Dữ liệu bảng
(Panel Data)

Đinh Công Khải

Tháng 5/2016

Nội dung

1. Giới thiệu chung về dữ liệu bảng
2. Những lợi thế khi sử dụng dữ liệu bảng
3. Ước lượng mô hình hồi qui dữ liệu bảng
   - Mô hình những ảnh hưởng cố định (FEM)
   - Mô hình những ảnh hưởng ngẫu nhiên (REM)
4. Các kiểm định phương sai thay đổi và tương quan chuỗi trong dữ liệu bảng.

Giới thiệu chung về dữ liệu bảng

- Thế nào là dữ liệu bảng?
  - Dữ liệu bảng là dữ liệu có 2 chiều: chiều không gian và chiều thời gian.
  - Là sự mở rộng dữ liệu chéo (cross section) theo thời gian (time series).
  - Là dữ liệu chéo theo chuỗi thời gian (cross sectional time-series data).

Bảng cân đối (Balanced panel)

Tỉnh	Năm	GDP	Dân số
1	2005		
1	2006		
1	2007		
2	2005		
2	2006		
2	2007		
……	……..		
…….	…….		
63	2005		
63	2006		
63	2007		

Bảng không cân đối (Unbalanced panel)

Tỉnh	Năm	GDP	Dân số
1	2005		
1	2006		
1	2007		
2	2005		
2	2006		
……	……		
10	2007		
…….	…….		
63	2005		
63	2006		
63	2007		

Những lợi thế của việc sử dụng dữ liệu bảng

- Dữ liệu bảng cung cấp nhiều thông tin hơn, biến thiên hơn, ít có sự đa cộng tuyến giữa các biến số, bậc tự do cao hơn, và hiệu quả hơn.
- Bằng cách nghiên cứu các dữ liệu chéo một cách lặp đi lặp lại qua thời gian, dữ liệu bảng thực hiện tốt hơn các nghiên cứu về những thay đổi xảy ra liên tục như tỷ lệ thất nghiệp, di chuyển lao động.

Những lợi thế của việc sử dụng số liệu bảng

- Cho phép kiểm soát sự khác biệt không quan sát được giữa các thực thể (entities), ví dụ như khác biệt văn hoá giữa các quốc gia hay sự khác biệt về triết lý kinh doanh giữa các công ty.
- Cho phép kiểm soát các biến không quan sát được nhưng thay đổi theo thời gian (chính sách quốc gia, thỏa thuận quốc tế, hành vi của con người).
- Cho phép nghiên cứu các mô hình phức tạp, ví dụ như tính kinh tế do quy mô hay thay đổi công nghệ.

Dữ liệu bảng

TABLE 16.1 INVESTMENT DATA FOR FOUR COMPANIES, 1935–1954

Observation	I	F_{-1}	C_{-1}	Observation	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

Ước lượng các mô hình hồi qui dữ liệu bảng: Phương pháp những ảnh hưởng cố định

- Mô hình ước lượng

$$Y_{it} = \beta_1 + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it} \quad (1)$$

i= 1, 2, 3, 4 và t = 1, 2, .., 20

trong đó

Y_{it} = tổng đầu tư thực của công ty i tại thời điểm t

X_{2it} = giá trị thực của công ty i tại thời điểm t

X_{3it} = trữ lượng vốn của công ty i tại thời điểm t

u_{it} = nhiễu trắng

Ước lượng các mô hình hồi qui dữ liệu bảng (tt)

- Xem xét việc ước lượng (1) trong 5 trường hợp sau đây:

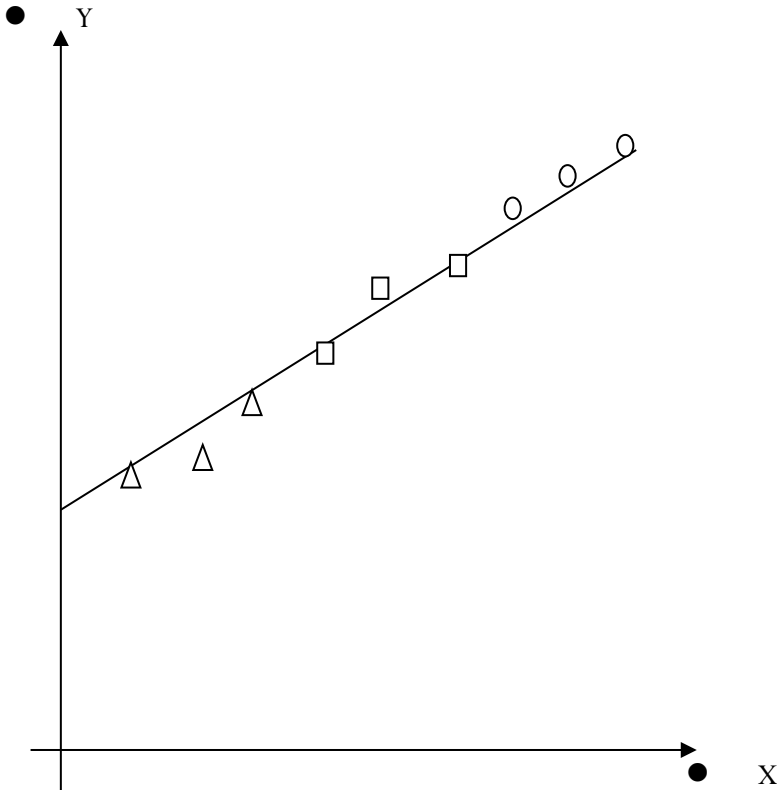
1. Tung độ gốc và hệ số góc giống nhau giữa các công ty và qua thời gian (phần dư thể hiện sự khác biệt giữa các công ty và qua thời gian).
2. Tung độ gốc khác nhau giữa các cty, hệ số góc là hằng số
3. Tung độ gốc khác nhau giữa các công ty và qua thời gian, hệ số góc là hằng số.

Ước lượng các mô hình hồi qui dữ liệu bảng (tt)

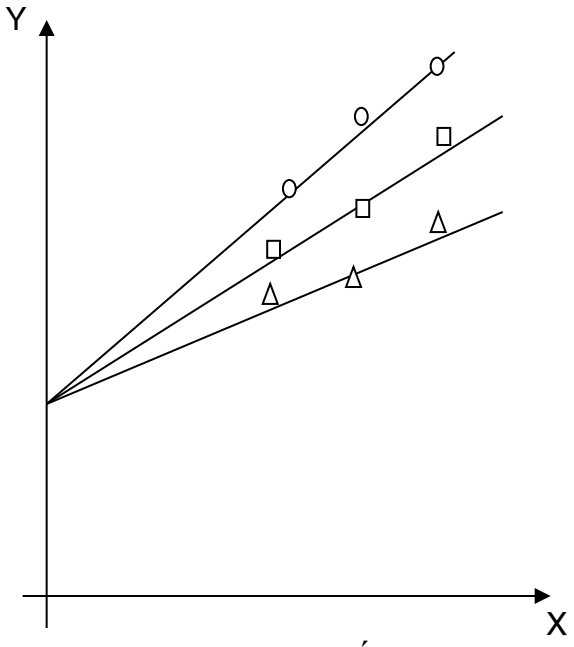
4. Tung độ gốc và hệ số góc thay đổi giữa các công ty.

5. Tung độ gốc và hệ số góc thay đổi giữa các công ty và qua thời gian.

Ước lượng các mô hình hồi qui dữ liệu bảng (tt)
Nguồn: Cao Hào Thi

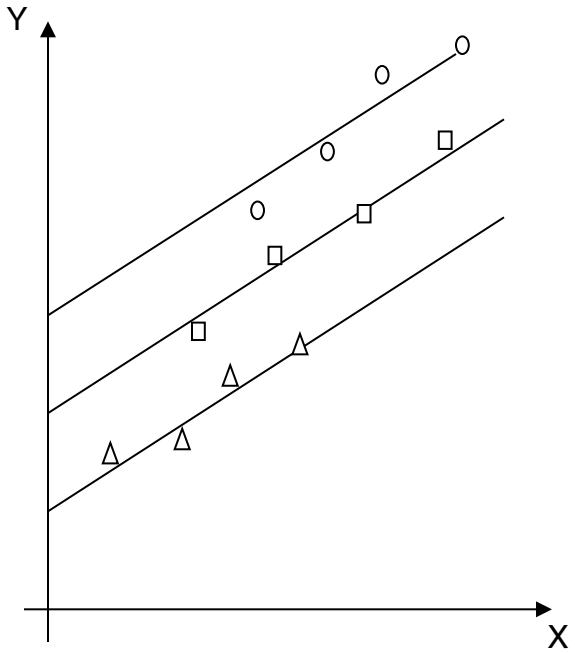


cùng tung độ gốc,
cùng hệ số góc

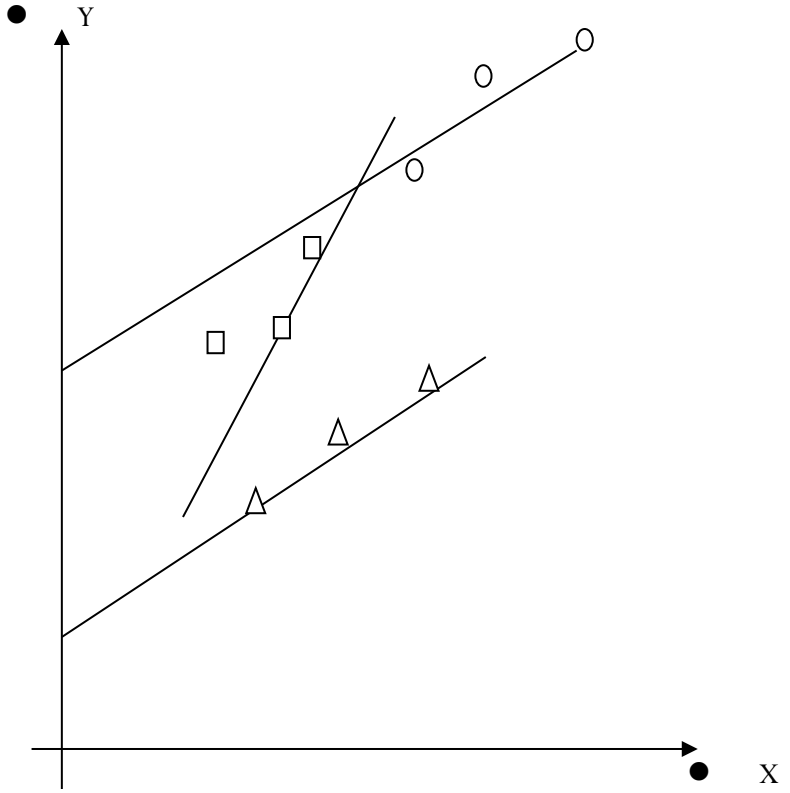


cùng tung độc gốc,
khác nhau về hệ số góc

Ước lượng các mô hình hồi qui dữ liệu bảng (tt)
Nguồn: Cao Hào Thi



Khác tung độ gốc
Cùng hệ số góc



Khác tung độ gốc
Khác hệ số góc

Ước lượng các mô hình hồi qui dữ liệu bảng (tt)

- TH 1: Tung độ gốc không đổi và hệ số góc không đổi (Pooled Regression)

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

- TH 2: Tung độ gốc thay đổi theo i và hệ số góc không đổi

Mô hình những các ảnh hưởng cố định (fixed effects) hay mô hình bình phương tối thiểu các biến giả (LSDV)

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

Mô hình những ảnh hưởng cố định (fixed effects) hay mô hình bình phương tối thiểu các biến giả (LSDV)

- Mỗi thực thể đều có những đặc điểm riêng biệt, có thể ảnh hưởng đến các biến giải thích.

  Ví dụ: Cách thức kinh doanh của một công ty có thể ảnh hưởng đến giá trị của công ty hay trữ lượng vốn của nó.

- Giả thiết rằng có sự tương quan giữa phần dư của mỗi thực thể (có chứa các đặc điểm riêng) với các biến giải thích.

Mô hình những ảnh hưởng cố định (fixed effects) hay mô hình bình phương tối thiểu các biến giả (LSDV)

- FE có thể kiểm soát và tách ảnh hưởng của các đặc điểm riêng biệt (không đổi theo thời gian) này ra khỏi các biến giải thích để chúng ta có thể ước lượng những ảnh hưởng thực (net effects) của biến giải thích lên biến phụ thuộc.
- Các đặc điểm riêng biệt (không đổi theo thời gian) này là đơn nhất đối với 1 thực thể và không tương quan với đặc điểm của các thực thể khác.

Ước lượng các mô hình hồi qui dữ liệu bảng (tt)

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

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

D_{1i} = 1 nếu quan sát thuộc GE; bằng 0 nếu không thuộc GE

D_{2i} = 1 nếu quan sát thuộc GM; bằng 0 nếu không thuộc GM

D_{3i} = 1 nếu quan sát thuộc US; bằng 0 nếu không thuộc US

D_{4i} = 1 nếu quan sát thuộc WEST, bằng 0 nếu không thuộc WEST

Phân tích dữ liệu bảng (tt)

- TH 3: Tung độ gốc thay đổi theo t và hệ số góc không đổi

  (sự thay đổi về công nghệ, chính sách của chính phủ, thuế)

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

$$Y_{it} = \alpha_1 + \alpha_2 t_{35} + \alpha_3 t_{36} + .. + \alpha_4 t_{54} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

t_{35} = 1 nếu quan sát ở năm 1935; bằng 0 nếu không phải

t_{36} = 1 nếu quan sát ở năm 1936; bằng 0 nếu không phải

..

t_{54} = 1 nếu quan sát ở năm 1954; bằng 0 nếu không phải .

Phân tích dữ liệu bảng (tt)

- TH 4: Tung độ gốc thay đổi theo i và t và hệ số góc không đổi

$$Y_{it} = \alpha_1 + \alpha_2 D_{2i} + \alpha_3 D_{3i} + \alpha_4 D_{4i} + \lambda_1 t_{35} + .. + \lambda_{19} t_{53} + \beta_2 X_{2it} + \beta_3 X_{3it} + u_{it}$$

- TH 5: Tung độ thay đổi và hệ số góc thay đổi

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

Những hạn chế của FEM hay LSDV

- Có quá nhiều biến được tạo ra trong mô hình, do đó có khả năng làm giảm bậc tự do và làm tăng khả năng sự đa cộng tuyến của mô hình.
- FEM không đo lường được tác nhân không thay đổi theo thời gian như giới tính, màu da, hay chủng tộc.
- Số hạng sai số uit vẫn có thể có các vấn đề về phương sai thay đổi, tương quan chuỗi hoặc tự tương quan.

Mô hình những tác động ngẫu nhiên (random effects model)

- Đặc điểm riêng giữa các thực thể được giả sử là ngẫu nhiên và không tương quan đến các biến giải thích thì chúng ta dùng REM.
- REM xem các phần dư của mỗi thực thể (không tương quan với biến giải thích) là một biến giải thích mới.
- Số hạng sai số vẫn có thể bị hiện tượng phương sai thay đổi hay tự tương quan

Mô hình những tác động ngẫu nhiên (random effects model)

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

$$\beta_{1i} = \beta_1 + \varepsilon_i$$

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

$$w_{it} = \varepsilon_i + u_{it}$$

$$E(w_{it}) = 0$$

$$\text{var}(w_{it}) = \sigma_\varepsilon^2 + \sigma_u^2; \qquad \text{cov}(w_{it}, w_{is}) = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_u^2}$$

$$\varepsilon_i \sim N(0, \sigma_\varepsilon^2)$$

$$u_{it} \sim N(0, \sigma_u^2)$$

$$E(\varepsilon_i u_{it}) = 0$$

$$E(\varepsilon_i \varepsilon_j) = 0$$

Mô hình FEM (LSDV) hay REM

- Nếu ε_i và Xs không có tương quan, sử dụng REM
- Nếu ε_i và Xs có tương quan, sử dụng FEM
- Nếu T lớn, N nhỏ, 2 phương pháp giống nhau
- Nếu N lớn, T nhỏ, kết quả ước lượng của 2 phương pháp khá khác nhau ➔ FEM phù hợp nếu các đơn vị **KHÔNG** được rút ra ngẫu nhiên từ mẫu lớn.
- Nếu N lớn, T nhỏ, các điều kiện trong REM được thỏa, ước lượng của REM có hiệu quả hơn FEM

Mô hình FEM (LSDV) hay REM

- Hausman test
  - Ho: Ước lượng của FEM và REM không khác nhau
  - p-value < 0.05, bác bỏ Ho
  - Nếu bác bỏ H0, REM không hợp lý, nên sử dụng FEM
- Breusch-Pagan Lagrange Multiplier cho REM

  Stata: xttest0

  Ho: Phương sai qua các thực thể là không đổi

  p-value < 0.05, bác bỏ Ho

Các test khác

- Phương sai thay đổi trong FEM

  Stata: xttest3

  Ho: Phương sai không thay đổi

  p-value < 0.05, bác bỏ Ho

➔ sử dụng robust trong lệnh reg hoặc xtreg

Các test khác

- Tự tương quan (kiểm định nhân tử Lagrange)

  Stata: xtserial y x

  Ho: Không có tương quan chuỗi

  p-value < 0.05, bác bỏ Ho (có tương quan chuỗi)

➔ Sử dụng cluster() trong lệnh reg; riêng lệnh xtreg đã có sử dụng cluster() trong lệnh này.