THỐNG KÊ ỨNG DỤNG

ĐỖ LÂN

dolan@tlu.edu.vn
Đại học Thủy Lợi

Ngày 11 tháng 12 năm 2018

Nội dung môn học

1. Tổng quan về Thống kê
2. Thu thập dữ liệu
3. Tóm tắt và trình bày dữ liệu bằng bảng và đồ thị
4. Tóm tắt dữ liệu bằng các đại lượng thống kê mô tả
5. Xác suất căn bản và biến ngẫu nhiên
6. Phân phối của tham số mẫu và ước lượng tham số tổng thể
7. Kiểm định giả thuyết về tham số một tổng thể
8. Kiểm định giả thuyết về tham số hai tổng thể
9. Phân tích phương sai
10. Kiểm định phi tham số
11. Kiểm định chi - bình phương
12. **Hồi quy đơn biến**
13. Hồi quy đa biến

Phần XI

Hồi quy đơn biến

Mục lục

Nội dung chính được trình bày trong chương

- Giới thiệu bài toán hồi quy;
- Trình bày định nghĩa mô hình hồi quy tuyến tính đơn biến tổng thể;
- Giới thiệu phương pháp bình phương tối thiểu tìm phương trình hồi quy tuyến tính mẫu;
- Trình bày các kiến thức liên quan đến hồi quy: Hệ số xác định, sai số chuẩn của ước lượng, kiểm định về hệ số độ dốc và khoảng tin cậy cho hệ số độ dốc;
- Trình bày bài toán dự báo giá trị của biến phụ thuộc theo các biến độc lập;
- Giới thiệu bài toán tương quan tuyến tính.

Những kiến thức sinh viên phải làm được trong chương

- Nắm được thế nào là bài toán hồi quy và phân biệt được mối liên hệ tuyến tính và mối liên hệ hồi quy;
- Nắm được định nghĩa mô hình hồi quy tuyến tính đơn biến tổng thể;
- Biết cách tìm phương trình hồi quy tuyến tính theo phương pháp bình phương tối thiểu;
- Hiểu các kiến thức liên quan đến hồi quy: Hệ số xác định, sai số chuẩn của ước lượng, kiểm định về hệ số độ dốc và khoảng tin cậy cho hệ số độ dốc;
- Biết cách dự báo giá trị của biến phụ thuộc theo các biến độc lập;
- Hiểu được thế nào là bài toán tương quan tuyến tính và phân biệt được bài toán tương quan tuyến tính với bài toán hồi quy tuyến tính.

1. Hồi quy tuyến tính đơn biến
   - Giới thiệu chung về hồi quy
   - Mô hình hồi quy tuyến tính đơn biến
   - Khả năng dự đoán của phương trình hồi quy tuyến tính
   - Suy diễn thống kê về hệ số độ dốc
   - Dự báo giá trị của biến phụ thuộc theo biến độc lập
2. Tương quan tuyến tính
   - Hệ số tương quan tuyến tính
   - Kiểm định hệ số tương quan tổng thể
3. Tương quan thứ hạng Spearman
   - Hệ số tương quan thứ hạng Spearman

Thuật ngữ Hồi quy (regression)

Thuật ngữ "**Hồi quy**" lần đầu tiên được Francis Galton sử dụng năm 1886. Trong nghiên cứu của mình, ông cho rằng có một xu hướng về chiều cao của những đứa trẻ do cha mẹ quá cao hoặc quá thấp! Xu hướng này được gọi là luật Galton.
Trong nghiên cứu của Galton, ông sử dụng thuật ngữ "**regression to mediocrity**", nghĩa là Quy về trung bình.
Từ đó, thuật ngữ "**hồi quy**" được sử dụng khi nghiên cứu những vấn đề tương tự.

Khái niệm hồi quy

Định nghĩa

Phân tích hồi quy là quá trình xây dựng mô hình toán học hoặc hàm toán học mà có thể miêu tả, dự đoán hoặc điều khiển một biến từ một hoặc nhiều biến khác.

Khái niệm hồi quy

Định nghĩa

Phân tích hồi quy là quá trình xây dựng mô hình toán học hoặc hàm toán học mà có thể miêu tả, dự đoán hoặc điều khiển một biến từ một hoặc nhiều biến khác.

Chú ý:

- Biến được dự đoán gọi là *biến phụ thuộc* (ký hiệu là Y), các biến dự đoán được gọi là các *biến độc lập* (ký hiệu X).

Khái niệm hồi quy

Định nghĩa

Phân tích hồi quy là quá trình xây dựng mô hình toán học hoặc hàm toán học mà có thể miêu tả, dự đoán hoặc điều khiển một biến từ một hoặc nhiều biến khác.

Chú ý:

- Biến được dự đoán gọi là *biến phụ thuộc* (ký hiệu là Y), các biến dự đoán được gọi là các *biến độc lập* (ký hiệu X).
- Nếu trong mô hình chỉ có một biến được dự đoán và một biến độc lập thì ta có mô hình *hồi quy đơn biến*.

Khái niệm hồi quy

Định nghĩa

Phân tích hồi quy là quá trình xây dựng mô hình toán học hoặc hàm toán học mà có thể miêu tả, dự đoán hoặc điều khiển một biến từ một hoặc nhiều biến khác.

Chú ý:

- Biến được dự đoán gọi là *biến phụ thuộc* (ký hiệu là Y), các biến dự đoán được gọi là các *biến độc lập* (ký hiệu X).
- Nếu trong mô hình chỉ có một biến được dự đoán và một biến độc lập thì ta có mô hình *hồi quy đơn biến*.
  Nếu trong mô hình có một biến được dự đoán và hai hay nhiều hơn hai biến dự đoán thì ta có mô hình *hồi quy đa biến*.

Ví dụ

Luật Galton Karl Pearson nghiên cứu sự phụ thuộc chiều cao của những bé trai vào chiều cao của bố chúng. Trong trường hợp này:

- Chiều cao của bố là biến độc lập;

Ví dụ

Luật Galton Karl Pearson nghiên cứu sự phụ thuộc chiều cao của những bé trai vào chiều cao của bố chúng. Trong trường hợp này:

- Chiều cao của bố là biến độc lập;
- Chiều cao của con là biến phụ thuộc.

Ví dụ

Luật Galton Karl Pearson nghiên cứu sự phụ thuộc chiều cao của những bé trai vào chiều cao của bố chúng. Trong trường hợp này:

- Chiều cao của bố là biến độc lập;
- Chiều cao của con là biến phụ thuộc.

Ông đã xây dựng được đồ thị (dạng đường thẳng) chỉ ra phân bố chiều cao của những bé trai ứng với chiều cao của người cha. Mô hình này cho ta thấy:

- Với chiều cao đã biết của cha, chiều cao của con sẽ dao động xung quanh giá trị trung bình;

Ví dụ

Luật Galton Karl Pearson nghiên cứu sự phụ thuộc chiều cao của những bé trai vào chiều cao của bố chúng. Trong trường hợp này:

- Chiều cao của bố là biến độc lập;
- Chiều cao của con là biến phụ thuộc.

Ông đã xây dựng được đồ thị (dạng đường thẳng) chỉ ra phân bố chiều cao của những bé trai ứng với chiều cao của người cha. Mô hình này cho ta thấy:

- Với chiều cao đã biết của cha, chiều cao của con sẽ dao động xung quanh giá trị trung bình;
- Chiều cao của cha tăng thì chiều cao của con cũng tăng;

Ví dụ

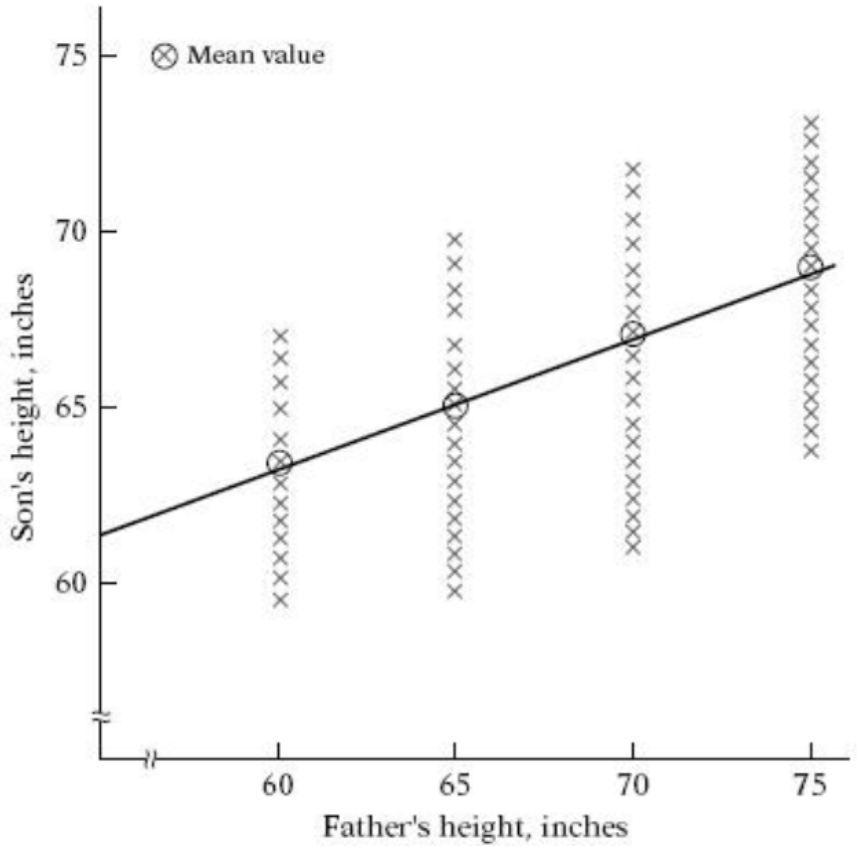
Luật Galton Karl Pearson nghiên cứu sự phụ thuộc chiều cao của những bé trai vào chiều cao của bố chúng. Trong trường hợp này:

- Chiều cao của bố là biến độc lập;
- Chiều cao của con là biến phụ thuộc.

Ông đã xây dựng được đồ thị (dạng đường thẳng) chỉ ra phân bố chiều cao của những bé trai ứng với chiều cao của người cha. Mô hình này cho ta thấy:

- Với chiều cao đã biết của cha, chiều cao của con sẽ dao động xung quanh giá trị trung bình;
- Chiều cao của cha tăng thì chiều cao của con cũng tăng;
- Chiều cao trung bình của nhóm bố cao nhỏ hơn chiều cao của bố và chiều cao trung bình của nhóm bố thấp thì lớn hơn chiều cao của bố.

Minh họa quan hệ chiều cao của cha con



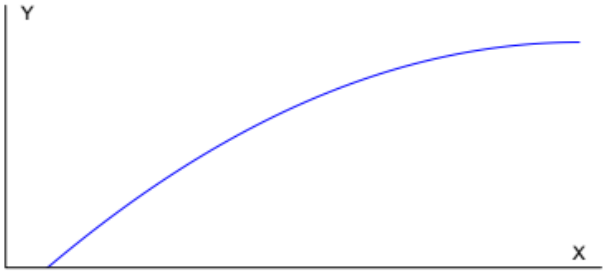
Liên hệ thống kê và liên hệ hàm số

- Nếu Y và X có liên hệ hàm số, giả sử hàm số bậc nhất $Y = aX + b$, thì với mỗi giá trị của X ta có duy nhất một giá trị của Y.

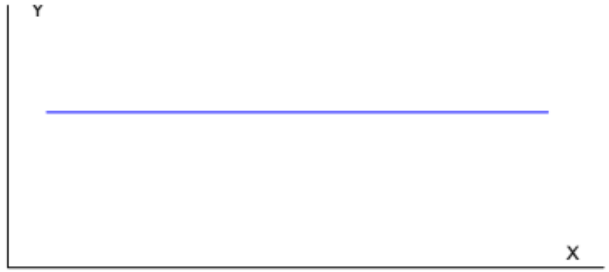
Liên hệ thống kê và liên hệ hàm số

- Nếu Y và X có liên hệ hàm số, giả sử hàm số bậc nhất $Y = aX + b$, thì với mỗi giá trị của X ta có duy nhất một giá trị của Y.
- Nếu Y và X có liên hệ thống kê, chẳng hạn theo mô hình hồi quy tuyến tính $Y = aX + b + \epsilon$, thì với mỗi giá trị của biến dự đoán X, ta không thể tính chính xác được biến dự đoán Y nhận giá trị bao nhiêu, vì có nhiều yếu tố khác cùng tác động đến Y mà không được đề cập đến trong mô hình.

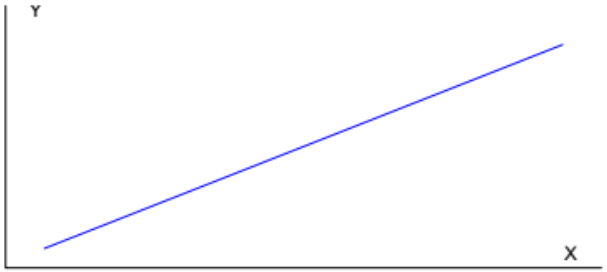
Các dạng liên hệ giữa hai biến X và Y



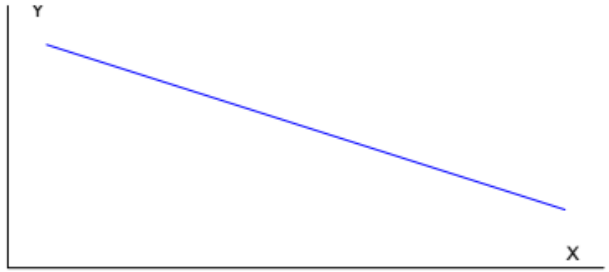
a. Lien he phi tuyen



b. Khong co lien he



c. Lien he tuyen tinh thuan



d. Lien he tuyen tinh nghich

Mô hình hồi quy tuyến tính đơn biến

Định nghĩa

Mô hình hồi quy tuyến tính đơn biến của biến phụ thuộc Y theo biến độc lập X là phương trình có dạng $Y = \beta_0 + \beta_1 X + \epsilon$, sao cho khi X nhận giá trị X_i thì giá trị tương ứng Y_i được xác định bởi

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

trong đó

Mô hình hồi quy tuyến tính đơn biến

Định nghĩa

Mô hình hồi quy tuyến tính đơn biến của biến phụ thuộc Y theo biến độc lập X là phương trình có dạng $Y = \beta_0 + \beta_1 X + \epsilon$, sao cho khi X nhận giá trị X_i thì giá trị tương ứng Y_i được xác định bởi

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

trong đó

- $\beta_0 + \beta_1 X_i$: là giá trị trung bình của biến phụ thuộc Y khi X nhận giá trị X_i và được kí hiệu là $E(Y|X_i)$.
- β_0, β_1 là các hệ số hồi quy, β_0 gọi là tung độ gốc, β_1 gọi là độ dốc;
- ϵ_i là biến ngẫu nhiên độ cách, được tính là chênh lệch giữa giá trị Y_i thực tế và giá trị trung bình $E(Y|X_i)$, tức là $\epsilon_i = Y_i - E(Y|X_i)$.

Các giả định liên quan đến biến ngẫu nhiên độ cách

Tại cùng một giá trị X_i có thể có nhiều giá trị Y_i khác nhau, điều này dẫn đến có nhiều giá trị ϵ_i khác nhau. Tại mỗi giá trị X_i, biến ngẫu nhiên ϵ_i miêu tả ảnh hưởng của các yếu tố khác ngoài biến độc lập X_i lên biến phụ thuộc Y_i và được giả định thỏa mãn các điều kiện sau:

Các giả định liên quan đến biến ngẫu nhiên độ cách

Tại cùng một giá trị X_i có thể có nhiều giá trị Y_i khác nhau, điều này dẫn đến có nhiều giá trị ϵ_i khác nhau. Tại mỗi giá trị X_i, biến ngẫu nhiên ϵ_i miêu tả ảnh hưởng của các yếu tố khác ngoài biến độc lập X_i lên biến phụ thuộc Y_i và được giả định thỏa mãn các điều kiện sau:

- Mỗi ϵ_i là một biến ngẫu nhiên tuân theo phân phối chuẩn;
- Các biến ngẫu nhiên ϵ_i có cùng trung bình bằng 0 và cùng phương sai;
- Các biến ngẫu nhiên ϵ_i là độc lập với nhau.

Ý nghĩa của các hệ số hồi quy

- β_1 là hệ số độ dốc, đo lường lượng thay đổi trung bình của biến phụ thuộc Y cho mỗi đơn vị thay đổi của X. Hệ số độ dốc có thể dương, âm hoặc bằng 0 phụ thuộc vào mối liên hệ của Y và X là dương, âm hay bằng 0.
- β_0 là hệ số tung độ gốc, cho biết giá trị trung bình của Y khi X bằng 0.

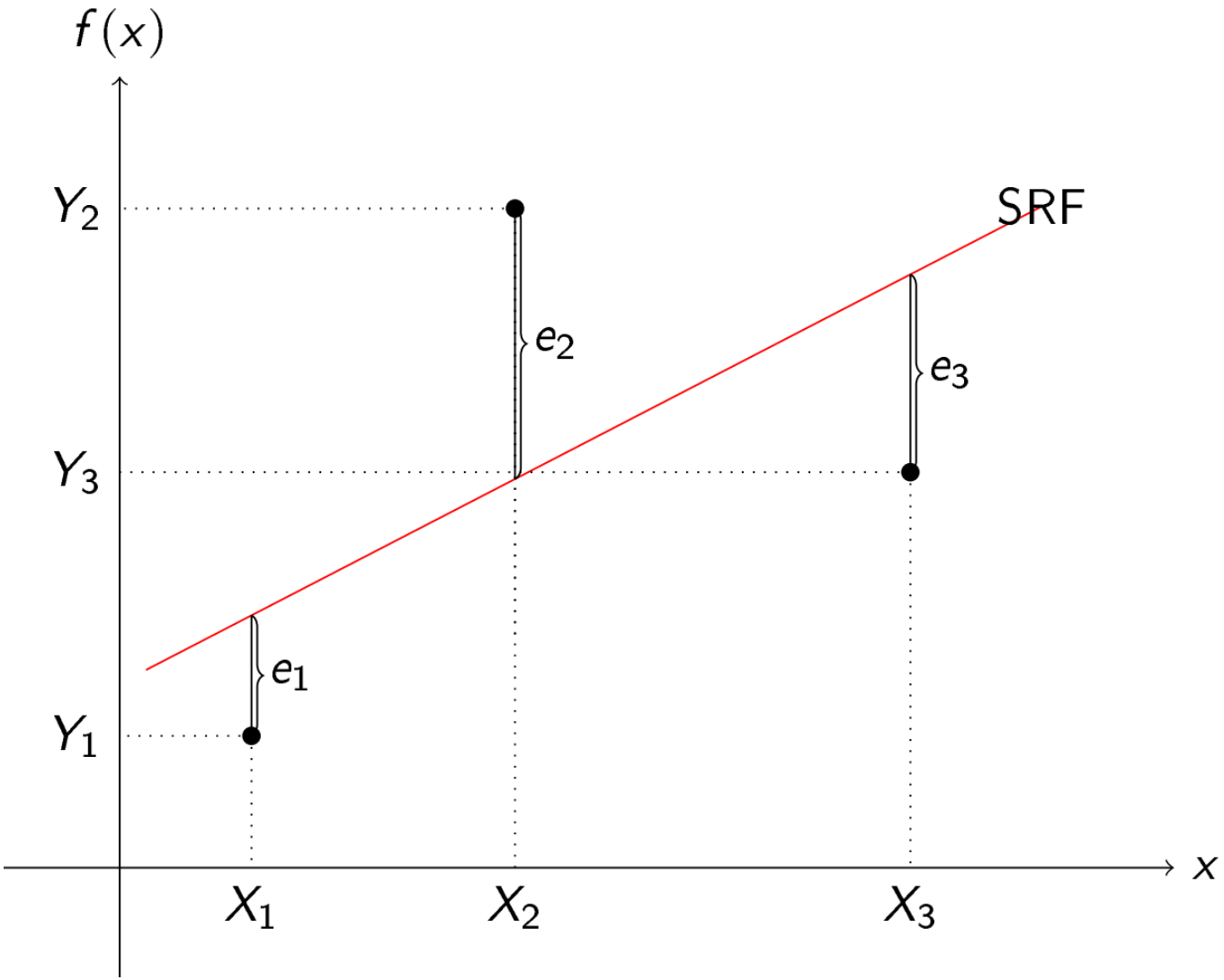
Phương pháp bình phương tối thiểu

- Với một tập dữ liệu mẫu, ta cần xác định đường thẳng dựa trên tập dữ liệu này là ước lượng tốt nhất cho mô hình hồi quy. Đường thẳng ước lượng này gọi là đường hồi quy mẫu của Y theo X.
- Để tìm đường hồi quy mẫu, ta cần xác định cặp hệ số $b_0; b_1$ lần lượt là các ước lượng điểm của hệ số tung độ gốc và hệ số độ dốc. Tức là, ta cần tìm một đường thẳng có dạng $Y = b_0 + b_1 X$ là ước lượng tốt nhất cho đường thẳng $E(Y|X) = \beta_0 + \beta_1 X$.
- Phương pháp bình phương nhỏ nhất tìm cặp b_0, b_1 của đường hồi quy mẫu sao cho sự khác biệt giữa giá trị thực Y_i và giá trị tìm thấy từ đường hồi quy $\hat{Y}_i = b_0 + b_1 X_i$ nhỏ nhất, theo nghĩa tổng bình phương

$$\sum (Y_i - \hat{Y}_i)^2 = \sum (Y_i - (b_0 + b_1 X_i))^2$$

đạt giá trị nhỏ nhất.

Minh họa phương pháp bình phương tối thiểu



Công thức xác định cặp hệ số của đường hồi quy mẫu

- Giả sử ta có mẫu ngẫu nhiên gồm n cặp $(X_1, Y_1), (X_1, Y_1), ..., (X_n, Y_n)$, theo phương pháp bình phương tối thiểu, ta sẽ xác định công thức của b_0, b_1 sao cho hàm hai biến b_0, b_1

$$\sum_{i=1}^{n} n(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X))^2$$

đạt giá trị nhỏ nhất.

Công thức xác định cặp hệ số của đường hồi quy mẫu

- Giả sử ta có mẫu ngẫu nhiên gồm n cặp $(X_1, Y_1), (X_1, Y_1), ..., (X_n, Y_n)$, theo phương pháp bình phương tối thiểu, ta sẽ xác định công thức của b_0, b_1 sao cho hàm hai biến b_0, b_1

$$\sum_{i=1}^{n} n(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - (b_0 + b_1 X))^2$$

đạt giá trị nhỏ nhất.

- Theo công thức tìm giá trị cực trị của hàm hai biến, ta có

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}$$

và $b_0 = \bar{Y} - b_1 \bar{X}$, trong đó $\bar{X}, \bar{Y}$ tương ứng là trung bình của X_i, Y_i.

Ví dụ

Bài toán

Để nghiên cứu mối liên hệ giữa chi phí trên mỗi chuyến bay và số lượng hành khách trên mỗi chuyến, nghiên cứu 12 chuyến máy bay thương mại đường bay 500km, sử dụng bằng loại máy bay Airbus và bay cùng mùa trong năm, ta thu được bảng dữ liệu như sau:

Chuyến bay	Số khách	Chi phí (1000$)
1	61	4.280
2	63	4.080
3	67	4.420
4	69	4.170
5	70	4.480
6	74	4.300
7	76	4.820
8	81	4.700
9	86	5.110
10	91	5.130
11	95	5.640
12	97	5.560

Tìm đường hồi quy mẫu của chi phí mỗi chuyến bay theo số hành khách.

Ví dụ

Lời giải:

- Gọi X là biến ngẫu nhiên chỉ số khách hàng trên mỗi chuyến bay và Y là biến ngẫu nhiên chỉ tổng chi phí cho mỗi chuyến bay. Từ bảng dữ liệu ta có

$$\sum_{i=1}^{12} X_i Y_i = 4462.22, \overline{X} = 77.5, \overline{Y} = 4.7242, \sum_{i=1}^{12} X_i^2 = 73764.$$

- Từ đó theo công thức tính các hệ số của đường hồi qui mẫu ta có:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{4462.22 - 12 \times 77.5 \times 4.7242}{73764 - 12 \times 77.5^2} = 0.0407$$

và $b_0 = \overline{Y} - b_1\overline{X} = 4.7242 - 0.0407 \times 77.5 = 1.5698.$

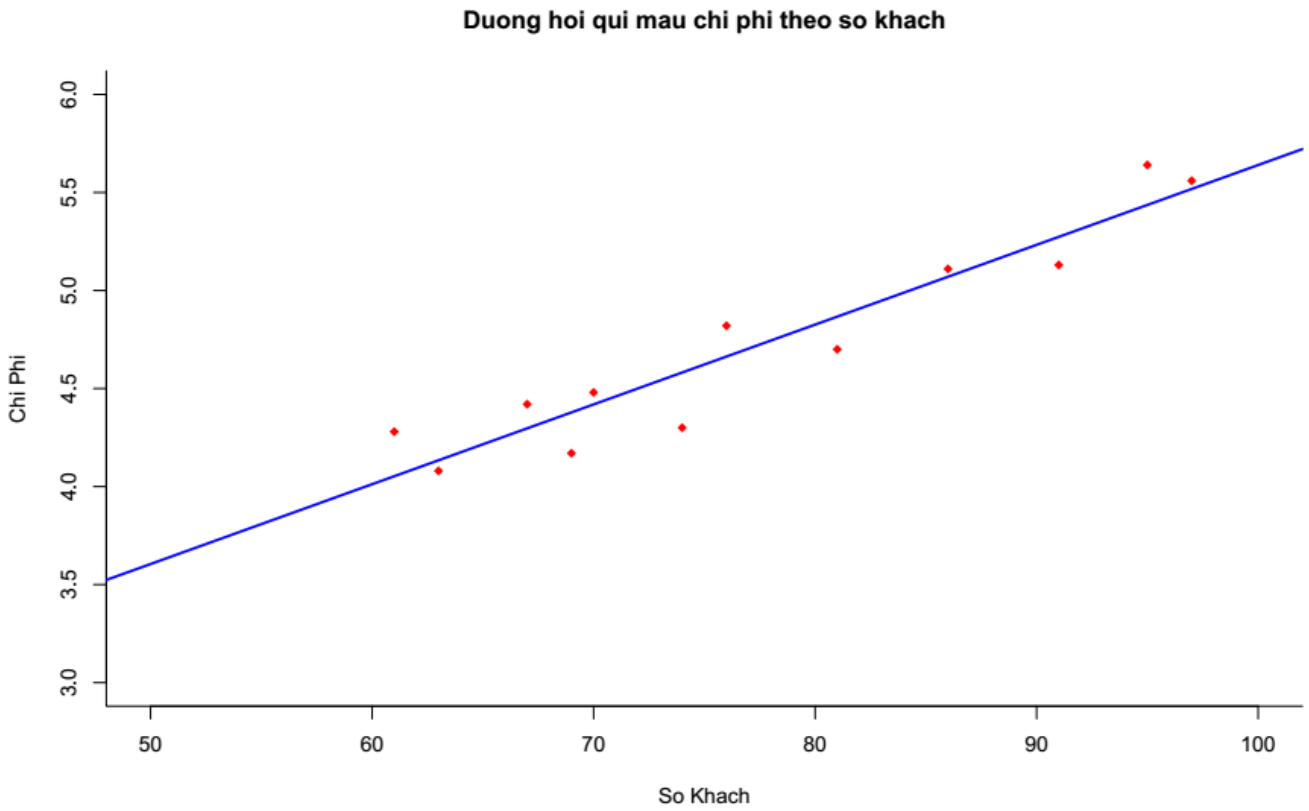
- Vậy đường hồi qui mẫu của chi phí mỗi chuyến bay theo số khách là:

$$Y = 1.5698 + 0.0407X.$$

Nhận xét

- Từ phương trình của đường hồi quy mẫu, ta thấy mối liên hệ giữa tổng chi phí và số khách trên mỗi chuyến bay biến thiên cùng chiều, có nghĩa là là số khách tăng thì chi phí tăng và ngược lại.
- Hệ số độ dốc $b_1 = 0,0407$ cho ta thấy:
- Hệ số tung độ gốc $b_0 = 1.5698$ cho ta thấy:

Minh họa bằng đồ thị



Tính hệ số đường hồi quy mẫu trong R

- Để tìm các hệ số của phương trình hồi qui tuyến tính mẫu của biến phụ thuộc Y theo biến độc lập X trong R, ta dùng hàm `lm(Y ~ X)`
- Chẳng hạn để tìm phương trình hồi qui tuyến tính mẫu của chi phí theo số hành khách, ta thực hiện trong R như sau:

```
> SoKhach = c(61, 63, 67, 69, 70,
74, 76, 81, 86, 91, 95, 97)
> ChiPhi = c(4.280, 4.080, 4.420, 4.170,
4.480, 4.300, 4.820, 4.700, 5.110, 5.130, 5.640, 5.560)
> lm(ChiPhi ~ SoKhach)
```

- Kết quả trong R

```
Call:
  lm(formula = ChiPhi ~ SoKhach)
Coefficients:
(Intercept)     SoKhach
     1.5698      0.0407
```

- Kết quả trong R cho ta $b_0 = 1.5698$ và $b_1 = 0.0407$ hay phương trình hồi qui mẫu có dạng $Y = 1.5698 + 0.0407X$.

1. Hồi quy tuyến tính đơn biến
   - Giới thiệu chung về hồi quy
   - Mô hình hồi quy tuyến tính đơn biến
   - Khả năng dự đoán của phương trình hồi quy tuyến tính
   - Suy diễn thống kê về hệ số độ dốc
   - Dự báo giá trị của biến phụ thuộc theo biến độc lập
2. Tương quan tuyến tính
   - Hệ số tương quan tuyến tính
   - Kiểm định hệ số tương quan tổng thể
3. Tương quan thứ hạng Spearman
   - Hệ số tương quan thứ hạng Spearman

Minh họa mức độ phù hợp của mô hình với tập dữ liệu

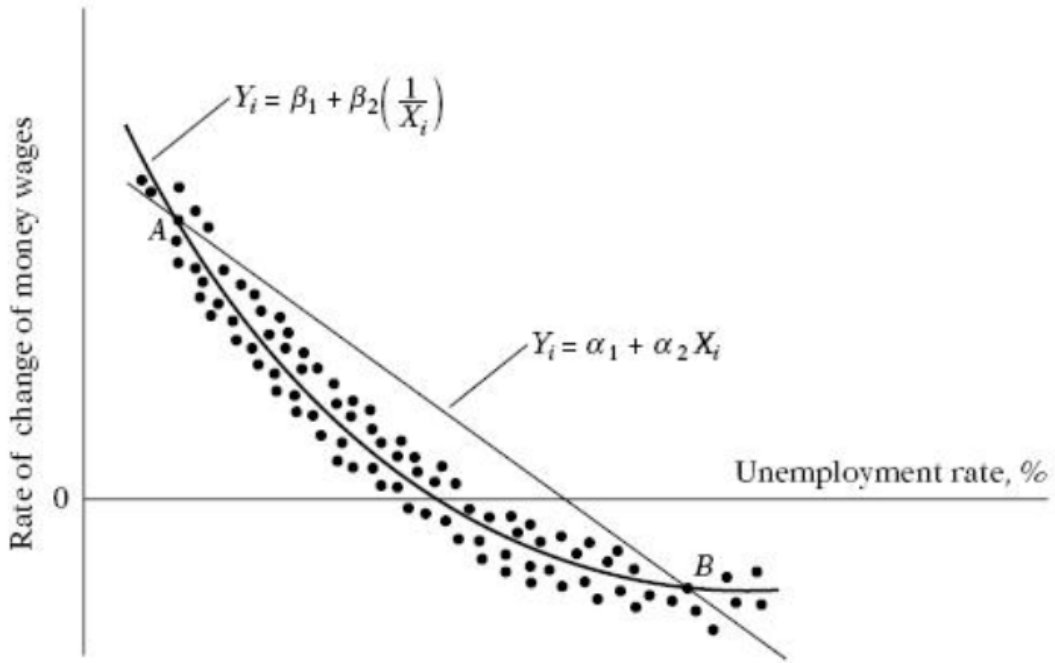


FIGURE 3.7 Linear and nonlinear Phillips curves.

Độ biến thiên dự đoán và không dự đoán được

- Ta có $\hat{Y}_i = b_0 + b_1 X_i$ và $\varepsilon = Y_i - \hat{Y}_i = Y_i - (b_0 + b_1 X_i)$.
- Từ đó $Y_i = \hat{Y}_i + \varepsilon_i$, Y_i được chia làm hai phần: phần $\hat{Y}_i = b_0 + b_1 X_i$ là phần dự đoán được bởi phương trình hồi qui, còn độ cách hay phần dư ε_i là phần không dự đoán được bởi phương trình hồi qui.
- Gọi $\overline{Y}$ là trung bình mẫu của những Y_i, khi đó $Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + \varepsilon_i$. Ta có thể chứng minh được

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n}\varepsilon_i^2.$$

  - $\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$: là độ biến thiên dự đoán được bởi phương trình hồi qui;
  - $\sum_{i=1}^{n}\varepsilon_i^2$: là độ biến thiên không dự đoán được bởi phương trình hồi qui.

Hệ số xác định

- $STT = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$: tổng bình phương toàn phần;
- $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$: tổng bình phương hồi qui;
- $SSE = \sum_{i=1}^{n}\varepsilon_i^2$: tổng bình phương độ cách (phần dư).
- Ta có $SST = SSR + SSE$.

Hệ số xác định

- $STT = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$: tổng bình phương toàn phần;
- $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$: tổng bình phương hồi qui;
- $SSE = \sum_{i=1}^{n}\varepsilon_i^2$: tổng bình phương độ cách (phần dư).
- Ta có $SST = SSR + SSE$.

Định nghĩa

Hệ số xác định của đường hồi qui mẫu, R^2 được xác định bởi công thức:

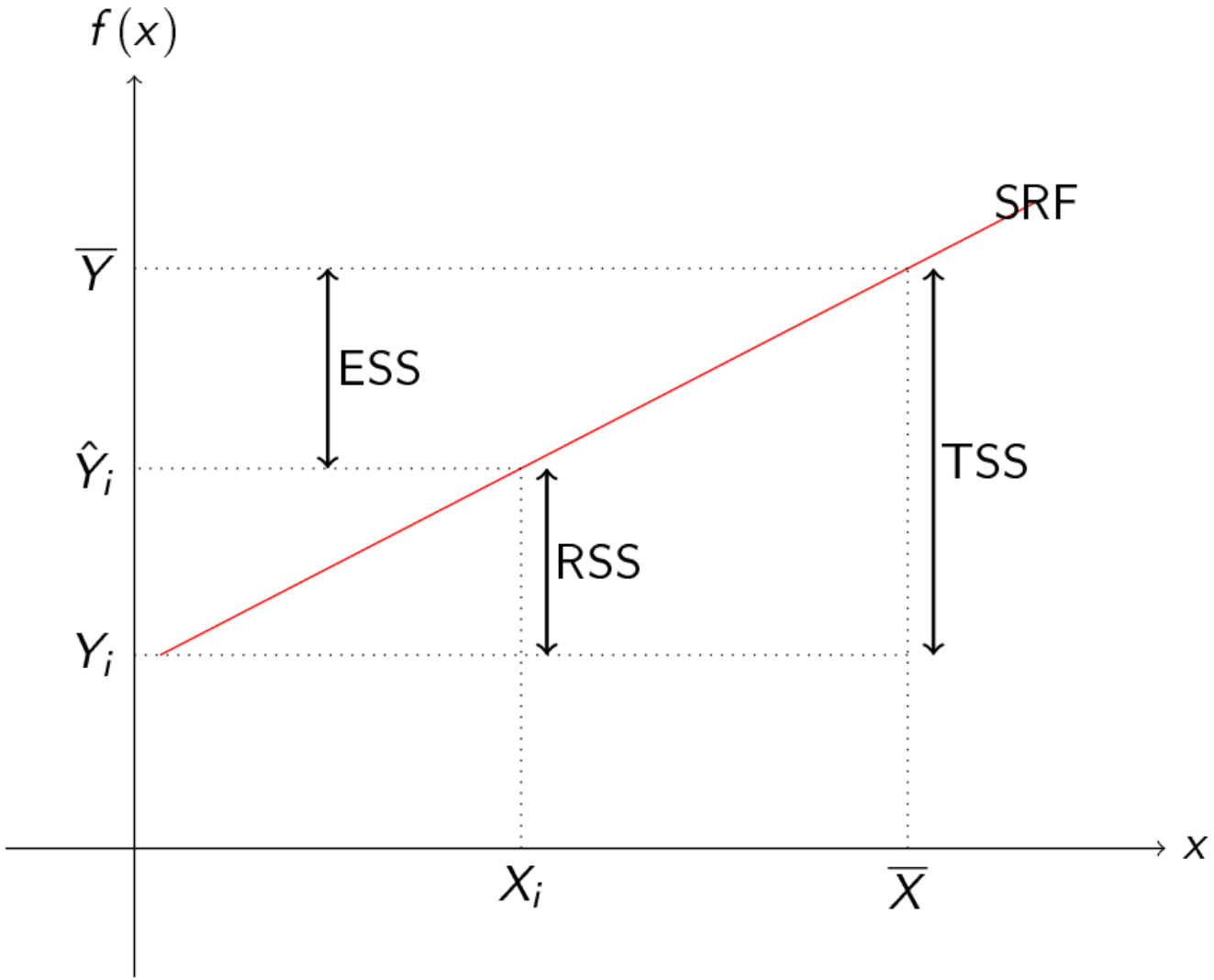
$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}.$$

Hệ số xác định

Nhận xét:

- Hệ số xác định R^2 luôn thuộc đoạn $[0,1]$: $0 \leqslant R^2 \leqslant 1$.
- Nếu $R^2 = 1$ thì Y được giải thích hoàn toàn qua X.
- Nếu $R^2 = 0$ thì Y hoàn toàn không được giải thích qua X hay Y và X không có quan hệ.
- Hệ số xác định đo mức độ phù hợp của mô hình so với tập số liệu và cho ta thấy tỉ lệ phần dự đoán được từ phương trình hồi qui. Hệ số xác định càng lớn thì khả năng dự đoán của hồi qui càng cao. Không có tiêu chuẩn chung để xác định R^2 bao nhiêu là cao hay thấp và ta không chỉ căn cứ vào R^2 để đánh giá mô hình là tốt hay không tốt. Để xem xét một mô hình là tốt hay không ta phải căn cứ vào nhiều yếu tố: R^2, dấu của hệ số hồi qui, kinh nghiệp thực tế, khả năng dự báo chính xác,...
  Theo kinh nghiệm, với số liệu chuỗi thời gian thì $R^2 > 0.9$ được xem là tốt, với số liệu chéo thì $R^2 > 0.7$ được xem là tốt.

Minh họa tổng bình phương các độ lệch



Định lý Gauss-Markov

Giả sử Y có phân phối chuẩn, khi đó các tham số mẫu $b_0; b_1$ cũng có phân phối chuẩn, và ta chứng minh được:

Định lý Gauss-Markov

Giả sử Y có phân phối chuẩn, khi đó các tham số mẫu $b_0; b_1$ cũng có phân phối chuẩn, và ta chứng minh được:

- Trung bình: $E(b_1) = \beta_1$.
- Phương sai: $s_{b_1}^2 = \dfrac{s_\epsilon^2}{\sum_{i=1}^{n}((X_i - \overline{X})^2}$

Định lý Gauss-Markov

Giả sử Y có phân phối chuẩn, khi đó các tham số mẫu $b_0; b_1$ cũng có phân phối chuẩn, và ta chứng minh được:

- Trung bình: $E(b_1) = \beta_1$.
- Phương sai: $s^2_{b_1} = \dfrac{s^2_\epsilon}{\sum_{i=1}^{n}((X_i - \overline{X})^2}$

Theorem

Trong các ước lượng tuyến tính không chệch cho hệ số hồi quy tổng thể, ước lượng tìm được bằng phương pháp bình phương tối thiểu có phương sai nhỏ nhất.

Định lý Gauss-Markov

Giả sử Y có phân phối chuẩn, khi đó các tham số mẫu $b_0; b_1$ cũng có phân phối chuẩn, và ta chứng minh được:

- Trung bình: $E(b_1) = \beta_1$.
- Phương sai: $s^2_{b_1} = \dfrac{s^2_\epsilon}{\sum_{i=1}^{n}((X_i - \overline{X})^2}$

Theorem

Trong các ước lượng tuyến tính không chệch cho hệ số hồi quy tổng thể, ước lượng tìm được bằng phương pháp bình phương tối thiểu có phương sai nhỏ nhất.

Nhận xét: Định lý Gauss-Markov cho ta thấy b_1 là ước lượng không chệch tuyến tính hữu hiệu nhất của β_1.

Khoảng tin cậy cho hệ số độ dốc

- Khi các giả định về các độ cách ϵ_i được thỏa mãn thì ta có thể chứng minh được biến ngẫu nhiên

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

tuân theo phân phối student với $n - 2$ bậc tự do.

- Khoảng tin cậy $100(1 - \alpha)\%$ cho hệ số độ dốc của đường hồi quy tổng thể là

$$b_1 - t_{n-2,\alpha/2} s_{b_1} < \beta_1 < b_1 + t_{n-2,\alpha/2} s_{b_1}$$

Ví dụ

Với bảng dữ liệu về số khách và chi phí trên mỗi chuyến máy bay trong ví dụ trước, ta có thể tính toán được

- $STT = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = 3.1121$, $SSR = \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 = 2.7980$, $SSE = \sum_{i=1}^{n}\varepsilon_i^2 = 0.3141$.
- Hệ số xác định $R^2 = \dfrac{SSR}{SST} = 0.8991$.
- Sai số chuẩn của ước lượng $s_\varepsilon = \sqrt{\dfrac{SSE}{n-2}} = 0.1772$.
- Độ lệch chuẩn của hệ số độ dốc $s_{b_1} = \sqrt{\dfrac{s_\varepsilon^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}} = 0.004212$.
- Khoảng tin cậy 95% cho độ dốc của đường hồi qui tổng thể là: (0.0311, 0.0503)

Kiểm định ý nghĩa kinh tế của bài toán chi phí theo số khách

- Để xét xem số hành khách và chi phí có liên hệ đồng biến với nhau không, ta kiểm định cặp giả thuyết $H_0 : \beta_1 \leqslant 0, H_1 : \beta_1 > 0$.
- Ta có thống kê $t = \dfrac{b_1 - \beta}{se(b_1)} = \dfrac{0.040702}{0.004312} = 9.439$
- Giá trị tới hạn $t_{n-2,\alpha} = t_{10,0.05} = 1.81$.
- Do $t = 9.439 > t_{10,0.05} = 1.81$ nên bác bỏ H_0. Tức là ta có đủ bằng chứng thống kê để cho rằng $\beta_1 > 0$, tức là số hành khách và chi phí liên hệ đồng biến hay bài toán có ý nghĩa kinh tế.

Kiểm định về độ dốc của đường hồi quy tổng thể

- Các cặp giả thuyết cho bài toán kiểm định giả thuyết cho hệ số độ dốc β_1

	Bài toán 2	Bài toán 2	Bài toán 3
H_0 :	$\beta_1 = \beta;\ \beta_1 \leqslant \beta$	$\beta_1 = \beta;\ \beta_1 \geqslant \beta$	$\beta_1 = \beta$
H_1 :	$\beta_1 > \beta$	$\beta_1 < \beta$	$\beta_1 \neq \beta$

- Do thống kê $t = \dfrac{b_1 - \beta}{s_{b_1}}$ tuân theo phân phối student với $n-2$ bậc tự do nên ta có các qui luật quyết định tại mức ý nghĩa α:
  - Bài toán 1: Bác bỏ H_0 nếu $\dfrac{b_1 - \beta}{s_{b_1}} > t_{n-2,\alpha}$.
  - Bài toán 2: Bác bỏ H_0 nếu $\dfrac{b_1 - \beta}{s_{b_1}} < -t_{n-2,\alpha}$.
  - Bài toán 3: Bác bỏ H_0 nếu $\left|\dfrac{b_1 - \beta}{s_{b_1}}\right| > t_{n-2,\alpha/2}$.

Tính toán một số đại lượng hồi quy trong R

- Để biết một số đại lượng liên quan đến hồi qui trong R, ta dùng hàm `summary(lm(Y ~ X))`
- Với dữ liệu về chi phí theo số hành khách, ta có thể tính toán:

```
> SoKhach = c(61, 63, 67, 69, 70,
74, 76, 81, 86, 91, 95, 97)
> ChiPhi = c(4.280, 4.080, 4.420, 4.170,
4.480, 4.300, 4.820, 4.700, 5.110, 5.130, 5.640, 5.560)
> summary(lm(ChiPhi ~ SoKhach))
```

Tính toán một số đại lượng hồi quy trong R

Kết quả trong *R* cho ta:

```
Call:
  lm(formula = ChiPhi ~ SoKhach)
Residuals:
      Min         1Q    Median        3Q       Max
 -0.28171  -0.14938   0.04101   0.13162   0.22741
Coefficients:
             Estimate  Std. Error  t value  Pr(>|t|)
 (Intercept) 1.569793    0.338083    4.643  0.000917  ***
    SoKhach  0.040702    0.004312    9.439  2.69e-06  ***
---
Signif.codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1' ' 1
Residual standard error: 0.1772 on 10 degrees of freedom
Multiple R-squared: 0.8991,   Adjusted R-squared: 0.889
F-statistic: 89.09 on 1 and 10 DF,   p-value: 2.692e-06
```

Tính toán một số đại lượng hồi quy trong R

Kết quả cho ta các thông tin sau:

- Hai hệ số của phương trình hồi qui tuyến tính mẫu: $b_0 = 1.569793$ và $b_1 = 0.040702$;
- Hệ số xác định $R^2 = 0.8991$;
- Sai số chuẩn của ước lượng $s_\varepsilon = 0.1772$;
- Độ lệch chuẩn của hệ số độ dốc $s_{b_1} = 0.004212$.
- Bài toán kiểm định độ dốc của đường hồi qui tổng thể bằng 0: $H_0 : \beta = 0, H_1 : \beta \neq 0$, có p-giá trị bằng 2.69e-06 và do đó ta bác bỏ giả thuyết H_0, chấp nhận H_1 tức là độ dốc của đường hồi qui tổng thể khác 0.

Thực hiện tính toán liên quan đến hồi quy trong R

- Muốn tìm khoảng tin cậy cho độ dốc của đường hồi qui Y theo X, ta dùng hàm `confint(lm(Y ~ X), level = 1 - α)`, trong đó `level` $= 1 - \alpha$ là tham số chỉ độ tin cậy bằng $1 - \alpha$, mặc định là 0.95.
- Với ví dụ về chi phi theo số hành khách, để tìm độ tin cậy 90% cho độ dốc của đường hồi qui tổng thể ta thực hiện lệnh:

```
> SoKhach = c(61, 63, 67, 69, 70,
74, 76, 81, 86, 91, 95, 97)
> ChiPhi = c(4.280, 4.080, 4.420, 4.170,
4.480, 4.300, 4.820, 4.700, 5.110, 5.130, 5.640, 5.560)
> confint(lm(ChiPhi ~ SoKhach), level = 0.9)
```

Thực hiện tính toán liên quan đến hồi quy trong R

- Kết quả trong *R* cho ta

	5 %	95 %
(Intercept)	0.95703055	2.18255500
SoKhach	0.03288603	0.04851716

- Kết quả này cho ta khoảng tin cậy 90% cho độ dốc của đường hồi qui tổng thể là $(0.03288603, 0.04851716)$.

1. Hồi quy tuyến tính đơn biến
   - Giới thiệu chung về hồi quy
   - Mô hình hồi quy tuyến tính đơn biến
   - Khả năng dự đoán của phương trình hồi quy tuyến tính
   - Suy diễn thống kê về hệ số độ dốc
   - Dự báo giá trị của biến phụ thuộc theo biến độc lập
2. Tương quan tuyến tính
   - Hệ số tương quan tuyến tính
   - Kiểm định hệ số tương quan tổng thể
3. Tương quan thứ hạng Spearman
   - Hệ số tương quan thứ hạng Spearman

Bài toán dự báo

Bài toán

Khi X nhận giá trị X_0 thì Y nhận giá trị thật tương ứng

$$Y_0 = \beta_0 + \beta_1 X_0 + \epsilon_0.$$

Khi X nhận giá trị X_0 thì Y nhận giá trị trung bình là $E(Y|X_0) = \beta_0 + \beta_1 X_0$. Ta sẽ dự báo giá trị thật và giá trị trung bình $E(Y|X_0$ theo ước lượng điểm và ước lượng khoảng.

Tính ước lượng điểm và khoảng của bài toán hồi quy trong R

- Để tính ước lượng điểm hoặc ước lượng khoảng ta dùng hàm `predict(lm(Y ~ X), newdata, interval = c("none", "confidence", "prediction"), level = 1 - α)`, trong đó
  - `newdata` là tham số chỉ giá trị mới cần tìm ước lượng;
  - `interval = c("none", "confidence", "prediction")` là tham số chỉ tương ứng ước lượng điểm, khoảng ước lượng cho giá trị trung bình và khoảng ước lượng cho giá trị thật;
  - `level` $= 1 - \alpha$ là tham số chỉ độ tin cậy nếu tìm ước lượng khoảng, mặc định là 0.95.

Tính ước lượng điểm và khoảng của bài toán hồi quy trong R

- Với ví dụ về chi phí theo số hành khách, để tìm một ước lượng điểm cho chi phí khi số hành khách là $X_0 = 75$, ta thực hiện lệnh:

```
> SoKhach = c(61, 63, 67, 69, 70,
74, 76, 81, 86, 91, 95, 97)
> ChiPhi = c(4.280, 4.080, 4.420, 4.170,
4.480, 4.300, 4.820, 4.700, 5.110, 5.130, 5.640, 5.560)
> predict(lm(ChiPhi SoKhach), data.frame(SoKhach = 75),
interval = "none")
hoặc
> predict(lm(ChiPhi ~ SoKhach), data.frame(SoKhach = 75)
```

- Kết quả trong *R* cho ta

```
[1]  4.622413
```

- Kết quả này cho ta $\hat{Y}_0 = 4.622413$.

Tính ước lượng điểm và khoảng của bài toán hồi quy trong R

- Để tìm một khoảng ước lượng 98% cho giá trị thật của chi phí khi số hành khách là $X_0 = 75$, ta thực hiện lệnh:

```
> SoKhach = c(61, 63, 67, 69, 70,
74, 76, 81, 86, 91, 95, 97)
> ChiPhi = c(4.280, 4.080, 4.420, 4.170,
4.480, 4.300, 4.820, 4.700, 5.110, 5.130, 5.640, 5.560)
> predict(lm(ChiPhi ~ SoKhach), data.frame(SoKhach = 75)
interval = "prediction", level = 0.98)
```

- Kết quả trong *R* cho ta

```
          fit       lwr      upr
[1,] 4.622413 4.111755 5.13307
```

- Kết quả này cho ta: khi số hành khách là 75, một ước lượng điểm cho chi phí là 4.622413 và khoảng tin cậy 98% cho giá trị thật của chi phí là $(4.111755, 5.13307)$.

Tính ước lượng điểm và khoảng của bài toán hồi quy trong R

- Để tìm một khoảng ước lượng 98% cho giá trị trung bình của chi phí khi số hành khách là $X_0 = 75$, ta thực hiện lệnh:

```
> SoKhach = c(61, 63, 67, 69, 70,
74, 76, 81, 86, 91, 95, 97)
> ChiPhi = c(4.280, 4.080, 4.420, 4.170,
4.480, 4.300, 4.820, 4.700, 5.110, 5.130, 5.640, 5.560)
> predict(lm(ChiPhi ~ SoKhach), data.frame(SoKhach = 75)
interval = "confidence", level = 0.98)
```

- Kết quả trong *R* cho ta

```
          fit       lwr       upr
[1,] 4.622413 4.477918 4.766907
```

- Kết quả này cho ta: khi số hành khách là 75, một ước lượng điểm cho chi phí là 4.622413 và khoảng tin cậy 98% cho giá trị trung bình của chi phí là $(4.477918, 4.766907)$.

Hệ số tương quan tuyến tính tổng thể

Định nghĩa

Hệ số tương quan giữa hai biến ngẫu nhiên X và Y với trung bình μ_X, μ_Y và phương sai σ_X^2, σ_Y^2, kí hiệu là ρ, được xác định bởi công thức:

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y},$$

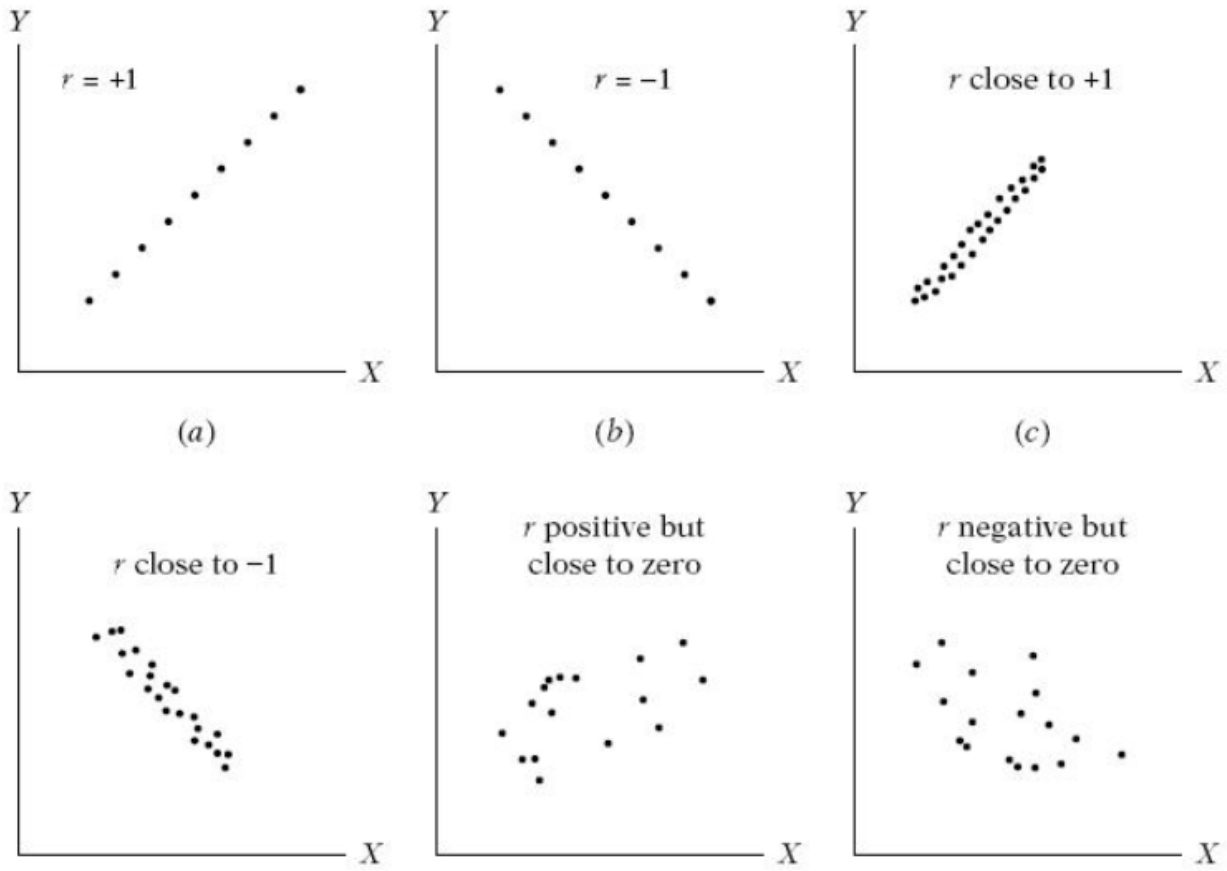
trong đó $Cov(X,Y)$ là tích sai giữa X, Y được xác định bởi công thức $Cov(X,Y) = E(XY) - \mu_X \mu_Y$.

Một số tính chất của hệ số tương quan

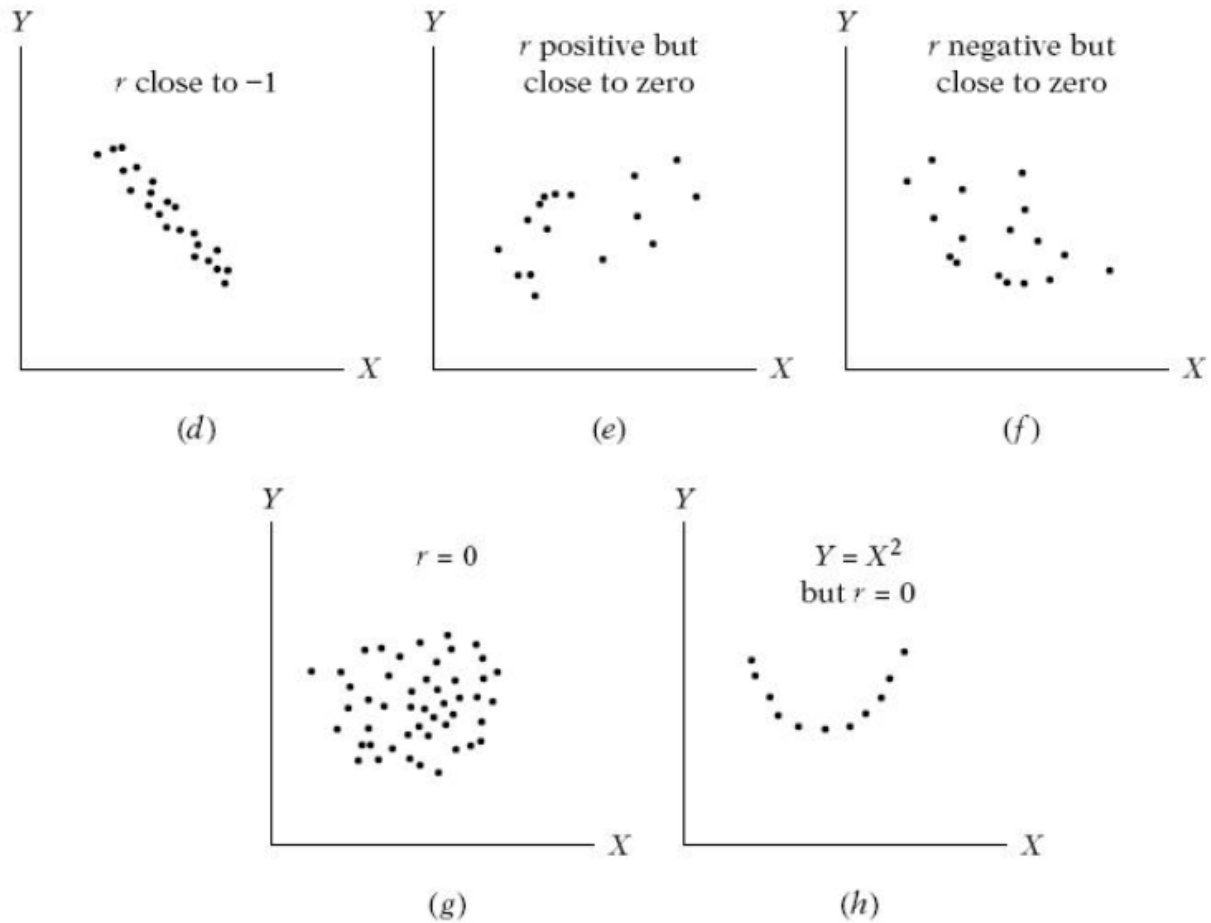
Hệ số tương quan có những tính chất sau:

- $-1 \leq \rho \leq 1$;
- Nếu $\rho = -1$ thì X, Y liên hệ tuyến tính hoàn toàn âm, tức là các điểm (x, y) nằm hoàn toàn trên một đường thẳng có độ dốc âm;
- Nếu $\rho = 1$ thì X, Y liên hệ tuyến tính hoàn toàn dương, tức là các điểm (x, y) nằm hoàn toàn trên một đường thẳng có độ dốc dương;
- Nếu $\rho < 0$ thì X, Y liên hệ tuyến tính âm;
- Nếu $\rho > 0$ thì X, Y liên hệ tuyến tính dương;
- Nếu $\rho = 0$ thì không có liên hệ tuyến tính giữa X và Y.
- $|x|$ càng lớn thì liên hệ tuyến tính càng mạnh.

Minh họa hệ số tương quan tuyến tính



Minh họa hệ số tương quan tuyến tính



Hệ số tương quan mẫu

Cho $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ là n cặp giá trị của các biến ngẫu nhiên X và Y. Khi đó hệ số tương quan tổng thể được ước lượng bằng hệ số tương quan mẫu (kí hiệu là r) được xác định bởi công thức

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}.$$

Ví dụ

Bài toán

Tính hệ số tương quan tuyến tính mẫu giữa chi phí và số khách cho bởi bảng số liệu sau:

Chuyến bay	Số khách	Chi phí (1000$)
1	61	4.280
2	63	4.080
3	67	4.420
4	69	4.170
5	70	4.480
6	74	4.300
7	76	4.820
8	81	4.700
9	86	5.110
10	91	5.130
11	95	5.640
12	97	5.560

Ví dụ

Lời giải:

- Gọi X, Y là hai biến ngẫu nhiên tương ứng chỉ số khách và chi phí trên mỗi chuyến bay. Khi đó hệ số tương quan tuyến tính mẫu r được tính theo công thức:

$$r = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2)(\sum_{i=1}^{n} y_i^2 - n\bar{y}^2)}}.$$

- Thay $\sum_{i=1}^{n} x_i y_i = 4462.22$, $\bar{x} = 77.5, \bar{y} = 4.72, n = 12$, $\sum_{i=1}^{n} x_i^2 = 73764, \sum_{i=1}^{n} y_i^2 = 270.9251$ vào ta được:

$$r = \frac{4462.22 - 12 \times 77.5 \times 4.72}{\sqrt{(73764 - 12 \times 77.5^2)(270.9251 - 12 \times 4.72^2)}} = 0.93.$$

Nhận xét

Hồi quy và tương quan khác nhau về mục đích và kĩ thuật:

- Phân tích tương quan đo mức độ liên hệ tuyến tính giữa hai biến và các biến có vai trò bình đẳng (đối xứng) với nhau;

Nhận xét

Hồi quy và tương quan khác nhau về mục đích và kĩ thuật:

- Phân tích tương quan đo mức độ liên hệ tuyến tính giữa hai biến và các biến có vai trò bình đẳng (đối xứng) với nhau;
- Phân tích hồi quy lại ước lượng hoặc dự báo một biến theo một một số biến khác và các biến không có tính chất đối xứng.

1. Hồi quy tuyến tính đơn biến
   - Giới thiệu chung về hồi quy
   - Mô hình hồi quy tuyến tính đơn biến
   - Khả năng dự đoán của phương trình hồi quy tuyến tính
   - Suy diễn thống kê về hệ số độ dốc
   - Dự báo giá trị của biến phụ thuộc theo biến độc lập

2. Tương quan tuyến tính
   - Hệ số tương quan tuyến tính
   - Kiểm định hệ số tương quan tổng thể

3. Tương quan thứ hạng Spearman
   - Hệ số tương quan thứ hạng Spearman

Kiểm định hệ số tương quan tổng thể

Bài toán

Khi hệ số tương quan tổng thể $\rho = 0$ thì hai biến ngẫu nhiên X và Y không có liên quan tuyến tính. Để xét mối liên hệ tuyến tính giữa hai biến ngẫu nhiên X, Y, ta thực hiện bài toán kiểm định hệ số tương quan tổng thể bằng không: $H_0 : \quad \rho = 0$.

Kiểm định hệ số tương quan tổng thể

Bài toán

Khi hệ số tương quan tổng thể $\rho = 0$ thì hai biến ngẫu nhiên X và Y không có liên quan tuyến tính. Để xét mối liên hệ tuyến tính giữa hai biến ngẫu nhiên X, Y, ta thực hiện bài toán kiểm định hệ số tương quan tổng thể bằng không: $H_0 : \quad \rho = 0$.

Nếu H_0 đúng và các biến ngẫu nhiên X, Y có phân phối xác suất hợp là phân phối chuẩn thì biến ngẫu nhiên

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}$$

tuân theo phân phối Student với $(n - 2)$ bậc tự do.

Quy luật quyết định trong bài toán kiểm định hệ số tương quan tổng thể

- Bài toán 1: $H_0 : \rho = 0, H_1 : \rho > 0$.
  Bác bỏ H_0 tại mức ý nghĩa α nếu $t = \dfrac{r}{\sqrt{(1-r^2)/(n-2)}} > t_{n-2,\alpha}$.
- Bài toán 2: $H_0 : \rho = 0, H_1 : \rho < 0$.
  Bác bỏ H_0 tại mức ý nghĩa α nếu $t = \dfrac{r}{\sqrt{(1-r^2)/(n-2)}} < -t_{n-2,\alpha}$.
- Bài toán 3: $H_0 : \rho = 0, H_1 : \rho \neq 0$.
  Bác bỏ H_0 tại mức ý nghĩa α nếu
  $|t| = \left|\dfrac{r}{\sqrt{(1-r^2)/(n-2)}}\right| > t_{n-2,\alpha/2}$.

Ví dụ về kiểm định hệ số tương quan giữa chi phí và số khách

- Trong ví dụ về chi phí theo số hành khách ta sẽ kiểm định mối liên hệ tuyến tính đồng biến giữa chi phí và số khách bằng cách kiểm định cặp giả thuyết $H_0 : \rho = 0, H_1 : \rho > 0$, ở mức ý nghĩa $\alpha = 1\%$.
- Tính giá trị thống kê:

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.933}{\sqrt{(1-(0.933)^2)/(12-2)}} = 8.2.$$

- Tính giá trị tới hạn $t_{n-2,\alpha} = t_{10,0.01} = 2.76$
- Do $8.2 > 2.76$ nên ta đưa ra quyết định bác bỏ H_0, tức là có mối liên hệ tuyến tính dương giữa chi phí và số hành khách.

1. Hồi quy tuyến tính đơn biến
   - Giới thiệu chung về hồi quy
   - Mô hình hồi quy tuyến tính đơn biến
   - Khả năng dự đoán của phương trình hồi quy tuyến tính
   - Suy diễn thống kê về hệ số độ dốc
   - Dự báo giá trị của biến phụ thuộc theo biến độc lập
2. Tương quan tuyến tính
   - Hệ số tương quan tuyến tính
   - Kiểm định hệ số tương quan tổng thể
3. Tương quan thứ hạng Spearman
   - Hệ số tương quan thứ hạng Spearman

Hệ số tương quan thứ hạng Spearman

- Hệ số tương quan tuyến tính đòi hỏi dữ liệu ít nhất phải là thang đo khoảng, trong trường hợp dữ liệu của chúng ta ở thang đo định danh hoặc thứ bậc, để tính mối liên hệ giữa hai biến ngẫu nhiên đo bằng thang đo định danh hoặc thứ bậc ta sử dụng hệ số tương quan thứ hạng Spearman.
- Cho $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ là n cặp giá trị của các biến ngẫu nhiên X và Y. Gọi $(u_1, v_1), (u_2, v_2), \ldots, (u_n, v_n)$ là n cặp thứ hạng tương ứng. Khi đó hệ số tương quan thứ hạng Spearman, kí hiệu là r_s được xác định bởi công thức sau:

$$r_s = \frac{\sum_{i=1}^{n} u_i v_i - n\bar{u}\bar{v}}{\sqrt{(\sum_{i=1}^{n} u_i^2 - n\bar{u})(\sum_{i=1}^{n} v_i^2 - n\bar{v})}}.$$

- Khi các cặp (x_i, x_j) và (y_i, y_j) đôi một khác nhau khi $i \neq j$ thì hệ số tương quan thứ hạng Spearman có thể tính bằng công thức:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

Ví dụ

Để nghiên cứu mối liên hệ giữa giá thịt bò và giá thịt cừu non, thu thập giá của hai loại thịt này từ năm 1988 đến năm 2000 ta thu được bảng dữ liệu, sắp xếp thứ hạng tương ứng cho ta kết quả sau:

Năm	Giá thịt bò	Giá thịt cừu	Hạng (thịt bò)	Hạng (thịt cừu)	d_i
1988	66.6	69.10	6	7	-1
1989	69.5	66.10	9	6	2
1990	74.60	55.50	13	2	11
1991	72.70	52.20	12	1	11
1992	71.30	59.50	10	3	7
1993	72.60	64.40	11	4	7
1994	66.70	65.60	7	5	2
1995	61.80	78.20	3	10	-7
1996	58.70	82.20	1	12	-11
1997	63.10	90.30	4	13	-9
1998	59.60	72.30	2	8	-6
1999	63.40	74.50	5	9	-4
2000	68.60	79.40	8	11	-3

Hãy tính hệ số tương quan thứ hạng spearman của giá thịt bò và giá thịt cừu.

Ví dụ

- Do các cặp giá trị của hai mẫu chọn ra đôi một khác nhau nên ta tính hệ số tương quan thứ hạng Spearman theo công thức:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}.$$

- Thay n $= 13$, $\sum d_i^2 = 666$ vào ta có:

$$r_s = 1 - \frac{666}{13(13^2-1)} = -0.830.$$

Thực hiện tính và kiểm định hệ số tương quan trên R

- Để tính hệ số tương quan ta dùng hàm `cor(x, y, method = c("pearson", "kendall", "spearman"))`
  trong đó,
  - x, y là hai véc tơ chỉ hai dữ liệu mẫu;
  - `method = c("pearson", "kendall", "spearman")` là tham số chỉ kiểu tính hệ số tương quan tương ứng là hệ số tương quan mẫu, hệ số tương quan kendall và hệ số tương quan thứ hạng spearman.
- Để thực hiện bài toán kiểm định hệ số tương quan, ta dùng hàm `cor.test(x, y, alternative = c("two.sided", "less", "greater"), method = c("pearson", "kendall", "pearman"), conf.level = 1 - α)`
  - x, y là hai véc tơ chỉ hai dữ liệu mẫu;
  - `alternative = c("two.sided", "less", "greater")` là tham số chỉ giả thuyết đối tương ứng là hai bên, nhỏ hơn và lớn hơn, mặc định là hai bên;
  - `method = c("pearson", "kendall", "spearman")` là tham số chỉ tương ứng kiểm định hệ số tương quan tổng thể, hệ số tương quan kendall và hệ số tương quan thứ hạng spearman.

Thực hiện tính và kiểm định hệ số tương quan trên R

```
> SoKhach = c(61, 63, 67, 69, 70, 74, 76, 81, 86, 91, 95, 97)
> ChiPhi = c(4.28, 4.08, 4.42, 4.17, 4.48, 4.30, 4.82, 4.70,
5.11, 5.13, 5.64, 5.56)
> cor(SoKhach, ChiPhi, method = "pearson")
[1] 0.9482003
```

Kết quả này cho ta hệ số tương quan mẫu $r = 0.9482003$.

Thực hiện tính và kiểm định hệ số tương quan trên R

```
> GiaThitBo = c(66.6, 69.50, 74.60, 72.70,
71.3, 72.6, 66.7, 61.8, 58.7, 63.1, 59.6, 63.4, 68.6)
> GiaThitCuu = c(69.1, 66.1, 55.5,
52.2, 59.5, 64.4, 65.6, 78.2, 82.2, 90.3, 72.3, 74.5, 79.4)
> cor(GiaThitBo, GiaThitCuu, method = "spearman")
[1] -0.8296703
```

Kết quả này cho ta hệ số tương quan thứ hạng spearman $r_s = -0.8296703$.

Thực hiện tính và kiểm định hệ số tương quan trên R

- Để kiểm định cặp giả thuyết liên quan đến hệ số tương quan tổng thể về chi phí và số khách: $H_0 : \rho = 0, H_1 : \rho > 0$, ở mức ý nghĩa $\alpha = 1\%$.
- ta thực hiện như sau:

```
> SoKhach = c(61, 63, 67, 69, 70, 74, 76, 81, 86,
91, 95, 97)
> ChiPhi = c(4.28, 4.08, 4.42, 4.17, 4.48, 4.30, 4.82,
4.70, 5.11, 5.13, 5.64, 5.56)
> cor.test(SoKhach,ChiPhi,method='pearson',alt='greater')
```

Thực hiện tính và kiểm định hệ số tương quan trên R

- kết quả trong *R* cho ta:

```
        Pearson's product-moment correlation
data: SoKhach and ChiPhi
t = 9.4389, df = 10, p-value = 1.346e-06
alternative hypothesis: true correlation is greater than
95 percent confidence interval:
  0.8525335 1.0000000
sample estimates:
      cor
0.9482003
```

- Với p-giá trị = 1.346e-06 nhỏ hơn $\alpha = 0.01$ nên ta đưa ra quyết định bác bỏ H_0 và chấp nhận hệ số tương quan tổng thể giữa chi phí và số khách là dương.