

## Chương 16

# Các mô hình hồi quy dữ liệu bảng

Trong Chương 1 chúng ta đã thảo luận qua về các loại dữ liệu nhìn chung có sẵn để phân tích thực nghiệm, đó là **dữ liệu theo chuỗi thời gian**, **dữ liệu chéo theo không gian**, và **dữ liệu bảng**. Trong dữ liệu theo chuỗi thời gian, ta quan sát giá trị của một hay nhiều biến trong một khoảng thời gian (ví dụ như GDP trong một vài quý hay vài năm). Trong dữ liệu chéo theo không gian, giá trị của một hay nhiều biến được thu thập cho một vài đơn vị mẫu, hay thực thể, vào cùng một thời điểm (ví dụ như tỷ lệ tội phạm trong 50 bang ở Hoa Kỳ trong một năm nhất định). *Trong dữ liệu bảng, đơn vị chéo theo không gian (ví dụ như hộ gia đình, doanh nghiệp, hay tiểu bang) được khảo sát theo thời gian.* Nói vắn tắt, dữ liệu bảng có cả bình diện không gian cũng như thời gian.

Ta đã thấy một ví dụ về dữ liệu bảng trong Bảng 1.1, trình bày dữ liệu về số trứng sản xuất ra và giá trứng ở 50 tiểu bang Hoa Kỳ trong các năm 1990 và 1991. Trong một năm cho trước, dữ liệu về trứng và giá trứng của 50 tiểu bang tiêu biểu cho một mẫu chéo theo không gian. Trong một bang cho trước, có hai quan sát chuỗi thời gian về trứng và giá trứng. Như vậy, ta có tổng cộng  $(50 \times 2) = 100$  quan sát (kết hợp) đối với trứng sản xuất ra và giá trứng.

Dữ liệu bảng còn có những cách gọi khác, như **dữ liệu kết hợp** (kết hợp các quan sát theo chuỗi thời gian và theo không gian), **kết hợp các dữ liệu theo chuỗi thời gian và không gian**, **dữ liệu vi bảng**, **dữ liệu theo chiều dọc** (nghiên cứu theo thời gian đối với một biến hay một nhóm đối tượng), **phân tích lịch sử biến cố** (ví dụ, nghiên cứu sự biến thiên theo thời gian của các đối tượng thông qua các trạng thái hay các điều kiện nối tiếp), **phân tích nhóm** (ví dụ, theo dõi diễn tiến sự nghiệp của 1965 sinh viên tốt nghiệp của một trường kinh doanh). Cho dù có nhiều biến thể tinh tế, tất cả các tên gọi này về thực chất đều tiêu biểu cho sự biến thiên theo thời gian của các đơn vị chéo theo không gian. Do đó, chúng ta sử dụng thuật ngữ dữ liệu bảng theo ý nghĩa tổng quát để bao gồm một hay nhiều thuật ngữ này. *Và ta sẽ gọi các mô hình hồi qui dựa vào các dữ liệu này là mô hình hồi qui dữ liệu bảng.*

Dữ liệu bảng ngày càng được sử dụng nhiều trong nghiên cứu kinh tế. Có một vài bộ dữ liệu bảng nổi tiếng như:

1. **Nghiên cứu bảng về Động học Thu nhập (PSID)** do Viện Nghiên cứu Khoa học thuộc Đại học Michigan thực hiện. Bắt đầu vào năm 1968, mỗi năm Viện lại thu thập dữ liệu về khoảng 5000 hộ gia đình với các biến số nhân khẩu và kinh tế xã hội khác nhau.
2. Văn phòng Điều tra dân số của Bộ Thương mại thực hiện việc điều tra khảo sát tương tự như PSID, gọi là **Khảo sát Tham gia Chương trình và Thu nhập (SIPP)**. Bốn lần trong một năm, những người tham gia được phỏng vấn về điều kiện kinh tế của họ.

Cũng có nhiều cuộc điều tra khảo sát khác được thực hiện bởi các cơ quan chính phủ khác nhau.

Ngay từ đầu ta cũng nên lưu ý một cảnh báo. Đề tài hồi qui dữ liệu bảng thì rộng lớn, và phần nào liên quan đến toán học và thống kê khá phức tạp. Chúng ta chỉ hy vọng chạm đến một phần những vấn đề then chốt của các mô hình hồi qui dữ liệu bảng, còn chi tiết để lại cho phần tài liệu

tham khảo.<sup>1</sup> Nhưng cũng nên được báo trước rằng một số tài liệu tham khảo này cũng có tính chất hết sức kỹ thuật. May thay, các gói phần mềm thân thiện với người sử dụng như Limdep, PcGive, SAS, STATA, Shazam và Eviews, cùng nhiều phần mềm khác, đã giúp cho việc thực hiện hồi qui dữ liệu trở nên khá dễ dàng.

### 16.1 Tại sao phải sử dụng dữ liệu bảng?

Các ưu điểm của dữ liệu bảng so với dữ liệu theo chuỗi thời gian và không gian là gì? Baltagi liệt kê các ưu điểm sau đây của dữ liệu bảng:<sup>2</sup>

1. Vì dữ liệu bảng liên quan đến các cá nhân, doanh nghiệp, tiểu bang, đất nước, v.v... theo thời gian, nên nhất định phải có tính dị biệt (không đồng nhất) trong các đơn vị này. Kỹ thuật ước lượng dữ liệu bảng có thể chính thức xem xét đến tính dị biệt đó bằng cách xem xét các biến số có tính đặc thù theo từng cá nhân, được trình bày ngay sau đây. Ta sử dụng thuật ngữ *cá nhân* theo ý nghĩa chung bao gồm các đơn vị vi mô như các cá nhân, các doanh nghiệp, tiểu bang, và đất nước.
2. Thông qua kết hợp các chuỗi theo thời gian của các quan sát theo không gian, dữ liệu bảng cung cấp ‘những dữ liệu có nhiều thông tin hơn, đa dạng hơn, ít cộng tuyến hơn giữa các biến số, nhiều bậc tự do hơn và hiệu quả hơn.’
3. Thông qua nghiên cứu các quan sát theo không gian lặp lại, dữ liệu bảng phù hợp hơn để nghiên cứu *tính động của thay đổi*. Tình trạng thất nghiệp, luân chuyển công việc, và tính lưu chuyển lao động sẽ được nghiên cứu tốt hơn với dữ liệu bảng.
4. Dữ liệu bảng có thể phát hiện và đo lường tốt hơn những ảnh hưởng mà không thể quan sát trong dữ liệu chuỗi thời gian thuần túy hay dữ liệu chéo theo không gian thuần túy. Ví dụ, ảnh hưởng của luật tiền lương tối thiểu đối với việc làm và thu nhập có thể được nghiên cứu tốt hơn nếu chúng ta xem xét các đợt gia tăng tiền lương tối thiểu liên tiếp nhau trong mức lương tối thiểu của liên bang và (hoặc) tiểu bang.
5. Dữ liệu bảng giúp ta nghiên cứu những mô hình hành vi phức tạp hơn. Ví dụ, các hiện tượng như lợi thế kinh tế theo qui mô và thay đổi kỹ thuật có thể được xem xét thông qua dữ liệu bảng tốt hơn so với dữ liệu theo chuỗi thời gian thuần túy hay theo không gian thuần túy.
6. Bằng cách thu thập những số liệu có sẵn cho vài nghìn đơn vị, dữ liệu bảng có thể tối thiểu hóa sự thiên lệch có thể xảy ra nếu ta tổng hợp các cá nhân hay các doanh nghiệp thành số liệu tổng.

---

<sup>1</sup>Một số tài liệu tham khảo như của G. Chamberlain, ‘Panel Data,’ trong *Handbook of Econometrics*, tập II, Z. Griliches và M. D. Intriligator chủ biên, North Hollans Publishers, 1984, chương 22; C. Hsiao, *Analysis of Panel Data*, Cambridge University Press, 1986; G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lukepohl, và T. C. Lee, *Introduction to the Theory and Practice of Econometrics*, xuất bản lần thứ hai, John Wiley & Sons, New York, 1985, chương 11; W. H. Greene, *Econometric Analysis*, xuất bản lần thứ 4, Prentice Hall, Englewood Cliffs, N. J., 2000, chương 14; Badi H. Baltagi, *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, Mass., 1999.

<sup>2</sup> Baltagi, tài liệu đã dẫn, trang 3-6.

Nói vắn tắt, dữ liệu bảng có thể làm phong phú các phân tích thực nghiệm theo những cách thức mà không chắc có thể đạt được nếu ta chỉ sử dụng các dữ liệu theo chuỗi thời gian hay không gian thuần túy. Điều này không có nghĩa rằng ta không có vấn đề gì với việc lập mô hình dữ liệu bảng. Ta sẽ thảo luận về những vấn đề này sau khi ta tìm hiểu ít nhiều lý thuyết và thảo luận một ví dụ.

## 16.2 Dữ liệu bảng: Một ví dụ minh họa

Để chuẩn bị, ta hãy xem xét một ví dụ cụ thể. Xem số liệu cho trong Bảng 16.1, được lấy từ một nghiên cứu nổi tiếng về lý thuyết đầu tư do Y. Grunfeld đề xuất.<sup>3</sup>

Grunfeld quan tâm đến việc tìm hiểu xem tổng đầu tư thực ( $Y$ ) phụ thuộc như thế nào vào giá trị thực của doanh nghiệp ( $X_2$ ) và trữ lượng vốn thực ( $X_3$ ). Cho dù nghiên cứu ban đầu bao gồm một số công ty, vì mục đích minh họa, ta thu thập dữ liệu cho bốn công ty, General Electric (GE), General Motor (GM), US Steel (US), và Westinghouse (WEST). Dữ liệu mỗi công ty về ba biến số trên đây có sẵn trong giai đoạn 1935-1954. Như vậy, ta có bốn đơn vị theo không gian và 20 thời đoạn. Do đó, tổng cộng ta có 80 quan sát. Tiên nghiệm,  $Y$  dự kiến có quan hệ đồng biến với  $X_2$  và  $X_3$ .

Trên nguyên tắc, ta có thể chạy bốn phép hồi qui theo chuỗi thời gian, một hồi qui cho mỗi công ty; hoặc ta cũng có thể chạy 20 phép hồi qui theo không gian, mỗi năm một phép hồi qui, cho dù trong trường hợp sau ta sẽ phải lo lắng về bậc tự do.<sup>4</sup>

Kết hợp tất cả 80 quan sát, ta có thể viết hàm đầu tư Grunfeld như sau:

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

$$i = 1, 2, 3, 4$$

$$t = 1, 2, \dots, 20$$
(16.2.1)

trong đó  $i$  tiêu biểu cho đơn vị thứ  $i$  (cá nhân thứ  $i$ ) và  $t$  tiêu biểu cho thời đoạn thứ  $t$ . Theo qui ước, ta chọn  $i$  là ký hiệu đơn vị theo không gian và  $t$  là ký hiệu theo thời gian. Ta giả định rằng có một số lượng tối đa  $N$  đơn vị chéo và một số lượng tối đa  $T$  thời đoạn. Nếu mỗi đơn vị theo không gian có cùng một số lượng quan sát như nhau theo chuỗi thời gian, thì dữ liệu bảng này được gọi là **bảng cân đối**. Trong bảng hiện đang xem xét, ta có một bảng cân đối, vì mỗi công ty trong mẫu đều có 20 quan sát. Nếu số quan sát khác nhau giữa các phần tử của bảng, ta gọi đó là **bảng không cân đối**. Trong chương này, nói chung ta chỉ quan tâm đến bảng cân đối.

Đầu tiên, ta giả định rằng các biến số  $X$  không ngẫu nhiên và các số hạng sai số tuân theo các giả định kinh điển, ấy là  $E(u_{it}) \sim N(0, \sigma^2)$ .

<sup>3</sup> Grunfeld, 'The Determinants of Corporate Investment,' luận án tiến sĩ không xuất bản, phòng Kinh tế, đại học Chicago, 1958. Dữ liệu được giới thiệu lại trong một vài quyển sách. Chúng tôi lấy từ nghiên cứu của H. D. Vinod và Aman Ullha, *Recent Advances in Regression Methods*, Marcel Dekker, New York, 1981, trang 259-261. Nghiên cứu Grunfeld đã trở thành một nghiên cứu được ưa thích của các tác giả viết sách giáo khoa vì dữ liệu dễ sử dụng cho mục đích minh họa.

<sup>4</sup> Đối với mỗi năm, ta chỉ có bốn quan sát đối với biến hồi qui phụ thuộc và các biến hồi qui độc lập. Nếu ta cũng cho phép có tung độ gốc, ta sẽ phải ước lượng ba thông số, chỉ còn lại một bậc tự do. Hiển nhiên, một phép hồi qui như vậy xem ra không chắc có ý nghĩa.

Cẩn thận lưu ý ký hiệu ghép đôi và ký hiệu ghép ba, mà tự chúng đã giải thích.

Làm thế nào ta ước lượng phương trình (16.2.1)? Câu trả lời như sau.

**Bảng 16.1 Dữ liệu đầu tư đối với bốn công ty, 1935-54**

Quan sát	I	F <sub>-1</sub>	C <sub>-1</sub>	Quan sát	I	F <sub>-1</sub>	C <sub>-1</sub>
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	1796.3	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	402.0	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	4926.9	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	226.3	1954	68.60	1188.9	213.5
------	--------	--------	-------	------	-------	--------	-------

*Chú thích:*  $Y = I =$  Tổng đầu tư = Bổ sung nhà máy thiết bị công bảo trì và sửa chữa; đơn vị tính: triệu USD giảm phát theo  $P_1$ .

$X_2 = F =$  Giá trị doanh nghiệp = Giá cổ phiếu phổ thông và cổ phiếu ưu đãi vào ngày 31-12 (hay giá bình quân của ngày 31-12 và ngày 31-1 của năm sau) nhân cho số cổ phiếu phổ thông và cổ phiếu ưu đãi đang lưu hành cộng tổng giá trị sổ sách của nợ vào ngày 31-12; đơn vị tính: triệu USD giảm phát theo  $P_2$ .

$X_2 = C =$  Trữ lượng máy móc thiết bị = Tổng lũy kế của bổ sung ròng máy móc thiết bị giảm phát theo  $P_1$  trừ đi khấu hao giảm phát theo  $P_3$  với các định nghĩa sau đây:

$P_1 =$  Hệ số giảm phát giá ngầm ẩn đối với thiết bị lâu bền của nhà sản xuất (1947 = 100).

$P_2 =$  Hệ số giảm phát giá ngầm ẩn đối với GNP (1947 = 100).

$P_3 =$  Hệ số giảm phát chi phí khấu hao = Bình quân di động 10 năm của chỉ số giá bán buôn kim loại và sản phẩm kim loại (1947 = 100).

Nguồn: Trình bày lại từ nghiên cứu của H. D. Vinod và Aman Ullah, *Recent Advances in Regression Methods*, Marcel Dekker, New York, 1981, trang 259-261.

### 16.3 Ước lượng các mô hình hồi qui dữ liệu bảng: Cách tiếp cận các ảnh hưởng cố định

Việc ước lượng phương trình (16.2.1) phụ thuộc vào những giả định mà ta nêu lên về tung độ gốc, các hệ số độ dốc, và số hạng sai số  $u_{it}$ . Có một vài khả năng có thể xảy ra:<sup>5</sup>

1. Giả định rằng các hệ số độ dốc và tung độ gốc là hằng số theo thời gian và không gian, và số hạng sai số thể hiện sự khác nhau theo thời gian và theo các cá nhân.
2. Các hệ số độ dốc là hằng số nhưng tung độ gốc thay đổi theo các cá nhân.
3. Các hệ số độ dốc là hằng số 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) đều thay đổi theo các cá nhân.
5. Tung độ gốc cũng như các hệ số độ dốc đều thay đổi theo các cá nhân và theo thời gian.

Như bạn có thể thấy, mỗi trường hợp này sẽ cho thấy tính phức tạp tăng dần (và có lẽ cũng sát thực tế hơn) trong việc ước lượng các mô hình hồi qui dữ liệu bảng, như mô hình (16.2.1). Lẽ dĩ nhiên, tính phức tạp sẽ gia tăng nếu ta bổ sung thêm các biến hồi qui độc lập vào mô hình do khả năng có thể xảy ra hiện tượng cộng tuyến giữa các biến hồi qui độc lập.

Việc tìm hiểu sâu xa từng khả năng trong các khả năng nêu trên sẽ đòi hỏi phải viết một quyển sách riêng, và hiện đã có một vài quyển sách như thế trên thị trường.<sup>6</sup> Trong những phần tiếp theo, chúng ta sẽ tìm hiểu một vài đặc điểm chính của các khả năng khác nhau này, đặc biệt là bốn khả năng đầu tiên. Thảo luận của chúng ta sẽ không đi sâu vào mặt kỹ thuật.

#### 1. Tất cả các hệ số đều không đổi theo thời gian và theo các cá nhân

Cách tiếp cận đơn giản nhất và có lẽ khá ngây thơ là bỏ qua bình diện không gian và thời gian của dữ liệu kết hợp và chỉ ước lượng hồi qui OLS thông thường. Nghĩa là, xếp chồng lên nhau 20

<sup>5</sup> Phần thảo luận này chịu ảnh hưởng của nghiên cứu của Judge và những người khác, tài liệu đã dẫn, và nghiên cứu của Hsiao, tài liệu đã dẫn, trang 9-10.

<sup>6</sup> Ngoài những quyển sách đã đề cập trong chú thích số 1, xem thêm sách của Terry E. Dielman, *Pooled Cross-sectional and Time Series Data Analysis*, Marcel Dekker, New York, 1989, và Lois W. Sayrs, *Pooled Time Series Analysis*, Sage Publications, Newbury Park, California, 1989.

quan sát của từng công ty, qua đó, ta có tổng cộng 80 quan sát cho từng biến số trong mô hình. Các kết quả hồi qui OLS là như sau:

$$\begin{aligned} \hat{Y} &= -63.3041 + 0.1101 X_2 + 0.3034 X_3 \\ se &= (29.6124) \quad (0.0137) \quad (0.0493) \\ t &= (-2.1376) \quad (8.0188) \quad (6.1545) \\ R^2 &= 0.7565 \quad \text{Durbin Watson} = 0.2187 \\ n &= 80 \quad df = 77 \end{aligned} \tag{16.3.1}$$

(se = sai số chuẩn; df = bậc tự do)

Nếu bạn xem xét các kết quả của **hồi qui kết hợp**, và áp dụng các tiêu chí thông thường, bạn sẽ thấy rằng tất cả các hệ số đều có ý nghĩa thống kê một cách riêng lẻ; các hệ số độ dốc có dấu dương như dự kiến và giá trị  $R^2$  cao một cách hợp lý. Như dự kiến,  $Y$  có quan hệ đồng biến với  $X_2$  và  $X_3$ . Con số ‘duy nhất’ làm rầu nồi canh là trị thống kê Durbin Watson ước lượng khá thấp, cho thấy rằng có lẽ có sự tự tương quan trong dữ liệu. Lẽ dĩ nhiên, như ta biết, trị thống kê Durbin Watson thấp cũng có thể do sai số về xác định qui cách mô hình. Ví dụ, mô hình ước lượng giả định rằng giá trị tung độ gốc của GE, GM, US và Westinghouse là như nhau. Mô hình cũng có thể giả định rằng các hệ số độ dốc của hai biến  $X$  hoàn toàn hết nhau đối với cả bốn công ty. Hiển nhiên, đó là những giả định hết sức hạn chế. Do đó, bất chấp tính đơn giản, hồi qui kết hợp (16.2.1) có thể bóp méo bức tranh thực tế về mối quan hệ giữa  $Y$  và các biến số  $X$  trong bốn công ty. Điều ta cần làm là tìm cách nào để xem xét bản chất cụ thể của bốn công ty. Phần tiếp theo sẽ giải thích cách làm điều này.

## 2. Các hệ số độ dốc là hằng số nhưng tung độ gốc thay đổi theo các cá nhân: Mô hình các ảnh hưởng cố định hay mô hình hồi qui biến giả bình phương tối thiểu (Least Square Dummy Variable, LSDV)

Một cách để xem xét ‘đặc điểm cá nhân’ của từng công ty hay từng đơn vị theo không gian là để cho tung độ gốc thay đổi theo từng công ty nhưng vẫn giả định rằng các hệ số độ dốc là hằng số đối với các công ty. Để thấy điều này, ta viết mô hình (16.2.1) là:

$$Y_{it} = \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \tag{16.3.2}$$

Lưu ý rằng ta đã đặt ký hiệu  $i$  vào số hạng tung độ gốc để cho thấy rằng các tung độ gốc của bốn công ty có thể khác nhau; sự khác biệt có thể là do các đặc điểm riêng của từng công ty, như phong cách quản lý hay triết lý quản lý.

Trong tư liệu nghiên cứu, mô hình (16.3.2) được gọi là **mô hình các ảnh hưởng cố định** (Fixed Effects Model, FEM). Thuật ngữ ‘các ảnh hưởng cố định’ này là do: cho dù tung độ gốc có thể khác nhau đối với các cá nhân (ở đây là bốn công ty), nhưng tung độ gốc của mỗi công ty không thay đổi theo thời gian; nghĩa là *bất biến theo thời gian*. Lưu ý là nếu ta viết tung độ gốc là  $\beta_{1it}$ , điều đó cho thấy rằng tung độ gốc của mỗi công ty hay cá nhân *thay đổi theo thời gian*. Có thể lưu ý rằng mô hình các ảnh hưởng cố định thể hiện qua phương trình (16.3.2) giả định rằng các hệ số (độ dốc) của các biến độc lập không thay đổi theo các cá nhân hay theo thời gian.

Trên thực tế ta cho phép tung độ gốc (ảnh hưởng cố định) khác nhau giữa các công ty như thế nào? Ta có thể dễ dàng làm điều đó thông qua kỹ thuật biến giả mà ta đã học trong Chương 9 mà cụ thể là **biến giả tung độ gốc khác biệt**. Do đó, ta viết (16.3.2) là:

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

Trong đó  $D_{2i} = 1$  nếu quan sát thuộc về GM hoặc bằng 0 trong những trường hợp khác;  $D_{3i} = 1$  nếu quan sát thuộc về US hoặc bằng 0 trong những trường hợp khác; và  $D_{4i} = 1$  nếu quan sát thuộc về WEST hoặc bằng 0 trong những trường hợp khác. Vì ta có 4 công ty, ta chỉ sử dụng ba biến giả để tránh rơi vào **bẫy biến giả** (nghĩa là tình huống cộng tuyến hoàn hảo). Ở đây không có biến giả cho GE. Nói cách khác,  $\alpha_1$  tiêu biểu cho tung độ gốc của GE và  $\alpha_2, \alpha_3, \alpha_4$  là các hệ số *tung độ gốc khác biệt* cho ta biết các tung độ gốc của GM, US, và WEST khác biệt như thế nào so với tung độ gốc của GE. Nói vắn tắt, GE trở thành công ty so sánh. Lẽ dĩ nhiên, bạn được tự do chọn bất kỳ công ty nào làm công ty so sánh.

Nhân thể, nếu bạn muốn mỗi công ty đều có các giá trị tung độ gốc cụ thể, bạn có thể đưa ra bốn biến giả, miễn là bạn chạy hồi qui thông qua gốc tọa độ, nghĩa là bạn phải bỏ đi tung độ gốc chung trong (16.3.3); nếu bạn không làm điều này, bạn sẽ rơi vào bẫy biến giả.

Vì bạn đang sử dụng các biến giả để ước lượng các ảnh hưởng cố định, trong tư liệu nghiên cứu, mô hình (16.3.3) còn được gọi là **mô hình biến giả bình phương tối thiểu (LSDV)**. Như vậy, các thuật ngữ ảnh hưởng cố định và LSDV có thể sử dụng với ý nghĩa như nhau. Nhân thể cũng lưu ý rằng mô hình LSDV (16.3.3) còn được gọi là **mô hình đồng phương sai**, và  $X_2$  và  $X_3$  còn gọi là *biến đồng phương sai*.

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

$$\begin{aligned} \hat{Y}_{it} &= -245.7924 + 161.5722 D_{2i} + 339.6328 D_{3i} + 186.5666 D_{4i} + 0.1079 X_{2i} + 0.3461 X_{3i} \\ se &= (35.8112) \quad (46.4563) \quad (23.9863) \quad (31.5068) \quad (0.0175) \quad (0.0266) \\ t &= (-6.8635) \quad (3.4779) \quad (14.1594) \quad (5.9214) \quad (6.1653) \quad (12.9821) \\ R^2 &= 0.9345 \quad d = 1.1076 \quad df = 74 \end{aligned} \quad (16.3.4)$$

So sánh hồi qui này với (16.3.1). Trong (16.3.4), tất cả các hệ số ước lượng đều có ý nghĩa thống kê khá cao một cách riêng lẻ, vì *các trị thống kê t* của các hệ số t ước lượng đều cực kỳ nhỏ. Giá trị tung độ gốc của bốn công ty đều khác nhau về mặt thống kê: tung độ gốc 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). Sự khác biệt về tung độ gốc này có thể do những đặc điểm riêng của từng công ty, như 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 sẽ bộc lộ hiển nhiên, được phán đoán bằng ý nghĩa thống kê của các hệ số ước lượng và sự kiện là giá trị  $R^2$  đã tăng đáng kể và sự kiện là trị thống kê Durbin Watson  $d$  cao hơn nhiều, cho thấy rằng mô hình (16.3.1) đã được xác định qui cách sai. Tuy nhiên, giá trị  $R^2$  tăng lên không có gì ngạc nhiên vì trong mô hình (16.3.4) có nhiều biến số hơn.

Ta cũng có thể đưa ra một phép kiểm định chính thức cho hai mô hình. Trong mối quan hệ với (16.3.4), mô hình (16.3.1) là một mô hình hạn chế ở chỗ nó áp đặt một tung độ gốc chung cho tất

cả các công ty. Do đó, ta có thể sử dụng **kiểm định F hạn chế** 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 này, trị thống kê F là:

$$F \frac{(R_{UR}^2 - R_R^2)/3}{(1 - R_{UR}^2)/74} = \frac{(0.9345 - 0.7565)/3}{(1 - 0.9345)/74} = 66.9980 \quad (16.3.5)$$

Trong đó giá trị  $R^2$  hạn chế được lấy từ mô hình (16.3.1) và  $R^2$  không hạn chế được lấy từ mô hình (16.3.4) và số điều kiện hạn chế là 3, vì mô hình (16.3.1) giả định rằng các tung độ gốc của GE, GM, US và WEST đều như nhau.

Rõ ràng, trị thống kê F bằng 66.9980 (đối với bậc tự do tử số 3 và bậc tự do mẫu số 74) hết sức có ý nghĩa và do đó, phép hồi qui hạn chế (16.3.1) xem ra không có giá trị.

**Ảnh hưởng của thời gian.** Cũng giống như ta sử dụng biến giả để xem xét ảnh hưởng cá nhân (công ty), ta cũng có thể xem xét *ảnh hưởng thời gian* theo ý nghĩa là hàm đầu tư Grunfeld dịch chuyển theo thời gian do những yếu tố như thay đổi công nghệ, thay đổi chính sách thuế hay qui định của chính phủ và các ảnh hưởng bên ngoài như chiến tranh và những xung đột khác. Những ảnh hưởng thời gian này có thể dễ dàng được xem xét nếu ta đưa vào mô hình các biến giả thời gian, mỗi năm một biến. Vì ta có số liệu cho 20 năm, từ 1935 đến 1954, nên ta có thể đưa vào 19 biến giả (tại sao?), và viết mô hình (16.3.3) là:

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

Trong đó Dum35 nhận giá trị bằng 1 đối với những quan sát trong năm 1935 và nhận giá trị bằng 0 đối với những quan sát trong những năm khác, v.v... Chúng ta xem năm 1954 là năm gốc, với giá trị tung độ gốc được cho bởi  $\lambda_0$  (tại sao?)

Ở đây chúng ta không trình bày các kết quả hồi qui dựa vào phương trình (16.3.6), vì không một biến giả thời gian nào có ý nghĩa thống kê một cách riêng lẻ. Giá trị  $R^2$  của (16.3.6) là 0.7697, trong khi giá trị  $R^2$  của (16.3.1) là 0.7565, chênh lệch chỉ có 0.0132. Chúng ta để dành cho độc giả phần này như một bài tập để chứng minh rằng, trên cơ sở kiểm định F hạn chế, chênh lệch này không có ý nghĩa thống kê, mà có thể cho thấy rằng, ảnh hưởng năm hay ảnh hưởng thời gian không có ý nghĩa. Điều này có thể cho thấy rằng có lẽ hàm đầu tư không thay đổi nhiều theo thời gian.

Chúng ta đã thấy rằng ảnh hưởng các công ty riêng lẻ có ý nghĩa thống kê, nhưng ảnh hưởng năm riêng lẻ không có ý nghĩa thống kê. Có lẽ nào ta đã xác định qui cách mô hình sai ở chỗ ta đã không xem xét đến cả ảnh hưởng cá nhân và ảnh hưởng thời gian cùng với nhau? Ta hãy xem xét khả năng này.

### 3. Các hệ số độ dốc là hằng số nhưng tung độ gốc khác nhau theo cá nhân cũng như theo thời gian

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

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

Khi ta chạy hồi qui này, ta thấy các biến giả công ty cũng như các hệ số của X đều có ý nghĩa thống kê một cách riêng lẻ nhưng không biến giả thời gian nào có ý nghĩa thống kê. Thực chất, ta quay lại với (16.3.4).



Kết luận chung ở đây là có lẽ có một ảnh hưởng riêng biệt của mỗi công ty thì sâu sắc nhưng không có ảnh hưởng thời gian. Nói cách khác, các hàm đầu tư của bốn công ty này là như nhau ngoại trừ đối với các tung độ gốc. Trong tất cả các trường hợp chúng ta đã xem xét, các biến X có một tác động mạnh đối với biến Y.

**4. Tất cả các hệ số đều thay đổi theo các cá nhân**

Ở đây chúng ta giả định rằng các tung độ gốc và các hệ số độ dốc đều khác nhau đối với mọi đơn vị riêng lẻ. Nói như thế có nghĩa là các hàm đầu tư của GE, GM, US, và WEST đều khác nhau hoàn toàn. Ta có thể dễ dàng mở rộng mô hình LSDV để xem xét tình huống này. Hãy xem lại mô hình (16.3.4). Ở đó ta đã giới thiệu các biến giả cá nhân theo cách thức *cộng thêm vào*. Nhưng trong Chương 9 về biến giả, ta đã trình bày *các biến giả độ dốc khác biệt* hay *tương tác* có thể giải thích sự khác biệt về hệ số độ dốc như thế nào. Để làm điều này trong bối cảnh hàm đầu tư Grunfeld, điều ta phải làm là nhân từng biến giả công ty cho từng biến số X [điều này sẽ làm tăng thêm 6 biến nữa cho mô hình (16.3.4)]. Nghĩa là ta ước lượng mô hình sau đây:

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

Bạn sẽ nhận thấy rằng các hệ số  $\gamma$  là *các hệ số độ dốc khác biệt*, cũng như  $\alpha_2, \alpha_3$  và  $\alpha_4$  là *các tung độ dốc khác biệt*. Nếu một hay nhiều hệ số  $\gamma$  có ý nghĩa thống kê, điều đó sẽ cho ta biết rằng một hay nhiều hệ số độ dốc là khác với nhóm gốc. Ví dụ, chẳng hạn như  $\beta_2$  và  $\gamma_1$  có ý nghĩa thống kê. Trong trường hợp này,  $(\beta_2 + \gamma_1)$  sẽ cho ta giá trị hệ số độ dốc của  $X_2$  đối với General Motor, cho thấy rằng hệ số độ dốc của  $X_2$  của GM khác với của General Electric, vốn là công ty so sánh của chúng ta.

Nếu tất cả các tung độ gốc khác biệt và tất cả các hệ số độ dốc khác biệt đều có ý nghĩa thống kê, ta có thể kết luận rằng hàm đầu tư của General Motor, United States Steel và Westinghouse đều khác với của General Electric. Nếu điều này là đúng trên thực tế, việc ước lượng hồi qui kết hợp (16.3.1) xem ra vô nghĩa.

Ta hãy xem các kết quả hồi qui dựa vào (16.3.8). Để dễ đọc, các kết quả hồi qui (16.3.8) được trình bày dưới dạng bảng trong Bảng 16.2.

Như các kết quả này cho thấy, Y có quan hệ có ý nghĩa thống kê với  $X_2$  và  $X_3$ . Tuy nhiên, một vài hệ số độ dốc khác biệt có ý nghĩa thống kê. Ví 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 một tung độ gốc khác biệt nào có ý nghĩa thống kê.

**Bảng 16.2 Các kết quả hồi qui (16.3.8)**

Biến số	Hệ số	Sai số chuẩn	Trị thống kê t	Trị thống kê p
Tung độ gốc	-9.9563	76.3518	-0.1430	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	0.3784	-0.1584	0.8745
	$R^2 = 0.9511$		$d = 1.0896$	

Nhìn chung, xem ra các hàm đầu tư của bốn công ty đều khác nhau. Điều này có thể cho thấy rằng dữ liệu của bốn công ty ‘không thể kết hợp’ được; trong trường hợp đó ta có thể ước lượng hàm đầu tư cho từng công ty một cách riêng biệt (xem bài tập 16.3). Điều này cũng nhắc ta nhớ rằng các mô hình hồi qui dữ liệu bảng không chắc phù hợp trong mọi tình huống, bất chấp sự sẵn có các dữ liệu cả theo không gian và thời gian.

**Thận trọng khi sử dụng mô hình các ảnh hưởng cố định hay mô hình LSDV.** Cho dù dễ sử dụng, mô hình LSDV có một vài vấn đề cần lưu ý.

*Thứ nhất*, nếu bạn đưa vào quá nhiều biến giả, như trong trường hợp mô hình (16.3.7), bạn sẽ vướng phải vấn đề 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 cho ba biến giả công ty, 19 bậc tự do cho 19 biến giả năm, 2 cho hai hệ số độ dốc, và 1 cho tung độ gốc chung.

*Thứ hai*, với nhiều biến số trong mô hình như thế, luôn luôn có khả năng đa cộng tuyến, làm cho việc ước lượng chính xác một hay nhiều thông số trở nên khó khăn.

*Thứ ba*, giả sử trong mô hình ảnh hưởng cố định (16.3.1), ta cũng bao gồm những biến như giới tính, màu da, và chủng tộc, cũng là những biến bất biến theo thời gian vì giới tính, màu da hay chủng tộc của một cá nhân không thay đổi theo thời gian. Vì thế, cách tiếp cận LSDV xem ra không thể nhận diện tác động của những biến số bất biến theo thời gian như vậy.

*Thứ tư*, ta phải suy nghĩ cẩn thận về số hạng sai số  $u_{it}$ . Tất cả các kết quả mà chúng ta trình bày cho đến giờ đều dựa vào giả định là số hạng sai số tuân theo các giả định kinh điển, ấy là,  $u_{it} \sim N(0, \sigma^2)$ . Vì chỉ số  $i$  tiêu biểu cho các quan sát theo không gian và chỉ số  $t$  tiêu biểu cho các quan sát theo thời gian nên giả định kinh điển đối với  $u_{it}$  có thể phải hiệu chỉnh. Có một vài khả năng:

1. Ta có thể giả định rằng phương sai sai số là như nhau đối với tất cả các đơn vị (các cá nhân), hay ta có thể giả định rằng phương sai sai số là không đồng nhất.
2. Đối với từng cá nhân, ta có thể giả định rằng không có tự tương quan theo thời gian. Như vậy, ví dụ, ta có thể giả định rằng số hạng sai số của hàm đầu tư đối với General Motor là không tự tương quan. Hay ta có thể giả định rằng nó tự tương quan, ví dụ như thuộc loại AR(1).

3. Trong một thời đoạn cho trước, có thể số hạng sai số của General Motor có tương quan với số hạng sai số của US Steel chẳng hạn hay với cả US Steel và Westinghouse.<sup>7</sup> Hoặc ta có thể giả định là không có tương quan như vậy.
4. Ta có thể suy nghĩ về các cách bố trí và kết hợp khác của số hạng sai số. Như bạn sẽ nhanh chóng nhận ra, cho phép một hay nhiều khả năng này xảy ra sẽ làm cho phân tích trở nên phức tạp hơn nhiều. Nhu cầu về không gian và toán học không cho phép ta xem xét mọi khả năng có thể xảy ra. Bạn đọc có thể tìm đọc những bài thảo luận khá dễ tiếp cận về các khả năng khác nhau trong các nghiên cứu của Dielman, Sayrs, và Kmenta.<sup>8</sup> Tuy nhiên, một vài vấn đề *có thể* được loại trừ nếu ta vận dụng cái gọi là **mô hình ảnh hưởng ngẫu nhiên** mà ta sẽ thảo luận sau đây.

#### 16.4 Ước lượng các mô hình hồi qui dữ liệu bảng: Cách tiếp cận ảnh hưởng ngẫu nhiên

Cho dù để áp dụng, việc lập mô hình ảnh hưởng cố định hay mô hình LSDV có thể làm giảm bậc tự do nếu ta có một vài đơn vị theo không gian. Ngoài ra, như Kmenta lưu ý:

Một vấn đề hiển nhiên liên quan đến mô hình đồng phương sai (tức là mô hình LSDV) là liệu việc thêm vào mô hình các biến giả - và vì thế mà mất đi một số bậc tự do - có thật sự cần thiết hay không. Lập luận làm nền tảng cho mô hình đồng phương sai là khi xác định qui cách mô hình, ta không thể bao gồm những biến giả thích phù hợp mà không thay đổi theo thời gian (và có thể những biến khác có thay đổi theo thời gian nhưng có cùng giá trị đối với tất cả các đơn vị theo không gian), và việc bao gồm biến giả là *sự che đậy tình trạng không hiểu biết của chúng ta* [phần nhân mạnh được bổ sung].<sup>9</sup>

Nếu quả thật các biến giả tiêu biểu cho tình trạng không hiểu biết của chúng ta về mô hình (thật sự), tại sao ta không biểu thị tình trạng không hiểu biết này thông qua số hạng nhiễu  $u_{it}$ ? Đây chính xác là cách tiếp cận được đề xuất bởi những người ủng hộ cái gọi là **mô hình các thành phần sai số (Error Components Model, ECM)** hay **mô hình ảnh hưởng ngẫu nhiên (Random Effects Model, REM)**.

Ý tưởng cơ bản là bắt đầu bằng (16.3.2):

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

Thay vì xem  $\beta_{1i}$  là cố định, ta giả định đó là một biến ngẫu nhiên với một giá trị trung bình là  $\beta_1$  (không có ký hiệu  $i$  ở đây). Và giá trị tung độ gốc cho một công ty riêng lẻ có thể được biểu thị là:

$$\beta_{1i} = \beta_1 + \varepsilon_i \quad i = 1, 2, \dots, N \quad (16.4.2)$$

<sup>7</sup> Điều này dẫn đến cái gọi là **lập mô hình hồi qui có vẻ không quan hệ** (seemingly unrelated regression, SURE) do Arnold Zellner đề xuất lần đầu. Tìm đọc thảo luận về mô hình này trong nghiên cứu của Terry E. Dielman, tài liệu đã dẫn.

<sup>8</sup> Dielman, tài liệu đã dẫn, Sayrs, tài liệu đã dẫn, Jan Kmenta, *Elements of Econometrics*, xuất bản lần thứ hai, Macmillan, New York, 1986, chương 12.

<sup>9</sup> Kmenta, tài liệu đã dẫn, trang 633.

trong đó  $\varepsilon_i$  là số hạng sai số ngẫu nhiên với một giá trị trung bình bằng 0 và phương sai bằng  $\sigma_\varepsilon^2$ .

Thực chất điều mà ta đang nói tới là: bốn công ty bao gồm trong mẫu của chúng ta được rút ra từ một tập hợp rộng lớn hơn nhiều gồm nhiều công ty như vậy và các công ty đó có một trị trung bình chung đối với tung độ gốc ( $= \beta_1$ ) và sự khác biệt cá nhân về giá trị tung độ gốc của từng công ty được phản ánh trong số hạng sai số  $\varepsilon_i$ .

Thay (16.4.2) vào (16.4.1), ta có:

$$\begin{aligned} Y_{it} &= \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + \varepsilon_i + u_{it} \\ &= \beta_{1i} + \beta_2 X_{2it} + \beta_3 X_{3it} + w_{it} \end{aligned} \quad (16.4.3)$$

Trong đó

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

Số hạng sai số kết hợp bao gồm hai thành phần:  $\varepsilon_i$  là thành phần sai số theo không gian, hay theo các cá nhân, và  $u_{it}$  là thành phần sai số theo không gian và chuỗi thời gian kết hợp. Thuật ngữ *mô hình các thành phần sai số* được đặt tên vì số hạng sai số kết hợp  $w_{it}$  gồm hai (hay nhiều) thành phần sai số.

Giả định thông thường mà mô hình các thành phần sai số đưa ra là:

$$\begin{aligned} \varepsilon_{it} &\sim N(0, \sigma_\varepsilon^2) \\ u_{it} &\sim N(0, \sigma_u^2) \\ E(\varepsilon_i u_{it}) &= 0 \quad E(\varepsilon_i \varepsilon_j) = 0 \quad (i \neq j) \\ E(u_{it} u_{is}) &= E(u_{it} u_{jt}) = E(u_{it} u_{js}) = 0 \quad (i \neq j; t \neq s) \end{aligned} \quad (16.4.5)$$

Nghĩa là, các thành phần sai số cá nhân không tương quan với nhau và không tự tương quan giữa các đơn vị theo không gian và theo chuỗi thời gian.

Cần thận lưu ý sự khác nhau giữa FEM và ECM. Trong FEM, mỗi đơn vị theo không gian có giá trị tung độ gốc (cố định) riêng, tổng cộng có N giá trị như vậy cho toàn bộ N đơn vị. Mặt khác, trong ECM, tung độ gốc  $\beta_1$  tiêu biểu cho trị trung bình của tất cả các tung độ gốc và số hạng sai số  $\varepsilon_i$  tiêu biểu cho sự sai lệch (ngẫu nhiên) của từng tung độ gốc so với trị trung bình này. Tuy nhiên, nên nhớ rằng  $\varepsilon_i$  không thể quan sát trực tiếp được; nó được gọi là **biến không thể quan sát**, hay **biến ẩn**.

Như một hệ quả của những giả định nêu lên trong (16.4.5), ta suy ra rằng:

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

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

Bây giờ nếu  $\sigma_\varepsilon^2 = 0$ , sẽ không có khác biệt gì giữa mô hình (16.2.1) và mô hình (16.4.3), trong trường hợp đó, ta có thể kết hợp tất cả các quan sát (theo không gian và theo chuỗi thời gian) và chỉ cần chạy hồi qui kết hợp, như ta đã làm trong mô hình (16.3.1).

Như (16.4.7) cho thấy, số hạng sai số  $w_{it}$  có phương sai mang tính đồng nhất. Tuy nhiên, ta có thể chứng minh rằng  $w_{it}$  và  $w_{is}$  ( $t \neq s$ ) tương quan với nhau; nghĩa là các số hạng sai số của một

đơn vị cho trước ở hai thời đoạn khác nhau thì tương quan với nhau. Hệ số tương quan  $corr(w_{it}, w_{it})$  là như sau:

$$corr(w_{it}, w_{it}) = \frac{\sigma_{\varepsilon}^2}{\sigma_{\varepsilon}^2 + \sigma_u^2} \quad (16.4.8)$$

Lưu ý hai thuộc tính đặc biệt của hệ số tương quan trên đây. *Thứ nhất*, đối với một đơn vị (cá nhân) cho trước, giá trị của tương quan giữa các số hạng sai số ở hai thời đoạn khác nhau vẫn hết như nhau bất kể hai thời đoạn cách nhau bao xa, như thể hiện rõ rệt từ (16.4.8). Điều này tương phản mạnh với cơ chế bậc một [AR(1)] mà chúng ta đã thảo luận trong chương 12, trong đó ta nhận thấy rằng mối tương quan giữa các thời đoạn giảm dần theo thời gian. *Thứ hai*, cơ cấu tương quan được cho trong (16.4.8) vẫn như nhau đối với mọi đơn vị; nghĩa là nó hết như nhau đối với mọi cá nhân.

Nếu ta không xem xét đến cơ cấu tương quan này, và ước lượng mô hình (16.4.3) bằng OLS, ước lượng đạt được sẽ không hiệu quả. Phương pháp thích hợp nhất ở đây là phương pháp  *bình phương tối thiểu tổng quát* (generalized least squares, GLS).

Ta sẽ không thảo luận về toán học phương pháp GLS trong bối cảnh này do tính phức tạp của nó.<sup>10</sup> Vì hầu hết các gói phần mềm thống kê hiện đại hiện nay đều có lộ trình ước lượng ECM (cũng như FEM), nên ở đây ta chỉ trình bày các kết quả cho ví dụ đầu tư của chúng ta mà thôi. Nhưng trước khi làm điều đó, xem ra nên lưu ý rằng ta có thể dễ dàng mở rộng (16.4.4) để cho một thành phần sai số ngẫu nhiên có thể xem xét đến sự biến thiên theo thời gian (xem bài tập 16.6).

Các kết quả của ước lượng ECM của hàm đầu tư Grunfeld được trình bày trong bảng 16.3. Ta nên lưu ý một vài khía cạnh của hồi qui này. *Thứ nhất*, nếu bạn cộng các giá trị ảnh hưởng ngẫu nhiên cho trước đối với bốn công ty, nó sẽ bằng không, vì nó phải như thế (tại sao?). *Thứ hai*, giá trị trung bình của thành phần sai số ngẫu nhiên  $\varepsilon_i$  là giá trị của tung độ gốc chung, bằng -73.0353. Giá trị ảnh hưởng ngẫu nhiên của GE là -169.9282 cho ta thấy thành phần sai số ngẫu nhiên của GE khác biệt bao nhiêu so với giá trị tung độ gốc chung. Cách lý giải tương tự cũng được áp dụng cho ba giá trị khác của ảnh hưởng ngẫu nhiên. *Thứ ba*, ta thu được giá trị  $R^2$  từ hồi qui GLS biến đổi.

Nếu bạn so sánh các kết quả của mô hình ECM trình bày trong Bảng 16.3 với các kết quả thu được từ mô hình FEM, bạn sẽ thấy rằng, nhìn chung, các giá trị hệ số của hai biến X xem ra không khác nhau nhiều, ngoại trừ những giá trị cho trong Bảng 16.2, trong đó ta cho các hệ số độ dốc của hai biến khác nhau giữa các cá nhân.

**Bảng 16.3 Ước lượng ECM của hàm đầu tư Grunfeld**

Biến số	Hệ số	Sai số chuẩn	Trị thống kê t	Trị thống kê p
Tung độ gốc	-73.0353	83.9495	-0.8699	0.3870
X <sub>2</sub>	0.1076	0.0168	6.4016	0.0000
X <sub>3</sub>	0.3457	0.0168	13.0235	0.0000

<sup>10</sup> Bạn đọc quan tâm có thể tham khảo phần thảo luận dễ tiếp cận trong nghiên cứu của Kmenta, tài liệu đã dẫn, trang 625-630.

---

Ảnh hưởng ngẫu nhiên:

GE	-169.9282
GM	-9.5078
USS	165.5613
Westinghouse	13.87475

$$R^2 = 0.9323 \quad (\text{GLS})$$


---

### 16.5 Mô hình ảnh hưởng cố định (hay mô hình LSDV) so với mô hình ảnh hưởng ngẫu nhiên

Nhà nghiên cứu đứng trước một thử thách là: Mô hình nào tốt hơn, FEM hay ECM? Câu trả lời cho câu hỏi này xoay quanh giả định mà ta đưa ra về mối tương quan khả dĩ giữa thành phần sai số theo cá nhân (hay theo đơn vị)  $\varepsilon_i$  và các biến hồi qui độc lập  $X$ .

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

Tại sao người ta dự kiến mối tương quan giữa thành phần sai số cá nhân  $\varepsilon_i$  và một hay nhiều biến hồi qui độc lập? Ta hãy xem một ví dụ. Giả sử ta có một mẫu ngẫu nhiên gồm một số lượng lớn các cá nhân và ta muốn lập mô hình hàm tiền lương hay thu nhập của họ. Giả sử thu nhập là một hàm số theo trình độ giáo dục, kinh nghiệm làm việc, v.v... Bây giờ nếu ta gọi  $\varepsilon_i$  tiêu biểu cho khả năng bẩm sinh, nền tảng gia đình v.v..., thì khi ta lập mô hình hàm thu nhập bao gồm  $\varepsilon_i$ , nó có thể tương quan với trình độ giáo dục, vì khả năng bẩm sinh và nền tảng gia đình thường là những yếu tố quan trọng xác định trình độ giáo dục. Như Wooldridge trình bày: ‘Trong nhiều ứng dụng, toàn bộ lý do khiến ta sử dụng dữ liệu bảng là để cho ảnh hưởng không quan sát được (nghĩa là  $\varepsilon_i$ ) được phép tương quan với các biến giải thích.’<sup>11</sup>

Các giả định làm nền tảng cho ECM là:  $\varepsilon_i$  được rút ra ngẫu nhiên từ một dân số lớn hơn nhiều. Nhưng đôi khi điều này không chắc đúng. Ví dụ, giả sử ta muốn nghiên cứu tỷ lệ tội phạm trong 50 tiểu bang Hoa Kỳ. Hiển nhiên trong trường hợp này, giả định cho rằng 50 tiểu bang là một mẫu ngẫu nhiên xem ra không thể biện hộ được.

Lưu ý sự khác biệt cơ bản này của hai cách tiếp cận, ta có thể nói gì thêm về việc chọn lựa giữa FEM và ECM? Ở đây các quan sát của Judge và những người khác có thể bổ ích.<sup>12</sup>

1. Nếu  $T$  (số thời đoạn của dữ liệu chuỗi thời gian) lớn và  $N$  (số đơn vị theo không gian) nhỏ, giá trị của các thống số ước lượng bằng FEM và ECM có thể sẽ không khác nhau nhiều. Vì thế, việc chọn lựa ở đây dựa vào sự thuận tiện trong tính toán. Về điểm này, FEM có thể đáng ưa chuộng hơn.
2. Khi  $N$  lớn và  $T$  nhỏ, các giá trị ước lượng thu được bằng hai phương pháp có thể khác nhau đáng kể. Nên nhớ rằng trong mô hình ECM,  $\beta_{1i} = \beta_1 + \varepsilon_i$ , trong đó  $\varepsilon_i$  là thành phần ngẫu nhiên theo cá nhân, trong khi trong mô hình FEM, ta xem  $\beta_{1i}$  là cố định và không ngẫu nhiên. Trong trường hợp mô hình FEM, suy luận thống kê được lập điều kiện theo các đơn vị được quan sát trong mẫu. Mô hình này sẽ phù hợp nếu ta tin tưởng mãnh liệt

---

<sup>11</sup> Wooldridge, tài liệu đã dẫn, trang 450.

<sup>12</sup> Judge và những người khác, tài liệu đã dẫn, trang 489-491.

rằng các cá nhân, hay các đơn vị riêng lẻ trong mẫu không phải được rút ra ngẫu nhiên từ một mẫu lớn hơn. Trong trường hợp đó, mô hình FEM là phù hợp. Tuy nhiên, nếu các đơn vị trong mẫu được xem là rút ra ngẫu nhiên, thì ECM sẽ thích hợp, trong trường hợp đó, suy luận thống kê có tính chất vô điều kiện.

3. Nếu thành phần sai số cá nhân  $\varepsilon_i$  và một hay nhiều biến độc lập tương quan với nhau, thì ước lượng ECM sẽ bị chệch, trong khi ước lượng thu được từ mô hình FEM sẽ không chệch.
4. Nếu  $N$  lớn và  $T$  nhỏ, và nếu các giả định làm nền tảng cho mô hình ECM được thỏa thì ước lượng ECM sẽ hiệu quả hơn so với ước lượng FEM.<sup>13</sup>

Có chăng một kiểm định chính thức giúp ta chọn lựa giữa mô hình FEM và ECM? Có, kiểm định này do Hausman xây dựng vào năm 1978.<sup>14</sup> Ta sẽ không thảo luận chi tiết kiểm định này, vì điều đó vượt ra ngoài phạm vi quyển sách này.<sup>15</sup> *Giả thiết không* làm nền tảng cho kiểm định Hausman là: các ước lượng FEM và ECM không khác nhau đáng kể. Trị thống kê kiểm định do Hausman xây dựng có một phân phối  $\chi^2$  tiệm cận. Nếu *giả thiết không* bị bác bỏ, kết luận là: ECM không thích hợp và tốt hơn xem ra ta nên sử dụng mô hình FEM, trong trường hợp đó, các suy luận thống kê sẽ lập điều kiện theo  $\varepsilon_i$  trong mẫu.

Bất chấp kiểm định Hausman, điều quan trọng là cần ghi nhớ cảnh báo của Johnston và DiNardo. Khi quyết định chọn lựa giữa mô hình ảnh hưởng cố định hay mô hình ảnh hưởng ngẫu nhiên, họ lập luận rằng: ‘... không có một qui tắc đơn giản nào giúp nhà nghiên cứu tránh chiếc vỏ dưa ảnh hưởng cố định hay chiếc vỏ dưa sai số đo lường và chọn lựa động học. Cho dù cải thiện hơn so với dữ liệu theo không gian, dữ liệu bảng không mang lại một phương thuốc bách bệnh cho tất cả các vấn đề của nhà kinh tế lượng.’<sup>16</sup>

### 16.6 Hồi qui dữ liệu bảng: Một vài nhận xét kết luận

Như đã lưu ý ngay từ đầu, đề tài lập mô hình dữ liệu bảng thì rộng lớn và phức tạp. Chúng ta chỉ mới lướt qua bề mặt. Trong những chủ đề mà ta chưa thảo luận, những vấn đề sau đây có thể được đề cập:

1. Kiểm định giả thiết với dữ liệu bảng.
2. Phương sai không đồng nhất và tự tương quan trong ECM.
3. Dữ liệu bảng không cân đối.
4. Các mô hình dữ liệu bảng động học, trong đó một hay nhiều giá trị trễ của biến 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.

<sup>13</sup> Taylor đã chứng minh rằng đối với  $T \geq 3$ , và  $(N - K) \geq 9$ , trong đó  $K$  là số lượng biến độc lập, thì phát biểu này được thỏa. Xem nghiên cứu của W. E. Taylor, ‘Small Sample Considerations in Estimation from Panel Data,’ *Journal of Econometrics*, tập 13, 1980, trang 203-223.

<sup>14</sup> J. A. Hausman, ‘Specification Tests in Econometrics,’ *Econometrica*, tập 46, 1978, trang 1251-1271.

<sup>15</sup> Xem chi tiết trong nghiên cứu của Baltagi, tài liệu đã dẫn, trang 68-73.

<sup>16</sup> Jack Johnston và John DiNardo, *Econometric Methods*, xuất bản lần thứ 4, McGraw Hill, 1997, trang 403.

Bạn đọc có thể tìm thấy một hay nhiều chủ đề này trong phần tài liệu tham khảo của chương này, và nên tham khảo để học hỏi thêm về đề tài này. Các tài liệu tham khảo này cũng trích dẫn một số nghiên cứu thực nghiệm trong những lĩnh vực kinh doanh và kinh tế học khác nhau đã sử dụng các mô hình hồi qui dữ liệu bảng. Chúng tôi khuyên những người mới bắt đầu nên đọc một số ứng dụng này để cảm nhận về cách thức các nhà nghiên cứu đã thật sự thực hiện những mô hình này như thế nào.

### 16.7 Tóm tắt và kết luận

1. Các mô hình hồi qui bảng dựa vào dữ liệu bảng. Dữ liệu bảng bao gồm các quan sát về các đơn vị riêng lẻ, hay các đơn vị theo không gian trong một số thời đoạn.
2. Việc sử dụng dữ liệu bảng có một vài ưu điểm. *Thứ nhất*, dữ liệu bảng làm tăng kích thước mẫu một cách đáng kể. *Thứ hai*, thông qua nghiên cứu các quan sát theo không gian lặp lại, dữ liệu bảng phù hợp hơn để nghiên cứu động học thay đổi. *Thứ ba*, dữ liệu bảng giúp ta nghiên cứu những mô hình hành vi phức tạp hơn.
3. Bất chấp các ưu điểm đáng kể, dữ liệu bảng đặt ra một số vấn đề ước lượng và suy luận. Vì dữ liệu này liên quan đến cả bình diện không gian và thời gian, nên những vấn đề cố hữu trong dữ liệu theo không gian (ví dụ như phương sai không đồng nhất) và dữ liệu theo chuỗi thời gian (ví dụ như tự tương quan) cần được giải quyết. Ngoài ra còn có thêm một số vấn đề, như tương quan chéo trong các đơn vị cá nhân trong cùng một thời đoạn.
4. Có một số kỹ thuật ước lượng để giải quyết một hay nhiều vấn đề này. Hai kỹ thuật nổi bật nhất là (1) mô hình các ảnh hưởng cố định (FEM) và (2) mô hình các ảnh hưởng ngẫu nhiên (REM) hay mô hình các thành phần sai số (ECM).
5. Trong mô hình FEM, tung độ gốc trong mô hình hồi qui được phép khác nhau giữa các cá nhân, khi thừa nhận sự kiện là mỗi cá nhân hay mỗi đơn vị có thể có những đặc điểm riêng nhất định. Để xem xét các tung độ gốc khác nhau, ta có thể sử dụng các biến giả. Mô hình FEM sử dụng biến giả được gọi là mô hình biến giả bình phương tối thiểu (LSDV). Mô hình FEM phù hợp trong những tình huống mà tung độ gốc của từng cá nhân có thể tương quan với một hay nhiều biến độc lập. Một nhược điểm của mô hình LSDV là nó làm mất đi nhiều bậc tự do khi số đơn vị  $N$  rất lớn, trong trường hợp đó ta sẽ phải đưa vào  $N$  biến giả (nhưng không chế số hạng tung độ gốc chung).
6. Một phương án khác thay cho mô hình FEM là mô hình ECM. Trong mô hình ECM, ta giả định rằng tung độ gốc của một đơn vị riêng lẻ được rút ngẫu nhiên từ một dân số lớn hơn nhiều với một trị trung bình không đổi. Tung độ gốc cá nhân khi đó được biểu thị như sự sai lệch so với trị trung bình không đổi này. Một ưu điểm của mô hình ECM so với mô hình FEM là: nó không làm mất bậc tự do vì ta không phải ước lượng  $N$  tung độ gốc riêng lẻ. Ta chỉ cần ước lượng trị trung bình của tung độ gốc và phương sai của nó. Mô hình ECM thích hợp trong những tình huống mà tung độ gốc (ngẫu nhiên) của từng đơn vị không tương quan với các biến độc lập.
7. Kiểm định Hausman có thể được sử dụng để quyết định chọn lựa giữa mô hình FEM và mô hình ECM.



8. Cho dù ngày càng được ưa chuộng trong nghiên cứu ứng dụng, và cho dù ngày càng sẵn có những dữ liệu này, hồi qui dữ liệu bảng không chắc phù hợp trong mọi tình huống. Ta phải vận dụng sự đánh giá thực tiễn trong từng trường hợp.

## Bài tập

### Câu hỏi

16.1. Các đặc điểm của (a) dữ liệu theo không gian, (b) dữ liệu theo chuỗi thời gian, và (c) dữ liệu bảng là gì?

16.2. Mô hình các ảnh hưởng cố định (FEM) nghĩa là gì? Vì dữ liệu bảng có cả bình diện không gian lẫn thời gian, mô hình FEM cho phép ta xem xét cả hai bình diện như thế nào?

16.3. Mô hình các thành phần sai số (ECM) nghĩa là gì? Mô hình này khác với mô hình FEM như thế nào? Khi nào mô hình ECM phù hợp? Và khi nào mô hình FEM phù hợp?

16.4. Có gì khác nhau giữa mô hình FEM, mô hình biến giả bình phương tối thiểu (LSDV), và mô hình đồng phương sai hay không?

16.5. Khi nào mô hình hồi qui số liệu bảng không phù hợp? Cho các ví dụ.

16.6. Bạn sẽ mở rộng mô hình (16.4.4) như thế nào để cho phép có một thành phần sai số theo thời gian? Trong trường hợp đó, điều gì sẽ xảy ra với công thức (16.3.6), (16.3.7) và (16.3.8)?

16.7. Tham khảo số liệu trúng và giá trúng đã cho trong Bảng 1.1. Mô hình nào có thể phù hợp ở đây, FEM hay ECM? Và tại sao?

16.8. Trong các kết quả hồi qui trong (16.3.4), các tung độ góc ảnh hưởng cố định của bốn công ty là bao nhiêu? Các ảnh hưởng này có khác nhau về mặt thống kê hay không?

16.9. Đối với ví dụ đầu tư thảo luận trong chương này, Bảng 16.3 cho ta các kết quả dựa vào mô hình ECM. Nếu bạn so sánh các kết quả này với những kết quả đã cho trong (16.3.4), bạn sẽ rút ra những kết luận tổng quát gì?

16.10. Dựa vào Nghiên cứu Động học Thu nhập Michigan, Hausman cố gắng ước lượng một mô hình tiền lương hay thu nhập, sử dụng mẫu gồm 629 học sinh tốt nghiệp phổ thông được theo dõi trong một khoảng thời gian 6 năm, vì thế ta có tổng cộng 3774 quan sát. Biến phụ thuộc trong nghiên cứu này là log tiền lương, và các biến giải thích là độ tuổi (được chia thành một số nhóm), thất nghiệp trong năm trước, sức khỏe kém trong năm trước, tự làm chủ, miền cư trú (miền nam = 1; hoặc bằng 0 ở những miền khác), khu vực cư trú (nông thôn = 1; hoặc bằng 0 ở khu vực khác). Hausman sử dụng cả hai mô hình FEM và ECM. Các kết quả được cho trong bảng 16.4 (sai số chuẩn trong dấu ngoặc đơn).

- Các kết quả có ý nghĩa kinh tế không?
- Có hay không sự khác biệt lớn giữa các kết quả đạt được bởi hai mô hình? Nếu có, điều gì có thể giải thích cho những khác biệt này?
- Trên cơ sở dữ liệu cho trong bảng, bạn sẽ chọn mô hình nào nếu được?

**Bảng 16.4 Các phương trình tiền lương (biến phụ thuộc: log tiền lương<sup>\*</sup>)**

Biến số	Ảnh hưởng cố định		Ảnh hưởng ngẫu nhiên	
1. Độ tuổi 1 (20-35)	0.0557	(0.0042)	0.0393	(0.0033)
2. Độ tuổi 2 (35-45)	0.0351	(0.0051)	0.0092	(0.0036)
3. Độ tuổi 3 (45-55)	0.0209	(0.0055)	-0.0007	(0.0042)
4. Độ tuổi 4 (55-65)	0.0209	(0.0078)	-0.0097	(0.0060)
5. Độ 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 làm chủ	-0.2190	(0.0297)	-0.2670	(0.0263)
9. Miền 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

**Bài toán**

16.11. Tham khảo dữ liệu trong Bảng 1.1.

- Gọi  $Y$  = số trứng sản xuất ra (triệu chiếc), và  $X$  = giá trứng (cents trên một chục trứng). Ước lượng mô hình:  $Y_i = \beta_1 + \beta_2 X_i + u_i$  cho năm 1990 và 1991 một cách riêng biệt.
- Kết hợp các quan sát trong hai năm và ước lượng hồi qui kết hợp. Bạn đưa ra những giả định gì khi kết hợp các dữ liệu?
- Sử dụng mô hình các ảnh hưởng cố định, phân biệt hai năm, và trình bày các kết quả hồi qui.
- Bạn có thể sử dụng mô hình các ảnh hưởng cố định, phân biệt 50 tiểu bang hay không? Tại sao có hoặc tại sao không?
- Việc phân biệt vừa ảnh hưởng tiểu bang vừa ảnh hưởng năm liệu có ý nghĩa hay không? Nếu có, bạn sẽ phải đưa ra bao nhiêu biến giả?
- Mô hình các thành phần sai số có thích hợp để lập mô hình việc sản xuất trứng hay không? Tại sao có hoặc tại sao không? Hãy xem liệu bạn có thể ước lượng mô hình như vậy thông qua sử dụng phần mềm Eviews chẳng hạn hay không.

16.12. Tiếp tục với bài tập 16.11. Trước khi quyết định chạy hồi qui kết hợp, bạn muốn tìm hiểu xem thử dữ liệu ‘có thể kết hợp được’ hay không. Vì mục đích này, bạn quyết định sử dụng kiểm định Chow như đã thảo luận trong chương 8. Trình bày những phép tính cần thiết liên quan và xác định xem liệu hồi qui kết hợp có ý nghĩa hay không.

16.3. Quay lại với hàm đầu tư Grunfeld thảo luận trong phần 16.2.

<sup>\*</sup> 3774 quan sát; sai số chuẩn trong dấu ngoặc đơn. Được trình bày lại từ nghiên cứu của Chen Hsiao, *Analysis of Panel Data*, Cambridge University Press, 1986, trang 42. Nguồn ban đầu: J. A. Hausman, ‘Specification Tests in Econometrics,’ *Econometrica*, tập 46, 1978, trang 1251-1271.

- a. Ước lượng hàm đầu tư Grunfeld cho GE, GM, US Steel và Westinghouse một cách riêng biệt. Các kết quả của việc kết hợp toàn bộ 80 quan sát đã được cho trong (16.3.1).
- b. Để xác định xem liệu hồi qui kết hợp (16.3.1) có thích hợp hay không, bạn quyết định thực hiện kiểm định Chow như đã thảo luận trong chương 8. Thực hiện kiểm định này. *Gợi ý:* Lấy RSS từ hồi qui kết hợp, lấy RSS từ từng hàm đầu tư trong bốn hàm đầu tư, rồi áp dụng kiểm định Chow.
- c. Từ kiểm định Chow, bạn rút ra kết luận gì? Nếu kết luận của bạn là không kết hợp dữ liệu, vậy bạn có thể nói gì về việc sử dụng các kỹ thuật hồi qui dữ liệu bảng?

16.14. Bảng 16.5 trình bày dữ liệu về tỷ lệ thất nghiệp dân sự Y (phần trăm) và tiền lương theo giờ trong công nghiệp chế tạo X, tính bằng USD (chỉ số, 1992 = 100) của Canada, Anh, và Hoa Kỳ trong giai đoạn 1980-1999. Hãy xem mô hình:

$$Y_{it} = \beta_1 + \beta_2 X_{it} + u_{it} \tag{1}$$

- a. Một cách tiên nghiệm, mối quan hệ dự kiến giữa Y và X là gì? Tại sao?
- b. Ước lượng mô hình đã cho trong (1) cho từng nước.
- c. Ước lượng mô hình, kết hợp toàn bộ 60 quan sát.
- d. Ước lượng mô hình các ảnh hưởng cố định (FEM).
- e. Ước lượng mô hình các thành phần sai số (ECM).
- f. Mô hình nào tốt hơn, FEM hay ECM? Giải thích câu trả lời của bạn.

**Bảng 16.5 Tỷ lệ thất nghiệp và tiền lương theo giờ trong công nghiệp chế tạo ở Hoa Kỳ, Canada và Anh, 1980-1999**

Quan sát	Hoa Kỳ		Canada		Anh	
	Tiền lương USD/giờ	Thất nghiệp, %	Tiền lương USD/giờ	Thất nghiệp, %	Tiền lương USD/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	10.0	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

Tiền lương theo giờ tính bằng USD, chỉ số 1992 = 100.

*Nguồn:* Báo cáo kinh tế của tổng thống, tháng 1-2001, bảng B109, trang 399.