Chương 8

## Giới thiệu về hồi quy tuyến tính

## 8.1 Phù hợp với đường thẳng, phần dư và tương quan

## 8.2 Hồi quy bình phương nhỏ nhất

## 8.3 Các loại giá trị ngoại lệ trong hồi quy tuyến tính

## 8.4 Suy luận cho hồi quy tuyến tính

Hồi quy tuyến tính là một kỹ thuật thống kê rất mạnh mẽ. Nhiều người có một số quen thuộc với hồi quy chỉ từ việc đọc tin tức, ở đâu các đường thẳng được phủ trên các biểu đồ phân tán. Mô hình tuyến tính có thể được sử dụng cho dự đoán hoặc để đánh giá liệu có mối quan hệ tuyến tính giữa hai biến số.

8.1 Phù hợp với đường thẳng, phần dư và tương quan

**Sẽ rất hữu ích nếu bạn suy nghĩ sâu sắc về quy trình điều chỉnh dây chuyền. Trong phần này, chúng tôi xác định biểu mẫu của một mô hình tuyến tính, khám phá các tiêu chí cho điều gì làm cho phù hợp và giới thiệu một thống kê mới được gọi là tương quan**.

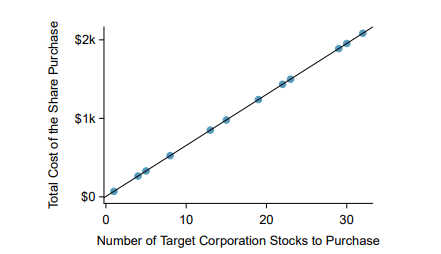
8.1.1 Nối một dòng với dữ liệu

Hình 8.1 cho thấy hai biến mà mối quan hệ của chúng có thể được mô hình hóa hoàn hảo bằng một đường thẳng.

Phương trình của đường thẳng là

y = 5 + 64,96x

Hãy xem xét mối quan hệ tuyến tính hoàn hảo có nghĩa là gì: chúng ta biết giá trị chính xác của y chỉ bằng cách biết giá trị của x. Điều này là không thực tế trong hầu hết mọi quá trình tự nhiên. Ví dụ: nếu chúng tôi lấy thu nhập của gia đình (x), giá trị này sẽ cung cấp một số thông tin hữu ích về mức hỗ trợ tài chính cho một trường đại học có thể cung cấp một sinh viên tương lai (y). Tuy nhiên, dự đoán sẽ không hoàn hảo, vì khác các yếu tố đóng một vai trò trong hỗ trợ tài chính ngoài khả năng tài chính của gia đình.



Hình 8.1: Yêu cầu từ mười hai người mua riêng biệt đã được đặt đồng thời với một công ty thương mại mua cổ phiếu Target Corporation (mã TGT, tháng 12

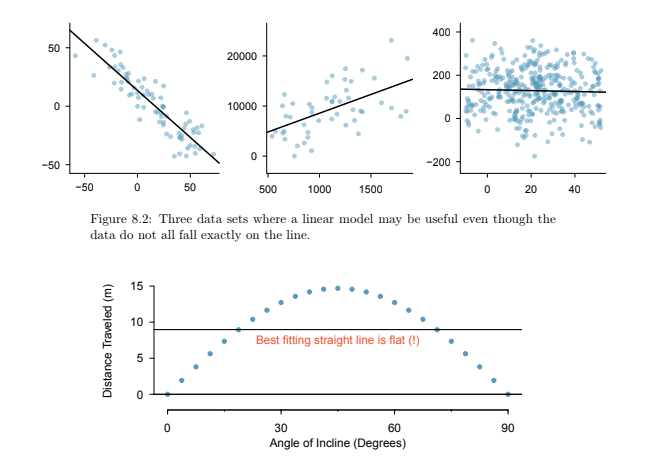
Ngày 28 năm 2018), và tổng giá vốn của cổ phiếu đã được báo cáo. Bởi vì chi phí là được tính bằng công thức tuyến tính, sự phù hợp tuyến tính là hoàn hảo.

Hồi quy tuyến tính là phương pháp thống kê để điều chỉnh một dòng với dữ liệu có mối quan hệ giữa hai biến, x và y, có thể được mô hình hóa bằng một đường thẳng với một số lỗi:

y = β0 + β1x + ε

Các giá trị β0 và β1 đại diện cho các tham số của mô hình (β là chữ cái Hy Lạp beta) và lỗi là được đại diện bởi ε (chữ cái Hy Lạp epsilon). Các tham số được ước tính bằng cách sử dụng dữ liệu và chúng tôi viết ước lượng điểm của chúng là b0 và b1. Khi chúng ta sử dụng x để dự đoán y, chúng ta thường gọi x là giải thích hoặc biến dự báo, và chúng tôi gọi y là phản hồi; chúng tôi cũng thường bỏ thuật ngữ khi viết ra mô hình vì trọng tâm chính của chúng tôi thường là dự đoán kết quả trung bình.

Rất hiếm khi tất cả dữ liệu nằm trên một đường thẳng. Thay vào đó, nó phổ biến hơn cho dữ liệu xuất hiện dưới dạng một đám mây các điểm, chẳng hạn như các ví dụ thể hiện trong Hình 8.2. Trong mỗi trường hợp, dữ liệu rơi xung quanh một đường thẳng, ngay cả khi không có quan sát nào rơi chính xác trên đường thẳng. Người đầu tiên biểu đồ cho thấy một xu hướng tuyến tính giảm tương đối mạnh, trong đó sự thay đổi còn lại trong dữ liệu xung quanh đường thẳng là nhỏ so với độ mạnh của mối quan hệ giữa x và y. Thư hai cốt truyện cho thấy một xu hướng đi lên, trong khi rõ ràng, không mạnh như phần đầu. Cốt truyện cuối cùng cho thấy



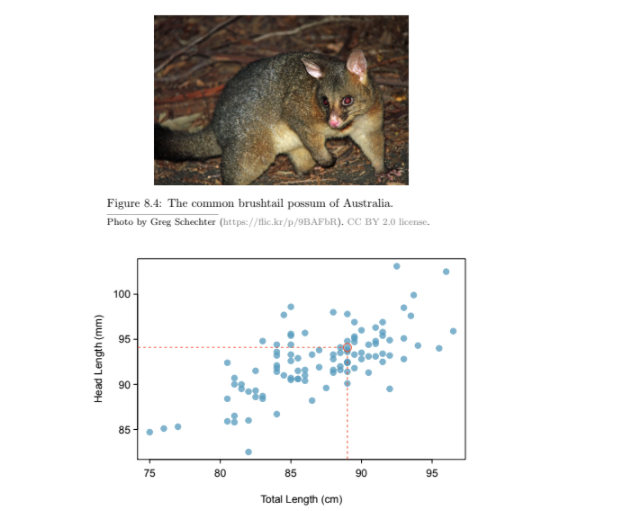
xu hướng giảm rất yếu trong dữ liệu, đến mức chúng ta khó có thể nhận thấy nó. Trong mỗi ví dụ này, chúng ta sẽ có một số không chắc chắn liên quan đến các ước tính của chúng ta về các tham số mô hình, β0 và β1. Vì ví dụ, chúng ta có thể tự hỏi, chúng ta nên di chuyển dòng lên hoặc xuống một chút, hay chúng ta nên nghiêng nó nhiều hơn hoặc ít hơn? Khi chúng ta tiếp tục trong chương này, chúng ta sẽ tìm hiểu về các tiêu chí để phù hợp với đường dây và chúng ta sẽ

cũng tìm hiểu về độ không đảm bảo đo liên quan đến các ước tính của các tham số mô hình.

Cũng có trường hợp khớp một đường thẳng với dữ liệu, ngay cả khi có mối quan hệ rõ ràng giữa các biến, không hữu ích. Một trường hợp như vậy được thể hiện trong Hình 8.3, nơi có một mối quan hệ giữa các biến mặc dù xu hướng không tuyến tính. Chúng tôi thảo luận về các xu hướng phi tuyến tính trong chương này và chương tiếp theo, nhưng chi tiết về cách lắp các mô hình phi tuyến được lưu lại cho khóa học sau.

8.1.2 Sử dụng hồi quy tuyến tính để dự đoán độ dài đầu có túi

Động vật có túi đuôi cọ là một loài thú có túi sống ở Úc, và ảnh của chúng được thể hiện trong Hình 8.4. Các nhà nghiên cứu đã bắt 104 con vật này và đo cơ thể trước khi thả chúng về tự nhiên. Chúng tôi xem xét hai trong số các phép đo này: tổng chiều dài của mỗi con thú có túi, từ đầu đến đuôi và chiều dài của đầu mỗi con thú có túi. Hình 8.5 cho thấy một biểu đồ phân tán cho chiều dài đầu và tổng chiều dài của các loài thú có túi. Mỗi điểm đại diện cho một khả năng duy nhất từ ​​dữ liệu. Các biến số về đầu và tổng chiều dài có liên quan: những con thú có tổng chiều dài trung bình trên cũng có xu hướng có chiều dài đầu trên trung bình. Mặc dù mối quan hệ không hoàn toàn tuyến tính, nhưng có thể hữu ích khi giải thích một phần mối liên hệ giữa các biến này với một đường thẳng.



Hình 8.5: Biểu đồ phân tán cho thấy chiều dài đầu so với tổng chiều dài cho 104 đuôi cọ thú có túi. Một điểm đại diện cho thú có túi với chiều dài đầu 94,1mm và tổng chiều dài 89cm được đánh dấu.

Chúng tôi muốn mô tả mối quan hệ giữa các biến chiều dài phần đầu và tổng chiều dài trong tập dữ liệu có thể sử dụng một dòng. Trong ví dụ này, chúng tôi sẽ sử dụng tổng độ dài làm biến dự đoán, x, để dự đoán chiều dài đầu của thú có túi, y. Chúng ta có thể điều chỉnh mối quan hệ tuyến tính bằng mắt, như trong Hình 8.6.

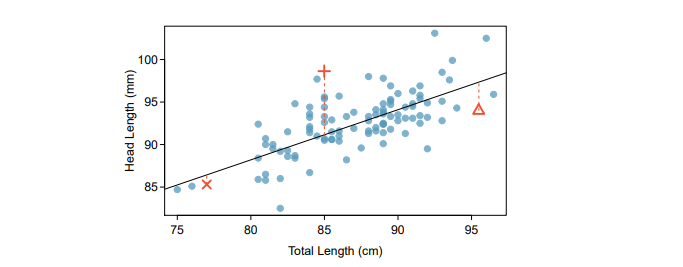
Phương trình cho dòng này là

yˆ = 41 + 0,59x

"Mũ" trên y được sử dụng để biểu thị rằng đây là một ước tính. Chúng ta có thể sử dụng dòng này để thảo luận về các thuộc tính của thú có túi. Ví dụ, phương trình dự đoán một con thú có tổng chiều dài là 80 cm sẽ có chiều dài đầu của

yˆ = 41 + 0,59 × 80 = 88,2

Ước tính có thể được xem là giá trị trung bình: phương trình dự đoán rằng những con thú có tổng chiều dài 80 cm sẽ có chiều dài đầu trung bình là 88,2 mm. Không có thông tin thêm về 80 cm có thể, dự đoán cho chiều dài đầu sử dụng giá trị trung bình là một ước tính hợp lý.



Hình 8.6: Một mô hình tuyến tính hợp lý phù hợp để thể hiện mối quan hệ giữa chiều dài đầu và tổng chiều dài.

VÍ DỤ 8.1

Những biến số nào khác có thể giúp chúng ta dự đoán chiều dài đầu của thú có túi ngoài chiều dài của nó?

Có lẽ mối quan hệ sẽ khác một chút đối với thú có túi đực so với thú có túi cái, hoặc có lẽ nó sẽ khác đối với thú có túi ở một vùng của Úc so với vùng khác. Trong Chương 9, chúng ta sẽ tìm hiểu về cách chúng tôi có thể bao gồm nhiều hơn một công cụ dự đoán. Trước khi đạt được điều đó, trước tiên chúng ta cần cải thiện hiểu cách xây dựng tốt nhất một mô hình tuyến tính đơn giản với một công cụ dự đoán.

8.1.3 Phần dư

Phần còn lại là biến thể còn lại trong dữ liệu sau khi tính toán cho phù hợp với mô hình:

Dữ liệu = Phù hợp + Dư lượng

Mỗi quan sát sẽ có một phần dư và ba phần dư cho mô hình tuyến tính mà chúng tôi phù hợp với

Dữ liệu về sở hữu thể được thể hiện trong Hình 8.6. Nếu một quan sát nằm trên đường hồi quy, thì phần dư của nó, khoảng cách thẳng đứng từ điểm quan sát đến đường thẳng, là số dương. Các quan sát bên dưới dòng có

dư âm. Một mục tiêu trong việc chọn mô hình tuyến tính phù hợp là để những phần dư này nhỏ đến mức khả thi.

Chúng ta hãy xem xét kỹ hơn ba phần dư được mô tả trong Hình 8.6. Quan sát được đánh dấu bởi một “×” có phần dư âm nhỏ khoảng -1; quan sát được đánh dấu bằng “+” có phần dư lớn trong khoảng +7; và quan sát được đánh dấu bằng "4" có dư lượng vừa phải khoảng -4. Kích thước của phần dư thường được thảo luận về giá trị tuyệt đối của nó. Ví dụ: phần dư cho "4" là lớn hơn của “×” bởi vì | - 4 | lớn hơn | - 1 |.

***DÂN CƯ: SỰ KHÁC BIỆT GIỮA ĐƯỢC QUAN SÁT VÀ DỰ KIẾN***

*Phần còn lại của i quan sát thứ (xi, yi) là sự khác biệt của phản hồi quan sát được (yi) và phản hồi mà chúng tôi sẽ dự đoán dựa trên mô hình phù hợp (ˆyi):*

*ei = yi - yˆi*

*Chúng tôi thường xác định ˆyi bằng cách cắm xi vào mô hình.*

VÍ DỤ 8.2

Sự phù hợp tuyến tính trong Hình 8.6 được cho là ˆy = 41 + 0,59x. Dựa trên dòng này, chính thức tính toán phần còn lại của quan sát (77,0, 85,3). Quan sát này được biểu thị bằng “×” trong Hình 8.6. Đánh dấu nó so với ước tính trực quan trước đó, -1.

Trước tiên, chúng tôi tính toán giá trị dự đoán của điểm “×” dựa trên mô hình:

yˆ × = 41 + 0,59x × = 41 + 0,59 × 77,0 = 86,4

Tiếp theo, chúng tôi tính toán sự khác biệt của chiều dài đầu thực tế và chiều dài đầu dự đoán:

e × = y × - yˆ × = 85,3 - 86,4 = −1,1

Sai số của mô hình là e × = −1,1mm, rất gần với ước tính trực quan là -1mm. Sự tiêu cực còn lại chỉ ra rằng mô hình tuyến tính có chiều dài đầu được dự đoán quá mức cho khả năng cụ thể này.

HƯỚNG DẪN THỰC HÀNH 8.3

Nếu một mô hình đánh giá thấp một quan sát, phần dư sẽ tích cực hay tiêu cực? Còn nếu nó đánh giá quá cao sự quan sát? 1

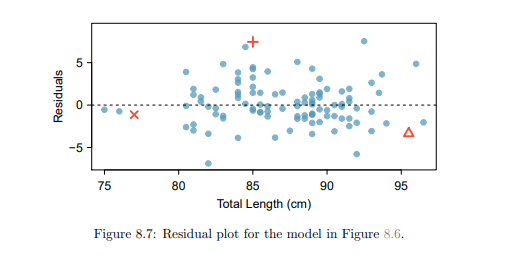
HƯỚNG DẪN THỰC HÀNH 8.4

Tính toán phần dư cho quan sát “+” (85.0, 98.6) và quan sát “4” (95.5, 94.0) trong

hình sử dụng mối quan hệ tuyến tính ˆy = 41 + 0,59x.2

Phần còn lại hữu ích trong việc đánh giá mức độ phù hợp của mô hình tuyến tính với một tập dữ liệu. Chúng tôi thường trưng bày chúng trong một đồ thị thặng dư, chẳng hạn như biểu đồ trong Hình 8.7 cho đường hồi quy trong Hình 8.6.

Các phần còn lại được vẽ ở vị trí nằm ngang ban đầu của chúng nhưng với tọa độ dọc là phần dư. Ví dụ: điểm (85,0, 98,6) + có phần dư là 7,45, vì vậy trong biểu đồ còn lại, nó được đặt ở (85.0, 7.45). Tạo một âm mưu còn lại giống như lật ngược biểu đồ phân tán để đường hồi quy nằm ngang.



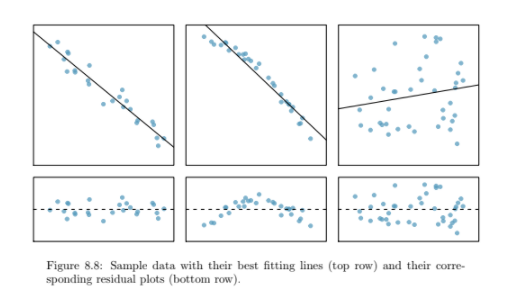
Nếu một mô hình đánh giá thấp một quan sát, thì ước tính của mô hình thấp hơn thực tế. Phần còn lại, là khi đó giá trị quan sát thực tế trừ đi ước tính của mô hình phải là giá trị dương. Điều ngược lại là đúng khi mô hình đánh giá quá cao sự quan sát: phần dư là âm.(+) Đầu tiên tính giá trị dự đoán dựa trên mô hình:

yˆ + = 41 + 0,59x + = 41 + 0,59 × 85,0 = 91,15

Sau đó, phần dư được cho bởi

e + = y + - yˆ + = 98,6 - 91,15 = 7,45

Con số này gần với ước tính trước đó là 7. (4) ˆy4 = 41 + 0,59x4 = 97,3. e4 = y4 - yˆ4 = −3,3, gần với ước lượng của -4.



VÍ DỤ 8.5

Một mục đích của các ô còn lại là xác định các đặc điểm hoặc mô hình vẫn còn rõ ràng trong dữ liệu sau lắp một mô hình. Hình 8.8 cho thấy ba biểu đồ phân tán với mô hình tuyến tính ở hàng đầu tiên và phần dư ô ở hàng thứ hai. Bạn có thể xác định bất kỳ mẫu nào còn lại trong phần dư không?

Trong tập dữ liệu đầu tiên (cột đầu tiên), các phần dư không hiển thị mẫu rõ ràng. Các phần còn lại dường như được rải rác ngẫu nhiên xung quanh đường đứt nét biểu thị 0.

Tập dữ liệu thứ hai cho thấy một mẫu trong phần dư. Có một số độ cong trong biểu đồ phân tán, điều này rõ ràng hơn trong cốt truyện còn lại. Chúng ta không nên sử dụng một đường thẳng để mô hình hóa những dữ liệu này.

Thay vào đó, một kỹ thuật tiên tiến hơn nên được sử dụng.

Biểu đồ cuối cùng cho thấy rất ít xu hướng đi lên, và các phần còn lại cũng không cho thấy các mô hình rõ ràng. Nó là

hợp lý để cố gắng điều chỉnh một mô hình tuyến tính với dữ liệu. Tuy nhiên, không rõ liệu có thống kê bằng chứng đáng kể rằng tham số độ dốc khác 0. Ước tính điểm của độ dốc tham số, có nhãn b1, không phải là 0, nhưng chúng tôi có thể tự hỏi liệu điều này có thể chỉ là do ngẫu nhiên hay không. Chúng tôi sẽ giải quyết loại tình huống này trong Phần 8.4.

8.1.4 Mô tả mối quan hệ tuyến tính với mối tương quan

Chúng tôi đã thấy các âm mưu có mối quan hệ tuyến tính mạnh và những âm mưu khác có mối quan hệ tuyến tính rất yếu. Nó sẽ hữu ích nếu chúng ta có thể định lượng độ mạnh của các mối quan hệ tuyến tính này bằng một thống kê.

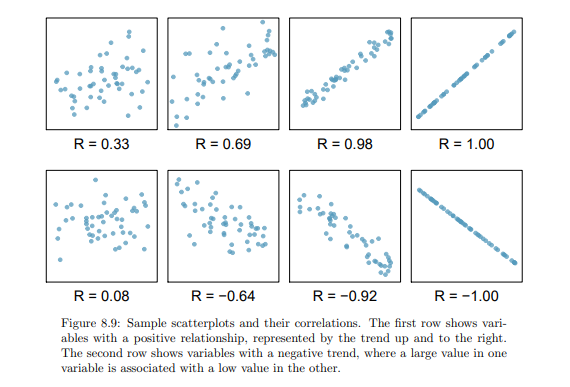
CORRELATION: SỨC MẠNH CỦA MỐI QUAN HỆ TUYẾN TÍNH

Tương quan, luôn nhận các giá trị từ -1 đến 1, mô tả độ mạnh của tuyến tính mối quan hệ giữa hai biến. Chúng tôi biểu thị mối tương quan bằng R.

Chúng ta có thể tính toán mối tương quan bằng cách sử dụng một công thức, giống như chúng ta đã làm với giá trị trung bình của mẫu và độ lệch chuẩn. Công thức này khá phức tạp, 3 và giống như các thống kê khác, chúng tôi thường thực hiện 3 Thông thường, chúng ta có thể tính toán mối tương quan cho các quan sát (x1, y1), (x2, y2), ..., (xn, yn) bằng công thức

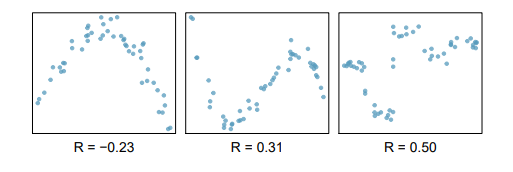


trong đó ¯x, ¯y, sx và sy là trung bình mẫu v à độ lệch chuẩn cho mỗi biến.



Các phép tính trên máy tính hoặc máy tính. Hình 8.9 cho thấy tám ô và tương ứng của chúng các mối tương quan. Chỉ khi mối quan hệ là hoàn toàn tuyến tính thì mối tương quan mới là -1 hoặc 1. Nếu mối quan hệ mạnh mẽ và tích cực, mối tương quan sẽ gần +1. Nếu nó mạnh và tiêu cực, nó sẽ gần -1. Nếu không có mối quan hệ tuyến tính rõ ràng giữa các biến, thì mối tương quan sẽ gần bằng không.

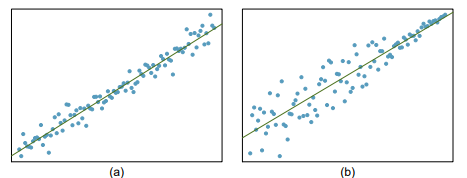
Mối tương quan nhằm định lượng sức mạnh của một xu hướng tuyến tính. Xu hướng phi tuyến tính, thậm chí khi mạnh mẽ, đôi khi tạo ra những mối tương quan không phản ánh sức mạnh của mối quan hệ; Thấy chưa ba ví dụ như vậy trong Hình 8.10.



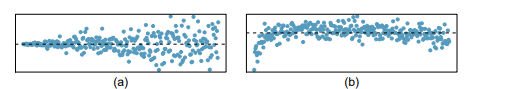
HƯỚNG DẪN THỰC HÀNH 8.6

Không có đường thẳng nào phù hợp cho các tập dữ liệu được biểu diễn trong Hình 8.10. Thử vẽ phi tuyến tính các đường cong trên mỗi ô. Sau khi bạn tạo đường cong cho từng đường cong, hãy mô tả điều gì quan trọng đối với sự phù hợp của bạn.4 4Chúng tôi sẽ để bạn vẽ các đường thẳng. Nói chung, các đường bạn vẽ phải gần với hầu hết các điểm và phản ánh xu hướng tổng thể trong dữ liệu.

**8.1 Hình dung phần dư.** Các biểu đồ phân tán được hiển thị bên dưới, mỗi biểu đồ có một đường hồi quy xếp chồng. Nếu nhưchúng tôi phải xây dựng một ô còn lại (phần dư so với x) cho mỗi ô, mô tả những ô đó trông như thế nào.



**8.2 Xu hướng của phần dư**. Dưới đây là hai đồ thị còn lại sau khi điều chỉnh mô hình tuyến tính thành hai bộ dữ liệu khác nhau. Mô tả các tính năng quan trọng và xác định xem mô hình tuyến tính có phù hợp không cho những dữ liệu này. Giải thích lý do của bạn.



**8.3 Xác định các mối quan hệ, Phần I**. Đối với mỗi ô trong số sáu biểu đồ, hãy xác định độ mạnh của mối quan hệ (ví dụ: yếu, trung bình hoặc mạnh) trong dữ liệu và liệu việc điều chỉnh mô hình tuyến tính có hợp lý hay không

