Chương 7

BIẾN ĐỘC LẬP ĐỊNH TÍNH (HOẶC BIẾN GIẢ)

Tất cả các biến chúng ta gặp trước đây đều có bản chất định lượng; nghĩa là các biến này có các đặc tính có thể đo lường bằng số. Tuy nhiên, hành vi của các biến kinh tế cũng có thể phụ thuộc vào các nhân tố ***định tính*** *như giới tính, trình độ học vấn, mùa, công cộng hay cá nhân v.v… Lấy một ví dụ cụ thể, hãy xem xét mô hình hồi qui tuyến tính đơn sau (để đơn giản ta bỏ qua chữ t nhỏ):*

$$Y = \alpha + \beta X + u \qquad \textbf{(7.1)}$$

Gọi Y là mức tiêu thụ năng lượng trong một ngày và X là nhiệt độ trung bình. Khi nhiệt độ tăng trong mùa hè, chúng ta sẽ kỳ vọng mức tiêu thụ năng lượng sẽ tăng. Vì vậy, hệ số độ dốc β có khả năng là số dương. Tuy nhiên, trong mùa đông, khi nhiệt độ tăng ví dụ từ 20 đến 40 độ, năng lượng được dùng để sưởi ấm sẽ ít hơn, và mức tiêu thụ sẽ có vẻ giảm khi nhiệt độ tăng. Điều này cho thấy β có thể âm trong mùa đông. Vì vậy, bản chất của quan hệ giữa mức tiêu thụ năng lượng và nhiệt độ có thể được kỳ vọng là phụ thuộc vào biến định tính "mùa". Trong chương này, chúng ta sẽ khảo sát các thủ tục để xem xét các biến định tính trong ước lượng và kiểm định giả thuyết. Chúng ta chỉ tập trung chú ý vào các biến độc lập định tính. Chương 12 thảo luận trường hợp các biến phụ thuộc định tính.

⭘ 7.1 Các Biến Định Tính Chỉ Có Hai Lựa Chọn

Chúng ta bắt đầu với việc xem xét trường hợp đơn giản nhất trong đó một biến định tính chỉ có hai lựa chọn. Ví dụ, giữa hai ngôi nhà có cùng các đặc trưng, một có thể có hồ bơi trong khi ngôi nhà còn lại không có. Tương tự, giữa hai nhân viên của một công ty có cùng tuổi, học vấn, kinh nghiệm v.v… một người là nam và người kia là nữ. Câu hỏi quan trọng trong những ví dụ này là làm thế nào để đo lường tác động của giới tính đến lương và tác động của sự hiện diện của hồ bơi đến giá nhà. Để phát triển lý thuyết, chúng ta xem xét ví dụ về lương và đặt Y_t là tiền lương hàng tháng của nhân viên thứ *t* trong một công ty. Để đơn giản về mặt sư phạm, ở đây, chúng ta bỏ qua các biến khác có ảnh hưởng đến lương và chỉ tập trung vào giới tính. Vì biến giới tính không phải là một biến định lượng một cách trực tiếp được nên chúng ta định nghĩa một **biến giả** (gọi là D), biến giả này là **biến nhị nguyên** chỉ nhận giá trị 1 đối với nhân viên nam và giá trị 0 đối với nhân viên nữ. Chúng ta sẽ thấy sau này là các định nghĩa trên cũng tương đương với việc định nghĩa biến D bằng 1 đối với nữ nhân viên và bằng 0 đối với nam nhân viên. Do đó cách chọn này là hoàn toàn ngẫu nhiên. Nhóm mà giá trị D bằng 0 gọi là **nhóm điều khiển.** Bảng 7.1 có dữ liệu lương tháng và giá trị của D cho 49 nhân viên trong tập tin DATA6-4 mà chúng ta đã gặp trong chương trước. Lưu ý rằng, có 26 nam và 23 nữ. Lương tháng trung bình chung là $1.820,20. Tuy nhiên, nếu chúng ta chia nhân viên thành hai nhóm theo giới tính, lương trung bình của nam là $2.086,93 và $1.518,70 đối với nữ (hãy chứng minh). Có phải điều này nghĩa là có "phân biệt giới tính" trung bình lên đến $568,23 mỗi tháng? Câu trả lời rõ ràng là không vì chúng ta không kiểm soát được các biến khác như kinh nghiệm, học vấn, v.v… Có thể là nhân viên nữ trong mẫu này có số năm học tập và kinh

nghiệm ít hơn và do đó nhận được lương trung bình thấp hơn. Chúng ta có thể thử xác định nhân viên nữ với nhân viên nam có kinh nghiệm như nhau hoặc có học vấn như nhau và sau đó tính lương trung bình. Việc này không những khó khăn mà còn có thể không khả thi vì có thể có nhiều đặc điểm khác như dân tộc hoặc loại nghề mà chúng ta phải xem xét. Đây là phạm vi mà phân tích kinh tế lượng trở thành một công cụ rất hiệu quả. Chúng ta sẽ thiết lập và ước lượng một mô hình sử dụng biến giả như một biến giải thích. Dạng đơn giản nhất của mô hình như sau:

$$Y_t = \alpha + \beta D_t + u_t \qquad \textbf{(7.2)}$$

với mô hình không có một biến giải thích nào khác (được gọi là **mô hình phân tích phương sai**). Chúng ta sẽ dần dần mở rộng mô hình này, thêm vào các đặc điểm của nhân viên thay vì chỉ có giới tính. Chúng ta giả sử là số hạng sai số thay đổi ngẫu nhiên và thỏa mãn tất cả các giả thiết trong Chương 3. Chúng ta có thể lấy kỳ vọng có điều kiện của Y với D cho trước và được các phương trình sau

Nam: $E(Y_t|D = 1) = \alpha + \beta$
Nữ: $E(Y_t|D = 0) = \alpha$

Bảng 7.1 Dữ liệu chéo về lương tháng và giới tính

Y	*D*	*Y*	*D*	*Y*	*D*
1345	0	1566	0	2533	1
2435	1	1187	0	1602	0
1715	1	1345	0	1839	0
1461	1	1345	0	2218	1
1639	1	2167	1	1529	0
1345	0	1402	1	1461	1
1602	0	2115	1	3307	1
1144	0	2218	1	3833	1
1566	1	3575	1	1839	1
1496	1	1972	1	1461	0
1234	0	1234	0	1433	1
1345	0	1926	1	2115	0
1345	0	2165	0	1839	1
3389	1	2365	0	1288	1
1839	1	1345	0	1288	0
981	1	1839	0		
1345	0	2613	1		

Vậy, α là lương trung bình của nhóm điều khiển và β là khác biệt kỳ vọng của lương trung bình của hai nhóm, cho cả tổng thể.

Chúng ta đã thấy trong Chương 3 là các phương trình chuẩn để ước lượng Phương trình (7.2) đã cho như sau:

$$\sum Y_t = n\hat{\alpha} + \hat{\beta}\sum D_t \qquad \textbf{(7.3)}$$

$$\sum Y_t D_t = \hat{\alpha}\sum D_t + \hat{\beta}\sum D_t^2 = \hat{\alpha}\sum D_t + \hat{\beta}\sum D_t \qquad \textbf{(7.4)}$$

Lưu ý rằng do *D* là biến giả và chỉ nhận giá trị 1 và 0, D^2 cũng có giá trị giống *D*. Trong Phương trình (7.4), $\sum D_t$ ở vế bên phải bằng số nam nhân viên (gọi là n_m) và $\sum Y_t D_t$ ở vế bên trái bằng tổng lương của họ. Chia hai vế cho n_m ta có

$$\hat{\alpha} + \hat{\beta} = \bar{Y}_m \quad \textbf{(7.5)}$$

với $\bar{Y}_m$ là lương trung bình của nam nhân viên. Vì vậy, tổng các hệ số hồi qui là một ước lượng của $E(Y_t|D = 1)$, trung bình tổng thể lương của nam nhân viên.

Vì $\sum D_t = n_m$, Phương trình (7.3) và (7.4) có thể viết lại thành

$$\sum Y_t = n\hat{\alpha} + n_m\hat{\beta}$$
$$\sum Y_t D_t = n_m(\hat{\alpha} + \hat{\beta})$$

Lấy phương trình thứ nhất trừ phương trình thứ hai và bỏ đi những số hạng chung ở vế bên phải, ta có

$$\sum Y_t - \sum Y_t D_t = (n - n_m)\hat{\alpha} = n_f\hat{\alpha}$$

với n_f là số nhân viên nữ. Lưu ý là vế bên trái của phương trình đơn giản là tổng lương của nữ nhân viên (tổng của toàn bộ lương trừ tổng lương của nam nhân viên). Vì vậy, chia hai vế cho n_f, chúng ta có $\hat{\alpha} = \bar{Y}_f$, trung bình mẫu của lương nữ nhân viên, đây là một ước lượng của trung bình tổng thể $E(Y_t|D = 0)$.

Tóm lại, nếu chúng ta hồi qui Y_t theo một số hạng không đổi và biến giả D_t, tung độ gốc $\hat{\alpha}$ ước lượng lương trung bình của nữ nhân viên và hệ số độ dốc $\hat{\beta}$ ước lượng khác biệt giữa lương trung bình của nam nhân viên và nữ nhân viên. Từ Bài thực hành máy tính Phần 7.1 (xem Phụ lục Bảng D.1), chúng ta có các ước lượng hồi qui là $\hat{\alpha}$ = 1.518,70 và $\hat{\beta}$ = 568,23. Chúng ta thấy là phương pháp hồi qui tương tự như việc chúng ta chia mẫu thành hai nhóm nam và nữ và tính lương trung bình tương ứng. Tuy nhiên, như chúng ta sẽ thấy trong những phần sau, phương pháp hồi qui này mạnh hơn vì phương pháp này có thể ứng dụng ngay cả khi các nhân viên khác nhau về các đặc điểm khác như kinh nghiệm và học vấn.

⊙ BÀI TẬP THỰC HÀNH 7.1+

Giả sử biến giả đã được định nghĩa là $D^* = 1$ đối với nữ và bằng 0 đối với nam và biến D_t^* được dùng thay cho biến D_t. Nói cách khác, xét mô hình mới $Y_t = \alpha^* + \beta^* D_t^* + u_t$. Lưu ý là $D^* = 1 - D$, tính các tương quan đại số giữa các hệ số hồi qui mới và các hệ số hồi qui cũ. Cụ thể hơn, chỉ ra bằng cách nào ta có thể ước lượng α^* và β^* *mà không cần thực hiện hồi qui.* Các sai số chuẩn, giá trị *t*, R^2, ESS, và trị thống kê F có bị ảnh hưởng hay không? Nếu có, ảnh hưởng như thế nào?

Thêm Các Biến Độc Lập Định Lượng

Bước tiếp theo trong phân tích là thêm vào các biến độc lập có thể định lượng được. Để minh họa, đặt *Y* là lương tháng như trước nhưng ngoài biến giả *D* đã giới thiệu trước, ta đưa thêm biến kinh nghiệm (gọi là X) vào như một biến giải thích. Lưu ý là bây giờ chúng ta có thể kiểm soát được kinh nghiệm và có thể hỏi "Giữa hai nhân viên có cùng kinh nghiệm, có sự khác biệt do giới tính không?" Một cách đơn giản để trả lời câu hỏi này là đặt tung độ gốc α trong Phương trình (7.1) khác nhau đối với nam và nữ. Thực hiện việc này bằng cách giả sử là $\alpha = \alpha_1 + \alpha_2 D$. Với nữ, $D = 0$ và vì vậy $\alpha = \alpha_1$. Với nam, $D = 1$ và vì vậy $\alpha = \alpha_1 + \alpha_2$. Dễ dàng thấy là α_2 đo lường khác biệt trong tung độ gốc của hai nhóm. Thay thế giá trị của α vào Phương trình (7.1) ta có mô hình kinh tế lượng

$$Y = \alpha_1 + \alpha_2 D + \beta X + u \qquad \textbf{(7.6)}$$

Lưu ý là α_1, α_2 và β được ước lượng bằng cách hồi qui Y theo một hằng số, D, và X. Các quan hệ được ước lượng cho hai nhóm là

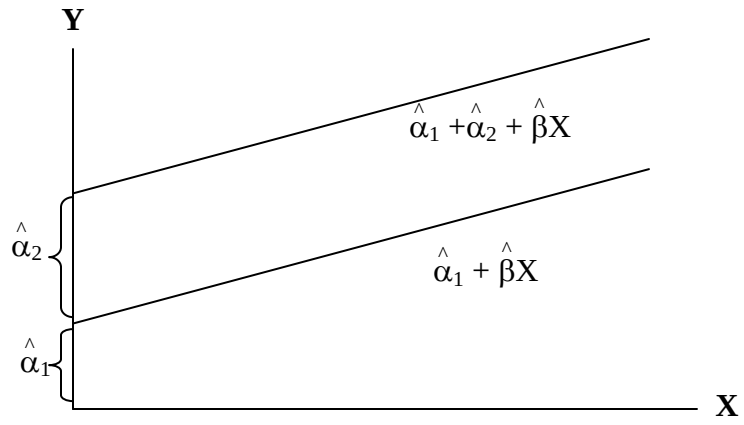
$$\text{Nữ:} \quad \hat{Y} = \hat{\alpha}_1 + \hat{\beta}X \qquad \textbf{(7.7)}$$

$$\text{Nam:} \quad \hat{Y} = (\hat{\alpha}_1 + \hat{\alpha}_2) + \hat{\beta}X \qquad \textbf{(7.8)}$$

Hình 7.1 vẽ các mối quan hệ này khi các α và β dương. Chúng ta lưu ý là các đường thẳng ước lượng song song với nhau. Đó là do chúng ta đã giả định là cả hai nhóm đều có cùng β. Giả thiết này được bỏ qua trong Phần 7.3.

Một giả thuyết tự nhiên cần kiểm định là "không có sự khác biệt trong quan hệ giữa hai nhóm". So sánh Phương trình (7.7) và (7.8), chúng ta thấy là các quan hệ sẽ như nhau nếu $\alpha_2 = 0$. Vì vậy, chúng ta có H_0: $\alpha_2 = 0$ và H_1 $\alpha_2 > 0$ hoặc $\alpha_2 \neq 0$. Kiểm định thích hợp là kiểm định t cho α_2 với bậc tự do là d.f. = $n - 3$.

⭘ Hình 7.1 Một Ví Dụ Về Dời Tung Độ Gốc Bằng Cách Sử Dung Một Biến Giả



⭘ VÍ DỤ 7.1

Sử dụng DATA7-2 mô tả trong Phụ lục D, chúng ta đã ước lượng Phương trình (7.6) như sau (các số trong dấu ngoặc là các giá trị p)

$$\widehat{\text{WAGE}} = 1.366{,}27 + 525{,}63D + 19{,}81\ \text{EXPER}$$
$$(<0{,}01) \qquad (0{,}003) \quad (0{,}152)$$

$$\bar{R}^2 = 0{,}197 \qquad n = 49 \qquad F_c(2, 46) = 6{,}90$$

Để tính lại được các kết quả này, hãy thực hiện Bài thực hành máy tính Phần 7.2. Giá trị p của biến giả là rất nhỏ, cho thấy mức ý nghĩa cao. Vì vậy, khi kiểm soát được biến kinh nghiệm, có sự khác biệt về lương trung bình có ý nghĩa theo giới tính. Giá trị p của kinh nghiệm cho thấy không có ý nghĩa ở mức 0,15. Tuy nhiên không nên nhìn vấn đề này quá nghiêm trọng bởi vì mô hình không

gồm các biến giải thích khác như trình độ học vấn và tuổi và do đó các ước lượng là thiên lệch. Biến phụ thuộc cũng vậy, nên được thể hiện dưới dạng logarít (xem lại Phần 6.8). Trong Phần 7.3, chúng ta trình bày một phân tích tổng quát hơn về các yếu tố quyết định của lương, bao gồm cả những tác động của một số biến định tính.

⭘ VÍ DỤ 7.2

DATA7-3 có dữ liệu bổ sung về 14 căn hộ gia đình đơn, tất cả các dữ liệu bổ sung đều là các biến giả. POOL nhận giá trị 1 nếu căn nhà có hồ bơi và giá trị 0 cho trường hợp ngược lại. Tương tự, FAMROOM đại diện cho việc căn nhà có phòng gia đình, và FIREPL là căn nhà có thiết bị báo cháy. Có người sẽ kỳ vọng là một căn nhà mà có những đặc trưng như vậy có lẽ giá sẽ cao hơn một căn nhà tương tự nhưng không có những đặc trưng này. Bảng 7.2 có các hệ số ước lượng và các trị thống kê liên quan của một số mô hình trong đó có Mô hình A, mà chúng ta đã ước lượng trước đây (các kết quả có thể tính lại bằng Bài thực hành máy tính Phần 7.3).

So sánh Mô hình A với Mô hình E là mô hình có tất cả các biến mới, chúng ta lưu ý là $\overline{R}^2$ tăng từ 0,806 lên 0,836, nhưng bốn trong số các tiêu chuẩn để lựa chọn mô hình lại xấu hơn. RICE thì không xác định được vì cần phải có số quan sát gấp đôi số hệ số được ước lượng, không phù hợp trong trường hợp này. Trị thống kê *t* của POOL là 2,411 có ý nghĩa ở mức thấp hơn 1 phần trăm. Tuy nhiên, các hệ số hồi qui của BEDRMS, BATHS, FAMROOM, và FIREPL không có ý nghĩa ở các mức ý nghĩa lớn hơn 25 phần trăm (hãy chứng minh). Trong Mô hình F những biến không có ý nghĩa này bị loại bỏ và mô hình được ước lượng lại. Sử dụng Mô hình E như là mô hình không giới hạn và Mô hình F là mô hình giới hạn, chúng ta thực hiện kiểm định Wald để kiểm định *giả thuyết không* là các hệ số hồi qui của BEDRMS, BATHS, FAMROOM và FIREPL bằng không. Trị thống kê F được tính bằng

$$F_c = \frac{(9.455 - 9.010) \div 4}{9.010 \div 7} = 0{,}086$$

giá trị này có phân phối *F* với bậc tự do là 4 và 7. Dễ thấy là F_c không có ý nghĩa ngay cả ở mức ý nghĩa trên 25 phần trăm. Do đó chúng ta kết luận là các hệ số hồi qui tương ứng không có ý nghĩa liên kết.

Nếu những biến này bị loại bỏ, chúng ta thấy là các trị thống kê của SQFT và POOL cao hơn. Tương tự, $\overline{R}^2$ tăng lên 0,89. Vì vậy, việc loại bỏ những biến không có ý nghĩa đã cải thiện kết quả chung của mô hình. Cần phải nhấn mạnh là kết luận này không có nghĩa là các biến loại bỏ không quan trọng, mà chỉ có nghĩa là *giữ* SQFT và POOL *không đổi,* việc thêm vào biến BEDRMS, BATHS, FAMROOM, và FIREPL không tăng thêm khả năng giải thích của mô hình. Ít nhất một số ảnh hưởng của các biến bị loại bỏ đã được các biến có trong mô hình thể hiện. Trong Mô hình F hệ số của POOL là 52,790, có nghĩa là giữa hai ngôi nhà có cùng diện tích sử dụng, căn nhà có hồ bơi được kỳ vọng sẽ bán ở mức giá cao hơn căn nhà không có hồ bơi một khoảng là $52.790. Xem xét chi phí xây dựng hồ bơi, giá trị này có vẻ quá cao. Có một cách giải thích là cùng với hồ bơi, những căn nhà này có thể còn có mạch nước ngầm, sân thượng hoặc một số đặc điểm khác. Vì vậy biến giả POOL có thể thật sự đại diện cho các nâng cấp khác.

⭘ Bảng 7.2 Ảnh Hưởng Của Các Biến Giả Đến Giá Nhà

Biến	Mô hình A	Mô hình E	Mô hình F
CONSTANT	52,351	39,057	22,673
	(1,404)	(0,436)	(0,768)
SQFT	0,13875	0,147	0,144
	(7,407)	(4,869)	(10,118)
BEDRMS		–7,046	
		(–0,245)	
BATHS		–0,264	
		(–0,006)	
POOL		53,196	52,790
		(2,411)	(3,203)
FAMROOM		–21,345	
		(–0,498)	
FIREPL		26,188	
		(0,486)	
$\overline{R}^2$	0,806	0,836	0,890
ESS	18.274	9.010	9.455
d.f.	12	7	11
SCMASQ	1.523	1.287	860*
AIC	1.737	1.749	1.037*
FPE	1.740	1.931	1.044*
HQ	1.722	1.698	1.024*
SCHWARZ	1.903	2.408	1.189*
SHIBATA	1.678	1.287	965*
CCV	1.777	2.574	1.094*
RICE	1.827	Không xác định	1.182*

Ghi chú: Mô hình B, C và D trong Bảng 4.2. Các giá trị trong ngoặc là các trị thống kê tương ứng.
* Đánh dấu mô hình tốt nhất xét về tiêu chuẩn tương ứng

⭘ BÀI THỰC HÀNH 7.2

Trong Phần 6.2, chúng ta đã lập luận là tác động biên tế của SQFT lên PRICE có thể giảm khi SQFT tăng. Điều này đưa đến việc sử dụng ln(SQFT) thay cho SQFT. Sử dụng chương trình hồi qui của bạn, ước lượng lại Mô hình A, E và F trong Bảng 7.2 sử dụng ln(SQFT) thay vì SQFT. Các kết quả có tốt không? Tiếp theo thử với SQFT. Các kết quả có được cải thiện hay xấu hơn không? Tìm một biểu thức cho ảnh hưởng *biên tế* của SQFT đến PRICE. (Bài thực hành máy tính Phần 7.4 sẽ có ích cho bài tập này)

Ví Dụ Thực Nghiệm: Thù Lao Và Thi Đấu Trong Liên Đoàn Bóng Chày

Sommers và Quinton (1982) tiến hành một nghiên cứu về thù lao và thi đấu trong liên đoàn bóng chày, trong nghiên cứu này, các biến giả được sử dụng để thể hiện các biến định tính như các đội trong liên đoàn quốc gia, các đội đoạt giải, một sân vận động cũ hay mới v.v… Trước khi thảo luận về kết quả của họ, cần phải giới thiệu vài nét tổng quan.

Ngành bóng chày có đặc điểm là *độc quyền (hoặc cạnh tranh độc quyền)* trong đó các ông chủ có khả năng kiểm soát lương của cầu thủ. Đến tận 1975, người chủ có thể ký lại hợp đồng mới

vĩnh viễn với một cầu thủ không ký hợp đồng. Tuy nhiên, trong năm đó trọng tài lao động Peter Seitz qui định là các cầu thủ có thể làm việc với nhiều chủ khác nhau sau khi chơi một năm không ký hợp đồng. Do lúc này cầu thủ có thể đi tìm những nơi trả giá cao cho dịch vụ họ cung cấp, chúng ta có thể kỳ vọng là lương của họ sẽ gần với kết quả doanh thu biên tế mong đợi (thu nhập tăng thêm trên một giờ lao động thêm). Cụ thể hơn, đặt R là tổng doanh thu của cả đội. Vậy lợi nhuận ròng là $\pi = R - wL - rK$, với L là lao động trong số giờ làm việc của công nhân, K đại diện cho tất cả những đầu vào khác, w là mức lương, và r là giá thuê. Vậy ông chủ đội muốn tối đa hóa lợi nhuận sẽ làm cho $\Delta\pi / \Delta L$ bằng không, dẫn đến điều kiện $\Delta R / \Delta L = w$. Vế trái là doanh thu biên tế. Vậy, để tối đa hóa lợi nhuận, lương phải bằng kết quả doanh thu biên tế.

Sommers và Quinton đã ước lượng đóng góp cá nhân của một số cầu thủ vào doanh thu biên tế và so sánh ước lượng này với lương của các cầu thủ tự do. Hai phương trình sau đã được ước lượng một cách riêng biệt, sử dụng các quan sát chéo của 50 đội trong số SMSA (Các khu vực thống kê của thành phố chuẩn) trong những năm 1976 và 1977:

$$\widehat{PCTWIN} = \underset{}{188{,}45} + \underset{(1{,}97)}{256{,}33}\ TSA^{*} + \underset{(2{,}90)}{80{,}87}\ TSW^{*} - \underset{(2{,}85)}{55{,}33}\ XPAN - \underset{(-2{,}62)}{}$$
$$+ \underset{(4{,}90)}{51{,}34}\ CONT - \underset{(-6{,}14)}{72{,}07}\ OUT$$
$$R^2 = 0{,}892 \quad d.f. = 44$$

$$\widehat{REVENUE} = \underset{}{-2.297.500} + (\underset{(-1{,}12)}{22.736} - \underset{(4{,}54)}{2.095}\ SMSA + \underset{(-1{,}65)}{415}\ SMSA^2)\ PCTWIN$$
$$- \underset{(-0{,}45)}{313.700}\ STD + \underset{(2{,}70)}{4.298.900}\ XPAN - \underset{(-3{,}74)}{3.750.200}\ TWOTM$$
$$- \underset{(-0{,}08)}{2.513}\ BBPCT$$
$$R^2 = 0{,}704 \quad d.f. = 42$$

với

PCTWIN = 100 lần tỷ số của số trận thắng trên số trận đã thi đấu

REVENUE = Số khách tham dự nhân giá vé trung bình cộng thu nhập giảm được ước lượng cộng doanh thu từ quyền truyền hình trận đấu

TSA^{*} = mức hoạt động trung bình của đội (tổng trận đấu chia cho tổng số cầu thủ) là một tỷ số trung bình của các bộ phận có liên quan của liên đoàn

TSW^{*} = tỷ số tấn công-xuất quân (số lần tấn công chia cho số lần ra quân) chia cho tỷ số tương tự của liên đoàn

XPAN = 1 nếu đội là một câu lạc bộ mở rộng, ngược lại bằng 0

CONT = 1 đối với đội đoạt giải hay đội thắng, ngoài ra sẽ bằng 0

OUT = 1 đối với các đội chơi 20 trận hoặc nhiều hơn từ khi bắt đầu đến kết thúc mùa bóng, ngược lại bằng 0

SMSA = dân số của SMSA

STD = 1 nếu sân vận động cũ, ngược lại bằng 0

TWOTM = 1 nếu đội có cùng SMSA nhà với một đội khác

BBPCT = phần trăm cầu thủ da đen chơi cho đội

Để cho phép có tương tác giữa PCTWIN và kích thước của SMSA trong hàm doanh thu, các tác giả đã giả định là $\Delta REVENUE / \Delta PCTWIN$ là hàm bậc hai của SMSA. Vì ví dụ này chỉ tập

trung vào các biến giả, chúng ta không diễn dịch bất kỳ kết quả nào khác. Sommers và Quinton đã sử dụng những phương trình được ước lượng này để tính kết quả doanh thu biên tế của 14 cầu thủ và so sánh các kết quả này với các mức lương tương ứng. Kết luận của họ là, trái với suy nghĩ phổ biến, các cầu thủ bóng chày bị trả lương thấp hơn nhiều so với mức họ đáng được hưởng.

Trong phương trình PCTWIN tất cả các biến giả đều có ý nghĩa. Câu lạc bộ được mở rộng, trung bình sẽ giảm 55 điểm. Các đội chơi trận đầu tiên trung bình thấp hơn 72 điểm. Trong phương trình REVENUE, STD không có ý nghĩa, cho thấy là sân vận động mới hay cũ không quan trọng. TWOTM có ý nghĩa và giá trị âm của biến này cho thấy việc có thêm một đội thứ hai trong cùng một thành phố sẽ gây thiệt hại đến doanh thu, điều này không có gì đáng ngạc nhiên.

⭘ 7.2 Biến Định Tính Với Nhiều Lựa Chọn

Số các lựa chọn có thể có của một biến định tính có thể nhiều hơn hai. Ví dụ, đặt Y là tiền tiết kiệm của một hộ gia đình và X là thu nhập của họ. Chúng ta kỳ vọng quan hệ giữa tiền tiết kiệm và thu nhập sẽ khác nhau đối với các nhóm tuổi khác nhau. Đối với một mức thu nhập cho trước, trung bình một hộ gia đình trẻ có thể tiêu dùng nhiều hơn so với một gia đình do một người trung niên làm chủ. Đó là do gia đình sau có thể tiết kiệm nhiều hơn dành cho việc giáo dục con cái và chuẩn bị khi về hưu. Một gia đình đã nghỉ hưu trung bình có vẻ tiêu xài nhiều hơn vì nhu cầu tiết kiệm cho tương lai lúc này sẽ giảm. Nếu chúng ta có tuổi chính xác của người chủ hộ, biến này có thể được đưa vào một mô hình như là biến định lượng. Tuy nhiên, nếu chúng ta chỉ có nhóm tuổi (ví dụ người chủ hộ thuộc nhóm tuổi dưới 25, từ 25 đến 55 hay trên 55), chúng ta xem xét biến định tính " nhóm tuổi của người chủ hộ" này như thế nào? Thủ tục ở đây là chọn một trong những nhóm này làm nhóm kiểm soát và xác định các biến giả cho hai nhóm còn lại. Cụ thể hơn, chúng ta xác định

$$A_1 = \begin{cases} 1 & \text{nếu chủ hộ từ 25 đến 55 tuổi} \\ 0 & \text{nếu điều kiện khác} \end{cases} \quad \textbf{(7.9)}$$

$$A_2 = \begin{cases} 1 & \text{nếu chủ hộ trên 55 tuổi} \\ 0 & \text{nếu điều kiện khác} \end{cases} \quad \textbf{(7.10)}$$

Nhóm kiểm soát (là nhóm mà cả A_1 và A_2 đều bằng 0) là tất cả những hộ gia đình mà người chủ hộ dưới 25 tuổi. Để α khác nhau đối với mỗi nhóm khác nhau, chúng ta giả định là $\alpha = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2$. Thay vào Phương trình (7.1) ta có

$$Y = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + \beta X + u \quad \textbf{(7.11)}$$

Đối với một hộ gia đình trẻ, $A_1 = A_2 = 0$. Đối với nhóm tuổi trung niên $A_1 = 1$ và $A_2 = 0$. Đối với nhóm lớn tuổi nhất, $A_1 = 0$ và $A_2 = 1$. Các mô hình được ước lượng cho ba nhóm này như sau:

Tuổi < 25: $\hat{Y} = \hat{\alpha}_0 + \hat{\beta} X$ **(7.12)**

Tuổi 25-55: $\hat{Y} = (\hat{\alpha}_0 + \hat{\alpha}_1) + \hat{\beta} X$ **(7.13)**

Tuổi > 55: $\hat{Y} = (\hat{\alpha}_0 + \hat{\alpha}_2) + \hat{\beta} X$ **(7.14)**

$\hat{\alpha}_1$ là một ước lượng của khác biệt trong tung độ gốc giữa hai nhóm hộ gia đình trẻ và trung niên. $\hat{\alpha}_2$ là ước lượng của khác biệt trong tung độ góc giữa nhóm hộ gia đình trẻ và hộ gia đình lớn tuổi. Vì vậy, *dịch chuyển tung độ gốc là những sai lệch so với nhóm kiểm soát.* Các đường thẳng ước lượng sẽ song song với nhau.

❶ Bảng 7.3 Giá Trị Dữ Liệu Mẫu Với Một Số Biến Định Tính

t	*Y*	*Const*	*X*	A_1	A_2	H	E_1	E_2	O_1	O_2	O_3	O_4
1	Y_1	1	X_1	1	0	1	1	0	0	1	0	0
2	Y_2	1	X_2	1	0	0	0	0	0	0	0	1
3	Y_3	1	X_3	0	0	0	0	1	0	0	0	0
4	Y_4	1	X_4	0	1	0	1	0	0	0	1	0
5	Y_5	1	X_5	0	1	0	1	0	0	1	0	0

Có một lý do đặc biệt để không định nghĩa một biến giả thứ ba, A_3, nhận giá trị 1 đối với nhóm gia đình trẻ và giá trị 0 cho các nhóm khác. Nếu chúng ta đã giả định là $\alpha = \alpha_0 + \alpha_1 A_1 + \alpha_2 A_2 + \alpha_3 A_3$, chúng ta sẽ gặp đa cộng tuyến chính xác vì $A_1 + A_2 + A_3$ luôn luôn bằng 1, là một số hạng không đổi (xem Bảng 7.3). Đây gọi là **bẫy biến giả.** Để tránh vấn đề này, *số các biến giả luôn luôn ít hơn một biến so với số các lựa chọn* (xem Bài thực hành 7.3 đối với một trường hợp ngoại lệ đối với vấn đề này). Vì vậy, nếu chúng ta muốn tính các sai biệt theo mùa giữa lượng điện tiêu thụ và nhiệt độ, chúng t sẽ định nghĩa ba biến giả (vì có tất cả bốn mùa). Để tính sai biệt theo tháng, chúng ta cần 11 biến giả.

Một số giả thuyết rất thú vị. Để kiểm định giả thuyết gia đình ở nhóm tuổi cao hơn có hành vi giống gia đình ở nhóm tuổi trẻ hơn, chúng ta đơn giản chỉ tiến hành kiểm định *t* đối với $\hat{\alpha}_2$. Để kiểm định giả thuyết "không có khác biệt trong hàm tiết kiệm theo độ tuổi", giả thuyết là H_0: $\alpha_1 = \alpha_2 = 0$ và giả thuyết ngược lại là H_1: ít nhất một trong các hệ số khác không. Giả thuyết này được kiểm định bằng kiểm định Wald được trình bày trong Phần 4.4. Mô hình không giới hạn là Phương trình (7.11), và mô hình giới hạn là $Y = \alpha_0 + \beta X + u$. Kiểm định Wald F từ các tổng bình phương tương ứng sẽ có bậc tự do d.f. là 2 và $n - 4$. Giả thuyết "không có khác biệt trong hành vi giữa hai nhóm tuổi trung niên và cao tuổi" nghĩa là $\alpha_1 = \alpha_2$. Giả thuyết này có thể được kiểm định bằng cách sử dụng ba phương pháp đã được mô tả trong Phần 4.4. Để áp dụng kiểm định Wald, đặt điều kiện này vào Phương trình (7.11). Chúng ta có mô hình giới hạn

$$\begin{aligned} Y &= \alpha_0 + \alpha_1 A_1 + \alpha_1 A_2 + \beta X + u \\ &= \alpha_0 + \alpha_1 (A_1 + A_2) + \beta X + u \end{aligned} \qquad \textbf{(7.15)}$$

Thủ tục để ước lượng mô hình giới hạn là tạo ra một biến mới, $Z = A_1 + A_2$, và hồi qui Y theo một hằng số, Z, và X. Một kiểm định Wald được thực hiện sau đó giữa mô hình này và Phương trình (7.11) bằng cách so sánh các tổng bình phương của các phần dư ước lượng. Trị thống kê F sẽ có bậc tự do d.f. là 1 và $n - 4$.

❶ BÀI TẬP THỰC HÀNH 7.3

Giả sử chúng ta đã dùng biến giả thứ ba A_3 như vừa định nghĩa và đã thiết lập mô hình $Y = \beta_1 A_1 + \beta_2 A_2 + \beta_3 A_3 + \beta X + u$, *không có số hạng không đổi.* Chứng tỏ là không có vấn đề đa cộng tuyến

chính xác ở đây. Hãy mô tả có thể tính được các ước lượng của các α từ các ước lượng của các β như thế nào.

⊙ BÀI TẬP THỰC HÀNH 7.4+

Chọn một nhóm tuổi khác làm nhóm kiểm soát – giả sử nhóm trung niên – và lập lại mô hình. Các giá trị ước lượng của mô hình mới quan hệ như thế nào với các ước lượng trong Phương trình (7.11)? Cụ thể hơn, tính các ước lượng của mô hình mới từ những ước lượng của Phương trình (7.11). Mô tả các kiểm định giả thuyết cụ thể có thể thực hiện trong mô hình mới này.

Một Số Các Biến Định Tính

Phân tích biến giả dễ dàng được mở rộng cho trường hợp trong đó có nhiều biến định tính, một số các biến này có thể có nhiều hơn một giá trị. Để minh họa, hãy xem xét hàm tiết kiệm được mô tả trước đây, trong đó, Y là tiết kiệm của hộ gia đình và X là thu nhập của hộ gia đình. Có thể đưa ra giả thuyết là ngoài tuổi của chủ hộ, các yếu tố khác như sở hữu nhà, trình độ học vấn, tình trạng nghề nghiệp v.v… cũng là những yếu tố xác định tiết kiệm của hộ gia đình. Ví dụ, giả sử ta có thông tin là chủ hộ có trình độ sau đại học, có trình độ đại học, chỉ tốt nghiệp trung học. Hơn nữa, giả sử ta biết là chủ hộ có thể làm một trong những nghề sau: quản lý, công nhân tay nghề cao, công nhân không có tay nghề, thư ký, kinh doanh tự do hoặc nhân viên chuyên nghiệp. Cũng tương tự, ta không biết chính xác tuổi của chủ hộ nhưng biết được ông ta/bà ta thuộc nhóm tuổi nào. Chúng ta đưa những biến này vào phân tích như thế nào? Thủ tục là định nghĩa tất cả các biến giả cần có và đưa chúng vào mô hình. Mô hình không giới hạn sẽ như sau:

$$Y = \beta_0 + \beta_1 A_1 + \beta_2 A_2 + \beta_3 H + \beta_4 E_1 + \beta_5 E_2 + \beta_6 O_1 + \beta_7 O_2 + \beta_8 O_3 + \beta_9 O_4 + \beta_{10} X + u \qquad \textbf{(7.16)}$$

với

$A_1 = \begin{cases} 1 & \text{nếu chủ hộ từ 25 đến 55 tuổi} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$A_2 = \begin{cases} 1 & \text{nếu chủ hộ trên 55 tuổi} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$H = \begin{cases} 1 & \text{nếu chủ hộ sở hữu căn nhà} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$E_1 = \begin{cases} 1 & \text{nếu chủ hộ có trình độ sau đại học} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$E_2 = \begin{cases} 1 & \text{nếu chủ hộ có trình độ đại học} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$O_1 = \begin{cases} 1 & \text{nếu chủ hộ là nhà quản lý} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$O_2 = \begin{cases} 1 & \text{nếu chủ hộ là công nhân lành nghề} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$O_3 = \begin{cases} 1 & \text{nếu chủ hộ là thư ký} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

$O_4 = \begin{cases} 1 & \text{nếu chủ hộ kinh doanh cá thể} \\ 0 & \text{nếu điều kiện khác} \end{cases}$

Nên lưu ý rằng đặc tính của nhóm điều khiển như sau: chủ hộ có độ tuổi dưới 25, là công nhân không có tay nghề, với trình độ học vấn chỉ ở bậc trung học. Bảng 7.3 là một ví dụ về ma trận dữ liệu. Ước lượng các tham số được thực hiện bằng việc lấy hồi qui Y theo một số hạng không đổi, A_1, A_2, H, E_1, E_2, O_1, O_2, O_3, O_4, và X (các biến định lượng cộng thêm thêm được đưa vào dễ dàng nếu mô hình cần chúng). Tình trạng sở hữu nhà được kiểm định bằng kiểm định t đối với b_3 (với bậc tự do df là n - 11). Trình độ học vấn được kiểm định bằng kiểm định Wald với *giả thuyết không* là $b_4 = b_5 = 0$. Mô hình không giới hạn là Phương trình (7.16), và mô hình giới hạn là mô hình có được từ việc loại bỏ E_1 và E_2 ra khỏi (7.16). Bậc tự do đối với trị thống kê F sẽ là 2 và n - 11. Tương tự, để kiểm định xem tình trạng việc làm có phải là vấn đề trong việc lý giải những biến động trong tiết kiệm, ta sử dụng kiểm định Wald với giả thuyết không là $b_6 = b_7 = b_8 = b_9 = 0$. Có thể sử dụng rất nhiều kiểm định khác nữa; những kiểm định này được dành lại cho người đọc trong phần bài tập.

⊙ BÀI THỰC HÀNH 7.5

Viết quan hệ ước lượng cho một hộ gia đình trung niên sở hữu nhà, chủ nhà có bằng đại học, và là một nhân viên văn phòng.

⊙ BÀI THỰC HÀNH 7.6 +

Mô tả từng bước việc thực hiện kiểm định những giả thuyết sau: (a) "hành vi tiết kiệm của những nhân viên văn phòng tương tự như hành vi tiết kiệm của những công nhân lành nghề," và (b) "tình trạng việc làm không có tác động ý nghĩa lên hành vi tiết kiệm." Cụ thể hơn, mô tả việc chạy (các) hồi qui, tính toán các kiểm định thống kê, phân phối thống kê theo giả thuyết không (bao gồm cả bậc tự do), và các tiêu chí để bác bỏ *giả thuyết không*.

Các Mô Hình Phân Tích Phương Sai*

Tất cả các biến độc lập trong một mô hình đều có thể là nhị nguyên. Những mô hình như vậy được gọi là **mô hình phân tích phương sai (ANOVA)**. Chúng rất phổ biến trong các ngành kinh tế nông nghiệp, nghiên cứu thị trường, xã hội học, và tâm lý học. Trong phần này, chúng ta chỉ giới thiệu các mô hình ANOVA một cách tóm tắt. Chi tiết hơn, tham khảo một cuốn sách về thống kê nào đó hay những thiết kế thực nghiệm.

Xem xét một thực nghiệm nông nghiệp mà nhà điều tra lên kế hoạch nghiên cứu sản lượng trung bình trên một mẫu do ba loại hạt giống ghép khác nhau được xử lý với bốn loại liều lượng thuốc trừ sâu khác nhau. Người thiết kế thực nghiệm này chia khoảnh đất rộng thành một số các mảnh đất nhỏ hơn và một cách ngẫu nhiên đưa vào những kết hợp khác nhau giữa hạt giống và liều lượng phân bón. Tiếp theo sản lượng quan sát được trên mỗi mảnh đất được liên hệ với loại hạt giống và liều lượng phân bón tương ứng. Nhà thiết kế thực nghiệm sẽ thiết lập nên mô hình như sau:

$$Y_{ijk} = m + a_j + b_k + e_{ijk}$$

với Y_{ijk} là sản lượng quan sát được trên mảnh đất thứ i sử dụng hạt giống thứ j (j = 1, 2, 3) và liều lượng phân bón thứ k (k = 1, 2, 3, 4), m là "trung bình lớn", a_j là "tác động của hạt giống", và b_k là "tác động của phân bón", e_{ijk} là số hạng sai số không quan sát được. Do vậy sản lượng trung bình được kết hợp lại từ tác động toàn bộ chung lên tất cả các mảnh đất, mà nó được hiệu chỉnh theo loại hạt giống và liều lượng phân bón trên từng mảnh đất . Bởi vì a_j và b_k là những thiên lệch từ trị trung bình tổng thể, chúng ta có điều kiện $\sum a_j = \sum b_k = 0$. Chính vì những ràng buộc này, tám tham số (m,

ba *a*, và bốn *b*) thực tế giảm xuống chỉ còn sáu tham số. Mô hình được viết lại như sau cho những kết hợp đã chọn:

$$Y_{i12} = m + a_1 + b_2 + e_{i12}$$

$$Y_{i34} = m + a_3 + b_4 + e_{i34}$$

Ta có thể thiết lập một mô hình tương tự chỉ với những biến giả. Đối với những loại hạt giống, định nghĩa hai biến giả: $S_1 = 1$ nếu loại hạt giống đầu tiên được chọn, nếu không sẽ là 0; $S_2 = 1$ nếu loại hạt giống thứ hai được chọn, nếu không sẽ là 0. Tương tự như vậy, định nghĩa ba biến giả cho liều lượng thuốc trừ sâu: $D_1 = 1$ khi liều lượng thứ nhất được sử dụng, $D_2 = 1$ cho liều lượng thứ hai, và $D_3 = 1$ cho liều lượng thứ ba. Lưu ý rằng nhóm kiểm soát là loại hạt giống thứ ba và liều lượng thuốc thứ tư. Phương trình kinh tế lượng là

$$Y = a_0 + a_1S_1 + a_2S_2 + b_1D_1 + b_2D_2 + b_3D_3 + u$$

Ở đây cũng có sáu tham số chưa biết để ước lượng. Đối với hai kết hợp ở trên, mô hình trở thành

$$Y = a_0 + a_1 + b_2 + u \qquad (S_1 = D_2 = 1, S_2 = D_1 = D_3 = 0)$$

$$Y = a_0 + u \qquad (S_1 = S_2 = D_1 = D_2 = D_3 = 0)$$

Trong khi so sánh hai phương pháp, chúng ta lưu ý rằng $a_0 + a_1 + b_2 = m + a_1 + b_2$ và $a_0 = m + a_3 + b_4$. Có thể chỉ rõ sự tương ứng một-một giữa mô hình kinh tế lượng và mô hình thiết kế thực nghiệm. Giả thuyết cho rằng không có sự khác biệt giữa các hạt giống có thể được diễn dịch như $a_1 = a_2 = a_3 = 0$, hay cũng tương đương như $a_1 = a_2 = 0$. Tương tự như vậy, giả thuyết cho rằng không có sự khác biệt về sản lượng do tác động của liều lượng thuốc trừ sâu có thể được kiểm định hoặc bằng $b_1 = b_2 = b_3 = b_4 = 0$ hoặc $b_1 = b_2 = b_3 = 0$.

⭘ BÀI THỰC HÀNH 7.7

Viết tất cả các quan hệ giữa các *a, b,* và *a, b*; Tìm a và b dưới dạng các *a* và *b*; và chỉ ra cách thiết lập thiết kế thực nghiệm từ phương trình kinh tế lượng.

⭘ 7.3 Tác động Của Các Biến Định Tính Lên Số Hạng Độ Dốc (Phân Tích Đồng Phương Sai)

Chỉ Dịch Chuyển Số Hạng Độ Dốc

Trong phần này, chúng ta cho phép khả năng của *b* có thể khác nhau cho những biến định tính khác nhau. Những mô hình như vậy được biết đến như **những mô hình phân tích đồng phương sai.** Chẳng hạn như trong ví dụ về tiền lương, làm sao chúng ta có thể kiểm định được giả thuyết cho rằng *b* là khác nhau giữa nam và nữ? Đầu tiên chúng ta giả định rằng hệ số tung độ gốc *a* là không thay đổi. (Điều này sẽ được nới lỏng trong phần kế tiếp.) Thủ tục tương tự với trường hợp mà tung độ gốc dịch chuyển giữa hai lựa chọn. Đặt $b = b_1 + b_2D$, với $D = 1$ cho nam và bằng 0 cho nữ. Phương trình (7.1) bây giờ trở thành

$$Y = a + (b_1 + b_2D)X + u \quad \textbf{(7.17)}$$

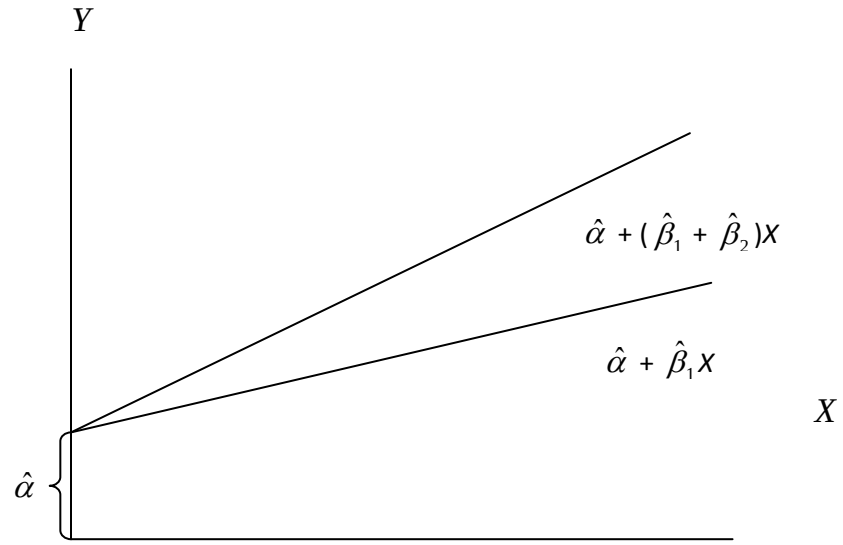
$$= a + b_1X + b_2(DX) + u$$

b_2DX biểu diễn số hạng tương tác được mô tả trong Phần 6.5. Để ước lượng mô hình này, chúng ta nhân biến giả với X và tạo một biến mới, $Z = DX$. Rồi chúng ta hồi qui Y theo một số hạng không đổi, X, và Z. Các quan hệ được ước lượng như sau (được biểu diễn trên Hình 7.2, với giả định rằng a và tất cả b dương):

Nữ: $$\hat{Y} = \hat{\alpha} + \hat{\beta}_1 X \quad \textbf{(7.18)}$$

Nam: $$\hat{Y} = \hat{\alpha} + (\hat{\beta}_1 + \hat{\beta}_2)X \quad \textbf{(7.19)}$$

Hình 7.2 Một Ví Dụ Của Việc Dịch Chuyển Độ Dốc Bằng Cách Sử Dụng Biến Giả



Bởi vì tung độ gốc được giả định là như nhau, nên những đoạn thẳng bắt đầu từ cùng một điểm nhưng có độ dốc khác nhau. Nếu một công nhân viên nữ tích lũy thêm một năm kinh nghiệm, thì cô ta sẽ mong đợi nhận được mức lương trung bình tăng lên $\hat{\beta}_1$ đô la. Nam nhân viên với thêm một năm kinh nghiệm sẽ kỳ vọng mức lương trung bình tăng lên $\hat{\beta}_1 + \hat{\beta}_2$ đô la một tháng. Do vậy, $\hat{\beta}_2$ đo lường sự khác biệt trong độ dốc ước lượng.

Thủ tục kiểm định giả thuyết cũng tương tự như trường hợp trước, tức là chỉ có tung độ gốc dịch chuyển. Một kiểm định t đối với b_2 (bậc tự do d.f là $n - 3$) sẽ kiểm định rằng không có sự khác biệt nào về độ dốc.

Dịch Chuyển Cả Số Hạng Tung Độ Gốc Và Độ Dốc

Cho phép dịch chuyển cả tung độ gốc và độ dốc là một thủ tục không mấy phức tạp. Chúng ta chỉ đơn giản cho $a = a_1 + a_2D$ và $b = b_1 + b_2D$. Thay thế hai giá trị này vào Phương trình (7.1), ta có mô hình không giới hạn là

$$Y = a_1 + a_2D + (b_1 + b_2D)X + u \quad \textbf{(7.20)}$$

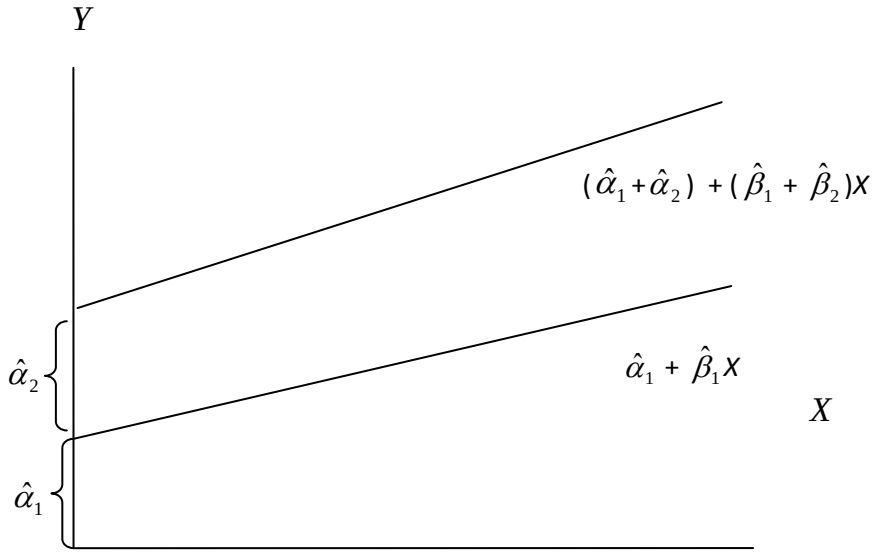
$$= a_1 + a_2D + b_1X + b_2(DX) + u$$

Hồi qui Y theo một hằng số, *D*, *X*, và số hạng tương tác *DX*. Các quan hệ được ước lượng cho hai nhóm là

Nam: $$\hat{Y} = (\hat{\alpha}_1 + \hat{\alpha}_2) + (\hat{\beta}_1 + \hat{\beta}_2)X \quad \textbf{(7.21)}$$

Nữ: $$\hat{Y} = \hat{\alpha}_1 + \hat{\beta}_1 X \quad \textbf{(7.22)}$$

Hình 7.3 Một Ví Dụ Của Việc Dịch Chuyển Tung Độ Gốc Và Độ Dốc



Hình 7.3 biểu diễn các mối quan hệ này khi tất cả *a* và *b* dương. Để kiểm định giả thuyết cho rằng không có sự khác biệt nào trong toàn bộ quan hệ, chúng ta có H_0: $a_2 = b_2 = 0$. Kiểm định là kiểm định *Wald F*, với Phương trình (7.20) là mô hình không giới hạn và $Y = a_1 + b_1X + u$ là mô hình giới hạn. Trị thống kê *F* sẽ có bậc tự do df là 2 và *n* - 4.

Diễn Dịch Các Hệ Số Biến Giả Trong Mô Hình Tuyến Tính-Lôgarít

Trong Phần 6.8 chúng ta đã giới thiệu mô hình tuyến tính-lôgarít mà theo đó biến phụ thuộc là ln*(Y)*. 100 nhân với một hệ số hồi qui được diễn dịch là *thay đổi phần trăm* trung bình của *Y* so với *thay đổi một đơn vị* của biến độc lập tương ứng. Tuy nhiên, nếu biến độc lập là một biến giả, thì việc diễn dịch sẽ không còn giá trị. Để thấy được điều này, xem xét mô hình

$$\ln (Y) = b_1 + b_2X + b_3D + u$$

với *D* là một biến giả. Lấy đối log của phương trình này, ta được $Y = \exp(b_1 + b_2X + b_3D + u)$, với exp là hàm mũ. Ký hiệu biến phụ thuộc là Y_1 khi $D = 1$, và Y_0 khi $D = 0$. Do đó phần trăm thay đổi giữa hai nhóm là $100\ (Y_1 - Y_0)/Y_0 = 100\ [\exp(b_3) - 1)]$. Việc đầu tiên là ước lượng $\exp^{\beta_3}$ theo $\exp^{\hat{\beta}_3}$. Tuy nhiên, đây không phải là phương pháp thích hợp, lý do tại sao sẽ được giải thích kỹ hơn

trong Phần 6.8. Phương pháp đúng để hiệu chỉnh thiên lệch ở $\widehat{\exp^{\beta_3}}$ là $\exp\left[\widehat{\beta_3 - \frac{1}{2}}\mathrm{Var}(\hat{\beta}_3)\right]$, với Var là phương sai ước lượng. Từ đó ta có

$$100(\hat{Y}_1 / \hat{Y}_0 - 1) = 100\left\{\exp\left[\widehat{\hat{\beta}_3 - \tfrac{1}{2}}\mathrm{Var}\left(\hat{\beta}_3\right)\right] - 1\right\}$$

Nếu mô hình có một số hạng tương tác thì mô hình sẽ trở thành

$$\text{Ln}\,(Y) = b_1 + b_2X + b_3D + b_4DX + u$$

biểu thức tương ứng phức tạp hơn nhiều. Trong trường hợp này, việc kiểm tra mô hình sẽ là

$$100(\hat{Y}_1 / \hat{Y}_0 - 1) = 100\left\{\exp\left[\hat{\beta}_3 + \widehat{\hat{\beta}_4 - \tfrac{1}{2}}\mathrm{Var}\left(\hat{\beta}_3 + \hat{\beta}_4 X\right)\right] - 1\right\}$$

Biểu thức phương sai phụ thuộc vào giá trị của X và nó cũng bao gồm một kết hợp tuyến tính giữa các biến ngẫu nhiên. Dễ dàng thấy rằng, khi mô hình có một số hạng tương tác giữa một biến giả và biến định lượng, việc diễn giải tác động của biến giả phức tạp hơn nhiều.

Mặc dù việc diễn giải tác động của biến giả đòi hỏi sự hiệu chỉnh trong trường hợp mô hình tuyến tính-lôgarít, tác động cận biên của một biến định lượng thì khá dễ hiểu. Ta có, $\partial \ln(\hat{Y})/\partial X = \hat{\beta}_2 + \hat{\beta}_4 D$. Sử dụng Tính chất 6.2c, cho ta,

$$100\frac{\Delta Y}{Y} = 100(\hat{\beta}_2 + \hat{\beta}_4 D)\Delta X$$

Dẫn đến $100\,\hat{\beta}_2$ là *phần trăm thay đổi* gần đúng của Y đối với sự *thay đổi một đơn vị* của X khi D = 0 và $100(\hat{\beta}_2 + \hat{\beta}_4)$ là phần trăm thay đổi gần đúng của Y đối với sự thay đổi một đơn vị của X khi D = 1.

⭘ 7.4 Ứng Dụng: Phân Tích Đồng Phương Sai Trong Mô Hình Tiền Lương

Ứng dụng xuyên suốt được chọn ở đây là ứng dụng đã được sử dụng trong Ví dụ 6.5, đó là quan hệ giữa tiền lương và đặc tính của nhân viên. Tuy nhiên, trong ví dụ đó, chúng ta chỉ sử dụng yếu tố học vấn, kinh nghiệm, lương bổng, và mức chi tiêu của họ. Mô hình tuyến tính-lôgarít căn bản là

(A) $\ln(\text{WAGE}) = a + b\,\text{EDUC} + g\,\text{EXPER} + d\,\text{AGE} + u$

Từ phân tích trước đó chúng ta thấy rằng giá trị của $\overline{R}^2$ của Mô hình A là 0,283, điều đó có nghĩa là ba biến giải thích chỉ giải thích được 28,3% mức biến động của ln(WAGE). Như đã chỉ ra trước đây, điều này là không thuyết phục lắm ngay cả đối với một nghiên cứu chéo, mà thường là có những giá trị $\overline{R}^2$ thấp. Quan hệ trên là trung bình đối với tất cả các nhóm nhân viên và có vẻ như khác nhau cho từng mức độ kỹ năng khác nhau cũng như giới tính và sắc tộc khác nhau. DATA 7-2, mô tả trong Phụ lục D, có dữ liệu hoàn chỉnh cho mẫu gồm 49 nhân viên ở một cơ quan nào đó. Các biến giải thích bao gồm số năm đi học trên lớp tám ở thời điểm mà người đó được thuê mướn (EDUC), số năm kinh nghiệm (EXPER) tại cơ quan đó, và độ tuổi của nhân viên (AGE). Đồng thời cũng có những biến giả về giới tính, sắc tộc, và loại công việc (VD: nhân viên văn phòng, bảo trì, hay thợ thủ công). Phân loại như sau: nam (GENDER = 1), da trắng (RACE = 1),

nhân viên văn phòng (CLERICAL = 1), nhân viên bảo trì (MAINT = 1), và thợ thủ công (CRAFTS = 1). Nhóm điều khiển là nữ, da màu, có tay nghề và chúng ta có các giá trị zero cho những biến giả này.

Giả sử giả thuyết rằng *a* không giống nhau cho tất cả các nhân viên, nhưng khác nhau tùy theo giới tính, sắc tộc, và tình trạng nghề nghiệp. Để kiểm định điều này, giả định rằng

$$a = a_1 + a_2\text{GENDER} + a_3\text{RACE} + a_4\text{CLERICAL} + a_5\text{MAINT} + a_6\text{CRAFTS}$$

và kiểm định giả thuyết $a_2 = a_3 = \ldots = a_6 = 0$.

Thay thế *a* trong Mô hình A, ta được Mô hình B, mô hình không giới hạn mà nó liên quan với ln(WAGE) đến một số các biến định tính cũng như các biến giả.

$$\text{(B)} \quad \ln(\text{WAGE}) = a_1 + a_2\,\text{GENDER} + a_3\,\text{RACE} + a_4\,\text{CLERICAL} + a_5\,\text{MAINT} + a_6\text{CRAFTS} + b\,\text{EDUC} + g\,\text{EXPER} + d\,\text{AGE} + u$$

Một câu hỏi dễ thấy là liệu những tác động cận biên của học vấn, kinh nghiệm, và độ tuổi có phụ thuộc vào loại công việc, giới tính, và sắc tộc hay không. Hay nói một cách khác, số năm đi học hay số năm kinh nghiệm góp phần vào mức lương của một nhân viên nam nhiều hơn góp phần vào một nhân viên nữ hay một nhân viên da màu hay không? Cũng như vậy, có "lợi nhuận giảm dần theo qui mô" đối với việc học tập và kinh nghiệm không? Cụ thể hơn, thu nhập tăng thêm cho việc có nhiều hơn một năm học tập có giảm khi học vấn tăng lên hay không? Để trả lời cho những câu hỏi này, chúng ta cho phép số hạng "độ dốc" *b, g,* và *d* phụ thuộc vào các đặc tính khác nhau của một người nhân viên. Do đó, ví dụ như chúng ta có thể giả định

$$b = b_1 + b_2\text{GENDER} + b_3\text{RACE} + b_4\text{CLERICAL} + b_5\text{MAINT} + b_6\text{CRAFTS} + b_7\text{EDUC}$$

và kiểm định xem $b_i = 0$ cho tất cả các i = 2 – 7 hay không. Những đặc trưng tương tự đều có thể sử dụng được cho *g* và *d*. Nếu ta thay thế *a* và *b* và những quan hệ tương tự là *g* và *d* vào mô hình cơ bản, ta được một mô hình hoàn chỉnh với nhiều số hạng bậc hai và số hạng tương tác. Để tiết kiệm khoảng trống, chúng ta sẽ không viết phương trình hoàn chỉnh này. Với sự gia tăng nhanh chóng của các biến như vậy, phương pháp "tổng quan đến đơn giản" sẽ không dễ dàng tí nào. Phương pháp kiểm định LM bắt đầu từ Mô hình A căn bản sẽ dễ kiểm soát hơn nhiều.

Bảng 7.4 cho thấy một kết quả vi tính riêng phần mà nó minh họa cách kiểm định nhân tử Lagrange có thể được sử dụng để xác định xem một vài hoặc tất cả các số hạng thêm vào có ý nghĩa hay không. Phần Thực Hành Máy Tính 7.5 sẽ hữu ích trong việc tái tạo lại các kết quả này và trong việc thực hiện những nghiên cứu về sau.

So sánh các hệ số và những trị thống kê liên quan đối với các biến thêm vào trong hồi qui phụ (xem Bảng 7.4) với những hệ số và trị thống kê của Mô hình 3 tổng quát nhất trong Bảng 7.5.

Lưu ý rằng chúng giống như nhau. Giá trị R bình phương đối với hồi qui phụ là 0,818, trị thống kê nR^2 hơi lớn hơn 40, và giá trị *p* tương ứng là 0,01506. Điều này có nghĩa là chúng ta *bác bỏ* giả thuyết không cho rằng tất cả các biến thêm vào đều có những hệ số hồi qui không có ý nghĩa, dẫn đến xác suất của sai lầm loại I (loại bỏ một giả thuyết đúng) chỉ là 1,5 phần trăm. Bởi vì con số này rất thấp, chúng ta khá "an toàn" trong việc loại bỏ giả thuyết không và không có gì ngạc nhiên khi kết luận rằng có ít nhất một vài biến có liên quan đến mô hình.

Câu hỏi đặt ra ở đây là, "Chúng ta nên đưa biến mới nào trong mô hình hồi qui phụ vào đặc trưng của mô hình?" Nếu chúng ta tuân theo ý nghĩa chặt chẽ (ở mức 10 phần trăm hoặc những mức thấp hơn), thì chỉ có sq_EDUC, (bình phương của EDUC). ED_CRAFT (EDUC*CRAFTS), và AGE_MAIN (AGA*MAINT) sẽ được đưa vào mô hình. Tuy nhiên, chúng ta có thể kỳ vọng một đa cộng tuyến giữa các biến giải thích, mà có thể làm cho các hệ số không còn ý nghĩa. Qui tắc kinh nghiệm bảo thủ là chọn những biến mà các giá trị *p* của các hệ số là nhỏ hơn 0,5 (những nhà nghiên cứu khác có thể sẽ ưa thích một vài qui tắc khác). Theo qui tắc này, chúng ta đưa các biến GENDER, RACE, sq_EDUC, sq_EXPER, sq_AGE, ED-GEN, ED_CLER, ED_MAINT, ED_CRAFT, AGE_GEN, AGE_RACE, AGE_MAIN, AGE_CRFT, EXP_RACE, và EXP_CRFT vào mô hình. Ta ước lượng mô hình này, kết quả được tóm tắt trong Bảng 7.5 với tiêu đề Mô hình 1.

❹ Bảng 7.4 Một Phần Kết Quả Có Kèm Chú Giải Của Ứng Dụng Kiểm Định LM Trong Phần 7.4

[Danh sách dưới đây bao gồm một số biến bình phương và tương tác của chúng được phát ra thông qua những biến đổi nội tại. sq_x là bình phương của x, và x_y là tích của x và y.]

0)	const	1)	WAGE	2)	EDUC	3)	EXPER	4)	AGE
5)	GENDER	6)	RACE	7)	CLERICAL	8)	MAINT	9)	CRAFTS
10)	sq_EDUC	11)	sq_EXPER	12)	sq_AGE	13)	ED_GEN	14)	ED_RACE
15)	ED_CLER	16)	ED_MAINT	17)	ED_CRAFT	18)	AGE_GEN	19)	AGE_RACE
20)	AGE_CLER	21)	AGE_MAIN	22)	AGE_CRFT	23)	EXP_GEN	24)	EXP_RACE
25)	EXP_CLER	26)	EXP_MAIN	27)	EXP_CRFT	28)	LWAGE		

[Đầu tiên lấy hồi qui của ln(WAGE) theo một hằng số, EDUC, EXPER, và AGE, và giữ lại các phần dư $\hat{u}_t$ xem như ut. Tiếp theo là hồi qui phụ tức là lấy hồi qui các phần dư theo tất cả các biến trong mô hình không giới hạn.]

❹ Bảng 7.4 (tiếp theo)

Dependent variable: ut = $\hat{u}_t$

	VARIABLE	COEFFICIENT	STDERROR	T STAT	Prob (t > |T|)
0)	const	-0.8801	1.0029	-0.878	0.389639
2)	EDUC	0.2627	0.1399	1.878	0.073766

3)	EXPER	0.0259	0.0354	0.732	0.471946
4)	AGE	0.0118	0.0290	0.408	0.686923
5)	GENDER	0.4091	0.4499	0.909	0.373031
6)	RACE	-0.3639	0.4476	-0.813	0.424849
7)	CLERICAL	0.3677	0.7374	0.499	0.622954
8)	MAINT	-0.2408	0.9154	-0.263	0.794947
9)	CRAFTS	0.1086	0.7718	0.141	0.889374
10)	sq_EDUC	-0.0222	0.0109	-2.038	0.053703
11)	sq_EXPER	-0.0008053	0.0011	-0.705	0.488381
12)	sq_AGE	-0.0002380	0.0003148	-0.756	0.457766
13)	ED_GEN	0.0700	0.0471	1.485	0.151642
14)	ED_RACE	0.0368	0.0560	0.658	0.517394
15)	ED_CLER	-0.0761	0.0506	-1.504	0.146848
16)	ED_MAINT	-0.2094	0.1305	-1.605	0.122645
17)	ED_CRAFT	-0.1245	0.0682	-1.826	0.081493
18)	AGE_GEN	-0.0151	0.0098	-1.533	0.139646
19)	AGE_RACE	0.0104	0.0100	1.044	0.307974
20)	AGE_CLER	-0.0041	0.0096	-0.426	0.674481
21)	AGE_MAIN	0.0255	0.0121	2.102	0.047225
22)	AGE_CRFT	0.0114	0.0118	0.968	0.343325
23)	EXP_GEN	-0.0048	0.0178	-0.268	0.791547
24)	EXP_RACE	-0.0229	0.0246	-0.928	0.363564
25)	EXP_CLER	-0.0077	0.0206	-0.375	0.711264
26)	EXP_MAIN	-0.0018	0.0272	-0.068	0.946544
27)	EXP_CRFT	0.0184	0.0188	0.975	0.340112

```
Unadjuste R-squared        0.818   Adjusted R-squared          0.603
Chi-square(23): area to the right of 40.078742 (LM statistic) = 0.015060
```

[Giá trị *p* thấp chỉ ra việc bác bỏ giả thuyết không cho rằng hệ số của các biến thêm vào là zero. Bước kế tiếp là chọn biến thêm vào mô hình cơ bản sử dụng qui tắc kinh nghiệm đơn giản nhưng tùy ý về việc bao gồm cả *các biến cộng thêm mới* mà có giá trị *p* nhỏ hơn 0,5. Kết quả là Mô Hình 1 trong Bảng 7.5. Để có được mô hình cuối cùng, tức Mô Hình 2 trong Bảng 7.5, chúng ta bỏ các biến mà có các hệ số không có ý nghĩa, một vài biến một lần, cho đến khi nào tất cả các hệ số đều có ý nghĩa ở mức 10 phần trăm. Cuối cùng, chúng ta ước lượng mô hình hoàn chỉnh với tất cả các số hạng bình phương và số hạng tương tác. Đó chính là Mô Hình 3 trong Bảng 7.5. Lưu ý rằng Mô Hình 3 có một số số hạng không có ý nghĩa chủ yếu là do hiện tượng đa cộng tuyến mạnh.]

Bảng 7.5 Kết Quả Mô Hình Chọn Lọc Cho Ứng Dụng

Biến	**Mô hình 1**	**Mô hình 2**	**Mô hình 3**
CONSTANT	6,25809	6,69328	5,95582
	(11,879)	(36,624)	(5,939)
EDUC	0,29236	0,22078	0,32721
	(3,347)	(3,618)	(2,339)
EXPER	0,04514	0,01794	0,04864
	(2,095)	(4,287)	(1,372)
AGE	0,00465		0,01223
	(0,219)		(0,422)
GENDER	0,18628		0,40913
	(0,671)		(0,909)
RACE	0,00089		-0,36392
	(0,004)		(-0,813)
CLERICAL			0,36774
			(0,499)
MAINT			-0,24081
			(-0,263)
CRAFTS			1,10860
			(0,141)
sq_EDUC	-0,01792	-0,01109	-0,02219
	(-2,874)	(-2,506)	(-2,038)
sq_EXPER	-0,00085		-0,00081
	(-1,098)		(-0,705)
sq_AGE	-0,00015	-0,00007	-0,00024
	(-0,633)	(-2,184)	(-0,756)

ED_GEN	0,06650	0,02960	0,06995
	(2,191)	(2,943)	(1,485)
ED_RACE			0,03682
			(0,658)
ED_CLER	-0,06110	-0,06169	-0,07611
	(-5,843)	(-6,525)	(-1,504)
ED_MAINT	-0,22790	-0,13896	-0,20945
	(-3,523)	(-3,119)	(-1,605)
ED_CRAFT	-0,11688	-0,10726	-0,12452
	(-3,857)	(-5,504)	(-1,826)
AGE_GEN	-0,01081		-0,01507
	(-1,492)		(-1,533)
AGE_RACE	0,00721		0,01043
	(1,127)		(1,044)
AGE_CLER			-0.00409
			(-0,426)
AGE_MAIN	0.02066	0,00758	0,02550
	(2,554)	(1,700)	(2,102)
AGE_CRFT	0,01059	0,01152	0,01145
	(1,945)	(3,649)	(0,968)
EXP_GEN			-0,00475
			(-0,268)
EXP_RACE	-0,02525		-0,02286
	(-1,881)		(-0,928)
EXP_CLER			-0,00771
			(-0,375)
EXP_MAIN			-0,00185
			(-0,068)
EXP_CRFT	0,02402		0,01837
	(2,058)		(0,975)
ESS	0,61473	0,78302	0,574710
$\overline{R}^2$	0,790	0,789	0,733
d.f	30	38	22
SGMASQ	0,020491*	0,020606	0,026123
AIC	0,027245	0,025036*	0,035307
FPE	0,028437	0,025231*	0,040518
HQ	0,035988	0,029413*	0,052436

SCHWARZ	0,056738	0,038283*	0,100135
SHIBATA	0,022275*	0,023155	0,024655
GCV	0,033469	0,026571*	0,058184
RICE	0,055885	0,029001*	không xác định

Lưu ý: dấu * ký hiệu mô hình là "tốt nhất" đối với tiêu chí đó. Giá trị trong ngoặc đơn là các trị thống kê *t*.

Chúng ta lưu ý một vài hệ số có trị thống kê *t* rất thấp, tức là không ý nghĩa. Những trị này sẽ được loại bỏ dần dần cho đến khi chỉ thu được một mô hình mà tất cả các hệ số đều có ý nghĩa ở 10 phần trăm hoặc thấp hơn (Xem Phần Thực Hành Máy Tính 7.5). Kết quả được thể hiện ở Mô hình 2 trong bảng. Cuối cùng, Mô hình 3 là một đặc trưng "bồn rửa chén", trong đó bao gồm tất cả các số hạng bậc hai và số hạng tương tác. Ai đó có thể hoài nghi rằng Mô Hình 3 bị vi phạm nặng nề do vấn đề đa cộng tuyến, mà thường thường nó có xu hướng làm cho các hệ số trở nên mất ý nghĩa. Dưới hình thức dùng trị thống kê để lựa chọn mô hình và mức ý nghĩa của các hệ số hồi qui, Mô Hình 2 rõ ràng là tốt hơn và được chọn là mô hình cuối cùng cho việc diễn dịch.

DIỄN DỊCH KẾT QUẢ Các biến trong mô hình giải thích 79 phần trăm sự thay đổi trong lôgarít của WAGE. Đối với một nghiên cứu chéo, điều này khá tốt. Bây giờ chúng ta xem xét những tác động cận biên của từng yếu tố riêng rẽ.

Học vấn: Số năm đi học (trên lớp tám) quan trọng trong việc giải thích về lương. Nó minh họa ý nghĩa phi tuyến và nó có ý nghĩa tương tác với giới tính và tình trạng nghề nghiệp. Tác động cục bộ là

$$\widehat{D\ln(WAGE)} / DEDUC = 0{,}221 - 0{,}022\ EDUC + 0{,}030\ GENDER$$
$$- 0{,}062\ CLERICAL - 0{,}139\ MAINT - 0{,}107\ CRAFTS$$

Khá thú vị khi nhận thấy rằng tác động cận biên của việc đi học giảm theo số năm đi học. Hay nói cách khác, tiền lương *thêm vào* cho một năm đi học *tăng thêm*, về mặt trung bình, là thấp hơn cho một người đã có trình độ học vấn ở mức cao khi so sánh với một người ít học hơn. Do vậy có hiện tượng lợi nhuận "giảm dần" theo số năm đi học. Giữa nam nhân viên và nữ nhân viên, mà họ đều có cùng những đặc tính khác, thì nam nhân viên kỳ vọng kiếm được trung bình 3 phần trăm nhiều hơn nữ nhân viên cho mỗi năm đi học nhiều hơn. Loại công việc tương tác một cách ý nghĩa với học vấn. Khi so sánh với những nhân viên chuyên nghiệp (nhóm điều khiển), một năm đi học tăng thêm có ý nghĩa 6,2 phần trăm lương ít hơn đối với những nhân viên văn phòng, 13,9 phần trăm ít hơn cho nhân viên bảo trì, và 10,7 phần trăm ít hơn đối với thợ thủ công.

Kinh nghiệm: Không có gì ngạc nhiên khi số năm kinh nghiệm cho một công việc nào đó có tác động tích cực lên tiền lương. Tuy nhiên, không có hiện tượng lợi nhuận giảm dần một cách có ý nghĩa nào cũng như không có bất kỳ tương tác nào với các biến giả. Một năm kinh nghiệm tăng thêm có ý nghĩa mức lương trung bình tăng thêm chỉ có 1,8 phần trăm.

Độ tuổi: Độ tuổi của một nhân viên có lợi nhuận giảm dần ý nghĩa, nó chỉ ra rằng với các yếu tố khác là như nhau, một người đứng tuổi hơn sẽ có mức lương trung bình thấp hơn. Tác động cận biên là

$$\widehat{\text{Dln(WAGE)}} / \text{DAGE} = -0{,}00014\ \text{AGE} + 0{,}00758\ \text{MAINT} + 0{,}01152\ \text{CRAFTS}$$

Anh hưởng của độ tuổi phần nào bù đắp cho những nhân viên bảo trì và thợ thủ công nhưng không cho những loại nhân viên khác.

Giới tính: Từ Bảng 7.4 ta thấy tác động cục bộ của giới tính phụ thuộc vào số năm đi học. Hệ số ước lượng là 0,02960 EDUC, và phương sai ước lượng tương ứng là $(0{,}0101\ \text{EDUC})^2$. Dấu dương cho biết một sự chênh lệch về giới tính có ý nghĩa thực sự hiện hữu (với thu nhập của người nhân viên nam có thể so sánh được ở mức trung bình là cao hơn) và nó cũng cho biết khoảng cách tăng lên với số năm đi học. Do đó những người phụ nữ có học vấn cao có mức lương trung bình thấp hơn một cách bất cân đối so với nam giới cũng với đặc tính như vậy. Tác động cận biên ước lượng có thể được xác định như sau (xem lại kết quả trong phần trước):

$$100\{\exp[\widehat{\beta\,\text{EDUC}} - \tfrac{1}{2}\text{Var}(\widehat{\beta\,\text{EDUC}})] - 1\} = 100\{\exp[\widehat{\beta\,\text{EDUC}} - \tfrac{1}{2}\widehat{\text{EDUC}^2}\text{Var}(\hat{\beta})] - 1\}$$

Đối với $\hat{\beta} = 0{,}02960$ và EDUC = 4 và 8 (trung học và đại học), những tác động này lần lượt là 12,5 phần trăm và 26,3 phần trăm, chúng cao một cách khó hiểu.

Sắc tộc: Không có các biến về sắc tộc nào là có ý nghĩa, cho thấy không có sự khác biệt có ý nghĩa về lương giữa các sắc tộc.

Loại công việc: Những nhân viên trong mẫu thuộc vào bốn nhóm nghề nghiệp khác nhau. Nhóm điều khiển bao gồm những nhân viên chuyên nghiệp, và những nhóm khác là nhân viên văn phòng, bảo trì và thợ thủ công. Những tác động riêng phần được ước lượng cho lôgarít của tiền lương cho từng nhóm không điều khiển là

Nhân viên văn phòng:	- 0,062 EDUC
Bảo trì:	- 0,139 EDUC + 0,008 AGE
Thợ thủ công:	- 0,107 EDUC + 0,012 AGE

Như đã đưa ra trước đây, loại công việc và học vấn tương tác với nhau rất mạnh. Tuổi tác có những tác động tích cực ý nghĩa lên nhân viên bảo trì và thợ thủ công nhưng không tác động lên những nhóm khác.

Đối với một ứng dụng trong thế giới thực của mô hình log-tiền lương, tham khảo bài viết của Tansel được trích dẫn ở Chương 6.

❶ 7.5 Ước Lượng Những Tác động Mùa

Một ví dụ khác về việc sử dụng biến giả xuất hiện trong ước lượng tác động mùa của các biến độc lập. Xem xét quan hệ $E = a + bT + u$, đã được giới thiệu trước đây, giữa việc tiêu thụ điện năng và nhiệt độ. Trong mùa hè, khi nhiệt độ tăng, nhu cầu máy lạnh sẽ đẩy việc tiêu thụ điện năng lên. Do vậy chúng ta kỳ vọng b có dấu dương, cho ra một quan hệ dương giữa E và T. Tuy nhiên, vào mùa đông, khi nhiệt độ tăng (từ 20 độ lên 40 độ), nhu cầu cho việc sưởi ấm nhà trở nên thấp hơn và từ đó chúng ta mong đợi b có dấu âm về mùa đông, cho ra một quan hệ âm giữa E và T. Bằng cách nào chúng ta có thể ghi nhận được tác động lên E của biến định tính "mùa" có bốn loại: xuân, hạ, thu, đông? Việc này thực hiện được bằng cách xác định ba biến giả; được gọi là: **biến giả theo**

mùa. Như đã giải thích trước đây, chúng ta không định nghĩa bốn biến giả để tránh hiện tượng đa cộng tuyến hoàn hảo. Mùa thu được sử dụng làm mùa điều khiển:

$$D_1 = \begin{cases} 1 & \text{Nếu là mùa đông} \\ 0 & \text{Nếu là mùa khác} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{Nếu là mùa xuân} \\ 0 & \text{Nếu là mùa khác} \end{cases}$$

$$D_3 = \begin{cases} 1 & \text{Nếu là mùa hạ} \\ 0 & \text{Nếu là mùa khác} \end{cases}$$

Bây giờ đặt $a = a_0 + a_1D_1 + a_2D_2 + a_3D_3$ và $b = b_0 + b_1D_1 + b_2D_2 + b_3D_3$. Đặc trưng tổng quát thu được bằng cách thay thế những biểu thức này vào quan hệ giữa E và T:

$$E = a_0 + a_1D_1 + a_2D_2 + a_3D_3 + b_0T + b_1D_1T + b_2D_2T + b_3D_3T + u \qquad \textbf{(7.23)}$$

Những mô hình ước lượng cho từng mùa như sau (Hình 7.4 minh họa điều này):

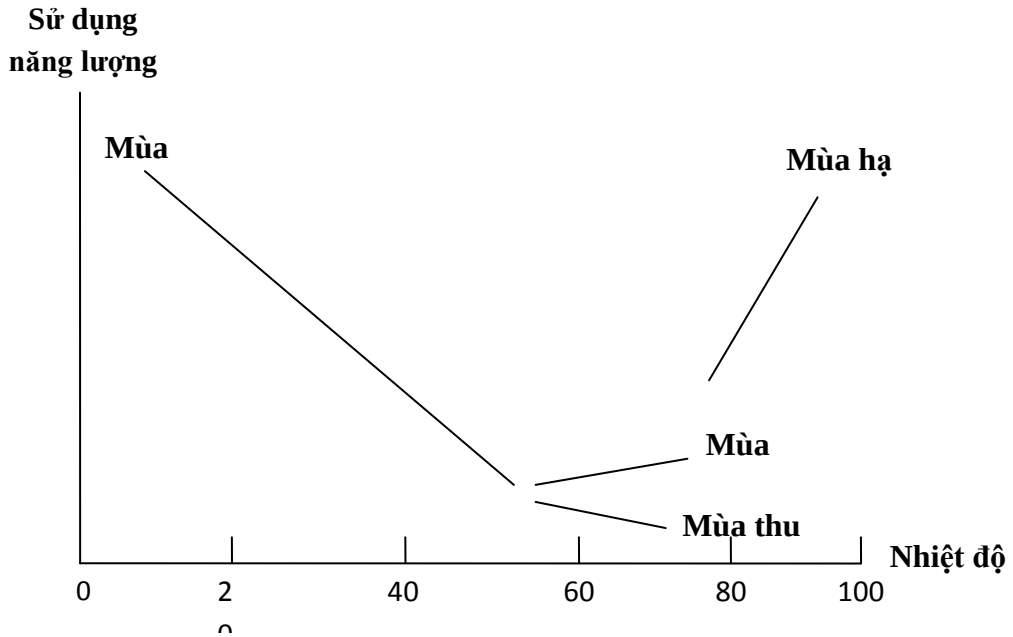
Thu: $\hat{E} = \hat{\alpha}_0 + \hat{\beta}_0 T$

Đông: $\hat{E} = (\hat{\alpha}_0 + \hat{\alpha}_1) + (\hat{\beta}_0 + \hat{\beta}_1)T$

Xuân: $\hat{E} = (\hat{\alpha}_0 + \hat{\alpha}_2) + (\hat{\beta}_0 + \hat{\beta}_2)T$

Hạ: $\hat{E} = (\hat{\alpha}_0 + \hat{\alpha}_3) + (\hat{\beta}_0 + \hat{\beta}_3)T$

⭘ Hình 7.4 Ví Dụ Về Yếu Tố Mùa



a_1 là độ lệch của hệ số tung độ gốc mùa đông so với hệ số tung độ gốc của mùa thu, và b_1 là độ lệch của hệ số độ dốc mùa đông so với hệ số độ dốc của mùa thu. Có thể thực hiện nhiều kiểm định đối với những mô hình này. Ví dụ, giả thuyết hợp lý là không có sự khác biệt trong quan hệ giữa mùa thu và mùa xuân. So sánh các phương trình của mùa thu và mùa xuân, giả thuyết hàm ý rằng $a_2 = b_2 = 0$. Điều này được kiểm định bằng kiểm định Wald trong đó Phương trình (7.23) là mô hình không giới hạn và mô hình giới hạn là

$$E = a_0 + a_1D_1 + a_3D_3 + b_0T + b_1D_1T + b_3D_3T + u \qquad \textbf{(7.24)}$$

Một số nhà điều tra thích giữ lại những biến giả mùa trong mô hình ngay cả khi nếu một vài biến trong số các biến giả đó là không có ý nghĩa. Tuy nhiên, sự ưa thích này không rõ ràng vì các biến giả dư ra có thể làm giảm một cách không cần thiết độ chính xác của các thông số khác. Nếu ta tìm thấy một vài mùa là như nhau (ví dụ mùa thu và mùa xuân), thì có thể hợp nhất chúng thành một mùa bằng cách định nghĩa một biến giả $D = 1$ cho mùa thu và mùa xuân, và bằng 0 cho các mùa còn lại. Để kiểm định giả thuyết cho rằng tất cả các mùa có cùng quan hệ, điều kiện là $a_1 = a_2 = a_3 = b_1 = b_2 = b_3 = 0$. Mô hình không giới hạn sẽ là (7.23), và mô hình giới hạn sẽ là $E = a_0 + b_0T + u$. Bậc tự do của trị thống kê F là 6 cho tử số và số quan sát trừ đi 8 cho mẫu số (giải thích tại sao). Chúng ta sẽ mong đợi việc bác bỏ giả thuyết này vì chúng ta đã chỉ ra trước đây là quan hệ giữa E và T sẽ đồng biến trong mùa hạ và nghịch biến trong mùa đông.

❶ BÀI THỰC HÀNH 7.8

Làm lại phân tích trước và sử dụng mùa hạ thay vì mùa thu làm mùa điều khiển.

Thay vì sử dụng ba biến giả, giả sử chúng ta định nghĩa một biến giả như sau:

$$D = \begin{cases} 1 & \text{Nếu là mùa thu} \\ 2 & \text{Nếu là mùa đông} \\ 3 & \text{Nếu là mùa xuân} \\ 4 & \text{Nếu là mùa hạ} \end{cases}$$

Kế đến giả sử rằng $\alpha = \alpha_0 D$ và $\beta = \beta_0 D$. Mô hình bây giờ trở thành

$$E = \alpha_0 D + \beta_0 DT + u$$

và các mối quan hệ được ước lượng bây giờ trở thành

Mùa Thu: $\hat{E} = \hat{\alpha}_0 + \hat{\beta}_0 T$

Mùa Đông: $\hat{E} = 2\hat{\alpha}_0 + 2\hat{\beta}_0 T$

Mùa Xuân: $\hat{E} = 3\hat{\alpha}_0 + 3\hat{\beta}_0 T$

Mùa Hè: $\hat{E} = 4\hat{\alpha}_0 + 4\hat{\beta}_0 T$

Phương pháp này so với phương pháp ba mùa mà chúng ta đã xem xét trước đây như thế nào? Xin lưu ý rằng mặc dù các phương trình cho mùa thu là đồng nhất, nhưng những phương trình khác thì rất khác nhau. Cụ thể là khi so sánh mùa thu với mùa đông, chúng ta thấy rằng sự chênh lệch về tung độ gốc giữa hai mùa này là α_0. Không có lý do tại sao sự chênh lệch này phải bằng với tung độ gốc của mùa thu. Thật ra, sự chênh lệch về tung độ gốc giữa bất kỳ hai mùa kế tiếp nhau nào cũng bằng α_0. Do đó chúng ta giả sử rằng sự dịch chuyển tung độ gốc là như nhau qua suốt các mùa kế tiếp nhau. Tương tự, sự chênh lệch về độ dốc giữa hai mùa kế nhau luôn bằng β_0. Giả thiết này là quá hạn chế và hầu như khó bảo đảm được trong trường hợp tổng quát. Vì vậy, phương pháp thay thế sẽ đưa đến đặc trưng sai mô hình nghiêm trọng và nên được bỏ qua nhường chỗ cho phương pháp tổng quát với ba biến giả đã được trình bày ở những phần trước.

Xem Phần 7.8 về một ứng dụng thực tế của việc mô hình hóa các tác động theo mùa.

❶ 7.6 Kiểm Định Sự Thay Đổi Về Cấu Trúc

Mối quan hệ giữa các biến phụ thuộc và độc lập có thể có một **sự thay đổi về cấu trúc** (còn được gọi là **sự bất ổn định về cấu trúc** hay **những gián đoạn về cấu trúc**); có nghĩa là, mối quan hệ có thể thay đổi từ thời đoạn này sang thời đoạn khác. Ví dụ như, giả sử *C* là lượng tiêu thụ gas ở Hoa Kỳ trong một thời đoạn cho trước và các biến độc lập là giá bán (*P*) và thu nhập (*Y*). Đã có ba thời đoạn trong suốt khoảng thời gian 1970-2000 khi giá gas tăng một cách trầm trọng, có thể gây ra những thay đổi trong mô hình hành vi tiêu thụ gas. Thay đổi đầu tiên xảy ra vào năm 1974 ngay sau khi tập đoàn OPEC (Tổ chức của các quốc gia xuất khẩu dầu hỏa) tuyên bố kiểm soát giá dầu trên thế giới. Đợt thay đổi thứ hai xảy ra vào năm 1979, ngay sau cuộc cách mạng ở Iran. Thay đổi cuối cùng xảy ra vào năm 1990 khi Iraq đánh chiếm Kuwait. Cũng có lý khi chúng ta kỳ vọng rằng các độ co giãn của giá bán và thu nhập đối với lượng tiêu thụ gas khác nhau qua bốn thời đoạn được phân chia bằng các năm kể trên. Kiểm định thống kê đối với thay đổi về cấu trúc được gọi là **Kiểm định Chow** (sau khi Gregory Chow [1960] lần đầu tiên công bố kỹ thuật này). Phần này trình bày

hai phương pháp kiểm định đối với thay đổi về cấu trúc. Phương pháp thứ nhất bao gồm việc chia mẫu thành hai hay nhiều nhóm, ước lượng mô hình một cách riêng biệt đối với từng thời đoạn và với cả mẫu chung lại, và sau đó xây dựng một trị thống kê F sử dụng để tiến hành kiểm định. Ở phương pháp thứ hai, chúng ta sử dụng các biến giả.

Kiểm định dựa trên việc phân cắt mẫu (Kiểm định Chow)

Giả sử chúng ta muốn kiểm định xem có một sự thay đổi về cấu trúc hay không vào thời điểm $t = n_1$. Thủ tục sẽ là phải chia mẫu gồm n quan sát thành hai nhóm – nhóm 1 gồm n_1 quan sát đầu tiên và nhóm 2 gồm những quan sát còn lại $n_2 = n - n_1$. Ước lượng mô hình một cách riêng biệt (với k hệ số hồi qui) đối với từng nhóm một và tính toán tổng số dư bình phương ESS_1 và ESS_2. Do đó, tổng các bình phương không giới hạn được tính bằng $ESS_U = ESS_1 + ESS_2$. Khi lấy số này chia cho σ^2, kết quả sẽ có phân phối chi-square với bậc tự do d.f. là $n_1 – k + n_2 – k = n – 2k$, bởi vì việc ước lượng mô hình một cách riêng biệt ngụ ý rằng mỗi phương trình có k hệ số hồi qui. Kế đến giả sử rằng các hệ số hồi qui là như nhau trước và sau thời đoạn n_1 (mà nó sẽ làm tăng lên k ràng buộc). Ước lượng mô hình lần nữa nhưng với chung cả mẫu, và thu được giá trị ESS_R. Trị thống kê kiểm định phù hợp bây giờ là

$$F_c = \frac{(ESS_R - ESS_1 - ESS_2) \div k}{(ESS_1 + ESS_2) \div (n - 2k)}$$

Thủ tục kiểm định là để bác bỏ *giả thuyết không* rằng không có thay đổi về cấu trúc nào nếu F_c vượt quá giá trị $F^*_{k,\, n-2k}$, điểm nằm trên phân phối F với bậc tự do d.f. là k và $n – 2k$ mà vùng từ đó tính sang bên phải bằng với mức ý nghĩa. Một giả thiết quan trọng đằng sau kiểm định này là các phương sai sai số của hai mẫu là như nhau.

Kiểm Định Dựa Trên Các Biến Giả

Kiểm định cũng có thể được tiến hành bằng cách sử dụng kỹ thuật dùng biến giả được giới thiệu trong chương này (xem chi tiết hơn trong sách của Franklin Fisher, 1970). Phương pháp này được minh họa ở đây cho lượng tiêu thụ gas vừa được mô tả (chỉ áp dụng đối với các mốc thời gian năm 1974 và 1979).

Mô hình cơ bản là

$$\ln C = \alpha + \beta \ln P + \gamma \ln Y + u$$

Đây là một mô hình log-hai lần mà trong đó β là độ co giãn về giá và γ là độ co giãn về thu nhập. Chúng ta định nghĩa hai biến giả như sau (1974.1 đề cập đến quí một của năm 1974; các quí khác được biểu diễn tương tự):

$$D_1 = \begin{cases} 1 & \text{đối với thời đoạn 1974.1 trở về sau} \\ 0 & \text{đối với thời đoạn khác} \end{cases}$$

$$D_2 = \begin{cases} 1 & \text{đối với thời đoạn 1979.1 trở về sau} \\ 0 & \text{đối với thời đoạn khác} \end{cases}$$

Xin lưu ý rằng đối với tất cả các thời đoạn từ 1979.1, cả D_1 và D_2 đều bằng 1. Để kiểm định xem các cấu trúc cho ba thời đoạn (từ trước đến 1974.1, 1974.1 đến 1978.4, và từ 1979.1 trở về sau) đều khác nhau, đặc trưng mẫu phải được giả sử như sau:

$$\alpha = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 \qquad \beta = \beta_0 + \beta_1 D_1 + \beta_2 D_2 \qquad \gamma = \gamma_0 + \gamma_1 D_1 + \gamma_2 D_2$$

Thay thế những thông số này vào phương trình (7.26), chúng ta có được mô hình không giới hạn

$$\begin{aligned} \ln C &= \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + (\beta_0 + \beta_1 D_1 + \beta_2 D_2)\ln P + (\gamma_0 + \gamma_1 D_1 + \gamma_2 D_2)\ln Y + u \\ &= \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \beta_0 \ln P + \beta_1 (D_1 \ln P) + \ldots + u \end{aligned}$$

Để ước lượng giá trị này, trước hết chúng ta tạo những biến mới $Z_1 = D_1 \ln P$, $Z_2 = D_2 \ln P$, $Z_3 = D_1 \ln Y$, và $Z_4 = D_2 \ln Y$. Kế đến chúng ta hồi qui $\ln C$ theo một hằng số, D_1, D_2 , $\ln P$, Z_1, Z_2 , $\ln Y$, Z_3, và Z_4. Các mô hình đã được ước lượng là:

Từ trước đến 1974.1: $\widehat{\ln C} = \hat{\alpha}_0 + \hat{\beta}_0 \ln P + \hat{\gamma}_0 \ln Y$

1974.1 – 1978.4 : $\widehat{\ln C} = \hat{\alpha}_0 + \hat{\alpha}_1 + (\hat{\beta}_0 + \hat{\beta}_1) \ln P + (\hat{\gamma}_0 + \hat{\gamma}_1) \ln Y$

1979.1 về sau: $\widehat{\ln C} = \hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 + (\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2) \ln P + (\hat{\gamma}_0 + \hat{\gamma}_1 + \hat{\gamma}_2) \ln Y$

Bằng cách so sánh những quan hệ này, chúng ta có thể kiểm định một loạt các giả thuyết khác nhau. Chẳng hạn như, giả thuyết rằng $\alpha_1 = \alpha_2 = \beta_1 = \beta_2 = \gamma_1 = \gamma_2 = 0$ cho thấy không có thay đổi về cấu trúc nào. Một kiểm định t đối với β_2 sẽ kiểm định xem độ co giãn về giá có không đổi trong thời đoạn từ 1974.1 – 1978.4 và 1979.1 về sau. Nhiều giả thuyết khác còn để lại ở dạng bài tập thực hành.

Phương pháp dùng biến giả có một thuận lợi hơn so với việc chia cắt mẫu; nói cách khác, chúng ta có thể kiểm định, nếu chúng ta mong muốn như vậy, chỉ một vài hệ số hồi qui đối với thay đổi về cấu trúc hơn là quan hệ toàn bộ, như phương pháp được trình bày sau này.

Chúng ta thấy từ trong mô hình không giới hạn đối với $\ln C$ rằng nếu tung độ gốc cũng như tất cả hệ số độ dốc được cho phép khác nhau qua các thời đoạn, thì số lượng các số hạng tương tác, và do đó cả số lượng các hệ số hồi qui để ước lượng có thể lớn. Điều này sẽ dẫn đến việc mất đi một vài bậc tự do và một sự giảm đi sức mạnh của các kiểm định. Vì vậy một nhà nghiên cứu thường được khuyên là phải cảnh giác với sự phát triển của các biến giả mà do đó dẫn đến tình trạng "khai thác dữ liệu" (data mining). Một phương pháp hữu ích để thiết lập nên một mô hình cơ bản không có biến giả và khi đó sẽ sử dụng kiểm định nhân tử Lagrange đã được mô tả ở Chương 6 để kiểm định xem các biến giả thêm vào và các số hạng tương tác có nên đưa vào mô hình hay không.

⊙ BÀI TOÁN THỰC HÀNH 7.9+

Mô tả cách kiểm định giả thuyết cho rằng độ co giãn về thu nhập không hề thay đổi trong ba thời đoạn.

⊙ BÀI TOÁN THỰC HÀNH 7.10

Mô tả cách kiểm định giả thuyết cho rằng tung độ gốc là như nhau đối với tất cả các thời đoạn.

⊙ BÀI TOÁN THỰC HÀNH 7.11

Giả sử biến giả D_3, được định nghĩa ở đây, được sử dụng thay cho D_1:

$$D_3 = \begin{cases} 1 & \text{đối với thời đoạn 1974.1 – 1978.4} \\ 0 & \text{đối với thời đoạn khác} \end{cases}$$

Làm lại phân tích có trước với giả sử này. Mối quan hệ giữa các hệ số đạt được theo cách này và những hệ số đạt được trước đó là gì?

Ứng Dụng: Thay Đổi Về Cấu Trúc Trong Lực Lượng Lao Động Nữ Các Tỷ Lệ Tham Dự

Trong Phần 4.7, chúng ta đã sử dụng DATA 4-5 và đã ước lượng một mô hình đối với tỷ lệ tham gia của lực lượng lao động nữ (WLFP). Tập dữ liệu đó được dùng cho năm 1990 và cho 50 tiểu bang. Trong DATA7-4 chúng ta có dữ liệu cho cả năm 1990 và 1980. Dữ liệu cho năm 1990 bị "sắp đặt" bên dưới các dữ liệu cho năm1980 với một cột mới được thêm vào, tức là cột D90. Đây là một biến giả có giá trị bằng 1 cho năm 1990 bằng không cho năm 1980. Biến này sẽ khá thú vị để kiểm tra xem có một sự thay đổi về cấu trúc trong mối quan hệ giữa WLFP và các yếu tố quyết định của nó hay không. Để có một thảo luận hoàn chỉnh về các biến độc lập và các tác động kỳ vọng của chúng lên WLFP, hãy xem Phần 4.7 trước. Bởi vì mối quan hệ toàn bộ có thể đã dịch chuyển giữa năm 1980 và 1990, chúng ta cần phải phát ra tất cả những số hạng tương tác bằng cách nhân D90 với từng biến độc lập. Như vậy, chúng ta sẽ phát được các biến như là D90YF, vốn là kết quả của D90 nhân với YF, và làm tương tự đối với các biến khác. Bảng 7.6 có một phần kết quả thu được từ máy tính (thu được từ Phần Thực hành trên máy tính 7.6). Xin lưu ý rằng kiểm định Chow đối với không có thay đổi về cấu trúc bị bác bỏ ngay ở những mức dưới 0,01 phần trăm. Mô hình tổng quát với tất cả những số hạng tương tác có một giá trị đã hiệu chỉnh của R^2 là 0,833, mà nó cao hơn giá trị đo được tương ứng (0,746) đối với mô hình 1990 trong Phần 4.7. Tuy nhiên, chúng ta có thể ngờ rằng hầu như có một lượng đáng kể tính đa cộng tuyến giữa các biến. Do đó chúng ta loại bỏ các biến có hệ số không ý nghĩa, nhưng phải bỏ từng biến một. Mô hình ước lượng sau cùng được cho bởi phương trình sau, với giá trị *p* trong ngoặc đơn:

$$\begin{aligned} \widehat{\text{WLFP}} = {} & 47{,}637 + \underset{(<0{,}01)}{0{,}00478\ \text{YF}} - \underset{(<0{,}01)}{0{,}00405\ (\text{D90 x YF})} + \underset{(<0{,}01)}{0{,}275\ \text{EDUC}} \\ & - \underset{(<,01)}{1{,}061\ \text{UE}} - \underset{(0{,}085)}{0{,}569\ (\text{D90 x UE})} - \underset{(0{,}051)}{0{,}207\ \text{MR}} \\ & + \underset{(0{,}015)}{0{,}126\ (\text{D90 x MR})} + \underset{(0{,}038)}{0{,}282\ \text{DR}} - \underset{(<0{,}01)}{0{,}078\ \text{URB}} - \underset{(<0{,}01)}{0{,}111\ \text{WH}} \end{aligned}$$

$$\bar{R}^2 = 0{,}842 \qquad \text{d.f.} = 89 \qquad \hat{\sigma} = 2{,}192$$

Để đạt được các mối liên hệ riêng biệt đối với hai thời đoạn, trước hết chúng ta cho D90 bằng không, như vậy sẽ có được phương trình đối với năm 1980 như sau:

$$\begin{aligned} \widehat{\text{WLFP}} = {} & 47{,}637 + 0{,}00478\ \text{YF} + 0{,}275\ \text{EDUC} - 1{,}061\ \text{UE} - 0{,}207\ \text{MR} \\ & + 0{,}282\ \text{DR} - 0{,}078\ \text{URB} - 0{,}111\ \text{WH} \end{aligned}$$

⭘ Bảng 7.6 Kết quả Từng phần đối với Ứng dụng Thay đổi về cấu trúc trong Phần 7.6

[Đầu tiên hồi qui WLFP theo một hằng số, YF, YM, EDUC, UE, MR, DR, URB, và WH, và thực hiện một kiểm định Chow đối với thay đổi về cấu trúc.]
Kiểm định Chow đối với gián đoạn về cấu trúc tại điểm quan sát thứ 50:
F(9, 82) = 6,903514 với giá trị p là 0,000000 (có nghĩa là rất nhỏ)
[Lưu ý rằng *giả thuyết không* về sự không có thay đổi về cấu trúc là bị bác bỏ hoàn toàn. Kế đến, ước lượng một mô hình với các biến gốc ban đầu cộng với các số hạng tương tác.]

	VARIABLE	COEFFICIENT	STDERROR	T STAT	2 Prob (t > |T|)	
0)	const	50.8808	11.6760	4.358	0.000038	***
10)	D90	- 6.3712	14.9234	- 0.427	0.670549	
2)	YF	0.0045	0.0012	3.757	0.000321	***
11)	D90YF	- 0.0035	0.0013	- 2.770	0.006939	***
3)	YM	- 0.0000111	0.0005489	- 0.020	0.983935	
12)	D90YM	- 0.0001633	0.0006339	- 0.258	0.797400	
4)	EDUC	0.2779	0.0674	4.121	0.000090	***
13)	D90EDUC	0.0072	0.1177	0.061	0.951319	
5)	UE	- 1.1191	0.2917	- 3.836	0.000244	***
14)	D90UE	- 0.4915	0.4365	- 1.126	0.263485	
15)	D90MR	0.1461	0.2427	0.602	0.548850	
6)	MR	- 0.2243	0.1636	- 1.371	0.174124	
7)	DR	0.2268	0.1876	1.209	0.230234	
16)	D90DR	0.2106	0.3268	0.645	0.521040	
8)	URB	- 0.0691	0.0317	- 2.180	0.032124	**
17)	D90URB	- 0.0236	0.0469	- 0.503	0.616474	
9)	WH	- 0.1284	0.0351	- 3.654	0.000455	***
18)	D90WH	0.0409	0.0542	0.755	0.452353	

Error Sum of Sq (ESS)	416.0265	Std Err of Resid. (sqmahat)	2.2524
Unadjusted R – squared	0.862	Adjusted R– squared	0.833
F – statistic (17, 82)	30.1406	p-value for F ( )	0.000000

MODEL SELECTION STATISTICS

SGMASQ	5.07349	AIC	5.96303	FPE	5.98672
HQ	7.20924	SCHWARZ	9.53062	SHIBATA	5.65796
GCV	6.18719	RICE	6.50041		

[Bây giờ bỏ từng biến một để thu được mô hình sau cùng với các hệ số có mức ý nghĩa 10%.]

⭘ Bảng 7.6 (Tiếp theo)

	VARIABLE	COEFFICIENT	STDERROR	T STAT	2 Prob (t > |T|)	
0)	const	47.6366	6.5784	7.241	0.000000	***
2)	YF	0.0048	0.0007339	6.512	0.000000	***
11)	D90YF	- 0.0041	0.0006821	- 5.943	0.000000	***
4)	EDUC	0.2751	0.0455	6.045	0.000000	***
5)	UE	- 1.0614	0.2456	- 4.322	0.000040	***
14)	D90UE	- 0.5694	0.3272	- 1.740	0.085324	*
6)	MR	- 0.2073	0.1049	- 1.976	0.051227	*
15)	D90MR	0.1264	0.0510	2.479	0.015066	**
7)	DR	0.2816	0.1337	2.106	0.037986	**

8)	URB	- 0.0785	0.0206	- 3.805	0.000260 ***
9)	WH	- 0.1115	0.0242	- 4.599	0.000014 ***

Mean of dep. var	53.869	S. D. of dep. variable	5.519
Error Sum of Sq (ESS)	427.5756	Std Err of Resid. (sqmahat)	2.1919
Unadjusted R – squared	0.858	Adjusted R– squared	0.842
F – statistic (10, 89)	53.8705	p -value for F ( )	0.000000
Durbin – Watson stat.	1.983	First-order autocorr. coeff	0.007

MODEL SELECTION STATISTICS

SGMASQ	4.80422	AIC	5.32792	FPE	5.33268
HQ	5.98311	SCHWARZ	7.09599	SHIBATA	5.21642
GCV	5.398	RICE	5.48174		

Mối quan hệ đối với năm 1990 thu được bằng cách cho D90 bằng 1 và kết hợp các số hạng cho các biến giống nhau. Chẳng hạn như, nếu D90 = 1, số hạng cho biến YF cần được kết hợp với số hạng cho D90 x YF. Do đó, mối quan hệ được ước lượng đối với năm1990 là

$$\begin{aligned} WLFP = {} & 47{,}637 + 0{,}00073\,YF + 0{,}275\,EDUC - 1{,}630\,UE - 0{,}081\,MR \\ & + 0{,}282\,DR - 0{,}078\,URB - 0{,}111\,WH \end{aligned}$$

Mô hình sau cùng giải thích được 84,2 phần trăm của sự biến đổi trong WLFP, điều này tương xứng với dữ liệu chéo. Các tác động của biến EDUC, DR, URB, và WH đã mang các dấu như kỳ vọng và giống nhau đối với năm 1980 và 1990. Tác động cận biên của tỷ lệ kết hôn (MR) có giá trị vào năm 1990 nhỏ hơn ở năm 1980. Một sự gia tăng 1 phần trăm ở MR làm giảm WLFP, trung bình khoảng 0,207 phần trăm vào năm 1980 nhưng chỉ giảm 0,081 phần trăm vào năm 1990. Điều này cho thấy rằng, so với năm 1980, có nhiều phụ nữ sau khi kết hôn ở trong lực lượng lao động hơn. Tác động của tỷ lệ thất nghiệp cũng khác nhau đáng kể giữa hai cuộc điều tra dân số này. Vào năm 1980, tác động cận biên của UE là –1,061, trong khi vào năm 1990 tác động đó là –1,630. Vì vậy, giả thuyết *người công nhân chán nản* đối với năm 1990 mạnh hơn đối với năm 1980. Sự khác nhau trong tác động của thu nhập của phụ nữ (YF) cũng có ý nghĩa — 0,00478 vào năm 1980 so với 0,00073 vào năm 1990 — một sự sụt giảm mạnh về giá trị, mà nguyên nhân của việc này không được rõ ràng. Một cách giải thích có thể có là do tính cộng tuyến gần hoàn hảo giữa YF và D90YF, mà nó làm cho khó đạt được các tác động riêng biệt.

⭘ 7.7 Ví Dụ Thực Nghiệm: Sự Bãi Bỏ Qui Định Vận Tải Mô-Tô

Blair, Kaserman, và McClave (1986) nghiên cứu tác động của việc bãi bỏ qui định về cấu trúc giá của các dịch vụ vận tải nội bộ tiểu bang ở Florida. Sự bãi bỏ qui định này có hiệu lực vào ngày 1 tháng bảy năm 1980, và dữ liệu của các tác giả tập hợp được hơn 27.000 quan sát, bao gồm 10 hãng vận tải và xuyên suốt bốn thời đoạn, một thời đoạn trước khi bãi bỏ qui định. Các tác giả đã giả sử rằng việc cung cấp các dịch vụ vận tải cho một nhà xuất nhập khẩu co giãn về giá rất nhiều theo tốc độ của thị trường. Biến phụ thuộc là ln(PTM), trong đó PTM là giá vận chuyển hàng trên một tấn-dặm theo đơn vị đô-la vào năm 1980. Các biến độc lập định lượng được là: ln(WT), trong đó WT là điểm giữa của các loại trọng lượng khác nhau; PD là giá dầu diesel vào năm 1980 tính theo cent trên một đơn vị gallon; và ln(DIST), trong đó DIST là số dặm đường vận chuyển. Nghiên cứu cũng bao gồm một vài biến giả: ORIGJ bằng 1 khi việc vận chuyển xuất phát từ Jacksonville, ORIGM bằng 1 nếu việc vận chuyển xuất phát từ Miami, $CLASS_i$ (i = 1, 2, 3, 4) biểu thị năm loại vận

chuyển khác nhau, và DEREG bằng 1 trong thời kỳ hậu-bãi bỏ qui định. Mô hình cơ bản được ước lượng như sau, với các trị thống kê *t* trong ngoặc đơn:

$$\widehat{\ln(PTM)} = \underset{(327,44)}{10,1805} + \underset{(6,31)}{0,0305}\,ORIGJ + \underset{(5,28)}{0,0254}\,ORIGM - \underset{(-133,74)}{0,1590}\ln(WT) - \underset{(-196,00)}{0,6398}\ln(DIST) + \underset{(16,21)}{0,2800}\,CLASS1 + \underset{(97,22)}{0,5871}\,CLASS2 + \underset{(150,45)}{0,9086}\,CLASS3 + \underset{(175,82)}{1,0923}\,CLASS4 + \underset{(10,42)}{0,0030}\,PD - \underset{(-35,08)}{0,1581}\,DEREG$$

$$\bar{R}^2 = 0,79$$

Hệ số hồi qui của mối quan tâm cơ bản là hệ số cho DEREG. Hệ số này vừa âm vừa có ý nghĩa ở mức ý nghĩa 1 phần trăm, cho thấy rằng giả thuyết việc bãi bỏ qui định tạo ra một sự giảm đáng kể đối với các tỷ lệ vận tải là đáng thuyết phục. Những điều kiện khác bằng nhau, việc bãi bỏ qui định vận tải nội bộ tiểu bang ở Florida cho ra một sự giảm tỷ lệ trung bình gần 16 phần trăm. Các biến còn lại cũng có ý nghĩa về mặt thống kê ở mức 1 phần trăm và có đúng dấu cho các hệ số.

Các tác giả cũng đã kiểm định sự tương tác giữa các biến giả và một số biến định lượng, cũng như giữa các biến giả với nhau, nhưng với kết quả hỗn hợp. Có thể đọc thêm chi tiết trong bài báo của các tác giả này.

⭘ 7.8 Ứng Dụng: Nhu Cầu Đối Với Một Loại Chất Chống Thấm (Sealant) Sử Dụng Trong Xây Dựng

Một công ty cụ thể làm một hợp chất chống thấm được sử dụng trong công việc đổ bê tông xây dựng và làm đường. Công ty tin rằng một đối thủ cạnh tranh đã tung ra tin đồn về chất lượng của sản phẩm của công ty, gây ra một khoảng mất mát về doanh thu và lợi nhuận trong suốt thời đoạn từ tháng bảy năm 1986 đến tháng mười năm 1988. Công ty đã phát đơn kiện đối thủ cạnh tranh và đòi đền bù thiệt hại. Một nhân chứng chuyên môn làm đại diện cho phía công ty với một thái độ hài hước gọi công ty là công ty Cement Overcome (COI), và bản thân ông ta Rodney Random, nhằm để bảo vệ sự cẩn mật của các chi tiết của phiên tòa.

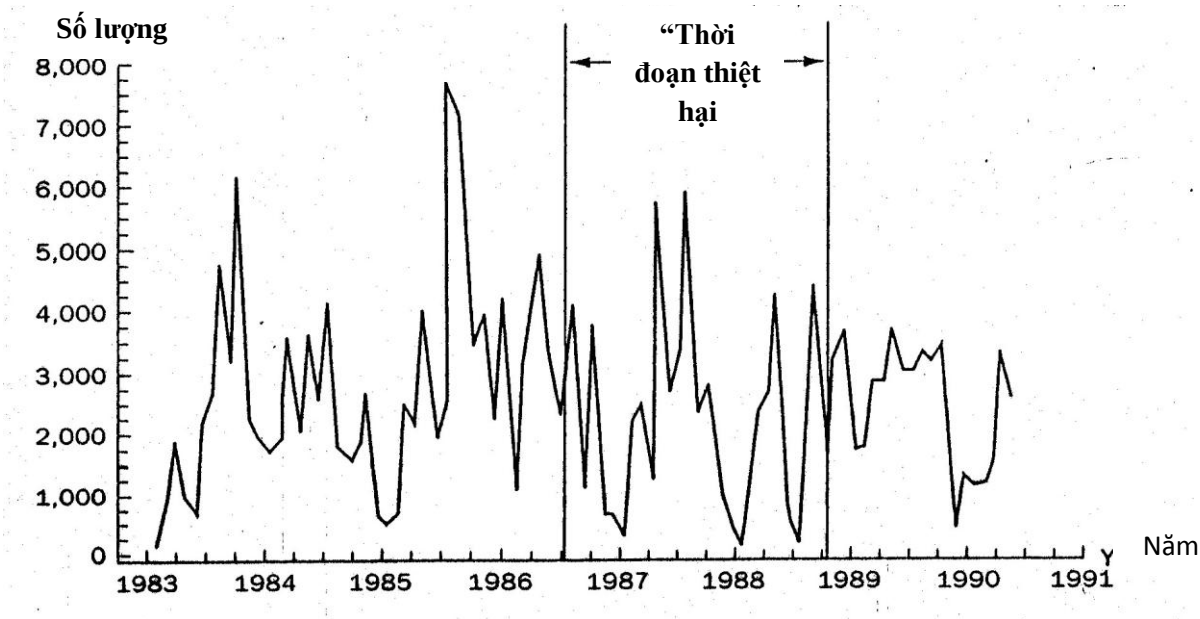
Hình 7.5 là một đồ thị biểu diễn số lượng (theo đơn vị gallon) của chất chống thấm đã được bán bởi COI mỗi tháng từ tháng giêng năm 1983 đến tháng năm năm 1990. Ba dạng đáng quan tâm xuất hiện trên đồ thị. Dạng thứ nhất, có tính chất mùa vụ trong số lượng, và có thể kỳ vọng rằng doanh số tháng giêng thấp một cách đặc trưng và doanh số trong suốt giai đoạn tháng tám-tháng chín nhìn chung là cao. thứ hai, doanh số trung bình thể hiện sự giảm sút trong thời đoạn "thiệt hại" (Tháng bảy năm 1986 – tháng mười năm 1988) và còn giảm hơn nữa trong thời đoạn hậu thiệt hại. Cuối cùng, chiều cao của thời cao điểm doanh số hè đã giảm xuống một cách đều đặn từ thời đoạn này sang thời đoạn khác. Như vậy, dường như có một biểu hiện ban đầu hỗ trợ đối với luận điểm cho rằng doanh số bán thấp hơn trong suốt thời đoạn bị thiệt hại. Thực ra, những thiệt hại vẫn có thể tiếp diễn sau thời đoạn kiện tụng.

Rodney Random có dữ liệu về một số biến có ảnh hưởng đến các chuyến vận chuyển hầu như hàng tháng. DATA 7-5 (xem Phụ lục D) cung cấp dữ liệu hàng tháng của các biến sau đây cho thời đoạn từ tháng giêng năm 1983 đến tháng năm năm 1990:

Q = Số chuyến vận chuyển hợp chất chống thấm sử dụng trong xây dựng, theo đơn vị gallon/tháng
P = Giá bán mỗi gallon, theo đơn vị đô-la
HS = Các địa điểm xuất phát, theo đơn vị ngàn địa điểm

SHC = Chỉ số danh mục của công trình xây dựng đường phố và đường cao tốc
OC = Chỉ số chung của công trình xây dựng tư nhân và công cộng
L = 1 cho thời đoạn từ tháng bảy năm 1986 đến tháng mười năm 1988, khi công ty chịu sự thiệt hại
PL = 1 cho thời đoạn từ tháng mười một năm 1988 trở về sau, thời đoạn hậu thiệt hại.

Hình 7.5 Các chuyến vận chuyển hợp chất chống thấm (gallon/tháng)



Random đã làm một phân tích rất kỹ lưỡng đối với tập dữ liệu, gồm việc thực hiện nhiều thủ tục kiểm định đã được mô tả ở Chương 8, 9, và 10. Ở đây chúng ta trình bày một phần phân tích đã có sửa đổi để minh họa cho sự hữu ích của các biến giả. Điểm khởi đầu là mô hình cơ bản:

$$(A) \qquad Q = \beta_1 + \beta_2 P + \beta_3 HS + \beta_4 SHC + \beta_5 OC + \beta_6 L + \beta_7 PL + u$$

Cần chú ý rằng số hạng L và PL là những biến giả làm dịch chuyển "tung độ gốc". Thời đoạn đầu tiên là sự kiểm soát, và β_6 và β_7 đo lường độ lệch của số hạng không đổi từ thời đoạn cơ bản (chú ý là L chỉ được xác định bằng 1 cho thời đoạn thiệt hại). Các ước lượng OLS của hệ số được cho tiếp theo cùng với giá trị p trong ngoặc đơn (Phần Thực hành trên Máy tính 7.7 có tất cả chi tiết cho việc sử dụng chương trình GRETL để cho ra kết quả như trong phần này).

$$\hat{Q} = \underset{(0,27)}{-2065} - \underset{(0,003)}{301,670P} + \underset{(0,047)}{14,423HS} + \underset{(0,124)}{0,629SHC}$$
$$+ \underset{(0,010)}{33,677OC} - \underset{(0,023)}{1.075,203L} - \underset{(0,223)}{733,934PL}$$

$\bar{R}^2 = 0,354$ d.f. = 82 $\hat{\sigma} = 1,258$

Các dấu của các hệ số hồi qui cho L và PL là âm, cho thấy rằng, tính trung bình, doanh số trong hai thời đoạn sau thấp hơn doanh số trong thời đoạn đầu, ngay cả sau khi có chỉnh sửa đối với

các tác động của các biến giải thích khác như các địa điểm xuất phát, công trình đường cao tốc của bang, và công trình xây dựng nói chung. Tuy nhiên, giá trị *p* đối với hệ số của PL là 0,223 cao lên một cách không chấp nhận được. Lưu ý rằng giá trị *p* đối với hệ số của L chỉ là 0,023, cho thấy các doanh số trung bình thấp hơn một cách đáng kể trong giai đoạn "thiệt hại" khi so sánh với thời đoạn đầu. Tuy nhiên, mô hình chỉ giải thích được có 35,4 phần trăm của những biến động trong các chuyến vận chuyển hàng tháng và có thể sử dụng một cải tiến nào đó trong đặc trưng.

Hình 7.5 có điểm cần lưu ý rằng có một dạng theo mùa trong dữ liệu của các chuyến hàng. Điều này gợi ý cho việc phối hợp các biến giả để giữ lại các tác động theo mùa. Theo đó, 11 biến giả đã được định nghĩa, từng biến một tương ứng cho các tháng từ tháng hai đến tháng mười hai (tháng giêng được bỏ qua để tránh "bẫy biến giả"). Các biến này sau đó được thêm vào Mô hình A, và một mô hình mới (B) đã được ước lượng. Do quá nhiều các số hạng hiện diện, kết quả không được trình bày ở đây, nhưng nó vẫn có thể thu được bằng cách sử dụng Phần Thực hành trên Máy tính 7.7. Người đã thấy rằng hệ số đối với L vẫn còn âm một cách đáng kể, nhưng hệ số đối với PL, mặc dù vẫn còn âm, chỉ âm một cách đáng kể ở mức ý nghĩa 48 phần trăm. Tuy nhiên, nhiều biến giả ở đây thậm chí càng không có ý nghĩa hơn. Chúng ta có thể bỏ qua những biến này và tái ước lượng mô hình để xem ý nghĩa của các biến còn lại có cải thiện hay không. Thay vì làm như vậy, chúng ta đã áp dụng một phương pháp mà nó nêu ngay được vấn đề thiệt hại.

Phân tích ban đầu cho thấy rằng có thể đã có một thiệt hại có ý nghĩa trong doanh số trong suốt thời đoạn thứ hai, và có lẽ ngay cả trong suốt thời đoạn trước đó. Phương cách hợp lý để đạt được một độ đo cho việc thiệt hại có thể có về doanh số bán là loại bỏ dữ liệu của các thời đoạn thiệt hại và hậu thiệt hại. Việc tính luôn chúng vào sẽ gây ảnh hưởng đến các ước lượng, vì thế, điều này chính là câu hỏi mà chúng ta đang cố gắng trả lời. Thủ tục này đã chấp nhận các ước lượng mà mô hình đã sử dụng 42 quan sát đối với thời đoạn 1983.01 – 1986.06. Khi đó chúng ta có thể phát ra các dự báo cho các thời đoạn thiệt hại và hậu thiệt hại và so sánh chúng với những giá trị thực tế đã biết. Nếu các chuyến vận chuyển đã được dự đoán nhiều hơn số chuyến thực tế một cách có hệ thống, thì có một bằng chứng mạnh mẽ về một thay đổi trong cấu trúc và những thiệt hại có ý nghĩa.

Thủ tục mà chúng ta vừa mô tả đã được áp dụng vào dữ liệu của thời đoạn đầu, và một mô hình thứ ba (C) đã được ước lượng bằng cách sử dụng các biến giải thích với số hạng không đổi, P, HS, SHC, OC, và 11 biến giả hàng tháng, nhưng *không kể đến* biến L và PL, cả hai đều bằng không đối với thời đoạn đầu. Như trước kia, các hệ số hồi qui đối với phần nhiều các biến giả không có ý nghĩa, cũng như cho các địa điểm xuất phát (HS). Nhằm để cải thiện tính chính xác của các hệ số còn lại, các biến này đã được loại bỏ và mô hình được tái thiết kế. Các ước lượng cho mô hình "sau cùng" (D) được đưa ra ở đây, với các giá trị *p* trong ngoặc đơn:

$$\underset{}{\hat{Q}} = \underset{(0,34)}{-1,915} - \underset{(0,096)}{1,157\ \text{dummy6}} - \underset{(0,002)}{499,986} + \underset{(0,0004)}{1,896\ \text{SHC}} + \underset{(0,0006)}{51,928\ \text{OC}}$$

$$\bar{R}^2 = 0,513 \qquad \text{d.f.} = 37 \qquad \hat{\sigma} = 1,202$$

Hệ số cho biến giả tháng sáu có ý nghĩa ở mức 9,6 phần trăm, nhưng tất cả những hệ số khác (không kể số hạng không đổi) có ý nghĩa ở mức dưới 1 phần trăm. Giá trị R^2 hiệu chỉnh tăng lên một cách đáng kể từ giá trị 0,354, nhưng ngay cả mô hình mới hơn cũng chỉ giải thích được một nửa sự biến động trong các chuyến hàng. Điều này có thể bởi vì dữ liệu hàng tháng thường hay thay đổi (nghĩa là thay đổi một lượng hàng lớn) và khó cho việc mô hình.

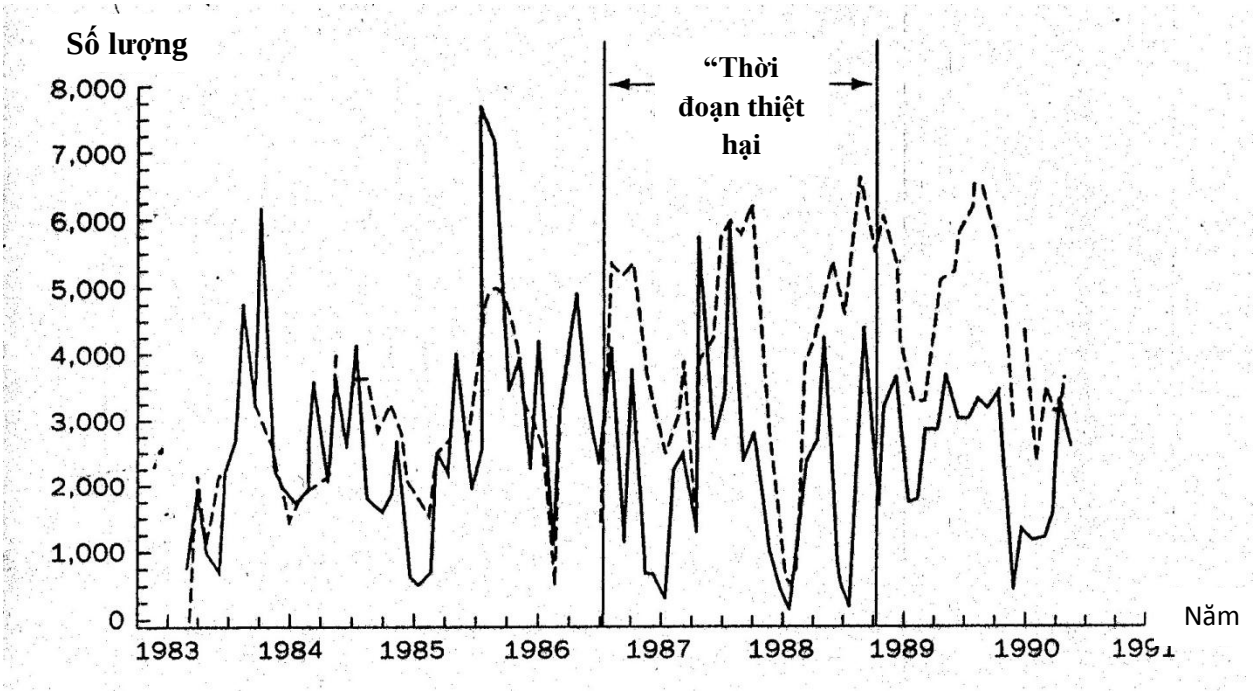
Mô hình D đã được sử dụng tiếp theo để dự báo các chuyến hàng cho thời đoạn 1986.07 – 1988.10 và 1988.11 – 1990.05. Hình 7.6 cho thấy đồ thị của các chuyến vận chuyển thực tế và dự báo đối với cả 89 tháng. (Phần Thực hành trên Máy tính 7.7 tính toán những giá trị bằng số.) Xin lưu ý rằng trong suốt thời đoạn đầu trước khi dư luận viện chứng về COI, mô hình bám sát theo các

giá trị thực tế, ngoại trừ đối với một số ít các giá trị cực đoan. Điều này không gây ngạc nhiên lắm vì OLS cho giá trị tổng sai số của các bình phương thấp nhất và có những sai số mà trung bình tiến tới không. Tuy nhiên, ngược lại, các dự báo cho thời đoạn "thiệt hại" lại cao hơn các giá trị thực tế một cách có hệ thống. Đối với thời đoạn hậu thiệt hại, sự khác biệt được phát biểu nhiều hơn. Đây là bằng chứng mạnh mẽ nhất rằng cấu trúc đã thay đổi thật sự đến sự tổn thất của doanh số và lợi nhuận của COI. Thực tế là những thiệt hại còn tiếp tục diễn ra trong suốt thời kỳ hậu thiệt hại. Các đại lượng đo lường của thiệt hại ròng trong doanh số và thu nhập đã đạt như sau (dấu $\sum$ đề cập đến tổng ròng của các thiệt hại):

$$\text{Doanh số bán} = \sum (\hat{Q}_t - Q_t) = 54.209 + 38.467 = 92.676$$

$$\text{Tổng thu nhập} = \sum P_t (\hat{Q}_t - Q_t) = 481.575 + 335.597 = 817.172$$

Hình 7.6 Số chuyến vận chuyển hợp chất chống thấm theo dự báo và trên thực tế
(Số liệu thực tế là đường nét liền còn số dự báo là đường nét đứt)



Rodney Random đã nộp các ước lượng về thiệt hại của mình (khác với những số liệu chỉ được cung cấp một phần bởi vì vụ kiện tụng chỉ được đề cập đến ở thời đoạn giữa mà thôi) trong một báo cáo chi tiết. Khi vụ việc được đưa ra xét xử, bên bị đơn đã bị áp đảo bởi những phân tích mạnh mẽ của Rodney mà tất cả những tổn thất đối với COI đã được xử lý bên ngoài tòa án và ông không bao giờ có một cơ hội để kiểm chứng. Điều này đã làm cho Rodney thất vọng tràn trề bởi vì ông sẽ phải trả 250 đô-la một giờ cho bằng chứng của ông ta.

Ví dụ này minh họa cho cách thức mà các biến giả có thể hữu dụng trong việc lập mô hình phù hợp cho hành vi thực tế.

7.9 Dự Án Thực Nghiệm

Nếu một dự án thực nghiệm là một yêu cầu trong chương trình học của các bạn, bạn sẽ phải tuân theo những chỉ dẫn trong Phần 4.9, tìm hiểu tổng quan về cơ sở lý thuyết có liên quan và thu thập dữ liệu thích hợp. Bây giờ khi bạn đã học được làm cách nào mà các biến phi tuyến tính và định tính có thể được xử lý trong một bối cảnh hồi qui, hãy chắc chắn để xác định các biến giả phù hợp cũng như các thành phần phi tuyến, nếu cần. Kế đến bạn hãy tạo một tập tin dữ liệu và nhập tất cả những dữ liệu vào máy tính, theo các biến hay theo các quan sát. Sau đó thiết lập mô hình (xem Phần 14.3), phát ra các biến mới có liên quan (Phần 14.4), và tiến hành phân tích dữ liệu sơ bộ (Phần 14.5) để bảo đảm rằng các dữ liệu đã được nhập một cách chính xác và các biến giải thích thay đổi qua các quan sát (như đã trình bày bằng hệ số của đại lượng biến đổi). Bạn thậm chí có thể thử ước lượng một mô hình ban đầu trước.

Các kết quả ước lượng ban đầu thường gây thất vọng vì các hệ số hồi qui có thể không có ý nghĩa hoặc các dấu của chúng có thể không như kỳ vọng (được đề cập phổ biến như là *các dấu sai*). Việc kiểm định chẩn đoán tiếp theo chắc chắn sẽ là cần thiết, nhưng chúng ta vẫn chưa phát biểu được những vấn đề đặc biệt nhất định mà có thể thu được những dạng kiểm định mới cũng như các thủ tục ước lượng. Do đó phân tích thực nghiệm sẽ chỉ hoàn tất được một phần trong giai đoạn này. Bạn chỉ nên tiếp tục việc phân tích sau khi đọc các chương tiếp theo.

Tóm tắt

Trong chương này chúng ta đã xem xét cách thức mà các biến độc lập *định tính* như biến về giới tính, trình độ học vấn hoặc nghề nghiệp, mùa vụ, công cộng hay tư nhân, v.v… có thể được xử lý như thế nào trong một công thức kinh tế lượng. Phương pháp này rất dễ tiếp cận. Đầu tiên chúng ta chọn một trong các lựa chọn của biến định tính xem như là *biến kiểm soát* (chọn lựa này tùy ý). Sau đó chúng ta xác định *các biến giả* (là những biến chỉ có giá trị 1 hoặc 0) đối với mỗi lựa chọn khác, *số lượng các biến giả ít hơn một biến so với số lựa chọn* (để tránh tính chất cộng tuyến hoàn toàn giữa các biến giả). Trong một mô hình có dạng $Y = \alpha + \beta X + u$, nếu số hạng không đổi (α) thay đổi tương ứng với các lựa chọn khác nhau, thì chúng ta giả sử như sau:

$$\alpha = \alpha_0 + \alpha_1 D_1 + \alpha_2 D_2 + \ldots + \alpha_{m-1} D_{m-1}$$

trong đó m là số các lựa chọn đối với các biến giả, và D_1 , D_2 , . . . , D_{m-1} là các biến giả. D_1 sẽ có giá trị bằng 1 đối với các quan sát tùy thuộc vào lựa chọn 1 và 0 đối với tất cả các quan sát khác. Các biến D khác được định nghĩa tương tự. Để dịch chuyển độ dốc (β) đối với các loại biến khác nhau, chúng ta giả sử rằng

$$\beta = \beta_0 + \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_{m-1} D_{m-1}$$

Để ước lượng các hệ số α và β, chúng ta phát ra các biến tương tác $Z_1 = D_1 X$, $Z_2 = D_2$ X, . . . , $Z_{m-1} = D_{m-1}$ X, trong đó X là một biến độc lập định lượng. Sau đó chúng ta hồi qui hàm Y theo một số hạng không đổi, $D_1, D_2, \ldots, D_{m-1}, X, Z_1, Z_2, \ldots, Z_{m-1}$. Các hệ số của các biến D là các $\hat{\alpha}$ và các hệ số của các biến Z là các $\hat{\beta}$. Xin lưu ý rằng số lượng các hệ số được ước lượng là $2m$ và do đó các bậc tự do là $n - 2m$. Điều này có nghĩa là trừ phi số lượng quan sát tối thiểu bằng $2m$, mô hình với các số hạng tương tác đầy đủ không thể ước lượng được. Nếu số lượng các lựa chọn lớn, biến giả và các số hạng tương tác có thể tăng lên nhanh chóng. Vấn đề càng trở nên rắc rối hơn nếu có nhiều biến định lượng hơn trong mô hình mà các hệ số của chúng đều phải dịch chuyển. Một phương pháp hữu dụng là thiết lập một mô hình cơ bản và sau đó tiến hành kiểm định LM đối với

các độ dịch chuyển tung độ gốc và độ dốc. Hoặc bằng cách khác, chúng ta có thể sử dụng việc giảm bớt dựa trên cơ sở dữ liệu Henry/LSE từ một mô hình tổng quát.

Thuật ngữ

Analysis of covariance model	Mô hình phân tích đồng phương sai
Analysis of variance (ANOVA) model	Mô hình phân tích phương sai
Binary variables	Biến nhị nguyên
Chow test	Kiểm định Chow
Control group	Nhóm kiểm soát
Dummy variables	Các biến giả
Dummy variables trap	Bẫy biến giả
Qualitative variables	Các biến định tính
Quantitative variables	Các biến định lượng
Seasonal dummies	Các biến giả theo mùa
Structural break	Gián đoạn về cấu trúc
Structural change	Thay đổi về cấu trúc
Structural instability	Bất ổn định về cấu trúc