Chương 1
Giới thiệu mô hình hồi quy tuyến tính

(Gujarati: Econometrics by example, 2011)[1].

Người dịch và diễn giải: Phùng Thanh Bình

http://vnp.edu.vn/

Như đã lưu ý ở phần Lời tựa, một trong những công cụ quan trọng của kinh tế lượng là **mô hình hồi quy tuyến tính** (LRM). Trong chương này, chúng ta thảo luận bản chất tổng quát của mô hình hồi quy tuyến tính và cung cấp kiến thức nền tảng sẽ được dùng để minh họa nhiều ví dụ khác nhau trong cuốn sách. Chúng ta không chứng minh, vì bạn có thể tìm hiểu những chứng minh ấy ở nhiều giáo trình kinh tế lượng[2].

1.1 Mô hình hồi quy tuyến tính

Gujarati bắt đầu bằng mô hình hồi quy bội (multiple regression model, dạng mô hình hồi quy tổng thể - population regression model) với k -1 biến giải thích có dạng như sau:

$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + ... + B_kX_{ki} + u_i \quad (1.1)$$

Với Y là biến phụ thuộc (dependent variable) hoặc còn gọi là **regressand**; X là các biến giải thích (explanatory variables) hoặc còn có những tên gọi khác như predictors, covariates, hoặc **regressors**; u là hạng nhiễu ngẫu nhiên (random hay stochastic error term); và i là ký hiệu cho quan sát thứ i trong tổng thể. [Diễn giải: Hàm ý dữ liệu chéo, và với mô hình tổng thể thì chúng ta không thể biết được có bao nhiêu quan sát]. Đôi khi để đơn giản hóa, phương trình (1.1) còn được viết ở dạng rút gọn như sau:

$$Y_i = BX + u_i \quad (1.2)$$

với BX là $B_1 + B_2X_{2i} + B_3X_{3i} + ... + B_kX_{ki}$.

Phương trình (1.1) hoặc hình thức rút gọn của nó là phương trình (1.2) được gọi là **mô hình tổng thể** (population model) hoặc **mô hình thực** (true model). Mô hình này gồm hai thành phần: (1) thành phần **tất định** (deterministic component), BX, và (2) thành phần **phi hệ thống** (nonsystematic component) hoặc thành phần **ngẫu nhiên** (random component), u_i. BX có thể được giải thích như **trung bình có điều kiện** (conditional

[1] Hiện nay đã có ấn bản mới (lần 2, năm 2015). Dữ liệu của phiên bản 2011: https://www.macmillanihe.com/companion/Gujarati-Econometrics-By-Example/student-zone/

[2] Ví dụ, xem Damodar N. Gujarati and Dawn C. Porter, *Basic Econometrics*, 5th edn, McGraw-Hill New York, 2009 (từ đây về sau, gọi là sách của Gujarati/Porter); Jeffrey M. Wooldridge, *Introductory Econometrics: A modern Approach*, 4th edn, South-Western, USA, 2009; James H. Stock and Mark W. Watson, *Introduction to Econometrics*, 2dn edn, Pearson, Boston, 2007; and R Carter Hill, William E. Griffiths and Guay C. Lim, *Principles of Econometrics*, 3rd edn, John Wiley & Sons, New York, 2008.

mean) của Y_i, tức là $E(Y_i|X)$: giá trị trung bình của Y khi X được cho trước là bao nhiêu[3]. Vì thế, phương trình (1.2) phát biểu rằng một giá trị Y_i của một cá nhân i bất kỳ sẽ bằng giá trị trung bình của tổng thể trong đó người này là một thành viên cộng hoặc trừ một con số ngẫu nhiên. Khái niệm tổng thể (population) có nghĩa là tổng quát (general) và đề cập đến một thực thể được xác định rõ (ví dụ con người, các công ty, các thành phố, các quốc gia, …) là trọng tâm của một phân tích kinh tế lượng hoặc thống kê.

Ví dụ, giả sử Y là chi tiêu cho thực phẩm của gia đình (food expenditure), X là thu nhập của gia đình (income), thì phương trình (1.2) cho biết rằng chi tiêu cho lương thực của một gia đình riêng lẻ bằng với chi tiêu cho lương thực trung bình của tất cả các gia đình với có cùng mức thu nhập, cộng hoặc trừ một thành phần ngẫu nhiên, thành phần này có thể khác nhau giữa các gia đình khác nhau và có thể phụ thuộc vào nhiều yếu tố.

[Diễn giải: Với nhóm có mức thu nhập thấp (ví dụ X = 1000) thì sẽ có rất nhiều mức chi tiêu khác nhau tùy vào hành vi chi tiêu của từng gia đình trong nhóm này (có nhà xài hết 1000, có nhà xài ít hơn 1000, có nhà xài nhiều hơn 1000 cho thực phẩm). Dĩ nhiên, chúng ta không thể biết có bao nhiêu gia đình trong nhóm này, nhưng chúng ta kỳ vọng sẽ tính được mức chi tiêu trung bình (mean or average expenditure) của toàn bộ các gia đình thuộc nhóm thu nhập này nếu có thể thu thập được dữ liệu. Tương tự, với nhóm có mức thu nhập cao (ví dụ X = 10.000) thì cũng sẽ có rất nhiều mức chi tiêu khác nhau tùy vào hành vi chi tiêu của từng gia đình trong nhóm này (có nhà xài hết 10.000, có nhà xài ít hơn 10.000, có nhà xài nhiều hơn 10.000). Nếu chúng ta kỳ vọng thu nhập tăng thì mức chi tiêu trung bình cho thực phẩm cũng tăng, thì hệ số hồi quy $B > 0$. Giả sử nhóm thu nhập thấp, giá trị trung bình của chi tiêu cho thực phẩm là 900, thì nếu một gia đình bất kỳ i nào đó có mức chi tiêu $Y_i = 700$, thì $u_i = -200$; và nếu một gia đình bất kỳ i nào khác có mức chi tiêu $Y_i = 1100$, thì $u_i = 200$; ... Như vậy, một quan sát Y_i bất kỳ trong nhóm thu nhập thấp, thì chi tiêu cho thực phẩm sẽ bằng mức chi tiêu trung bình của tất cả các gia đình trong nhóm thu nhập thấp này cộng hoặc trừ một thành phần ngẫu nhiên. Tất nhiên, thành phần ngẫu nhiên của mỗi gia đình riêng lẻ sẽ khác nhau tùy thuộc vào rất nhiều yếu tố. Muốn biết các yếu tố đó là gì thì chúng ta phải tìm hiểu lý thuyết về hành vi người tiêu dùng (consumer behavior theory) và các bài nghiên cứu trước đây về vấn đề này để xác định danh sách các biến thích hợp. Và, nếu cộng tất cả các thành phần ngẫu nhiên trong cùng nhóm thu nhập thấp, chúng ta chắc chắn sẽ có $\Sigma u_i = 0$, và điều này cũng đúng cho mọi nhóm thu nhập khác, ví dụ X = 2000, X = 3000, ..., X = 30.000.

Các hạng nhiễu u_i trong cùng một nhóm là khác nhau, nên u_i được xem như một biến ngẫu nhiên (random variable). Và một biến ngẫu nhiên thì phải theo một phân phối xác suất (probability distribution) nào đó. Đúng không? Ở đây, nếu Y là một biến liên tục, thì người ta kỳ vọng u_i theo phân phối chuẩn (normal distribution), với trung bình = 0 ($=\Sigma u_i/n$, với n là số gia đình trong một nhóm thu nhập nhất định) và phương sai không đổi (homoscedasticity), ký hiệu là σ^2. Tại sao người ta kỳ vọng u_i có phân phối chuẩn? Một biến ngẫu nhiên có phân phối chuẩn khi nào? Khi giá trị của biến đó phụ thuộc vào

[3] Nhớ lại thống kê căn bản rằng giá trị trung bình không có điều kiện của Y_i được ký hiệu là E(Y), nhưng trung bình có điều kiện, điều kiện theo X cho trước được ký hiệu là E(Y|X).

rất nhiều yếu tố, nhưng không có yếu tố nào là quan trọng nhất. Ví dụ, chi tiêu không thể là một biến có phân phối chuẩn, bởi vì chi tiêu phụ thuộc vào rất nhiều yếu tố khác nhau, nhưng ai cũng biết thu nhập là một yếu tố mang tính quyết định của chi tiêu. Cân nặng của một đứa trẻ 5 tuổi là một biến ngẫu nhiên có phân phối chuẩn bởi vì cân nặng phụ thuộc vào rất nhiều yếu tố, nhưng không biết yếu tố nào là quan trọng nhất. Trở lại với hạng nhiễu u_i. Giả sử mô hình (1.1) được xác định đúng (well-specified model), nghĩa là k-1 biến giải thích là đầy đủ, không thừa biến không quan trọng cũng không bỏ sót biến quan trọng nào khác; dạng hàm (mối quan hệ hàm số giữa Y và từng biến giải thích) được xác định đúng; và các biến Y và Xs được đo lường chính xác. Phương trình (1.1) có thể được triển khai như sau:

$$Y_i = B_1 + B_2X_{2i} + B_3X_{3i} + ... + B_kX_{ki} + \alpha_1z_1 + \alpha_1z_1 + ... \alpha_\infty z_\infty \quad (*)$$

Như vậy, u_i là một biến gộp (composite variable) đại diện cho tất cà các yếu tố Zs có ảnh hưởng lên Y_i nhưng từng yếu tố Z riêng lẻ là không có ảnh hưởng đáng kể. Nếu X_3 là một biến quan trọng mà vô tình bị bỏ sót (có thể do lười tham khảo tài liệu hoặc không có dữ liệu) thì X_3 sẽ "gia nhập" nhóm Zs và nằm trong u_i. Nếu như thế, giá trị của u_i sẽ phụ thuộc vào rất nhiều yếu tố, nhưng X_3 là yếu tố mang tính quyết định, và u_i sẽ không còn phân phối chuẩn nữa. Tóm lại, người ta giả sử u_i có phân phối chuẩn là hợp lý. Và giả định này rất quan trọng trong việc suy diễn thống kê, nhất là kiểm định giả thuyết về các hệ số hồi quy từ mẫu.

Nghe giải thích tiếp nhé. Một biến ngẫu nhiên theo phân phối chuẩn thì cần hai thông tin: trung bình μ và phương sai σ^2. Trung bình của u_i đã được nói ở trên, bằng 0. Còn phương sai là gì? Giả sử chỉ xét hai nhóm thu nhập thôi (thấp, X_i = 1000; và cao, X_j = 10.000). Chi tiêu cho thực phẩm của từng gia đình trong nhóm thu nhập thấp là khác nhau, và chênh lệch (difference) của từng gia đình so với mức trung bình của nhóm thu nhập thấp là u_i cũng sẽ khác nhau [tức là $(Y_i - E(Y_i|X_i)]$. Giả sử độ lệch chuẩn của u_i (standard deviation) của nhóm thu nhập thấp là σ_i = 100. Tính sao? Lấy từng chênh lệch bình phương lên, cộng lại, rồi chia cho số quan sát trong nhóm này (dĩ nhiên mình đang giả sử là có thể biết được có bao nhiêu gia đình trong nhóm này). Tương tự, chi tiêu cho thực phẩm của từng gia đình trong nhóm thu nhập cao cũng khác nhau, và chênh lệch của từng gia đình so với mức trung bình của nhóm thu nhập cao là u_j cũng sẽ khác nhau [tức là $(Y_j - E(Y_j|X_j)]$. Và, người ta cũng giả định rằng độ lệch chuẩn u_j của nhóm thu nhập cao cũng là σ_j = 100. Nghĩa là, $\sigma^2_i = \sigma^2_j = \sigma^2$ (tức phương sai đồng nhất). Đây là điều bất hợp lý, nhưng bước đầu chúng ta cần giả định như thế để giúp việc suy diễn thống kê (statistical inference) của các hệ số hồi quy được dễ dàng. Sau này, nếu σ của nhóm có thu nhập thấp $\neq$ σ của nhóm có thu nhập cao (gọi là phương sai thay đổi, heteroscedasticity) thì chúng ta sẽ có một số cách khắc phục].

Trở lại phương trình (1.1), B_1 là hệ số cắt hay tung độ gốc (intercept), B_2, B_3, ..., B_k là các hệ số độ dốc (slope coefficients). Nói chung, các hệ số này được gọi là **hệ số hồi quy** hay **tham số hồi quy tổng thể** (regression coefficients or parameters). Trong phân tích hồi quy, mục tiêu chính yếu của chúng ta là nhằm giải thích hành vi trung bình (mean or average behavior) của Y theo các biến giải thích. Nghĩa là, trung bình của Y (mean Y) sẽ

phản ứng theo những thay đổi trong các giá trị của các biến X như thế nào. Một giá trị Y riêng lẻ (individual Y value) sẽ xoay quanh giá trị trung bình của nó.

Cần nhấn mạnh rằng mối quan hệ nhân quả (causal relationship) giữa Y và các X, nếu có, nên được dựa trên lý thuyết thích hợp (relevant theory). [Diễn giải: Nghĩa là để xác định biến nào nên được đưa vào mô hình, dạng hàm giữa chúng với biến Y, và dấu kỳ vọng âm hay dương, … đều phải dựa vào lược khảo lý thuyết (literature review). Tức là phải đọc và đọc thật nhiều].

Mỗi hệ số B_2, B_3, ..., B_k là hệ số hồi quy riêng (partial coefficient): Hệ số hồi quy riêng đo lường mức độ thay đổi trong giá trị trung bình của Y theo một sự thay đổi đơn vị của biến giải thích khi giữ nguyên giá trị của các biến giải thích khác. [Diễn giải: Việc giải thích chính xác ý nghĩa hệ số hồi quy tùy vào dạng hàm (functional form). Vấn đề này sẽ được bàn ở chương 2 của cuốn sách này. Còn để hiểu 'hệ số hồi quy riêng' là gì thì nên tham khảo phần Ôn tập # 2 trong Tóm lược kinh tế lượng căn bản của Phùng Thanh Bình, sau đây gọi là Tóm lược kinh tế lượng căn bản]. Có bao nhiêu biến giải thích trong mô hình tùy vào bản chất của vấn đề đang nghiên cứu và sẽ khác nhau giữa các vấn đề nghiên cứu.

Hạng nhiễu u_i là một biến gộp (catchall) của tất cả các biến không thể được đưa vào mô hình vì nhiều lý do. Tuy nhiên, ảnh hưởng trung bình của tất cả các biến này lên biến phụ thuộc được giả định là không đáng kể.

Bản chất của biến phụ thuộc Y

Y nói chung được giả định là một biến ngẫu nhiên, và có thể được đo lường bằng một trong bốn thước đo sau đây: thang đo tỷ lệ, thang đo khoảng, thang đo thứ bậc, và thang đo danh nghĩa. [Diễn giải: Xem chương 1, Thống kê trong kinh tế và kinh doanh (sách dịch của Khoa toán – thống kê, UEH), sau đây gọi là Giáo trình thống kê UEH), hoặc chương 1, Kinh tế lượng căn bản của Phùng Thanh Bình, sau đây gọi là Kinh tế lượng căn bản), hoặc chương 1, Giáo trình kinh tế lượng căn bản của Wooldridge, ấn bản lần 5 do Khoa toán – thống kê, UEH dịch, sau đây gọi là Giáo trình kinh tế lượng UEH].

- Thang đo tỷ lệ (ratio scale): Một thang đo tỷ lệ có 3 tính chất: (1) tỷ số của hai biến, (2) khoảng cách giữa hai biến, và (3) xếp hạng các biến. Với thang đo tỷ lệ, ví dụ Y có hai giá trị, Y_1 và Y_2 thì tỷ số Y_1/Y_2 và khoảng cách $(Y_2 - Y_1)$ là các đại lượng có ý nghĩa (meaningful quantities); và có thể so sánh hoặc xếp thứ tự như $Y_2 \leq Y_1$ hoặc $Y_2 \geq Y_1$. Hầu hết các biến kinh tế thuộc loại thang đo này. Vì thế, chúng ta có thể nói về GDP năm nay lớn hơn hay nhỏ hơn năm trước, hoặc tỷ số GDP của năm nay so với năm trước lớn hơn hay nhỏ hơn một.
- Thang đo khoảng (interval scale): Thang đo khoảng không thỏa mãn tính chất đầu tiên của các biến có thang đo tỷ lệ. Ví dụ, khoảng cách giữa hai giai đoạn như 1997 và 2017 thì có ý nghĩa, nhưng tỷ số 2017/1997 thì không có ý nghĩa.
- Thang đo thứ bậc (ordinal scale): Các biến chỉ thỏa mãn tính chất xếp hạng của thang đo tỷ lệ, chứ việc lập tỷ số hay tính khoảng cách giữa hai giá trị không có ý nghĩa. Ví dụ, xếp hạng điểm A, B, C, D; phân loại thu nhập thấp, trung bình và

cao là thang đo thứ bậc, nhưng đại lượng A/B hay thu nhập cao - thu nhập thấp không có ý nghĩa.

- Thang đo danh nghĩa (nominal scale): Các biến thuộc nhóm này không thỏa mãn bất kỳ tính chất nào của các biến theo thang đo tỷ lệ. Các biến như giới tính (gender), tình trạng hôn nhân (martial status), tôn giáo (religion), có tham gia lực lượng lao động hay không (labor force participation), có sở hữu nhà hay không (house ownership), nghèo hay không nghèo (poverty), ... là các biến theo thang đo danh nghĩa. Các biến như thế thường gọi là **biến giả** (dummy variables) hoặc **biến phân loại** (categorical variables). Các biến này thường được lượng hóa bằng 1 và 0; trong đó, 1 chỉ sự hiện diện của thuộc tính và 0 chỉ không có sự hiện diện của thuộc tính.

Mặc dù hầu hết các biến kinh tế được đo theo thang đo tỷ lệ hoặc thang đo khoảng, nhưng có một số trường hợp cũng sử dụng hai thang đo thứ bậc hoặc thang đo định danh. Điều đó đòi hỏi các kỹ thuật kinh tế lượng chuyên biệt khác với mô hình LRM chuẩn [Diễn giải: Phương pháp hồi quy OLS không sử dụng được mà phải dùng phương pháp hợp lý tối đa, ML. Phương pháp này có trình bày ở phần Phụ lục cuối chương này].

[Diễn giải: Trong phần kinh tế lượng căn bản, mô hình hồi quy tuyến tính được ước lượng theo phương pháp OLS thì biến Y chỉ ở dạng thang đo tỷ lệ hoặc thang đo khoảng (gọi chung là biến ngẫu nhiên liên tục). Do hạng nhiễu u_i là phản chiếu của Y_i, nên Y dạng thang đo gì thì u cũng có thang đo. Phân phối xác suất của u_i tùy thuộc vào phân phối xác suất của Y_i. Chính vì thế mà chúng ta cần nắm rõ bản chất của các loại phân phối xác suất đã được trình bày ở Giáo trình thống kê UEH: ít nhất là các phân phối nhị thức, phân phối Poisson, và phân phối chuẩn].

Bản chất của các biến giải thích X

Các biến giải thích có thể được đo theo bất kỳ một trong bốn thang đo vừa nêu trên, mặc dù trong nhiều ứng dụng thực tế thì các biến giải thích được đo theo thang đo tỷ số và thang đo khoảng. Trong **mô hình hồi quy tuyến tính cổ điển** (CLRM - classical linear regression model), các biến giải thích được giả định là phi ngẫu nhiên (nonrandom); nghĩa là, các giá trị của biến giải thích được giữ cố định khi lấy mẫu lặp đi lặp lại (repeated sampling). [Diễn giải: Xem lại chương 5 ở Kinh tế lượng căn bản]. Chính vì thế mà phân tích hồi quy là có điều kiện (conditional), nghĩa là tính giá trị trung bình của Y khi cho trước các giá trị của biến giải thích (conditional on the given value of the regressors).

Chúng ta có thể cho phép các biến giải thích là ngẫu nhiên giống như biến Y, nhưng trong trường hợp đó cần lưu ý cách giải thích các kết quả hồi quy. Chúng ta sẽ minh họa điểm này trong Chương 7 và xem xét kỹ hơn ở Chương 19 của cuốn sách này.

Bản chất của hạng nhiễu ngẫu nhiên u

Như đã nói ở trên, hạng nhiễu ngẫu nhiên đại diện cho tất cả các biến không được đưa vào mô hình vì những lý do như không có sẵn dữ liệu [lack of data availability [Diễn giải: Ví dụ những yếu tố thuộc về tâm lý (psychological), tâm linh (spiritual) có ảnh hưởng đến chi tiêu thực phẩm, nhưng khó mà thu thập được dữ liệu khi tiến hành điều tra hộ

gia đình (household survey)], các lỗi đo lường trong dữ liệu [errors of measurement in the data [Diễn giải: Ví dụ năng lực (ability) của chủ hộ có ảnh hưởng đến năng suất sản xuất (productivity), nhưng nếu đo năng lực bằng các biến đại diện (proxy variables) như số năm đi học (schooling years), số năm kinh nghiệm (tenure), hay có tham gia các khóa tập huấn (participation in training courses), ... thì cũng đâu thể phản ánh hết năng lực của họ; hoặc rất khó đo lường chính xác thu nhập của cá nhân hay hộ gia đình nếu không thể tiếp cận được tài khoản ngân hàng của họ hoặc nếu được thì những khoản thu không qua ngân hàng thì làm sao mình biết được, còn nếu hỏi trực tiếp thì chắc gì họ móc ruột ra nói thiệt để mình ghi chép, ... nên có khi người ta dùng biến tổng chi tiêu (total expenditure variable) làm biến đại diện, và như thế biến chi tiêu chỉ là một đại diện xấp xỉ đúng (approximately correct) của thu nhập mà thôi], hoặc bản chất ngẫu nhiên nội tại của hành vi con người (intrinsic randomness of human behavior). Và cho dù nguồn tạo ra hạng nhiễu u là gì đi nữa, thì người ta giả định rằng ảnh hưởng trung bình của hạng nhiễu ngẫu nhiên lên Y là không đáng kể (whatever the source of the random term u, it is assumed that the average effect of the error term on the regressand is marginal at best).

Bản chất của các hệ số hồi quy Bs

Trong CLRM, các hệ số hồi quy (tổng thể), Bs, là những con số cố định (fixed numbers) và không ngẫu nhiên (not random), mặc dù mình không thể biết giá trị thực của các Bs là bao nhiêu. [Diễn giải: Giả sử chúng ta có thể thu thập đầy đủ và chính xác các thông tin về chi tiêu cho thực phẩm, thu nhập, học vấn chủ hộ (household head), nghề nghiệp, sở thích ăn uống, mối quan hệ bạn bè (social networking), ... của tất cả mọi gia đình ở thành phố Cà Mau; thì chúng ta sẽ có được giá trị của từng hệ số B ở phương trình (1.1), và mỗi hệ số B là duy nhất. Nhưng điều này là chắc chắn là bất khả thi]. Mục đích của phân tích hồi quy (regression analysis) là ƯỚC LƯỢNG (estimate) các giá trị B dựa trên dữ liệu mẫu (on the basis of sample data), và các ƯỚC LƯỢNG (estimators) bs của Bs là các biến ngẫu nhiên vì giá trị của từng b sẽ thay đổi khi mẫu thay đổi (vary from sample to sample). [Diễn giải: Xem chương 5 ở Kinh tế lượng căn bản hoặc Ôn tập # 1 trong Tóm lược kinh tế lượng căn bản để biết tính chất của các hệ số hồi quy OLS; tại sao từng hệ số bs có phân phối chuẩn, và tại sao khi kiểm định ý nghĩa của từng hệ số bs chúng ta sử dụng thống kê t chứ không phải thống kê z]. Một nhánh của thống kê được biết là thống kê Bayes (Bayesian statistics) xử lý các hệ số hồi quy (tổng thể) là ngẫu nhiên. Trong cuốn sách này, chúng ta sẽ không theo đuổi cách tiếp cận Bayes đối với các mô hình hồi quy tuyến tính[4].

Ý nghĩa của hồi quy tuyến tính

Đối với mục đích của chúng ta, thuật ngữ tuyến tính (linear) trong mô hình hồi quy tuyến tính nghĩa là **tuyến tính ở các hệ số hồi quy** (linearity in the regression coefficients), Bs, và không phải tuyến tính ở các biến Y và X. [Diễn giải: Nghĩa là Y và X có thể ở các dạng phi tuyến (nonlinear)]. Ví dụ, các biến Y và X có thể ở dạng logarít tự nhiên như $\ln(X_2)$

[4] Ví dụ tham khảo Gary Koop, Bayesian Econometrics, John Wiley & Sons, West Sussex, England, 2003.

(natural logarithm)[5], dạng tỷ lệ nghịch như $1/X_3$ (reciprocal), hoặc dạng bình phương như X_2^2 (square), lập phương như X_2^3 (cube), hay bất kỳ dạng nào khác.

Tuyến tính ở các hệ số Bs, nghĩa là Bs không ở dạng bình phương như B_2^2, tỷ lệ B_2/B_3, hay $\ln(B_4)$. Có các trường hợp ở đó chúng ta phải xem xét các mô hình hồi quy không tuyến tính ở các hệ số hồi quy[6].

1.2 Bản chất và các nguồn dữ liệu

Để thực hiện phân tích hồi quy, chúng ta cần dữ liệu. Nói chung, có ba loại dữ sẵn có cho phân tích: (1) chuỗi thời gian (time series), (2) dữ liệu chéo (cross-sectional), và (3) dữ liệu bảng (panel data) (một loại đặc biệt của dữ liệu gộp, pooled data). [Diễn giải: Xem chương 1, Giáo trình thống kê UEH; chương 1, Kinh tế lượng căn bản; hoặc chương 1, Giáo trình kinh tế lượng UEH].

Dữ liệu chuỗi thời gian

Dữ liệu chuỗi thời gian là tập hợp các quan sát của một biến tại các thời gian khác nhau, như theo ngày [daily - như giá chứng khoán (stock prices), tỷ giá hối đoái (exchange rate), báo cáo thời tiết (weather reports)], theo tuần [weekly - như cung tiền (money supply), tiền lương (wage)], theo tháng [monthly - như tỷ lệ thất nghiệp (the unemployment rate), chỉ số giá tiêu dùng (the consumer price index)], theo quý [quarterly - như GDP, sản lượng công nghiệp (industrial production)], theo năm [annually - như GDP, ngân sách chính phủ (government budgets)], theo năm năm [quinquenially - như tổng điều tra công nghiệp (the census of manufactures)], theo mười năm [decennially - như tổng điều tra dân số (the census of population)]. Đôi khi dữ liệu được thu thập cả theo quý hoặc theo năm (ví dụ GDP). Dữ liệu được gọi là có **tần suất cao** (high-frequency) được thu thập qua một giai đoạn cực kỳ ngắn. Trong **giao dịch chớp nhoáng** (flash trading) ở các thị trường chứng khoán và thị trường ngoại hối thì dữ liệu có tần suất cao như thế bây giờ càng trở nên phổ biến.

[Diễn giải: Hai vấn đề thường thấy với dữ liệu chuỗi thời gian là: (1) vì các quan sát liên tục (successive observations) theo thời gian có thể tương quan với nhau dẫn đến hiện tượng **tự tương quan** (autocorrelation, sẽ bàn ở chương 6 của cuốn sách này); và (2) các chuỗi thời gian trong kinh tế và tài chính (financial and economic time series) thường là các chuỗi **không dừng** (nonstationarity, sẽ bàn ở chương 13 của cuốn sách này) nên có thể dẫn đến hiện tượng hồi quy giả mạo (spurious regression). Hồi quy giả mạo hay còn gọi là hồi quy vô nghĩa (nonsense regression) là một hồi quy giữa hai chuỗi thời gian không dừng (non-stationary series) bất kỳ (ví dụ cung tiền của Fiji và GDP của Việt Nam) nhưng hệ số số hồi quy vẫn đúng và có ý nghĩa thống kê (statistically significant). Nhưng điều này không có hàm ý gì về khía cạnh chính sách kinh tế. Chẳng qua, mối tương quan này là do yếu tố xu thế (trend) chứa đựng trong hai chuỗi dữ liệu tạo ra mà thôi. Tuy nhiên, nếu hai chuỗi không dừng có một xu thế chung (common trend), thì chúng có

[5] Ngược lại, logarít cơ số 10 được gọi là log. Nhưng có một mối quan hệ cố định giữa các log tự nhiên và log thông thường, đó là $\ln_e X = 2.3026 \log_{10} X$.

[6] Vì đây là một chủ đề đặc biệt đòi hỏi kiến thức toán nâng cao (advanced mathematics), chúng ta sẽ không trình bày trong phạm vi cuốn sách này. Nhưng một thảo luận có thể tiếp cận, xem Gujarati/Porter, Chương 14.

thể đồng liên kết (đồng tích hợp, cointegration); và điều này giúp chúng ta xem xét cả mối quan hệ ngắn hạn và dài hại (short-term and long-term relationships). Đây là chủ đề đoạt giải Nobel kinh tế năm 2003. Chủ đề này sẽ được bàn ở chương 14 của cuốn sách này]. Các chuỗi thời gian thường được ký hiện là Y_t, X_t.

Dữ liệu chéo

Dữ liệu chéo là dữ liệu về một hoặc nhiều biến được thu thập tại cùng một thời điểm. Các ví dụ là tổng điều tra dân số được thực hiện bởi Cục dân số, lấy ý kiến cử tri được thực hiện bởi nhiều tổ chức bầu cử khác nhau, và nhiệt độ tại một thời điểm nhất định ở nhiều nơi khác nhau.

Giống dữ liệu chuỗi thời gian, dữ liệu chéo cũng có các vấn đề đặc thù, đặc biệt là vấn đề **phương sai thay đổi** (heteroscedasticity/heterogeneity). [Diễn giải: Hiện tượng này xảy ra là do ảnh hưởng quy mô (size or scale effect)]. Ví dụ, khi thu thập về tiền lương của một số công ty trong cùng một ngành công nghiệp (industry) tại cùng một thời điểm, hiện tượng phương sai thay đổi xảy ra bởi vì dữ liệu thu được từ nhiều công ty có quy mô rất khác nhau (nhỏ, vừa, và lớn) với những đặc điểm riêng của chúng. Vấn đề này sẽ được bàn tới ở chương 5 của cuốn sách này. Các biến dữ liệu chéo thường được ký hiện là Y_i, X_i.

Dữ liệu bảng

Dữ liệu bảng kết hợp các tính chất của cả dữ liệu chéo và dữ liệu chuỗi thời gian. Chẳng hạn, để ước lượng một hàm sản xuất (production function), chúng ta có thể sử dụng số liệu của một số công ty (khía cạnh chéo - the cross-sectional aspect) qua một giai đoạn thời gian (khía cạnh chuỗi thời gian - the time series aspect). Dữ liệu bảng cũng có một số thách thức khi phân tích hồi quy. Các quan sát của dữ liệu bảng sẽ được ký hiệu là Y_{it}, X_{it}.

Nguồn dữ liệu

[Diễn giải: Trong mục 1.2, Gujarati cũng đề cập đến các nguồn dữ liệu và chất lượng dữ liệu (sources of data and the quality of data). Tuy nhiên, nội dung không có gì khác so với chương 1, Giáo trình thống kê UEH và/hoặc chương 1, Kinh tế lượng căn bản], cho nên tôi xin phép bỏ qua cho đỡ mất thời gian].

Sự thành công của bất kỳ phân tích hồi quy nào phụ thuộc vào sự sẵn có của dữ liệu (availability of data). Dữ liệu có thể được thu thập bởi một cơ quan chính phủ (như Bộ ngân khố Hoa Kỳ), một cơ quan quốc tế (như Quỹ tiền tệ quốc tế - International Monetary Fund, IMF; hoặc Ngân hàng thế giới – World Bank), một tổ chức tư nhân (như Syandard & Poor's Corporation), hoặc các cá nhân hoặc các tổ chức tư nhân.

Ngày nay, nguồn dữ liệu tiềm năng nhất (most potent source of data) là từ Internet. Mọi thứ bạn phải làm là 'Google' một chủ đề bạn quan tâm và thật tuyệt vời làm sao vì bạn có thể tìm thấy rất nhiều nguồn dữ liệu trên đó.

Chất lượng dữ liệu

Sự thật rằng chúng ta có thể tìm kiếm dữ liệu ở rất nhiều nơi không có nghĩa rằng đó là dữ liệu tốt. Bạn phải kiểm tra cẩn thận chất lượng của cơ quan thu thập dữ liệu, vì dữ liệu rất thường chứa đựng các lỗi do đo lường (errors of measurement), các lỗi do bỏ sót biến quan trọng (errors of omission), hoặc các lỗi do làm tròn số (errors of rounding), và vân vân. Đôi khi dữ liệu có sẵn chỉ ở mức tổng gộp cao (highly aggregated level), dữ liệu gộp như thế có thể không cho chúng ta nhiều thông tin về các thực thể riêng lẻ (individual entities). Các nhà nghiên cứu phải luôn ghi nhớ rằng các kết quả nghiên cứu chỉ tốt khi chất lượng của dữ liệu là tốt.

Không may, một nhà nghiên cứu riêng lẻ không đủ xa xỉ để thu thập lại dữ liệu, và phải phụ thuộc vào các nguồn thứ cấp (secondary sources). Nhưng mọi nỗ lực nên được thực hiện là phải thu thập được dữ liệu đáng tin cậy.

1.3 Ước lượng mô hình hồi quy tuyến tính

[Diễn giải: Trong mục này, Gujarati trình bày ngắn gọn phương pháp bình phương bé nhất thông thường (OLS - the method of **O**rdinary **L**east **S**quares) mà chúng ta đã học ở chương 6 và 8 - Kinh tế lượng căn bản. Cho nên, mục này không có gì mới cả].

Sau khi đã thu thập dữ liệu, câu hỏi quan trọng là: chúng ta ước lượng mô hình hồi quy tuyến tính được cho ở phương trình (1.1) như thế nào? Giả sử chúng ta muốn ước lượng hàm tiền lương (wage function) của một nhóm công nhân. Để giải thích mức tiền lương theo giờ (Y), chúng ta có thể có các biến giải thích như giới tính (gender), dân tộc (ethnicity), tình trạng tham gia nghiệp đoàn (union status), kinh nghiệm làm việc (work experience), và nhiều biến khác, đó là các biến giải thích X. Hơn nữa, giả sử rằng chúng ta có một mẫu ngẫu nhiên gồm 1000 công nhân. Chúng ta ước lượng phương trình (1.1) như thế nào? Câu trả lời như sau.

Phương pháp bình phương bé nhất (OLS)

Một phương pháp được sử dụng phổ biến để ước lượng các hệ số hồi quy là phương pháp bình phương bé nhất thông thường (OLS)[7]. Để giải thích phương pháp này, chúng ta viết lại phương trình (1.1) như sau:

$$u_i = Y_i - (B_1 + B_2X_{2i} + B_3X_{3i} + ... + B_kX_{ki}) \quad (1.3)$$

$$= Y_i - BX$$

Phương trình (1.3) cho rằng hạng nhiễu là chênh lệch giữa giá trị thực của Y và giá trị Y thu được từ mô hình hồi quy.

Một cách để thu được các giá trị ước lượng (estimate) của các hệ số B có thể được thực hiện bằng cách là cho tổng các hạng nhiễu u_i ($=\sum u_i$) càng nhỏ càng tốt, lý tưởng là bằng

[7] OLS là một trường hợp đặc biệt của phương pháp bình phương bé nhất tổng quát (generalized least squares method - GLS). Mặc dù OLS có nhiều tính chất thú vị, như sẽ được thảo luận ở phần dưới. Một phương pháp thay thế OLS có khả năng áp dụng tổng quát là phương pháp hợp lý tối đa (method of maximum likelihood - ML), mà chúng ta sẽ thảo luận ngắn gọn ở Phụ lục của chương này.

0. Vì nhiều lý do về mặt lý thuyết và thực tiễn, nên phương pháp OLS không tối thiểu hóa tổng các hạng nhiễu, mà tối thiểu hóa tổng bình phương của hạng nhiễu như sau:

$$\sum u^2_i = \sum(Y_i - B_1 - B_2X_{2i} - B_3X_{3i} - ... - B_kX_{ki})^2 \quad (1.4)$$

Ở đây tổng được tính cho tất cả các quan sát. Chúng ta gọi $\sum u^2_i$ là **tổng bình phương hạng nhiễu** (error sum of squares, ESS). [Diễn giải: Tổng bình phương hạng nhiễu không quan sát được. Tí nữa sẽ thay bằng tổng bình phương phần dư (residual sum of squares, RSS) với dữ liệu mẫu. Và ESS tình cờ cũng là viết tắt của Explained Sum of Squares (tổng bình phương phần giải thích), nên chúng ta cần lưu ý để không bị nhầm lẫn khi đọc các sách kinh tế lượng nhé].

Bây giờ, trong phương trình (1.4) chúng ta biết các giá trị mẫu của Y_i và các X_s, nhưng chúng ta không biết các giá trị của các hệ số B. Vì thế, để tối thiểu hóa ESS, chúng ta phải tìm các giá trị của các hệ số B sao cho ESS càng nhỏ càng tốt. Hiển nhiên, ESS bây giờ là một hàm của các hệ số B.

Việc tối thiểu hóa thực sự ESS cần đến các phương pháp giải tích (calculus techniques). Chúng ta lấy đạo hàm riêng phần của ESS theo mỗi hệ số B, cho các phương trình từ kết quả lấy đạo hàm này bằng 0, và giải các phương trình này đồng thời để có k các hệ số hồi quy[8]. Vì chúng ta có k hệ số hồi quy, nên chúng ta sẽ giải k phương trình đồng thời. Chúng ta không cần giải các phương trình này ở đây, vì các phần mềm làm điều đó theo cách đã được lập trình sẵn[9].

Chúng ta sẽ ký hiệu các hệ số ước lượng của B bằng chữ b thường, và vì thế phương trình ước lượng có thể được viết lại như sau:

$$Y_i = b_1 + b_2X_{2i} + b_3X_{3i} + ... + b_kX_{ki} + e_i \quad (1.5)$$

Mô hình này có thể được gọi là **mô hình hồi quy mẫu** (sample regression model), bản sao của mô hình hồi quy tổng thể được cho ở phương trình (1.1). [Diễn giải: Phương trình (1.5) và (1.1) khác nhau chỗ ký hiệu b (hệ số hồi quy mẫu) và B (hệ số hồi quy tổng thể), b là một biến ngẫu nhiên vì giá trị sẽ thay đổi từ mẫu này sang mẫu khác, còn B là các hằng số, nhưng mình không thể biết được là bao nhiêu vì không thể thu thập được toàn bộ dữ liệu của tổng thể].

Cho

$$\widehat{Y} = b_1 + b_2X_{2i} + b_3X_{3i} + ... + b_kX_{ki} = b\mathbf{X} \quad (1.6)$$

Chúng ta có thể viết lại phương trình (1.5) như sau:

$$Y_i = \widehat{Y} + e_i = b\mathbf{X} + e_i \quad (1.7)$$

Ở đây $\widehat{Y}$ là một ước lượng (estimator) của $\mathbf{B}X$. Cũng như $\mathbf{B}X$ [tức là $E(Y|X)$] có thể được giải thích như một **hàm hồi quy tổng thể** (population regression function, PRF), chúng ta có thể giải thích bX như **hàm hồi quy mẫu** (sample regression function, SRF).

[8] Những ai biết giải tích sẽ nhớ rằng để tìm giá trị bé nhất hoặc lớn nhất của một hàm có nhiều biến, điều kiện bậc một (first-order condition) là cho các đạo hàm của hàm số theo mỗi biến bằng 0.

[9] Về mặt toán học, những bạn đọc quan tâm có thể tham khảo Gujarati/Porter, Chương 2.

Chúng ta gọi các hệ số b là các **ước lượng** (estimators) của các hệ số B và e_i, được gọi là phần dư (residual), là một ước lượng của hạng nhiễu u_i. *Một ước lượng là một công thức hoặc một quy tắc (formula or rule) cho chúng ta biết chúng ta sẽ đi tìm các giá trị của các tham số tổng thể như thế nào*. Một giá trị bằng số (numerical value) có được bởi một ước lượng trong một mẫu cụ thể được gọi là **giá trị ước lượng** (estimate). Lưu ý cẩn thận là các ước lượng, tức là các hệ số bs, là **các biến ngẫu nhiên** (random variables), vì giá trị của chúng sẽ thay đổi từ mẫu này qua mẫu khác. Trái lại, các hệ số hoặc tham số hồi quy tổng thể, tức là các hệ số Bs, là các con số cố định, mặc dù chúng ta không biết chính xác chúng là bao nhiêu. Trên cơ sở mẫu, chúng ta cố gắng có được các dự đoán tốt nhất về giá trị của chúng.

Khoảng cách giữa hàm hồi quy mẫu và hàm hồi quy tổng thể là quan trọng, vì trong hầu hết các áp dụng chúng ta không thể nghiên cứu toàn bộ tổng thể vì nhiều lý do, kể cả các xem xét về mặt chi phí. Điều đáng lưu ý là trong các cuộc bầu cử tổng thống ở Mỹ, số phiếu bầu dựa trên một mẫu ngẫu nhiên, ví dụ 1000 người, thường dự đoán gần đúng với số phiếu thực trong các lần bầu cử.

Trong phân tích hồi quy, mục tiêu của chúng ta là nhằm rút ra các suy diễn (inferences) về hàm hồi quy tổng thể trên cơ sở hàm hồi quy mẫu, vì trong thực tế chúng ta hiếm khi quan sát được hàm hồi quy tổng thể; chúng ta chỉ dự đoán điều gì có thể diễn ra. Điều này là quan trọng bởi vì mục tiêu cuối cùng của chúng ta là tìm ra các giá trị thực của các hệ số Bs có thể là bao nhiêu. Vì lý do này, chúng ta cần dựa nhiều hơn vào lý thuyết, được cung cấp bởi mô hình hồi quy tuyến tính cổ điển, mô hình này được thảo luận ngay dưới đây.

1.4 Mô hình hồi quy tuyến tính cổ điển

Ở mục này, Gujarati nhắc lại 8 giả định (**a**ssumptions) mà chúng ta đã biết ở chương 6 - Kinh tế lượng căn bản.

- **A-1:** Mô hình hồi quy là **tuyến tính ở các tham số** như trong phương trình (1.1); có thể hoặc không tuyến tính ở các biến Y và Xs.
- **A-2:** Các biến giải thích được giả định là **cố định hoặc không ngẫu nhiên** (nonstochastic) theo nghĩa là các giá trị của biến giải thích được giữ cố định khi lấy mẫu lặp đi lặp lại. Giả định này có thể không thích hợp cho tất cả các dữ liệu kinh tế, nhưng như chúng ta sẽ thấy trong chương 7 và chương 19, nếu X và u được phân phối độc lập (independently distributed) thì các kết quả dựa trên giả định cổ điển được thảo luận dưới đây sẽ đúng miễn là phân tích của chúng ta có điều kiện theo các giá trị X cụ thể được rút ra từ mẫu. Tuy nhiên, nếu X và u không tương quan, thì các kết quả cổ điển sẽ tiệm cận (asymptotically) đúng (tức trong các mẫu lớn)[10].

[10] Lưu ý rằng sự độc lập hàm ý là không có tương quan, nhưng không có tương quan không nhất thiết hàm ý sự độc lập.

- **A-3:** Khi cho trước các giá trị của các biến X, giá trị kỳ vọng hoặc trung bình của hạng nhiễu bằng không, nghĩa là[11]:

$$E(u_i|\mathbf{X}) = 0 \quad (1.8)$$

Trong đó, để biểu thức được viết ngắn gọn, **X** (chữ X đậm) đại diện cho tất cả các biến X trong mô hình. Nói cách khác, **kỳ vọng có điều kiện** (conditional expectation) của hạng nhiễu, khi cho trước các giá trị của các biến X, là bằng không. Vì hạng nhiễu đại diện cho ảnh hưởng của tất cả các yếu tố [khác X, có ảnh hưởng không đáng kể lên Y], về cơ bản nó có thể là ngẫu nhiên, nên giả định giá trị trung bình của hạng nhiễu bằng không là hợp lý.

Gujarati gọi A-3 là *giả định tối quan trọng* (critical assumption), vì nhờ đó mà chúng ta có thể viết phương trình (1.2) như sau:

$$E(Y_i|\mathbf{X}) = \mathbf{BX} + E(u_i|\mathbf{X})$$

$$= \mathbf{BX} \quad (1.9)$$

Phương trình này được giải thích như mô hình cho giá trị trung bình của Y_i với điều kiện các giá trị X cho trước. Đây là **hàm hồi quy trung bình tổng thể** (PRF) như đã đề cập ở trên. Trong phân tích hồi quy, mục tiêu chính của chúng ta là ước lượng phương trình này. Nếu chỉ có một biến X, bạn có thể hình dung nó như một đường hồi quy tổng thể. Nếu có nhiều hơn một biến X, bạn sẽ tưởng tượng nó là một đường cong trong một đồ thị đa chiều. Hàm PRF ước lượng, tức bản sao từ dữ liệu mẫu của phương trình (1.9), được ký hiệu là $\widehat{Y}_i$ = bX. Nghĩa là, $\widehat{Y}_i$ = bX là một ước lượng của $E(Y_i|X)$.

A-4: Phương sai của mỗi hạng nhiễu u_i, khi các giá trị X cho trước, là hằng số, hoặc **phương sai không đổi** (homoscedastic; homo là bằng nhau và scedastic là phương sai). [Diễu giải: Nhớ là ứng với mỗi giá trị của X chúng ta có rất nhiều giá trị có thể có của Y_i và vì thế chúng ta có rất nhiều giá trị u_i tại mỗi giá trị $X = X_i$ nào đó và trung bình của u_i tại mỗi giá trị X cho trước được giả định bằng 0, và phương sai của u_i tại mỗi giá trị X cho trước được giả định là bằng nhau, cho dù các giá trị X khác nhau thì trung bình của Y sẽ khác nhau]. Với giả định này, chúng ta có thể viết như sau:

$$var(u_i \mid X) = \sigma^2 \quad (1.10)$$

Lưu ý: Không có chỉ số dưới (subscript) trong đại lượng σ^2.

A-5: Không có tương quan giữa hai hạng nhiễu. Nghĩa là, không có **tự tương quan** (autocorrelation). Ký hiệu như sau:

$$cov(u_i , u_j \mid \mathbf{X}) = 0 \quad (1.11)$$

Ở đây Cov là hiệp phương sai (covariance) và i và j là hai hang nhiễu khác nhau. Dĩ nhiên, nếu i = j thì phương trình (1.11) là phương sai của u_i như ở phương trình (1.10).

[11] Ký hiệu | sau u_i nhắc chúng ta rằng phân tích là có điều kiện theo các giá trị cho trước của X.

A-6: Không có các mối quan hệ tuyến tính hoàn hảo giữa các biến X. Đây là giả định không có **đa cộng tuyến** (multicollinearity). Ví dụ, các mối quan hệ như $X_5 = 2X_3 + 4X_4$ bị loại trừ.

A-7: Mô hình hồi quy **được xác định đúng** (correctly specified). Nói cách khác, không có **chệch do sai dạng mô hình** (specification bias) hoặc **lỗi sai dạng mô hình** (specification error) được sử dụng trong phân tích thực nghiệm. Chúng ta cũng ngầm giả định rằng số quan sát, n, phải lớn hơn số hệ số được ước lượng.

A-8: Mặc dù không phải là một phần của CLRM, nhưng ta cũng giả định là hạng nhiễu có phân phối chuẩn với trung bình bằng 0 và phương sai không đổi là σ^2. [Diễn giải: Giả định A-8 chỉ là kết quả từ giả định A-3 và A-4].

$$u_i \sim N(0, \sigma^2) \tag{1.12}$$

Trên cơ sở các giả định từ A-1 đến A-7, chúng ta có thể thấy rằng phương pháp bình phương bé nhất thông thường (OLS), phương pháp được sử dung phổ biến nhất trên thực tế, cho chúng ta các ước lượng của tham số phương trình hồi quy tổng thể có các tính chất thống kế đáng mong muốn như sau:

1. Các ước lượng là **tuyến tính**, tức là các hàm tuyến tính của biến phụ thuộc Y. Các ước lượng tuyến tính thì dễ hiểu và dễ xử lý hơn so với các ước lượng phi tuyến. [Diễn giải: Xem Ôn tập # 1, trong Tóm lược kinh tế lượng căn bản để hiểu tại sao các ước lượng OLS là hàm theo Y hoặc u; từ đó suy ra phân phối xác suất của các ước lượng OLS].
2. Các ước lượng **không chệch** (unbiased), tức là, trong các áp dụng lặp đi lặp lại của phương pháp, trung bình, các ước lượng tiến tới giá trị thực của tổng thể [tức là, E(bs) = Bs].
3. Trong số các ước lượng không chệch tuyến tính, các ước lượng OLS có phương sai bé nhất. Vì thế, các giá trị tham số thực có thể được ước lượng với sự không chắc chắn có thể có là ít nhất; một ước lượng không chệch với phương sai bé nhất được gọi là một ước lượng **hiệu quả** (efficient estimator).

Tóm lại, dưới các điều kiện giả định, các ước lượng OLS được gọi với cái tên rất dễ thương là **BLUE** (xanh hay buồn?): **B**est **L**inear **U**nbiased **E**stimators. Đây là nội dung cốt lõi của **định lý nổi tiếng Gauss-Markov**, định lý này cung cấp nền tảng lý thuyết (theoretical justification) cho phương pháp bình phương bé nhất.

Với giả định thứ 8 **A-8**, chúng ta có thể thấy rằng các ước lượng OLS có phân phối chuẩn [Diễn giải: Xem Ôn tập # 1, Tóm lược kinh tế lượng căn bản để hiểu tại sao các ước lượng OLS theo phân phối chuẩn, rất quan trọng]. Vì thế, chúng ta có thể rút ra các suy diễn về giá trị thực của các hệ số hồi quy tổng thể và kiểm định các giả thuyết thống kê. Với giả định thứ 8 về phân phối chuẩn, các ước lượng OLS là các **ước lượng không chệch tốt nhất** (best unbiased estimators) trong toàn bộ các ước lượng không chệch, bất kể tuyến tính hay không. Với giả định thứ 8 này, CLRM được biết như **mô hình hồi quy tuyến tính cổ điển phân phối chuẩn** (normal classical linear regression model, NCLRM).

Trước khi đi tiếp, một số câu hỏi có thể cần được nêu ra. Các giả định này thực tế như thế nào? Điều gì xảy ra nếu một hoặc nhiều hơn một trong số giả định này không được thỏa mãn? Trong trường hợp đó, có các ước lượng nào khác thay thế hay không? Tại sao chúng ta chỉ giới hạn trong các ước lượng tuyến tính? Tất cả các câu hỏi này sẽ được trả lời khi chúng ta chuyển sang phần II. Nhưng cần nói thêm rằng khi mới bắt đầu bất kỳ một lĩnh vực mới nào, chúng ta cần một số kiến thức nền tảng. CLRM sẽ cung cấp cho chúng ta một kiến thức nền tảng như thế.

1.5 Phương sai và sai số chuẩn của các ước lượng OLS

[Diễn giải: Trong mục 1.5 này, Gujarati trình bày rất ngắn gọn về phương sai và sai số chuẩn của các ước lượng OLS. Nếu một người chưa học qua kinh tế lượng căn bản sẽ rất mù mờ với đôi dòng vắn tắt như thế. Nhắc lại rằng, vấn đề này được soạn rất tỉ mỉ trong các chương 6, 7, và 8 - Kinh tế lượng căn bản; hoặc chương 7 trong Phân tích dữ liệu và dự báo trong kinh tế và tài chính của Hoài-Bình-Duy (2009). Ở đó, chúng ta dễ dàng hiểu được tại sao các ước lượng OLS (tức là các hệ số bs) là các biến ngẫu nhiên theo phân phối chuẩn với E(bs) = Bs, và phương sai của các ước lượng OLS có mối quan hệ như thế nào với phương sai của hạng nhiễu ngẫu nhiên u_i, và rồi có quan hệ như thế nào với phương sai của phần dư (tức RSS/bậc tự do); và chúng ta lý giải tại sao các ước lượng OLS theo phân phối chuẩn nhưng lại sử dụng thống kê t để xây dựng khoảng tin cậy và kiểm định các giả thuyết về các tham số hồi quy tổng thể. Nói chung, bạn nên đọc kỹ các chương đó trước].

Như đã lưu ý trước đây, các ước lượng OLS, tức các hệ số bs, là các biến ngẫu nhiên, vì giá trị của chúng sẽ thay đổi từ mẫu này qua mẫu khác. [Diễn giải: Nếu chúng ta có thể lấy nhiều mẫu khác nhau (ví dụ 500 mẫu), thì mỗi mẫu như thế sẽ cho các giá trị ước lượng của các hệ số bs, và các giá trị ước lượng này sẽ khác nhau giữa 500 mẫu này. Như thế, mỗi hệ số b là một biến ngẫu nhiên với 500 giá trị khác nhau]. Vì thế, chúng ta cần một thước đo về sự biến thiên của các ước lượng này. Trong thống kê, sự biến thiên của một biến ngẫu nhiên được đo bằng **phương sai σ^2** (variance) hoặc bằng căn bậc hai của phương sai, tức là **độ lệch chuẩn σ** (standard deviation). Trong ngữ cảnh của phân tích hồi quy, độ lệch chuẩn của một ước lượng được gọi là **sai số chuẩn** [standard error, ký hiệu là $se(b_k)$], nhưng về mặt khái niệm thì nó hoàn toàn giống như độ lệch chuẩn vậy. Đối với mô hình hồi quy tuyến tính, một giá trị ước lượng của phương sai của hạng nhiễu u_i được tính như sau: [Diễn giải: Hãy nhớ là giá trị ước lượng (estimate) chỉ là một giá trị bằng số (numerical value) của một ước lượng (estimator): một mẫu nhất định cho một giá trị ước lượng cụ thể, khi thay đổi mẫu khác thì giá trị lượng sẽ thay đổi, nhưng công thức (tức là ước lượng) thì vẫn không thay đổi].

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n-k} \tag{1.13}$$

Đó là, **tổng bình phương phần dư** (RSS) chia cho (n - k), gọi là **bậc tự do** (df), n là cỡ mẫu và k là số tham số hồi quy ước lượng, bao gồm một hệ số cắt (b_1) và (k - 1) hệ số độ dốc (slope coefficients). Và $\hat{\sigma}$ là **sai số chuẩn của hồi quy** (standard error of the regression, SER). Nó đơn giản là độ lệch chuẩn của các giá trị Y xoay quanh đường hồi quy và thường được sử dụng như một thước đo tóm tắt về "**mức độ phù hợp**" (goodness of fit) của

đường hồi quy ước lượng (xem mục 1.6). *Lưu ý rằng, dấu ^ ở trên một tham số là ký hiệu một ước lượng của tham số đó.*

[Diễn giải: Cần phải hiểu tại sao bậc tự do ở đây là n – k? Có vài cách để hiểu bậc tự do, nhưng có lẽ cách dễ hiểu nhất '**bậc tự do của RSS là số nguồn thông tin của RSS**' (sources of information). Để đơn giản, trước hết chúng ta xét một mẫu chỉ có 2 quan sát và ước lượng hàm hồi quy đơn: Y = a + bX + e, nghĩa là phương trình đường thẳng qua hai điểm. Ở đây, chúng ta có các giá trị Y và X. Để xác định a và b chúng ta cần cả hai quan sát này, và các giá trị $\widehat{Y}$ = Y, nên cả hai quan sát của phần dư e = Y - $\widehat{Y}$ = 0, và vì thế RSS = 0. Như vậy, df = 2 -2 = 0, tức là không có nguồn thông tin nào về RSS. Bây giờ, tăng lên 3 quan sát, thì 2 trong 3 quan sát này dùng để xác định vị trí đường thẳng, tức là xác định a và b; và tại 2 quan sát đó $\widehat{Y}$ ~ Y, nên phần dư e ~ 0, nên chỉ còn 1 quan sát giúp giải thích RSS là bao nhiêu. Nếu mở rộng cho mô hình có k hệ số và số quan sát n = k, thì chúng ta cần hết k quan sát để xác định k hệ số hồi quy, tức $\widehat{Y}$ = Y và RSS = 0. Nếu ta tăng thêm 1 quan sát thì RSS sẽ khác 0, và việc RSS là bao nhiêu là nhờ n – k = 1 bậc tự do đó tạo nên. Nếu quan sát, chúng ta thấy trong n quan sát, thì có k quan sát có $\widehat{Y}$ ~ Y. Ý nghĩa của xác định đúng số bậc tự do là làm cho ước lượng của RSS là không chệch, nghĩa là E(RSS) = ESS (tức error sum of squares): Xem chứng minh ở chương 7, Kinh tế lượng căn bản. Đối với ESS (explained sum of squares), thì bậc tự do là k – 1, tức với hồi quy đơn thì df của ESS là 1, với hồi quy 3 biến (Y, X_1 và X_2) thì df của ESS là 2, … Tại sao? Vì trong hồi quy Y = a + bX + e, thì ESS = b$\sum y.x$, với y = Y - $\overline{Y}$, và x = X - $\overline{X}$, nghĩa là df = 1, tức chỉ có một nguồn thông tin về ESS. Trong hồi quy Y = a + bX + cZ + e, thì ESS = b$\sum y.x$ + c$\sum y.z$, nghĩa là df = 2, tức chỉ có hai nguồn thông tin về ESS; tương tự chúng ta mở rộng cho mô hình với k – 1 biến giải thích].

Xem ví dụ:

Giả sử chúng ta chỉ có 3 quan sát (tức n = 3) và ước lượng mô hình hồi quy 3 biến Y, X, và Z (tức 3 hệ số hồi quy, k = 3). Như vậy, bậc tự do của tổng bình phương phần giải thích (ESS) sẽ là 3 – 1 = 2; và bậc tự do của tổng bình phương phần dư (RSS) sẽ là 3 – 3 =0. Quan sát bảng dưới đây ta thấy rằng RSS = 0, và df của nó là 0.

```
. reg Y X Z
```

Source	SS	df	MS	Number of obs	=	3
				F(2, 0)	=	.
Model	200	2	100	Prob > F	=	.
Residual	0	0	.	R-squared	=	1.0000
				Adj R-squared	=	.
Total	200	2	100	Root MSE	=	0

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X	-3.333333	.	.	.	.	.
Z	-1.43e-13	.	.	.	.	.
_cons	76.66667	.	.	.	.	.

Bây giờ chúng ta tăng thêm một quan sát (n = 4), thì kết quả sẽ khác: RSS khác 0, và df = 1.

```
. reg Y X Z
```

Source	SS	df	MS	Number of obs	=	4
				F(2, 1)	=	6394.50
Model	499.960907	2	249.980453	Prob > F	=	0.0088
Residual	.039093041	1	.039093041	R-squared	=	0.9999
				Adj R-squared	=	0.9998
Total	500	3	166.666667	Root MSE	=	.19772

Y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
X	-1.806099	.1477283	-12.23	0.052	-3.683164	.0709669
Z	.4808444	.0578524	8.31	0.076	-.2542403	1.215929
_cons	44.60907	3.197871	13.95	0.046	3.976272	85.24187

Điều quan trọng cần nhớ là độ lệch chuẩn của các giá trị của Y, ký hiệu là S_Y, được kỳ vọng lớn hơn SER, trừ khi mô hình hồi quy không giải thích nhiều cho biến thiên trong các các giá trị Y[12]. [Diễn giải: Trong kết quả hồi quy trên Eviews, đó là 'S.D dependent var']. Nếu điều đó xảy ra (tức mô hình hồi quy không giải thích được gì ...), thì thực hiện phân tích hồi quy không có ích gì, vì trong trường hợp đó các biến giải thích X không có tác động gì lên Y. Thì giá trị ước lượng tốt nhất của Y đơn giả chính là giá trị trung bình của nó, tức $\bar{Y}$. [Diễn giải: Trong kết quả hồi quy trên Eviews, đó là 'mean dependent var']. Dĩ nhiên, chúng ta sử dụng mô hình hồi quy đơn giản là vì các biến X được đưa vào mô hình sẽ giúp giải thích tốt hơn hành vi của Y mà một mình $\bar{Y}$ không thể làm được.

Với các giả định của mô hình hồi quy tuyến tính cổ điển, ta có thể dễ dàng suy ra các phương sai và sai số chuẩn của các hệ số hồi quy b, nhưng ta sẽ không trình bày các công thức tính ở đây bởi vì các phần mềm thống kê tính toán một cách dễ dàng, như chúng ta sẽ thấy ở phần ví dụ minh họa dưới đây.

Phân phối xác suất của các ước lượng OLS

Nếu chúng ta sử dụng giả định số 8 [Nghĩa là, hạng nhiễu u_i có phân phối chuẩn: $u_i \sim N(0, \sigma^2)$], thì chúng ta có thể thấy rằng mỗi ước lượng OLS của các hệ số hồi quy (tức các hệ số bs) bản thân nó cũng theo phân phối chuẩn [Diễn giải: Đã được nói rất kỹ ở chương 6 và 7 - Kinh tế lượng căn bản] với trung bình bằng với giá trị tổng thể tương ứng của nó (tức Bs) và phương sai thì có liên quan đến phương sai của hạng nhiễu σ^2 và giá trị của các biến X [Diễn giải: Xem lại công thức ở chương 6-8, Kinh tế lượng căn bản: $\sigma^2_{b_k} = \frac{\sigma^2}{\sum x_k^2}$]. Trên thực tế, σ^2 (phương sai của u_i) được thay thế bằng ước lượng của nó, tức $\hat{\sigma}^2$ (phương sai của phần dư e_i) như ở phương trình (1.13). Cho nên, trong các nghiên cứu thực nghiệm chúng ta sử dụng **phân phối t** (t probability distribution) thay vì phân phối chuẩn cho việc suy diễn thống kê như kiểm định giả thuyết chẳng hạn. Nhưng nhớ rằng khi cỡ mẫu tăng, thì phân phối t tiến về phân phối chuẩn. Việc biết các

[12] Phương sai mẫu của Y được định nghĩa $S_y^2 = \sum(Y_i - \bar{Y})^2/(n-1)$, trong đó $\bar{Y}$ là trung bình mẫu. Căn bậc hai của phương sai là độ lệch chuẩn của Y, ký hiệu là S_y.

ước lượng OLS tuân theo phân phối chuẩn rất hữu ích trong việc thiết lập các khoảng tin cậy và rút ra các suy diễn thống kê về các giá trị trị thực của các tham số tổng thể. Điều này được thực hiện như thế nào sẽ được trình bày ngay sau đây.

1.6 Kiểm định giả thuyết về các hệ số hồi quy thực hay các hệ số hồi quy tổng thể

Giả sử chúng ta muốn kiểm định giả thuyết cho rằng hệ số hồi quy tổng thể $B_k = 0$. Để kiểm định giả thuyết này, chúng ta sử dụng kiểm định t[13], đó là: [Diễn giải: Giả thuyết này nghĩa là biến X_k không có ảnh hưởng lên Y hay X_k không có giải thích gì cho sự biến thiên của Y].

$$t = \frac{b_k}{se(b_k)} \qquad (*)$$

[Diễn giải: Đúng ra, công thức đầy đủ là $\mathbf{t} = \frac{\mathbf{b_k} - \mathbf{B_k}}{\mathbf{se(b_k)}}$ (**), nhưng với giả thuyết H_0: $B_k = 0$, nên (**) thành (*). Công thức này gần giống với $\mathbf{z} = \frac{\mathbf{b_k} - \mathbf{B_k}}{\boldsymbol{\sigma}(\mathbf{b_k})}$ (***), nhưng do chúng ta không có thông tin về $\sigma(b_k)$ nên chúng ta thay $\sigma(b_k)$ bằng ước lượng từ mẫu của nó, $\hat{\sigma}(b_k)$, tức là $se(b_k)$; và biến chuẩn hóa z trở thành t. Trong các kết quả hồi quy trên Eviews hoặc Stata, t-stat hoặc t được tính theo (*), hàm ý với giả thuyết H_0: $B_k = 0$, tức chúng ta kiểm định xem từng hệ số hồi quy có khác 0 một cách có ý nghĩa thống kê hay không. Có 3 cách kiểm định giả thuyết này: (1) Xây dựng khoảng tin cậy 99%, 95%, hoặc 90% (thường Stata cung cấp sẵn khoảng tin cậy 95%) và xem hệ số B_k nằm trong hay nằm ngoài khoảng tin cậy đó (nếu khoảng tin cậy chứa số 0 thì chúng ta chấp nhận giả thuyết H_0, ngược lại thì chúng ta bác bỏ H_0); (2) So sánh giá trị (tuyệt đối) của thống kê t tính toán từ công thức (*) với giá trị t phê phán (critical t value) hoặc hay quen gọi là t tra bảng ở một mức ý nghĩa α được chọn (thường là 5%), nếu |t tính toán| < t tra bảng, thì chúng ta chấp nhận H_0, ngược lại, nếu |t tính toán| > t tra bảng thì chúng ta bác bỏ H_0; (3) Chúng ta so sánh giá trị xác suất p (trên Stata là p > |t|, và Eviews là prob.) với mức ý nghĩa α được chọn, nếu $p > \alpha$ thì chúng ta chấp nhận H_0, ngược lại, nếu $p < \alpha$ thì chúng ta bác bỏ H_0. Như vậy, chỉ có cách thứ 3 là nhanh gọn nhẹ nhất vì chúng ta không cần phải mất thời gian xây dựng khoảng tin cậy hoặc tra bảng thống kê t. Dĩ nhiên, cả ba cách đều đưa ra cùng một kết luận giống nhau].

Trở lại công thức (*). Ở đây, $se(b_k)$ là sai số chuẩn của hệ số b_k. Giá trị t này có (n - k) bậc tự do (df); nhớ lại rằng gắn liền với một thống kê t là bậc tự do của nó. Trong mô hình hồi quy có k biến. [Diễn giải: Tính cả biến Y nhé, thì df bằng số quan sát trừ số hệ số được ước lượng (tức số bs, kể cả hệ số cắt). Tại sao bậc tự do của $se(b_k)$ là (n - k), giống như df của RSS? Bởi vì $se(b_k) = \sqrt{\frac{\frac{RSS}{n-k}}{\sum x_k^2}}$. Hiểu tại sao rồi chứ?].

Một khi thống kê t được tính toán [Diễn giải: Sau khi chạy hồi quy là chúng ta có sẵn trong bảng kết quả], thì chúng ta nhìn vào bảng t để tìm xác suất để có một giá trị t bằng

[13] Nếu biết giá trị σ^2 thực, thì chúng ta có thể sử dụng phân phối chuẩn chuẩn hóa (standard normal distribution) để kiểm định giả thuyết. Vì chúng ta ước lượng phương sai thực của hạng nhiễu bằng ước lượng của nó, tức $\hat{\sigma}^2$, nên lý thuyết thống kê cho thấy rằng chúng ta nên sử dụng phân phối t.

hoặc lớn hơn giá trị t tính toán đó là bao nhiêu. [Diễn giải: Như vừa nói ở trên, chúng ta không nhất thiết phải nhìn vào bảng t gì hết, vì các phần mềm Stata và Eviews đã cho sẵn giá trị xác suất p]. Nếu xác suất để có giá trị t tính toán là nhỏ, ví dụ nhỏ hơn hoặc bằng 5%, thì chúng ta bác bỏ giả thuyết H_0 cho rằng $B_k = 0$. Trong trường hợp đó, ta nói rằng giá trị b_k ước lượng [Diễn giải: Trong sách Gujarati ghi là giá trị t ước lượng là không đúng] có ý nghĩa thống kê, nghĩa là, khác 0 một cách có ý nghĩa.

Các giá trị xác suất được chọn phổ biến là 10%, 5%, và 1%. Các giá trị này được biết như là các **mức ý nghĩa** (levels of significance) (thường được ký hiệu bằng ký tự Hy Lạp là α và cũng được biết như Sai lầm loại I), vì thế có tên là **kiểm định ý nghĩa t** (t tests of significance).

Ta không cần tốn công sức thao tác bằng tay, vì phần mềm thống kê cung cấp kết quả cần thiết. Các phần mềm này không chỉ cho ra các giá trị t ước lượng (hay quen gọi là t tính toán), mà còn các giá trị (xác suất) p, tức là **mức ý nghĩa chính xác** (exact level of significance) của các giá trị t. Nếu một giá trị p được tính toán, thì không cần thiết sử dụng các giá trị α được chọn một cách tùy ý nữa. Trên thực tế, một giá trị p thấp cho biết rằng hệ số ước lượng (tức b_k) có ý nghĩa thống kê[14]. Điều này sẽ cho biết một biến cụ thể đang được xem xét có một tác động có ý nghĩa thống kê lên biến phụ thuộc, khi giữ nguyên giá trị của tất cả các biến giải thích khác.

Một số phần mềm, như Excel và Stata, cũng tính các **khoảng tin cậy** cho từng hệ số hồi quy - thường là một khoảng tin cậy 95% (confidence interval, CI). Các khoảng tin cậy như thế đưa ra một khoảng các giá trị có xác suất chứa giá trị thực của tổng thể. 95% (hoặc một thước đo tương tự) được gọi là **hệ số tin cậy** (confidence coefficient, CC), và CC đơn giản là bằng 1 trừ giá trị của mức ý nghĩa, α, nhân 100 - tức là CC = 100(1 - α).

Khoảng tin cậy (1 - α) của bất kỳ hệ số hồi quy tổng thể B_k nào được thiết lập như sau:

$$Pr[b_k \pm t_{\alpha/2} se(b_k)] = (1 - \alpha) \qquad (1.14)$$

Trong đó, Pr là xác suất và $t_{\alpha/2}$ là giá trị của thống kê t từ bảng phân phối t ở mức ý nghĩa α/2 với bậc tự do thích hợp, và $se(b_k)$ là sai số chuẩn của b_k. Nói cách khác, chúng ta trừ hoặc cộng $t_{\alpha/2}$ nhân với sai số chuẩn của b_k vào b_k để có được khoảng tin cậy (1 - α) cho giá trị thực của B_k. $[b_k - t_{\alpha/2} se(b_k)]$ được gọi là **giới hạn dưới** (lower limit) và $[b_k + t_{\alpha/2} se(b_k)]$ được gọi là **giới hạn trên** (upper limit) của khoảng tin cậy. Đây được gọi là khoảng tin cậy hai phía.

Các khoảng tin cậy cần được giải thích cẩn thận. Cụ thể cần lưu ý những điểm sau đây:

1. Khoảng tin cậy ở phương trình (1.14) không nói rằng xác suất của giá trị thực B_k nằm trong khoảng giới hạn cho sẵn là (1 - α). Mặc dù ta không biết giá trị thực của B_k là bao nhiêu, nhưng nó được giả định là một con số cố định.
2. Khoảng tin cậy ở phương trình (1.14) là một **khoảng ngẫu nhiên** - nghĩa là, nó thay đổi từ mẫu này sang mẫu khác bởi vì nó dựa vào giá trị của b_k, mà b_k là ngẫu nhiên.

[14] Một số người nghiên cứu chọn các giá trị α và bác bỏ giả thuyết H_0 nếu giá trị p thấp hơn giá trị α được chọn.

3. Vì khoảng tin cậy là ngẫu nhiên, một phát biểu xác suất như ở phương trình (1.14) nên được hiểu theo nghĩa trong dài hạn - đó là, khi lấy mẫu lặp đi lặp lại: nếu, khi lấy mẫu lặp đi lặp lại, các khoảng tin cậy như ở phương trình (1.14) được xây dựng rất nhiều lần trên cơ sở xác suất là (1 - α), thì trong dài hạn, trung bình, các khoảng như thế sẽ có (1 - α) trường hợp chứa đựng giá trị thực B_k. Bất cứ một khoảng riêng lẻ nào dựa trên một mẫu riêng lẻ có thể hoặc không chứa giá trị thực B_k.
4. Như đã lưu ý, các khoảng tin cậy như trong phương trình (1.14) là ngẫu nhiên. Nhưng một khi ta có một mẫu cụ thể và một khi ta có được một giá trị bằng số cụ thể của B_k, khoảng tin cậy dựa vào giá trị này là không ngẫu nhiên mà là cố định. Vì thế ta không thể nói rằng xác suất là (1 - α) mà khoảng tin cậy cố định cho trước chứa tham số thực. Trong trường hợp này, B_k hoặc nằm trong khoảng này hoặc không nằm trong khoảng này. Vì thế, xác suất là 1 hoặc 0.

Ý nghĩa tổng thể của hồi quy

Giả sử ta muốn kiểm định giả thuyết rằng tất cả các hệ số độ đốc ở phương trình (1.1) đồng thời bằng không. Điều này nghĩa là tất cả các biến giải thích trong mô hình không có tác động gì lên biến phụ thuộc. Nói gọn lại, mô hình không giúp giải thích được gì về hành vi của biến phụ thuộc. Kiểm định này được biết trong lý thuyết như là **kiểm định ý nghĩa tổng thể của hồi quy** (overall significance of the regression). Giả thuyết này được kiểm định bằng **kiểm định thống kê F**. Phát biểu bằng lời, thống kê F được định nghĩa như sau:

$$F = (ESS/df) / (RSS/df) \quad (1.15)$$

[Diễn giải: df của ESS khác với df của RSS].

Với ESS (tổng bình phương được giải thích) là phần biến thiên trong biến phụ thuộc Y được giải thích bởi mô hình và RSS (tổng bình phương phần dư) là phần biến thiên trong biến phụ thuộc Y không được giải thích bởi mô hình. Tổng của hai phần này là tổng biến thiên trong Y, và được gọi là tổng bình phương tổng (TSS).

Như phương trình Eq.(1.15) cho thấy, thống kê F có hai bậc tự do, một ở tử số và một ở mẫu số. Bậc tự do ở mẫu số luôn luôn là (n - k), nghĩa là bằng số quan sát trừ số hệ số được ước lượng, kể cả hệ số cắt, và bậc tự do ở tử số luôn là (k - 1), nghĩa là bằng tổng số biến giải thích trong mô hình không tính hệ số cắt, đó chính là tổng số hệ số độ dốc được ước lượng.

Giá trị F tính toán [theo công thức (1.15)] có thể được kiểm định cho ý nghĩa của nó bằng cách so sánh giá trị F tính toán với giá trị F từ bảng thống kê F [thường gọi là giá trị F tra bảng hay **giá trị F phê phán** (critical F value)]. Nếu giá trị F tính toán lớn hơn giá trị F phê phán ở một mức ý nghĩa α được chọn, ta có thể bác bỏ giả thuyết H_0 và kết luận rằng ít nhất có một biến giải thích có ý nghĩa thống kê. Giống như giá trị xác suất p trong thống kê t, hầu hết các phần mềm đều có trình bày giá trị xác suất p của thống kê F. Tất cả các thông tin này có thể được gặp trong bảng **phân tích phương sai** (AOV, hoặc có

thể viết khác là ANOVA) thường kèm theo trong kết quả hồi quy; tí nữa chúng ta sẽ thấy ngay trong phần ví dụ minh họa.

Điều rất quan trọng cần lưu ý là việc sử dụng các kiểm định t và F rõ ràng phải dựa trên giả định rằng hạng nhiễu u_i có phân phối chuẩn, như ở giả định số 8. Nếu giả định này không thể đứng vững, thì thủ tục kiểm định t và F không có hiệu lực trong các mẫu nhỏ, mặc dù các kiểm định này vẫn có thể được sử dụng nếu như mẫu đủ lớn, đây là một điểm sẽ được quay lại xem xét ở chương 7 khi bàn về các lỗi do sai dạng mô hình.

[Diễn giải: Một cách khác để hiểu giá trị F tính toán, và cách này y chang như thống kê Wald F trong phần kiểm định một ràng buộc tuyến tính (linear restriction)].

Sử dụng ví dụ minh họa về tiền lương theo giờ (xem mục 1.8):

Bước 1: Chúng ta hồi quy mô hình đầy đủ các biến, gọi là mô hình U (tức là unrestricted model), lưu RSS_U = 54342.5442 và df = 1283:

```
. reg wage female nonwhite union education exper
```

Source	SS	df	MS		
				Number of obs	= 1,289
				F(5, 1283)	= 122.61
Model	25967.2805	5	5193.45611	Prob > F	= 0.0000
Residual	54342.5442	1,283	42.3558411	R-squared	= 0.3233
				Adj R-squared	= 0.3207
Total	80309.8247	1,288	62.3523484	Root MSE	= 6.5081

Bước 2: Chúng ta hồi quy mô hình chỉ có hệ số cắt (tức ràng buộc bởi giả thuyết H_0: $B_2 = B_3 = \ldots = B_6 = 0$), gọi là mô hình R (tức là restricted model), lưu RSS_R = 80309.8247 và df = 1288:

```
. reg wage
```

Source	SS	df	MS		
				Number of obs	= 1,289
				F(0, 1288)	= 0.00
Model	0	0	.	Prob > F	= .
Residual	80309.8247	1,288	62.3523484	R-squared	= 0.0000
				Adj R-squared	= 0.0000
Total	80309.8247	1,288	62.3523484	Root MSE	= 7.8964

Bước 3: Tính giá trị F theo công thức sau đây:

$$F = \frac{\frac{(RSS_R - RSS_U)}{(df_R - df_U)}}{\frac{RSS_U}{df_U}} = \frac{\frac{80309.8247 - 54342.5442}{1288 - 1283)}}{\frac{54342.5442}{1283}} = 122.61$$

1.7 R^2: thước đo mức độ phù hợp của mô hình hồi quy được ước lượng

Hệ số xác định, ký hiệu là R^2, là một thước đo tổng quát về mức độ phù hợp của đường hồi quy được ước lượng (hoặc mặt phẳng, nếu có là mô hình hồi quy bội), nghĩa là, R^2 cho biết tỷ số hay phần trăm của tổng biến thiên trong biến phụ thuộc Y (TSS) được giải thích bởi tất cả các biến giải thích. Để biết R^2 được tính như thế nào, ta hãy định nghĩa như sau:

Tổng bình phương tổng (TSS) $= \Sigma y_i^2 = \Sigma(Y_i - \overline{Y})^2$

Tổng bình phương phần giải thích (ESS) = $\Sigma\hat{y}_i^2 = \Sigma(\hat{Y}_i - \overline{Y})^2$

Tổng bình phương phần dư (RSS) = Σe_i^2

Bây giờ, ta có thể thấy rằng:

$y_i = \hat{y}_i + e_i$

$y_i^2 = (\hat{y}_i + e_i)^2$

$= \hat{y}_i^2 + e_i^2 + 2\hat{y}_i * e_i$

$\Sigma y_i^2 = \Sigma\hat{y}_i^2 + \Sigma e_i^2 + 2\Sigma\hat{y}_i * e_i$

Bởi vì $2\Sigma\hat{y}_i * e_i = 0$ [bởi vì tổng phần dư luôn bằng 0]

Nên ta có:

$$\Sigma y_i^2 = \Sigma\hat{y}_i^2 + \Sigma e_i^2 \quad (1.16)^{15}$$

Phương trình này cho thấy tổng biến thiên của các giá trị thực tế của Y xoay quanh trung bình mẫu (TSS) bằng tổng biến thiên của các giá trị Y ước lượng xoay quanh giá trị trung bình (cũng là $\overline{Y}$) và tổng các phần dư bình phương. Nói cách khác,

$$TSS = ESS + RSS \quad (1.17)$$

Bây giờ, ta định nghĩa R^2 như sau:

$$R^2 = ESS/TSS \quad (1.18)$$

Như đã định nghĩa, hệ số xác định đơn giản là tỷ số hay phần trăm của tổng biến thiên của Y được giải thích bởi mô hình hồi quy.

Vì thế R^2 nằm giữa 0 và 1, với điều kiện là mô hình hồi quy có hệ số cắt. R^2 càng gần 1, thì mô hình càng phù hợp, và R^2 càng gần 0, thì mô hình càng không phù hợp. Nhớ rằng trong phân tích hồi quy, một trong những mục tiêu là giải thích sự biến thiên trong biến phụ thuộc càng nhiều càng tốt với sự hỗ trợ của các biến giải thích.

R^2 có thể được định nghĩa một cách khác như sau:

$$R^2 = 1 - RSS/TSS \quad (1.19)^{16}$$

Một nhược điểm của R^2 là nó là một hàm tăng theo số biến giải thích. Nghĩa là, nếu ta đưa thêm một biến giải thích vào mô hình (bất kể biến đó có thích hợp hay không), thì giá trị R^2 tăng. Vì thế, thỉnh thoảng các nhà nghiên cứu trả giá cho trò chơi "tối đa hóa" R^2, để có mô hình tốt hơn.

---

[15] Gợi ý: Bắt đầu với $y_i = \hat{y}_i + e_i$, lấy tổng bình phương hai phía của phương trình này và nhớ là $\Sigma\hat{y}_i e_i = 0$ vì kết quả của ước lượng OLS.

[16] TSS = ESS + RSS. Vì thế, 1 = ESS/TSS + RSS/TSS. Nghĩa là, 1 = R^2 - RSS/TSS. Sắp xếp lại chúng ta có phương trình (1.19).

Để tránh cám dỗ này, người ta đề nghị nên sử dụng một thước đo R^2 đã có tính đến số biến giải thích được đưa vào mô hình. R^2 như thế gọi là **R^2 điều chỉnh** (adjusted R^2), ký hiệu là $\overline{R}^2$ [đọc là R ba que :))))], và được tính từ R^2 chưa điều chỉnh như sau:

$$\overline{R}^2 = 1 - (1 - R^2)[(n - 1)/(n - k)] \quad (1.20)$$

Thuật ngữ "điều chỉnh" nghĩa là điều chỉnh số bậc tự do, mà số bậc tự do này phụ thuộc vào số biến giải thích (k) trong mô hình.

Lưu ý hai tính chất (đặc điểm) của R^2 điều chỉnh:

1. Nếu k > 1, thì adj. $R^2 < R^2$, nghĩa là, khi số biến giải thích trong mô hình tăng, thì R^2 điều chỉnh càng trở nên nhỏ hơn R^2 không điều chỉnh. Vì thế, R^2 điều chỉnh gán một mức "phạt" cho việc đưa thêm biến giải thích vào mô hình.
2. R^2 không điều chỉnh luôn dương, R^2 điều chỉnh đôi khi có thể âm.

R^2 điều chỉnh thường được sử dung để so sánh hai hoặc nhiều mô hình hồi quy có CÙNG BIẾN PHỤ THUỘC. Dĩ nhiên, còn có các thước đo khác để so sánh giữa các mô hình hồi quy, các thước đo này sẽ được trình bày ở chương 7 của cuốn sách này.

1.8 Ví dụ minh họa: các nhân tố quyết tiền lương theo giờ

Cuộc điều tra dân số hiện tại (CPS), được thực hiện bởi Cục Thống kê dân số Hoa Kỳ, định kỳ thực hiện nhiều cuộc điều tra về nhiều lĩnh vực khác nhau. Trong ví dụ này, ta xem xét một mẫu dữ liệu chéo gồm 1.289 người được phỏng vấn vào tháng 3 năm 1995 để nghiên cứu các yếu tố quyết định tiền lương theo giờ (tính bằng đô la) trong mẫu này[17]. Nhớ rằng 1.289 quan sát này là một mẫu được thu thập từ một tổng thể lớn hơn rất nhiều. Các biến được dùng trong phân tích được định nghĩa như sau:

Biến phụ thuộc:

- Wage: tiền lương theo giờ tính bằng đô la, là biến phụ thuộc.

Biến giải thích:

- Female: Giới tính, mã hóa 1 cho nữ, 0 cho nam.
- Nonwhite: Sắc tộc, mã hóa 1 cho các công nhân da màu, 0 cho các công nhân da trắng.
- Union: Tình trạng tham gia công đoàn, mã hóa 1 nếu công việc có công đoàn, 0 nếu không.
- Education: Giáo dục (số năm).
- Exper: Kinh nghiệm làm việc tiềm năng, được định nghĩa bằng số tuổi trừ số năm đi học trừ 6. (Giả định rằng số năm đi học bắt đầu lúc 6 tuổi).

[17] Dữ liệu được sử dụng ở đây là từ Điều tra dân số hiện tại được lấy từ Tổng cục điều tra dân số. Dữ liệu này cũng xuất hiện trong cuốn sách của Paul A. Rudd. An Introduction to Classical Econometric Theory, Oxford University Press, New York, 2000.

Sử dụng lệnh **des** trên Stata, chúng ta có:

variable name	storage type	display format	value label	variable label
obs	int	%8.0g		**Worker ID number**
wage	float	%9.0g		**Hourly wage in dollars**
female	byte	%8.0g		**Female (dummy)**
nonwhite	byte	%8.0g		**Nonwhite worker (dummy)**
union	byte	%8.0g		**Union status (dummy)**
education	byte	%8.0g		**Education in years**
exper	byte	%8.0g		**Potential work experience in years, age-schooling-6**
age	byte	%8.0g		**Age in years**
wind	byte	%8.0g		**Not paid by hour (dummy)**
femalenonw	byte	%9.0g		**Interaction between female and nonwhite**
lnwage	float	%9.0g		**ln(wage)**

Sử dụng lệnh **sum** trên Stata, chúng ta có:

Variable	Obs	Mean	Std. Dev.	Min	Max
obs	**1,289**	**645**	**372.2466**	**1**	**1289**
wage	**1,289**	**12.36585**	**7.89635**	**.84**	**64.08**
female	**1,289**	**.4972847**	**.5001867**	**0**	**1**
nonwhite	**1,289**	**.1528317**	**.3599648**	**0**	**1**
union	**1,289**	**.159038**	**.3658535**	**0**	**1**
education	**1,289**	**13.14507**	**2.813823**	**0**	**20**
exper	**1,289**	**18.78976**	**11.66284**	**0**	**56**
age	**1,289**	**37.93483**	**11.49428**	**18**	**65**
wind	**1,289**	**.4072925**	**.4915208**	**0**	**1**
femalenonw	**1,289**	**.0837859**	**.277174**	**0**	**1**
lnwage	**1,289**	**2.342416**	**.5863556**	**-.1743534**	**4.160132**

Mặc dù nhiều biến giải thích khác có thể được đưa vào mô hình, nhưng bây giờ ta sẽ tiếp tục với các biến này để minh họa một mô hình hồi quy bội nguyên mẫu (a prototype multiple regression model).

Lưu ý rằng tiền lương, giáo dục, và kinh nghiệm làm việc là các biến với thang đo tỷ lệ, và nữ, da màu, và công đoàn là các biến với thang đo định danh, các biến này được mã hóa dạng các **biến giả**. Cũng lưu ý rằng dữ liệu ở đây là dữ liệu chéo. Dữ liệu được cho sẵn trong **Bảng 1.1**, có thể được tìm thấy ở trang web đồng hành cùng cuốn sách này.

Trong cuốn sách này, ta sẽ sử dụng các phần mềm Eviews và Stata để ước lượng các mô hình hồi quy. Mặc dù với một bộ dữ liệu cho sẵn các phần mềm cho các kết quả tương tự, nhưng có một số điểm khác biệt trong cách trình bày kết quả giữa các phần mềm. Để giúp đọc giả làm quen với các phần mềm này, trong chương này chúng tôi sẽ trình bày kết quả dựa trên cả hai phần mềm. Ở các chương tiếp sau, ta có thể sử dung một trong hai phần mềm, nhưng hầu hết là Eviews bởi vì khả năng dễ tiếp cận của phần mềm này.

Sử dung Eviews 6, ta có kết quả như ở Bảng 1.2.

[Diễn giải: Hiện tại UEH ít sử dụng phần mềm Eviews trong giảng dạy thống kê và kinh tế lượng, thay vào đó là các phần mềm R và Stata. Do cuốn sách này sử dụng Eviews, nên tôi vẫn trình bày các ví dụ bằng Eviews. Tuy nhiên, phần mềm không quan trọng, mà quan trọng là chúng ta hiểu được kết quả].

Bảng 1.2: Hồi quy hàm tiền lương (Eviews 8)

LS WAGE C FEMALE NONWHITE UNION EDUCATION EXPER

Dependent Variable: WAGE
Method: Least Squares
Date: 10/21/17 Time: 23:13
Sample: 1 1289
Included observations: 1289

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-7.183338	1.015788	-7.071691	0.0000
FEMALE	-3.074875	0.364616	-8.433184	0.0000
NONWHITE	-1.565313	0.509188	-3.074139	0.0022
UNION	1.095976	0.506078	2.165626	0.0305
EDUCATION	1.370301	0.065904	20.79231	0.0000
EXPER	0.166607	0.016048	10.38205	0.0000

R-squared	0.323339	Mean dependent var	12.36585
Adjusted R-squared	0.320702	S.D. dependent var	7.896350
S.E. of regression	6.508137	Akaike info criterion	6.588627
Sum squared resid	54342.54	Schwarz criterion	6.612653
Log likelihood	-4240.370	Hannan-Quinn criter.	6.597646
F-statistic	122.6149	Durbin-Watson stat	1.897513
Prob(F-statistic)	0.000000		

Định dạng kết quả hồi quy trên Eviews được chuẩn hóa cao. Phần thứ nhất của bảng kết quả cho biết tên của biến phụ thuộc, phương pháp ước lượng (bình phương bé nhất), số quan sát, và khoảng mẫu được sử dụng cho ước lượng hiện tại. Thỉnh thoảng ta có thể không sử dung hết tất cả các quan sát của mẫu (nhất là các mô hình chuỗi thời gian), và để dành một vài quan sát, được gọi là **các quan sát giữ lại** (holdover observation), cho các mục đích dự báo. [Diễn giải: Thường đối với dữ liệu chuỗi thời gian, như các mô hình ARIMA, ARCH].

Phần thứ hai của bảng kết quả hồi quy cung cấp tên các biến giải thích, hệ số ước lượng của từng biến giải thích (kể cả hệ số cắt), sai số chuẩn của từng hệ số, giá trị thống kê t của mỗi hệ số, giá trị thống kê t này đơn giản là tỷ số của hệ số ước lượng và sai số chuẩn tương ứng[18] [Diễn giải: Tức là, giả thuyết H_0 ở đây là: $B_k = 0$. Chúng ta có thể viết tỷ số t như sau: $t = (b_k - B_k)/se(b_k)$, và khi $B_k = 0$ thì tỷ số này sẽ là $t = (b_k)/se(b_k)$. Nhưng chúng ta có thể kiểm định bất kỳ giả thiết nào khác của B_k (ví dụ $B_k = -2$) và tính lại giá trị của tỷ số t. Dĩ nhiên, phải tính toán bằng tay rồi], và giá trị xác suất p, hoặc **mức ý nghĩa chính xác** của thống kê t tương ứng của từng hệ số trong bằng kết quả. Đối với mỗi hệ số, giả thuyết không là giá trị của tổng thể của hệ số hồi quy (B lớn) là bằng 0, nghĩa là biến giải thích cụ thể không có ảnh hưởng gì đến biến phụ thuộc, sau khi giữ nguyên giá trị của các biến giải thích khác.

Giá trị xác suất p càng nhỏ, thì càng có bằng chứng bác bỏ giả thuyết không. Ví dụ, lấy biến kinh nghiệm, Exper. Giá trị hệ số của biến kinh nghiệm là 0.17 có giá trị t khoảng 10.38 (~ 0.17/0.016). Nếu giả thuyết rằng giá trị hệ số của biến này trong hàm hồi quy tổng thể (PRF) là 0, ta có thể dễ dàng bác bỏ giả thuyết đó bởi vì giá trị xác suất p để có giá trị $t \geq 10.38$ hầu như bằng 0 [Diễn giải: =TDIST(10.38,1233,2) trên Excel]. Trong trường hợp này, ta nói rằng hệ số của biến kinh nghiệm làm việc có ý nghĩa thống kê rất cao, nghĩa là nó khác không một cách có ý nghĩa thống kê rất cao. Nói theo một cách khác, ta có thể nói rằng kinh nghiệm làm việc là một nhân tố quyết định quan trọng của tiền lương theo giờ, sau khi cho phép ảnh hưởng của các biến khác trong mô hình - đây là một phát hiện không có gì đáng ngạc nhiên.

Nếu ta chọn giá trị p là 5%, thì Bảng 1.2 cho thấy mỗi hệ số ước lượng điều khác 0 một cách có ý nghĩa thống kê, nghĩa là, mỗi biến đều là một nhân tố quyết định quan trọng của tiền lương theo giờ.

Phần thứ ba của Bảng 1.2 cung cấp một số thống kê mô tả. Giá trị R^2 (hệ số xác định) khoảng 0.32 có nghĩa là khoảng 32% biến thiên trong tiền lương theo giờ được giải thích bởi sự biến thiên trong năm biến giải thích. Dường như có thể rằng giá trị R^2 hơi thấp, nhưng hay nhớ rằng ta có tới 1.289 quan sát với các giá trị thay đổi của biến phụ thuộc và biến giải thích. Trong một môi trường đa dạng như thế, các giá trị R^2 điển hình là thấp, và chúng thường thấp khi ta phân tích dữ liệu ở cấp độ cá nhân riêng lẻ. [Diễn giải: Thường R^2 khoảng 0.4 là đủ cao trong dữ liệu chéo với cỡ mẫu lớn]. Phần này cũng cung cấp giá trị R^2 điều chỉnh, giá trị này thấp hơn chút ít so với các giá trị R^2 chưa điều chỉnh, như đã được lưu ý trước đây. Vì ta không so sánh mô hình tiền lương này với bất kỳ mô hình nào khác, nên R^2 điều chỉnh không có đặc biệt quan trọng.

Nếu chúng ta muốn kiểm định giả thuyết rằng tất cả các hệ số độ dốc trong mô hình hồi quy tiền lương đồng thời bằng không [Nghĩa là: giả thuyết H_0: $B_2 = B_3 = B_4 = B_5 = B_6 = 0$], thì chúng ta sử dung kiểm định F như đã được thảo luận ở phần trước. Trong ví dụ hiện tại, giá trị F này khoảng 123. Giả thuyết H_0 này có thể bị bác bỏ nếu giá trị xác suất p của giá trị F ước lượng là rất thấp. Trong ví dụ của chúng ta, giá trị p thực tế là bằng 0, điều

[18] Giả thuyết H_0 ngầm ẩn ở đây là hệ số hồi quy thực của tổng thể là bằng 0. Chúng ta có thể viết tỷ số t như sau: $t = (b_k - B_k) / se(b_k)$, và giản lược bằng $t = b_k / se(b_k)$ nếu B_k thực sự bằng 0. Nhưng bạn có thể kiểm định bất kỳ giả thuyết nào khác về B_k bằng cách thay giá trị bạn muốn kiểm định vào công thức tỷ số t.

này cho thấy rằng chúng ta có thể mạnh dạn bác bỏ giả thuyết H_0 cho rằng tất cả các biến giải thích đồng thời không có tác động gì lên biến phụ thuộc, ở đây là tiền lương theo giờ. Ít nhất có một biến giải thích nào đó có ảnh hưởng có ý nghĩa thống kê lên biến phụ thuộc.

Bảng 1.2 cũng liệt kê một số thống kê khác như AIC, SIC, và HQ, các thống kê này được sử dung để lựa chọn giữa các mô hình [Diễn giải: Các tiêu chí này đặc biệt cần thiết trong các mô hình chuỗi thời gian như ARIMA, ARCH), và được trình bày khá chi tiết trong chương 7, Kinh tế lượng căn bản; hoặc chương 7, Hoài-Bình-Duy], thống kê d (tức *d* Durbin-Watson), là một thước đo mức độ tương quan trong hạng nhiễu (cũng thường thấy trong các mô hình chuỗi thời gian), và thống kê **log likelihood**, thống kê này hữu ích nếu chúng ta sử dung phương pháp hợp lý tối đa (ML) (xem Phụ lục cuối chương này). Chúng ta sẽ thảo luận về cách sử dụng các thống kê này ở các chương sau[19].

Mặc dù Eviews không trình bày bảng **phân tích phương sai** (analysis of variance, AOV), nhưng các phần mềm khác thì có. Bảng AOV có thể được suy ra một cách dễ dàng từ thông tin được cung cấp ở phần thứ ba của Bảng 1.2. Tuy nhiên, Stata không chỉ cho ta các hệ số hồi quy, các sai số chuẩn, và các thông tin như đã được đề cập ở trên, mà còn cho ta bảng AOV. Stata cũng cho ta khoảng tin cậy 95% cho mỗi hệ số ước lượng, như thấy trong Bảng 1.3.

Bảng 1.3: Kết quả ước lượng Stata của hàm tiền lương

```
. reg wage female nonwhite union education exper
```

Source	SS	df	MS	Number of obs	=	1,289
				F(5, 1283)	=	122.61
Model	25967.2805	5	5193.45611	Prob > F	=	0.0000
Residual	54342.5442	1,283	42.3558411	R-squared	=	0.3233
				Adj R-squared	=	0.3207
Total	80309.8247	1,288	62.3523484	Root MSE	=	6.5081

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-3.074875	.3646162	-8.43	0.000	-3.790185	-2.359566
nonwhite	-1.565313	.5091875	-3.07	0.002	-2.564245	-.5663817
union	1.095976	.5060781	2.17	0.031	.1031443	2.088807
education	1.370301	.0659042	20.79	0.000	1.241009	1.499593
exper	.1666065	.0160476	10.38	0.000	.1351242	.1980889
_cons	-7.183338	1.015788	-7.07	0.000	-9.176126	-5.190551

[19] Eviews cũng đưa ra tiêu chí thông tin Hannan – Quinn, tiêu chí này ở giữa hai tiêu chí AIC và SIC.

```
reg wage female nonwhite union education exper, level(99)
```

Source	SS	df	MS		
Model	25967.2805	5	5193.45611	Number of obs	= 1,289
Residual	54342.5442	1,283	42.3558411	F(5, 1283)	= 122.61
				Prob > F	= 0.0000
				R-squared	= 0.3233
				Adj R-squared	= 0.3207
Total	80309.8247	1,288	62.3523484	Root MSE	= 6.5081

wage	Coef.	Std. Err.	t	P>|t|	[99% Conf.	Interval]
female	-3.074875	.3646162	-8.43	0.000	-4.015464	-2.134287
nonwhite	-1.565313	.5091875	-3.07	0.002	-2.878848	-.2517792
union	1.095976	.5060781	2.17	0.031	-.2095371	2.401489
education	1.370301	.0659042	20.79	0.000	1.20029	1.540312
exper	.1666065	.0160476	10.38	0.000	.1252092	.2080039
_cons	-7.183338	1.015788	-7.07	0.000	-9.803732	-4.562944

Như bạn có thể thấy, không có sự khác biệt giữa Eviews và Stata trong các giá trị ước lượng của các hệ số hồi quy. Một đặc điểm duy nhất của Stata là nó đưa ra khoảng tin cậy 95% cho mỗi hệ số, được tính từ phương trình (1.14). Ví dụ, hãy xem biến giáo dục (education). Mặc dù giá trị ước lượng tốt nhất của hệ số thực của biến giá dục là 1.3703, khoảng 95% là (1.2410 đến 1.4995). Vì thế, chúng ta có thể nói rằng chúng ta tin cậy 95% rằng tác động của một năm đi học tăng them lên thu nhập theo giờ ít nhất là 1.24 đôla và cao nhất là 1.49 đôla, khi các yếu tố khác được giữ nguyên.

Vì vậy, nếu bạn giả định rằng hệ số thực của biến giáo dục, ví dụ, là 1.43, như được lưu ý trước đây, chúng ta không thể nói rằng 1.43 nằm trong khoảng này bởi vì khoảng tin cậy này là cố định. Vì thế, 1.43 hoặc là nằm trong khoảng này hoặc là không. Tất cả mà chúng ta có thể nói là nếu chúng ta theo thủ tục thiết lập các khoảng tin cậy theo như cách trong phương trình (1.14) trong các lần lấy mẫu lặp đi lặp lại, chúng ta có thể sẽ tin chắc một cách hợp lý rằng khoảng tin cậy chứa giá trị B_k thực. Dĩ nhiên, số lần chúng ta mất sai lầm sẽ là 5%.

Tác động lên tiền lương trung bình của một thay đổi đơn vị trong giá trị của một biến giải thích

Hệ số của biến nữ (female) ≈ -3.07 có nghĩa, khi giữ nguyên các biến khác không đổi, là tiền lương theo giờ trung bình của nữ thấp hơn tiền lương trung bình của nam khoảng 3 đôla. Tương tự, khi giữ nguyên các biến khác không đổi (ceteris paribus), tiền lương theo giờ trung bình của một công nhân da màu thấp hơn khoảng 1.56 đôla so với tiền lương của một công nhân da trắng. Hệ số của biến giáo dục cho biết rằng tiền lương theo giờ trung bình tăng khoảng 1.37 đôla cho mỗi một năm giáo dục tăng thêm, khi

giữ nguyên các biến khác không đổi. Tương tự, đối với mỗi năm kinh nghiệm làm việc tăng thêm, tiền lương theo giờ trung bình tăng thêm khoảng 17 cent, khi giữ nguyên các biến khác không đổi.

Kiểm định ý nghĩa chung của hồi quy

Để kiểm định giả thuyết rằng tất cả các hệ số độ dốc (tức hệ số của các biến X_k) đồng thời bằng 0 (tức là tất cả các biến giải thích không có tác động gì lên tiền lương theo giờ), Stata cho kết quả như ở Bảng 1.4.

Bảng 1.4: Bảng phân tích phương sai

Source	SS	df	MS		
				Number of obs	= 1,289
				F(5, 1283)	= 122.61
Model	25967.2805	5	5193.45611	Prob > F	= 0.0000
Residual	54342.5442	1,283	42.3558411	R-squared	= 0.3233
				Adj R-squared	= 0.3207
Total	80309.8247	1,288	62.3523484	Root MSE	= 6.5081

Hoặc sau khi hồi quy, chúng ta thực hiện như sau:

```
. reg wage female nonwhite union education exper
```

Source	SS	df	MS		
				Number of obs	= 1,289
				F(5, 1283)	= 122.61
Model	25967.2805	5	5193.45611	Prob > F	= 0.0000
Residual	54342.5442	1,283	42.3558411	R-squared	= 0.3233
				Adj R-squared	= 0.3207
Total	80309.8247	1,288	62.3523484	Root MSE	= 6.5081

wage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-3.074875	.3646162	-8.43	0.000	-3.790185	-2.359566
nonwhite	-1.565313	.5091875	-3.07	0.002	-2.564245	-.5663817
union	1.095976	.5060781	2.17	0.031	.1031443	2.088807
education	1.370301	.0659042	20.79	0.000	1.241009	1.499593
exper	.1666065	.0160476	10.38	0.000	.1351242	.1980889
_cons	-7.183338	1.015788	-7.07	0.000	-9.176126	-5.190551

```
. test female=nonwhite=union=education=exper=0

 ( 1)  female - nonwhite = 0
 ( 2)  female - union = 0
 ( 3)  female - education = 0
 ( 4)  female - exper = 0
 ( 5)  female = 0

       F(  5,  1283) =  122.61
            Prob > F =    0.0000
```

Bảng AOV đưa ra một phân tách của **tổng bình phương tổng** (TSS) thành hai thành phần: một được giải thích bởi mô hình, được gọi là **tổng bình phương phần giải thích** (ESS), và một phần khác không được giải thích bởi mô hình, được gọi là **tổng bình phương phần dư** (RSS), các thuật ngữ này chúng ta đã gặp ở phần trên.

Bây giờ mỗi tổng bình phương có bậc tự do riêng của nó. TSS có (n - 1) bậc tự do, vì chúng ta mất một bậc tự do để tính giá trị trung bình của biến phụ thuộc Y từ dữ liệu mẫu. ESS có (k - 1) bậc tự do, tức k biến giải thích không tính hệ số cắt, và RSS có (n - k) bậc tự do, nghĩa là bằng số quan sát, n, trừ số tham số được ước lượng (bao gồm cả hệ số cắt).

Bây giờ nếu bạn chia ESS cho bậc tự do của nó và chia RSS cho bậc tự do của nó, bạn sẽ có tổng bình phương trung bình (ký hiệu là MS – mean sums of squares) của ESS và RSS. Và nếu bạn lấy tỷ số của hai MS, bạn sẽ có giá trị F. Nó có thể được hiểu rằng dưới giả thuyết H_0 là tất cả các hệ số độ dốc đồng thời bằng 0, và giả định rằng hạng nhiễu u_i theo phân phối chuẩn, thì giá trị F tính toán sẽ theo phân phối F với bậc tự do của tử (numerator df) là (k - 1) và bậc tự do của mẫu (denominator df) là (n - k).

Trong ví dụ của chúng ta, giá trị F này là khoảng 123, giống với kết quả thu được từ hồi quy trên Eviews. Như bảng kết quả AOV cho thấy, xác suất để để có một giá trị lớn hơn hoặc bằng F là bằng 0, điều này gợi ý rằng giả thuyết H_0 có thể bị bác bỏ. Như thế ít nhất có một biến giải thích khác không một cách có ý nghĩa thống kê.

Nếu bảng AOV không có sẵn, chúng ta có thể kiểm định giả thuyết không rằng tất cả các hệ số độ dốc đồng thời bằng 0, nghĩa là, $B_2 = B_3 = \ldots = B_k = 0$, bằng cách sử dụng một mối quan hệ thú vị giữa F và R^2 như sau:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \qquad (1.21)^{20}$$

Vì giá trị R^2 được tạo ra bởi tất cả các phần mềm, nên nó có thể dễ dàng hơn để sử dụng phương trình (1.21) cho việc kiểm định giả thuyết không. Đối với ví dụ của chúng ta, giá trị R^2 tính toán là 0.3233. Sử dụng giá trị này, chúng ta có:

$$F = \frac{0.3233/5}{(1-0.3233)/1283} = 122.60 \qquad (1.22)$$

Giá trị này giống với giá trị như thấy trong bảng AOV.

Như đã được lưu ý trước đây, R^2 là tỷ lệ của biến thiên trong biến phụ thuộc được giải thích bởi các biến giải thích trong mô hình. Điều này có thể được xác nhận nếu bạn lấy tỷ số của ESS/TSS từ bảng AOV (= 25969.2805/8030.4247) = R^2 = 0.3233.

[20] Xem chứng minh ở Gujarati/Porter, p. 241.

1.9 Dự báo

[Mục này Gujarati chỉ viết vài dòng, thôi cho qua].

1.10 Nhìn về phía trước

Đến đây, chúng ta đã trình bày các vấn đề cơ bản của mô hình hồi quy tuyến tính cổ điển, chúng ta sẽ đi đâu từ đây? Sau đây là câu trả lời:

Mô hình hồi quy tiền lương ở Bảng 1.2 dựa vào các giả định của mô hình hồi quy tuyến tính cổ điển. Câu hỏi phát sinh một cách tự nhiên là: làm sao chúng ta biết mô hình này thỏa mãn các giả định của mô hình hồi quy tuyến tính cổ điển? Chúng ta cần biết các câu trả lời của các câu hỏi sau đây:

1. Mô hình tiền lương ở Bảng 1.2 là tuyến tính ở các biến cũng như ở các hệ số. Ví dụ, có thể biến tiền lương ở dạng logarít hay không? Có thể các biến giáo dục và kinh nghiệm cũng ở dạng logarít hay không? Vì tiền lương không được kỳ vọng tăng một cách tuyến tính với kinh nghiệm mãi mãi, nên có thể ta đưa biến kinh nghiệm bình phương như một biến giải thích thêm vào mô hình? Tất cả các câu hỏi này liên quan đến dạng hàm của mô hình hồi quy, và có nhiều dạng hàm khác nhau. Chúng ta sẽ xem xét chủ đề này ở Chương 2 của cuốn sách này.
2. Giả sử một số biến giải thích là biến định lượng và một số là biến định tính hoặc dạng biến có thang đo định danh. Có những vấn đề gì đặc biệt khi xử lý các biến giả hay không? Chúng ta sẽ xử lý sự tương tác giữa các biến định lượng và biến giả trong một tình huống nhất định như thế nào? Trong ví dụ về hàm tiền lương, chúng ta có ba biến giả: nữ, da màu, và công đoàn. Có phải các công nhân nữ có tham gia công đoàn nhận tiền lương cao hơn các công naân nữ không có tham gia công đoàn? Chúng ta sẽ giải quyết vấn đề này và các khía cạnh khác của **các biến giải thích định tính** ở Chương 3 của cuốn sách này.
3. Nếu chúng ta có một số biến giải thích trong mô hình hồi quy, làm sao chúng ta phát hiện rằng chúng ta không gặp phải vấn đề **đa cộng tuyến?** Nếu chúng ta gặp vần đền này, hậu quả là gì? Và chúng ta sẽ xử lý vấn đề này như thế nào? Chúng ta sẽ thảo luận vấn đề này ở Chương 4 của cuốn sách này.
4. Trong dữ liệu chéo, phương sai của hạng nhiễu có thể thay đổi hơn là không đổi. Chúng ta phát hiện vấn đề này như thế nào? Và hậu quả là gì? Các ước lượng OLS có còn là các ước ượng tuyến tính không chệch tốt nhất nữa không? Chúng ta sửa chữa phương sai thay đổi như thé nào? Chúng ta sẽ trả lời các câu hỏi này ở Chương 5 của cuốn sách này.
5. Trong dữ liệu chuỗi thời gian, giả định không có hiện tượng **tự tương quan** trong hạng nhiễu là điều không thể thỏa mãn. Chúng ta phát hiện tự tương quan như thế nào? Những hậu quả của tự tương quan là gì? Làm thế nào để khắc phục tự tương quan? Chúng ta sẽ trả lời các câu hỏi này ở Chương 6 của cuốn sách này.

6. Một trong những giả định của mô hình hồi quy tuyến tính cổ điển là mô hình được sử dụng trong phân tích thực nghiệm được ''xác định đúng'' theo cách mà tất cả các biến thích hợp đã được đưa vào mô hình, không có biến thừa nào được đưa vào mô hình, phân phối xác suất của hạng nhiễu được xác định đúng, và không có các lỗi trong đo lường các biến giải thích và biến phụ thuộc. Rõ ràng đây là một đòi hỏi quá mức. Nhưng quan trọng là chúng ta phát hiện ra những hậu quả của một hoặc nhiều hơn những tình huống này nếu chúng được phát hiện trong một ứng dụng cụ thể. Chúng ta thảo luận **vấn đề xác định dạng mô hình** chi tiết hơn ở Chương 7 của cuốn sách này. Chúng ta cũng thảo luận ngắn gọn trong chương này trường hợp khi mà các biến giải thích là ngẫu nhiên thay vì cố định như trong giả định của mô hình hồi quy tuyến tính cổ điển.

7. Giả sử biến phụ thuộc không phải là các biến có thang đo tỷ lệ hoặc thang đo khoảng mà là một biến có thang đo định danh, có giá trị 1 và 0. Liệu chúng ta có thể vẫn áp dụng các kỹ thuật OLS thông thường để ước lượng các mô hình như thế? Nếu không, thì các kỹ thuật ước lượng thay thế là gì? Trả lời cho các câu hỏi này có thể được tìm thấy ở Chương 8 của cuốn sách này, khi chúng ta thảo luận các mô hình **logit** và **probit**, các mô hình có thể xử lý biến phụ thuộc có thang đo định danh.

8. Chương 9 của cuốn sách này mở rộng các mô hình logit and probit hai thuộc tính sang các biến có thang đo định danh đa thuộc tính, ở đó biến phụ thuộc có nhiều hơn hai giá trị định danh. Ví dụ, xem xét phương tiện vận chuyển đến nơi làm việc. Giả sử ta có ba lựa chọn: xe tư, xe buýt công cộng, hoặc đi tàu. Chúng ta sẽ quyết định giữa các lựa chọn này như thế nào? Liệu chúng ta vẫn có thể sử dụng OLS? Như chúng ta sẽ thấy ở chương này, các vấn đề như thế yêu cầu các kỹ thuật ước lượng phi tuyến. **Các mô hình logit và probit đa thức** được thảo luận ở chương này sẽ cho thấy các biến với thang đo định danh đa thuộc tính có thể được mô hình hóa như thế nào.

9. Mặc dù các biến có thang đo định danh không thể được lượng hóa sẵn, các biến loại này đôi khi được xếp hạng (chắc ý là thang đo thứ bậc đó). **Các mô hình logit** và **probit thứ bậc** (thường người ta không dịch ra tiếng Việt) được thảo luận ở Chương 10 của cuốn sách này, cho thấy các mô hình này được ước lượng như thế nào.

10. Đôi khi biến phụ thuộc bị giới hạn trong các giá trị nó nhận được bởi do thiết kế của vấn đề đang được nghiên cứu. Giả sử chúng ta muốn nghiên cứu chi tiêu cho nhà cửa của các gia đình có thu nhập dưới $50.000 một năm. Rõ ràng là điều này loại trừ các gia đình có thu nhập cao hơn mức giới hạn này. Việc **mô hình hóa mẫu bị chặn** được thảo luận ở Chương 11 của cuốn sách này cho thấy chúng ta có thể mô hình hóa các hiện tượng như thế ra sao.

11. Thỉnh thoảng chúng ta gặp số liệu dạng số đếm, chẳng hạn như số lần đi khám bệnh, số bằng sáng chế một công ty nhận được, số khách hàng tại quầy thanh toán trong khoảng thời gian 15 phút, vân vân. Để mô hình hóa dữ liệu

dạng số đếm, **phân phối xác suất Poisson** (PPD) thường được sử dụng. Bởi vì giả định quan trọng của phân phối Poisson không phải lúc nào cũng được thỏa mãn, nên chúng ta sẽ thảo luận ngắn gọn một mô hình thay thế, được gọi là **phân phối nhị thức âm (NBD)**. Chúng ta sẽ thảo luận các chủ đề này ở Chương 12 của cuốn sách này.

12. Trong nhiều trường hợp dữ liệu chuỗi thời gian, một giả định nền tảng của mô hình hồi quy tuyến tính cổ điển thì các chuỗi thời gian là **chuỗi dừng**. Nếu không phải là chuỗi dừng, thì phương pháp luận OLS thông thường có còn có thể áp dụng được hay không? Các phương pháp thay thế là gì? Chúng ta sẽ thảo luận chủ đề này trong Chương 13 của cuốn sách này.

13. Mặc dù phương sai thay đổi nói chung thường liên quan đến dữ liệu chéo, hiện tượng này cũng có thể phát sinh trong chuỗi thời gian được gọi là hiện tượng **biến động nhóm** thường thấy trong dữ liệu chuỗi thời gian. Các mô hình ARCH và GARCH được thảo luận trong Chương 14 của cuốn sách này sẽ cho thấy chúng ta sẽ mô hình hóa biến động nhóm như thế nào.

14. Nếu chúng ta hồi quy một chuỗi thời gian không dừng với một hoặc nhiều chuỗi không dừng, điều này dẫn đến hiện tượng gọi là **hồi quy vô nghĩa** hoặc **hồi quy giả mạo**. Tuy nhiên, nếu có một mối quan hệ ổ định dài hạn giữa các biến, nghĩa là nếu các biến có **đồng liên kết**, thì đó không phải là một hồi quy giả mạo. Trong Chương 15 của cuốn sách này, chúng ta sẽ biết cách phát hiện hồi quy giả mạo và điều điều gì xảy ra nếu các biến không đồng liên kết (đồng tích hợp).

15. Dự báo là một lĩnh vực đặc biệt trong kinh tế lượng chuỗi thời gian. Trong Chương 16 của cuốn sách này, chúng ta thảo luận chủ đề dự báo kinh tế sử dụng mô hình hồi quy tuyến tính cũng như hai phương pháp dự báo ARIMA (trung bình trượt kết hợp tự hồi quy) và VAR (véctơ tự hồi quy). Với các ví dụ minh họa, chúng ta sẽ thấy các mô hình này được thực hiện như thế nào.

16. Các mô hình được thảo luận ở các chương trước đề cập đến dữ liệu chéo hoặc dữ liệu chuỗi thời gian. Chương 17 của cuốn sách này đề cập đến các mô hình kết hợp giữa dữ liệu chéo và dữ liệu chuỗi thời gian. Các mô hình này được biết với tên gọi là các mô hình hồi quy dữ liệu bảng. Chúng ta sẽ thấy trong chương này các mô hình như thế được ước lượng và giải thích như thế nào.

17. Trong Chương 18 của cuốn sách này, chúng ta thảo luận chủ đề về **phân tích thời gian chịu đựng** (trong tài chính thì gọi là thời gian đáo hạn trung bình) hoặc phân tích sống sót. Thời gian kéo dài dài của cuộc hôn nhân, thời gian kéo dài của cuộc đình công, thời gian kéo dài của cơn đau bệnh, và thời gian kéo dài của tình trạng thất nghiệp là các ví dụ về dữ liệu này.

18. Trong Chương 19 của cuốn sách này, chương cuối cùng, chúng ta thảo luận một chủ đề nhận được sự quan tâm đáng kể trong lý thuyết, đó là phương pháp **biến công cụ** (IV). Hầu như phần lớn nội dung cuốn sách này dành cho

trường hợp các biến giải thích dạng phi ngẫu nhiên hoặc có giá trị cố định, nhưng có nhiều tình huống ở đó ta phải xem xét các biến giải thích ngẫu nhiên. Nếu các biến giải thích ngẫu nhiên có tương quan với hạng nhiễu, thì các ước lượng OLS không chỉ bị chệch mà còn không vững - tức là, sự chệch không giảm cho dung cỡ mẫu lớn đến đâu. Nguyên tắc cơ bản của phương pháp IV là nó thay thế các biến giải thích ngẫu nhiên bằng các biến giải thích khác, gọi là các biến công cụ, các biến này có đặc điểm là có tương quan với biến giải thích ngẫu nhiên nhưng không có tương quan với hạng nhiễu. Nhờ đó chúng ta thu được các giá trị ước lượng vững của các tham số hồi quy. Trong chương này, chúng ta sẽ thấy phương pháp biến công cụ có thể được thực hiện như thế nào.

Phụ lục

Phương pháp ước lượng hợp lý tối đa (ML)

Lưu ý: Trước khi học phương pháp ML, bạn nên xem lại chương 6 trong Giáo trình thống kê UEH, vì công thức (thật ra cũng không cần nhớ, mà chỉ nắm ý tưởng là được) dưới đây chính là từ phân phối xác suất chuẩn.

Như đã lưu ý ở trước đây (ở mục 1.3), một phương pháp thay thế phương pháp bình phương bé nhất thông thường là **phương pháp ước lượng hợp lý tối đa** (the method of maximum likelihood). Phương pháp này đặc biệt hữu ích trong việc ước lượng các tham số của các mô hình hồi quy phi tuyến (ở tham số, lưu ý: OLS chỉ ước lượng được các mô hình tuyến tính ở tham số thôi) như mô hình logit, probit, logit đa thức (MNL), và probit đa thức. Ta sẽ gặp phương pháp ML ở nhiều chương khi thảo luận về các mô hình này (Lưu ý thêm: môn kinh tế lượng ứng dụng, phân tích hành vi người tiêu dùng, và rất nhiều nghiên cứu trong thực tế sử dụng phương pháp ML thay vì OLS).

Để tối thiểu hóa về mặt đại số (đỡ nhức đầu), ta xem xét một mô hình hồi quy hai biến:

$$Y_i = B_1 + B_2X_i + u_i \quad (1)$$

Trong đó

$$u_i \sim IIDN(0,\sigma^2) \quad (2)$$

Nghĩa là, hạng nhiễu có *phân phối độc lập và giống nhau như một phân phối chuẩn* (*IIDN*) với trung bình bằng 0 và phương sai không đổi (tức là theo phân phối chuẩn như đã giả định của CLRM).

Vì B_1 và B_2 là hằng số và X được giả định là cố định trong lấy mẫu lặp đi lặp lại, phương trình (2) hàm ý là:

$$Y_i \sim IIDN(B_1 + B_2X_i, \sigma^2) \quad (3)^{21}$$

[21] Nhớ lại thống kê căn bản rằng hàm mật độ của một biến X ngẫu nhiên có phân phối chuẩn với trung bình là μ và phương sai σ^2 là:

$$f(X) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma^2}(X - \mu)^2\right] \text{ với } -\infty < X < \infty, \sigma^2 > 0$$

[Lưu ý: Do *Y* là một hàm theo *u*, *u* có phân phối chuẩn nên *Y* cũng có phân phối chuẩn; và dễ dàng chứng minh được trung bình và phương sai của *Y* như ở phương trình (3)].

Nghĩa là, Y_i cũng có phân phối độc lập và giống nhau như một phân phối chuẩn với các tham số như đã nói [ở phương trình (3)]. Vì thế, ta có thể viết:

$$f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} exp\left[-\frac{1}{2\sigma^2}(Y_i - B_1 - B_2X_i)^2\right] \quad (4)$$

Phương trình (4) là hàm mật độ (density function) của biến ngẫu nhiênY_i có phân phối chuẩn với trung bình và phương sai được cho ở phương trình (3). Lưu ý: *exp* nghĩa là *e* lũy thừa của biểu thức trong ngoặc móc, *e* là cơ số của *ln*.

Vì mỗi Y_i phân phối như ở phương trình (4), nên hàm mật độ kết hợp (tức là, xác suất kết hợp) của các quan sát *Y* có thể được viết như một tích của *n* số hạng như thế, mỗi số hạng cho mỗiY_i. [Giải thích thêm: Do quan sát Y_i được xem như một biến cố độc lập, nên xác suất kết hợp, tức dựa vào phép giao của các biến cố độc lập, là tích của xác suất từng biến cố xảy ra. Xem lại chương 4, Giáo trình thống kê UEH, mục Quy tắc nhân cho các biến cố độc lập, ở chương 4 chỉ giới hạn cố 2 biến cố, nhưng trong trường hợp ta đang xét có đến *n* biến cố, tức *i = n*. Hiểu chứ?). Tích các số hạng từ phương trình (4) cho ra kết quả sau đây:

$$f(Y_1, Y_2, ..., Y_n) = \frac{1}{\sigma^n(\sqrt{2\pi})^n} exp\left[-\frac{1}{2}\sum\frac{(Y_i - B_1 - B_2X_i)^2}{\sigma^2}\right] \quad (5)$$

[Giải thích: Phương trình (5) là tích của *n* phương trình (4), nên từng số hạng như $1/\sigma^n$, $sqrt(2\pi)^n$ được lũy thừa lên *n* lần, và tích của *e* lũy thừa sẽ thành *e* lũy thừa của tổng các lũy thừa].

Nếu Y_1, Y_2, Y_3, ..., Y_n được cho trước hoặc được biết (tức có thể thu thập được dữ liệu), nhưng B_1, B_2, và σ^2 thì không được biết là bao nhiêu, hàm ở phương trình (5) được gọi là một **hàm khả năng** (likelihood function), ký hiệu là LF.

Phương pháp ước lượng hợp lý tối đa (hay khả năng tối đa), như cái tên đã cho biết, bao gồm việc ước lượng các tham số chưa biết theo cách mà xác suất để quan sát các *Y*s từ mẫu là cao nhất có thể có. Vì thế, chúng ta phải tìm cực đại của phương trình (5). Dễ dàng để tìm cực đại nếu chúng ta lấy log hai vế của hàm này, và kết quả là:

$$-\frac{n}{2}\ln\sigma^2 - \frac{n}{2}\ln(2\pi) - \frac{1}{2}\sum\frac{(Y_i - B_1 - B_2X_i)^2}{\sigma^2} \quad (6)$$

Vì số hạng cuối cùng trong phương trình (6) có dấu âm, nên để tối đa hóa phương trình (6), ta phải tối thiểu hóa số hạng cuối này. Ngoài σ^2, thì số hạng này chẳng qua là hạng nhiễu bình phương như của OLS. Nếu chúng ta lấy đạo hàm số hạng cuối này theo hệ số cắt và hệ số độ dốc (nghĩa là với dữ liệu mẫu, tức là các b_1 và b_2), chúng ta sẽ tìm ra các ước lượng của B_1 và B_2 (tức là, các hệ số b_1 và b_2) giống y như các ước lượng OLS đã được nói trong chương này (và đặc biệt là ở kinh tế lượng căn bản).

Tuy nhiên, có một khác biệt duy nhất là ở ước lượng của phương sai hạng nhiễu, tức là σ^2. Có thể thấy rằng ước lượng này là:

$$\hat{\sigma}^2_{ML} = \frac{\sum e_i^2}{n} \quad (7)$$

Trong khi ước lượng OLS là:

$$\hat{\sigma}^2_{OLS} = \frac{\sum e_i^2}{n-k} \quad (8)$$

Nói cách khác, ước lượng ML của phương sai chưa biết không được điều chỉnh bậc tự do, trong khi đó ước lượng của OLS thì có. Tuy nhiên, trong các mẫu lớn, hai ước lượng cho giá trị như nhau, mặc dù trong mẫu nhỏ, ước lượng ML là một ước lượng bị chệch của phương sai thực của hạng nhiễu.

Nếu chúng ta xem lại kết quả hồi quy về ví dụ tiền lương ở Bảng 1.2 (Mục 1.8), chúng ta sẽ thấy giá trị ln(LF) là -4240.37. Đây là giá trị lớn nhất của làm log likelihood. Nếu chúng ta lấy anti-log của giá trị này, chúng ta sẽ thấy nó gần bằng 0 [=exp(-4240.37) = 0, hàm Excel]. Cũng lưu ý rằng, các giá trị của tất cả các hệ số hồi quy được cho trong bảng đó cũng là các giá trị ước lượng theo phương pháp ML dưới giả định là hạng nhiễu có phân phối chuẩn.

Vì vậy, với tất cả các mục đích thực tế, các giá trị ước lượng ML và OLS của các hệ số hồi quy là giống nhau, khi giả định là hạng nhiễu có phân phối chuẩn trong bất kỳ ứng dung nào. Chúng ta sẽ thảo luận chủ đề này sâu hơn ở chương 7 của cuốn sách này.

Các ước lượng ML có nhiều tính chất đáng mong muốn đối với mẫu lớn: (1) chúng là các ước lượng "xấp xỉ" không chệch (từ asymptotic nghĩa là "tiệm cận", ở đây chúng ta nên hiểu là khi mẫu nhỏ thì các ước lượng ML bị chệch, nhưng khi tăng dần cỡ mẫu lên thì các ước lượng ML sẽ tiệm cận về các ước lượng không chệch); (2) chúng là các ước lượng vững (đôi khi gọi là nhất quán); (3) chúng là các ước lượng xấp xỉ hiệu quả - nghĩa là, trong các mẫu lớn, chúng có phương sai nhỏ nhất trong số các ước lượng vững; và (4) chúng là các ước lượng xấp xỉ phân phối chuẩn (nghĩa là khi cỡ mẫu lớn thì các ước lượng ML sẽ có phân phối chuẩn).

Nhớ rằng cần phân biệt một ước lượng không chệch và một ước lượng vững. Không chệch là một tính chất của việc lấy mẫu lặp đi lặp lại: giữ nguyên cỡ mẫu ở một lần lấy mẫu, chúng ta rút một số mẫu và từ mỗi mẫu ta có được một giá trị ước lượng của tham số chưa biết. Nếu giá trị trung bình của tất cả cả giá trị ước lượng (nhớ là giá trị ước lượng khác với ước lượng nhé) bằng với giá trị thực của tham số chưa biết, thì ước lượng đó (hoặc đúng hơn là phương pháp ước lượng đó) cho chúng ta một ước lượng không chệch (unbiased estimator).

Một ước lượng được gọi là vững (hay nhất quán) nếu (các giá trị ước lượng của) nó tiến gần về giá trị thực của tham số tổng thể khi cỡ mẫu (của cùng một mẫu) tăng lên.

Như đã lưu ý trước đây, trong OLS ta sử dung R^2 như một thước đo mức độ phù hợp của đường hồi quy được ước lượng. Một thước đo tương đương của R^2 trong phương pháp ML là pseudo R^2, được định nghĩa như sau[22]:

$$\text{pseudo } R^2 = 1 - \frac{lfL}{lfL_0} \quad (9)$$

[22] Thảo luận về vấn đề này có thể xem Christopher Dougherty, Introduction to Econometrics, 3rd edn, Oxford University Press, Oxford, 2007, pp. 320 – 321.

Trong đó *lf*L là giá trị log likelihood của mô hình đang được xem xét và lfL_0 là giá trị log likelihood của mô hình không có bất kỳ biến giải thích nào (ngoại trừ hệ số cắt). Vì thế, pseudo R^2 đo tỷ phần mà *lf*L nhỏ (theo giá trị tuyệt đối) hơn so với lfL_0.

Do log likelihood là xác suất kết hợp, nên nó phải nằm giữa 0 và 1. Vì thế, giá trị *lf*L phải là âm, như chúng ta thấy trong ví dụ minh họa ở Bảng 1.2.

Trong OLS ta kiểm định giả thuyết về ý nghĩa tổng quát (hay gọi là giả thuyết đồng thời - joint hypothesis) của mô hình hồi quy bằng kiểm định *F*. Một kiểm định tương đương ở phương pháp ML là **thống kê phân số khả năng** λ. Thống kê này được định nghĩa như sau:

$$\lambda = 2(lfL - lfL_0) \quad (10)$$

Dưới giả thuyết không rằng các hệ số hồi quy của tất cả các biến giải thích đồng thời bằng 0, thống kê này được phân phối như một phân phối χ^2 (chi-square) với (k - 1) bậc tự do, trong đó (k - 1) là số biến giải thích trong mô hình. Cũng như các kiểm định ý nghĩa khác, nếu giá trị χ^2 tính toán lớn hơn giá trị χ^2 phê phán ở mức ý nghĩa được chọn =CHISQ.INV(α, df), chúng ta bác bỏ giả thuyết không./.