Nhập môn Kinh tế lượng
(Introduction to Econometrics)

Lê Việt Phú
Trường Chính sách Công và Quản lý Fulbright

07-10/12/2021

Tình huống nghiên cứu điển hình

Giả dụ bạn được chính phủ giao nhiệm vụ đánh giá tác động của chính sách giáo dục lên thu nhập của người dân. Chính sách giáo dục được thể hiện thông qua việc cung cấp các chương trình giảng dạy, từ phổ cập tiểu học bắt buộc đến các cấp trung học cơ sở, trung học phổ thông, cao đẳng/đại học và sau đại học.

Nhiệm vụ cụ thể của bạn là đánh giá hiệu quả kinh tế (economic return). Hiệu quả kinh tế thường được đo lường bằng tiền lương theo giờ hoặc tổng thu nhập. Chính sách giáo dục được đo lường xấp xỉ bằng tổng số năm đi học từ tất cả các cấp học.

Bạn cần chuẩn bị gì để hoàn thành nhiệm vụ trên?

1. Tìm lý thuyết để giải thích mối quan hệ giữa chính sách can thiệp với kết quả đạt được:
   - Tăng vốn con người (đại diện bởi số năm đi học) làm tăng năng suất công việc/tiền lương tại điểm cân bằng thị trường (Lý thuyết Mincer)

2. Mô hình hóa lý thuyết: Xây dựng mô hình kinh tế lượng kết nối giữa biến số kết quả với biến số chính sách

$$y = f(x_1, x_2, ..., x_k)$$

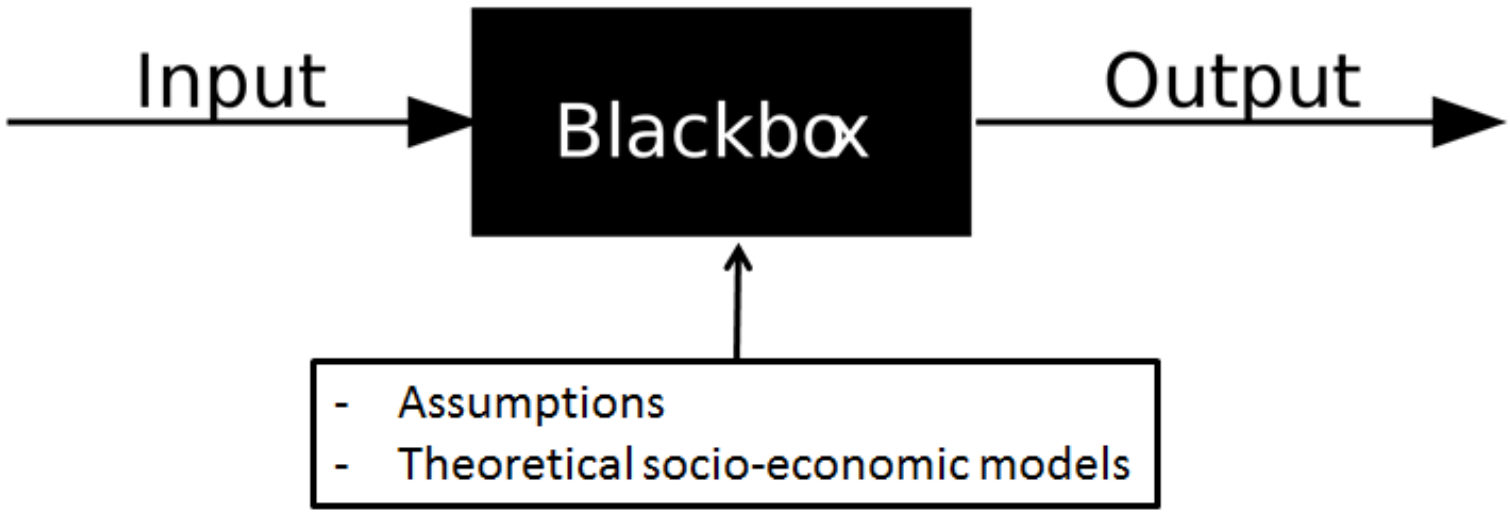
trong đó: y là thu nhập, x_1 là số năm đi học – đại diện cho biến chính sách, x_2 là số năm kinh nghiệm làm việc, x_3, x_4,... là các nhân tố ảnh hưởng đến tiền lương như nhân khẩu học, yếu tố kinh tế xã hội, môi trường kinh doanh...

3. Tìm kiếm dữ liệu phù hợp
   - Nghiên cứu định lượng yêu cầu làm chủ nhiều khía cạnh khác nhau liên quan đến dữ liệu: thu thập ở đâu, của ai, chi tiết đặc tính của dữ liệu, cấp độ chi tiết của thông tin, tần suất thu thập, phạm vi thu thập, hình thức thu thập...
   - Dữ liệu có vai trò quyết định đối với lựa chọn mô hình, và mức độ tin cậy của kết quả.

4. Ước lượng mô hình, thực hiện các bước kiểm định và kiểm chứng để đảm bảo mô hình vững
   - Giải bài toán tối ưu hóa bằng công cụ thống kê xác suất với sự hỗ trợ của các phần mềm chuyên dụng để tìm ra được các tham số tối ưu
   - Cho phép bóc tách được tác động mong muốn ra khỏi các nhân tố gây nhiễu
   - Kết quả không phụ thuộc/nhạy cảm với các giả định sử dụng

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.



Thế nào là một thiết kế nghiên cứu hợp lý

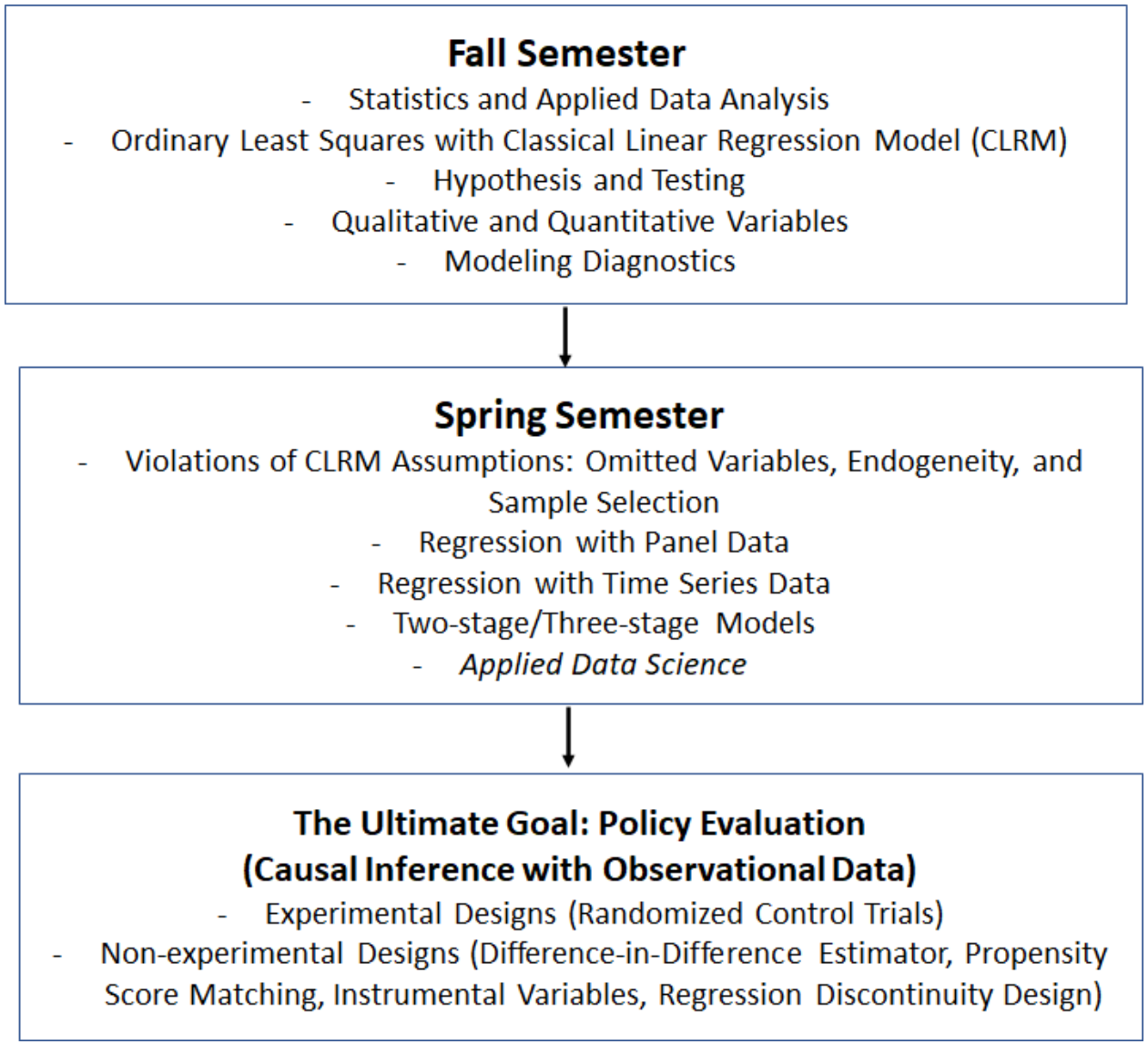
Các nghiên cứu chính sách vững chắc cần phải dựa trên các thiết kế nghiên cứu (research design) hợp lý để bóc tách được tác động của chính sách can thiệp đến kết quả.

- Thiết kế nghiên cứu phải có mô hình lý thuyết vững chắc – thay vì chỉ sử dụng các **thuật toán** tính toán để tìm ra mô hình có khả năng dự báo cao nhất.
- Phải nhận định được các hạn chế của mô hình và dữ liệu, và đề xuất phương án xử lý nhằm đảm bảo độ vững của mô hình khi các giả định căn bản bị vi phạm.

Một số câu hỏi liên quan đến chính sách có thể trả lời bằng công cụ kinh tế lượng

- 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?

Cấu trúc của các môn học định lượng trong chương trình MPP-PA



- Học phần kinh tế lượng 1 chỉ tập trung vào việc xây dựng mô hình vững chắc với các giả định căn bản được đảm bảo.
- Học phần 2 giới thiệu các mô hình nâng cao khi các giả định căn bản bị vi phạm.
- Phần ứng dụng sẽ giúp học viên làm quen và thiết kế các mô hình cho nghiên cứu đánh giá tác động chính sách (impact evaluation/program evaluation).

Nội dung của học phần nhập môn kinh tế lượng

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

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:

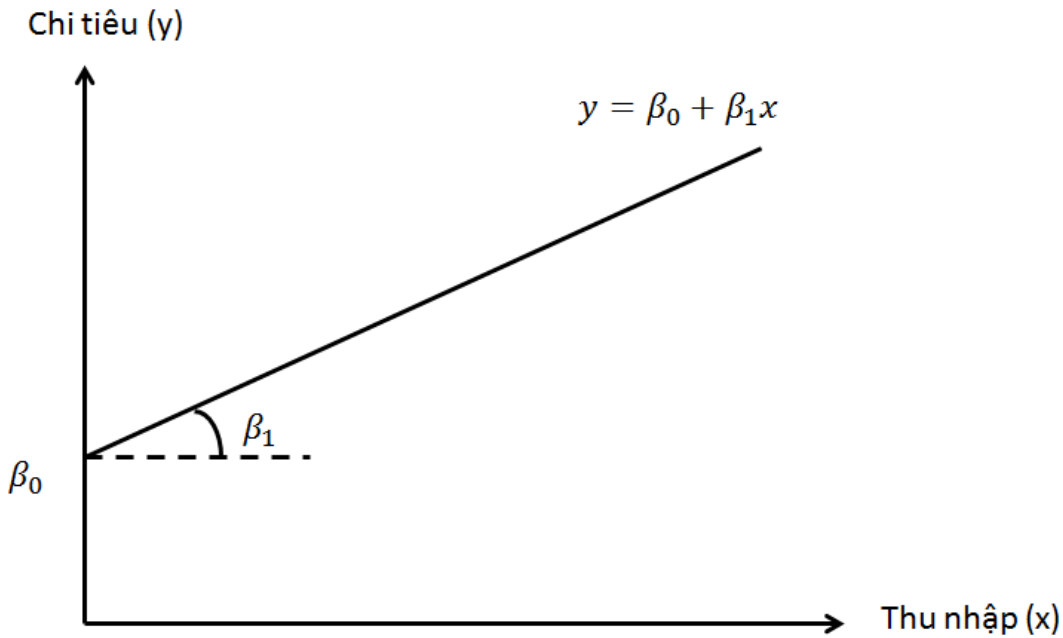
$$y_i = \beta_0 + \beta_1 x_i + u_i$$

- i đại diện cho quan sát thứ i trong tổng thể gồm có n quan sát.
- y gọi là biến phụ thuộc/biến được giải thích/biến phản ứng/biến được dự báo
- x là biến độc lập/biến giải thích/biến kiểm soát/biến dự báo
- u là sai số (số hạng nhiễu), không quan sát được, 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 – cần phải ước lượng.

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):

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

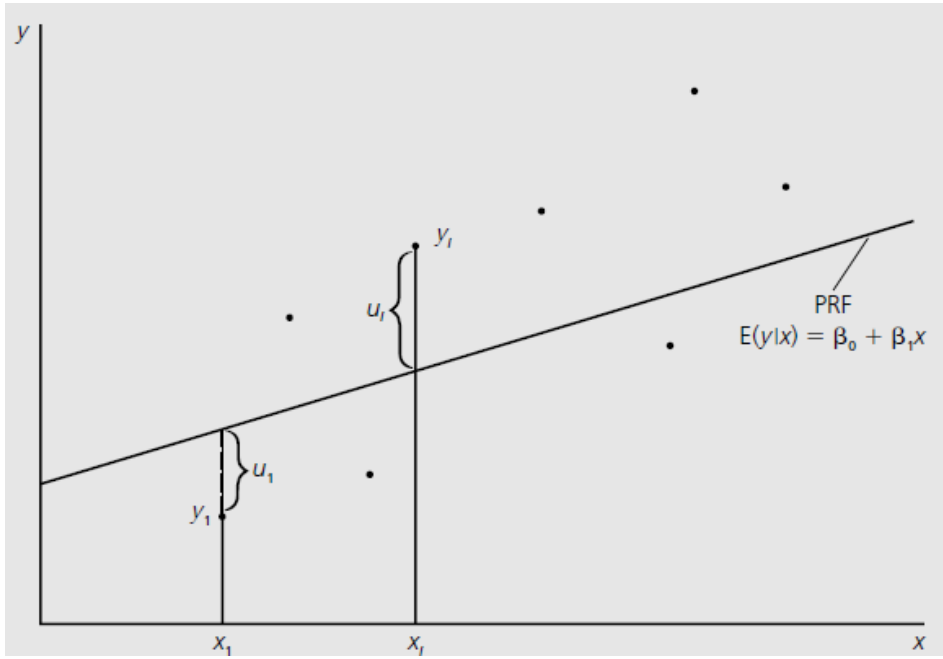
Phương pháp bình phương tối thiểu thông thường (Ordinary Least Square - OLS)

- 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 (least absolute deviations-LAD).

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:

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

và

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

Điều kiện của ước lượng OLS

Hai điều kiện bậc nhất (1) và (2) tương ứng với:

$$\mathbf{E}(\mathbf{u}) = 0 \tag{3}$$

$$\mathbf{E}(\mathbf{xu}) = 0 \tag{4}$$

Với $\mathbf{E}(\mathbf{u}) = 0$ thì $\mathbf{E}(\mathbf{xu}) = Cov(x, u)$.

Giải các điều kiện bậc nhất để tìm giá trị tối ưu $\hat{\beta}_0$ và $\hat{\beta}_1$

Chứng minh rằng:

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

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

Sau khi ước lượng được $\hat{\beta}_0$ và $\hat{\beta}_1$, ta có thể tính được các giá trị dự báo của y và u tại các giá trị cho trước 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).

Diễn giải điều kiện bậc nhất của phương pháp ước lượng bằng OLS

Trung bình của sai số u bằng không và sai số u không tương quan với biến giải thích x.

- Điều kiện sai số trung bình bằng không thực ra được tự động thỏa khi hàm hồi quy có chứa tung độ gốc (tham số β_0).
- Điều kiện sai số không được tương quan với biến giải thích là điều kiện rất khó giải thích và khó đảm bảo trên thực tế.
  - u chứa những nhân tố không quan sát được. Vậy u là gì?
  - Làm sao có thể đảm bảo nhân tố không quan sát được u không tương quan với phần quan sát được x?

Đặc tính của ước lượng bằng OLS

- Ước lượng OLS là không chệch (unbiased),

$$\mathbf{E}(\hat{\beta}) = \beta$$

- Ước lượng OLS là nhất quán (consistent),

$$\mathbf{plim}(\hat{\beta}) \to \beta$$

khi cỡ mẫu tiến đến vô cùng, $n \to \infty$

Học viên tự tìm hiểu và chứng minh!

Diễn giải

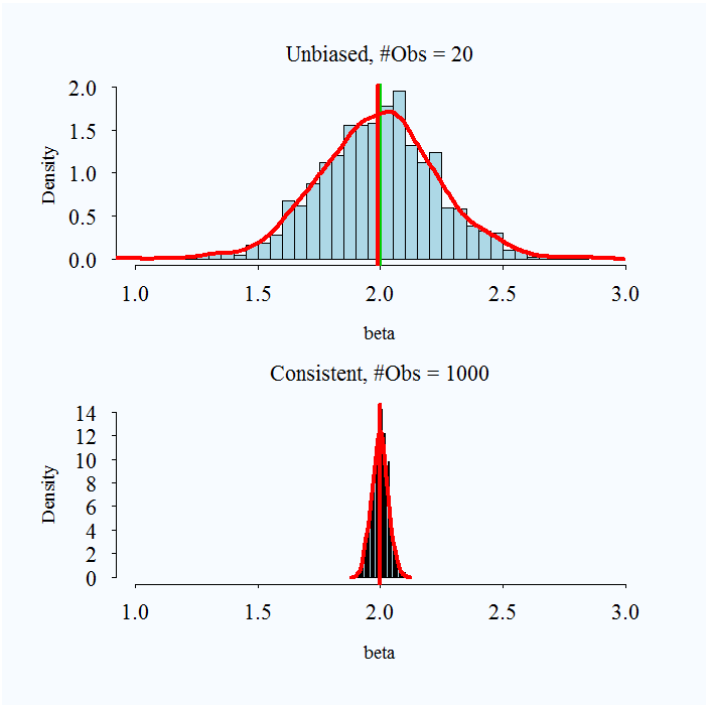
- Một ước lượng được gọi là không chệch nếu kỳ vọng của ước lượng của tham số β bằng với giá trị thực sau nhiều lần lấy mẫu lặp. Điều này có nghĩa là khi chúng ta lặp lại ước lượng nhiều lần với các nhóm mẫu khác nhau, và lấy giá trị trung bình của các ước lượng, thì giá trị trung bình đó sẽ bằng với giá trị thực (mặc dù không bao giờ biết được giá trị thực).

- Ước lượng được coi là nhất quán nếu giá trị của ước lượng hội tụ về mặt xác suất (convergence in probability) về giá trị thực nếu tăng cỡ mẫu tiến đến vô cùng.

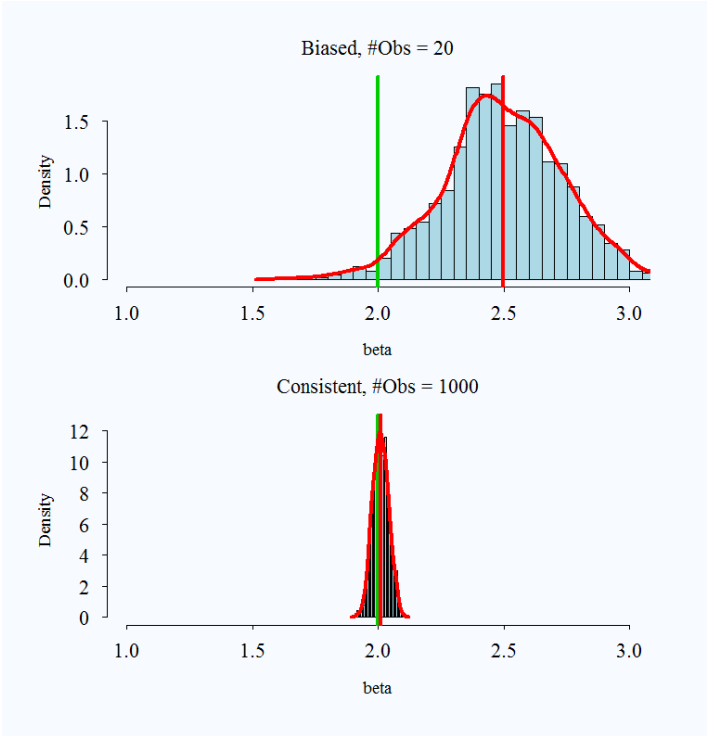
Nếu điều kiện (4) bị vi phạm thì ước lượng bằng OLS sẽ mất các thuộc tính này.

Extra: Khái niệm Bias và Consistency của một ước lượng

F1



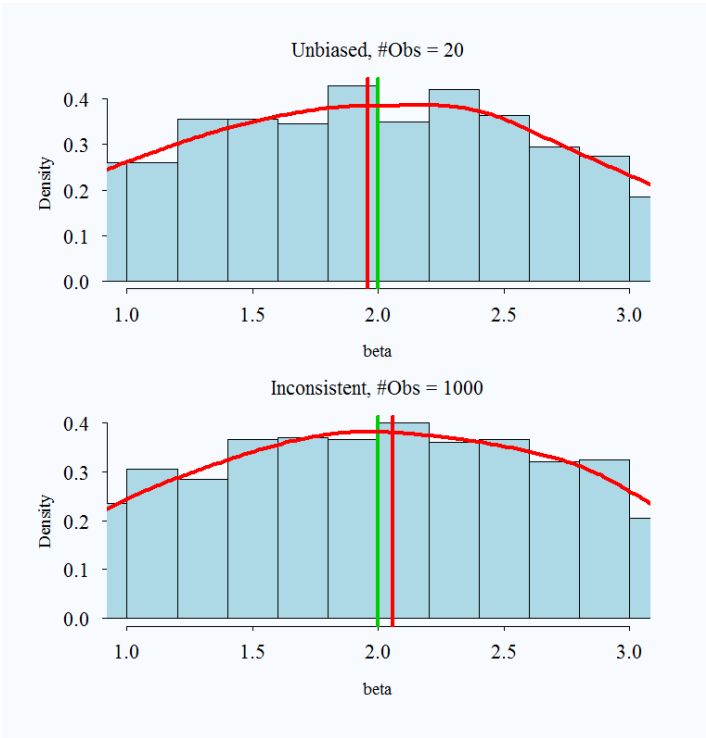
F2



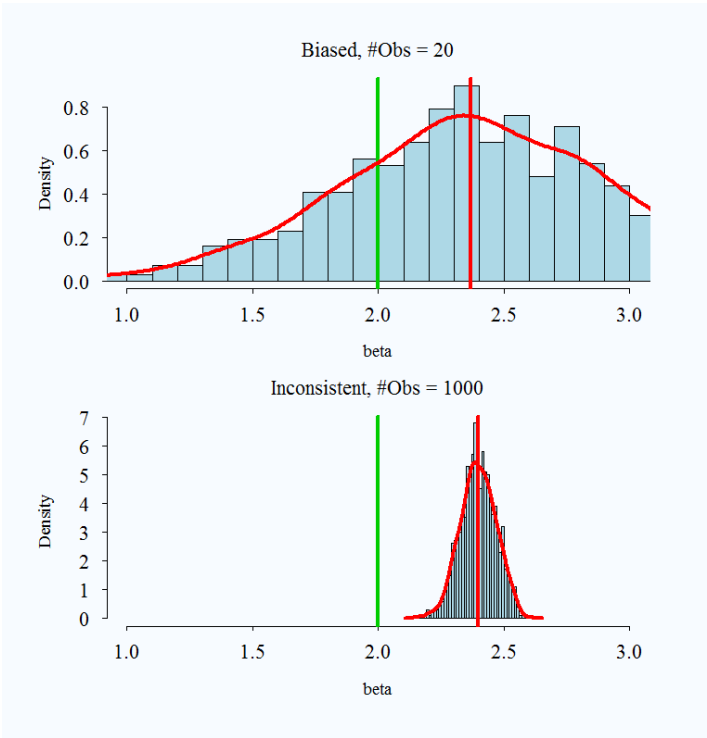
F1. Ước lượng không chệch và nhất quán.

F2. Ước lượng chệch nhưng nhất quán.

F3



F4



F3. Ước lượng không chệch nhưng không nhất quán.

F4. Ước lượng chệch và không nhất quán.

Ví dụ 1: Ước lượng tác động của tỷ suất sinh lợi của doanh nghiệp lên mức 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:

$$salary = \beta_0 + \beta_1 roe + u$$

- Kỳ vọng gì về giá trị của β_0 và β_1?

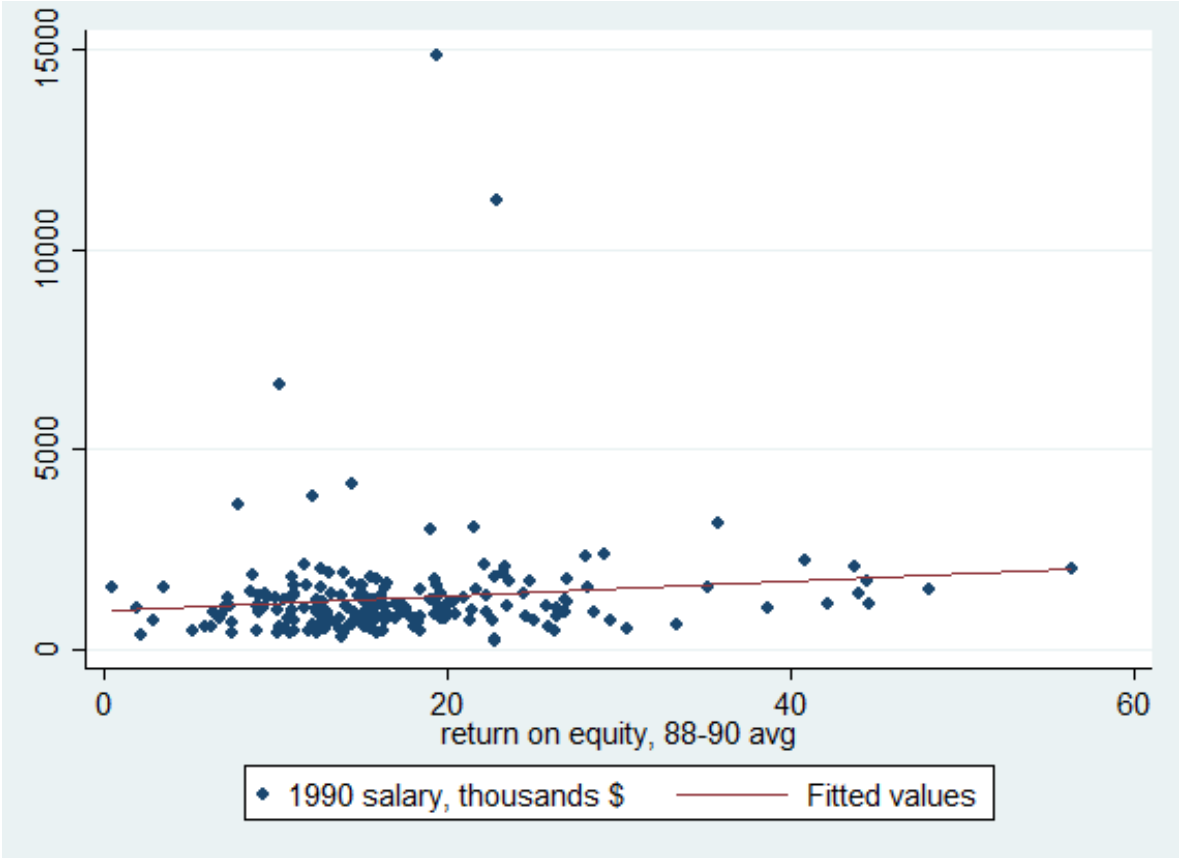
Ước lượng tác động của tỷ suất thu nhập lên tiền lương của CEO

Source	SS	df	MS	Number of obs	=	209
				F(1, 207)	=	2.77
Model	5166419.04	1	5166419.04	Prob > F	=	0.0978
Residual	386566563	207	1867471.32	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

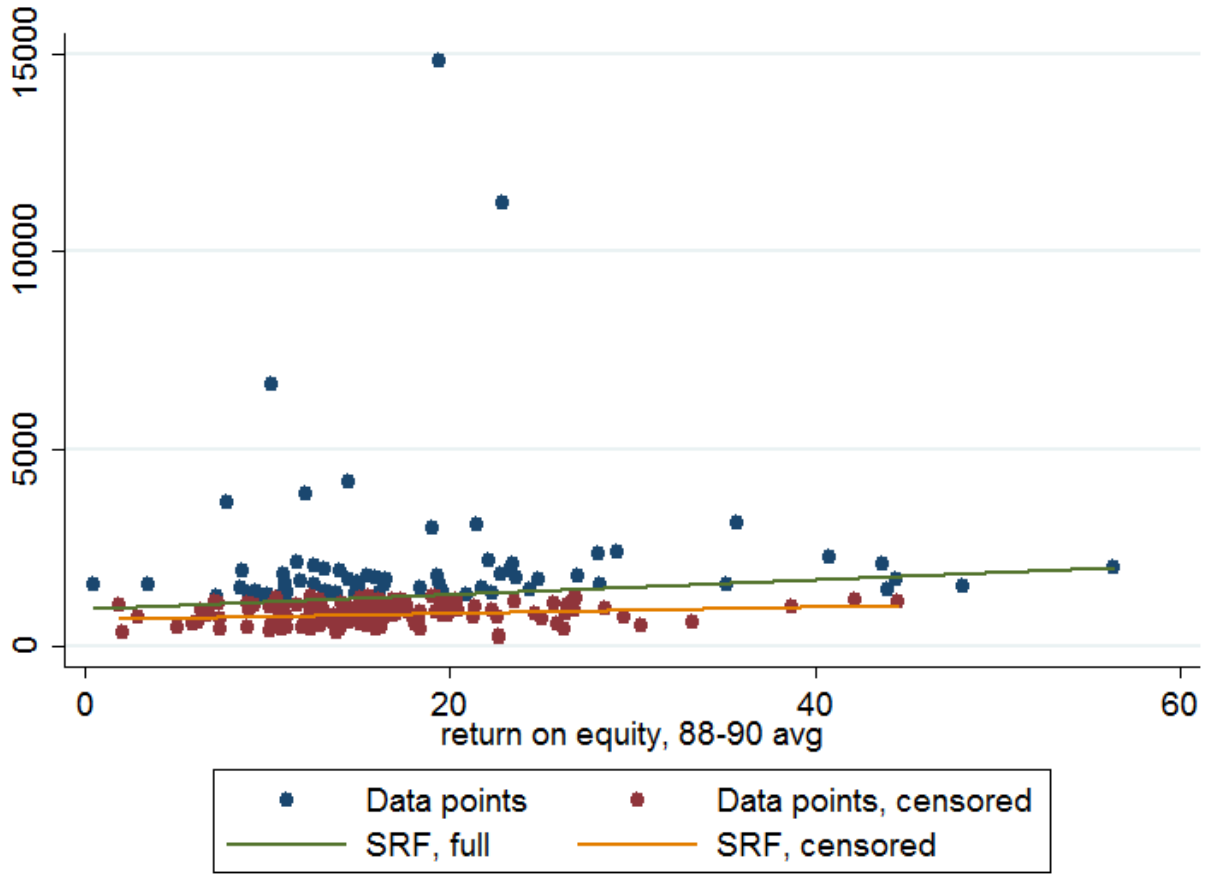
	salary	roe	salaryhat	uhat
1	1095	14.1	1224.058	-129.0581
2	1001	10.9	1164.854	-163.8543
3	1122	23.5	1397.969	-275.9692
4	578	5.9	1072.348	-494.3483
5	1368	13.8	1218.508	149.4923
6	1145	20	1333.215	-188.2151
7	1078	16.4	1266.611	-188.6108

Hình dạng đường hồi quy



So sánh đường hồi quy mẫu với tổng thể

Giả sử chúng ta chỉ có dữ liệu của những CEO có mức lương từ trung bình trở xuống (salary < 1.281 triệu đô la/năm). Ước lượng tương ứng với đồ thị màu cam.



$\Rightarrow$ 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.

Thực hành ước lượng OLS theo các bước thủ công

Ước lượng các tham số thủ công theo công thức sau:

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

và

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

Vai trò của các giả định trong mô hình OLS

$$y_i = \beta_0 + \beta_1 x_i + u_i$$

1. Tuyến tính theo tham số
2. Quá trình lấy mẫu dữ liệu là ngẫu nhiên
3. Có sự thay đổi trong các giá trị của biến giải thích x
4. Sai số u không tương quan với biến giải thích x, $E(u|x) = 0$

Bằng toán học, giả định (4) tương đương với:

$$\mathbf{E}(\mathbf{u}) = 0 \qquad (4.1)$$

$$\mathbf{E}(\mathbf{xu}) = 0 \qquad (4.2)$$

Vai trò của các giả định trong mô hình OLS

- **Giả định (4.2) là giả định quan trọng nhất trong mô hình OLS**. Rất khó chứng minh trong thực tế. Cần thiết phải hiểu sâu về lý thuyết kinh tế và quá trình thu thập dữ liệu để giải thích.
- Nếu giả định (4.2) bị vi phạm, ước lượng OLS sẽ bị chệch và không nhất quán.
- Toàn bộ nội dung của môn KTL 2 chỉ tập trung để giải quyết vấn đề này.

Ví dụ về tính hợp lý của giả định sai số không tương quan với biến giải thích

Giả định chúng ta ước lượng mô hình tỷ suất thu nhập của việc đi học với một biến giải thích là số năm đi học:

$$log(income) = \beta_0 + \beta_1 * educ + u$$

Sai số u có thể gồm những nhân tố gì không quan sát được và có tương quan với biến số năm đi học?

Lựa chọn biến và cấu trúc hàm trong mô hình hồi quy

- Cách sử dụng biến số ảnh hưởng đến ý nghĩa của mô hình.
- Sử dụng đơn vị (level), logarithm, hay tỷ lệ thay đổi được quyết định bởi mô hình kinh tế.
- Có thể lấy logarithm của biến số khi dữ liệu có phân phối lệch.

Model	Dependent Variable	Independent Variable	Interpretation of β_1
Level-level	y	x	$\Delta y = \beta_1 \Delta x$
Level-log	y	$\log(x)$	$\Delta y = (\beta_1/100)\%\Delta x$
Log-level	$\log(y)$	x	$\%\Delta y = (100\beta_1)\Delta x$
Log-log	$\log(y)$	$\log(x)$	$\%\Delta y = \beta_1 \%\Delta x$

Đánh giá 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 (SSR):

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

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

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

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?*

Ví dụ 2: So sánh các mô hình tiền lương của CEO

So sánh hai mô hình với biến phụ thuộc lần lượt là tiền lương và logarithm của tiền lương. Mô hình nào phù hợp hơn? Giải thích.

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 3: Mô hình giá nhà

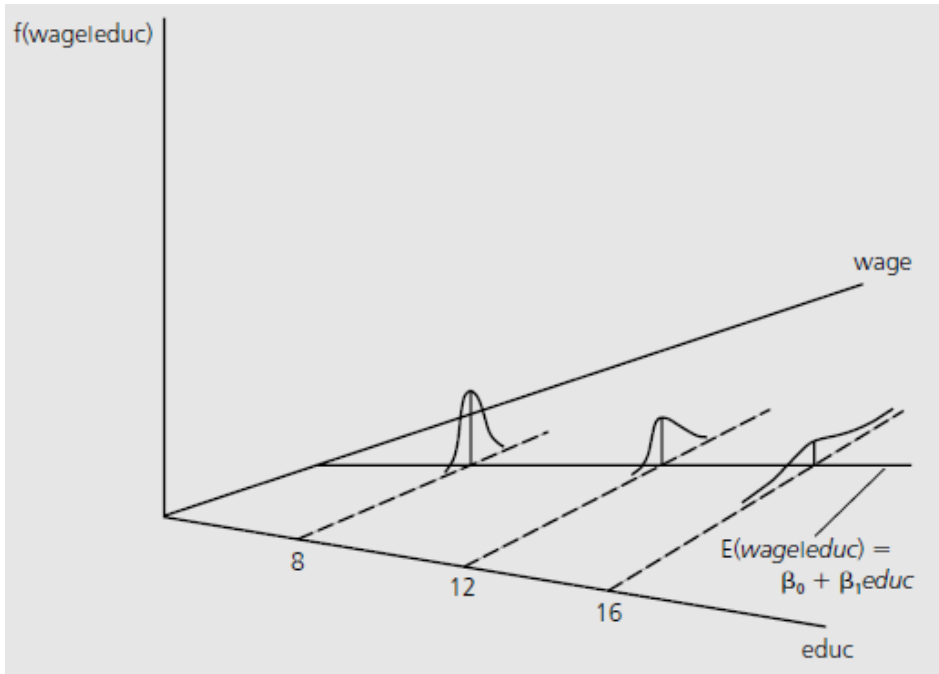
Sử dụng bộ dữ liệu hprice1.dta.

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?

Giả định 5: 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).

Hồi quy Đa biến
(Multivariate Regression)

Mô hình hồi quy đa biến

Tương tự như mô hình hồi quy đơn biến, tuy nhiên với nhiều biến giải thích. Ví dụ mô hình hồi quy với hai biến giải thích:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$$

- i là quan sát thứ i trong mẫu bao gồm n quan sát
- y gọi là biến phụ thuộc/biến được giải thích
- x_1, x_2 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_1, x_2.
- β_0, β_1, β_2 là các tham số trong mô hình – cần phải ước lượng.

Phương pháp bình phương tối thiểu thông thường OLS với hồi quy đa biến

- Tìm $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ để tối thiểu hóa tổng bình phương của sai số u_i:

$$U = min \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

  với ký hiệu i đại diện cho quan sát thứ i.

- $\hat{\beta}_1$ và $\hat{\beta}_2$ là tác động riêng phần của các biến giải thích x_1 và x_2 lên biến phụ thuộc.
- Ý nghĩa của các trị thống kê R^2, SST, SSE, SSR tương tự như mô hình SLR.

Điều kiện của ước lượng OLS

Tương tự như các điều kiện của mô hình SLR:

- Hai điều kiện bậc nhất tương ứng với $\mathbf{E}(\mathbf{u}) = \mathbf{0}$ và $\mathbf{E}(\mathbf{xu}) = \mathbf{0}$ 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 tất cả các biến giải thích x_1, x_2.

Diễn giải ý nghĩa của hồi quy đa biến

Với hàm hồi quy mẫu:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$$

- $\hat{\beta}_1$ và $\hat{\beta}_2$ là tác động riêng phần của biến x_1 và x_2 lên biến phụ thuộc, *trong điều kiện các yếu tố khác không đổi*.
- $\hat{y}$ là giá trị thích hợp (hoặc giá trị dự báo) của biến phụ thuộc với điều kiện x_1 và x_2 cho trước.
- Phần dư là chênh lệch giữa giá trị thực tế và giá trị dự báo của biến phụ thuộc, $\hat{u} = y - \hat{y}$.

Ví dụ 1: Ước lượng các nhân tố ảnh hưởng đến điểm GPA

Sử dụng bộ dữ liệu GPA1.dta. Ước lượng mô hình điểm GPA học đại học *colGPA* với một và hai biến giải thích là điểm GPA cho giai đoạn học trung học *hsGPA* và điểm thành tích *ACT*.

```
. reg colGPA hsGPA
```

Source	SS	df	MS	Number of obs	=	141
				F(1, 139)	=	28.85
Model	3.33506006	1	3.33506006	Prob > F	=	0.0000
Residual	16.0710394	139	.115618988	R-squared	=	0.1719
				Adj R-squared	=	0.1659
Total	19.4060994	140	.138614996	Root MSE	=	.34003

colGPA	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsGPA	.4824346	.0898258	5.37	0.000	.304833	.6600362
_cons	1.415434	.3069376	4.61	0.000	.8085635	2.022304

```
.
. reg colGPA hsGPA ACT
```

Source	SS	df	MS	Number of obs	=	141
				F(2, 138)	=	14.78
Model	3.42365506	2	1.71182753	Prob > F	=	0.0000
Residual	15.9824444	138	.115814814	R-squared	=	0.1764
				Adj R-squared	=	0.1645
Total	19.4060994	140	.138614996	Root MSE	=	.34032

colGPA	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
hsGPA	.4534559	.0958129	4.73	0.000	.2640047	.6429071
ACT	.009426	.0107772	0.87	0.383	-.0118838	.0307358
_cons	1.286328	.3408221	3.77	0.000	.612419	1.960237

Ví dụ 2: Ước lượng mô hình tiền lương

Sử dụng bộ dữ liệu WAGE1.dta. Ước lượng tác động của số năm đi học *educ*, số năm thâm niên *exper*, số năm kinh nghiệm làm việc hiện tại *tenure* lên tiền lương *lwage*.

```
. reg lwage educ exper tenure
```

Source	SS	df	MS	Number of obs	=	526
				F(3, 522)	=	80.39
Model	46.8741776	3	15.6247259	Prob > F	=	0.0000
Residual	101.455574	522	.194359337	R-squared	=	0.3160
				Adj R-squared	=	0.3121
Total	148.329751	525	.28253286	Root MSE	=	.44086

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.092029	.0073299	12.56	0.000	.0776292	.1064288
exper	.0041211	.0017233	2.39	0.017	.0007357	.0075065
tenure	.0220672	.0030936	7.13	0.000	.0159897	.0281448
_cons	.2843595	.1041904	2.73	0.007	.0796756	.4890435

Ví dụ 3: Ước lượng mô hình tiền lương với tác động phi tuyến của giáo dục

Cũng với mô hình trên, nhưng giả sử số năm đi học có tác động phi tuyến (bình phương) lên thu nhập.

```
. gen educsq = educ^2

. reg lwage educ educsq exper tenure
```

Source	SS	df	MS	Number of obs	=	526
				F(4, 521)	=	65.87
Model	49.8213265	4	12.4553316	Prob > F	=	0.0000
Residual	98.5084249	521	.189075672	R-squared	=	0.3359
				Adj R-squared	=	0.3308
Total	148.329751	525	.28253286	Root MSE	=	.43483

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	-.0316271	.0321443	-0.98	0.326	-.0947755	.0315213
educsq	.0052535	.0013306	3.95	0.000	.0026394	.0078676
exper	.0037126	.0017028	2.18	0.030	.0003673	.0070579
tenure	.0216263	.0030534	7.08	0.000	.0156279	.0276247
_cons	.9776996	.2034733	4.81	0.000	.5779708	1.377428

Tác động biên của học thêm một năm lên thu nhập là (%):

$$\frac{\Delta y}{\Delta educ} \approx \beta_1 + 2\beta_2 \times educ$$

Những vấn đề cần lưu ý với hồi quy đa biến

- Chọn biến số đưa vào mô hình theo tiêu chí gì?
- Hậu quả gì nếu đưa biến không liên quan vào mô hình?
- Hậu quả gì nếu bỏ sót biến quan trọng trong mô hình?
- Hậu quả gì nếu đưa các biến tương quan với nhau vào cùng một mô hình?

Chọn biến đưa vào mô hình

- R^2 luôn luôn tăng khi đưa thêm biến vào mô hình, kể cả những biến không liên quan.
- Do đó, để tránh lạm dụng đưa quá nhiều biến vào mô hình, sử dụng R^2-điều chỉnh:

$$R^2_{adj} = 1 - \frac{(1 - R^2)(n - 1)}{n - k - 1}$$

với n và k là số quan sát và số biến giải thích trong mô hình.

- R^2_{adj} có thể tăng hoặc giảm khi đưa biến mới vào mô hình.

Ví dụ 4: Ước lượng mô hình tiền lương với nhiều biến giải thích

Sử dụng bộ dữ liệu WAGE1.dta. Ước lượng mô hình lần lượt với các biến giải thích là (1) số năm đi học, số năm đi học bình phương, kinh nghiệm; (2) thêm biến màu da, giới tính, và hôn nhân; (3) thêm biến số người phụ thuộc. Kiểm tra R^2 và R^2_{adj} thay đổi như thế nào khi thêm biến.

Regression Results

	Model 1 b/se	Model 2 b/se	Model 3 b/se
educ	-0.0340	-0.0136	-0.0134
	(0.0336)	(0.0314)	(0.0315)
educsq	0.0056***	0.0043**	0.0043**
	(0.0014)	(0.0013)	(0.0013)
exper	0.0098***	0.0072***	0.0072***
	(0.0015)	(0.0015)	(0.0016)
nonwhite		-0.0089	-0.0094
		(0.0609)	(0.0612)
female		-0.3097***	-0.3098***
		(0.0377)	(0.0378)
married		0.1404***	0.1394**
		(0.0409)	(0.0422)
numdep			0.0016
			(0.0156)
Constant	0.9571***	1.0261***	1.0219***
	(0.2128)	(0.1987)	(0.2031)
Obs	526.0000	526.0000	526.0000
R2	0.2719	0.3792	0.3793
R2-adj	0.2678	0.3721	0.3709
df(r)	522.0000	519.0000	518.0000
SSR	107.9936	92.0757	92.0739

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

Sử dụng hệ số phóng đại phương sai (Variance Inflation Factor) để lựa chọn biến

Hệ số *VIF* dùng để kiểm tra mức độ tương quan của một biến giải thích với các biến còn lại. Biến số càng ít tương quan với các biến khác càng tốt.

- Hồi quy lần lượt biến x^j lên các biến còn lại. Tính hệ số thích hợp R_j^2.
- Tính hệ số *VIF*:

$$VIF_j = \frac{1}{1 - R_j^2}$$

- Nếu R_j^2 lớn chứng tỏ biến x^j tương quan nhiều với các biến giải thích khác.
- Quy tắc chung: Loại biến có $VIF > 10$

Ví dụ 5: Chọn biến sử dụng hệ số *VIF*

Ước lượng lại ví dụ (4), tính VIF và giải thích.

- Nếu có một biến cộng tuyến hoàn hảo trong mô hình thì VIF của biến đó là bao nhiêu?

Tình huống 1: Đưa biến không liên quan vào mô hình

Giả sử mô hình chuẩn là $\mathbf{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 \mathbf{x_1}$, nhưng chúng ta ước lượng mô hình $\mathbf{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x_1} + \hat{\beta}_2 \mathbf{x_2}$.

- Mối quan hệ giữa $\tilde{\beta}_1$ và $\hat{\beta}_1$ là:

$$\tilde{\beta}_1 = \hat{\beta}_1 + \hat{\beta}_2 \tilde{\sigma_1}$$

  Với $\tilde{\sigma_1}$ là hệ số góc của hàm hồi quy của biến x_2 lên biến x_1.

- Độ chệch của một ước lượng **(được tính bằng giá trị ước lượng được trừ đi giá trị thực)**,

$$Bias(\beta_1) = \hat{\beta}_1 - \tilde{\beta}_1 = -\hat{\beta}_2 \tilde{\sigma_1}$$

- Nếu biến x_2 không quan trọng, $\hat{\beta}_2 = 0$, do đó $\hat{\beta}_1$ vẫn không chệch, $\hat{\beta}_1 = \tilde{\beta}_1$.

- Tuy nhiên phương sai của các ước lượng sẽ thay đổi $\Rightarrow$ Ảnh hưởng đến việc xây dựng khoảng tin cậy hay độ chính xác của ước lượng!

Tình huống 2: Thiếu biến quan trọng trong mô hình

Giả sử mô hình chuẩn là $\mathbf{Y} = \tilde{\beta}_0 + \tilde{\beta}_1 \mathbf{x_1} + \tilde{\beta}_2 \mathbf{x_2}$, nhưng chúng ta ước lượng mô hình $\mathbf{Y} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x_1}$.

- Với công thức:

$$\hat{\beta}_1 = \tilde{\beta}_1 + \tilde{\beta}_2 \tilde{\sigma}_1$$

- Mức độ chệch của ước lượng khi xảy ra vấn đề thiếu biến quan trọng là:

$$Bias(\beta_1) = \hat{\beta}_1 - \tilde{\beta}_1 = \tilde{\beta}_2 \tilde{\sigma}_1$$

Đánh giá hướng chệch trong mô hình thiếu biến quan trọng

- Nếu $\beta_2 = 0$ (nghĩa là biến x_2 không phải là biến quan trọng) thì ước lượng của β_1 không chệch.
- Nếu $\tilde{\sigma_1} = 0$ (nghĩa là x_1 và x_2 không tương quan) thì β_1 cũng không chệch.
- Nếu không phải 2 trường hợp trên, β_1 bị chệch, với hướng và mức độ chệch tùy thuộc vào giá trị của β_2 và $\tilde{\sigma_1}$.

	$\text{Corr}(x_1, x_2) > 0$	$\text{Corr}(x_1, x_2) < 0$
$\beta_2 > 0$	Positive bias	Negative bias
$\beta_2 < 0$	Negative bias	Positive bias

Nếu nghi ngờ mô hình thiếu biến thì khi giải thích kết quả phải nhận định hướng chệch của tác động!

Ví dụ 6: Ước lượng phương trình tiền lương theo số năm đi học

Sử dụng bộ dữ liệu WAGE1.dta

- Giả sử mô hình chuẩn có hai biến là giáo dục (educ) và tố chất cá nhân (ability):

$$log(wage) = \beta_0 + \beta_1 educ + \beta_2 ability + u$$

- Chúng ta không quan sát được tố chất cá nhân, do đó chúng ta chỉ ước lượng được mô hình:

$$log(wage) = \beta_0 + \beta_1 educ + u$$

- Ước lượng của β_1 có bị chệch không? và chệch theo hướng nào?

```
. reg lwage educ
```

Source	SS	df	MS	Number of obs	=	526
				F(1, 524)	=	119.58
Model	27.5606288	1	27.5606288	Prob > F	=	0.0000
Residual	120.769123	524	.230475425	R-squared	=	0.1858
				Adj R-squared	=	0.1843
Total	148.329751	525	.28253286	Root MSE	=	.48008

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
educ	.0827444	.0075667	10.94	0.000	.0678796	.0976091
_cons	.5837727	.0973358	6.00	0.000	.3925563	.7749891

Tỷ suất thu nhập của một năm đi học ước lượng được là 8.3%. Ước lượng này có bị chệch không?

- Tham khảo nghiên cứu về tỷ suất thu thập của việc đi học từ các các quốc gia trên thế giới.
- Tìm cách điều chỉnh hàm hồi quy để xử lý vấn đề thiếu biến, và nhận định kết quả thay đổi như thế nào.

Mô hình thiếu biến quan trọng trong trường hợp tổng quát

- Mô hình tổng quát với nhiều biến giải thích:

$$Y = \beta_0 + \beta_1 x^1 + \beta_2 x^2 + ... + \beta_k x^k + u$$

- Nếu thiếu một biến quan trọng nào đó, tất cả các ước lượng $\hat{\beta}$ đều bị chệch.
- Xác định hướng chệch khó hơn nhiều do tương quan giữa các biến giải thích với biến bị thiếu, và giữa các biến giải thích với nhau.

Tóm tắt các giả định đối với hồi quy đa biến

Tương tự như các điều kiện của hồi quy đơn biến:

1. Tuyến tính theo tham số.
2. Chọn mẫu ngẫu nhiên.
3. **Không có cộng tuyến hoàn hảo.**
4. Trung bình có điều kiện của sai số bằng 0:

$$E(u|x^1, ..., x^k) = 0$$

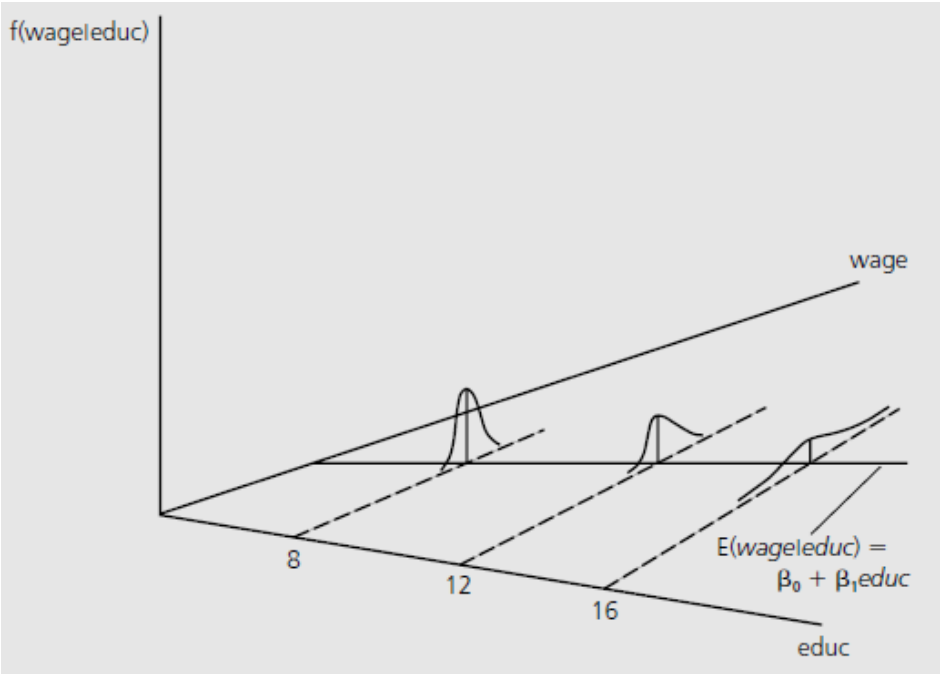
$\Rightarrow$ Ước lượng OLS của các tham số β là không chệch.

$$E(\hat{\beta}) = \beta$$

Giả định phương sai của sai số không đổi (homoskedasticity)

5. Với các giá trị của các biến giải thích cho trước, phương sai của sai số là một hằng số:

$$Var(u|x_1, ..., x_k) = \sigma^2$$



Đặc tính của ước lượng OLS

Với các giả định 1-5, ước lượng của OLS là ước lượng tuyến tính, không chệch, và hiệu quả nhất (Best Linear Unbiased Estimator - BLUE):

- Trong tất cả các ước lượng tuyến tính, OLS có phương sai của ước lượng là nhỏ nhất.
- Không chệch.