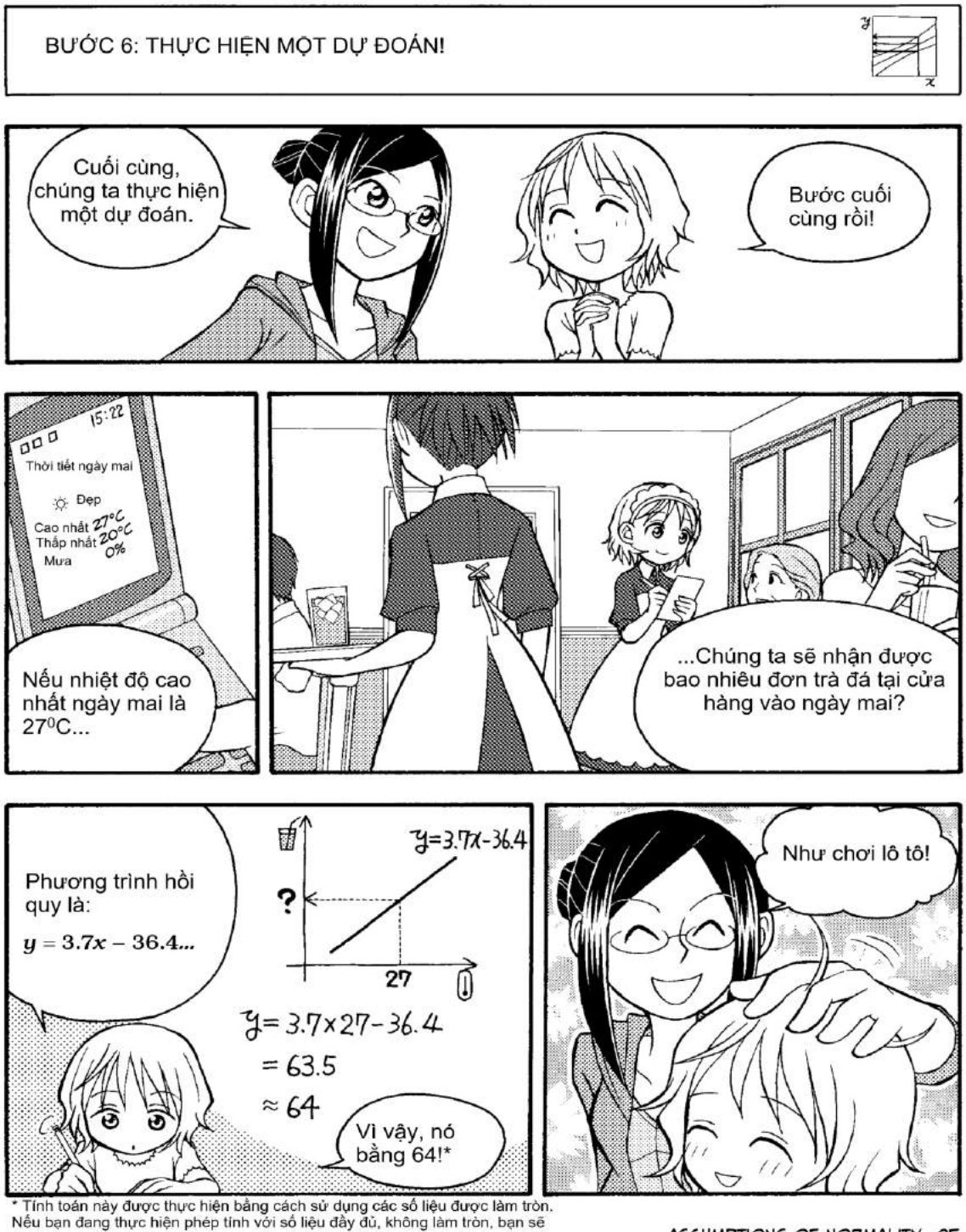
HỌC THỐNG KÊ QUA TRUYỆN TRANH

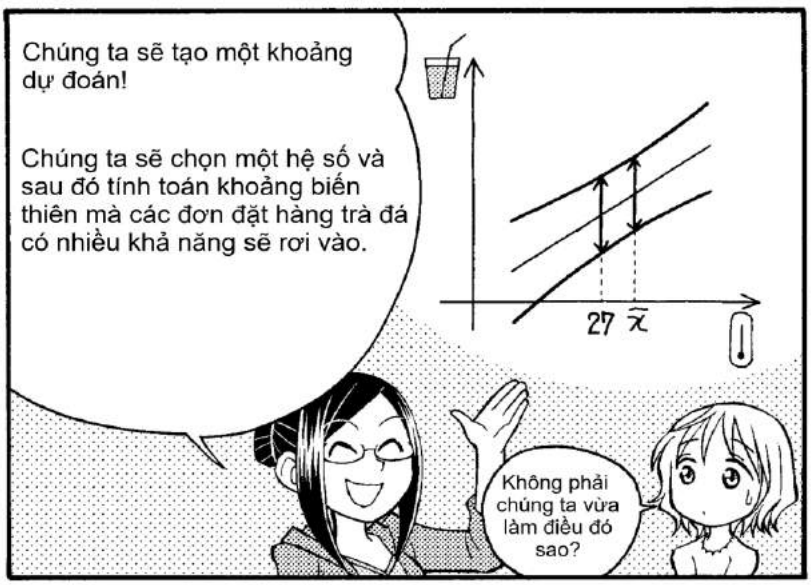


* Tính toán này được thực hiện bằng cách sử dụng các số liệu được làm tròn. Nếu bạn đang thực hiện phép tính với số liệu đầy đủ, không làm tròn, bạn sẽ nhận được 64,6.

ASSUMPTIONS OF NORMALITY 95

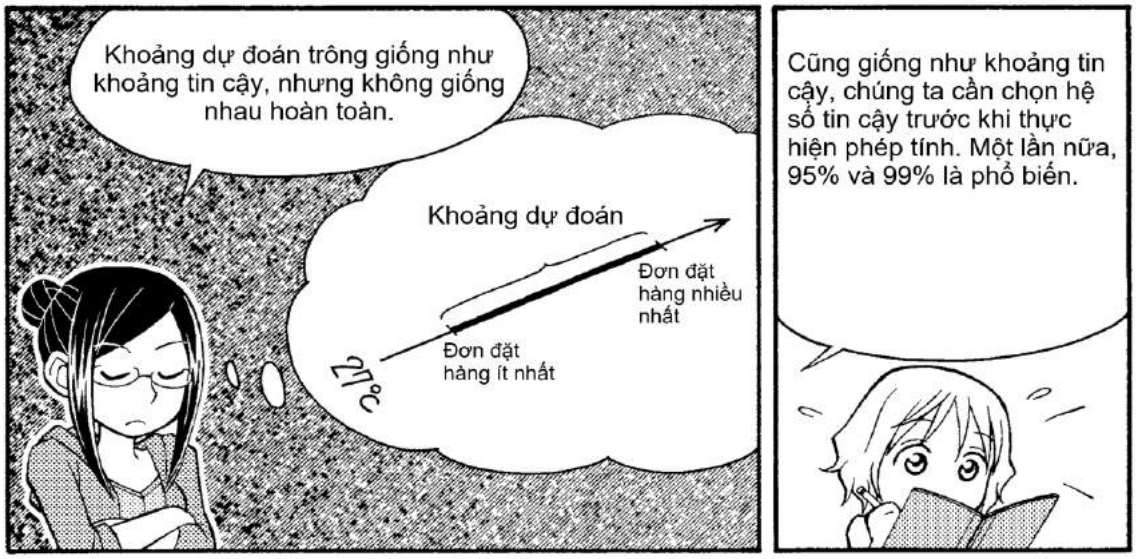

Nhưng sẽ có chính xác 64 đơn đặt hàng?
Làm thế nào chúng ta có thể biết chắc chắn?

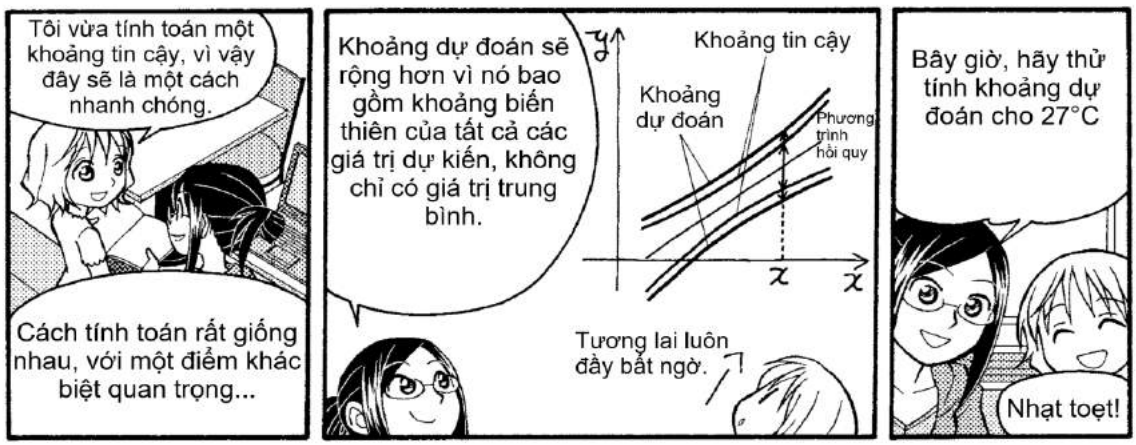

Đó là một câu hỏi lớn.
Chúng ta đạt gần 64 đơn hàng vì giá trị của R2 là 0,8225, nhưng... gần là bao nhiêu?

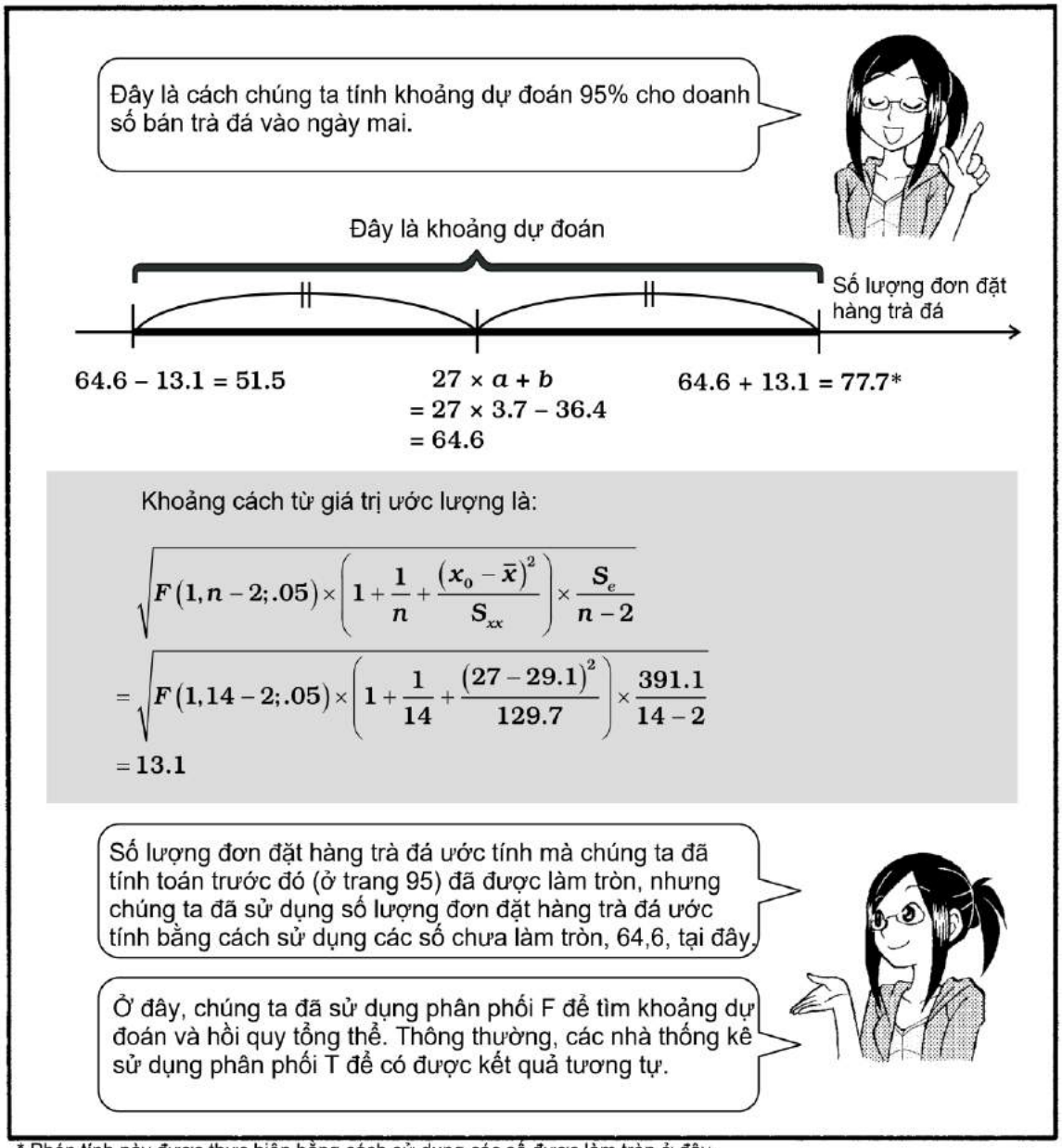

Chúng ta sẽ tạo một khoảng dự đoán!
Chúng ta sẽ chọn một hệ số và sau đó tính toán khoảng biến thiên mà các đơn đặt hàng trà đá có nhiều khả năng sẽ rơi vào.
27
Không phải chúng ta vừa làm điều đó sao?


Không hẳn thế. Trước đây, chúng ta đã dự đoán số lượng đơn đặt hàng trà đá trung bình cho tổng thể trong những ngày có nhiệt độ cao nhất định, nhưng hiện tại, chúng ta đang dự đoán số lượng đơn đặt hàng trà đá có thể có trong một ngày nhất định với nhiệt độ nhất định.
Tớ không thấy sự khác biệt gì.

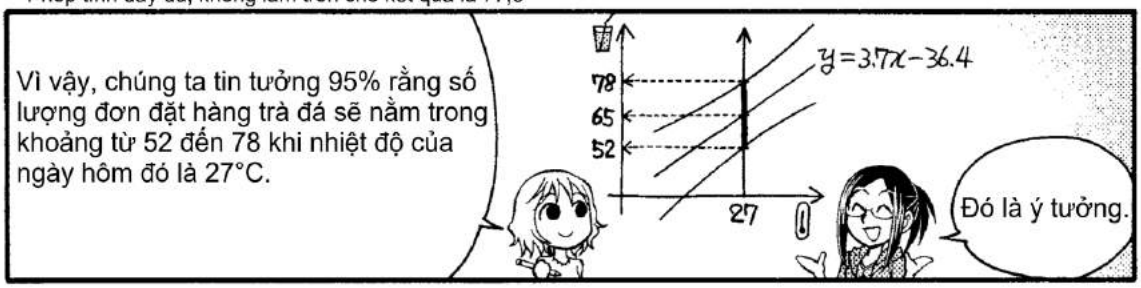

Tổng thể như thế nào?
y=ax+b
31
Khoảng tin cậy giúp chúng ta đánh giá tổng thể.
Chúng ta sẽ có bao nhiêu đơn đặt hàng
y=ax+b
27
P
Khoảng dự đoán đưa ra một khoảng biến thiên có thể xảy ra của các giá trị trong tương lai.


Khoảng dự đoán trông giống như khoảng tin cậy, nhưng không giống nhau hoàn toàn.
Khoảng dự đoán
Đơn đặt hàng nhiều nhất
Đơn đặt hàng ít nhất
27°C
Cũng giống như khoảng tin cậy, chúng ta cần chọn hệ số tin cậy trước khi thực hiện phép tính. Một lần nữa, 95% và 99% là phổ biến.


Tôi vừa tính toán một khoảng tin cậy, vì vậy đây sẽ là một cách nhanh chóng.
Cách tính toán rất giống nhau, với một điểm khác biệt quan trọng...
Khoảng dự đoán sẽ rộng hơn vì nó bao gồm khoảng biến thiên của tất cả các giá trị dự kiến, không chỉ có giá trị trung bình.
y
Khoảng tin cậy
Khoảng dự đoán
Phương trình hồi quy
x̄
x
Tương lai luôn đầy bất ngờ.
Bây giờ, hãy thử tính khoảng dự đoán cho 27°C
Nhạt toẹt!



* Phép tính này được thực hiện bằng cách sử dụng các số được làm tròn ở đây.
Phép tính đầy đủ, không làm tròn cho kết quả là 77,6




Cậu đang nhìn chăm chăm vào cái gì?
Ồ! Tớ đang mơ màng
Cậu đã thực hiện nó thông qua bài học ngày hôm nay.
Bạn thấy thế nào?
Đã có lúc khó khăn...
...nhưng tớ sẽ bắt kịp. Tớ nghĩ rằng tớ có thể làm được.
Và dự đoán tương lai thực sự thú vị!
Ừ, thật sự thấy thú vị đấy.
Chúng ta có thể thực hiện tất cả các loại dự đoán về tương lai.
Giống như, bao nhiêu ngày cho đến khi bạn nói chuyện với anh ấy.

Bước nào là cần thiết?

Bạn có nhớ quy trình phân tích hồi quy được giới thiệu ở trang 68 không?

1. Vẽ biểu đồ phân tán của biến độc lập so với biến phụ thuộc. Nếu các dấu chấm thẳng hàng, các biến có thể tương quan với nhau.
2. Tính phương trình hồi quy.
3. Tính hệ số tương quan (R) và đánh giá tổng thể và các giả định.
4. Tiến hành phân tích phương sai.
5. Tính khoảng tin cậy.
6. Thực hiện dự đoán!

Trong chương này, chúng ta đã đi qua từng bước trong số sáu bước, nhưng không phải lúc nào cũng cần thiết phải thực hiện từng bước. Nhớ lại ví dụ về tuổi và chiều cao của Miu ở trang 25.

- Thực tế: Trên đời này chỉ có một Miu.
- Thực tế: Chiều cao của Miu khi cô ấy 10 tuổi là 137,5 cm. Với hai dữ kiện này, thật vô lý khi nói rằng "Chiều cao của Miu khi cô ấy 10 tuổi tuân theo phân phối chuẩn với trung bình Ax + B và độ lệch chuẩn σ." Nói cách khác, thật vô nghĩa khi phân tích tổng thể về chiều cao của Miu lúc 10 tuổi. Cô ấy chỉ có một, và chúng ta biết chiều cao của cô ấy là bao nhiêu. Trong phân tích hồi quy, chúng ta có thể phân tích toàn bộ tổng thể hoặc phổ biến hơn là phân tích một mẫu của tổng thể. Khi bạn phân tích một mẫu, bạn nên thực hiện tất cả các bước. Tuy nhiên, vì Bước 4 và 5 đánh giá mức độ mẫu đại diện cho tổng thể nên bạn có thể bỏ qua chúng nếu đang sử dụng dữ liệu từ toàn bộ tổng thể thay vì chỉ một mẫu.

Lưu ý: Chúng ta sử dụng thuật ngữ thống kê để mô tả phép đo đặc tính từ một mẫu, chẳng hạn như giá trị trung bình mẫu và tham số để mô tả phép đo xuất phát từ tổng thể, chẳng hạn như giá trị trung bình tổng thể hoặc hệ số.

Phần dư chuẩn hóa

Hãy nhớ rằng *phần dư* là sự khác biệt giữa *giá trị đo được* và *giá trị ước tính* bằng phương trình hồi quy. *Phần dư chuẩn hóa* là phần dư chia cho độ lệch chuẩn ước tính của nó. Chúng tôi sử dụng phần dư chuẩn hóa để đánh giá liệu một phép đo cụ thể có đi chệch hướng đáng kể so với xu hướng hay không.

Ví dụ: giả sử một nhóm người chạy bộ khát nước dừng lại ở Norns vào ngày 4, nghĩa là mặc dù đơn đặt hàng trà đá dự kiến là khoảng 76 dựa trên nhiệt độ của ngày hôm đó, nhưng thực tế khách hàng đã đặt 84 đơn hàng trà đá. Một sự kiện như vậy sẽ dẫn đến một số dư tiêu chuẩn hóa lớn.

Phần dư chuẩn hóa được tính bằng cách chia mỗi phần dư cho ước tính độ lệch chuẩn của nó, được tính bằng cách sử dụng tổng bình phương phần dư. Phép tính hơi phức tạp và hầu hết các phần mềm thống kê đều thực hiện tự động, vì vậy chúng ta sẽ không đi vào chi tiết phép tính ở đây.

Bảng 2-1 cho thấy phần dư chuẩn hóa cho dữ liệu Norns được sử dụng trong chương này.

Bảng 2-1: Tính toán phần dư chuẩn hóa

	Nhiệt độ x	Đo lường số lượng đơn hàng trà đá y	Ước lượng số lượng đơn hàng trà đá $\hat{y} = 3.7x - 36.4$	Phần dư $y - \hat{y}$	Phần dư chuẩn hóa
22nd (Mon.)	29	77	72.0	5.0	0.9
23rd (Tues.)	28	62	68.3	–6.3	–1.2
24th (Wed.)	34	93	90.7	2.3	0.5
25th (Thurs.)	31	84	79.5	4.5	0.8
26th (Fri.)	25	59	57.1	1.9	0.4
27th (Sat.)	29	64	72.0	–8.0	–1.5
28th (Sun.)	32	80	83.3	–3.3	–0.6
29th (Mon.)	31	75	79.5	–4.5	–0.8
30th (Tues.)	24	58	53.3	4.7	1.0
31st (Wed.)	33	91	87.0	4.0	0.8
1st (Thurs.)	25	51	57.1	–6.1	–1.2
2nd (Fri.)	31	73	79.5	–6.5	–1.2
3rd (Sat.)	26	65	60.8	4.2	0.8
4th (Sun.)	30	84	75.8	8.2	1.5

Như bạn có thể thấy, phần dư chuẩn hóa vào ngày thứ 4 là 1,5. Nếu đơn đặt hàng trà đá là 76, như mong đợi, thì phần dư tiêu chuẩn hóa sẽ là 0.

Đôi khi một giá trị được đo có thể đi chệch khỏi xu hướng đến mức nó ảnh hưởng xấu đến phân tích. Nếu phần dư chuẩn hóa lớn hơn 3 hoặc nhỏ hơn –3, phép đo được coi là ngoại lệ. Có một số cách để xử lý các giá trị ngoại lệ, như loại bỏ chúng, thay đổi chúng thành bộ giá trị hoặc chỉ giữ nguyên như vậy trong phân tích. Để xác định cách tiếp cận nào là phù hợp nhất, hãy điều tra nguyên nhân cơ bản của các giá trị ngoại lệ.

Nội suy và ngoại suy

Nếu bạn nhìn vào giá trị x (nhiệt độ cao) ở trang 64, bạn có thể thấy rằng giá trị cao nhất là 34°C và giá trị thấp nhất là 24°C. Sử dụng phân tích hồi quy, bạn có thể *nội suy* số lượng đơn đặt hàng trà đá vào những ngày có nhiệt độ cao từ 24°C đến 34°C và *ngoại suy* số lượng đơn đặt hàng trà đá vào những ngày có nhiệt độ cao dưới 24°C hoặc trên 34°C. Nói cách khác, phép ngoại suy là phép ước tính các giá trị nằm ngoài phạm vi dữ liệu được quan sát của bạn.

Vì chúng ta chỉ quan sát thấy xu hướng giữa 24°C và 34°C nên chúng ta không biết liệu doanh số bán trà đá có theo xu hướng tương tự khi thời tiết cực lạnh hay cực nóng hay không. Do đó phép ngoại suy kém tin cậy hơn phép nội suy và một số nhà thống kê hoàn toàn tránh nó.

Để sử dụng hàng ngày, bạn có thể ngoại suy—miễn là bạn biết rằng kết quả của mình không hoàn toàn đáng tin cậy. Tuy nhiên, tránh sử dụng phép ngoại suy trong nghiên cứu học thuật hoặc để ước tính một giá trị vượt xa phạm vi của dữ liệu đo được.

Tự tương quan

Biến độc lập được sử dụng trong chương này là nhiệt độ; điều này được sử dụng để dự đoán doanh số bán trà đá. Ở hầu hết các nơi, ít có khả năng nhiệt độ sẽ là 20°C vào một ngày nào đó rồi tăng vọt lên 30°C vào ngày hôm sau. Thông thường, nhiệt độ tăng hoặc giảm dần trong khoảng thời gian vài ngày, vì vậy nếu hai biến số này có liên quan với nhau thì số lượng đơn hàng trà đá cũng sẽ tăng hoặc giảm dần. Tuy nhiên, giả định của chúng ta là các giá trị độ lệch (lỗi) là ngẫu nhiên. Do đó, các giá trị dự đoán của chúng ta không thay đổi từ ngày này sang ngày khác một cách suôn sẻ như chúng có thể xảy ra trong cuộc sống thực.

Khi phân tích các biến có thể bị ảnh hưởng bởi thời gian, bạn nên kiểm tra tính tự tương quan. Tự tương quan xảy ra khi sai số tương quan theo thời gian và nó có thể chỉ ra rằng bạn cần sử dụng một loại mô hình hồi quy khác.

Có một chỉ số để mô tả hiện tượng tự tương quan—*thống kê Durbin-Watson*, được tính như sau:

$$d = \frac{\sum_{t=2}^{T} (e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

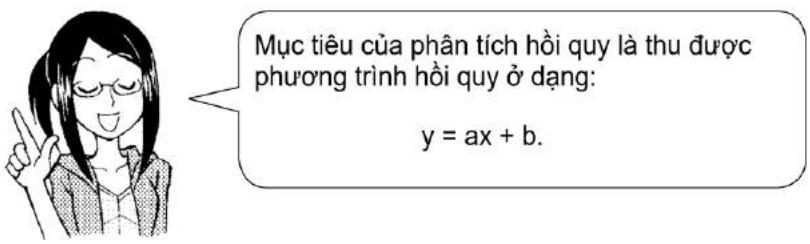
Phương trình có thể được đọc là "tổng bình phương của mỗi số dư trừ đi số dư trước đó, chia cho tổng của mỗi bình phương số dư." Bạn có thể tính giá trị của thống kê Durbin-Watson cho ví dụ trong chương này:

$$\frac{(-6.3-5.0)^2+(2.3-(-6.3))^2+\cdots+(8.2-4.2)^2}{5.0^2+(-6.3)^2+\cdots+8.2^2}=1.8$$

Giá trị tới hạn chính xác của phép thử Durbin-Watson khác nhau đối với từng phân tích và bạn có thể sử dụng bảng để tìm giá trị này, nhưng nhìn chung chúng ta sử dụng 1 làm ngưỡng: kết quả nhỏ hơn 1 có thể cho biết có sự hiện diện của tự tương quan. Nhưng kết quả ở đây gần bằng 2, vì vậy chúng ta có thể kết luận rằng không có tự tương quan trong ví dụ trên.

Hồi quy phi tuyến tính

Ở trang 66, Risa nói:



Phương trình này là tuyến tính, nhưng phương trình hồi quy không nhất thiết phải là tuyến tính. Ví dụ, các phương trình này cũng có thể được sử dụng làm phương trình hồi quy:

- $y=\frac{a}{x}+b$
- $y=a\sqrt{x}+b$
- $y=ax^2+bx+c$
- $y=a\times\log x+b$

Phương trình hồi quy cho tuổi và chiều cao của Miu được giới thiệu ở trang 26 thực ra có dạng: $y=\frac{a}{x}+b$ chứ không phải $y = ax + b$.

Tất nhiên, điều này đặt ra câu hỏi bạn nên chọn loại phương trình nào khi thực hiện phân tích hồi quy trên dữ liệu của chính mình. Dưới đây là một số bước có thể giúp bạn quyết định.

1. Vẽ biểu đồ phân tán của các điểm dữ liệu, với các giá trị biến phụ thuộc trên trục x và các giá trị biến độc lập trên trục y. Kiểm tra mối quan hệ giữa các biến được gợi ý bởi sự trải rộng của các dấu chấm: Chúng có nằm trên một đường thẳng không? Hay chúng nằm theo một đường cong? Nếu đường cong, thì hình dạng của đường cong là gì?

2. Hãy thử phương trình hồi quy được gợi ý bởi hình dạng trong các biến được vẽ ở Bước 1. Vẽ phần dư (hoặc phần dư chuẩn hóa) trên trục y và biến độc lập trên trục x. Phần dư có vẻ là ngẫu nhiên, vì vậy nếu có một mẫu rõ ràng trong phần dư, chẳng hạn như hình cong, thì điều này cho thấy rằng phương trình hồi quy không khớp với hình dạng của mối quan hệ.

3. Nếu biểu đồ phần dư từ Bước 2 hiển thị một mẫu trong phần dư, hãy thử một phương trình hồi quy khác và lặp lại Bước 2. Thử hình dạng của một số phương trình hồi quy và chọn một phương trình có vẻ khớp với dữ liệu nhất. Thông thường, tốt nhất là chọn phương trình đơn giản nhất phù hợp với dữ liệu.

Biến đổi phương trình phi tuyến tính thành phương trình tuyến tính

Có một cách khác để xử lý các phương trình phi tuyến tính: chỉ cần biến chúng thành phương trình tuyến tính. Ví dụ, hãy xem phương trình tuổi và chiều cao của Miu (từ trang 26):

$$y = -\frac{326.6}{x} + 173.3$$

Bạn có thể biến điều này thành một phương trình tuyến tính. Hãy nhớ:

$$\text{If } \frac{1}{x} = X \text{ , thì } \frac{1}{X} = x.$$

Vì vậy, chúng tôi sẽ xác định một biến mới X, đặt nó bằng 1/x và sử dụng X trong phương trình hồi quy y = aX + b thông thường. Như đã trình bày ở trang 76, giá trị của a và b trong phương trình hồi quy y = aX + b có thể được tính như sau:

$$\begin{cases} a = \dfrac{S_{Xy}}{S_{XX}} \\ b = \bar{y} - \bar{X}a \end{cases}$$

Chúng ta tiếp tục phân tích như thường lệ. Xem Bảng 2-2.

Bảng 2-2: Tính phương trình hồi quy

	Tuổi x	$\frac{1}{\text{Tuổi}}$ $\frac{1}{x} = X$	Chiều cao y	$(X-\bar{X})$	$y-\bar{y}$	$(X-\bar{X})^2$	$(y-\bar{y})^2$	$(X-\bar{X})(y-\bar{y})$
	4	0.2500	100.1	0.1428	−38.1625	0.0204	1456.3764	−5.4515
	5	0.2000	107.2	0.0928	−31.0625	0.0086	964.8789	−2.8841
	6	0.1667	114.1	0.0595	−24.1625	0.0035	583.8264	−1.4381
	7	0.1429	121.7	0.0357	−16.5625	0.0013	274.3164	−0.5914
	8	0.1250	126.8	0.0178	−11.4625	0.0003	131.3889	−0.2046
	9	0.1111	130.9	0.0040	−7.3625	0.0000	54.2064	−0.0292
	10	0.1000	137.5	−0.0072	−0.7625	0.0001	0.5814	−0.0055
	11	0.0909	143.2	−0.0162	4.9375	0.0003	24.3789	−0.0802
	12	0.0833	149.4	−0.0238	11.1375	0.0006	124.0439	−0.2653
	13	0.0769	151.6	−0.0302	13.3375	0.0009	177.889	−0.4032
	14	0.0714	154.0	−0.0357	15.7375	0.0013	247.6689	−0.5622
	15	0.0667	154.6	−0.0405	16.3375	0.0016	266.9139	−0.6614
	16	0.0625	155.0	−0.0447	16.7375	0.0020	280.1439	−0.7473
	17	0.0588	155.1	−0.0483	16.8375	0.0023	283.5014	−0.8137
	18	0.0556	155.3	−0.0516	17.0375	0.0027	290.2764	−0.8790
	19	0.0526	155.7	−0.0545	17.4375	0.0030	304.0664	−0.9507
Tổng:	184	1.7144	2212.2	0.0000	0.0000	0.0489	5464.4575	−15.9563
Trung bình:	11.5	0.1072	138.3					

Theo Bảng trên, ta có:

$$\begin{cases} a = \dfrac{S_{xy}}{S_{xx}} = \dfrac{-15.9563}{0.0489} = -326.6^{*} \\ b = \bar{y} - \bar{X}a = 138.2625 - 0.1072 \times (-326.6) = 173.3 \end{cases}$$

Vì vậy, phương trình hồi quy là:

$$y = -326.6X + 173.3$$

↑ Chiều cao ↑ $\frac{1}{\text{Tuổi}}$

* Nếu kết quả của bạn hơi khác so với 326,6, sự khác biệt có thể là do làm tròn. Nếu thế, khác biệt sẽ rất nhỏ.

Tương tự như này:

$$y = -\frac{326.6}{x} + 173.3$$

↑ Chiều cao ↑ Tuổi

Chúng ta đã chuyển đổi phương trình phi tuyến tính ban đầu thành phương trình tuyến tính!

Biên dịch: Anh Tuấn

(còn tiếp)