CHƯƠNG 5: LỰA CHỌN MÔ HÌNH VÀ VẤN ĐỀ KIỂM ĐỊNH

Trên thực tế, việc lập mô hình và ước lượng không phải là một vấn đề đơn giản. Chẳng hạn như trong ví dụ 4.2 về nhu cầu đầu tư ở Mỹ (1968 – 82). Cho dù lý thuyết kinh tế vĩ mô đã gợi ý rằng, cầu về đầu tư chịu ảnh hưởng bởi hai yếu tố chính là GNP và lãi suất. Tuy nhiên, việc Ngân hàng trung ương Mỹ sử dụng chính sách tiền tệ chặt trong thời kỳ đó đã đòi hỏi ta phải đưa thêm biến xu thế vào mô hình để giải thích cho cầu về đầu tư. Việc thêm hoặc bớt biến giải thích như vậy làm nẩy sinh một loạt các câu hỏi: Liệu ta nên thêm hoặc bớt những biến nào trong phương trình hồi quy? Chẳng hạn, liệu việc chỉ đưa thêm biến xu thế vào mô hình như vậy là đã đủ chưa? Hay cần phải đưa thêm nhiều biến giải thích khác nữa, như tỷ lệ lạm phát, số lượng quân nhân giải ngũ, vân vân? Trong rất nhiều sự lựa chọn như vậy, mô hình nào là tốt nhất? Và dựa trên tiêu chuẩn đánh giá nào? Ngược lại, nếu giả sử ta áp dụng một cách máy móc lý thuyết ghi trong sách giáo khoa, và bỏ quên, không đưa biến xu thế vào mô hình, thì hậu quả gì sẽ xẩy ra cho ước lượng và dự báo? Đó là những câu hỏi chúng ta muốn trả lời trong chương này.

5.1 Phân tích kết quả hồi quy

Chúng ta hãy bắt đầu bằng ví dụ phân tích một kết quả hồi quy đưa ra trong Ramanathan (1989):

Ví dụ 5.1: Một công ty bất động sản nghiên cứu giá các căn hộ cho những gia đình trẻ. Họ lập mô hình hồi quy như sau:

$$PRICE = \beta_1 + \beta_2 SQFT + \beta_3 BEDRMS + \beta_4 BATHS + \varepsilon \qquad (5.1)$$

Ở đó, *PRICE* là giá căn hộ tính theo nghìn dollars; bên cạnh diện tích sử dụng *SQFT*, (tính theo đơn vị tương tự như mét vuông), giá căn hộ còn chịu ảnh hưởng bởi số lượng phòng ngủ *BEDRMS*, và số nhà tắm *BATHS*. Vì đây đều là các đặc trưng về tính tốt của căn hộ, ta kỳ vọng rằng các hệ số $\beta_2, \beta_3, \beta_4$ đều dương.

Một trong ích lợi cơ bản của phương pháp hồi quy đa biến là nó cho phép đánh giá **tác động riêng phần** của từng yếu tố giải thích lên biến được giải thích. Chẳng hạn, nếu ta có hai căn hộ giống hệt nhau về diện tích sử dụng (*SQFT*) và số nhà tắm (*BATHS*). Nhưng chúng khác nhau về số phòng ngủ (*BEDRMS*). Khi đó, hệ số ước lượng $\hat{\beta}_3$ sẽ cho phép

chúng ta đánh giá liệu giá căn hộ có thêm một phòng ngủ sẽ đắt hơn là bao nhiêu so với căn hộ còn lại.

Để làm những so sánh đó, ta cần tiến hành ước lượng mô hình hồi quy (5.1). Dữ liệu điều tra cho việc ước lượng được ghi ở bảng 5.1 dưới đây:

Bảng 5.1: Dữ liệu điều tra về giá cả các căn hộ

obs	PRICE	CONSTANT	SQFT	BEDRMS	BATHS
	Y	X1	X2	X3	X4
1	199.9	1	1065	3	1.75
2	228	1	1254	3	2
3	235	1	1300	3	2
4	285	1	1577	4	2.5
5	239	1	1600	3	2
6	293	1	1750	4	2
7	285	1	1800	4	2.75
8	365	1	1870	4	2
9	295	1	1935	4	2.5
10	290	1	1948	4	2
11	385	1	2254	4	3
12	505	1	2600	3	2.5
13	425	1	2800	4	3
14	415	1	3000	4	3

Sau đây là kết quả ước lượng mô hình hồi quy mô hình (5.1):

$$PRICE = 129.062 + 0.1548 SQFT - 21.588 BEDRMS - 12.193 BATHS$$

Điều chúng ta nhận thấy ngay là dấu của các hệ số đi kèm với *BEDRMS* và *BATHS* là không giống với kỳ vọng. Thông thường, ta sẽ nghĩ rằng, nếu tăng thêm số lượng phòng ngủ hoặc nhà tắm, thì giá trị căn hộ phải đắt lên. Liệu kết quả ước lượng trên đây có phải là một điều bất hợp lý hay không?

Nhìn kỹ hơn, chúng ta vẫn có thể tìm được một cách diễn giải hợp lý, nếu xét đến **tác động riêng phần** của từng biến giải thích lên giá cả. Giả sử ta giữ nguyên diện tích sử dụng (*SQFT*) và số lượng phòng tắm (*BATHS*). Kết quả ước lượng nói lên rằng, nếu tăng thêm một phòng ngủ, thì về trung bình, giá của căn hộ sẽ giảm đi là 21,588 (21 nghìn 588) dollars. Vấn đề là, cũng vẫn cùng một diện tích sử dụng như vậy, nhưng nay bị chia nhỏ ra để có thêm phòng ngủ. Do vậy, từng phòng ngủ sẽ trở nên chật trội hơn. Và người tiêu dùng không thích việc làm như vậy. Họ chỉ sẵn sàng chi trả ở mức thấp hơn.

Tương tự như vậy, nếu số lượng nhà tắm tăng thêm một, mà diện tích và số phòng ngủ vẫn giữ nguyên, thì giá trị căn hộ sẽ giảm đi là 12,193 (12 nghìn 193) dollars.

Những phân tích trên đây về tác động riêng phần của các nhân tố cho thấy, những điều mà xem ra có vẻ là không hợp lý, thì bây giờ lại là có lý.

Bây giờ nếu giả sử chúng ta đồng thời tăng thêm một phòng ngủ và diện tích sử dụng lên 300. Khi đó, tác động đồng thời của những thay đổi đó lên giá cả sẽ là:

$$\begin{aligned}\Delta PRICE &= 0.1548\Delta SQFT - 21.588\Delta BEDRMS \\ &= 0.1548\times 300 - 21.588\times 1 = 24.852\end{aligned}$$

Nói khác đi, về trung bình, giá căn hộ sẽ tăng thêm là 24, 852 (24 nghìn 852) dollars.

Chúng ta cũng có thể tiến hành dự báo cho giá của một căn hộ, chẳng hạn có 4 phòng ngủ (*BEDRMS*), 3 nhà tắm (*BATHS*), với diện tích (*SQFT*) là 2500:

$$\begin{aligned}PRICE &= 129.062 + 0.1548\times 2500 - 21.588\times 4 - 12.193\times 3 \\ &= 391{,}163 \text{ (391 nghìn 163) dollars.}\end{aligned}$$

Như chúng ta thấy, kết quả dự báo là không tồi so với dữ liệu điều tra (rất gần với mẫu quan sát thứ11).

5.2 Lựa chọn mô hình

Bây giờ chúng ta hãy đưa thêm yếu tố tâm lý của người mua vào việc phân tích. Việc người tiêu dùng không thích căn hộ có phòng ngủ hoặc nhà tắm quá chật hẹp thể hiện rằng họ có những đòi hỏi về tiện nghi. Tức là họ yêu cầu phải có một sự phù hợp giữa diện tích sử dụng với số lượng phòng ngủ và phòng tắm trong căn hộ. Khi những đòi hỏi về tính phù hợp đó được chấp nhận bởi số đông, nó trở thành chuẩn mực chi phối cách thiết kế các căn hộ. Vì vậy, thông tin về diện tích có thể là đủ để cho người tiêu dùng đánh giá được giá trị của căn hộ. Điều đó đặt ra vấn đề là, ngoài mô hình đã xét, ta cần phải thử nghiệm nhiều mô hình khác nữa, và chọn ra đâu là cái tốt nhất.

Trong bảng 5.2 có 3 mô hình khác nhau. Mô hình C giống hệt như cái đã phân tích. Ta đưa thêm vào mô hình A và B, theo đó, mô hình A chỉ còn mỗi biến giải thích là diện tích (*SQFT*); trong khi mô hình B vẫn còn giữ lại số phòng ngủ (*BEDRMS*).

Ta quan tâm trước tiên tới độ phù hợp của từng mô hình với dữ liệu điều tra. Nhắc lại là từ chương 4, chúng ta đo mức độ phù hợp đó bởi quan hệ sau:

$$\sum_n (y_n - \bar{y})^2 = \sum_n (\hat{y}_n - \bar{y})^2 + \sum_n e_n^2$$

$$TSS = RSS + ESS$$

Bảng 5.2 đưa ra các con số so sánh giữa các mô hình. Nhìn từ A sang B và C, ta nhận thấy việc đưa thêm biến giải thích vào mô hình làm **tăng** mức độ giải thích của mô hình, thể hiện bởi tổng bình phương các sai số ước lượng (*ESS*) **giảm xuống**. Một cách trực quan, ta có thể lý giải việc *ESS* giảm như sau: Thay vì chỉ có yếu tố diện tích, việc đưa thêm những tính chất tốt khác của căn hộ vào (như số lượng phòng ngủ, nhà tắm, độ dịu của mầu vôi, độ thoáng gió, vân vân) sẽ làm cho việc diễn giải độ khác biệt của giá căn hộ so với trung bình sẽ tốt hơn lên. Vì vậy, việc tăng số biến giải thích trong mô hình luôn làm cho tổng bình phương sai số *ESS* giảm. Và vì vậy, hệ số đánh giá độ phù hợp của mô hình hồi quy là $R^2 = 1 - \frac{ESS}{TSS}$ **luôn luôn tăng**. [Xem hàng thứ nhất và thứ hai ở sau vạch ngang đầu tiên trong bảng 5.2].

Bảng 5.2: Những mô hình ước lượng cho giá các căn hộ

Variable	**model A**	**model B**	**model C**
Constant	52.351	121.179	129.062
	(38.28)	(80.17)	(88.3)
SQFT	0.13875 ***	0.14831 ***	0.1548 ***
	(0.018)	(0.021)	(0.031)
BEDRMS		-23.911	-21.588
		(24.64)	(27.029)
BATHS			-12.193
			(4.25)
ESS	18,274	16,833	16,700
R^2	0.821	0.835	0.836
$\bar{R}^2$	0.806	0.805	0.787
F-STAT	54.861	27.767	16.989
d.f (N-K)	12	11	10
AIC	1,737	1,846	2,112
SCHWAR	1,903	2,177	2,535

Chú thích: số trong ngoặc là standard error. * là ở mức ý nghĩa 0.1; ** là ở mức ý nghĩa 0.05; *** là ở mức ý nghĩa 0.001.

Tuy nhiên việc làm phức tạp hóa mô hình như vậy, nói chung là không được khuyến khích, bởi vì logic của việc lập mô hình là chỉ quan tâm đến việc đánh giá cái chính, chủ yếu, và lọc bỏ những cái không quan trọng ra khỏi phân tích. Ta không muốn đưa vào bức tranh phân tích tất cả mọi thứ trên đời, vì nó sẽ làm mờ đi yếu tố chính mà ta muốn đánh giá.

Về mặt kỹ thuật, việc đưa thêm các biến giải thích ít có ý nghĩa vào mô hình sẽ làm giảm mức độ chính xác của ước lượng, như chỉ ra vắn tắt dưới đây:

Như đã nêu, đi kèm với ước lượng tham số $\hat{\beta}_k$ là thống kê $t_k = \dfrac{\hat{\beta}_k - \beta_k}{\sqrt{s^2/S_{kk}}} \sim t(N-K)$, [tuân theo phân bố *t-student* với (*N-K*) bậc tự do].

Lưu ý là ở mẫu số của thống kê t_k, độ lớn của $s^2 = \dfrac{1}{N-K}\sum_n e_n^2 = \dfrac{ESS}{N-K}$ sẽ có ảnh hưởng trực tiếp tới giá trị của thống kê t_k. Việc tăng thêm số biến giải thích (*K* tăng) sẽ làm số bậc tự do (*N-K*) giảm, tức là làm s^2 có xu hướng bị đẩy lên. Ước lượng do vậy trở nên kém chính xác, vì sai số của ước lượng: $se(\hat{\beta}_k) = \sqrt{s^2 / S_{kk}}$ bị tăng lên. Hệ quả là, giá trị thống kê t_k sẽ trở nên nhỏ đi. Do đó, t_k dễ bị rơi vào vùng không bác bỏ giả thuyết ($DNRH_0$). Và ta dễ bị mắc phải sai lầm là chấp nhận một giả thuyết sai, mà đáng ra ta cần phải bác bỏ nó.

Nhìn chung, việc thêm biến giải thích vào mô hình có cái lợi là làm giảm tổng bình phương sai số, hay phần chưa được giải thích bởi mô hình, *ESS*. Nhưng cái thiệt là nó cũng làm giảm bậc tự do *(N-K)* [tức là làm cho việc phân tích có độ chính xác kém đi, như vừa nêu ở trên]. Nói một cách ẩn dụ, với việc đưa thêm các yếu tố mới vào mô hình, ta sẽ có cái nhìn đầy đủ hơn về mọi chi tiết, nhưng với cái giá là bức tranh không có điểm nhấn (thiếu *focus*). Chính vì vậy, thay vì sử dụng R^2, người ta thường dùng hệ số hiệu chỉnh của nó: $\bar{R}^2 = 1 - \dfrac{ESS/(N-K)}{TSS/(N-1)}$. Việc hiệu chỉnh như vậy là để tránh khuynh hướng đưa quá nhiều biến giải thích không cần thiết vào mô hình. Cụ thể là, nếu việc đưa thêm biến giải thích **có ý nghĩa** vào mô hình, thì phần lợi [tức là làm giảm *ESS*] phải vượt quá phần thiệt [tức là làm giảm bậc tự do *(N-K)*]. Khi đó, $\bar{R}^2$ tăng lên, thể hiện rằng đó là việc nên làm. Trong hoàn cảnh ngược lại, lợi không đủ bù phần mất mát, thì $\bar{R}^2$ bị giảm xuống, thể hiện rằng ta không nên đưa thêm biến giải thích đó vào mô hình, vì đấy là việc làm ít có ý nghĩa.

Ví dụ, trong bảng 5.2, dòng thứ 3, sau vạch ngang thứ nhất, ta thấy việc đưa thêm biến giải thích là số phòng ngủ và số nhà tắm vào làm giảm $\bar{R}^2$. Theo tiêu chuẩn này, mô hình tốt nhất sẽ là mô hình A: chỉ có duy nhất biến diện tích căn hộ (*SQFT*) là có ý nghĩa giải thích cho giá cả của căn hộ đó.

Người ta có thể chỉ ra rằng $\bar{R}^2$ không phạt đủ nặng việc đưa thêm các biến giải thích ít có ý nghĩa vào mô hình. Vì vậy, bên cạnh tiêu chuẩn đó, người ta còn sử dụng một số đánh giá khác, chẳng hạn như $AIC = \left(\dfrac{ESS}{N}\right)e^{2K/N}$ và $SCHWARZ = \left(\dfrac{ESS}{N}\right)N^{K/N}$. Nhìn chung, khi biến giải thích không có ý nghĩa được đưa vào mô hình, thì các tiêu chuẩn này bị đẩy lên. Vì vậy, mô hình lý tưởng nhất là mô hình có $\bar{R}^2$ cao hơn, và các tiêu chuẩn *AIC* và

SCHWARZ thấp hơn so với những mô hình khác. Ví dụ, trong bảng 5.2, mô hình A là tốt nhất theo mọi tiêu chuẩn đánh giá, bao gồm cả $\bar{R}^2$, *AIC* và *SCHWARZ*.

Trên thực tế, không phải bao giờ ta cũng may mắn như vậy. Rất có thể ta thấy một mô hình tốt hơn các cái còn lại về tiêu chuẩn này, nhưng lại tồi hơn về tiêu chuẩn khác. Khi đó, mô hình có nhiều tiêu chuẩn tốt nhất sẽ được lựa chọn.

5.3 Kiểm định các giả thuyết thống kê

Nhận xét vừa nêu cho thấy, việc chọn ra mô hình tốt nhất không phải lúc nào cũng thuyết phục cho lắm, nếu các tiêu chuẩn $\bar{R}^2$, *AIC* và *SCHWARZ* không đồng thời chỉ ra đâu là mô hình ưu việt nhất. Chính vì vậy, ta cần phải kiểm định lại xem quyết định của chúng ta có phù hợp về mặt thống kê hay không. Chẳng hạn, việc chọn mô hình A thay vì mô hình B hàm ý rằng, ta đã coi giả thuyết $H_0 : \beta_3 = 0$ là đúng. Trong khi việc loại bỏ mô hình C lại bao hàm rằng, ta coi giả thuyết đồng thời: $H_0 : \beta_3 = \beta_4 = 0$ là đúng. Việc kiểm định mức độ có ý nghĩa của từng tham số mô hình, như đã đề cập, được tiến hành bởi t-test. Trong khi đó, việc kiểm định giả thuyết đồng thời lại được thực hiện bởi Wald-test, như sẽ chỉ ra dưới đây.

Trong chương 4, chúng ta đã nói rằng, kiểm định t-test về mức độ có ý nghĩa của tham số ước lượng có thể được làm đơn giản bởi việc sử dụng *p-value*. Trong mô hình phân tích ở đây, ta thấy, chỉ có hệ số hồi quy của *SQFT* là **có ý nghĩa** giải thích trong cả 3 mô hình. Trong khi *p-value* của *BEDRMS* và *BATHS* trong cả hai mô hình B và C đều quá cao: $p - value > 0.05$. Tức là các hệ số hồi quy đi kèm với các biến giải thích này là **không có ý nghĩa** ở mức $\lambda = 5\%$. Vì vậy, xét một cách riêng biệt, ta nên loại từng biến này ra khỏi mô hình. Nhưng liệu ta có nên loại cả hai biến đó ra, và chỉ giữ lại duy nhất biến giải thích là diện tích căn hộ (*SQFT*) hay không? Điều đó dẫn đến vấn đề kiểm định giả dưới đây.

Việc loại bỏ cùng một lúc hai biến *BEDRMS* và *BATHS* là tương đương với việc chấp nhận giả thuyết đồng thời: $H_0 : \beta_3 = \beta_4 = 0$. Ta muốn nhấn mạnh rằng, giả thuyết đó là hoàn toàn **khác** với việc, cùng một lúc, xảy ra hai giả thuyết riêng biệt: $H_0 : \beta_3 = 0$ và $H_0 : \beta_4 = 0$. Ví dụ, nếu xét riêng biệt, từng yếu tố nhỏ như mầu vôi, cách bố cục căn bếp, nhà tắm, vân vân, có thể là không có ý nghĩa quyết định tới sự sẵn lòng chi trả của người đi mua nhà. Nhưng một cách **đồng thời**, chúng vẫn có thể ảnh hưởng tới cái giá mà người mua sẵn lòng bỏ ra. Nói khác đi, từng giả thuyết riêng biệt đúng, không có nghĩa là giả thuyết đồng thời cũng đúng.

Bây giờ, ta hãy xét xem làm thế nào để kiểm định giả thuyết đồng thời $H_0 : \beta_3 = \beta_4 = 0$. Hãy nhìn lại hai mô hình sau:

$$\text{(U):} \quad PRICE = \beta_1 + \beta_2 SQFT + \beta_3 BEDRMS + \beta_4 BATHS + \varepsilon \qquad (5.1)$$

(R): $PRICE = \beta_1 + \beta_2 SQFT + \varepsilon$ (5.2)

Mô hình (U) [tức là mô hình C trong bảng 5.2] được gọi là **mô hình không bị ràng buộc** (*unrestricted model*). Mô hình (R) [tức là mô hình A trong bảng 5.2] được gọi là **mô hình bị ràng buộc** (*restricted model*). Sở dĩ như vậy là vì mô hình (R) chính là mô hình (U), nhưng chịu ràng buộc là $H_0 : \beta_3 = \beta_4 = 0$. Việc lựa chọn xem mô hình nào là đúng, về thực chất quy về việc kiểm định giả thuyết kép sau:

$$H_0 : \begin{cases} \beta_2 = 0 \\ \beta_3 = 0 \end{cases} \quad \text{.vs.} \quad H_1 : \text{không phải là } H_0$$

Chúng ta đã nhận xét rằng, việc đưa thêm biến giải thích vào mô hình luôn làm tăng mức độ giải thích của mô hình, tức là làm giảm tổng bình phương sai số *ESS*. Vì vậy, ta luôn có: $ESS_R > ESS_U$. Trong đó, ESS_R là tổng bình phương sai số ước lượng của mô hình (R), và ESS_U là của mô hình (U).

Chúng ta nhận xét thêm rằng, nếu các biến *BEDRMS* và *BATHS* là không có ý nghĩa lắm cho việc giải thích giá căn hộ (*PRICE*), thì việc đưa chúng vào mô hình sẽ làm tổng bình phương sai số ước lượng giảm đi, nhưng **không nhiều**. Nói khác đi, nếu giả thuyết H_0 là đúng, thì hiệu $(ESS_R - ESS_U)$ là dương, nhưng với độ lớn không đáng kể. Ngược lại, nếu H_0 là sai, thì việc đưa thêm các biến *BEDRMS* và *BATHS* sẽ cải thiện đáng kể mức độ giải thích của mô hình. Do vậy, độ lệch $(ESS_R - ESS_U)$ sẽ rất lớn. Như vậy, chúng ta có thể đi đến nhận định rằng, khi hiệu $(ESS_R - ESS_U)$ là lớn, thì ta sẽ bác bỏ giả thuyết H_0 (RH_0). Tuy nhiên, như thế nào thì hiệu $(ESS_R - ESS_U)$ được coi là lớn? Điều đó dẫn đến việc lập thống kê *F*, mà ta sẽ trình bày dưới dạng tổng quát như sau. Xét hai lựa chọn về mô hình khác nhau:

(U): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon$ (5.3)

(R): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_{K-J} X_{K-J} + \varepsilon$ (5.4)

Mô hình (5.4) chính là mô hình (5.3), với *J* ràng buộc: $\beta_{K-J+1} = \beta_{K-J+2} = ... = \beta_K = 0$. Nói khác đi, ta muốn kiểm định giả thuyết đồng thời sau:

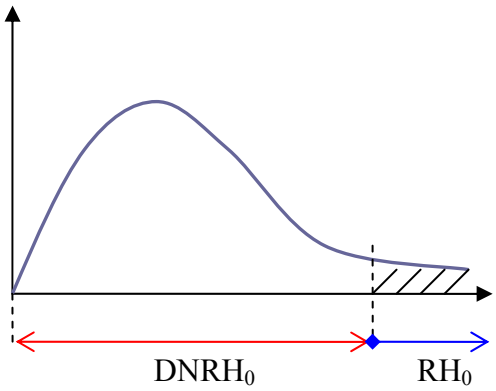
$$H_0 : \begin{cases} \beta_{K-J+1} = 0 \\ \beta_{K-J+2} = 0 \\ ..... \\ \beta_K = 0 \end{cases} \quad \text{.vs.} \quad H_1 : \text{không phải là } H_0$$

Người ta có thể chứng minh được rằng, đại lượng sau có phân bố F với J và $(N\text{-}K)$ bậc tự do:

$$F_c = \frac{(ESS_R - ESS_U)/J}{ESS_U/(N-K)} \sim F(J, N-K) \qquad (5.5)$$

Từ lập luận nêu trên, nếu F_c lớn hơn giá trị F-tra bảng: $F_c > F_\lambda(J, N-K)$, khi đó ta bác bỏ giả thuyết (RH_0). Ngược lại, nếu $F_c < F_\lambda(J, N-K)$ thì ta sẽ không bác bỏ giả thuyết đó $(DNRH_0)$.

Đồ thị 5.1: kiểm định giả thuyết với F-test.



Ví dụ 5.1 (tiếp theo): trong ví dụ về giá căn hộ, với việc chọn giữa mô hình (5.1) và (5.2), ta có $ESS_R = 18{,}274$, $ESS_U = 16{,}700$ [xem bảng 5.2], $J = 2, (N-K) = 10$. Vì vậy:

$$F_c = \frac{(18{,}274 - 16{,}700)/2}{16{,}700/10} = 0.471$$

Ta có thể tra bảng F-statistic: $F_{0.05}(2,10) = 4.1$. Vì vậy, ta có: $F_c < F_{0.05}(2,10)$. Tức là ta sẽ không bác bỏ giả thuyết H_0. Khi đó, mô hình với chỉ một biến giải thích là diện tích sử dụng *(SQFT)* được coi là mô hình đúng nhất theo kiểm định Wald –test.

5.4 Kiểm định tính có ý nghĩa của cả mô hình (overall significance test)

Một trường hợp đặc biệt của Wald test (hay F-test) vừa nêu trên là đánh giá hai mô hình sau:

(U): $Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon$ (5.6)

(R): $Y = \beta_1 + \varepsilon$ (5.7)

Trong mô hình bị ràng buộc (R), tất cả các biến giải thích, ngoại trừ hằng số (constant term), bị loại bỏ. Tức là chúng ta muốn kiểm định giả thuyết H_0:

$H_0 : \beta_2 = \beta_3 = ... = \beta_K = 0$ .vs. H_1 :không phải là H_0

Nói khác đi, ta muốn kiểm tra nhận định là: "không có bất cứ một biến giải thích nào trong mô hình, ngoại trừ *constant term*, là có ý nghĩa cả". Wald-test cho kiểm định như vậy có dạng đơn giản như sau:

$$F_c = \frac{R^2 / (K-1)}{(1-R^2)/(N-K)} \sim F(K-1, N-K)$$

Trong đó, R^2 là độ phù hợp của mô hình (5.6).

Nếu ta **không** bác bỏ giả thuyết H_0, thì không có biến giải thích nào, ngoại trừ *constant term* trong (5.6) là có ý nghĩa cả. Chúng ta có một mô hình tồi và cần phải xây dựng lại mô hình hồi quy.

Thông thường các *sofware* như eviews sẽ cho ra thông báo về việc kiểm định giả thuyết về tính có ý nghĩa chung của cả mô hình (*overall significance*). Giá trị của F_c, tính theo công thức (5.5), lúc này được gọi là *F-stat*. Đi kèm theo nó, eviews cũng cho ra *p-value* của *F-stat*.

Ví dụ 5.1 (tiếp theo): ứng với mô hình (C), máy tính sẽ kiểm định giả thuyết: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. Và cho ra thông báo *F-stat* = 16.98, [*p-value* =0.000]. Nhìn vào bảng 5.2, các giá trị của *F-stat* cho mô hình (A) và (B) lần lượt là 54.86 và 27.7. Tức là trong cả 3 mô hình, một cách đồng thời, các biến là có ý nghĩa cho việc giải thích những biến động của giá căn hộ *PRICE*.

5.5 Những ứng dụng khác của Wald test

Ứng dụng của Wald test là khá rộng và đa dạng hơn nhiều so với những ví dụ đã nêu ở trên. Nhưng nhìn chung, chúng có cùng chung một cách tiếp cận là so sánh độ tốt về mặt thống kê giữa hai dạng mô hình: bị ràng buộc và không bị ràng buộc. Chúng ta xem lại một số cải biên của ví dụ đơn giản về nhu cầu đầu tư ở Mỹ (1968 -82):

(U): $$INV = \beta_1 + \beta_2 T + \beta_3 G + \beta_4 INT + \beta_5 INF + \varepsilon \quad (5.8)$$

Mô hình này giả định rằng các nhà đầu tư nhậy cảm với lãi suất (*INT)* và lạm phát (*INF)*. Một giả định khác là các nhà đầu tư chỉ nhậy cảm với lãi suất thực. Mô hình biểu diễn sẽ là:

(R): $$INV = \beta_1 + \beta_2 T + \beta_3 G + \beta_4 (INT - INF) + \varepsilon \quad (5.9)$$

Chúng ta nhận xét rằng mô hình (5.9) là bị ràng buộc (restricted) so với mô hình (5.8) bởi giả định là: $H_0 : \beta_4 + \beta_5 = 0$. Hay cũng vậy, ta kiểm định:

$$H_0 : \beta_4 = -\beta_5 \quad \text{.vs.} \quad H_1 : \text{không phải là } H_0 \quad (5.10)$$

Các bước tiến hành kiểm định (Wald test) là như sau:

Bước 1: Xác định rõ đâu là mô hình bị ràng buộc (restricted model: R) , bằng cách nhận dạng yêu cầu cần kiểm định là gì, hay cũng vậy, giả thuyết H_0 bao gồm những ràng buộc gì.

Bước 2: Tiến hành chạy hồi quy mô hình không bị ràng buộc (U) và mô hình bị ràng buộc (R).

Bước 3: Tính thống kê F_c, sử dụng phương trình (5.5), với các bậc tự do *J* [là số các ràng buộc nêu bởi H_0] và (*N-K*).

Bước 4: Từ bảng thống kê *F*, tìm giá trị F-tra bảng [tức là tìm critical value: $F(J, N-K)_\lambda$]. Một cách khác nữa, ta có thể tính $p - value = \Pr ob(F(J, N-K) > F_c)$

Bước 5: Loại bỏ giả thuyết (RH_0) , nếu $F_c > F(J, N-K)_\lambda$, hoặc $p - value < \lambda$.

5.6 Lỗi lầm trong việc lập mô hình (Specification errors)

Chúng ta đã nêu lên dạng tổng quát của mô hình hồi quy như sau:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_K X_K + \varepsilon$$

Tuy nhiên, người lập mô hình có thể phạm phải rất nhiều loại sai lầm trong việc xác định dạng mô hình cụ thể. Có lẽ hai loại lỗi phổ biến nhất là: bỏ qua những biến có ý nghĩa, không đưa chúng vào mô hình; và ngược lại, đưa quá nhiều biến giải thích không có ý nghĩa vào mô hình. Lỗi sau cùng đã được bàn ở trên. Nó làm ước lượng trở nên mất chính xác. Mặc dù ước lượng vẫn là không chệch. Việc né tránh lỗi lầm đó, như đã nói, được thực hiện dựa trên xem xét các chỉ tiêu đo lường $\bar{R}^2$, *AIC*, và *SCHWARZ*, cũng như sử dụng các kiểm định F-test và t-test. Đối với lỗi lầm thứ nhất, việc phát hiện trở nên khó khăn hơn.

Chủ yếu là do người lập mô hình thường sử dụng cách tiếp cận máy móc, học từ sách giáo khoa (text-book approach), mà không có những phân tích thấu đáo về đối tượng nghiên cứu. Như trong ví dụ về đầu tư của Mỹ (1968-82), nếu chỉ dựa vào sách giáo khoa, người lập mô hình có thể sẽ bỏ quên, không đưa biến xu thế vào phân tích. Ở đây ta muốn nêu lên hậu quả khá tai hại của cách tiếp cận text-book đó là như thế nào.

Để đơn giản, chúng ta giả sử mô hình **đúng** là như sau:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \varepsilon \qquad (5.11)$$

Nhưng chúng ta phạm **sai lầm**, và chạy hồi quy mô hình sau:

$$Y = \beta_1 X_1 + \tilde{\varepsilon} \qquad (5.12)$$

Trong (5.12), ta bỏ quên X_2, nên về thực chất, hay về mặt tổng thể, $\tilde{\varepsilon} = \beta_2 X_2 + \varepsilon$.

Bây giờ, do sự lầm tưởng, ta chạy hồi quy (5.12), thay vì chạy mô hình đúng (5.11). Sử dụng phương trình (3.9), định lý 3.1, ta có ước lượng sau:

$$\hat{\beta}_1 = \beta_1 + \sum c_{1n} \tilde{\varepsilon}_n \qquad (5.13)$$

Thế giá trị $\tilde{\varepsilon} = \beta_2 X_2 + \varepsilon$ vào (5.13), và lấy kỳ vọng cả hai vế, ta có:

$$E\hat{\beta}_1 = E[\beta_1 + \sum c_{1n} \tilde{\varepsilon}_n]$$

$$= E[\beta_1 + \sum c_{1n}(\beta_2 x_{n2} + \varepsilon_n)]$$
$$= \beta_1 + \beta_2 \sum c_{1n} x_{n2} + \sum c_{n1} E\varepsilon_n$$

Trong đó, x_{n2} là quan sát thứ *n* của biến giải thích X_2. Sử dụng giả thuyết A1, ta có:

$$E\hat{\beta}_1 = \beta_1 + \beta_2 \sum c_{1n} x_{n2} \neq \beta_1 \qquad (5.14)$$

Phương trình (5.14) nói lên rằng, nhìn chung, việc bỏ quên biến giải thích có ý nghĩa sẽ làm ước lượng **bị chệch (**biased estimation**)**. Vì vậy, mọi kiểm định thống kê trở nên vô giá trị, và việc dự báo trở nên vô nghĩa. Đây có lẽ là lời cảnh tỉnh nghiêm khắc nhất với những nghiên cứu máy móc, dựa trên *text books*. Để tránh tình huống này, việc đánh giá thực tiễn kỹ lưỡng trước khi lập mô hình là một việc làm không thể thiếu.