CHƯƠNG *4*

Mô Hình Hồi Qui Bội

Trong Chương 3 chúng ta giới hạn trong trường hợp đơn giản của mô hình hồi qui hai biến. Bây giờ, chúng ta sẽ xem xét ***hồi qui bội****, nghĩa là liên hệ biến phụ thuộc Y cho trước với nhiều biến độc lập X_1, X_2, ..., X_k. Mô hình hồi qui tuyến tính đa biến có công thức tổng quát như sau:*

$$Y_t = \beta_1 + \beta_2 X_{t2} + ... + \beta_k X_{tk} + u_t \qquad \textbf{(4.1)}$$

X_{t1} được đặt bằng 1 để có được "tung độ gốc". Chữ t nhỏ biểu thị số lần quan sát và có giá trị từ 1 đến n. Các giả thiết về số hạng nhiễu, u_t, hoàn toàn giống những giả thiết đã xác định trong Chương 3. Trong các đặc trưng tổng quát của một mô hình hồi qui bội, Việc lựa chọn các biến độc lập và biến phụ thuộc xuất phát từ các lý thuyết kinh tế, trực giác, và kinh nghiệm quá khứ. Trong ví dụ về ngành bất động sản ở Chương 3, biến phụ thuộc là giá của căn nhà một hộ gia đình. Chúng ta đã đề cập ở đó là ***chỉ số giá - hưởng thụ*** *phụ thuộc vào đặc điểm của căn nhà. Bảng 4.1 trình bày dữ liệu bổ sung cho 14 căn nhà mẫu đã bán. Lưu ý rằng, dữ liệu cho X_1 chỉ đơn giản là một cột gồm các số 1 và tương ứng với số hạng không đổi. Tính cả số hạng không đổi, có tất cả là k biến độc lập và vì vậy có k hệ số tuyến tính chưa biết cần ước lượng.*

Mô hình tuyến tính bội trong ví dụ này như sau:

$$\text{PRICE} = \beta_1 + \beta_2 \text{SQFT} + \beta_3 \text{BEDRMS} + \beta_4 \text{BATHS} + u \qquad \textbf{(4.2)}$$

Cũng như trước, giá được tính bằng đơn vị ngàn đô la. Ngoài diện tích sử dụng, giá còn liên hệ với số phòng ngủ cũng như số phòng tắm.

Ảnh hưởng của thay đổi trong Y_t khi chỉ có X_{ti} thay đổi được xác định bởi $\Delta Y_t / \Delta X_{ti} = \beta_i$. Vì vậy, ý nghĩa của hệ số hồi qui β_i là, giữ giá trị của tất cả các biến khác không đổi, nếu X_{ti} thay đổi một đơn vị thì Y_t kỳ vọng thay đổi, trung bình là, β_i đơn vị. Do đó, β_4 trong phương trình (4.2) được diễn giải như sau: Giữa hai căn nhà có cùng diện tích sử dụng (SQFT) và số phòng ngủ (BEDRMS), căn nhà nào có thêm một phòng tắm được kỳ vọng sẽ bán với giá cao hơn, trung bình, khoảng β_4 ngàn đô la. Vì vậy, phân tích hồi qui bội giúp chúng ta kiểm soát được một tập hợp con các biến giải thích và kiểm tra ảnh hưởng của một biến độc lập đã chọn.

❹ Bảng 4.1 Dữ liệu về nhà một hộ gia đình (giá tính bằng ngàn đô la)

t	Giá (Y)	Hằng số (X_1)	SQFT (X_2)	BEDRMS (X_3)	BATHS (X_4)
1	199,9	1	1.065	3	1,75
2	228	1	1.254	3	2

3	235	1	1.300	3	2
4	285	1	1.577	4	2,5
5	239	1	1.600	3	2
6	293	1	1.750	4	2
7	285	1	1.800	4	2,75
8	365	1	1.870	4	2
9	295	1	1.935	4	2,5
10	290	1	1.948	4	2
11	385	1	2.254	4	3
12	505	1	2.600	3	2,5
13	425	1	2.800	4	3
14	415	1	3.000	4	3

4.1 Phương trình chuẩn

Trong trường hợp mô hình hồi qui bội, Giả thiết 3.4 được hiệu chỉnh như sau: *Mỗi X cho trước sao cho* $\text{Cov}(X_{si}, u_t) = E(X_{si}\, u_t) = 0$ *với mỗi i từ 1 đến k và mỗi s, t từ 1 đến n.* Vì vậy, *mỗi biến độc lập được giả định là không liên hệ với tất cả các số hạng sai số.* Trong trường hợp của thủ tục bình phương tối thiểu thông thường (OLS), chúng ta định nghĩa tổng của bình phương sai số là

$$\text{ESS} = \sum_{t=1}^{n} \hat{u}_t^2 = \sum_{t=1}^{n} (Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{t2} - ... - \hat{\beta}_k X_{tk})^2$$

Thủ tục OLS cực tiểu ESS theo $\hat{\beta}_1, \hat{\beta}_2 ..., \hat{\beta}_k$. Bằng cách thực hiện như trong Phần 3.A.3, chúng ta có thể có được các phương trình chuẩn, số phương trình chuẩn bằng số hệ số tuyến tính ước lượng. Do đó chúng ta có *k* phương trình trong đó *k* hệ số hồi qui chưa biết (các tổng được tính theo chỉ số *t* – nghĩa là số lần quan sát):

$$\Sigma Y_t = n(\hat{}) \hat{\beta}_1 + \hat{\beta}_2 \Sigma X_{t2} + ... + \hat{\beta}_k \Sigma X_{tk}$$

$$\Sigma Y_t X_{t2} = \hat{\beta}_1 \Sigma X_{t2} + \hat{\beta}_2 \Sigma X^2_{t2} + ... + \hat{\beta}_k \Sigma X_{tk} X_{t2}$$

.....

$$\Sigma Y_t X_{ti} = \hat{\beta}_1 \Sigma X_{ti} + \hat{\beta}_2 \Sigma X_{t2} X_{ti} + ... + \hat{\beta}_k \Sigma X_{tk} X_{ti}$$

$$\Sigma Y_t X_{tk} = \hat{\beta}_1 \Sigma X_{tk} + \hat{\beta}_2 \Sigma X_{t2} X_{tk} + ... + \hat{\beta}_k \Sigma X^2_{tk}$$

k phương trình chuẩn trên có thể giải được các nghiệm đơn β (chỉ trừ một vài trường hợp ngoại lệ trình bày trong Chương 5). Các chương trình máy tính chuẩn thực hiện được mọi tính toán này khi nhập dữ liệu vào và xác định các biến độc lập, biến phụ thuộc. Phụ lục 4.A.1 mô tả các bước đối với mô hình ba biến trong đó Y hồi qui theo một số hạng không đổi, X_2 và X_3.

Các tính chất 3.1 đến 3.3 cũng đúng trong trường hợp hồi qui tuyến tính bội. Do đó, các ước lượng OLS là BLUE, không thiên lệch, hiệu quả và nhất quán. Phần dư và các giá trị dự đoán có được từ các liên hệ sau:

$$\hat{u}_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{t2} - ... - \hat{\beta}_k X_{tk}$$

$$\hat{Y}_t = \hat{\beta}_1 + \hat{\beta}_2 X_{t2} + ... + \hat{\beta}_k X_{tk} = Y_t - \hat{u}_t$$

❹ VÍ DỤ 4.1

Đối với mô hình đã nêu trong Phương trình (4.2), liên hệ ước lượng là (xem phần Thực hành máy tính 4.1)

$$\widehat{PRICE} = 129{,}062 + 0{,}1548SQFT - 21{,}588BEDRMS - 12{,}193BATHS$$

Lập tức chúng ta lưu ý là các hệ số hồi qui của BEDRMS và BATHS đều âm, trái với chúng ta mong đợi. Chúng ta có thể cảm thấy theo trực giác là thêm phòng tắm hoặc phòng ngủ sẽ tăng giá trị của căn nhà. Tuy nhiên, hệ số hồi qui có ý nghĩa đúng chỉ khi mọi biến khác đều không thay đổi. Do đó, nếu chúng ta tăng số phòng ngủ lên một, *giữ nguyên* SQFT *và* BATHS *không đổi*, giá trung bình được kỳ vọng sẽ hạ xuống khoảng $21.588. Nếu cùng một diện tích sử dụng được chia nhỏ để có thêm một phòng ngủ thì mỗi phòng ngủ sẽ có diện tích nhỏ hơn. Dữ liệu cho thấy là, trung bình, người mua đánh giá thấp việc chia nhỏ diện tích này và vì vậy họ sẽ chỉ sẵn lòng trả một mức giá thấp hơn.

Lý luận tương tự cho BATHS. Giữ nguyên SQFT và BEDRMS không đổi, nếu ta tăng thêm một phòng tắm, giá trung bình kỳ vọng sẽ giảm khoảng $12.193. Một lần nữa, tăng thêm phòng tắm nhưng vẫn giữ nguyên diện tích sử dụng cũng có nghĩa là phòng ngủ sẽ nhỏ hơn. Kết quả cho thấy sự không đồng ý của khách hàng và vì vậy chúng ta quan sát thấy giá trung bình giảm. Từ lập luận này chúng ta lưu ý là những dấu có vẻ không như mong đợi lúc đầu (thường được gọi là "dấu sai") lại được giải thích hợp lý.

Giả sử chúng ta tăng thêm một phòng ngủ và tăng thêm diện tích sử dụng khoảng 300 (cho thêm hành lang và các yếu tố liên quan khác). BEDRMS sẽ tăng thêm 1 và SQFT tăng thêm 300. Thay đổi giá trung bình (ΔPRICE) là kết quả của tác động kết hợp như sau:

$$\Delta\widehat{PRICE} = \hat{\beta}_2 \Delta SQFT + \hat{\beta}_3 \Delta BEDRMS = 300\hat{\beta}_2 + \hat{\beta}_3$$

Trong mô hình, phần này thể hiện một khoảng tăng $24.852 trong giá trung bình ước lượng [được tính như sau (300 x 0,1548) – 21,588; đơn vị ngàn đô la], mức giá này có vẻ hợp lý.

---

❹ BÀI TẬP THỰC HÀNH 4.1

Giả sử tăng thêm một phòng tắm và một phòng ngủ, với diện tích sử dụng tăng thêm 350 bộ vuông. Mức giá trung bình kỳ vọng tăng thêm bao nhiêu? Giá trị này có đáng tin không?

❹ BÀI TẬP THỰC HÀNH 4.2

Dự báo giá trung bình của một căn nhà với 4 phòng ngủ, 3 phòng tắm và diện tích sử dụng là 2.500 bộ vuông. Dự báo có hợp lý so với dữ liệu trong Bảng 4.1 không?

Một ước lượng không thiên lệch của phương sai phần dư σ^2 được tính bằng $s^2 = \hat{\sigma}^2 = \Sigma\hat{u}_t^2/(n-k)$, với n là số lần quan sát sử dụng trong ước lượng và k là số hệ số hồi qui ước lượng, gồm cả số hạng không đổi. Chứng minh phát biểu này về nguyên tắc tương tự như đã trình bày trong phần 3.A.7, nhưng phức tạp hơn nhiều vì có đến *k* phương trình chuẩn ở đây (xem Johnston, 1984, trang 180-181). Trong Chương 3 chúng ta chia tổng bình phương sai số cho *n – 2* để được ước lượng không thiên lệch của σ^2. Ở đây, *k* phương trình chuẩn đặt ra k ràng buộc, điều này dẫn đến việc "mất đi" k bậc tự do. Vì vậy, chúng ta chia cho *n – k*. Bởi vì $\hat{\sigma}^2$ phải không âm, *n* phải lớn hơn *k*. Thủ tục để tính sai số chuẩn của các $\hat{\beta}$ là tương tự, nhưng các phép tính bây giờ sẽ nhàm chán hơn nhiều. Các chương trình máy tính cung cấp các phép toán thống kê cần thiết để ước lượng các thông số và kiểm định giả thuyết về chúng. Có thể thấy là $\Sigma\hat{u}_t^2/\sigma^2$ có phân phối Chi bình phương với bậc tự do *n – k* (xem Johnston, 1984, trang 181). Các kết quả này được tóm tắt trong tính chất 4.1.

Tính Chất 4.1

a. Một ước lượng không thiên lệch của phương sai sai số (σ^2) được tính bằng

$$s^2 = \hat{\sigma}^2 = \frac{ESS}{n-k} = \frac{\Sigma\hat{u}_t^2}{n-k}$$

với ESS là tổng bình phương của các phần dư

b. ESS/σ^2 có phân phối Chi bình phương với bậc tự do *n – k*. Lưu ý rằng tính chất này phụ thuộc đặc biệt vào Giả thiết 3.8 là số hạng sai số u_t tuân theo phân phối chuẩn $N(0,\sigma^2)$.

Các Giá Trị Dự Báo Và Sai Số Chuẩn

Cũng như trong mô hình hồi qui đơn biến, chúng ta sẽ quan tâm đến tạo ra các dự báo có điều kiện của biến phụ thuộc với các giá trị cho trước của các biến độc lập. Giả sử X_{fi} là giá trị cho trước của biến độc lập thứ i với i = 2, ..., k, và t = f, với các giá trị này chúng ta muốn dự báo Y. Định nghĩa

$$\beta = \beta_1 + \beta_2 X_{f2} + \ldots + \beta_k X_{fk}$$

Và $\hat{\beta} = \hat{Y}_f$, định nghĩa trước đó *t = f*, và vì vậy dự báo cần có là giá trị ước lượng của β, và sai số chuẩn tương ứng sẽ giúp chúng ta xây dựng một khoảng tin cậy cho dự báo. Giải β_1 từ phương trình trên và thay vào mô hình ban đầu, chúng ta có

$$Y_t = \beta - \beta_2 X_{f2} - \ldots - \beta_k X_{fk} + \beta_2 X_{t2} + \ldots + \beta_k X_{tk} + u_t$$

Nhóm số hạng một cách thích hợp, ta có thể viết lại như sau:

$$\begin{aligned} Y_t &= \beta + \beta_2 (X_{t2} - X_{f2}) + \ldots + \beta_k (X_{tk} - X_{fk}) + u_t \\ &= \beta + \beta_2 Z_{t2} + \ldots + \beta_k Z_{tk} + u_t \end{aligned}$$

với $Z_{ti} = X_{ti} – X_{fi}$, cho i = 2, ..., k. Việc viết lại công thức này chỉ ra các bước sau để tiến hành dự báo

Bước 1 Với giá trị X_{fi} cho trước của biến độc lập thứ *i* và $t = f$, tạo một biến mới $Z_{ti} = X_{ti} – X_{fi}$ với i = 2, ..., k.

Bước 2 Hồi qui Y_t theo một số hạng và các biến mới Z_{t2}, ..., Z_{tk}.

Bước 3 Số hạng không đổi được ước lượng là một dự báo điểm cần có. Khoảng tin cậy tương ứng (xem phần 3.8) được tính bằng $\hat{\beta} - t^*s_f$, $\hat{\beta} + t^*s_f$), với t* là giá trị tới hạn của phân phối *t* với bậc tự do $n – k$ và mức ý nghĩa cho trước, và s_f là sai số chuẩn của số hạng không đổi được ước lượng có được từ bước 2.

❍ VÍ DỤ 4.2

Trong ví dụ về bất động sản, đặt SQFT = 2.000, BEDRMS = 4 và BATHS = 2,5. Bước thứ nhất tạo các biến mới, SQFT2 = SQFT – 2000, BEDRMS2 = BEDRMS – 4 và BATHS2 = BATHS – 2,5. Kế đến hồi qui PRICE theo một số hạng không đổi và SQFT2, BEDRMS2 và BATHS2. Từ bài thực hành máy tính phần 4.1 chúng ta lưu ý là giá trung bình dự báo của căn nhà này là $321.830 và sai số chuẩn của dự báo là $13.865. Điều này cho khoảng tin cậy 95% là 321.830 ± (2,201 x 13.865) tính được khoảng tin cậy là (291.313; 352.347).

❍ 4.2 Độ Thích Hợp

Khi đánh giá mức độ thích hợp, tổng bình phương toàn phần, tổng bình phương hồi qui, và tổng bình phương của sai số có cùng dạng như đã trình bày trước, và ở đây cũng có TSS = RSS + ESS (miễn là mô hình có một số hạng không đổi). Vì vậy,

$$TSS = \Sigma (Y_t - \bar{Y})^2 \qquad RSS = \Sigma(\hat{Y}_t - \bar{Y})^2 \qquad ESS = \Sigma \hat{u}_t^2$$

Mức độ thích hợp được đo như trước đây bằng $R^2 = 1 –$ (ESS/TSS). Nếu có số hạng không đổi trong mô hình, R^2 cũng bằng với bình phương của hệ số tương quan giữa Y_t và $\hat{Y}_t$. Tuy nhiên, định nghĩa R^2 theo cách này sẽ phát sinh một vấn đề. Có thể thấy là việc thêm vào bất kỳ một biến nào (dù biến này có ý nghĩa hay không) thì R^2 cũng sẽ không bao giờ giảm. Chứng minh bằng đại số phát biểu này rất nhàm chán, nhưng chúng ta có thể lý luận theo trực giác. Khi một biến mới được thêm vào và ESS được cực tiểu, chúng ta đang cực tiểu theo một tập rất nhiều biến số và vì vậy ESS mới có vẻ sẽ nhỏ hơn (ít nhất thì cũng không lớn hơn). Cụ thể hơn, giả sử số hạng $\beta_{k+1}X_{tk+1}$ được thêm vào phương trình (4.1) và ta có được một mô hình mới. Nếu giá trị *cực tiểu* của tổng bình phương của mô hình mới này *lớn hơn* giá trị của mô hình cũ, thì ta đặt β_{k+1} bằng không và sử dụng các ước lượng cũ cho các giá trị β khác sẽ tốt hơn, và vì vậy các ước lượng mới không thể có ESS cực tiểu. Điều này kéo theo khi một biến mới được thêm vào, giá trị R^2 tương ứng không thể giảm đi mà còn có thể tăng thêm. Do vậy, người ta thường cố gắng thêm một biến mới vào chỉ để tăng R^2 không kể đến mức độ quan trọng của biến đó đối với vấn đề đang giải quyết.

Để ngăn chặn tình trạng "có đưa thêm biến vào mô hình" như đã nêu trên, một phép đo khác về mức độ thích hợp được sử dụng thường xuyên hơn. Phép đo này gọi là **R^2 hiệu chỉnh** hoặc **R^2 hiệu chỉnh theo bậc tự do** (chúng ta thấy kết quả này trong kết quả in ra của máy tính ở Chương 3). Để phát triển phép đo này, trước hết phải nhớ là R^2 đo lường tỷ số giữa phương sai của Y "được giải thích" bằng mô hình; một cách tương đương, nó bằng một trừ tỷ số "không được giải thích" do phương sai của sai số Var(u). Phép đo tự nhiên gọi là $\bar{R}^2$ (R-ngang bình phương), bằng

$$\bar{R}^2 = 1 - \frac{\widehat{\text{Var}(u)}}{\widehat{\text{Var}(Y)}}$$

Chúng ta biết rằng một ước lượng không thiên lệch của σ^2 = Var (u) được tính bằng ESS/($n - k$), và một ước lượng không thiên lệch của Var (Y) được tính bằng TSS/($n - 1$). Thay vào phương trình trên ta có

$$\bar{R}^2 = 1 - \frac{\text{ESS}/(n-k)}{\text{TSS}/(n-1)} = 1 - \frac{\text{ESS}(n-1)}{\text{TSS}(n-k)}$$

$$= 1 - \frac{n-1}{n-k}(1 - R^2) = 1 - \frac{\hat{\sigma}^2 (n-1)}{\text{TSS}}$$

Việc thêm vào một biến dẫn đến tăng R^2 nhưng cũng làm giảm đi một bậc tự do, bởi vì chúng ta đang ước lượng thêm một tham số nữa. R^2 hiệu chỉnh là một phép đo độ thích hợp tốt hơn bởi vì nó cho phép đánh đổi giữa việc tăng R^2 và giảm bậc tự do. Cũng cần lưu ý là vì $(n-1) / (n-k)$ không bao giờ nhỏ hơn 1. $\bar{R}^2$ sẽ không bao giờ lớn hơn R^2. Tuy nhiên, mặc dù R^2 không thể âm, $\bar{R}^2$ có thể nhỏ hơn không. Ví dụ, khi n = 26, k = 6, và R^2 = 0,1, chúng ta có $\bar{R}^2$ = – 0,125. $\bar{R}^2$ âm cho thấy là mô hình không mô tả đầy đủ quá trình phát dữ liệu.

VÍ DỤ 4.3

Bảng 4.2 trình bày các hệ số hồi qui ước lượng và các trị thống kê liên quan của bốn mô hình khác nhau (Phần thực hành máy tính 4.1 có hướng dẫn các tạo những số này). Các dữ liệu thấp hơn bậc tự do (d.f.) được thảo luận trong phần tiếp theo. Mô hình A giống như mô hình đã được trình bày trong Chương 3. Trong mô hình B, BEDRMS được thêm vào và trong mô hình C cả BEDRMS và BATHS đều được thêm vào. Mô hình D không có các biến giải thích, chỉ có số hạng không thay đổi. Nó sẽ được sử dụng trong phần 4.4. Rõ ràng từ Bảng 4.2, khi càng nhiều biến được thêm vào, tổng bình phương phần dư giảm và R^2 tăng. Tuy nhiên, $\bar{R}^2$ lại giảm khi thêm các biến. Điều này có nghĩa là lợi ích trong việc R^2 tăng ít hơn so với mất mát do giảm bậc tự do, dẫn đến mất mát ròng trong "mức độ thích hợp". Mô hình D có một giá trị R^2 bằng không vì các giá trị ESS và TSS của nó là như nhau. Điều này không lạ gì bởi vì không có phần nào trong

mô hình giải thích thay đổi về PRICE. Nó được đề cập ở đây vì nó sẽ có ích trong việc kiểm định giả thuyết (đề cập ở phần 4.4)

Trong mô hình A. SQFT giải thích 80,6 phần trăm của các thay đổi về giá nhà. Tuy nhiên, khi tất cả ba biến đều được đưa vào, mô hình giải thích được 78,7 phần trăm thay đổi về giá, điều này hợp lý đối với nghiên cứu chéo. Nếu các biến bổ sung được thêm vào, khả năng giải thích của mô hình sẽ cao hơn. Ví dụ, kích thước, số lượng và loại các đồ gia dụng … v.v. cũng là những biến có thể thêm vào. Tuy nhiên, khi các dữ liệu này không có sẵn trong mẫu dữ liệu, chúng ta không thể thêm nhiều biến nữa vào. Trong Chương 7, chúng ta thảo luận về tác động của hồ bơi đến giá nhà.

Bảng 4.2 Các Mô Hình Ước Lượng Cho Dữ Liệu Giá Nhà

Biến số	Mô hình A	Mô hình B	Mô hình C	Mô hình D
HẰNG SỐ	52,351	121,179	129,062	317,493
	(1,404)	(1,511)	(1,462)	(13,423)
SQFT	0,13875	0,14831	0,1548	
	(7,407)	(6,993)	(4,847)	
BEDRMS		– 23,911	– 21,588	
		(– 0,970)	(– 0,799)	
BATHS			– 12,193	
			(– 0,282)	
ESS	18.274	16.833	16.700	101.815
R^2	0,821	0,835	0,836	0,000
$\bar{R}^2$	0,806	0,805	0,787	0,000
F	54,861	27,767	16,989	180,189
d.f.	12	11	10	13
SGMASQ	1.523*	1.530	1.670	7.832
AIC	1.737*	1.846	2.112	8.389
FPE	1.740*	1.858	2.147	8.391
HQ	1.722*	1.822	2.077	8.354
SCHWARZ	1.903*	2.117	2.535	8.781
SHIBATA	1.678*	1.718	1.874	8.311
GCV	1.777*	1.948	2.338	8.434
RICE	1.827*	2.104	2.783	8.485

Ghi chú: các giá trị trong ngoặc là những trị thống kê *t* tương ứng, đó là các hệ số chia cho sai số chuẩn của chúng.

* Đánh dấu mô hình "tốt nhất" đối với tiêu chuẩn, nghĩa là, có giá trị nhỏ nhất

⭘ BÀI THỰC HÀNH 4.3

Chứng minh rằng $\bar{R}^2$ và $\hat{\sigma}^2$ chuyển động ngược chiều nhau; nghĩa là nếu $\bar{R}^2$ tăng, thì $\hat{\sigma}^2$ nhất thiết phải giảm. (Vì vậy, chọn một mô hình có $\bar{R}^2$ cao hơn đồng nghĩa với chọn một mô hình có $\hat{\sigma}^2$ thấp hơn.)

Tính R^2 và $\bar{R}^2$ khi không có số hạng không đổi *

Tổng bình phương gộp TSS = RSS + ESS chỉ có giá trị khi và chỉ khi mô hình có số hạng không đổi. Nếu mô hình không có số hạng không đổi, tổng bình phương gộp thích hợp là $\Sigma Y_t^2 = \Sigma \hat{Y}_t^2 + \Sigma \hat{u}_t^2$. Lưu ý là giá trị trung bình $\bar{Y}$ không được trừ ra ở đây. Một số chương trình máy tính tính R^2 bằng $1 - (ESS/\Sigma Y_t^2)$ khi không có số hạng tung độ gốc. Công thức này được Viện Tiêu chuẩn và Công nghệ Quốc gia đề nghị sử dụng. Tuy nhiên, có thể chỉ ra là giá trị tính theo cách này không tương thích với giá trị tính bằng TSS vì các mẫu số khác nhau giữa hai mô hình. Nếu mục tiêu là so sánh các mô hình có và không có số hạng không thay đổi, về mặt mức độ thích hợp, công thức tính R^2 không thể độc lập với mô hình. Tốt hơn nên dùng $1 - (ESS/TSS)$ trong cả hai trường hợp để có thể so sánh được R^2. Nếu R^2 được tính bằng TSS trong mẫu số, có thể nó sẽ có giá trị âm khi số hạng không đổi không có mặt trong mô hình. Giá trị âm như vậy thể hiện mô hình có thể không được đặc trưng tốt. Một lựa chọn khác và có lẽ là một phép đo tốt hơn của R^2 là bình phương của hệ số tương quan giữa Y_t và $\hat{Y}_t$, giá trị luôn luôn không âm.

Chúng ta đã lập luận trước đây là $\bar{R}^2 = 1 - [\text{Var}(u) / \text{Var}(Y)]$ là phép đo tốt hơn của thay đổi trong biến Y được giải thích bởi mô hình. Điều này cho công thức

$$\bar{R}^2 = 1 - \frac{\text{ESS} \div (n - k)}{\text{TSS} \div (n - 1)}$$

trong *mọi* trường hợp.

Vì các chương trình máy tính khác nhau về cách tính R^2 và $\bar{R}^2$ trong trường hợp không có số hạng không đổi, vì vậy đề nghị độc giả kiểm tra bất kỳ chương trình nào được sử dụng và xác định xem các phép đo có tương thích giữa các mô hình hay không. Các nhà điều tra thường loại số hạng không đổi ra nếu nó không có ý nghĩa để làm tăng mức ý nghĩa thống kê của các biến còn lại (ví dụ, mô hình giá tài sản vốn của Ví dụ 1.3 không có số hạng không đổi), việc thực hành này không được khuyến khích vì nó có thể dẫn đến mô hình không đặc trưng (xem thêm ở phần 4.5)

⭘ 4.3 Các Tiêu Chuẩn Chung Để Chọn Mô Hình

Chúng ta đã chứng minh trước đây bằng cách tăng số biến trong một mô hình, tổng bình phương phần dư $\Sigma \hat{u}_t^2$ sẽ giảm và R^2 sẽ tăng, nhưng đổi lại bậc tự do sẽ giảm. $\bar{R}^2$ và sai số chuẩn của phần dư, $[ESS / (n - k)]^{1/2}$, tính đến việc đánh đổi giữa giảm ESS và giảm bậc tự do. Đây là những tiêu chuẩn thông dụng nhất để so sánh các mô hình.

Nhìn chung, mô hình đơn giản hơn được ưa thích hơn vì hai lý do kỹ thuật sau. Thứ nhất, đưa quá nhiều biến vào mô hình khiến cho độ chính xác tương đối của riêng mỗi hệ số giảm. Điều này sẽ được nghiên cứu kỹ trong Chương 5. Thứ hai, việc giảm bậc tự do sẽ giảm năng lực của kiểm định trên các hệ số. Vì vậy, xác suất của việc không bác bỏ giả thuyết sai (sai lầm loại II) tăng khi bậc tự do giảm. Các mô hình đơn giản cũng dễ hiểu hơn các mô hình phức tạp. Vì vậy, lý tưởng nên thiết lập những tiêu chuẩn hạn chế những mô hình lớn nhưng cũng không luôn luôn chọn mô hình đơn giản.

Trong những năm gần đây, nhiều tiêu chuẩn chọn mô hình được đề nghị. Tất cả những tiêu chuẩn này có dạng của tổng bình phương phần dư (ESS) nhân với một nhân tố bất lợi phụ thuộc vào mức độ phức tạp của mô hình. Mô hình càng phức tạp ESS càng giảm nhưng lại tăng tính bất lợi. Các tiêu chuẩn vì vậy phải cung cấp các loại đánh đổi khác giữa mức độ thích hợp và độ phức tạp của mô hình. Một mô hình có trị thống kê tiêu chuẩn thấp được ưa chuộng hơn. Trong phần này, chúng ta trình bày tóm tắt tổng quát các nhân tố bất lợi mà không đi sâu vào phần kỹ thuật của mỗi yếu tố. Nếu độc giả quan tâm đến một tóm tắt đầy đủ chi tiết hơn cùng với những ứng dụng, bạn có thể tham khảo bài báo của Engle và Brown (1985).

Akaike (1970, 1974) xây dựng hai phương pháp, một được gọi là **sai số hoàn toàn xác định trước (FPE)** và phương pháp thứ hai gọi là **tiêu chuẩn thông tin Akaike (AIC)**. Hannan và Quinn (1979) đề nghị một phương pháp khác (được gọi là **tiêu chuẩn HQ**). Các tiêu chuẩn khác gồm của Schwarz (1978), Shibata (1981), và Rice (1984), và phương pháp **tính chính xác chéo tổng quát (GCV)** được Craven và Wahba (1979) phát triển và được Engle, Graner, Rice, và Weiss (1986) sử dụng. Mỗi một trị thống kê này đều dựa trên vài tính chất tối ưu, chi tiết về các phương pháp này được đề cập trong các bài báo liệt kê trên (lưu ý là các bài báo này đòi hỏi kiến thức về đại số tuyến tính). Bảng 4.3 tóm tắt những tiêu chuẩn này (*n* là số lần quan sát và *k* là số thông số ước lượng).

Không cần thiết phải đưa $\bar{R}^2$ vào trong tiêu chuẩn vì $\bar{R}^2$ và SGMASQ $\hat{\sigma}^2$) quan hệ nghịch, và vì vậy giá trị SGMASQ thấp cũng có nghĩa là $\bar{R}^2$ sẽ có giá trị cao. $\bar{R}^2$ chỉ có ích khi xác định tỷ số của biến đổi trong Y được giải thích bởi các biến X.

Bảng 4.3 Tiêu Chuẩn Chọn Mô Hình

SGMASQ:	$\left(\frac{ESS}{n}\right)\left(1-\left(\frac{k}{n}\right)\right)^{-1}$	HQ:	$\left(\frac{ESS}{n}\right)(\ln n)^{2k/n}$
AIC:	$\left(\frac{ESS}{n}\right)e^{(2k/n)}$	RICE:	$\left(\frac{ESS}{n}\right)\left(1-\left(\frac{2k}{n}\right)\right)^{-1}$
FPE:	$\left(\frac{ESS}{n}\right)\frac{n+k}{n-k}$	SCHWARZ:	$\left(\frac{ESS}{n}\right)n^{k/n}$
GVC:	$\left(\frac{ESS}{n}\right)\left(1-\left(\frac{k}{n}\right)\right)^{-2}$	SHIBATA:	$\left(\frac{ESS}{n}\right)\frac{n+2k}{n}$

Một cách lý tưởng, chúng ta muốn có một mô hình có các giá trị của các trị thống kê đều thấp, khi so sánh với một mô hình khác. Mặc dù có thể xếp hạng một vài tiêu chuẩn này đối với một giá trị ESS, *n*, và *k* cho trước, thứ tự này sẽ không còn ý nghĩa nữa bởi vì các mô hình đều có ESS và *k* khác nhau. Ramanathan (1992) khảo sát kỹ hơn một số trường hợp đặc biệt. Trong những trường hợp dặc biệt này, một số tiêu chuẩn trở nên dư thừa – nghĩa là, một mô hình ưu việt hơn theo một tiêu chuẩn cũng sẽ ưu việt hơn xét theo các tiêu chuẩn khác. Tuy nhiên, một cách tổng quát, có thể tìm được một mô hình ưu việt theo một tiêu chuẩn nhưng lại không ưu việt theo tiêu chuẩn khác. Ví dụ, tiêu chuẩn Schwarz coi trọng về tính phức tạp của mô hình hơn là các yếu tố khác và vì vậy có thể dẫn đến một kết luận khác. Một mô hình tốt hơn một mô hình khác theo một số tiêu chuẩn sẽ được ưa chuộng hơn. Tuy nhiên, tiêu chuẩn AIC là tiêu chuẩn được sử dụng phổ biến nhất trong phân tích chuỗi thời gian.

❹ VÍ DỤ 4.4

Đối với dữ liệu giá nhà ở, Bảng 4.2 có 8 trị thống kê lựa chọn mô hình đối với mỗi một trong ba mô hình. Tất cả các tiêu chuẩn đều đánh giá cao mô hình đơn giản nhất, trong mô hình đó chỉ có một biến giải thích duy nhất là SQFT. Điều này có nghĩa là việc giảm ESS do tính phức tạp của mô hình không đủ để đánh đổi với nhân tố bất lợi gắn liền với mô hình phức tạp. Kết quả này thật sự không quá bất ngờ đối với chúng ta. Diện tích sử dụng phụ thuộc vào số phòng ngủ và phòng tắm trong nhà. Mô hình A vì vậy không trực tiếp đề cập đến BEDRMS và BATHS. Do đó, chúng ta không nên kỳ vọng mô hình B và C sẽ tốt hơn khi giảm ESS đủ thấp.

❹ 4.4 Kiểm Định Giả Thuyết

Trong phần này chúng ta thảo luận ba loại kiểm định giả thuyết: (1) kiểm định mức ý nghĩa thống kê của các hệ số riêng lẻ, (2) kiểm định một số hệ số hồi qui liên kết, và (3) kiểm định tổ hợp tuyến tính của các hệ số hồi qui.

Kiểm Định Các Hệ Số Riêng Lẻ

Như trong Chương 3, kiểm định giả thuyết về một hệ số hồi qui đơn được tiến hành bằng kiểm định *t*. Các tính chất mà mỗi $\hat{\beta}_i$ tuân theo phân phối chuẩn và $ESS/\sigma^2 = (n - k)\hat{\sigma}^2/\sigma^2$ tuân theo phân phối chi bình phương cũng được mở rộng cho trường hợp đa biến. Chỉ có một hiệu chỉnh là ESS/σ^2 phân phối chi bình phương với $n - k$ d.f. Các bước tiến hành kiểm định một hệ số riêng biệt như sau:

KIỂM ĐỊNH T MỘT PHÍA

Bước 1 $H_o: \beta = \beta_0, H_1: \beta > \beta_0.$

Bước 2 Thiết lập trị thống kê $t_c = (\hat{\beta} - \beta_0) / \hat{\beta}$, với $\hat{\beta}$ là giá trị ước lượng và $\hat{\beta}$ là sai số chuẩn ước lượng của nó. Nếu $\beta_0 = 0$, giá trị *t* này sẽ giảm đến tỷ số của hệ số hồi qui chia cho sai số chuẩn của nó. Với giả thuyết H_0, nó tuân theo phân phối *t* với $n - k$ d.f.

Bước 3 Tìm trong bảng tra *t* giá trị tương ứng với bậc tự do bằng $n-k$ và tìm điểm $t^*_{n-k}(\alpha)$ sao cho diện tích của phần bên phải điểm này bằng mức ý nghĩa (α).

Bước 4 Bác bỏ *giả thuyết không* nếu $t_c > t^*$. Nếu trường hợp $H_1 : \beta < \beta_0$, H_0 sẽ bị bác bỏ nếu $t_c < -t^*$. Một cách tương đương cho cả hai trường hợp, bác bỏ nếu $|t_c| > t^*$.

Để sử dụng phương pháp giá trị *p*, tính $p = P(t > |t_c|$, với H_0 cho trước) và bác bỏ H_0 nếu giá trị *p* nhỏ hơn mức ý nghĩa.

❹ VÍ DỤ 4.5

Chúng ta hãy áp dụng với Mô hình B và C trong Bảng 4.2. Mô hình B có bậc tự do là 11 d.f. (14 – 3) và Mô hình C có bậc tự do bằng 10. Từ Bảng A.2, $t^*_{11}(0,05) = 1,796$ *và* $t^*_{10}(0,05) = 1,812$ đối với kiểm định 5%. Vì vậy, để một hệ số hồi qui dương hoặc âm có ý nghĩa thống kê, giá trị tuyệt đối của trị thống kê *t* cho trong Bảng 4.2 phải lớn hơn 1,796 đối với Mô hình B và lớn hơn 1,812 đối với Mô hình C. Chúng ta lưu ý là trong mỗi mô hình hệ số hồi qui của SQFT là có ý nghĩa. Điều này có nghĩa là trong những trường hợp đó chúng ta không thể bác bỏ *giả thuyết không* là hệ số tương ứng bằng không.

Có hay không một mức ý nghĩa nào khác 5 phần trăm tại đó ta có thể bác bỏ *giả thuyết không* được? Sau cùng, không có gì đặc biệt đối với mức 5 phần trăm. Nếu mức ý nghĩa thực sự cao hơn một chút, chúng ta vẫn có thể sẵn sàng bác bỏ *giả thuyết không*. Chúng ta lưu ý từ Bảng A.2 là đối với mức ý nghĩa 10 phần trăm, $t^*_{10}(0,1) = 1,372$. Trị thống kê *t* của BEDRMS trong Mô hình C là 0,799 về trị tuyệt đối, nhỏ hơn 1,372. Do đó, chúng ta kết luận là BEDRMS không có ý nghĩa trong Mô hình C, ở mức ý nghĩa 10 phần trăm.

Sử dụng chương trình GRETL, chúng ta đã tính giá trị *p* cho các hệ số của BEDRMS và BATHS (xem phần thực hành máy tính 4.1). Các hệ số này xếp từ 0,175 đến 0,39, ngụ ý là nếu chúng ta bác bỏ *giả thuyết không* là các hệ số này bằng không, có một cơ hội từ 17,5 đến 39 phần trăm phạm sai lầm loại I. Khi các hệ số này cao hơn một mức chấp nhận thông thường, chúng ta không bác bỏ H_0 nhưng thay vì vậy, kết luận là các hệ số này không khác không một cách có ý nghĩa.

KIỂM ĐỊNH *t* HAI PHÍA

Bước 1 $H_0: \beta = \beta_0$, $H_1: \beta \neq \beta_0$.

Bước 2 Thiết lập trị thống kê *t*, $t_c = (\hat{\beta} - \beta_0)/\hat{\beta}$, với $\hat{\beta}$ là giá trị ước lượng và $\hat{\beta}$ là sai số chuẩn của nó. Theo giả thuyết H_0, $\hat{\beta}$ tuân theo phân phối *t* với bậc tự do $n-k$

Bước 3 Tìm trong Bảng *t* A.2 giá trị tương ứng với bậc tự do $n-k$ và tìm $t^*_{n-k}(\alpha/2)$ sao cho diện tích bên phải của nó bằng phân nửa mức ý nghĩa.

Bước 4 Bác bỏ *giả thuyết không* nếu $|t_c| > t^*$.

Để sử dụng giá trị *p*, tính giá trị $p = 2P(t > |t_c|$, với H_0 cho trước) và bác bỏ H_0 nếu *p* nhỏ hơn mức ý nghĩa.

Tóm tắt, giá trị *p* (giống như xác suất của sai lầm loại I bác bỏ giả thuyết đúng) thấp nghĩa là chúng ta "an toàn" khi bác bỏ *giả thuyết không* là hệ số bằng không (đối với

$\beta_0 = 0$) và kết luận là hệ số này khác không đáng kể. Nếu giá trị *p* cao, thì chúng ta không thể bác bỏ *giả thuyết không* nhưng thay vào đó kết luận là hệ số không có ý nghĩa thống kê.

❍ VÍ DỤ 4.6

Chúng ta áp dụng kiểm định hai phía với Mô hình B và C. Trong Mô hình B, bậc tự do là 11 vì vậy $t^*_{11}(0{,}025)$ là 2,201 đối với mức ý nghĩa 5 phần trăm. Trong Mô hình C, $t^*_{10}(0{,}025) = 2{,}228$. Vì vậy, để một hệ số hồi qui khác không có ý nghĩa tại mức ý nghĩa 5 phần trăm, trị thống kê *t* cho trong bảng 4.2 phải lớn hơn 2,201 về giá trị tuyệt đối ở Mô hình B và lớn hơn 2,228 về giá trị tuyệt đối ở Mô hình C. Chúng ta lưu ý là trong mỗi mô hình hệ số hồi qui của SQFT đều có ý nghĩa, trong khi tất cả các hệ số hồi qui khác không có ý nghĩa. Điều này có nghĩa là trong những trường hợp đó chúng ta không thể bác bỏ *giả thuyết không* là hệ số tương ứng bằng không.

Có hay không một mức ý nghĩa khác ngoài mức 5 phần trăm có thể bác bỏ được *giả thuyết không*? Giá trị *p* bây giờ bằng hai lần các giá trị có trước đây (đó là 0,35 đến 0,78). Khi các giá trị này cao, kết luận là các giá trị khác không quan sát được của những hệ số hồi qui này có thể là do sai số mẫu ngẫu nhiên. Vì vậy, với giá trị SQFT cho trước, các biến BEDRMS và BATHS không ảnh hưởng quan trọng đến giá căn nhà. Kết quả này khẳng định kết quả trước đó trong Mô hình A đã được đánh giá là tốt theo tất cả 8 tiêu chuẩn.

❍ BÀI TẬP THỰC HÀNH 4.4

Sử dụng chương trình hồi qui của bạn, ước lượng Mô hình B và C, và kiểm tra kết quả trong Bảng 4.2.

Có thể thiết lập được tính chất sau (xem Haitovsky, 1969):

Tính chất 4.2

Nếu giá trị tuyệt đối của trị thống kê *t* của một hệ số hồi qui nhỏ hơn 1, thì việc loại hệ số này ra khỏi mô hình sẽ làm tăng R^2 hiệu chỉnh. Tương tự, bỏ một biến có trị thống kê *t* lớn hơn 1 (về giá trị tuyệt đối) sẽ làm giảm $\bar{R}^2$.

Điều này có thể chỉ ra là, bên cạnh trị thống kê *t* tới hạn, chúng ta có thể sử dụng giá trị *t* bằng 1 như là hướng dẫn trong việc xác định xem có thể bỏ bớt một biến hay không. Tuy nhiên, vì $\bar{R}^2$ chỉ là một trong nhiều tiêu chuẩn nên các giá trị *p* riêng lẻ, giá trị thống kê chọn mô hình và tầm quan trọng về lý thuyết của các biến nên được dùng để xác định các biến nào có thể loại bỏ (xem ví dụ phần 4.6 và 4.7)

Kiểm định một số hệ số liên kết (kiểm định Wald)

Kiểm định *t* về các hệ số riêng lẻ dùng cho mức ý nghĩa của các hệ số cụ thể. Ta cũng có thể kiểm định **ý nghĩa liên kết** của một số hệ số hồi qui, ví dụ như các mô hình dưới đây:

$$\text{(U)} \qquad \text{PRICE} = \beta_1 + \beta_2\text{SQFT} + \beta_3\text{BEDROOMS} + \beta_4\text{BATHS} + u$$

(R) $\text{PRICE} = \gamma_1 + \gamma_2 \text{SQFT} + v$

Mô hình U (là mô hình C trong Bảng 4.2) được gọi là **mô hình không giới hạn**, và Mô hình R (là Mô hình A trong Bảng 4.2) được gọi là **mô hình giới hạn.** Đó là do β_3 và β_4 buộc phải bằng không trong Mô hình R. Ta có thể kiểm định giả thuyết liên kết $\beta_3 = \beta_4 = 0$ với giả thuyết đối là ít nhất một trong những hệ số này không bằng không. Kiểm định giả thuyết liên kết này được gọi là **kiểm định Wald** (Wald, 1943). Thủ tục như sau.

Kiểm định Wald tổng quát Đặt các mô hình giới hạn và không giới hạn là (bỏ qua ký hiệu *t* ở dưới):

(U) $Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \beta_{m+1} X_{m+1} + \ldots + \beta_k X_k + u$
(R) $Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_m X_m + v$

Mặc dù Mô hình U có vẻ khác nhưng nó hoàn toàn giống Phương trình (4.1). Mô hình R có được bằng cách bỏ bớt một số biến ở Mô hình U, đó là X_{m+1}, X_{m+2}, …X_k. Vì vậy, *giả thuyết không* là $\beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0$. Lưu ý rằng (U) chứa *k* hệ số hồi qui chưa biết và (R) chứa *m* hệ số hồi qui chưa biết. Do đó, Mô hình R có ít hơn $k - m$ thông số so với U. Câu hỏi chúng ta sẽ nêu ra là $k - m$ biến bị loại ra có ảnh hưởng *liên kết* có ý nghĩa đối với Y hay không.

Giả sử những biến bị loại này không có ảnh hưởng có ý nghĩa đối với Y. Chúng ta sẽ không kỳ vọng tổng bình phương sai số của Mô hình R (ESS_R) quá khác biệt với tổng bình phương sai số của Mô hình U (ESS_U). Nói cách khác, sai biệt $\text{ESS}_R - \text{ESS}_U$ có vẻ rất nhỏ. Nhưng giá trị này nhỏ như thế nào? Chúng ta biết là ESS rất nhạy với đơn vị đo lường, và vì vậy có thể làm giá trị này lớn hơn hay nhỏ hơn chỉ đơn giản bằng cách thay đổi thang đo. "Nhỏ" hoặc "lớn" được xác định bằng cách so sánh sai biệt trên với ESS_U, tổng bình phương sai số của mô hình hoàn toàn không giới hạn. Vì vậy, $\text{ESS}_R - \text{ESS}_U$ được so sánh với ESS_U. Nếu giá trị đầu "nhỏ" tương đối so với giá trị sau, chúng ta kết luận là việc loại bỏ các biến X_{m+1}, X_{m+2}, …, X_k không thay đổi ESS đủ để có thể tin là các hệ số của chúng có ý nghĩa.

Chúng ta biết là các tổng của những bình phương độc lập có phân phối chi bình phương (xem phần 2.7). Vì vậy, ESS_U/σ^2 là phân phối chi bình phương với $n - k$ bậc tự do (*n* quan sát trừ *k* thông số trong Mô hình U). Có thể thấy trong *giả thuyết không* là vì tính chất cộng của chi bình phương (Tính chất 2.12b), $(\text{ESS}_R - \text{ESS}_U)/\sigma^2$ cũng là phân phối chi bình phương với bậc tự do bằng số biến số loại bỏ trong (R). Trong phần 2.7, chúng ta thấy là tỷ số của hai phân bố chi bình phương độc lập có phân phối *F* có hai thông số: bậc tự do cho tử số của tỷ số, bậc tự do cho mẫu số. Trị thống kê sẽ căn cứ trên tỷ số *F*.

Các bước thông thường để kiểm định Wald (thường được gọi là kiểm định *F)* như sau:

Bước 1 *Giả thuyết không* là H_0: $\beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0$. Giả thuyết ngược lại là H_1: có ít nhất một trong những giá trị β không bằng không. *Giả thuyết không* có $k - m$ ràng buộc.

Bước 2 Trước tiên hồi qui Y theo một biến không đổi, X_2, X_3, …, X_k, và tính tổng bình phương sai số ESS_U. Kế đến hồi qui Y theo một biến không đổi, X_2, X_3, …, X_m

và tính ESS_R. Chúng ta biết từ Tính chất 4.1b là ESS_U/σ^2 tuân theo phân phối chi bình phương với bậc tự do $DF_U = n - k$ (nghĩa là n số quan sát trừ k hệ số ước lượng). Tương tự, với *giả thuyết không*, ESS_R/σ^2 tuân theo phân phối chi bình phương với bậc tự do $DF_R = n - m$. Có thể thấy là chúng độc lập và với tính chất cộng được của phân phối chi bình phương, sai biệt của chúng $(ESS_R - ESS_U) / \sigma^2$ cũng phân phối chi bình phương, với bậc tự do bằng sai biệt về bậc tự do, nghĩa là, $DF_R - DF_U$. Lưu ý là $DF_R - DF_U$ cũng bằng $k - m$, là số ràng buộc trong *giả thuyết không* (đó là số biến bị loại bỏ). Trong phần 2.7, chúng ta đã định nghĩa phân phối F là tỷ số của hai biến ngẫu nhiên phân phối chi bình phương độc lập. Điều này cho ta trị thống kê

$$F_c = \frac{(ESS_R - ESS_U) \div (DF_R - DF_U)}{ESS_U \div DF_U} \qquad \textbf{(4.3)}$$

$$= \frac{(ESS_R - ESS_U) / (k - m)}{ESS_U / (n - k)}$$

$$= \frac{\text{(sai biệt trong ESS} \div \text{số ràng buộc)}}{\text{(tổng bình phương sai số của Mô hình U} \div \text{d.f. của Mô hình U)}}$$

$$= \frac{(R_U^2 - R_R^2)/ (k - m)}{(1 - R_U^2) / (n - k)}$$

với R^2 là số đo độ thích hợp không hiệu chỉnh. Chia cho bậc tự do ta được tổng bình phương trên một bậc tự do. Với *giả thuyết không*, F_c có phân phối F với $k - m$ bậc tự do đối với tử số và $n - k$ bậc tự do đối với mẫu số.

Bước 3 Từ số liệu trong bảng F tương ứng với bậc tự do $k - m$ cho tử số và $n - k$ cho mẫu số, và với mức ý nghĩa cho trước (gọi là α), ta có $F^*_{k-m,n-k}(\alpha)$ sao cho diện tích bên phải của F^* là α.

Bước 4 Bác bỏ *giả thuyết không* ở mức ý nghĩa α nếu $F_c > F^*$. Đối với phương pháp giá trị p, tính giá trị $p = P(F > F_c|H_0)$ và bác bỏ *giả thuyết không* nếu giá trị p *nhỏ hơn* mức ý nghĩa.

❹ VÍ DỤ 4.7

Trong ví dụ về bất động sản của chúng ta, H_0: $\beta_3 = \beta_4 = 0$ và H_1: có ít nhất một giá trị β không bằng không. Vì vậy, Mô hình U giống như Mô hình C trong Bảng 4.2, và Mô hình R chính là Mô hình A. Số ràng buộc sẽ là 2. Cũng vậy, ESS_R = 18.274 và ESS_U = 16.700 (xem Bảng 4.2). Bậc tự do của Mô hình U là 10. Vì vậy, trị thống kê F được tính

$$F_c = \frac{(18.274 - 16.700) / 2}{16.700 / 10} = 0{,}471$$

Từ bảng F (Bảng A.4b), $F^*_{2,10}(0{,}05) = 4{,}1$. Vì F_c không lớn hơn F^*, chúng ta không thể bác bỏ *giả thuyết không*, và vì vậy chúng ta kết luận là β_3 và β_4 thật sự không có ý nghĩa ở mức 5 phần trăm. Ngay cả nếu mức ý nghĩa là 10 phần trăm (xem Bảng A.4c), $F^*_{2,10}(0{,}1) = 2{,}92 > F_c$. Điều này có nghĩa là về phương diện mức ý nghĩa của các biến độc lập, Mô hình A đơn giản hơn và tốt hơn. Kiểm định tương tự cũng có thể thực hiện để so sánh Mô hình A và B, nhưng việc này không cần thiết vì sai biệt giữa hai mô

hình này chỉ do một biến, đó là BEDRMS. Trong trường hợp này, phân phối *F* chỉ có một bậc tự do ở tử số. Khi điều này xảy ra, giá trị của *F* đơn giản chỉ là bình phương của trị thống kê *t* đối với BEDRMS (xem Tính chất 2.14b). Chứng minh điều này rất dễ. Mô hình B bây giờ là không giới hạn và vì vậy

$$F_c = \frac{(18.274 - 16.700) / 1}{16.700 / 11} = 0{,}942$$

Có căn bậc hai là 0,97, bằng với trị thống kê *t* trong Bảng 4.2. Vì vậy, *kiểm định Wald cần phải tiến hành chỉ khi có hai hoặc nhiều hơn hai hệ số hồi qui bằng không trong giả thuyết không.*

Giá trị *p* trong ví dụ này là $P(F > 0{,}471) = 0{,}64$. Bởi vì có 64 phần trăm cơ hội bác bỏ một giả thuyết đúng H_0 (là các hệ số của BEDRMS và BATHS bằng không) là quá cao không thể chấp nhận được, nên chúng ta không thể bác bỏ H_0 nhưng thay vào đó ta kết luận là các hệ số có giá trị khác không, không có ý nghĩa thống kê.

Chúng ta thấy từ Bảng 4.2 là số hạng không đổi không có ý nghĩa trong bất kỳ mô hình nào (trừ Mô hình D). Tuy nhiên, thật không khôn ngoan khi loại bỏ số hạng không đổi khỏi mô hình. Đó là do số hạng không đổi thể hiện một cách không gián tiếp một số các ảnh hưởng trung bình của các biến bị loại bỏ (vấn đề này được thảo luận đầy đủ hơn trong phần 4.5). Do đó, việc loại bỏ số hạng không thay đổi có thể dẫn đến sai nghiêm trọng trong đặc trưng của mô hình.

Kiểm định Wald đặc biệt về độ thích hợp tổng quát Hãy xem xét một trường hợp đặc biệt của kiểm định Wald trong hai mô hình sau:

(U) $Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_k X_k + u$
(SR) $Y = \beta_1 + w$

Mô hình U là mô hình hồi qui bội trong phương trình (4.1), với X_1 là số hạng không thay đổi. Trong Mô hình SR (thật giới hạn), tất cả các biến ngoại trừ số hạng không thay đổi đều bị loại khỏi mô hình; nghĩa là, chúng ta đặt $k - 1$ ràng buộc $\beta_2 = \beta_3 = \ldots = \beta_k = 0$. Giả thuyết này sẽ kiểm định phát biểu "Không một hệ số nào trong mô hình (ngoại trừ số hạng không thay đổi) có ý nghĩa thống kê." Có thể thực hiện kiểm định Wald cho giả thuyết này. Nếu giả thuyết không bị bác bỏ, chúng ta kết luận là không có biến nào có thể giải thích một cách liên kết thay đổi của Y. Điều này có nghĩa là chúng ta có một mô hình xấu và phải thiết lập lại mô hình này. ESS_U là tổng bình phương sai số của mô hình đầy đủ.

Để có ESS_{SR}, trước hết chúng ta cực tiểu $\Sigma w^2_t = \Sigma (Y_t - \beta_1)^2$ theo β_1. Dễ dàng chứng minh được là $\hat{\beta}_1 = \bar{Y}$ (xem chứng minh ở phần 2.5). Do đó, ta có $ESS_{SR} = \Sigma(Y_t - \bar{Y})^2$ giống như tổng bình phương toàn phần (TSS_U) của Mô hình U (đây cũng là tổng bình phương của Mô hình SR). Trị thống kê *F* trở thành

$$F_c = \frac{(TSS_U - ESS_U)/(k-1)}{ESS_U/(n-k)} = \frac{RSS_U/(k-1)}{ESS_U/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} \qquad \textbf{(4.4)}$$

giá trị này có thể được tính từ R^2 không hiệu chỉnh của mô hình đầy đủ. Các chương trình hồi qui đều cung cấp trị thống kê F này trong phần tóm tắt thống kê của một mô hình. Nhiệm vụ đầu tiên là phải đảm bảo rằng *giả thuyết không* của kiểm định F này bị bác bỏ, nghĩa là, $F_c > F^*_{k-1,\, n-k}(\alpha)$. Nếu không, chúng ta có một mô hình trong đó không có biến độc lập nào giải thích được những thay đổi trong biến phụ thuộc, và vì vậy mô hình cần được thiết lập lại.

○ VÍ DỤ 4.8

Bảng 4.2 cung cấp trị thống kê F kiểm định Wald, cho trước trong phương trình (4.4), đối với ví dụ về giá nhà. Với Mô hình C, $k = 4$, và vì vậy $k - 1 = 3$ và $n - k = 14 - 4 = 10$. Bậc tự do của trị thống kê F là 3 đối với tử số và 10 đối với mẫu số. Từ bảng F, A.4b, giá trị tới hạn đối với kiểm định ở 5 phần trăm là $F^*_{3,10}(0{,}05) = 3{,}71$. Vì giá trị F trong Bảng 4.2 là 16,989 đối với Mô hình C, chúng ta bác bỏ *giả thuyết không* là tất cả hệ số hồi qui ngoại trừ số hạng không đổi bằng không. Vì vậy, có ít nhất một hệ số hồi qui khác không có ý nghĩa thống kê. Từ kiểm định t đối với hệ số của SQFT, chúng ta đã biết được trường hợp này. Dễ dàng chứng minh được là $F^*_{2,11}(0{,}05) = 3{,}98$ đối với Mô hình B và $F^*_{1,12}(0{,}05) = 4{,}75$ đối với Mô hình A, và vì vậy tất cả các mô hình đều bác bỏ *giả thuyết không* là không có biến giải thích nào là có ý nghĩa.

Chúng ta lưu ý rằng các trị thống kê F của Mô hình B và C thấp hơn nhiều so với Mô hình A. Điều này là do các sai biệt trong R^2 khá nhỏ, trong khi tỷ số $(n-1)/(n-k)$ tăng đáng kể khi k tăng. Do đó chúng ta thấy từ Phương trình (4.4) có thể giải thích sai biệt lớn về F. Tuy nhiên, nói chung, các sai biệt về F giữa các mô hình là không quan trọng. Chỉ có kết quả của kiểm định Wald là đáng quan tâm.

○ BÀI TẬP THỰC HÀNH 4.5

Trong Bảng 4.2, Mô hình D là mô hình thật giới hạn về hồi qui PRICE chỉ theo số hạng không đổi. So sánh mô hình này với Mô hình C là mô hình không giới hạn, và chứng minh giá trị F của kiểm định Wald được báo cáo trong Bảng 4.2 của Mô hình C. Sau đó thực hiện đúng như vậy cho Mô hình A và B. Cuối cùng, giải thích tại sao $R^2 = \bar{R}^2 = 0$ đối với Mô hình D.

Khác biệt giữa hai loại kiểm định F cần được ghi chú cẩn thận. *Công thức cho trong Phương trình (4.4) không thể ứng dụng chỉ khi một số ít các biến bị loại bỏ. Nó có thể ứng dụng được khi mô hình giới hạn chỉ có một số hạng không đổi.* Trị thống kê F in từ chương trình máy tính kiểm định tính thích hợp chung, trong khi trị thống kê F tính được từ Phương trình (4.3) kiểm định xem một nhóm các hệ số có khác không một cách có ý nghĩa thống kê hay không. Cũng lưu ý là kiểm định F luôn luôn là kiểm định một phía.

Tính trị thống kê *F* khi mô hình không có số hạng không đổi* Trong phần 4.2, chúng ta đã thảo luận về các sai biệt của các số đo R^2 giữa hai mô hình, một với số hạng không

đổi và mô hình thứ hai không có số hạng không đổi, và lập luận rằng có thể sử dụng cùng một công thức cho cả hai trường hợp để so sánh mức độ thích hợp tương đối của chúng. Tuy nhiên, khi tính tỷ số *F* công thức được sử dụng sẽ khác. Để giải thích vì sao lại như vậy, chúng ta hãy xem xét hai mô hình sau:

(A) $Y = \beta_2 X_2 + \beta_3 X_3 + \ldots \beta_k X_k + u$
(B) $Y = w$

Với số hạng không thay đổi X_1 (=1) bị loại bỏ. Lưu ý là Mô hình không giới hạn A bây giờ chỉ có $k - 1$ thông số (có nghĩa là số bậc tự do là $n - k + 1$) và Mô hình giới hạn B không có thông số nào (với d.f. *n*). Để kiểm định độ thích hợp chung của mô hình, *giả thuyết không* lại là H_0: $\beta_2 = \beta_3 = \ldots = \beta_k = 0$, và giả thuyết ngược lại tương tự như trước. Kiểm định Wald cũng có thể áp dụng ở đây và công thức thích hợp là Phương trình (4.3). Đặt $ESS_A = \Sigma \hat{u}_t^2$ là tổng bình phương sai số của Mô hình A. Trong Mô hình B, tổng bình phương sai số sẽ là $ESS_B = \Sigma Y^2_t$. Giá trị *F* được tính bởi:

$$F_c = \frac{(ESS_B - ESS_A)/(k-1)}{ESS_A/(n-k+1)} = \frac{(\Sigma Y_t^2 - \Sigma \hat{u}_t^2)/(k-1)}{ESS_A/(n-k+1)} = \frac{\Sigma \hat{Y}_t^2/(k-1)}{ESS_A/(n-k+1)} \quad \textbf{(4.4a)}$$

bởi vì khai triển $\Sigma Y_t^2 = \Sigma \hat{Y}_t^2 + \Sigma \hat{u}_t^2$ trong đó không có số hạng không đổi. Với *giả thuyết không*, tổng này có phân phối *F* với $k - 1$ và $n - k + 1$ bậc tự do. Tiêu chuẩn để chấp nhận/bác bỏ H_0 cũng tương tự. Giá trị thống kê *F* đại diện cho Mô hình D kiểm định giả thuyết là số hạng không đổi bằng không. Vì chỉ có một hệ số sẽ bị loại khỏi đây, giá trị *F* là bình phương của trị thống kê *t*. Do đó, *F* = 180,189 mặc dù $R^2 = 0$. Lưu ý công thức này chỉ được dùng để kiểm định độ thích hợp chung hoàn toàn khác với công thức trong Phương trình (4.4).

Kiểm Định Tổ Hợp Tuyến Tính Của Các Hệ Số

Chúng ta rất thường gặp những giả thuyết được phát biểu dưới dạng tổ hợp tuyến tính của các hệ số hồi qui. Một ví dụ minh họa như hàm tiêu thụ tổng hợp sau:

$$C_t = \beta_1 + \beta_2 W_t + \beta_3 P_t + u_t$$

Với *C* là chi tiêu cho tiêu dùng tổng hợp trong một vùng cho trước, W là tổng tiền lương thu nhập, và P là tất cả các thu nhập khác, phần lớn là từ lợi nhuận hoặc thu hồi từ vốn. β_2 là xu hướng cận biên chi tiêu ngoài lương thu nhập, và β_3 là xu hướng cận biên chi tiêu ngoài những thu nhập khác. Giả thuyết $\beta_2 = \beta_3$ ngụ ý là một đô la *thêm vào* của thu nhập tiền lương và một đô la *thêm vào* của thu nhập khác đều đóng góp cùng một khoảng *thêm vào* tiêu thụ bình quân. Kiểm định *t* về các hệ số riêng lẻ không thể áp dụng trong trường hợp này nữa vì giả thuyết là một tổ hợp tuyến tính của hai hệ số hồi qui. Giả thuyết H_0: $\beta_2 = \beta_3$ đối lại H_1: $\beta_2 \neq \beta_3$ có thể được kiểm định bằng ba cách khác nhau, mọi cách đều đưa đến cùng một kết luận.

Trong những phần sau, chúng ta sẽ gặp phải những loại tổ hợp tuyến tính khác như là $\beta_2 + \beta_3 = 1$ hoặc $\beta_2 + \beta_3 = 0$. Bây giờ chúng ta thiết lập thủ tục để kiểm định tổ

hợp tuyến tính như vậy của các hệ số hồi qui. Việc này thực hiện đối với mô hình (không giới hạn) sau, với hai biến độc lập (X_2 và X_3):

$$(U) \quad Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t \qquad \textbf{(4.5)}$$

PHƯƠNG PHÁP 1 (KIỂM ĐỊNH WALD)

Bước 1 Sử dụng ràng buộc, giải để tìm một trong những hệ số theo các hệ số còn lại, và thế vào mô hình không giới hạn để có được mô hình giới hạn. Vì vậy, để kiểm định $\beta_2 = \beta_3$, thay cho β_3 trong Phương trình (4.5) và có được mô hình sau:

$$(R) \quad Y_t = \beta_1 + \beta_2 X_{t2} + \beta_2 X_{t3} + u_t \qquad \textbf{(4.6)}$$
$$= \beta_1 + \beta_2 (X_{t2} + X_{t3}) + u_t$$

Viết lại mô hình giới hạn bằng cách nhóm các số hạng thích hợp. Trong trường hợp của chúng ta, chúng ta sẽ tạo một biến mới $Z_t = X_{t2} + X_{t3}$ và viết mô hình như sau:

$$(R) \quad Y_t = \beta_1 + \beta_2 Z_t + u_t$$

Bước 2 Ước lượng các mô hình giới hạn và không giới hạn, và có được các tổng bình phương sai số, ESS_R và ESS_U.

Bước 3 Tính giá trị thống kê *F* Wald (F_c), dùng Phương trình (4.3), và bậc tự do đối với tử số và mẫu số

Bước 4 Từ bảng F, có được điểm F^* sao cho diện tích phần bên phải bằng mức ý nghĩa. Một cách khác, tính giá trị $p = P(F > F_c)$.

Bước 5 Bác bỏ H_0 nếu $F_c > F^*$ hoặc nếu giá trị *p* nhỏ hơn mức ý nghĩa.

❍ BÀI TẬP THỰC HÀNH 4.6

Xuất phát từ các mô hình giới hạn để kiểm định $\beta_2 + \beta_3 = 1$ và $\beta_2 + \beta_3 = 0$

❍ VÍ DỤ 4.9

Tập tin DATA 4-2 (xem Phụ lục D) chứa dữ liệu hàng năm về Hoa Kỳ trong thời kỳ 1959-1994 (với $n = 36$). Các định nghĩa của các biến như sau:

CONS (C_t) = Chi tiêu thực cho tiêu dùng tính bằng tỷ đô la năm 1992
GDP (Y_t) = Tổng sản phẩm quốc dân thực tính bằng tỷ đô la năm 1992
WAGES = Tổng tiền trả cho nhân viên (lương, và các khoản phụ trợ) tính bằng tỷ đô la hiện hành
PRDEFL = Giá giảm phát đối với tiêu dùng, 1992 = 100 (đây là chỉ số giá của hàng hóa tiêu dùng)

Mô hình chúng ta sẽ ước lượng là hàm tiêu thụ sau đã được trình bày ở phần trên:

$$(U) \quad C_t = \beta_1 + \beta_2 W_t + \beta_3 P_t + u_t \qquad \textbf{(4.5)}$$

Với các biến đã được mô tả trước. Trước khi ước lượng mô hình, chúng ta phải thực hiện một số chuyển đổi dữ liệu để có được tất cả các biến tài chính ở dạng "thực" (nghĩa là đồng đô la không đổi được hiệu chỉnh đối với lạm phát).

Tiêu dùng đã ở dạng thực. Để có thu nhập tiền lương ở dạng thực (W_t), chúng ta chia WAGES với PRDEFL và nhân với 100. Tổng lợi nhuận và các thu nhập khác từ vốn có được bằng cách trừ thu nhập tiền lương thực ra khỏi GDP.

$$W_t = \frac{100\ WAGES_t}{PRDEFL_t} \qquad P_t = Y_t - W_t$$

Trong Phương trình (4.5), đặt ràng buộc $\beta_2 = \beta_3$. Chúng ta có

$$\text{(R)}\quad C_t = \beta_1 + \beta_2 W_t + \beta_2 P_t + u_t = \beta_1 + \beta_2(W_t + P_t) + u_t \qquad \textbf{(4.6)}$$
$$= \beta_1 + \beta_2 Y_t + u_t$$

với $Y_t = W_t + P_t$ là thu nhập tổng hợp. Phương trình (4.5) là mô hình không giới hạn (với *n* bậc tự do) và Phương trình (4.6) là mô hình giới hạn. Do đó chúng ta có thể tính trị thống kê *F* Wald cho trong Phương trình (4.3) (với $k - m = 1$ bởi vì chỉ có duy nhất một ràng buộc). Vì vậy,

$$F_c = \frac{(ESS_R - ESS_U) / 1}{ESS_U / (n - 3)}$$

sẽ được kiểm định với $F^*_{1,\,n-3}(0{,}05)$ và bác bỏ *giả thuyết không* nếu $F_c > F^*$.

Ap dụng vào dữ liệu tiêu dùng tổng hợp, ta có Phương trình ước lượng (4.5) và (4.6). (Xem phần thực hành máy tính 4.2)

$$\hat{C}_t = -222{,}16 + 0{,}69W_t + 0{,}47P_t \qquad ESS_U = 38.977$$
$$\hat{C}_t = -221{,}4 + 0{,}71Y_t \qquad ESS_R = 39.305$$
$$F_c = \frac{(39.305 - 38.977)}{38.977 / 33} = 0{,}278$$

Từ Bảng A.4c, $F^*_{1,33}(0{,}10)$ nằm giữa 2,84 và 2,88. Vì $F_c < F^*$, chúng ta không thể bác bỏ *giả thuyết không* và kết luận là các xu hướng biên tế tiêu dùng ngoài lương và lợi nhuận không khác nhau một cách có ý nghĩa ở mức ý nghĩa 10 phần trăm. Vì vậy, mặc dù giá trị số học của chúng hoàn toàn khác nhau, về mặt thống kê khác biệt này là do ngẫu nhiên.

PHƯƠNG PHÁP 2 (KIỂM ĐỊNH *t* GIÁN TIẾP) Trong phương pháp thứ hai, mô hình được thay đổi theo cách khác và kiểm định *t* gián tiếp được tiến hành. Các bước thực hiện như sau:

Bước 1 Xác định một thông số mới, gọi là δ, có giá trị bằng không khi *giả thuyết không* là đúng. Do đó khi H_0 là $\beta_2 = \beta_3$, chúng ta sẽ định nghĩa $\delta = \beta_2 - \beta_3$, và khi giả thuyết H_0 là $\beta_2 + \beta_3 = 1$ thì $\delta = \beta_2 + \beta_3 - 1$.

Bước 2 Diễn tả một trong những tham số theo δ và các tham số còn lại, thay vào mô hình và nhóm các số hạng một cách hợp lý.

Bước 3 Tiến hành kiểm định *t* sử dụng $\hat{\delta}$, ước lượng của δ.

❹ VÍ DỤ 4.10

Trong trường hợp hàm tiêu thụ, $\delta = \beta_2 - \beta_3$. Giả thuyết không bây giờ trở thành H_0: $\delta = 0$ đối với H_1: $\delta \neq 0$. Cũng có $\beta_3 = \beta_2 - \delta$. Thay vào mô hình ta có

$$C_t = \beta_1 + \beta_2 W_t + (\beta_2 - \delta)P_t + u_t$$
$$= \beta_1 + \beta_2 (W_t + P_t) - \delta P_t + u_t$$

Vì $Y_t = W_t + P_t$, mô hình này trở thành

$$C_t = \beta_1 + \beta_2 Y_t - \delta P_t + u_t \qquad \textbf{(4.7)}$$

Mô hình này về mặt khái niệm hoàn toàn tương đương với Phương trình (4.5). Bây giờ hồi qui *C* theo một số hạng không đổi, *Y*, và *P*, và sử dụng trị thống kê *t* cho δ để kiểm định giả thuyết mong muốn. Trong trường hợp này, kiểm định giảm đến kiểm định *t* chuẩn nhưng theo mô hình hiệu chỉnh. (Xem như bài tập thực hành, hãy áp dụng kỹ thuật này đối với $\beta_2 + \beta_3 = 1$)

Đối với dữ liệu của chúng ta, Phương trình ước lượng (4.7) là (xem phần thực hành máy tính 4.2)

$$\hat{C}_t = -222{,}16 + 0{,}69Y_t + 0{,}04P_t$$
$$(-11{,}4) \quad (21{,}3) \quad (0{,}5)$$

Các giá trị trong ngoặc đơn là trị thống kê *t* tương ứng. Đối với $\hat{\delta}$, giá trị *t* là 0,5, nhỏ hơn $t^*_{33}(0{,}05)$ ở giữa 2,021 và 2,042. Do đó, ở đây cũng không bác bỏ *giả thuyết không*.

PHƯƠNG PHÁP 3 (KIỂM ĐỊNH *t* TRỰC TIẾP) Phương pháp cuối cùng áp dụng một kiểm định *t* trực tiếp và không đòi hỏi ước lượng của một hệ số hồi qui nào khác.

Bước 1 Như trong phương pháp 2, xác định một thông số mới – gọi là δ – có giá trị bằng không khi *giả thuyết không* là đúng. Do đó khi H_0 là $\beta_2 = \beta_3$, chúng ta sẽ định nghĩa $\delta = \beta_2 - \beta_3$, và khi giả thuyết H_0 là $\beta_2 + \beta_3 = 1$ thì $\delta = \beta_2 + \beta_3 - 1$.
Bước 2 Trực tiếp lấy phân phối thống kê của δ, và sử dụng để tính trị thống kê *t*.
Bước 3 Tiến hành kiểm định *t* trên δ sử dụng trực tiếp để tính trị thống kê.

Kiểm định trước được minh họa ở đây chỉ cho ví dụ chúng ta sử dụng, H_0: $\beta_2 = \beta_3$. (Xem như bài tập thực hành, hãy áp dụng phương pháp này đối với giả thuyết $\beta_2 + \beta_3 = 1$)

Vì các ước lượng OLS là tổ hợp tuyến tính của các quan sát trên biến phụ thuộc và do đó là tổ hợp tuyến tính của các số hạng sai số phân phối chuẩn, chúng ta biết là

$$\hat{\beta}_2 \sim N(\beta_2, \sigma^2_{\hat{\beta}_2}) \qquad \hat{\beta}_3 \sim N(\beta_3, \sigma^2(_{\hat{\beta}_3})$$

với σ^2 là phương sai tương ứng. Hơn nữa, một tổ hợp tuyến tính của các biến chuẩn cũng phân phối chuẩn. Do đó,

$$\hat{\beta}_2 - \hat{\beta}_3 \sim [\beta_2 - \beta_3, \text{Var}(\hat{\beta}_2 - \hat{\beta}_3)]$$

Từ Tính chất 2.8a, phương sai của $\hat{\beta}_2 - \hat{\beta}_3$ tính bằng $\text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\beta}_3) - 2\,\text{Cov}(\hat{\beta}_2,\hat{\beta}_3)$. Chuyển những số trên về phân phối chuẩn chuẩn hóa (bằng cách trừ đi giá trị trung bình và chia cho độ lệch chuẩn), chúng ta có

$$\frac{\hat{\beta}_2 - \hat{\beta}_3 - (\beta_2 - \beta_3)}{[\text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\beta}_3) - 2\,\text{Cov}(\hat{\beta}_2,\hat{\beta}_3)]^{1/2}} \sim N(0,1)$$

Với *giả thuyết không*, H_0: $\beta_2 - \beta_3 = 0$. Cũng vậy, chúng ta không biết chính xác các phương sai và đồng phương sai, nhưng có thể ước lượng được chúng (hầu hết các chương trình máy tính đều có lựa chọn cung cấp các giá trị này). Nếu chúng ta thay các ước lượng của các phương sai và đồng phương sai này, trị thống kê trên không còn tuân theo phân phối $N(0,1)$ mà theo phân phối thống kê t_{n-k} ($n - 3$ trong ví dụ của chúng ta). Vì vậy, có thể sử dụng cùng kiểm định t cho trị thống kê tính từ đẳng thức trên với các ước lượng phù hợp được thay vào. Trị thống kê t được tính bằng

$$t_c = \frac{\hat{\beta}_2 - \hat{\beta}_3}{[\text{Var}(\hat{\beta}_2) + \text{Var}(\hat{\beta}_3) - 2\,\text{Cov}(\hat{\beta}_2,\hat{\beta}_3)]^{1/2}}$$

Vì $\beta_2 = \beta_3$ theo *giả thuyết không*. Với mức ý nghĩ 5%, H_0 bị bác bỏ và giả thuyết H_1: $\beta_2 - \beta_3 > 0$ được củng cố nếu giá trị t_c lớn hơn $t^*_{n-k}(0,05)$. Đối với trường hợp giả thiết ngược lại có dạng hai phía, H_1: $\beta_2 \neq \beta_3$, ta tra giá trị $t^*_{n-k}(0,025)$ và bác bỏ H_0 nếu $|t_c| > t^*$. Vì phương pháp này đòi hỏi phải thực hiện một số tính toán phụ, nên một trong các phương khác thường được đề nghị sử dụng hơn phương pháp 3.

⭘ VÍ DỤ 4.11:

Để minh họa, chúng ta xem phương trình (4.5), phương trình này được ước lượng từ tập dữ liệu DATA4-2 ở phụ lục D. Phương trình ước lượng cùng với các trị phương sai và đồng phương sai được trình bày dưới đây (xem Phần Thực Hành Máy Tính 4.2):

$$\hat{C}_t = -222{,}16 + 0{,}693W_t + 0{,}736P_t$$

$$\bar{R}^2 = 0{,}999 \qquad \text{d.f.} = 33 \qquad \text{ESS} = 38.977$$

$$\widehat{\text{Var}\hat{\beta}_2} = (0{,}032606)^2 \qquad \widehat{\text{Var}\hat{\beta}_3} = (0{,}048822)^2$$

$$\widehat{\text{Cov}(\hat{\beta}_2, \hat{\beta}_3)} = -0{,}001552$$

Trị thống kê t được tính theo:

$$t_c = \frac{0{,}693 - 0{,}736}{[(0{,}032606)^2 + (0{,}048822)^2 - 2(-0{,}001552)]^{1/2}} = -0{,}53$$

Vì $t^*_{33}(0,05)$ có giá trị nằm giữa 2,021 và 2,042, và giá trị này lớn hơn nhiều so với giá trị tính toán nhiều nên chúng ta không bác bỏ giả thuyết H_0 cho rằng khuynh hướng cận biên chi tiêu từ tiền lương và thu nhập khác là như nhau. Kết quả này giữ nguyên cho dù giả thuyết ngược lại H_1 là một phía hay hai phía.

Chúng ta thấy rằng cả ba phương pháp đều cho ra cùng một kết quả. Trong ba phương pháp được trình bày, Phương pháp 2 thực hiện dễ nhất vì nó không đòi hỏi các tính toán phụ nhưng lại có thể sử dụng để kiểm định giả thuyết bằng phép kiểm định t trực tiếp theo một mô hình được điều chỉnh một tí. Tuy nhiên, kiểm định Wald được trình bày trong phương pháp 1 có thể được áp dụng trong nhiều trường hợp tổng quát hơn.

4.5. Các Sai Số Đặc Trưng

Như đã đề cập trước đây, việc lựa chọn các biến độc lập và phụ thuộc trong mô hình kinh tế lượng phải được dựa trên lý thuyết kinh tế, kiến thức về các hành vi tiềm ẩn, và kinh nghiệm quá khứ. Tuy nhiên, các bản chất các quan hệ giữa các biến kinh tế là không bao giờ biết, và vì vậy chúng ta có thể mong đợi những sai số trong việc xác định các đặc trưng của mô hình kinh tế lượng. Sai số đặc trưng xảy ra nếu chúng ta xác định sai mô hình theo các loại như chọn biến, dạng hàm số, hoặc cấu trúc sai số (nghĩa là số hạng ngẫu nhiên u_t và các tính chất của nó). Trong phần này chúng ta sẽ khảo sát sai số đặc trưng loại thứ nhất. Trong chương 6 chúng ta sẽ xem xét đến việc lựa chọn các dạng hàm số và các sai số đặc trưng của số hạng ngẫu nhiên sẽ được thảo luận ở chương 8 và 9.

Khi chọn các biến độc lập của mô hình, ta có thể phạm phải hai loại sai số sau: (1) bỏ qua một biến thuộc về mô hình và (2) đưa vào một biến không liên quan. Trong hàm cầu, nếu chúng ta bỏ qua biến giá cả hàng hóa hoặc thu nhập của hộ gia đình, chúng ta có thể gây ra trường hợp sai số đặc trưng loại thứ nhất. Trong ví dụ về bất động sản trước đây, giả sử các biến về loại mái lợp hoặc thiết bị điện sử dụng hoặc khoảng cách đến các trường học lân cận không tác động đáng kể đến giá bán ngôi nhà. Nếu chúng ta vẫn tiếp tục đưa những biến này vào mô hình, chúng ta sẽ phạm phải sai số đặc trưng loại thứ hai, nghĩa là, đưa thừa biến vào mô hình. Trong những phần sau, chúng ta sẽ xem xét các hệ quả lý thuyết của từng loại sai số đặc trưng này đồng thời trình bày các bằng chứng thực nghiệm.

Bỏ qua biến quan trọng.

Đầu tiên chúng ta khảo sát trường hợp trong đó một biến thuộc về mô hình bị bỏ qua. Giả sử mô hình *thật* là:

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t$$

Nhưng chúng ta ước lượng được mô hình

$$Y_t = \beta_1 + \beta_2 X_{t2} + v_t$$

Nói cách khác, giá trị thật của β_3 là khác 0, nhưng chúng ta lại giả định rằng nó bằng 0 và vì vậy đã loại bỏ biến X_3 ra khỏi mô hình. Các số hạng sai số của mô hình thật được giả định là đáp ứng được các giả thiết từ 3.2 đến 3.8. Các hệ quả của loại sai số xác định này được tóm tắt qua các tính chất sau:

Tính chất 4.3

a. Nếu một biến độc lập mà hệ số hồi qui thật của nó khác không bị loại ra khỏi mô hình, các giá trị ước lượng của tất cả các hệ số hồi qui còn lại sẽ bị thiên lệch trừ phi biến bị loại ra không tương quan với mọi biến được đưa vào.
b. Ngay cả khi điều kiện này được thỏa mãn, số hạng hằng số được ước lượng nói chung cũng bị thiên lệch, và vì vậy các giá trị dự báo cũng bị thiên lệch.
c. Ước lượng phương sai của hệ số hồi qui của một biến được đưa vào nói chung sẽ bị thiên lệch, và vì vậy các kiểm định giả thuyết sẽ không có ý nghĩa.

Có thể thấy từ Tính chất 4.3 rằng hệ quả của việc bỏ qua một biến quan trọng là rất nghiêm trọng. Các ước lượng và trị dự báo sẽ bị thiên lệch, và các kiểm định giả thuyết sẽ không còn có ý nghĩa nữa. Nguyên nhân của sự thiên lệch (được gọi là **thiên lệch biến bị bỏ sót**) là dễ dàng nhận thấy. So sánh hai mô hình, chúng ta thấy rằng $v_t = \beta_3 X_{t3} + u_t$. Giá trị kỳ vọng của số hạng sai số trong mô hình sai là $E(v_t) = \beta_3 X_{t3} \neq 0$. Vì vậy, v_t vi phạm Giả sử 3.3. Nghiêm trọng hơn, đồng phương sai giữa X_{t2} và v_t được tính theo (xem Phần 2.3 về đồng phương sai):

$$\begin{aligned} \text{Cov}(X_{t2}, v_t) &= \text{Cov}(X_{t2}, \beta_3 X_{t3} + u_t) = \beta_3 \text{Cov}(X_{t2}, X_{t3}) + \text{Cov}(X_{t2}, u_t) \\ &= \beta_3 \text{Cov}(X_{t2}, X_{t3}) \end{aligned}$$

Vì X_2 và u không tương quan. Như vậy, trừ phi đồng phương sai giữa X_2 và X_3 bằng 0 – nghĩa là, trừ phi X_2 và X_3 là không tương quan – đồng phương sai giữa X_2 và v sẽ khác không, và như vậy cũng vi phạm Giả thiết 3.4. Tính chất không thiên lệch và nhất quán phụ thuộc vào hai giả thiết này. Như vậy, $\hat{\beta}_2$ sẽ không bị không thiên lệch và nhất quán.

Khẳng định trên có thể được nhận ra một cách rõ ràng hơn. Gọi $\hat{\beta}_1$ và $\hat{\beta}_2$ là các ước lượng của số hạng hằng số và hệ số độ dốc của X_{t2} khi chúng ta hồi qui Y_t theo số hạng hằng số và một biến X_{t2}, nghĩa là loại bỏ ra X_{t3}. Các giá trị ước lượng thực của hai ước lượng này được chứng minh ở Phụ Lục Phần 4.2 như sau:

$$E(\hat{\beta}_2) = \beta_2 + \beta_3 \left[\frac{S_{23}}{S_{22}} \right] \text{ và } E(\hat{\beta}_1) = \beta_1 + \beta_3 \left[\overline{X}_3 - \overline{X}_3 \frac{S_{23}}{S_{22}} \right]$$

Trong đó các biến có gạch ngang trên đầu là các giá trị trung bình tương ứng, $S_{23} = \sum (X_{t2} - \overline{X}_2)(X_{t3} - \overline{X}_3)$ và $S_{22} = \sum (X_{t2} - \overline{X}_2)^2$. Từ đây chúng ta có thể thấy rằng, trừ phi $S_{23} = 0$, tức là, trừ phi X_2 và X_3 không tương quan, thì $E(\hat{\beta}_2) \neq \beta_2$ và vì vậy nói chung $\hat{\beta}_2$ là thiên lệch. Cũng lưu ý rằng $\hat{\beta}_2$ bao gồm một số hạng liên quan đến β_3, đó là ảnh hưởng của biến bị loại bỏ. Vì vậy, chúng ta không thể diễn dịch $\hat{\beta}_2$ là ảnh hưởng cận biên của riêng X_2. Một phần ảnh hưởng của biến bị loại bỏ ra khỏi mô hình

cũng được kể đến. Như vậy, hệ số trong mô hình sẽ đo lường ảnh hưởng trực tiếp của biến được đưa vào mô hình cũng như ảnh hưởng gián tiếp của biến bị loại bỏ. Điều này cũng đúng với các ước lượng của số hạng hằng số. Lưu ý rằng ngay cả khi $S_{23} = 0$, $\hat{\beta}_1$ sẽ thiên lệch trừ phi có thêm giá trị trung bình của $X_3 = 0$. Bởi vì các điều kiện đưa ra ở đây là rất khó thỏa mãn, nên nhìn chung các ước lượng và các giá trị dự báo là thiên lệch.

SỰ NGUY HIỂM CỦA VIỆC LOẠI BỎ SỐ HẠNG HẰNG SỐ. Như đã thấy ở trên $\hat{\beta}_1$ và $\hat{\beta}_2$ có kể đến một phần ảnh hưởng của biến bị loại bỏ X_3. Do đó cần thiết phải đưa số hạng hằng số vào mô hình. Nếu số hạng hằng số bị bỏ qua, đường hồi qui bị ép phải đi qua gốc tọa độ, điều này có thể dẫn đến việc đặc trưng sai nghiêm trọng hàm hồi qui. Chúng ta có thể thấy từ biểu đồ phân tán ở Hình 3.1 hay Hình 3.11 rằng sự ràng buộc đường hồi qui đi qua gốc tọa độ sẽ làm cho các ước lượng của độ dốc bị thiên lệch và các sai số sẽ lớn hơn. Một lần nữa, kết luận từ phần thảo luận này là số hạng hằng số luôn luôn nên được đưa vào mô hình trừ phi có một lý do lý thuyết vững chắc để không làm điều đó (trong Chương 6 chúng ta sẽ gặp một trường hợp trong đó lý thuyết bắt buộc không có số hạng hằng số)

❹ BÀI TẬP THỰC HÀNH 4.7

Trong mô hình tuyến tính đơn, giả sử rằng bạn đã nhầm lẫn loại bỏ số hạng hằng số; nghĩa là, giả sử rằng mô hình thật là $Y_t = \alpha + \beta X_t + u_t$, nhưng bạn ước lượng ra thành $Y_t = \beta X_t + v_t$. Đầu tiên kiểm chứng ước lượng OLS của β khi sử dụng mô hình sai đó là $\hat{\beta} = [\Sigma(X_t Y_t)]/[\Sigma(X_t^2)]$. Kế đến thay vào Y_t trong biểu thức này bằng Y_t từ mô hình thật, và tính $E(\hat{\beta})$. Và sau đó chứng minh rằng $\hat{\beta}$ là thiên lệch. Cuối cùng tìm điều kiện để $\hat{\beta}$ là không thiên lệch mặc dù sử dụng mô hình sai. Nêu các diễn dịch trực giác về các điều kiện bạn tìm ra.

[Trong bài tập dạng này và các bài tương tự ở cuối chương, tiến hành như sau: (1) sử dụng mô hình ước lượng và tìm ra biểu thức đại số cho các trị ước lượng thông số; (2) thay vào Y_t từ *mô hình thật* theo số hạng X_t, u_t, và các thông số của mô hình thật (chúng ta sử dụng mô hình thật vì Y_t được xác định thông qua nó mà không phải bằng mô hình sai); (3) tính giá trị kỳ vọng của các ước lượng; và (4) so sánh các giá trị kỳ vọng với giá trị thật, kiểm tra tính không thiên lệch, và nếu cần thiết, xác định điều kiện để có sự không thiên lệch]

❹ Ví dụ 4.12:

Đến đây cần phải có một minh họa thực tiễn về các thiên lệch xác định do việc loại bỏ các biến quan trọng. Tập tin DATA4-3 mô tả ở Phụ lục D chứa các dữ liệu hàng năm về việc xây mới nhà ở Mỹ. Quan hệ ước lượng giữa việc mua nhà (HOUSING) (đơn vị nghìn đô la), GNP (theo tỉ đô la 1982), và lãi suất cầm cố (%) là như sau (xem chi tiết ở Phần Thực Tập Máy Tính 4.3)

Mô hình A: $\widehat{HOUSING} = 687,898 + 0,905GNP - 169,658\ INTRATE$
(1,80) (3,64) (-3,87)
$\bar{R}^2 = 0,375$ F(2, 20) = 7,609 d.f. = 20

Từ lý thuyết cơ bản về nhu cầu chúng ta kỳ vọng rằng nhu cầu về nhà ở sẽ tăng khi thu nhập tăng. Trái lại, khi lãi suất cầm cố tăng, chi phí sở hữu nhà sẽ tăng, và nhu cầu về nhà ở sẽ giảm. Nhận thấy rằng các dấu của các hệ số ước lượng phù hợp với cảm nhận trực giác của chúng ta. Chúng ta cũng thấy từ các trị thống kê t trong ngoặc đơn rằng GNP và INTRATE và rất có ý nghĩa. Tuy nhiên, $\bar{R}^2$ có giá trị không cao lắm đối với tập dữ liệu theo thời gian. Giả sử chúng ta bỏ qua biến quan trọng INTRATE. Mô hình ước lượng sẽ trở thành như sau:

Mô hình B: $\widehat{HOUSING} = 1.442,209 + 0,058GNP$
(3,39) (0,38)
$\bar{R}^2 = - 0,04$ F(1, 21) = 0,144 d.f. = 21

Các kết quả thay đổi rất lớn. Đầu tiên, $\bar{R}^2$ bây giờ có giá trị âm, cho thấy rằng một sự thích hợp kém. Điều này được củng cố thêm bằng trị thống kê F với giá trị nhỏ và không có ý nghĩa. Trị thống kê t của GNP không có ý nghĩa, cho thấy GNP có tác động không đáng kể đến việc mua nhà. Cuối cùng giá trị ước lượng của hệ số GNP bị thay đổi đáng kể. Các kết quả này là hoàn toàn không chấp nhận được và là hậu quả của việc bỏ qua lãi suất thế chấp, là một biến quan trọng tiên quyết trong việc xác định nhu cầu nhà ở.

Đưa Vào Mô Hình Một Biến Không Liên Quan

Giả sử rằng mô hình *thật* là

$$Y_t = \beta_1 + \beta_2 X_{t2} + u_t$$

Nhưng chúng ta thêm nhầm biến X_3 và ước lượng được mô hình

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + v_t$$

Như trước đây số dư *thật* u_t được giả định tuân theo giả thiết 3.2 đến 3.8 ở Chương 3. Hậu quả của loại đặc trưng sai này là gì? Ước lượng của β_2 có thiên lệch hay không? Liệu nó vẫn là BLUE? Các kiểm định giả thuyết có hợp lệ không? Câu trả lời cho các câu hỏi này được tóm tắt ở các tính chất sau:

Tính chất 4.4

a. Nếu một biến độc lập có giá trị hệ số hồi qui thật bằng không (nghĩa là, biến này là thừa) được đưa vào mô hình, các giá trị ước lượng của tất cả các hệ số hồi qui khác vẫn sẽ không thiên lệch và nhất quán.
b. Tuy nhiên phương sai của chúng sẽ cao hơn các giá trị khi không có biến không liên quan, và vì vậy các hệ số sẽ không hiệu quả.
c. Vì các phương sai ước lượng của các hệ số hồi qui là không thiên lệch, các kiểm định giả thuyết vẫn có hiệu lực.

Như vậy hậu quả của việc đưa vào mô hình một biến không liên quan là ít nghiêm trọng hơn so với trường hợp bỏ sót một biến quan trọng.

CHỨNG MINH*

Ở phần 4.A.3 ta đã chứng minh được rằng

$$E(\hat{\beta}_2) = \beta_2 \text{ và } E(\hat{\beta}_3) = 0$$

Như vậy, $\hat{\beta}_2$ là không thiên lệch và kỳ vọng của $\hat{\beta}_3$ bằng 0. Tính nhất quán được giữ nguyên. Các kết quả này được tổng quát hóa cho trường hợp hồi qui bội với nhiều biến giải thích. Do vậy, việc đưa vào các biến không liên quan không làm thiên lệch các ước lượng của các hệ số của các biến còn lại. Vì các ước lượng là không thiên lệch và nhất quán, các giá trị dự báo dựa trên chúng cũng vậy.

Bước kế tiếp là tính phương sai của $\hat{\beta}_2$ để xác định tính chất hiệu quả. Từ phần 4.A.3 (sử dụng các ký hiệu ở đó) ta có:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{S_{22}(1-r^2)}$$

Trong đó r^2 là bình phương của phương sai đơn (xem Phương trình 2.11) giữa X_2 và X_3 được định nghĩa như là $r^2 = S^2_{23}/(S_{22}S_{33})$. Chúng ta so sánh kết quả này với phương sai của ước lượng theo OLS (gọi giá trị này là β^*_2) mà lẽ ra sẽ thu được nếu mô hình thật được sử dụng. Từ các phương trình (3.12) và (3.19) ở chương 3 ta có:

$$\beta^*_2 = \frac{S_{y2}}{S_{22}} \text{ và } \text{Var}(\beta^*_2) = \frac{\sigma^2}{S_{22}}$$

Độ hiệu quả tương đối (xem định nghĩa 2.8b) của $\hat{\beta}_2$ đối với β^*_2 là

$$\frac{\text{Var}(\hat{\beta}_2)}{\text{Var}(\beta^*_2)} = \frac{1}{1-r^2} \geq 1$$

Vì vậy rõ ràng ước lượng của β_2 khi sử dụng mô hình sai là không hiệu quả trừ phi $r^2 = 0$ – nghĩa là, trừ phi X_2 và X_3 không tương quan với nhau. Vì tính không hiệu quả này, trị thống kê t có khuynh hướng nhỏ hơn, và do đó chúng ta có thể kết luận sai rằng những biến này là không có ý nghĩa về mặt thống kê nhưng thực sự chúng lại hoàn toàn khác không. Có thể chứng minh rằng (xem Johnston, 1984, trang 262) ước lượng của phương sai của $\hat{\beta}_2$ là không thiên lệch và do đó các kiểm định giả thuyết vẫn có hiệu lực.

❍ VÍ DỤ 4.13:

Tập tin DATA4-3 cũng có chứa dữ liệu về dân số (POP) và tỉ lệ thất nghiệp (UNEMP). Dân số được đo theo đơn vị hàng triệu và thất nghiệp được đo bằng tỉ lệ phần trăm. Chúng ta có thể kỳ vọng rằng dân số càng cao thì số lượng nhà xây mới càng cao. Tỉ lệ thất nghiệp là số đo hợp lý đối với chu kỳ kinh doanh. Khi tỉ lệ thất nghiệp cao, người tiêu dùng có khuynh hướng hoãn việc mua nhà. Chính vì vậy việc đưa biến POP và

UNEMP vào làm biến giải thích là hợp lý. Mô hình hiệu chỉnh như sau: (trị thống kê t ở trong ngoặc đơn)

Mô hình C: $\widehat{HOUSING} = 5.087,434 + 1,756GNP - 174,692\ INTRATE$
$(0,5) \quad (0,8) \quad (-2,9)$
$- 33,434\ POP + 79,720\ UNEMP$
$(-0,4) \quad (0,7)$
$\bar{R}^2 = 0,328 \quad F(4, 18) = 3,681 \quad d.f. = 18$

Khi so sánh với mô hình A chúng ta thấy có nhiều sự khác biệt đáng kể. GNP trước đó là có ý nghĩa thì bây giờ không còn ý nghĩa nữa. Trị thống kê t của biến INTRATE cũng giảm mặc dù nó vẫn còn có ý nghĩa. Điều này đúng với những điều phân tích lý thuyết đã được dự đoán. Tính chất 4.4b nói rằng phương sai của các hệ số có khả năng lớn hơn, điều này hàm ý rằng các trị thống kê t sẽ có thể nhỏ hơn. Các trị thống kê t của các biến POP và INTRATE là rất nhỏ, cho thấy các biến này có thể là không quan trọng trong vai trò các biến *thêm vào* chi phối nhu cầu về nhà ở, cho trước rằng GNP và INTRATE đo lường quy mô của nền kinh tế và chu kỳ kinh doanh. Thực ra, chúng ta có thể thực hiện kiểm định Wald đối với việc loại bỏ POP và UNEMP. Xem Mô hình C là mô hình không giới hạn và Mô hình A là mô hình giới hạn, trị thống kê F trong kiểm định Wald (xem Phương Trình 4.3) được tính theo:

$$F_c = \frac{(ESS_A - ESS_C) \div (d.f._A - d.f._C)}{ESS_C \div d.f._C}$$
$$= \frac{(1.491.140 - 1.444.274)/2}{1.444.274/18} = 0,292$$

Giá trị quan sát F_c là rất nhỏ và không có ý nghĩa ngay cả ở mức 25% (*p-value* là 0.75). Vì vậy, kiểm định Wald sẽ không bác bỏ *giả thuyết không* cho rằng các hệ số hồi qui của POP và UNEMP bằng không. Chúng ta cũng lưu ý rằng dấu của POP và UNEMP ngược với những gì chúng ta đã kỳ vọng. Tuy nhiên, trong trường hợp các hệ số không có nghĩa, thì dấu của chúng không liên quan và có thể được chọn tùy ý.

❹ Bài tập thực hành 4.8

Thay vì đưa cả hai biến POP và UNEMP vào, như đã làm trên đây, chỉ đưa tỉ lệ thất nghiệp vào mô hình A thôi (được gọi là mô hình D). Hãy so sánh các kết quả nhận được với các kết quả của mô hình A. Các kết quả có khác biệt nhiều không?

Chúng ta quan sát sự đánh đổi bằng cách so sánh giữa việc thêm vào một biến không liên quan với việc loại bỏ một biến quan trọng. Sai số đặc trưng loại thứ nhất tạo ra các ước lượng không hiệu quả, mặc dù không bị thiên lệch. Loại sai thứ hai gây ra sự thiên lệch trong các ước lượng và các kiểm định giả thiết. Bởi vì không thể biết được các mối quan hệ thật sự, nên chúng ta phải đối mặt với vấn đề nan giải trong việc lựa chọn dạng thức thích hợp nhất. Một nhà khảo cứu đặt nặng sự không thiên lệch, nhất quán, và tin cậy của các kiểm định sẽ nắm giữ các biến không liên quan hơn là đối mặt với những hậu quả của việc bỏ mất các biến quan trọng. Mặt khác, nếu một nhà nghiên cứu không

thể chấp nhận những ước lượng kém hiệu quả, thì việc xóa một hay nhiều biến gây khó chịu sẽ dễ được chọn lựa. Lý thuyết về kinh tế học và sự hiểu biết về các hành vi đang diễn ra có thể giúp thoát khỏi tình trạng tiến thoái lưỡng nan này. Các ràng buộc chọn lựa mô hình thảo luận ở phần trước cũng có thể trợ giúp được. Trong chương 6 chúng ta sẽ thấy rằng các kiểm định đối với các đặc trưng cũng sẽ trợ giúp cho chúng ta. Tất cả điều này cần rất nhiều sự cân nhắc. Sự gắn bó mù quáng đối với các tiêu chuẩn cứng nhắc phải được ngăn ngừa bằng mọi giá.

4.6 Ứng dụng: Các Yếu Tố Quyết Định Số Người Đi Xe Buýt

Ứng dụng đầu tiên liên quan đến số người sẽ di chuyển bằng xe buýt với nhiều yếu tố ảnh hưởng khác nhau. DATA 4-4 được mô tả trong phụ lục D có dữ liệu chéo cho 40 thành phố khắp nước Mỹ. Các biến như sau:

BUSTRAVL = Mức độ giao thông bằng xe buýt ở đô thị tính theo ngàn hành khách mỗi giờ
FARE = Giá vé xe buýt tính bằng Mỹ kim
GASPRICE = Giá một ga lông nhiên liệu tính bằng Mỹ kim
INCOME = Thu nhập bình quân đầu người tính bằng Mỹ kim
POP = Dân số thành phố tính bằng ngàn người
DENSITY = Mật độ dân số tính (người/dặm vuông)
LANDAREA = Diện tích thành phố (dặm vuông)

Đặc trưng tổng quát của mô hình, thường được xem như mô hình "bồn rửa chén", được cho dưới đây (không có chỉ số t):

$$\text{BUSTRAV} = \beta_1 + \beta_2\text{FARE} + \beta_3\text{GASPRICE} + \beta_4\text{INCOME} + \beta_5\text{POP} + \beta_6\text{DENSITY} + \beta_7\text{LANDAREA} + u$$

Trước khi ước lượng mô hình, chúng ta sẽ xác định dấu của các biến, mức độ ưu tiên, cho các hệ số hồi qui. Trong phần thảo luận này, những tiềm ẩn về phía cung không được xem là quan trọng. Bởi vì một sự gia tăng giá vé xe buýt có thể làm giảm nhu cầu đi xe buýt, nên chúng ta kỳ vọng β_2 sẽ âm. Trong lĩnh vực di chuyển, xe hơi sẽ là một thay thế đối với xe buýt, và vì vậy một sự gia tăng giá nhiên liệu có thể khiến một số người tiêu thụ chuyển sang đi xe buýt. Vì vậy chúng ta kỳ vọng một hiệu ứng tích cực ở đây; nghĩa là, β_3 sẽ dương. Khi thu nhập tăng, chúng ta kỳ vọng nhu cầu đối với hàng tiêu dùng cũng tăng lên, và vì vậy như thường lệ chúng ta kỳ vọng β_4 sẽ dương. Tuy nhiên, nếu hàng tiêu dùng thuộc loại hàng hóa "thấp cấp", thì hiệu ứng thu nhập (nghĩa là, β_4) sẽ âm. Một sự gia tăng kích thước dân số hay mật độ dân số thường làm gia tăng nhu cầu di chuyển bằng xe buýt. Vì vậy, chúng ta kỳ vọng β_5 và β_6 sẽ dương. Nếu diện tích đất tăng cao, thì thành phố sẽ trải rộng ra hơn và người tiêu thụ có thể thích dùng xe hơi như là phương tiện giao thông chính hơn. Nếu đây là một tình huống, β_7 được kỳ vọng sẽ âm.

Bảng 4.4 trích một phần kết quả chạy máy tính sử dụng chương trình GRETL (Xem phần Thực hành máy tính 4.4). Các nhận xét cần phải chi tiết và nên được nghiên cứu cẩn thận trước khi phát triển xa hơn. Tất cả các chủ đề mà chúng ta đã nghiên cứu

được gắn kết lại với nhau trong dự án thực nghiệm nhỏ này, và Bảng 4.4 sẽ giúp bạn lắp ghép những mảnh ráp hình khác nhau thành một hình ảnh hoàn chỉnh. Ngay cả nếu bạn sử dụng chương trình của riêng mình để kiểm tra lại các kết quả, thì cũng đáng để nghiên cứu các lưu ý trong Bảng 4.4.

Bảng 4.4 Trích một phần kết quả chạy máy tính đối với Số người đi xe buýt

```
MODEL 1: OLS estimates using the 40 observations 1-40
Dependent variable: BUSTRAVL

      VARIABLE      COEFFICIENT       STDERROR      T STAT    2PROB(t > |T|)

 0)   Const           2744.6797      2641.6715       1.039    0.306361
 2)   FARE            -238.6544       451.7281      -0.528    0.600816
 3)   GASPRICE         522.1132      2658.2276       0.196    0.845491
 4)   INCOME            -0.1947         0.0649      -3.001    0.005090 ***
 5)   POP                1.7114         0.2314       7.397    0.000000 ***
 6)   DENSITY            0.1164         0.0596       1.954    0.059189 *
 7)   LANDAREA          -1.1552         1.8026      -0.641    0.526043

Mean of dep. var            1933.175   S.D. of dep. variable         2431.757
Error Sum of Sp (ESS)    1.8213e+007   Std Err of Resid. (sgmahat)   742.9113
Unadjusted R-squared           0.921   Adjusted R-squared               0.907
F-statistic (6,33)           64.1434   p-value for F()               0.000000
Durbin-Watson stat             2.083   First-order autocorr. coeff     -0.156

MODEL SELECTION STATISTICS

SGMASQ          551917     AIC             646146     FPE             648503
HQ              719020     SCHWARZ         868337     SHIBATA         614698
GCV             668991     RICE            700510
Excluding the constant, p-value was highest for variable 3 (GASPRICE)
```

[R bình phương hiệu chỉnh là 0,907, hiển thị rằng 90,7% phương sai của BUSTRAVL được giải thích chung bởi các biến trong mô hình. Đối với một nghiên cứu chéo, điều này hoàn toàn tốt. Cột cuối cùng cho giá trị *p-value* đối với kiểm định 2 đuôi cho *giả thuyết không* tương ứng với các hệ số hồi qui bằng không. Ba dấu sao (***) hiển thị rằng giá trị *p-value* nhỏ hơn 1%, ** có nghĩa nó nằm giữa 1 và 5%, * ám chỉ *p-value* trong khoảng 5 đến 10%, và không có * nghĩa là *p-value* trên 10%. Hãy nhớ rằng giá trị *p-value* cao nghĩa là xác suất sai lầm loại I bác bỏ *giả thuyết không* sẽ cao. Nếu điều này cao hơn mức ý nghĩa đã được chọn (0,10, chẳng hạn), thì chúng ta sẽ không bác bỏ *giả thuyết không* cho rằng hệ số bằng 0. Nói cách khác, khi giữ các biến khác cố định, biến này sẽ không có ảnh hưởng có ý nghĩa lên BUSTRAVL. Dựa theo điều này, chỉ INCOME, POP, và DENSITY có các hệ số có nghĩa ở mức 10%. Hằng số và các hệ số của FARE, GASPRICE, và LANDAREA không có ý nghĩa về mặt thống kê ngay cả ở mức 25%.

Sự phù hợp của các trị thống kê chọn lựa mô hình (đã thảo luận trong Phần 4.3) sau này sẽ trở nên hiển nhiên. Trị thống kê Durbin-Watson và tự tương quan bậc nhất sẽ được thảo luận trong chương 9, nhưng không liên quan lắm cho mục đích của chúng ta.

Chúng ta nên làm điều gì kế tiếp? Chúng ta đã một vài lần nhắc đến hằng số không có nghĩa thực tế và không tham gia giải thích ý nghĩa của biến phụ thuộc cũng như

các hiệu ứng trung bình của các biến bị loại bỏ. Do đó, qui tắc chung là bỏ qua ý nghĩa của hằng số hoặc là không cần nó. Tuy nhiên, FARE, GASPRICE, và LANDAREA là những "ứng cử viên đầu tiên" cho sự loại bỏ khỏi mô hình bởi vì không có bằng chứng chứng tỏ chúng có những ảnh hưởng có nghĩa lên BUSTRAVL. Chúng ta có thể thực hiện một bước lớn, bỏ tất cả chúng, ước lượng một mô hình được giới hạn, và thực hiện kiểm định Wald F-test như đã được mô tả ở Phần 4.4. Để tạo thuận lợi cho việc này, chúng ta lấy ra tổng bình phương sai số và số bậc tự do cho mô hình không giới hạn vừa mới được ước lượng. Tuy nhiên, bạn hãy cẩn trọng, một số biến đồng thời bị loại bỏ không phải là việc làm khôn ngoan. Bởi vì bạn sẽ nhìn thấy điều này trong ví dụ kế tiếp và những ví dụ sau, việc cùng lúc loại bỏ một vài biến cũng có thể bỏ mất những biến có ý nghĩa hoặc là những biến quan trọng về mặt lý thuyết. Do đó, cách làm thận trọng và nhạy bén hơn là loại bỏ dần từng biến. Có một vài lý do đối với việc loại bỏ các biến với các hệ số không có nghĩa. Thứ nhất, một mô hình đơn giản hơn dễ diễn giải hơn một mô hình phức tạp. Thứ hai, việc bỏ bớt một biến làm tăng bậc tự do và vì vậy cải thiện sự chính xác của các hệ số còn lại. Cuối cùng, như chúng ta sẽ thấy trong chương tiếp theo, nếu các biến giải thích có tương quan chặt với nhau nó sẽ gây khó khăn cho sự diễn giải riêng từng hệ số. Việc loại trừ các biến làm giảm cơ hội nảy sinh những tương quan này và vì vậy nó làm cho việc diễn giải có ý nghĩa hơn.

Điểm bắt đầu cho quá trình loại bỏ là nhận diện biến có hệ số hồi qui ít có nghĩa nhất. Điều này được thực hiện bằng cách nhìn vào giá trị *p-value* cao nhất trong mô hình ước lượng không có hằng số. Về trung bình, hệ số tương ứng được kỳ vọng gần bằng không, và vì vậy chúng ta tin rằng bất cứ thiên lệch nào bị gây ra do sự loại bỏ sẽ là rất nhỏ. Từ kết quả mô hình A, chúng ta để ý rằng hệ số cho GASPRICE có giá trị *p-value* cao nhất và vì vậy ít có ý nghĩa nhất. Do đó, biến này bị loại bỏ khỏi đặc trưng mô hình và chúng ta hãy xem điều gì xảy ra. Dựa trên đó chúng ta có thể loại bỏ nhiều biến hơn. Quá trình này được gọi là **Đơn giản mô hình dựa trên số liệu**.]

Bảng 4.4 (Tiếp theo)

```
MODEL 2: OLS estimates using the 40 observations 1-40
Dependent variable: BUSTRAVL

      VARIABLE      COEFFICIENT      STDERROR      T STAT     2PROB(t > |T|)

 0)   const           3215.8565     1090.4692       2.949      0.005730 ***
 2)   FARE            -225.6595      440.4936      -0.512      0.611762
 4)   INCOME            -0.1957        0.0638      -3.069      0.004203 ***
 5)   POP                1.7168        0.2265       7.581      0.000000 ***
 6)   DENSITY            0.1182        0.0580       2.037      0.049453 **
 7)   LANDAREA          -1.1953        1.7656      -0.677      0.502980

Mean of dep. var           1933.175   S.D. of dep. variable          2431.757
Error Sum of Sp (ESS)   1.8235e+007   Std Err of Resid. (sgmahat)    732.3323
Unadjusted R-squared          0.921   Adjusted R-squared               0.909
F-statistic (5,34)           79.204   p-value for F()                0.000000
Durbin-Watson stat            2.079   First-order autocorr. coeff      -0.155
```

```
MODEL SELECTION STATISTICS

SGMASQ          536311      AIC             615352      FPE             616757
HQ              674378      SCHWARZ         792765      SHIBATA         592623
GCV             630954      RICE            651234

Excluding the constant, p-value was highest for variable 2 (FARE) of the
8 model selection statistics, 8 have improved.
```

[Lưu ý rằng tất cả 8 ràng buộc lựa chọn mô hình đã được cải thiện, nghĩa là, giảm đi. Cũng vậy, việc loại bỏ GASPRICE đã cải thiện độ chính xác của các hệ số còn lại bằng cách làm cho chúng có ý nghĩa nhiều hơn – chẳng hạn, hằng số và DENSITY. Biến có hệ số ít ý nghĩa nhất, nghĩa là, giá trị *p-value* cao nhất, bây giờ là FARE. Nhưng vé xe buýt là một thước đo giá cả mà theo cách nói lý thuyết kinh tế là một yếu tố quan trọng của nhu cầu. Do đó, chúng ta không nên loại bỏ nó ngay cả khi giá trị *p-value* cho rằng chúng ta có thể bỏ. Do vậy bước kế tiếp là loại bỏ LANDAREA, biến có giá trị *p-value* cao nhất kế tiếp.]

```
MODEL 3: OLS estimates using the 40 observations 1-40
Dependent variable: BUSTRAVL

      VARIABLE     COEFFICIENT       STDERROR       T STAT    2PROB(t > |T|)

 0)   const          3111.1805      1071.0669        2.905     0.006330 ***
 2)   FARE             -295.7306       424.8354       -0.696     0.490959
 4)   INCOME             -0.2022         0.0626       -3.232     0.002680 ***
 5)   POP                 1.5883         0.1227       12.950     0.000000 ***
 6)   DENSITY             0.1490         0.0357        4.173     0.000189 ***

Mean of dep. var           1933.175  S.D. of dep. variable           2431.757
Error Sum of Sp (ESS)   1.848e+007  Std Err of Resid. (sgmahat)      726.6434
Unadjusted R-squared         0.920  Adjusted R-squared                 0.911
F-statistic (5,34)         100.445  p-value for F()                 0.000000
Durbin-Watson stat           1.995  First-order autocorr. coeff       -0.102
```

Bảng 4.4 (Tiếp theo)

```
MODEL SELECTION STATISTICS

SGMASQ          528011      AIC             593232      FPE             594012
HQ              640287      SCHWARZ         732670      SHIBATA         577512
GCV             603441      RICE            616012

Excluding the constant, p-value was highest for variable 2 (FARE) of the
8 model selection statistics, 8 have improved.
```

[Biến DENSITY đã gia tăng đáng kể. Tuy nhiên, biến FARE có giá trị *p-value* là 49%, quá cao không thể chấp nhận được với bất cứ mức ý nghĩa hợp lý nào. Điều này gợi ý rằng, với sự có mặt của các biến khác, giá cả có thể không ảnh hưởng lên nhu cầu đi xe buýt. Nói cách khác, khi có nhu cầu đi xe buýt, người tiêu thụ có thể không nhạy cảm lắm với giá cả. Do vậy, loại bỏ FARE là cần thiết và xem điều gì xảy ra.]

```
MODEL 4: OLS estimates using the 40 observations 1-40
Dependent variable: BUSTRAVL

      VARIABLE     COEFFICIENT        STDERROR        T STAT     2PROB(t > |T|)

 0)   const          2815.7032        976.3007         2.884      0.006589 ***
 4)   INCOME            -0.2013          0.0621        -3.241      0.002566 ***
 5)   POP                1.5766          0.1206        13.071      0.000000 ***
 6)   DENSITY            0.1534          0.0349         4.396      0.000093 ***

Mean of dep. var              1933.175   S.D. of dep. variable          2431.757
Error Sum of Sp (ESS)     1.8736e+007   Std Err of Resid. (sgmahat)    721.4228
Unadjusted R-squared             0.919   Adjusted R-squared                0.912
F-statistic (5,34)             135.708   p-value for F()               0.000000
Durbin-Watson stat               1.879   First-order autocorr. coeff      -0.043

MODEL SELECTION STATISTICS

SGMASQ           520451   AIC              572112   FPE              572496
HQ               608137   SCHWARZ          677373   SHIBATA          562087
GCV              578279   RICE             585507

Of the 8 model selection statistics, 8 have improved.
```

[Lưu ý rằng mô hình 4 có các trị thống kê lựa chọn mô hình thấp nhất và tất cả các hệ số đều có nghĩa rất lớn. Cũng vậy, các hệ số đối với INCOME, POP, và DENSITY không khác với các hệ số giữa mô hình 3 và mô hình 4. Vì vậy sự thiên lệch trong việc loại bỏ FARE không quá nghiêm trọng.

Vì lợi ích của sự hoàn tất, thật đáng để xem mô hình 1 như một mô hình không giới hạn và Mô hình 4 như một mô hình giới hạn và để thực hiện một kiểm định *F*-test để kiểm tra xem liệu các hệ số của GASPRICE, LANDAREA, và FARE là đồng thời khác với không. Kết quả cho ở dưới đây.]

```
F(3,33): area to the right of 0.315845 = 0.813800
```

[*Giả thuyết không* đối với kiểm định *F* Wald phát biểu rằng các hệ số của tất cả các biến bị loại bỏ đều bằng không, nghĩa là, hệ số $\beta_2 = \beta_3 = \beta_7 = 0$. Vì giá trị *p-value* trong trường hợp này là 0.8138, giá trị này cao trong bất cứ tiêu chuẩn hợp lý nào, chúng ta không thể bác bỏ *giả thuyết không*. Sử dụng tính toán và phương trình (4.3), kiểm tra lại trị thống kê *F* cho các biến bị loại bỏ đã cho ở trên là 0,315845 (lưu ý rằng mô hình 1 là mô hình không bị giới hạn và mô hình 4 là mô hình giới hạn đối với kiểm định này). Sau đó dùng bảng *F* với mức 10% được cho trong Bảng A.4c và kiểm tra lại rằng bạn không thể bác bỏ *giả thuyết không* ở mức 10%. Vì vậy, các hệ số của FARE, GASPRICE, và LANDAREA thì đồng thời không có nghĩa ở mức này. Dựa trên tất cả các ràng buộc, mô hình 4 dường như là "tốt nhất" và được chọn như là mô hình cuối cùng cho việc diễn dịch.

Các hệ số của thu nhập, kích thước dân số, và mật độ dân số có ý nghĩa rất lớn. Lý thuyết kinh tế chuẩn cho rằng ảnh hưởng thu nhập lên nhu cầu đối với bất cứ hàng hóa nào đều dương, nhưng hệ số ước lượng của INCOME thì lại âm. Điều này, không gây ngạc nhiên, gợi ý rằng đi xe buýt là một loại hàng hóa "thấp cấp". Khi thu nhập tăng lên,

người ta có khuynh hướng sử dụng xe hơi để di chuyển, và vì vậy lượng đi xe buýt sẽ giảm xuống. Nếu thu nhập đầu người tăng lên khoảng 100 đô la, thì, về trung bình, đi xe buýt được kỳ vọng giảm khoảng $100|\beta_4|$, nghĩa là, khoảng 20,13 ngàn người mỗi giờ. Như kỳ vọng, hệ số của POP và DENSITY dương. Nói cách khác, khi kích thước dân số hay mật độ dân số tăng lên, thì có nhiều người di chuyển bằng xe buýt hơn. Tuy nhiên, mặc dù giá trị số của DENSITY rất nhạy, nhưng đối với POP thì lại không bởi vì nó lớn hơn 1 (chú ý rằng cả hai DENSITY và POP đều được đo lường cùng đơn vị). Điều này gợi ý một khả năng đặc trưng sai mô hình.

Khi ước lượng các mối quan hệ nhu cầu, người ta thường đặt câu hỏi liệu nhu cầu "co giãn" hay "không co giãn" đối với giá cả và thu nhập. Việc trả lời cho câu hỏi đó đòi hỏi ước lượng mối quan hệ phi tuyến tính, một chủ đề được khảo sát chi tiết ở Chương 6.]

4.7 Ứng dụng: Sự tham gia lực lượng lao động của nữ giới

Ứng dụng thứ hai xuyên suốt được dùng ở đây là nghiên cứu kinh tế lượng xác định tỷ lệ tham gia lực lượng lao động của nữ giới – phần trăm nữ giới trên 16 tuổi trong lực lượng lao động thực sự đang làm việc hay tìm việc. DATA4-5 đã mô tả trong phụ lục D trình bày dữ liệu điều tra dân số năm 1990 cho 50 bang trên nhiều biến (biến đầu tiên là biến phụ thuộc):

WLFP =	Tỷ lệ tham gia (%) của mọi phụ nữ trên 16 tuổi (phần trăm phụ nữ trong lực lượng lao động)
YF =	Mức lương trung vị (ngàn đô-la) của nữ
YM =	Mức lương trung vị (ngàn đô-la) của nam
EDUC =	Phần trăm nữ giới tốt nghiệp trung học trên 24 tuổi
UE =	Tỷ lệ thất nghiệp (%)
MR =	Tỷ lệ kết hôn (%) của nữ giới từ 16 tuổi trở lên
DR =	Tỷ lệ ly hôn
URB =	Phần trăm dân số thành thị trong nước
WH =	Phần trăm phụ nữ da trắng trên 16 tuổi

Mô hình kinh tế lượng dùng tất cả các biến giải thích như sau:

$$\text{WLFP} = \beta_1 + \beta_2\text{YF} + \beta_3\text{YM} + \beta_4\text{EDUC} + \beta_5\text{UE} + \beta_6\text{MR} + \beta_7\text{DR} + \beta_8\text{URB} + \beta_9\text{WH} + u$$

Trước khi thực sự ước lượng mô hình, việc thảo luận các dấu của các hệ số hồi qui kỳ vọng là rất hữu ích. Sự thảo luận được rút ra dựa trên "lý thuyết kinh tế" tương phản với "lý thuyết kinh tế lượng". Bạn đọc có thể tham khảo bài viết của O'Neill (1981), Kelley và Da Silva (1980), và King (1978) để biết thêm chi tiết về vài lý thuyết này.

YF: đây là độ đo lường tiền trả cho người lao động nữ, ta kỳ vọng nó có hiệu ứng dương lên biến WLFP. Nói cách khác, lương càng cao, nữ giới càng tham gia lao động. Tuy nhiên, ta nên nhớ rằng lý thuyết lao động nói rằng "hiệu ứng của thu nhập" lên lao động là âm; nghĩa là khi thu nhập tăng, người lao động mong muốn thư nhàn hơn (ít

việc). Với tiền lương hiện hành, hiệu ứng này có thể yếu; và do đó, khi cân bằng, ta kỳ vọng biến này có hệ số dương.

YM: Khi người chồng làm ra tiền nhiều hơn, người vợ không cần làm việc nhiều. Do đó, ta kỳ vọng hệ số này là âm. Cũng có thể bởi vì nhiều phụ nữ có khả năng chuyên môn tốt, cho nên thu nhập của nam giới càng cao khiến càng nhiều phụ nữ tìm những việc như vậy. Tuy nhiên, điều này tác động đến loại công việc và hầu như không tác động đến việc nhiều phụ nữ tham gia lực lượng lao động hơn hay không.

EDUC: Sự giáo dục càng nhiều ngụ ý càng có nhiều cơ hội việc làm (mong ước) sẵn có cho nữ. Vậy, ta kỳ vọng hệ số này dương.

UE: Tỷ lệ thất nghiệp có cả hiệu ứng âm và dương. "giả thuyết người lao động chán nản" nói rõ rằng tỷ lệ thất nghiệp càng cao là một dấu hiệu cho phụ nữ (và bộ phận người thiểu số) biết rằng tìm việc là công việc vô ích. Điều này làm cho họ rời khỏi lực lượng lao động, vậy hệ số này có dấu âm. Cũng có thể có hiệu ứng dương. Nếu người chồng mất việc, người vợ có thể phải tham gia lao động để bù vào khoản tiền bị mất. Nếu hiệu ứng này không mạnh, thì dấu âm sẽ chiếm ưu thế.

MR: Nếu một phụ nữ kết hôn, cô ta có xu hướng có ít cơ hội làm việc (đặc biệt khi họ có con) và có thể giảm mong muốn và sự cần thiết có việc. Vậy tỷ lệ kết hôn cao có thể giảm tỷ lệ tham gia lao động của nữ – WLFP.

DR: Ta kỳ vọng dấu dương cho biến này bởi vì khi tỷ lệ ly hôn cao, nhiều phụ nữ có thể tham gia lực lượng lao động nhằm tự chu cấp cho họ.

URB: Tại các khu vực thành thị cơ hội việc làm nhiều hơn tại nông thôn. Ta kỳ vọng rằng những tiểu bang có phần dân số sống ở thành thị nhiều hơn sẽ có tỷ lệ tham gia lao động nữ cao hơn. Mặt khác, phụ nữ nông thôn có chiều hướng tự sống bằng nuôi thú nuôi và gia cầm và làm những việc đồng áng khác. Vậy, họ đã là một phần lực lượng lao động. Điều này có nghĩa rằng nếu một tiểu bang có dân số nông thôn đông hơn (nghĩa là ít URB), thì sự tham gia lao động nữ sẽ cao hơn, kết quả là hệ số âm. Hiệu ứng sau cùng có thể được xác định chỉ theo kinh nghiệm.

WH: Không có dấu rõ ràng kỳ vọng trước cho biến này. Nếu phụ nữ da màu tương đối không giỏi chuyên môn và tìm loại việc như giúp việc hay quản gia, ta kỳ vọng dấu âm cho hệ số này bởi vì tỷ lệ phụ nữ da trắng (WH) cao hơn thì số phụ nữ da màu thấp hơn. Cũng vậy, nếu phụ nữ da trắng tương đối giàu có, họ có thể không tham gia lực lượng lao động. Điều này cũng sẽ dẫn đến dấu âm. Nếu những giả thiết này không đúng, kết quả sẽ là dấu dương hoặc bằng 0.

Bảng 4.5 cho thấy kết quả chạy máy tính từng phần với những chú thích (xem Phần 4.5 Thực hành Máy tính). Dùng chương trình hồi qui của chính bạn và DATA 4-5 để mô phỏng các kết quả. Sau đó nghiên cứu kỹ các kết quả trước khi tiến hành tiếp.

✪ Bảng 4.5 Kết quả chạy máy tính từng phần có chú thích tỷ lệ tham gia lực lượng lao động của nữ giới

[Mô hình với tất cả các biến (thường được xem là mô hình "bồn rửa chén")]

```
MODEL 1: OLS estimates using the 50 observations 1-50
Dependent variable: wlfp

     VARIABLE     COEFFICIENT      STDERROR      T STAT    2Prob(t > |T|)

0)      const         44.5096        8.9750       4.959       0.000013 ***
2)         yf          0.9880        0.4076       2.424       0.019847 **
3)         ym         -0.1743        0.3062      -0.569       0.572212
4)       educ          0.2851        0.0932       3.060       0.003888 ***
5)         ue         -1.6106        0.3136      -5.136       0.000007 ***
6)         mr         -0.0782        0.1731      -0.452       0.653835
7)         dr          0.4374        0.2583       1.693       0.098035 *
8)        urb         -0.0926        0.0333      -2.776       0.008195 ***
9)         wh         -0.0875        0.0398      -2.196       0.033819 **
```

❍ Bảng 4.5 (tiếp theo)

```
Mean of dep. var.           57.474   S.D. of dep. variable           4.249
Error Sum of Sq (ESS)     193.9742   Std Err of Resid. (sgmahat)     2.1751
Unadjusted R-squared         0.781   Adjusted R-squared              0.738
F-statistic (8, 41)        18.2459   p-value for F( )             0.000000
Durbin-Watson stat.          1.637   First-order autocorr. coeff     0.179

MODEL SELECTION STATISTICS

SGMASQ         4.73108      AIC            5.56058      FPE         5.58267
HQ             6.33926      SCHMARZ        7.84492      SHIBATA      5.2761
GCV            5.76961      RICE           6.06169

Excluding the constant, p-value was highest for variable 6 (mr).
```

[Lưu ý rằng ym và mr có giá trị p cao và là các biến ưu tiên để loại ra khỏi mô hình. Bây giờ ta bỏ các biến mỗi lần một biến, bắt đầu với mr, có giá trị p cao nhất]

```
MODEL 2: OLS estimates using the 50 observations 1-50
Dependent variable: wlfp

     VARIABLE     COEFFICIENT      STDERROR      T STAT    2Prob (t > |T|)


0)      const         41.3460        5.5598       7.437       0.000000 ***
2)         yf          1.0671        0.3645       2.927       0.005497 ***
3)         ym         -0.1984        0.2987      -0.664       0.510097
4)       educ          0.2582        0.0709       3.643       0.000734 ***
5)         ue         -1.5910        0.3076      -5.171       0.000006 ***
7)         dr          0.3916        0.2354       1.664       0.103626
8)        urb         -0.0876        0.0311      -2.814       0.007420 ***
9)         wh         -0.0851        0.0391      -2.175       0.035271 **
```

```
Mean of dep. var.           57.474   S.D. of dep. variable           4.249
Error Sum of Sq (ESS)     194.9397   Std Err of Resid. (sgmahat)    2.1544
Unadjusted R-squared         0.781   Adjusted R-squared              0.743
F-statistic (7, 42)        21.2255   p-value for F( )             0.000000
Durbin-Watson stat.          1.649   First-order autocorr. coeff     0.173


MODEL SELECTION STATISTICS


SGMASQ         4.64142      AIC            5.36914      FPE         5.38405
HQ             6.03252      SCHMARZ        7.29064      SHIBATA     5.14641
GCV             5.5255      RICE           5.73352
```

Bảng 4.5 (tiếp theo)

```
Excluding the constant, p-value was highest for variable 3 (ym).
Of the 8 model selection statistics, 8 have improved
```

[Bỏ biến ym, là biến vẫn còn giá trị p cao, và chú ý rằng bây giờ dr trở nên có ý nghĩa ở mức 10 phần trăm]

```
MODEL 3: OLS estimates using the 50 obsetvations 1-50
Dependent variable: wlfp

     VARIABLE     COEFFICIENT      STDERROR      T STAT    2Prob (t > |T|)

0)      const         41.8336        5.4753       7.640       0.000000 ***
2)         yf          0.8493        0.1582       5.370       0.000003 ***
4)       educ          0.2492        0.0691       3.606       0.000804 ***
5)         ue         -1.6776        0.2769      -6.059       0.000000 ***
7)         dr          0.4341        0.2251       1.929       0.060390 *
8)        urb         -0.0942        0.0293      -3.212       0.002500 ***
9)         wh         -0.0961        0.0352      -2.729       0.009156 ***

Mean of dep. var.           57.474   S.D. of dep. variable           4.249
Error Sum of Sq (ESS)     196.9882   Std Err of Resid. (sgmahat)    2.1404
Unadjusted R-squared         0.777   Adjusted R-squared              0.746
F-statistic (6, 43)        25.0145   p-value for F( )             0.000000
Durbin-Watson stat.          1.668   First-order autocorr. coeff     0.165


MODEL SELECTION STATISTICS


SGMASQ         4.58112      AIC            5.21282      FPE         5.22248
HQ             5.77222      SCHMARZ        6.81281      SHIBATA      5.0429
GCV            5.32688      RICE           5.47189

Of the 8 model selection statistics, 8 have improved
```

[Dùng Mô hình 3 làm mô hình giới hạn và Mô hình 1 làm mô hình không giới hạn, ta có thể thực hiện F-test. Kết quả cho như sau.]

```
F(2, 41): area to the right of 0.318535 = 0.728997
```

[Dùng một máy tính, thực hiện thống kê kiểm định Wald khi bỏ các biến ym và dr. *Giả thuyết không* cho kiểm định này là $\beta_3 = \beta_7 = 0$. Như trên, giá trị p là xác suất của sai lầm loại I nếu ta bác bỏ *giả thuyết không*. Vì 0,279 là quá cao cho bất cứ mức ý nghĩa hợp lý nào, ta không nên bác bỏ *giả thuyết không* mà thay vào đó kết luận rằng ym và dr cùng không có ý nghĩa liên kết. Bạn nên chứng minh điều này bằng cách dùng bảng F trong Phụ lục A.4c với mức ý nghĩa 10 phần trăm. Tất cả các trị thống kê chọn lựa mô hình là thấp nhất trong Mô hình 3. Do đó, ta chọn Mô hình 3 là mô hình cuối cùng "tốt nhất" để khảo sát tiếp. Để giải thích các kết quả, xem bài đọc.]

Trong Mô hình 3, được chọn là mô hình cuối cùng "tốt nhất", dấu dương tại biến YF chỉ ra rằng "hiệu ứng đường cung bẻ ngược" lên lao động – nghĩa là, khi tiền lương tăng người lao động thích thư nhàn hơn và ít tham gia vào lực lượng lao động – là yếu. Mọi điều khác như nhau, lương của một phụ nữ tăng lên $1.000 thì tỷ lệ tham gia lao động của cô ta được kỳ vọng tăng trung bình 0,849 phần trăm.

Tiền lương của nam giới (YM) đã không có ý nghĩa. Điều này có thể bởi vì biến này được liên kết chặt chẽ với biến YF và bị bao gộp bởi hệ số của biến YF.

Như đã kỳ vọng, giáo dục tăng làm cho nhiều phụ nữ tìm việc hơn. Tỷ lệ phụ nữ tốt nghiệp trung học tăng 1 phần trăm sẽ tăng tỷ lệ tham gia lao động trung bình 0,249 phần trăm.

Dấu âm của biến UE xác nhận cho "giả thuyết người lao động chán nản", nói rõ rằng khi tỷ lệ thất nghiệp cao, phụ nữ đang tìm việc có thể chán nản và rời khỏi lực lượng lao động. Mức quan trọng của hệ số này hoàn toàn cao. Biến UE tăng 1 phần trăm đồng nghĩa với tỷ lệ tham gia lao động WLFP giảm trung bình 1,678 phần trăm.

Tỷ lệ ly hôn có dấu dương. Trung bình, tỷ lệ ly hôn tăng 1 phần trăm sẽ kỳ vọng làm cho tỷ lệ tham gia lao động WLFP tăng 0,434 phần trăm. Tuy nhiên, tỷ lệ kết hôn (MR) không có ý nghĩa về mặt thống kê.

Hệ số âm của biến URB (-0,094) xác nhận luận điểm trước đây rằng dân số nông thôn cao (nghĩa là, URB thấp) có thể làm cho WLFP cao bởi vì phụ nữ nông thôn làm nhiều công việc đồng áng và nghĩa là tham gia lực lượng lao động.

Phần trăm phụ nữ da trắng tăng 1 phần trăm làm cho tỷ lệ tham gia lao động nữ giảm trung bình 0,096 phần trăm.

Giá trị $\bar{R}^2$ cho biết chỉ khoảng 74 phần trăm thay đổi trong tỷ lệ tham gia lao động liên bang được giải thích bởi Mô hình C. Vậy, ta có thể bỏ vài biến để tăng khả năng giải thích của mô hình. Tuy nhiên, dữ liệu chéo giữa các lớp cho ra $\bar{R}^2$ thấp là hoàn toàn đặc trưng. Bởi vì dữ liệu theo chuỗi nói chung nhiều lần phát triển quá mức, các mô hình dựa trên dữ liệu này có chiều hướng cho độ thích hợp một cách tương đối. Có thể thấy điều này qua giá trị của $\bar{R}^2$ (0,999) đối với hàm tiêu dùng được trình bày trong Ví dụ 4.11. Với dữ liệu thêm vào, ta có thể có sự giải thích tốt hơn về tỷ lệ tham gia lao động của nữ giới. Các biến có thể tính đến khi hồi qui như sau:

1. Quy mô gia đình, tỷ lệ sinh sản, và số trẻ em dưới một "ngưỡng" tuổi; các yếu tố này có chiều hướng làm giảm cơ hội việc làm của nữ giới.
2. Một biến đo lường số phụ nữ tốt nghiệp đại học.

3. Phân phối tuổi của nữ giới
4. Trợ cấp trả cho phụ nữ độc thân có trẻ em; yếu tố này có thể khiến cho phụ nữ đi làm hoặc ở nhà (Tính sẵn có của sự chăm sóc hàng ngày cũng có cùng hiệu ứng)
5. Độ đo thể hiện sự khác nhau giữa các vùng; vùng trang trại và vùng công nghiệp có thể có các kiểu hành vi khác nhau.

Những nhận xét quan trọng khi diễn giải các hệ số hồi qui

Khi diễn giải các hệ số hồi qui ước lượng cần phải thật thận trọng. Trước hết, dấu của một hệ số hồi qui có thể trái ngược với những gì bạn kỳ vọng ban đầu. Nếu hệ số không có ý nghĩa về mặt thống kê (nghĩa là, bạn không thể bác bỏ *giả thuyết không* cho rằng hệ số bằng không), thì sự sai dấu là không thích hợp bởi vì về mặt thống kê, giá trị bằng số có thể mang dấu dương hoặc âm ngang nhau và đó đơn thuần là tình cờ ngẫu nhiên bạn thu được dấu sai. Trong các nhận xét có chú thích trong bảng 4.4, ta đã nói rõ một số lý do hợp lý để loại bỏ một biến có một hệ số không có ý nghĩa (diễn giải dễ hơn, ý nghĩa hơn và chính xác hơn). Trong một trường hợp như vậy, đơn giản là bạn nên bỏ biến số ra và ước lượng lại mô hình với sự tin rằng độ thiên lệch của biến vừa loại bỏ là không đáng kể. Nên chú ý rằng khi bỏ một biến ra không có nghĩa bạn nói biến đó không có hiệu ứng lên biến Y, mà phải hiểu là, *mọi thứ khác như nhau*, biến đang bàn đến không có hiệu ứng *riêng lẻ*. Hiệu ứng của nó được thể hiện qua sự hiện diện của biến khác có tương quan (Chương kế tiếp sẽ đề cập nhiều hơn)

Khi quyết định chọn sự ý nghĩa hoặc không của một thông số hồi qui, một câu hỏi đáng quan tâm là "Với mức giá trị nào của p ta cho rằng là cao để bác bỏ *giả thuyết không* của hiệu ứng 0?" Hầu hết các nhà phân tích dùng mức 5 phần trăm (hoặc 0.05) làm chuẩn. Mức ưa thích của cá nhân tôi là 10 phần trăm. Một ưu điểm khi dùng giá trị cao hơn là có nhiều biến được giữ lại trong mô hình hơn (giải thích vì sao đây là tình huống), do vậy giảm bất cứ sự thiên lệch nào của biến bỏ đi. Không giống như những thí nghiệm y học, khi mà sai lầm có thể trả giá rất đắt, hành vi kinh tế phải chịu nhiều yếu tố không chắc chắn, và do đó mức dung sai phải cao hơn. Tuy nhiên, nếu cỡ mẫu (n) là rất lớn, ta nên dùng giá trị p ngặt hơn. Bởi vì khi n lớn, các độ lệch chuẩn sẽ nhỏ, làm cho hầu hết mọi hệ số đều có ý nghĩa.

Ta nên làm gì nếu hệ số có dấu ngược có ý nghĩa về mặt thống kế? Ta nên tìm câu giải thích. Lấy ví dụ, trong ví dụ 4.1 về giá nhà, ta phát hiện những dấu âm khác thường của 2 biến BEDRMS và BATHS. Tuy nhiên, theo ý nghĩa hợp lý của một hệ số hồi qui – nghĩa là hiệu ứng từng phần, *khi tất cả các biến khác không đổi giá trị* – ta thấy rằng các hệ số âm xét cho cùng là không quá ngạc nhiên. Ở ví dụ thứ hai, trong mô hình du lịch xe buýt trong Phần 4.6 (xem Bảng 4.4), ta phát hiện hệ số thu nhập có dấu âm, trái ngược với điều mọi người thường kỳ vọng. Trong trường hợp này, ta có thể đi đến một sự giải thích hợp lý khi nhận thấy rằng dấu âm cho biết du lịch xe buýt là một "hàng hóa thấp cấp". Chương 5 cung cấp những ví dụ khác của các trường hợp mà ta bắt gặp những dấu khác thường và đề xuất những biện pháp xử lý. Các ví dụ này nên được nghiên cứu kỹ lưỡng.

Một lưu ý quan trọng khác là phải chú ý đến các đơn vị đo của các biến khi diễn giải các giá trị bằng số của các hệ số hồi qui (xem Phần 3.6 về chuyển đổi đơn vị để nhớ lại). Bạn cũng thật thận trọng khi diễn giải những biến được thể hiện bằng phân số hoặc phần trăm (ví dụ, tỷ lệ thất nghiệp và lãi suất). Nếu bạn thực hiện dự án thực nghiệm của chính bạn, nhìn chung nên tránh dạng phân số hoặc tỷ lệ mà hãy biểu diễn những biến như

vậy theo phần trăm. Lý do bởi vì sẽ dễ dàng hơn khi diễn dịch hiệu ứng của 1 phần trăm thay vì 0,01 thay đổi trong 1 biến số. Tuy nhiên, trong một bài viết thực nghiệm nhà điều tra nghiên cứu có thể biểu diễn vài biến dưới dạng tỷ lệ. Trong những trường hợp như thế, phải thật thận trọng khi diễn dịch các giá trị bằng số. Ở phần này, bạn đọc sẽ hiểu hơn khi xem lại sự diễn dịch của các hệ số ước lượng của các biến dạng phần trăm trong ví dụ tham gia lực lượng lao động của nữ giới vừa mới thảo luận.

4.8 Ví dụ thực nghiệm: Tỷ lệ di trú ròng và chất lượng cuộc sống

Liu (1975) đã nghiên cứu mối quan hệ giữa sự thay đổi tỷ lệ di trú ròng giữa các bang và một số các biến giải thích, trong đó gồm "chất lượng cuộc sống". Dữ liệu chéo trên 50 bang, và mô hình cơ bản được dùng như sau:

$$MIGRATE = f(QOL, Y, E, IS, ES, AP, ED, HW)$$

Trong đó

MIGRATE =	Tỷ lệ di trú ròng giữa năm 1960 và 1970 (số chuyển đến trừ số chuyển đi chia cho dân số)
QOL =	Chỉ số chất lượng cuộc sống
Y =	Chỉ số thu nhập bang trên thu nhập quốc gia
E =	Tỷ lệ giữa số việc làm của bang trên số việc làm quốc gia
IS =	Chỉ số tình trạng cá nhân
ES =	Chỉ số tình trạng nền kinh tế
AP =	Chỉ số sản xuất nông nhgiệp
ED =	Chỉ số phát triển giáo dục
HW =	Chỉ số trợ cấp phúc lợi và chăm sóc sức khỏe

Dựa trên những chỉ tiêu được phát triển bởi Ủy ban Mục tiêu Quốc gia của Chủ tịch Eisenhower. Liu đã xây dựng mỗi chỉ số liệt kê như trên. QOL là trung bình số học của các chỉ số khác của chất lượng cuộc sống. Bảng 4.6 có các hệ số ước lượng và các thống kê liên quan cho một số mô hình hồi qui bội liên kết tỷ lệ di trú với các chỉ số chất-lượng-cuộc-sống. Để thưởng thức bài nghiên cứu về di trú của tác giả, sinh viên nên đọc nguyên bản bài viết. Mặc dù những chủ đề đề cập trong chương này đủ để hiểu rõ các mô hình và các kết quả, ở đây ta chỉ trình bày tóm tắt các kết quả này.

Tác giả đã không cung cấp thông tin các tổng bình phương phần dư cho các mô hình, và do vậy ta không thể so sánh các mô hình bằng cách dùng các tiêu chuẩn chung đã được trình bày trước đây. Độ thích hợp có thể được đánh giá chỉ bởi $\overline{R}^2$. Ta lưu ý rằng thu nhập và việc làm tự bản thân không giải thích bất kỳ sự thay đổi nào trong biến di trú. Giá trị của $\overline{R}^2$ là âm trong Mô hình 2. QOL tự thân giải thích khoảng 6 phần trăm thay đổi trong biến di trú. Nếu thu nhập và việc làm được thêm vào QOL (Mô hình 3), $\overline{R}^2$ giảm một cách đáng kể. Điều này hàm ý rằng 2 biến này hầu như không thuộc về mô hình. Trong Mô hình 4, tác giả loại trừ Y và E. Ta lưu ý rằng, khi loại bỏ yếu tố phúc lợi và chăm sóc sức khỏe (HW), mọi biến chất-lượng-cuộc-sống khác đều có ý nghĩa hoặc hầu như có ý nghĩa ở mức ý nghĩa 5 phần trăm. Bởi vì HW không có ý nghĩa trong các mô hình, tốt hơn nên bỏ biến

này ra và ước lượng lại mô hình để ước lượng các hệ số còn lại hiệu quả hơn. Nhưng tác giả đã quyết định giữ biến số lại để tránh sự thiên lệch có thể có của biến bị bỏ đi.

Tất cả các biến chất lượng-cuộc-sống có dấu kỳ vọng dương khi loại bỏ biến phát triển giáo dục (dấu âm của biến HW có thể bỏ qua bởi vì nó không có ý nghĩa về mặt thống kê). Sự hợp lý của Liu trong các kết quả khác thường này được tái diễn lại ở đây (Liu, 1975, trang 333):

Bảng 4.6 Tương quan ước lượng giữa Di trú và Chất lượng của cuộc sống

Biến độc lập	Mô hình 1	Mô hình 2	Mô hình 3	Mô hình 4	Mô hình 5
CONSTANT	-23.05	104.62	55.94	-16.46	-62.50
QOL	24.06		23.40		
	(2.05)		(1.93)		
Y		0.36	-0.74		7.19
		(0.05)	(-0.10)		(1.11)
E		103.47	-77.26		41.76
		(-0.48)	(-0.37)		(0.23)
IS				28.68	30.21
				(2.02)	(2.14)
ES				20.03	20.49
				(2.24)	(2.28)
AP				18.73	19.13
				(2.87)	(2.89)
ED				-31.56	-33.48
				(-3.46)	(-3.59)
HW				-18.45	-21.69
				(-1.41)	(-1.57)
$\overline{R}^2$	0.06	-0.03	0.02	0.37	0.36
D.F.	48	47	46	44	42

Lưu ý: Các giá trị trong () là thông kê t

Nguồn: Liu (1975), Tái bản với sự cho phép của Hiệu trưởng và hội viên của trường Harvard

> Trước hết, trong khi sự di trú là biến động thì biến giáo dục đại diện cho khái niệm tĩnh. Điều này dẫn đến một tiến trình hiệu chỉnh cân bằng giữa khối người được giáo dục và dòng di trú tại Mỹ, nghĩa là những bang được biết là có sự phát triển giáo dục đáng kể đang xuất khẩu nhân lực có trình độ cao sang những bang mà nhân lực có kỹ năng cao tương đối khan hiếm và kết quả là, di dân có trình độ cao tìm thấy ở những bang này cơ hội nghề nghiệp nhiều hơn cũng như nhiều công việc thích đáng hơn. Thứ hai, những di dân không đồng nhất về quá trình giáo dục, và quyết định di trú của họ thường bị tác động bởi bạn bè hay người thân tại nơi đến, những người này thường có cùng trình độ giáo dục như họ. Hệ quả là, những bang có cư dân không đồng nhất về trình độ văn hóa thì kỳ vọng sẽ có tỷ lệ di trú ròng cao hơn những bang tương đối đồng nhất. Tuy nhiên, cần phải nghiên cứu bổ sung thêm để đánh giá hiệu ứng của biến giáo dục này lên sự di trú.

4.9 Dự án thực nghiệm

Nếu dự án thực nghiệm là một phần trong khóa học kinh tế lượng của bạn, bạn nên theo hướng dẫn trong phần 1.3 và thu thập vài dữ liệu. Nếu bạn có thông tin đủ các biến, bạn nên nhập dữ liệu vào máy tính và chắc rằng dữ liệu được nhập một cách chính xác (nếu bạn đang dùng GRELT, hãy đọc sách hướng dẫn để sắp đặt file dữ liệu của chính bạn). Sau đó bạn có thể thử mô hình đầu tiên, loại bỏ các biến và thực hiện kiểm định Wald, và kế tiếp áp dụng kỹ thuật đơn giản hóa mô hình dựa trên dữ liệu để khử các biến. Tuy nhiên tất cả các bước này đơn thuần là để thực hành và hiểu rõ thêm những chủ đề được đề cập trong chương này. Bạn không nên quá xem trọng các kết quả, bởi vì cần phải có nghiên cứu lý thuyết đáng kể trước khi đảm nhận một mô hình ý nghĩa và phân tích.

Tóm tắt

Trong mô hình hồi qui tuyến tính bội, biến phụ thuộc (Y) được hồi qui dựa vào k biến độc lập X_1, X_2,…, X_k. X_1 thông thường đặt là 1 để có thể bao gộp một số hạng tung độ gốc không đổi. Như trước đây, thủ tục OLS cực tiểu tổng bình phương sai số $\sum \hat{u}_t^2$ và cho ra k phương trình chuẩn. Những phương trình này nói chung duy nhất được giải cho các hệ số, với điều kiện là số quan sát lớn hơn k.

Ước lượng không thiên lệch của phương sai sai số (σ^2) được xác định bởi $s^2 = \hat{\sigma}^2 = (\sum \hat{u}_t^2)/(n-k)$. Với giả thiết rằng số hạng sai số u_t là phân phối độc lập và đồng nhất như $N(0, \sigma^2)$, trị thống kê $[(n-k)\hat{\sigma}^2]/\sigma^2$ có phân phối chi bình phương với n-k bậc tự do.

Độ thích hợp được đo lường theo một trong 2 cách tương đương. Từ phương trình ước lượng, phần dư được đo là $\hat{u}_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{t2} - ... - \hat{\beta}_k X_{tk}$. Tổng bình phương sai số (ESS) là $\sum \hat{u}_t^2$, và tổng bình phương toàn phần (TSS) là $\sum (Y_t - \bar{Y})^2$. Độ lệch chuẩn hồi qui được xác định bởi $\hat{\sigma} = [ESS/(n-k)]^{1/2}$ có thể so sánh với $\hat{\sigma}_Y = [TSS/(n-1)]^{1/2}$ để thấy độ biến giảm như thế nào. Một độ đo lường không tự do đơn vị được xác định bởi *bình phương R có hiệu chỉnh* (ký hiệu bằng $\bar{R}^2$), được tính như sau

$$\bar{R}^2 = 1 - \frac{ESS(n-1)}{TSS(n-k)} = 1 - \frac{n-1}{n-k}(1-R^2) = 1 - \frac{\hat{\sigma}^2(n-1)}{TSS}$$

$\bar{R}^2$ có thể được diễn giải là sự thay đổi của Y_t được giải thích bởi mô hình. Không giống R^2, bằng 1–(ESS/TSS), $\bar{R}^2$ có tính đến sự đánh đổi giữa sự tăng thêm của R^2 do biến được thêm vào và sự giảm đi trong các bậc tự do.

Trong chương này, ta cũng thảo luận 8 tiêu chuẩn khác nhau để chọn các mô hình tốt nhất. Một mô hình đơn giản hơn được ưa thích vì (1) sự gộp quá nhiều biến làm cho độ chính xác tương đối của các hệ số riêng lẻ kém đi (sẽ thấy chi tiết hơn trong chương kế tiếp), (2) Thêm các biến đồng nghĩa với giảm bậc tự do, làm cho khả năng kiểm định kém đi, và (3) một mô hình đơn giản hơn thì dễ hiểu hơn một mô hình phức tạp. Tiêu chuẩn chọn lựa mô hình có dạng của tổng bình phương sai số nhân với hệ số bất lợi, hệ số này phụ thuộc vào tính phức tạp của mô hình. Một mô hình được đánh giá là tốt hơn nếu các trị thống kê tiêu chuẩn trong phần lớn các đặc trưng có giá trị thấp hơn. Tuy nhiên, trong vài trường hợp đặc biệt nào đó, một hay vài tiêu chuẩn trở nên không cần thiết.

Để kiểm định một hệ số riêng lẻ (β) khác không một cách ý nghĩa hay không, trước tiên ta tính thống kê t (t_c), là tỷ số của hệ số ước lượng với độ lệch chuẩn ước lượng. Nếu $|t_c| > t^*_{n-k}(\alpha/2)$, với t^* là điểm trong phân phối t với bậc tự do n-k theo đó xác suất để $t > t^*$ bằng một nửa của mức ý nghĩa α, thì giả thuyết không H_0: $\beta = 0$ bị bác bỏ và giả thuyết H_1: $\beta \neq 0$ được củng cố. Nếu giả thuyết củng cố được kiểm định một phía, ta thu được t^* mà vùng bên phải của giá trị này bằng với mức ý nghĩa. Vậy ta bác bỏ H_0 và chấp nhận $\beta > 0$ nếu $t_c > t^*$ hoặc $\beta < 0$ nếu $t_c < -t^*$.

Để áp dụng phương pháp p-value, trước tiên tính toán 2 lần vùng bên phải của $|t_c|$ trong phân phối t với bậc tự do n-k. Bác bỏ H_0 nếu giá trị p nhỏ hơn mức ý nghĩa, và kết luận rằng hệ số có ý nghĩa.

Để kiểm định bộ hệ số hồi qui có bằng không hay không, phải thực hiện kiểm định F-test, còn được gọi là kiểm định Wald. Cụ thể hơn, để kiểm định H_0: $\beta_{m+1} = \beta_{m+2} = \ldots = \beta_k = 0$ đối lại giả thuyết rằng có ít nhất một hệ số khác không, trước tiên ta ước lượng mô hình không giới hạn (U):

$$(U) \qquad Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \beta_{m+1} X_{m+1} + \ldots + \beta_k X_k + u$$

Tiếp theo ta bỏ k-m biến cuối cùng và ước lượng mô hình giới hạn (R):

$$(R) \qquad Y = \beta_1 + \beta_2 X_2 + \ldots + \beta_m X_m + v$$

Kế đến ta tính trị thống kê F Wald

$$F_c = \frac{(ESS_R - ESS_U)/(k-m)}{ESS_U/(n-k)} = \frac{(R^2_U - R^2_R)/(k-m)}{(1-R^2_U)/(n-k)}$$

Trong đó R^2 là độ thích hợp chưa hiệu chỉnh. Giả thuyết không bị bác bỏ nếu $F_c > F^*_{k-m,\, n-k}(\alpha)$, trong đó F^* là điểm trong phân phối F với k-m và n-k bậc tự do theo đó xác suất để $F > F^*$ là α (ví dụ, 0,05 hoặc 0,01). Kiểm định Wald không cần thực hiện nếu chỉ có một hệ số hồi qui bị bỏ ra khỏi mô hình. Lý do vì một kiểm định t-test trên hệ số tương ứng là tương đương.

Trị thống kê kiểm định Wald cho độ thích hợp tổng quát được xác định như sau

$$F_c = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

có phân phối F với bậc tự do k-1 và n-k.

Kiểm định tổ hợp tuyến tính của các hệ số hồi qui có thể thực hiện theo 3 cách tương đương. Thống kê t dựa trên tổ hợp tuyến tính của các ước lượng có bậc tự do n-k và có thể dùng trong kiểm định t tương tự như dựa trên hệ số hồi qui riêng lẻ. Hoặc tổ hợp tuyến tính có thể được sáp nhập vào mô hình và thực hiện kiểm định t hoặc F-test.

Khoảng tin cậy cho các hệ số riêng lẻ tương tự như những điều rút ra từ Chương 3. Khoảng tin cậy cho dự báo $\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k$ dễ dàng có được bằng các ước lượng một mô hình có sửa đổi nhỏ.

Nên tránh "khai thác dữ liệu" không cẩn thận để tìm "độ thích hợp tốt nhất" bởi vì điều này thường dẫn đến sự chứng minh của bất kỳ giả thuyết nào mà ta nghĩ đến tuy nhiên những chứng minh như vậy có thể trái ngược. Không nên áp dụng mù quáng những chỉ tiêu cứng nhắc mà không xét đến lý thuyết hoặc sự hiểu biết của hành vi cơ bản.

Hệ quả của việc đưa vào một biến không liên quan (nghĩa là biến có hệ số hồi qui bằng không) như sau:

1. Các hệ số hồi qui ước lượng do dùng mô hình sai và những dự báo dựa trên các ước lượng này là không thiên lệch và nhất quán.
2. Những ước lượng là không hiệu quả và không phải ước lượng không thiên lệch tuyến tính tốt nhất (BLUE) bởi vì ước lượng dựa trên mô hình đúng là BLUE.
3. Những kiểm định của các giả thuyết vẫn hợp lệ bởi vì các phương sai ước lượng cũng không thiên lệch. Tuy nhiên, khả năng kiểm định bị giảm. Nói cách khác, khả năng chấp nhận một giả thuyết sai lầm (sai lầm loại II) là cao hơn khi dùng mô hình sai.

Hệ quả của việc loại bỏ biến đáng ra thuộc về mô hình là:

1. Các hệ số hồi qui ước lượng do dùng mô hình sai và các dự báo dựa trên các ước lượng này là thiên lệch và không nhất quán
2. Phương sai ước lượng cũng thiên lệch, và do đó các kiểm định của các giả thuyết không còn hợp lệ.

So sánh những hệ quả theo lý thuyết giữa việc thêm biến không liên quan với việc loại bỏ một biến quan trọng, ta quan sát thấy có sự đánh đổi. Sai số đặc trưng của việc thêm biến vào làm cho các ước lượng không hiệu quả, cho dù là không thiên lệch. Dạng sai số của việc bỏ biến ra làm cho các ước lượng và các kiểm định các giả thuyết thiên lệch. Bởi vì chưa thể biết mối quan hệ thực, ta lâm vào tình thế khó khăn để chọn công thức thích hợp nhất. Một nhà điều tra nghiên cứu cho rằng tính không thiên lệch, tính thích hợp và tin cậy của các kiểm định là quan trọng thì sẽ giữ một biến không liên quan hơn là nhận hậu quả của việc loại bỏ một biến quan trọng. Ngược lại, nếu một nhà nghiên cứu không thể chấp nhận các ước lượng không hiệu quả, thì sẽ thích loại bỏ các biến không liên quan hơn. Lý thuyết kinh tế và sự hiểu biết hành vi cơ bản thường giúp ích trong tình thế khó khăn như vầy. Tiêu chuẩn lựa chọn mô hình được thảo luận trước đây cũng có thể giúp ích. Các kiểm định của các đặc trưng (Chương 6) cũng sẽ giúp ích.

Bởi vì số hạng không đổi bao gộp những hiệu ứng trung bình của các biến bị loại bỏ, nên nhìn chung không nên bỏ số hạng này ra khỏi đặc trưng, ngay cả khi nó rất không có ý nghĩa và / hoặc có dấu không như kỳ vọng.

Thuật ngữ

Adjusted R^2	Bình phương R có hiệu chỉnh
Akaike information criterion (AIC)	Tiêu chuẩn thông tin Akaike (AIC)
Data-based model simplification	Đơn giản hóa mô hình dựa trên dữ liệu
Finite prediction error (FPE)	Sai số dự báo hữu hạn (FPE)
F-test	Kiểm định F-test
Generalized cross validation (GCV)	Tính hợp lệ chéo suy rộng (GCV)
Hedonic price index	Chỉ số giá hưởng thụ
HQ criterion	Tiêu chuẩn HQ
Joint significance	Ý nghĩa liên kết
Nodel in deviation form	Mô hình ở dạng sai lệch
Multiple regression	Hồi qui bội
Omitted variable bias	Thiên lệch của biến bị loại bỏ
Restricted model	Mô hình giới hạn
R^2 adjusted for degrees of freedom	Bình phương R có hiệu chỉnh đối với các bậc tự do
Specification error	Sai số đặc trưng
Unrestricted model	Mô hình không giới hạn
Wald test	Kiểm định Wald

4.A PHỤ LỤC

Các Kết Quả Tính Toán Khác

4.A.1 Mô Hình Hồi Quy Ba Biến

Mô hình hồi quy 3 biến diễn tả mối quan hệ giữa biến phụ thuộc Y với một hằng số và hai biến độc lập X_2, X_3. Mô hình chính thức được cho như sau:

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t \qquad (4.A.1)$$

Lấy trung bình mỗi số hạng của mô hình, ta có được:

$$\overline{Y} = \beta_1 + \beta_2 \overline{X}_2 + \beta_3 \overline{X}_3 + \overline{u} \qquad (4.A.2)$$

Lấy hiệu số với mô hình (4.A.1), ta có được mô hình ở dạng sai lệch như sau:

$$y_t = \beta_2 X_{t2} + \beta_3 X_{t3} + e_t \qquad (4.A.3)$$

Trong đó $y_t = Y_t - \overline{Y}$, $x_{t2} = X_{t2} - \overline{X}_2$, $x_{t3} = X_{t3} - \overline{X}_3$, và $e_t = u_t - \overline{u}$. Các ký tự ở dạng chữ thường diễn tả giá trị sai lệch giữa biến với giá trị trung bình tương ứng của biến đó. Lợi điểm trong việc biểu diễn mô hình dưới dạng sai lệch là chỉ còn hai thông số cần được ước lượng (β_2 và β_3). Nếu $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$ là giá trị ước lượng của hệ số tương quan hồi qui, $\hat{\beta}_1$ được ước lượng như sau:

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 - \hat{\beta}_3 \overline{X}_3$$

giá trị ước lượng của số dư là

$$\hat{u}_t = Y_t - \hat{\beta}_1 - \hat{\beta}_2 X_{t2} - \hat{\beta}_3 X_{t3}$$

Nguyên tắc OLS sẽ làm cực tiểu hoá tổng bình phương sai số ESS = $\sum u_t^2$ theo $\hat{\beta}_1$, $\hat{\beta}_2$, và $\hat{\beta}_3$. Điều này tương đương với việc cực tiểu hoá (không chứng minh) $\sum \hat{e}_t^2 = \sum (y_t - \hat{\beta}_2 X_{t2} - \hat{\beta}_3 X_{t3})^2$. Cho đạo hàm từng phần theo $\hat{\beta}_2$, và $\hat{\beta}_3$ của đẳng thức trên bằng 0, dễ dàng chứng minh điều kiện trên trở thành

$$\sum x_{t2}\hat{e}_t = 0 = \sum x_{t2}(y_t - \hat{\beta}_2 x_{t2} - \hat{\beta}_3 x_{t3})$$

$$\sum x_{t3}\hat{e}_t = 0 = \sum x_{t3}(y_t - \hat{\beta}_2 x_{t2} - \hat{\beta}_3 x_{t3})$$

Kết quả trên dẫn đến hai phương trình như sau (bỏ qua chỉ số t nhỏ).

$$\hat{\beta}_2 \sum x_2^2 + \hat{\beta}_3 \sum x_2 x_3 = \sum y x_2 \qquad (4.A.4)$$

$$\hat{\beta}_2 \sum x_2 x_3 + \hat{\beta}_3 \sum x_3^2 = \sum y x_3 \qquad (4.A.5)$$

Dùng các ký hiệu đơn giản hơn, hai phương trình này có thể viết lại như sau:

$$\hat{\beta}_2 S_{22} + \hat{\beta}_3 S_{23} = S_{y2} \qquad (4.A.6)$$

$$\hat{\beta}_2 S_{23} + \hat{\beta}_3 S_{33} = S_{y3} \qquad (4.A.7)$$

Trong đó

$$S_{22} = \sum x_{t2}^2 = \sum (X_{t2} - \overline{X}_2)^2 \qquad (4.A.8)$$

$$S_{23} = \sum x_{t2} x_{t3} = \sum (X_{t2} - \overline{X}_2)(X_{t3} - \overline{X}_3) \qquad (4.A.9)$$

$$S_{33} = \sum x_{t3}^2 = \sum (X_{t3} - \overline{X}_3)^2 \qquad (4.A.10)$$

$$S_{y2} = \sum y_t x_{t2} = \sum (Y_t - \overline{Y})(X_{t2} - \overline{X}_2) \qquad (4.A.11)$$

$$S_{y3} = \sum y_t x_{t3} = \sum (Y_t - \overline{Y})(X_{t3} - \overline{X}_3) \qquad (4.A.12)$$

Lời giải cho phương trình (4.A.6) và (4.A.7) như sau

$$\hat{\beta}_2 = (S_{y2} S_{33} - S_{y3} S_{23})/\Delta \qquad (4.A.13)$$

$$\hat{\beta}_3 = (S_{y3} S_{22} - S_{y2} S_{23})/\Delta \qquad (4.A.14)$$

Với

$$\Delta = S_{22} S_{33} - S_{23}^2 \qquad (4.A.15)$$

Cách tính phương sai của $\hat{\beta}_s$ được trình bày ở phụ lục 5.A.

4.A.2 Độ Thiên Lệch Do Việc Bỏ Qua Một Số Biến Liên Quan

Mô hình đúng và ước lượng được cho như sau

Mô hình đúng: $Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t$

Mô hình ước lượng: $Y_t = \beta_1 + \beta_2 X_{t2} + v_t$

Các giá trị ước lượng theo phương pháp OLS đối với những thông số trong mô hình ước lượng được cho như sau (xem phương trình 3.9 và 3.10)

$$\hat{\beta}_2 = S_{y2} / S_{22} \text{ và } \hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2 \qquad (4.A.16)$$

Trong đó S_{y2} và S_{22} được định nghĩa theo phương trình (4.A.11) và (4.A.8). Giá trị kỳ vọng của $\hat{\beta}_2$ được cho bởi $E(S_{y2}) / S_{22}$ vì S_{22} là không ngẫu nhiên:

$$S_{y2} = \sum (Y_t - \overline{Y})(X_{t2} - \overline{X}_2) = \sum Y_t (X_{t2} - \overline{X}_2) - \sum \overline{Y}(X_{t2} - \overline{X}_2)$$

$$= \sum Y_t (X_{t2} - \overline{X}_2)$$

Vì giá trị $\overline{Y}$ có thể rút ra được từ phép tính tổng và $\sum (X_{t2} - \overline{X}_2) = 0$ theo tính chất 2.A.4. Thay thế Y_t từ mô hình đúng (vì đó là quá trình đúng để tạo ra Y_t):

$$S_{y2} = \sum (X_{t2} - \overline{X}_2)(\beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + u_t)$$

$$= 0 + \beta_2 \sum (X_{t2} - \overline{X}_2) X_{t2} + \beta_3 \sum (X_{t2} - \overline{X}_2) X_{t3} + \sum (X_{t2} - \overline{X}_2) u_t$$

Số hạng đầu tiên bằng zero rút ra từ tính chất 2.A.4. Số hạng thứ hai như sau:

$$\sum (X_{t2} - \overline{X}_2) X_{t2} = \sum (X_{t2} - \overline{X}_2)(X_{t2} - \overline{X}_2 + \overline{X}_2)$$

$$= \sum (X_{t2} - \overline{X}_2)^2 + \overline{X}_2 (X_{t2} - \overline{X}_2) = \sum (X_{t2} - \overline{X}_2)^2$$

Vì số hạng thứ hai bằng zero , và theo cách tính tương tự, ta có:

$$\sum (X_{t2} - \overline{X}_2) X_{t3} = \sum (X_{t2} - \overline{X}_2)(X_{t3} - \overline{X}_3)$$

Sử dụng các kết quả này, ta có được:

$$S_{y2} = \beta_2 \sum (X_{t2} - \overline{X}_2)^2 + \beta_3 \sum (X_{t2} - \overline{X}_2)(X_{t3} - \overline{X}_3) + \sum (X_{t2} - \overline{X}_2) u_t$$

$$= \beta_2 S_{22} + \beta_3 S_{23} + S_{u2}$$

Trong đó, việc ký hiệu đối với các số hạng S cũng tương tự như những số hạng cho trong phương trình (4.A.8) cho đến phương trình (4.A.12). Vì X_2 và X_3 là không ngẫu nhiên và không tương quan với u và vì E(u) = 0 nên ta có:

$$E(S_{y2}) = \beta_2 S_{22} + \beta_3 S_{23} + E(S_{u2}) = \beta_2 S_{22} + \beta_3 S_{23}$$

Theo sau đẳng thức trên, ta có:

$$E(\hat{\beta}_2) = \beta_2 + \beta_3 \left[\frac{S_{23}}{S_{22}} \right]$$

Vì $\beta_3 \neq 0$ nên $\hat{\beta}_2$ sẽ có sai số trừ khi $S_{23} = 0$ – nghĩa là trừ khi X_2 và X_3 không tương quan nhau. Điều này chứng minh cho phương trình 4.4a được sử dụng trong các mô hình ở đây. Độ sai số của các biến bị bỏ qua được cho bằng $\beta_3 (S_{23}/S_{22})$. Hướng của độ thiên lệch phụ thuộc vào giá trị âm hay dương của β_3 cũng như sự tương quan giữa X_2 và X_3 là thuận hay nghịch. Vì cỡ mẫu tăng lên một cách không xác định nên β_2 sẽ không hội tụ về β_2 (nếu $S_{23} \neq 0$), và do đó giá trị ước lượng có được sẽ không nhất quán.

Từ phương trình (4.A.16), ta có $\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}_2$, và do đó $E(\hat{\beta}_1) = E(\overline{Y}) - \overline{X}_2 E(\hat{\beta}_2)$. Vì $\overline{Y} = \beta_1 + \beta_2 \overline{X}_2 + \beta_3 \overline{X}_3 + \overline{u}$, nên suy ra $E(\overline{Y}) = \beta_1 + \beta_2 \overline{X}_2 + \beta_3 \overline{X}_3$. Thế giá trị kỳ vọng này và giá trị kỳ vọng của $\hat{\beta}_2$ vào đẳng thức trên, ta có:

$$E(\hat{\beta}_1) = \beta_1 + \beta_2 \overline{X}_2 + \beta_3 \overline{X}_3 - \overline{X}_2 \left[\beta_2 + \beta_3 \frac{S_{23}}{S_{22}} \right]$$

$$= \beta_1 + \beta_3 \left[\overline{X}_3 - \overline{X}_2 \frac{S_{23}}{S_{22}} \right]$$

Lưu ý rằng điều kiện cần và đủ cho $\hat{\beta}_1$ không bị thiên lệch là $\left[\overline{X}_3 - \overline{X}_2 \frac{S_{23}}{S_{22}} \right] = 0$. Điều kiện hai biến X_2 và X_3 không tương quan nhau cũng không đủ để bảo đảm cho giá trị ước lượng của số hạng tung độ gốc không bị thiên lệch. Ngoài ra, giá trị trung bình của X_3 phải bằng zero. Từ các giá trị ước lượng của $\hat{\beta}_1$ và $\hat{\beta}_2$, có thể nhận thấy rằng các giá trị này cũng chịu một phần các ảnh hưởng do việc loại bỏ biến X3. Điểm nhận xét này có ý nghĩa rất quan trọng và nên được nhấn mạnh. Do hệ quả này mà giá trị số học của hệ số tương quan hồi qui có thể khác so với những phát biểu trước đây. Điều này chỉ ra rằng các vấn đề đặt ra cho hệ số tương quan không chỉ là các tác động trực tiếp của biến tương ứng mà còn là các tác động của những biến bị lược bỏ nhưng có tương quan với các biến đang xem xét.

Tác giả Kamenta (1986, p. 394) đã chứng minh rằng ngay cả khi $S_{23} = 0$ thì giá trị phương sai ước lượng của $\hat{\beta}_2 (s^2_{\hat{\beta}_2})$ cũng bị thiên lệch theo phía dương. Điều này có nghĩa là $E(s^2_{\hat{\beta}_2})$ = Var $(\hat{\beta}_2)$ + Q, trong đó giá trị Q là không âm. Vì thế mà những kiểm tra giả định thông thường sẽ không đem lại kết quả. Hậu quả cho việc lược bỏ một biến có liên quan là khá nghiêm trọng.

4.A.3 Chứng Minh Tính Chất 4.4

Mô hình ước lượng được cho như sau

$$Y_t = \beta_1 + \beta_2 X_{t2} + \beta_3 X_{t3} + v_t$$

Từ phương trình (4.A.13) và (4.A.14) – được đề cập lại cùng với (4.A.15) – giá trị ước lượng của β_2 và β_3 theo phương pháp OLS là:

$$\hat{\beta}_2 = (S_{y2} S_{33} - S_{y3} S_{23}) / \Delta \qquad (4.A.13)$$

$$\hat{\beta}_3 = (S_{y3}S_{22} - S_{y2}S_{23}) / \Delta \quad (4.A.14)$$

Trong đó

$$\Delta = S_{22}S_{33} - S_{23}^2 \quad (4.A.15)$$

Để kiểm tra xem giá trị $\hat{\beta}_2$ có bị thiên lệch hay không, ta cần có các giá trị kỳ vọng *đúng* của S_{y2} và S_{y3}. Mô hình đúng như sau (dưới dạng độ lệch):

$$y_t = \hat{\beta}_2 x_{t2} + u_t - \bar{u}$$

Thế giá trị y_t từ mô hình đúng vào trong S_{y2}, ta có:

$$S_{y2} = \sum y_t x_{t2} = \sum x_{t2}(\beta_2 x_{t2} + u_t - \bar{u}) = \beta_2 S_{22} + S_{u2}$$

$$E(S_{y2}) = \beta_2 S_{22}$$

vì x_{t2} là không ngẫu nhiên hoặc cho trước và $E(S_{u2}) = 0$. Mô hình đúng phải được sử dụng vì y_t được phát ra bởi mô hình đúng này chứ không phải bởi phương trình ước lượng. Tương tự, ta có:

$$S_{y3} = \sum y_t x_{t3} = \sum x_{t3}(\beta_2 x_{t2} + u_t - \bar{u}) = \beta_2 S_{23} + S_{u3}$$

$$E(S_{y3}) = \beta_2 S_{23}$$

Lấy giá trị kỳ vọng của phương trình (4.A.13) và (4.A.14) và thế vào $E(Sy_2)$ và $E(Sy_3)$, ta có được:

$$E(\hat{\beta}_2) = [S_{33}\beta_2 S_{22} - S_{23}\beta_2 S_{23}] / \Delta = \beta_2$$

$$E(\hat{\beta}_3) = [S_{22}\beta_2 S_{23} - S_{23}\beta_2 S_{22}] / \Delta = 0$$

Suy ra, giá trị $\hat{\beta}_2$ không bị thiên lệch và giá trị kỳ vọng của $\hat{\beta}_3$ sẽ bằng không. Đó cũng là kết quả của tích chất 4.5a. Theo nguyên tắc luật số đông thì tính chất nhất quán dễ dàng được thiết lập.

Tính toán phương sai của $\hat{\beta}_2$

Bước tiếp theo là tính toán giá trị phương sai của $\hat{\beta}_2$. Ta có:

$$\text{Var}\ (S_{y2}) = \text{Var}\ (\beta_2 S_{22} + S_{u2}) = \text{Var}\ (S_{u2}) = \sigma^2 S_{22}$$

$$\text{Var}\ (S_{y3}) = \text{Var}\ (\beta_2 S_{23} + S_{u3}) = \text{Var}\ (S_{u3}) = \sigma^2 S_{33}$$

$$\text{Cov}\ (S_{y2}, S_{y3}) = \text{Cov}\ (\beta_2 S_{22} + S_{u2}, \beta_2 S_{23} + S_{u3}) = \sigma^2 S_{23}$$

Trong việc đạo hàm các vế trên, ta đã sử dụng tính chất không ngẫu nhiên của biến S_{22} và S_{33}. Áp dụng tính chất 2.4a, ta có

$$\text{Var}(\hat{\beta}_2) = [S_{33}^2 \text{Var}(S_{y2}) + S_{23}^2 \text{Var}(S_{y3}) - 2S_{33}S_{23}\text{Cov}(S_{y2}, S_{y3})] / \Delta^2$$

$$= \sigma^2[S_{33}^2 S_{22} + S_{23}^2 S_{33} - 2S_{33}S_{23}S_{23}]/\Delta^2$$

$$= \frac{\sigma^2 S_{33}}{S_{22}S_{33} - S_{23}^2} = \frac{\sigma^2}{S_{22} - (S_{23}^2 / S_{33})}$$

Do $r^2 = S_{23}^2 / (S_{22}S_{33})$ (r^2 là bình phương của giá trị tương quan đơn giữa biến x_2 và x_3), phương trình trên có thể rút gọn lại như sau:

$$\text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{S_{22}(1 - r^2)}$$

Từ đây suy ra được phát biểu về giá trị phương sai của $\hat{\beta}_2$.