

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/hoặc 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 chi 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 chi 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 **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 có được bất kỳ một trong số những chi tiêu tiêu dùng này là $1/5$. Biểu thị bằng các ký hiệu toán

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.

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 kỹ thuật 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 2.1Thu nhập gia đình hàng tuần X , \$

$Y \downarrow$ $X \rightarrow$	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 chi tiêu tiêu dùng của từng gia đình, Hình 2.1 cho thấy một cách rất rõ ràng là chi 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** .

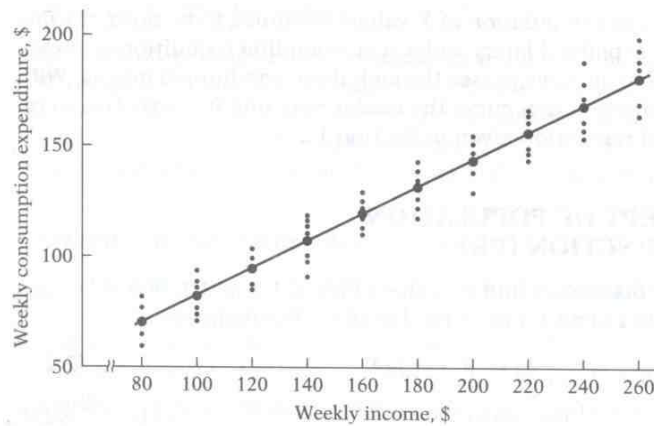
BẢNG 2.2Xác suất có Điều kiện $p(Y | X_i)$ của dữ liệu trong Bảng 2.1

$p(Y X_i)$ $X \rightarrow$ $\downarrow$	80	100	120	140	160	180	200	220	240	260
Xác suất	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
có điều kiện	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7

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.

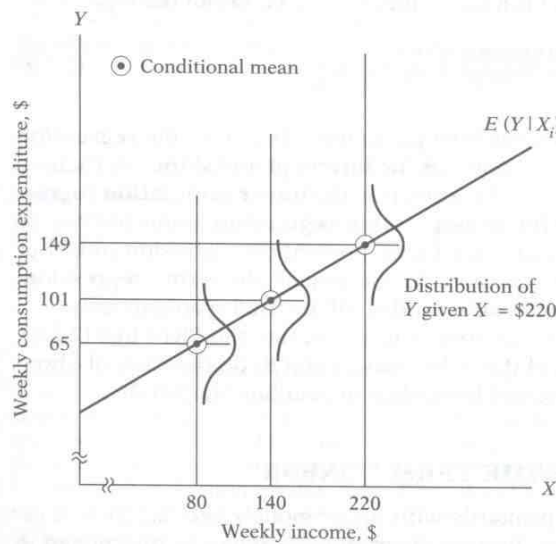
$p(Y X_i)$	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	1/5	1/6	1/5	1/7	1/6	1/6	1/5	1/7	1/6	1/7
	—	1/6	—	1/7	1/6	1/6	—	1/7	1/6	1/7
	—	—	—	1/7	—	—	—	1/7	—	1/7
Trung bình có điều kiện của Y	65	77	89	101	113	125	137	149	161	173

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 phân 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

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, $X * Z$ hay X/Z , trong đó Z là một biến khác). Nếu Y chỉ phụ thuộc vào một mình 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

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

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).

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.)

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

$$\begin{aligned} Y_1 &= 55 = \beta_1 + \beta_2(80) + u_1 \\ Y_2 &= 60 = \beta_1 + \beta_2(80) + u_2 \\ Y_3 &= 65 = \beta_1 + \beta_2(80) + u_3 \\ Y_4 &= 70 = \beta_1 + \beta_2(80) + u_4 \\ Y_5 &= 75 = \beta_1 + \beta_2(80) + u_5 \end{aligned} \quad (2.4.3)$$

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

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

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.

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.

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.

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.

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

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.

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
---	---

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

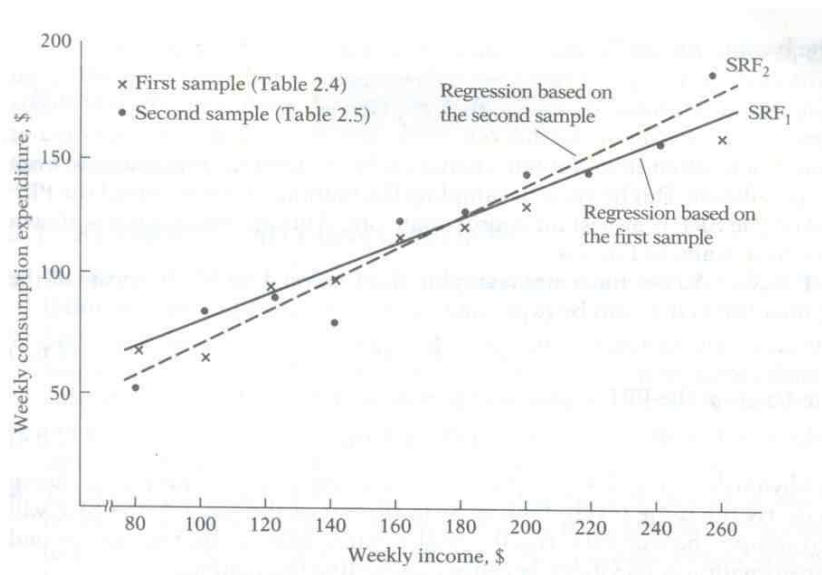
¹⁰ "Nên giữ cho sự diễn tả càng đơn giản càng tốt cho đến khi nào tỏ ra không thoả đáng thì thôi," *The World of Mathematics* (Thế giới toán học), tập 2, J. R. Newman, Simon & Schuster, New York, 1956, trang 1247, hay "Không nên nhân các đối tượng vượt quá mức cần thiết," Donald F. Morrison, *Applied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

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 đạt 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. Regression lines based on two different samples

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 mẫu"

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 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

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

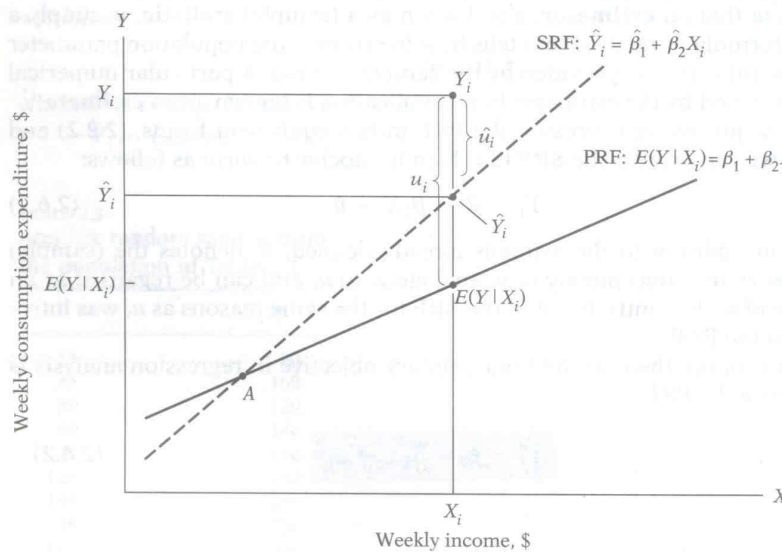
và theo PRF nó có thể được thể hiện như sau

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

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. Sample and population regression lines

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? (Gợi 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

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

* Tùy ý

- Tính các hàm mật độ biên, $f(X)$ và $f(Y)$.
- Tính các hàm mật độ có điều kiện $f(X|Y)$ và $f(Y|X)$.
- Tính $E(X)$ và $E(Y)$.
- Tính $E(X|Y = 0.4)$

2.7 Xem xét các dữ liệu dưới đây.

Lương trung vị của các nhà kinh tế học 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

- Các dữ liệu này cho ta thấy gì?
- Tuổi tác hay kinh nghiệm có quan hệ gần hơn đối với mức lương hay không? Làm sao Anh/Chị biết?
- 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.

- 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.
- 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:

- Về 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," *American 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?

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

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

- a) Chứng minh rằng đối với phân phối này $E(X) = \sum X_i (1/k)$ và phương sai $\sigma_X^2 = \sum [X_i - E(X)]^2 \cdot (1/k)$ trong đó $E(X)$ 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.

- a) 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.
- b) Chúng ta có thể rút ra được những kết luận gì?
- c) Khi đã biết điểm văn đáp của nam và nữ, làm cách nào bạn có thể tiên đoán được điểm toán của họ?
- d) Vẽ đồ thị điểm SAT tổng cộng của nữ so với điểm SAT tổng cộng của nam. Về 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, 1967-1 990*

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?