SỬ DỤNG DỮ LIỆU PHÂN LOẠI TRONG PHÂN TÍCH HỒI QUY ĐA BIẾN

Quay lại Chương 1 rằng dữ liệu phân loại là dữ liệu không thể đo lường được bằng số. Ví dụ: màu mắt của người quản lý cửa hàng có tính phân loại (và có thể là một biến dự báo khủng khiếp về doanh thu hàng tháng). Mặc dù các biến phân loại có thể được *biểu thị* bằng số (1 = xanh dương, 2 = xanh lục), chúng rời rạc - không có thứ gọi là "xanh lục rưỡi". Ngoài ra, không thể nói rằng 2 (mắt xanh lục) lớn hơn 1 (mắt xanh dương). Cho đến nay chúng ta đang sử dụng dữ liệu số (có thể được biểu diễn một cách có ý nghĩa bằng các giá trị số liên tục - 110 m tính từ ga tàu là xa hơn 109,9 m) được trình bày trong Bảng 3-2, cũng xuất hiện ở phần trước đây.

Bảng 3-2: Ví dụ về dữ liệu của Kazami Bakery

Cửa hàng	Diện tích sàn cửa hàng (tsubo)	Khoảng cách đến ga gần nhất (mét)	Doanh thu hàng tháng **(¥10,000)**
Yumenooka Shop	10	80	469
Terai Station Shop	8	0	366
Sone Shop	8	200	371
Hashimoto Station Shop	5	200	208
Kikyou Town Shop	7	300	246
Post Office Shop	8	230	297
Suidobashi Station Shop	7	40	363
Rokujo Station Shop	9	0	436
Wakaba Riverside Shop	6	330	198
Misato Shop	9	180	364

Biến dự đoán *diện tích sàn* được đo bằng tsubo, *khoảng cách đến ga gần nhất* tính bằng mét và *doanh thu hàng tháng* bằng yên. Rõ ràng, tất cả những điều này đều có thể đo lường được bằng số. Trong phân tích hồi quy đa biến, biến kết quả phải là biến số có thể đo lường được, nhưng biến dự đoán có thể là

- tất cả các biến số;
- một số biến số và một số biến phân loại, hoặc
- tất cả các biến phân loại.

Bảng 3-3 và 3-4 đều hiển thị các bộ dữ liệu hợp lệ. Trong trường hợp thứ nhất, cả hai biến phân loại và số đều xuất hiện, và trong trường hợp thứ hai, tất cả các biến dự đoán đều biến phân loại.

Bảng 3-3: Sự kết hợp giữa dữ liệu phân loại và dữ liệu số

Cửa hàng	Diện tích sàn cửa hàng (tsubo)	Khoảng cách đến ga gần nhất (mét)	Mẫu tự do	Doanh thu hàng tháng **(¥10,000)**
Yumenooka Shop	10	80	1	469
Terai Station Shop	8	0	0	366
Sone Shop	8	200	1	371
Hashimoto Station Shop	5	200	0	208
Kikyou Town Shop	7	300	0	246
Post Office Shop	8	230	0	297
Suidobashi Station Shop	7	40	0	363
Rokujo Station Shop	9	0	1	436
Wakaba Riverside Shop	6	330	0	198
Misato Shop	9	180	1	364

Trong Bảng 3-3, chúng ta đã đưa vào các *mẫu tự do* có biến dự báo phân loại. Một số địa điểm của Kazami Bakery cung cấp một khay mẫu tự do (1) và những địa điểm khác thì không (0). Khi đưa dữ liệu này vào phân tích, chúng ta sẽ nhận được phương trình hồi quy đa biến sau:

$$y = 30.6x_1 - 0.4x_2 + 39.5x_3 + 135.9$$

Trong đó y đại diện cho doanh số hàng tháng, x_1 đại diện cho diện tích sàn, x_2 đại diện cho khoảng cách đến ga gần nhất và x_3 đại diện cho các mẫu tự do.

Bảng 3-4: Chỉ dữ liệu dự đoán phân loại

Cửa hàng	Diện tích sàn cửa hàng (tsubo)	Khoảng cách đến ga gần nhất (mét)	Mẫu hằng ngày	Mẫu chỉ ngày cuối tuần	Doanh thu hàng tháng **(¥10,000)**
Yumenooka Shop	1	0	1	0	469
Terai Station Shop	1	0	0	0	366
Sone Shop	1	1	1	0	371
Hashimoto Station Shop	0	1	0	0	208
Kikyou Town Shop	0	1	0	0	246
Post Office Shop	1	1	0	0	297
Suidobashi Station Shop	0	0	0	0	363
Rokujo Station Shop	1	0	1	1	436
Wakaba Riverside Shop	0	1	0	0	198
Misato Shop	1	0	1	1	364
	↑	↑	↑	↑	
	Ít hơn 8 tsubo = 0 Từ 8 tsubo trở lên = 1	Ít hơn 200m = 0 Từ 200m trở lên = 1	Không cung cấp mẫu = 0 Cung cấp mẫu = 1		

Trong Bảng 3-4, chúng ta đã chuyển đổi dữ liệu số (diện tích sàn và khoảng cách đến ga gần nhất) thành dữ liệu phân loại bằng cách tạo một số phân loại chung. Sử dụng dữ liệu này, chúng ta tính toán phương trình hồi quy bội như sau:

$$y = 50.2x_1 - 110.1x_2 + 13.4x_3 + 75.1x_4 + 336.4$$

Trong đó y đại diện cho doanh số hàng tháng, x_1 đại diện cho diện tích sàn, x_2 đại diện cho khoảng cách đến nhà ga gần nhất, x_3 đại diện cho mẫu hằng ngày và x_4 đại diện cho mẫu chỉ ngày cuối tuần.

ĐA CỘNG TUYẾN

Đa cộng tuyến xảy ra khi hai trong số các biến dự đoán có mối tương quan chặt chẽ với nhau. Khi điều này xảy ra, thật khó để phân biệt giữa tác động của các biến này đối với biến kết quả và điều này có thể có những tác động sau đối với phân tích của các cậu:

- Ước tính kém chính xác hơn về tác động của một biến nhất định đến biến kết quả
- Sai số chuẩn lớn bất thường của các hệ số hồi quy
- Thất bại trong việc bác bỏ giả thuyết không
- *Quá khớp*, có nghĩa là phương trình hồi quy mô tả mối quan hệ giữa biến kết quả và sai số ngẫu nhiên, thay vì biến dự đoán.

Sự hiện diện của đa cộng tuyến có thể được đánh giá bằng cách sử dụng một chỉ số như dung sai hoặc nghịch đảo của dung sai, được gọi là *hệ số lạm phát phương sai* (VIF). Nói chung, dung sai nhỏ hơn 0,1 hoặc VIF lớn hơn 10 được cho là biểu thị hiện tượng đa cộng tuyến đáng kể, nhưng đôi khi người ta sử dụng các ngưỡng thận trọng hơn.

Khi mới bắt đầu phân tích hồi quy đa biến, các cậu không cần phải lo lắng quá nhiều về điều này. Chỉ cần nhớ rằng hiện tượng đa cộng tuyến có thể gây ra vấn đề khi nó nghiêm trọng. Do đó, khi các biến dự đoán có mối tương quan chặt chẽ với nhau, tốt hơn nên loại bỏ một trong các biến có mối tương quan cao và sau đó phân tích lại dữ liệu.

XÁC ĐỊNH ẢNH HƯỞNG TƯƠNG ĐỐI CỦA CÁC BIẾN DỰ ĐOÁN ĐẾN BIẾN KẾT QUẢ

Một số người sử dụng phân tích hồi quy đa biến để kiểm tra ảnh hưởng tương đối của từng biến dự đoán đến biến kết quả. Đây là cách sử dụng khá phổ biến và được chấp nhận của phân tích hồi quy đa biến, nhưng không phải lúc nào cũng là cách sử dụng khôn ngoan.

Câu chuyện dưới đây minh họa cách một nhà nghiên cứu sử dụng phân tích hồi quy đa biến để đánh giá tác động tương đối của các yếu tố khác nhau đến sự hài lòng chung của những người mua một loại kẹo nhất định.

Ông Torikoshi là nhân viên nghiên cứu phát triển sản phẩm của một công ty bánh kẹo. Gần đây ông ấy đã phát triển một loại kẹo có hương soda mới, Magic Fizz, có thể sủi bọt khi ướt. Kẹo đang bán chạy một cách đáng ngạc nhiên. Để tìm hiểu điều gì khiến nó trở nên phổ biến, công ty đã tặng mẫu kẹo miễn phí cho sinh viên tại trường đại học địa phương và yêu cầu họ đánh giá sản phẩm bằng bảng câu hỏi sau.

Bảng câu hỏi về kẹo Magic Fizz

Vui lòng cho chúng tôi biết suy nghĩ của bạn về Magic Fizz bằng cách trả lời các câu hỏi sau. Hãy khoanh tròn câu trả lời thể hiện đúng nhất quan điểm của bạn.

Hương vị	1. Không hài lòng 2. Hài lòng 3. Tuyệt vời
Quy cách	1. Không hài lòng 2. Hài lòng 3. Tuyệt vời
Cảm giác sủi bọt	1. Không hài lòng 2. Hài lòng 3. Tuyệt vời
Thiết kế bao bì	1. Không hài lòng 2. Hài lòng 3. Tuyệt vời
Đánh giá chung	1. Không hài lòng 2. Hài lòng 3. Tuyệt vời

Hai mươi sinh viên trả lại bảng câu hỏi và kết quả được tổng hợp trong Bảng 3-5. Lưu ý rằng không giống như trong ví dụ của Kazami Bakery, các giá trị của biến kết quả - sự hài lòng chung - đã được biết trước. Trong bài toán tiệm bánh, mục tiêu là dự đoán biến kết quả (lợi nhuận) của một cửa hàng chưa tồn tại dựa trên xu hướng được hiển thị bởi các cửa hàng hiện tại. Còn trong trường hợp này, mục đích của phân tích là kiểm tra tác động tương đối của các biến dự đoán khác nhau để tìm hiểu xem mỗi biến dự đoán (hương vị, quy cách, cảm giác sủi bọt, thiết kế) ảnh hưởng như thế nào đến kết quả (sự hài lòng).

Bảng 3-5: Kết quả bảng hỏi về kẹo Magic Fizz

Người trả lời	Hương vị	Quy cách	Cảm giác sủi bọt	Thiết kế bao bì	Đánh giá chung
1	2	2	3	2	2
2	1	1	3	1	3
3	2	2	1	1	3
3	2	2	1	1	1
4	3	3	3	2	2
5	1	1	2	2	1
6	1	1	1	1	1
7	3	3	1	3	3
8	3	3	1	2	2
9	3	3	1	2	3
10	1	1	3	1	1
11	2	3	2	1	3
12	2	1	1	1	1
13	3	3	3	1	3
14	3	3	1	3	3
15	3	2	1	1	2
16	1	1	3	3	1
17	2	2	2	1	1
18	1	1	1	3	1
19	3	1	3	3	3
20	3	3	3	3	3

Mỗi biến được chuẩn hóa trước khi tính phương trình hồi quy đa biến. Chuẩn hóa làm giảm ảnh hưởng của sai số hoặc tỷ lệ, cho phép nhà nghiên cứu so sánh hai biến chính xác hơn. Phương trình có kết quả là:

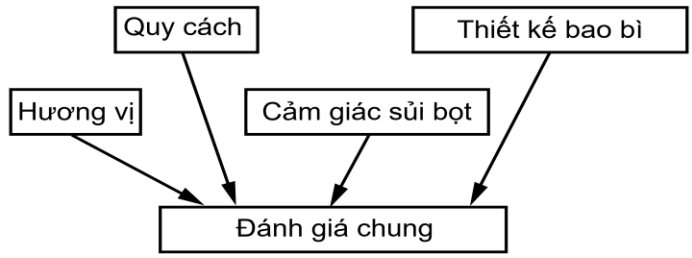
$$y = 0.41x_1 + 0.32x_2 + 0.26x_3 + 0.11x_4$$

Trong đó y thể hiện sự hài lòng chung, x_1 thể hiện hương vị, x_2 thể hiện quy cách, x_3 thể hiện cảm giác sủi bọt và x_4 thể hiện thiết kế bao bì.

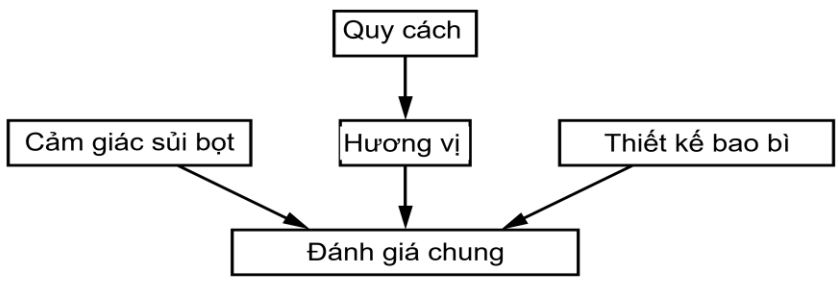
Nếu các cậu so sánh các hệ số hồi quy từng phần của bốn biến dự đoán, các cậu có thể thấy rằng hệ số hương vị là lớn nhất. Dựa trên thực tế đó, ông Torikoshi kết luận rằng hương vị có ảnh hưởng mạnh nhất đến sự hài lòng chung.

Lý luận của ông Torikoshi có lý. Biến kết quả bằng tổng của các biến dự đoán nhân với hệ số hồi quy từng phần của chúng. Nếu các cậu nhân một biến dự đoán với một số cao hơn, nó sẽ có tác động lớn hơn đến kết quả kiểm đếm cuối cùng, phải không? Vâng, đôi khi - nhưng không phải lúc nào cũng đơn giản như vậy.

Chúng ta hãy xem xét kỹ hơn lý luận của ông Torikoshi khi được mô tả ở đây:



Nói cách khác, ông giả định rằng tất cả các biến số đều có mối liên hệ độc lập và trực tiếp với sự hài lòng chung. Tuy nhiên, điều này không hẳn là đúng. Có thể trong thực tế, kết cấu ảnh hưởng đến mức độ hài lòng của mọi người với hương vị, như thế này:



Mô hình phương trình cấu trúc (SEM) là một phương pháp tốt hơn để so sánh tác động tương đối của các biến dự đoán khác nhau đến kết quả. Cách tiếp cận này đưa ra các giả định linh hoạt hơn so với hồi quy tuyến tính và thậm chí nó có thể được sử dụng để phân tích các tập dữ liệu có đa cộng tuyến. Tuy nhiên, SEM không phải là thuốc chữa bách bệnh. Nó dựa trên giả định rằng dữ liệu có liên quan để trả lời câu hỏi được đặt ra.

SEM cũng giả định rằng dữ liệu được mô hình hóa chính xác. Điều đáng chú ý là các câu hỏi trong cuộc khảo sát này yêu cầu mỗi người đánh giá có cách giải thích chủ quan. Nếu Miu cho chiếc kẹo hai điểm "Hài lòng" và hai điểm "Tuyệt vời", cô ấy có thể đánh giá mức độ hài lòng chung của mình là "Hài lòng" hoặc "Tuyệt vời". Xếp hạng nào cô ấy chọn có thể phụ thuộc vào tâm trạng của cô ấy vào ngày hôm đó!

Risa có thể đánh giá bốn hạng mục chính giống như Miu, đưa ra đánh giá mức độ hài lòng chung khác với Miu và vẫn tự tin rằng mình đang đưa ra đánh giá khách quan. Vì Miu và Risa có suy nghĩ khác nhau về đánh giá cuối cùng nên dữ liệu của chúng ta có thể không được mô hình hóa chính xác. Tuy nhiên, mô hình phương trình cấu trúc vẫn có thể mang lại kết quả hữu ích bằng cách cho chúng ta biết biến nào có tác động đến các biến còn lại thay vì kết quả cuối cùng.

Biên dịch: Anh Tuấn

(còn tiếp)