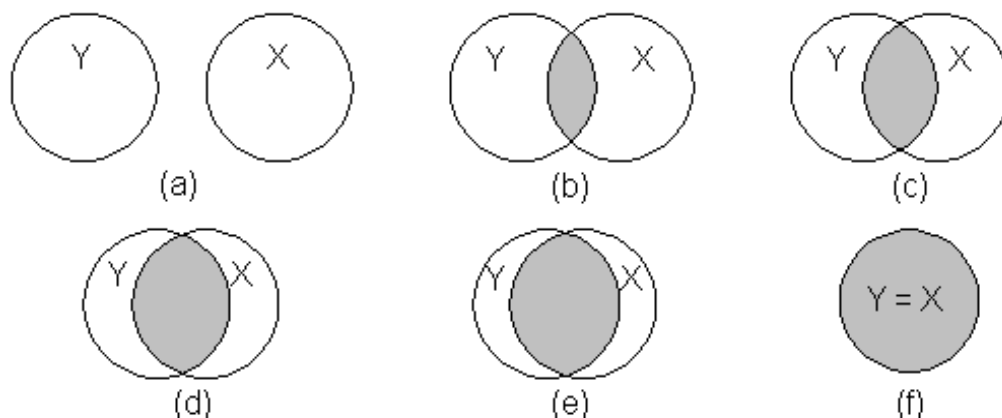
Các tính chất thống kê mà ta vừa thảo luận được biết như là các **tính chất mẫu hữu hạn**: Các tính chất này thỏa mãn mong muốn về cỡ mẫu, trên nền tảng của các hàm ước lượng. Sau này ta sẽ có dịp xét các **tính chất tiệm cận**, nghĩa là các tính chất chỉ áp dụng nếu cỡ của mẫu rất lớn (nghĩa là vô hạn). Một sự thảo luận chung về các tính chất mẫu hữu hạn và mẫu lớn của các hàm ước lượng được đưa vào phụ lục A.

3.5 HỆ SỐ XÁC ĐỊNH r^2 : ĐẠI LƯỢNG ĐO “SỰ THÍCH HỢP”

Cho đến giờ ta đã đề cập đến vấn đề ước lượng các hệ số hồi quy, các sai số chuẩn của chúng, và một số tính chất của chúng. Bây giờ ta xét đến **sự thích hợp** của các đường hồi quy thích hợp với bộ dữ liệu; nghĩa là, ta sẽ tìm ra rằng đường hồi quy mẫu sẽ thích hợp “tốt” như thế nào với dữ liệu. Từ hình 3.1 rõ ràng là ta có thể thu được sự thích hợp “hoàn hảo” nếu như tất cả các quan sát nằm trên đường hồi quy, nhưng trường hợp đó thật hiếm. Nói chung, sẽ có vài u_i dương

và vài $\hat{u}_i$ âm. Điều mà ta hy vọng là những phần dư xung quanh đường hồi quy này sẽ càng nhỏ càng tốt. **Hệ số xác định r^2** (trường hợp hai biến) hay là R^2 (hồi quy đa biến) là đại lượng chỉ cho ta rằng đường hồi quy mẫu thích hợp tốt như thế nào với dữ liệu.

Trước khi chỉ rõ r^2 được tính như thế nào ta hãy xét sự giải thích có tính khai phá đối với r^2 bằng đồ thị, đó là **phương pháp đồ thị Venn, hay là Ballentine**, như trên hình 3.9³⁴



Hình 3.9

Quan điểm Ballentine đối với r^2 : (a) $r^2 = 0$; (f) $r^2 = 1$

Trong hình này vòng tròn Y tượng trưng cho biến thiên trong biến phụ thuộc Y và vòng X tượng trưng cho biến thiên trong biến giải thích X^{35} . Vùng chồng lên nhau của hai vòng tròn (vùng tối) chỉ rõ phạm vi mà độ biến thiên trong Y được giải thích bởi biến thiên trong X (cho là theo hướng hồi quy các bình phương tối thiểu thông thường OLS). Phạm vi vùng chồng lên càng lớn, độ biến thiên trong Y được giải thích bởi X càng lớn. r^2 đơn giản là đại lượng đo bằng số cho vùng tối này. Trong hình, khi ta di chuyển từ trái sang phải, vùng tối tăng dần nghĩa là tỷ lệ biến thiên trong Y được giải thích bởi X liên tục tăng. Nói ngắn gọn, r^2 tăng. Khi không có vùng tối, r^2 rõ ràng bằng 0, nhưng khi vùng tối đã hoàn chỉnh, r^2 bằng 1, và 100% độ biến thiên của Y được giải thích bởi X. Ta thấy ngắn gọn rằng r^2 nằm giữa 0 và 1.

Để tính r^2 , ta làm như sau. Nhắc lại rằng:

$$Y_i = \hat{Y}_i + \hat{u}_i \quad (2.6.3)$$

hay là trong dạng độ lệch:

$$y_i = \hat{y}_i + \hat{u}_i \quad (3.5.1)$$

trong đó đã sử dụng (3.1.13) và (3.1.14). Bình phương (3.5.1) cho cả hai vế và lấy tổng đối với mẫu, ta có:

³⁴ Xem cuốn “Ballentine: A Graphical Aid for Econometrics” (Ballentine: Một hỗ trợ bằng đồ thị cho Kinh tế lượng) của Peter Kennedy, *Australian Economics Papers*, Vol. 20, 1981, 414-416. Tên gọi Ballentine xuất phát từ huy hiệu bìa Ballentine nổi tiếng với các vòng của nó.

³⁵ Thuật ngữ *biến thiên* và *phương sai* là khác nhau. Biến thiên là tổng các bình phương của độ lệch giữa biến số với giá trị trung bình của nó. Phương sai là tổng các bình phương chia cho bậc tự do thích ứng. Nói ngắn gọn, phương sai bằng biến thiên chia cho bậc tự do (df)

$$\begin{aligned}
\sum y_i^2 &= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 + 2 \sum \hat{y}_i \hat{u}_i \\
&= \sum \hat{y}_i^2 + \sum \hat{u}_i^2 \\
&= \hat{\beta}_2^2 \sum x_i^2 + \sum \hat{u}_i^2
\end{aligned}
\tag{3.5.2}$$

vì $\sum y_i \hat{u}_i = 0$ (tại sao) và $\hat{y}_i = \hat{\beta}_2 x_i$

Các tổng khác nhau của các bình phương xuất hiện trong (3.5.2) có thể được mô tả như sau: $\sum y_i^2 = \sum (Y_i - \bar{Y})^2$ = độ lệch tổng cộng của giá trị thực của Y so với trung bình mẫu của chúng, nó có thể được gọi là **tổng bình phương toàn phần (TSS)**. $\sum \hat{y}_i^2 = \sum (\hat{Y}_i - \bar{\hat{Y}})^2 = \sum (\hat{Y}_i - \bar{Y})^2 = \hat{\beta}_2^2 \sum x_i^2$ = chênh lệch của giá trị ước lượng của Y với trung bình của chúng ($\bar{\hat{Y}} = \bar{Y}$), nó có thể được gọi một cách gần đúng là tổng của các bình phương do hồi quy [nghĩa là do (các) biến giải thích] hay là được giải thích bởi hồi quy, hay đơn giản **tổng bình phương giải thích được (tổng bình phương hồi qui) (ESS)**. $\sum \hat{u}_i^2$ = phần dư hay là biến thiên **không giải thích** của giá trị Y với đường hồi quy, hay đơn giản là **tổng bình phương phần dư (tổng bình phương sai số (RSS))**. Vì vậy, (3.5.2) là:

$$TSS = ESS + RSS \tag{3.5.3}$$

và chỉ ra rằng, độ lệch tổng cộng trong các giá trị Y được quan sát so với giá trị trung bình có thể được phân ra hai phần, một phần là do đường hồi quy và phần khác là do sự bất buộc ngẫu nhiên vì không phải tất cả các quan sát thực tế Y nằm trên đường thích hợp. Một cách hình học, ta có hình 3.10:

Bây giờ ta chia 2 vế (3.5.3) cho TSS, ta được:

$$\begin{aligned}
1 &= \frac{ESS}{TSS} + \frac{RSS}{TSS} \\
&= \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} + \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2}
\end{aligned}
\tag{3.5.4}$$

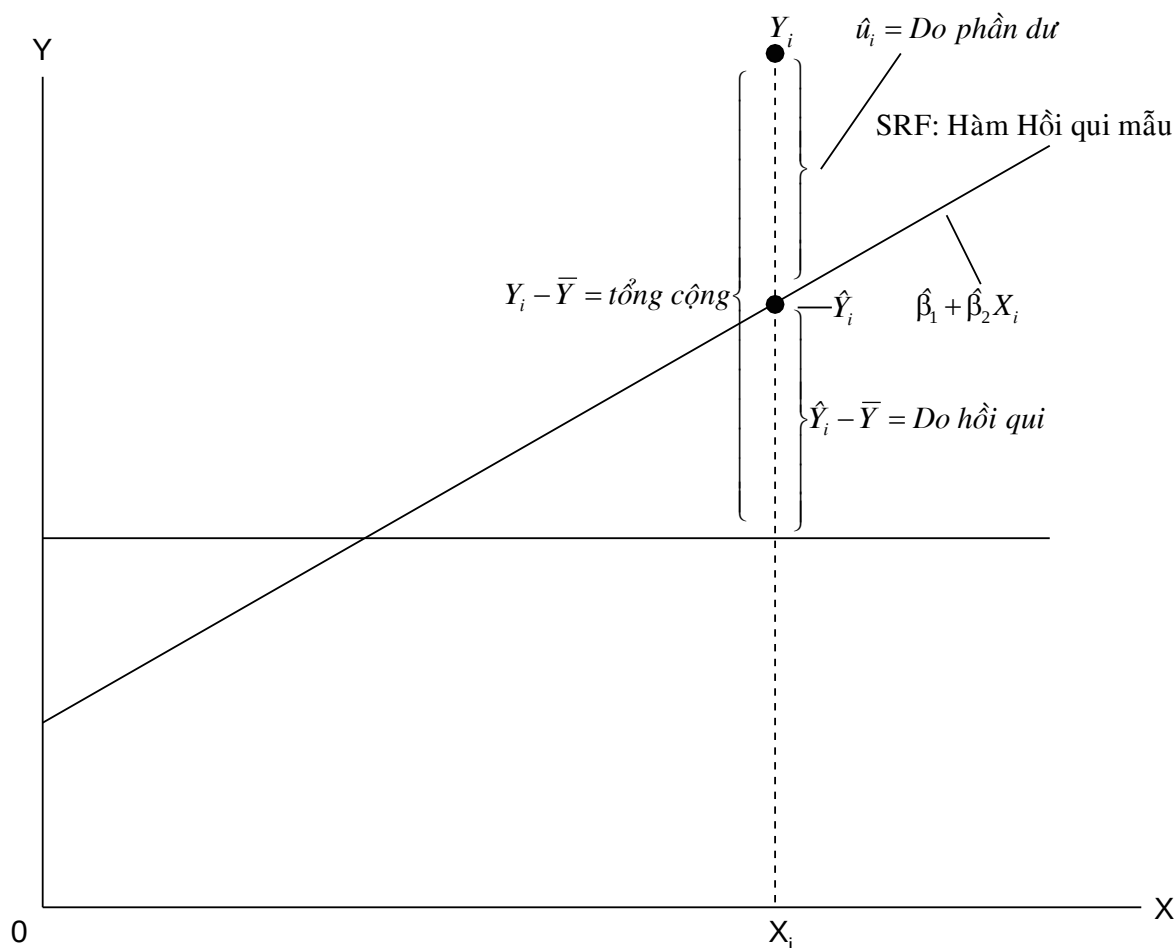
Ta định nghĩa r^2 là:

$$r^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} = \frac{ESS}{TSS} \tag{3.5.5}$$

Hay ta có thể viết

$$r^2 = 1 - \frac{\sum \hat{u}_i^2}{\sum (Y_i - \bar{Y})^2} \quad (3.5.5a)$$

$$= 1 - \frac{RSS}{TSS}$$



Hình 3.10

Sự chia độ biến thiên của Y_i ra hai thành phần

Số lượng r^2 được xác định như vậy gọi là **hệ số xác định** và là đại lượng được sử dụng chung để đo tính thích hợp tốt của đường hồi quy. Bằng lời, r^2 đo tỷ số hay là phần trăm của độ lệch tổng cộng trong Y được giải thích bởi mô hình hồi quy.

Ta lưu ý 2 tính chất của r^2 :

1. Nó là số không âm (Tại sao?)
2. Giới hạn của nó $0 \leq r^2 \leq 1$. r^2 bằng 1 nghĩa là hoàn toàn phù hợp, nghĩa là $\hat{Y}_i = Y_i$ với mỗi i . Ở đầu khác, $r^2 = 0$ nghĩa là dù thế nào đi nữa (nghĩa là $\hat{\beta}_2 = 0$) cũng không có liên quan giữa biến hồi qui phụ thuộc và biến hồi qui độc lập. Trong trường hợp này, như (3.1.9) chỉ rõ, $\hat{Y}_i = \hat{\beta}_1 = \bar{Y}$, nghĩa là, dự báo tốt nhất của giá trị Y bất kỳ đơn giản là giá trị trung bình của nó. Vì vậy, đường hồi quy sẽ là đường nằm ngang so với trục X .

Tuy r^2 có thể được tính trực tiếp từ định nghĩa trong (3.5.5), nó có thể tính được nhanh hơn từ công thức sau:

$$\begin{aligned} r^2 &= \frac{ESS}{TSS} \\ &= \frac{\sum \hat{y}_i^2}{\sum y_i^2} \\ &= \frac{\hat{\beta}_2^2 \sum x_i^2}{\sum y_i^2} \\ &= \hat{\beta}_2^2 \left(\frac{\sum x_i^2}{\sum y_i^2} \right) \end{aligned} \quad (3.5.6)$$

Nếu ta chia cả tử số và mẫu số của (3.5.6) cho cỡ mẫu n (hay $n-1$ nếu cỡ mẫu nhỏ), ta có:

$$r^2 = \hat{\beta}_2^2 \left(\frac{S_x^2}{S_y^2} \right) \quad (3.5.7)$$

trong đó S_y^2 và S_x^2 tương ứng là các phương sai mẫu của Y và X .

Vì $\hat{\beta}_2 = \sum x_i y_i / \sum x_i^2$, phương trình (3.5.6) có thể biểu thị như là

$$r^2 = \frac{(\sum x_i y_i)^2}{\sum x_i^2 \sum y_i^2} \quad (3.5.8)$$

và biểu thức trên có thể tính dễ dàng.

Cho trước định nghĩa r^2 , ta có thể biểu thị ESS và RSS đã được thảo luận trước đây như sau:

$$\begin{aligned} ESS &= r^2 \cdot TSS \\ &= r^2 \sum y_i^2 \end{aligned} \quad (3.5.9)$$

$$\begin{aligned} RSS &= TSS - ESS \\ &= TSS(1 - ESS/TSS) \\ &= \sum y_i^2 \cdot (1 - r^2) \end{aligned} \quad (3.5.10)$$

Do đó, ta có thể viết:

$$\begin{aligned} TSS &= ESS + RSS \\ \sum y_i^2 &= r^2 \sum y_i^2 + (1 - r^2) \sum y_i^2 \end{aligned} \quad (3.5.11)$$

biểu thức mà sẽ rất bổ ích sau này.

Giá trị bằng số thì quan hệ rất gần, nhưng về khái niệm, r^2 khác xa với **hệ số tương quan**, là đại lượng đo bậc kết hợp giữa hai biến (như Chương 1 đã lưu ý). Nó cũng có thể tính từ biểu thức:

$$r = \pm \sqrt{r^2} \quad (3.5.12)$$

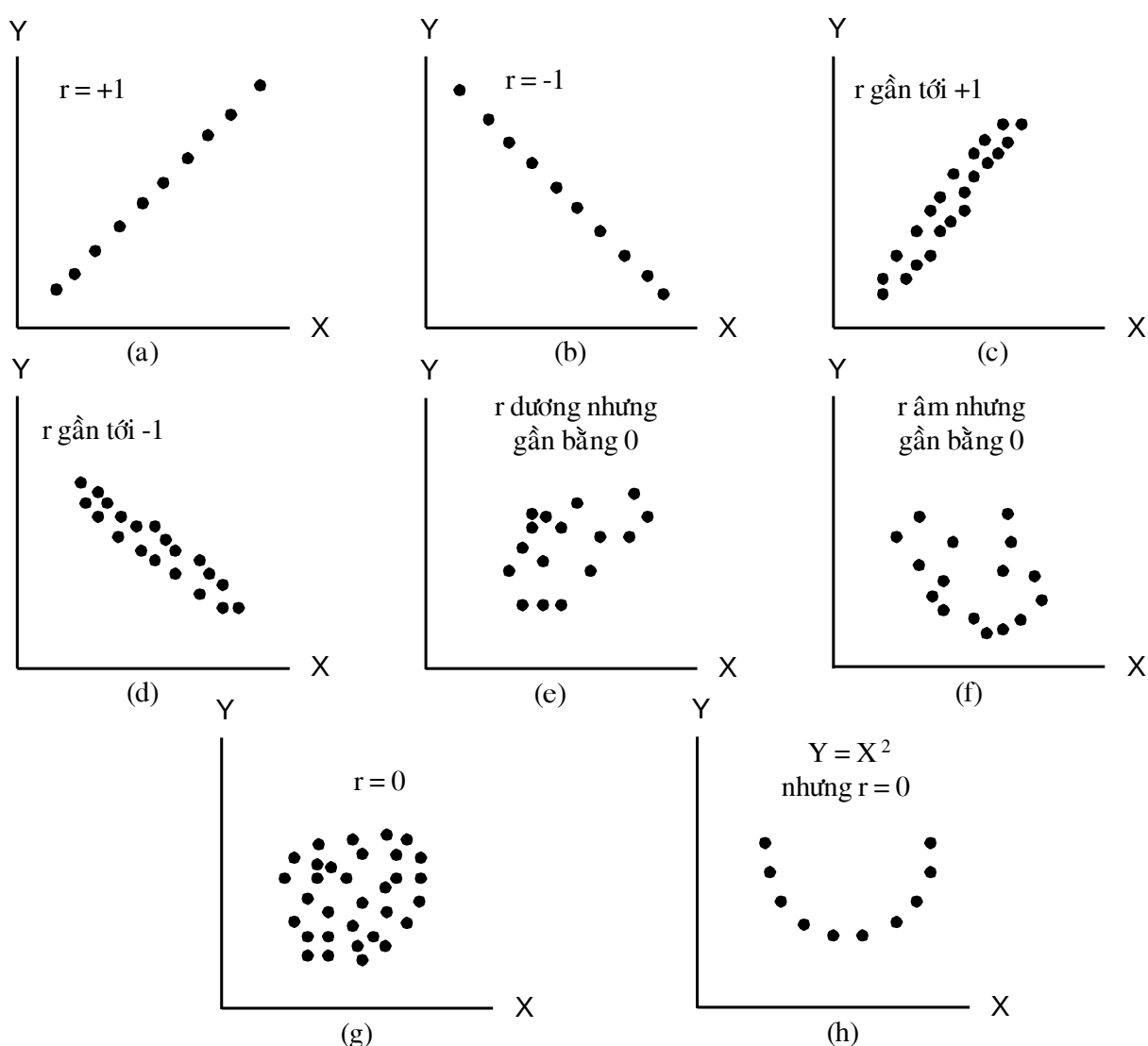
hay từ định nghĩa

$$r = \frac{\sum x_i y_i}{\sqrt{(\sum x_i^2)(\sum y_i^2)}} \quad (3.5.13)$$

$$= \frac{n \sum X_i Y_i - (\sum X_i)(\sum Y_i)}{\sqrt{[n \sum X_i^2 - (\sum X_i)^2][n \sum Y_i^2 - (\sum Y_i)^2]}}$$

được xem là **hệ số tương quan mẫu**³⁶

Một vài tính chất của r như sau (xem hình 3.11):



Hình 3.11

Các kiểu tương quan (theo Henri Theil, Nhập môn Kinh tế lượng, Prentice Hall, Englewood Cliffs, N.J, 1978, trang 86)

³⁶ Hệ số tương quan tổng thể, được biểu thị bởi ρ được định nghĩa trong phụ lục A.

1. r có thể dương hoặc âm, dấu của r phụ thuộc vào dấu của số hạng trong tử số của (3.5.13), đo *đồng phương sai* mẫu của hai biến.
2. r nằm từ -1 đến $+1$, nghĩa là $-1 \leq r \leq +1$
3. Bản chất của r là đối xứng; nghĩa là hệ số tương quan giữa X và Y (r_{XY}) cũng bằng hệ số đó giữa Y và X (r_{YX}).
4. r độc lập đối với gốc tọa độ và các tỷ lệ; nghĩa là nếu ta định nghĩa $X_i^* = aX_i + c$ và $Y_i^* = bY_i + d$, trong đó $a > 0$, $b > 0$ và c, d là hằng số, thì giữa X^* và Y^* cũng có một giá trị r giống như giá trị r giữa các biến nguyên thủy X và Y .
5. Nếu X và Y là độc lập theo quan điểm thống kê (xem phụ lục A để có khái niệm), hệ số tương quan giữa chúng bằng 0; nhưng nếu $r = 0$, điều đó không có nghĩa là hai biến này độc lập. Nói cách khác, **hệ số tương quan zero không ngụ ý là có tính độc lập** (xem hình 3.11(h)).
6. r chỉ là đại lượng đo sự *kết hợp tuyến tính* hay là *phụ thuộc tuyến tính*; r không có ý nghĩa để mô tả quan hệ phi tuyến tính. Vì vậy, trong hình 3.11(h), $Y = X^2$ là một quan hệ chính xác nhưng $r = 0$. (Tại sao?)
7. Mặc dù r là đại lượng đo sự kết hợp tuyến tính giữa hai biến, r không ngụ ý là có bất kỳ mối liên quan nhân quả nào, như ta đã lưu ý ở Chương 1.

Trong nội dung hồi quy, r^2 là đại lượng có đủ ý nghĩa hơn r , nó cho ta biết tỷ lệ độ biến thiên trong các biến phụ thuộc được giải thích bởi (các) biến giải thích và do đó, nó cũng cho ta thước đo toàn diện của phạm vi mà độ biến thiên trong một biến xác định độ biến thiên trong các biến khác. Đại lượng sau không thể có cùng giá trị đó³⁷. Hơn thế nữa, như ta sẽ thấy sau này, việc chứng minh $r (= R)$ trong mô hình hồi quy đa biến là giá trị hơi mơ hồ. Tuy nhiên ta sẽ còn phải nói nhiều về r^2 trong Chương 7

Nhân tiện đây, xin lưu ý rằng r^2 , như định nghĩa ở trên có thể được *tính bằng bình phương của hệ số tương quan giữa giá trị thực của Y_i và giá trị ước lượng của Y_i* , gọi là $\hat{Y}_i$. Nghĩa là khi sử dụng (3.5.13) ta có thể viết:

$$r^2 = \frac{[\sum (Y_i - \bar{Y})(\hat{Y}_i - \bar{Y})]^2}{\sum (Y_i - \bar{Y})^2 \sum (\hat{Y}_i - \bar{Y})^2}$$

Nghĩa là

$$r^2 = \frac{(\sum y_i \hat{y}_i)^2}{(\sum y_i^2)(\sum \hat{y}_i^2)} \quad (3.5.14)$$

trong đó $Y_i = Y$ thực, $\hat{Y}_i = Y$ ước lượng, và $\bar{Y} = \bar{\hat{Y}}$ = giá trị trung bình của Y . Để có thêm bằng chứng xin xem bài tập 3.15. Biểu thức (3.5.14) chứng minh rằng r^2 là đại lượng đo của sự thích hợp, vì nó chỉ rõ giá trị Y ước lượng sẽ gần như thế nào tới các giá trị thực của chúng.

3.6 MỘT VÍ DỤ BẰNG SỐ

Ta minh họa lý thuyết kinh tế lượng đã được phát triển cho tới nay bởi sự xem xét hàm giả thiết Keynes đã thảo luận ở Phần Giới thiệu. Nhắc lại là Keynes đã phát biểu: “Luật

³⁷ Trong việc mô hình các hồi quy, lý thuyết nền tảng sẽ chỉ hướng của nguyên nhân giữa Y và X , đại lượng là tổng quát từ Y đến X trong các bài về các mô hình phương trình đơn giản.

tâm lý cơ bản ... là đàn ông [phụ nữ] sẽ sẵn sàng, như một quy tắc và về mặt trung bình, tăng chi tiêu khi thu nhập của họ tăng, nhưng

BẢNG 3.2

**Dữ liệu giả thiết về mức chi tiêu tiêu dùng
và thu nhập hàng tuần của một gia đình .**

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

BẢNG 3.3

Dữ liệu thô dựa trên Bảng 3.2

Y_i	X_i	$Y_i X_i$	X_i^2	$X_i - \bar{X}$	$Y_i - \bar{Y}$	X_i^2	$X_i Y_i$	$\hat{Y}_i$	$Y_i - \hat{Y}_i$	$\hat{Y}_i \hat{u}_i$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
70	80	5600	6400	-90	-41	8100	3690	65.1818	4.8181	314.0524
65	100	6500	10000	-70	-46	4900	3220	75.3636	-10.3636	-781.0382
90	120	10800	14400	-50	-21	2500	1050	85.5454	4.4545	381.0620
95	140	13300	19600	-30	-16	900	480	95.7272	-0.7272	-69.6128
110	160	17600	25600	-10	-1	100	10	105.9090	4.0909	433.2631
115	180	20700	32400	10	4	100	40	116.0909	-1.0909	-126.6434
120	200	24000	40000	30	9	900	270	125.2727	-6.2727	-792.0708
140	220	30800	48400	50	29	2500	1450	136.4545	3.5454	483.7858
155	240	37200	57600	70	44	4900	3080	145.6363	8.3636	1226.4073
150	260	39000	67600	90	39	8100	3510	156.8181	-6.8181	-1069.2014

TC	1110	1700	205500	322000	0	0	33000	16800	1109.9995	0	0.0040
									≈1110.0		≈ 0.0

TB	111	170	nc	nc	0	0	nc	nc	110	0	0
----	-----	-----	----	----	---	---	----	----	-----	---	---

$$\hat{\beta}_2 = \frac{\sum X_i Y_i}{\sum X_i^2} = \frac{16,800}{33,000} = 0.5091$$

$$\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X} = 111 - 0.5091(170) = 24.4545$$

Lưu ý : ≈ nghĩa là “xấp xỉ bằng”; nc được hiểu là “không tính được”

không nhiều như là sự gia tăng trong thu nhập”, nghĩa là xu hướng cận biên đối với người tiêu dùng (MPC) lớn hơn 0 nhưng nhỏ hơn 1. Tuy Keynes chưa chỉ rõ dạng hàm chính xác của tương quan giữa chi tiêu và thu nhập, để cho đơn giản, ta giả thiết rằng liên quan là tuyến tính như trong (2.4.2). Như cách kiểm định của hàm tiêu dùng Keynes, ta sử dụng dữ liệu mẫu trong Bảng 2.4, bảng này được tái tạo thành Bảng 3.2. Dữ liệu thô cần có để thu được ước lượng của các hệ số hồi quy, các sai số chuẩn của chúng v.v... được

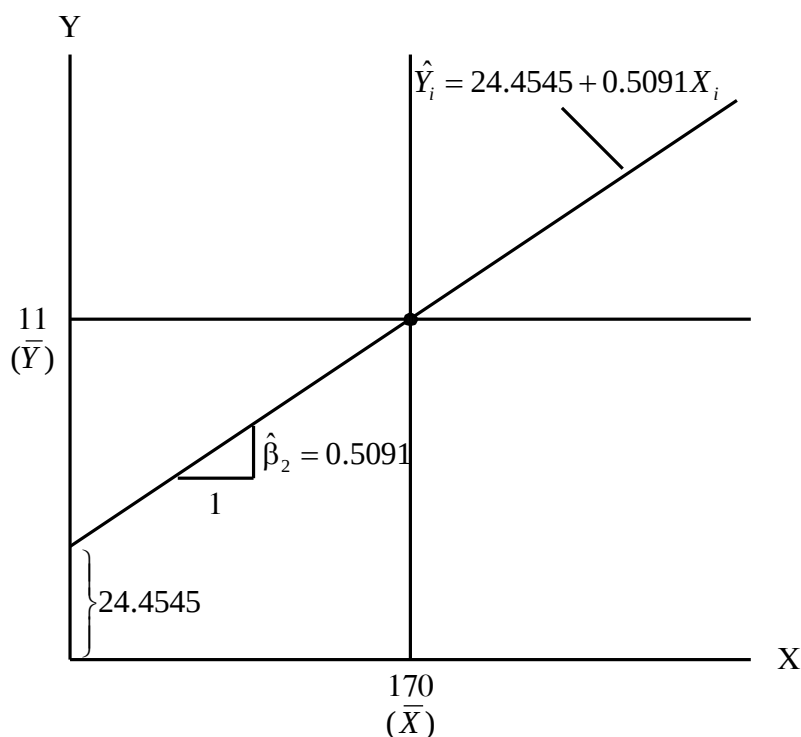
cho bởi Bảng 3.3. Dựa vào dữ liệu thô này, ta có các tính toán như sau và bạn đọc nên kiểm tra lại.

$$\begin{aligned}\hat{\beta}_1 &= 24.4545 & \text{var}(\hat{\beta}_1) &= 41.1370 & \text{và} & \text{se}(\hat{\beta}_1) &= 6.4138 \\ \hat{\beta}_2 &= 0.5091 & \text{var}(\hat{\beta}_2) &= 0.0013 & \text{và} & \text{se}(\hat{\beta}_2) &= 0.0357 \\ \text{cov}(\hat{\beta}_1, \hat{\beta}_2) &= -0.2172 & \hat{\sigma}^2 &= 42.1591 \\ r^2 &= 0.9621 & r &= 0.9809 & df &= 8\end{aligned}\quad (3.6.1)$$

Do đó đường hồi quy ước lượng là:

$$\hat{Y}_i = 24.4545 + 0.5091X_i \quad (3.6.2)$$

đã được trình bày bằng hình học trên hình 3.12



Hình 3.12
Đường hồi quy mẫu dựa trên dữ liệu Bảng 3.2

Tiếp theo Chương 2, hàm hồi quy mẫu SRF [Phương trình (3.6.2)] và đường hồi quy kết hợp được giải thích như sau: mỗi điểm trên đường hồi quy cho *một ước lượng* của giá trị trung bình hay giá trị kỳ vọng của Y tương ứng với giá trị X đã chọn; nghĩa là, $\hat{Y}_i$ là một ước lượng của $E(Y|X_i)$. Giá trị của $\hat{\beta}_2 = 0.5091$ đo độ dốc của đường, chỉ ra rằng, trong dải mẫu của các giá trị X nằm giữa 80 đô la và 260 đô la cho mỗi tuần, khi X tăng, cho là 1 đô la, thì lượng gia tăng được ước lượng một cách trung bình hàng tuần về chi tiêu tiêu dùng sẽ vào khoảng 51 cent. Giá trị $\hat{\beta}_1 = 24.4545$ là tung độ gốc của đường, chỉ mức chi tiêu tiêu dùng trung bình hàng tuần khi mà thu nhập hàng tuần bằng 0. Tuy nhiên, đây là sự giải thích một cách máy móc số hạng tung

độ gốc. Trong phép phân tích hồi quy, kiểu giải thích theo nghĩa đen của số hạng tung độ gốc như thế này không phải lúc nào cũng có ý nghĩa, mặc dù trong ví dụ hiện tại, nó có thể được lập luận rằng một gia đình không có bất cứ thu nhập nào (do thất nghiệp, do bị sa thải,...) có thể duy trì mức chi tiêu dùng tối thiểu hoặc là từ vay mượn, hoặc là từ tiết kiệm. Nhưng nói chung, người ta cần phải sử dụng độ nhạy cảm chung trong việc giải thích số hạng tung độ gốc đối với cả dải mẫu của các giá trị X vốn có thể không bao gồm số 0 như là một trong các giá trị quan sát.

Có lẽ, tốt nhất là giải thích số hạng tung độ gốc như trị trung bình hay là ảnh hưởng trung bình lên Y của tất cả các biến đã được bỏ qua từ mô hình hồi quy. Giá trị r^2 bằng 0.9621 nghĩa là khoảng 96 phần trăm độ biến thiên trong chỉ tiêu tiêu dùng hàng tuần được giải thích bởi thu nhập. Khi r^2 có thể gần như bằng 1, giá trị r^2 có được từ mẫu quan sát cho thấy rằng đường hồi quy mẫu thích hợp rất tốt với dữ liệu³⁸. Hệ số tương quan 0.9809 nói lên rằng hai biến chỉ tiêu tiêu dùng và thu nhập tương quan đồng biến cao. Các sai số mẫu được ước lượng của các hệ số tương quan sẽ được giải thích trong Chương 5.

3.7 CÁC VÍ DỤ MINH HỌA

Sự tiêu thụ Cà phê ở Mỹ năm 1970-1980

Xét dữ liệu đã cho trong Bảng 3.4³⁹

Từ môn kinh tế vi mô, ta đã biết rằng nhu cầu đối với mỗi loại hàng nói chung phụ thuộc vào giá hàng đó, các giá của các hàng khác đang cạnh tranh hay là bổ sung đối với hàng đó, và thu nhập của người tiêu dùng. Để ghép tất cả các biến này vào hàm nhu cầu, ta cho rằng dữ liệu đã có, đòi hỏi ta phải tiến tới mô hình hồi quy đa biến. Chúng ta còn chưa được chuẩn bị cho bước này. Do đó, điều mà ta sẽ làm là giả thiết một hàm cầu *riêng phần* (các yếu tố khác được giữ cho không đổi), trong đó lượng cầu chỉ liên quan với giá của chính nó. Vì lúc này, ta giả sử rằng các biến khác nhập vào hàm cầu đều là hằng số.

BẢNG 3.4

Tiêu thụ cà phê ở Mỹ (Y) trong tương quan với giá bán lẻ thực tế trung bình (X)*, 1970-1980.

Năm	Y (số tách 1 người uống mỗi ngày)	X (\$ mỗi lb)
1970	2.57	0.77
1971	2.50	0.74
1972	2.35	0.72
1973	2.30	0.73
1974	2.25	0.76
1975	2.20	0.75
1976	2.11	1.08
1977	1.94	1.81
1978	1.97	1.39
1979	2.06	1.20
1980	2.02	1.17

³⁸ Cách kiểm định chính thức đối với mức ý nghĩa của r^2 sẽ đề cập ở Chương 8.

³⁹ Tôi biết ơn Scott E. Sandberg vì đã thu thập các dữ liệu này.

*Lưu ý: giá danh nghĩa được lấy từ chỉ số giá tiêu dùng (CPI) cho thực phẩm và đồ uống, 1967=100

Nguồn: Dữ liệu Y lấy từ tóm lược của công trình nghiên cứu Quốc gia về uống Cà phê, Nhóm dữ liệu, Elkins Park, Penn., 1981 và dữ liệu về X danh nghĩa (nghĩa là X tính theo giá hiện tại) lấy từ Niealsen Food Index A.C.Nielsen, New York, 1981.

Sau đó nếu ta dùng mô hình tuyến tính hai biến để làm thích hợp với dữ liệu đã cho trong Bảng 3.4, ta thu được các kết quả như sau (bản in từ máy tính SAS cho trong phụ lục 3A, Phần 3A.7)

$$\begin{aligned}\hat{Y}_t &= 2.6911 - 0.4795X_t \\ \text{var}(\hat{\beta}_1) &= 0.0148; \text{se}(\hat{\beta}_1) = 0.1216 \\ \text{var}(\hat{\beta}_2) &= 0.0129; \text{se}(\hat{\beta}_2) = 0.01140; \hat{\sigma}^2 = 0.01656 \\ r^2 &= 0.6628\end{aligned}\tag{3.7.1}$$

Có thể giải thích hồi quy được ước lượng như sau: nếu giá bán lẻ trung bình thực tế của một pound cà phê tăng, cho là 1 đô la, lượng cà phê tiêu thụ trung bình trong ngày sẽ kỳ vọng giảm trong khoảng một nửa tách. Nếu giá cà phê đã là 0, lượng cà phê kỳ vọng tiêu thụ trung bình cho mỗi người sẽ là vào khoảng 2.69 tách trong một ngày. Đương nhiên, như đã nói trước đây, đa phần ta không thể gắn bất kỳ nghĩa vật lý nào vào tung độ gốc. Tuy nhiên, hãy nhớ rằng thậm chí nếu giá cà phê bằng 0, con người cũng không thể sử dụng lượng cà phê quá mức do các ảnh hưởng xấu của cafein tới sức khỏe. Giá trị r^2 có nghĩa là vào khoảng 66 phần trăm độ biến thiên của mức tiêu thụ cà phê cho mỗi người mỗi ngày được giải thích bởi độ biến thiên trong giá bán lẻ của cà phê.

Mô hình mà ta vừa làm thích hợp với dữ liệu có tính thực tế như thế nào? Lưu ý rằng nó không bao gồm tất cả các biến liên quan, ta không thể nói rằng nó là hàm cầu hoàn chỉnh về cà phê. Mô hình đơn giản được chọn cho ví dụ này đương nhiên chỉ là cho mục đích sư phạm tại giai đoạn này trong quá trình nghiên cứu của chúng ta. Trong Chương 7, ta sẽ giới thiệu hàm cầu hoàn chỉnh hơn. (Xem bài tập 7.23, cho ta hàm cầu về tiêu dùng gà ở Mỹ).

Hàm tiêu thụ Keynes cho Hoa Kỳ, 1980-1991.

Trở lại với dữ liệu trong Bảng I.1 của Phần Giới thiệu. Trên nền tảng của dữ liệu này, hồi quy bình phương tối thiểu thông thường OLS đã được ước lượng, trong đó Y đại diện cho chi tiêu tiêu dùng cá nhân (P.C.E) tính bằng tỷ đô la năm 1987 và X đại diện cho Tổng sản phẩm nội địa (GDP), một đại lượng đo mức thu nhập, tính bằng tỷ đô la năm 1987 (các kết quả thu được khi sử dụng **SHAZAM**TM kiểu 7.0):

$$\begin{aligned}\hat{Y}_t &= -231.80 + 0.71943X_t \\ \text{se}(\hat{\beta}_1) &= 0.9453; \text{se}(\hat{\beta}_2) = 0.02175 \\ r^2 &= 0.9909\end{aligned}\tag{3.7.2}$$

Như các kết quả này cho thấy, trong thời gian từ 1980-1991 sự chi tiêu tiêu dùng trung bình tăng trong khoảng 72 cent cho 1 đô la tăng của GDP; nghĩa là, xu hướng cận biên đối với tiêu dùng (MPC) bằng khoảng 72 cent. Giải thích theo nghĩa đen, giá trị tung độ gốc khoảng -232 cho thấy rằng khi GDP bằng 0, mức chi tiêu tiêu dùng trung bình sẽ vào khoảng -232 tỷ đô la. Một lần nữa sự giải thích máy móc như thế này về tung độ gốc không thể tạo ra cảm nhận về kinh tế nào

trong trường hợp này vì nó ở ngoài dãy giá trị mà ta quan tâm đến, và do đó nó không thể đại diện cho kết quả thực tế. Giá trị r^2 vào khoảng 0.99 nghĩa là GDP giải thích khoảng 99% của độ lệch trong chỉ tiêu tiêu dùng trung bình, đó là giá trị cao.

Tuy giá trị r^2 cao, người ta thường hay hỏi: Liệu hàm tiêu dùng Keynes đơn giản kia có phải là mô hình thích hợp để giải thích cơ cấu sự chi tiêu tiêu dùng ở Mỹ. Đôi khi, các mô hình hồi quy rất đơn giản (2 biến) có thể cho thông tin bổ ích. Các ước lượng của xu hướng cận biên đối với tiêu dùng (MPC) cho Hoa Kỳ dựa trên các mô hình phức tạp cũng chỉ ra rằng MPC vào khoảng 0.7. Nhưng ta sẽ phải nói nhiều hơn về mô hình đầy đủ trong chương sau.

3.8. KẾT QUẢ CỦA MÁY VI TÍNH ĐỐI VỚI HÀM CẦU VỀ CÀ PHÊ

Như đã lưu ý ở phần giới thiệu, trong suốt cuốn sách này, ta sẽ sử dụng máy vi tính nhiều để trả lời cho các ví dụ minh họa để cho bạn đọc quen với một vài chương trình hồi quy. Trong phụ lục C ta sẽ thảo luận chi tiết về vài chương trình này. Các ví dụ minh họa trong cuốn sách này sẽ sử dụng một hay vài chương trình này. Đối với hàm cầu cà phê, kết quả máy vi tính SAS được trình bày trong Phụ lục 3A, Phần 3A.7.

3.9 LƯU Ý VỀ CÁC THỬ NGHIỆM MONTE CARLO

Trong chương này, ta đã rõ rằng dưới các giả thiết về các mẫu hồi quy tuyến tính cổ điển các hàm ước lượng bình phương tối thiểu có các đặc tính thống kê mong muốn nhất định được tóm lược trong tính chất BLUE. Trong Phụ lục của Chương này, ta sẽ chứng minh tính chất này một cách chính thức hơn. Nhưng trong thực tế, làm sao người ta biết các tính chất trên áp dụng như thế nào? Ví dụ như làm thế nào để tìm ra các hàm ước lượng bình phương tối thiểu thông thường là không bị thiên lệch? Lời giải đáp cho câu hỏi này là cái gọi là các thử nghiệm Monte Carlo, về bản chất đó là các mô phỏng hay lấy mẫu hay thử nghiệm bằng máy vi tính.

Để giới thiệu ý tưởng cơ bản, ta xét hàm hồi quy tổng thể hai biến PRF:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \quad (3.9.1)$$

Thử nghiệm Monte Carlo tiến hành như sau:

1. Giả sử rằng các giá trị thực của các thông số như sau: $b_1 = 20$ và $b_2 = 0.6$.
2. Bạn chọn cỡ mẫu n , cho là $n = 25$
3. Bạn cố định các giá trị của X cho mỗi quan sát. Trong tất cả các quan sát đó bạn có 25 giá trị X .
4. Giả sử bạn lấy một bảng số ngẫu nhiên, chọn 25 giá trị, và gọi chúng là u_i (hiện nay hầu hết các phần mềm thống kê đều có xây dựng các bộ phận phát số ngẫu nhiên)⁴⁰.
5. Khi đã biết b_1 , b_2 , X_i và u_i , sử dụng (3.9.1) ta sẽ có 25 giá trị Y_i
6. Bây giờ, ta sử dụng 25 giá trị Y_i đã sinh ra, ta hồi quy chúng trên 25 giá trị X đã được chọn ở bước 3, thu được $\hat{\beta}_1$ và $\hat{\beta}_2$ các hàm ước lượng bình phương tối thiểu.
7. Giả sử bạn lặp lại thí nghiệm này 99 lần, mỗi lần lại sử dụng cùng giá trị b_1 , b_2 và các X . Đương nhiên, các giá trị u_i sẽ khác nhau tại các thí nghiệm khác nhau. Do đó bạn có tất cả

⁴⁰ Trong thực tế, người ta cho rằng u_i tuân theo phân phối xác suất nào đó, giả sử là phân phối chuẩn, với các thông số nào đó (ví dụ như trị trung bình và phương sai). Một khi các thông số đã được xác định, người ta có thể dễ dàng phát ra u_i khi sử dụng các phần mềm thống kê.

100 thí nghiệm thì chúng sinh ra 100 giá trị của mỗi b_1 và b_2 . (Trong thực tế, nhiều thí nghiệm như thế này đã được thực hiện, có khi tới 1000 hay 2000)

8. Bạn lấy trung bình cộng của 100 ước lượng này và gọi chúng là $\bar{\beta}_1$ và $\bar{\beta}_2$.
9. Nếu các giá trị trung bình này gần giống với các giá trị thực của b_1 và b_2 đã giả thiết trong bước 1, thử nghiệm Monte Carlo này “thiết lập” rằng các hàm ước lượng bình phương tối thiểu là không thiên lệch. Nhắc lại rằng dưới mô hình hồi quy tuyến tính cổ điển $E(\hat{\beta}_1) = \beta_1$ và $E(\hat{\beta}_2) = \beta_2$.

Các bước này đặc trưng cho bản chất chung của các thử nghiệm Monte Carlo. Các thử nghiệm kiểu này thường được sử dụng để nghiên cứu các tính chất thống kê của các phương pháp ước lượng thông số tổng thể khác nhau. Chúng rất có ích để nghiên cứu diễn biến các hàm ước lượng trong các mẫu nhỏ hay các mẫu hữu hạn. Các thử nghiệm này cũng là phương tiện cực kỳ tốt để đưa về khái niệm **lấy mẫu lặp lại**, nền tảng của hầu hết các kết luận thống kê cổ điển, như ta sẽ thấy trong Chương 5. Chúng tôi sẽ cung cấp nhiều ví dụ thử nghiệm Monte Carlo trong các bài tập cho trên lớp. (Xem bài tập 3.26).

3.10 TÓM TẮT VÀ KẾT LUẬN:

Các đề mục và khái niệm quan trọng được phát triển trong chương này có thể được tóm tắt lại như sau:

1. Cái khung cơ bản của phép phân tích hồi quy là **mô hình hồi quy tuyến tính cổ điển (CRLM)**.
2. Các mô hình hồi quy tuyến tính cổ điển dựa trên một tập hợp các giả thiết.
3. Dựa trên các giả thiết này, các hàm ước lượng bình phương tối thiểu có các tính chất nhất định, các tính chất này đã được tóm tắt trong định lý Gauss-Markov, phát biểu rằng trong nhóm các hàm ước lượng không thiên lệch tuyến tính, các hàm ước lượng các bình phương tối thiểu có phương sai nhỏ nhất. Ngắn gọn hơn, chúng là các hàm ước lượng không thiên lệch tuyến tính tốt nhất (BLUE).
4. *Tính chính xác* của các hàm bình phương tối thiểu thông thường OLS được đo bởi các **sai số chuẩn**. Trong Chương 4 và 5 ta sẽ thấy các sai số chuẩn đưa ta đến việc rút ra các suy diễn về các thông số tổng thể, các hệ số b , như thế nào.
5. Độ thích hợp toàn diện của mô hình hồi quy được đo bởi **hệ số xác định r^2** . Nó chỉ ra tỷ lệ mà độ biến thiên trong mỗi biến phụ thuộc hay là các biến hồi qui phụ thuộc được giải thích bởi các biến giải thích, bởi các biến hồi qui độc lập. Đại lượng r^2 này nằm giữa 0 và 1; r^2 càng gần tới 1 độ thích hợp càng tốt.
6. Khái niệm liên quan đến hệ số xác định là **hệ số tương quan r** . Nó là đại lượng đo *sự kết hợp tuyến tính* giữa hai biến và nó nằm giữa -1 và $+1$.

BẢNG 3.5

Điều gì sẽ xảy ra nếu các giả thiết về mô hình hồi quy tuyến tính cổ điển bị vi phạm?

Số thứ tự của giả thiết	Loại vi phạm	Nghiên cứu ở đâu ?
1	Phi tuyến tính trong các thông số	Không có trong sách này
2	biến hồi qui độc lập ngẫu nhiên	Giới thiệu cho Phần II
3	Giá trị trung bình của u_i khác 0	Giới thiệu cho Phần II
4	Phương sai của sai số thay đổi	Chương 11
5	Các nhiễu tự tương quan	Chương 12
6	Đồng phương sai của các nhiễu và biến hồi qui độc	Giới thiệu cho Phần II và Phần IV

	lập khác 0	
7	Các quan sát mẫu nhỏ hơn số các biến hồi qui độc lập	Chương 10
8	Tính biến thiên không hiệu quả trong các các biến hồi qui độc lập	Chương 10
9	Độ thiên lệch đặc trưng	Chương 13,14
10	Đa cộng tuyến	Chương 10
11*	Tính không theo qui luật chuẩn của các nhiễu	Giới thiệu cho Phần I

* Lưu ý: Giả thiết rằng các nhiễu u_i phân phối chuẩn không phải là một phần của CLRM. Nhưng có thể biết nhiều hơn trong Chương 4.

7. Mô hình hồi quy tuyến tính cổ điển CLRM là phép xây dựng lý thuyết hay là sự trừu tượng bởi vì nó dựa trên tập hợp các giả thiết có thể là rất nghiêm ngặt hoặc “không thực tế”. Nhưng phép trừu tượng thể này là luôn cần thiết trong các giai đoạn đầu tiên bước vào con đường tìm hiểu bất cứ lĩnh vực nào của kiến thức. Một khi CLRM đã được lập ra, người ta có thể biết được điều gì xảy ra nếu một hoặc vài giả thiết của nó không được thỏa mãn. Phần đầu của cuốn sách này giành cho việc tìm hiểu mô hình hồi quy tuyến tính cổ điển. Trong các phần khác của sách là việc xem xét các cải tiến của CLRM. Bảng 3.5 cho ta bản đồ đường đi tới phía trước.

BÀI TẬP

Các câu hỏi

- 3.1. Cho trước các giả thiết trong cột 1 của bảng sau, hãy chỉ ra rằng các giả thiết trong cột 2 là tương đương với chúng

Các giả thiết của mô hình cổ điển	
(1)	(2)
$E(u_i X_i) = 0$	$E(Y_i X_i) = \beta_0 + \beta_1 X_i$
$\text{cov}(u_i, u_j) = 0, i \neq j$	$\text{cov}(Y_i, Y_j) = 0, i \neq j$
$\text{var}(u_i X_i) = \sigma^2$	$\text{var}(Y_i X_i) = \sigma^2$

- 3.2. Chứng minh rằng các ước lượng $\hat{\beta}_1 = 1.572$ và $\hat{\beta}_2 = 1.357$ đã được sử dụng trong thử nghiệm 1 ở Bảng 3.1 chính là các hàm ước lượng bình phương tối thiểu thông thường OLS.
- 3.3. Theo Malinvaud (xem chú thích 11), các giả thiết $E(u_i | X_i) = 0$ là vô cùng quan trọng. Để thấy điều đó, xét hàm hồi qui tổng thể PRF $Y_i = b_1 + b_2 X_i + u_i$. Bây giờ hãy xét 2 trường hợp: (i) $b_1 = 0, b_2 = 1$; và $E(u_i) = 0$; và (ii) $b_1 = 1, b_2 = 0$; và $E(u_i) = (X_i - 1)$. Bây giờ ta hãy lấy giá trị dự tính của hàm PRF có điều kiện theo với X trong 2 trường hợp trên và ta xem liệu có thể đồng ý với Malinvaud về ý nghĩa của giả thiết $E(u_i | X_i) = 0$ hay không.
- 3.4. Xét hồi quy mẫu:

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

Đặt các giới hạn (i) $\sum \hat{u}_i = 0$ và (ii) $\sum \hat{u}_i X_i = 0$, xác định được các hàm ước lượng $\hat{\beta}_1$ và $\hat{\beta}_2$ và chỉ ra rằng chúng là đồng nhất với các hàm ước lượng bình phương tối thiểu cho trong (3.1.6) và (3.1.7). Phương pháp xác định các hàm ước lượng này được gọi là **nguyên tắc tương tự**. Hãy cho một sự biện giải bằng trực giác đối với các giới hạn được áp đặt (i) và (ii). (Gợi ý: Hãy nhắc lại các giả thiết mô hình hồi quy tuyến tính cổ điển CLRM về u_i). Nhân đây, lưu ý rằng nguyên tắc tương đồng về việc ước lượng các thông số chưa biết cũng được gọi là **phương pháp các momen** mà trong đó các momen mẫu (ví dụ trung bình mẫu) được sử dụng để ước lượng các momen tổng thể (ví dụ trung bình tổng thể). Như được lưu ý ở Phụ lục A, momen là một trị thống kê tổng hợp của phân phối xác suất, như là các giá trị kỳ vọng và phương sai.

- 3.5. Chứng tỏ rằng r^2 được định nghĩa trong (3.5.5) biến đổi giữa 0 và 1. Bạn có thể sử dụng bất đẳng thức Cauchy-Schwartz, cho rằng đối với các biến X và Y ngẫu nhiên mỗi quan hệ sau là đúng:

$$[E(XY)]^2 \leq E(X^2)E(Y^2)$$

- 3.6. Gọi $\hat{\beta}_{YX}$ và $\hat{\beta}_{XY}$ là các độ dốc trong hồi quy tương ứng của Y trên X và của X trên Y . Hãy chỉ ra rằng :

$$\hat{\beta}_{YX}\hat{\beta}_{XY} = r^2$$

trong đó r là hệ số tương quan giữa X và Y .

- 3.7. Trong bài tập 3.6 giả sử rằng $\hat{\beta}_{YX}\hat{\beta}_{XY} = 1$. Điều gì sẽ xảy ra nếu ta hồi quy Y trên X hay là hồi quy X trên Y ? Hãy giải thích một cách chi tiết.
- 3.8. Hệ số tương quan dãy sắp hạng của Spearman r_s được định nghĩa như sau:

$$r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

trong đó d = khác biệt trong các hạng được quy cho cá thể hay hiện tượng giống nhau.

n = số lượng các cá thể hay là các hiện tượng được sắp hạng
 r_s được lấy từ r đã được xác định trong (3.5.13). Gợi ý: Sắp hạng các giá trị X và Y từ 1 đến n . Lưu ý rằng tổng của các hạng của mỗi X và Y là $n(n+1)/2$ và do đó các giá trị trung bình của chúng là $(n+1)/2$.

- 3.9. Xét các công thức sau của hàm PRF hai biến:

$$\text{Mô hình 1: } Y_i = \beta_1 + \beta_2 X_i + u_i$$

$$\text{Mô hình 2: } Y_i = \alpha_1 + \alpha_2 (X_i - \bar{X}) + u_i$$

- Tìm các hàm ước lượng b_1 và α_1 . Chúng có đồng nhất không? Phương sai của chúng có đồng nhất không?
- Tìm các hàm ước lượng b_2 và α_2 . Chúng có đồng nhất không? Các phương sai của chúng có đồng nhất không?
- Mô hình II có lợi thế đối với mô hình 1 không? Nếu có thì đó là lợi thế gì?

3.10. Giả sử bạn đang tiến hành hồi quy sau:

$$y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \hat{u}_i$$

trong đó, như thường lệ, y_i và x_i là các độ lệch so với các giá trị trung bình tương ứng của chúng. Giá trị $\hat{\beta}_1$ sẽ như thế nào? Tại sao? $\hat{\beta}_2$ có giống như đại lượng thu được từ phương trình (3.1.6) không? Tại sao?

3.11. Cho r_1 = hệ số tương quan giữa n cặp giá trị (Y_i, X_i) và r_2 = hệ số tương quan giữa n cặp giá trị $(aX_i + b, cY_i + d)$ trong đó a, b, c, d là hằng số. Chứng tỏ rằng $r_1 = r_2$ và do đó hãy thiết lập nguyên tắc cho rằng hệ số tương quan là bất biến theo sự thay đổi của thang tỷ lệ và thay đổi của gốc tọa độ.

Gợi ý: Ứng dụng định nghĩa về r cho trong (3.5.13).

Lưu ý: Các toán tử $aX_i, X_i + b$ và $aX_i + b$ được gọi tương ứng là thay đổi về thang tỷ lệ, thay đổi gốc tọa độ và thay đổi cả thang tỷ lệ lẫn gốc tọa độ.

3.12. Nếu r , hệ số tương quan giữa n cặp giá trị (X_i, Y_i) là dương, thì hãy xác định các phát biểu sau đây là đúng hay sai:

(a) r giữa $(-X_i, -Y_i)$ cũng có giá trị dương.

(b) r giữa $(-X_i, Y_i)$ và r giữa $(X_i, -Y_i)$ có thể hoặc dương hoặc âm.

(c) Cả hai hệ số độ dốc của b_{yx} và b_{xy} đều có giá trị dương, trong đó b_{yx} = hệ số độ dốc trong hồi quy của Y trên X và b_{xy} = hệ số độ dốc trong hồi quy của X trên Y .

3.13. Nếu X_1, X_2 và X_3 là các biến không tương quan mỗi biến có độ lệch chuẩn như nhau. Hãy chứng tỏ rằng hệ số tương quan giữa $X_1 + X_2$ và $X_2 + X_3$ bằng $1/2$. Tại sao hệ số tương quan lại khác 0?

3.14. Trong hồi quy $Y_i = b_1 + b_2 X_i + u_i$ giả sử rằng ta đã nhân giá trị của mỗi X với hằng số, giả sử là 2. Nó có làm thay đổi các phần dư và giá trị Y không? Giải thích. Sẽ ra sao nếu ta thêm giá trị hằng số, cho là 2, vào mỗi giá trị X ?

3.15. Hãy chứng tỏ rằng (3.5.14) thực chất là đo hệ số xác định. Gợi ý: Áp dụng định nghĩa r cho bởi (3.5.13) và nhắc lại rằng $\sum y_i \hat{y}_i = \sum (\hat{y}_i + \hat{u}_i) \hat{y}_i = \sum \hat{y}_i^2$ và nhớ biểu thức (3.5.6)

Các vấn đề

3.16. Bạn đã được cho dãy sắp hạng điểm thi giữa kỳ và cuối kỳ của 10 sinh viên về môn thống kê. Hãy tính hệ số Spearman's của tương quan sắp hạng và giải thích.

Dãy	Sinh viên									
	A	B	C	D	E	F	G	H	I	J
Giữa Kỳ	1	3	7	10	9	5	4	8	2	6
Cuối Kỳ	3	2	8	7	9	6	5	10	1	4

3.17. Bảng sau đây cho biết dữ liệu về tỷ lệ bỏ việc với mỗi 100 công nhân trong sản xuất và tỷ lệ thất nghiệp trong sản xuất, ở Hoa Kỳ trong giai đoạn 1960-1972.

Lưu ý: thuật ngữ “bỏ việc” để chỉ những người tự nguyện bỏ việc.

**Tỷ lệ thất nghiệp và bỏ việc trong sản xuất
ở Hoa Kỳ, năm 1960-1972.**

Năm	Tỷ lệ bỏ việc trong 100 công nhân, Y	Tỷ lệ thất nghiệp (%), X
1960	1.3	6.2
1961	1.2	7.8
1962	1.4	5.8
1963	1.4	5.7
1964	1.5	5.0
1965	1.9	4.0
1966	2.6	3.2
1967	2.3	3.6
1968	2.5	3.3
1969	2.7	3.3
1970	2.1	5.6
1971	1.8	6.8
1972	2.2	5.6

Nguồn: Báo cáo nguồn nhân lực của tổng thống, 1973, Các Bảng C-10 và A-18.

- (a) Vẽ các dữ liệu lên đồ thị phân tán.
 (b) Giả sử rằng tỷ lệ bỏ việc Y tương quan tuyến tính với tỷ lệ thất nghiệp X như là $Y_i = b_1 + b_2 X_i + u_i$. Xác định b_1 , b_2 và các sai số chuẩn của chúng.
 (c) Tính r^2 và r .
 (d) Giải thích các kết quả của bạn.
 (e) Vẽ các phần dư $\hat{u}_i$. Bạn rút ra điều gì từ các phần dư này?
 (f) Bằng cách sử dụng số liệu hàng năm cho giai đoạn 1966-1978 và bằng cách sử dụng mô hình như trong (b) ở trên, ta có thể thu được kết quả sau :

$$\hat{Y}_i = 3.1237 - 0.1714X_i$$

$$se(\hat{\beta}_2) = 0.0210 \quad \text{và} \quad r^2 = 0.8575$$

Nếu các kết quả này không giống với những gì ta có ở (b), bạn có thể giải thích như thế nào cho sự khác biệt ấy.

3.18. Dựa trên một mẫu của 10 quan sát, ta có các kết quả sau:

$$\begin{aligned} \sum Y_i &= 1110 & \sum X_i &= 1700 & \sum X_i Y_i &= 205,500 \\ \sum X_i^2 &= 322,000 & \sum Y_i^2 &= 132,100 \end{aligned}$$

với hệ số tương quan $r = 0.9758$. Nhưng khi kiểm tra lại các tính toán này, ta thấy hai cặp quan sát được ghi như sau :

Y	X		Y	X
90	120	thay vì	80	110
140	220		150	200

Sai số này sẽ ảnh hưởng như thế nào lên r ? Hãy tìm r đúng.

- 3.19. Bảng sau đây cho ta dữ liệu về giá vàng, chỉ số giá tiêu dùng (CPI), và chỉ số trao đổi cổ phiếu ở New York (NYSE) ở Mỹ cho giai đoạn 1977-1991. Chỉ số NYSE bao gồm hầu hết các cổ phiếu liệt kê trong các NYSE, có khoảng 1500 giá trị.

Năm	Giá vàng ở New York \$ cho 1 troy ounce	Chỉ số giá tiêu thụ (CPI), 1982-84=100	Chỉ số trao đổi cổ phiếu New York (NYSE), 31 tháng 12-1965=100
1977	147.98	60.6	53.69
1978	193.44	65.2	53.70
1979	307.62	72.6	58.32
1980	612.51	82.4	68.10
1981	459.61	90.9	74.02
1982	376.01	96.5	68.93
1983	423.83	99.6	92.63
1984	360.29	103.9	92.46
1985	317.30	107.6	108.90
1986	367.87	109.6	136.00
1987	446.50	113.6	161.70
1988	436.93	118.3	149.91
1989	381.28	124.0	180.02
1990	384.08	130.7	183.46
1991	362.04	136.2	206.33

Nguồn: Dữ liệu trên chỉ số CPI và NYSE lấy từ báo cáo kinh tế của tổng thống, tháng 1/93. tương ứng bảng B-59 và B-91. Giá vàng thì lấy từ Phòng Thương mại Hoa Kỳ, Văn phòng phân tích Kinh tế, *Thống kê kinh doanh*, 1963-1991, trang 68.

- (a) Trên cùng một đồ thị phân tán, vẽ đồ thị chỉ số giá vàng, CPI và NYSE.
(b) Một việc đầu tư được cho là hàng rào ngăn lạm phát nếu giá vàng hay là suất thu lợi của việc đầu tư ít ra cũng kim giữ được nhịp độ lạm phát. Để kiểm định giả thiết này, giả sử bạn quyết định làm thích hợp bằng mô hình sau đây, và giả thiết rằng đồ thị trong (a) gợi ý rằng mô hình thích hợp là:

$$\begin{aligned}\text{Giá vàng}_t &= \beta_1 + \beta_2 \text{CPI}_t + u_t \\ \text{Chỉ số NYSE}_t &= \beta_1 + \beta_2 \text{CPI}_t + u_t\end{aligned}$$

Nếu giả thiết là đúng, ta có thể kỳ vọng gì về giá trị β_2 .

- (c) Hàng rào nào chống lại lạm phát tốt hơn? Giá vàng hay thị trường chứng khoán?

- 3.20. Làm cho mô hình tuyến tính thích hợp với các dữ liệu tương quan đến chỉ số giá tiêu dùng và cung tiền ở Nhật cho giai đoạn quý 1/1988 đến quý 2/1992, và bình luận các kết quả thu được của bạn.

Giá tiêu dùng và cung tiền ở Nhật cho giai đoạn quý 1/1988 đến quý 3/1992

Năm và quý	CPI Chỉ số giá tiêu dùng (1985 = 100)	Lượng tiền (M_1) (tỷ yên)
1988-1	101.0	101,587
1988-2	101.1	102,258
1988-3	101.6	104,653
1988-4	102.1	107,561
1989-1	102.1	109,525
1989-2	103.7	108,442
1989-3	104.4	109,176
1989-4	104.7	107,660

1990-1	105.7	111,600
1990-2	106.3	111,929
1990-3	107.1	112,753
1990-4	108.5	112,155
1991-1	109.7	113,150
1991-2	109.9	115,827
1991-3	110.5	120,718
1991-4	111.5	125,891
1992-1	111.7	123,589
1992-2	112.4	125,583
1992-3	112.5	126,816

Nguồn: Ngân hàng dự trữ liên bang của St.Louis, *Các điều kiện Kinh tế Quốc tế* tháng 2/1993, trang 26,28.

- 3.21.** Bảng sau đây cho dữ liệu về số lượng máy điện thoại cho 1000 người (Y) và cho tổng sản phẩm nội địa theo đầu người (GDP), tại mức giá cơ cấu (X) (tính theo đồng đô la Singapore năm 1968), ở Singapore trong khoảng thời gian 1960-1981. Có mối quan hệ gì giữa hai biến trên hay không? Làm thế nào để bạn biết được?

**Sự sở hữu máy điện thoại và chỉ số GDP theo đầu người
Tại Singapore, 1960-1981**

Năm	Y	X	Năm	Y	X
1960	36	1299	1971	90	2723
1961	37	1365	1972	102	3033
1962	38	1409	1973	114	3317
1963	41	1549	1974	126	3487
1964	42	1416	1975	141	3575
1965	45	1473	1976	163	3784
1966	48	1589	1977	196	4025
1967	54	1757	1978	223	4286
1968	59	1974	1979	262	4628
1969	67	2204	1980	291	5038
1970	78	2462	1981	317	5472

Nguồn: Lim Chong-Yah, *Economic Restructuring in Singapore* (Cấu trúc lại Kinh tế ở Singapore), Federal Publications, Pvt Ltd., 1984, trang 110-113

- 3.22.** Bảng sau cho biết tổng giá trị sản phẩm nội địa (GDP) ở Hoa Kỳ cho các năm 1972-1991

**Tổng giá trị sản phẩm nội địa (GDP) tính theo đô la hiện hành
và đô la 1987, năm 1972-1991**

Năm	GDP (đô la hiện hành, tỷ)	GDP (đô la 1987, tỷ)
1972	1207.0	3107.1
1973	1349.6	3268.6
1974	1458.6	3248.1
1975	1585.9	3221.7
1976	1768.4	3380.8
1977	1974.1	3533.3
1978	2232.7	3703.5
1979	2488.6	3796.8
1980	2708.0	3776.3

1981	3030.6	3843.1
1982	3149.6	3760.3
1983	3405.0	3906.6
1984	3777.2	4148.5
1985	4038.7	4279.8
1986	4268.6	4404.5
1987	4539.9	4539.9
1988	4900.4	4718.6
1989	5250.8	4838.0
1990	5522.2	4877.5
1991	5677.5	4821.0

Nguồn: Báo cáo Kinh tế của Tổng thống, tháng 1/1993, Bảng B-1 và B-2, trang 348-349.

- (a) Vẽ đồ thị GDP bằng đô la hiện hành và đô la không đổi (năm 1987) theo thời gian.
 (b) Gọi Y là GDP, X là thời gian (theo chiều thời gian bắt đầu từ 1 cho năm 1972, 2 cho 1973 cho đến 20 cho năm 1991), và hãy xem mô hình sau có thích hợp với dữ liệu GDP không:

$$Y_t = \beta_1 + \beta_2 X_t + u_t$$

Ước lượng mô hình này cho cả GDP theo đô la hiện hành và đô la không đổi.

- (c) Bạn có thể giải thích β_2 như thế nào?
 (d) Nếu có một sự khác biệt giữa β_2 được ước lượng cho GDP theo đô la hiện hành và β_2 ước lượng cho GDP đô la không đổi, Điều gì giải thích sự khác biệt đó?
 (e) Từ kết quả của mình, bạn có thể nói gì về bản chất của sự lạm phát ở Hoa Kỳ qua những thập niên mẫu?

3.23. Dùng dữ liệu cho ở Bảng I.1 của Phần giới thiệu, kiểm chứng lại phương trình (3.7.2)

3.24. Với ví dụ S.A.T cho trong bài 2.16, hãy thực hiện những công việc sau:

- (a) Vẽ đồ thị thể hiện điểm văn đáp của nữ theo điểm văn đáp của nam.
 (b) Nếu đồ thị phân tán gợi ý rằng quan hệ tuyến tính giữa hai đại lượng hầu như thích hợp, hãy tìm hồi quy của điểm văn đáp của nữ trên điểm văn đáp của nam.
 (c) Nếu có một mối liên hệ giữa hai điểm văn đáp, thì đây có phải là quan hệ *nhân quả* không?

3.25. Cũng giống như bài tập 3.24 nhưng thay điểm văn đáp bằng điểm Toán.

3.26. Bài tập trên lớp về nghiên cứu Monte Carlo:

Tham khảo 10 giá trị X đã cho trên Bảng 3.2, coi $\beta_1 = 25$ và $\beta_2 = 0.5$. Giả sử $u_i \sim N(0,9)$, nghĩa là u_i tuân theo phân phối chuẩn với giá trị trung bình bằng 0 và phương sai bằng 9. Hãy phát ra 100 mẫu, bằng cách sử dụng các giá trị này sẽ thu được 100 ước lượng của β_1 và β_2 . Vẽ đồ thị các giá trị ước lượng này. Bạn rút ra được kết luận gì từ nghiên cứu Monte Carlo? Lưu ý: Hầu hết các phần mềm thống kê hiện nay đều có thể phát ra các biến ngẫu nhiên từ hầu hết các phân phối xác suất nổi tiếng. Hãy nhờ thầy hướng dẫn bạn trong trường hợp bạn gặp khó khăn khi muốn tạo ra các biến như thế.

PHỤ LỤC 3A..**3A.1 ĐẠO HÀM CỦA CÁC ƯỚC LƯỢNG BÌNH PHƯƠNG TỐI THIỂU.**

Lấy vi phân (3.1.2) từng phần theo $\hat{\beta}_1$ và $\hat{\beta}_2$, ta có:

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_1} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) = -2 \sum \hat{u}_i \quad (1)$$

$$\frac{\partial(\sum \hat{u}_i^2)}{\partial \hat{\beta}_2} = -2 \sum (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i) X_i = -2 \sum \hat{u}_i X_i \quad (2)$$

Cho phương trình này bằng 0, sau các quá trình biến đổi đại số, sẽ cho ta các hàm ước lượng đã cho trong phương trình (3.1.6) và (3.1.7)

3A.2. CÁC TÍNH CHẤT TUYẾN TÍNH VÀ KHÔNG THIÊN LỆCH CỦA CÁC HÀM ƯỚC LƯỢNG BÌNH PHƯƠNG TỐI THIỂU

Từ (3.1.8) ta có:

$$\hat{\beta}_2 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i \quad (3)$$

trong đó:

$$k_i = \frac{x_i}{(\sum x_i^2)}$$

chứng tỏ rằng $\hat{\beta}_2$ là **hàm ước lượng tuyến tính** bởi vì đó là hàm tuyến tính của Y ; thực ra nó là trung bình trọng số của Y_i với k_i đóng vai trò như là trọng số. Bằng cách tương tự nó có thể được chỉ ra rằng $\hat{\beta}_1$ cũng là hàm ước lượng tuyến tính.

Nhân đây, hãy lưu ý các tính chất của các trọng số k_i :

1. Vì X_i giả sử là không ngẫu nhiên, k_i cũng là không ngẫu nhiên.
2. $\sum k_i = 0$
3. $\sum k_i^2 = 1 / \sum x_i^2$
4. $\sum k_i x_i = \sum k_i X_i = 1$. Các tính chất này có thể được kiểm chứng trực tiếp từ định nghĩa của k_i .

Ví dụ:

$$\sum k_i = \sum \left(\frac{x_i}{\sum x_i^2} \right) = \frac{1}{\sum x_i^2} \sum x_i, \quad \text{vì đối với một mẫu cho trước, } \sum x_i^2 \text{ đã biết}$$

$$= 0, \quad \text{vì } \sum x_i \text{ (tổng các độ lệch đối với giá trị trung bình) luôn bằng 0.}$$

Bây giờ ta thế hàm hồi quy tổng thể $Y_i = \beta_1 + \beta_2 X_i + u_i$ vào (3) để thu được

$$\begin{aligned} \hat{\beta}_2 &= \sum k_i (\beta_1 + \beta_2 X_i + u_i) \\ &= \beta_1 \sum k_i + \beta_2 \sum k_i X_i + \sum k_i u_i \\ &= \beta_2 + \sum k_i u_i \end{aligned} \quad (4)$$

trong đó ứng dụng các tính chất của k_i như đã lưu ý trước đây.

Bây giờ lấy kỳ vọng của (4) trên cả 2 vế và lưu ý rằng k_i không ngẫu nhiên, nó có thể được xử lý như một hằng số, ta có:

$$\begin{aligned} E(\hat{\beta}_2) &= \beta_2 + \sum k_i E(u_i) \\ &= \beta_2 \end{aligned} \quad \begin{array}{l} \text{vì} \\ \text{sử dụng phương trình (4) ở} \\ \text{trên} \end{array} \quad (5)$$

vì theo giả thiết $E(u_i) = 0$. Do đó $\hat{\beta}_2$ là hàm ước lượng không thiên lệch của β_2 . Tương tự như vậy, có thể chứng minh rằng $\hat{\beta}_1$ cũng là hàm ước lượng không thiên lệch của β_1 .

3A.3 CÁC PHƯƠNG SAI VÀ SAI SỐ CHUẨN CỦA CÁC HÀM ƯỚC LƯỢNG BÌNH PHƯƠNG TỐI THIỂU

Bây giờ, theo định nghĩa phương sai, ta viết:

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= E[\hat{\beta}_2 - E(\hat{\beta}_2)]^2 \\ &= E(\hat{\beta}_2 - \beta_2)^2, \quad E(\hat{\beta}_2) = \beta_2 \\ &= E\left(\sum k_i u_i\right)^2, \\ &= E(k_1^2 u_1^2 + k_2^2 u_2^2 + \dots + k_n^2 u_n^2 + 2k_1 k_2 u_1 u_2 + \dots + 2k_{n-1} k_n u_{n-1} u_n) \end{aligned} \quad (6)$$

Vì theo giả thiết, $E(u_i^2) = \sigma^2$ cho mỗi i và $E(u_i, u_j) = 0, i \neq j$, nó tiếp theo rằng

$$\begin{aligned} \text{var}(\hat{\beta}_2) &= \sigma^2 \sum k_i^2 \\ &= \frac{\sigma^2}{\sum x_i^2}, \quad \text{sử dụng định nghĩa của } k_i^2 \\ &= \text{phương trình (3.3.1)} \end{aligned} \quad (7)$$

Phương sai của $\hat{\beta}_1$ có thể tính được khi tuân theo lý luận giống như trên. Một khi xác định được các phương sai của $\hat{\beta}_1$ và $\hat{\beta}_2$, các căn bậc hai dương của chúng sẽ cho ta các sai số chuẩn tương ứng.

3A.4. ĐỒNG PHƯƠNG SAI GIỮA $\hat{\beta}_1$ VÀ $\hat{\beta}_2$

Từ định nghĩa :

$$\begin{aligned}\text{cov}(\hat{\beta}_1, \hat{\beta}_2) &= E\left\{\left[\hat{\beta}_1 - E(\hat{\beta}_1)\right]\left[\hat{\beta}_2 - E(\hat{\beta}_2)\right]\right\} \\ &= E(\hat{\beta}_1 - \beta_1)(\hat{\beta}_2 - \beta_2) \quad (\text{Vì sao?}) \\ &= -\bar{X}E(\hat{\beta}_2 - \beta_2)^2 \\ &= -\bar{X} \text{var}(\hat{\beta}_2) \\ &= \text{phương trình (3.3.9)}\end{aligned}\tag{8}$$

trong đó ứng dụng các dữ kiện là $\hat{\beta}_1 = \bar{Y} - \hat{\beta}_2 \bar{X}$ và $E(\hat{\beta}_1) = \bar{Y} - \hat{\beta}_2 \bar{X}$ khi cho, $\hat{\beta}_1 - E(\hat{\beta}_1) = -\bar{X}(\hat{\beta}_2 - \beta_2)$. Lưu ý: $\text{var}(\hat{\beta}_2)$ đã cho trong (3.3.1)

3A.5. HÀM ƯỚC LƯỢNG BÌNH PHƯƠNG TỐI THIỂU CỦA σ^2

Nhắc lại rằng:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{9}$$

Do đó:

$$\bar{Y} = \beta_1 + \beta_2 \bar{X} + \bar{u} \tag{10}$$

Lấy (9) trừ đi (10) ta có:

$$y_i = \beta_2 x_i + (u_i - \bar{u}) \tag{11}$$

Cũng nhắc lại rằng:

$$\hat{u}_i = y_i - \hat{\beta}_2 x_i \tag{12}$$

Do đó, khi thế (11) vào (12) ta được:

$$\hat{u}_i = \beta_2 x_i + (u_i - \bar{u}) - \hat{\beta}_2 x_i \tag{13}$$

Thu thập các số hạng, bình phương lên và lấy tổng hai vế, ta được:

$$\sum \hat{u}_i^2 = (\hat{\beta}_2 - \beta_2)^2 \sum x_i^2 + \sum (u_i - \bar{u})^2 - 2(\hat{\beta}_2 - \beta_2) \sum x_i (u_i - \bar{u}) \tag{14}$$

Lấy các kỳ vọng của cả hai vế cho:

$$\begin{aligned}E(\sum \hat{u}_i^2) &= \sum x_i^2 E(\hat{\beta}_2 - \beta_2)^2 + E\left[\sum (u_i - \bar{u})^2\right] - 2E\left[(\hat{\beta}_2 - \beta_2) \sum x_i (u_i - \bar{u})\right] \\ &= \quad A \quad + \quad B \quad + \quad C\end{aligned}\tag{15}$$

Bây giờ, từ các giả thiết về mô hình hồi quy tuyến tính cổ điển và một vài trong số các kết quả đã thiết lập nên, nó có thể được kiểm chứng rằng:

$$A = \sigma^2$$

$$B = (n-1)\sigma^2$$

$$C = -2\sigma^2$$

Vì vậy khi thế các giá trị này vào (15) ta được

$$E\left(\sum \hat{u}_i^2\right) = (n-2)\sigma^2 \quad (16)$$

Do đó, nếu ta định nghĩa

$$\hat{\sigma}_2^2 = \frac{\sum \hat{u}_i^2}{n-2} \quad (17)$$

giá trị kỳ vọng của nó là

$$E(\hat{\sigma}_2^2) = \frac{1}{n-2} E\left(\sum \hat{u}_i^2\right) = \sigma^2 \quad \text{khi sử dụng (16)} \quad (18)$$

nó chỉ ra rằng $\hat{\sigma}_2^2$ là hàm ước lượng không thiên lệch của σ^2 thực.

3A.6 TÍNH CHẤT PHƯƠNG SAI NHỎ NHẤT CỦA CÁC HÀM ƯỚC LƯỢNG BÌNH PHƯƠNG TỐI THIỂU:

Trong Phụ lục 3A, Phần 3A.2, ta đã trình bày hàm ước lượng bình phương tối thiểu $\hat{\beta}_2$ là tuyến tính và không thiên lệch (điều này cũng đúng với $\hat{\beta}_1$). Để chứng tỏ rằng các hàm ước lượng này cũng là phương sai nhỏ nhất trong nhóm tất cả các hàm ước lượng không thiên lệch tuyến tính, ta xét hàm ước lượng bình phương tối thiểu $\hat{\beta}_2$:

$$\hat{\beta}_2 = \sum k_i Y_i$$

trong đó:

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} = \frac{x_i}{\sum x_i^2} \quad (\text{xem phụ lục 3A.2}) \quad (19)$$

nó chứng tỏ rằng $\hat{\beta}_2$ là trung bình trọng số của các Y , với k_i đóng vai trò các trọng lượng.

Ta hãy định nghĩa hàm ước lượng tuyến tính thay thế của β_2 như sau:

$$\beta_2^* = \sum w_i Y_i \quad (20)$$

trong đó w_i cũng là các trọng lượng, không nhất thiết bằng k_i . Bây giờ:

$$\begin{aligned} E(\beta_2^*) &= \sum w_i E(Y_i) \\ &= \sum w_i (\beta_1 + \beta_2 X_i) \\ &= \beta_1 \sum w_i + \beta_2 \sum w_i X_i \end{aligned} \quad (21)$$

Do đó, đối với β_2^* không thiên lệch ta phải có:

$$\sum w_i = 0 \quad (22)$$

và

$$\sum w_i X_i = 1 \quad (23)$$

Ta cũng có thể viết:

$$\begin{aligned} \text{var}(\beta_2^*) &= \text{var} \sum w_i Y_i \\ &= \sum w_i^2 \text{var } Y_i \quad [\text{Lưu ý: } \text{var } Y_i = \text{var } u_i = \sigma^2] \\ &= \sigma^2 \sum w_i^2 \quad [\text{Lưu ý: } \text{cov}(Y_i, Y_j) = 0 \text{ (} i \neq j \text{)}] \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} + \frac{x_i}{\sum x_i^2} \right)^2 \quad (\text{Lưu ý bẫy toán học}) \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \frac{\sum x_i^2}{(\sum x_i^2)^2} + 2\sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right) \left(\frac{x_i}{\sum x_i^2} \right) \\ &= \sigma^2 \sum \left(w_i - \frac{x_i}{\sum x_i^2} \right)^2 + \sigma^2 \left(\frac{1}{\sum x_i^2} \right) \end{aligned} \quad (24)$$

vì số hạng cuối cùng trong biểu thức áp chót triệt tiêu (Tại sao?)

Vì số hạng cuối cùng trong (24) là hằng số, phương sai của (β_2^*) có thể là cực tiểu chỉ khi ta biến đổi số hạng thứ nhất. Nếu ta coi:

$$w_i = \frac{x_i}{\sum x_i^2}$$

Phương trình (24) giảm tới

$$\begin{aligned} \text{var}(\beta_2^*) &= \frac{\sigma^2}{\sum x_i^2} \\ &= \text{var}(\hat{\beta}_2) \end{aligned} \quad (25)$$

Nói gọn lại, với các trọng số $w_i = k_i$ là các trọng số bình phương tối thiểu, phương sai của hàm ước lượng tuyến tính β_2^* bằng phương sai của hàm ước lượng bình phương tối thiểu $\hat{\beta}_2$; ngược lại, $\text{var}(\beta_2^*) > \text{var}(\beta_2)$. Để đặt nó khác đi, nếu có các hàm ước lượng không thiên lệch tuyến tính với phương sai nhỏ nhất β_2 , nó sẽ phải là hàm ước lượng bình phương tối thiểu. Tương tự, nó có thể được chứng tỏ rằng $\hat{\beta}_1$ là hàm ước lượng không thiên lệch tuyến tính với phương sai nhỏ nhất của β_1 .

3A.7. KẾT QUẢ SAS CỦA HÀM CẦU VỀ CÀ PHÊ (3.7.1)

Vì đây là lần đầu tiên ta trình bày đến kết quả SAS, sẽ rất bổ ích khi ta giới thiệu ngắn gọn về kết quả. Các kết quả đã được thu từ quá trình HỒI QUY của SAS. Biến phụ thuộc là Y (số tách cho một người trong một ngày) và biến hồi qui độc lập là X_2 [giá lẻ thực tế trung bình, tính bằng \$ cho một pound. Lưu ý rằng đây là biến X trong (3.7.1)]. Với mục đích trình bày, kết quả đề cập trong trang sau được chia làm 6 phần. Lưu ý rằng rất nhiều các số thập phân được chỉ rõ trong kết quả nhưng trong thực tế, ta chỉ cần lấy 4 hoặc 5 số.

- Phần I:** Phần này cho ta Bảng của phép phân tích phương sai (ANOVA) mà ta đã thảo luận trong Chương 5.
- Phần II:** *Căn* MSE nghĩa là căn bậc 2 của sai số bình phương trung bình ($=\sigma^2$), nghĩa là nó cho ta sai số chuẩn của ước lượng $\hat{\sigma}$.
Trung bình Dep nghĩa là giá trị trung bình của biến phụ thuộc $Y (= \bar{Y})$.
C.V là hệ số biến thiên được xác định như là $(\sigma^2 / \bar{Y}) \times 100$, và nó biểu thị tính biến thiên không giải thích được duy trì trong dữ liệu (nghĩa là biến Y) liên quan tới giá trị trung bình $\bar{Y}$.
 R^2 = hệ số xác định
 $\bar{R}^2 = R^2$ đã điều chỉnh (xem Chương 7)
- Phần III:** Phần này cho các giá trị ước lượng của các thông số, các sai số chuẩn của chúng, các tỷ số t của chúng và mức ý nghĩa của các tỷ số t . Hai đại lượng sau cùng sẽ được trình bày đầy đủ trong Chương 5.
- Phần IV:** Phần này cho ta cái được gọi là ma trận phương sai- đồng phương sai của các thông số ước lượng. các phần tử trên đường chéo chạy từ góc trái-trên đến góc phải-dưới cho các phương sai (nghĩa là bình phương của các sai số chuẩn đã cho trong Phần III)⁴¹ và các phần tử không nằm trên đường chéo cho các đồng phương sai giữa các thông số ước lượng, ở đây $\text{cov}(\hat{\beta}_1, \hat{\beta}_2)$ như đã định nghĩa ở (3.3.9).
- Phần V:** Phần này cho các giá trị thực của Y_i và X_i , các giá trị ước lượng của $Y (= \hat{Y}_i)$, và các phần dư $\hat{u}_i = (Y_i - \hat{Y}_i)$
- Phần VI:** Phần này cho thống kê d Durbin-Watson và hệ số tương quan bậc nhất, các chủ đề đã được thảo luận trong Chương 12.

BIẾN DEP: Y

I	Nguồn	DF	Tổng bình phương	Bình phương trung bình	Giá trị F	PROB>F
	Mô hình	1	0.292975	0.292975	17.687	0.0023
	Sai số	9	0.149080	0.016564		
	Tổng số C	10	0.442055			
II	Căn MSE		0.128703	R-bình phương	0.6628	
	TBình DEP		2.206364	ADJ R-SC	0.6253	
	C.V		5.833255			
III	Biến	DF	Thông số ước lượng	Sai số chuẩn	T cho HO Thông số=0	PROB> T

⁴¹ Vì vậy, 0.01479 là phương sai của $\hat{\beta}_1$ và 0.0130 là phương sai của $\hat{\beta}_2$; lấy căn bậc hai của các số này ta được 0.1216 và 0.1140, chúng là các sai số chuẩn tương ứng của hai hệ số như đã được ghi ở Phần III, ngoại trừ đối với các sai số làm tròn.

	Tung độ gốc	1	2.691124	0.121622	22.127	0.0001
	X	1	-0.479529	0.114022	-4.206	0.0023
IV	Đồng phương sai của các ước lượng					
	COVB					X
	Tung độ gốc		0.01479203	-0.0131428		
	X		-0.0131428	0.01300097		
V	OBS	Y	X	YHAT	YRESID = $\hat{u}_i$	
	1	2.57	0.77	2.32189	0.24811	
	2	2.50	0.74	2.33627	0.16373	
	3	2.35	0.72	2.34586	0.00414	
	4	2.25	0.73	2.34107	-0.04107	
	5	2.20	0.76	2.32668	-0.07668	
	6	2.20	0.75	2.33148	-0.13148	
	7	2.11	1.08	2.17323	-0.06323	
	8	1.94	1.81	1.82318	0.11682	
	9	1.97	1.39	2.02458	-0.05458	
	10	2.06	1.20	2.11569	-0.05569	
	11	2.02	1.17	2.13007	-0.11007	
VI	d DURBIN-WATSON					0.727
	Tự tương quan bậc 1					0.390