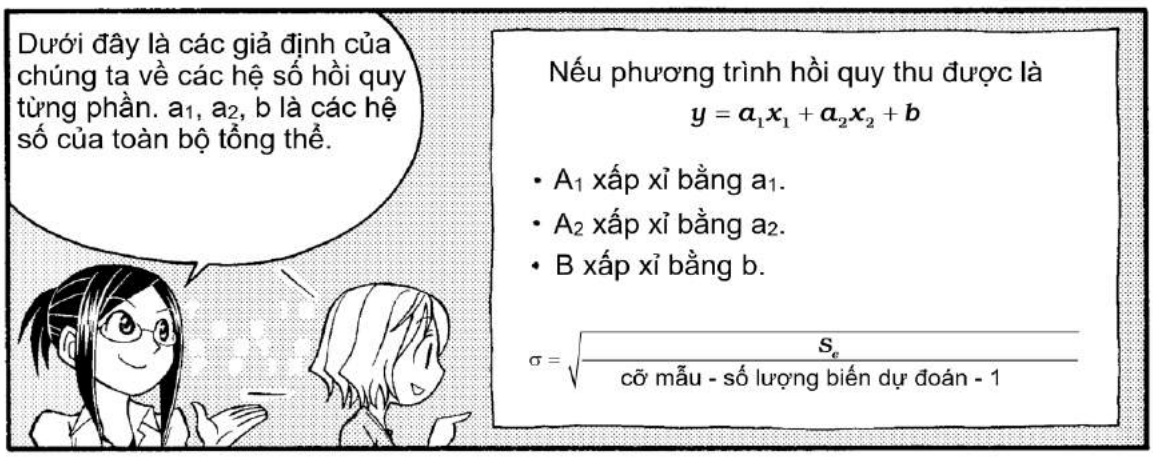
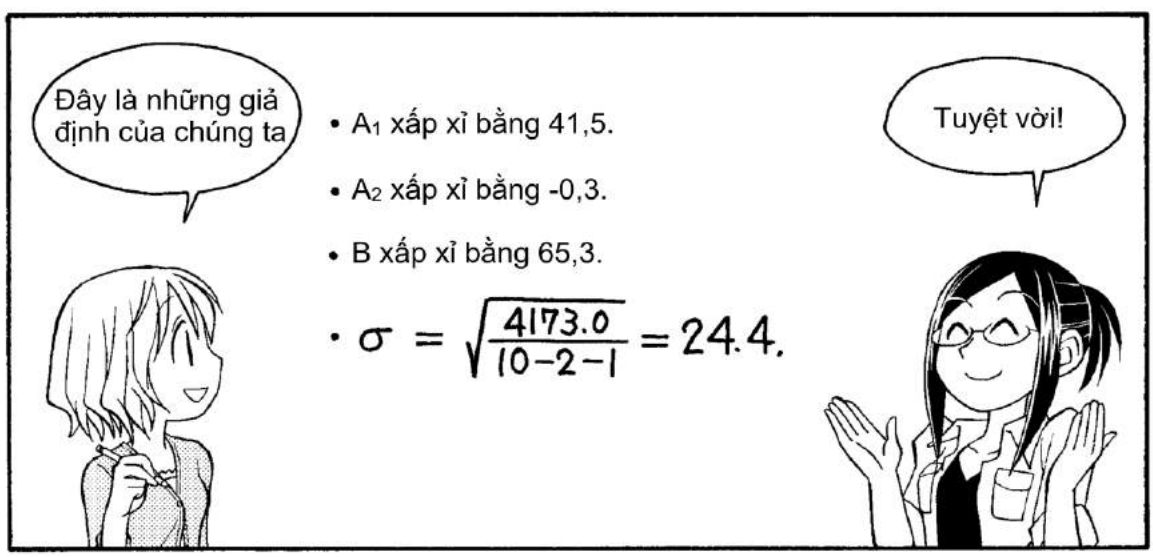
Bước 4: Tiến hành kiểm tra phân tích phương sai (ANOVA).





Một là kiểm định tất cả các hệ số hồi quy từng phần cùng nhau.

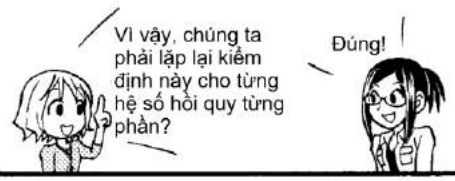
Giả thuyết không	$A_1 = 0$ và $A_2 = 0$
Giả thuyết thay thế	Không phải $A_1 = A_2 = 0$

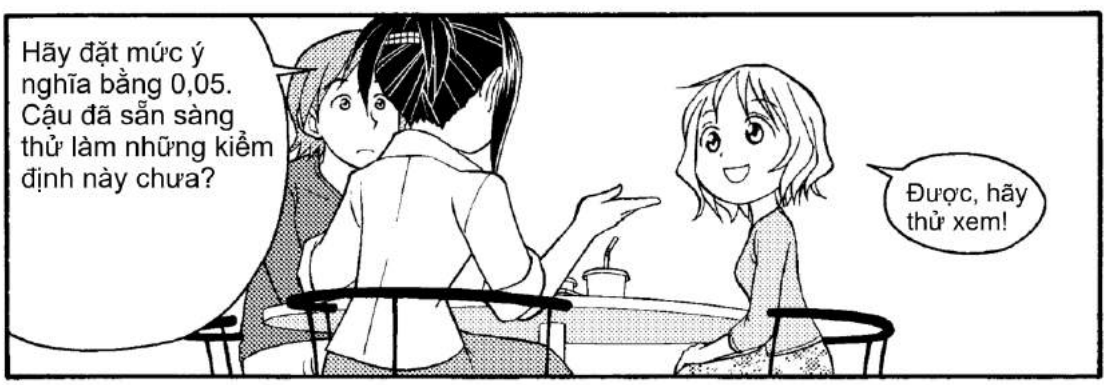
Nói cách khác, một trong những giả thuyết sau là đúng:

- $A_1 \neq 0$ $A_2 \neq 0$
- $A_1 \neq 0$ $A_2 = 0$
- $A_1 = 0$ $A_2 \neq 0$

Hai là kiểm định các hệ số hồi quy từng phần riêng lẻ.

Giả thuyết không	$A_i = 0$
Giả thuyết thay thế	$A_i \neq 0$





Các bước của ANOVA

Bước 1	Xác định tổng thể.	Tổng thể là tất cả các cửa hàng Kazami Bakery.
Bước 2	Thiết lập một giả thuyết không và một giả thuyết thay thế.	Giả thuyết không là $A_1 = 0$ và $A_2 = 0$. Giả thuyết thay thế là A_1 hoặc A_2 hoặc cả hai $\neq 0$.
Bước 3	Lựa chọn kiểm định giả thuyết để thực hiện.	Chúng ta sẽ sử dụng kiểm định F.
Bước 4	Chọn mức ý nghĩa.	Chúng ta sẽ sử dụng mức ý nghĩa bằng 0,05
Bước 5	Tính toán kiểm định thống kê từ dữ liệu mẫu.	Kiểm định thống kê là: $\frac{\frac{S_{yy} - S_e}{\text{Số lượng biến dự đoán}}}{\frac{S_e}{\text{cỡ mẫu - số lượng biến dự đoán - 1}}} =$ $\frac{76199.6 - 4173.0}{2} \div \frac{4173.0}{10-2-1} = 60.4$ Kiểm định thống kê, 60,4, sẽ tuân theo phân phối F với bậc tự do thứ nhất là 2 (số biến dự đoán) và bậc tự do thứ hai là 7 (cỡ mẫu trừ đi số biến dự đoán trừ 1), nếu giả thuyết không là đúng.
Bước 6	Xác định xem giá trị p đối với kiểm định thống kê thu được ở Bước 5 có nhỏ hơn mức ý nghĩa hay không.	Ở mức ý nghĩa 0,05, với d là 2 và d là 7 (10 - 2 - 1), giá trị tới hạn là 4,7374. Kiểm định thống kê của chúng ta là 60,4.
Bước 7	Quyết định xem cậu có thể bác bỏ giả thuyết không hay không.	Vì kiểm định thống kê của chúng ta lớn hơn giá trị tới hạn nên chúng ta bác bỏ giả thuyết không.

Tiếp theo, chúng ta sẽ kiểm định các hệ số hồi quy từng phần riêng lẻ. Tớ sẽ làm ví dụ với A1.

Các bước của ANOVA

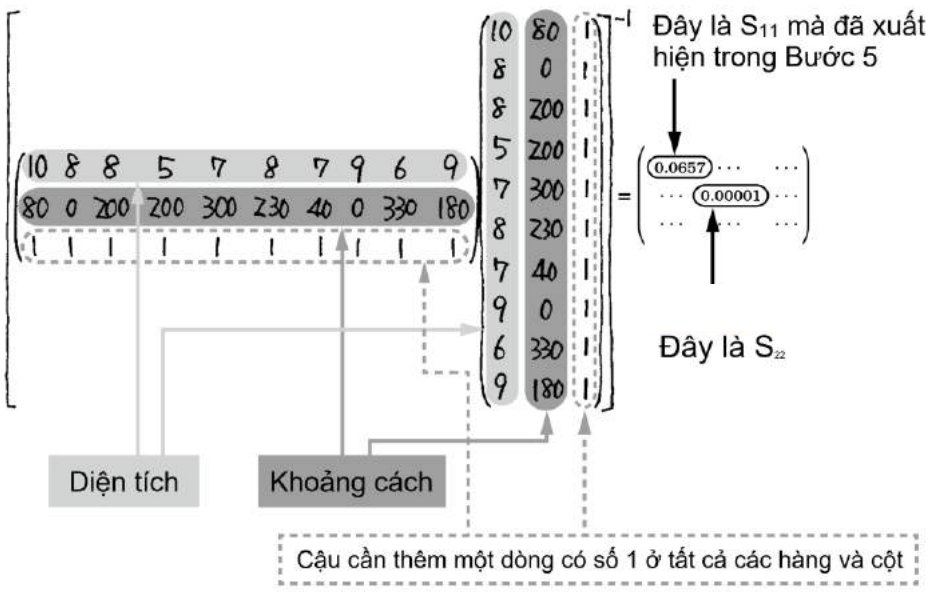
Bước 1	Xác định tổng thể.	Tổng thể là tất cả các cửa hàng Kazami Bakery.
Bước 2	Thiết lập một giả thuyết không và một giả thuyết thay thế.	Giả thuyết không là A = 0. Giả thuyết thay thế là A ≠ 0.
Bước 3	Lựa chọn kiểm định giả thuyết để thực hiện.	Chúng ta sẽ sử dụng kiểm định F.
Bước 4	Chọn mức ý nghĩa.	Chúng ta sẽ sử dụng mức ý nghĩa bằng 0,05.
Bước 5	Tính toán kiểm định thống kê từ dữ liệu mẫu.	Kiểm định thống kê là: $\frac{a_1^2}{S_{11}} \div \frac{S_e}{\text{cỡ mẫu - số lượng biến dự đoán - 1}} =$ $\frac{41.5^2}{0.0657} \div \frac{4173.0}{10-2-1} = 44$ Kiểm định thống kê sẽ tuân theo phân phối F với bậc tự do thứ nhất là 1và bậc tự do thứ hai là 7 (cỡ mẫu trừ đi số biến dự đoán trừ 1), nếu giả thuyết không là đúng. (Giá trị của S_{11} sẽ được giải thích ở trang tiếp theo)
Bước 6	Xác định xem giá trị p đối với kiểm định thống kê thu được ở Bước 5 có nhỏ hơn mức ý nghĩa hay không.	Ở mức ý nghĩa 0,05, với d là 1 và d là 7, giá trị tới hạn là 5,5914. Kiểm định thống kê của chúng ta là 44.
Bước 7	Quyết định xem cậu có thể bác bỏ giả thuyết không hay không.	Vì kiểm định thống kê của chúng ta lớn hơn giá trị tới hạn nên chúng ta bác bỏ giả thuyết không.

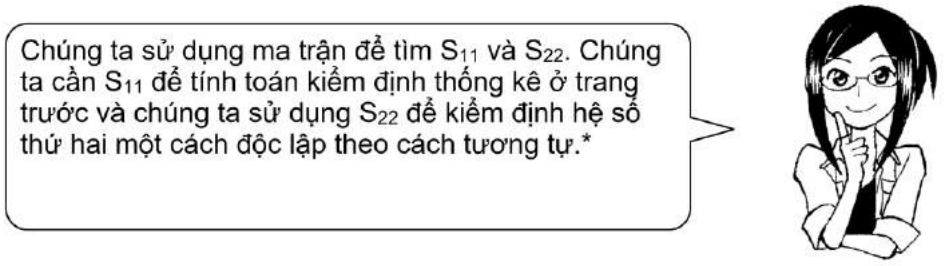
Bất kể kết quả của bước 7 là gì, nếu giá trị của kiểm định thống kê

$$\frac{a_1^2}{S_{11}} \div \frac{S_e}{\text{cỡ mẫu - số lượng biến dự đoán - 1}}$$

bằng 2 trở lên, chúng ta vẫn coi biến dự đoán tương ứng với hệ số hồi quy từng phần đó là hữu ích cho việc dự đoán biến kết quả.

Tìm S_{11} và S_{22}





*Một số người sử dụng phân phối t thay vì phân phối F khi giải thích "kiểm định hệ số hồi quy từng phần". Kết quả cuối cùng của cậu sẽ giống nhau cho dù cậu chọn phương pháp nào.

132

Bước 5: Tính khoảng tin cậy cho tổng thể.

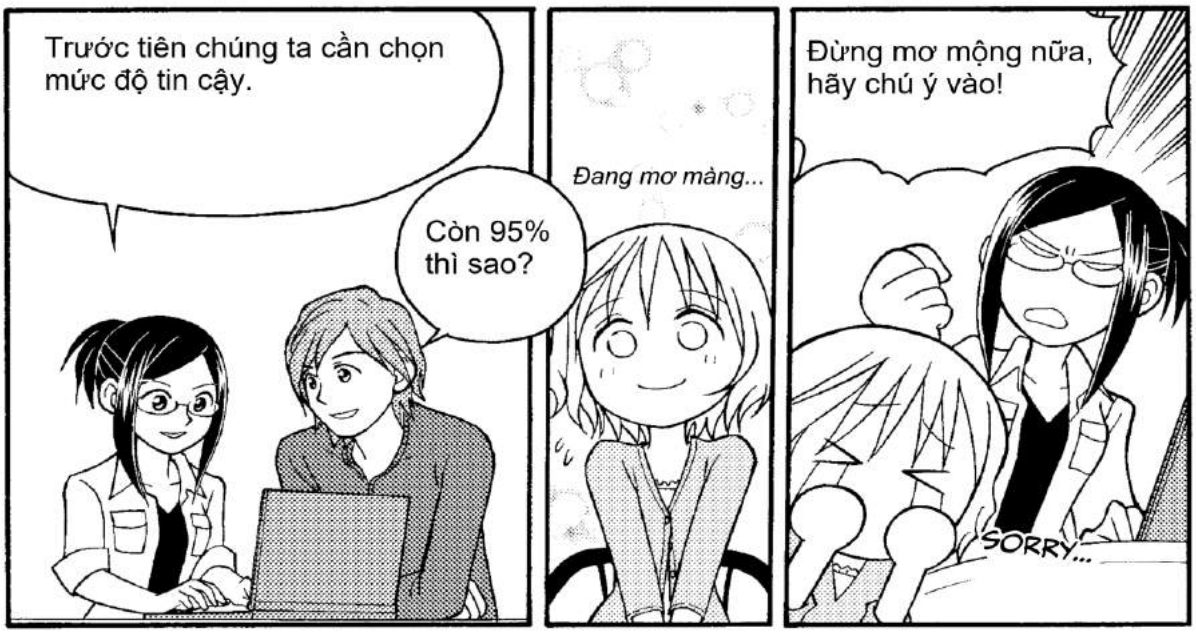
* Nhà toán học P.C. Mahalanobis đã phát minh ra cách sử dụng khoảng cách đa biến để so sánh các tổng thể.

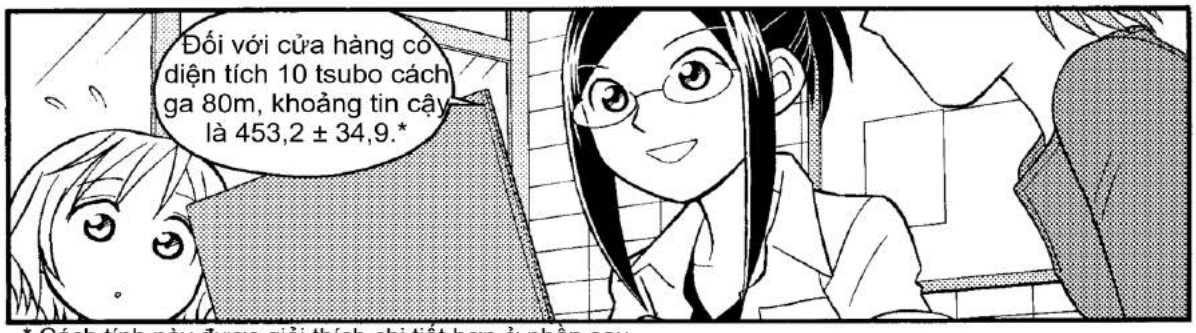

Chúng ta hãy đánh nhau bằng gối nhé!
Hứa là không ngủ nhé?
Tớ tham gia.
Cho tớ mượn bộ đồ ngủ nhé?
Cậu thực sự muốn điều đó?
Được rồi, chúng ta hãy thức cả đêm để làm toán nhé!
Không! Tớ đổi ý rồi! Không là toán nữa.


Vậy thì, tớ đoán chúng ta sẽ phải tìm ra khoảng tin cậy bằng phần mềm phân tích dữ liệu.
Cảm ơn, máy tính.

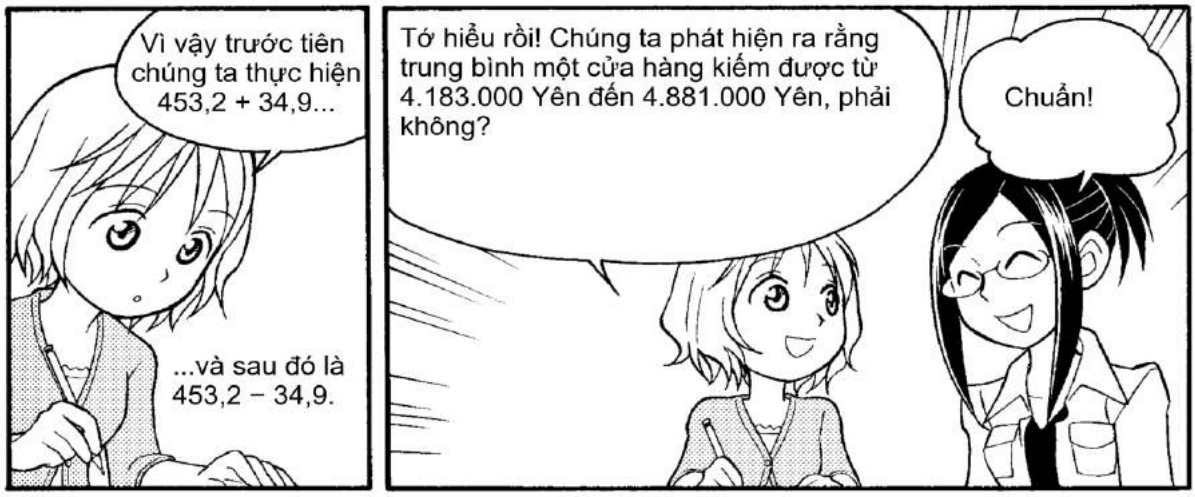

Lần này thì không sao, nhưng bạn không nên lúc nào cũng dựa vào máy tính. Thực hiện các phép tính bằng tay giúp bạn học hỏi được.
Nó sẽ giúp bạn học hỏi thêm.


Đôi khi bạn như một kẻ ngốc nghếch!
Xin lỗi!
Ô, buồn cười quá.

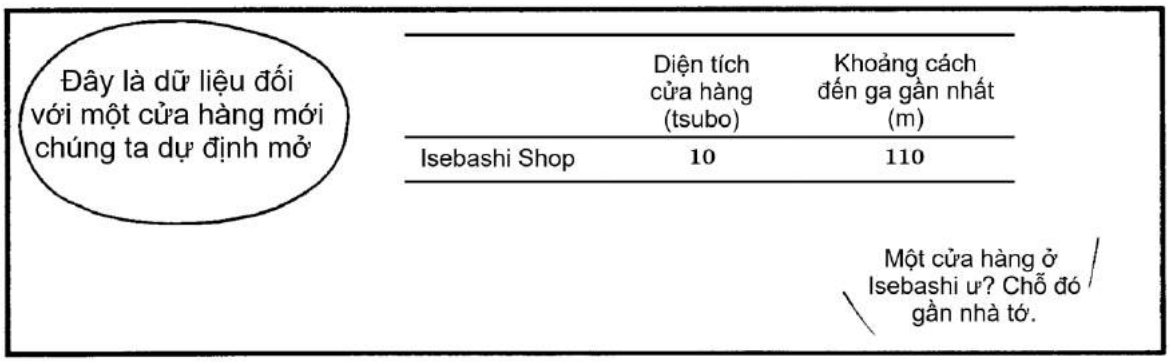




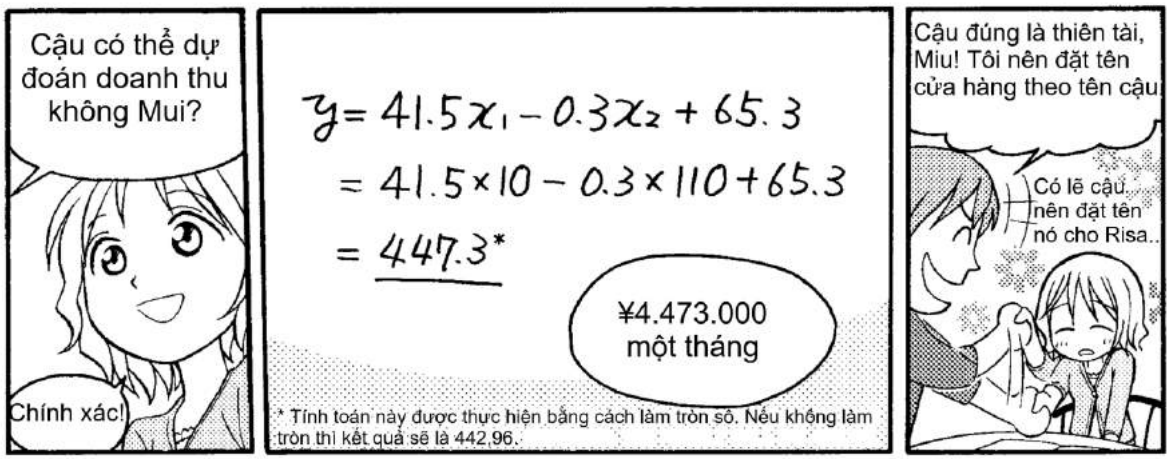
* Cách tính này được giải thích chi tiết hơn ở phần sau.



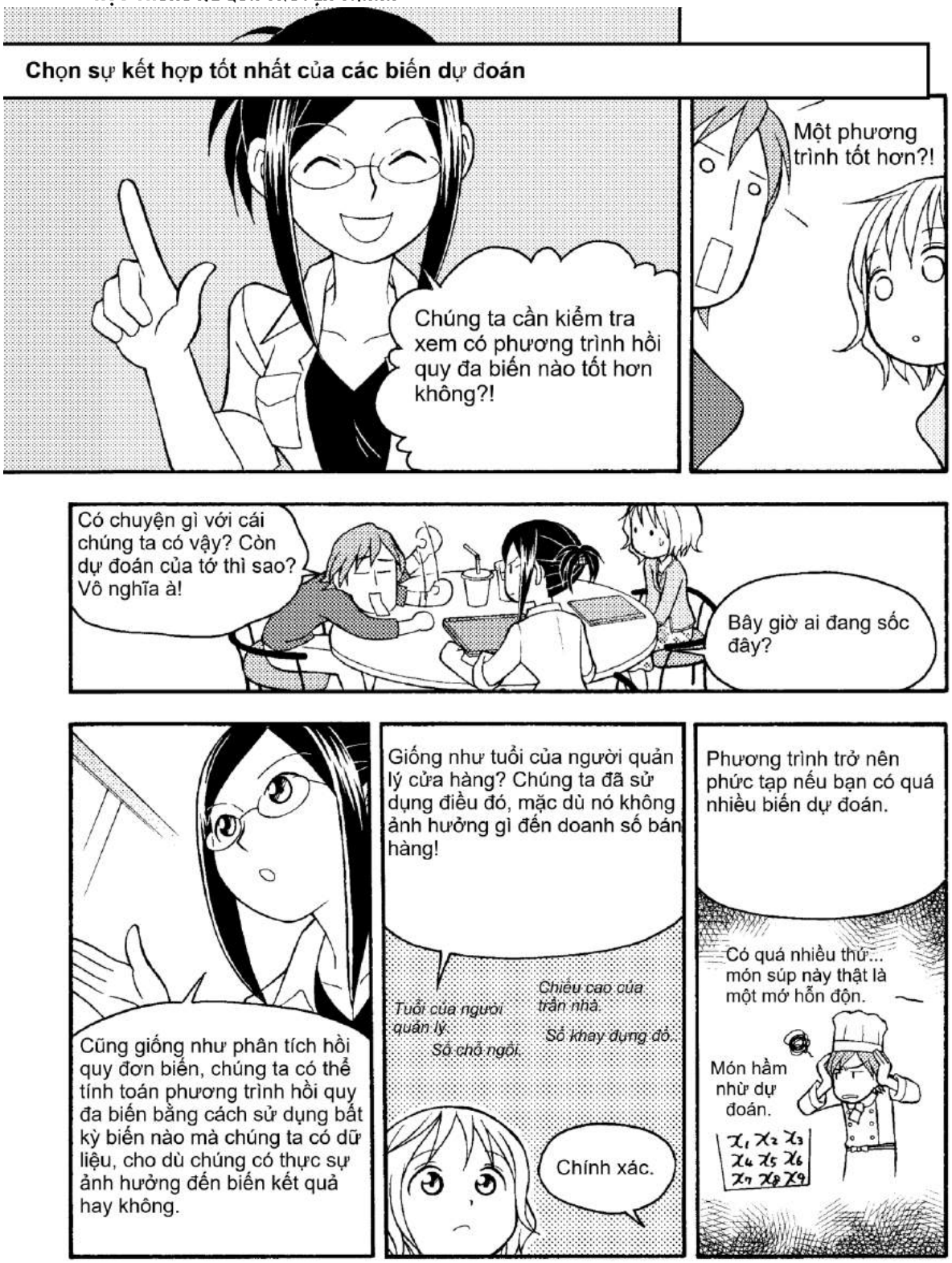
Bước 6: Thực hiện dự đoán



	Diện tích cửa hàng (tsubo)	Khoảng cách đến ga gần nhất (m)
Isebashi Shop	10	110




Vậy chúng ta sẽ sử dụng maha...maha... thứ gì đó nữa à?
Đúng, bla bla khoảng cách.
Nhếch mép cười
Đó là khoảng cách Mahalanobis. Đúng, chúng ta cần sử dụng nó để tìm khoảng dự đoán.
Chúng tớ có thể...có thể, sử dụng lại máy tính của cậu được không? Chỉ một lần nữa thôi à?
Nếu cậu cứ nhất quyết.
Mức độ tin cậy là 95%, do đó doanh số dự đoán...
...nằm trong khoảng từ ¥3.751.000 đến ¥5.109.000
Không tệ!
Vậy cậu có nghĩ mình sẽ mở thêm cửa hàng nữa không?
Những con số này là khá tốt. Cậu biết đấy, tớ nghĩ chúng tôi có thể xem xét!
Đây đã là át chủ bài. Cảm ơn hai cậu!
Ôi, ôi! chờ đã, chỉ còn một điều nữa thôi.


Chọn sự kết hợp tốt nhất của các biến dự đoán
Chúng ta cần kiểm tra xem có phương trình hồi quy đa biến nào tốt hơn không?!
Một phương trình tốt hơn?!
Có chuyện gì với cái chúng ta có vậy? Còn dự đoán của tớ thì sao? Vô nghĩa à!
Bây giờ ai đang sốc đây?
Cũng giống như phân tích hồi quy đơn biến, chúng ta có thể tính toán phương trình hồi quy đa biến bằng cách sử dụng bất kỳ biến nào mà chúng ta có dữ liệu, cho dù chúng có thực sự ảnh hưởng đến biến kết quả hay không.
Giống như tuổi của người quản lý cửa hàng? Chúng ta đã sử dụng điều đó, mặc dù nó không ảnh hưởng gì đến doanh số bán hàng!
Tuổi của người quản lý.
Chiều cao của trần nhà.
Số chỗ ngồi.
Số khay đựng đồ.
Chính xác.
Phương trình trở nên phức tạp nếu bạn có quá nhiều biến dự đoán.
Có quá nhiều thứ... món súp này thật là một mớ hỗn độn.
Món hầm như dự đoán.

Phương trình hồi quy đa biến tốt nhất là cân bằng độ chính xác và độ phức tạp bằng cách chỉ bao gồm các biến dự đoán cần thiết để đưa ra dự đoán tốt nhất.

Khó: $y=a_1x_1+a_2x_2+a_3x_3+a_4x_4+a_5x_5+a_6x_6+\ldots\ldots\ldots+b$

Dễ: $y=a_1x_1+a_2x_2+b$

Chính xác: $y=a_1x_1+a_2x_2+b$, $\bar{R}^2$ > Không chính xác: $y=a_1x_1+a_3x_3+b$, $\bar{R}^2$

Ngắn thì ngọt ngào.

Có một số cách để tìm ra phương trình mang lại cho cậu nhiều lợi ích nhất.

- Lựa chọn chuyển tiếp
- Loại bỏ ngược
- Lựa chọn từng bước tiến-lùi
- Hỏi chuyên gia về miền xem biến nào là quan trọng nhất.

Đây là một số cách phổ biến.

Phương pháp chúng ta sẽ sử dụng hôm nay đơn giản hơn bất kỳ phương pháp nào trong số đó. Nó được gọi là hồi quy tập hợp con tốt nhất hoặc đôi khi là phương pháp quay vòng (Round - Robin).

Quay vòng ư?

Một con chim béo?

Cái quái gì vậy?

x_1 x_2 x_3

Tớ sẽ cho các cậu thấy. Giả sử x_1, x_2 và x_3 là các biến dự báo tiềm năng.

Đầu tiên, chúng ta tính toán phương trình hồi quy đa biến cho mọi sự kết hợp của các biến dự đoán!

- x_1
- x_2
- x_3
- x_1 và x_2
- x_2 và x_3
- x_1 và x_3
- x_1 và x_2 và x_3

Haha. Điều này chắc chắn là tròn trịa?

HAHA. THIS SURE IS ROUND-ABOUT.

Biên dịch: Anh Tuấn

(còn tiếp)