Chương 16

CÁC MÔ HÌNH HỒI QUY DỮ LIỆU BẢNG

Trong Chương 1, chúng ta đã thảo luận ngắn gọn về các loại dữ liệu thường có cho phân tích thực nghiệm, đó là **dữ liệu chuỗi thời gian, dữ liệu chéo** và **dữ liệu bảng**. Đối với dữ liệu chuỗi thời gian, chúng ta quan sát các giá trị của một hoặc nhiều biến theo thời gian (ví dụ, quan sát chỉ tiêu GDP trong nhiều quí hay nhiều năm). Trong dữ liệu chéo, các giá trị của một hoặc nhiều biến được thu thập cho nhiều đơn vị mẫu hoặc nhiều đại diện mẫu ở tại cùng một thời điểm (ví dụ, tỷ lệ tội phạm của 50 tiểu bang ở Mỹ trong một năm nào đó). *Trong dữ liệu bảng, cùng một đơn vị chéo nào đó (theo không gian) (thí dụ một gia đình hay một doanh nghiệp hay một tiểu bang) được điều tra theo thời gian.* Nói ngắn gọn, dữ liệu bảng có *qui mô về thời gian lẫn không gian*.

Chúng ta đã xem một thí dụ về dữ liệu bảng trong Bảng 1.1. Bảng này cho thấy dữ liệu về số trứng được sản xuất ra và các giá của chúng đối với 50 tiểu bang ở Mỹ trong các năm 1990 và 1991. Đối với một năm cho trước, dữ liệu về số trứng và các giá của chúng thể hiện một mẫu dữ liệu chéo. Đối với bất kỳ một tiểu bang cho trước nào, có hai quan sát chuỗi thời gian về số trứng và các giá của chúng. Như thế, chúng ta có tất cả là (50 x 2) = 100 quan sát (*gộp chung*) về số trứng được sản xuất ra và các giá của chúng.

Dữ liệu bảng còn được gọi bằng các tên khác, như là **dữ liệu gộp chung** (gộp chung các quan sát chéo và chuỗi thời gian), là **sự kết hợp của dữ liệu chéo và chuỗi thời gian, dữ liệu bảng vi mô (micropanel data), dữ liệu dọc** (longitudinal data) (đó là một nghiên cứu nào đó theo thời gian về một biến hay một nhóm đối tượng), **phân tích lịch sử sự kiện** (thí dụ, nghiên cứu sự thay đổi theo thời gian của những đối tượng qua các tình trạng hay các điều kiện được tiếp diễn theo thời gian), phân tích theo tổ (cohort analysis) (ví dụ, theo dõi con đường sự nghiệp của 1965 sinh viên tốt nghiệp một trường kinh doanh). Mặc dù có những sự thay đổi tinh tế, nhưng tất cả các tên gọi này *thực chất muốn nói đến sự thay đổi theo thời gian của các đơn vị chéo*. Vì thế, chúng ta sẽ sử dụng thuật ngữ dữ liệu bảng theo nghĩa chung để bao gồm một hay nhiều hơn các thuật ngữ nói trên. *Và chúng ta sẽ gọi các mô hình hồi quy dựa trên dữ liệu như thế là các mô hình hồi quy dữ liệu bảng*.

Dữ liệu bảng hiện đang được sử dụng ngày càng nhiều trong nghiên cứu kinh tế. Một số tập dữ liệu bảng nổi tiếng là:

1. **Panel Study of Income Dynamics (PSID)** (Nghiên cứu dữ liệu bảng Sự thay đổi theo Thời gian của Thu nhập) do Viện Nghiên cứu Xã hội tại Đại học Michigan tiến hành. Bắt đầu vào năm 1968, mỗi năm Viện này thu thập dữ liệu đối với khoảng 5.000 gia đình về các biến nhân khẩu học và kinh tế xã hội khác nhau.

2. Cục Điều tra Dân số của Bộ Thương mại Mỹ tiến hành một cuộc điều tra tương tự như PSID, được gọi là **Survey of Income and Program Participation (SIPP)** (Điều tra về Thu nhập và Sự Tham gia Chương trình). Những người tham gia trả lời phỏng vấn được phỏng vấn mỗi năm bốn lần về điều kiện kinh tế của họ.

Nhiều cơ quan chính phủ khác nhau ở Mỹ cũng tiến hành nhiều cuộc điều tra khác nhau. Ngay từ đầu, đưa ra một lời cảnh báo là điều phù hợp. Đề tài các hồi quy dữ liệu bảng thật là rộng, phần nội dung liên quan đến toán học và thống kê rất phức tạp. Chúng ta chỉ hy vọng đề cập đến một số nội dung cơ bản của các mô hình hồi quy dữ liệu bảng, các chi tiết của vấn đề này nằm ở phần tài liệu tham khảo.[1] Xin cảnh báo trước rằng một số tài liệu tham khảo này có tính kỹ thuật chuyên môn cao. Rất may là trong số các phần mềm quen thuộc với chúng ta như Limdep, PcGive, SAS, STATA, Shazam, và Eviews đã làm cho công việc thực hiện các hồi quy dữ liệu bảng trên thực tế hoàn toàn dễ dàng.

16.1. TẠI SAO LẠI LÀ DỮ LIỆU BẢNG?

Những ưu điểm của dữ liệu bảng so với dữ liệu chéo hay dữ liệu chuỗi thời gian là gì? Baltagi liệt kê những ưu điểm sau đây của dữ liệu bảng. [2]

1. Bởi vì dữ liệu bảng liên hệ đến các cá nhân, các doanh nghiệp, các tiểu bang, các quốc gia v.v theo thời gian, nên chắc chắn có tính không đồng nhất trong các đơn vị này. Các kỹ thuật ước lượng dựa trên dữ liệu bảng có thể tính đến tính không đồng nhất đó một cách rõ ràng bằng cách bao gồm các biến chuyên biệt theo cá nhân, như chúng tôi sắp cho thấy. Chúng tôi sử dụng thuật ngữ *cá nhân* ở đây theo nghĩa chung nhất để bao gồm các đơn vị vi mô như các cá nhân, doanh nghiệp, tiểu bang và quốc gia.

2. Bằng cách kết hợp chuỗi thời gian của các quan sát chéo, dữ liệu bảng cho chúng ta "dữ liệu chứa nhiều thông tin hữu ích hơn, tính biến thiên nhiều hơn, ít hiện tượng đa cộng tuyến giữa các biến hơn, nhiều bậc tự do hơn và hiệu quả cao hơn."

3. Bằng cách nghiên cứu quan sát lập đi lập lại của các đơn vị chéo, dữ liệu bảng phù hợp hơn cho việc nghiên cứu sự *động thái thay đổi theo thời gian của các đơn vị chéo này*. Những tác động của thất nghiệp, tốc độ quay vòng việc làm, tính dịch chuyển của lao động được nghiên cứu tốt hơn khi có dữ liệu bảng.

4. Dữ liệu bảng có thể phát hiện và đo lường tốt hơn các tác động mà người ta không thể quan sát được trong dữ liệu chuỗi thời gian hay dữ liệu chéo thuần túy. Thí dụ, tác động của các luật về mức lương tối thiểu đối với việc làm và thu nhập có thể được nghiên cứu tốt hơn nếu chúng ta bao gồm các đợt gia tăng mức lương tối thiểu liên tiếp trong các mức lương tối thiểu của liên bang và/hoặc tiểu bang.

5. Dữ liệu bảng làm cho chúng ta có thể nghiên cứu các mô hình hành vi phức tạp hơn. Thí dụ, chúng ta có thể xử lý tốt hơn bằng dữ liệu bảng các hiện tượng như lợi thế kinh tế theo qui mô và thay đổi công nghệ so với dữ liệu chéo hay dữ liệu chuỗi thời gian.

6. Bằng cách cung cấp dữ liệu đối với vài nghìn đơn vị, dữ liệu bảng có thể giảm đến mức thấp nhất hiện tượng chệch có thể xảy ra nếu chúng ta gộp các cá nhân hay các doanh nghiệp theo những biến số có mức tổng hợp cao.

Nói tóm lại, dữ liệu bảng có thể làm cho phân tích thực nghiệm phong phú hơn so với cách chúng ta chỉ sử dụng dữ liệu chéo hay dữ liệu chuỗi thời gian. Điều này không

có ý cho rằng không có vấn đề khó khăn gì với việc lập mô hình dựa trên dữ liệu bảng. Chúng ta sẽ thảo luận về chúng sau khi trình bày một vài lý thuyết và thảo luận một ví dụ.

16.2. DỮ LIỆU BẢNG: MỘT VÍ DỤ MINH HỌA

Để chuẩn bị, chúng ta hãy xét một ví dụ cụ thể. Hãy xét dữ liệu được cho trong Bảng 16.1, dữ liệu này được lấy từ một nghiên cứu nổi tiếng về lý thuyết đầu tư do Y. Grunfeld đề xuất. [3]

Grunfeld quan tâm đến việc tìm hiểu xem tổng đầu tư (Y) phụ thuộc như thế nào vào giá trị thực của doanh nghiệp (X_2) và trữ lượng vốn thực (X_3). Mặc dù nghiên cứu đầu tiên bao gồm nhiều công ty, nhưng nhằm mục đích minh họa chúng tôi chỉ thu nhận dữ liệu về bốn công ty, đó là General Electric (GE), General Motor (GM), U.S. Steel (US), và Westinghouse. Dữ liệu đối với mỗi công ty về ba biến nói trên có sẵn cho thời kỳ 1935-1954. Như thế, có bốn đơn vị chéo (theo không gian) và 20 thời đoạn. Vì thế, tính tổng cộng chúng ta có 80 quan sát. Y được kỳ vọng có quan hệ đồng biến với X_2 và X_3.

Trên nguyên tắc, chúng ta có thể chạy bốn hồi quy chuỗi thời gian, tức là một hồi quy cho mỗi công ty, hay chúng ta có thể chạy 20 hồi quy chéo, tức là một hồi quy cho mỗi năm. Trong trường hợp chạy hồi quy chéo, chúng ta sẽ phải lo lắng đến số bậc tự do. [4]

BẢNG 16.1 DỮ LIỆU VỀ ĐẦU TƯ CHO BỐN CÔNG TY, 1935-1954

Quan sát	*I*	F_{-1}	C_{-1}	*Quan sát*	*I*	F_{-1}	C_{-1}
	GE				US		
1935	33,1	1170,6	97,8	1935	209,9	1362,4	53,8
1936	45,0	2015,8	104,4	1936	355,3	1807,1	50,5
1937	77,2	2803,3	118,0	1937	469,9	2673,3	118,1
1938	44,6	2039,7	156,2	1938	262,3	1801,9	260,2
1939	48,1	2256,2	172,6	1939	230,4	1957,3	312,7
1940	74,4	2132,2	186,6	1940	361,6	2202,9	254,2
1941	113,0	1834,1	220,9	1941	472,8	2380,5	261,4
1942	91,9	1588,0	287,8	1942	445,6	2168,6	298,7
1943	61,3	1749,4	319,9	1943	361,6	1985,1	301,8
1944	56,8	1687,2	321,3	1944	288,2	1813,9	279,1
1945	93,6	2007,7	319,6	1945	258,7	1850,2	213,8
1946	159,9	2208,3	346,0	1946	420,3	2067,7	232,6
1947	147,2	1656,7	456,4	1947	420,5	1796,7	264,8
1948	146,3	1604,4	543,4	1948	494,5	1625,8	306,9
1949	98,3	1431,8	618,3	1949	405,1	1667,0	351,1
1950	93,5	1610,5	647,4	1950	418,8	1677,4	357,8
1951	135,2	1819,4	671,3	1951	588,2	2289,5	341,1
1952	157,3	2079,7	726,1	1952	645,2	2159,4	444,2
1953	179,5	2371,6	800,3	1953	641,0	2031,3	623,6
1954	189,6	2759,9	888,9	1954	459,3	2115,5	669,7
	GM				WEST		
1935	317,6	3078,5	2,8	1935	12,93	191,5	1,8
1936	391,8	4661,7	52,6	1936	25,90	516,0	0,8
1937	410,6	5387,1	156,9	1937	35,05	729,0	7,4
1938	257,7	2792,2	209,2	1938	22,89	560,4	18,1
1939	330,8	4313,2	203,4	1939	18,84	519,9	23,5
1940	461,2	4643,9	207,2	1940	28,57	628,5	26,5
1941	512,0	4551,2	255,2	1941	48,51	537,1	36,2
1942	448,0	3244,1	303,7	1942	43,34	561,2	60,8
1943	499,6	4053,7	264,1	1943	37,02	617,2	84,4
1944	547,5	4379,3	201,6	1944	37,81	626,7	91,2
1945	561,2	4840,9	265,0	1945	39,27	737,2	92,4
1946	688,1	4900,0	402,2	1946	53,46	760,5	86,0
1947	568,9	3526,5	761,5	1947	55,56	581,4	111,1
1948	529,2	3245,7	922,4	1948	49,56	662,3	130,6
1949	555,1	3700,2	1020,1	1949	32,04	583,8	141,8
1950	642,9	3755,6	1099,0	1950	32,24	635,2	136,7
1951	755,9	4833,0	1207,7	1951	54,38	732,8	129,7
1952	891,2	4924,9	1430,5	1952	71,78	864,1	145,5
1953	1304,4	6241,7	1777,3	1953	90,08	1193,5	174,8
1954	1486,7	5593,6	2226,3	1954	68,60	1188,9	213,5

Ghi chú:

Y = I = tổng đầu tư = những đầu tư bổ sung vào nhà máy và thiết bị cộng với bảo trì và sửa chữa, tính bằng triệu đô la Mỹ đã khử lạm phát bởi chỉ số giá P_1.

X_2 = F = giá trị của doanh nghiệp = giá của cổ phiếu thường và cổ phiếu ưu đãi vào ngày 31 tháng 12 (hay giá trung bình của ngày 31 tháng 12 và ngày 31 tháng 1 của năm sau) nhân với số cổ phiếu thường và cổ phiếu ưu đãi còn lưu hành cộng với tổng giá trị trên sổ sách của vốn vay vào ngày 31 tháng 12, tính bằng triệu đô la Mỹ đã khử lạm phát bởi P_2.

X_3 = C = trữ lượng nhà máy và thiết bị = tổng số tích lũy của những đầu tư bổ sung vào nhà máy và thiết bị đã được khử lạm phát bởi P_1 trừ đi khoản tiền khấu hao đã khử lạm phát bởi P_3 trong các định nghĩa này.

P_1 = Chỉ số khử lạm phát tiềm ẩn của thiết bị lâu bền của các nhà sản xuất (1947 = 100)

P_2 = Chỉ số khử lạm phát tiềm ẩn của GDP (1947 = 100)

P_3 = Chỉ số khử lạm phát chi phí khấu hao = trung bình trượt 10-năm của chỉ số giá bán buôn của kim loại và các sản phẩm từ kim loại (1947 = 100)

Nguồn: Trích từ H.D. Vinod và Aman Ullah, *Những Tiến bộ Gần đây trong Các Phương pháp Hồi quy*, Nhà Xuất bản Marcel Dekker, New York, 1981, các trang 259-261

Gộp chung tất cả 80 quan sát, chúng ta có thể viết hàm đầu tư của Grunfeld như sau:

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$
$$i = 1, 2, 3, 4$$
$$t = 1, 2, \ldots, 20 \qquad \textbf{(16.2.1)}$$

trong đó i là đơn vị chéo thứ i và t là thời đoạn thứ t. Theo qui ước, chúng ta sẽ cho i là ký hiệu cho đơn vị chéo và t là ký hiệu theo thời gian. Chúng ta giả định rằng có một số tối đa đơn vị chéo hay quan sát N và một số tối đa thời đoạn T. Nếu mỗi đơn vị chéo có cùng số quan sát chuỗi thời gian như nhau, thì bảng dữ liệu đó được gọi là **bảng cân bằng**. Trong ví dụ đang dùng chúng ta có bảng cân bằng, vì mỗi công ty trong mẫu đều có 20 quan sát. Nếu số quan sát khác nhau giữa các thành viên của bảng, chúng ta gọi bảng như thế là **bảng không cân bằng**. Trong chương này, chúng ta sẽ quan tâm phần lớn đến bảng cân bằng.

Ban đầu, chúng ta giả định rằng các giá trị X là không ngẫu nhiên và rằng số hạng sai số theo đúng các giả định cổ điển, đó là, $E(u_{it}) \sim N(0, \sigma^2)$. Hãy cẩn thận lưu ý hai và ba ký hiệu dưới dòng, những ký hiệu này không cần giải thích chắc người đọc cũng hiểu.

Làm sao chúng ta ước lượng (16.2.1)? Câu trả lời được trình bày sau đây.

16.3 ƯỚC LƯỢNG CÁC MÔ HÌNH HỒI QUI DỮ LIỆU BẢNG: PHƯƠNG PHÁP TÁC ĐỘNG CỐ ĐỊNH

Việc ước lượng (16.2.1) phụ thuộc vào các giả định chúng ta đưa ra về tung độ gốc, các hệ số độ dốc, và số hạng sai số u_{it}. Có nhiều khả năng xảy ra[5]:

1. Giả định rằng tung độ gốc và các hệ số độ dốc không đổi theo thời gian và không gian và số hạng sai số thể hiện những khác biệt theo thời gian và các cá nhân.

2. Các hệ số độ dốc không đổi nhưng tung độ gốc thay đổi theo các cá nhân.

3. Các hệ số độ dốc không đổi nhưng tung độ gốc thay đổi theo các cá nhân và thời gian.

4. Tất cả các hệ số (tung độ gốc cũng như các hệ số độ dốc) thay đổi theo các cá nhân.

5. Tung độ gốc cũng như các hệ số độ dốc thay đổi theo các cá nhân và thời gian.

Như bạn có thể thấy, trong mỗi trường hợp này thể hiện mức độ phức tạp tăng dần (và có lẽ thực tế hơn) trong việc ước lượng các mô hình hồi quy dữ liệu bảng, như mô hình (16.2.1). Dĩ nhiên, mức độ phức tạp sẽ gia tăng nếu chúng ta thêm nhiều biến hồi quy độc lập hơn vào mô hình này, do khả năng xảy ra hiện tượng đa cộng tuyến giữa các biến độc lập.

Để trình bày đầy đủ nội dung của mỗi loại nói trên sẽ cần một cuốn sách riêng biệt, và trên thị trường hiện đã có vài cuốn sách như thế [6]. Trong phần sau đây, chúng tôi sẽ trình bày một số đặc điểm chính của các khả năng khác nhau này, đặc biệt là bốn khả năng đầu. Nội dung thảo luận của chúng tôi sẽ không đi sâu và kỹ thuật.

1. Tất cả hệ số không đổi qua thời gian và giữa các cá nhân.

Phương pháp đơn giản nhất, và có lẽ ngây ngô, là không kể đến các kích thước không gian và thời gian của dữ liệu kết hợp và chỉ ước lượng hồi quy Bình phương Nhỏ nhất Thông thường (OLS) thường lệ. Đó là, cứ xếp 20 quan sát của mỗi công ty lên trên các quan sát của công ty kia, như thế cho ta tổng cộng là 80 quan sát đối với mỗi biến trong mô hình. Các kết quả OLS như sau:

$$\begin{aligned} Y &= -63{,}3041 + 0{,}1101X_2 + 0{,}3034X_3 \\ \text{se} &= (29{,}6124) \quad (0{,}0137) \quad (0{,}0493) \\ \text{t} &= (-2{,}1376) \quad (8{,}0188) \quad (6{,}1545) \end{aligned} \qquad \textbf{(16.3.1)}$$

$R^2 = 0{,}7565$ Durbin–Watson = 0,2187

n = 80 df = 77

se: sai số chuẩn

df: bậc tự do

Nếu bạn xem xét các kết quả của **hồi quy kết hợp**, và áp dụng các tiêu chuẩn thông thường, bạn sẽ thấy rằng tất cả hệ số đều có ý nghĩa thống kê, các hệ số độ dốc có dấu dương kỳ vọng và giá trị R^2 tương đối cao. Như đã kỳ vọng, Y có quan hệ đồng biến với X_2 và X_3. Con sâu "duy nhất" làm rầu nồi canh là trị thống kê Durbin-Watson ước lượng rất thấp, gợi ý có lẽ có hiện tượng tự tương quan trong dữ liệu. Dĩ nhiên, như chúng ta biết, giá trị Durbin-Watson thấp cũng có thể do các sai lầm khi nhận dạng mô hình. Thí dụ, mô hình ước lượng giả định giá trị tung độ gốc của GE, GM, US, và Westinghouse giống nhau. Nó cũng giả định các hệ số độ dốc của hai biến X đều giống hệt nhau đối với cả bốn doanh nghiệp. Rõ ràng đó là những giả định rất hạn chế. Vì thế cho nên, cho dù mô hình trên rất đơn giản, hồi quy kết hợp (16.1.2) có thể làm biến dạng bức tranh đích thực của mối quan hệ giữa Y và các biến X giữa bốn công ty nêu trên. Điều chúng ta cần làm là tìm một cách nào đó để tính đến bản chất cụ thể của bốn công ty. Phần tiếp theo sẽ giải thích làm thế nào thực hiện điều này.

2. Các hệ số độ dốc không đổi, nhưng tung độ gốc thay đổi giữa các cá nhân: Mô hình tác động cố định hay hồi quy biến giả bình phương nhỏ nhất (LSDV)

Một cách để tính đến "tính đặc trưng" của mỗi công ty hay mỗi đơn vị chéo là để cho tung độ gốc thay đổi đối với mỗi công ty nhưng vẫn giả định các hệ số độ dốc không đổi giữa các doanh nghiệp. Để thấy được điều này, chúng ta viết mô hình (16.2.1) như sau:

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \qquad \textbf{(16.3.2)}$$

Lưu ý rằng chúng ta đã đặt ký hiệu dưới dòng *i* vào số hạng tung độ gốc để cho thấy rằng các tung độ gốc của bốn doanh nghiệp này có thể khác nhau; những khác biệt có thể do những đặc điểm đặc biệt của mỗi công ty, như là phong cách quản lý hay phong cách quản lý.

Trong các tài liệu, mô hình (16.3.2) được biết đến dưới tên gọi là mô hình (hồi quy) **tác động cố định (FEM)**. Thuật ngữ tác động cố định được sử dụng là do thực tế là mặc dù tung độ gốc có thể khác nhau giữa các cá nhân (ở đây là bốn công ty), nhưng mỗi tung độ gốc của cá nhân không thay đổi theo thời gian; nghĩa là nó *bất biến theo thời gian*. Lưu ý rằng nếu chúng ta phải viết tung độ gốc là β_{1it}, thì nó sẽ gợi ý rằng tung độ gốc của mỗi công ty hay cá nhân là *thay đổi theo thời gian*. Có thể lưu ý rằng FEM được cho trong (16.3.2) giả định các hệ số độ dốc của các biến hồi quy độc lập là không thay đổi giữa các cá nhân hay theo thời gian.

Làm thế nào chúng ta có thể thực sự tính đến tung độ gốc (tác động cố định) thay đổi giữa các công ty? Chúng ta có thể làm điều đó một cách dễ dàng bằng kỹ thuật biến giả mà chúng ta đã học trong Chương 9, đặc biệt là **các biến giả tung độ gốc chênh lệch**. Vì thế, chúng ta viết (16.3.2) thành:

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \qquad \textbf{(16.3.3)}$$

trong đó D_{2i} = 1 nếu quan sát thuộc về GM, 0 nếu khác đi; D_{3i} = 1 nếu quan sát thuộc về US, 0 nếu khác đi; và D_{4i} = 1 nếu quan sát thuộc về WEST, 0 nếu khác đi. Bởi vì chúng ta có bốn công ty, nên chúng ta chỉ sử dụng ba biến giả để tránh rơi vào **bẫy biến giả** (nghĩa là tình huống có hiện tượng đa cộng tuyến hoàn hảo). Ở đây, không có biến giả cho GE. Nói cách khác, α_1 biểu hiện tung độ gốc của GE và α_2, α_3 và α_4 là các hệ số *tung độ gốc chênh lệch*, cho biết các tung độ gốc của GM, US, và WEST chênh lệch với tung độ gốc của GE bao nhiêu. Nói ngắn gọn là GE trở thành công ty so sánh. Tất nhiên bạn được tùy ý chọn bất kỳ công ty nào làm công ty so sánh.

Nhân đây cũng xin nói rằng nếu bạn muốn các giá trị tung độ gốc rõ ràng cho mỗi công ty, bạn có thể đưa vào bốn biến giả, với điều kiện bạn chạy hồi quy của mình qua gốc tọa độ, nghĩa là, bỏ tung độ gốc chung trong (16.3.3); nếu bạn không làm thế, bạn sẽ rơi vào bẫy biến giả.

Bởi vì chúng ta sử dụng các biến giả để ước lượng các tác động cố định nên trong các tài liệu, mô hình (16.3.3) còn được gọi là **mô hình biến giả bình phương nhỏ nhất (LSDV)**. Vì thế các thuật ngữ các tác động cố định và LSDV có thể được sử dụng thay thế cho nhau. Nhân tiện, chú ý rằng mô hình LSDV (16.3.3) cũng được gọi là **mô hình hiệp biến (covariance model)** và X_2 và X_3 được gọi là *hiệp biến*.

Các kết quả dựa trên (16.3.3) là như sau:

$$\begin{array}{rlllllll}
Y = & -245{,}7924 & + 161{,}5722 D_{2i} & + 339{,}6328 D_{3i} & + 186{,}5666 D_{3i} & + 0{,}1079 X_{2i} + & 0{,}3461 X_{3i} \\
\text{se} = & (35{,}8112) & (46{,}4563) & (23{,}9863) & (31{,}5068) & (0{,}0175) & (0{,}0266) \\
t = & (-6{,}8635) & (3{,}4779) & (14{,}1594) & (5{,}9214) & (6{,}1653) & (12{,}9821)
\end{array}$$

$$R^2 = 0{,}9345 \quad d = 1{,}1076 \quad \text{df} = 74 \qquad \textbf{(16.3.4)}$$

Hãy so sánh hồi quy này với (16.3.1). Trong (16.3.4), tất cả hệ số ước lượng đều có ý nghĩa thống kê cao, vì các *giá trị p* của các hệ số *t* ước lượng cực kỳ nhỏ. Các giá trị

tung độ gốc của bốn công ty này khác nhau đáng kể về thống kê; của GE là –245,7924, của GM là –84,220 (= –245,7924 + 161,5722), của US là 93,8774 (= –245,7924 + 339,6328), và của WEST là –59,2258 (= –245,7924 + 186,5666). Những chênh lệch của các tung độ gốc này có thể do các đặc điểm độc đáo của mỗi công ty, như những khác biệt về phong cách quản lý hay tài năng quản lý.

Mô hình nào tốt hơn: (16.3.1) hay (16.3.4)? Câu trả lời thật là hiển nhiên, xem xét dựa vào ý nghĩa thống kê của các hệ số ước lượng, và dựa vào giá trị R^2 tăng đáng kể và giá trị *d* Durbin-Watson tăng lên, cho thấy rằng mô hình (16.3.1) đã được xác định sai. Tuy nhiên, giá trị R^2 gia tăng chẳng đáng ngạc nhiên bởi vì chúng ta có nhiều biến hơn trong mô hình (16.3.4).

Chúng ta có thể tạo ra một kiểm định chính thức về hai mô hình này. Trong quan hệ với mô hình (16.3.4), mô hình (16.3.1) là một mô hình giới hạn, theo nghĩa là nó áp đặt một tung độ gốc chung lên tất cả công ty. Vì thế cho nên chúng ta có thể sử dụng **kiểm định *F* giới hạn** đã thảo luận trong Chương 8. Sử dụng công thức (8.7.10), độc giả có thể dễ dàng kiểm tra rằng trong ví dụ hiện tại, giá trị *F* tính tóan được:

$$F\ \frac{\left(R^2_{UR} - R^2_R\right)/3}{\left(1 - R^2_{UR}\right)/74} = \frac{(0{,}9345 - 0{,}7565)/3}{(1 - 0{,}9345)/74} = 66{,}9980 \qquad \textbf{(16.3.5)}$$

trong đó giá trị R^2 giới hạn là từ (16.3.1) và R^2 không giới hạn là từ (16.3.4) và trong đó số ràng buộc bằng 3 do mô hình (16.3.1) giả định rằng các tung độ gốc của GE, GM, US, và WEST giống nhau.

Rõ ràng giá trị *F* bằng 66,9980 (đối với 3 bậc tự do ở tử số và 74 bậc tự do ở mẫu số) là có ý nghĩa cao và vì thế mô hình hồi quy giới hạn (16.3.1) dường như không có giá trị.

Tác động thời gian. Giống như chúng ta sử dụng các biến giả để giải thích cho tác động cá nhân (công ty), chúng ta có thể giải thích cho *tác động thời gian* theo nghĩa là hàm đầu tư Grunfeld dịch chuyển theo thời gian bởi vì các thay đổi về công nghệ, thay đổi về kiểm soát của chính phủ và/hoặc các chính sách thuế, và các tác động bên ngoài như chiến tranh hay các xung đột khác. Những tác động thời gian như thế có thể được giải thích dễ dàng nếu chúng ta đưa vào các biến giả thời gian, một biến cho mỗi năm. Bởi vì chúng ta có dữ liệu cho 20 năm, từ 1935 đến 1954, nên chúng ta có thể đưa vào 19 biến giả thời gian (tại sao?), và viết mô hình (16.3.3) thành:

$$Y_{it} = \lambda_0 + \lambda_1 \text{Dum35} + \lambda_2 \text{Dum36} + \ldots + \lambda_{19} \text{Dum53} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \qquad \textbf{(16.3.6)}$$

trong đó Dum35 (Biến giả thời gian 35) có giá trị 1 đối với quan sát trong năm 1935 và 0 nếu khác đi, v.v. Chúng ta xem năm 1954 là năm gốc, mà giá trị tung độ gốc của nó được cho trước bởi λ_0 (tại sao?)

Chúng ta không trình bày các kết quả hồi quy dựa trên (16.3.6), vì không một biến giả thời gian nào có ý nghĩa thống kê riêng biệt. Giá trị R^2 của mô hình (16.3.6) là 0,7697, trong khi giá trị đó của mô hình (16.3.1) là 0,7565, một lượng tăng thêm chỉ có 0,0132. Độc giả có thể tự làm phần sau đây như là một bài tập: hãy chỉ ra rằng, trên cơ sở kiểm định *F* giới hạn, lượng tăng thêm này không có ý nghĩa thống kê, mà có lẽ gợi ý

rằng tác động của năm hay tác động thời gian không có ý nghĩa về thống kê. Điều này có thể đề xuất rằng có lẽ hàm đầu tư không thay đổi nhiều theo thời gian.

Chúng ta đã thấy rằng các tác động của từng công ty là có ý nghĩa về thống kê, nhưng tác động của từng năm thì không. Phải chăng có thể là mô hình của chúng ta bị xác định sai, theo nghĩa là chúng ta đã không tính đến cả hai tác động thời gian và cá nhân kết hợp với nhau? Chúng ta hãy xem xét khả năng này.

Các hệ số độ dốc không đổi nhưng tung độ gốc thay đổi theo các cá nhân và thời gian

Để xét khả năng này, chúng ta có thể kết hợp (16.3.4) và (16.3.6), như sau:

$$Y_{it} = \alpha_1 + \alpha_2 D_{GMi} + \alpha_3 D_{USi} + \alpha_4 D_{WESTi} + \lambda_0 + \lambda_1 \text{Dum35} + \ldots + \lambda_{19} \text{Dum53} + \beta_2 X_{2i} + \beta_3 X_{3i} + u_{it} \quad \textbf{(16.3.7)}$$

Khi chúng ta chạy hồi quy này, chúng ta nhận thấy các biến giả công ty cũng như các hệ số của *X* đều có ý nghĩa về thống kê riêng biệt, nhưng không có biến giả thời gian nào có ý nghĩa thống kê cả. Thực chất là chúng ta trở về mô hình (16.3.4).

Kết luận chung xuất hiện là có lẽ có tác động của từng công ty rõ rệt nhưng không có tác động thời gian. Nói cách khác, các hàm đầu tư của bốn công ty này giống nhau, ngoại trừ các tung độ gốc của chúng. Trong tất cả trường hợp chúng ta đã xét, các biến *X* có tác động mạnh đến *Y*.

Tất cả hệ số thay đổi giữa các cá nhân

Ở đây, chúng ta giả định các tung độ gốc và các hệ số độ dốc khác nhau đối với tất cả đơn vị cá nhân hay là các đơn vị chéo. Điều này có nghĩa là các hàm đầu tư của GE, GM, US và WEST đều khác nhau. Chúng ta có thể dễ dàng mở rộng mô hình LSDV của chúng ta để bao hàm cả tình huống này. Hãy xét lại phương trình (16.3.4). Ở đó chúng ta đưa các biến giả cá nhân vào bằng cách *cộng thêm vào*. Nhưng trong Chương 9 về các biến giả, chúng ta đã cho thấy làm thế nào *các biến giả độ dốc, chênh lệch* hay *tương tác* có thể giải thích những chênh lệch trong các hệ số độ dốc. Trong bối cảnh hàm đầu tư Grunfeld, để làm được điều này thì chúng ta phải nhân mỗi biến giả công ty với mỗi biến *X* [làm như thế sẽ thêm sáu biến nữa vào mô hình (16.3.4)]. Đó là, chúng ta ước lượng mô hình sau đây:

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \gamma_1 (D_{2i} X_{2it}) + \gamma_2 (D_{2i} X_{3it}) + \gamma_3 (D_{3i} X_{2it}) + \gamma_4 (D_{3i} X_{3it}) + \gamma_5 (D_{4i} X_{2it}) + \gamma_6 (D_{4i} X_{3it}) + u_{it} \quad \textbf{(16.3.8)}$$

Bạn sẽ lưu ý rằng các hệ số γ là *các hệ số độ dốc chênh lệch* (*differential slope coefficients*), cũng như α_2, α_3 và α_4 là các *tung độ gốc chênh lệch* (*differential intercepts*). Nếu một hay nhiều hơn một hệ số γ có ý nghĩa về thống kê, thì điều này sẽ cho chúng ta biết rằng một hay nhiều hơn một hệ số độ dốc khác với nhóm cơ sở. Thí dụ , cho β_2 và γ_1 có ý nghĩa về thống kê. Trong trường hợp này, $(\beta_2+\gamma_1)$ sẽ cho ta giá trị của hệ số độ dốc của X_2 đối với General Motors, gợi ý rằng hệ số độ dốc của X_2 đối với GM khác với hệ số độ dốc của General Electric (General Electric là công ty so sánh của chúng ta).

Nếu tất cả tung độ gốc chênh lệch và tất hệ số độ dốc chênh lệch đều có ý nghĩa về thống kê, thì chúng ta có thể kết luận rằng các hàm đầu tư của General Motors, United

States Steel, và Westinghouse đều khác với hàm đầu tư của General Electric. Nếu điều này thực ra là đúng, thì có thể chẳng có lý trong việc ước lượng hồi quy kết hợp (16.3.1).

Chúng ta hãy xem xét các kết quả hồi quy dựa trên (16.3.8). Để dễ đọc, các kết quả hồi quy của (16.3.8) được cho dưới dạng bảng trong Bảng 16.2.

Như các kết quả này bộc lộ, Y có quan hệ đáng kể với X_2 và X_3. Tuy nhiên, nhiều hệ số độ dốc chênh lệch có ý nghĩa thống kê. Thí dụ, hệ số độ dốc của X_2 là 0,0902 đối với GE, nhưng là 0,1828 (0,0902 + 0,092) đối với GM. Điều thú vị là không có tung độ gốc chênh lệch nào có ý nghĩa về thống kê.

BẢNG 16.2 CÁC KẾT QUẢ HỒI QUI (16.3.8)

Biến số	Hệ số	Sai số chuẩn	giá trị *t*	giá trị *p*
Tung độ gốc	–9,9563	76,3518	–0,1304	0,8966
D_{2i}	–139,5104	109,2808	–1,2766	0,2061
D_{3i}	–40,1217	129,2343	–0,3104	0,7572
D_{4i}	9,3759	93,1172	0,1006	0,9201
X_{2i}	0,0926	0,0424	2,1844	0,0324
X_{3i}	0,1516	0,0625	2,4250	0,0180
$D_{2i}X_{2i}$	0,0926	0,0424	2,1844	0,0324
$D_{2i}X_{3i}$	0,2198	0,0682	3,2190	0,0020
$D_{3i}X_{2i}$	0,1448	0,0646	2,2409	0,0283
$D_{3i}X_{3i}$	0,2570	0,1204	2,1333	0,0365
$D_{4i}X_{2i}$	0,0265	0,1114	0,2384	0,8122
$D_{4i}X_{3i}$	–0,0600	0,3785	–0,1584	0,8745
		R^2 = 0,9511	d = 1,0896	

Nói chung, dường như các hàm đầu tư của bốn công ty này là khác nhau. Điều này có thể gợi ý rằng dữ liệu của bốn công ty này "không thể kết hợp lại". Trong trường hợp này người ta có thể ước lượng các hàm đầu tư của mỗi công ty một cách riêng biệt. (Xem bài tập 16.13.). Điều này nhắc nhở chúng ta rằng trong từng tình huống, các mô hình hồi quy dữ liệu bảng có thể không thích hợp, bất kể khả năng có sẵn cả dữ liệu chuỗi thời gian lẫn dữ liệu chéo.

Cảnh báo về việc sử dụng Mô hình Các Tác động Cố định hay LSDV. Mặc dù dễ sử dụng nhưng mô hình LSDV có một số vấn đề cần phải luôn ghi nhớ.

Thứ nhất, nếu bạn đưa vào mô hình quá nhiều biến giả, như trong trường hợp mô hình (16.3.7), bạn sẽ chạm trán với vấn đề khó khăn về số bậc tự do. Trong trường hợp mô hình (16.3.7), chúng ta có 80 quan sát, nhưng chỉ có 55 bậc tự do – chúng ta mất 3 bậc tự do đối với ba biến giả công ty, 19 bậc tự do đối với 19 biến giả năm, 2 bậc tự do đối với hai hệ số độ dốc, và 1 bậc tự do đối với tung độ gốc chung.

Thứ hai, với quá nhiều biến trong mô hình, luôn luôn có khả năng xảy ra hiện tượng đa cộng tuyến, vốn có thể gây khó khăn cho việc ước lượng chính xác (precise) một hoặc nhiều hơn một thông số.

Thứ ba, giả sử trong FEM (16.3.1), chúng ta cũng bao gồm các biến như giới tính, màu da, và sắc tộc. Những biến này cũng bất biến theo thời gian bởi vì giới tính, màu da,

hay sắc tộc của một cá nhân không thay đổi theo thời gian. Như thế, phương pháp LSDV có thể không có khả năng xác định tác động của các biến số bất biến theo thời gian.

Thứ tư, chúng ta phải suy nghĩ cẩn thận về số hạng sai số u_{it}. Tất cả kết quả chúng ta trình bày cho đến bây giờ được dựa trên giả định rằng số hạng sai số theo đúng các giả định cổ điển, đó là $u_{it} \sim N(0, \sigma^2)$. Do chỉ số *i* chỉ các quan sát chéo và *t* chỉ các quan sát chuỗi thời gian, nên có thể phải điều chỉnh giả định cổ điển về u_{it}. Có nhiều khả năng.

1. Chúng ta có thể giả định phương sai của sai số giống như nhau đối với tất cả đơn vị chéo hay chúng ta có thể giả định phương sai thay đổi.

2. Đối với mỗi cá nhân, chúng ta có thể giả định không có hiện tượng tự tương quan. Như thế, thí dụ, chúng ta có thể giả định rằng số hạng sai số của hàm đầu tư của General Motors là không tự tương quan. Hoặc chúng ta có thể giả định nó tự tương quan, thí dụ là tự tương quan bậc I (AR(1)).

3. Đối với một thời điểm định trước, có thể là số hạng sai số của General Motors tương quan với số hạng sai số thí dụ như của U.S. Steel hay với cả U.S. Steel lẫn Westinghouse[7]. Hay chúng ta có thể giả định không có sự tương quan như thế.

4. Chúng ta có thể nghĩ đến những cách hoán vị và những cách kết hợp khác đối với số hạng sai số. Như bạn có thể nhanh chóng nhận ra, tính đến một, hay nhiều hơn, các khả năng này sẽ làm cho phép phân tích phức tạp hơn nhiều. Các yêu cầu về toán học và chỗ để trình bày làm cho chúng ta không thể xét đến tất cả khả năng này. Bạn có thể tìm thấy nội dung thảo luận có phần dễ tiếp cận về các khả năng khác nhau này trong Dielman, Sayrs, và Kmenta[8]. Tuy nhiên, một số vấn đề khó khăn *có thể* được giảm nhẹ khi chúng ta cầu viện đến cái gọi là **mô hình các tác động ngẫu nhiên** mà chúng ta sẽ thảo luận tiếp theo đây.

16.4. ƯỚC LƯỢNG CÁC MÔ HÌNH HỒI QUI DỮ LIỆU BẢNG: PHƯƠNG PHÁP TÁC ĐỘNG NGẪU NHIÊN.

Mặc dầu ứng dụng dễ dàng, nhưng việc lập mô hình tác động cố định, hay LSDV có thể tốn nhiều chi phí nếu chúng ta xét đến bậc tự do khi chúng ta có nhiều đơn vị chéo. Bên cạnh đó, Kmenta lưu ý chúng ta là:

> Một câu hỏi hiển nhiên liên quan đến mô hình hiệp biến (nghĩa là mô hình LSDV) được đề cập đến là liệu việc thêm vào các biến giả, hậu quả là bậc tự do giảm, điều này có thật sự cần thiết hay không. Lý luận làm cơ sở cho mô hình hiệp biến là trong việc xác định mô hình hồi quy chúng ta đã không đưa vào các biến giải thích phù hợp vốn không thay đổi theo thời gian (và có thể các biến giải thích khác thực sự thay đổi theo thời gian nhưng có cùng giá trị đối với tất cả các đơn vị chéo), và việc đưa vào các biến giả là để biểu hiện sự ngu dốt của chúng ta [nhấn mạnh thêm] [9].

Nếu các biến giả thực sự biểu hiện sự thiếu kiến thức về mô hình (đúng) tại sao không biểu thị sự ngu dốt này thông qua số hạng nhiễu u_{it}? Đây đúng là phương pháp được đề nghị bởi những người ủng hộ cái gọi là **mô hình các thành phần sai số (error components model – ECM) hay mô hình các tác động ngẫu nhiên (Random Effects Model – REM)**.

Ý tưởng cơ bản là bắt đầu với phương trình (16.3.2):

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \qquad \textbf{(16.4.1)}$$

Thay vì coi β_{1i} như là hằng số, chúng ta giả định rằng đây là một biến ngẫu nhiên với giá trị trung bình là β_1 (không có ký hiệu dưới dòng *i* ở đây). Và giá trị tung độ gốc đối với một công ty đơn lẻ có thể được biểu thị như sau:

$$\beta_{1i} + \beta_1 + \varepsilon_i \qquad i = 1, 2, \ldots, N \qquad \textbf{(16.4.2)}$$

trong đó ε_i là một số hạng sai số ngẫu nhiên có giá trị trung bình là 0 và phương sai σ_ε^2.

Thực chất những gì chúng ta đề cập ở đây là rằng bốn doanh nghiệp được đưa vào mẫu của chúng ta là một mẫu lấy ra từ một tổng thể lớn hơn nhiều của những công ty như vậy và rằng chúng có một giá trị trung bình chung của tung độ gốc (=β_1) và những chênh lệch riêng lẻ trong các giá trị tung độ gốc của mỗi công ty được thể hiện trong số hạng sai số ε_i.

Thay (16.4.2) vào (16.4.1), chúng ta có:

$$\begin{aligned} Y_{it} &= \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + u_{it} \\ &= \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + w_{it} \end{aligned} \qquad \textbf{(16.4.3)}$$

trong đó

$$w_{it} = \varepsilon_i + u_{it} \qquad \textbf{(16.4.4)}$$

Số hạng sai số tổng hợp w_{it} gồm có hai thành phần, đó là ε_i là thành phần sai số chéo hay theo cá nhân, và u_{it}, thành phần sai số chéo và chuỗi thời gian kết hợp. Thuật ngữ *mô hình các thành phần sai số* trở thành tên của mô hình này bởi vì số hạng sai số tổng hợp w_{it} gồm có hai (hay nhiều hơn) thành phần sai số.

ECM đưa ra các giả định thông thường sau đây:

$$\begin{aligned} \varepsilon_i &\sim N(0, \sigma_\varepsilon^2) \\ u_{it} &\sim N(0, \sigma_u^2) \end{aligned} \qquad \textbf{(16.4.5)}$$

$$E(\varepsilon_i u_{it}) = 0 \qquad E(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j)$$

$$E(u_{it} u_{is}) = E(u_{it} u_{jt}) = E(u_{it} u_{js}) = 0 \qquad (i \neq j \,;\, t \neq s)$$

nghĩa là các thành phần sai số đơn lẻ không tương quan với nhau và không tự tương quan giữa các đơn vị chéo lẫn chuỗi thời gian.

Hãy cẩn thận lưu ý sự khác biệt giữa FEM và ECM. Trong FEM, mỗi đơn vị chéo có giá trị tung độ gốc (cố định) riêng của nó, cả thảy là *N* giá trị như thế cho *N* đơn vị chéo. Trái lại, trong ECM, tung độ gốc β_1 là giá trị trung bình của tất cả tung độ gốc (chéo) và thành phần sai số ε_i biểu hiện độ lệch (ngẫu nhiên) của từng tung độ gốc khỏi giá trị trung bình này. Tuy nhiên, hãy luôn ghi nhớ rằng ε_i không thể quan sát được một cách trực tiếp; nó là biến được gọi là **biến không thể quan sát** hay **tiềm ẩn**.

Kết quả của các giả định được phát biểu trong (16.4.5) là:

$$E(w_{it}) = 0 \qquad \textbf{(16.4.6)}$$

$$\text{var}(w_{it}) = \sigma_\varepsilon^2 + \sigma_u^2 \qquad \textbf{(16.4.7)}$$

Bây giờ, nếu $\sigma_\varepsilon^2 = 0$ thì không có sự khác biệt giữa các mô hình (16.2.1) và (16.4.3), trong trường hợp này chúng ta có thể đơn thuần kết hợp (gộp chung) tất cả quan sát (chuỗi thời gian và chéo) và chỉ chạy hồi quy kết hợp, như chúng ta đã làm trong (16.3.1).

Như (16.4.7) cho thấy, số hạng sai số w_{it} có phương sai không đổi. Tuy nhiên, chúng ta có thể chỉ ra rằng w_{it} và w_{is} ($t \neq s$) tương quan với nhau; nghĩa là các số hạng sai số của một đơn vị chéo cho trước tại hai thời điểm khác nhau tương quan với nhau. Hệ số tương quan, corr (w_{it}, w_{is}), là như sau:

$$\text{corr}\,(w_{it}, w_{is}) = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_u^2} \qquad \textbf{(16.4.8)}$$

Hãy lưu ý hai đặc điểm đặc biệt của hệ số tương quan trên đây. *Thứ nhất*, đối với bất kỳ đơn vị chéo cho trước nào, giá trị của hệ số tương quan giữa các số hạng sai số tại hai thời đoạn khác nhau vẫn giống như nhau bất kể hai thời đoạn cách xa nhau bao lâu, như có thể thấy rõ từ (16.4.8). Điều này tương phản mạnh với dạng bậc nhất [AR(1)] mà chúng ta đã thảo luận trong Chương 12, trong đó chúng ta đã tìm thấy tương quan giữa các thời đoạn sụt giảm theo thời gian. *Thứ hai*, cấu trúc tương quan đã cho trong (16.4.8) vẫn giống nhau đối với tất cả đơn vị chéo; nghĩa là, nó giống nhau đối với tất cả cá nhân.

Nếu chúng ta không tính đến cấu trúc tương quan này, và ước lượng (16.4.3) bằng phương pháp OLS, thì các toán tử ước lượng được tạo ra sẽ không hiệu quả. Phương pháp thích hợp nhất ở đây là phương pháp *bình phương nhỏ nhất tổng quát* (GLS).

Chúng ta sẽ không thảo luận về nội dung toán học của GLS trong bối cảnh hiện tại vì tính phức tạp của nó[10]. Bởi vì hiện nay hầu hết các phần mềm thống kê hiện đại đều có các thủ tục để ước lượng ECM (cũng như FEM), nên chúng ta sẽ chỉ trình bày các kết quả cho thí dụ về đầu tư của chúng ta. Nhưng trước khi chúng ta làm điều đó, xin lưu ý rằng chúng ta có thể dễ dàng mở rộng (16.4.4) để cho phép thành phần sai số ngẫu nhiên tính đến biến thiên theo thời gian (xem bài tập 16.6).

Các kết quả của việc ước lượng ECM về hàm đầu tư Grunfeld được trình bày trong Bảng 16.3. Chúng ta cần lưu ý vài khía cạnh của hồi quy này. *Thứ nhất*, nếu bạn tính tổng cộng các giá trị của tác động ngẫu nhiên đã cho đối với bốn công ty này, nó sẽ là zero, như nó phải như thế (tại sao?). *Thứ hai*, giá trị trung bình của thành phần sai số ngẫu nhiên, ε_i, là giá trị tung độ gốc chung –73,0353. Giá trị tác động ngẫu nhiên của GE là –169,9282, giá trị này cho chúng ta biết thành phần sai số ngẫu nhiên của GE khác với giá trị tung độ gốc chung bao nhiêu. Chúng ta có thể diễn giải tương tự cho ba giá trị khác của các tác động ngẫu nhiên. *Thứ ba*, chúng ta thu được giá trị R^2 từ hồi quy GLS biến đổi.

Nếu bạn so sánh các kết quả của mô hình ECM được cho trong Bảng 16.3 với các kết quả thu được từ FEM, bạn sẽ thấy rằng nhìn chung các giá trị hệ số của hai biến X dường như không khác nhau nhiều, ngoại trừ các giá trị được cho trong Bảng 16.2. Trong Bảng 16.2 chúng ta cho phép các hệ số độ dốc của hai biến này khác nhau giữa các đơn vị chéo.

BẢNG 16.3 ƯỚC LƯỢNG ECM VỀ HÀM ĐẦU TƯ GRUNFELD

Biến số	Hệ số	Sai số chuẩn	trị thống kê *t*	giá trị *p*
Tung độ gốc	–73,0353	83,9495	–0,8699	0,3870
X_2	0,1076	0,0168	6,4016	0,0000
X_3	0,3457	0,0168	13,0235	0,0000
Tác động ngẫu nhiên:				
GE	–169,9282			
GM	–9,5078			
USS	165,5613			

West	13,87475	
	R2 = 0,9323	(GLS)

16.5. MÔ HÌNH TÁC ĐỘNG CỐ ĐỊNH (LSDV) SO VỚI MÔ HÌNH TÁC ĐỘNG NGẪU NHIÊN

Thách thức mà một nhà nghiên cứu phải đối mặt là: Mô hình nào tốt hơn, FEM hay ECM? Câu trả lời cho câu hỏi này phụ thuộc vào giả định người ta đưa ra về tương quan có thể có giữa thành phần sai số chuyên biệt chéo hay cá nhân ε_i và các biến hồi quy độc lập *X*.

Nếu người ta giả định rằng ε_i và các biến *X không tương quan*, thì ECM có thể thích hợp, trong khi nếu ε_i và các biến *X có tương quan*, thì FEM có thể thích hợp.

Tại sao người ta kỳ vọng có mối tương quan giữa thành phần sai số cá nhân ε_i và một hay nhiều hơn một biến hồi quy độc lập? Hãy xét thí dụ sau đây. Giả sử chúng ta có một mẫu ngẫu nhiên lấy ra từ một số lượng nhiều cá nhân và chúng ta muốn lập mô hình hàm tiền lương hay thu nhập của họ. Giả sử thu nhập là một hàm phụ thuộc vào trình độ giáo dục, kinh nghiệm làm việc v.v. Bây giờ nếu chúng ta cho ε_i đại diện cho khả năng bẩm sinh, hoàn cảnh gia đình xuất thân, v.v thì khi chúng ta lập mô hình hàm thu nhập có bao gồm ε_i, ε_i rất có thể có tương quan với giáo dục, vì khả năng bẩm sinh và hoàn cảnh gia đình xuất thân thường là các yếu tố quyết định quan trọng của trình độ giáo dục. Như Wooldridge khẳng định "Trong nhiều ứng dụng, toàn bộ lý do sử dụng dữ liệu bảng là cho phép tác động không quan sát được [nghĩa là ε_i] tương quan với các biến giải thích."[11]

Các giả định làm cơ sở cho ECM là rằng ε_i là một mẫu lấy ra ngẫu nhiên từ một tổng thể lớn hơn nhiều. Nhưng đôi khi có thể không đúng như thế. Thí dụ, giả sử chúng ta muốn nghiên cứu tỷ lệ tội phạm giữa 50 tiểu bang ở Mỹ. Rõ ràng là, trong trường hợp này, giả định rằng 50 tiểu bang này không thể là một mẫu ngẫu nhiên.

Luôn ghi nhớ sự khác biệt cơ bản này trong hai phương pháp, chúng ta có thể nói gì thêm về sự chọn lựa giữa FEM và ECM? Ở đây các nhận định do Judge và các đồng sự đưa ra có thể hữu ích[12]:

1. Nếu *T* (số dữ liệu chuỗi thời gian) lớn và *N* (số đơn vị chéo) nhỏ, rất có thể chẳng có khác biệt trong các giá trị của các thông số được ước lượng bởi FEM và ECM. Như thế, sự chọn lựa ở đây dựa trên sự tiện lợi về sử dụng máy điện toán. Đối với điều đó thì FEM có thể được ưa thích hơn.

2. Khi *N* lớn và *T* nhỏ, các ước lượng thu nhận được bởi hai phương pháp này có thể khác nhau đáng kể. Hãy nhớ lại rằng trong ECM, $\beta_{1i} = \beta_1 + \varepsilon_i$, trong đó ε_i là thành phần ngẫu nhiên chéo, trong khi trong FEM, chúng ta xem β_{1i} là cố định và không ngẫu nhiên. Trong trường hợp thứ hai, sự suy luận thống kê phụ thuộc vào các đơn vị chéo quan sát được trong mẫu. Điều này thích hợp nếu chúng ta tin tưởng mạnh mẽ rằng các đơn vị cá nhân hay chéo trong mẫu của chúng ta không phải là những đơn vị được lấy ra ngẫu nhiên từ một mẫu lớn hơn. Trong trường hợp đó, FEM là thích hợp. Tuy nhiên, nếu các đơn vị chéo trong mẫu không được xem là những đơn vị được lấy ra ngẫu nhiên, thì ECM là thích hợp, vì trong trường hợp này sự suy luận thống kê là không có điều kiện.

3. Nếu thành phần sai số cá nhân ε_i và một hay nhiều hơn một biến hồi quy độc lập tương quan với nhau, thì các toán tử ước lượng ECM bị chệch, trong khi đó các toán tử ước lượng thu được từ FEM thì không chệch.

4. Nếu N lớn và T nhỏ, và nếu các giả định cơ bản cho ECM được giữ đúng, thì các toán tử ước lượng ECM hiệu quả lớn hơn các toán tử ước lượng FEM.[13]

Có phải là có một kiểm định chính thức sẽ giúp chúng ta chọn lựa giữa FEM và ECM? Có, đó là kiểm định do Hausman xây dựng năm 1978.[14] Chúng ta sẽ không thảo luận về các chi tiết của kiểm định này vì chúng vượt quá phạm vi cuốn sách này.[15] Giả thuyết 'không' làm cơ sở cho kiểm định Hausman là các toán tử ước lượng FEM và ECM không khác nhau đáng kể. Trị thống kê kiểm định do Hausman xây dựng xấp xỉ tuân theo phân phối λ^2. Nếu giả thuyết 'không' bị bác bỏ, thì kết luận là ECM không thích hợp và sử dụng FEM chúng ta sẽ được thuận lợi hơn, trong trường hợp này, những suy luận thống kê sẽ phụ thuộc vào ε_i trong mẫu.

Bất kể kiểm định Hausman, điều quan trọng là luôn ghi nhớ lời cảnh báo của Johnston và DiNardo. Trong việc quyết định chọn giữa mô hình các tác động cố định và mô hình các tác động ngẫu nhiên, họ lập luận rằng, ". . . không có một qui tắc đơn giản nào giúp nhà nghiên cứu tìm cách vượt qua được "Vỏ dưa" của các tác động cố định và "Vỏ dừa" của sai số đo lường và chọn lựa năng động. Mặc dù chúng tốt hơn so với dữ liệu chéo, nhưng dữ liệu bảng không phải là phương thuốc trị bá bệnh cho tất cả các vấn đề của một nhà kinh tế lượng.

16.6. CÁC HỒI QUI DỮ LIỆU BẢNG: MỘT SỐ NHẬN XÉT ĐỂ KẾT LUẬN

Như đã lưu ý từ đầu, đề tài lập mô hình dữ liệu bảng rất rộng và phức tạp. Chúng ta chỉ mới thảo luận sơ qua. Trong số các đề tài mà chúng ta chưa thảo luận, có thể đề cập các đề tài sau đây.

1. Kiểm định giả thuyết với dữ liệu bảng.

2. Phương sai thay đổi và tự tương quan trong ECM.

3. Dữ liệu bảng không cân bằng

4. Các mô hình dữ liệu bảng động trong đó (các) giá trị trễ của biến hồi quy phụ thuộc (Y_{it}) xuất hiện như một biến giải thích.

5. Các phương trình đồng thời liên quan đến dữ liệu bảng.

6. Các biến phụ thuộc định tính và dữ liệu bảng.

Chúng ta có thể tìm thấy một hay nhiều hơn một đề tài này trong các tài liệu tham khảo được trích dẫn trong chương này, và độc giả nên tham khảo chúng để học thêm về đề tài này. Các tài liệu tham khảo này cũng trích dẫn nhiều nghiên cứu thực nghiệm trong nhiều lĩnh vực kinh doanh và kinh tế học khác nhau đã sử dụng các mô hình hồi quy dữ liệu bảng này. Những người mới bắt đầu nghiên cứu đề tài này được khuyên nên đọc một số ứng dụng này để cảm nhận được các nhà nghiên cứu thực sự thực hiện các mô hình như thế nào.

16.7. TÓM TẮT VÀ KẾT LUẬN

1. Các mô hình hồi quy dữ liệu bảng dựa vào dữ liệu bảng. Dữ liệu bảng gồm các quan sát về các đơn vị chéo hay cá nhân trong nhiều thời đoạn.

2. Sử dụng dữ liệu bảng có nhiều lợi điểm. *Thứ nhất*, chúng làm tăng qui mô mẫu đáng kể. *Thứ hai*, bằng cách nghiên cứu các quan sát chéo lập đi lập lại, dữ liệu bảng phù

hợp hơn với nghiên cứu về dynamics của thay đổi. *Thứ ba*, dữ liệu bảng làm cho chúng ta có thể nghiên cứu các mô hình hành vi phức tạp hơn.

3. Mặc dù có các lợi điểm quan trọng, nhưng dữ liệu bảng cũng đặt ra nhiều vấn đề về ước lượng và suy luận. Bởi vì dữ liệu như thế bao gồm các kích thước thời gian và chéo (không gian) nên người ta cần phải giải quyết các vấn đề gây trở ngại cho dữ liệu chéo (thí dụ, phương sai thay đổi) và dữ liệu chuỗi thời gian (thí dụ, hiện tượng tự tương quan). Ngoài ra còn có một số vấn đề nữa, như tương quan chéo trong các đơn vị cá nhân ở cùng một thời điểm.

4. Có nhiều kỹ thuật ước lượng để giải quyết một hay nhiều hơn một vấn đề này. Hai kỹ thuật nổi bật là (1) mô hình các tác động cố định (FEM) và (2) mô hình các tác động ngẫu nhiên (REM) hay mô hình các thành phần sai số (ECM).

5. Trong FEM, tung độ gốc trong mô hình hồi quy được phép khác nhau giữa các cá nhân do công nhận sự thực là mỗi đơn vị chéo hay cá nhân có thể có một số đặc điểm đặc biệt riêng của nó. Để tính đến các tung độ gốc khác nhau, người ta có thể sử dụng các biến giả. FEM sử dụng các biến giả được gọi là mô hình biến giả bình phương nhỏ nhất (LSDV). FEM thích hợp trong những tình huống mà tung độ gốc chuyên biệt theo cá nhân có thể tương quan với một hay nhiều hơn một biến hồi quy độc lập. Một bất lợi điểm của LSDV là nó dùng hết nhiều bậc tự do khi số đơn chéo, *N*, rất lớn. Trong trường hợp này chúng ta sẽ phải đưa vào *N* biến giả (nhưng kìm hãm số hạng tung độ gốc chung).

6. Một mô hình thay thế cho FEM là ECM. Trong ECM, người ta giả định rằng tung độ gốc của một đơn vị cá nhân được lấy ra ngẫu nhiên từ một tổng thể lớn hơn nhiều, với giá trị trung bình không đổi. Sau đó, tung độ gốc của cá nhân được thể hiện như một sự lệch khỏi giá trị trung bình không đổi này. Một ưu điểm của ECM so với FEM là nó tiết kiệm được bậc tự do, bởi vì chúng ta không phải ước lượng *N* tung độ gốc chéo. Chúng ta chỉ cần ước lượng giá trị trung bình của tung độ gốc và phương sai của nó. ECM thích hợp trong các tình huống mà tung độ gốc (ngẫu nhiên) của mỗi đơn vị chéo không tương quan với các biến hồi quy độc lập.

7. Kiểm địnhh Hausman có thể được sử dụng để chọn giữa FEM và ECM.

8. Bất kể tính phổ biến ngày càng tăng trong nghiên cứu ứng dụng, và bất kể khả năng có sẵn ngày càng tăng dữ liệu như thế, các hồi quy dữ liệu bảng có thể không thích hợp trong mọi tình huống. Người ta phải sử dụng một cách phán đoán thực tiễn nào đó trong mỗi trường hợp.

BÀI TẬP

Câu hỏi

16.1. Những đặc tính đặc biệt của (**a**) dữ liệu chéo, (**b**) dữ liệu chuỗi thời gian, và (**c**) dữ liệu bảng là gì?

16.2. Mô hình các tác động cố định (FEM) nghĩa là gì? Bởi vì dữ liệu bảng có cả kích thước thời gian lẫn kích thước không gian, FEM tính đến cả hai kích thước này như thế nào?

16.3. Mô hình các thành phần sai số (ECM) có nghĩa là gì? Nó khác với FEM như thế nào? Khi nào thì ECM thích hợp? Và khi nào FEM thích hợp?

16.4. Có sự khác biệt giữa FEM, mô hình biến giả bình phương nhỏ nhất (LSDV), và mô hình hiệp biến hay không?

16.5. Khi nào thì các mô hình hồi quy dữ liệu bảng không thích hợp? Hãy cho các thí dụ.

16.6. Làm thế nào bạn có thể mở rộng mô hình (16.4.4) để tính đến một thành phần sai số thời gian.

16.7. Tham chiếu dữ liệu về trứng và giá của chúng được cho trong Bảng 1.1. Mô hình nào có thể thích hợp ở đây, FEM hay ECM? Giải thích tại sao?

16.8. Trong các kết quả hồi quy trong (16.3.4), các tung độ gốc tác động cố định của bốn công ty này là gì? Các tác động này có khác nhau theo ý nghĩa thống kê không?

16.9. Đối với thí dụ về đầu tư đã thảo luận trong chương này, Bảng 16.3 cho ra các kết quả dựa trên ECM. Nếu bạn so sánh các kết quả này với những kết quả được cho trong (16.3.4), bạn rút ra được các kết luận tổng quát gì?

16.10. Dựa trên Michigan Income Dynamics Study (Nghiên cứu Sự Vận động theo thời gian của Thu nhập ở Michigan), Hausman đã cố gắng ước lượng một mô hình tiền lương, hay thu nhập, sử dụng một mẫu gồm 629 học sinh tốt nghiệp phổ thông trung học. Những người này được theo dõi trong một thời kỳ 6 năm, như thế cho chúng ta tất cả là 3.774 quan sát. Biến phụ thuộc trong nghiên cứu này là lôgarít của tiền lương, và các biến giải thích là tuổi (được chia thành nhiều nhóm tuổi), thất nghiệp trong năm trước đó, sức khỏe kém trong năm trước đó, tự tuyển dụng, miền cư trú (Nam = 1; 0 nếu khác đi), khu vực cư trú (nông thôn = 1; 0 nếu khác đi). Hausman đã sử dụng cả FEM lẫn ECM. Các kết quả được trình bày trong Bảng 16.4 (các sai số chuẩn ở trong ngoặc đơn):

- **a.** Các kết quả này có ý nghĩa kinh tế không?
- **b.** Có sự khác biệt lớn trong các kết quả do hai mô hình này tạo ra hay không? Nếu có, điều gì có thể giải thích cho những khác biệt này?
- **c.** Trên cơ sở dữ liệu được cho trong bảng nói trên, bạn sẽ chọn mô hình nào, nếu có.

BẢNG 16.4 CÁC PHƯƠNG TRÌNH TIỀN LƯƠNG
(BIẾN PHỤ THUỘC: LOG TIỀN LƯƠNG)

Biến số	Các tác động cố định		Các tác động ngẫu nhiên	
1. Nhóm tuổi 1 (20–35)	0,0557	(0,0042)	0,0393	(0,0033)
2. Nhóm tuổi 2 (35–45)	0,0351	(0,0051)	0,0092	(0,0036)
3. Nhóm tuổi 3 (45–55)	0,0209	(0,0055)	–0,0007	(0,0042)
4. Nhóm tuổi 4 (55–65)	0,0209	(0,0078)	–0,0097	(0,0060)
5. Nhóm tuổi 5 (65–)	–0,0171	(0,0155)	–0,0423	(0,0121)
6. Thất nghiệp năm trước	–0,0042	(0,0153)	–0,0277	(0,0151)
7. Sức khỏe kém năm trước	–0,0204	(0,0221)	–0,0250	(0,0215)
8. Tự tuyển dụng	–0,2190	(0,0297)	–0,2670	(0,0263)
9. Nam	–0,1569	(0,0656)	–0,0324	(0,0333)
10. Nông thôn	–0,0101	(0,0317)	–0,1215	(0,0237)
11. Hằng số	—	—	0,8499	(0,0433)
S^2	0,0567		0,0694	
Bậc tự do	3.135		3.763	

* 3774 quan sát; các sai số chuẩn trong ngoặc đơn.

Sao lại từ Cheng Hsiao, *Phân tích Dữ liệu Bảng*, Nhà Xuất bản Đại học Cambridge, 1986, trang 42.

Nguồn nguyên thủy: J. A. Hausman, "Các Kiểm định Đặc trưng trong Kinh tế lượng" Econometrica, tập 46, 1978, các trang 1251-1271.

Bài tập tình huống

16.11. Dựa vào dữ liệu trong Bảng 1.1.

- **a.** Cho Y = số trứng sản xuất ra (tính bằng triệu) và X = giá của trứng (xu Mỹ mỗi tá). Ước lượng mô hình: $Y_i = \beta_1 + \beta_2 X_i + u_i$ cho các năm 1990 và 1991 một cách riêng biệt.
- **b.** Gộp chung các quan sát của 2 năm này và ước lượng mô hình hồi quy kết hợp. Bạn đưa ra các giả định nào trong việc kết hợp dữ liệu?
- **c**. Sử dụng mô hình các tác động cố định, phân biệt 2 năm này, và trình bày các kết quả hồi quy.
- **d.** Phải chăng bạn có thể dùng mô hình các tác động cố định nhằm phân biệt 50 tiểu bang? Tại sao lại có thể? Tại sao lại không?
- **e.** Có hợp lý khi phân biệt cả tác động của tiểu bang lẫn tác động của năm không? Nếu có, bạn phải đưa vào bao nhiêu biến giả?
- **f.** Mô hình các thành phần sai số có thích hợp để mô hình hóa việc sản xuất trứng hay không? Tại sao và tại sao không? Xem thử bạn có thể ước lượng một mô hình như vậy bằng cách sử dụng, thí dụ như, Eviews.

16.12. Tiếp tục với bài tập 16.11. Trước khi quyết định chạy hồi quy kết hợp, bạn muốn tìm hiểu xem liệu dữ liệu "có thể kết hợp" hay không. Nhằm mục đích này, bạn quyết định sử dụng kiểm định Chow, đã thảo luận trong Chương 8. hãy cho thấy những tính toán cần thiết liên quan và xác định xem hồi quy kết hợp này có nghĩa không.

16.13. Hãy trở lại với hàm đầu tư Grunfeld được thảo luận trong Phần 16.2.

- **a.** Ước lượng hàm đầu tư Grunfeld cho GE, GM, U.S. Steel, và Westinghouse một cách riêng biệt. Các kết quả của việc kết hợp tất cả 80 quan sát đã được cho trong (16.3.1)
- **b.** Để xác định liệu hồi quy kết hợp (16.3.1) có thích hợp hay không, bạn quyết định tiến hành kiểm định Chow, đã thảo luận trong Chương 8. Hãy thực hiện kiểm định này. *Gợi ý*: Lấy RSS từ hồi quy kết hợp, lấy RSS từ mỗi trong bốn hàm đầu tư, và sau đó áp dụng kiểm định Chow.
- **c.** Từ kiểm định Chow, bạn rút ra được các kết luận gì? Nếu kết luận của bạn là không kết hợp dữ liệu này, thì bạn có thể nói gì về tính hữu dụng của các kỹ thuật hồi quy dữ liệu bảng?

16.14. Bảng 16.5 đưa ra dữ liệu về tỷ lệ thất nghiệp thường dân Y(%) và mức thù lao hàng giờ trong công nghiệp chế tạo tính bằng đô la Mỹ X (chỉ số, 1992 = 100) cho Canada, Anh và Mỹ trong thời kỳ 1980-1999. Hãy xét mô hình:

$$Y_{it} = \beta_1 + \beta_2 X_{it} + u_{it} \qquad (1)$$

- **a.** Tiên đóan quan hệ kỳ vọng giữa Y và X là gì? Tại sao?
- **b.** Ước lượng mô hình đã cho trong (1) cho mỗi quốc gia.
- **c.** Ước lượng mô hình, kết hợp tất cả 60 quan sát.
- **d.** Ước lượng mô hình các tác động cố định.
- **e.** Ước lượng mô hình các thành phần sai số.
- **f.** Mô hình nào tốt hơn, FEM hay ECM? Biện minh cho câu trả lời của bạn.

BẢNG 16.5 TỶ LỆ THẤT NGHIỆP VÀ MỨC THÙ LAO HÀNG GIỜ TRONG CÔNG NGHIỆP CHẾ TẠO Ở MỸ , CANADA, ANH, 1980-1999.

	Mỹ		Canada		Anh Quốc	
Quan sát	Thù lao $/giờ	Thất nghiệp %	Thù lao $/giờ	Thất nghiệp %	Thù lao $/giờ	Thất nghiệp %
1980	55,6	7,1	49,0	7,2	43,7	7,0
1981	61,1	7,6	54,1	7,3	44,1	10,5
1982	67,0	9,7	59,6	10,6	42,2	11,3
1983	68,8	9,6	63,9	11,5	39,0	11,8
1984	71,2	7,5	64,3	10,9	37,2	11,7
1985	75,1	7,2	63,5	10,2	39,0	11,2
1986	78,5	7,0	63,3	9,2	47,8	11,2
1987	80,7	6,2	68,0	8,4	60,2	10,3
1988	84,0	5,5	76,0	7,3	68,3	8,6
1989	86,6	5,3	84,1	7,0	67,7	7,2
1990	90,8	5,6	91,5	7,7	81,7	6,9
1991	95,6	6,8	100,1	9,8	90,5	8,8
1992	100,0	7,5	100,0	10,6	100,0	10,1
1993	102,7	6,9	95,5	10,7	88,7	10,5
1994	105,6	6,1	91,7	9,4	92,3	9,7
1995	107,9	5,6	93,3	8,5	95,9	8,7
1996	109,3	5,4	93,1	8,7	95,6	8,2
1997	111,4	4,9	94,4	8,2	103,3	7,0
1998	117,3	4,5	90,6	7,5	109,8	6,3
1999	123,2	4,0	91,9	5,7	112,2	6,1

Mức thù lao hàng giờ tính bằng đô la Mỹ, chỉ số 1992 = 100.
Nguồn: *Báo cáo về Kinh tế của Tổng thống Mỹ*, tháng 1 năm 2001, Bảng B109, trang 399.