

Trường Chính sách Công và Quản lý Fulbright

Kinh tế lượng nâng cao

Bài tập 1

Ngày Phát: Thứ ba 14/2/2023 - Ngày Nộp: 8h20 Thứ ba 28/2/2023

Giới thiệu

Trong bài tập này học viên sẽ sử dụng dữ liệu điều tra doanh nghiệp để ước lượng hàm sản xuất. Học viên được cung cấp bộ dữ liệu đã xử lý thô cho hai năm 2015-2016, bao gồm:

Tên biến	Giải thích
province	mã tỉnh
district	mã huyện
commune	mã xã
taxid	mã số thuế
lhdn	loại hình doanh nghiệp
industry	ngành kinh tế
year	năm điều tra
revenue	doanh thu kinh doanh trong năm
capital	giá trị tài sản cố định ròng theo sổ sách
labor	là chi phí lao động tại thời điểm cuối năm, bao gồm cả lao động thời vụ
energy	tổng chi phí cho năng lượng
material	tổng chi phí cho nguyên vật liệu trung gian

Học viên được cấp bộ dữ liệu điều tra doanh nghiệp 2016 và bảng câu hỏi để có thể lựa chọn bổ sung thêm các biến hay thông tin nếu cần thiết¹.

Mô hình

Mô hình 4 nhân tố đầu vào (KLEM) của hàm sản xuất giả định rằng giá trị sản phẩm được tạo ra từ bốn yếu tố chính gồm: vốn (K), lao động (L), năng lượng (E), và nguyên vật liệu trung gian (M). Lý thuyết về sản xuất của doanh nghiệp thường sử dụng hàm Cobb-Douglas để mô

¹Dữ liệu điều tra kinh tế có đặc thù trễ một năm, do đó điều tra 2016 tương ứng với dữ liệu hoạt động trong năm 2015

tả quá trình chuyển đổi từ nguyên liệu đầu vào và vốn sản xuất thành sản phẩm đầu ra như sau:

$$Y_{it} = AK_{it}^{\beta_1} L_{it}^{\beta_2} E_{it}^{\beta_3} M_{it}^{\beta_4} \quad (1)$$

Lấy logarithm hai vế, và thêm các đặc tính ngẫu nhiên của hàm sản xuất thông qua phần dư u , chúng ta có phương trình sau:

$$\ln Y_{it} = \beta_0 + \beta_1 \ln K_{it} + \beta_2 \ln L_{it} + \beta_3 \ln E_{it} + \beta_4 \ln M_{it} + u_{it} \quad (2)$$

Với $\beta_0 = \ln A$, $\beta_1, \beta_2, \beta_3, \beta_4$ là các tham số cần phải ước lượng trong mô hình.

Câu hỏi

1. Xử lý và thống kê mô tả bộ dữ liệu để phục vụ cho phân tích. Hãy giải thích rõ cơ sở lý thuyết và việc lựa chọn các biến trong mô hình. So sánh các nhóm doanh nghiệp theo các đặc tính sau đây:
 - a. Thống kê so sánh dữ liệu giữa các ngành kinh tế và loại hình sở hữu của doanh nghiệp.
 - b. Thống kê so sánh giữa các doanh nghiệp có quy mô khác nhau.
 - c. Thống kê so sánh giữa các doanh nghiệp ở các khu vực địa lý khác nhau (Bắc/Trung/Nam, và/hoặc khu vực Hà Nội/Hồ Chí Minh).

Học viên tự thực hiện. Học viên có thể nhận thấy có phân phối lệch đối với biến doanh thu, hàm ý rằng có một số doanh nghiệp rất lớn, nhưng đại đa số là các doanh nghiệp vừa, nhỏ và rất nhỏ trong mẫu. Tuy nhiên, nếu phương trình hồi quy sử dụng logarithm của các biến thì việc có phân phối lệch không ảnh hưởng lớn đến kết quả.

2. Nêu rõ các giả định cần có, sau đó ước lượng mô hình (2) bằng hồi quy đa biến với dữ liệu từng năm và dữ liệu gộp, trình bày kết quả thành bảng biểu rõ ràng và diễn giải ngắn gọn kết quả. Kiểm định liệu hàm sản xuất thể hiện đặc tính lợi suất không đổi theo quy mô (constant return to scale).

Đối với hồi quy sử dụng dữ liệu gộp thì chúng ta cần đảm bảo các giả định từ (1-4) đã học trong học kỳ trước để ước lượng không chệch và nhất quán. Sau đó, nếu áp thêm các điều kiện 5 (phương sai không đổi) và 6 (phân phối chuẩn của phần dư) thì chúng ta có thêm các thuộc tính ước lượng hiệu quả nhất và có thể kiểm định với mô hình CLRM.

Regression Results

	Pooled OLS b/se	Robust STD b/se	2015 b/se	2016 b/se
lCapital	0.0525*** (0.0019)	0.0525*** (0.0076)	0.0724*** (0.0106)	0.0312*** (0.0065)
lLabor	0.4756*** (0.0030)	0.4756*** (0.0158)	0.4725*** (0.0194)	0.4767*** (0.0160)
lEnergy	0.1007*** (0.0021)	0.1007*** (0.0070)	0.0883*** (0.0075)	0.1142*** (0.0128)
lMaterial	0.4685*** (0.0025)	0.4685*** (0.0163)	0.4611*** (0.0196)	0.4775*** (0.0168)
year	-0.0856*** (0.0070)	-0.0856* (0.0366)		
Constant	173.1856*** (14.1862)	173.1856* (73.7644)	0.7385*** (0.1299)	0.6737*** (0.1485)
Obs	115067.0000	115067.0000	55738.0000	59329.0000
R2	0.7458	0.7458	0.7155	0.7701
R2-adj	0.7458	0.7458	0.7155	0.7701
df (r)	115061.0000	62.0000	62.0000	62.0000
SSR	156210.2746	156210.2746	80414.1945	75624.0856

* p<0.05, ** p<0.01, *** p<0.001

Trong các điều kiện trên thì điều kiện 4.2 (biến giải thích không tương quan với phần dư) là điều kiện khó đảm bảo nhất bởi một trong các nhân tố đầu vào, chẳng hạn như quy mô đầu tư vốn, có thể tương quan với nhân tố quan trọng không quan sát được (ví dụ đặc tính của chủ doanh nghiệp như năng khiếu kinh doanh). Do đó chúng ta có tình huống thiếu biến trong trọng không quan sát được và ước lượng sẽ bị chệch do thiếu biến (omitted variables bias - OVB). Nhà nghiên cứu phải đánh giá được hướng chệch của ước lượng. Trong ví dụ ở đây thì do khiếu kinh doanh tương quan dương với cả doanh thu lẫn quy mô vốn nên ước lượng tham số của vốn sẽ bị chệch lên.

So sánh các ước lượng cho thấy các ước lượng khá tương đồng giữa các năm. Do chúng ta sử dụng cấu trúc hàm log-log nên các tham số sẽ được diễn giải là độ do dẫn của doanh thu theo các nhân tố đầu vào hoặc tỷ phần chi phí của các yếu tố đầu vào. Tỷ phần chi phí cho lao động là lớn nhất, tiếp theo đó là vật liệu trung gian, năng lượng, và vốn. Kết quả này cũng khá phù hợp với thực trạng sản xuất kinh doanh của các doanh nghiệp ở Việt Nam là chi phí lớn nhất là cho lao động và nguyên vật liệu sản xuất, trong khi chi phí cho vốn rất ít. Chi phí cho năng lượng chiếm khoảng 8-10% tổng chi phí, khá phù hợp với các báo cáo đã được công bố.

Chi phí vốn thấp cho thấy thực trạng là doanh nghiệp Việt Nam đầu tư ít cho tài sản cố định, trong khi thâm dụng lao động, nguyên liệu hay năng lượng. Nếu kết quả này là phù hợp thì định hướng chính sách để giảm sử dụng năng lượng là cần tạo các điều kiện cho doanh nghiệp đầu tư máy móc thiết bị sản xuất tiên tiến để giảm thâm dụng năng lượng.

3. Ước lượng mô hình bằng hồi quy dữ liệu bảng với tất cả các phương pháp có thể thực hiện được (fixed effects, random effects, lsdv, difference in difference). Kiểm tra và đối chiếu kết quả với ước lượng trong câu (2) và giải thích kỹ nếu có sự khác biệt.

Kiểm tra thấy có 22,318 doanh nghiệp có quan sát cả ở hai kỳ, do đó bảng dữ liệu là bảng không cân bằng và khi ước lượng thì kết quả có thể khác biệt giữa các phương pháp do vấn đề dữ liệu bị thiếu. Học viên cần kiểm tra liệu doanh nghiệp có trong cả hai kỳ có sự khác biệt mang tính hệ thống với các doanh nghiệp chỉ có dữ liệu một kỳ hay không.

Khi thực hiện các ước lượng, lưu ý với ước lượng bằng sai phân bậc nhất. Do chỉ có 22,318 doanh nghiệp có dữ liệu lặp nên DiD chỉ có 22,318 quan sát dựa trên dữ liệu sai phân. Nếu học viên xử lý không kỹ sẽ sử dụng sai dữ liệu.

Regression Results

	FE b/se	RE b/se	LSDV b/se	DID b/se
lCapital	0.0520*** (0.0047)	0.0687*** (0.0019)	0.0520*** (0.0096)	
lLabor	0.4369*** (0.0063)	0.4922*** (0.0029)	0.4369*** (0.0198)	
lEnergy	0.0346*** (0.0026)	0.0742*** (0.0018)	0.0346*** (0.0046)	
lMaterial	0.1803*** (0.0080)	0.4479*** (0.0026)	0.1803*** (0.0217)	
year	0.0162** (0.0054)	-0.0553*** (0.0045)	0.0162 (0.0090)	
DlK				0.0520*** (0.0072)
DlL				0.4369*** (0.0171)
DlE				0.0346*** (0.0041)
DlM				0.1803*** (0.0181)
Constant	-28.8916** (10.8840)	112.1477*** (9.0918)	-28.8916 (17.9803)	0.0162 (0.0085)
Obs	115067.0000	115067.0000	115067.0000	22295.0000
R2	0.2437		0.9904	0.2317
R2-adj	-2.9044		0.9505	0.2316
dF(r)	22290.0000		22290.0000	62.0000
SSR	5890.3306		5890.3306	11780.6611

* p<0.05, ** p<0.01, *** p<0.001

So sánh kết quả cho thấy kết quả khá tương đồng giữa FE, LSDV và DID. Tuy nhiên, khác biệt ở mức ý nghĩa của các tham số (thể hiện qua sai số chuẩn). FE có sai số chuẩn nhỏ nhất (do đó khoảng tin cậy của tham số nhỏ nhất và ước lượng chính xác nhất). LSDV có R2 rất cao bởi đưa vào (n-1) biến giả. Trong khi đó, DID có R2 thấp nhất do ước lượng này chỉ sử dụng các dữ liệu sai phân của 22,318 quan sát.

Ước lượng RE thì khá tương đồng với ước lượng bằng hồi quy gộp trong phần 2.

Đánh giá phương sai cho thấy tỷ phần phương sai của nhân tố không quan sát được a_i chiếm 81% trong tổng phương sai của phần dư gộp (gồm cả a_i và nhiều u_i). Đây là một tỷ lệ khá lớn. Đồng thời kiểm định Hausman bác bỏ giả thuyết Ho (ước lượng FE và RE giống nhau). Do đó, chúng ta sẽ lựa chọn ước lượng FE bởi nhân tố cố định quan trọng trong mô hình và sử dụng ước lượng FE sẽ có hiệu lực nội tại.

4. Ước lượng mô hình mở rộng bằng cách lần lượt thêm các biến kiểm soát đại diện cho ngành kinh tế (D_1), vùng miền (D_2), và loại hình sở hữu (D_3):

$$\ln Y_{it} = \beta_0 + \beta_1 \ln K_{it} + \beta_2 \ln L_{it} + \beta_3 \ln E_{it} + \beta_4 \ln M_{it} + \gamma_1 D_1 + \gamma_2 D_2 + \gamma_3 D_3 + u_{it} \quad (3)$$

Dựa trên các kết quả phía trên thì chúng ta phải dùng ước lượng FE. Tuy nhiên, do các nhân tố như ngành nghề kinh tế, sở hữu, vùng miền không hoặc rất ít biến động giữa các năm điều tra nên ước lượng bằng FE truyền thống sẽ không thực hiện được. Do đó, cách làm phù hợp nhất là dùng DiD với điều kiện ban đầu. Tuy nhiên, do các nhân tố này không hoặc ít thay đổi giữa các năm nên chúng ta có thể thực hiện khá dễ bằng cách mở rộng ước lượng DiD và đưa thêm thông tin trong năm 2016.

Học viên lưu ý với các biến nhóm (categorical variable) không xếp hạng được thì phải đưa vào mô hình dưới dạng biến nhóm nhị phân và so với nhóm tham chiếu. Mặc dù chúng ta ít khi quan tâm đến các tham số ước lượng được của các biến này, chúng có thể giúp đánh giá sự khác biệt nhóm giữa các thành phần kinh tế, khu vực, loại hình doanh nghiệp...

Kết quả ước lượng cho thấy không khác biệt nhiều so với bảng kết quả trong câu 3. Do đó chúng ta có thể kết luận là kết quả ước lượng vững.

Regression Results		
	FE In.. b/se	DiD I.. b/se
lCapital	0.0528*** (0.0060)	
lLabor	0.4357*** (0.0123)	
lEnergy	0.0351*** (0.0029)	
lMaterial	0.1789*** (0.0134)	
year	0.0166**	
DlK		0.0533*** (0.0066)
DlL		0.4303*** (0.0166)
DlE		0.0452*** (0.0042)
DlM		0.1772*** (0.0179)
Constant	-29.5417** (11.1255)	0.0353 (0.0665)
Obs	115067.0000	22295.0000
R2	0.2479	0.2395
R2-adj	0.2477	0.2383
df (r)	92771.0000	62.0000
SSR	5857.2239	11661.4319

5. Thực hiện các kiểm định cần thiết, lựa chọn mô hình tối ưu, và giải thích kết quả. So sánh kết quả thu được với một số nghiên cứu đã thực hiện trên thế giới và rút ra hàm ý chính sách từ bài tập.

Đại đa số khi có dữ liệu bảng thì ước lượng tối ưu nhất là FE estimator. Ước lượng này giúp xử lý vấn đề vi phạm giả định 4.2 và đảm bảo hiệu lực nội tại. So sánh giữa các ước lượng pooled OLS, RE với các hình thức ước lượng FE (thông qua hồi quy dữ liệu bảng xtreg fe, lsdv, did, và thêm các điều kiện ban đầu), nếu có khác biệt thì FE đáng tin cậy hơn.

Tuy nhiên, học viên cần lưu ý về mặt dữ liệu. Bởi trong ví dụ này bảng cân bằng chỉ có khoảng 22,000 doanh nghiệp, do đó ước lượng FE chỉ có hiệu lực nội tại trong nhóm doanh nghiệp có dữ liệu lặp. Với các doanh nghiệp chỉ có một kỳ dữ liệu, chúng ta phải sử dụng pooled OLS và kết quả cho thấy có sự khác biệt đáng kể trong kết quả câu 2 và câu 3-4. Nếu doanh nghiệp gặp vấn đề thiếu dữ liệu có sự khác biệt mang tính hệ thống với doanh nghiệp có đủ dữ liệu thì rõ ràng việc áp dụng FE chỉ có hiệu lực rất hạn chế trong trường hợp này.

Lưu ý với bài tập nộp:

Học viên nộp bài viết, chất lọc các thông tin quan trọng và trình bày kết quả theo yêu cầu của đề bài, tránh in tất cả dữ liệu đầu ra từ phần mềm để nộp bài. Học viên cần gửi kèm do-file và log-file (bản điện tử) để trợ giảng đối chiếu chấm điểm.

Tài liệu tham khảo

Le, PV 2019. Energy Demand and Factor Substitution in Vietnam: Evidence from Two Recent Enterprise Surveys. *Journal of Economic Structures*, 8(35).

<https://journalofeconomicstructures.springeropen.com/articles/10.1186/s40008-019-0168-9>