**Mô hình tuyến tính**

Cho đến thời điểm này, cuốn sách này đã tập trung chủ yếu vào các bộ dữ liệu bao gồm một biến duy nhất. Tuy nhiên, trong các thách thức phân tích dữ liệu, rất phổ biến để quan tâm đến mối quan hệ giữa hai hoặc nhiều biến. Trong phần này của cuốn sách, chúng tôi giới thiệu các mô hình tuyến tính, một khuôn khổ chung thống nhất các phương pháp được sử dụng để phân tích mối liên hệ giữa các biến, chẳng hạn như hồi quy đơn giản và đa biến, *mô hình* hiệu quả điều trị và kiểm tra liên kết. Chúng tôi sẽ minh họa những điều này bằng cách sử dụng các nghiên cứu điển hình liên quan đến việc học tập nếu chiều cao là di truyền, được mô tả chi tiết trong Chương Chương 13, sử dụng dữ liệu để xây dựng một đội bóng chày với ngân sách, được mô tả chi tiết trong Chương Chương 14, xác định xem chế độ ăn nhiều chất béo có làm cho chuột nặng hơn hay không, được mô tả chi tiết trong [Chương 16](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/regression.html) và kiểm tra xem có sự thiên vị giới tính trong tài trợ nghiên cứu ở Hà Lan hay không, được mô tả chi tiết trong Chương [Chương 17](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/association-tests.html).

# 13 Hồi quy

## 13.1 Nghiên cứu điển hình: chiều cao có di truyền không?

Để hiểu các khái niệm về tương quan và hồi quy đơn giản, chúng tôi thực sự sử dụng tập dữ liệu mà từ đó hồi quy được sinh ra. Ví dụ là từ di truyền học. Francis Galton[1](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/regression.html#fn1) nghiên cứu sự biến đổi và di truyền của các đặc điểm của con người. Trong số nhiều đặc điểm khác, Galton đã thu thập và nghiên cứu dữ liệu chiều cao từ các gia đình để cố gắng hiểu về di truyền. Trong khi làm điều này, ông đã phát triển các khái niệm về tương quan và hồi quy, cũng như kết nối với các cặp dữ liệu tuân theo phân phối chuẩn. Tất nhiên, tại thời điểm dữ liệu này được thu thập, kiến thức của chúng ta về di truyền học khá hạn chế so với những gì chúng ta biết ngày nay. Một câu hỏi rất cụ thể mà Galton đã cố gắng trả lời là: chúng ta có thể dự đoán chiều cao của một đứa trẻ dựa trên chiều cao của cha mẹ tốt như thế nào? Kỹ thuật mà ông đã phát triển để trả lời câu hỏi này, hồi quy, cũng có thể được áp dụng cho câu hỏi bóng chày của chúng tôi. Hồi quy cũng có thể được áp dụng trong nhiều trường hợp khác.

|  |
| --- |
| Galton đã có những đóng góp quan trọng cho thống kê và di truyền học, nhưng ông cũng là một trong những người đầu tiên đề xuất thuyết ưu sinh, một phong trào triết học thiếu sót về mặt khoa học được nhiều nhà sinh vật học thời Galton ưa chuộng nhưng với những hậu quả lịch sử khủng khiếp. Bạn có thể đọc thêm về nó ở đây: <https://pged.org/history-eugenics-and-genetics/>. |

Galton đã có những đóng góp quan trọng cho thống kê và di truyền học, nhưng ông cũng là một trong những người đầu tiên đề xuất thuyết ưu sinh, một phong trào triết học thiếu sót về mặt khoa học được nhiều nhà sinh vật học thời Galton ưa chuộng nhưng với những hậu quả lịch sử khủng khiếp. Bạn có thể đọc thêm về nó ở đây: <https://pged.org/history-eugenics-and-genetics/>.

Chúng tôi có quyền truy cập vào dữ liệu chiều cao gia đình của Galton thông qua gói **HistData**. Dữ liệu này chứa chiều cao của vài chục gia đình: mẹ, cha, con gái và con trai. Để bắt chước phân tích của Galton, chúng tôi sẽ tạo một tập dữ liệu với chiều cao của người cha và con trai được chọn ngẫu nhiên của mỗi gia đình:

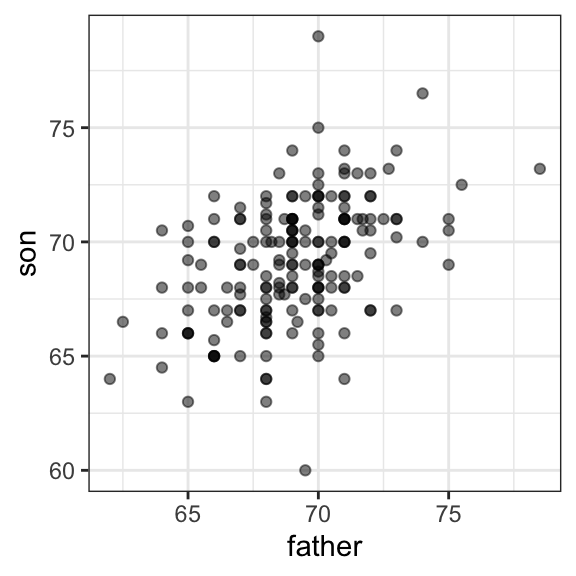
|  |
| --- |
| library(tidyverse)  library(HistData)  set.seed(1983)  galton\_heights <- GaltonFamilies |>    filter(gender == "male") |>    group\_by(family) |>    sample\_n(1) |>    ungroup() |>    select(father, childHeight) |>    rename(son = childHeight) |

Giả sử chúng ta được yêu cầu tóm tắt dữ liệu cha và con trai. Vì cả hai phân phối đều được xấp xỉ tốt bởi phân phối chuẩn, chúng ta có thể sử dụng hai trung bình và hai độ lệch chuẩn làm tóm tắt:

|  |
| --- |
| galton\_heights |>    summarize(mean(father), sd(father), mean(son), sd(son))  #> # A tibble: 1 × 4  #>   `mean(father)` `sd(father)` `mean(son)` `sd(son)`  #>            <dbl>        <dbl>       <dbl>     <dbl>  #> 1           69.1         2.55        69.2      2.71 |

Tuy nhiên, bản tóm tắt này không mô tả được một đặc điểm quan trọng của dữ liệu: xu hướng người cha càng cao, con trai càng cao.

|  |
| --- |
| galton\_heights |> ggplot(aes(father, son)) +    geom\_point(alpha = 0.5) |



Chúng ta sẽ biết rằng hệ số tương quan là một bản tóm tắt thông tin về cách hai biến di chuyển cùng nhau và sau đó thúc đẩy hồi quy đơn giản bằng cách lưu ý cách điều này có thể được sử dụng để dự đoán một biến bằng cách sử dụng biến kia.

## 13.2 Hệ số tương quan

Hệ số tương quan được xác định cho một danh sách các cặp (x1,y1),...,(xn,yn) Là trung bình cộng của sản phẩm của các giá trị được tiêu chuẩn hóa:

với  Trung bình của ,......   và   , tương ứng, và   độ lệch chuẩn. Chữ cái Hy Lạp  thường được sử dụng trong sách thống kê để biểu thị mối tương quan. Bức thư Hy Lạp cho , bởi vì nó là chữ cái đầu tiên của hồi quy. Chẳng mấy chốc chúng ta tìm hiểu về mối liên hệ giữa tương quan và hồi quy. Chúng ta có thể biểu diễn công thức trên bằng mã R bằng cách sử dụng:

|  |
| --- |
| rho <- mean(scale(x) \* scale(y)) |

Để hiểu tại sao phương trình này thực sự tóm tắt cách hai biến di chuyển cùng nhau, hãy xem xét . Mục thứ của  Là SD cách xa mức trung bình. Tương tự, các tôi được ghép nối với . Là  SD cách xa mức trung bình Nếu  và  không liên quan, sản phẩm sẽ tích cực ( +×+ và −×− ) t hường xuyên như tiêu cực (+×− và −×+) và sẽ trung bình ra khoảng 0. Mối tương quan này là trung bình và do đó các biến không liên quan sẽ có mối tương quan bằng 0. Thay vào đó, nếu số lượng thay đổi cùng nhau, thì chúng tôi đang tính trung bình hầu hết các sản phẩm tích cực ( +×+ và −×−) và chúng ta có được một mối tương quan tích cực. Nếu chúng khác nhau theo hướng ngược nhau, chúng ta sẽ có mối tương quan âm.

Hệ số tương quan luôn nằm trong khoảng từ -1 đến 1. Chúng ta có thể chỉ ra điều này bằng toán học: xem xét rằng chúng ta không thể có mối tương quan cao hơn so với khi chúng ta so sánh một danh sách với chính nó (tương quan hoàn hảo) và trong trường hợp này mối tương quan là:

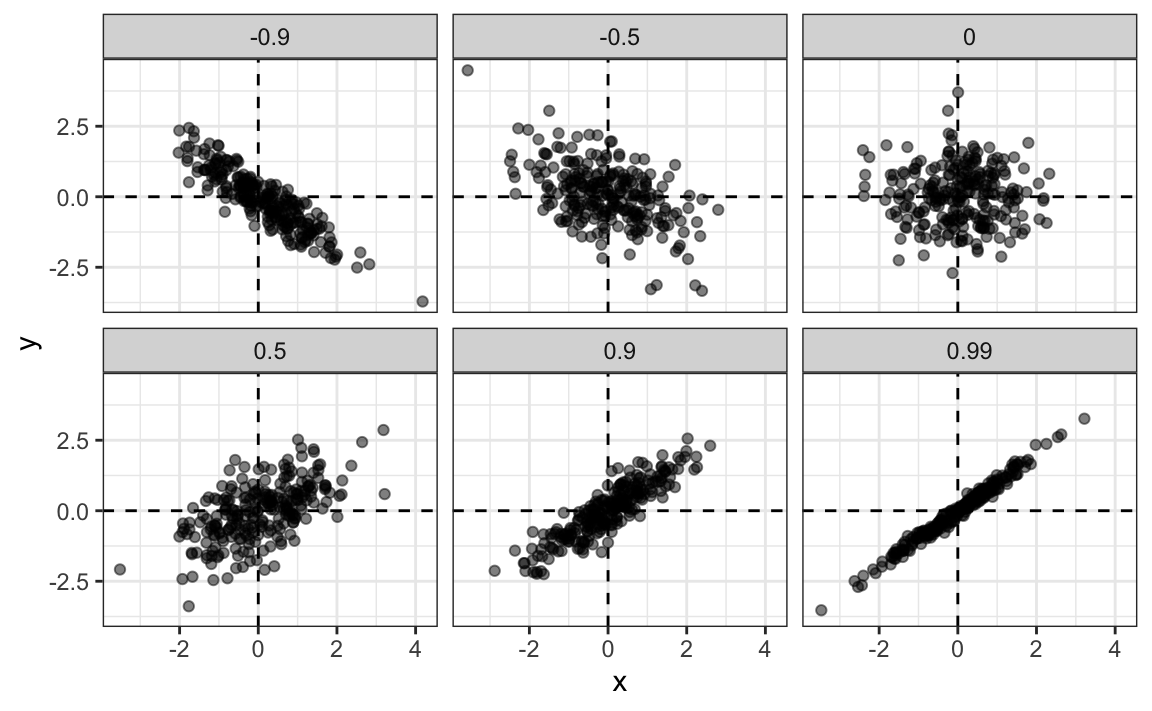
Một dẫn xuất tương tự, nhưng với  và hoàn toàn ngược lại, chứng tỏ mối tương quan phải lớn hơn hoặc bằng -1.

Đối với các cặp khác, mối tương quan nằm trong khoảng từ -1 đến 1. Mối tương quan, được tính với hàm , giữa chiều cao của cha và con trai là khoảng 0, 5:cor

|  |
| --- |
| galton\_heights |> summarize(r = cor(father, son)) |> pull(r)  #> [1] 0.433 |

Vì những lý do tương tự như được giải thích trong Phần [10.2.1](https://rafalab.dfci.harvard.edu/dsbook-part-2/inference/models.html#sec-population-sd) cho độ lệch chuẩn, chia cho thay vì .cor(x,y)length(x)-1length(x)

Để xem dữ liệu trông như thế nào đối với các giá trị khác nhau của �, đây là sáu ví dụ về các cặp có mối tương quan từ -0,9 đến 0,99:



13.2.1 Tương quan mẫu là một biến ngẫu nhiên

Trước khi chúng ta tiếp tục kết nối mối tương quan với hồi quy, chúng ta hãy nhắc nhở bản thân về sự biến đổi ngẫu nhiên.

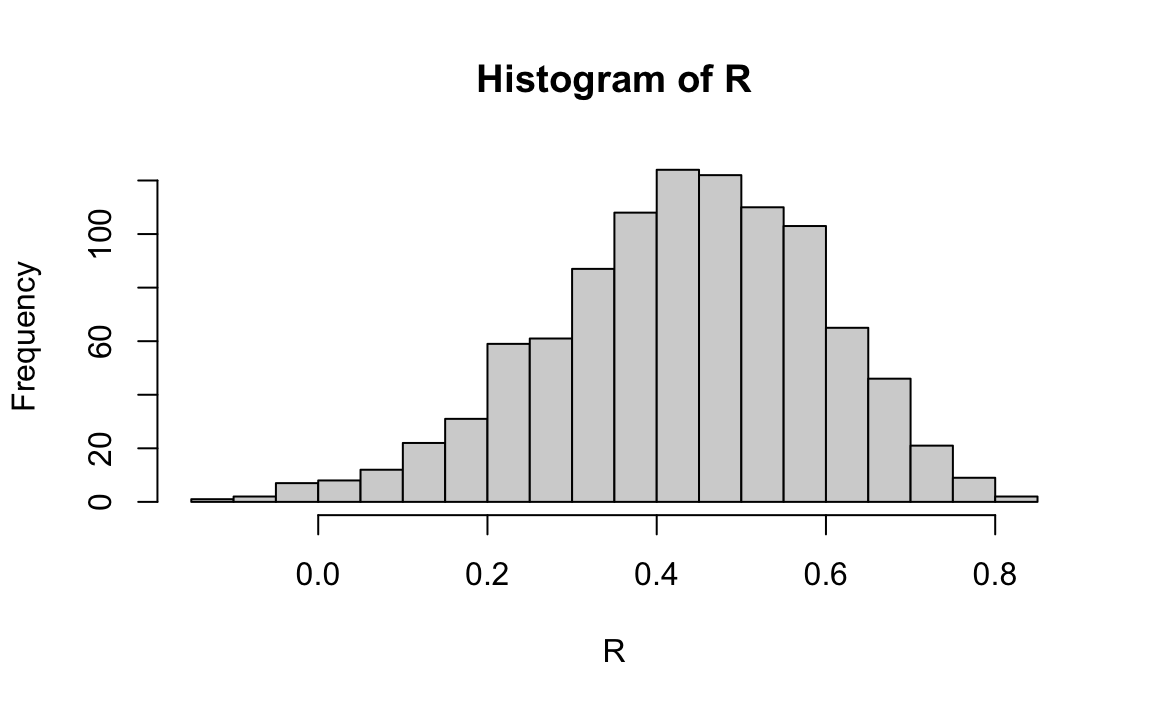
Trong hầu hết các ứng dụng khoa học dữ liệu, chúng tôi quan sát dữ liệu bao gồm các biến thể ngẫu nhiên. Ví dụ, trong nhiều trường hợp, chúng tôi không quan sát dữ liệu cho toàn bộ dân số quan tâm mà là cho một mẫu ngẫu nhiên. Cũng như độ lệch trung bình và độ lệch chuẩn, tương *quan mẫu* là ước tính được sử dụng phổ biến nhất về tương quan dân số. Điều này ngụ ý rằng mối tương quan mà chúng ta tính toán và sử dụng làm tóm tắt là một biến ngẫu nhiên.

Bằng cách minh họa, hãy giả sử rằng 179 cặp cha và con trai là toàn bộ dân số của chúng ta. Một nhà di truyền học kém may mắn hơn chỉ có thể đủ khả năng đo lường từ một mẫu ngẫu nhiên gồm 25 cặp. Tương quan mẫu có thể được tính với:

|  |
| --- |
| R <- sample\_n(galton\_heights, 25, replace = TRUE) |>    summarize(r = cor(father, son)) |> pull(r) |

R là một biến ngẫu nhiên. Chúng ta có thể chạy mô phỏng Monte Carlo để xem sự phân bố của nó:

|  |
| --- |
| B <- 1000  N <- 25  R <- replicate(B, {    sample\_n(galton\_heights, N, replace = TRUE) |>      summarize(r = cor(father, son)) |>      pull(r)  })  hist(R, breaks = 20) |



Chúng ta thấy rằng giá trị kỳ vọng của là tương quan dân số:R

|  |
| --- |
| mean(R)  #> [1] 0.431 |

và nó có sai số tiêu chuẩn tương đối cao so với phạm vi giá trị có thể lấy:R

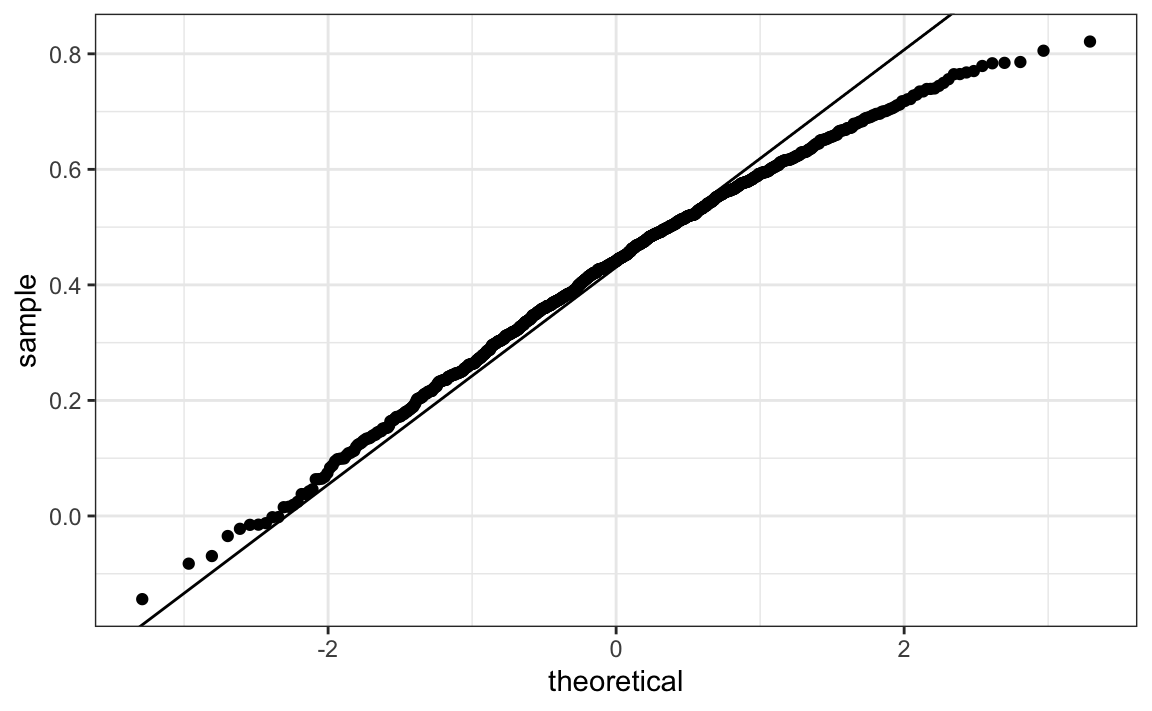
|  |
| --- |
| sd(R)  #> [1] 0.161 |

Vì vậy, khi giải thích các mối tương quan, hãy nhớ rằng các mối tương quan bắt nguồn từ các mẫu là các ước tính có chứa sự không chắc chắn.

Ngoài ra, lưu ý rằng vì mối tương quan mẫu là trung bình của các lần rút độc lập, giới hạn trung tâm thực sự được áp dụng. Do đó, cho đủ lớn , phân bố xấp xỉ bình thường với giá trị kỳ vọng R. Độ lệch chuẩn, hơi phức tạp để suy ra, là

Trong ví dụ của chúng tôi,  =25 dường như không đủ lớn để làm cho xấp xỉ trở thành một xấp xỉ tốt:

|  |
| --- |
| ggplot(aes(sample = R), data = data.frame(R)) +    stat\_qq() +    geom\_abline(intercept = mean(R), slope = sqrt((1 - mean(R)^2)/(N - 2))) |

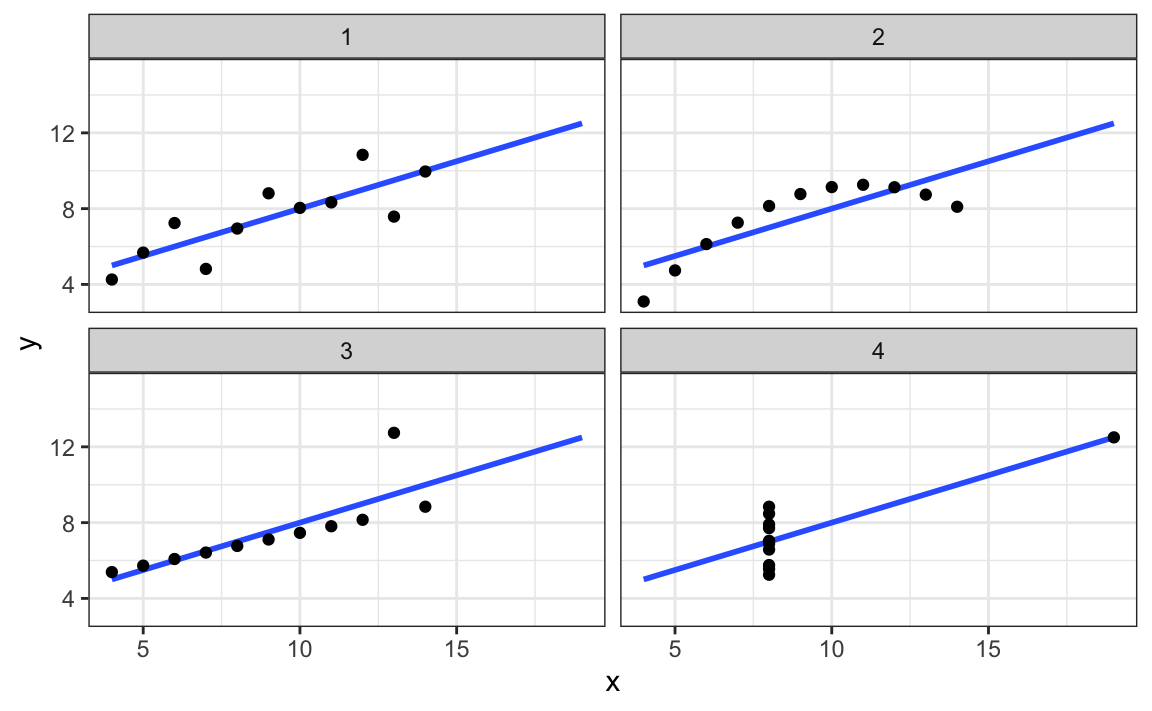


Nếu bạn tăng , bạn sẽ thấy sự phân bố hội tụ về mức bình thường.

13.2.2 Tương quan không phải lúc nào cũng là một bản tóm tắt hữu ích

Tương quan không phải lúc nào cũng là một bản tóm tắt tốt về mối quan hệ giữa hai biến. Bốn bộ dữ liệu nhân tạo sau đây, được gọi là bộ tứ của Anscombe, minh họa nổi tiếng cho điểm này. Tất cả các cặp này có mối tương quan 0,82:

#> `geom\_smooth()` using formula = 'y ~ x'



Tương quan chỉ có ý nghĩa trong một bối cảnh cụ thể. Để giúp chúng ta hiểu khi nào mối tương quan đó có ý nghĩa như một thống kê tóm tắt, chúng ta sẽ trở lại ví dụ về dự đoán chiều cao của con trai bằng cách sử dụng chiều cao của cha mình. Điều này sẽ giúp thúc đẩy và xác định hồi quy tuyến tính. Chúng tôi bắt đầu bằng cách chứng minh mối tương quan có thể hữu ích như thế nào để dự đoán.

## 13.3 Kỳ vọng có điều kiện

Giả sử chúng ta được yêu cầu đoán chiều cao của một đứa con trai được chọn ngẫu nhiên và chúng ta không biết chiều cao của cha nó. Bởi vì sự phân bố chiều cao của con trai là xấp xỉ bình thường, chúng ta biết chiều cao trung bình, 69,2, là giá trị có tỷ lệ cao nhất và sẽ là dự đoán có cơ hội giảm thiểu sai số cao nhất. Nhưng điều gì sẽ xảy ra nếu chúng ta được cho biết rằng người cha cao hơn mức trung bình, giả sử cao 72 inch, chúng ta vẫn đoán 69.2 cho con trai?

Nó chỉ ra rằng nếu chúng ta có thể thu thập dữ liệu từ một số lượng rất lớn các ông bố là 72 inch, sự phân bố chiều cao của con trai họ sẽ được phân phối bình thường. Điều này ngụ ý rằng trung bình của phân phối được tính trên tập con này sẽ là dự đoán tốt nhất của chúng tôi.

Nói chung, chúng tôi gọi cách tiếp cận này *là điều hòa*. Ý tưởng chung là chúng tôi phân tầng dân số thành các nhóm và tính toán tóm tắt trong mỗi nhóm. Để cung cấp một mô tả toán học về điều hòa, hãy xem xét chúng ta có một quần thể các cặp giá trị , ví dụ như tất cả chiều cao của cha và con trai ở Anh. Trong chương trước, chúng ta đã học được rằng nếu bạn lấy một cặp ngẫu nhiên , giá trị kỳ vọng và công cụ dự đoán tốt nhất của  Là , dân số trung bình  tôi. Tuy nhiên, chúng tôi không còn quan tâm đến dân số nói chung, thay vào đó chúng tôi chỉ quan tâm đến tập hợp con của một quần thể với một dân số cụ thể tôi giá trị, 72 inch trong ví dụ của chúng tôi. Tập hợp con này của dân số, cũng là một dân số và do đó các nguyên tắc và tính chất tương tự mà chúng ta đã học được áp dụng. Các tôi Trong phân quần thể có một phân phối, được gọi là phân phối có điều kiện và *phân phối* này có một giá trị kỳ vọng được gọi là *kỳ vọng có điều kiện*. Trong ví dụ của chúng tôi, kỳ vọng có điều kiện là chiều cao trung bình của tất cả con trai ở Anh có cha là 72 inch. Ký hiệu thống kê cho kỳ vọng có điều kiện là

với  đại diện cho giá trị cố định xác định tập hợp con đó, ví dụ 72 inch. Tương tự, chúng tôi biểu thị độ lệch chuẩn của địa tầng với

Bởi vì kỳ vọng có điều kiện  là công cụ dự đoán tốt nhất cho biến ngẫu nhiên  cho một cá nhân trong các tầng được xác định bởi , nhiều thách thức khoa học dữ liệu giảm xuống để ước tính số lượng này. Độ lệch chuẩn có điều kiện định lượng độ chính xác của dự đoán.

Trong ví dụ chúng tôi đã xem xét, chúng tôi quan tâm đến việc tính toán chiều cao trung bình của con trai với *điều kiện* người cha là 72 inch. Chúng tôi muốn ước tính sử dụng mẫu do Galton thu thập. Trước đây chúng tôi đã học được rằng trung bình mẫu là cách tiếp cận ưa thích để ước tính trung bình dân số. Tuy nhiên, một thách thức khi sử dụng phương pháp này để ước tính các kỳ vọng có điều kiện là đối với dữ liệu liên tục, chúng tôi không có nhiều điểm dữ liệu khớp chính xác với một giá trị trong mẫu của chúng tôi. Ví dụ: chúng tôi chỉ có:

|  |
| --- |
| sum(galton\_heights$father == 72)  #> [1] 8 |

những người cha chính xác là 72 inch. Nếu chúng tôi thay đổi số thành 72,5, chúng tôi thậm chí còn nhận được ít điểm dữ liệu hơn:

|  |
| --- |
| sum(galton\_heights$father == 72.5)  #> [1] 1 |

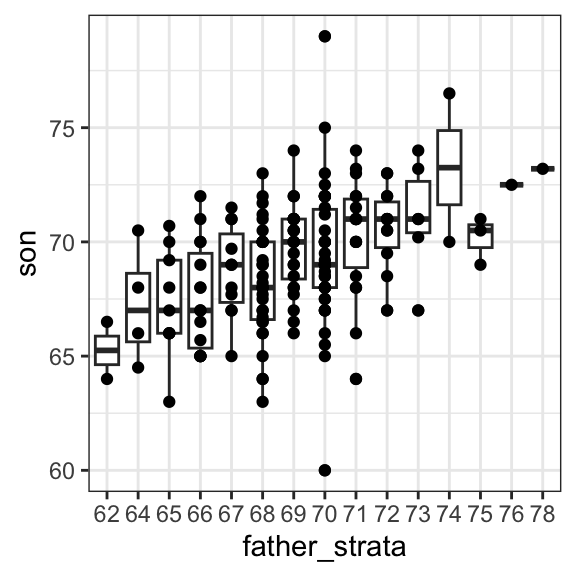
Một cách thực tế để cải thiện các ước tính này về các kỳ vọng có điều kiện, là xác định các tầng có giá trị tương tự của . Trong ví dụ của chúng tôi, chúng tôi có thể làm tròn chiều cao của cha đến inch gần nhất và giả sử rằng tất cả chúng đều là 72 inch. Nếu chúng ta làm điều này, chúng ta sẽ kết thúc với dự đoán sau đây cho con trai của một người cha cao 72 inch:

|  |
| --- |
| conditional\_avg <- galton\_heights |>    filter(round(father) == 72) |>    summarize(avg = mean(son)) |>    pull(avg)  conditional\_avg  #> [1] 70.5 |

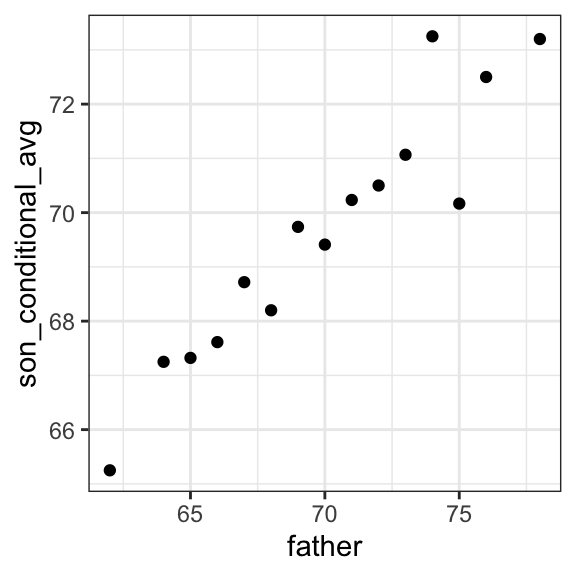
Lưu ý rằng một người cha 72 inch cao hơn mức trung bình - cụ thể, (72,0 - 69,1) / 2,5 = 1,1 độ lệch chuẩn cao hơn người cha trung bình. Dự đoán của chúng tôi 70,5 cũng cao hơn mức trung bình, nhưng chỉ lớn hơn 0,49 độ lệch chuẩn so với con trai trung bình. Con trai của những người cha 72 inch đã *thụt lùi* một số về chiều cao trung bình. Chúng tôi nhận thấy rằng việc giảm bao nhiêu SD cao hơn là khoảng 0,5, đó là mối tương quan. Như chúng ta sẽ thấy trong phần sau, đây không phải là một sự trùng hợp ngẫu nhiên.

Nếu chúng ta muốn đưa ra dự đoán về bất kỳ độ cao nào, không chỉ 72, chúng ta có thể áp dụng cùng một cách tiếp cận cho từng tầng. Sự phân tầng theo sau là các biểu đồ hộp cho phép chúng ta xem sự phân bố của từng nhóm:

|  |
| --- |
| galton\_heights |> mutate(father\_strata = factor(round(father))) |>    ggplot(aes(father\_strata, son)) +    geom\_boxplot() +    geom\_point() |



Không có gì đáng ngạc nhiên, trung tâm của các nhóm đang tăng theo chiều cao. Hơn nữa, các trung tâm này dường như tuân theo một mối quan hệ tuyến tính. Dưới đây chúng tôi vẽ mức trung bình của mỗi nhóm. Nếu chúng ta tính đến việc các trung bình này là các biến ngẫu nhiên với sai số chuẩn, dữ liệu phù hợp với các điểm này theo một đường thẳng:



Thực tế là các trung bình có điều kiện này tuân theo một đường không phải là một sự trùng hợp ngẫu nhiên. Trong phần tiếp theo, chúng tôi giải thích rằng đường trung bình này tuân theo là cái mà chúng tôi gọi là *đường hồi quy*, giúp cải thiện độ chính xác của các ước tính của chúng tôi. Tuy nhiên, không phải lúc nào cũng thích hợp để ước tính các kỳ vọng có điều kiện với đường hồi quy, vì vậy chúng tôi cũng mô tả lý thuyết biện minh của Galton cho việc sử dụng đường hồi quy.

## 13.4 Đường hồi quy

Nếu chúng ta đang dự đoán một biến ngẫu nhiên  Biết giá trị của người khác  sử dụng một đường hồi quy, sau đó chúng tôi dự đoán rằng đối với mọi độ lệch chuẩn,  đó  tăng trên mức trung bình , dự đoán của chúng tôi  tăng  Độ lệch chuẩn  trên mức trung bình  với  Mối tương quan giữa  và . Do đó, công thức cho hồi quy là:

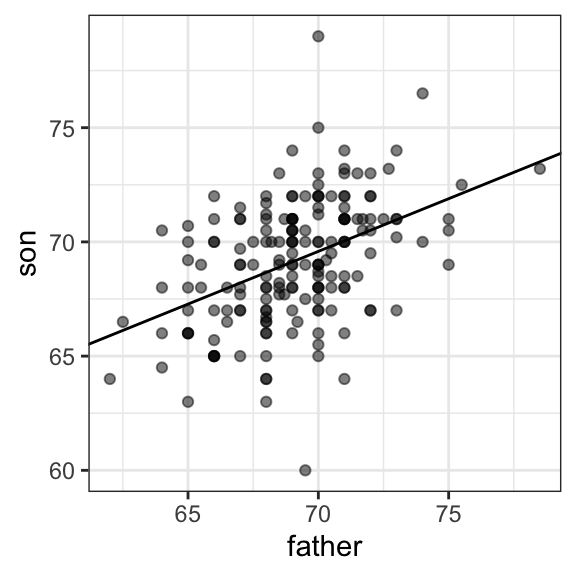
Chúng ta có thể viết lại nó như thế này:

Nếu có mối tương quan hoàn hảo, đường hồi quy dự đoán mức tăng có cùng số SD. Nếu có 0 tương quan, thì chúng ta không sử dụng  ở tất cả cho dự đoán và chỉ đơn giản là dự đoán mức trung bình . Đối với các giá trị từ 0 đến 1, dự đoán nằm ở giữa. Nếu mối tương quan là âm, chúng tôi dự đoán giảm thay vì tăng.

Lưu ý rằng nếu mối tương quan là dương và thấp hơn 1, dự đoán của chúng tôi gần hơn, theo đơn vị tiêu chuẩn, với chiều cao trung bình so với giá trị được sử dụng để dự đoán, , là trung bình của s. Đây là lý do tại sao chúng tôi gọi nó là *hồi quy*: con trai thoái lui về chiều cao trung bình. Trên thực tế, tiêu đề bài báo của Galton là: *Hồi quy về sự tầm thường trong tầm vóc di truyền*. Để thêm các đường hồi quy vào các ô, chúng ta sẽ cần công thức trên dưới dạng:

Ở đây chúng tôi thêm dòng hồi quy vào dữ liệu gốc:

|  |
| --- |
| mu\_x <- mean(galton\_heights$father)  mu\_y <- mean(galton\_heights$son)  s\_x <- sd(galton\_heights$father)  s\_y <- sd(galton\_heights$son)  r <- cor(galton\_heights$father, galton\_heights$son)  galton\_heights |>    ggplot(aes(father, son)) +    geom\_point(alpha = 0.5) +    geom\_abline(slope = r \* s\_y/s\_x, intercept = mu\_y - r \* s\_y/s\_x \* mu\_x) |



Công thức hồi quy ngụ ý rằng nếu trước tiên chúng ta chuẩn hóa các biến, nghĩa là trừ trung bình và chia cho độ lệch chuẩn, thì đường hồi quy có giao nhau 0 và độ dốc bằng tương quan . Bạn có thể tạo cùng một cốt truyện, nhưng sử dụng các đơn vị tiêu chuẩn như thế này:

|  |
| --- |
| galton\_heights |>    ggplot(aes(scale(father), scale(son))) +    geom\_point(alpha = 0.5) +    geom\_abline(intercept = 0, slope = r) |

## 13.5 Hồi quy cải thiện độ chính xác

Hãy so sánh hai cách tiếp cận dự đoán mà chúng tôi đã trình bày:

1. Làm tròn chiều cao của các ông bố đến inch gần nhất, phân tầng, và sau đó lấy trung bình.

2. Tính toán đường hồi quy và sử dụng nó để dự đoán.

Chúng tôi sử dụng lấy mẫu mô phỏng Monte Carlo Gia đình:

|  |
| --- |
| set.seed(1983)  conditional\_avg <- replicate(B, {    dat <- sample\_n(galton\_heights, N)    dat |> filter(round(father) == 72) |>      summarize(avg = mean(son)) |>      pull(avg)    })  regression\_prediction <- replicate(B, {    dat <- sample\_n(galton\_heights, N)    mu\_x <- mean(dat$father)    mu\_y <- mean(dat$son)    s\_x <- sd(dat$father)    s\_y <- sd(dat$son)    r <- cor(dat$father, dat$son)    mu\_y + r\*(72 - mu\_x)/s\_x\*s\_y  }) |

Mặc dù giá trị kỳ vọng của hai biến ngẫu nhiên này là như nhau:

|  |
| --- |
| mean(conditional\_avg, na.rm = TRUE)  #> [1] 70.5  mean(regression\_prediction)  #> [1] 70.5 |

Sai số tiêu chuẩn cho dự đoán hồi quy nhỏ hơn đáng kể:

|  |
| --- |
| sd(conditional\_avg, na.rm = TRUE)  #> [1] 0.964  sd(regression\_prediction)  #> [1] 0.452 |

Do đó, đường hồi quy ổn định hơn nhiều so với giá trị trung bình có điều kiện. Có một lý do trực quan cho việc này. Trung bình có điều kiện được tính trên một tập hợp con tương đối nhỏ: những người cha cao khoảng 72 inch. Trên thực tế, trong một số hoán vị, chúng tôi không có dữ liệu, đó là lý do tại sao chúng tôi sử dụng . Hồi quy luôn sử dụng tất cả dữ liệu.na.rm=TRUE

Vậy tại sao không luôn luôn sử dụng hồi quy để dự đoán? Bởi vì nó không phải lúc nào cũng thích hợp. Ví dụ: Anscombe cung cấp các trường hợp mà dữ liệu không có mối quan hệ tuyến tính. Vậy chúng ta có hợp lý khi sử dụng đường hồi quy để dự đoán không? Galton đã trả lời điều này trong dữ liệu tích cực cho chiều cao. Lời biện minh, mà chúng tôi đưa vào phần tiếp theo, có phần nâng cao hơn phần còn lại của chương.

## 13.6 Phân bố chuẩn hai biến

Tương quan và độ dốc hồi quy là một thống kê tóm tắt được sử dụng rộng rãi, nhưng chúng thường bị lạm dụng hoặc hiểu sai. Các ví dụ của Anscombe cung cấp các trường hợp tập dữ liệu được đơn giản hóa quá mức, trong đó tóm tắt với mối tương quan sẽ là một sai lầm. Nhưng có nhiều ví dụ thực tế hơn.

Cách chính chúng ta thúc đẩy việc sử dụng tương quan liên quan đến cái được gọi là *phân phối chuẩn hai biến*.

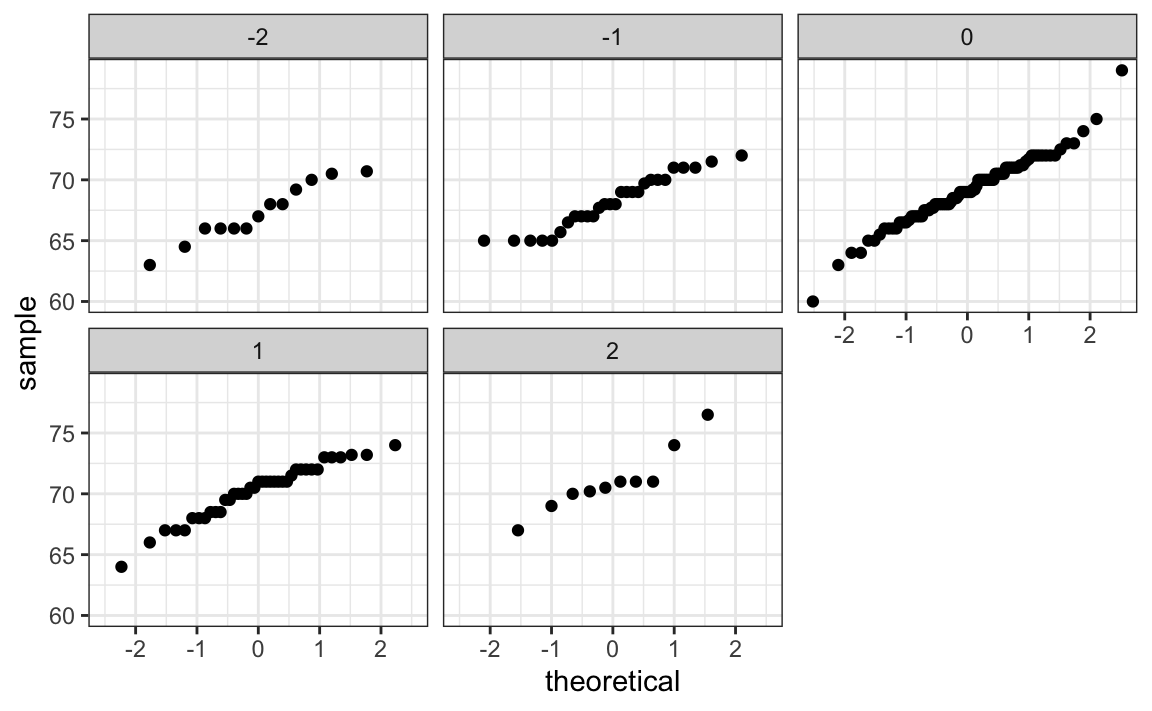
Khi một cặp biến ngẫu nhiên được xấp xỉ bởi phân phối chuẩn hai biến, các biểu đồ tán xạ trông giống như hình bầu dục. Như chúng ta đã thấy trong Phần [13.2](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/regression.html#sec-corr-coef)), chúng có thể mỏng (tương quan cao) hoặc hình tròn (không có mối tương quan.

Một cách kỹ thuật hơn để xác định phân phối chuẩn hai biến như sau: nếu  là một biến ngẫu nhiên phân bố thông thường,  cũng là một biến ngẫu nhiên phân phối bình thường và phân phối có điều kiện của  cho bất kỳ  là xấp xỉ bình thường, sau đó cặp này xấp xỉ hai biến bình thường. Khi ba hoặc nhiều biến có thuộc tính là mỗi cặp là hai biến bình thường, chúng ta nói rằng các biến tuân theo phân phối chuẩn *đa biến* hoặc chúng cùng *bình thường*.

* hoặc đơn giản là các biến là *cùng bình thường*

Nếu chúng ta nghĩ rằng dữ liệu chiều cao được xấp xỉ tốt bởi phân bố chuẩn hai biến, thì chúng ta sẽ thấy xấp xỉ bình thường cho mỗi tầng. Ở đây chúng ta phân tầng chiều cao con trai theo chiều cao của người cha được tiêu chuẩn hóa và thấy rằng giả định dường như đúng:

|  |
| --- |
| galton\_heights |>    mutate(z\_father = round((father - mean(father)) / sd(father))) |>    filter(z\_father %in% -2:2) |>    ggplot() +    stat\_qq(aes(sample = son)) +    facet\_wrap( ~ z\_father) |



Bây giờ chúng ta quay trở lại để xác định mối tương quan. Galton đã sử dụng thống kê toán học để chứng minh rằng, khi hai biến tuân theo phân phối chuẩn hai biến, việc tính toán đường hồi quy tương đương với tính toán kỳ vọng có điều kiện. Chúng tôi không hiển thị đạo hàm ở đây, nhưng chúng tôi có thể chỉ ra rằng theo giả định này, cho bất kỳ giá trị nhất định nào của , giá trị kỳ vọng của  theo cặp  Là:

Đây là đường hồi quy, có độ dốc

và đánh chặn . Nó tương đương với phương trình hồi quy mà chúng tôi đã trình bày trước đó có thể được viết như thế này:

Điều này ngụ ý rằng, nếu dữ liệu của chúng ta xấp xỉ hai biến, đường hồi quy cho xác suất có điều kiện. Do đó, chúng ta có thể có được ước tính ổn định hơn nhiều về kỳ vọng có điều kiện bằng cách tìm đường hồi quy và sử dụng nó để dự đoán.

Tóm lại, nếu dữ liệu của chúng tôi xấp xỉ hai biến, thì kỳ vọng có điều kiện, dự đoán tốt nhất về  cho chúng ta biết giá trị của X, được cho bởi đường hồi quy.

## 13.7 Giải thích phương sai

Lý thuyết chuẩn hai biến cũng cho chúng ta biết rằng độ lệch chuẩn của phân phối *có điều kiện* được mô tả ở trên là:

Để xem tại sao điều này là trực quan, hãy chú ý rằng không có điều kiện, , chúng tôi đang xem xét sự biến đổi của tất cả các con trai. Nhưng một khi chúng ta điều kiện, chúng ta chỉ nhìn vào sự biến đổi của những đứa con trai với một người cha cao, 72 inch. Nhóm này đều sẽ có xu hướng hơi cao nên độ lệch chuẩn giảm.

Cụ thể, nó được giảm xuống  so với ban đầu. Chúng ta có thể nói rằng chiều cao của cha "giải thích" 13% sự thay đổi quan sát được ở chiều cao con trai.

Tuyên bố " giải thích như vậy và như vậy phần trăm của sự thay đổi" thường được sử dụng trong các bài báo học thuật. Trong trường hợp này, phần trăm này thực sự đề cập đến phương sai (bình phương SD). Vì vậy, nếu dữ liệu là hai biến bình thường, phương sai sẽ giảm đi , vì vậy chúng tôi nói rằng  Giải thích  (bình phương tương quan) của phương sai.

Nhưng điều quan trọng cần nhớ là câu lệnh "giải thích phương sai" chỉ có ý nghĩa khi dữ liệu được xấp xỉ bởi phân phối chuẩn hai biến.

## 13.8 Có hai dòng hồi quy

Chúng tôi đã tính toán một đường hồi quy để dự đoán chiều cao của con trai từ chiều cao của cha. Chúng tôi đã sử dụng các tính toán sau:

|  |
| --- |
| mu\_x <- mean(galton\_heights$father)  mu\_y <- mean(galton\_heights$son)  s\_x <- sd(galton\_heights$father)  s\_y <- sd(galton\_heights$son)  r <- cor(galton\_heights$father, galton\_heights$son)  m\_1 <-  r \* s\_y / s\_x  b\_1 <- mu\_y - m\_1\*mu\_x |

cung cấp cho chúng ta chức năng

Điều gì sẽ xảy ra nếu chúng ta muốn dự đoán chiều cao của người cha dựa trên con trai? Điều quan trọng cần biết là điều này không được xác định bằng cách tính toán hàm nghịch đảo: 37.3 }/ 0.5.

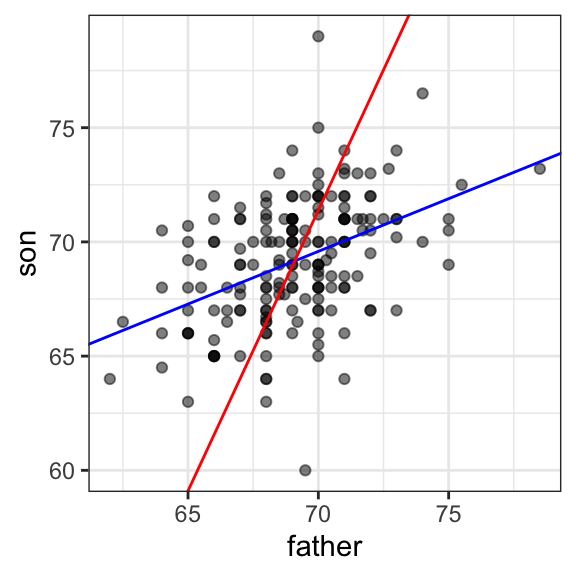
Chúng ta cần tính toán . Vì dữ liệu xấp xỉ bình thường hai biến, lý thuyết được mô tả ở trên cho chúng ta biết rằng kỳ vọng có điều kiện này sẽ tuân theo một đường có độ dốc và chặn:

|  |
| --- |
| m\_2 <-  r \* s\_x / s\_y  b\_2 <- mu\_x - m\_2 \* mu\_y |

Vì vậy, chúng tôi nhận được 40,9 + 0,41y. Một lần nữa chúng ta thấy hồi quy về mức trung bình: dự đoán cho người cha gần với mức trung bình của người cha hơn chiều cao của con trai  là con trai trung bình.

Dưới đây là biểu đồ cho thấy hai đường hồi quy, với màu xanh lam cho chiều cao con trai dự đoán với chiều cao của cha và màu đỏ để dự đoán chiều cao của cha với chiều cao con trai:

|  |
| --- |
| galton\_heights |>    ggplot(aes(father, son)) +    geom\_point(alpha = 0.5) +    geom\_abline(intercept = b\_1, slope = m\_1, col = "blue") +    geom\_abline(intercept = -b\_2/m\_2, slope = 1/m\_2, col = "red") |



## 13.9 Mô hình tuyến tính

Bây giờ chúng ta đã sẵn sàng để hiểu tiêu đề của phần này của cuốn sách. Cụ thể, kết nối giữa *các mô hình* hồi quy và tuyến tính. Chúng tôi đã mô tả làm thế nào nếu dữ liệu là hai biến bình thường thì các kỳ vọng có điều kiện tuân theo đường hồi quy. Thực tế là kỳ vọng có điều kiện là một dòng không phải là một giả định bổ sung mà là một kết quả có nguồn gốc. Tuy nhiên, trong thực tế, người ta thường viết rõ ràng một mô hình mô tả mối quan hệ giữa hai hoặc nhiều biến bằng *mô hình tuyến tính*.

Chúng tôi lưu ý rằng tuyến tính ở đây không chỉ đề cập đến các dòng, mà là thực tế là kỳ vọng có điều kiện là sự kết hợp *tuyến* tính của các đại lượng đã biết. Trong toán học, khi chúng ta nhân mỗi biến với một hằng số và sau đó cộng chúng lại với nhau, chúng ta nói rằng chúng ta đã hình thành một *tổ hợp tuyến tính* của các biến. Chẳng hạn  là sự kết hợp tuyến tính của ,  và . Chúng ta cũng có thể thêm một hằng số như vậy  cũng là sự kết hợp tuyến tính của �, �và �.

Trước đây chúng tôi đã mô tả làm thế nào nếu  và  là hai biến bình thường, sau đó nếu chúng ta chỉ nhìn vào các cặp với X=x sau đó  tuân theo phân phối chuẩn với giá trị kỳ vọng , là một hàm tuyến tính của  và độ lệch chuẩn Điều đó không phụ thuộc vào . Lưu ý rằng nếu chúng ta viết

thì nếu chúng ta giả định  theo một phân phối chuẩn với giá trị kỳ vọng 0 và độ lệch chuẩn cố định, sau đó  có các thuộc tính tương tự như thiết lập hồi quy đã cho chúng ta: nó tuân theo một phân phối chuẩn, giá trị kỳ vọng là một hàm tuyến tính và độ lệch chuẩn không phụ thuộc vào .

|  |
| --- |
| Trong sách giáo khoa thống kê, các S được gọi là "lỗi", ban đầu đại diện cho các lỗi đo lường trong các ứng dụng ban đầu của các mô hình này. Những lỗi này có liên quan đến sự không chính xác trong việc đo chiều cao, cân nặng hoặc khoảng cách. Tuy nhiên, thuật ngữ "lỗi" hiện được sử dụng rộng rãi hơn, ngay cả khi s không nhất thiết biểu thị một lỗi thực tế. Ví dụ, trong trường hợp chiều cao, nếu ai đó cao hơn 2 inch so với dự kiến dựa trên chiều cao của cha mẹ họ, 2 inch đó không nên được coi là lỗi. Mặc dù thiếu độ chính xác mô tả, thuật ngữ "lỗi" được sử dụng để làm sáng tỏ sự thay đổi không giải thích được trong mô hình, không liên quan đến các thuật ngữ được bao gồm khác. |

Nếu chúng ta chỉ định một mô hình tuyến tính cho dữ liệu của Galton, chúng ta sẽ biểu thị  quan sát chiều cao của cha với , sau đó chúng tôi mô hình hóa  Son Heights chúng tôi đang cố gắng dự đoán với:

Ở đây tôi là chiều cao của người cha, được cố định (không phải ngẫu nhiên) do điều hòa, và tôi là chiều cao của con trai ngẫu nhiên mà chúng tôi muốn dự đoán. Chúng ta có thể giả định thêm rằng  độc lập với nhau và tất cả đều có cùng độ lệch chuẩn.

Trong mô hình trên, chúng ta biết , nhưng để có một mô hình hữu ích để dự đoán, chúng ta cần  và . Chúng tôi ước tính những điều này từ dữ liệu. Một khi chúng ta làm điều này, chúng ta có thể dự đoán chiều cao của con trai cho bất kỳ chiều cao nào của người cha . Chúng tôi chỉ ra cách làm điều này trong phần tiếp theo.

Mặc dù mô hình này hoàn toàn giống với mô hình mà chúng tôi đã rút ra trước đó bằng cách giả định dữ liệu bình thường hai biến, một sự khác biệt hơi sắc thái là trong cách tiếp cận đầu tiên, chúng tôi giả định dữ liệu là bình thường hai biến và mô hình tuyến tính được suy ra, không giả định. Trong thực tế, các mô hình tuyến tính chỉ được giả định mà không nhất thiết phải giả định tính bình thường: sự phân bố của s không nhất thiết phải được chỉ định. Tuy nhiên, nếu dữ liệu của bạn là hai biến bình thường, mô hình tuyến tính ở trên sẽ giữ. Nếu dữ liệu của bạn không phải là hai biến bình thường, thì bạn sẽ cần phải có những cách khác để biện minh cho mô hình.

Một lý do khiến các mô hình tuyến tính phổ biến là chúng *có thể diễn giải được*. Trong trường hợp dữ liệu của Galton, chúng ta có thể giải thích dữ liệu như sau: do gen di truyền, dự đoán chiều cao của con trai tăng lên  Đối với mỗi inch, chúng tôi tăng chiều cao của người cha . Bởi vì không phải con trai nào cũng có bố chiều cao  có chiều cao bằng nhau, chúng ta cần thuật ngữ , điều này giải thích sự thay đổi còn lại. Sự thay đổi còn lại này bao gồm hiệu ứng di truyền của người mẹ, các yếu tố môi trường và sự ngẫu nhiên sinh học khác.

Với cách chúng tôi viết mô hình trên, đánh chặn  không dễ hiểu lắm vì nó là chiều cao dự đoán của một đứa con trai với một người cha không có chiều cao. Do hồi quy về giá trị trung bình, dự đoán thường sẽ lớn hơn 0 một chút. Để làm cho tham số độ dốc dễ hiểu hơn, chúng ta có thể viết lại mô hình một chút thành:

với tôi trung bình của . Trong trường hợp này  đại diện cho chiều cao khi là chiều cao của con trai của một người cha trung bình.

Sau đó, cụ thể là trong Chương [Chương 14](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html) và các mô hình hiệu ứng @treatment, chúng ta sẽ thấy cách biểu diễn mô hình tuyến tính cho phép chúng ta sử dụng cùng một khung toán học trong các bối cảnh khác và đạt được các mục tiêu phức tạp hơn so với dự đoán một biến từ một biến khác.

## 13.10 Ước tính bình phương nhỏ nhất

Để các mô hình tuyến tính trở nên hữu ích, chúng ta phải ước tính những điều chưa biết s. Cách tiếp cận tiêu chuẩn trong khoa học là tìm các giá trị giảm thiểu khoảng cách của mô hình được trang bị đến dữ liệu. Sau đây được gọi là phương trình bình phương nhỏ nhất (LS) và chúng ta sẽ thấy nó thường xuyên trong chương này. Đối với dữ liệu của Galton, chúng tôi sẽ viết:

Đại lượng này được gọi là tổng dư của bình phương (RSS). Khi chúng ta tìm thấy các giá trị thu nhỏ RSS, chúng ta sẽ gọi các giá trị là ước tính bình phương nhỏ nhất (LSE) và biểu thị chúng bằng và . Hãy chứng minh điều này với tập dữ liệu đã xác định trước đó:

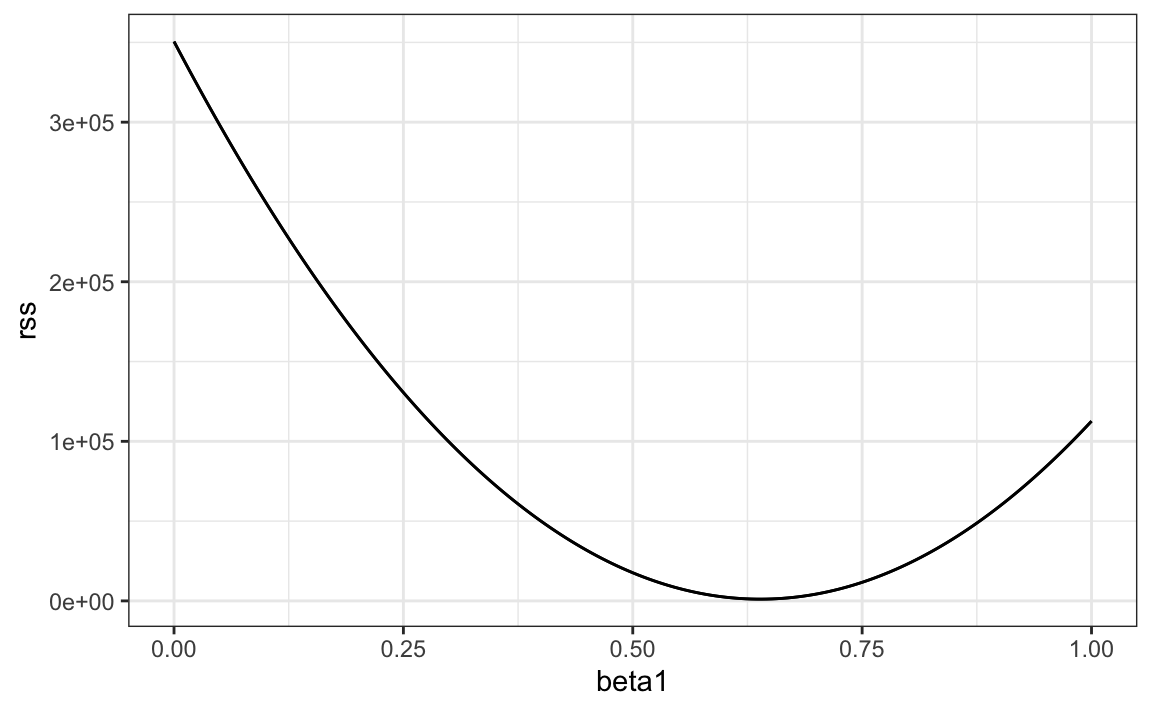
|  |
| --- |
| library(HistData)  set.seed(1983)  galton\_heights <- GaltonFamilies |>    filter(gender == "male") |>    group\_by(family) |>    sample\_n(1) |>    ungroup() |>    select(father, childHeight) |>    rename(son = childHeight) |

Hãy viết một hàm tính toán RSS cho bất kỳ cặp giá trị nào  và .

|  |
| --- |
| rss <- function(beta0, beta1, data){    resid <- galton\_heights$son - (beta0 + beta1\*galton\_heights$father)    return(sum(resid^2))  } |

Vì vậy, đối với bất kỳ cặp giá trị nào, chúng tôi nhận được RSS. Dưới đây là một biểu đồ của RSS như một chức năng của  Khi chúng tôi giữ  cố định ở tuổi 25.

|  |
| --- |
| beta1 = seq(0, 1, length = nrow(galton\_heights))  results <- data.frame(beta1 = beta1,                        rss = sapply(beta1, rss, beta0 = 25))  results |> ggplot(aes(beta1, rss)) + geom\_line() +    geom\_line(aes(beta1, rss)) |



Chúng ta có thể thấy mức tối thiểu rõ ràng cho  vào khoảng 0,65. Tuy nhiên, mức tối thiểu này cho  là cho khi , một giá trị chúng tôi tùy tiện chọn. Chúng ta không biết liệu (25, 0,65) có phải là cặp giảm thiểu phương trình trên tất cả các cặp có thể hay không.

Thử và sai sẽ không hoạt động trong trường hợp này. Chúng ta có thể tìm kiếm mức tối thiểu trong một lưới tốt  và giá trị, nhưng điều này tốn thời gian không cần thiết vì chúng ta có thể sử dụng phép tính: lấy đạo hàm từng phần, đặt chúng thành 0 và giải cho  và . Tất nhiên, nếu chúng ta có nhiều tham số, các phương trình này có thể khá phức tạp. Nhưng có những hàm trong R thực hiện các phép tính này cho chúng ta. Chúng ta sẽ tìm hiểu những điều này tiếp theo. Để tìm hiểu toán học đằng sau điều này, bạn có thể tham khảo một cuốn sách về các mô hình tuyến tính.

## 13.11 Chức năng lm

Trong R, chúng ta có thể thu được các ước tính bình phương nhỏ nhất bằng cách sử dụng hàm. Để phù hợp với mô hình: lm

với chiều cao của con trai và  Chiều cao của người cha, chúng ta có thể sử dụng mã này để có được ước tính bình phương nhỏ nhất.

|  |
| --- |
| fit <- lm(son ~ father, data = galton\_heights)  fit$coef  #> (Intercept)      father  #>      37.288       0.461 |

Cách phổ biến nhất chúng ta sử dụng là sử dụng ký tự để cho biết đâu là biến chúng ta đang dự đoán (bên trái ) và biến nào chúng ta đang sử dụng để dự đoán (bên phải của ). Việc đánh chặn được thêm tự động vào mô hình sẽ phù hợp. lm~lm~~

Đối tượng bao gồm nhiều thông tin hơn về sự phù hợp. Chúng ta có thể sử dụng hàm để trích xuất thêm thông tin này (không được hiển thị): fitsummary

|  |
| --- |
| summary(fit)  #>  #> Call:  #> lm(formula = son ~ father, data = galton\_heights)  #>  #> Residuals:  #>    Min     1Q Median     3Q    Max  #> -9.354 -1.566 -0.008  1.726  9.415  #>  #> Coefficients:  #>             Estimate Std. Error t value Pr(>|t|)  #> (Intercept)  37.2876     4.9862    7.48  3.4e-12 \*\*\*  #> father        0.4614     0.0721    6.40  1.4e-09 \*\*\*  #> ---  #> Signif. codes:  0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  #>  #> Residual standard error: 2.45 on 177 degrees of freedom  #> Multiple R-squared:  0.188,  Adjusted R-squared:  0.183  #> F-statistic: 40.9 on 1 and 177 DF,  p-value: 1.36e-09 |

Để hiểu một số thông tin có trong bản tóm tắt này, chúng ta cần nhớ rằng LSE là các biến ngẫu nhiên. Thống kê toán học cho chúng ta một số ý tưởng về sự phân bố của các biến ngẫu nhiên này.

Trong Chương @, sau khi mô tả một nghiên cứu điển hình phức tạp hơn, chúng ta tìm hiểu thêm về cách áp dụng hồi quy trong R.

## 13.12 LSE là các biến ngẫu nhiên

LSE được lấy từ dữ liệu , là sự hiện thực hóa các biến ngẫu nhiên Điều này ngụ ý rằng ước tính của chúng tôi là các biến ngẫu nhiên. Để thấy điều này, chúng ta có thể chạy mô phỏng Monte Carlo, trong đó chúng ta giả sử dữ liệu chiều cao của con trai và cha xác định dân số, lấy một mẫu kích thước ngẫu nhiên và tính hệ số độ dốc hồi quy cho từng hệ số:

|  |
| --- |
| B <- 1000  N <- 50  lse <- replicate(B, {    sample\_n(galton\_heights, N, replace = TRUE) |>      lm(son ~ father, data = \_) |>      coef()  })  lse <- data.frame(beta\_0 = lse[1,], beta\_1 = lse[2,]) |

Chúng ta có thể thấy sự thay đổi của các ước tính bằng cách vẽ các phân phối của chúng:

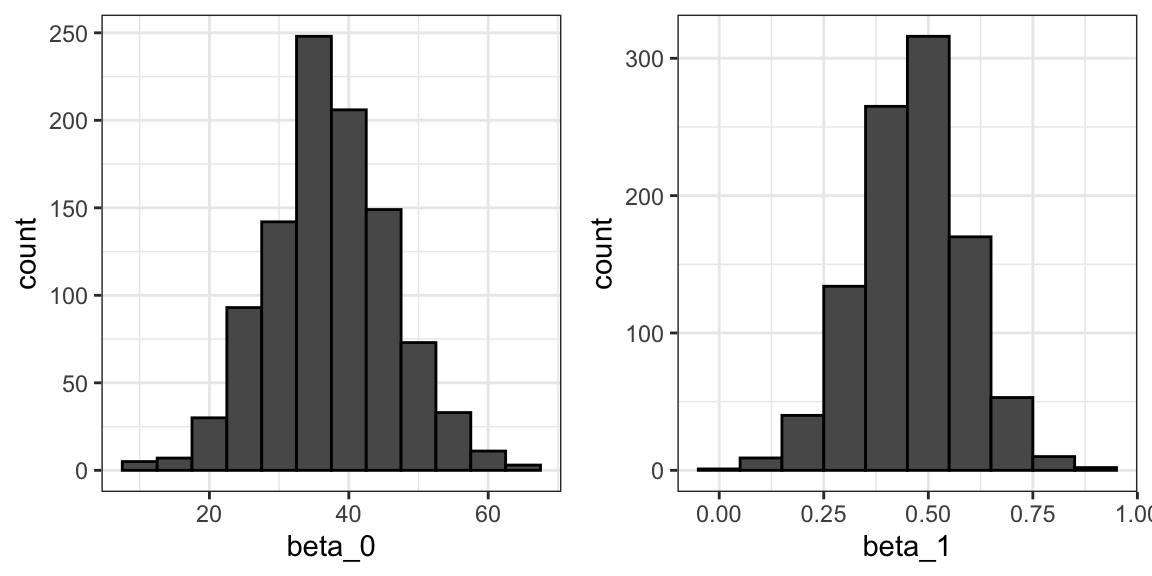
#>

#> Attaching package: 'gridExtra'

#> The following object is masked from 'package:dplyr':

#>

#>     combine



Lý do chúng trông bình thường là vì định lý giới hạn trung tâm cũng được áp dụng ở đây: đủ lớn , ước tính bình phương nhỏ nhất sẽ xấp xỉ bình thường với giá trị kỳ vọng và Tương ứng. Các lỗi tiêu chuẩn hơi phức tạp để tính toán, nhưng lý thuyết toán học cho phép chúng ta tính toán chúng và chúng được bao gồm trong bản tóm tắt được cung cấp bởi hàm. Đây là một trong những bộ dữ liệu mô phỏng của chúng tôi: lm

|  |
| --- |
| sample\_n(galton\_heights, N, replace = TRUE) |>    lm(son ~ father, data = \_) |>    summary() |>    coef()  #>             Estimate Std. Error t value Pr(>|t|)  #> (Intercept)    19.28     11.656    1.65 1.05e-01  #> father          0.72      0.169    4.25 9.79e-05 |

Bạn có thể thấy rằng các ước tính lỗi tiêu chuẩn được báo cáo bởi các lỗi gần với các lỗi tiêu chuẩn từ mô phỏng: summary

|  |
| --- |
| lse |> summarize(se\_0 = sd(beta\_0), se\_1 = sd(beta\_1))  #>   se\_0  se\_1  #> 1 8.84 0.128 |

Hàm này cũng báo cáo t-statistics () và p-values (). Thống kê t không thực sự dựa trên định lý giới hạn trung tâm mà dựa trên giả định rằng summaryt valuePr(>|t|) s tuân theo một phân phối bình thường. Theo giả định này, lý thuyết toán học cho chúng ta biết rằng LSE chia cho sai số chuẩn của chúng,  và , theo phân phối t với  mức độ tự do, với  số lượng tham số trong mô hình của chúng tôi. Trong trường hợp chiều cao , hai giá trị p đang kiểm tra giả thuyết null rằng và  Tương ứng.

Hãy nhớ rằng, như chúng tôi đã mô tả trong Phần [10.2.3](https://rafalab.dfci.harvard.edu/dsbook-part-2/inference/models.html#sec-t-dist) cho đủ lớn , CLT hoạt động và phân phối t trở nên gần giống như phân phối bình thường. Ngoài ra, lưu ý rằng chúng ta có thể xây dựng các khoảng tin cậy, nhưng chúng ta sẽ sớm tìm hiểu về **chổi**, một gói tiện ích bổ sung giúp việc này trở nên dễ dàng.

Mặc dù chúng tôi không đưa ra các ví dụ trong cuốn sách này, thử nghiệm giả thuyết với các mô hình hồi quy thường được sử dụng trong dịch tễ học và kinh tế học để đưa ra các tuyên bố như "ảnh hưởng của A đối với B có ý nghĩa thống kê sau khi điều chỉnh X, Y và Z". Tuy nhiên, một số giả định phải giữ cho những tuyên bố này là đúng.

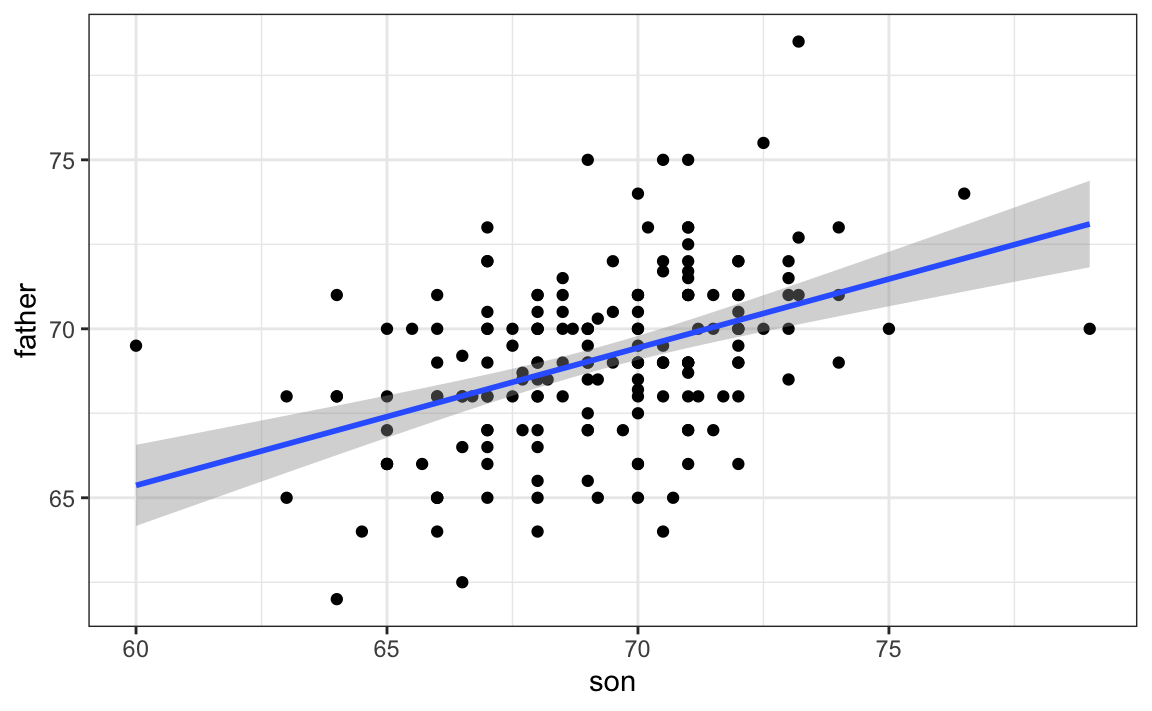
## 13.13 Giá trị dự đoán là các biến ngẫu nhiên

Khi chúng tôi phù hợp với mô hình của mình, chúng tôi có thể có được dự đoán về  bằng cách cắm các ước tính vào mô hình hồi quy. Ví dụ, nếu chiều cao của người cha là , sau đó là dự đoán của chúng tôi  Đối với chiều cao của con trai sẽ là:

Khi chúng tôi âm mưu đấu với , chúng ta thấy đường hồi quy.

Hãy nhớ rằng dự đoán  cũng là một biến ngẫu nhiên và lý thuyết toán học cho chúng ta biết các lỗi tiêu chuẩn là gì. Nếu chúng ta giả định các lỗi là bình thường hoặc có kích thước mẫu đủ lớn, chúng ta cũng có thể sử dụng lý thuyết để xây dựng các khoảng tin cậy. Trên thực tế, layer **ggplot2** mà trước đây chúng ta đã sử dụng plot geom\_smooth(method = "lm")và bao quanh nó bằng các khoảng tin cậy:

|  |
| --- |
| galton\_heights |> ggplot(aes(son, father)) +    geom\_point() +    geom\_smooth(method = "lm")  #> `geom\_smooth()` using formula = 'y ~ x' |



Hàm R lấy một đối tượng làm đầu vào và trả về dự đoán. Nếu được yêu cầu, các lỗi tiêu chuẩn và thông tin khác mà từ đó chúng tôi có thể xây dựng khoảng tin cậy được cung cấp: predictlm

|  |
| --- |
| fit <- galton\_heights |> lm(son ~ father, data = \_)  y\_hat <- predict(fit, se.fit = TRUE)  names(y\_hat)  #> [1] "fit"            "se.fit"         "df"             "residual.scale" |

## 13.14 Biểu đồ chẩn đoán

Khi mô hình tuyến tính được giả định thay vì suy ra, tất cả các cách giải thích phụ thuộc vào tính hữu ích của mô hình. Hàm sẽ phù hợp với mô hình và trả về tóm tắt ngay cả khi mô hình sai và không hữu ích.lm

Kiểm tra trực quan dư lượng, được định nghĩa là sự khác biệt giữa các giá trị quan sát được và các giá trị dự đoán

Và tóm tắt phần còn lại, là một cách mạnh mẽ để chẩn đoán xem mô hình có hữu ích hay không. Lưu ý rằng phần còn lại có thể được coi là ước tính các lỗi kể từ khi

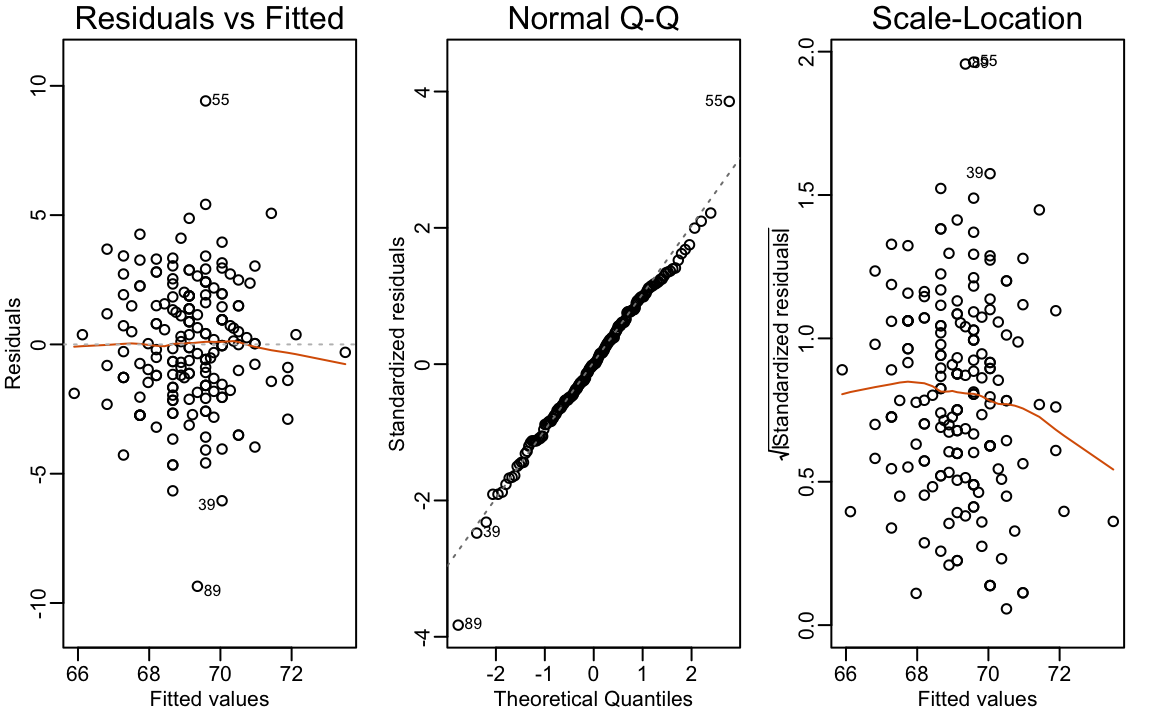
Trong thực tế, dư lượng thường được ký hiệu là . Điều này thúc đẩy một số âm mưu *chẩn đoán*. Becasue chúng tôi obervere,  Nhưng đừng quan sát , chúng tôi dựa trên các lô đất trên phần còn lại.

Bởi vì các lỗi được giả định không phụ thuộc vào giá trị kỳ vọng của , một cốt truyện của  so với các giá trị được trang bị  nên không thể hiện mối quan hệ.

Trong trường hợp chúng tôi giả định các lỗi tuân theo một phân phối bình thường, một qqplot được tiêu chuẩn hóa  nên rơi vào một đường thẳng khi vẽ đối với các lượng tử lý thuyết.

Bởi vì chúng ta giả sử độ lệch chuẩn của các lỗi là không đổi, nếu chúng ta vẽ giá trị tuyệt đối của phần dư, nó sẽ xuất hiện không đổi.

Chúng tôi thích các biểu đồ hơn là tóm tắt dựa trên, ví dụ, mối tương quan bởi vì, như đã lưu ý trong Phần @ascombe, mối tương quan không phải lúc nào cũng là bản tóm tắt tốt nhất về sự liên kết. Hàm được áp dụng cho một đối tượng sẽ tự động vẽ những thứ này . plotlm



|  |
| --- |
| plot(fit, which = 1:3) |

Hàm này có thể tạo ra sáu ô khác nhau và đối số cho phép bạn chỉ định ô nào bạn muốn xem. Bạn có thể tìm hiểu thêm bằng cách đọc tệp trợ giúp. Tuy nhiên, một số cốt truyện dựa trên các khái niệm nâng cao hơn ngoài phạm vi của cuốn sách này. Để tìm hiểu thêm, chúng tôi khuyên bạn nên đọc một cuốn sách nâng cao về phân tích hồi quy.whichplot.lm

Trong Chương Chương [14 và](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html) [Chương 16](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/treatment-effect-models.html), chúng tôi giới thiệu các thách thức phân tích dữ liệu trong đó có nhiều hơn một biến, một số không có trong mô hình. Trong những trường hợp này, một xét nghiệm chẩn đoán quan trọng để thêm kiểm tra xem phần dư có liên quan đến các biến không có trong mô hình hay không.

## 13.15 Ngụy biện hồi quy

Wikipedia định nghĩa sự *sụt giảm năm thứ hai* là:

Một sự sụt giảm năm thứ hai hoặc jinx năm thứ hai hoặc bồn chồn năm thứ hai đề cập đến một trường hợp trong đó một nỗ lực thứ hai, hoặc năm hai, không đáp ứng các tiêu chuẩn của nỗ lực đầu tiên. Nó thường được sử dụng để chỉ sự thờ ơ của học sinh (năm thứ hai trung học, cao đẳng hoặc đại học), hiệu suất của vận động viên (mùa thứ hai của vở kịch), ca sĩ / ban nhạc (album thứ hai), chương trình truyền hình (mùa thứ hai) và phim (phần tiếp theo / tiền truyện).

Trong Major League Baseball, giải thưởng tân binh của năm (ROY) được trao cho cầu thủ năm nhất được đánh giá là có thành tích tốt nhất. Cụm từ *slump sophmore* được sử dụng để mô tả quan sát rằng những người chiến thắng giải thưởng ROY không làm tốt trong năm thứ hai của họ. Ví dụ, bài viết này của Fox Sports[2](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/regression.html#fn2) "Liệu lớp tân binh khủng khiếp của MLB năm 2015 có bị sa sút năm thứ hai?".

Dữ liệu có xác nhận sự tồn tại của sự sụt giảm năm thứ hai không? Chúng ta hãy xem. Kiểm tra dữ liệu để đo lường thành công được sử dụng rộng rãi, mức trung bình đánh bóng, chúng tôi thấy rằng quan sát này đúng với các Roy hoạt động hàng đầu:

| **nameFirst** | **tênCuối cùng** | **rookie\_year** | **Rc** | **Thứ hai** |
| --- | --- | --- | --- | --- |
| Willie | McCovey | 1959 | 0.354 | 0.238 |
| Ichiro | Suzuki | 2001 | 0.350 | 0.321 |
| Al | Bùm bùm | 1973 | 0.337 | 0.233 |
| Fred | Lynn | 1975 | 0.331 | 0.314 |
| Albert | Phốc phốc phốc | 2001 | 0.329 | 0.314 |

Trên thực tế, tỷ lệ người chơi có điểm trung bình đánh bóng thấp hơn trong năm thứ hai của họ là 0.6981132.

Vậy đó là "jitters" hay "jinx"? Để trả lời câu hỏi này, chúng ta hãy chuyển sự chú ý của chúng tôi đến tất cả những người chơi đã chơi mùa giải 2013 và 2014 và đánh bóng hơn 130 lần (tối thiểu để giành giải Tân binh của năm).

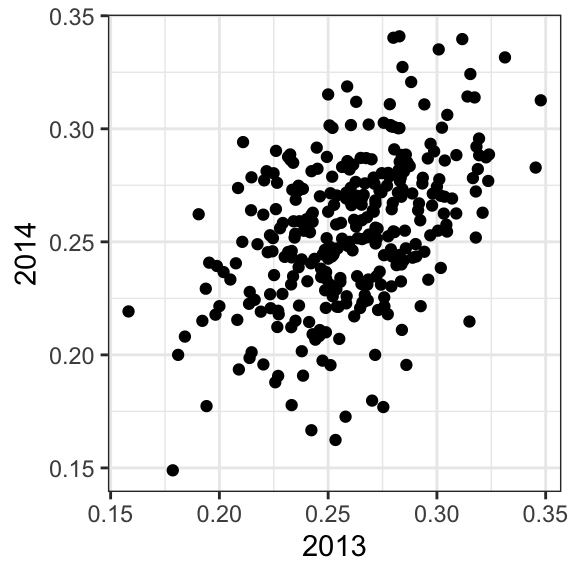
Mô hình tương tự cũng phát sinh khi chúng ta nhìn vào những người biểu diễn hàng đầu: trung bình đánh bóng giảm đối với hầu hết những người biểu diễn hàng đầu.

| **nameFirst** | **tênCuối cùng** | **2013** | **2014** |
| --- | --- | --- | --- |
| Miguel | Cabrera | 0.348 | 0.313 |
| Hanley | Ramirez | 0.345 | 0.283 |
| Michael | Cuddyer | 0.331 | 0.332 |
| Xe xcutơ | Gennett | 0.324 | 0.289 |
| Joe | Mauer | 0.324 | 0.277 |

Nhưng đây không phải là tân binh! Ngoài ra, hãy nhìn vào những gì xảy ra với những người hoạt động kém nhất năm 2013:

| **nameFirst** | **tênCuối cùng** | **2013** | **2014** |
| --- | --- | --- | --- |
| Danny | Espinosa | 0.158 | 0.219 |
| Dan | Uggla | 0.179 | 0.149 |
| Jeff | Mathis | 0.181 | 0.200 |
| B. J. | Upton | 0.184 | 0.208 |
| Adam | Rosales | 0.190 | 0.262 |

Trung bình đánh bóng của họ chủ yếu tăng lên! Đây có phải là một loại sụt giảm năm thứ hai ngược? Nó không phải là. Không có cái gọi là sự sụt giảm năm thứ hai. Tất cả điều này được giải thích với một thực tế thống kê đơn giản: mối tương quan về hiệu suất trong hai năm riêng biệt là cao, nhưng không hoàn hảo:



Mối tương quan là 0,460254 và dữ liệu trông rất giống phân phối chuẩn hai biến, có nghĩa là chúng tôi dự đoán mức trung bình đánh bóng năm 2014  cho bất kỳ cầu thủ nào có điểm trung bình đánh bóng năm 2013  với:

Bởi vì mối tương quan không hoàn hảo, hồi quy cho chúng ta biết rằng, trung bình, mong đợi những người có hiệu suất cao từ năm 2013 sẽ làm tồi tệ hơn một chút trong năm 2014. Nó không phải là một jinx; Đó chỉ là do tình cờ. ROY được chọn từ các giá trị hàng đầu của  Vì vậy, dự kiến rằng  sẽ thoái lui về giá trị trung bình.

## 13.16 Bài tập

1. Tải dữ liệu từ **HistData**. Những đứa trẻ trong mỗi gia đình được liệt kê theo giới tính và sau đó theo chiều cao. Tạo một tập dữ liệu được gọi bằng cách chọn ngẫu nhiên một nam và nữ. GaltonFamiliesgalton\_heights

2. Lập biểu đồ phân tán độ cao giữa mẹ và con gái, mẹ và con trai, cha và con gái, cha và con trai.

3. Tính toán mối tương quan về chiều cao giữa mẹ và con gái, mẹ và con trai, cha và con gái, cha và con trai.

# 14 Hồi quy đa biến

Kể từ sự phát triển ban đầu của Galton, hồi quy đã trở thành một trong những công cụ được sử dụng rộng rãi nhất trong phân tích dữ liệu. Một lý do liên quan đến thực tế là sự thích ứng của phương pháp hồi quy ban đầu, dựa trên các mô hình tuyến tính, cho phép chúng ta tìm mối quan hệ giữa hai biến có tính đến tác động của các biến khác ảnh hưởng đến cả hai. Điều này đặc biệt phổ biến trong các lĩnh vực mà các thí nghiệm ngẫu nhiên khó thực hiện, chẳng hạn như kinh tế và dịch tễ học.

Khi chúng ta không thể chỉ định ngẫu nhiên mỗi cá nhân vào một nhóm điều trị hoặc kiểm soát, sự nhầm lẫn đặc biệt phổ biến. Ví dụ: xem xét ước tính ảnh hưởng của việc ăn thức ăn nhanh đối với tuổi thọ bằng cách sử dụng dữ liệu được thu thập từ một mẫu ngẫu nhiên của những người trong khu vực pháp lý. Người tiêu dùng thức ăn nhanh có nhiều khả năng là người hút thuốc, uống rượu và có thu nhập thấp hơn. Do đó, một mô hình hồi quy ngây thơ có thể dẫn đến việc đánh giá quá cao tác động tiêu cực đến sức khỏe của thức ăn nhanh. Vậy làm thế nào để chúng ta giải thích cho sự nhầm lẫn trong thực tế? Trong chương này, chúng ta tìm hiểu làm thế nào *hồi quy đa* biến có thể giúp với các tình huống như vậy và có thể được sử dụng để mô tả cách một hoặc nhiều biến ảnh hưởng đến một biến kết quả. Chúng tôi minh họa bằng một ví dụ thực tế, trong đó dữ liệu được sử dụng để giúp chọn những người chơi bị đánh giá thấp để cải thiện một đội thể thao hạn chế về tài nguyên.

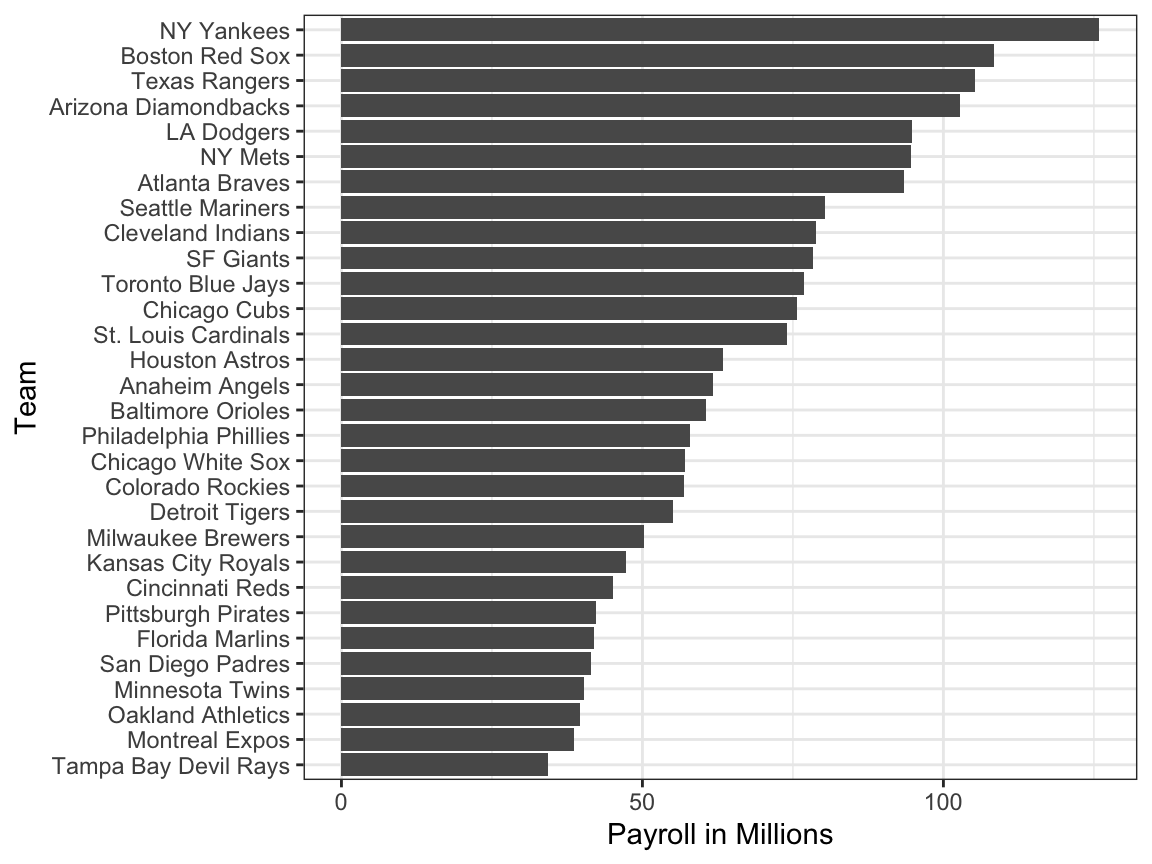
## 14.1 Nghiên cứu điển hình: Moneyball

*Moneyball: The Art of Winning an Unfair Game* là một cuốn sách của Michael Lewis về đội bóng chày Oakland Athletics (A's) và tổng giám đốc của nó, người được giao nhiệm vụ xây dựng đội, Billy Beane.

Theo truyền thống, các đội bóng chày sử dụng *các tuyển trạch viên* để giúp họ quyết định thuê cầu thủ nào. Những tuyển trạch viên này đánh giá người chơi bằng cách quan sát họ thực hiện. Các tuyển trạch viên có xu hướng ưu tiên các cầu thủ thể thao có khả năng thể chất quan sát được. Vì lý do này, các tuyển trạch viên có xu hướng đồng ý ai là cầu thủ giỏi nhất và kết quả là, những cầu thủ này có xu hướng có nhu cầu cao. Điều này lần lượt làm tăng lương của họ.

Từ năm 1989 đến năm 1991, A có một trong những bảng lương cao nhất trong bóng chày. Họ đã có thể mua những cầu thủ giỏi nhất và trong thời gian đó, họ là một trong những đội bóng tốt nhất. Tuy nhiên, vào năm 1995, chủ sở hữu đội A đã thay đổi và ban quản lý mới cắt giảm ngân sách đáng kể, khiến tổng giám đốc lúc đó, Sandy Alderson, có một trong những biên chế thấp nhất trong bóng chày. Ông không còn đủ khả năng mua những cầu thủ được săn đón nhất. Alderson bắt đầu sử dụng một phương pháp thống kê để tìm ra sự thiếu hiệu quả trên thị trường. Alderson là cố vấn cho Billy Beane, người kế nhiệm ông vào năm 1998 và hoàn toàn chấp nhận khoa học dữ liệu, trái ngược với các tuyển trạch viên, như một phương pháp để tìm kiếm những cầu thủ chi phí thấp mà dữ liệu dự đoán sẽ giúp đội giành chiến thắng. Ngày nay, chiến lược này đã được điều chỉnh bởi hầu hết các đội bóng chày. Như chúng ta sẽ thấy, hồi quy đóng một vai trò lớn trong phương pháp này.

Để tạo động lực cho phần này của cuốn sách, chúng tôi sẽ giả vờ đó là năm 2002 và cố gắng xây dựng một đội bóng chày với ngân sách hạn chế, giống như A phải làm. Để đánh giá cao những gì bạn đang chống lại, hãy lưu ý rằng vào năm 2002, bảng lương của Yankees là $ 125,928,583 nhiều hơn gấp ba lần $ 39,679,746 của Oakland A:



Thống kê đã được sử dụng trong bóng chày kể từ khi bắt đầu. Bộ dữ liệu chúng tôi sẽ sử dụng, bao gồm trong thư viện **Lahman**, có từ thế kỷ 19. Ví dụ, một số liệu thống kê tóm tắt mà chúng tôi sẽ mô tả sớm, *mức trung bình đánh bóng*, đã được sử dụng trong nhiều thập kỷ để tóm tắt thành công của người đánh bóng. Thống kê khác[1](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn1) chẳng hạn như chạy trên sân nhà (HR), chạy đánh bóng (RBI) và căn cứ bị đánh cắp (SB) được báo cáo cho mỗi người chơi trong phần tóm tắt trò chơi có trong phần thể thao của báo, với người chơi được thưởng cho số lượng cao. Mặc dù số liệu thống kê tóm tắt như thế này đã được sử dụng rộng rãi trong bóng chày, nhưng bản thân phân tích dữ liệu thì không. Những thống kê này được quyết định một cách tùy tiện mà không cần suy nghĩ nhiều về việc liệu chúng có thực sự dự đoán bất cứ điều gì hoặc có liên quan đến việc giúp một đội giành chiến thắng hay không.

Điều này đã thay đổi với Bill James[2](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn2). Vào cuối những năm 1970, nhà văn và người hâm mộ bóng chày đầy tham vọng này bắt đầu xuất bản các bài báo mô tả phân tích sâu hơn về dữ liệu bóng chày. Ông đặt tên cho cách tiếp cận sử dụng dữ liệu để dự đoán kết quả nào được dự đoán tốt nhất nếu một đội giành chiến thắng trong *lĩnh vực đo lường*[3](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn3). Cho đến khi Billy Beane biến sabermetrics thành trung tâm hoạt động bóng chày của mình, công việc của Bill James hầu như bị thế giới bóng chày bỏ qua. Hiện tại, sự phổ biến của sabermetrics không còn chỉ giới hạn ở bóng chày; Các môn thể thao khác cũng đã bắt đầu sử dụng phương pháp này.

Để đơn giản hóa bài tập, chúng tôi sẽ tập trung vào việc ghi điểm và bỏ qua hai khía cạnh quan trọng khác của trò chơi: ném bóng và ném bóng. Chúng ta sẽ thấy phân tích hồi quy có thể giúp phát triển các chiến lược để xây dựng một đội bóng chày cạnh tranh với ngân sách hạn chế như thế nào. Cách tiếp cận có thể được chia thành hai phân tích dữ liệu riêng biệt. Trong lần đầu tiên, chúng tôi xác định số liệu thống kê cụ thể của người chơi được ghi lại dự đoán số lần chạy. Trong lần thứ hai, chúng tôi kiểm tra xem người chơi có bị định giá thấp hay không dựa trên những gì phân tích đầu tiên của chúng tôi dự đoán.

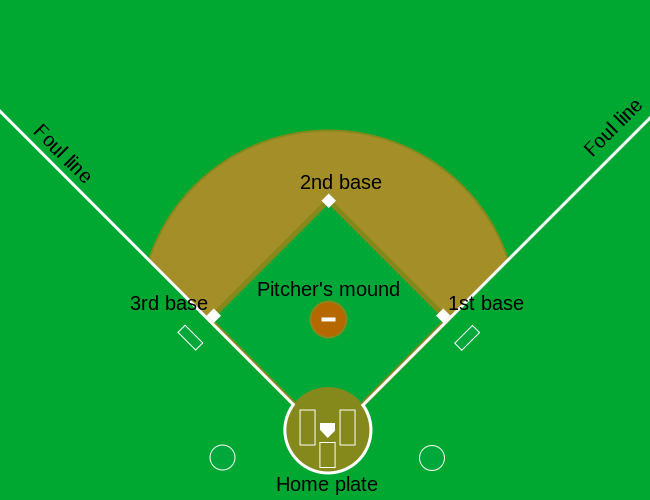
### 14.1.1 Khái niệm cơ bản về bóng chày

Để xem hồi quy sẽ giúp chúng ta tìm thấy những cầu thủ bị đánh giá thấp như thế nào, chúng ta thực sự không cần phải hiểu tất cả các chi tiết về trò chơi bóng chày, có hơn 100 quy tắc. Ở đây, chúng tôi chắt lọc môn thể thao này thành kiến thức cơ bản mà người ta cần để biết cách tấn công hiệu quả vấn đề khoa học dữ liệu.

Mục tiêu của một trận đấu bóng chày là ghi được nhiều lần chạy (điểm) hơn đội khác. Mỗi đội có 9 người đánh bóng có cơ hội đánh bóng bằng gậy theo thứ tự định trước. Sau khi người đánh bóng thứ 9 đã đến lượt, người đánh bóng đầu tiên lại đánh bóng, sau đó là người đánh thứ hai, v.v. Mỗi khi một người đánh bóng có cơ hội đánh bóng, chúng tôi gọi đó là sự xuất hiện của đĩa (PA). Tại mỗi PA, *người ném* bóng của đội khác ném bóng và người đánh bóng cố gắng đánh nó. PA kết thúc với một kết quả nhị phân: người đánh bóng hoặc ra *ngoài* (thất bại) và trở lại băng ghế dự bị hoặc người đánh bóng không (thành công) và có thể chạy xung quanh các căn cứ, và có khả năng ghi điểm chạy (đạt được cả 4 cơ sở). Mỗi đội có chín lần thử, được gọi là *hiệp*, để ghi điểm chạy và mỗi hiệp kết thúc sau ba lần ra ngoài (ba thất bại).

Dưới đây là video cho thấy sự thành công: <https://www.youtube.com/watch?v=HL-XjMCPfio>. Và đây là một trong những cho thấy một thất bại: <https://www.youtube.com/watch?v=NeloljCx-1g>. Trong những video này, chúng ta thấy may mắn tham gia vào quá trình này như thế nào. Khi đánh bóng, người đánh bóng muốn đánh bóng mạnh. Nếu người đánh bóng đánh đủ mạnh, đó là HR, kết quả tốt nhất có thể vì người đánh bóng có ít nhất một lần chạy tự động. Nhưng đôi khi, do cơ hội, người đánh bóng đánh bóng rất mạnh và một hậu vệ bắt được nó, dẫn đến một out. Ngược lại, đôi khi người đánh bóng chạm bóng nhẹ nhàng, nhưng nó hạ cánh đúng chỗ. Thực tế là có cơ hội liên quan gợi ý về lý do tại sao các mô hình xác suất sẽ tham gia.

Bây giờ, có một số cách để thành công. Hiểu được sự khác biệt này sẽ rất quan trọng đối với phân tích của chúng tôi. Khi người đánh bóng chạm bóng, người đánh bóng muốn vượt qua càng nhiều *cơ sở* càng tốt. Có bốn cơ sở với cơ sở thứ tư được gọi là *tấm nhà*. Tấm nhà là nơi người đánh bắt đầu bằng cách cố gắng đánh, vì vậy các cơ sở tạo thành một chu kỳ.



(Ảnh: Cburnett[4](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn4). Giấy phép CC BY-SA 3.0[5](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn5).)

Một người đánh bóng *đi vòng quanh các căn cứ* và về đến nhà, ghi được một cuộc chạy.

Chúng tôi đang đơn giản hóa một chút, nhưng có năm cách để một người đánh bóng có thể thành công, đó là, không tạo ra:

Căn cứ trên quả bóng (BB) - người ném bóng không ném bóng qua một khu vực được xác định trước được coi là có thể đánh được (khu vực tấn công), vì vậy người đánh bóng được phép đi đến căn cứ đầu tiên.

Đơn - Batter đánh bóng và đến căn cứ đầu tiên.

Đôi (2B) - Batter đánh bóng và đến căn cứ thứ hai.

Triple (3B) - Batter đánh bóng và đến căn cứ thứ ba.

Home Run (HR) - Batter đánh bóng và đi về nhà và ghi bàn chạy.

Dưới đây là một ví dụ về nhân sự: <https://www.youtube.com/watch?v=xYxSZJ9GZ-w>. Nếu một người đánh bóng đến một căn cứ, người đánh bóng vẫn có cơ hội về nhà và ghi điểm chạy nếu người đánh bóng tiếp theo đánh thành công. Trong khi bột ở trên cơ sở, người đánh bóng cũng có thể cố gắng ăn cắp một *cơ sở* (SB). Nếu một người đánh bóng chạy đủ nhanh, người đánh bóng có thể cố gắng đi từ cơ sở này sang cơ sở tiếp theo mà không cần đội khác gắn thẻ người chạy. Dưới đây là một ví dụ về một cơ sở bị đánh cắp: <https://www.youtube.com/watch?v=JSE5kfxkzfk>.

Tất cả những sự kiện này được theo dõi trong mùa giải và có sẵn cho chúng tôi thông qua gói **Lahman**. Bây giờ chúng ta sẽ bắt đầu thảo luận về cách phân tích dữ liệu có thể giúp chúng ta quyết định cách sử dụng các số liệu thống kê này để đánh giá người chơi.

### 14.1.2 Không có giải thưởng cho BB

Trong lịch sử, *trung bình đánh bóng* được coi là thống kê tấn công quan trọng nhất. Để xác định mức trung bình này, chúng tôi xác định một cú *đánh* (H) và một *cú đánh* (AB). Đơn, đôi, ba và chạy trên sân nhà là những cú đánh. Cách thứ năm để thành công, BB, không phải là một cú đánh. AB là số lần bạn nhận được một hit hoặc thực hiện một out; BB bị loại trừ. Trung bình đánh bóng chỉ đơn giản là H / AB và được coi là thước đo chính của tỷ lệ thành công. Ngày nay, tỷ lệ thành công này dao động từ 20% đến 38%. Chúng tôi đề cập đến mức trung bình đánh bóng tính bằng hàng nghìn, ví dụ: nếu tỷ lệ thành công của bạn là 28%, chúng tôi gọi đó là *đánh bóng 280*.



(Ảnh: Keith Allison[6](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn6). Giấy phép CC BY-SA 2.0[7](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn7).)

Một trong những hiểu biết quan trọng đầu tiên của Bill James là mức trung bình đánh bóng bỏ qua BB, nhưng BB là một thành công. Ông đề xuất chúng tôi sử dụng *tỷ lệ phần trăm cơ bản* (OBP) thay vì đánh bóng trung bình. Ông định nghĩa OBP là (H + BB) / (AB + BB) chỉ đơn giản là tỷ lệ xuất hiện của tấm không dẫn đến ra ngoài, một thước đo rất trực quan. Ông lưu ý rằng một người chơi nhận được nhiều BB hơn người chơi trung bình có thể không được công nhận nếu người đánh bóng không xuất sắc trong việc đánh bóng trung bình. Nhưng cầu thủ này không giúp tạo ra những bước chạy? Không có giải thưởng nào được trao cho người chơi có nhiều BB nhất. Tuy nhiên, những thói quen xấu rất khó phá vỡ và bóng chày đã không ngay lập tức áp dụng OBP như một thống kê quan trọng. Ngược lại, tổng số căn cứ bị đánh cắp được coi là quan trọng và là một giải thưởng[8](https://rafalab.dfci.harvard.edu/dsbook-part-2/linear-models/multivariate-regression.html#fn8) được trao cho người chơi có nhiều nhất. Nhưng những người chơi có tổng SB cao cũng kiếm được nhiều tiền hơn vì không phải lúc nào họ cũng thành công. Người chơi có tổng SB cao có giúp tạo ra các lần chạy không? Chúng ta có thể sử dụng khoa học dữ liệu để xác định xem tốt hơn là trả tiền cho những người chơi có BB hoặc SB cao không?

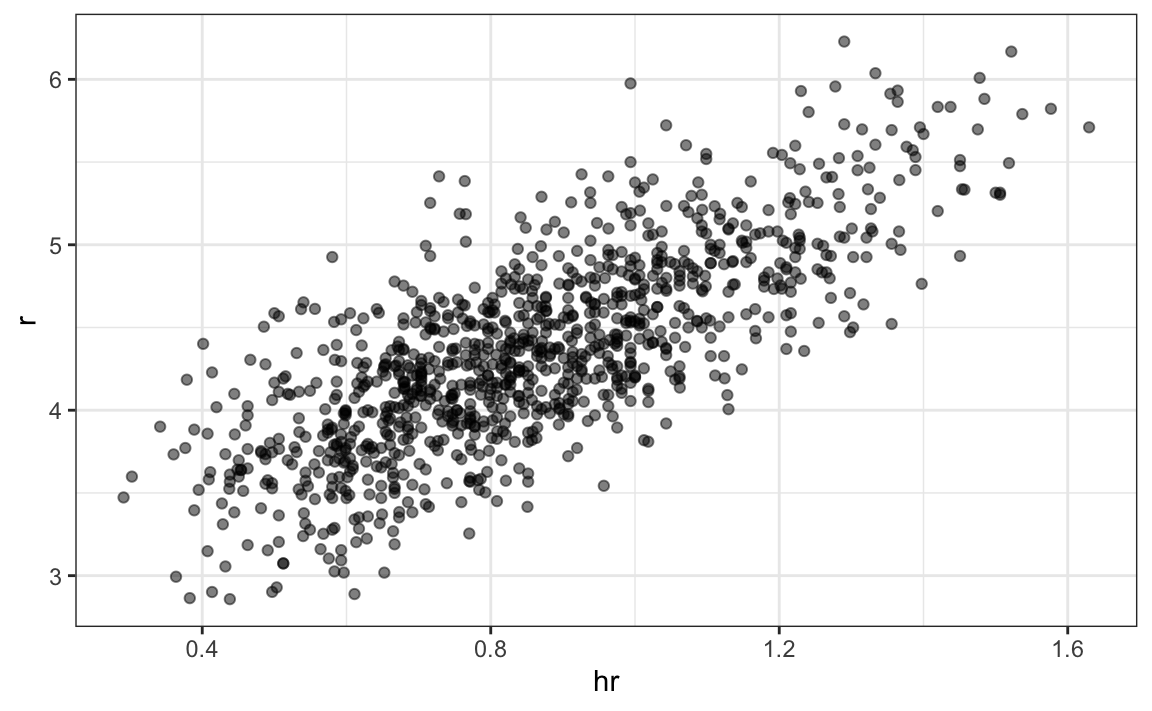
### 14.1.3 Căn cứ vào quả bóng hoặc căn cứ bị đánh cắp?

Một trong những thách thức trong phân tích này là không rõ làm thế nào để xác định xem một cầu thủ có tạo ra những bước chạy hay không vì quá nhiều phụ thuộc vào đồng đội của anh ta. Chúng tôi theo dõi số lần chạy được ghi bởi một cầu thủ. Tuy nhiên, hãy nhớ rằng nếu một người chơi X đánh ngay trước một người đánh nhiều HR, người đánh bóng X sẽ ghi được nhiều lần chạy. Nhưng những lần chạy này không nhất thiết phải xảy ra nếu chúng ta thuê cầu thủ X chứ không phải đồng đội đánh nhân sự của anh ta. Tuy nhiên, chúng tôi có thể kiểm tra số liệu thống kê cấp đội. Làm thế nào để các đội có nhiều SB so sánh với các đội có ít? Còn BB thì sao? Chúng tôi có dữ liệu! Hãy kiểm tra một số. Chúng tôi bắt đầu bằng cách tạo ra với số liệu thống kê từ năm 1962, năm đầu tiên tất cả các đội chơi 162 trận (như ngày nay) thay vì 154, đến năm 2001, năm trước năm mà chúng tôi sẽ xây dựng một đội. Chúng tôi chuyển đổi dữ liệu thành tỷ lệ mỗi trận đấu vì một tỷ lệ nhỏ các mùa giải có ít trận đấu hơn bình thường do đình công và một số đội đã chơi thêm các *trận* đấu do hòa break.

|  |
| --- |
| library(tidyverse)  #> ── Attaching core tidyverse packages ──────────────── tidyverse 2.0.0 ──  #> ✔ dplyr     1.1.1     ✔ readr     2.1.4  #> ✔ forcats   1.0.0     ✔ stringr   1.5.0  #> ✔ lubridate 1.9.2     ✔ tibble    3.2.1  #> ✔ purrr     1.0.1     ✔ tidyr     1.3.0  #> ── Conflicts ────────────────────────────────── tidyverse\_conflicts() ──  #> ✖ dplyr::filter() masks stats::filter()  #> ✖ dplyr::lag()    masks stats::lag()  #> ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors  library(Lahman)  dat <- Teams |> filter(yearID %in% 1962:2002) |>    mutate(team = teamID, year = yearID, r = R/G,           singles = (H - X2B - X3B - HR)/G, doubles = X2B/G, triples = X3B/G, hr = HR/G,           sb = SB/G, bb = BB/G) |>    select(team, year, r, singles, doubles, triples, hr, sb, bb) |

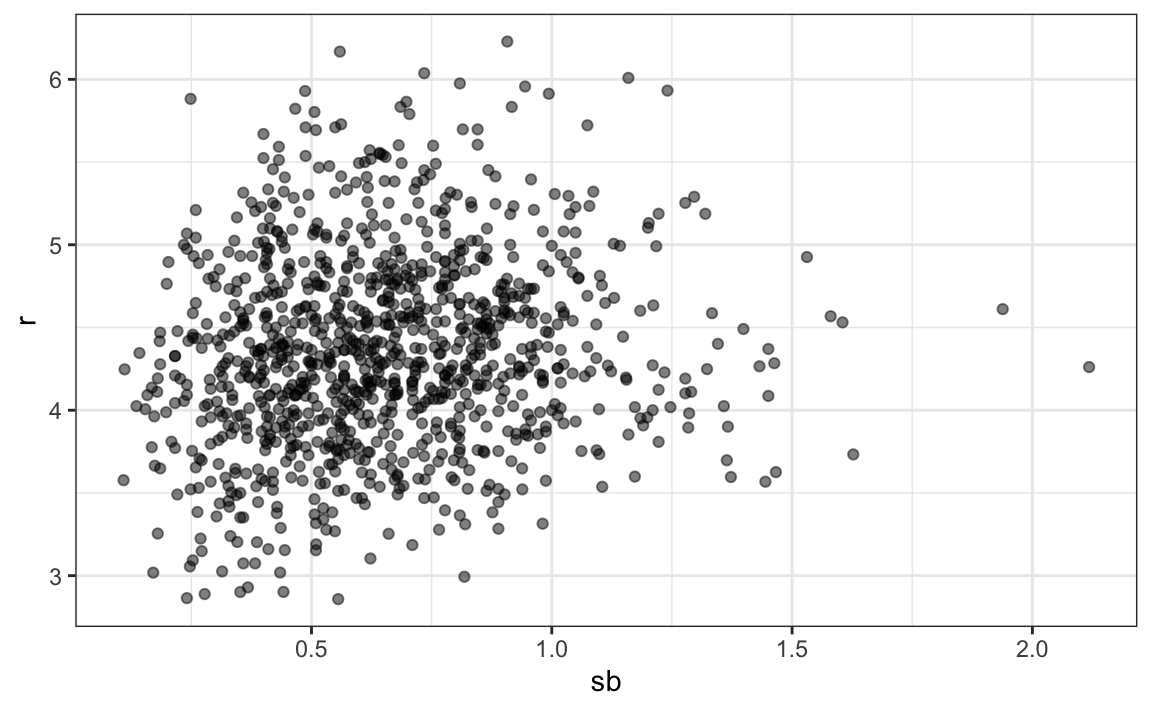
Bây giờ hãy bắt đầu với một câu hỏi rõ ràng: các đội nào đạt được nhiều lần chạy trên sân nhà hơn có ghi được nhiều lần chạy hơn không? Hình dung của sự lựa chọn khi khám phá mối quan hệ giữa hai biến là một biểu đồ phân tán.

|  |
| --- |
| p <- dat |> ggplot(aes(hr, r)) + geom\_point(alpha = 0.5)  p |



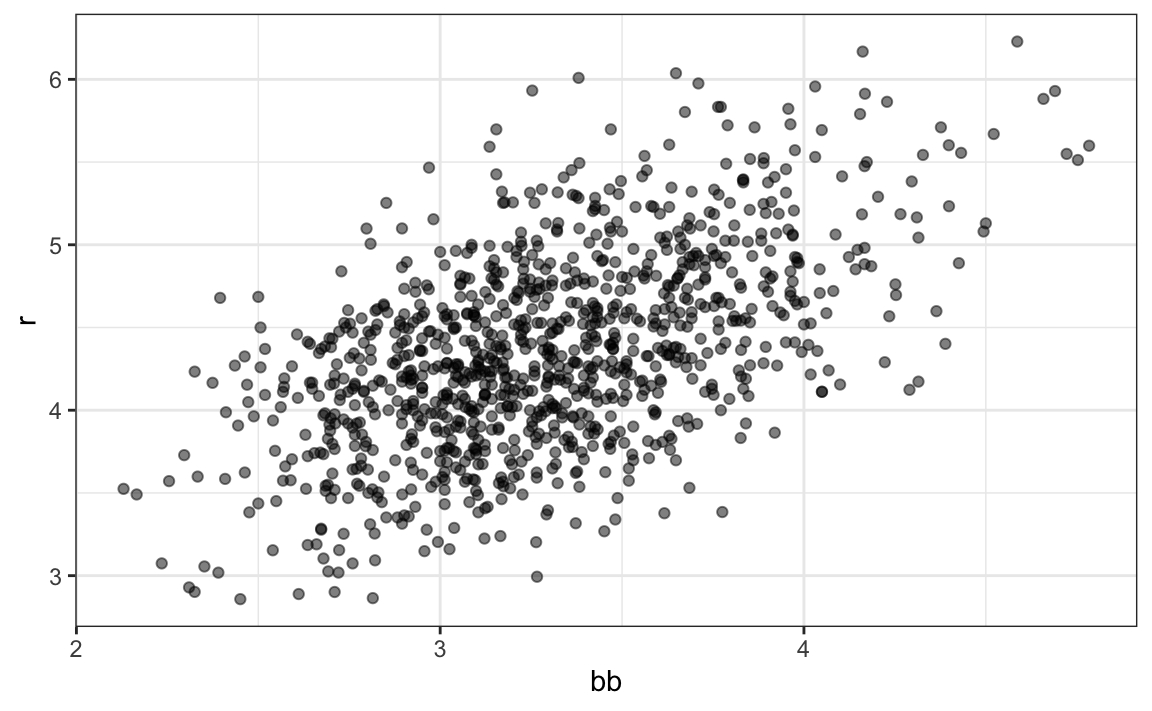
Chúng tôi xác định bởi vì chúng tôi sẽ thêm vào cốt truyện này sau. Cốt truyện cho thấy một mối liên hệ mạnh mẽ: các đội có nhiều nhân sự hơn có xu hướng ghi được nhiều lần chạy hơn. Bây giờ chúng ta hãy kiểm tra mối quan hệ giữa các căn cứ bị đánh cắp và chạy: p

|  |
| --- |
| dat |> ggplot(aes(sb, r)) + geom\_point(alpha = 0.5) |



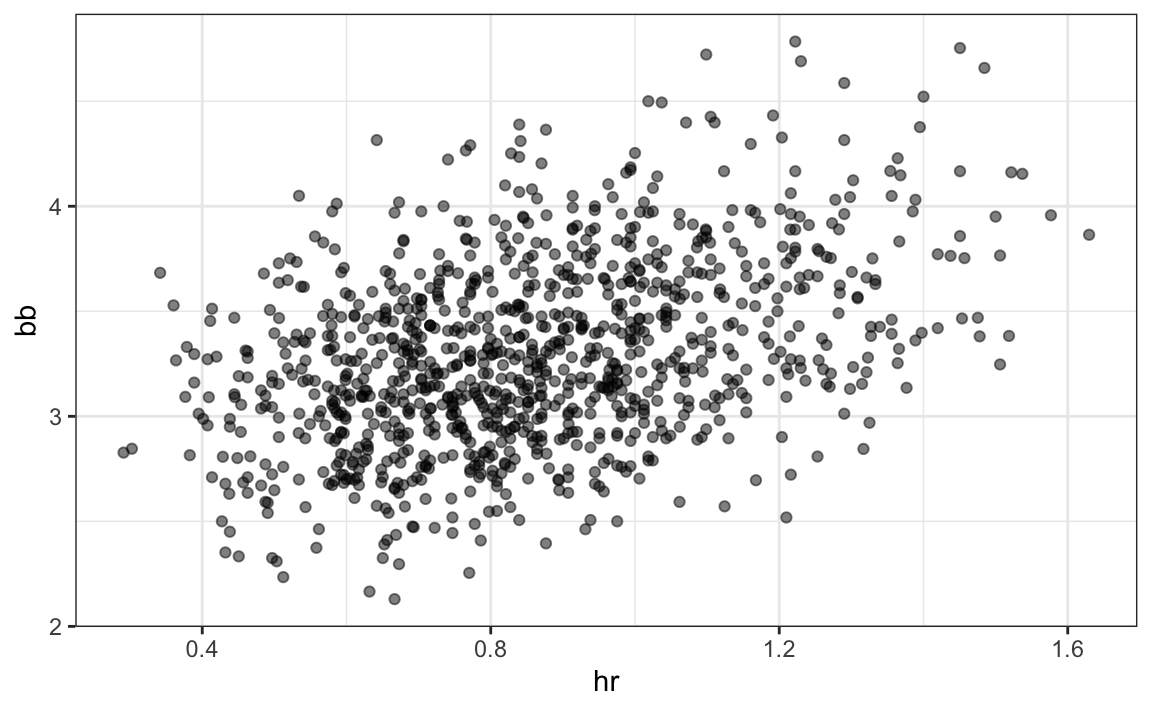
Ở đây mối quan hệ không rõ ràng. Cuối cùng, hãy xem xét mối quan hệ giữa BB và runs:

|  |
| --- |
| dat |> ggplot(aes(bb, r)) + geom\_point(alpha = 0.5) |



Ở đây một lần nữa chúng ta thấy một hiệp hội rõ ràng. Nhưng điều này có nghĩa là việc tăng BB của một đội **gây ra** sự gia tăng số lần chạy? Một trong những bài học quan trọng nhất bạn học được trong cuốn sách này là **sự liên kết không phải là quan hệ nhân quả.** Trên thực tế, có vẻ như BB và HR cũng được liên kết:

|  |
| --- |
| dat |> ggplot(aes(hr, bb)) + geom\_point(alpha = 0.5) |



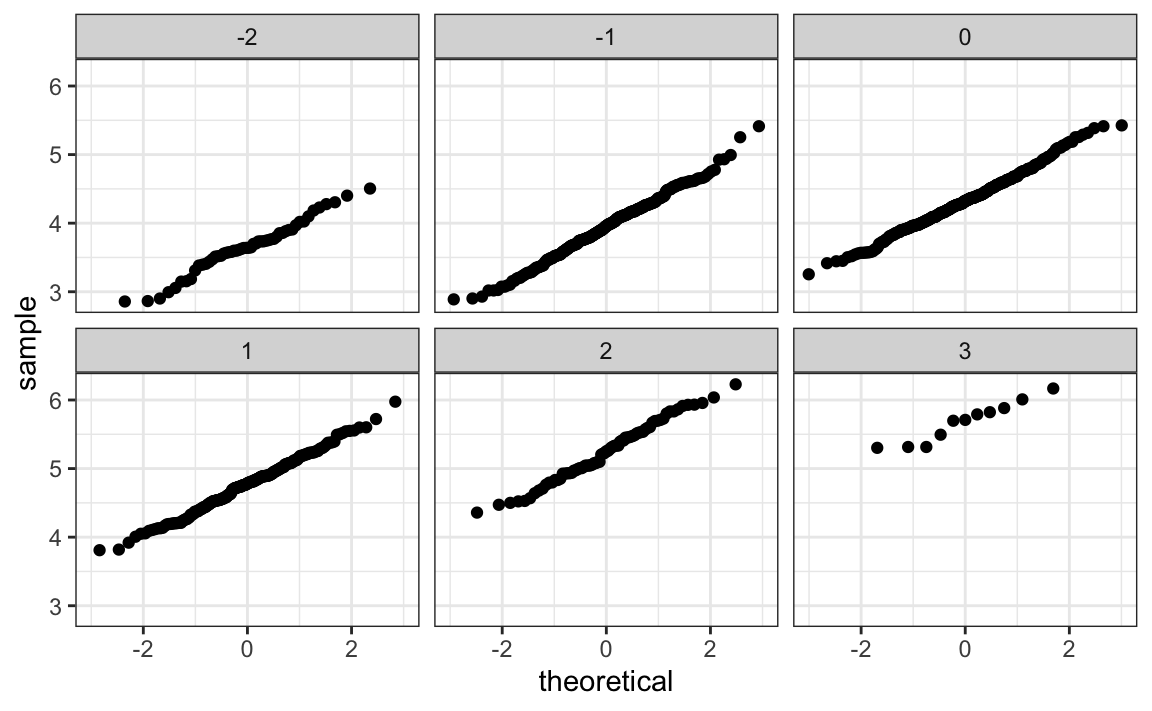
Chúng tôi biết rằng nhân sự gây ra các lần chạy bởi vì khi một người chơi đánh vào nhân sự, họ được đảm bảo ít nhất một lần chạy. Có thể là HR cũng gây ra BB và điều này làm cho nó xuất hiện như thể BB gây ra chạy? Khi điều này xảy ra, chúng tôi nói rằng có *sự nhầm lẫn*, một khái niệm quan trọng mà chúng tôi sẽ tìm hiểu thêm trong suốt chương này.

Hồi quy tuyến tính sẽ giúp chúng ta phân tích tất cả những điều này và định lượng các hiệp hội. Điều này sau đó sẽ giúp chúng tôi xác định những cầu thủ cần chiêu mộ. Cụ thể, chúng tôi sẽ cố gắng dự đoán những thứ như một đội sẽ ghi thêm bao nhiêu lần chạy nữa nếu chúng tôi tăng số lượng BB, nhưng giữ cho nhân sự cố định? Hồi quy sẽ giúp chúng ta trả lời những câu hỏi như thế này.

### 14.1.4 Hồi quy áp dụng cho thống kê bóng chày

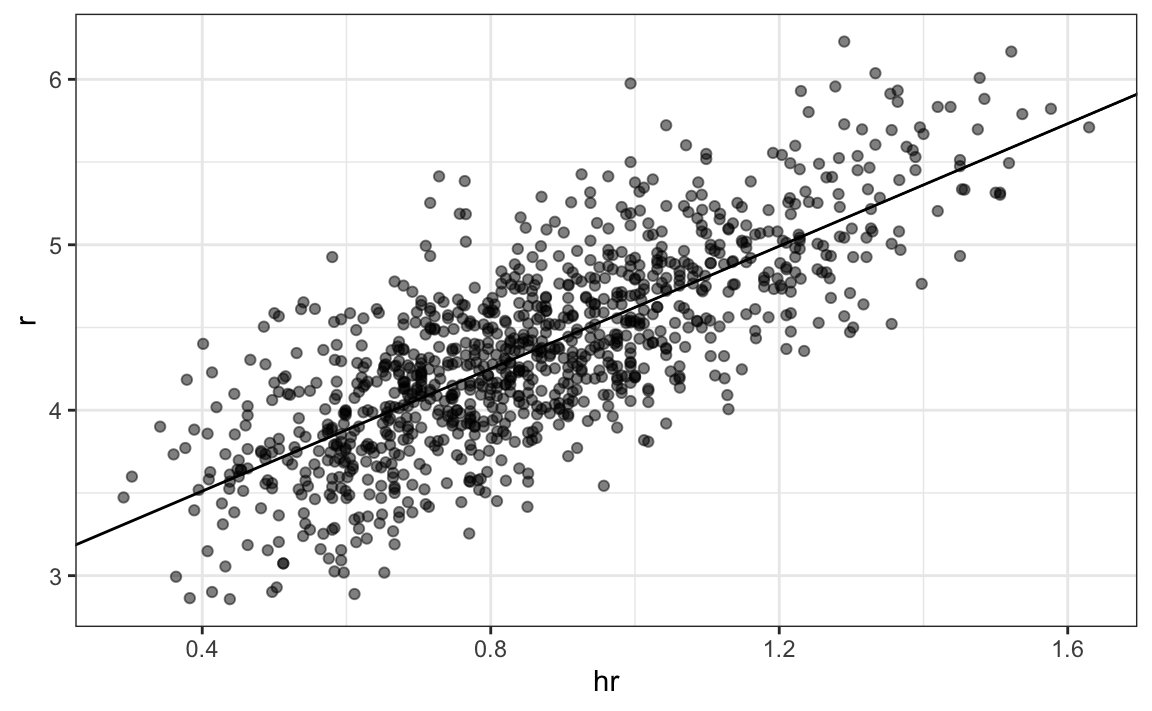
Chúng ta có thể sử dụng hồi quy với những dữ liệu này không? Đầu tiên, lưu ý rằng dữ liệu HR và Run, được hiển thị ở trên, dường như là bình thường hai biến. Cụ thể, các biểu đồ qq xác nhận rằng xấp xỉ bình thường cho mỗi tầng nhân sự là hữu ích ở đây:

|  |
| --- |
| dat |> mutate(z\_hr = round(scale(hr))) |>    filter(z\_hr %in% -2:3) |>    ggplot() +    stat\_qq(aes(sample = r)) +    facet\_wrap(~z\_hr) |



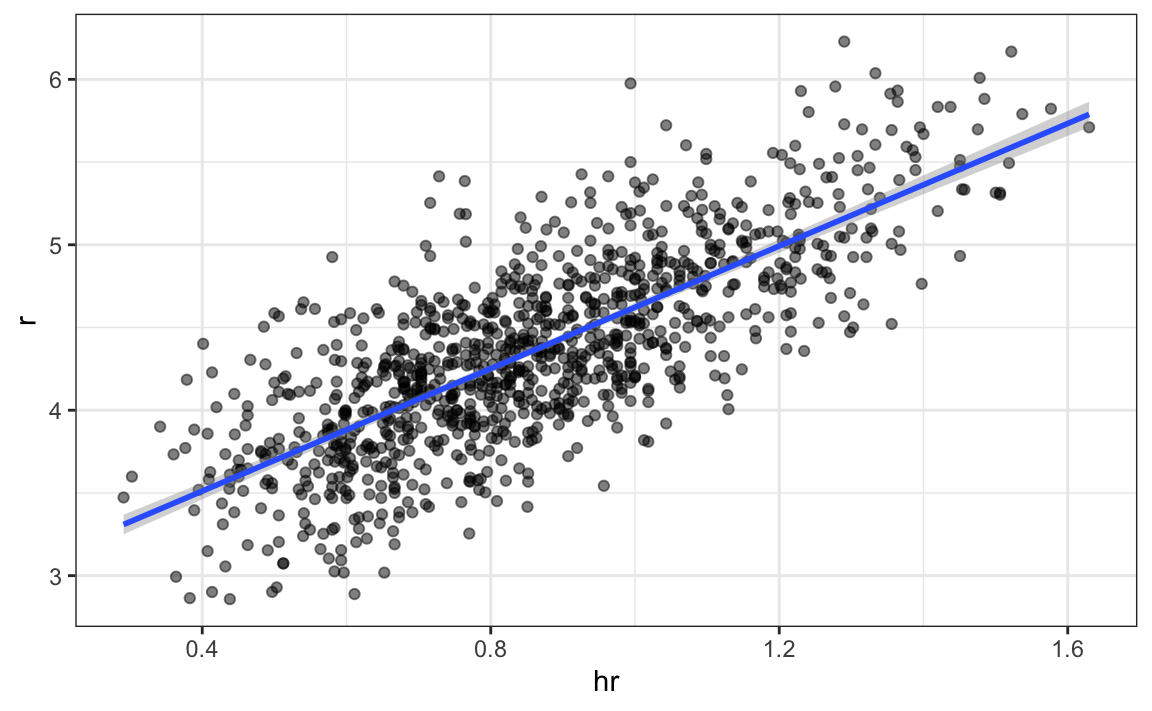
Bây giờ chúng ta đã sẵn sàng sử dụng hồi quy tuyến tính để dự đoán số lần chạy mà một đội sẽ ghi được nếu chúng ta biết đội đó đánh bao nhiêu lần trên sân nhà bằng cách sử dụng hồi quy:

|  |
| --- |
| hr\_fit  <- lm(r ~ hr, data = dat)$coef  p + geom\_abline(intercept = hr\_fit[[1]], slope = hr\_fit[[2]]) |



Lưu ý rằng chúng ta có thể có được cùng một biểu đồ nhanh hơn bằng cách sử dụng hàm **ggplot2** để tính toán và thêm một đường hồi quy vào biểu đồ cùng với các khoảng tin cậy. Chúng tôi sử dụng đối số viết tắt của *mô hình tuyến tính*, tiêu đề của một phần sắp tới. Vì vậy, chúng ta có thể đơn giản hóa code ở trên như thế này:geom\_smoothmethod = "lm"

|  |
| --- |
| p + geom\_smooth(method = "lm") |



Trong ví dụ trên, độ dốc là 1.8517449. Vì vậy, điều này cho chúng ta biết rằng các đội đạt 1 HR nhiều hơn mỗi trận so với đội trung bình, ghi được nhiều hơn 1.8517449 lần chạy mỗi trận so với đội trung bình. Cho rằng điểm số cuối cùng phổ biến nhất là sự khác biệt của một lần chạy, điều này chắc chắn có thể dẫn đến sự gia tăng lớn trong chiến thắng. Không có gì đáng ngạc nhiên, những người đánh nhân sự rất tốn kém. Bởi vì chúng tôi đang làm việc trên một ngân sách, chúng tôi sẽ cần phải tìm một số cách khác để tăng chiến thắng. Trong chương tiếp theo, chúng tôi giới thiệu các *mô hình tuyến tính*, cung cấp một khuôn khổ để thực hiện phân tích này. Trong chương @ref{@moneyball} chúng ta áp dụng những gì đã học được để xây dựng một đội bóng chày.

## 14.2 Gói chổi

Gói **chổi** tạo điều kiện thuận lợi cho việc sử dụng chức năng R như trong vũ trụ gọn gàng. Hãy nhớ lại điều đó không lấy khung dữ liệu làm đối số đầu tiên và không trả về khung dữ liệu, điều này làm cho việc sử dụng kết hợp với **vũ trụ gọn gàng** trở nên khó khăn. Nó có ba chức năng chính, tất cả đều trích xuất thông tin từ đối tượng được trả về và trả về nó trong một khung dữ liệu thân thiện với **vũ trụ gọn gàng**. Các hàm này là , , và . Hàm trả về ước tính và thông tin liên quan dưới dạng khung dữ liệu:lmlmlmlmtidyglanceaugmenttidy

|  |
| --- |
| library(broom)  fit <- lm(r ~ bb, data = dat)  tidy(fit)  #> # A tibble: 2 × 5  #>   term        estimate std.error statistic  p.value  #>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>  #> 1 (Intercept)    1.93     0.116       16.7 1.91e-55  #> 2 bb             0.739    0.0348      21.2 1.90e-83 |

Chúng tôi có thể thêm các tóm tắt quan trọng khác, chẳng hạn như khoảng tin cậy:

|  |
| --- |
| tidy(fit, conf.int = TRUE)  #> # A tibble: 2 × 7  #>   term        estimate std.error statistic  p.value conf.low conf.high  #>   <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>  #> 1 (Intercept)    1.93     0.116       16.7 1.91e-55    1.70      2.15  #> 2 bb             0.739    0.0348      21.2 1.90e-83    0.671     0.807 |

Bởi vì kết quả là một khung dữ liệu, chúng ta có thể ngay lập tức sử dụng nó để xâu chuỗi các lệnh tạo ra bảng mà chúng ta đang theo đuổi. Vì một khung dữ liệu được trả về, chúng ta có thể lọc và chọn các hàng và cột mà chúng ta muốn, như chúng ta sẽ thấy trong phần tiếp theo.summarize

Bây giờ chúng ta trở lại thảo luận về nhiệm vụ ban đầu của chúng ta là xác định xem độ dốc có thay đổi hay không. Sơ đồ chúng tôi vừa thực hiện, sử dụng và , cho thấy các khoảng tin cậy chồng lên nhau, điều này cung cấp một xác nhận trực quan tốt đẹp rằng giả định của chúng tôi rằng độ dốc không thay đổi là an toàn.summarizetidy

Các chức năng khác được cung cấp bởi **chổi**, và , liên quan đến các kết quả cụ thể của mô hình và quan sát cụ thể, tương ứng. Ở đây, chúng ta có thể thấy các bản tóm tắt phù hợp với mô hình trả về:glanceaugmentglance

|  |
| --- |
| glance(fit)  #> # A tibble: 1 × 12  #>   r.squared adj.r.squared sigma statistic  p.value    df logLik   AIC  #>       <dbl>         <dbl> <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl>  #> 1     0.304         0.303 0.493      451. 1.90e-83     1  -737. 1480.  #> # ℹ 4 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>,  #> #   nobs <int> |

Bạn có thể tìm hiểu thêm về các tóm tắt này trong bất kỳ sách giáo khoa hồi quy nào.

## 14.3 Gây nhiễu

Trước đây, chúng tôi đã ghi nhận mối quan hệ bền chặt giữa Runs và BB. Nếu chúng ta tìm thấy đường hồi quy để dự đoán các lần chạy từ các cơ sở trên quả bóng, chúng ta sẽ có độ dốc của:

|  |
| --- |
| bb\_slope <- lm(r ~ bb, data = dat)$coef[2]  bb\_slope  #>    bb  #> 0.739 |

Vì vậy, điều này có nghĩa là nếu chúng tôi đi và thuê những cầu thủ lương thấp với nhiều BB, và do đó tăng số lần đi bộ mỗi trận lên 2, đội của chúng tôi sẽ ghi thêm 1,5 lần chạy mỗi trận?

Chúng ta một lần nữa được nhắc nhở rằng sự liên kết không phải là quan hệ nhân quả. Dữ liệu cung cấp bằng chứng mạnh mẽ rằng một đội có nhiều BB hơn hai lần mỗi trận so với đội trung bình, ghi được 1,5 lần chạy mỗi trận. Nhưng điều này không có nghĩa là BB là nguyên nhân.

Lưu ý rằng nếu chúng ta tính toán độ dốc đường hồi quy cho người độc thân, chúng ta nhận được:

|  |
| --- |
| lm(r ~ singles, data = dat)$coef[2]  #> singles  #>   0.432 |

đó là một giá trị thấp hơn so với những gì chúng tôi nhận được cho BB. Lưu ý rằng một đĩa đơn đưa bạn đến cơ sở đầu tiên giống như BB. Những người biết về bóng chày sẽ cho bạn biết rằng với một đĩa đơn, người chạy trên cơ sở có cơ hội ghi bàn tốt hơn so với BB. Vậy làm thế nào BB có thể dự đoán nhiều hơn về các lần chạy? Lý do điều này xảy ra là vì sự nhầm lẫn. Ở đây chúng tôi chỉ ra mối tương quan giữa HR, BB và người độc thân:

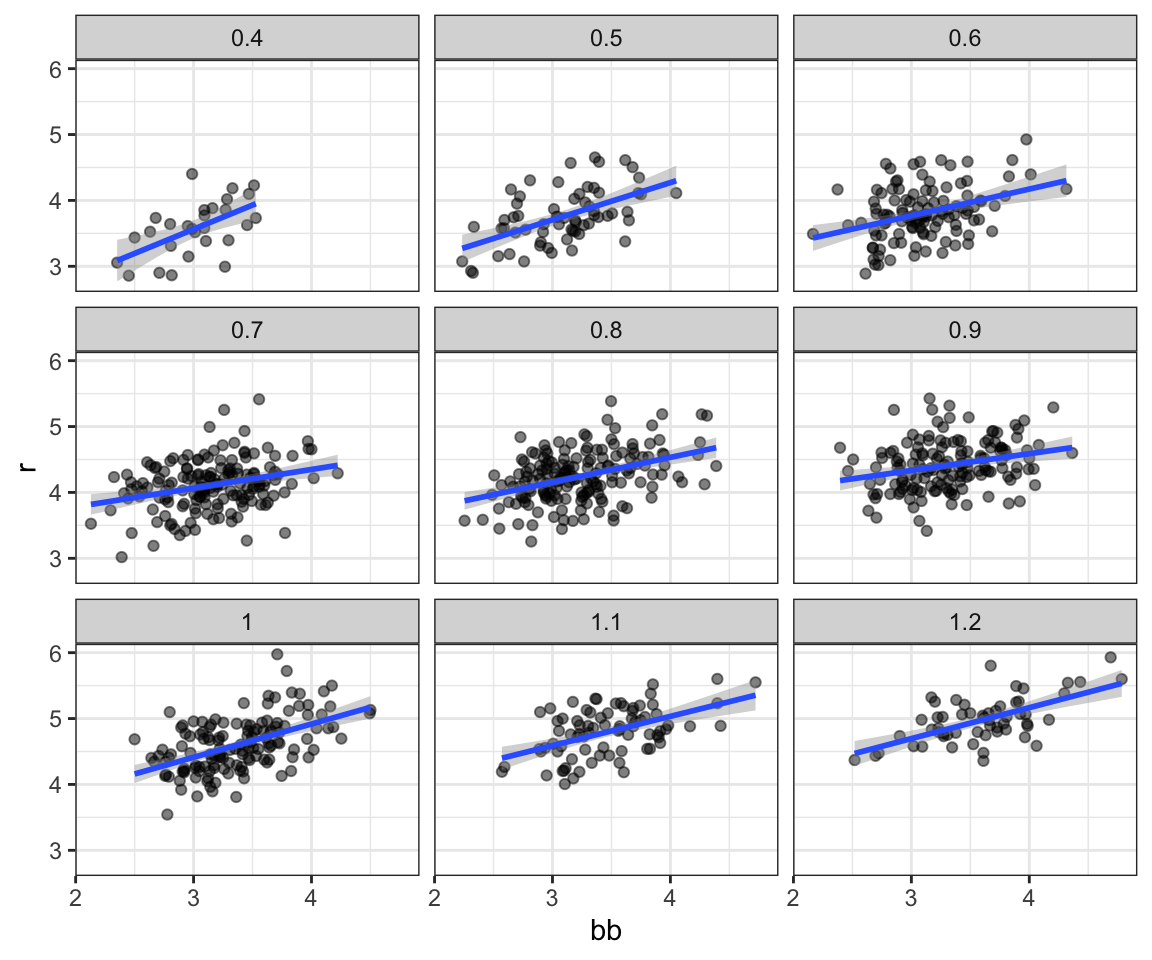
|  |
| --- |
| dat |> summarize(cor(bb, hr), cor(singles, hr), cor(bb, singles))  #>   cor(bb, hr) cor(singles, hr) cor(bb, singles)  #> 1       0.406           -0.186          -0.0513 |

Nó chỉ ra rằng những người ném bóng, sợ nhân sự, đôi khi sẽ tránh ném các cuộc đình công cho những người đánh nhân sự. Do đó, những người đánh nhân sự có xu hướng có nhiều BB hơn và một nhóm có nhiều nhân sự cũng sẽ có nhiều BB hơn. Mặc dù có vẻ như BB gây ra các lần chạy, nhưng thực sự chính các nhân sự gây ra hầu hết các lần chạy này. Chúng tôi nói rằng BB bị *nhầm lẫn* với nhân sự. Tuy nhiên, có thể là BB vẫn giúp đỡ? Để tìm hiểu, bằng cách nào đó chúng ta phải điều chỉnh hiệu ứng nhân sự. Hồi quy cũng có thể giúp với điều này.

### 14.3.1 Hiểu nhiễu thông qua phân tầng

Cách tiếp cận đầu tiên là giữ cho nhân sự cố định ở một giá trị nhất định và sau đó kiểm tra mối quan hệ giữa BB và chạy. Như chúng tôi đã làm khi chúng tôi phân tầng các ông bố bằng cách làm tròn đến inch gần nhất, ở đây chúng tôi có thể phân tầng HR mỗi trò chơi thành mười gần nhất. Chúng tôi lọc ra các địa tầng với một vài điểm để tránh các ước tính biến đổi cao và sau đó tạo một biểu đồ phân tán cho mỗi tầng:

|  |
| --- |
| dat |> mutate(hr\_strata = round(hr, 1)) |>    filter(hr\_strata >= 0.4 & hr\_strata <= 1.2) |>    ggplot(aes(bb, r)) +    geom\_point(alpha = 0.5) +    geom\_smooth(method = "lm") +    facet\_wrap(~hr\_strata) |



Hãy nhớ rằng độ dốc hồi quy để dự đoán các lần chạy với BB là 0,7. Khi chúng tôi phân tầng theo nhân sự, những độ dốc này sẽ giảm đáng kể:

|  |
| --- |
| dat |> mutate(hr\_strata = round(hr, 1)) |>    filter(hr\_strata >= 0.5 & hr\_strata <= 1.2) |>    group\_by(hr\_strata) |>    reframe(tidy(lm(r ~ bb))) |>    filter(term == "bb")  #> # A tibble: 8 × 6  #>   hr\_strata term  estimate std.error statistic      p.value  #>       <dbl> <chr>    <dbl>     <dbl>     <dbl>        <dbl>  #> 1       0.5 bb       0.566    0.110       5.14 0.00000302  #> 2       0.6 bb       0.405    0.0984      4.12 0.0000746  #> 3       0.7 bb       0.284    0.0717      3.96 0.000113  #> 4       0.8 bb       0.378    0.0638      5.92 0.0000000175  #> 5       0.9 bb       0.254    0.0762      3.33 0.00108  #> # ℹ 3 more rows |

|  |
| --- |
| Lưu ý chúng ta sử dụng thay vì vì trả về một khung dữ liệu có hai hàng.reframesummarizetidy |

Các độ dốc được giảm, nhưng chúng không phải là 0, điều này cho thấy BB rất hữu ích cho việc tạo ra các lần chạy, chỉ là không nhiều như suy nghĩ trước đây. Trên thực tế, các giá trị trên gần với độ dốc mà chúng ta thu được từ những người độc thân, 0,4, phù hợp hơn với trực giác của chúng ta. Vì cả đĩa đơn và BB đều đưa chúng ta đến căn cứ đầu tiên, chúng nên có cùng sức mạnh dự đoán.

Mặc dù sự hiểu biết của chúng tôi về ứng dụng cho chúng tôi biết rằng HR gây ra BB nhưng không phải ngược lại, chúng tôi vẫn có thể kiểm tra xem việc phân tầng theo BB có làm giảm hiệu ứng của BB hay không. Để làm điều này, chúng tôi sử dụng cùng một mã ngoại trừ việc chúng tôi trao đổi HR và BB. Trong trường hợp này, các sườn dốc không thay đổi nhiều so với ban đầu:

|  |
| --- |
| dat |> mutate(bb\_strata = round(bb, 1)) |>    filter(bb\_strata >= 3 & bb\_strata <= 4) |>    group\_by(bb\_strata) |>    reframe(tidy(lm(r ~ hr))) |>    filter(term == "hr")  #> # A tibble: 11 × 6  #>   bb\_strata term  estimate std.error statistic  p.value  #>       <dbl> <chr>    <dbl>     <dbl>     <dbl>    <dbl>  #> 1       3   hr        1.51     0.182      8.31 1.47e-12  #> 2       3.1 hr        1.49     0.168      8.87 3.10e-14  #> 3       3.2 hr        1.61     0.150     10.8  6.96e-18  #> 4       3.3 hr        1.57     0.167      9.39 5.73e-15  #> 5       3.4 hr        1.55     0.153     10.1  3.77e-16  #> # ℹ 6 more rows |

Chúng giảm một chút từ 1.8517449, điều này phù hợp với thực tế là BB thực sự gây ra một số lần chạy.

Bất kể, có vẻ như nếu chúng ta phân tầng theo nhân sự, chúng ta có phân phối hai biến cho các lần chạy so với BB. Tương tự, nếu chúng ta phân tầng theo BB, chúng ta có các phân phối bình thường hai biến gần đúng cho HR so với chạy.

## 14.4 Hồi quy đa biến

Nó hơi phức tạp để tính toán các đường hồi quy cho từng tầng. Về cơ bản, chúng tôi đang phù hợp với các mô hình như thế này:

với các sườn dốc cho  thay đổi cho các giá trị khác nhau của  và ngược lại. Nhưng có cách tiếp cận nào dễ dàng hơn không?

Nếu chúng ta tính đến sự thay đổi ngẫu nhiên, các sườn dốc trong các tầng dường như không thay đổi nhiều. Nếu những sườn dốc này trên thực tế giống nhau, điều này ngụ ý rằng  và  là hằng số. Điều này đến lượt nó ngụ ý rằng kỳ vọng của các lần chạy có điều kiện trên HR và BB có thể được viết như thế này:

Mô hình này cho thấy rằng nếu số lượng nhân sự được cố định tại , chúng tôi quan sát thấy mối quan hệ tuyến tính giữa chạy và BB với sự chặn của . Phân tích dữ liệu thăm dò của chúng tôi cho thấy rằng đây là trường hợp. Mô hình cũng cho thấy rằng khi số lượng nhân sự tăng lên, tăng trưởng đánh chặn cũng tuyến tính và được xác định bởi . Trong phân tích này, được gọi là *hồi quy đa biến*, bạn sẽ thường nghe mọi người nói rằng độ dốc BB  được *điều chỉnh* theo hiệu ứng nhân sự.

Bởi vì dữ liệu xấp xỉ bình thường và phân phối có điều kiện cũng bình thường, chúng tôi hợp lý khi sử dụng mô hình tuyến tính:

với  Số lần chạy mỗi trận cho đội ,  đi bộ mỗi trò chơi, và . Để sử dụng ở đây, chúng ta cần cho hàm biết chúng ta có hai biến dự đoán. Vì vậy, chúng tôi sử dụng biểu tượng như sau:lm+

|  |
| --- |
| tidy(lm(r ~ bb + hr, data = dat), conf.int = TRUE)  #> # A tibble: 3 × 7  #>   term        estimate std.error statistic   p.value conf.low conf.high  #>   <chr>          <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>  #> 1 (Intercept)    1.74     0.0820      21.2 3.38e- 83    1.58      1.90  #> 2 bb             0.387    0.0269      14.4 8.41e- 43    0.334     0.440  #> 3 hr             1.57     0.0488      32.1 1.39e-157    1.47      1.66 |

Khi chúng tôi lắp mô hình chỉ với một biến, độ dốc ước tính lần lượt là 0,7388725 và 1,8517449 cho BB và HR. Lưu ý rằng khi lắp mô hình đa biến đều đi xuống, với hiệu ứng BB giảm hơn nhiều.

|  |
| --- |
| Bạn đã sẵn sàng để làm bài tập 1-12 nếu bạn muốn thực hành trước khi tiếp tục. |

### 14.4.1 Xây dựng đội bóng chày

Bây giờ chúng tôi muốn xây dựng một số liệu để chọn người chơi, và chúng tôi cũng cần xem xét đơn, đôi và ba. Chúng ta có thể xây dựng một mô hình dự đoán chạy dựa trên tất cả các kết quả này không? Chúng tôi phần nào thực hiện một "bước nhảy vọt của niềm tin" và cho rằng năm biến này là bình thường chung. Điều này có nghĩa là nếu chúng ta chọn bất kỳ một trong số chúng và giữ bốn giá trị còn lại cố định, mối quan hệ với kết quả là tuyến tính và độ dốc không phụ thuộc vào bốn giá trị được giữ không đổi. Nếu điều này là đúng, thì một mô hình tuyến tính cho dữ liệu của chúng tôi là:

với với  đại diện cho BB, đơn, đôi, ba và HR tương ứng.

Sử dụng , chúng ta có thể nhanh chóng tìm thấy LSE cho các tham số bằng cách sử dụng:lm

đại diện cho BB, đơn, đôi, ba và HR tương ứng.

Sử dụng , chúng ta có thể nhanh chóng tìm thấy LSE cho các tham số bằng cách sử dụng:lm

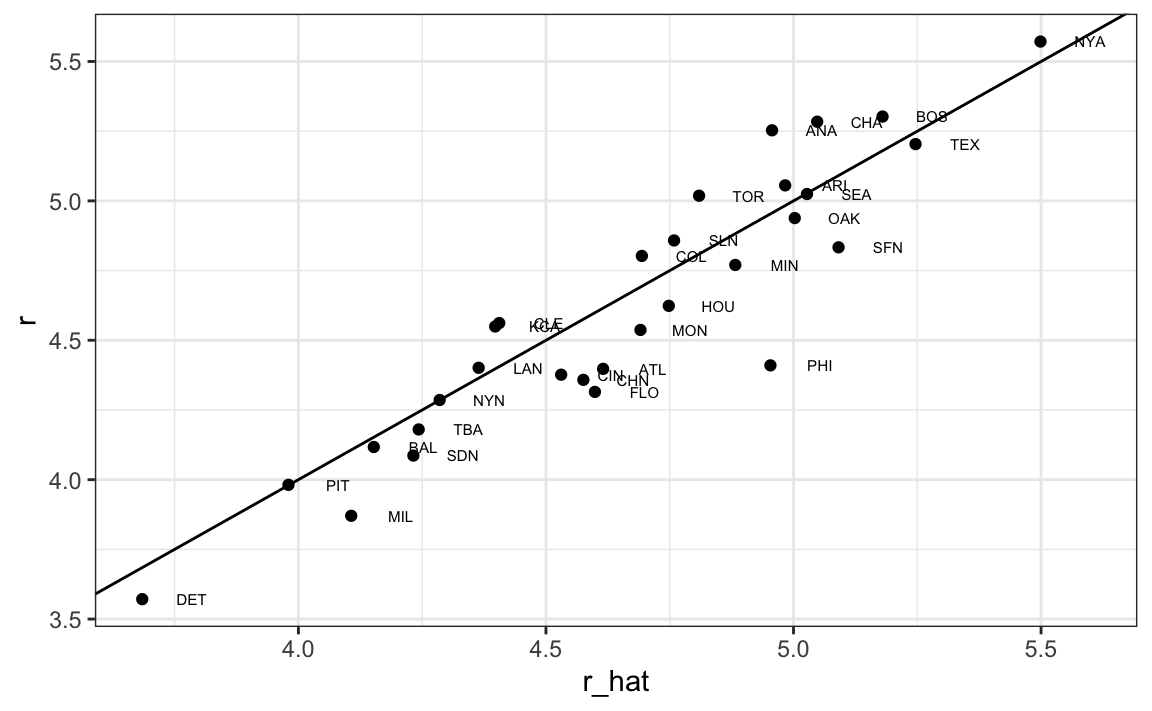
|  |
| --- |
| fit <- dat |>  filter(year <= 2001) |> lm(r ~ bb + singles + doubles + triples + hr, data = \_) |

Lưu ý rằng chúng tôi phù hợp với mô hình để dữ liệu cho đến năm 2001, năm trước khi chúng tôi sẽ xây dựng nhóm của mình. Chúng ta có thể thấy các hệ số bằng cách sử dụng :tidy

|  |
| --- |
| tidy(fit, conf.int = TRUE) |> filter(term != "(Intercept)")  #> # A tibble: 5 × 7  #>   term    estimate std.error statistic   p.value conf.low conf.high  #>   <chr>      <dbl>     <dbl>     <dbl>     <dbl>    <dbl>     <dbl>  #> 1 bb         0.370    0.0119      31.2 1.00e-149    0.347     0.393  #> 2 singles    0.517    0.0128      40.5 5.29e-213    0.492     0.543  #> 3 doubles    0.775    0.0229      33.8 7.09e-168    0.730     0.820  #> 4 triples    1.24     0.0778      15.9 4.62e- 51    1.09      1.39  #> 5 hr         1.44     0.0248      58.1 1.98e-323    1.39      1.49 |

Để xem số liệu của chúng tôi thực sự dự đoán số lần chạy tốt như thế nào, chúng tôi có thể dự đoán số lần chạy cho mỗi đội trong năm 2002 bằng cách sử dụng hàm , sau đó tạo một biểu đồ:predict

|  |
| --- |
| dat |> mutate(r\_hat = predict(fit, newdata = dat)) |>    filter(year == 2002) %>%    ggplot(aes(r\_hat, r, label = team)) +    geom\_point() +    geom\_text(nudge\_x = 0.1, cex = 2) +    geom\_abline() |



Mô hình của chúng tôi thực hiện khá tốt công việc được chứng minh bằng thực tế là các điểm từ biểu đồ quan sát được so với dự đoán rơi gần với đường nhận dạng.

Vì vậy, thay vì sử dụng mức trung bình đánh bóng, hoặc chỉ số lượng nhân sự, làm thước đo chọn người chơi, chúng tôi có thể sử dụng mô hình được trang bị của mình để tạo thành một số liệu liên quan trực tiếp hơn đến việc điều hành sản xuất. Cụ thể, để xác định số liệu cho người chơi A, chúng tôi tưởng tượng một đội gồm những người chơi giống như người chơi A và sử dụng mô hình hồi quy được trang bị của chúng tôi để dự đoán số lần chạy mà đội này sẽ tạo ra. Công thức sẽ như thế này: -2.7580763 + 0.3699921 × BB + 0.5174284 × Đĩa đơn + 0.7750757 × Đôi + 1.2387738 × Bộ ba + 1.4419724 × HR.

Để xác định số liệu cụ thể cho người chơi, chúng tôi còn một chút việc phải làm. Một thách thức ở đây là chúng tôi đã lấy được số liệu cho các đội, dựa trên số liệu thống kê tóm tắt cấp nhóm. Ví dụ: giá trị nhân sự được nhập vào phương trình là HR mỗi trò chơi cho toàn đội. Nếu chúng ta tính HR mỗi trận cho một người chơi, nó sẽ thấp hơn nhiều vì tổng số được tích lũy bởi 9 người đánh bóng. Hơn nữa, nếu một người chơi chỉ chơi một phần của trò chơi và nhận được ít cơ hội hơn mức trung bình, nó vẫn được coi là một trò chơi đã chơi. Đối với người chơi, một tỷ lệ có tính đến các cơ hội là tỷ lệ xuất hiện trên mỗi đĩa.

Để làm cho tỷ lệ đội trên mỗi trận đấu có thể so sánh với tỷ lệ người chơi trên mỗi đĩa xuất hiện, chúng tôi tính số lần xuất hiện trung bình của đội trên mỗi trò chơi:

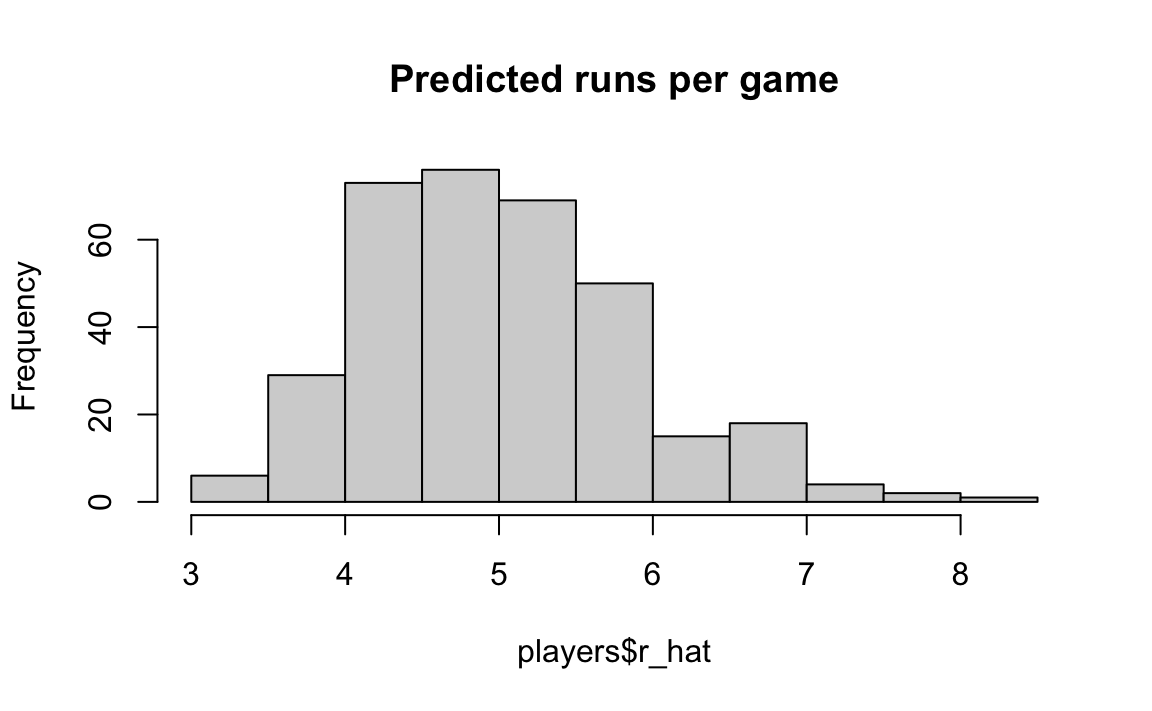
|  |
| --- |
| pa\_per\_game <- Batting |> filter(yearID == 2002) |>    group\_by(teamID) |>    summarize(pa\_per\_game = sum(AB + BB)/162) |>    pull(pa\_per\_game) |>    mean() |

Chúng tôi tính toán tỷ lệ xuất hiện trên mỗi tấm cho người chơi có sẵn trong năm 2002 trên dữ liệu từ năm 1997-2001. Để tránh các hiện vật mẫu nhỏ, chúng tôi lọc những người chơi có ít hơn 1.000 lần xuất hiện đĩa mỗi năm. Dưới đây là toàn bộ tính toán trong một dòng:

|  |
| --- |
| players <- Batting |>    filter(yearID %in% 1997:2001) |>    group\_by(playerID) |>    mutate(pa = BB + AB) |>    summarize(g = sum(pa)/pa\_per\_game,      bb = sum(BB)/g,      singles = sum(H - X2B - X3B - HR)/g,      doubles = sum(X2B)/g,      triples = sum(X3B)/g,      hr = sum(HR)/g,      avg = sum(H)/sum(AB),      pa = sum(pa)) |>    filter(pa >= 1000) |>    select(-g)  players$r\_hat = predict(fit, newdata = players) |

Số lần chạy dự đoán cụ thể của người chơi được tính ở đây có thể được hiểu là số lần chạy mà chúng tôi dự đoán một đội sẽ ghi bàn nếu tất cả các tay đập đều giống hệt như người chơi đó. Bản phân phối cho thấy có sự thay đổi lớn giữa những người chơi:

|  |
| --- |
| hist(players$r\_hat, main = "Predicted runs per game") |



Để thực sự xây dựng đội bóng, chúng tôi sẽ cần biết mức lương cũng như vị trí phòng ngự của họ. Đối với điều này, chúng tôi sử dụng hàm để kết hợp khung dữ liệu mà chúng tôi vừa tạo với khung dữ liệu thông tin người chơi có trong một số bảng dữ liệu Lahman khác.righ\_joinplayers

Bắt đầu bằng cách cộng mức lương năm 2002 của mỗi cầu thủ:

|  |
| --- |
| players <- Salaries |>    filter(yearID == 2002) |>    select(playerID, salary) |>    right\_join(players, by = "playerID") |

Tiếp theo, chúng tôi thêm vị trí phòng thủ của họ. Đây là một nhiệm vụ hơi phức tạp vì người chơi chơi nhiều hơn một vị trí mỗi năm. Bảng gói **Lahman** cho biết mỗi cầu thủ đã chơi bao nhiêu trận ở mỗi vị trí, vì vậy chúng tôi có thể chọn vị trí được chơi nhiều nhất bằng cách sử dụng trên mỗi hàng. Chúng tôi sử dụng để làm điều này. Tuy nhiên, vì một số người chơi được giao dịch, họ xuất hiện nhiều hơn một lần trên bàn, vì vậy trước tiên chúng tôi tổng hợp số lần xuất hiện của họ giữa các đội. Ở đây, chúng tôi chọn một vị trí mà người chơi chơi nhiều nhất bằng cách sử dụng chức năng. Để đảm bảo chúng tôi chỉ chọn một vị trí, trong trường hợp ràng buộc, chúng tôi chọn hàng đầu tiên của khung dữ liệu kết quả. Chúng tôi cũng loại bỏ vị trí viết tắt của tiền vệ, khái quát hóa ba vị trí: trường trái (LF), trường trung tâm (CF) và trường bên phải (RF). Chúng tôi cũng loại bỏ các cầu thủ ném bóng vì họ không đánh trong giải đấu mà A chơi.Appearanceswhich.maxapplytop\_nOF

|  |
| --- |
| position\_names <-    paste0("G\_", c("p","c","1b","2b","3b","ss","lf","cf","rf", "dh"))  tmp <- Appearances |>    filter(yearID == 2002) |>    group\_by(playerID) |>    summarize\_at(position\_names, sum) |>    ungroup()    pos <- tmp |>    select(all\_of(position\_names)) |>    apply(X = \_, 1, which.max)  players <- tibble(playerID = tmp$playerID, POS = position\_names[pos]) |>    mutate(POS = str\_to\_upper(str\_remove(POS, "G\_"))) |>    filter(POS != "P") |>    right\_join(players, by = "playerID") |>    filter(!is.na(POS)  & !is.na(salary)) |

Cuối cùng, chúng tôi thêm họ và tên của họ:

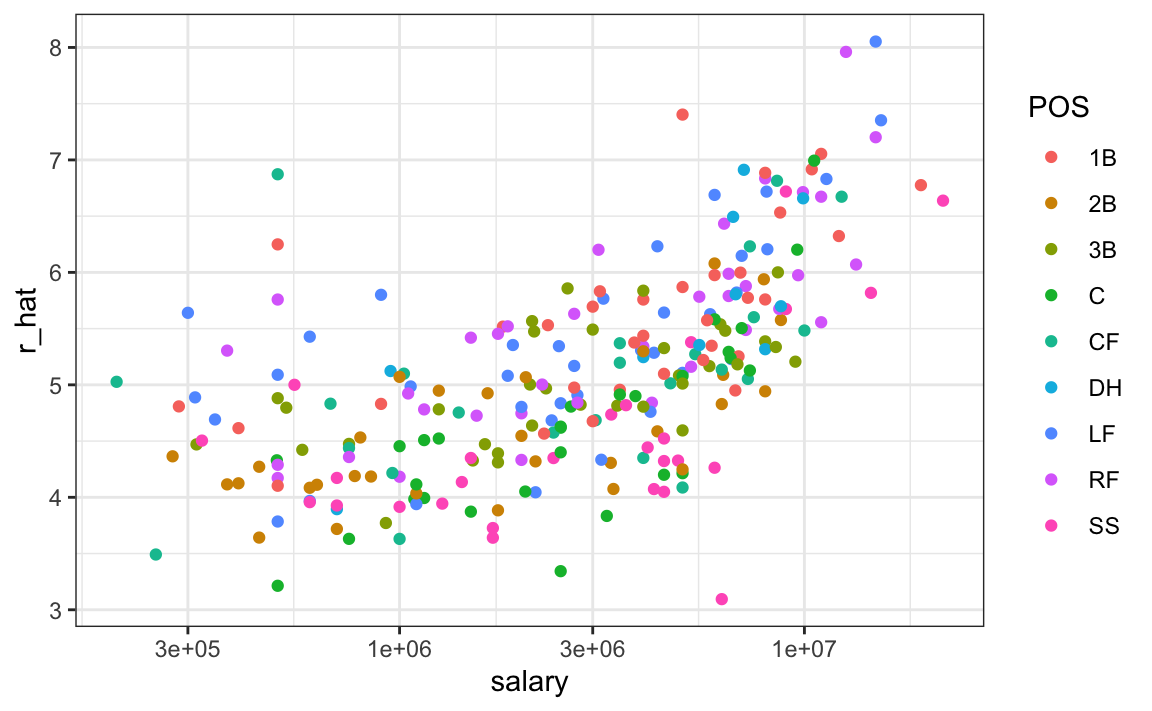
|  |
| --- |
| players <- People |>    select(playerID, nameFirst, nameLast, debut) |>    mutate(debut = as.Date(debut)) |>    right\_join(players, by = "playerID") |

Nếu bạn là một fan hâm mộ bóng chày, bạn sẽ nhận ra 10 cầu thủ hàng đầu:

|  |
| --- |
| players |> select(nameFirst, nameLast, POS, salary, r\_hat) |> arrange(desc(r\_hat)) |> head(10)  #>    nameFirst nameLast POS   salary r\_hat  #> 1      Barry    Bonds  LF 15000000  8.05  #> 2      Larry   Walker  RF 12666667  7.96  #> 3       Todd   Helton  1B  5000000  7.40  #> 4      Manny  Ramirez  LF 15462727  7.35  #> 5      Sammy     Sosa  RF 15000000  7.20  #> 6       Jeff  Bagwell  1B 11000000  7.05  #> 7       Mike   Piazza   C 10571429  6.99  #> 8      Jason   Giambi  1B 10428571  6.92  #> 9      Edgar Martinez  DH  7086668  6.91  #> 10       Jim    Thome  1B  8000000  6.89 |

Trung bình, những cầu thủ có chỉ số cao hơn có mức lương cao hơn:

|  |
| --- |
| players |> ggplot(aes(salary, r\_hat, color = POS)) +    geom\_point() +    scale\_x\_log10() |



Chúng ta có thể tìm kiếm những thỏa thuận tốt bằng cách xem xét những cầu thủ tạo ra nhiều lần chạy hơn những người khác có mức lương tương tự. Chúng tôi có thể sử dụng bảng này để quyết định chọn cầu thủ nào và giữ tổng mức lương của chúng tôi dưới mức 40 triệu đô la mà Billy Beane phải làm việc cùng. Điều này có thể được thực hiện bằng cách sử dụng những gì các nhà khoa học máy tính gọi là lập trình tuyến tính. Đây không phải là điều chúng tôi dạy, nhưng đây là những người chơi vị trí được chọn với phương pháp này:

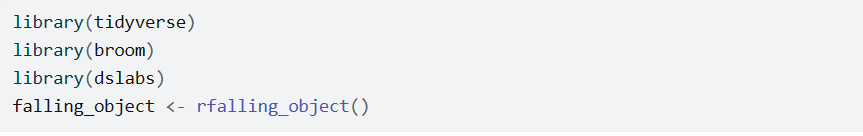
| **nameFirst** | **tênCuối cùng** | **POS** | **lương** | **r\_hat** |
| --- | --- | --- | --- | --- |
| Todd | Sân bay trực thăng | 1 tỷ | 5000000 | 7.40 |
| Micrô | Quảng trường | C | 10571429 | 6.99 |
| Edgar | Martinez | DH | 7086668 | 6.91 |
| Jim | Edmonds | CF | 7333333 | 6.23 |
| Jeff | Kent | 2 tỷ | 6000000 | 6.08 |
| Phil | Nevin | 3 tỷ | 2600000 | 5.86 |
| Matt | Cầu thang | RF | 500000 | 5.76 |
| Henry | Rodriguez | LF | 300000 | 5.64 |
| John | Valentin | SS | 550000 | 5.00 |

Chúng tôi thấy rằng tất cả những người chơi này đều có BB trên trung bình và hầu hết đều có tỷ lệ nhân sự trên trung bình, trong khi điều tương tự không đúng với người đánh đơn và đánh bóng trung bình. Dưới đây là bảng với số liệu thống kê được chuẩn hóa giữa những người chơi để, ví dụ, những người đánh nhân sự trên trung bình có giá trị trên 0.

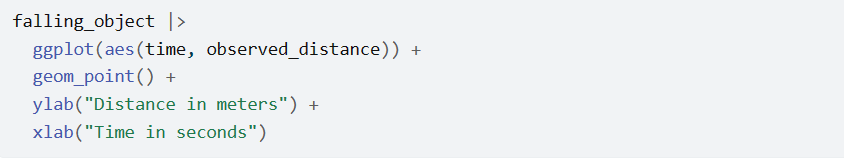
| **tênCuối cùng** | **Bb** | **Đĩa đơn** | **Đôi** | **gấp ba** | **Hr** | **Avg** | **r\_hat** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Sân bay trực thăng | 0.909 | -0.215 | 2.649 | -0.311 | 1.522 | 2.670 | 2.542 |
| Quảng trường | 0.328 | 0.423 | 0.204 | -1.418 | 1.825 | 2.199 | 2.093 |
| Martinez | 2.135 | -0.005 | 1.265 | -1.224 | 0.808 | 2.203 | 2.004 |
| Edmonds | 1.071 | -0.558 | 0.791 | -1.152 | 0.973 | 0.854 | 1.259 |
| Kent | 0.232 | -0.732 | 2.011 | 0.448 | 0.766 | 0.787 | 1.093 |
| Nevin | 0.307 | -0.905 | 0.479 | -1.191 | 1.193 | 0.105 | 0.850 |
| Cầu thang | 1.100 | -1.513 | -0.046 | -1.129 | 1.121 | -0.561 | 0.742 |
| Rodriguez | 0.201 | -1.596 | 0.332 | -0.782 | 1.320 | -0.672 | 0.613 |
| Valentin | 0.180 | -0.929 | 1.794 | -0.435 | -0.045 | -0.472 | -0.088 |

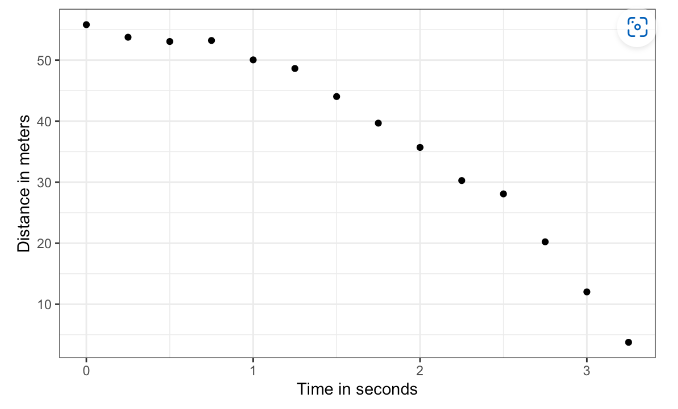
# 15.Mô hình lỗi đo lường

Một ứng dụng quan trọng khác của mô hình tuyến tính đến từ các mô hình lỗi đo lường. Trong những ứng dụng này, thường xuyên xuất hiện một biến độc lập không ngẫu nhiên, chẳng hạn như thời gian, và sự ngẫu nhiên được giới thiệu từ lỗi đo lường thay vì từ sự lấy mẫu hoặc biến động tự nhiên.

Để hiểu rõ về những mô hình này, hãy tưởng tượng bạn là Galileo trong thế kỷ 16 cố gắng mô tả vận tốc của một vật rơi. Một trợ lý leo lên Tháp Pisa và thả một quả cầu, trong khi một số trợ lý khác ghi lại vị trí tại các thời điểm khác nhau. Hãy mô phỏng một số dữ liệu bằng cách sử dụng các phương trình chúng ta biết ngày nay và thêm vào đó một số lỗi đo lường. Hàm rfalling\_object của gói dslabs tạo ra những mô phỏng này. 

Các trợ lý truyền dữ liệu cho Galileo và đây là những gì ông ta nhìn thấy:





Galileo không biết chính xác phương trình, nhưng nhìn vào đồ thị ở trên, ông ta suy luận rằng vị trí nên tuân theo một hàm bậc hai (parabol), mà chúng ta có thể viết như sau:



Dữ liệu không chính xác rơi trên một đường parabol. Galileo biết rằng điều này là do lỗi đo lường. Những người trợ lý của ông ấy mắc sai lầm khi đo khoảng cách. Để tính đến điều này, ông ấy mô hình dữ liệu bằng cách sử dụng:

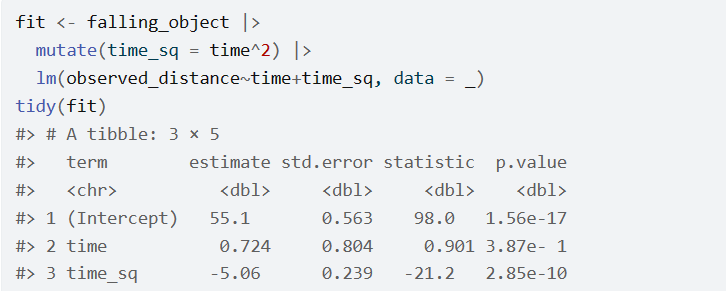


với yi đại diện cho khoảng cách trong mét, xi đại diện cho thời gian trong giây, và e đại diện cho lỗi đo lường. Giả sử lỗi đo lường là ngẫu nhiên, độc lập với nhau và có phân phối giống nhau cho mỗi i. Chúng ta cũng giả định rằng không có độ chệch, điều này có nghĩa là giá trị kỳ vọng của

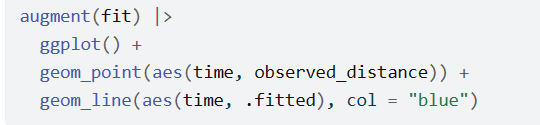
Chú ý rằng đây là một mô hình tuyến tính vì nó là một tổ hợp tuyến tính của các lượng đã biết (xi và yi là đã biết) và các tham số không biết (các βs là các tham số không biết đến Galileo). Khác với các ví dụ trước, ở đây xi là một lượng cố định; chúng ta không đang điều kiện.

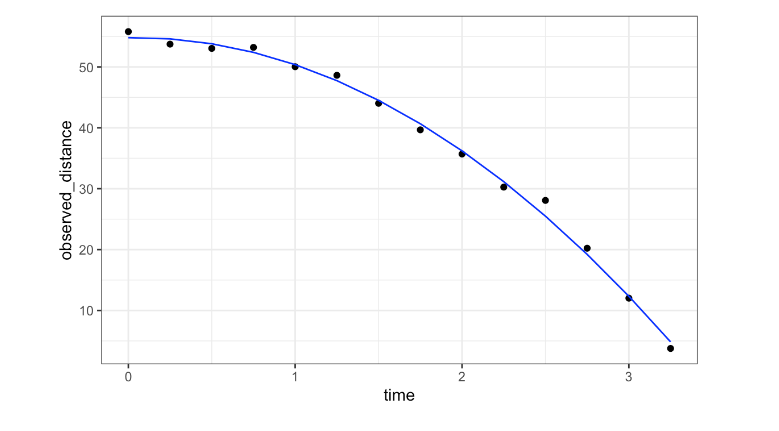
Để đưa ra một lý thuyết vật lý mới và bắt đầu đưa ra dự đoán về các vật rơi khác, Galileo cần các con số thực tế, thay vì các tham số không biết. Sử dụng phương pháp LSE dường như là một cách tiếp cận hợp lý. Làm thế nào chúng ta tìm LSE?

Các tính toán LSE không yêu cầu lỗi phải gần như tuân theo phân phối chuẩn. Hàm lm sẽ tìm ra các βs mà sẽ làm giảm thiểu tổng bình phương dư thừa:

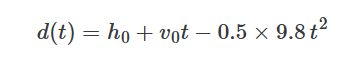


Hãy kiểm tra xem đường parabol được ước lượng có phù hợp với dữ liệu hay không. Hàm augment của gói broom giúp chúng ta thực hiện điều này một cách dễ dàng:





Cảm ơn vì thông tin từ giáo viên vật lý trung học của bạn, tôi biết rằng phương trình cho quỹ đạo của một vật rơi là:

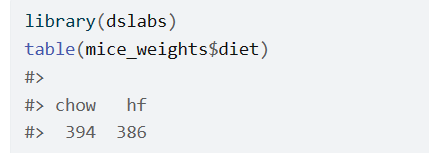


# 16. Mô hình hiệu quả điều trị

Cho đến nay, tất cả các mô hình tuyến tính của chúng ta đã được áp dụng cho hai hoặc nhiều biến ngẫu nhiên liên tục. Chúng ta giả định các biến ngẫu nhiên là đa biến phân phối chuẩn và sử dụng điều này để thúc đẩy một mô hình tuyến tính. Phương pháp này bao gồm nhiều ví dụ thực tế về hồi quy tuyến tính. Tuy nhiên, mô hình tuyến tính còn có nhiều ứng dụng khác. Một trong những ứng dụng phổ biến nhất là đo lường ảnh hưởng của liệu pháp trong các thử nghiệm ngẫu nhiên và kiểm soát. Một trong những ứng dụng đầu tiên là trong nông nghiệp, nơi các miếng đất khác nhau được xử lý bằng các kết hợp khác nhau của phân bón để xem liệu chúng có hiệu quả hay không. Trong thực tế, việc sử dụng cho kết quả trong Thống kê xuất phát từ việc phát triển lý thuyết toán học cho sản lượng mùa vụ làm kết quả.

Kể từ đó, các ý tưởng tương tự đã được áp dụng trong các lĩnh vực khác như thử nghiệm ngẫu nhiên phát triển để xác định liệu thuốc có chữa trị hoặc ngăn chặn bệnh không, hoặc liệu chính sách có ảnh hưởng đến kết quả xã hội hoặc giáo dục không. Trong ví dụ sau, chúng ta nghĩ về can thiệp chính sách như một liệu pháp và tuân theo quy trình toán học tương tự. Các phân tích được sử dụng trong thử nghiệm A/B, rất phổ biến ngày nay trong các công ty internet, dựa trên mô hình ảnh hưởng của liệu pháp. Hơn nữa, việc sử dụng những mô hình này đã được mở rộng sang các nghiên cứu quan sát nơi các nhà phân tích cố gắng sử dụng mô hình tuyến tính để ước lượng các ảnh hưởng mong muốn trong khi tính đến các yếu tố gây nhiễu tiềm ẩn. Ví dụ, để ước lượng ảnh hưởng của chế độ dinh dưỡng giàu trái cây và rau củ đối với huyết áp, chúng ta cần điều chỉnh cho các yếu tố như tuổi, giới tính và tình trạng hút thuốc.

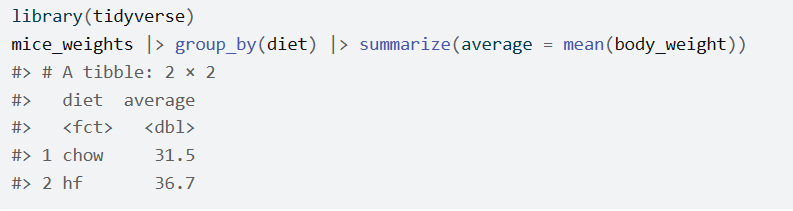
Trong chương này, chúng ta xem xét một thí nghiệm kiểm tra ảnh hưởng của chế độ ăn giàu chất béo đối với sinh lý của chuột. Chuột được chọn và chia ngẫu nhiên thành hai nhóm, một nhóm nhận chế độ ăn giàu chất béo được coi là liệu pháp và nhóm khác để lại làm nhóm kiểm soát và nhận chế độ ăn thông thường. Dữ liệu được bao gồm trong gói dslabs:



Một biểu đồ boxplot cho thấy rằng những con chuột được cho chế độ ăn giàu chất béo, trung bình, nặng hơn.



Các trung bình mẫu cho hai nhóm, chế độ ăn giàu chất béo và chế độ ăn thông thường, là khác nhau:



## 16.1 Thiết kế một yếu tố

Mặc dù kiểm định t tốt cho các trường hợp chỉ có hai liệu pháp, thường xuyên có các biến khác ảnh hưởng đến kết quả của chúng ta. Mô hình tuyến tính cho phép kiểm tra giả thuyết trong các tình huống tổng quát hơn. Chúng tôi bắt đầu mô tả cách sử dụng mô hình tuyến tính để ước lượng hiệu ứng điều trị bằng cách minh họa cách chúng có thể được sử dụng để thực hiện kiểm định t.

Nếu chúng ta giả định rằng phân phối trọng lượng cho cả chế độ ăn thông thường và chế độ ăn giàu chất béo đều tuân theo phân phối chuẩn, chúng ta có thể viết mô hình tuyến tính sau để biểu diễn dữ liệu:



với1 nếu con chuột thứ

được nuôi chế độ ăn giàu chất béo và trong trường hợp ngược lại, và các lỗi e i

​ là độc lập và tuân theo phân phối chuẩn với giá trị kỳ vọng bằng 0 và độ lệch chuẩn là

σ. Lưu ý rằng công thức toán học này trông hoàn toàn giống như mô hình mà chúng ta đã viết cho chiều cao cha-con trai. Tuy nhiên, việcx igiờ đây là 0 hoặc 1 thay vì là biến liên tục cho phép chúng ta sử dụng nó trong ngữ cảnh khác nhau này. Đặc biệt, hãy chú ý rằng bây giờβ

đại diện cho chiều cao trung bình của chuột trên chế độ ăn thông thường và β

​đại diện cho chiều cao trung bình của chuột trên chế độ ăn giàu chất béo.

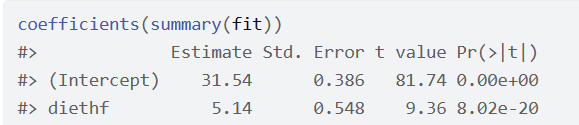
Một đặc điểm tốt của mô hình này là β

đại diện cho hiệu ứng điều trị của việc nhận chế độ ăn giàu chất béo. Nếu giả thuyết không có tác động của chế độ ăn giàu chất béo có thể được định lượng là β =0, chúng ta sau đó có thể ước lượng β và trả lời câu hỏi liệu sự khác biệt quan sát được có thực sự là đáng tin cậy hay không bằng cách tính toán ước lượng này so với giả thuyết. Vậy làm thế nào để ước lượng β và một độ lệch chuẩn cho ước lượng?

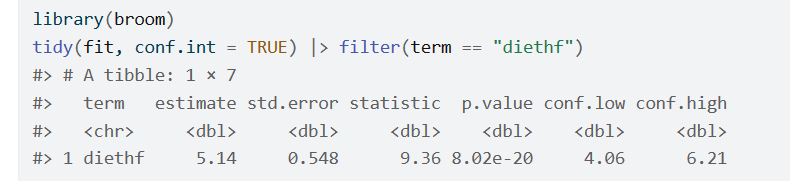
Một đặc điểm mạnh mẽ của mô hình tuyến tính là chúng ta có thể ước lượng các tham số β và độ lệch chuẩn của chúng với cùng một bộ máy LSE:



Vì chế độ ăn là một yếu tố có hai giá trị, hàm lm biết cách điều chỉnh mô hình trên với một biến chỉ số. Hàm summary cho chúng ta thấy các ước lượng, độ lệch chuẩn và giá trị p kết quả:



Hoặc sử dụng chúng ta có thể viết:broom



## 16.2 Thiết kế 2 yếu tố

Lưu ý rằng thí nghiệm này bao gồm cả chuột đực và cái, và chuột đực được biết là nặng hơn. Điều này giải thích tại sao các dư thừa phụ thuộc vào biến giới tính:

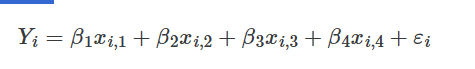


Sự sai mô tả này có thể có tác động thực sự vì nếu có nhiều chuột đực hơn nhận chế độ ăn giàu chất béo, điều này có thể giải thích sự tăng cường. Hoặc nếu có ít hơn nhận nó, thì chúng ta có thể đánh giá thấp hiệu ứng của chế độ ăn. Giới tính có thể là một biến gây nhiễu. Mô hình của chúng ta nhất định có thể được cải thiện.

Từ việc xem xét dữ liệu:

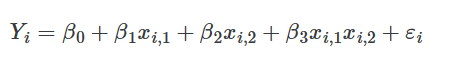


Chúng ta thấy rằng hiệu ứng của chế độ ăn được quan sát cho cả hai giới và rằng chuột đực nặng hơn chuột cái. Mặc dù không rõ ràng nhưng có vẻ hiệu ứng của chế độ ăn mạnh mẽ hơn ở chuột đực. Một mô hình tuyến tính cho phép giá trị kỳ vọng khác nhau cho bốn nhóm, 1) cái trên chế độ ăn thông thường, 2) cái trên chế độ ăn giàu chất béo, 3) đực trên chế độ ăn thông thường và 4) đực trên chế độ ăn giàu chất béo, có thể được biểu diễn như sau:



với các biến chỉ số xi1,xi2,xi3 cho mỗi trong bốn nhóm. Tuy nhiên, với biểu diễn này, không có β nào đại diện cho hiệu ứng quan trọng: hiệu ứng của chế độ ăn. Hơn nữa, chúng ta bây giờ đang xem xét khả năng rằng hiệu ứng của chế độ ăn có thể khác nhau đối với đực và cái và có thể kiểm tra giả thuyết đó luôn.

Một đặc điểm mạnh mẽ của mô hình tuyến tính là chúng ta có thể viết lại mô hình sao cho chúng ta vẫn có giá trị kỳ vọng khác nhau cho mỗi nhóm, nhưng các tham số đại diện cho những ảnh hưởng mà chúng ta quan tâm. Ví dụ, trong biểu diễn:



với xi1 là biến chỉ số cho việc bạn có điều trị và xi2 là biến chỉ số cho việc bạn là đực, các β có thể được giải thích như sau:

​

β1 : hiệu ứng điều trị cho cái

​

β2 : sự khác biệt trọng lượng giữa đực và cái khi không có điều trị

​

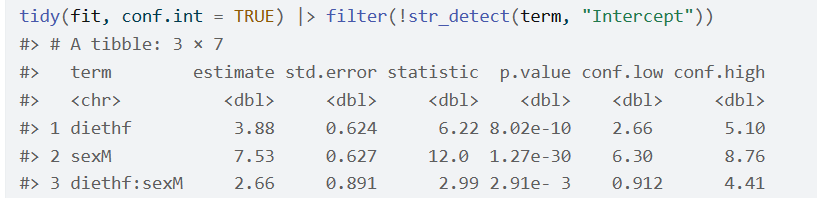
β3 : hiệu ứng điều trị bổ sung cho đực

Trong Thống kê, hiệu ứng cuối cùng này được xem xét là một hiệu ứng tương tác. Giá trị cơ sở (β0) được coi là giá trị trung bình cơ sở, là trung bình trọng lượng của cái trên chế độ ăn thông thường.

Sách thống kê mô tả nhiều cách khác nhau để viết lại mô hình để đạt được các loại giải thích khác nhau. Ví dụ, chúng ta có thể muốn β đại diện cho hiệu ứng điều trị trung bình giữa cái và đực, thay vì các hiệu ứng điều trị của cái. Điều này được đạt được bằng cách xác định rõ chúng ta quan tâm đến những so sánh nào.

Trong R, chúng ta có thể chỉ định mô hình này bằng cách sử dụng công thức sau:

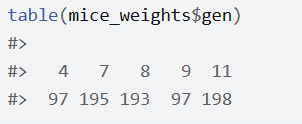




## 16.3 Phân tích phương sai

Trong ví dụ chúng ta đã xem xét, mỗi liệu pháp chỉ có hai cấp: chế độ ăn thông thường và chế độ ăn giàu chất béo, giới tính có cái và đực. Tuy nhiên, thường xuyên chúng ta có các biến quan trọng có nhiều hơn một cấp. Ví dụ, chúng ta có thể đã thử nghiệm một chế độ ăn thứ ba cho chuột. Trong sách thống kê, những biến này được gọi là yếu tố. Trong những trường hợp như vậy, thường muốn biết không chỉ về ảnh hưởng của mỗi cấp của yếu tố, mà còn muốn biết về một lượng tổng quát hơn đối với biến thiên qua các cấp. Phân tích phương sai hoặc ANOVA chính làm điều này. Tóm lược được sử dụng để đo lường sự biến thiên của một yếu tố là sai số trung bình của ước lượng ảnh hưởng của mỗi cấp.

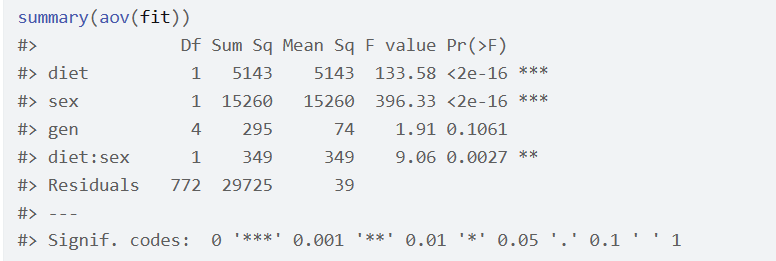
Ví dụ, giả sử rằng chuột trong tập dữ liệu của chúng ta thực sự thuộc vào nhiều thế hệ khác nhau:



Chúng ta có thể điều chỉnh một mô hình tuyến tính mà điều chỉnh một ảnh hưởng cho mỗi thế hệ này cùng với mô hình về chế độ ăn và giới tính đã được điều chỉnh trước đó:



We can then perform an analysis of variance with the R aov function:



Phân tích này cho thấy rằng phần lớn sự biến động được giải thích bởi giới tính và sau đó là chế độ ăn. Yếu tố thế hệ giải thích rất ít biến động so với hai yếu tố trước và không được xác định là có ý nghĩa thống kê.

# 17. Bài kiểm tra liên kết

Các mô hình thống kê được nghiên cứu cho đến nay là phù hợp cho kết quả liên tục. Chúng tôi chưa thảo luận về suy luận cho dữ liệu nhị phân, phân loại và thứ tự. Để đưa ra một ví dụ rất cụ thể, chúng tôi sẽ xem xét một nghiên cứu điển hình kiểm tra tỷ lệ tài trợ thành công ở Hà Lan, theo giới tính.

## 17.1 Nghiên cứu điển hình: tỷ lệ tài trợ thành công

Một bài báo PNAS năm 20141 đã phân tích tỷ lệ thành công từ các cơ quan tài trợ ở Hà Lan và kết luận rằng:

Kết quả cho thấy sự thiên lệch giới tính ưa thích ứng viên nam hơn ứng viên nữ trong việc ưu tiên đánh giá "chất lượng của người nghiên cứu" (nhưng không phải "chất lượng của đề xuất") và tỷ lệ thành công, cũng như trong cách sử dụng ngôn ngữ trong tài liệu hướng dẫn và đánh giá.

Bằng chứng chính cho kết luận này là so sánh tỷ lệ phần trăm. Bảng S1 trong bài báo bao gồm thông tin chúng ta cần. Dưới đây là ba cột hiển thị kết quả tổng thể:

[library](https://rdrr.io/r/base/library.html)([tidyverse](https://tidyverse.tidyverse.org/))

[library](https://rdrr.io/r/base/library.html)(dslabs)

research\_funding\_rates |> [select](https://dplyr.tidyverse.org/reference/select.html)(discipline, applications\_total,

success\_rates\_total) |> [head](https://rdrr.io/r/utils/head.html)()

#> discipline applications\_total success\_rates\_total

#> 1 Chemical sciences 122 26.2

#> 2 Physical sciences 174 20.1

#> 3 Physics 76 26.3

#> 4 Humanities 396 16.4

#> 5 Technical sciences 251 17.1

#> 6 Interdisciplinary 183 15.8

Chúng tôi có những giá trị này cho từng giới tính:

[names](https://rdrr.io/r/base/names.html)(research\_funding\_rates)

#> [1] "discipline" "applications\_total" "applications\_men"

#> [4] "applications\_women" "awards\_total" "awards\_men"

#> [7] "awards\_women" "success\_rates\_total" "success\_rates\_men"

#> [10] "success\_rates\_women"

Chúng ta có thể tính tổng số thành công và tổng số không như sau:

totals <- research\_funding\_rates |>

[select](https://dplyr.tidyverse.org/reference/select.html)(-discipline) |>

[summarize\_all](https://dplyr.tidyverse.org/reference/summarise_all.html)(sum) |>

[summarize](https://dplyr.tidyverse.org/reference/summarise.html)(yes\_men = awards\_men,

no\_men = applications\_men - awards\_men,

yes\_women = awards\_women,

no\_women = applications\_women - awards\_women)

Vì vậy, chúng ta thấy rằng một tỷ lệ lớn hơn nam giới so với phụ nữ nhận được giải thưởng:

totals |> [summarize](https://dplyr.tidyverse.org/reference/summarise.html)(percent\_men = yes\_men/(yes\_men+no\_men),

percent\_women = yes\_women/(yes\_women+no\_women))

#> percent\_men percent\_women

#> 1 0.177 0.149

Nhưng điều này có thể chỉ là do sự biến đổi ngẫu nhiên? Ở đây chúng ta tìm hiểu cách thực hiện suy luận cho loại dữ liệu này.

## 17.3 Quý bà nếm trà

R.A Fisher là một trong những người đầu tiên chính thức hóa thử nghiệm giả thuyết. "Lady Tasting Tea" là một trong những ví dụ nổi tiếng nhất.

Câu chuyện như sau: một người quen của Fisher tuyên bố rằng cô ấy có thể biết sữa được thêm vào trước hay sau khi trà được rót. Fisher tỏ ra hoài nghi. Ông đã thiết kế một thí nghiệm để kiểm tra tuyên bố này. Anh đưa cho cô bốn cặp trà: một tách với sữa rót trước, một sau đó. Thứ tự được sắp xếp ngẫu nhiên. Giả thuyết vô giá trị ở đây là cô ấy đang đoán. Fisher rút ra sự phân phối cho số lượng lựa chọn chính xác với giả định rằng các lựa chọn là ngẫu nhiên và độc lập.

Ví dụ, giả sử cô ấy chọn đúng 3 trong số 4. Chúng ta có tin rằng cô ấy có một khả năng đặc biệt không? Câu hỏi cơ bản mà chúng tôi đặt ra là: nếu người kiểm tra thực sự đoán, cơ hội mà cô ấy nhận được 3 điểm trở lên là bao nhiêu? Giống như chúng ta đã làm trước đây, chúng ta có thể tính toán một xác suất theo giả thuyết null rằng cô ấy đang đoán 4 trong số mỗi xác suất. Theo giả thuyết null này, chúng ta có thể nghĩ về ví dụ cụ thể này như chọn 4 quả bóng ra khỏi một chiếc bình với 4 quả bóng màu xanh lam (câu trả lời đúng) và 4 quả bóng màu đỏ (câu trả lời sai). Hãy nhớ rằng, cô ấy biết rằng có bốn trước khi uống trà và bốn sau đó.

Theo giả thuyết vô giá trị rằng cô ấy chỉ đơn giản là đoán, mỗi quả bóng đều có cơ hội được chọn như nhau. Sau đó, chúng ta có thể sử dụng các kết hợp để tìm ra từng xác suất. Xác suất chọn 3 là = 16/70 Xác suất chọn đúng cả 4 là = 1/70. Do đó, cơ hội quan sát 3 hoặc một cái gì đó cực đoan hơn, theo giả thuyết null, là Đây là giá trị p. Quy trình tạo ra giá trị p này được gọi là thử nghiệm chính xác của Fisher và nó sử dụng phân phối siêu hình học.

## 17.3 Hai bảng cùng một lúc

Dữ liệu từ thí nghiệm thường được tóm tắt bằng một bảng như thế này:

tab <- [matrix](https://rdrr.io/r/base/matrix.html)([c](https://rdrr.io/r/base/c.html)(3,1,1,3),2,2)

[rownames](https://rdrr.io/r/base/colnames.html)(tab)<-[c](https://rdrr.io/r/base/c.html)("Poured Before","Poured After")

[colnames](https://rdrr.io/r/base/colnames.html)(tab)<-[c](https://rdrr.io/r/base/c.html)("Guessed before","Guessed after")

tab

#> Guessed before Guessed after

#> Poured Before 3 1

#> Poured After 1 3

Chúng được gọi là bảng 2x2. Đối với mỗi kết hợp trong số bốn kết hợp, người ta có thể nhận được với một cặp biến nhị phân, chúng hiển thị số lượng quan sát được cho mỗi lần xuất hiện.

Hàm *fisher.test* thực hiện các phép tính suy luận ở trên:

[fisher.test](https://rdrr.io/r/stats/fisher.test.html)(tab, alternative="greater")$p.value

#> [1] 0.243

## 17.4 Thử nghiệm chi bình phương

Lưu ý rằng, theo một cách nào đó, ví dụ về tỷ lệ tài trợ của chúng tôi tương tự như Lady Tasting Tea. Tuy nhiên, trong ví dụ Lady Tasting Tea, số lượng hạt màu xanh và đỏ được cố định bằng thực nghiệm và số lượng câu trả lời được đưa ra cho mỗi danh mục cũng được cố định. Điều này là do Fisher đảm bảo có bốn cốc sữa được rót trước khi trà và bốn cốc sữa rót sau và người phụ nữ biết điều này, vì vậy câu trả lời cũng sẽ phải bao gồm bốn trước và bốn sau. Nếu đúng như vậy, tổng của các hàng và tổng của các cột là cố định. Điều này xác định các ràng buộc về các cách có thể chúng ta có thể điền vào hai bảng bằng hai và cũng cho phép chúng ta sử dụng phân phối siêu hình học. Nói chung, đây không phải là trường hợp. Tuy nhiên, có một cách tiếp cận khác, thử nghiệm Chi-bình phương, được mô tả dưới đây.

Hãy tưởng tượng chúng ta có 290, 1.345, 177, 1.011 ứng viên, một số là nam giới và một số là phụ nữ và một số được tài trợ, trong khi những người khác thì không. Chúng tôi thấy rằng tỷ lệ thành công của nam giới và phụ nữ là:

totals |> [summarize](https://dplyr.tidyverse.org/reference/summarise.html)(percent\_men = yes\_men/(yes\_men+no\_men),

percent\_women = yes\_women/(yes\_women+no\_women))

#> percent\_men percent\_women

#> 1 0.177 0.149

Tương ứng. Chúng ta sẽ thấy điều này một lần nữa nếu chúng ta chỉ định ngẫu nhiên tài trợ theo tỷ lệ tổng thể:

rate <- [with](https://rdrr.io/r/base/with.html)(totals, (yes\_men + yes\_women))/[sum](https://rdrr.io/r/base/sum.html)(totals)

rate

#> [1] 0.165

Bài kiểm tra Chi-bình phương trả lời câu hỏi này. Bước đầu tiên là tạo bảng dữ liệu 2x2:

two\_by\_two <- [with](https://rdrr.io/r/base/with.html)(totals, [data.frame](https://rdrr.io/r/base/data.frame.html)(awarded = [c](https://rdrr.io/r/base/c.html)("no", "yes"),

men = [c](https://rdrr.io/r/base/c.html)(no\_men, yes\_men),

women = [c](https://rdrr.io/r/base/c.html)(no\_women, yes\_women)))

two\_by\_two

#> awarded men women

#> 1 no 1345 1011

#> 2 yes 290 177

Ý tưởng chung của bài kiểm tra Chi-bình phương là so sánh bảng 2x2 này với những gì bạn mong đợi sẽ thấy, đó sẽ là:

[with](https://rdrr.io/r/base/with.html)(totals, [data.frame](https://rdrr.io/r/base/data.frame.html)(awarded = [c](https://rdrr.io/r/base/c.html)("no", "yes"),

men = (no\_men + yes\_men) \* [c](https://rdrr.io/r/base/c.html)(1 - rate, rate),

women = (no\_women + yes\_women) \* [c](https://rdrr.io/r/base/c.html)(1 - rate, rate)))

#> awarded men women

#> 1 no 1365 991

#> 2 yes 270 197

Chúng ta có thể thấy rằng nhiều nam giới hơn dự kiến và ít phụ nữ hơn dự kiến nhận được tài trợ. Tuy nhiên, theo giả thuyết null, những quan sát này là các biến ngẫu nhiên. Thử nghiệm Chi-square cho chúng ta biết khả năng nhìn thấy độ lệch lớn hoặc lớn hơn như thế nào. Thử nghiệm này sử dụng kết quả tiệm cận, tương tự như CLT, liên quan đến tổng kết quả nhị phân độc lập. Hàm R *chisq.test* lấy bảng 2x2 và trả về kết quả từ bài kiểm tra:

chisq\_test <- [chisq.test](https://rdrr.io/r/stats/chisq.test.html)(two\_by\_two[, -1])

Ta sẽ thấy p\_value = 0.0509:

chisq\_test$p.value

#> [1] 0.0509

## 17.5 Mô hình tuyến tính tổng quát

Chúng tôi đã trình bày một cách để thực hiện kiểm tra giả thuyết để xác định xem có mối liên hệ giữa hai kết quả nhị phân hay không. Nhưng chúng tôi chưa mô tả làm thế nào để định lượng hiệu ứng. Chúng ta có thể ước tính ảnh hưởng của việc trở thành phụ nữ trong việc tài trợ thành công ở Hà Lan không? Lưu ý rằng nếu kết quả của chúng ta là nhị phân, thì các mô hình tuyến tính được trình bày trong Chương 16 không phù hợp vì và liên tục. Tuy nhiên, việc điều chỉnh các phương pháp này, được sử dụng rộng rãi, chẳng hạn như trong nghiên cứu y học, cho chúng ta một cách để ước tính tác động cùng với sai số chuẩn của chúng.

Ý tưởng là mô hình hóa sự chuyển đổi giá trị kỳ vọng của kết quả bằng mô hình tuyến tính. Việc chuyển đổi được chọn sao cho bất kỳ giá trị liên tục nào cũng có thể. Phương trình toán học cho một mô hình có một biến trông như thế này:

Để kết thúc mô tả mô hình, chúng tôi áp đặt một phân phối trên *Y* chẳng hạn như nhị thức hoặc Poisson. Chúng được gọi là \_generalized mô hình tuyến tính.

Chúng ta chứng minh với ví dụ về tỷ lệ tài trợ. Ta xác định sẽ là 1 nếu người nhận được tài trợ và nếu không thì 0 và sẽ là 1 dành cho người khi họ là phụ nữ, 0 khi họ là đàn ông. Đối với dữ liệu mong đợi là xác suất tài trợ của người Chúng ta cho rằng giá trị xuất ra là nhị thức với N = 1 và xác suất . Với dữ liệu nhị thức, phép chuyển đổi được sữ dụng rộng rãi nhất là hàm LOGIT thứ sẽ cho ra giá trị số giữa 0 và 1 sang bất kì số liên tục nào. Mô hình sẽ có dạng:

### 17.5.1 Tỷ lệ chênh lệch

Để hiểu được tại sao có thể sử dụng để định lượng sự ảnh hưởng của việc trở thành phụ nữ với tỉ lệ thành công, lưu ý đầu rằng là một chênh lệch của người khi nhận được tài trợ tỷ lệ xác suất thành công và xác suất thất bại. Điều này ngụ ý rằng

là tỷ lệ cược cho nam giới và là tỷ lệ cược cho phụ nữ, ngụ ý là tỷ lệ cược cho phụ nữ chia cho tỷ lệ cược cho nam giới. Số lượng này được gọi là tỷ lệ chênh lệch. Để thấy điều này không phải là nếu sử dụng và để biểu thị xác suất thành công cho phụ nữ và nam giới, tương ứng, sau đó có thể được viết lại là:

do đó định lượng tỷ lệ chênh lệch log.

Bây giờ làm thế nào để chúng ta ước tính các tham số này? Mặc dù các chi tiết không được mô tả trong cuốn sách này, bình phương tối thiểu không còn là cách tối ưu để ước tính các tham số và thay vào đó chúng tôi sử dụng một phương pháp gọi là ước tính khả năng tối đa (MLE). Các dẫn xuất toán học tiên tiến hơn cho thấy một phiên bản của định lý giới hạn trung tâm được áp dụng và các ước tính thu được theo cách này là xấp xỉ bình thường khi số lượng quan sát lớn. Lý thuyết cũng cung cấp một cách để tính toán các lỗi tiêu chuẩn cho các ước tính của .

### 17.5.2 Sắp đặt mô hình

Để có được các ước tính khả năng tối đa bằng cách sử dụng R, chúng ta có thể sử dụng chức năng *glm* với đối số *family* đặt thành *binomial* . Điều này mặc định sử dụng chuyển đổi logit. Lưu ý rằng chúng tôi không có dữ liệu cấp độ cá nhân, nhưng vì mô hình của chúng tôi giả định xác suất thành công là như nhau đối với tất cả phụ nữ và tất cả nam giới, nên số lượng thành công có thể được mô hình hóa là nhị thức với

thử nghiệm và xác suất cho phụ nữ và nhị thức với Thử nghiệm và xác suất cho nam giới, với và tổng số phụ nữ và nam giới. Trong trường hợp này, hàm *glm* được sử dụng như thế này:

success <- [with](https://rdrr.io/r/base/with.html)(totals, [c](https://rdrr.io/r/base/c.html)(yes\_men, yes\_women))

failure <- [with](https://rdrr.io/r/base/with.html)(totals, [c](https://rdrr.io/r/base/c.html)(no\_men, no\_women))

gender <- [factor](https://rdrr.io/r/base/factor.html)([c](https://rdrr.io/r/base/c.html)("men", "women"))

fit <- [glm](https://rdrr.io/r/stats/glm.html)([cbind](https://rdrr.io/r/base/cbind.html)(success, failure) ~ gender, family = "binomial")

[coefficients](https://rdrr.io/r/stats/coef.html)([summary](https://rdrr.io/r/base/summary.html)(fit))

#> Estimate Std. Error z value Pr(>|z|)

#> (Intercept) -1.534 0.0647 -23.7 3.83e-124

#> genderwomen -0.208 0.1041 -2.0 4.54e-02

Ước tính tỷ lệ chênh là 0,811982, được hiểu là tỷ lệ cược giảm 20% đối với phụ nữ so với nam giới. Nhưng điều này có phải do cơ hội? Chúng tôi đã lưu ý rằng giá trị p là khoảng 0,05, nhưng phương pháp GLM cũng cho phép chúng tôi tính toán khoảng tin cậy bằng cách sử dụng hàm *confint*. Để hiển thị khoảng thời gian cho tỷ lệ chênh dễ hiểu hơn, chúng ta chỉ cần cấp số nhân:

[exp](https://rdrr.io/r/base/Log.html)([confint](https://rdrr.io/r/stats/confint.html)(fit, 2))

#> 2.5 % 97.5 %

#> 0.661 0.995

### 17.5.3 Xấp xỉ sai số tiêu chuẩn đơn giản cho tỷ lệ chênh lệch bảng hai x hai

Việc sử dụng *glm* giúp chúng tôi có thể thu được ước tính, lỗi tiêu chuẩn và khoảng tin cậy cho một loạt các mô hình. Để làm điều này, chúng tôi sử dụng một thuật toán khá phức tạp. Trong trường hợp 2x2, chúng ta có thể nhận được lỗi tiêu chuẩn cho tỷ lệ chênh lệch nhật ký bằng cách sử dụng một xấp xỉ đơn giản.

Nếu bảng 2x2 của chúng tôi có các mục sau:

|  |  |  |
| --- | --- | --- |
|  | Men | Women |
| Awarded | a | b |
| Not Awarded | c | d |

Trong trường hợp này, giá trị chênh lệch đơn giản là . Chúng tôi có thể xác nhận rằng chúng tôi có được ước tính tương tự như khi sử dụng *glm*:

or <- [with](https://rdrr.io/r/base/with.html)(two\_by\_two, women[2]/[sum](https://rdrr.io/r/base/sum.html)(women) / (women[1]/[sum](https://rdrr.io/r/base/sum.html)(women)) / ((men[2]/[sum](https://rdrr.io/r/base/sum.html)(men)) / (men[1]/[sum](https://rdrr.io/r/base/sum.html)(men))))

[c](https://rdrr.io/r/base/c.html)([log](https://rdrr.io/r/base/Log.html)(or), fit$coef[2])

#> genderwomen

#> -0.208 -0.208

Lý thuyết thống kê cho chúng ta biết rằng khi cả bốn mục của bảng hai x hai đủ lớn, thì tỷ lệ chênh lệch nhật ký xấp xỉ bình thường với sai số chuẩn

Điều này ngụ ý rằng khoảng tin cậy 95% cho tỷ lệ chênh lệch nhật ký có thể được hình thành bởi:

Bằng cách lũy thừa hai con số này, chúng ta có thể xây dựng một khoảng tin cậy của tỷ lệ chênh lệch.

Sử dụng R, chúng ta có thể tính khoảng tin cậy này như sau:

se <- two\_by\_two |> [select](https://dplyr.tidyverse.org/reference/select.html)(-awarded) |>

[summarize](https://dplyr.tidyverse.org/reference/summarise.html)(se = [sqrt](https://rdrr.io/r/base/MathFun.html)([sum](https://rdrr.io/r/base/sum.html)(1/men) + [sum](https://rdrr.io/r/base/sum.html)(1/women))) |>

[pull](https://dplyr.tidyverse.org/reference/pull.html)(se)

[exp](https://rdrr.io/r/base/Log.html)([log](https://rdrr.io/r/base/Log.html)(or) + [c](https://rdrr.io/r/base/c.html)(-1,1) \* [qnorm](https://rdrr.io/r/stats/Normal.html)(0.975) \* se)

#> [1] 0.662 0.996

Lưu ý rằng 1 không được bao gồm trong khoảng tin cậy, điều đó có nghĩa là giá trị p nhỏ hơn 0,05. Chúng tôi có thể xác nhận điều này bằng cách sử dụng:

2\*(1 - [pnorm](https://rdrr.io/r/stats/Normal.html)([abs](https://rdrr.io/r/base/MathFun.html)([log](https://rdrr.io/r/base/Log.html)(or)), 0, se))

#> [1] 0.0454

## 17.6 Lấy mẫu giá trị lớn, giá trị p nhỏ

Như đã đề cập trước đó, chỉ báo cáo giá trị p không phải là một cách thích hợp để báo cáo kết quả phân tích dữ liệu. Ví dụ, trong các tạp chí khoa học, một số nghiên cứu dường như nhấn mạnh quá mức các giá trị p. Một số nghiên cứu này có kích thước mẫu lớn và báo cáo giá trị p nhỏ ấn tượng. Tuy nhiên, khi nhìn kỹ vào kết quả, chúng ta nhận ra tỷ lệ chênh khá khiêm tốn: hầu như không lớn hơn 1. Trong trường hợp này, sự khác biệt có thể không có ý nghĩa thực tế hoặc có ý nghĩa khoa học.

Lưu ý rằng mối quan hệ giữa tỷ lệ chênh và giá trị p không phải là một-một. Nó phụ thuộc vào kích thước mẫu. Vì vậy, giá trị p rất nhỏ không nhất thiết có nghĩa là tỷ lệ chênh rất lớn. Lưu ý những gì xảy ra với giá trị p nếu chúng ta nhân bảng hai với hai với 10, điều này không làm thay đổi tỷ lệ chênh lệch:

two\_by\_two\_x\_10 <- two\_by\_two |>

[select](https://dplyr.tidyverse.org/reference/select.html)(-awarded) |>

[mutate](https://dplyr.tidyverse.org/reference/mutate.html)(men = men\*10, women = women\*10)

[chisq.test](https://rdrr.io/r/stats/chisq.test.html)(two\_by\_two\_x\_10)$p.value

#> [1] 2.63e-10

Lưu ý rằng tỷ lệ chênh lệch nhật ký không được xác định nếu bất kỳ ô nào của bảng hai x hai là 0. Điều này là do nếu , hoặc *d* là 0, các là nhật ký của 0 hoặc có 0 trong mẫu số. Đối với tình huống này, thông thường tránh số 0 bằng cách thêm 0,5 vào mỗi ô. Điều này được gọi là hiệu chỉnh Haldane-Anscombe và đã được chứng minh, cả trong thực tế và lý thuyết, hoạt động tốt.

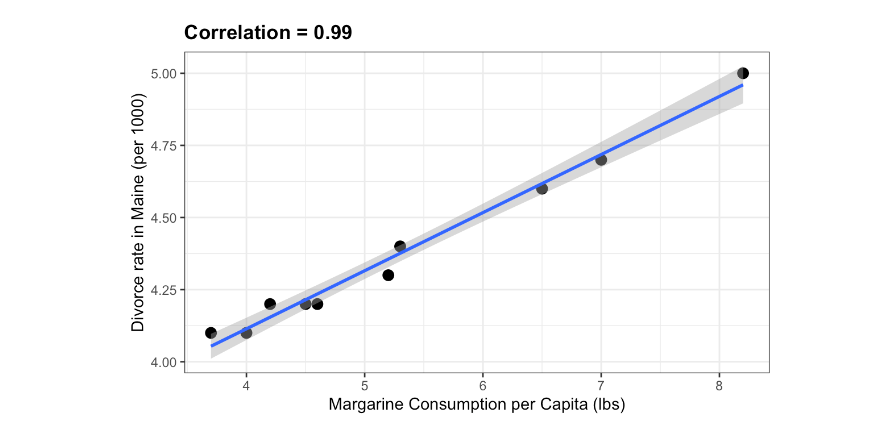
# 18 Hiểu rõ mối liên kết không đồng nghĩa với quan hệ nhân quả

Hiểu rõ mối liên kết không đồng nghĩa với quan hệ nhân quả có lẽ là bài học quan trọng nhất mà mọi người học được trong một lớp học thống kê. "Tương quan không đồng nghĩa với nhân quả" là cách khác để diễn đạt điều này. Trong phần Thống kê của cuốn sách, chúng ta đã mô tả các công cụ hữu ích để đo lường mối quan hệ giữa các biến. Tuy nhiên, chúng ta phải cẩn thận để không quá mức giải thích quá mức về những mối quan hệ này.

Có nhiều lý do mà một biến *X* có thể có mối quan hệ với một biến *Y* mà không có bất kỳ ảnh hưởng trực tiếp nào đối với biến *Y*. Dưới đây, chúng ta sẽ xem xét bốn cách phổ biến có thể dẫn đến việc hiểu lầm dữ liệu.

## 18.1 Tương quan giả mạo

Ví dụ hài hước sau đây làm nổi bật điều rằng tương quan không đồng nghĩa với nhân quả. Nó chỉ ra một mối quan hệ rất mạnh mẽ giữa tỷ lệ ly hôn và việc tiêu thụ bơ.



Điều này có nghĩa là bơ gây ra việc ly hôn không? Hoặc liệu việc ly hôn làm người ta ăn nhiều bơ hơn không? Tất nhiên, câu trả lời cho cả hai câu hỏi này đều là không. Đây chỉ là một ví dụ về những gì chúng ta gọi là một tương quan giả mạo.

Bạn có thể thấy nhiều ví dụ ngớ ngẩn hơn trên trang web về Tương Quan Giả Mạo.

Những trường hợp được trình bày trên trang web tương quan giả mạo đều là các trường hợp của điều được gọi chung là khai thác dữ liệu, đánh cá dữ liệu hoặc theo dõi dữ liệu. Đây về cơ bản là một hình thức của những gì họ gọi là chọn lựa cherry ở Hoa Kỳ. Một ví dụ về khai thác dữ liệu sẽ là nếu bạn xem xét nhiều kết quả được tạo ra bởi một quy trình ngẫu nhiên và chọn ra kết quả hiển thị một mối quan hệ hỗ trợ một lý thuyết bạn muốn bảo vệ.

Một mô phỏng Monte Carlo có thể được sử dụng để thể hiện cách khai thác dữ liệu có thể dẫn đến việc tìm thấy các tương quan cao giữa các biến không tương quan. Chúng tôi sẽ lưu kết quả của mô phỏng của chúng tôi vào một tibble:

[library](https://rdrr.io/r/base/library.html)([tidyverse](https://tidyverse.tidyverse.org/))

N <- 25

g <- 1000000

sim\_data <- [tibble](https://tibble.tidyverse.org/reference/tibble.html)(group = [rep](https://rdrr.io/r/base/rep.html)(1:g, each = N),

x = [rnorm](https://rdrr.io/r/stats/Normal.html)(N\*g),

y = [rnorm](https://rdrr.io/r/stats/Normal.html)(N\*g))

Cột đầu tiên chỉ định nhóm. Chúng tôi đã tạo các nhóm và cho mỗi nhóm, chúng tôi tạo ra một cặp vector độc lập, *X* và *Y*, mỗi vector có 25 quan sát, được lưu trữ trong cột thứ hai và thứ ba. Vì chúng tôi xây dựng mô phỏng, chúng tôi biết rằng *X* và *Y* không có tương quan.

Tiếp theo, chúng tôi tính toán tương quan giữa X và Y cho mỗi nhóm và xem xét giá trị lớn nhất:

res <- sim\_data |>

[group\_by](https://dplyr.tidyverse.org/reference/group_by.html)(group) |>

[summarize](https://dplyr.tidyverse.org/reference/summarise.html)(r = [cor](https://rdrr.io/r/stats/cor.html)(x, y)) |>

[arrange](https://dplyr.tidyverse.org/reference/arrange.html)([desc](https://dplyr.tidyverse.org/reference/desc.html)(r))

res

#> # A tibble: 1,000,000 × 2

#> group r

#> <int> <dbl>

#> 1 606777 0.789

#> 2 949026 0.781

#> 3 752659 0.774

#> 4 815223 0.773

#> 5 890876 0.768

#> # ℹ 999,995 more rows

Chúng tôi thấy một tương quan tối đa là 0.7892932 và nếu bạn chỉ vẽ biểu đồ dữ liệu từ nhóm đạt được tương quan này, nó sẽ cho thấy một biểu đồ thuyết phục rằng *X* và *Y* thực sự có mối quan hệ tương quan:

sim\_data |> [filter](https://dplyr.tidyverse.org/reference/filter.html)(group == res$group[[which.max](https://rdrr.io/r/base/which.min.html)(res$r)]) |>

[ggplot](https://ggplot2.tidyverse.org/reference/ggplot.html)([aes](https://ggplot2.tidyverse.org/reference/aes.html)(x, y)) +

[geom\_point](https://ggplot2.tidyverse.org/reference/geom_point.html)() +

[geom\_smooth](https://ggplot2.tidyverse.org/reference/geom_smooth.html)(method = "lm")

#> `geom\_smooth()` using formula = 'y ~ x'

A graph with a line and dots

Description automatically generated

Hãy nhớ rằng tóm tắt về tương quan là một biến ngẫu nhiên. Đây là phân phối được tạo ra bởi mô phỏng Monte Carlo:

res |> [ggplot](https://ggplot2.tidyverse.org/reference/ggplot.html)([aes](https://ggplot2.tidyverse.org/reference/aes.html)(x=r)) + [geom\_histogram](https://ggplot2.tidyverse.org/reference/geom_histogram.html)(binwidth = 0.1, color = "black")

A graph of a graph with Willis Tower in the background

Description automatically generated

Đó chỉ là một sự thật toán học rằng nếu chúng ta quan sát các tương quan ngẫu nhiên được dự kiến là 0, nhưng có độ lệch chuẩn là 0.2041507, thì tương quan lớn nhất sẽ gần với 1.

Nếu chúng ta thực hiện phân tích hồi quy trên nhóm này và giải thích giá trị p, chúng ta sẽ sai lầm khi khẳng định rằng đây là một mối quan hệ có ý nghĩa thống kê:

[library](https://rdrr.io/r/base/library.html)([broom](https://broom.tidymodels.org/))

sim\_data |>

[filter](https://dplyr.tidyverse.org/reference/filter.html)(group == res$group[[which.max](https://rdrr.io/r/base/which.min.html)(res$r)]) |>

[summarize](https://dplyr.tidyverse.org/reference/summarise.html)([tidy](https://generics.r-lib.org/reference/tidy.html)([lm](https://rdrr.io/r/stats/lm.html)(y ~ x))) |>

[filter](https://dplyr.tidyverse.org/reference/filter.html)(term == "x")

#> # A tibble: 1 × 5

#> term estimate std.error statistic p.value

#> <chr> <dbl> <dbl> <dbl> <dbl>

#> 1 x 0.690 0.112 6.16 0.00000274

Hình thức cụ thể của việc khai thác dữ liệu này được gọi là p-hacking. P-hacking là một chủ đề được thảo luận nhiều vì nó là một vấn đề trong các xuất bản khoa học. Do các nhà xuất bản thường ưu tiên các kết quả có ý nghĩa thống kê hơn so với các kết quả tiêu cực, có động lực để báo cáo các kết quả quan trọng. Trong dịch tễ học và các ngành khoa học xã hội, ví dụ, các nhà nghiên cứu có thể tìm kiếm các mối quan hệ giữa một kết quả tiêu cực và một số yếu tố tác động và chỉ báo cáo yếu tố tác động nào đó mà kết quả trong giá trị p nhỏ. Hơn nữa, họ có thể thử nghiệm nhiều mô hình khác nhau để giải quyết vấn đề nhiễu loạn và chọn ra mô hình mà cho kết quả p nhỏ nhất. Trong các lĩnh vực thí nghiệm, một thí nghiệm có thể được lặp lại nhiều lần, nhưng chỉ kết quả của một thí nghiệm có giá trị p nhỏ được báo cáo. Điều này không nhất thiết xảy ra do hành vi không đạo đức, mà thay vào đó là do sự không hiểu biết về thống kê hoặc tư duy mong muốn. Trong các khóa học thống kê nâng cao, bạn có thể học các phương pháp để điều chỉnh cho những so sánh đa nguyên này.

## 18.2 Ngoại lai

Giả sử chúng ta thực hiện đo lường từ hai kết quả độc lập, *X* và *Y*, và chúng ta chuẩn hóa các đo lường này. Tuy nhiên, hãy tưởng tượng rằng chúng ta mắc một sai lầm và quên chuẩn hóa mục nhập thứ 23. Chúng ta có thể mô phỏng dữ liệu như vậy bằng cách sử dụng:

[set.seed](https://rdrr.io/r/base/Random.html)(1985)

x <- [rnorm](https://rdrr.io/r/stats/Normal.html)(100,100,1)

y <- [rnorm](https://rdrr.io/r/stats/Normal.html)(100,84,1)

x[-23] <- [scale](https://rdrr.io/r/base/scale.html)(x[-23])

y[-23] <- [scale](https://rdrr.io/r/base/scale.html)(y[-23])

Dữ liệu sẽ như thế này:

[qplot](https://ggplot2.tidyverse.org/reference/qplot.html)(x, y)

#> Warning: `qplot()` was deprecated in ggplot2 3.4.0.

A graph with lines and numbers

Description automatically generated

Không có gì đáng ngạc nhiên, mối tương quan rất cao:

[cor](https://rdrr.io/r/stats/cor.html)(x,y)

#> [1] 0.988

Nhưng điều này được thúc đẩy bởi một ngoại lai. Nếu chúng ta loại bỏ ngoại lai này, mối tương quan sẽ giảm đáng kể xuống gần như bằng 0, đó là những gì nó phải là:

[cor](https://rdrr.io/r/stats/cor.html)(x[-23], y[-23])

#> [1] -0.0442

Có một giải pháp thay thế cho mối tương quan mẫu để ước tính mối tương quan dân số mạnh mẽ với các ngoại lệ. Nó được gọi là tương quan Spearman. Ý tưởng rất đơn giản: tính toán mối tương quan trên thứ hạng của các giá trị. Dưới đây là một âm mưu của các cấp bậc được âm mưu chống lại nhau:

[qplot](https://ggplot2.tidyverse.org/reference/qplot.html)([rank](https://rdrr.io/r/base/rank.html)(x), [rank](https://rdrr.io/r/base/rank.html)(y))

A graph with black dots

Description automatically generated

Ngoại lai không còn được liên kết với một giá trị rất lớn và mối tương quan đi xuống:

[cor](https://rdrr.io/r/stats/cor.html)([rank](https://rdrr.io/r/base/rank.html)(x), [rank](https://rdrr.io/r/base/rank.html)(y))

#> [1] 0.00251

Tương quan Spearman cũng có thể được tính toán như thế này:

[cor](https://rdrr.io/r/stats/cor.html)(x, y, method = "spearman")

#> [1] 0.00251

Ngoài ra còn có các phương pháp để phù hợp mạnh mẽ với các mô hình tuyến tính mà bạn có thể tìm hiểu trong, ví dụ, cuốn sách này: Thống kê mạnh mẽ: Phiên bản 2 của Peter J. Huber &; Elvezio M. Ronchetti.

## 18.3 Đảo ngược nguyên nhân và kết quả

Một cách khác mà sự liên kết bị nhầm lẫn với quan hệ nhân quả là khi nguyên nhân và kết quả bị đảo ngược. Một ví dụ về điều này là tuyên bố rằng dạy kèm làm cho học sinh thực hiện kém hơn vì họ kiểm tra thấp hơn so với các đồng nghiệp không được dạy kèm. Trong trường hợp này, việc dạy kèm không gây ra điểm kiểm tra thấp, mà ngược lại.

Một hình thức của tuyên bố này thực sự đã biến nó thành một op-ed trên tờ New York Times có tiêu đề Sự tham gia của cha mẹ được đánh giá quá cao. Hãy xem xét trích dẫn này từ bài viết:

Khi chúng tôi kiểm tra xem việc giúp đỡ thường xuyên với bài tập về nhà có tác động tích cực đến kết quả học tập của trẻ em hay không, chúng tôi khá giật mình bởi những gì chúng tôi tìm thấy. Bất kể tầng lớp xã hội, chủng tộc hay dân tộc của một gia đình, hoặc cấp lớp của trẻ, bài tập về nhà nhất quán hầu như không bao giờ cải thiện điểm kiểm tra hoặc điểm số... Thậm chí đáng ngạc nhiên hơn đối với chúng tôi là khi cha mẹ thường xuyên giúp làm bài tập về nhà, trẻ em thường thực hiện kém hơn.

Một khả năng rất có thể là những đứa trẻ cần sự giúp đỡ thường xuyên của cha mẹ, nhận được sự giúp đỡ này vì chúng không học tốt ở trường.

Chúng ta có thể dễ dàng xây dựng một ví dụ về sự đảo ngược nguyên nhân và kết quả bằng cách sử dụng dữ liệu chiều cao của cha và con trai. Nếu chúng tôi phù hợp với mô hình:

đến dữ liệu chiều cao của cha và con trai, với

chiều cao của người cha và chiều cao con trai, chúng tôi nhận được một kết quả có ý nghĩa thống kê. Chúng tôi sử dụng *galton\_heights* tập dữ liệu quy định tại Chương 13:

galton\_heights |> [summarize](https://dplyr.tidyverse.org/reference/summarise.html)([tidy](https://generics.r-lib.org/reference/tidy.html)([lm](https://rdrr.io/r/stats/lm.html)(father ~ son)))

#> Warning: Returning more (or less) than 1 row per `summarise()` group was

#> deprecated in dplyr 1.1.0.

#> ℹ Please use `reframe()` instead.

#> ℹ When switching from `summarise()` to `reframe()`, remember that

#> `reframe()` always returns an ungrouped data frame and adjust

#> accordingly.

#> # A tibble: 2 × 5

#> term estimate std.error statistic p.value

#> <chr> <dbl> <dbl> <dbl> <dbl>

#> 1 (Intercept) 40.9 4.40 9.29 5.47e-17

#> 2 son 0.407 0.0636 6.40 1.36e- 9

Mô hình phù hợp với dữ liệu rất tốt. Nếu chúng ta nhìn vào công thức toán học của mô hình trên, nó có thể dễ dàng được giải thích không chính xác để gợi ý rằng con trai cao khiến người cha cao. Nhưng với những gì chúng ta biết về di truyền học và sinh học, chúng ta biết đó là cách khác. Mô hình là chính xác về mặt kỹ thuật. Các ước tính và giá trị p cũng thu được một cách chính xác. Điều sai ở đây là cách giải thích.

## 18.4 Những yếu tố lẫn lộn (Confounders)

Những yếu tố lẫn lộn có lẽ là lý do phổ biến nhất dẫn đến việc hiểu lầm về các mối quan hệ.

Nếu X và Y có mối quan hệ tương quan, chúng ta gọi Z là một yếu tố lẫn lộn nếu thay đổi trong Z gây ra thay đổi cả trong X và Y. Trước đó, khi nghiên cứu dữ liệu về bóng chày, chúng ta đã thấy Home Runs là một yếu tố lẫn lộn dẫn đến một tương quan cao hơn so với dự kiến khi nghiên cứu mối quan hệ giữa Bases on Balls và Runs. Trong một số trường hợp, chúng ta có thể sử dụng mô hình tuyến tính để tính toán cho những yếu tố lẫn lộn. Tuy nhiên, điều này không phải luôn luôn đúng.

Hiểu lầm không đúng do sự hiện diện của yếu tố lẫn lộn rất phổ biến trong truyền thông thông thường và chúng thường khó nhận biết. Dưới đây, chúng ta sẽ trình bày một ví dụ được sử dụng phổ biến liên quan đến việc đăng ký đại học.

### 18.4.1 Ví dụ: Tuyển sinh UC Berkeley

Dữ liệu nhập học từ sáu chuyên ngành UC Berkeley, từ năm 1973, cho thấy nhiều nam giới được nhận hơn phụ nữ: 44% nam giới được nhận so với 30% phụ nữ. PJ Bickel, EA Hammel và JW O'Connell. Khoa học (1975). Chúng ta có thể tải dữ liệu và tính toán một bài kiểm tra thống kê, trong đó bác bỏ rõ ràng giả thuyết rằng giới tính và nhập học là độc lập:

two\_by\_two <- admissions |> [group\_by](https://dplyr.tidyverse.org/reference/group_by.html)(gender) |>

[summarize](https://dplyr.tidyverse.org/reference/summarise.html)(total\_admitted = [round](https://rdrr.io/r/base/Round.html)([sum](https://rdrr.io/r/base/sum.html)(admitted / 100 \* applicants)),

not\_admitted = [sum](https://rdrr.io/r/base/sum.html)(applicants) - [sum](https://rdrr.io/r/base/sum.html)(total\_admitted)) |>

[select](https://dplyr.tidyverse.org/reference/select.html)(-gender)

[chisq.test](https://rdrr.io/r/stats/chisq.test.html)(two\_by\_two)$p.value

#> [1] 1.06e-21

Nhưng kiểm tra kỹ hơn cho thấy một kết quả nghịch lý. Dưới đây là tỷ lệ phần trăm tuyển sinh theo chuyên ngành:

admissions |> [select](https://dplyr.tidyverse.org/reference/select.html)(major, gender, admitted) |>

[pivot\_wider](https://tidyr.tidyverse.org/reference/pivot_wider.html)(names\_from = "gender", values\_from = "admitted") |>

[mutate](https://dplyr.tidyverse.org/reference/mutate.html)(women\_minus\_men = women - men)

#> # A tibble: 6 × 4

#> major men women women\_minus\_men

#> <chr> <dbl> <dbl> <dbl>

#> 1 A 62 82 20

#> 2 B 63 68 5

#> 3 C 37 34 -3

#> 4 D 33 35 2

#> 5 E 28 24 -4

#> # ℹ 1 more row

Bốn trong số sáu chuyên ngành ủng hộ phụ nữ. Quan trọng hơn, tất cả sự khác biệt nhỏ hơn nhiều so với sự khác biệt 14,2 mà chúng ta thấy khi kiểm tra tổng số.

Nghịch lý là việc phân tích tổng số cho thấy sự phụ thuộc giữa nhập học và giới tính, nhưng khi dữ liệu được nhóm theo chuyên ngành, sự phụ thuộc này dường như biến mất. Điều gì đang xảy ra? Điều này thực sự có thể xảy ra nếu một kẻ gây nhiễu không đếm được đang thúc đẩy hầu hết các biến đổi.

Vì vậy, hãy xác định ba biến: X là 1 đối với nam và 0 đối với nữ, Y là 1 cho những người được thừa nhận và 0 nếu không, và Z định lượng tính chọn lọc của chuyên ngành. Một tuyên bố thiên vị giới tính sẽ dựa trên thực tế là cao hơn cho hơn . Tuy nhiên Z là một yếu tố gây nhiễu quan trọng cần xem xét. Rõ ràng Z được liên kết với Y, vì chuyên ngành càng chọn lọc thì càng thấp . Nhưng là sự chọn lọc lớn Z gắn liền với giới tính X?

Một cách để thấy điều này là vẽ tổng tỷ lệ phần trăm được nhận vào một chuyên ngành so với tỷ lệ phần trăm phụ nữ tạo nên những người nộp đơn:

admissions |>

[group\_by](https://dplyr.tidyverse.org/reference/group_by.html)(major) |>

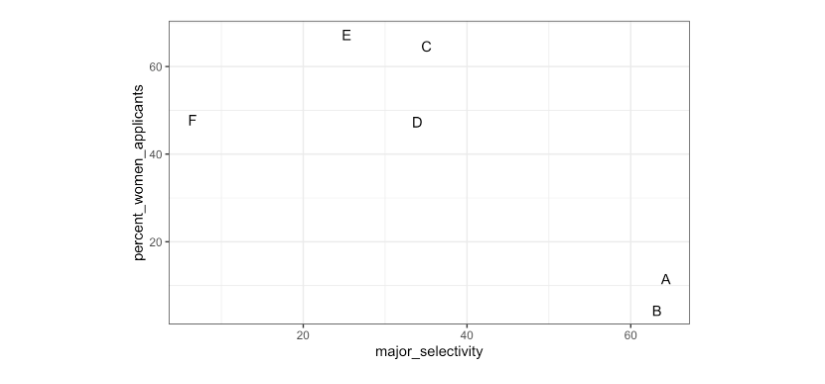
[summarize](https://dplyr.tidyverse.org/reference/summarise.html)(major\_selectivity = [sum](https://rdrr.io/r/base/sum.html)(admitted \* applicants)/[sum](https://rdrr.io/r/base/sum.html)(applicants),

percent\_women\_applicants = [sum](https://rdrr.io/r/base/sum.html)(applicants \* (gender=="women")) /

[sum](https://rdrr.io/r/base/sum.html)(applicants) \* 100) |>

[ggplot](https://ggplot2.tidyverse.org/reference/ggplot.html)([aes](https://ggplot2.tidyverse.org/reference/aes.html)(major\_selectivity, percent\_women\_applicants, label = major)) +

[geom\_text](https://ggplot2.tidyverse.org/reference/geom_text.html)()



Dường như có sự liên kết. Cốt truyện cho thấy phụ nữ có nhiều khả năng nộp đơn vào hai chuyên ngành "khó": giới tính và tính chọn lọc của chuyên ngành bị nhầm lẫn. So sánh, ví dụ, chuyên ngành B và chuyên ngành E. Thiếu tá E khó vào hơn nhiều so với chuyên ngành B và hơn 60% ứng viên vào chuyên ngành E là phụ nữ, trong khi ít hơn 30% ứng viên của chuyên ngành B là phụ nữ.

### 18.4.2 Giải thích nhiễu bằng đồ họa

Biểu đồ sau đây cho thấy số lượng ứng viên được nhận và những người không được nhận:

A graph of different colored squares

Description automatically generated with medium confidence

Nó cũng phá vỡ sự chấp nhận theo chuyên ngành. Sự phân tích này cho phép chúng ta thấy rằng phần lớn những người đàn ông được chấp nhận đến từ hai chuyên ngành: A và B. Nó cũng cho chúng ta thấy rằng rất ít phụ nữ nộp đơn vào các chuyên ngành này.

### 18.4.3 Trung bình sau khi phân tầng

Trong biểu đồ này, chúng ta có thể thấy rằng nếu chúng ta điều kiện hoặc phân tầng theo chính, và sau đó nhìn vào sự khác biệt, chúng ta kiểm soát sự nhầm lẫn và hiệu ứng này biến mất:

admissions |>

[ggplot](https://ggplot2.tidyverse.org/reference/ggplot.html)([aes](https://ggplot2.tidyverse.org/reference/aes.html)(major, admitted, col = gender, size = applicants)) +

[geom\_point](https://ggplot2.tidyverse.org/reference/geom_point.html)()

A graph with red and blue dots

Description automatically generated

Bây giờ chúng ta thấy rằng chuyên ngành theo chuyên ngành, không có nhiều khác biệt. Kích thước của dấu chấm đại diện cho số lượng ứng viên và giải thích nghịch lý: chúng ta thấy các chấm lớn màu đỏ và các chấm nhỏ màu xanh lam cho các chuyên ngành dễ nhất, A và B.

Nếu chúng ta tính trung bình sự khác biệt theo chuyên ngành, chúng ta thấy rằng tỷ lệ phần trăm thực sự cao hơn 3,5% đối với phụ nữ.

admissions |> [group\_by](https://dplyr.tidyverse.org/reference/group_by.html)(gender) |> [summarize](https://dplyr.tidyverse.org/reference/summarise.html)(average = [mean](https://rdrr.io/r/base/mean.html)(admitted))

#> # A tibble: 2 × 2

#> gender average

#> <chr> <dbl>

#> 1 men 38.2

#> 2 women 41.7

## 18.5 Nghịch lý của Simpson

Trường hợp chúng tôi vừa đề cập là một ví dụ về nghịch lý của Simpson. Nó được gọi là nghịch lý vì chúng ta thấy dấu hiệu của sự đảo ngược tương quan khi so sánh toàn bộ ấn phẩm và các tầng lớp cụ thể. Như một ví dụ minh họa, giả sử bạn có ba biến ngẫu nhiên X, Y và Z và rằng chúng ta quan sát sự chứng ngộ về những điều này. Dưới đây là sơ đồ các quan sát mô phỏng cho X và Y cùng với mối tương quan mẫu:

A graph with black dots

Description automatically generated

Bạn có thể thấy rằng X và Y có tương quan nghịch với nhau. Tuy nhiên, một khi chúng ta phân tầng bằng Z (hiển thị bằng các màu khác nhau bên dưới) Một mô hình khác xuất hiện:

A graph with colored dots

Description automatically generated

Nó thực sự là Z có tương quan nghịch với X. Nếu chúng ta phân tầng bởi Z, các X và Y thực sự có mối tương quan thuận như đã thấy trong cốt truyện trên.