

Nhập môn Kinh tế lượng (Introduction to Econometrics)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

Ngày 23 tháng 11 năm 2017

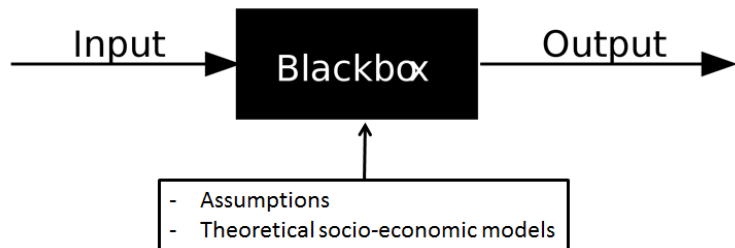
Học kinh tế lượng để làm gì?

Để xác lập và lượng hóa các mối quan hệ kinh tế-xã hội-môi trường sử dụng trong nghiên cứu kinh tế và phân tích chính sách.

- ▶ Tác động của việc đi học đến thu nhập như thế nào?
- ▶ Chương trình xóa đói giảm nghèo có giúp tăng thu nhập của người dân không?
- ▶ Tham nhũng có thực sự cản trở doanh nghiệp phát triển hay không?
- ▶ Biến đổi khí hậu có ảnh hưởng như thế nào đến năng suất mùa màng?
- ▶ Nhân tố nào ảnh hưởng đến hành vi sử dụng phương tiện đi lại (xe buýt, xe máy, xe đạp, đi bộ) của người dân ở các thành phố lớn?
- ▶ Tăng thuế xăng dầu từ 3000 đồng lên 8000 đồng/lít ảnh hưởng như thế nào đến nhu cầu đi lại của người dân?

Mục đích của môn học

- ▶ Hiểu bản chất của các mô hình kinh tế lượng căn bản.
- ▶ Sử dụng Stata để tiến hành các phân tích định lượng.
- ▶ Diễn giải, phân tích, và phê phán các kết quả nghiên cứu thực nghiệm.



Giáo trình, phần mềm, tài liệu tham khảo

- ▶ Sách giáo trình: Introductory Econometrics: A Modern Approach của Jeffrey Wooldridge, có bản dịch tiếng Việt phiên bản 5th và sách gốc tiếng Anh.
- ▶ Tham khảo: A Guide to Modern Econometrics, 4nd edition, của Marno Verbeek.
- ▶ Phần mềm thống kê Stata, phiên bản 11 hoặc cao hơn.
- ▶ Tham khảo cách mô phỏng lại các ví dụ tại:
<http://fmwww.bc.edu/gstat/examples/wooldridge/wooldridge.html>

Yêu cầu của môn học

- ▶ 7 bài tập về nhà (50%), gồm các bài tập lý thuyết và thực hành trên máy tính.
- ▶ Hoàn thiện đề án môn học theo nhóm (50%).
- ▶ Không thi cuối kỳ.

Nội dung môn học (cập nhật mới nhất)

- ▶ Bài 1: Nhập môn kinh tế lượng (JW Ch1)
- ▶ Bài 2: Hồi quy đơn biến (JW Ch2)
- ▶ Bài 3: Hồi quy đa biến (JW Ch3)
- ▶ Bài 4-5: Hướng dẫn sử dụng Stata và khai thác các bộ dữ liệu kinh tế xã hội.
- ▶ Bài 6: Giả thuyết và kiểm định giả thuyết (JW Ch4).
- ▶ Bài 7: Cấu trúc hàm và lựa chọn mô hình (JW Ch6).
- ▶ Bài 8: Hồi quy với biến định tính (JW Ch7).
- ▶ Bài 9: Phương sai thay đổi và tự tương quan (JW Ch8 và MV Ch4).
- ▶ Bài 10: Chuẩn đoán và xử lý các vấn đề liên quan đến dạng hàm số và dữ liệu (JW Ch9).
- ▶ Bài 11-12: Mô hình với biến phụ thuộc bị giới hạn (JW Ch17).

Phân tích định lượng (phân tích thực nghiệm) bao gồm những gì?

- ▶ Dữ liệu: loại dữ liệu gì, đặc tính của dữ liệu
- ▶ Mô hình lý thuyết: xây dựng mối quan hệ giữa các biến số
- ▶ Phương pháp ước lượng: mô hình thống kê để thiết lập quan hệ giữa các biến số
- ▶ Diễn giải và kiểm định: giải thích kết quả
- ▶ Chuẩn đoán và xử lý các vấn đề có liên quan: kiểm tra tính vững của kết quả

Hình thức thu thập dữ liệu

- ▶ Dữ liệu thử nghiệm/thí nghiệm (experimental data): thu được trong nghiên cứu khoa học cơ bản dựa trên các điều kiện có kiểm soát chặt chẽ trong phòng thí nghiệm.
- ▶ Dữ liệu phi thử nghiệm: là các dữ liệu quan sát/điều tra (observational/surveyed data). Hầu hết các dữ liệu kinh tế xã hội đều là dữ liệu phi thử nghiệm.

Hình thức thu thập dữ liệu sẽ quyết định liệu mô hình có xác lập được quan hệ nhân quả hay không. Xác lập quan hệ nhân quả phức tạp hơn nhiều so với ước lượng một quan hệ tương quan giữa các biến số kinh tế xã hội.

Nhân quả (Causation) hay Tương quan (Correlation)?

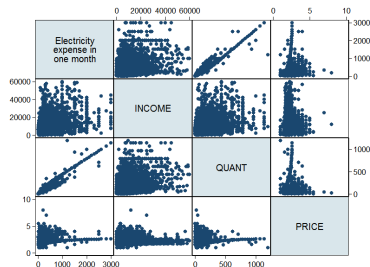
- ▶ Nhân quả: A gây ra B. Ví dụ: tăng giá điện dẫn đến giảm tiêu thụ.
- ▶ Tương quan: A và B biến động cùng hoặc ngược chiều. Ví dụ: giá điện và thu nhập người dân trong 10 năm qua đều tăng. Nhưng không có quan hệ nhân quả, giá điện tăng không phải do thu nhập tăng hay ngược lại.
- ▶ Cảnh báo với quan hệ nhân quả: giá điện tăng chưa chắc là nguyên nhân dẫn đến việc giảm tiêu thụ điện. Có thể do các nguyên nhân khác như sự xuất hiện của các thiết bị tiết kiệm điện năng, hay thu nhập bị giảm, dẫn đến tiêu thụ điện giảm. Để xác lập quan hệ nhân quả phải xử lý được các nguyên nhân khác có thể gây ra kết luận sai, dẫn đến đề xuất chính sách sai.

Ví dụ về tương quan và nhân quả

Xem bộ dữ liệu electricity.dta của 5000 hộ gia đình tại Việt Nam trong năm 2015. Sử dụng ma trận tương quan và đồ thị phân phối điểm (scatter plot) để nhận diện các mối quan hệ giữa các biến số.

```
. corr CONS INCOME QUANT PRICE  
(obs=4,820)
```

	CONS	INCOME	QUANT	PRICE
CONS	1.0000			
INCOME	0.3830	1.0000		
QUANT	0.9787	0.3869	1.0000	
PRICE	0.0350	-0.0040	-0.1162	1.0000



- ▶ Nếu dựa trên bộ dữ liệu này thì chính sách nào là phù hợp nhất nếu muốn giảm tiêu thụ điện?
- ▶ Các nguyên nhân có thể làm sai lệch mối quan hệ trên?

Các loại dữ liệu kinh tế

- ▶ Dữ liệu chéo (cross-sectional data): thông tin của cá nhân, hộ gia đình, công ty, đơn vị địa lý tại một thời điểm.
- ▶ Dữ liệu chuỗi thời gian (time-series data): các quan sát được theo dõi lặp lại qua nhiều kỳ, tập trung vào đặc tính tần suất và độ trễ của tác động.
- ▶ Dữ liệu chéo gộp (chéo kết hợp - pooled cross-sectional data).
- ▶ Dữ liệu bảng (panel/longitudinal data): các cá thể kinh tế được theo dõi lặp lại ít nhất là 2 kỳ, tập trung vào đặc tính thay đổi giữa các kỳ quan sát.

Bộ dữ liệu chéo WAGE1.dta về tiền lương bình quân theo giờ và đặc điểm cá nhân

	wage	educ	exper	female	married
1	3.1	11	2	1	0
2	3.2	12	22	1	1
3	3	11	2	0	0
4	6	8	44	0	1
5	5.3	12	7	0	1
6	8.8	16	9	0	1
7	11	18	16	0	0
8	5	12	5	1	0
9	3.6	12	26	1	0
10	18	17	22	0	1
11	6.3	16	8	1	0
12	8.1	13	3	1	0
13	8.8	12	15	0	1
14	5.5	12	18	0	0
15	22	12	31	0	1
16	17	16	14	0	1
17	7.5	12	10	1	1
18	11	13	16	1	0

- ▶ Mỗi dòng dữ liệu là một quan sát (cá thể). Thứ tự dữ liệu không ảnh hưởng đến kết quả phân tích.
- ▶ Dễ phân tích nhất nhưng cũng gặp nhiều vấn đề nhất.
- ▶ *Câu hỏi: So sánh tiền lương giữa phụ nữ và đàn ông, giữa người có gia đình và độc thân, người có trình độ giáo dục và số năm kinh nghiệm khác nhau.*

Dữ liệu chuỗi thời gian PHILLIPS.dta về tỷ lệ lạm phát và thất nghiệp ở Mỹ

	year	unem	inf	unem_1	inf_1
3	1960	5.3	1.3	5.9	-1.2
4	1961	3.3	7.9	5.3	1.3
5	1962	3	1.9	3.3	7.9
6	1963	2.9	.8	3	1.9
7	1964	5.5	.7	2.9	.8
8	1965	4.4	-.4	5.5	.7
9	1966	4.1	1.5	4.4	-.4
10	1967	4.3	3.3	4.1	1.5
11	1968	6.8	2.8	4.3	3.3
12	1969	5.5	.7	6.8	2.8
13	1960	5.5	1.7	5.5	.7
14	1961	6.7	1	5.5	1.7
15	1962	5.5	1	6.7	1
16	1963	5.7	1.3	5.5	1
17	1964	5.2	1.3	5.7	1.3
18	1965	4.5	1.6	5.2	1.3
19	1966	3.8	2.9	4.5	1.6
20	1967	3.8	3.1	3.8	2.9

- ▶ Thứ tự của dữ liệu rất quan trọng. Cần lưu ý đến tần suất thu dữ liệu và tính chu kỳ.
- ▶ Khó phân tích nhất trong các loại dữ liệu.
- ▶ *Câu hỏi: vẽ đồ thị xu hướng biến động của tỷ lệ thất nghiệp và lạm phát theo thời gian. Dữ liệu trên có phù hợp với lý thuyết về đường Phillips không?*

Dữ liệu chéo gộp HPRICE3.dta

	year	age	cbd	lnac	price	rooms
172	1978	0	14000	19000	107900	7
173	1978	0	14000	14000	120000	7
174	1978	1	14000	14000	120000	7
175	1978	0	14000	14000	121000	7
176	1978	0	14000	18000	107000	7
177	1978	0	14000	18000	119900	7
178	1978	0	14000	19000	142500	7
179	1978	0	14000	19000	94211	7
180	1981	81	4000	1000	49000	6
181	1981	71	3000	2000	52000	5
182	1981	31	3000	2000	68000	6
183	1981	41	3000	2000	54000	6
184	1981	31	4000	2000	70000	6
185	1981	81	3000	2000	47000	6

- ▶ Hai hoặc nhiều bộ dữ liệu chéo gộp lại thành một bộ dữ liệu lớn, có thêm biến thời gian để xác định thời điểm thu thập dữ liệu. Các quan sát không nhất thiết phải trùng lặp lại giữa các thời kỳ.
- ▶ Có thể cung cấp nhiều thông tin hơn dữ liệu chéo thuần túy, nhưng không mạnh như dữ liệu bảng.
- ▶ *Câu hỏi: các nhân tố nào ảnh hưởng đến giá nhà? và các nhân tố đó thay đổi theo thời gian như thế nào?*

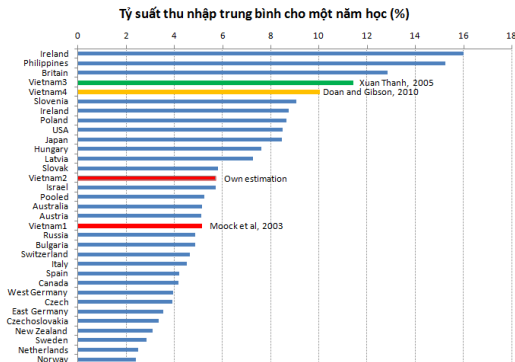
Dữ liệu bảng CRIME2.dta

	pop	crimes	unem	officers	year
1	229528	17136	8.2	326	82
2	246815	17906	9.7	321	87
3	814054	75654	8.1	1621	82
4	933177	83960	5.4	1803	87
5	374974	31352	9	633	82
6	406297	31364	5.9	685	87
7	176496	15698	12.6	245	82
8	201723	16953	5.7	259	87
9	288446	31202	12.6	504	82
10	331728	34355	7.4	563	87
11	122768	16806	13.9	186	82
12	142124	15931	5.7	232	87

- ▶ Mỗi cá thể được thu dữ liệu trong ít nhất 2 kỳ, kỳ trước và kỳ sau.
- ▶ Cho phép thực hiện các phân tích tốt nhất trong các loại dữ liệu. Tuy nhiên yêu cầu cao về kỹ thuật.

Ví dụ về ước lượng tỷ suất thu nhập của việc đi học

- ▶ Chính sách giáo dục là một trong những ưu tiên hàng đầu của mọi quốc gia trên thế giới.
- ▶ Ước lượng tác động của giáo dục lên thu nhập sẽ cho biết hiệu quả của các chính sách khuyến học đối với cá nhân.
- ▶ Với toàn bộ nền kinh tế tác động còn có thể cao hơn.



Mô hình lý thuyết ước lượng tỷ suất thu nhập của việc đi học

Mô hình nguồn lực con người (human capital) của Mincer (1974), theo đó $\log Y_i$ là logarit của thu nhập là hàm số của số năm đi học và số năm kinh nghiệm làm việc:

$$\log Y_i = \beta_0 + \beta_1 \times EDUC_i + \beta_2 \times EXP_i + \beta_3 \times EXP_i^2 + \varepsilon_i$$

Các vấn đề có thể xảy ra làm sai lệch kết quả

- ▶ Nhân tố không quan sát được có tương quan với thu nhập và trình độ giáo dục → kết quả bị sai lệch. Ảnh hưởng như thế nào?

⇒ Các chương tiếp theo chúng ta sẽ làm quen với các mô hình căn bản, các giả định, và ước lượng các mô hình từ đơn giản đến phức tạp.

Hồi quy tuyến tính đơn biến (Simple Linear Regression - SLR)

Giới thiệu mô hình SLR

Chúng ta có 2 biến số x và y và muốn tìm hiểu x ảnh hưởng như thế nào đến y . Mô hình đơn giản nhất được viết dưới dạng một hàm số tuyến tính của y theo x :

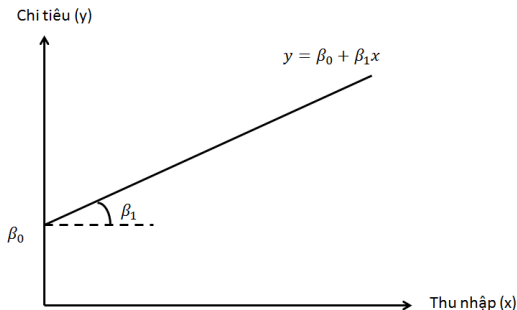
$$y = \beta_0 + \beta_1 x + u$$

- ▶ y gọi là biến phụ thuộc/biến được giải thích
- ▶ x là biến độc lập/biến giải thích
- ▶ u là sai số, bao gồm tất cả những yếu tố khác ảnh hưởng đến y nhưng không nằm trong x .
- ▶ β_0 và β_1 là các tham số trong mô hình.

Diễn giải mô hình

- ▶ β_0 là tung độ gốc
- ▶ β_1 là độ dốc của đường hồi quy
- ▶ Nếu các yếu tố khác (u) giữ nguyên không đổi, x tác động tuyến tính tới y thông qua phương trình:

$$\Delta y = \beta_1 \Delta x$$



Hàm hồi quy tổng thể và Hàm hồi quy mẫu

- ▶ Với giả định sai số bình quân $E(u)$ trong tổng thể bằng không, $E(u) = 0$, hàm hồi quy tổng thể (Population Regression Function - PFR) được viết dưới dạng:

$$y = \beta_0 + \beta_1 x$$

- ▶ Chúng ta không bao giờ biết chính xác giá trị của β_0 và β_1 từ tổng thể.
- ▶ Các phương pháp hồi quy sẽ ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ từ dữ liệu, từ đó chúng ta có mô hình hồi quy mẫu (Sample Regression Function - SRF):

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

Ví dụ: Tỷ suất sinh lợi ảnh hưởng như thế nào đến lương của CEO

- ▶ Xem bộ dữ liệu CEOSAL1.dta.
- ▶ Giả sử tiền lương CEO được quyết định do kết quả hoạt động của doanh nghiệp (đại diện bởi tỷ suất sinh lợi trên vốn, *roe*) mang lại:

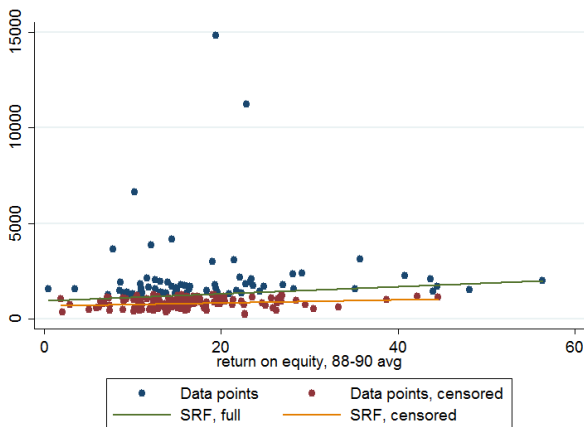
$$\text{salary} = \beta_0 + \beta_1 \text{roe} + u$$

- ▶ Kỳ vọng gì về giá trị của β_0 và β_1 ?
- ▶ Tìm hiểu bộ dữ liệu:

```
. sum salary roe
```

Variable	Obs	Mean	Std. Dev.	Min	Max
salary	209	1281.12	1372.345	223	14822
roe	209	17.18421	8.518509	.5	56.3

```
graph twoway (scatter salary roe) (scatter salary roe if salary < 1281) (lfit salary roe) (lfit
salary roe if salary < 1281), legend(label(1 "Data points") label(2 "Data points, censored")
label(3 "SRF, full") label(3 "SRF, full") label(4 "SRF, censored")) graphregion(fcolor(white))
```

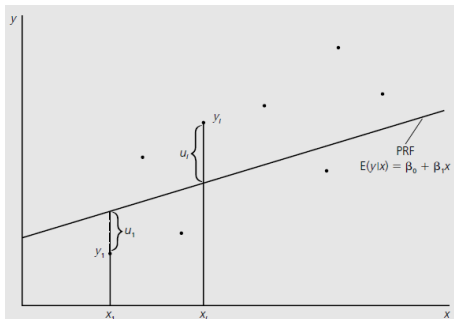


⇒ Mục tiêu là ước lượng được $\hat{\beta}_0$ và $\hat{\beta}_1$ của SRF càng gần với β_0 và β_1 của PRF càng tốt.

Phương pháp bình phương tối thiểu thông thường (Ordinary Least Square - OLS)

- ▶ Ký hiệu i đại diện cho quan sát thứ i của dữ liệu gồm n quan sát. Từ phương trình hồi quy ta có thể viết lại là $u_i = y_i - \beta_0 - \beta_1 x_i$.
- ▶ Cơ chế của phương pháp OLS là tìm $\hat{\beta}_0$ và $\hat{\beta}_1$ để tối thiểu hóa tổng bình phương của u_i .

$$U = \min \sum_{i=1}^n u_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$



- ▶ Dựa vào hình vẽ: Bản chất của OLS là tìm phương trình đường thẳng đi qua phân phối điểm của dữ liệu sao cho tổng bình phương khoảng cách từ các điểm dữ liệu đến đường thẳng là tối thiểu. Tại sao phải dùng bình phương của khoảng cách?
- ▶ Các phương pháp khác có thể sử dụng giá trị tuyệt đối của khoảng cách.

Cơ chế của phương pháp OLS

Để tìm giá trị $\hat{\beta}_0$ và $\hat{\beta}_1$ để tối thiểu hóa tổng bình phương của u_i , ta sử dụng điều kiện bậc nhất là đạo hàm của hàm mục tiêu bằng không tại các giá trị cực trị:

$$\frac{\partial U}{\partial \beta_0} = -2 \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (1)$$

và

$$\frac{\partial U}{\partial \beta_1} = -2 \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \quad (2)$$

Giải các điều kiện bậc nhất ta thu được giá trị của $\hat{\beta}_0$ và $\hat{\beta}_1$:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}$$

và

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Điều kiện của ước lượng OLS

- ▶ Hai điều kiện bậc nhất tương ứng với $E(u) = 0$ và $E(xu) = 0$ sẽ đảm bảo ước lượng OLS là không chệch (unbiased) và nhất quán (consistent).
- ▶ Diễn giải: trung bình của sai số u bằng không và sai số u không tương quan với biến giải thích x .

Sau khi ước lượng được $\hat{\beta}_0$ và $\hat{\beta}_1$, ta có thể tính được các giá trị của y và u tại các giá trị của x như sau:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

và

$$\hat{u}_i = y_i - \hat{y}_i$$

$\hat{y}_i$ được gọi là giá trị thích hợp (fitted value) hoặc giá trị dự báo (predicted value) của biến phụ thuộc tại mỗi giá trị của x_i cho trước. $\hat{u}_i$ gọi là phần dư (residual).

Một số ví dụ

Sử dụng bộ dữ liệu CEOSAL1.dta. Chúng ta muốn ước lượng tiền lương của CEO theo tỷ suất thu nhập trên vốn, *roe*. Giả sử hai điều kiện về sai số và không tương quan được thỏa.

```
reg salary roe
predict salaryhat, xb
predict uhat, resid
```

Source	SS	df	MS	Number of obs	=	209
Model	5166419.04	1	5166419.04	F(1, 207)	=	2.77
Residual	386566563	207	1867471.32	Prob > F	=	0.0978
				R-squared	=	0.0132
				Adj R-squared	=	0.0084
Total	391732982	208	1883331.64	Root MSE	=	1366.6

salary	Coef	Std. Err.	t	P> t	[95% Conf. Interval]	
roe	18.50119	11.12325	1.66	0.098	-3.428196	40.43057
_cons	963.1913	213.2403	4.52	0.000	542.7902	1383.592

	salary	roe	salaryhat	
1	1095	14.1	1224.058	-
2	1001	10.9	1164.854	-
3	1122	23.5	1397.969	-
4	578	5.9	1072.348	-
5	1368	13.8	1218.508	-
6	1145	20	1333.215	-
7	1078	16.4	1266.611	-

Thực hiện ước lượng tuần tự bằng Stata hoặc MS Excel

Dựa trên công thức

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} = \frac{Cov(X, Y)}{Var(X)}$$

và

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Độ thích hợp của các mô hình hồi quy

Dựa trên tổng bình phương (SST, còn được gọi là tổng biến thiên), tổng bình phương được giải thích (SSE), và tổng bình phương phần dư:

$$SST = \sum (y_i - \bar{y})^2$$

$$SSE = \sum (\hat{y}_i - \bar{y})^2$$

$$SSR = \sum \hat{u}_i^2$$

và

$$SST = SSE + SSR$$

Hệ số thích hợp R-bình phương được tính bằng tỷ số giữa biến thiên được giải thích và tổng biến thiên:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

Hiểu thế nào về hệ số thích hợp R^2 ?

- ▶ Mô hình gồm có phần quan sát được x và phần không quan sát được u .
- ▶ Phần quan sát được giải thích được càng nhiều các nhân tố ảnh hưởng đến y càng tốt. Ví dụ $R^2 = 0.5$ có nghĩa là mô hình giải thích được 50% độ biến thiên của mẫu.
- ▶ $\hat{y}_i$ và $\hat{u}_i$ sẽ có quan hệ nghịch biến vì tổng biến thiên là cố định đối với mỗi mẫu.

$$0 \leq R^2 \leq 1$$

- ▶ Trên thực tế, hệ số xác định luôn $0 < R^2 < 1$.
- ▶ *Câu hỏi: Nếu $R^2 = 0$ hoặc $R^2 = 1$ thì hình dạng đường hồi quy mẫu sẽ như thế nào?*

Lưu ý về hệ số thích hợp R^2

- ▶ Nhìn chung những người mới nghiên cứu hay có xu hướng chọn mô hình hay biến số để tăng R^2 . Điều này không sai nhưng không được khuyến khích để xây dựng mô hình.
- ▶ Sử dụng R^2 để chọn biến có thể dẫn đến những sai sót rất nghiêm trọng, đặc biệt khi biến giải thích là không ngẫu nhiên.
- ▶ Không có tiêu chí để xác định R^2 khi nào cao hay thấp.
- ▶ Với hồi quy đa biến, tăng số biến số trong mô hình làm tăng R^2 , do đó cần phải cân đối giữa số biến với độ thích hợp của mô hình.

Ví dụ mô hình giá nhà

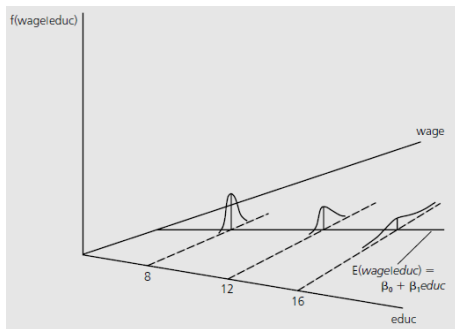
Sử dụng bộ dữ liệu `hprice1.dta`.

Hãy lựa chọn một mô hình hồi quy đơn biến giải thích các nhân tố ảnh hưởng đến giá nhà. Biến số nào giải thích tốt nhất? Cấu trúc hàm nào phù hợp nhất?

Phương sai của sai số trong mô hình hồi quy

Nếu phương sai của sai số là $Var(u) = \sigma^2$ là một hằng số, không phụ thuộc vào các biến giải thích x , khi này ta có mô hình hồi quy đơn biến với phương sai của sai số không đổi (homoskedasticity).

- Phương sai không đổi là gì?



- Ước lượng bằng OLS có tính chất đặc biệt gọi là ước lượng tuyến tính không chệch hiệu quả nhất (Best Linear Unbiased Estimator - BLUE).

Tóm tắt tuần 1

- ▶ Hiểu các loại dữ liệu kinh tế và sự khác biệt giữa chúng.
- ▶ Nhận biết được các thành phần của một phân tích định lượng.
- ▶ Cấu trúc của mô hình hồi quy đơn biến, các giả định, và các công thức ước lượng các tham số trong mô hình.
- ▶ Sử dụng Stata để thực hiện các mô hình hồi quy đơn giản, lựa chọn mô hình, diễn giải kết quả, và phân tích phê phán độ thích hợp của mô hình.