Chương 3

Mô Hình Hồi Quy Tuyến Tính Đơn

Ở chương 1 phát biểu rằng bước đầu tiên trong phân tích kinh tế lượng là việc thiết lập mô hình mô tả được hành vi của các đại lượng kinh tế. Tiếp theo đó nhà phân tích kinh tế/ kinh doanh sẽ thu thập những dữ liệu thích hợp và ước lược mô hình nhằm hỗ trợ cho việc ra quyết định. Trong chương này sẽ giới thiệu mô hình đơn giản nhất và phát triển các phương pháp ước lượng, phương pháp kiểm định giả thuyết và phương pháp dự báo. Mô hình này đề cập đến biến độc lập (Y) và một biến phụ thuộc (X). Đó chính là mô hình hồi quy tuyến tính đơn. Mặc dù đây là một mô hình đơn giản, và vì thế phi thực tế, nhưng việc hiểu biết những vấn đề cơ bản trong mô hình này là nền tảng cho việc tìm hiểu những mô hình phức tạp hơn. Thực tế, mô hình hồi quy đơn tuyến tính có thể giải thích cho nhiều phương pháp kinh tế lượng. Trong chương này chỉ đưa ra những kết luận căn bản về mô hình hồi quy tuyến tính đơn biến. Còn những phần khác và phần tính toán sẽ được giới thiệu ở phần phụ lục. Vì vậy, đối với người đọc có những kiến thức căn bản về toán học, nếu thích, có thể đọc phần phụ lục để hiểu rõ hơn về những kết quả lý thuyết.

3.1 Mô Hình Cơ Bản

Chương 1 đã trình bày ví dụ về mô hình hồi quy đơn đề cập đến mối liên hệ giữa giá của một ngôi nhà và diện tích sử dụng (xem Hình 1.2). Chọn trước một số loại diện tích, và sau đó liệt kê số lượng nhà có trong tổng thể tương ứng với từng diện tích đã chọn. Sau đó tính giá bán trung bình của mỗi loại nhà và vẽ đồ thị (quy ước các điểm được biểu thị là *X*). Giả thuyết cơ bản trong mô hình hồi quy tuyến tính đơn là các trị trung bình này sẽ nằm trên một đường thẳng (biểu thị bằng $\alpha + \beta SQFT$), đây là **hàm hồi quy của tổng thể** và là *trung bình có điều kiện* (kỳ vọng) của GIÁ theo SQFT *cho trước*. Công thức tổng quát của mô hình hồi quy tuyến tính đơn dựa trên Giả thiết 3.1 sẽ là

GIẢ THIẾT 3.1 (Tính Tuyến Tính của Mô Hình)

$$Y_t = \alpha + \beta X_t + u_t \qquad \textbf{(3.1)}$$

trong đó, X_t và Y_t là trị quan sát thứ t (t = 1 đến n) của biến độc lập và biến phụ thuộc, tiếp theo α và β là các tham số chưa biết và sẽ được ước lượng; và u_t là số hạng sai số không quan sát được và được giả định là biến ngẫu nhiên với một số đặc tính nhất định mà sẽ được đề cập kỹ ở phần sau. α và β được gọi là **hệ số hồi quy**. (*t* thể hiện thời điểm trong chuỗi thời gian hoặc là trị quan sát trong một chuỗi dữ liệu chéo.)

Thuật ngữ *đơn* trong mô hình hồi quy tuyến tính đơn được sử dụng để chỉ rằng chỉ có duy nhất một biến giải thích (*X*) được sử dụng trong mô hình. Trong chương tiếp theo khi nói về *mô hồi quy đa biến* sẽ bổ sung thêm nhiều biến giải thích khác. Thuật ngữ *hồi quy* xuất phát từ Fraccis Galton (1886), người đặt ra mối liên hệ giữa chiều cao của nam với

chiều cao của người cha và quan sát thực nghiệm cho thấy có một xu hướng giữa chiều cao trung bình của nam với chiều cao của những người cha của họ để "hồi quy" (hoặc di chuyển) cho chiều cao trung bình của toàn bộ tổng thể. $\alpha + \beta X_b$ gọi là *phần xác định* của mô hình và là **trung bình có điều kiện của *Y* theo *X*,** đó là $E(Y_t | X_t) = \alpha + \beta X_t$. Thuật ngữ *tuyến tính* dùng để chỉ rằng bản chất của các *thông số của* tổng thể α và β là tuyến tính (bậc nhất) chứ *không phải là* X_t *tuyến tính*. Do đó, mô hình $Y_t = \alpha + \beta X_t^2 + u_t$ vẫn được gọi là hồi quy quyến tính đơn mặc dầu có *X* bình phương. Sau đây là ví dụ về phương trình **hồi quy phi tuyến tính** $Y_t = \alpha + X^{\beta} + u_t$. Trong cuốn sách này sẽ không đề cập đến mô hình hồi quy phi tuyến tính mà chỉ tập trung vào những mô hình có tham số có tính tuyến tính mà thôi. Những mô hình tuyến tính này có thể bao gồm các số hạng phi tuyến tính đối với biến giải thích (Chương 6). Để nghiên cứu sâu hơn về mô hình hồi quy phi tuyến tính, có thể tham khảo các tài liệu: Greene (1997), Davidson và MacKinnon (1993), và Griffths, Hill, và Judg (1993).

Số hạng sai số u_t (hay còn gọi là số hạng *ngẫu nhiên*) là thành phần ngẫu nhiên không quan sát được và là sai biệt giữa Y_t và phần xác định $\alpha + \beta X_t$. Sau đây một tổ hợp của bốn nguyên nhân ảnh hưởng khác nhau:

1. *Biến bỏ sót.* Giả sử mô hình thực sự là $Y_t = \alpha + \beta X_t + \gamma Z_t + v_t$ trong đó, Z_t là một biến giải thích khác và v_t là số hạng sai số thực sự, nhưng nếu ta sử dụng mô hình là $Y = \alpha + \beta X_t + u_t$ thì $u_t = \gamma Z_t + v_t$. Vì thế, u_t bao hàm cả ảnh hưởng của biến *Z* bị bỏ sót. Trong ví dụ về địa ốc ở phần trước, nếu mô hình thực sự bao gồm cả ảnh hưởng của phòng ngủ và phòng tắm và chúng ta đã bỏ qua hai ảnh hưởng này mà chỉ xét đến diện tích sử dụng thì số hạng *u* sẽ bao hàm cả ảnh hưởng của phòng ngủ và phòng tắm lên giá bán nhà.
2. *Phi tuyến tính.* u_t có thể bao gồm ảnh hưởng phi tuyến tính trong mối quan hệ giữa *Y* và *X*. Vì thế, nếu mô hình thực sự là $Y_t = \alpha + \beta X_t + \gamma X_t^2 + u_t$, nhưng lại được giả định bằng phương trình $Y = \alpha + \beta X_t + u_t$, thì ảnh hưởng của X_t^2 sẽ được bao hàm trong u_t.
3. *Sai số đo lường.* Sai số trong việc đo lường *X* và *Y* có thể được thể hiện qua *u*. Ví dụ, giả sử Y_t giá trị của việc xây dựng mới và ta muốn ước lượng hàm $Y_t = \alpha + \beta r_t + v_t$ trong đó r_t là lãi suất nợ vay và v_t là sai số thật sự (để đơn giản, ảnh hưởng của thu nhập và các biến khác lên đầu tư đều được loại bỏ). Tuy nhiên khi thực hiện ước lượng, chúng ta lại sử dụng mô hình $Y_t = \alpha + \beta X_t + u_t$ trong đó $X_t = r_t + Z_t$ là lãi suất căn bản. Như vậy thì lãi suất được đo lường trong sai số Z_t thay $r_t = X_t - Z_t$ vào phương trình ban đầu, ta sẽ được
$$Y_t = \alpha + \beta(X_t - Z_t) + v_t = \alpha + \beta X_t - \beta Z_t + v_t = \alpha + \beta X_t + u_t$$
Cần luôn lưu ý rằng tính ngẫu nhiên của số hạng u_t bao gồm sai số khi đo lường lãi suất nợ vay một cách chính xác.
4. *Những ảnh hưởng không thể dự báo*. Dù là một mô hình kinh tế lượng tốt cũng có thể chịu những ảnh hưởng ngẫu nhiên không thể dự báo được. Những ảnh hưởng này sẽ luôn được thể hiện qua số hạng sai số u_t.

Như đã đề cập ban đầu, việc thực hiện điều tra toàn bộ tổng thể để xác định hàm hồi quy của tổng thể là không thực tế. Vì vậy, trong thực tế, người phân tích thường chọn một mẫu bao gồm các căn nhà một cách ngẫu nhiên và đo lường các đặc tính của mẫu này để thiết lập **hàm hồi quy cho mẫu**. Bảng 3.1 trình bày dữ liệu của một mẫu gồm 14

nhà bán trong khu vực San Diego. Số liệu này có sẵn trong đĩa mềm với tên tập tin là DATA3-1. Trong Hình 3.1, các cặp giá trị (X_t, Y_t) được vẽ trên đồ thị. Đồ thị này được gọi là **đồ thị phân tán của mẫu** cho các dữ liệu. Hình 3.1 tương tự như Hình 1.2, nhưng trong Hình 1.2 liệt kê toàn bộ các giá trị (X_t, Y_t) của tổng thể, còn trong Hình 3.1 chỉ liệt kê dữ liệu của mẫu mà thôi. Giả sử, tại một thời điểm, ta biết được giá trị của α và β. Ta có thể vẽ được đường thẳng $\alpha + \beta X$ trên biểu đồ. Đây chính là **đường hồi quy của tổng thể**. Khoảng cách chiếu thẳng xuống từ giá thực (Y_t) đến đường hồi quy $\alpha + \beta X$ là sai số ngẫu nhiên u_t. Độ dốc của đường thẳng (β) cũng là $\Delta Y/\Delta X$, là *lượng thay đổi của Y trên một đơn vị thay đổi của X*. Vì vậy β được diễn dịch là **ảnh hưởng cận biên của *X* lên *Y***. Do đó, nếu là β là 0.14, điều đó có nghĩa là một mét vuông diện tích tăng thêm sẽ làm tăng giá bán nhà lên, ở mức trung bình, 0.14 ngàn đô la (lưu ý đơn vị tính) hay 140 đô la. Một cách thực tế hơn, khi diện tích sử dụng nhà tăng thêm 100 mét vuông thì hy vọng rằng giá bán trung bình của ngôi nhà sẽ tăng thêm $14.000 đô la. Mặc dầu α là tung độ gốc và là giá trị của trị trung bình *Y* khi *X* bằng 0, số hạng này vẫn không thể được hiểu như là giá trung bình của một lô đất trống. Nguyên nhân là vì α cũng ẩn chứa biến bỏ sót và do đó không có cách giải thích cho α (điều này được đề cập kỹ hơn trong Phần 4.5).

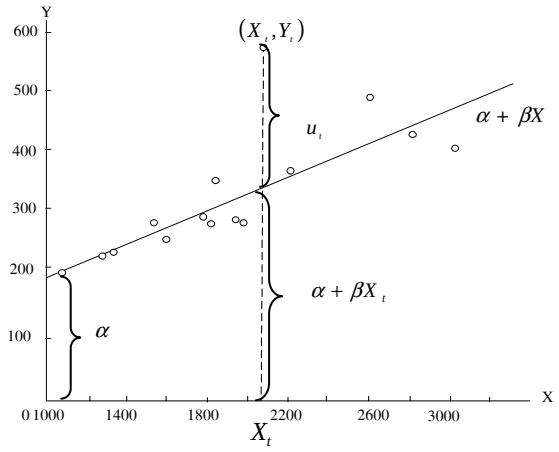
BẢNG 3.1 Giá trị trung bình ước lượng và trung bình thực tế của giá nhà và diện tích sử dụng (mét vuông)

t	SQFT	Giá bán[1]	Giá trung bình ước lượng[2]
1	1.065	199,9	200,386
2	1.254	288	226,657
3	1.300	235	233,051
4	1.577	285	271,554
5	1.600	239	274,751
6	1.750	293	295,601
7	1.800	285	302,551
8	1.870	365	312,281
9	1.935	295	321,316
10	1.948	290	323,123
11	2.254	385	365,657
12	2.600	505	413,751
13	2.800	425	441,551
14	3.000	415	469,351

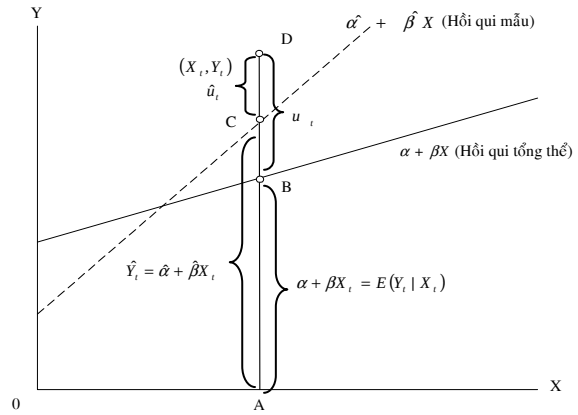
HÌNH 3.1 Biểu Đồ Phân Tán Của Mẫu Trình Bày Mối Liên Hệ Giữa Giá và SQFT

[1] Đơn vị tính: 1.000 đô la

[2] Phương pháp tính giá trung bình ước lượng sẽ được trình bày ở Phần 3.2



HÌNH 3.2 Phương Trình Hồi Quy của Tổng Thể và của Mẫu



Mục tiêu đầu tiên của một nhà kinh tế lượng là làm sao sử dụng dữ liệu thu thập được để ước lượng hàm hồi quy của tổng thể, đó là, ước lượng tham số của tổng thể α và β. Ký hiệu $\hat{\alpha}$ là ước lượng mẫu của α và $\hat{\beta}$ là ước lượng mẫu của β. Khi đó mối quan hệ trung bình ước lượng là $\hat{Y} = \hat{\alpha} + \hat{\beta}X$. Đây được gọi là **hàm hồi quy của mẫu**. Ứng với một giá trị quan sát cho trước *t*, ta sẽ có $\hat{Y}_t = \hat{\alpha} + \hat{\beta}X_t$. Đây là giá trị dự báo của *Y* với một giá trị cho trước là X_t. Lấy giá trị quan sát được Y_t trừ cho giá trị này, ta sẽ được ước lượng của u_t được gọi là **phần dư ước lượng**, hoặc đơn giản là **phần dư**, và ký hiệu là $\hat{u}_t$ [1]và được thể hiện trong phương trình sau:

$$\hat{u}_t = Y_t - \hat{Y}_t = Y_t - \hat{\alpha} - \hat{\beta}X_t$$

Sắp xếp lại các số hạng trên, ta có

[1] Một số tác giả và giảng viên thích sử dụng *a* thay cho $\hat{\alpha}$, *b* thay cho $\hat{\beta}$ và e_t thay cho $\hat{u}_t$. Chúng ta sử dụng dấu hiệu ^ theo qui định trong lý thuyết thống kê vì nó giúp phân biệt rõ ràng giữa giá trị thật và giá trị ước lượng và cũng xác định được thông số đang được ước lượng.

$$Y_t = \hat{\alpha} + \hat{\beta}X_t + \hat{u}_t \quad \textbf{(3.3)}$$

Việc phân biệt giữa *hàm hồi quy của* tổng thể $Y = \alpha + \beta X$ và *hàm hồi quy của mẫu* $\hat{Y}_t = \hat{\alpha} + \hat{\beta}X$ là rất quan trọng. Hình 3.2 trình bày cả hai đường và sai số và phần dư (cần nghiên cứu kỹ vấn đề này). Lưu ý rằng u_t là ký hiệu chỉ "sai số", và $\hat{u}_t$ là ký hiệu chỉ "phần dư".

BÀI TẬP 3.1

Xem xét các phương trình sau đây:

a. $Y_t = \alpha + \beta X + u_t$
b. $Y_t = \hat{\alpha} + \hat{\beta}X + \hat{u}_t$
c. $Y_t = \hat{\alpha} + \hat{\beta}X + u_t$
d. $\hat{Y}_t = \alpha + \beta X$
e. $\hat{Y}_t = \alpha + \beta X + \hat{u}_t$
f. $\hat{Y}_t = \hat{\alpha} + \hat{\beta}X + \hat{u}_t$

Giải thích kỹ tại sao phương trình (a) và (b) đúng, nhưng (c), (d), (e) và (f) sai. Hình 3.2 rất có ích trong việc trả lời câu hỏi này.

3.2 Ước lượng mô hình cơ bản bằng phương pháp bình phương tối thiểu thông thường

Trong phần trước, đã nêu rõ mô hình hồi quy tuyến tính cơ bản và phân biệt giữa hồi quy của tổng thể và hồi quy của mẫu. Mục tiêu tiếp theo sẽ là sử dụng các dữ liệu *X* và *Y* và tìm kiếm ước lượng "tốt nhất" của hai tham số của tổng thể là α và β. Trong kinh tế lượng, thủ tục ước lượng được dùng phổ biến nhất là **phương pháp bình phương tối thiểu**. Phương pháp này thường được gọi là **bình phương tối thiểu thông thường**, để phân biệt với những phương pháp bình phương tối thiểu khác sẽ được thảo luận trong các chương sau. Ký hiệu ước lượng của α và β là $\hat{\alpha}$ và $\hat{\beta}$, phần dư ước lượng thì bằng $\hat{u}_t = Y_t - \hat{\alpha} - \hat{\beta}X_t$. Tiêu chuẩn tối ưu được sử dụng bởi phương pháp bình phương tối thiểu là cực tiểu hóa hàm mục tiêu

$$ESS(\hat{\alpha}, \hat{\beta}) = \sum_{t=1}^{t=n} \hat{u}_t^2 = \sum_{t=1}^{t=n} (Y_t - \hat{\alpha} - \hat{\beta}X_t)^2$$

với các tham số chưa biết là $\hat{\alpha}$ và $\hat{\beta}$. ESS là tổng các phần dư bình phương và phương pháp OLS cực tiểu tổng các phần dư bình phương[2]. Cần nên lưu ý rằng ESS là khoảng

[2] Rất dễ nhầm khi gọi ESS là tổng của các phần dư bình phương, nhưng ký hiệu này được sử dụng phổ biến trong nhiều chương trình máy tính nổi tiếng và có từ tài liệu về Phân tích phương sai

cách bình phương được đo lường từ đường hồi quy. Sử dụng khoảng cách đo lường này, có thể nói rằng phương pháp OLS là tìm đường thẳng "gần nhất" với dữ liệu trên đồ thị.

Trực quan hơn, giả sử ta chọn một tập hợp những giá trị $\hat{\alpha}$ và $\hat{\beta}$, đó là một đường thẳng $\hat{\alpha}-\hat{\beta}X$. Có thể tính được độ lệch của Y_t từ đường thẳng được chọn theo phần dư ước lượng $\hat{u}_t = Y_t - \hat{\alpha} - \hat{\beta}X$. Sau đó bình phương giá trị này và cộng tất cả các giá trị bình phương của toàn bộ mẫu quan sát. Tổng các phần dư bình phương của các *trị quan sát* [được xem như **tổng bình phương sai số (ESS)**] do đó sẽ bằng $\sum \hat{u}_t^2$. Tương ứng với một điểm trên đường thẳng sẽ có một một trị tổng bình phương sai số. Phương pháp bình phương tối thiểu chọn những giá trị $\hat{\alpha}$ và $\hat{\beta}$ sao cho ESS là nhỏ nhất.

Việc bình phương sai số đạt được hai điều sau. Thứ nhất, bình phương giúp loại bỏ dấu của sai số và do đó xem sai số dương và sai số âm là như nhau. Thứ hai, bình phương tạo ra sự bất lợi cho sai số lớn một cách đáng kể. Ví dụ, giả sử phần dư của mẫu là 1, 2, –1 và –2 của hệ số hồi quy chọn trước trị $\hat{\alpha}$ và $\hat{\beta}$ chọn trước. So sánh các giá trị này với một mẫu khác có phần dư là –1, –1, –1 và 3. Tổng giá trị sai số tuyệt đối ở cả hai trường hợp là như nhau. Mặc dù mẫu chọn thứ hai có sai số tuyệt đối thấp hơn từ 2 đến 1, điều này dẫn đến sai số lớn không mong muốn là 3. Nếu ta tính ESS cho cả hai trường hợp thì ESS của trường hợp đầu là 10 ($1^2 + 2^2 + 1^2 + 2^2$), ESS cho trường hợp sau là 12 ($1^2 + 1^2 + 1^2 + 3^2$). Phương pháp bình phương tối thiểu áp đặt sự bất lợi lớn cho sai số lớn và do đó đường thẳng trong trường hợp đầu sẽ được chọn. Phần 3.3 sẽ tiếp tục trình bày những đặc tính cần thiết khác của phương pháp cực tiểu ESS.

Phương Pháp Thích Hợp Cực Đại

Phần này chỉ đề cập sơ về phương pháp thích hợp cực đại. Phương pháp này sẽ được trình bày chi tiết ở phần 2.A.4. Phần 3.A.5 sẽ trình bày nguyên tắc áp dụng mô hình hồi quy tuyến tính đơn. Mặc dù phương pháp thích hợp cực đại dựa trên một tiêu chuẩn tối ưu khác, nhưng các thông số ước lượng vẫn giống như các thông số ước lượng ở phương pháp OLS. Nói đơn giản, phương pháp thích hợp cực đại chọn ước lượng sao cho xác suất xảy ra của mẫu quan sát là lớn nhất.

Phần thảo luận trước cho thấy nếu thực hiện hai phương pháp ước lượng α và β khác nhau một cách chính xác thì đều dẫn đến cùng một kết quả. Như vậy thì tại sao cần phải xem xét cả hai phương pháp? Câu trả lời là trong các chương sau, ta sẽ thấy rằng khi một số giả thiết của mô hình được giảm nhẹ, thì thực tế, hai phương pháp ước lượng khác nhau sẽ cho kết quả khác nhau. Một phương pháp khác có thể cho kết quả khác nữa, đó là phương pháp cực tiểu tổng sai số tuyệt đối $\sum \left|\hat{u}_t\right|$. Nhưng phương pháp này không được dùng phổ biến trong kinh tế lượng vì khó tính toán.

Phương Trình Chuẩn

Trong phần 3.A.3 của phụ lục, phương pháp OLS được chính thức áp dụng. Phần này cho thấy rằng điều kiện để cực tiểu ESS với $\hat{\alpha}$ và $\hat{\beta}$ sẽ theo hai phương trình sau đây, được gọi là **phương trình chuẩn** (không có liên hệ gì đến phân phối chuẩn).

$$\sum \hat{u}_t = 0 = \sum (Y_t - \hat{\alpha} - \hat{\beta} X_t) = \sum Y_t - (n\hat{\alpha}) - \hat{\beta} \sum X_t \quad \textbf{(3.4)}$$

$$\sum (X_t \hat{u}_t) = \sum [X_t (Y_t - \hat{\alpha} - \hat{\beta} X_t)] = 0 \quad \textbf{(3.5)}$$

Trong Phương trình (3.4), cần lưu ý rằng $\sum \hat{\alpha} = n\hat{\alpha}$ bởi vì mỗi số hạng sẽ có một $\hat{\alpha}$ và có *n* số hạng. Chuyển vế các số hạng âm trong Phương trình (3.4) sang phải và chia mọi số hạng cho *n*, ta được

$$\frac{1}{n} \sum Y_t = \alpha + \beta \frac{1}{n} \sum X_t \quad \textbf{(3.6)}$$

(1/n)ΣY_t là trung bình mẫu của *Y*, ký hiệu là $\bar{Y}$, và *(1/n)*ΣY_t là trung bình mẫu của *X*, ký hiệu là $\bar{X}$. Sử dụng kết quả này thay vào Phương trình (3.6), ta được phương trình sau

$$\bar{Y} = \alpha + \beta \bar{X} \quad \textbf{(3.7)}$$

Đường thẳng $\hat{\alpha} + \hat{\beta} X$ là đường *ước lượng* và là **đường hồi quy của mẫu**, hoặc **đường thẳng thích hợp**. Có thể thấy rằng từ Phương trình (3.7) đường hồi quy của mẫu đi qua điểm trung bình $(\bar{X}, \bar{Y})$. Trong Bài tập 3.12c, ta sẽ thấy rằng tính chất này không đảm bảo trừ khi số hạng hằng số α có trong mô hình.

Từ Phương trình (3.5), cộng tất cả theo từng số hạng, và đưa $\hat{\alpha}$ và $\hat{\beta}$ ra làm thừa số chung, ta được

$$\sum (X_t Y_t) - \hat{\alpha} \sum X_t - \hat{\beta} \sum X_t^2 = 0$$

hay

$$\sum (X_t Y_t) = \hat{\alpha} \sum X_t + \hat{\beta} \sum X_t^2 \quad \textbf{(3.8)}$$

Lời Giải về Phương Trình Chuẩn

Để thuận lợi cho việc đáp án về hai phương trình chuẩn, các tính chất sau đây là rất cần thiết. Những tính chất này được chứng minh trong Phụ lục Phần 3.A.2

TÍNH CHẤT 3.1

$$S_{xx} = \sum (X_t - \bar{X})^2 = \sum X_t^2 - n\bar{X})^2 = \sum X_t^2 - \frac{1}{n} (\sum X_t)^2$$

TÍNH CHẤT 3.2

$$S_{xy} = \sum (X_t - \bar{X})(Y_t - \bar{Y}) = (\sum X_t Y_t) - n\,\bar{X}\bar{Y}$$
$$= \sum X_t Y_t - [(\sum X_t)(\sum Y_t) / n]$$

Từ Phương trình (3.7),

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{X} = \frac{1}{n}\sum Y_t - \hat{\beta}\frac{1}{n}\sum X_t \quad \textbf{(3.9)}$$

Thay $\hat{\alpha}$ vào (3.8)

$$\sum X_t Y_t = \left[\frac{1}{n}\sum Y_t - \hat{\beta}\frac{1}{n}\sum X_t\right](\sum X_t) + \hat{\beta}\sum X_t^2$$

Nhóm các số hạng có thừa số $\hat{\beta}$:

$$\sum X_t Y_t = \left[\frac{(\sum X_t)(\sum Y_t)}{n}\right] + \hat{\beta}\left[\sum X_t^2 - \frac{(\sum X_t)^2}{n}\right]$$

Tìm $\hat{\beta}$ ta được

$$\hat{\beta} = \frac{\sum X_t Y_t - \frac{(\sum X_t)(\sum Y_t)}{n}}{\sum X_t^2 - \frac{(\sum X_t)^2}{n}}$$

Sử dụng ký hiệu đơn giản đã được giới thiệu ở Tính chất 3.1 và 3.2, có thể được diễn tả như sau

$$\hat{\beta} = \frac{S_{xy}}{S_{xx}} \quad \textbf{(3.10)}$$

trong đó

$$S_{xx} = \sum X_t^2 - \frac{(\sum X_t)^2}{n} \quad \textbf{(3.11)}$$

và

$$S_{xy} = \sum X_t Y_t - \frac{(\sum X_t)(\sum Y_t)}{n} \quad \textbf{(3.12)}$$

Ký hiệu S_{xx} và S_{xy} có thể được nhớ một cách trực quan như sau, định nghĩa $x_t = X_t - \overline{X}$ và $y_t = Y_t - \overline{Y}$, trong đó ký hiệu thanh ngang chỉ trung bình của mẫu. Do đó x_t và y_t ký hiệu độ lệch giữa X và Y so với giá trị X và Y trung bình. Kết quả sau đây sẽ được chứng minh ở phần Phụ lục Phần 2.A.1 và 3.A.2.

$$\sum x_t = 0$$

$$S_{xx} = \sum x_t^2 = \sum (X_t - \overline{X})^2 = \sum X_t^2 - \frac{1}{n}(\sum X_t)^2 \quad \textbf{(3.13)}$$

$$S_{xy} = \sum x_t y_t = \sum (X_t - \overline{X})(Y_t - \overline{Y}) = \sum X_t Y_t - \frac{1}{n}\left[\left(\sum X_t\right)\left(\sum Y_t\right)\right] \quad \textbf{(3.14)}$$

S_{xy} là "tổng các giá trị của x_t nhân y_t ". Tương tự, S_{xx} "tổng các giá trị của x_t nhân x_t , hay tổng của x_t bình phương

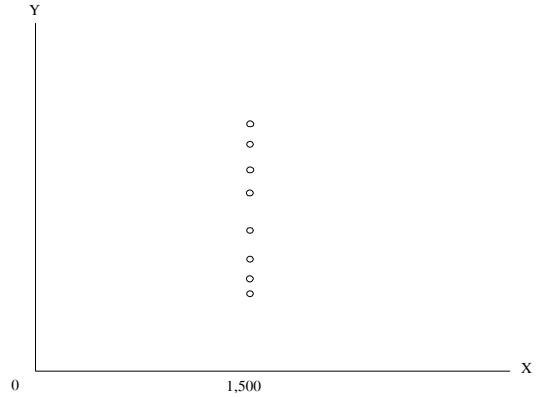
Phương trình (3.9) và (3.10) là lời giải cho phương trình chuẩn [(3.4) và (3.5)] và cho ta ước lượng $\hat{\alpha}$ và $\hat{\beta}$ của mẫu cho tham số α và β của tổng thể.

Cần lưu ý rằng không thể xác định được ước lượng của β trong Phương trình (3.10) nếu $S_{xx} = \sum x_t^2 = \sum (X_t - \overline{X})^2 = 0$. S_{xx} bằng không khi và chỉ khi mọi x_t bằng không, có nghĩa là khi và chỉ khi mọi X_t bằng nhau. Điều này dẫn đến giả thuyết sau đây

GIẢ THIẾT 3.2 (Các Giá Trị Quan Sát X Là Khác Nhau)
Không phải là tất cả giá trị X_t là bằng nhau. Có ít nhất một giá trị X_t khác so với những giá trị còn lại. Nói cách khác, phương sai của mẫu $Var\ (X) = \frac{1}{n-1}\sum (X_t - \overline{X})^2$ không được bằng không.

Đây là một giả thiết rất quan trọng và luôn luôn phải tuân theo bởi vì nếu không mô hình không thể ước lượng được. Một cách trực quan, nếu X_t không đổi, ta không thể giải thích được tại sao Y_t thay đổi. Hình 3.3 minh họa giả thuyết trên bằng hình ảnh. Trong ví dụ về địa ốc, giả sử thông tin thu thập chỉ tập trung một vào loại nhà có diện tích sử dụng là 1.500 mét vuông. Đồ thị phân tán của mẫu sẽ được thể hiện như ở Hình 3.3. Từ đồ thị có thể thấy rõ rằng dữ liệu này không đầy đủ cho việc ước lượng đường hồi quy tổng thể $\alpha + \beta X$.

HÌNH 3.3 Ví Dụ về Giá Trị X Không Đổi



Ví dụ 3.1

Theo thuật ngữ được dùng phổ biến trong kinh tế lượng, nếu ta sử dụng dữ liệu trong Bảng 3.1 và thực hiện "hồi quy *Y* (GIÁ) theo số hạng hằng số và *X* (SQFT)", ta có thể xác định được mối quan hệ ước lượng (hay **hàm hồi quy của mẫu**) là $\hat{Y}_t = 52{,}351 + 0{,}13875351 X_t$. $\hat{Y}_t$ là giá ước lượng trung bình (ngàn đô la) tương ứng

với X_t. (xem Bảng 3.1). Hệ số hồi quy của X_t là ảnh hưởng *cận biên* ước lượng của diện tích sử dụng đến giá nhà, ở mức trung bình. Do vậy, nếu diện tích sử dụng tăng lên một đơn vị, giá trung bình ước lượng kỳ vọng sẽ tăng thêm 0,13875 ngàn đô la ($138.75). Một cách thực tế, cứ mỗi 100 mét vuông tăng thêm diện tích sử dụng, giá bán ước lượng được kỳ vọng tăng thêm, mức trung bình, $ 13.875.

Hàm hồi quy của mẫu có thể được dùng để ước lượng giá nhà trung bình dựa trên diện tích sử dụng cho trước (Bảng 3.1 có trình bày giá trung bình ở cột cuối.) Do đó, một căn nhà có diện tích 1.800 mét vuông thì giá bán kỳ vọng trung bình là $302.551[= 52,351 + (0,139 × 1.800)]. Nhưng giá bán thực sự của căn nhà là $285.000. Mô hình đã ước lượng giá bán vượt quá $17.551. Ngược lại, đối với một căn nhà có diện tích sử dụng là 2.600 mét vuông, giá bán trung bình ước lượng là $413.751, thấp hơn giá bán thực sự $505.000 một cách đáng kể. Sự khác biệt này có thể xảy ra bởi vì chúng ta đã bỏ qua các yếu tố ảnh hưởng khác lên giá bán nhà. Ví dụ, một ngôi nhà có sân vườn rộng và/ hay hồ bơi, sẽ có giá cao hơn giá trung bình. Điều này nhấn mạnh tầm quan trọng trong việc nhận diện được các biến giải thích có thể ảnh hưởng đến giá trị của biến phụ thuộc và đưa các ảnh hưởng này vào mô hình được thiết lập. Ngoài ra, rất cần thiết trong việc phân tích độ tin cậy của các ước lượng của tung độ và hệ số độ dốc trong Phương trình (3.1), và mức độ "thích hợp" của mô hình đối với dữ liệu thực tế.

BÀI TẬP 3.2

Sao chép hai cột số liệu trong Bảng 3.1 vào một bảng mới. Trong cột đầu tiên của bảng tính sao chép các giá trị về Y_t (GIÁ) và X_t (SQFT) trong cột thứ hai. Sử dụng máy tính và tính thêm giá trị cho hai cột khác. Bình phương từng giá trị trong cột thứ hai và điền giá trị đó vào cột thứ ba (x). Nhân lần lượt từng giá trị ở cột thứ nhất với giá trị tương ứng ở cột hai và điền kết qua vào cột thứ tư (X_tY_t). Tiếp theo, tính tổng của từng cột và đánh giá các tổng sau đây:

$$\sum X_t = 26.753 \qquad \sum X_t^2 = 55.462.515$$

$$\sum Y_t = 4.444,9 \qquad \sum Y_t^2 = 9.095.985,5$$

Để tránh tình trạng quá nhiều và sai số làm tròn, cần sử dụng càng nhiều số thập phân càng tốt. Sau đó, tính S_{xy} từ Phương trình (3.12) và S_{xx} từ Phương trình (3.11). Cuối cùng, tính $\hat{\beta}$ theo (3.10) và $\hat{\alpha}$ theo (3.9) và kiểm tra lại những giá trị đã trình bày ban đầu.

3.3 Tính chất của các ước lượng

Mặc dù phương pháp bình phương cho ra kết quả ước lượng về mối quan hệ tuyến tính có thể phù hợp với dữ liệu sẵn có, chúng ta cần trả lời một số câu hỏi sau. Ví dụ, Đặc tính thống kê của $\hat{\alpha}$ và $\hat{\beta}$? Thông số nào được dùng để đo độ tin cậy của $\hat{\alpha}$ và $\hat{\beta}$? Bằng cách nào để có thể sử dụng $\hat{\alpha}$ và $\hat{\beta}$ để kiểm định giả thuyết thống kê và thực hiện dự báo? Sau đây chúng ta sẽ đi vào thảo luận từng vấn đề trên. Sẽ rất hữu ích nếu bạn ôn lại Phần 2.6, phần này đưa ra tóm tắt về những tính chất cần thiết của thông số ước lượng.

Tính chất đầu tiên cần xem xét là *độ không thiên lệch.* Cần lưu ý rằng trong Phần 2.4 các thông số ước lượng $\hat{\alpha}$ và $\hat{\beta}$? tự thân chúng là biến ngẫu nhiên và do đó tuân theo phân phối thống kê. Nguyên nhân là vì những lần thử khác nhau của một cuộc nghiên cứu sẽ cho các kết quả ước lượng thông số khác nhau . Nếu chúng ta lặp lại nghiên cứu với số lần thử lớn, ta có thể đạt được nhiều giá trị ước lượng. Sau đó chúng ta có thể tính tỷ số số lần mà những ước lượng này rơi vào một khoảng giá trị xác định. Kết quả sẽ sẽ cho ra phân phối của các ước lượng của mẫu. Phân phối này có giá trị trung bình và phương sai. Nếu trung bình của phân phối mẫu là thông số thực sự (trong trường hợp này là α hoặc β), thì đây là ước lượng không thiên lệch. Độ không thiên lệch rõ ràng là điều luôn được mong muốn bởi vì, điều đó có nghĩa là, ở mức trung bình, giá trị ước lượng sẽ bằng với giá trị thực tế, mặc dù trong một số trường hợp cá biệt thì điều này có thể không đúng.

Có thể nói rằng thông số ước lượng OLS của α và β đưa ra trong Phần 3.2 có tính chất không thiên lệch. Tuy nhiên, để chứng minh điều này, chúng ta cần đặt ra một số giả thuyết bổ sung về X_t và u_t. Cần nhớ rằng, mặc dù Giả thiết 3.1 có thể và được giảm nhẹ ở phần sau, nhưng Giả thuyết 3.2 và 3.3 là luôn luôn cần thiết và phải tuân theo. Sau đây là các giả thiết bổ sung cần thiết.

GIẢ THIẾT 3.3 (Sai Số Trung Bình bằng Zero)
Mỗi là *u* một biến ngẫu nhiên với E(u) = 0

Trong Hình 3.1 cần lưu ý rằng một số điểm quan sát nằm trên đường $\alpha + \beta X$ và một số điểm nằm dưới. Điều này có nghĩa là có một giá trị sai số mang dấu dương và một số sai số mang dấu âm. Do $\alpha + \beta X$ là đường trung bình, nên có thể giả định rằng các sai số ngẫu nhiên trên sẽ bị loại trừ nhau, ở mức trung bình, *trong* tổng thể. Vì thế, giả định rằng u_t là biến ngẫu nhiên với giá trị kỳ vọng bằng 0 là hoàn toàn thực tế.

GIẢ THIẾT 3.4 (Các Giá Trị *X* Được Cho Trước và Không Ngẫu Nhiên)
Mỗi giá trị X_t được cho trước và không là biến ngẫu nhiên. Điều này ngầm chỉ rằng đồng phương sai của tổng thể giữa X_t và u_t, $\text{Cov}(X_t, u_t) = E(X_t, u_t) - E(X_t)E(u_t) = X_tE(u_t) - X_tE(u_t) = 0$. Do đó giữa X_t và u_t không có mối tương quan (xem Định nghĩa 2.4 và 2.5).

Theo trực giác, nếu *X* và u có mối tương quan, thì khi *X* thay đổi, u cũng sẽ thay đổi. Trong trường hợp này, giá trị kỳ vọng của *Y* sẽ không bằng $\alpha + \beta X$. Nếu giá trị *X* là không ngẫu nhiên thì giá trị kỳ vọng có điều kiện của *Y* theo giá trị *X* sẽ bằng $\alpha + \beta X$. Kết quả của việc vi phạm Giả thiết 3.4 sẽ được trình bày trong phần sau, đặc biệt là khi nghiên cứu mô hình hệ phương trình (Chương 13). Tính chất 3.3 phát biểu rằng khi hai giả thiết được bổ sung, thông số ước lượng OLS là không thiên lệch.

TÍNH CHẤT 3.3
(Độ Không Thiên Lệch)

Trong hai giả thiết bổ sung 3.3 và 3.4, [$E(u_t) = 0$, $\text{Cov}(X_t, u_t) = 0$], thông số ước lượng, thông số ước lượng bình phương tối thiểu $\hat{\alpha}$ và $\hat{\beta}$ là không thiên lệch; nghĩa là $E(\hat{\alpha}) = \alpha$, và $E(\hat{\beta}) = \hat{\beta}$.

CHỨNG MINH (Nếu độc giả không quan tâm đến chứng minh, có thể bỏ qua phần).

Từ Phương trình (3.10), $E(\hat{\beta}) = E(S_{xy}/S_{xx})$. Nhưng theo Giả thuyết 3.4, X_t là không ngẫu nhiên và do đó S_{xx} cũng không ngẫu nhiên. Điều này có nghĩa là khi tính giá trị kỳ vọng, các số hạng liên quan đến X_t có thể được đưa ra ngoài giá trị kỳ vọng. Vì vậy, ta có $E(\hat{\beta}) = \frac{1}{S_{xx}} E(S_{xy})$. Trong Phương trình (3.12), thay Y_t từ Phương trình (3.1) và thay $\sum \alpha$ bằng $n\alpha$.

$$S_{xy} = \sum X_t(\alpha + \beta X_t + u_t) - \left[\frac{(\sum X_t)(n\alpha + \beta \sum X_t + \sum u_t)}{n}\right] \quad \textbf{(3.15)}$$

$$= \alpha \sum X_t + \beta \sum X_t^2 + \sum X_t u_t - \alpha \sum X_t - \beta \left[\frac{(\sum X_t)^2}{n}\right] - \left[\frac{(\sum X_t)(\sum u_t)}{n}\right]$$

$$= \beta \left[\sum X_t - \frac{(\sum X_t)^2}{n}\right] + \left[\sum X_t u_t - \frac{(\sum X_t)(\sum u_t)}{n}\right]$$

$$= \beta S_{xx} + S_{xu}$$

trong đó S_{xx} được cho bởi Phương trình (3.13) và

$$S_{xu} = \sum X_t u_t - \frac{(\sum X_t)(\sum u_t)}{n} \quad \textbf{(3.16)}$$

$$= \sum X_t u_t - \bar{X} \sum u_t = \sum (X_t - \bar{X}) u_t$$

$\bar{X}$ là trung bình mẫu của X, X_t là không ngẫu nhiên, $\bar{X}$ xuất hiện ở mọi số hạng, và kỳ vọng của tổng các số hạng thì bằng tổng các giá trị kỳ vọng. Do vậy,

$$E(S_{xu}) = \sum E(X_t u_t) - \bar{X} \sum E(u_t) = \sum X_t E(u_t) - \bar{X} \sum E(u_t) = 0$$

theo Giả thiết 3.3. Do đó, $E(S_{xy}) = \beta S_{xx}$, nghĩa là $E(\hat{\beta}) = E(S_{xy})/S_{xx} = \beta$. Như vậy β là ước lượng không thiên lệch của β. Chứng minh tương tự cho $\hat{\alpha}$. Cần nhận thấy rằng việc chứng minh độ không thiên lệch phụ thuộc chủ yếu vào Giả thiết 3.4. Nếu $E(X_t u_t) \neq 0$, $\hat{\beta}$ có thể bị thiên lệch.

BÀI TẬP 3.3

Sử dụng Phương trình (3.9) để chứng minh rằng $\hat{\alpha}$ là không thiên lệch. Nêu rõ các giả thuyết cần thiết khi chứng minh.

Mặc dầu độ không thiên lệch luôn là một tính chất luôn được mong muốn, nhưng tự bản thân độ không thiên lệch không làm cho thông số ước lượng "tốt", và một ước lượng không thiên lệch không chỉ là trường hợp cá biệt. Hãy xem xét ví dụ sau về một thông số ước lượng khác là $\tilde{\beta} = (Y_2 - Y_1)/(X_2 - X_1)$. Lưu ý rằng $\tilde{\beta}$ đơn giản là độ dốc của đường thẳng nối hai điểm (X_1, Y_1) và (X_2, Y_2). Rất dễ nhận thấy rằng $\tilde{\beta}$ là không thiên lệch

$$\tilde{\beta} = \frac{Y_2 - Y_1}{X_2 - X_1} = \frac{(\alpha + \beta X_2 + u_2) - (\alpha + \beta X_1 + u_1)}{X_2 - X_1} = \beta + \frac{u_2 - u_1}{X_2 - X_1}$$

Như đã nói trước đây, các giá trị X là không ngẫu nhiên và $E(u_2) = E(u_1) = 0$. Do đó, $\tilde{\beta}$ là không thiên lệch. Thực ra, ta có thể xây dựng một chuỗi vô hạn của các thông số ước lượng không thiên lệch như trên. Bởi vì $\tilde{\beta}$ loại bỏ các giá trị quan sát từ 3 đến n, một cách trực quan đây không thể là một thông số ước lượng "tốt". Trong Bài tập 3.6, tất cả các giá trị quan sát được sử dụng thể thiết lập các thông số ước lượng không thiên lệch khác, nhưng tương tự như trên đây không phải là là thông số ước lượng không thiên lệch tốt nhất. Do đó, rất cần có những tiêu chuẩn bổ sung để đánh giá "độ tốt" của một thông số ước lượng.

Tiêu chuẩn thứ hai cần xem xét là *tính nhất quán*, đây là một tính chất của mẫu lớn đã được định nghĩa trong Phần 2.6 (Định nghĩa 2.10). Giả sử ta chọn ngẫu nhiên một mẫu có n phần tử và đi tìm $\hat{\alpha}$ và $\hat{\beta}$. Sau đó chọn một mẫu lớn hơn và ước lượng lại các thông số này. Lặp lại quá trình này nhiều lần để có được một chuỗi những thông số ước lượng. Tính nhất quán là tính chất đòi hỏi các thông số ước lượng vẫn phù hợp khi cỡ mẫu tăng lên vô hạn. Ước lượng $\tilde{\beta}$ được trình bày ở trên rõ ràng là không đạt được tính nhất quán bởi vì khi cỡ mẫu tăng lên không ảnh hưởng gì đến thông số này. Tính chất 3.4 phát biểu các điều kiện để một ước lượng có tính nhất quán.

TÍNH CHẤT 3.4
(Tính Nhất Quán)

Theo Giả thiết (3.2), (3.3) và (3.4), ước lượng bình phương tối thiểu có tính chất nhất quán. Do đó, điều kiện để đạt được tính nhất quán là $E(u_t) = 0$, $\text{Cov}(X_t, u_t) = 0$ và $\text{Var}(X_t) \neq 0$.

CHỨNG MINH (Nếu độc giả không quan tâm, có thể bỏ qua phần này.)

Từ Phương trình (3.15) và (3.10)

$$\hat{\beta} = \beta + \frac{S_{xu}/n}{S_{xx}/n} \qquad \textbf{(3.17)}$$

Theo quy luật số lớn (Tính chất 2.7a), S_{xu}/n đồng quy với kỳ vọng của chính nó, đó là $\text{Cov}(X, u)$. Tương tự, S_{xx}/n đồng quy với Var(X). Do vậy dẫn tới điều, nếu n hội tụ đến vô

cùng, β sẽ đồng quy với β + [Cov*(X,u)*/Var*(X)*, và sẽ bằng β nếu Cov*(X,u)* = 0 – nghĩa là nếu X và u không tương quan. Như vậy, $\hat{\beta}$ là ước lượng nhất quán của β.

Mặc dù $\hat{\beta}$ là không thiên lệch và nhất quán, vẫn có những tiêu chuẩn cần bổ sung bởi để có thể xây dựng ước lượng nhất quán và không thiên lệch khác. Bài tập 3.6 là một ví dụ về loại ước lượng đó. Tiêu chuẩn sử dụng tiếp theo là *tính hiệu quả* (định nghĩa trong Phần 2.6). Nói một cách đơn giản, ước lượng không thiên lệch có tính hiệu quả hơn nếu ước lượng này có phương sai nhỏ hơn. Để thiết lập tính hiệu quả, cần có các giả thiết sau về u_t.

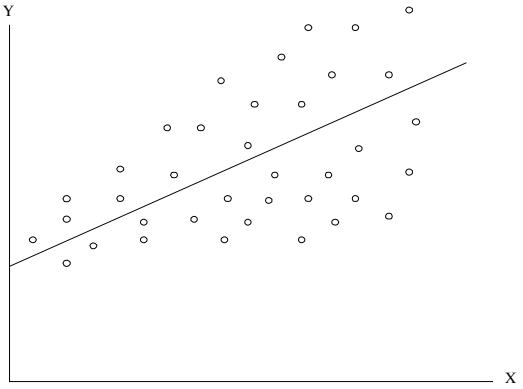
GIẢ THIẾT 3.5 (Phương sai của sai số không đổi)
Tất cả giá trị *u* được phân phối giống nhau với cùng phương sai σ^2, sao cho $Var(u_t) = E(u_t^2) = \sigma^2$. Điều này được gọi là phương sai của sai số không đổi (phân tán đều).

GIẢ THIẾT 3.6 (Độc Lập Theo Chuỗi)
Giá trị *u* được phân phối độc lập sao cho Cov*(u_t, u_s)* = *E($u_t u_s$)* = 0 đối với mọi $t \neq s$. Đây được gọi là chuỗi độc lập.

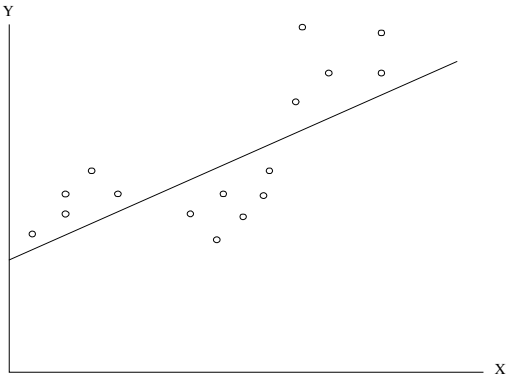
Các giả thiết trên ngầm chỉ rằng các phần dư phân có phân phối giống nhau và phân phối độc lập (iid). Từ Hình 1.2 ta thấy rằng ứng với một giá trị *X* sẽ có một giá trị phân phối Y để xác định phân phối *có điều kiện*. Sai số u_t là độ lệch từ *trung bình có điều kiện* $\alpha + \beta X_t$. Giả thiết 3.5 ngầm định rằng phân phối của u_t có cùng phương sai (σ^2) với phân phối của u_s cho một quan sát khác *s*. Hình 3.4a là một ví dụ về **phương sai của sai số thay đổi** (hoặc không phân tán đều) khi phương sai thay đổi tăng theo giá trị quan sát *X*. Giả thuyết 3.5 được giảm nhẹ trong Chương 8. Phần 3.6 Phụ chương có trình bày mô tả ba chiều của giả thuyết này.

Giả thiết 3.6 (sẽ được giảm nhẹ trong Chương 9) ngầm định rằng là u_t và u_s độc lập và do vậy không có mối tương quan. Cụ thể là, các sai số liên tiếp nhau không tương quan nhau và không tập trung. Hình 3.4b là một ví dụ về **tự tương quan** khi giả thuyết trên bị vi phạm. Chú ý rằng khi các giá trị quan sát kế tiếp nhau tập trung lại, thì có khả năng các sai số sẽ có tương quan.

HÌNH 3.4 Ví Dụ về Phương Sai Của Sai Số Thay Đổi và Tự Hồi Quy



a. Phương sai của sai số thay đổi



b. Tự hồi quy

TÍNH CHẤT 3.5
(Hiệu quả, BLUE và Định lý Gauss-Markov)

Theo Giả thiết 3.2 đến 3.6, ước lượng bình phương tối thiểu thông thường (OLS) là ước lượng tuyến tính không thiên lệch có hiệu quả nhất trong các ước lượng. Vì thế phương pháp OLS đưa ra **Ước Lượng Không Thiên lệch Tuyến Tính Tốt Nhất (BLUE).**

Kết quả này (được chứng minh trong Phần 3.A.4) được gọi là **Định lý Gauss–Markov**, theo lý thuyết này ước lượng OLS là BLUE; nghĩa là trong tất cả các tổ hợp tuyến tính không thiên lệch của *Y*, ước lượng OLS của α và β có phương sai bé nhất.

Tóm lại, áp dụng phương pháp bình phương tối thiểu (OLS) để ước lượng hệ số hồi quy của một mô hình mang lại một số tính chất mong muốn sau: ước lượng là (1) không thiên lệch, (2) có tính nhất quán và (3) có hiệu quả nhất. Độ không thiên lệch và tính nhất quán đòi hỏi phải kèm theo Giả thuyết $E(u_t) = 0$ và $\text{Cov}(X_t, u_t) = 0$. Yêu cầu về tính hiệu quả và BLUE, thì cần có thêm giả thuyết, $\text{Var}(u_t) = \sigma^2$ và $\text{Cov}(u_t, u_s) = 0$, với mọi $t \neq s$.

3.4 Độ Chính Xác của Ước Lượng và Mức Độ Thích Hợp của Mô Hình

Sử dụng các dữ liệu trong ví dụ về địa ốc ta ước lượng được thông số như sau $\hat{\alpha} = 52.351$ và $\hat{\beta} = 0{,}13875$. Câu hỏi cơ bản là các ước lượng này tốt như thế nào và mức độ thích hợp của hàm hồi quy mẫu $\hat{Y}_t = 52{,}351 + 0{,}13875351\,X$ với dữ liệu ra sao. Phần này sẽ thảo luận phương pháp xác định thông số đo lường độ chính xác của các ước lượng cũng như **độ phù hợp**.

Độ Chính Xác của Các Ước Lượng

Từ lý thuyết xác suất ta biết rằng phương sai của một biến ngẫu nhiên đo lường sự phân tán xung quanh giá trị trung bình. Phương sai càng bé, ở mức trung bình, từng giá trị riêng biệt càng gần với giá trị trung bình. Tương tự, khi đề cập đến khoảng tin cậy, ta biết rằng phương sai của biến ngẫu nhiên càng nhỏ, khoảng tin cậy của các tham số càng bé. Như vậy, phương sai của một ước lượng là thông số để chỉ độ chính xác của một ước lượng. Do đó việc tính toán phương sai của $\hat{\alpha}$ và $\hat{\beta}$ là luôn cần thiết.

Do $\hat{\alpha}$ và $\hat{\beta}$ thuộc vào các giá trị *Y*, mà *Y* lại phụ thuộc vào các biến ngẫu nhiên u_1, u_2, …, u_n, nên chúng cũng là biến ngẫu nhiên với phân phối tương ứng. Sau đây các phương trình được rút ra trong Phần 3.A.6 ở phần phụ lục của chương này.

$$Var(\hat{\beta}) = \sigma_{\hat{\beta}}^2 = E\left[\left(\hat{\beta} - \beta\right)^2\right] = \frac{\sigma^2}{S_{xx}} \quad \textbf{(3.18)}$$

$$Var(\hat{\alpha}) = \sigma_{\hat{\alpha}}^2 = E\left[(\hat{\alpha} - \alpha)^2\right] = \frac{\sum X_t^2}{nS_{xx}}\sigma^2 \quad \textbf{(3.19)}$$

$$Cov(\hat{\alpha}, \hat{\beta}) = \sigma_{\hat{\alpha}\hat{\beta}} = E\left[(\hat{\alpha} - \alpha)(\hat{\beta} - \beta)\right] = -\frac{\overline{X}}{S_{xx}}\sigma^2 \quad \textbf{(3.20)}$$

trong đó S_{xx} được định nghĩa theo Phương trình (3.11) và σ^2 là phương sai của sai số. Cần lưu ý rằng nếu S_{xx} tăng, giá trị phương sai và đồng phương sai (trị tuyệt đối) sẽ giảm. Điều này cho thấy *sự biến thiên ở X càng cao và cỡ mẫu càng lớn thì càng tốt bởi vì điều đó cho chứng tỏ độ chính của các thông số được ước lượng*.

Các biểu thức trên là **phương sai của tổng thể** và là ẩn số bởi vì σ^2 là ẩn số. Tuy nhiên, các thông số này có thể được ước lượng bởi vì σ^2 có thể được ước lượng dựa trên mẫu. Lưu ý rằng $\hat{Y}_t = \hat{\alpha} + \hat{\beta}X_t$ là đường thẳng ước lượng. Do đó, $\hat{u}_t = \hat{Y}_t - \hat{\alpha} - \hat{\beta}X_t$ là một ước lượng của u_t, và là *phần dư ước lượng*. Một ước lượng dễ thấy của σ^2 là $\sum \hat{u}_t^2 / n$ nhưng ước lượng này ngẫu nhiên bị thiên lệch. Một ước lượng khác của σ^2 được cho sau đây (xem chứng minh ở Phần 3.A.7)

$$s^2 = \hat{\sigma}^2 = \frac{\sum \hat{u}_t^2}{n-2} \quad \textbf{(3.21)}$$

Lý do chia tử số cho n – 2 thì tương tự như trường hợp chia chi-square cho n – 1, đã được thảo luận trong Phần 2.7. n – 1 được áp dụng do $\sum(x_i - \bar{x})$ có điều kiện là bằng 0. Để áp dụng chia cho n – 2, cần có hai điều kiện bởi Phương trình (3.4) và (3.5). Căn bậc hai của phương sai ước lượng được gọi là **sai số chuẩn của phần dư** hay **sai số chuẩn của hồi quy.** Sử dụng ước lượng này, ta tính được các ước lượng của phương sai và đồng phương sai của $\hat{\alpha}$ và $\hat{\beta}$. Căn bậc hai của phương sai được gọi là **sai số chuẩn của hệ số hồi quy** và ký hiệu $s_{\hat{\alpha}}$ và $s_{\hat{\beta}}$. Phương sai ước lượng và đồng phương sai của hệ số hồi quy ước lượng bằng

$$s_{\hat{\beta}}^2 = \frac{\hat{\sigma}^2}{S_{xx}} \quad \textbf{(3.22)}$$

$$s_{\hat{\alpha}}^2 = \frac{\sum X_t^2}{nS_{xx}}\hat{\sigma}^2 \quad \textbf{(3.23)}$$

$$s_{\alpha\beta} = -\frac{\overline{X}}{S_{xx}}\sigma^2 \quad \textbf{(3.24)}$$

Tóm lại: Trước tiên, cần tính hệ số hồi quy ước lượng $\hat{\alpha}$ và $\hat{\beta}$ bằng cách áp dụng Phương trình (3.9) và (3.10). Kết quả cho cho mối quan hệ ước lượng giữa *Y* và *X*. sau đó tính giá trị dự báo của Y_t theo $\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$. Từ đó, ta có thể tính được phần dư $\hat{u}_t$ theo $Y_t - \hat{Y}_t$. Sau đó tính toán ước lượng của phương sai của u_t dựa theo Phương trình (3.21). Thay kết quả vào Phương trình (3.18), (3.19) và (3.20), ta được giá trị phương sai và đồng phương sai của $\hat{\alpha}$ và $\hat{\beta}$.

Cần lưu ý rằng để công thức tính phương sai của phần dư s^2 được cho trong Phương trình 3.21 có ý nghĩa, cần có điều kiện n > 2. Không có giả thuyết này, phương sai được ước lượng có thể không xác định được hoặc âm. Điều kiện tổng quát hơn được phát biểu trong Giả thuyết 3.7, và bắt buộc phải tuân theo.

GIẢ THIẾT 3.7 (n > 2)
Số lượng quan sát (*n)* phải lớn hơn số lượng các hệ số hồi quy được ước lượng (*k*). Trong trường hợp hồi quy tuyến tính đơn biến, thì điều kiện n > 2 không có.

Ví dụ 3.2

Sau đây là sai số chuẩn trong ví dụ về giá nhà,
Sai số chuẩn của phần dư = s = $\hat{\sigma}$ = 39,023
Sai số chuẩn của $\hat{\alpha} = s_{\hat{\alpha}} = 37{,}285$
Sai số chuẩn của $\hat{\beta} = s_{\hat{\beta}} = 0{,}01873$
Đồng phương sai giữa $\hat{\alpha}$ và $\hat{\beta} = s_{\hat{\alpha}\hat{\beta}} = -0{,}671$
Thực hành máy tính Phần 3.1 của Phụ chương D sẽ cho kết quả tương tự.

Mặc dù có các đại lượng đo lường số học về độ chính xác của các ước lượng, tự thân các đo lường này không sử dụng được bởi vì các đo lường này có thể lớn hoặc nhỏ một cách tùy tiện bằng cách đơn giản là thay đổi đơn vị đo lường (xem thêm ở Phần 3.6). Các đo lường này được sử dụng chủ yếu trong việc kiểm định giả thuyết, đề tài này sẽ được thảo luận chi tiết ở Phần 3.5.

Độ Thích Hợp Tổng Quát

Hình 3.1 cho thấy rõ rằng không có đường thẳng nào hoàn toàn "thích hợp" với các dữ liệu bởi vì có nhiều giá trị dự báo bởi đường thẳng cách xa với giá trị thực tế. Để có thể đánh giá một mối quan hệ tuyến tính mô tả những giá trị quan sát có tốt hơn một mối quan hệ tuyến tính khác hay không, cần phải có một đo lường toán học **độ thích hợp.** Phần này sẽ phát triển các thông số đo lường đó.

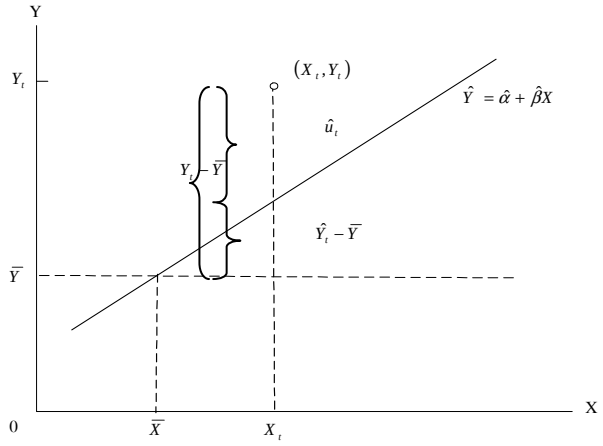
Khi thực hiện dự báo về một biến phụ thuộc *Y*, nếu ta chỉ có những thông tin về các giá trị quan sát của *Y* có được từ một số phân phối xác suất, thì có lẽ cách tốt nhất có thể là là ước lượng giá trị trung bình $\overline{Y}$ và phương sai sử dụng $\hat{\sigma}_Y^2 = \left[\sum \left(Y_t - \overline{Y}\right)^2\right] / (n-1)$. Nếu cần dự báo, một cách đơn giản, ta có thể sử dụng giá trị trung bình bởi vì không còn thông tin nào khác. Sai số khi dự báo quan sát thứ t bằng $Y_t - \overline{Y}$. Bình phương giá trị này

và tính tổng bình phương cho tất cả mẫu, ta tính được **tổng phương sai** của Y_t so với $\overline{Y}$ là $\sum\left(Y - \overline{Y}\right)^2$. Đây là **tổng bình phương toàn phần (TSS).** Độ lệch chuẩn của mẫu của *Y* đo lường độ phân tán của Y_t xung quanh giá trị trung bình của *Y*, nói cách khác là độ phân tán của sai số khi sử dụng $\overline{Y}$ làm biến dự báo, và được cho như sau $\hat{\sigma}_Y = \sqrt{TSS/(n-1)}$

Giả sử ta cho rằng *Y* có liên quan đến một biến X khác theo Phương trình (3.1). Ta có thể hy vọng rằng biết trước giá trị *X* sẽ giúp dự báo *Y* tốt hơn là chỉ dùng $\overline{Y}$. Cụ thể hơn là, nếu ta có các ước lượng $\hat{\alpha}$ và $\hat{\beta}$ và biết được giá trị của *X* là X_t, như vậy ước lượng của Y_t sẽ là $\hat{Y}_t = \hat{\alpha} + \hat{\beta}X_t$. Sai số của ước lượng này là $\hat{u}_t = Y_t - \hat{Y}_t$. Bình phương giá trị sai số này và tính tổng các sai số cho toàn bộ mẫu, ta có được **tổng bình phương sai số (ESS),** hay **tổng các bình phương phần dư,** là ESS = $\sum \hat{u}_t^2$. Sai số chuẩn của các phần dư là $\hat{\sigma} = \sqrt{ESS(n-2)}$. Giá trị này đo lường độ phân tán của sai số khi sử dụng $\hat{Y}_t$ làm biến dự báo và thường được so sánh với $\hat{\sigma}_Y$ được cho ở trên để xem xét mức độ giảm xuống là bao nhiêu. Bởi vì ESS càng nhỏ càng tốt, và mức độ giảm xuống càng nhiều. Trong ví dụ đưa ra, $\hat{\sigma}_Y = 88{,}498$ và $\hat{\sigma} = 39{,}023$, giảm hơn phân nửa so với giá trị ban đầu.

Phương pháp này không hoàn toàn tốt lắm, tuy nhiên bởi vì các sai số chuẩn rất nhạy cảm đối với đơn vị đo lường Y nên rất cần có một thông số đo lường khác không nhạy cảm với đơn vị đo lường. Vấn đề này sẽ được đề cập sau đây.

HÌNH 3.5 Các Thành Phần của Y



Thông số đo lường tổng biến thiên của $\hat{Y}_t$ so với $\overline{Y}$ (là giá trị trung bình của $\hat{Y}_t$) cho toàn mẫu là $\sum\left(\hat{Y}_t - \overline{Y}\right)^2$. Được gọi là **tổng bình phương hồi quy** (RSS). Phần 3.A.8 cho thấy

$$\sum\left(Y_t - \overline{Y}\right)^2 = \sum\left(\hat{Y}_t - \overline{Y}\right)^2 + \sum \hat{u}_t^2 \qquad \textbf{(3.25)}$$

Do vậy, TSS = RSS + ESS. Lưu ý rằng $(Y_t - \bar{Y}) = (\hat{Y}_t - \bar{Y}) + \hat{u}_t$. Hình 3.5 minh họa các thành phần trên. Phương trình (3.25) phát biểu rằng các thành phần cũng được bình phương. Nếu mối quan hệ giữa *X* và *Y* là "chặt chẽ", các điểm phân tán (X_t, Y_t) sẽ nằm gần đường thẳng $\hat{\alpha} + \hat{\beta}X$. nói cách khác ESS sẽ càng nhỏ và RSS càng lớn. Tỷ số

$$\frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

được gọi là **hệ số xác định đa biến** và ký hiệu là R^2. Thuật ngữ *đa biến* không áp dụng trong hồi quy đơn biến bởi vì chỉ có duy nhất một biến phụ độc lập *X*. Tuy nhiên, do biểu thức R^2 trong hồi quy đơn biến cũng giống như trong hồi quy đa biến nên ở đây chúng ta dùng cùng thuật ngữ

$$R^2 = 1 - \frac{\sum \hat{u}_t^2}{\sum (Y_t - \bar{Y})} = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS} \qquad 0 \le R^2 \le 1 \qquad \textbf{(3.26)}$$

Rõ ràng rằng, R^2 nằm giữa khoảng từ 0 đến 1. R^2 không có thứ nguyên vì cả tử số và mẫu số đều có cùng đơn vị. Điểm quan sát càng gần đường thẳng ước lượng, "độ thích hợp" càng cao, nghĩa là ESS càng nhỏ và R^2 càng lớn. Do vậy, R^2 là thông số đo lường độ thích hợp, R^2 càng cao càng tốt. ESS còn được gọi là **biến thiên không giải thích được** bởi vì $\hat{u}_t$ là ảnh hưởng của những biến khác ngoài X_t và không có trong mô hình. RSS là **biến thiên giải thích được**. Như vậy, TSS, là tổng biến thiên của *Y*, có thể phân thành hai thành phần: (1) RSS, là phần giải thích được theo *X*; và (2) ESS, là phần không giải thích được. Giá trị R^2 nhỏ nghĩa là có nhiều sự biến thiên ở *Y* không thể giải thích được bằng *X*. Ta cần phải thêm vào những biến khác có ảnh hưởng đến Y.

Ngoài ý nghĩa là một tỷ lệ của tổng biến thiên của *Y* được giải thích qua mô hình, R^2 còn có một ý nghĩa khác. Đó là thông số đo lường mối tương quan giữa giá trị quan sát Y_t và giá trị dự báo $\hat{Y}_t (r_{Y_t\hat{Y}_t})$. Cần xem lại phần trình bày về hệ số tương quan của mẫu và của tổng thể ở Phần 2.3 và 3.5. Phần 3.A.9 trình bày

$$r_{Y\hat{Y}}{}^2 = \frac{\widehat{Cov}^2(Y_t \hat{Y}_t)}{\widehat{Var}(Y_t)\widehat{Var}(\hat{Y}_t)} = \frac{RSS}{TSS} = R^2 \qquad \textbf{(3.26a)}$$

Như vậy, bình phương hệ số tương quan đơn biến giữa giá trị quan sát Y_t và giá trị dự báo $\hat{Y}_t$ bằng phương trình hồi quy thì sẽ cho ra kết quả bằng với giá trị R^2 được định nghĩa trong Phương trình (3.26a). Kết quả này vẫn đúng trong trường hợp có nhiều biến giải thích, *miễn là trong hồi quy có một số hạng hằng số.*

Có một thắc mắc phổ biến về độ thích hợp tổng thể, đó là "bằng cách nào để xác định rằng R^2 là cao hay thấp?". Không có một quy định chuẩn hay nhanh chóng để kết luận về R^2 như thế nào là cao hay thấp. Với chuỗi dữ liệu theo thời gian, kết quả R^2 thường lớn bởi vì có nhiều biến theo thời gian chịu ảnh hưởng xu hướng và tương quan với nhau rất nhiều. Do đó, giá trị quan sát R^2 thường lớn hơn 0.9. R^2 bé hơn 0.6 và 0.7 được xem là thấp. Tuy nhiên, đối với dữ liệu chéo, đại diện cho dạng của một yếu tố thay đổi vào một

thời điểm nào đó, thì R^2 thường thấp. Trong nhiều trường hợp, R^2 bằng 0.6 hoặc 0.7 thì chưa hẳn là xấu. Đây đơn giản chỉ là thông số đo lường về tính đầy đủ của mô hình. Điều quan trọng hơn là nên đánh giá mô hình xem dấu của hệ số hồi quy có phù hợp với các lý thuyết kinh tế, trực giác và kinh nghiệm của người nghiên cứu hay không.

Ví dụ 3.3

Trong bài tập về giá nhà, TSS, ESS và R^2 có các giá trị sau (xem lại kết quả ở Phần thực hành máy tính 3.1):

TSS = 101.815 ESS = 18.274 R^2 = 0,82052

Như vậy, 82,1% độ biến thiên của giá nhà trong mẫu được giải thích bởi diện tích sử dụng tương ứng. Trong chương 4, sẽ thấy rằng thêm vào các biến giải thích khác, như số lượng phòng ngủ và phòng tắm sẽ cải thiện độ thích hợp của mô hình.

3.5 Kiểm Định Giả Thuyết Thống Kê

Như đã đề lúc đầu, kiểm định giả thuyết thống kê là một trong những nhiệm vụ chính của nhà kinh tế lượng. Trong mô hình hồi quy (3.1), nếu β bằng 0, giá trị dự báo của Y sẽ độc lập với X, nghĩa là X không có ảnh hưởng đối với Y. Do đó, cần có giả thuyết $\beta = 0$, và ta kỳ vọng rằng giả thuyết này sẽ bị bác bỏ. Hệ số tương quan (ρ) giữa hai biến X và Y đo lường độ tương ứng giữa hai biến. Ước lượng mẫu của ρ được cho trong Phương trình (2.11). Nếu $\rho = 0$, các biến không có tương quan nhau. Do đó cũng cần kiểm định giả thuyết $\rho = 0$. Phần này chỉ thảo luận phương pháp kiểm định giả thuyết đối với α và β. Kiểm định giả thuyết đối với p sẽ được trình bày ở phần sau. Cần lưu ý rằng, trước khi tiếp tục phần tiếp theo, bạn nên xem lại Phần 2.8 về kiểm định giả thuyết và Phần 2.7 về các loại phân phối.

Kiểm định giả thuyết bao gồm ba bước cơ bản sau: (1) thiết lập hai giả thuyết trái ngược nhau (Giả thuyết không và Giả thuyết ngược lại), (2) đưa ra kiểm định thống kê và phân phối xác suất cho giả thuyết không, và (3) đưa ra quy luật ra quyết định để bác bỏ hay chấp nhận giả thuyết không. Trong ví dụ về giá nhà, Giả thuyết không là $H_o : \beta = 0$. Bởi vì chúng ta kỳ vọng rằng β sẽ dương, Giả thuyết ngược lại là H_1: β ≠0. Để thực hiện kiểm định này, $\hat{\beta}$ và sai số chuẩn ước lượng s được sử dụng để đưa ra thống kê kiểm định. Để đưa ra phân phối mẫu cho α và β, mà điều này ảnh hưởng gián tiếp đến các số hạng sai số ngẫu nhiên $u_1, u_2, \ldots u_n$ (xem Phương trình 3.15), cần bổ sung một giả thuyết về phân phối của u_t.

GIẢ THIẾT 3.8 (Tính Chuẩn Tắc của Sai Số)
Mọi giá trị sai số u_t tuân theo phân phối chuẩn $N(0, \sigma^2)$, nghĩa là mật độ có điều kiện của Y theo X tuân theo phân phối $N(\alpha + \beta X, \sigma^2)$.

Như vậy, các số hạng sai số $u_1, u_2, \ldots u_n$ được giả định là độc lập và có phân phối chuẩn giống nhau với giá trị trung bình bằng không và phương sai bằng σ^2. Giả thiết 3.8 là giả thiết căn bản trong kiểm định giả thuyết thống kê. Bảng 3.2 sẽ trình bày tóm tắt tất cả các

giả thiết đã được đưa ra. Những số hạng sai số thỏa các Giả thiết từ 3.2 đến 3.8 thì được xem là sai số ngẫu nhiên hay sai số do nhiễu trắng.

BẢNG 3.2 Các Giả Thiết của Mô Hình Hồi Quy Tuyến Tính Đơn Biến

3.1 Mô hình hồi quy là đường thẳng với ẩn số là các hệ số α và β; đó là $Y_t = \alpha + \beta X_t + u_t$, với $t = 1, 2, 3\ldots, n$.

3.2 Tất cả các giá trị quan sát X không được giống nhau; phải có ít nhất một giá trị khác biệt.

3.3 Sai số u_t là biến ngẫu nhiên với trung bình bằng không; nghĩa là, $E(u_t) = 0$.

3.4 X_t được cho và không ngẫu nhiên, điều này ngầm định rằng không tương quan với u_t; nghĩa là Cov $(X_t, u_t) = E(X_t u_t) - E(X_t)E(u_t) = 0$.

3.5 u_t có phương sai không đổi với mọi t; nghĩa là $\text{Var}(u_t) = E\left(u_t^2\right) = \sigma^2$

3.6 u_t và u_s có phân phối độc lập đối với mọi $t \neq s$, sao cho $\text{Cov}(u_t, u_s) = E(u_t u_s)$.

3.7 Số lượng quan sát (n) phải lớn hơn số lượng hệ số hồi quy được ước lượng (ở đây n > 2).

3.8 u_t tuân theo phân phối chuẩn $u_t \sim N(0, \sigma^2)$, nghĩa là ứng với giá trị X_t cho trước, $Y_t \sim N(\alpha + \beta X_t, \sigma^2)$.

Xác Định Trị Thống Kê Kiểm Định

Phần này chứng minh rằng kiểm định thống kê $t_c = \left(\hat{\beta} - \beta_0\right) \big/ s_{\hat{\beta}}$ tuân theo phân phối Student t, theo giả thuyết không, với bậc tự do là n – 2 (bởi vì ta đang ước lượng hai tham số α và β). Lưu ý rằng Giả thuyết 3.7 rất cần để chắc chắn rằng bậc tự do là dương.

CHỨNG MINH (Độc giả không quan tâm đến nguồn gốc vấn đề, có thể bỏ qua phần này).

Trước hết cần xem xét các tính chất sau

TÍNH CHẤT 3.6

a. $\hat{\alpha}$ và $\hat{\beta}$ có phân phối chuẩn.

b. $\left(\sum \hat{u}_t^2\right) \big/ \sigma^2 = \left[(n-2)\hat{\sigma}^2\right] \big/ \sigma^2$ có phân phối chi-bình phương với bậc tự do n–2.

c. $\hat{\alpha}$ và $\hat{\beta}$ được phân phối độc lập với $\hat{\sigma}^2$.

Tính chất 3.6a xuất phát từ thực tế là $\hat{\alpha}$ và $\hat{\beta}$ là những tổ hợp tuyết tính của u_t và u_t có phân phối chuẩn. Để chứng minh tính chất b và c, nên tham khảo tài liệu Hogg và Graig (1978, trang 296-298). Tận dụng các kết qua đó ta được

$$\hat{\alpha} \sim N(\alpha, \sigma_{\hat{\alpha}}^2), \qquad \hat{\beta} \sim N(\beta, \sigma_{\hat{\beta}}^2), \qquad \frac{\sum \hat{u}_t^2}{\sigma^2} \sim X_{n-2}^2$$

trong đó $\sigma^2_{\hat{\alpha}}$ và $\sigma^2_{\hat{\beta}}$ là phương sai của $\hat{\alpha}$ và $\hat{\beta}$ theo Phương trình (3.18) và (3.19). Bằng cách chuẩn hóa phân phối của thông số ước lượng – nghĩa là trừ cho trung bình và chia cho độ lệch chuẩn) – ta được

$$\frac{(\hat{\alpha} - \alpha)}{\hat{\sigma}_{\hat{\alpha}}} \sim N(0,1), \qquad \frac{(\hat{\beta} - \beta)}{\hat{\sigma}_{\hat{\beta}}} \sim N(0,1), \quad \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim X^2_{n-2}$$

Trong phần 2.7, phân phối t được định nghĩa là tỷ số của số chuẩn chuẩn hóa trên căn bậc hai của một chi-square độc lập với nó. Thay vào cho β và áp dụng phương trình (3.18), (3.19) và (3.22), ta được

$$t = \frac{(\hat{\beta} - \beta)}{\hat{\sigma}_{\hat{\beta}}} \div \left[\frac{\hat{\sigma}^2}{\sigma^2}\right]^{1/2} = \frac{\sigma(\hat{\beta} - \beta)}{\hat{\sigma}\sigma_{\hat{\beta}}} = \frac{(\hat{\beta} - \beta)}{s_{\hat{\beta}}} \sim t_{n-2}$$

trong đó

$$s_{\hat{\beta}} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}} = \frac{\hat{\sigma}}{\sigma}\frac{\sigma}{\sqrt{S_{xx}}} = \frac{\hat{\sigma}\sigma_{\hat{\beta}}}{\sigma}$$

$s_{\hat{\beta}}$ là sai số chuẩn ước lượng của $\hat{\beta}$ theo Phương trình (3.22).

t được trình bày ở trên là trị thống kê kiểm định dựa trên quy luật ra quyết định được thiết lập sau này. Kiểm định này được gọi là kiểm định *t*. Các bước kiểm định thống kê phân ra trong hai trường hợp kiểm định một phía và kiểm định hai phía được trình bày sau đây.

Quy Tắc Ra Quyết Định

Kiểm định t-test một phía

BƯỚC 1 H_0: β = β_0 H_1: β ≠β_0

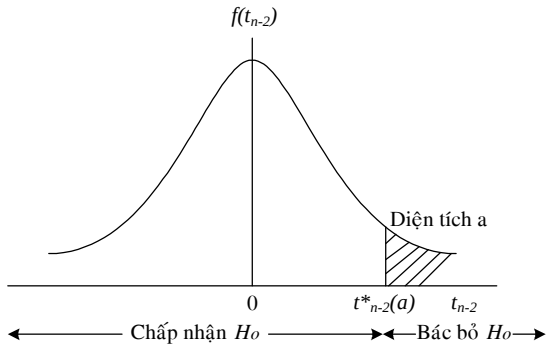
BƯỚC 2 Kiểm định thống kê là $t_c = (\hat{\beta} - \beta_0)/s_{\hat{\beta}}$, được tính dựa trên mẫu. Theo giả thuyết không, kiểm định thống kê có phân phối *t* với bậc tự do là n – 2. Nếu t_c tính được là "lớn", ta có thể nghi ngờ rằng β sẽ không bằng β_0. Điều này dẫn đến bước tiếp theo.

BƯỚC 3 Trong bảng tra phân phối t ở trang bìa trước của sách, tra bậc tự do là n – 2. Và chọn mức ý nghĩa (α) và xác định điểm $t^*_{n-2}(\alpha)$ sao cho $P(t > t^*) = \alpha$.

BƯỚC 4 Bác bỏ H_0 nếu $t_c > t^*$. Nếu giả thuyết ngược lại $\beta < \beta_0$, tiêu chuẩn kiểm định để bác bỏ H_0 là nếu $t_c < -t^*$.

Kiểm định trên được minh họa bằng hình ảnh qua Hình 3.6 (ký hiệu α được sử dụng để chỉ mức ý nghĩa để tránh nhầm lẫn với α chỉ tung độ). Nếu t_c rơi vào diện tích in đậm trong hình vẽ (được gọi là **vùng tới hạn)** nghĩa là t_c >t*. Trong trường hợp đó, giả thuyết không sẽ bị bác bỏ và kết luận được rằng β lớn hơn β_0 rất nhiều.

HÌNH 3.6 Kiểm Định Một Phía với H_0: $\beta = \beta_0$ H_1: $\beta \neq \beta_0$



Ví dụ 3.4

Trong ví dụ về giá nhà, ta có $\beta_0 = 0$. Do đó, $t_c = \hat{\beta} / s_{\hat{\beta}}$, là kiểm định thống kê đơn giản và là tỷ số giữa hệ số hồi quy ước lượng trên sai số chuẩn. Tỷ số được gọi là **trị thống kê *t*.** Các ước lượng là $\hat{\beta} = 0{,}13875$, và theo ví dụ 3.2 ta biết $s_{\hat{\beta}} = 0{,}01873$. Do đó, trị thống kê t được tính sẽ là t_c = 0,13875/0,01873 = 7.41. Bậc tự do bằng n – 2 = 14 – 2 = 12. Cho mức ý nghĩa là 1%, nghĩa là α = 1%. Tra bảng phân phối t, ta được $t^*_{n-2}=2{,}681$. Do t_c > t*, giả thuyết H_0 bị bác bỏ và kết luận được rằng β lớn hơn zero một cách đáng kể với mức ý nghĩa 1%. Lưu ý rằng hệ số này vẫn có ý nghĩa trong trường hợp mức ý nghĩa chỉ là 0,05% bởi vì $t^*_{12}(0{,}0005) = 4{,}318$.

Trị thống kê *t* đối với $\hat{\alpha}$ được cho bởi t_c = 52,351/37,285 = 1.404 nhỏ hơn $t^*_{12}(0{,}0005)$ = 1.782. Do đó không thể bác bỏ H_0 nhưng thay vào đó có thể có thể kết luận rằng α **không lớn hơn zero xét về mặt thống kê** với mức ý nghĩa 5%. Các điểm $\hat{\alpha}$ không nghĩa ở hai điểm sau. Thứ nhất, X = 0 thì hoàn toàn nằm ngoài khoảng mẫu và do đó ước lượng $\hat{Y}$ khi X = 0 không đáng tin cậy (xem thêm Phần 3.9). Thứ nhì, từ Hình 3.1 có thể thấy rằng đặc điểm hai biến là không đầy đủ để giải thích độ biến thiên giá của các giá trị quan sát. Trong chương 4 sẽ cho thấy $\hat{\alpha}$ bao hàm cả ảnh hưởng trung bình của biến bị bỏ sót và tính phi tuyến, khi X bằng 0. Các ảnh hưởng trên sẽ làm cho α không có ý nghĩa.

Một Số Lưu Ý khi Sử Dụng Kiểm Định t-Test

Mặc dù kiểm định *t*-test rất hữu ích trong việc xác định ý nghĩa thống kê của các hệ số, tuy nhiên rất dễ nhầm lẫn giữa các ý nghĩa của kiểm định. Ví dụ, ở Ví dụ 3.4 kiểm định *t*-test đối với α không thể bác bỏ giả thuyết không là α = 0. Như vậy có phải kiểm định này "chứng minh" rằng α = 0 hay không? Câu trả lời là không. Có thể chắc chắn rằng, *theo tập dữ liệu và mô hình được mô tả,* không có bằng chứng nào cho thấy α > 0. Trong chương 4, sẽ đề cập kiểm định *t*-test cho nhiều hệ số hồi quy. Nếu một trong những hệ số này không có ý nghĩa (nghĩa là, không thể bác bỏ giả thuyết rằng hệ số bằng 0), điều đó không có nghĩa là biến tương ứng không có ảnh hưởng gì đến biến phụ thuộc hoặc biến đó không quan trọng. Vấn đề này sẽ được thảo luận đầy đủ trong chương sau. Trong chương 5 sẽ thấy rằng khi mô hình thay đổi, mức ý nghĩa của hệ số cũng thay đổi. Do đó, cần thực hiện kỹ các kiểm định giả thuyết đưa ra và không nên vội vã kết luận mà không

xét đến mô hình và những phân tích thêm về các kiểm định chuẩn đoán cần thiết để đưa ra một kết luận ý nghĩa (ổn định theo đặc điểm mô hình).

Phương Pháp *p*-value trong Kiểm Định Giả thuyết

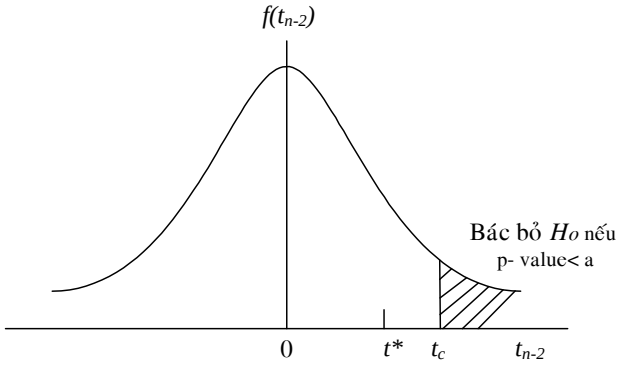
Kiểm định *t*-test có thể được thực hiện theo một phương pháp khác tương đương. Trước tiên tính xác suất để biến ngẫu nhiên *t* lớn hơn trị quan sát t_c, nghĩa là

$$p\text{-}value = P(t>t_c) = P \text{ (sai lầm loại I)}$$

Xác suất này (được gọi là *p*-**value**) là phần diện tích bên phải t_c trong phân phối *t* (xem Hình 3.7) và là xác suất sai lầm loại I – nghĩa là xác suất loại bỏ giả thuyết H_0. Xác suất này càng cao cho thấy hậu quả của việc loại bỏ sai lầm giả thuyết đúng H_0 càng nghiêm trọng. *p*-value bé nghĩa là hậu quả của việc loại bỏ giả thuyết đúng H_0 là không nghiêm trọng (nghĩa là, xác suất xảy ra sai lầm loại I là thấp) và do đó có thể yên tâm khi bác bỏ H_0. Như vậy, quy luật ra quyết định là *không bác bỏ* H_0 nếu *p* -value quá lớn, ví dụ: lớn hơn 0,1, 0,2, 0,3. Nói cách khác, nếu *p*-value lớn hơn mức ý nghĩa α, có thể kết luận rằng hệ số hồi quy không lớn hơn β_0 ở mức ý nghĩa α. Nếu *p*-value nhỏ hơn α, giả thuyết H_0 bị bác bỏ và kết luận được rằng β lớn hơn β_0 một cách đáng kể.

Để thấy được sự tương đương của hai phương pháp, lưu ý rằng trên Hình 3.7 nếu xác suất $P(t>t_c)$ bé hơn mức ý nghĩa α, thì điểm tương ứng là t_c phải nằm bên phải điểm $t^*_{n-2}(\alpha)$. Nghĩa là t_c rơi vào miền bác bỏ. Tương tự, nếu xác suất $P(t>t_c)$ lớn hơn mức ý nghĩa α, thì điểm tương ứng là t_c phải nằm bên trái điểm $t^*_{n-2}(\alpha)$ và do đó rơi vào miền chấp nhận. Sau đây là các bước bổ sung trong phương pháp *p*-value như sau:

HÌNH 3.7 Kiểm Định Giả thuyết theo Phương Pháp *p*-value



BƯỚC 3a Tính xác suất (ký hiệu *p*-value) để *t* lớn hơn t_c , nghĩa là tính phần diện tích bên phải giá trị t_c.

BƯỚC 4a Bác bỏ H_0 và kết luận rằng hệ số có ý nghĩa nếu *p*-value bé hơn mức ý nghĩa được chọn.

Tóm lại, β được xem là lớn hơn β_0 một cách đáng kể nếu trị thống kê *t* lớn hay *p*-value là bé, mức độ như thế nào là lớn và bé sẽ được quyết định bởi người nghiên cứu. Phương pháp phổ biến trong kiểm định giả thuyết là xác định giá trị mốc *t**. Tuy nhiên theo hương pháp tính *p*-value, lại cần tính toán phần diện tích một đầu ứng với giá trị t_c cho trước. Ngày càng có nhiều phần mềm máy tính tính toán sẵn *p*-value (chương trình SHAZAM và ESL được giới thiệu trong sách này) và do đó phương pháp này dễ ứng dụng dễ dàng. Tuy nhiên, cần cẩn thận kiểm tra lại giá trị *p*-value là dùng cho kiểm một phía hay kiểm định hai phía.

Ví dụ 3.4a

Để áp dụng phương pháp *p*-value cho ví dụ về giá nhà, ta tính xác suất để *t* lớn hơn giá trị quan sát β = 7.41. Sử dụng ESL để tính toán ta được *p* < 0,0001 (tham khảo phần kết quả trong phần Thực hành máy tính 3.1). Điều đó có nghĩa là, nếu ta bác bỏ giả thuyết không, thì cơ hội để xảy ra sai lầm loại I bé hơn 0,01%, và do đó hoàn toàn yên tâm khi bác bỏ H_o và kết luận được rằng β lớn hơn 0. Đối với tham số α, *p*-value bằng 0,093, nghĩa là $P(t>1{,}404) = 0{,}093$. Nếu H_0: $\alpha = 0$ bị bác bỏ, xác suất để xảy ra sai lầm loại I là 9,3%, lớn hơn 5%. Do đó, không thể bác bỏ H_0 ở mức ý nghĩa 5%, nghĩa là ta có cùng kết luận như trong phương pháp đầu, đó là ở mức ý nghĩa 5%, α không lớn hơn zero xét về mặt thống kê. Như vậy phương pháp *p*-value có một ưu điểm là, ta biết được chính xác mức độ mà hệ số có ý nghĩa và có thể đánh giá xem mức ý nghĩa này đủ thấp hay không để xem xét bác bỏ H_0. Cuối cùng, không cần lo lắng đối với các giá trị 0,01, 0,05 và 0,1.

Kiểm Định *t*-test Hai Phía

Bao gồm các bước sau:

BƯỚC 1 $H_0: \beta = \beta_0 \qquad H_1: \beta \neq \beta_0$

BƯỚC 2 Kiểm định thống kê là $t_c = \left(\hat{\beta} - \beta_0\right) \Big/ s_{\hat{\beta}}$, được tính dựa trên mẫu. Theo giả thuyết không, kiểm định thống kê có phân phối *t* là $t_{n-2.}$

BƯỚC 3 Trong bảng tra phân phối t ở trang bìa trước của sách, tra bậc tự do là n – 2 và chọn mức ý nghĩa (α) và xác định điểm $t^*_{n-2}(\alpha)$ sao cho $P(t>t^*) = \alpha/2$ (phân nửa mức ý nghĩa).

BƯỚC 3a Áp dụng phương pháp ρ- value, tính giá trị *p*

$$\rho\text{- value} = P(t > t_c \text{ hoặc } t < -t_c) = 2P(t > |t_c|)$$

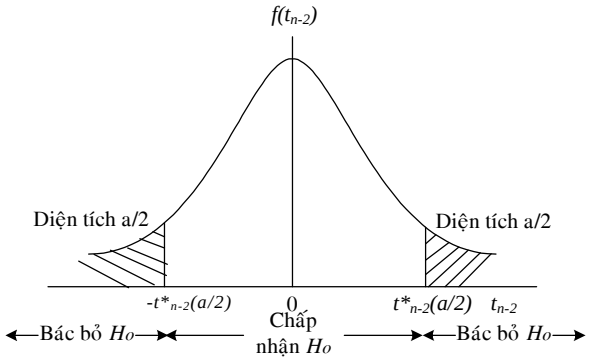
do phân phối t đối xứng.

BƯỚC 4 Bác bỏ H_0 nếu $|t_c| > t^*$ và kết luận **β khác với β_0 một cách đáng kể ở mức ý nghĩa α.**

BƯỚC 4a Bác bỏ H_0 nếu *p*-value < α, ở mức ý nghĩa này.

Kiểm định trên được minh họa bằng hình ảnh qua Hình 3.8. Bậc tự do trong trường hợp này bằng n–2. Nếu trị thống kê t (t_c) rơi vào vùng diện tích đen, giả thuyết không bị bác bỏ và kết luận được rằng β khác với β_0. giá trị t* = 2 được sử dụng là quy luật để đánh giá mức ý nghĩa của trị thống kê *t* ở mức 5% (kiểm định hai phía). Bởi vì *t** gần bằng 2 với bậc tự do là 25.

HÌNH 3.8 Kiểm Định Hai Phía với H_0: $\beta = \beta_0$ H_1: $\beta \neq \beta_0$



Ví dụ 3.5

Theo cách tính này t_c trong ví dụ giá nhà có giá trị như cách tính theo *t*-test, $\hat{\beta} = 7.41$ và $\hat{\alpha} = 1.404$. Tra bảng giá trị *t*, ta có $t^*_{12}(0.005) = 3.055$, điều này có nghĩa là diện tích của cả 2 phía tương ứng với giá trị 3.055 là 0.01. Bởi đối với $\hat{\beta}$ thì $t_c > t^*$ do đó ta có thể loại giả thuyết H_0 và kết luận được rằng β khác với ở mức ý nghĩa 1%. Đối với $\hat{\alpha}$ thì $t^*_{12}(0.025) = 2.179$ lớn hơn giá trị t_c. Do đó ta không thể bác bỏ giả thuyết H_0 (lưu ý rằng ta đang dùng kiểm định giá trị α ở mức ý nghĩa 5%). Từ bước 3a ta có thể suy ra được giá trị p-value đối với $\hat{\alpha} = 2P(t > 1.404) = 0.186$ (lưu ý giá trị *p*-value tương ứng với t_c trong trường hợp kiểm định 2 phía sẽ gấp 2 lần giá trị của nó trong trường hợp kiểm định 1 phía). Do sai lầm loại I có giá trị 18.6% là không thể chấp nhận được nên ta không thể bác bỏ giả thuyết H_0: $\alpha = 0$. Điều này có nghĩa là α không có ý nghĩa về thống kê trong khi β lại có.

BÀI TẬP 3.4

Trong ví dụ giá nhà, hãy kiểm định giả thuyết H_0: $\beta = 0.1$ và giả thuyết H_1: $\beta \neq 0.1$ lần lượt ở mức ý nghĩa 0.05 và 0.01.

BÀI TẬP 3.5

Chứng minh rằng nếu một hệ số có ý nghĩa ở mức 1% thì hệ số này cũng sẽ có ý nghĩa ở mức cao hơn.

BÀI TẬP 3.6

Hãy chứng minh rằng nếu một hệ số không có ý nghĩa ở mức 10% thì hệ số này cũng sẽ không có ý nghĩa ở bất kỳ mức ý nghĩa nào thấp hơn 10%.

Kiểm Định σ^2

Mặc dù thống kê kiểm định mức ý nghĩa phương sai sai số σ^2 không phổ biến nhưng vẫn được trình bày đầy đủ trong phần này. Kiểm định σ^2 gồm các bước sau:

BƯỚC 1 H_0: $\sigma^2 = \sigma^2_0$ H_1: $\sigma^2 \neq \sigma^2_0$

BƯỚC 2 Trị kiểm định là $Q_c = (n-2)\frac{\hat{\sigma}^2}{\hat{\sigma}^2_0}$. Sau đó tra bảng phân phối Chi-square với bậc tự do n-2. Nếu Q có giá trị "lớn" ta có thể nghi ngờ rằng σ^2 không bằng σ^2_0

BƯỚC 3 Trong bảng tra phân phối Chi-square ở trang bìa trước của sách, tra giá trị của $Q^*_{n-2}(\alpha)$ sao cho diện tích bên phải bằng α.

BƯỚC 4 Bác bỏ H_0 ở mức ý nghĩa α nếu $Q_c > Q^*_{n-2}(\alpha)$.

Nguyên nhân tổng quát làm cho kiểm định này không phổ biến là do người kiểm định không có thông tin sơ cấp ban đầu về giá trị của σ^2 sử dụng trong giả thuyết H_0.

Kiểm Định Độ Thích Hợp

Ta có thể thực hiện kiểm định độ thích hợp. Gọi p là hệ số tương quan tổng thể giữa X và Y được định nghĩa ở Phương trình (2.7). Theo phương trình (2.11), ta thấy giá trị ước lượng p^2 được xác định bởi $r^2_{xy} = S^2_{xy}/(S_{xx}S_{yy})$ trong đó S_{xx} và S_{xy} được định nghĩa theo Phương trình (3.8) và (3.9), và

$$S_{yy} = \sum Y_t^2 - \left[\frac{\left(\sum Y_t\right)^2}{n}\right] = \sum (Y_t - \bar{Y})^2 = TSS \qquad \textbf{(3.27)}$$

Ở Phần 3.A.10 người ta đã chứng minh rằng r^2_{xy} bằng với R^2 (điều này chỉ đúng trong trường hợp hồi qui đơn biến mà thôi). Ở Phần kiểm định giả thuyết 2.8 trình bày phương pháp kiểm định giả thuyết cho rằng X và Y không có mối tương quan. Kiểm định này gọi là **kiểm định F** (F-test)**.** Kiểm định F-test gồm các bước sau:

BƯỚC 1 H_0: $\rho_{xy} = 0$ H_1: $\rho_{xy} \neq 0$

BƯỚC 2 Trị thống kê kiểm định là $F_c = R^2(n-2)/(1-R^2)$. F_c cũng có thể được tính theo công thức sau F_c = RSS(n – 2)/ESS. Theo giả thuyết H_0, trị thống kê này tuân theo phân phối F với 1 bậc tự do ở tử số và n – 2 bậc tự do ở mẫu số.

BƯỚC 3 Tra bảng F theo 1 bậc tự ở tử số và n – 2 bậc tự do ở mẫu số tìm giá trị $F^*_{1,\,n-2}(\alpha)$ sao cho phần diện tích về phía phải của F^* là α, mức ý nghĩa.

BƯỚC 4 Bác bỏ giả thuyết H_0 (tại mức ý nghĩa α) nếu $F_c > F^*$.

Nên lưu ý rằng giả thuyết H_0 ở trên sẽ không hợp lệ khi có nhiều giá trị X. Như sẽ được trình bày ở chương 4, kiểm định F vẫn được sử dụng nhưng H_0 sẽ khác.

Ví dụ 3.6

Trong ví dụ giá nhà, R^2 = 0,82052. F_c = 0,82052(14 – 2)/(1 – 0,82052) = 54,86. Theo ví dụ 3.5, ESS = 18.274, và RSS = TSS – ESS = 83.541. Vì vậy F_c còn có thể được tính theo công thức khác như ở bước 2: F_c = 83.541 (14 – 2)/18.274 = 54,86. Bậc tự do của tử

số là 1, của mẫu số là 12. Với mức ý nghĩa $\alpha = 5\%$, tra bảng A.4b ta được $F^*_{1, 12}(0.05) = 4{,}75$. Vì $F_c > F^*$ chúng ta bác bỏ (tại mức ý nghĩa 5%) giả thuyết H_0 cho rằng X và Y không tương quan. Thực ra, vì $F_c > F^*_{1, 12}(0.01)$ (tra bảng A.4a), giả thuyết H_0 cũng bị bác bỏ tại mức ý nghĩa 1%. Như vậy, mặc dù giá trị R^2 khá nhỏ hơn 1, nó cũng khác 0 một đáng kể.

Trình Bày Các Kết Quả Hồi Quy

Các kết quả của phân tích hồi quy được trình bày theo nhiều cách. Theo cách thông thường, người ta sẽ viết phương trình ước lượng kèm với các trị thống kê t ở dưới mỗi hệ số hồi quy như sau:

$$\widehat{\text{GIÁ}} = \underset{(1,404)}{52{,}351} + \underset{(7,41)}{0{,}13875}\,\text{SQFT}$$

$$R^2 = 0.821 \qquad \text{d.f.} = 12 \qquad \sigma = 39.023$$

Một cách khác là điền các sai số chuẩn dưới các hệ số hồi quy:

$$\widehat{\text{GIÁ}} = \underset{(37.29)}{52{,}351} + \underset{(0.019)}{0{,}13875}\,\text{SQFT}$$

Nếu nhiều mô hình hồi quy được ước lượng, việc trình bày kết quả ở dạng bảng như Bảng 4.2 sẽ thuận tiện hơn.

Việc tách tổng các bình phương toàn phần ra thành các thành phần thường được tóm tắt ở dạng bảng **Phân Tích Phương Sai (ANOVA)** Bảng 3.3.

3.6 Thang Đo và Đơn Vị Đo

Giả sử chúng ta đã tính GIÁ theo đơn vị đồng đôla thay vì theo ngàn đồng đôla. Cột GIÁ ở bảng 3.1 sẽ chứa các giá trị như 199.900, 228.000, v.v. Những ước lượng của hệ số hồi quy, các sai số chuẩn của chúng, R^2, v.v. sẽ bị ảnh hưởng như thế nào bởi sự thay đổi đơn vị này? Câu hỏi này sẽ được khảo sát ở đây vì GIÁ và SQFT được tính ở các đơn vị khác nhau. Đầu tiên chúng ta chạy lại mô hình.

$$\text{GIÁ} = \alpha + \beta\,\text{SQFT} + u$$

Gọi GIÁ^* là giá tính theo đô la thường. Như vậy $\text{GIÁ}^* = 1.000\ \text{GIÁ}$. Nhân mọi số hạng trong phương trình với 1.000 và thay GIÁ^* vào vế trái. Chúng ta có

$$\text{GIÁ}^* = 1.000\alpha + 1.000\beta\,\text{SQFT} + 1.000\text{u} = \text{GIÁ}^* = \alpha^* + \beta^*\,\text{SQFT} + u^*$$

Nếu chúng ta áp dụng phương pháp OLS cho phương trình này và cực tiểu hóa $\Sigma (u^*_i)^2$, chúng ta sẽ tìm được các giá trị ước lượng của α^* và β^*. Dễ dàng nhận thấy rằng các hệ số hồi quy mới sẽ bằng các hệ số cũ nhân với 1,000. *Như vậy, thay đổi thang đo của chỉ biến phụ thuộc trong mô hình hồi quy làm cho thang đo của mỗi hệ số hồi quy thay đổi theo tương ứng*. Vì $u^* = 1{,}000u$, các phần dư và sai số chuẩn cũng sẽ được nhân

lên 1.000. Tổng các bình phương sẽ được nhân thêm 1 triệu (1.000 bình phương). Cần lưu ý rằng các trị thống kê *t, F,* và R^2 sẽ không bị ảnh hưởng vì chúng là các tỉ số trong đó yếu tố thang đo sẽ triệt tiêu.

BẢNG 3.3 Phân Tích Phương Sai

Nguồn	Tổng bình phương (SS)	Bậc tự do (d.f.)	Bình phương trung bình (SS÷d.f.)	F
Hồi quy (RSS)	$\sum(\hat{Y}_t - \overline{Y})^2 = 83.541$	1	83.541	$\frac{\text{RSS}(n-2)}{\text{ESS}} = 54{,}86$
Sai số (ESS)	$\sum \hat{u}_t^2 = 18.274$	N – 2 = 12	1.523	
Tổng (TSS)	$\sum(Y_t - \overline{Y})^2 = 101.815$	N – 1 = 13	7.832	

Tác động của việc thay đổi thang đo của một *biến độc lập* sẽ ra sao? Giả sử SQFT được tính theo đơn vị trăm mét vuông thay vì theo mét vuông thông thường, nhưng GIÁ được tính theo đơn vị ngàn đôla như trước. Gọi SQFT' là biến tính theo trăm mét vuông. Vậy SQFT= 100SQFT'. Thay vào phương trình ban đầu ta có:

$$\text{GIÁ} = \alpha + \beta 100\text{SQFT}' + u$$

Rõ ràng theo phương trình này, nếu chúng ta hồi quy GIÁ theo một hằng số và SQFT', hệ số duy nhất sẽ bị ảnh hưởng là hệ số của SQFT. Nếu β là hệ số của SQFT', thì $\hat{\beta}' = 100\hat{\beta}$. Sai số chuẩn của nó cũng sẽ nhân với 100. Tuy nhiên, tất cả các số đo khác – ESS, giá trị thống kê *t, F,* R^2 chẳng hạn sẽ không bị ảnh hưởng. *Tóm lại, trong một mô hình hồi quy tuyến tính, nếu thang đo của một biến độc lập thay đổi các hệ số hồi quy của nó và các sai số chuẩn tương ứng sẽ thay đổi tương ứng nhưng các trị thống kê khác sẽ không thay đổi.*

Có lý do chính đáng để thay đổi thang đo của các giá trị sao cho các số sau khi thay đổi sẽ không lớn cũng không quá nhỏ và tương tự với các giá trị của các biến khác. Điều này là vì các số có giá trị lớn sẽ lấn át các sai số và các số nhỏ sẽ gây ra sai số làm tròn, đặc biệt là khi tính giá trị tổng bình phương, việc này sẽ làm ảnh hưởng xấu đến độ chính xác của kết quả.

Để hiểu một cách thực tế hậu quả của việc thay đổi đơn vị, hãy Thực Hành Máy Tính phần 3.2 ở phụ lục D.

BÀI TẬP 3.7

Giả sử chúng ta đặt một biến mới X^* = SQFT – 1.000 (nghĩa là, X^* là phần diện tích vuông trên 1.000) và ước lượng mô hình GIÁ = $a + bX^* + v$. Giải thích bằng cách nào bạn có thể tìm được $\hat{a}$ và $\hat{b}$ từ $\hat{\alpha}$ và $\hat{\beta}$ *mà không phải ước lượng lại mô hình mới*.

3.7 Ứng dụng: Ước Lượng Đường Engel Biểu Diễn Quan Hệ Giữa Chi Tiêu cho Chăm Sóc Sức Khỏe và Thu Nhập.

Trong phần này, chúng ta sẽ trình bày một ứng dụng "tập dượt" với mô hình hồi quy hai biến. Dữ liệu được sử dụng là chuỗi dữ liệu chéo cho 50 bang và quận Columbia (*n* = 51), dữ liệu được thu thập từ cuốn *Tóm Lược Thống Kê Mỹ năm 1995 (Statistical Abstract of the US).* Các giá trị của dữ liệu thực có ở tập tin DATA3-2. Các biến là:

EXPHLTH = Chi tiêu tổng hợp (đơn vị tỷ đôla) cho chăm sóc sức khỏe của bang vào năm 1993, Bảng 153, trang 111, khoảng từ 0,998-9,029.

INCOME = Thu nhập cá nhân (đơn vị tỷ đôla) của bang vào năm 1993, Bảng 712, trang 460, khoảng từ 9,3-64,1.

Mô hình là đường Engel tìm được ở ví dụ 1.4 và được áp dụng với tổng chi tiêu cho chăm sóc sức khỏe của Mỹ là hàm số theo tổng thu nhập cá nhân. Phần Ứng Dụng Máy Tính 3.3 (xem phụ lục bảng D.1) trình bày hướng dẫn để tìm ra kết quả. Bản chú thích của báo cáo in từ máy tính, sử dụng chương trình ESL và tập tin PS3-3.ESL, được trình bày ở bảng 3.4. Phần được in đậm là nhập lượng của chương trình và các phần in nghiêng là các nhận xét về kết quả. Bạn nên tìm hiểu các chú thích này cẩn thận và sử dụng chương trình hồi quy bạn có để chạy lại các kết quả này (tập tin PS3-3.SHZ chứa các dòng lệnh để sử dụng phần mềm SHAZAM). Dưới đây là mô hình ước lượng cùng với trị thống kê mẫu t trong ngoặc đơn, và *p*-value (giá trị xác suất *p*) trong ngoặc vuông:

$$\widehat{\text{EXPHLTH}} = 0{,}176496 + 0{,}141652\,\text{INCOME}$$

(0.378) (49.272)

[0.707] [<0.0001]

$R^2 = 0{,}98$ d.f. = 49 F = 2.428 $\hat{\sigma} = 2{,}547$

Mô hình rất thích hợp với số liệu vì 98% sự biến đổi của chi tiêu cho chăm sóc sức khỏe được giải thích bởi biến thu nhập. Như đã giải thích ở Bảng 3.3, số hạng hằng số không có ý nghĩa về mặt thống kê và phù hợp với tiêu chuẩn lý thuyết đề ra trong ví dụ 1.4, chỉ ra rằng $\alpha = 0$. Để biết thêm chi tiết, xem các chú thích ở Bảng 3.4

3.8 Khoảng Tin Cậy

Như đã được chỉ ra ở Phần 2.9, một cách để xem xét trực tiếp đến việc ước lượng α và β trong điều kiện không chắc chắn là xác định khoảng tin cậy. Như vậy, ví dụ, thay vì nói $\hat{\beta} = 0{,}139$ chúng ta có thể nói rằng với mức xác suất cho trước, $\hat{\beta}$ sẽ nằm trong khoảng từ 0,09 đến 0,17. Từ kết quả các giá trị thống kê kiểm định ở phần 3.5 ta có:

$$\frac{\hat{\alpha}-\alpha}{s_\alpha} \sim t_{n-2} \quad \text{và} \quad \frac{\hat{\beta}-\beta}{s_\beta} \sim t_{n-2}$$

Đặt $t^*_{n-2}(0{,}025)$ là điểm name trên phân phối t với n-2 bậc tự do sao cho $P(t>t^*) = 0{,}025$. Điều này tương đương với $P(-t^* \le t \le t^*) = 0{,}95$. Như vậy,

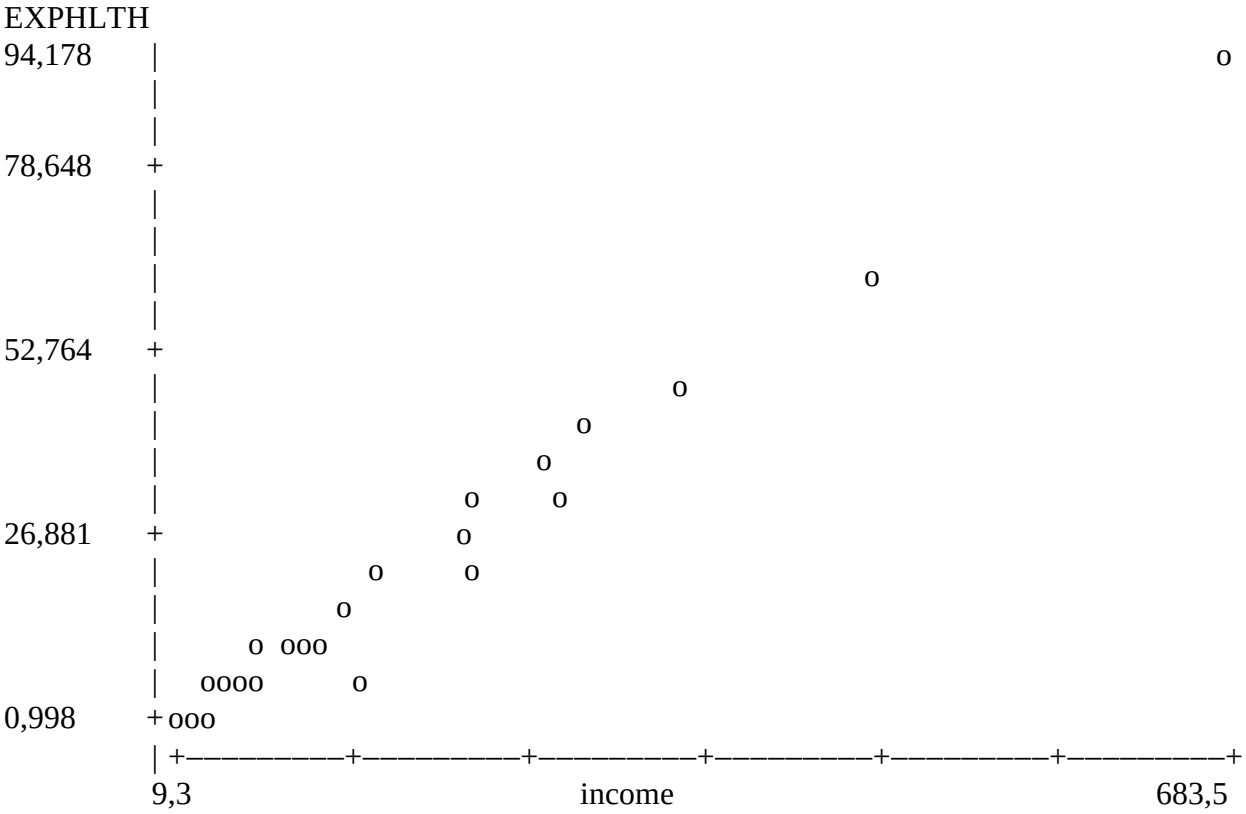
BẢNG 3.4 Báo Cáo từ Máy Tính Kèm Theo Chú Giải cho Phần 3.7

Các lệnh ESL được in đậm và các nhận xét được in nghiêng.

Danh sách biến

(0) Hằng số (1) exphlth (2) income

(Đồ thị của mức chi tiêu theo thu nhập cho thấy có sự quan hệ chặt chẽ giữa hai biến)



Ước lượng OLS với 51 quan sát 1-51
Biến phụ thuộc EXPHLTH

Biến	Hệ số	Sai số chuẩn	T-stat	2 Prob(t > | T |)
(0) hằng	0,176496	0,467509	0,377525	0,707414
(1) income	0,141652	0,002875	49,271792	<0,0001***

Giá trị ước lượng của hệ số của biến thu nhập là $\hat{\beta}$ = 0,141652 và ước lượng của số hạng hằng số là $\hat{\alpha}$ = 0,176496. Trị thống kê t (hệ số chia cho sai số chuẩn) của biến thu nhập là 49,271792, đây là giá trị rất ý nghĩa. 2Prob(t > |T|) là vùng diện tích ở hai đầu phân phối t chặn bởi giá trị kiểm định t và là giá trị p-value hoặc xác suất sai lầm loại I (đối với kiểm định 2 phía). Nếu p-value nhỏ (trong trường hợp này, nhỏ hơn 0,0001), chúng ta "an toàn" khi bác bỏ giả thuyết Ho rằng $\beta = 0$, và kết luận rằng hệ số của biến thu nhập là khác 0 đáng kể. Giá trị p-value của số hạng hằng số bằng 0,707414 gợi ý rằng nếu chúng ta bác bỏ giả thuyết Ho cho rằng $\alpha = 0$, chúng ta có thể phạm phải sai lầm loại I trong 70,7 % số lần. Vì mức sai lầm này quá cao, chúng ta không thể bác bỏ giả thuyết Ho. Như vậy chúng ta kết luận rằng số hạng hằng số không khác 0 đáng kể. Lưu ý rằng trong ví dụ 1.4, việc suy diễn lý thuyết ra đường Engel ám chỉ rằng không có số hạng hằng số. Số hạng hằng số không có ý nghĩa là phù hợp với kết quả theo lý thuyết. Xu hướng chi tiêu cận biên cho việc chăm sóc sức khỏe lấy từ thu nhập là

0,141652; nghĩa là, với mỗi khoản tăng thu nhập 100 đôla, chúng ta có thể kỳ vọng các cá nhân sẽ chi trung bình 14,17 đôla cho chăm sóc sức khỏe.

Giá trị R^2 (R-square) chỉ ra rằng 98% sự biến đổi của chi tiêu được giải thích bởi biến thu nhập. Sự khác nhau giữa giá trị R^2 Hiệu chỉnh và Không hiệu chỉnh sẽ được giải thích ở chương 4 cùng với các giá trị thống kê mẫu để chọn mô hình.

Giá trị thống kê mẫu Durbin-Watson và hệ số tương quan chuỗi bậc nhất sẽ được giải thích ở chương 9, nhằm giải quyết sự vi phạm giả thiết 3.6 cho rằng các số hạng sai số của hai quan sát là không tương quan. Giá trị trung bình của biến phụ thuộc là $\overline{Y}$ và S.D. là độ lệch chuẩn của S_y

Giá trị trung bình của biến phụ thuộc	15,068863	S.D. của biến phụ thuộc	17,926636
Tổng bình phương sai số (ESS)	317,898611	Sai số chuẩn của phần dư	2,547102
R- bình phươngkhông hiệu chỉnh	0,980	*R*- hiệu chỉnh	0,980
Trị thống kê F	2427,709468	*p*-value = Prob(F>2427.709)	<0,0001
Trị thống kê Durbin-Watson	2,209485	Hệ số tự tương quan bậc nhất	-0,121

Giá trị thống kê mẫu để chọn mô hình

SGMASQ	6,487727	AIC	6,741876	FPE	6,742147
HQ	6,939901	SCHWARZ	7,272471	SHIBATA	6,722193
GCV	6,752532	RICE	6,7638		

?genr ut=uhat (*lưu các ước lượng phần dư trong máy vào ut*.)
Generated var. no. 3 (ut)
?genr =exphlth-ut *(giá trị "thích hợp" = exphlth quan sát trừ phần dư)*
Generated var. no. 4 (yhat)
?print –o exphlth yhat ut; *(In giá trị chi tiêu thực và dự báo, giá trị phần dư. Dấu hiệu –o chỉ in ra ở dạng bảng)*

Obs	exphlth	yhat	ut
1	0,998	1,493862	-0,49586172
2	1,499	1,763001	-0,26400087
3	4,285	2,598749	1,686251
4	1,573	2,131297	-0,55829655
5	2,021	1,720505	0,30049479
6	2,26	2,343775	-0,08377483
7	1,953	1,989644	-0,03664435
8	2,103	2,244618	-0,1416183
9	3,428	3,179523	0,24847729
10	2,277	2,910384	-0,63338356
11	3,452	3,731965	-0,27996523

12	3,485	4,057766	-0,57276526
13	3,433	3,476992	-0,0439923
14	3,747	4,652705	-0,90570543
15	4,4	4,666871	-0,26687065
16	3,878	3,916114	-0,03811407
17	5,197	4,341071	0,85592937
18	4,118	4,426062	-0,30806194
19	6,111	5,672601	0,43839884
20	6,903	7,301601	-0,39850129
21	6,187	5,686766	0,50023362
22	7,341	7,485749	-0,14474913
23	7,999	8,533975	-0,53497529
24	8,041	7,967367	0,07353344
25	12,216	13,250993	-1,034993
26	10,066	11,027054	-0,96105374
27	9,029	8,84561	0,1833899
28	10,384	9,256401	1,127599
29	10,635	10,276297	0,35870284
30	12,06	10,318793	1,741207
31	13,014	10,276297	2,737703
32	14,194	13,619289	0,57471128
33	15,154	16,96228	-1,80828
34	14,502	14,32755	0,17445035
35	16,203	13,477637	2,725363
36	15,949	14,68168	1,26732
37	15,129	16,395672	-1,256672
38	16,401	15,701576	0,69942416
39	23,421	20,985202	2,435798
40	6,682	20,036133	-13,354133
41	20,104	19,002072	1,101928
42	18,241	18,56295	-0,32194997
43	25,741	30,093438	-4,352438
44	27,136	27,756177	-0,62017675
45	33,456	31,042507	2,413493
46	34,747	37,516012	-2,769012
47	41,521	36,439456	5,081544
48	44,811	40,320726	4,490274
49	49,816	49,0465	0,7694999
50	67,033	64,004971	3,028029
51	94,178	96,995765	-2,817765

$$P(-t^* \le \frac{\hat{\alpha}-\alpha}{s_\alpha} \le t^*) = 0.95 = P(\hat{\alpha} - t^* s_\alpha \le \alpha \le \hat{\alpha} + t^* s_\alpha$$

Từ đây có thể rút ra rằng khoảng tin cậy 95% của α và β lần lượt là $\hat{\alpha} \pm t^* s_\alpha$ và $\hat{\beta} \pm t^* s_\beta$

Ví dụ 3.7

Trong ví dụ về giá nhà, sai số chuẩn của $\hat{\alpha}$ và $\hat{\beta}$ là $s_{\hat{\alpha}}$ = 37,285 và $s_{\hat{\beta}}$ = 0,18373. Đồng thời, từ bảng *t*, ta có $t^*_{12}(0{,}025)$ = 2,179. Do đó, khoảng tin cậy 95% là

Đối với α: 52,351 ± (2,179x37,285) = (-28,893; 133,595)

Đối với β: 0,13875 ± (2,179x0,018373) = (0,099; 0,179)

Lưu ý rằng các khoảng tin cậy này là tương đối rộng. Đây là dấu hiệu cho thấy mô hình hồi quy tuyến tính thích hợp rất kém với tập dữ liệu. Một mô hình hồi quy thích hợp sẽ cho khoảng tin cậy hẹp hơn.

BÀI TẬP 3.8

Xác định khoảng tin cậy của α và β trong Phần Ứng Dụng 3.7

3.9 Dự Báo

Như đã đề cập trước đây, một trong các ứng dụng phổ biến của mô hình hồi quy là để dự báo (chủ đề này sẽ được thảo luận chi tiết hơn ở chương 11). Trong ví dụ giá nhà, chúng ta có thể đặt câu hỏi giá bán dự báo của một ngôi nhà có diện tích 2,000 mét vuông sẽ là bao nhiêu. Mô hình hồi quy ước lượng là $\hat{Y} = 52.351 + 0.13875X$. Như vậy, khi *X* = 2,000, giá trị dự báo của*Y* là 52,351 + (2,000x0,13875) = 329,851. Vì giá được tính theo đơn vị ngàn đôla, giá trị dự báo này cũng có đơn vị ngàn đôla. Vì vậy, theo mô hình, giá trung bình ước lượng của một căn hộ diện tích 2,000 mét vuông là 329.851 đôla. Một cách tổng quát, dễ dàng nhận thấy nếu *X* có giá trị X_0 thì giá trị dự báo của Y_0 sẽ là $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$. Giá trị trung bình có điều kiện của biến dự đoán Y cho trước $X = X_0$ là

$$E(\hat{Y}\,|\,X = X_0) = E(\hat{\alpha}) + X_0 E(\hat{\beta}) = \alpha + \beta X_0 = E(Y\,|\,X = X_0)$$

Như vậy $\hat{Y}_0$ là giá trị dự báo có điều kiện không thiên lệch của giá bán trung bình tại X_0.

Khoảng Tin Cậy cho Giá Trị Dự Báo Trung Bình

Vì α và β được ước lượng có sai số, giá trị dự báo $\hat{Y}_0$ cũng chịu sai số. Để xét đến yếu tố này, chúng ta tính sai số chuẩn và khoảng tin cậy cho giá trị dự báo trung bình. Dưới đây là ước lượng của phương sai của giá trị dự báo (xem chứng minh ở Phần 3.A.11)

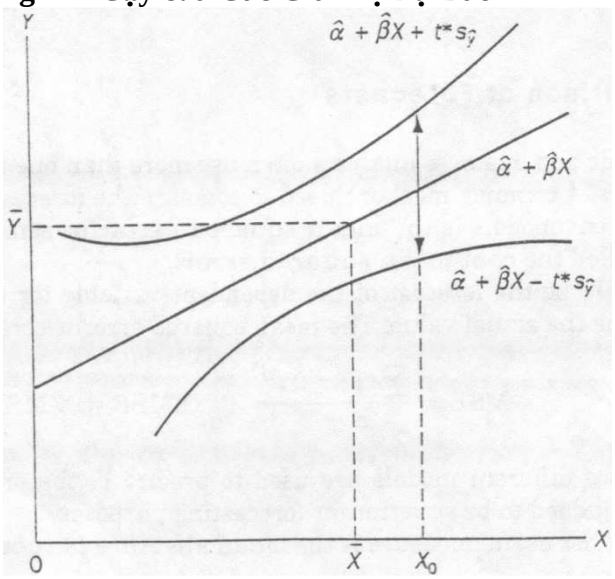
$$s^2_{Y_0} = \hat{\sigma}^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{S_{xx}}\right] \qquad \textbf{(3.28)}$$

Khoảng tin cậy của giá trị dự báo trung bình là

$$[\hat{Y}_0 - t^* s_{\hat{Y}_0}, \hat{Y}_0 + t^* s_{\hat{Y}_0}]$$

trong đó t^* là giá trị ngưỡng của phân phối t. Lưu ý rằng khi X_0 càng lệch xa giá trị trung bình $\overline{X}$, thì $s_{\hat{Y}_0}$ càng lớn và khoảng tin cậy tương ứng càng rộng. Điều này có nghĩa rằng nếu dự báo được thực hiện quá xa khỏi phạm vi của mẫu, độ tin cậy của dự báo sẽ giảm đi. Nếu $X_0 = \overline{X}$, khoảng tin cậy sẽ hẹp nhất. Hình 3.9 cho ý niệm về "dải tin cậy" với các giá trị X_0.

HÌNH 3.9 Dải Khoảng Tin Cậy của Các Giá Trị Dự Báo



Khoảng Tin Cậy cho Dự Báo Điểm

Phương sai mẫu trình bày ở phần trước dùng để dự báo giá trị trung bình. Bên cạnh đó chúng ta cũng muốn tìm phương sai của sai số dự báo cho các giá trị thực Y_0 tương ứng với X_0. Công thức dưới đây được lấy từ Phụ lục 3.A.12:

$$s^2_{\hat{u}_0} = Var(\hat{u}_0) = \hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{S_{xx}}\right] > s^2_{\hat{Y}_0} \qquad \textbf{(3.29)}$$

trong đó $\hat{u}_0 = Y_0 - \hat{Y}_0$ là sai số của dự báo điểm. Khoảng tin cậy được tính theo $s_{\hat{u}_0}$ thay vì $s_{\hat{Y}_0}$. Khi cỡ mẫu lớn, số hạng thứ hai và thứ ba ở trên sẽ không đáng kể so với $s_{\hat{u}_0}$ một giá trị gần bằng $\hat{\sigma}$. Ngoài ra, t^* cũng gần bằng 2 trong trường hợp mức ý nghĩa 95%. Như vậy, khoảng tin cậy của mẫu có kích thước lớn là $\hat{Y}_0 \pm 2\hat{\sigma}$

Ví dụ 3.8

Trong ví dụ giá nhà, chúng ta có $s^2_{\hat{Y}_0}$ = 111,555 và $s^2_{\hat{u}_0}$ =1634,353 và khoảng tin cậy tương ứng khi X_0 = 2.000 sẽ là (307, 353) và (242, 418). Khoảng tin cậy với cỡ mẫu lớn là (252,408). (Xem phần Thực Hành Máy Tính 3.4 để chạy lại kết quả này).

Chúng ta nên chọn loại khoảng tin cậy nào trong số hai loại trên? Vì quan tâm chính là sai số dự báo đối với giá trị thực Y_0, phương trình (3.29) thường được sử dụng. Lưu ý rằng khoảng tin cậy theo phương trình này rộng hơn nhiều khoảng tin cậy dựa trên phương trình (3.28)

So Sánh Các Giá Trị Dự Báo

Các nhà phân tích kinh tế và kinh doanh thường sử dụng nhiều hơn một mô hình để dự báo. Một số đo thường dùng để so sánh năng lực dự báo của các mô hình khác nhau là sai số **bình phương trung bình** (hoặc đôi khi người ta sử dụng căn bậc hai của nó, và được gọi là **căn bậc hai sai số bình phương trung bình**).

Gọi Y^f_t là giá trị dự báo của biến phụ thuộc cho quan sát *t*, và Y_t là giá trị thực. Sai số bình phương trung bình được tính như sau:

$$\text{MSE} = \frac{\sum(\text{Y}_\text{t}^\text{f} - \text{Y}_\text{t})^2}{\text{n} - 2} \qquad \text{RMSE} = \sqrt{\text{MSE}}$$

Nếu hai mô hình được sử dụng để dự báo Y, mô hình nào có MSE nhỏ hơn sẽ được đánh giá là mô hình tốt hơn cho mục đích dự báo.

Một số đo hữu ích khác là **sai số phần trăm tuyệt đối trung bình (MAPE)**

$$\text{MAPE} = \frac{1}{\text{n}}\sum 100\frac{\left|\text{Y}_\text{t} - \text{Y}_\text{t}^\text{f}\right|}{\text{Y}_\text{t}}$$

Số đo này chỉ có ý nghĩa nếu tất cả các giá trị *Y* đều dương (xem Phần Ứng Dụng 3.11). Một cách khác, chúng ta có thể tính **sai số phần trăm bình phương trung bình (MSPE)** hoặc căn của nó

$$\text{MSPE} = \frac{1}{\text{n}}\sum\left[100\frac{\text{Y}_\text{t} - \text{Y}_\text{t}^\text{f}}{\text{Y}_\text{t}}\right]^2 \qquad \text{RMSPE} = \sqrt{\text{MSPE}}$$

Một phương pháp khác để đánh giá mô hình và năng lực dự báo của nó là thực hiện **dự báo hậu mẫu**. Theo phương pháp này, người phân tích sẽ không sử dụng một số quan sát cuối cùng (chẳng hạn, 10% số quan sát cuối cùng) trong việc ước lượng mô hình, nhưng sẽ sử dụng các ước lượng thông số từ tập quan sát đầu tiên để dự báo Y_t cho phần mẫu để dành. Sau đó chúng ta có thể tính MSE và MAPE cho giai đoạn hậu mẫu. Mô hình nào có các giá trị đo lường này thấp hơn sẽ tốt hơn cho mục đích dự báo.

3.10 Tính Nhân Quả trong Mô Hình Hồi Quy

Khi định mô hình ở dạng $Y = \alpha + \beta X + u$, chúng ta ngầm giả định rằng *X gây ra Y*. Mặc dù R^2 đo độ thích hợp, nó không thể được sử dụng để *xác định tính nhân quả*. Nói cách

khác, việc *X* và *Y* tương quan chặt với nhau không có nghĩa rằng sự thay đổi *X* dẫn đến sự thay đổi *Y* hay ngược lại. Ví dụ, hệ số tương quan giữa số lượng kăngurư của Úc và tổng dân số nước này có thể là rất cao. Phải chăng điều này có nghĩa rằng sự thay đổi một biến sẽ làm cho biến kia thay đổi? Rõ ràng là không, vì ở đây chúng ta có một trường hợp **tương quan giả tạo**. Nếu chúng ta hồi quy một trong các biến với biến còn lại, chúng ta sẽ có sự **hồi qui giả tạo**. Lấy một ví dụ khác thực tế hơn, giả sử chúng ta hồi quy số lượng vụ trộm trong một thành phố với số hạng hằng số và số nhân viên cảnh sát (*X*) và sau đó quan sát thấy hệ số góc ước lượng có giá trị dương, có nghĩa rằng có tương quan thuận giữa *X* và *Y*. Phải chăng điều này có nghĩa rằng việc tăng số lượng cảnh sát sẽ làm tăng số vụ trộm, do đó ngầm kéo theo phải có chính sách giảm lực lượng cảnh sát? Rõ ràng kết luận này là không thể chấp nhận được. Điều xảy ra có thể là mối quan hệ nhân quả là ngược lại, có nghĩa là thành phố nên thuê thêm cảnh sát vì số vụ trộm tăng lên, và như vậy việc hồi quy *X* theo *Y* là hợp lý hơn. Tuy nhiên, trong thực tế, hai biến sẽ được **xác định kết hợp** và do đó chúng ta nên định rõ hai phương trình, một với *Y* theo *X* và các biến khác và phương trình còn lại với *X* theo *Y* và các biến khác. Việc xác định *đồng thời* các biến sẽ được trình bày chi tiết ở chương 13. Như sẽ thấy ở chương này các ước lượng thu được bằng cách bỏ qua tính đồng thời sẽ bị sai lệch và không nhất quán. Cũng có thể là sự tương quan cao quan sát được giữa *X* và *Y* có thể hoàn toàn là do các biến khác và không biến nào trong số chúng có thể trực tiếp gây ra các biến còn lại. Những ví dụ này nhấn mạnh tầm quan trọng của việc cân nhắc kỹ lưỡng bản chất cơ chế hành vi tiềm ẩn là gì, tức là**, quá trình phát dữ liệu** là gì (DGP), và lập mô hình một cách phù hợp. Lý thuyết kinh tế, kiến thức của nhà phân tích về các hành vị tiềm ẩn, kinh nghiệm quá khứ, v.v. phải gợi ý mô hình nên phải được xác định như thế nào. Tuy nhiên, có thể kiểm định phương hướng của sự nhân quả một cách rõ ràng (chi tiết sẽ trình bày ở chương 10). Độc giả quan tâm đến vấn đề này có thể tham khảo bài viết của Granger (1969) và Sims (1972).

Để minh họa tầm quan trọng của việc xác định chính xác sự nhân quả, giả sử chúng ta đảo ngược vị trí của *X* và *Y* và ước lượng mô hình:

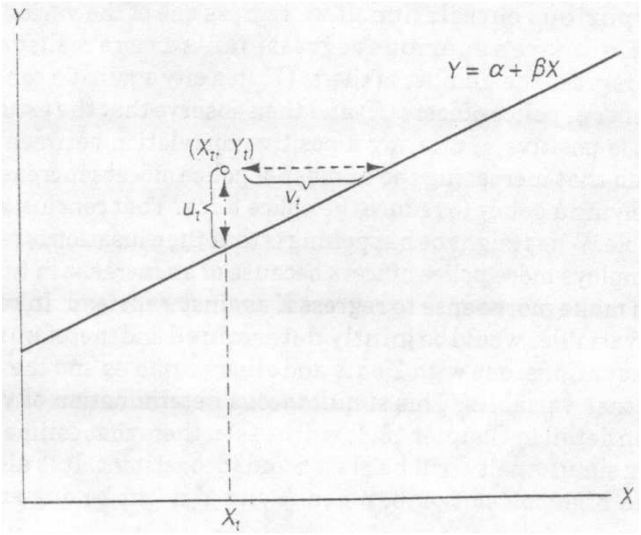
$$X_t = \alpha^* + \beta^* Y_t + v_t \qquad \textbf{(3.1')}$$

Liệu chúng ta có thể tìm được đường thẳng giống như trước không? Câu trả lời, nói chung, là không. Vì thủ tục bình phương nhỏ nhất được áp dụng cho phương trình (3.1) sẽ cực tiểu hóa tổng bình phương của các độ lệch đứng từ đường thẳng (xem hình 3.10). Trái lại, đường thẳng nghịch cực tiểu hóa tổng bình phương của các độ lệch ngang v_t. Tìm Y_t theo X_t, Phương trình (3.1') có thể được viết lại như sau:

$$Y_t = -\left(\frac{\alpha^*}{\beta^*}\right) + \left(\frac{1}{\beta^*}\right) X_t - \left(\frac{v^t}{\beta^*}\right) = \alpha' + \beta' X_t + v_t'$$

Việc cực tiểu hóa $\sum \hat{u}_t^2$, làm tương tự như với phương trình (3.1), và cực tiểu hóa $\sum \hat{v}_t^2$ sẽ thường cho ra các kết quả khác nhau. Cụ thể hơn, giá trị ước lượng của β' sẽ khác với giá trị β từ phương trình (3.1).

HÌNH 3.10 Cực Tiểu Hóa Tổng Bình Phương theo Trục Tung và Trục Hoành



Ví dụ 3.9

Quan hệ ước lượng khi $\sum \hat{u}_t^2$ được cực tiểu hóa là (xem Phần Thực Hành Máy Tính 3.5)

$$\widehat{\text{GIÁ}} = 52{,}351 + 0{,}13875\,\text{SQFT}$$

Khi quan hệ nhân quả được đảo ngược và $\sum \hat{v}_t^2$ được cực tiểu hóa, chúng ta có

$$\widehat{\text{SQFT}} = 33{,}385 + 5{,}913666\,\text{GIÁ}$$

Nghịch đảo quan hệ ước lượng thứ hai và biểu diễn $\widehat{\text{GIÁ}}$ là hàm của SQFT, ta có

$$\widehat{\text{GIÁ}} = -\frac{33.385}{5.913666} + \frac{1}{5.913666}\text{SQFT} = -5.645 + 0.169\text{SQFT}$$

Lưu ý rằng dấu của số hạng hằng số bị nghịch đảo và độ dốc là hoàn toàn khác.

Như vậy dưới điều kiện gì thì hai đường ước lượng sẽ như nhau? Để trả lời câu hỏi này, đầu tiên áp dụng OLS cho phương trình (3.1'); nghĩa là cực tiểu hóa $\sum \hat{v}_t^2$. Hoán đổi X và Y trong phương trình 3.10, ta có:

$$\hat{\beta}^* = \frac{S_{xy}}{S_{yy}} = \frac{1}{\hat{\beta}}$$

Và do đó $\hat{\beta}' = S_{yy}/S_{xy}$. Ước lượng bình phương nhỏ nhất làm cực tiểu $\sum \hat{u}_t^2$ là $\hat{\beta} = S_{xy}/S_{xx}$. Để β' bằng β, điều kiện là

$$\frac{S_{xy}}{S_{xx}} = \frac{S_{yy}}{S_{xy}} \qquad hoaëc \quad \frac{S_{xy}^2}{S_{xx}S_{yy}} = 1$$

Nhưng vế trái của phương trình thứ hai là r_{xy}^2, bình phương của hệ số hồi quy đơn giữa *X* và *Y* (định nghĩa ở phương trình 2.11). Như vậy, điều kiện cần là X và Y phải tương quan hoàn hảo. Tính chất 2.4d nói rằng nếu tồn tại sự tương quan hoàn hảo giữa hai biến, thì phải tồn tại một quan hệ tuyến tính chính xác giữa chúng. Vì vậy, sự thích hợp giữa X và Y phải hoàn hảo thì chúng ta mới nhận được cùng một đường hồi quy cho dù chúng ta áp dụng OLS cho phương trình (3.1) hay (3.1'). Nhìn chung, sự tương quan giữa *X* và *Y* sẽ không hoàn hảo, chính vì vậy chúng ta sẽ không nhận được cùng một đường thẳng hồi quy. Điều này nhấn mạnh tầm quan trọng của việc xác định đúng hướng quan hệ nhân quả thay vì việc chọn thiếu suy xét biến X và Y.

Như đã được minh họa trước đây trong ví dụ về tội phạm, quan hệ nhân quả có thể theo cả hai chiều, tình huống này được gọi là **phản hồi**. Quan hệ giữa giá bán và lượng bán cũng là ví dụ của hiện tượng này. Vì giá và lượng bán được xác định cùng lúc bởi quan hệ tương tác giữa cung và cầu, cho nên cái này có thể ảnh hưởng cái kia. Tương tự, hiện tượng phản hồi cũng được tìm thấy trong quan hệ giữa thu nhập tổng hợp và tiêu dùng hay đầu tư. Những tình huống này sẽ được trình bày ở chủ đề mô hình hồi quy hệ phương trình ở chương 13.

3.11 Ứng Dụng: Quan Hệ giữa Bằng Sáng Chế và Chi Phí cho Hoạt Động Nghiên Cứu và Phát Triển (R&D)

Phần này sẽ trình bày một ví dụ "diễn tập" khác về phân tích hồi quy. Dữ liệu dùng trong ví dụ này ở tập tin DATA3.3, mà sẽ đề cập đến các biến sau:

PATENTS = Số ứng dụng bằng sáng chế được ghi nhận, đơn vị ngàn, giao động từ 84,5 - 189,4

R&D = Chi phí cho nghiên cứu và phát triển, đơn vị tỉ đôla 1992, được xác định bằng tỉ số giữa chi phí theo đôla hiện hành và chỉ số giảm phát tổng sản phẩm quốc nội gộp (GDP), giao động từ 57,94 đến 166,7.

Dữ liệu theo năm lấy trong vòng 34 năm từ 1960 đến 1993 cho toàn bộ nước Mỹ. Nguồn được trình bày ở phụ lục D.

Nếu một quốc gia chi nhiều hơn cho hoạt động nghiên cứu và phát triển, chúng ta có thể kỳ vọng rằng quốc gia này sẽ đạt được nhiều cải tiến được bảo vệ thông qua luật bằng sáng chế hơn. Do đó, chúng ta kỳ vọng tồn tại một quan hệ dương giữa số lượng bằng sáng chế được ban bố và chi tiêu cho R&D. Mặc dù hiệu quả của hoạt động nghiên cứu và phát triển sẽ trễ vài năm sau khi dự án được bắt đầu, để đơn giản hóa chúng ta bỏ qua hiện tượng này. Ở những chương sau chúng ta sẽ khảo sát hiệu ứng trễ của các biến độc lập và sẽ quay lại ví dụ này.

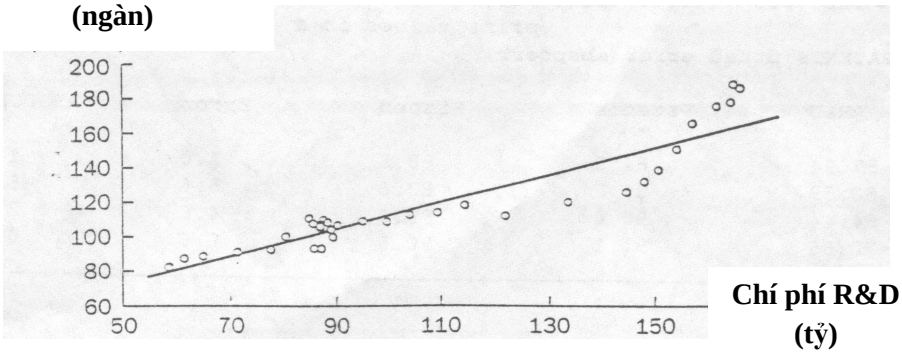
Mô hình hồi quy tuyến tính ước lượng được trình bày dưới đây kèm với các trị thống kê mẫu t trong ngoặc đơn (Phần Thực Hành Máy Tính 3.6 hướng dẫn cách lập lại kết quả của phần này và Bảng 3.5 trình bày kết quả.)

$$\widehat{\text{SÁNGCHẾ}} = 34{,}571 + 0{,}792\text{R \& D}$$
$$(5{,}44) \qquad (13{,}97)$$
$$R^2 = 0{,}859 \qquad \text{d.f.} = 32$$
$$F_c\,(1{,}32) = 195{,}055 \qquad \hat{\sigma} = 11{,}172$$

Để kiểm định mô hình về sự ý nghĩa tổng thể, chúng ta sử dụng trị thống kê *F*, có giá trị bằng 195,055. Theo giả thuyết H_0 thì số bằng sáng chế và chi phí cho R&D là không tương quan, F_c tuân theo phân phối *F* với bậc tự do ở tử số là 1 và bậc tự do ở mẫu số là 32 (= 34 – 2). Từ bảng A.4a (cũng ở trong bìa sau) chúng ta có nhận xét rằng giá trị ngưỡng *F*(1,32) ở mức ý nghĩa 1% nằm giữa 7,31 và 7,56. Vì F_c cao hơn nhiều so với giá trị này, chúng ta kết luận rằng số bằng sáng chế và chi phí cho R&D là tương quan đáng kể. Kết luận này được cũng cố thêm thông qua giá trị thống kê mẫu t. Kiểm định hai đầu với mức ý nghĩa 1%, bảng t trong bìa trước của quyển sách (hay Bảng A.2) cho thấy giá trị ngưỡng với 32 bậc tự do nằm giữa 2,704 và 2,75. Vì giá trị quan sát t_c cao hơn những giá trị này nhiều chúng ta kết luận rằng cả số hạng tung độ gốc và độ dốc có giá trị khác 0 một cách đáng kể. Số đo độ thích hợp R^2 cho biết mô hình giải thích được 85,9% sự biến đổi của biến phụ thuộc. Mặc dù đây dường như là một sự thích hợp tốt, tuy nhiên chúng ta thấy từ hình 3.11 rằng mô hình không hoàn toàn thể hiện sự biến đổi thực tế về số bằng sách chế. Đường thẳng hồi quy là đường liền và nó không đại diện đầy đủ bản chất đường cong của dữ liệu quan sát. Chính vì điều này mô hình sẽ dự báo rất kém số lượng bằng sáng chế tại nhiều năm.

Điểm này được nêu ra rõ hơn ở Bảng 3.5, bảng này có nhiều trị thống kê hữu ích khác. Cột thứ tư là giá trị trung bình ước lượng $\left(\hat{Y}_t\right)$, cột năm là giá trị phần dư được tính bằng giá trị quan sát trừ đi giá trị trung bình ước lượng $\left(\hat{u}_t = Y_t - \hat{Y}_t\right)$ và cột cuối cùng là sai số phần trăm tuyệt đối (APE), được tính bằng $100\left|\hat{u}_t\right|/Y_t$. Giá trị dự báo trình bày ở bảng 3.5 được làm tròn đến 1 chữ số thập phân. Vì dữ liệu gốc về số bằng sáng chế chỉ có một số thập phân, nên việc cố gắng có được các giá trị dự báo có độ chính xác đến hơn một số thập phân là không có ý nghĩa.

HÌNH 3.11 Số Bằng Sáng Chế Theo Chi Phí cho R&D của Nước Mỹ



BẢNG 3.5 Báo Cáo Máy Tính có Chú Thích cho Phần Ứng Dụng của Phần 3.11.

Các lệnh ELS được in đậm và các lời nhận xét được in nghiêng
Danh sách các biến
(0) Hằng số (1) Năm (2) R&D (3) PATENTS (SÁNG CHẾ)

Thời đoạn: 1, quan sát lớn nhất: 34, phạm vi quan sát: suốt 1960-1993, hiện hành 1960-1993 *(Ước lượng mô hình theo OLS)*

Ước lượng theo OLS sử dụng 34 quan sát từ 1960-1993
Biến phụ thuộc – PATENTS

Biến	Hệ số	Sai số chuẩn	T stat	2Prob(t > |T|)
(0) Hằng	34,571064	6,357873	5,437521	< 0,0001***
(2) R&D	0,791935	0,056704	13,966211	< 0,0001***

Giá trị trung bình của biến phụ thuộc	119,238235	S.D. của biến phụ thuộc	29,305827
Tổng bình phương sai số (ESS)	3994,300257	Sai số chuẩn của phần dư	11,172371
R-bình phương không hiệu chỉnh	0,859	R- bình phương hiệu chỉnh	0,855
Trị thống kê F	195,055061	p-value = Prob(F>2427.709)	<0,0001
Trị Durbin-Watson	0,233951	Hệ số tự tương quan bậc nhất	0,945

Các giá trị thống kê để chọn mô hình

SGMASQ	124,821883	AIC	132,146377	FPE	132,164347
HQ	136,255226	SCHWARZ	144,560215	SHIBATA	131,300527
GCV	132,623251	RICE	133,143342		

?genr ut=uhat *(lưu các ước lượng phần dư)*
?genr temp = PATENTS -ut *(tính giá trị "gắn")*
genr fitted = int (0.5+ (10*temp))/10 *(làm tròn đến một số thập phân)*
?genr error = PATENTS – fitted *(tính sai số dự báo)*
?genr abspcerr = int (0.5 + (1000*abs(error)/PATENTS))/100 *(tính sai số % tuyệt đối và làm tròn đến hai chữ số thập phân)*
?print –o R&D PATENTS fitted error abspcerr; *(in các giá trị ở dạng bảng)*

OBS	R&D	Patens	Fited	Error	ABSPCERR
1960	57,94	84,5	80,5	4,0	4,73
1961	60,59	S6,2	82,6	5,6	6,35
1962	64,44	90,4	85,6	4,8	5,31
1963	70,66	91,1	90,5	0,6	0,66
1964	76,83	93,2	95,4	-2,2	2,36

1965	80,00	100,4	97,9	2,5	2,49
1966	84,82	93,5	101,7	-8,2	8,77
1967	86,84	93,0	103,3	-10,3	11,08
1968	88,81	98,7	104,9	-6,2	6,28
1969	88,28	104,4	104,5	-0,1	0,10
1970	85,29	109,4	102,1	7,3	6,67
1971	83,18	111,1	100,4	10,7	9,63
1972	85,07	105,3	101,9	3,4	3,23
1973	86,72	109,6	103,2	6,4	5,84
1974	85,45	107,4	102,2	5,2	4,84
1975	83,41	108,0	100,6	7,4	6,85
1976	87,44	110,0	103,8	6,2	5,64
1977	90,11	109,0	105,9	3,1	2,84
1978	94,50	109,3	109,4	-0,1	0,09
1979	99,28	108,9	113,2	-4,3	3,95
1980	103,64	113,0	116,6	-3,5	3,19
1981	108,77	114,5	120,7	-6,2	5,41
1982	113,96	118,4	124,8	-6,4	5,41
1983	121,72	112,4	131,0	-18,5	16,55
1984	133,33	120,6	140,2	-19,6	-16,25
1985	144,78	127,1	149,2	-22,1	17,39
1986	148,39	133,0	152,1	-19,1	14,36
1987	150,90	139,8	154,1	-14,3	10,23
1988	154,36	151,9	156,8	-4,9	3,23
1989	157,19	166,3	159,1	7,2	4,33
1990	161,86	176,7	162,8	13,9	7,87
1991	164,54	178,4	164,9	13,5	7,57
1992	166,70	187,2	166,6	20,6	11,00
1993	165,20	189,4	155,4	24,0	12,67

Nhiều giá trị APE lớn hơn 5%, và trong một số năm chúng vượt qua 10%, đây là tỉ lệ khá lớn. Chúng ta cũng quan sát thấy rằng các điểm phân tán co cụm lại với nhau trong các năm từ 1966-1977, chỉ ra rằng một yếu tố nào đó khác hơn là chi phí R&D gây ra sự thay đổi về số bằng sáng chế. Do đó, quan sát kỹ hơn các kết quả chỉ cho thấy sự xác định sai mô hình. Trong chương 6, chúng ta sẽ dùng tập dữ liệu này để ước lượng mô hình đường cong và sẽ xem xét xem liệu việc xác định này có thể hiện tốt hơn các biến đổi quan sát được về số bằng sáng chế không.

TÓM TẮT

Mặc dù mô hình hồi quy tuyến tính đơn hai biến được sử dụng trong chương này, nhưng hầu hết các khía cạnh cơ bản của việc tiến hành phân tích thực nghiệm đã được đề cập. Thật hữu ích khi tóm tắt lại các kết quả đã được thảo luận từ đầu đến giờ.

Một mô hình hồi quy tuyến tính đơn là $Y_t = \alpha + \beta X_t + u_t$ (t = 1, 2, …, n). X_t và Y_t là quan sát thứ t lần lượt của biến độc lập và biến phụ thuộc, α và β là các thông số của tổng thể không biết sẽ được ước lượng từ dữ liệu của X và Y, u_t số hạng sai số không quan sát được, đây là các biến ngẫu nhiên với các tính chất được đề cập dưới đây, n là tổng số quan sát. Độ dốc (β) được diễn dịch là ảnh hưởng cận biên của sự tăng một đơn vị giá trị X_t lên Y_t, $\alpha + \beta X_t$ là trị trung bình có điều kiện của Y cho trước $X = X_t$.

Thủ tục bình phương nhỏ nhất thông thường (OLS) cực tiểu hóa tổng bình phương sai số $\sum \hat{u}_t^2$ và tính toán các ước lượng (ký hiệu $\hat{\alpha}$ và $\hat{\beta}$) của số hạng tung độ gốc α và độ

dốc β. Yêu cầu duy nhất để thực hiện việc ước lượng các thông số theo OLS là *n* có giá trị nhỏ nhất bằng 2 và ít nhất một trong những giá trị của X là khác nhau – nghĩa là, không phải tất cả các giá trị của X là như nhau.

Nếu u_t là biến ngẫu nhiên có giá trị trung bình bằng 0, và X_t cho trước và không ngẫu nhiên, thì $E(u_t) = 0$ và $E(X_t u_t) = 0$. Các phương trình chuẩn là $\sum \hat{u}_t = 0$ và $\sum X_t \hat{u}_t = 0$. Lời giải của các phương trình này cho kết quả là các ước lượng theo OLS của α và β.

Dưới các giả định vừa nêu ra, các ước lượng theo OLS là không thiên lệch và nhất quán. Sự nhất quán được giữ nguyên ngay cả nếu X_t là ngẫu nhiên, miễn là $Cov(X, u) = 0$ và $0 < Var(X) < \infty$ - nghĩa là, miễn là *X* và *u* không tương quan và *X* không là hằng số.

Nếu các giá trị u tuân theo phân phối độc lập và tương tự nhau (iid) với một phương sai xác định, $\hat{\alpha}$ và $\hat{\beta}$ cũng sẽ là các ước lượng không thiên lệch tuyến tính tốt nhất (BLUE); tức là, trong số tất cả tổ hợp tuyến tính không thiên lệch của các giá trị của Y, $\hat{\alpha}$ và $\hat{\beta}$ có phương sai nhỏ nhất. Kết quả này được gọi là định lý Gauss-Markov và có nghĩa rằng, ngoài tính chất không thiên lệch và nhất quán, các ước lượng theo OLS cũng là các ước lượng hiệu quả nhất. Nếu các giá trị của u tuân theo phân phối chuẩn độc lập và tương tự nhau $N(0, \sigma^2)$, các ước lượng theo OLS cũng là các ước lượng thích hợp nhất (MLE).

Từ $\hat{\alpha}$ và $\hat{\beta}$, giá trị dự báo của Y_t (ký hiệu là $\hat{Y}_t$) thu được bằng $\hat{Y}_t = \hat{\alpha} + \hat{\beta} X_t$, và phần dư được ước lượng bằng $\hat{u}_t = Y_t - \hat{Y}_t$. Sai số chuẩn của các phần dư là một ước lượng của độ lệch chuẩn σ và được tính theo công thức $\hat{\sigma} = \left[\sum \hat{u}_t^2 (n-2)\right]^{1/2}$. Từ các kết quả này, ta có thể suy ra sai số chuẩn của $\hat{\alpha}$ và $\hat{\beta}$ ($s_{\hat{\alpha}}$ và $s_{\hat{\beta}}$). Các sai số chuẩn càng nhỏ, độ chính xác của các ước lượng của các thông số càng lớn. Sự biến đổi của *X* càng lớn càng tốt vì điều này có khuynh hướng cải thiện độ chính xác của các ước lượng riêng lẻ.

Các bước tiến hành kiểm định đối thuyết một đầu về β được tiến hành như sau:

BƯỚC 1 $H_0: \beta = \beta_0 \quad H_1: \beta > \beta_0$

BƯỚC 2 Trị thống kê kiểm định là $t_c = \left(\hat{\beta} - \beta_0\right) / s_{\hat{\beta}}$, trong đó $s_{\hat{\beta}}$ là sai số chuẩn ước lượng của $\hat{\beta}$. Theo giả thuyết H_0, giá trị này tuân theo phân phối t với n – 2 bậc tự do.

BƯỚC 3 Tra bảng t với giá trị ứng với n – 2 bậc tự do và một mức ý nghĩa cho trước (chẳng hạn α), và tìm điểm $t^*_{n-2}(\alpha)$ sao cho $P(t > t^*) = \alpha$.

BƯỚC 4 Bác bỏ H_0 tại mức ý nghĩa α nếu $t_c > t^*$. Nếu giả thuyết ngược lại H_1 là $\beta < \beta_0$, H_0 sẽ bị bác bỏ nếu $t_c < -t^*$.

Kiểm định có thể được thực hiện theo một cách tương đương. Các bước 3 và 4 được điều chỉnh như sau:

BƯỚC 3a Tính xác suất (ký hiệu là p-value) sao cho $t > |t_c|$.

BƯỚC 4a Bác bỏ H_0 và kết luận là hệ số có ý nghĩa nếu p-value nhỏ hơn một mức ý nghĩa nào đó (α).

Các bước kiểm định giả thuyết ngược lại H_1 có tính hai phía được thực hiện như sau:

BƯỚC 1 H_0: $\beta = \beta_0$ H_1: $\beta \neq \beta_0$

BƯỚC 2 Trị thống kê kiểm định là $t_c = \left(\hat{\beta} - \beta_0\right) / s_{\hat{\beta}}$. Theo giả thuyết H_0, giá trị tuân theo phân phối t với n – 2 bậc tự do.

BƯỚC 3 Tra bảng t với giá trị ứng với n – 2 bậc tự do và một mức ý nghĩa cho trước (chẳng hạn α), và tìm điểm $t^*_{n-2}(\alpha/2)$ sao cho $P(t > t^*) = \alpha/2$ (một nữa của mức ý nghĩa).

BƯỚC 4 Bác bỏ H_0 tại mức ý nghĩa α nếu $|t_c| > t^*$.

Các bước hiệu chỉnh để thực hiện kiểm định theo phương pháp *p*-value như sau:

BƯỚC 3a Tính *p*-value = $2P(t > |t_c|)$.

BƯỚC 4a Bác bỏ H_0 nếu *p*-value nhỏ hơn một mức ý nghĩa nào đó (α).

Trị thống kê đo lường độ thích hợp của một mô hình là R^2 = 1- (ESS/TSS), trong đó $ESS = \sum \hat{u}_t^2$ và $\text{TSS} = \sum \left(\text{Y}_\text{t} - \overline{\text{Y}} \right)^2$. R^2 có giá trị từ 0 đến 1. Giá trị này càng cao độ thích hợp càng tốt. R^2 mang hai ý nghĩa: (1) nó là tỷ lệ của tổng phương sai của Y mà mô hình giải thích, và (2) nó là bình phương của hệ số tương quan giữa giá trị quan sát (Y_t) của biến phụ thuộc và giá trị dự báo $\left(\hat{\text{Y}}_\text{t}\right)$.

Kiểm định về độ thích hợp tổng thể của mô hình có thể được thực hiện bằng cách sử dụng giá trị R^2. Các bước được tiến hành như sau (ρ_{xy} là hệ số tương quan của tổng thể của hai biến *X* và *Y*):

BƯỚC 1 H_0: $\rho_{xy} = 0$ H_1: $\rho_{xy} \neq 0$

BƯỚC 2 Trị thống kê kiểm định là $F_c = R^2(n - 2)/(1 - R^2)$. Theo giả thuyết H_0, trị thống kê này tuân theo phân phối F với 1 bậc tự do ở tử số và n – 2 bậc tự do ở mẫu số.

BƯỚC 3 Tra bảng F theo tử số 1 bậc tự do và mẫu số n – 2 bậc tự do và một mức ý nghĩa cho trước (chẳng hạn α) tìm gí trị F^* sao cho: $P(F>F^*) = \alpha$.

BƯỚC 4 Bác bỏ giả thuyết H_0 (tại mức ý nghĩa α) nếu $F_c > F^*$.

Khoảng tin cậy 95% của β được xác định như sau:

$$\left(\hat{\beta} - t^* s_{\hat{\beta}}, \hat{\beta} + t^* s_{\hat{\beta}}\right)$$

Dự báo có điều kiện của Y, cho trước X bằng X_0, là $Y = \hat{\alpha} + \hat{\beta} X_0$. Phương sai của nó (phép đo độ tin cậy của dự báo) tỉ lệ thuận với khoảng cách của X_0 so với giá trị trung bình $\overline{X}$. Như vậy, X_0 càng xa khỏi giá trị trung bình của *X*, giá trị dự báo càng kém tin cậy.

Thay đổi thang đo của biến phụ thuộc dẫn đến thay đổi tương ứng thang đo của mỗi hệ số hồi quy. Tuy nhiên, các giá trị R^2 và trị thống kê t sẽ không đổi. Nếu thang đo của một biến độc lập thay đổi, hệ số hồi quy của nó và các hệ sai số chuẩn tương ứng bị thay đổi cùng thang đo, tuy nhiên tất cả các trị thống kê khác không thay đổi.

Việc xác định chính xác quan hệ nhân quả là hết sức quan trọng trong mô hình hồi quy. Giả thiết chuẩn là X gây ra Y. Tuy nhiên, nếu X và Y được tráo đổi, và mô hình được ước lượng bằng $X_t = \alpha^* + \beta^* Y_t + v_t$, đường thẳng hồi quy nói chung sẽ khác với đường được xác định từ mô hình $Y_t = \alpha + \beta X_t + u_t$.

THUẬT NGỮ

Analysis of variance (ANOVA)	Phân tích phương sai
Best linear unbiased estimator (BLUE)	Ước lượng không thiên lệch tuyến tính tốt nhất
Coefficient of multiple determination	Hệ số xác định bội
Conditional mean of Y given X	Giá trị trung bình điều kiện của Y biết trước X
Critical region	Vùng ngưỡng (vùng tới hạn)
Data-generating process (DGP)	Quá trình phát dữ liệu
Engel curve	Đường cong Engel
Error sum of square (ESS)	Tổng bình phương sai số
Estimated residual	Phần dư ước lượng
Explained variation	Sự biến đổi giải thích được
Feedback	Phản hồi
Fitted straight line	Đường thẳng thích hợp
F-test	Kiểm định F
Gauss-Markov theorem	Định lý Gauss-Markov
Goodness of fit	Độ khớp
Heteroscedasticity	Phương sai của sai số thay đổi
Homoscedasticity	Đồng phương sai sai số (tính chất phương sai của sai số không thay đổi)
Joinly determined	Được xác định cùng lúc
Linear estimator	Ước lượng tuyến tính
Marginal effect of X on Y	Hiệu ứng cận biên của X lên Y
Mean absolute percent error (MAPE)	Sai số phần trăm tuyệt đối trung bình
Mean squared error (MSE)	Sai số bình phương trung bình
Mean squared percentage error (MSPE)	Sai số phần trăm bình phương trung bình
Method of least square	Phương pháp bình phương tối thiểu
Nonlinear regression model	Mô hình hồi quy phi tuyến
Normal equation	Phương trình chuẩn
Ordinary least squares (OLS)	Bình phương tối thiểu thường
Population parameters	Tham số của tổng thể
Population regression function	Hàm hồi quy của tổng thể
Population regression line	Đường hồi quy của tổng thể
Population variance	Phương sai của tổng thể
Postsample forecast	Dự báo hậu mẫu

p-value	Giá trị p
Regression coefficients	Các hệ số hồi quy
Regression sum of squares (RSS)	Tổng bình phương hồi quy
Residual	Phần dư
Root mean squared error	Căn bậc hai của sai số bình phương trung bình
Sample estimate	Ước lượng của mẫu
Sample regression line	Đường hồi quy của mẫu
Sample regression function	Hàm hồi quy của mẫu
Sample scatter diagram	Biểu đồ phân tán của mẫu
Serial correlation	Tương quan chuỗi
Serial independence	Độc lập chuỗi
Significanly different from zero	Khác 0 một cách đáng kể
Significanly greater from zero	Lớn hơn 0 một cách đáng kể
Simple linear regression model	Mô hình hồi quy tuyến tính đơn
Spurious correlation	Tương quan giả tạo
Spurious regression	Hồi quy giả tạo
Standard error of a regression coefficient	Sai số chuẩn của hệ số hồi quy
Standard error of the regression	Sai số chuẩn của hồi quy
Standard error of the residuals	Sai số chuẩn của phần dư
Statistically insignificant	Không có ý nghĩavề thống kê
Statistically not greater than zero	Không lớn hơn 0 về mặt thống kê
Statistically significant	Có ý nghĩa về thống kê
Sum of squares of the residuals (ESS)	Tổng bình phương của các phần dư
Total sum of squares (TSS)	Tổng bình phương toàn phần
Total variance	Phương sai tổng
t-statistic	Trị thống kê t
t-test	Kiểm định t
Unexplained variation	Biến đổi không giải thích được
Well-behaved errors	Sai số thay đổi ngẫu nhiên
White-noise errors	Sai số do nhiễu trắng

3.A PHỤ LỤC

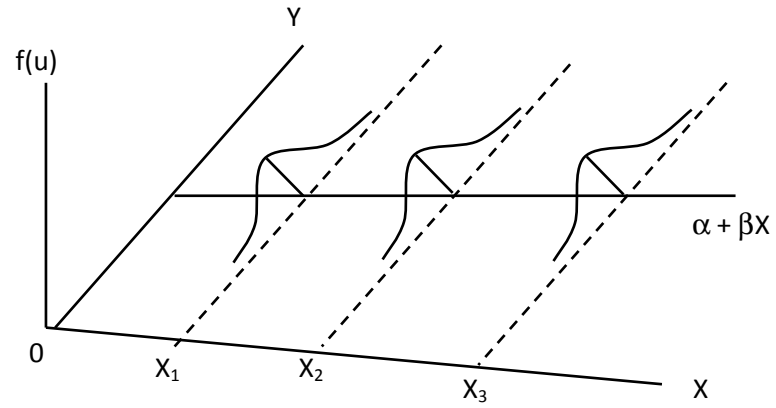
Chứng Minh Các Phương Trình

❍ 3.A.1 Biểu diễn 3 chiều của mô hình tuyến tính đơn

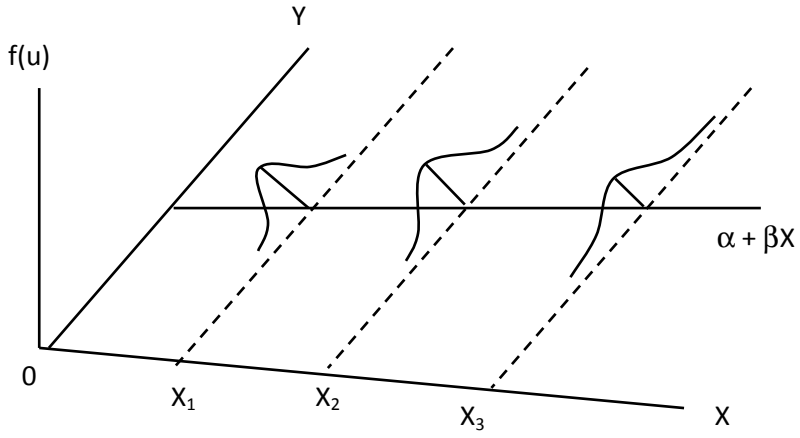
Hình 3.A.1 biểu diễn bằng đồ thị các giả thiết liệt kê trong bảng 3.2 cho trường hợp của mô hình hồi qui 2 biến đơn. Trục X và Y đại diện cho các giá trị của các biến X và Y. Trục Z là hàm mật độ xác suất f(u) của sai số ngẫu nhiên u. Đường thẳng $\alpha + \beta X$ là trung bình có điều kiện của Y với X cho trước, được giả sử là tuyến tính. Các phân phối thống kê được vẽ xung quanh đường trung bình cho 3 giá trị X_1, X_2 và X_3 là các phân phối có điều kiện tương ứng. Như đã đề cập trong bài, giả thiết rằng $Var(u_t) = \sigma^2$ được gọi là **phương sai của sai số không đổi**, có nghĩa "phân tán như nhau". Hình 3.A.1 mô tả tính bất biến của phương sai của sai số cho tất cả các quan sát. Nếu các phương sai này không bất biến mà thay đổi theo t [như vậy, $Var(u_t) \neq \sigma^2_t$], ta có

phương sai của sai số thay đổi (phân tán không như nhau). Hình 3.A.2 minh họa trường hợp phương sai của sai số thay đổi trong đó phương sai tăng khi X tăng. Trường hợp này được xem xét chi tiết hơn trong chương 8.

Hình 3.A.1 Biểu diễn đồ thị của Mô hình Hồi Qui Tuyến Tính Đơn



Hình 3.A.2 Minh họa về phương sai của sai số không đổi



3.A.2 Các Kết Quả Của Phép Tính Tổng

Các tính chất 3.1 và 3.2 được chứng minh ở đây

TÍNH CHẤT 3.1

$$S_{xx} = \sum (X_t - \overline{X})^2 = \sum X_t^2 - n(\overline{X})^2 = \sum X_t^2 - \frac{1}{n}\left(\sum X_t\right)^2$$

Chứng minh

$$\sum (X_t - \overline{X})^2 = \sum \left[X_t^2 - 2X_t\overline{X} + (\overline{X})^2\right] = \sum X_t^2 - \sum 2\overline{X}X_t + \sum (\overline{X})^2$$

Như trước đây, $\overline{X}$ như nhau với mỗi giá trị t. Do vậy, biểu thức trên = $\sum X_t^2 - 2\overline{X}\sum X_t + n(\overline{X})^2$. Hơn nữa $\sum X_t = n\overline{X}$. Do đó, biểu thức trở thành $\sum X_t^2 - 2\overline{X}n\overline{X} + n(\overline{X})^2$. Kết hợp số hạng thứ hai và ba trong biểu thức ta được phần thứ nhất của tính chất. Ta biết rằng $\overline{X} = (\sum X_t)/n$. Thay vào, ta có phần thứ hai của tính chất.

TÍNH CHẤT 3.2

$$S_{xy} = \sum (X_t - \overline{X})(Y_t - \overline{Y}) = \sum X_t Y_t - n\overline{X}\overline{Y} = \sum X_t Y_t - \left[\left(\sum X_t\right)\left(\sum Y_t\right)/n\right]$$

Chứng minh

$$\begin{aligned}\sum (X_t - \overline{X})(Y_t - \overline{Y}) &= \sum (X_t Y_t - X_t\overline{Y} - Y_t\overline{X} + \overline{X}\overline{Y}) \\ &= \sum X_t Y_t - \overline{Y}\sum X_t - \overline{X}\sum Y_t + n\overline{X}\overline{Y} \\ &= \sum X_t Y_t - \overline{Y}n\overline{X} - \overline{X}n\overline{Y} + n\overline{X}\overline{Y} \\ &= \sum X_t Y_t - n\overline{X}\overline{Y}\end{aligned}$$

Thay $\overline{X} = (\sum X_t)/n$ và $\overline{Y} = (\sum Y_t)/n$, ta có đẳng thức thứ hai.

3.A.3 Chứng Minh Các Phương Trình Chuẩn Bằng Phép Bình Phương Nhỏ Nhất

Trong phần này ta áp dụng phương pháp bình phương nhỏ nhất, được trình bày trong phần 3.2 và chứng minh các phương trình chuẩn (3.4) và (3.5). Tiêu chuẩn bình phương nhỏ nhất là chọn giá trị của $\hat{\alpha}$ và $\hat{\beta}$ làm tối thiểu tổng bình phương sai số:

$$ESS(\hat{\alpha},\hat{\beta}) = \sum_{t=1}^{t=n} \hat{u}_t^2 = \sum_{t=1}^{t=n} (Y_t - \hat{\alpha} - \hat{\beta}X_t)^2$$

Để tối thiểu ESS với $\hat{\alpha}$ và $\hat{\beta}$, ta cho đạo hàm riêng (xem phần 2.A.3 về đạo hàm riêng) ∂ESS/∂ $\hat{\alpha}$ và ∂ESS/∂ $\hat{\beta}$ bằng 0 và giải phương trình này. Ta có

$$\frac{\partial ESS}{\partial \hat{\alpha}} = \frac{\sum \partial(\hat{u}_t^2)}{\partial \hat{\alpha}} = \sum 2\hat{u}_t \frac{\partial \hat{u}_t}{\partial \hat{\alpha}} = 2\sum \hat{u}_t(-1) = 2\sum (Y_t - \hat{\alpha} - \hat{\beta}X_t)(-1) = 0$$

$$\frac{\partial ESS}{\partial \hat{\beta}} = \frac{\sum \partial(\hat{u}_t^2)}{\partial \hat{\beta}} = \sum 2\hat{u}_t \frac{\partial \hat{u}_t}{\partial \hat{\beta}} = 2\sum \hat{u}_t(-X_t) = 2\sum (Y_t - \hat{\alpha} - \hat{\beta}X_t)(-X_t) = 0$$

Từ đó ta thu được các phương trình sau:

$$\sum (Y_t - \hat{\alpha} - \hat{\beta}X_t) = 0$$
$$\sum (Y_t - \hat{\alpha} - \hat{\beta}X_t)X_t = 0$$

Lấy tổng từng số hạng và lưu ý rằng $\hat{\alpha}$ và $\hat{\beta}$ không phụ thuộc vào t và là thừa số chung có thể đưa ra ngoài các tổng, ta được

$$\sum Y_t = n\hat{\alpha} + \hat{\beta}\sum X_t$$
$$\sum Y_t X_t = \hat{\alpha}\sum X_t + \hat{\beta}\sum X_t^2$$

Phương trình đầu tiên tương đương với phương trình (3.4) và phương trình thứ 2 tương đương với phương trình (3.5).

❍ 3.A.4 Ước Lượng Không Thiên Lệch Tuyến Tính Tốt Nhất (Blue) Và Định Lý Gauss-Markov

Từ lý thuyết thống kê ta biết rằng một trong những tính chất mong muốn cho một ước lượng là ước lượng tuyến tính không thiên lệch phương sai nhỏ nhất (xem định nghĩa 2.8). Nói cách khác, giữa các tổ hợp tuyến tính của biến phụ thuộc không thiên lệch, ta chọn một biến có phương sai nhỏ nhất. Đây là ước lượng không thiên lệch tốt nhất (BLUE). Trong phần này ta chứng minh định lý Gauss-Markov, định lý này cho rằng ước lượng OLS rút ra trong phần 3.2 cũng có tính chất BLUE.

Đầu tiên lưu ý rằng ước số OLS $\hat{\beta}$ thực sự có thể được biểu diễn như là tổ hợp tuyến tính của Y_t. Để thấy điều này, ta viết lại phương trình (3.12) dưới đây.

(3.12) $$S_{xy} = \left[\sum X_t Y_t - \frac{(\sum X_t)(\sum Y_t)}{n}\right]$$

Lưu ý $\overline{X} = \sum X_t / n$, kết quả này có thể được biểu diễn như

$$\sum X_t Y_t - \overline{X}\sum Y_t = \sum (X_t - \overline{X})Y_t$$

Vì $\hat{\beta} = S_{xy}/S_{xx}$ từ phương trình (3.10), ta có

$$\hat{\beta} = \sum \left[\frac{X_t - \overline{X}}{S_{xx}} \right] Y_t = \sum \omega_t Y_t$$

Đây là tổ hợp tuyến tính của Y_t với trọng số $\omega_t = \left[\frac{X_t - \overline{X}}{S_{xx}} \right]$ phụ thuộc vào X_t. Bây giờ xem tổng tổ hợp tuyến tính của các giá trị của Y có dạng $\tilde{\beta} = \sum a_t Y_t$, với a_t có tính không ngẫu nhiên. Ước lượng không thiên lệch tốt nhất (BLUE) có 2 tính chất: (1) $\tilde{\beta}$ không thiên lệch và (2) Var($\tilde{\beta}$) là nhỏ nhất.

Chứng minh

Gọi $d_t = a_t - \omega_t$ là hiệu số các trọng số (lưu ý rằng d_t chỉ phụ thuộc các biến X và do đó được xem là không ngẫu nhiên). Vậy $a_t = \omega_t + d_t$. Tiếp theo là

$$\tilde{\beta} = \sum (\omega_t + d_t) Y_t = \hat{\beta} + \sum d_t Y_t$$

$$E(\tilde{\beta}) = \beta + \sum E(d_t Y_t) = \beta + \sum d_t E(Y_t) = \beta + \sum d_t (\alpha + \beta X_t)$$

$$= \beta + \alpha \sum d_t + \beta \sum d_t X_t$$

Để có tính không thiên lệch, kết quả này phải bằng β, điều này xảy ra khi và chỉ khi

$$\sum d_t = 0 \qquad \text{và} \qquad \sum d_t X_t = 0$$

Phương sai của ước lượng $\tilde{\beta}$ được xác định bởi $\text{Var}[\sum(\omega_t + d_t)Y_t]$. Từ tính chất 2.A.5c, phương sai của tổng các biến ngẫu nhiên độc lập là tổng các phương sai (tính độc lập được đảm bảo bởi giả thiết 3.6). Hơn nữa, do giả thiết 3.5 về phương sai của sai số không đổi, u_t và do vậy Y_t có phương sai không đổi σ^2. Từ đó ta có

$$\text{Var}(\tilde{\beta}) = \sigma^2 \sum(\omega_t + d_t)^2 = \sigma^2 \sum \omega_t^2 + \sigma^2 \sum d_t^2 + 2\sigma^2 \sum \omega_t d_t$$

Số hạng thứ 3 bằng 0 vì $\sum \omega_t d_t = \sum \left(\frac{X_t - \overline{X}}{S_{xx}} \right) d_t = 0$, do các điều kiện về tính không thiên lệch $\sum d_t = 0$ và $\sum d_t X_t = 0$ làm cho mỗi số hạng trong tổng bằng 0. Trong biểu thức trên về phương sai của $\tilde{\beta}$, số hạng đầu độc lập với các biến chọn d_t. Bởi vì số hạng thứ hai là tổng các bình phương,

chỉ có cách duy nhất để tối thiểu số hạng này là chọn mỗi giá trị của ds bằng 0. Điều này làm cho $a_t = \omega_t$ và do đó $\tilde{\beta} \equiv \hat{\beta}$, như vậy ý nói rằng ước lượng OLS thực sự là BLUE và do vậy sẽ có hiệu quả nhất. Điều này xác minh định lý Gauss-Markov. Cũng nên lưu ý rằng việc chứng minh định lý cần đến các giả thiết 3.5 và 3.6 về phương sai của sai số không đổi và tính độc lập theo chuỗi. Nếu một trong 2 giả thiết này bị vi phạm, thì phương pháp OLS không cho ước lượng hiệu quả.

3.A.5 Ước Lượng Thích Hợp Nhất

Lý do của phương pháp ước lượng thích hợp nhất được diễn tả chi tiết trong phần 2.A.4. Bạn đọc có thể xem phần đó trước khi bắt đầu phần này. Trong phần đó, phương pháp này đã được áp dụng cho trường hợp ước lượng giá trị trung bình và phương sai của một phân phối chuẩn. Ở đây ta áp dụng kỹ thuật tương tự vào bài toán hồi qui. Bởi vì nguyên lý thích hợp nhất đòi hỏi kiến thức về các phân phối trong bài toán, nên ta cần giả thiết 3.7. Các bước để xác định một ước lượng thích hợp nhất rất dễ hiểu. Trước tiên, lập hàm thích hợp liên kết hàm mật độ của các quan sát với các thông số chưa biết. Để cực đại hàm này, lấy vi phân riêng phần logarit của hàm thích hợp cho mỗi thông số chưa biết và cho bằng 0. Kế đến giải các điều kiện bậc nhất để tìm các ước lượng thích hợp nhất. Hàm mật độ của u được xác định theo [xem phương trình (2.4)]

$$f(u) = \frac{1}{(\sigma\sqrt{2\pi})} e^{-u^2/2\sigma^2}$$

Bởi vì các quan sát là độc lập nhau, hàm thích hợp của $u_1, u_2, \ldots, u_n$ là

$$L(\alpha, \beta, \sigma^2) = f(u_1)\, f(u_2)\, f(u_3) \ldots\ldots f(u_n)$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum u_i^2/(2\sigma^2)}$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum (Y_t - \alpha - \beta X_t)^2/(2\sigma^2)}$$

Thực hiện cực đại hóa logarit của hàm thích hợp thì dễ hơn, giá trị cực đại sẽ bằng với giá trị lớn nhất L bởi vì loga có tính chất tăng đều; nghĩa là nếu a>b, thì ln(a)>ln(b).

$$\ln L = -n \ln \sigma - n \ln(\sqrt{2\pi}) - \sum \left[\frac{(Y_t - \alpha - \beta X_t)^2}{2\sigma^2} \right]$$

$$= -n \ln \sigma - n \ln(\sqrt{2\pi}) - \frac{SSE}{2\sigma^2}$$

Trong đó SSE = $\sum (Y_t - \alpha - \beta X_t)^2$. α và β chỉ xuất hiện trong số hạng SSE. Do đó, lnL lớn nhất bằng với SEE nhỏ nhất (bởi vì có dấu âm trước SSE). Nhưng SSE nhỏ nhất nghĩa là các ước lượng bình phương nhỏ nhất. Do đó, các ước lượng bình phương nhỏ nhất cũng là MLE với điều kiện các sai số của u tuân theo phân phối $N(0,\sigma^2)$. Bởi vì các ước lượng thích hợp nhất là đồng nhất và hiệu quả một cách tiệm cận, nên các ước lượng OLS cũng vậy.

Để có MLE của σ^2, lấy vi phân riêng phần lnL theo σ và cho bằng 0. Ta có

$$\frac{\partial(\ln L)}{\partial\sigma} = -\frac{n}{\sigma} + \frac{SEE}{\sigma^3} = 0$$

Giải phương trình tìm σ^2 ta được σ^2 = SSE/n. Nhưng SSE phụ thuộc vào α và β. Tuy nhiên, ta có thể dùng các ước lượng của chúng $\hat{\alpha}$ và $\hat{\beta}$. Do đó ta thu được MLE của phương sai của u_t bằng với $\tilde{\sigma}^2 = \sum \hat{u}_t^2 / n$. Như đã phát biểu trước đó, giá trị này không thiên lệch. Một ước lượng không thiên lệch có thể tìm được bằng cách chia $\sum \hat{u}_t^2$ cho n-2 và dùng $\hat{\sigma}^2$ đã xác định trong phương trình (3.21). Điều kiện không thiên lệch được chứng minh trong phục lục phần 3.A.7.

3.A.6 Tìm Các Phương Sai Của Các Ước Lượng

Từ phương trình (3.10), ta có $\hat{\beta} = S_{xy}/S_{xx}$. Vì *X* là không ngẫu nhiên theo giả thiết 3.4, S_{xx} cũng không ngẫu nhiên và do đó $Var(\hat{\beta}) = Var(S_{xy})S^2_{xx}$. Từ phương trình (3.15), $S_{xy} = \beta S_{xx} + S_{xu}$ và do đó $Var(S_{xy}) = Var(S_{xu})$. Từ phương trình (3.16) ta lưu ý rằng $S_{xu} = \sum(X_t - \overline{X})u_t$. Tính chất 2.A.5c cho thấy phương sai của tổng các biến ngẫu nhiên là tổng của các phương sai với điều kiện đồng phương sai (covariance) các số hạng bằng 0. Theo giả thiết 3.6, u_t và u_s là không tương quan với mọi t≠s và đồng phương sai bằng 0. Do đó,

$$Var(S_{xu}) = Var\left[\sum(X_t - \overline{X})u_t\right] = \sum Var\left[(X_t - \overline{X})u_t\right] = \sum(X_t - \overline{X})^2 Var(u_t)$$

Với giả thiết 3.5, $Var(u_t) = \sigma^2$. Do đó, $Var(S_{xu}) = \sigma^2\sum(X_t - \overline{X})^2 = \sigma^2 S_{xx}$. Từ đó sẽ có

$$Var(\hat{\beta}) = \frac{Var(S_{xy})}{S_{xx}^2} = \frac{\sigma^2 S_{xx}}{S_{xx}^2} = \frac{\sigma^2}{S_{xx}}$$

Vậy ta đã chứng minh phương trình (3.18). Thủ tục để chứng minh các phương trình (3.19) và (3.20) cũng tương tự và sẽ là bài tập cho bạn đọc.

3.A.7 Ước Lượng Không Thiên Lệch Của Phương Sai Của Số Hạng Sai Số

Theo phương trình (3.21), $s^2 = \hat{\sigma}^2 = (\sum \hat{u}_t^2)/(n-2)$ là một ước lượng không thiên lệch của σ^2. Điều này được chứng minh như sau.

$$\hat{u}_t = Y_t - \hat{\alpha} - \hat{\beta}X_t = Y_t - (\overline{Y} - \hat{\beta}\overline{X}) - \hat{\beta}X_t$$

Dùng phương trình (3.9) cho $\hat{\alpha}$. Vì Y_t được xác định bởi phương trình (3.1), $\overline{Y} = \alpha + \beta\overline{X} + \overline{u}$ với $\overline{u}$ bằng $\sum u_t/n$. Do đó, nhóm tất cả các số hạng β ta có,

$$\begin{aligned}\hat{u}_t &= (\alpha + \beta X_t + u_t) - (\alpha + \beta\overline{X} + \overline{u}) + \hat{\beta}\overline{X} - \hat{\beta}X_t \\ &= (u_t - \overline{u}) - (\hat{\beta} - \beta)(X_t - \overline{X})\end{aligned}$$

Tổng bình phương của $\hat{u}_t$ được xác định theo

$$\begin{aligned}\sum \hat{u}_t^2 &= \sum (u_t - \overline{u})^2 + (\hat{\beta} - \beta)^2 \sum (X_t - \overline{X})^2 - 2(\hat{\beta} - \beta)\sum (X_t - \overline{X})(u_t - \overline{u}) \\ &= S_{uu} + (\hat{\beta} - \beta)^2 S_{xx} - 2(\hat{\beta} - \beta)S_{xu}\end{aligned}$$

Dùng ký hiệu tương tự như trong phương trình (3.11) và (3.16). Từ phương trình (3.15), $S_{xu} = S_{xy} - \beta S_{xx} = S_{xx}(\hat{\beta} - \beta)$. Thay kết quả này vào phương trình trên và kết hợp các số hạng thứ hai và ba ta có

$$\sum \hat{u}_t^2 = S_{uu} - (\hat{\beta} - \beta)^2 S_{xx}$$

Để tính giá trị kỳ vọng của tổng bình phương của sai số, ta cần $E(S_{uu})$và $E[(\hat{\beta} - \beta)^2]$. Từ tính chất 2.11b ta lưu ý rằng

$E(S_{uu}) = (n-1)Var(u) = (n-1)\sigma^2$. Hơn nữa,

$$E[(\hat{\beta} - \beta)^2] = Var(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}$$

Từ phương trình (3.18). Đặt tất cả các kết quả, ta có

$$E\left(\sum \hat{u}_t^2\right) = E(S_{uu}) - S_{xx}E[(\hat{\beta} - \beta)^2] = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

Chia cho *n-2* ta có kết quả mong muốn

$$E(\hat{\sigma}^2) = E\left[\frac{\sum \hat{u}_t^2}{n-2}\right] = \sigma^2$$

Vậy, $\hat{\sigma}^2$ là ước lượng không thiên lệch của σ^2.

✪ 3.A.8 Chứng Minh Phương Trình 3.25

Giá trị tổng bình phương được viết lại như sau:

$$\begin{aligned}\sum (Y_t - \overline{Y})^2 &= \sum (Y_t - \hat{Y}_t + \hat{Y}_t - \overline{Y})^2 \\ &= \sum (Y_t - \hat{Y}_t)^2 + \sum (\hat{Y}_t - \overline{Y})^2 + 2\sum (Y_t - \hat{Y}_t)(\hat{Y}_t - \overline{Y})\end{aligned}$$

Với $\hat{u}_t = Y_t - \hat{Y}_t$, hai số hạng đầu tiên là hai số hạng có trong phương trình (3.25). Bây giờ tất cả điều ta cần là phải chứng minh rằng $\sum (Y_t - \hat{Y}_t)(\hat{Y}_t - \overline{Y}) = \sum \hat{u}_t (\hat{Y}_t - \overline{Y}) = 0$.

$$\sum \hat{u}_t(\hat{Y}_t - \overline{Y}) = \sum \hat{u}_t(\hat{\alpha} + \hat{\beta}X_t - \overline{Y}) = \hat{\alpha}\sum \hat{u}_t + \hat{\beta}\sum \hat{u}_t X_t - \overline{Y}\sum \hat{u}_t$$

Từ phương trình chuẩn đầu tiên (3.4), $\sum \hat{u}_t = \sum (Y_t - \hat{\alpha} - \hat{\beta}X_t) = 0$. Từ phương trình (3.5), $\sum \hat{u}_t X_t = \sum (Y_t - \hat{\alpha} - \hat{\beta}X_t)X_t = 0$, vậy kết quả được chứng minh.

✪ 3.A.9 Chứng Minh Phương Trình 3.26a

Để chứng minh phương trình (3.26a), trước tiên ta tìm đồng phương sai mẫu (ký hiệu bởi $\widehat{\text{Cov}}$) giữa Y_t và $\hat{Y}_t$. Từ phương trình (2.10),

$$\widehat{\text{Cov}}(Y_t, \hat{Y}_t) = \frac{1}{n-1}\sum (Y_t - \overline{Y})(\hat{Y}_t - \overline{Y})$$

Lưu ý rằng trung bình của $\hat{Y}_t$ cũng là $\overline{Y}$ bởi vì $\hat{\alpha} + \hat{\beta}\overline{X} = \overline{Y}$. Vậy,

$$Y_t - \overline{Y} = (Y_t - \hat{Y}_t) + (\hat{Y}_t - \overline{Y}) = \hat{u}_t + (\hat{Y}_t - \overline{Y})$$

Do đó,

$$\widehat{\text{Cov}}(Y_t, \hat{Y}_t) = \frac{\sum \hat{u}_t(\hat{Y}_t - \overline{Y})}{n-1} + \frac{\sum (\hat{Y}_t - \overline{Y})^2}{n-1}$$

Phần trước đã cho thấy số hạng thứ nhất bằng 0. Do vậy, đồng phương sai của Y_t và $\hat{Y}_t$ bằng với số hạng thứ hai, là RSS/(n-1);

$$\widehat{\text{Cov}}(Y_t, \hat{Y}_t) = \frac{\text{RSS}}{n-1}$$

Ta cũng có

$$\widehat{\text{Var}}(Y_t) = \frac{\text{TSS}}{n-1} \qquad \text{và} \qquad \widehat{\text{Var}}(\hat{Y}_t) = \frac{\sum (\hat{Y}_t - \overline{Y})^2}{n-1} = \frac{\text{RSS}}{n-1}$$

Từ phương trình (2.7) ta nhớ lại bình phương của hệ số tương quan đơn giữa Y_t và $\hat{Y}_t$ được xác định bởi

$$r^2_{Y\hat{Y}} = \frac{\widehat{\text{Cov}^2}(Y_t, \hat{Y}_t)}{\widehat{\text{Var}}(Y_t)\widehat{\text{Var}}(\hat{Y}_t)}$$

Thay thế đồng phương sai và phương sai từ biểu thức vừa rút ra và bỏ *n-1*, ta có

$$r^2_{Y\hat{Y}} = \frac{RSS^2}{TSS\,RSS} = \frac{RSS}{TSS} = R^2$$

Vậy, bình phương của tương quan đơn giữa giá trị quan sát Y_t và giá trị $\hat{Y}_t$ được dự báo bởi mô hình hồi qui là như nhau và là R^2 được định nghĩa trong phương trình (3.26).

3.A.10 Chứng Minh Rằng $r^2_{xy} = R^2$ Cho Mô Hình Hồi Qui Đơn

Trong phần này ta sẽ chứng minh rằng trong trường hợp mô hình hồi qui đơn, R^2 cũng bằng với bình phương của tương quan đơn giữa X và Y. Từ phương trình (2.11), $r^2_{xy} = S^2_{xy}/(S_{xx}S_{yy})$. S_{yy} bằng với tổng bình phương TSS. Hơn nữa, RSS = $\sum(\hat{Y}_t - \overline{Y})^2$. Vì $\hat{Y}_t = \hat{\alpha} + \hat{\beta}X_t$ và $\overline{Y} = \hat{\alpha} + \hat{\beta}\overline{X}$, ta có $\hat{Y}_t - \overline{Y} = \hat{\beta}(X_t - \overline{X})$. Do đó,

$$RSS = \sum(\hat{Y}_t - \overline{Y})^2 = \hat{\beta}^2 \sum(X_t - \overline{X})^2 = \hat{\beta}^2 S_{xx}$$

Từ phương trình (3.10), $\hat{\beta} = S_{xy}/S_{xx}$. Thay kết quả này cho một số hạng $\hat{\beta}$ ở trên, ta thu được

$$RSS = \hat{\beta}\left(\frac{S_{xy}}{S_{xx}}\right)(S_{xx}) = \hat{\beta}S_{xy}$$

Thay thế S_{xy} từ kết quả này và lưu ý rằng S_{yy} = TSS, ta có

$$r^2_{xy} = \frac{S^2_{xy}}{S_{xx}S_{yy}} = \frac{S_{xy}S_{xy}}{S_{xx}TSS} = \frac{\hat{\beta}S_{xy}}{TSS} = R^2$$

Kết quả đã được chứng minh.

3.A.11 Chứng Minh Phương Trình 3.28

$$\begin{aligned} Var(\hat{Y}_0) &= E[\hat{Y}_0 - E(\hat{Y}|X_0)]^2 \\ &= E[\hat{\alpha} + \hat{\beta}X_0 - \alpha - \beta X_0]^2 = E[(\hat{\alpha} - \alpha) + X_0(\hat{\beta} - \beta)]^2 \\ &= Var(\hat{\alpha}) + X_0^2 Var(\hat{\beta}) + 2X_0 Cov(\hat{\alpha}, \hat{\beta}) \end{aligned}$$

Trong phép biến đổi trên, ta đã dùng tính chất 2.4a. Thay từ phương trình (3.18), (3.19) và (3.20), ta được

$$Var(\hat{Y}_0) = \sigma^2\left[\frac{\sum X_t^2}{nS_{xx}} + X_o^2\frac{1}{S_{xx}} - 2\frac{X_0\overline{X}}{S_{xx}}\right]$$

Với $S_{xx} = \sum X_t^2 - n\overline{X}^2$. Lưu ý rằng $\sum X_t^2 = S_{xx} + n\overline{X}^2$ và thay vào biểu thức phương sai, ta được

$$\text{Var}(\hat{Y}_0) = \sigma^2\left[\frac{1}{n} + \frac{\overline{X}^2 + X_0^2 - 2X_0\overline{X}}{S_{xx}}\right] = \sigma^2\left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{S_{xx}}\right]$$

Kết quả này là phương trình (3.28).

3.A.12 Chứng minh phương trình 3.29

Gọi $\hat{u}_0 = Y_0 - \hat{Y}_0$ là sai số tại điểm dự báo của Y_0, với $\hat{Y}_0 = \hat{\alpha} + \hat{\beta}X_0$ là giá trị dự báo của trung bình. Do đó ta có

$$\text{Var}(\hat{u}_0) = \text{Var}(Y_0) + \text{Var}(\hat{Y}_0) - 2\text{Cov}(Y_0, \hat{Y}_0)$$

Vì $Y_0 = \alpha + \beta X_0 + u_0$, $\text{Var}(Y_0) = \sigma^2$. Mặt khác, $\text{Var}(\hat{Y}_0)$ được xác định bởi phương trình (3.26). Cuối cùng, $\text{Cov}(Y_0, \hat{Y}_0) = 0$, bởi vì u_0 không tương quan với các số dư khác và do đó không tương quan với $\hat{\alpha}$ và $\hat{\beta}$. Vậy ta có

$$\text{Var}(\hat{u}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(X_0 - \overline{X})^2}{S_{xx}}\right]$$

BÀI TẬP

Câu Hỏi Lý Thuyết

3.1. Tất cả các tổng ở bốn biểu thức dưới đây được dùng để tính các *dữ liệu mẫu*, không phải dành cho tập hợp hoàn chỉnh. Hãy chỉ ra những biểu thức sai và đúng. Giải thích tại sao những biểu thức đó đúng hay sai.

a. $\sum_{t=1}^{n} \hat{u}_t = 0$

b. $\sum_{t=1}^{n} X_t \hat{u}_t = 0$

c. $\sum_{t=1}^{n} u_t = 0$

d. $\sum_{t=1}^{n} X_t u_t = 0$

3.2. Có sự khác biệt gì giữa số hạng sai số và phần dư? Hãy giải thích sự khác biệt giữa u_t và E(u_t). Sau đó, hãy chứng minh rằng $E(\hat{u}_t) = 0$. Và giải thích giá trị kỳ vọng nghĩa là gì cũng như nêu ra những giả thiết cần thiết để chứng minh biểu thức trên.

3.3. Cho mô hình tuyến tính đơn biến $Y_t = \alpha + \beta X_t + u_t$, hãy chứng minh rằng dưới những giả thiết nhất định, phương pháp ước lượng OLS cho kết quả các ước lượng không chệch? Có nghĩa là cần phải chứng minh rằng $E(\hat{Y}_t = E(Y_t)$. Hãy nêu ra những giả thiết cần thiết cho việc chứng minh đó.

3.4. Nêu ra những giả thiết cần thiết cho mỗi phát biểu sau. Đồng thời giải thích lý do tại sao những giả thiết này cần cho phát biểu đó.

e. Để ước lượng α và β bằng phương pháp OLS
f. Để chứng minh rằng các ước lượng của các thông số theo phương pháp OLS là không chệch và nhất quán.
g. Để chứng minh rằng các ước lượng theo phương pháp OLS là hiệu quả
h. Để thực hiện kiểm định *t* và *F*

3.5. Những câu hỏi sau là đúng hay sai? Nếu những câu hỏi này chỉ đúng một phần, bạn hãy chỉ ra phần đúng đó. Giải thích lý do tại sao những câu (phần) đó đúng.

i. Các ước lượng hệ số góc theo phương pháp OLS sẽ chính xác hơn nếu các giá trị *X* gần với trị trung bình mẫu của chúng hơn.
j. Nếu X_t và u_t tương quan, các ước lượng vẫn sẽ không chệch.
k. Các ước lượng không thể là ước lượng không lệch tuyến tính tốt nhất (BLUE) trừ phi tất cả các giá trị u_t tuân theo phân bố chuẩn.
l. Nếu các số hạng sai số không tuân theo phân phối chuẩn thì các kiểm định t và F không thể được thực hiện.
m. Nếu phương sai của ut lớn thì các khoảng tin cậy của các ước lượng sẽ lớn (rộng) hơn.
n. Nếu phương sai của *X* lớn thì các khoảng tin cậy ước lượng sẽ hẹp hơn.
o. Khi trị số *p*-value lớn thì hệ số sẽ khác 0 một cách đáng kể.
p. Nếu bạn chọn một mức ý nghĩa cao hơn thì hệ số hồi qui sẽ có khả năng có ý nghĩa hơn.