Chương 7: Hồi quy tuyến tính đơn

Bài giảng trực tuyến Xác suất - Thống kê

Hoàng Văn Hà
hvha@hcmus.edu.vn
Trường Đại học Khoa học tự nhiên
Đại học Quốc gia TP. Hồ Chí Minh

Nội dung

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

1 Giới thiệu bài toán hồi quy tuyến tính đơn

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,
- Độ giãn nở của một loại vật liệu theo nhiệt độ môi trường,

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,
- Độ giãn nở của một loại vật liệu theo nhiệt độ môi trường,
- Hàm lượng thuốc gây mê và thời gian ngủ của bệnh nhân,

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,
- Độ giãn nở của một loại vật liệu theo nhiệt độ môi trường,
- Hàm lượng thuốc gây mê và thời gian ngủ của bệnh nhân,
- Doanh thu khi bán 1 loại sản phẩm và số tiền chi cho quảng cáo và khuyến mãi,

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,
- Độ giãn nở của một loại vật liệu theo nhiệt độ môi trường,
- Hàm lượng thuốc gây mê và thời gian ngủ của bệnh nhân,
- Doanh thu khi bán 1 loại sản phẩm và số tiền chi cho quảng cáo và khuyến mãi,
- . . .

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,
- Độ giãn nở của một loại vật liệu theo nhiệt độ môi trường,
- Hàm lượng thuốc gây mê và thời gian ngủ của bệnh nhân,
- Doanh thu khi bán 1 loại sản phẩm và số tiền chi cho quảng cáo và khuyến mãi,
- . . .

Phân tích hồi quy

Bài toán: trong các hoạt động về khoa học - kỹ thuật, y học, kinh tế - xã hội, . . . ta có nhu cầu xác định mối liên giữa hai hay nhiều biến ngẫu nhiên với nhau.

Ví dụ:

- Mối liên hệ giữa chiều cao và cỡ giày của một người, từ đó một cửa hàng bán giầy dép có thể xác định chính xác cỡ giầy của một khách hàng khi biết chiều cao,
- Độ giãn nở của một loại vật liệu theo nhiệt độ môi trường,
- Hàm lượng thuốc gây mê và thời gian ngủ của bệnh nhân,
- Doanh thu khi bán 1 loại sản phẩm và số tiền chi cho quảng cáo và khuyến mãi,
- . . .

Để giải quyết các vấn đề trên, ta sử dụng kỹ thuật phân tích hồi quy (Regression Analysis).

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.
  - Mối liên hệ giữa $X_1, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính, tức là

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \text{ sai số }.$$

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.
  - Mối liên hệ giữa $X_1, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính, tức là

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \text{ sai số }.$$

  - Sự thay đổi trong Y được giả sử do những thay đổi trong $X_1, \ldots, X_p$ gây ra.

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.
  - Mối liên hệ giữa $X_1, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính, tức là

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \text{ sai số }.$$

  - Sự thay đổi trong Y được giả sử do những thay đổi trong $X_1, \ldots, X_p$ gây ra.

- Trên cơ sở xác định mối liên hệ giữa biến phụ thuộc Y và các biến giải thích X_1, X_2, $\ldots$, X_p, ta có thể:

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.
  - Mối liên hệ giữa $X_1, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính, tức là

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \text{ sai số }.$$

  - Sự thay đổi trong Y được giả sử do những thay đổi trong $X_1, \ldots, X_p$ gây ra.

- Trên cơ sở xác định mối liên hệ giữa biến phụ thuộc Y và các biến giải thích X_1, X_2, $\ldots$, X_p, ta có thể:
  - dự đoán, dự báo giá trị của Y,

Phân tích hồi quy

- Phân tích hồi quy được sử dụng để xác định mối liên hệ giữa:
  - một biến phụ thuộc Y, và
  - một hay nhiều biến độc lập $X_1, X_2, \ldots, X_p$. Các biến này còn được gọi là biến giải thích.
    - Biến phụ thuộc Y phải là biến liên tục (trong bối cảnh ta đang xét là hồi quy tuyến tính),
    - Các biến độc lập $X_1, X_2, \ldots, X_p$ có thể là biến liên tục, rời rạc hoặc phân loại.
  - Mối liên hệ giữa $X_1, \ldots, X_p$ và Y được biểu diễn bởi một hàm tuyến tính, tức là

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \text{ sai số }.$$

  - Sự thay đổi trong Y được giả sử do những thay đổi trong $X_1, \ldots, X_p$ gây ra.

- Trên cơ sở xác định mối liên hệ giữa biến phụ thuộc Y và các biến giải thích X_1, X_2, $\ldots$, X_p, ta có thể:
  - dự đoán, dự báo giá trị của Y,
  - giải thích tác động của sự thay đổi trong các biến giải thích lên biến phụ thuộc.

Mô hình hồi quy tuyến tính đơn

Định nghĩa 1

Một mô hình thống kê tuyến tính đơn (simple linear regression model) liên quan đến một biến ngẫu nhiên Y và một biến giải thích X là phương trình có dạng

$$Y = \beta_0 + \beta_1 X + \epsilon, \tag{1}$$

trong đó

- β_0, β_1 *là các tham số chưa biết, gọi là các hệ số hồi quy,*
- X *là biến độc lập, giải thích cho* Y*,*
- ϵ *là thành phần sai số.*

Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) được giả sử thỏa các điều kiện sau

Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) được giả sử thỏa các điều kiện sau
  - Các sai số ϵ_i độc lập với nhau,

Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) được giả sử thỏa các điều kiện sau
  - Các sai số ϵ_i độc lập với nhau,
  - $\mathbb{E}(\epsilon_i) = 0$ và $\mathbb{V}ar(\epsilon_i) = \sigma^2$,

Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) được giả sử thỏa các điều kiện sau
  - Các sai số ϵ_i độc lập với nhau,
  - $\mathbb{E}(\epsilon_i) = 0$ và $\mathbb{V}ar(\epsilon_i) = \sigma^2$,
  - Các sai số có phân phối chuẩn: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ với phương sai không đổi.

Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) được giả sử thỏa các điều kiện sau
  - Các sai số ϵ_i độc lập với nhau,
  - $\mathbb{E}(\epsilon_i) = 0$ và $\mathbb{V}ar(\epsilon_i) = \sigma^2$,
  - Các sai số có phân phối chuẩn: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ với phương sai không đổi.
- Cho trước $X = x$, ta có:

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x. \tag{2}$$

Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) được giả sử thỏa các điều kiện sau
  - Các sai số ϵ_i độc lập với nhau,
  - $\mathbb{E}(\epsilon_i) = 0$ và $\mathbb{V}ar(\epsilon_i) = \sigma^2$,
  - Các sai số có phân phối chuẩn: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ với phương sai không đổi.
- Cho trước $X = x$, ta có:

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x. \tag{2}$$

Các giả định về sai số ngẫu nhiên

- Các sai số ngẫu nhiên $\epsilon_i, i = 1, \ldots, n$ trong mô hình (6) được giả sử thỏa các điều kiện sau
  - Các sai số ϵ_i độc lập với nhau,
  - $\mathbb{E}(\epsilon_i) = 0$ và $\mathbb{V}ar(\epsilon_i) = \sigma^2$,
  - Các sai số có phân phối chuẩn: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ với phương sai không đổi.

- Cho trước $X = x$, ta có:

$$\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x. \tag{2}$$

Suy ra phân phối có điều kiện của Y cho trước $X = x$ là

$$Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2) \tag{3}$$

Mô hình hồi quy tuyến tính đơn

- Trong mô hình (5), sự thay đổi của Y được giả sử ảnh hưởng bởi 2 yếu tố:

Mô hình hồi quy tuyến tính đơn

- Trong mô hình (5), sự thay đổi của Y được giả sử ảnh hưởng bởi 2 yếu tố:
  - Mối liên hệ tuyến tính của X và Y: $\beta_0 + \beta_1 X$. Trong đó, β_0 được gọi là hệ số chặn (intercept) và β_1 gọi là hệ số góc (slope).

Mô hình hồi quy tuyến tính đơn

- Trong mô hình (5), sự thay đổi của Y được giả sử ảnh hưởng bởi 2 yếu tố:
  - Mối liên hệ tuyến tính của X và Y: $\beta_0 + \beta_1 X$. Trong đó, β_0 được gọi là hệ số chặn (intercept) và β_1 gọi là hệ số góc (slope).
  - Tác động của các yếu tố khác (không phải X): thành phần sai số ϵ.

Mô hình hồi quy tuyến tính đơn

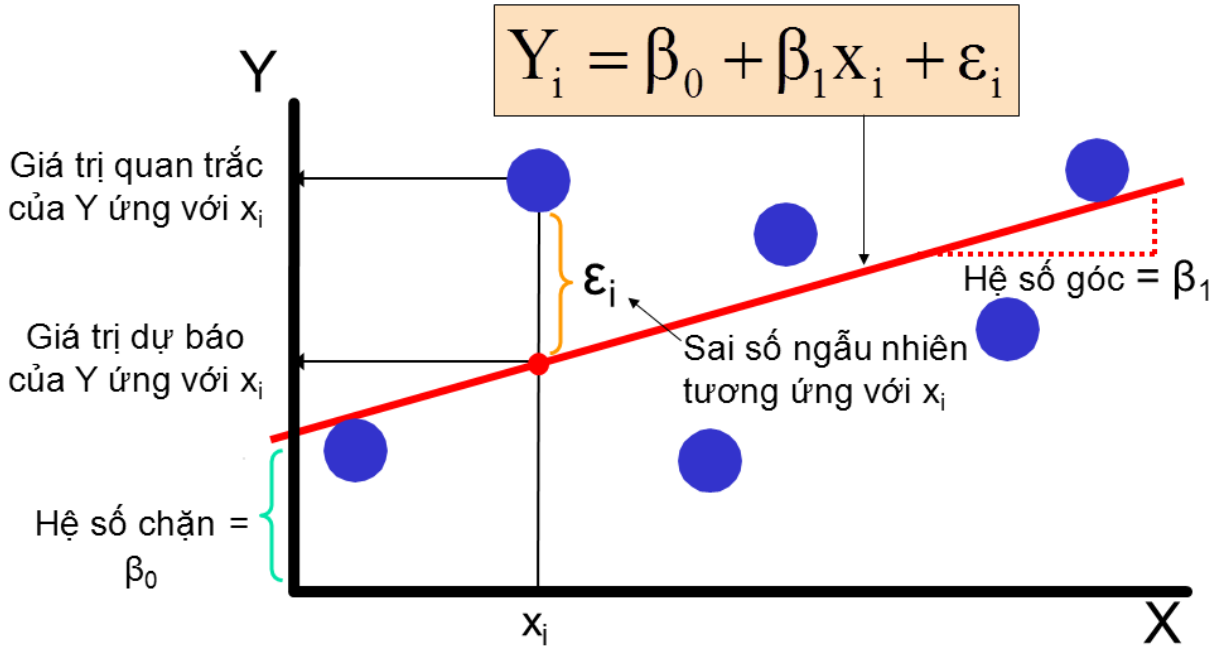
- Trong mô hình (5), sự thay đổi của Y được giả sử ảnh hưởng bởi 2 yếu tố:
  - Mối liên hệ tuyến tính của X và Y: $\beta_0 + \beta_1 X$. Trong đó, β_0 được gọi là hệ số chặn (intercept) và β_1 gọi là hệ số góc (slope).
  - Tác động của các yếu tố khác (không phải X): thành phần sai số ϵ.

- Với $(x_1, y_1), \ldots, (x_n, y_n)$ là n cặp giá trị quan trắc của một mẫu ngẫu nhiên cỡ n, từ (5) ta có

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{4}$$

Mô hình hồi quy tuyến tính đơn

- Sử dụng đồ thị phân tán (scatter plot) để biểu diễn các cặp giá trị quan trắc (x_i, y_i) trên hệ trục tọa độ Oxy.



ĐẠI HỌC
QUỐC GIA
TP. HỒ CHÍ MINH

2 Phương trình đường thẳng hồi quy

Ước lượng các hệ số hồi quy

- Gọi $\hat{\beta}_1$ và $\hat{\beta}_0$ là các ước lượng của β_0 và β_1.

Ước lượng các hệ số hồi quy

- Gọi $\hat{\beta}_1$ và $\hat{\beta}_0$ là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy với các hệ số ước lượng (fitted regression line):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{5}$$

Ước lượng các hệ số hồi quy

- Gọi $\hat{\beta}_1$ và $\hat{\beta}_0$ là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy với các hệ số ước lượng (fitted regression line):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{5}$$

- Một đường thẳng ước lượng tốt phải "gần với các điểm dữ liệu".

Ước lượng các hệ số hồi quy

- Gọi $\hat{\beta}_1$ và $\hat{\beta}_0$ là các ước lượng của β_0 và β_1.
- Đường thẳng hồi quy với các hệ số ước lượng (fitted regression line):

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x. \tag{5}$$

- Một đường thẳng ước lượng tốt phải "gần với các điểm dữ liệu".
- Tìm $\hat{\beta}_0$ và $\hat{\beta}_1$: dùng phương pháp bình phương bé nhất (method of least squares).

Phương pháp bình phương bé nhất

- Với dữ liệu $(x_i, y_i), i = 1, \ldots, n$, từ (5) ta có

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{6}$$

Phương pháp bình phương bé nhất

- Với dữ liệu $(x_i, y_i), i = 1, \ldots, n$, từ (5) ta có

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{6}$$

- Độ sai khác giữa giá trị quan trắc y_i và giá trị dự đoán $\hat{y}_i$ gọi là thặng dư (residual) thứ i, xác định như sau

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{7}$$

Phương pháp bình phương bé nhất

- Với dữ liệu $(x_i, y_i), i = 1, \ldots, n$, từ (5) ta có

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{6}$$

- Độ sai khác giữa giá trị quan trắc y_i và giá trị dự đoán $\hat{y}_i$ gọi là thặng dư (residual) thứ i, xác định như sau

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{7}$$

Phương pháp bình phương bé nhất

- Với dữ liệu $(x_i, y_i), i = 1, \ldots, n$, từ (5) ta có

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i. \tag{6}$$

- Độ sai khác giữa giá trị quan trắc y_i và giá trị dự đoán $\hat{y}_i$ gọi là thặng dư (residual) thứ i, xác định như sau

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{7}$$

Định nghĩa

Tổng bình phương sai số (Sum of Squares for Errors - SSE) hay tổng bình phương thặng dư cho n điểm dữ liệu được định nghĩa như sau

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right]^2. \tag{8}$$

Nội dung của phương pháp bình phương bé nhất là tìm các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ sao cho SSE đạt giá trị bé nhất.

Phương pháp bình phương bé nhất

Từ (8), lấy đạo hàm theo β_0 và β_1,

$$\frac{\partial \text{SSE}}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0,$$
$$\frac{\partial \text{SSE}}{\partial \beta_1} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] x_i = 0,$$

Phương pháp bình phương bé nhất

Từ (8), lấy đạo hàm theo β_0 và β_1,

$$\frac{\partial \text{SSE}}{\partial \beta_0} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] = 0,$$
$$\frac{\partial \text{SSE}}{\partial \beta_1} = -2 \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)] x_i = 0,$$

ta thu được hệ phương trình

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i, \tag{9}$$
$$\beta_0 \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

Ước lượng bình phương bé nhất

Giả hệ (9), ta tìm được các ước lượng bình phương bé nhất của β_0 và β_1 là

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}, \tag{10}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{11}$$

Ước lượng bình phương bé nhất

Giả hệ (9), ta tìm được các ước lượng bình phương bé nhất của β_0 và β_1 là

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}}{\sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}} = \frac{S_{xy}}{S_{xx}}, \tag{10}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{11}$$

với S_{xx} và S_{xy} xác định bởi

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - \frac{\left(\sum_{i=1}^n x_i\right)^2}{n}, \tag{12}$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{\left(\sum_{i=1}^n x_i\right)\left(\sum_{i=1}^n y_i\right)}{n}. \tag{13}$$

Ước lượng bình phương bé nhất

- Các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ tìm được gọi là các ước lượng bình phương bé nhất.
- Đường thẳng $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ gọi là đường thẳng bình phương bé nhất, thỏa các tính chất sau:

(1)

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

đạt giá trị bé nhất,

(2)

$$\text{SE} = \sum_{i=1}^{n} (y_i - \hat{y}_i) = \sum_{i=1}^{n} e_i = 0,$$

với SE là tổng các thặng dư (Sum of Errors).

Hồi quy tuyến tính đơn: ví dụ

Ví dụ

Một nhà thực vật học khảo sát mối liên hệ giữa tổng diện tích bề mặt (đv: cm^2) của các lá cây đậu nành và trọng lượng khô (đv: g) của các cây này. Nhà thực vật học trồng 13 cây trong nhà kính và đo tổng diện tích lá và trọng lượng của các cây này sau 16 ngày trồng, kết quả cho bởi bảng sau:

x_i	411	550	471	393	427	431	492	371	470	419	407	489	439
y_i	2.00	2.46	2.11	1.89	2.05	2.30	2.46	2.06	2.25	2.07	2.17	2.32	2.12

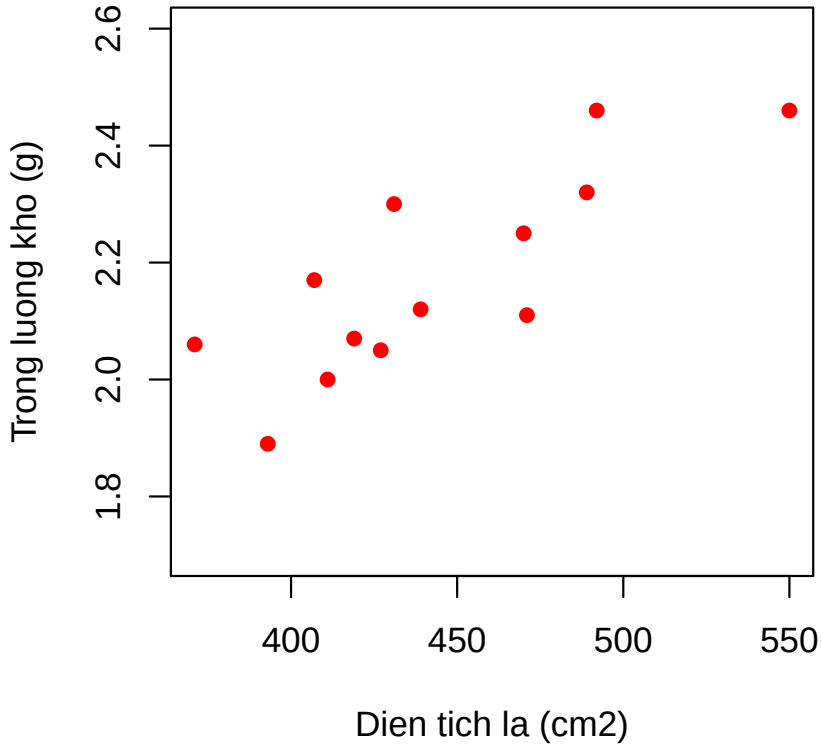
(a) Vẽ biểu đồ phân tán biểu diễn diện tích lá X và trọng lượng khô Y của cây đậu nành với mẫu quan sát đã cho.

(b) Tìm đường thẳng hồi quy biểu diễn mối liên hệ giữa trọng lượng cây Y theo diện tích lá X. Vẽ đường thẳng hồi quy tìm được trên đồ thị phân tán.

Hồi quy tuyến tính đơn: ví dụ

Giải:

(a) Vẽ đồ thị phân tán:



Hồi quy tuyến tính đơn: ví dụ

(b) Tìm đường thẳng hồi quy ước lượng

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1.$$

Nhắc lại, các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ được tính bởi:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

với

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\Sigma_{i=1}^{n} x_i\right)^2}{n},$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\Sigma_{i=1}^{n} x_i\right)\left(\Sigma_{i=1}^{n} y_i\right)}{n}.$$

Hồi quy tuyến tính đơn: ví dụ

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Hồi quy tuyến tính đơn: ví dụ

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Suy ra,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 443.8462, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.1738,$$

Hồi quy tuyến tính đơn: ví dụ

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Suy ra,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 443.8462, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.1738,$$

và

$$S_{xx} = 28465.69, \quad S_{xy} = 82.8977.$$

Hồi quy tuyến tính đơn: ví dụ

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Suy ra,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 443.8462, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.1738,$$

và

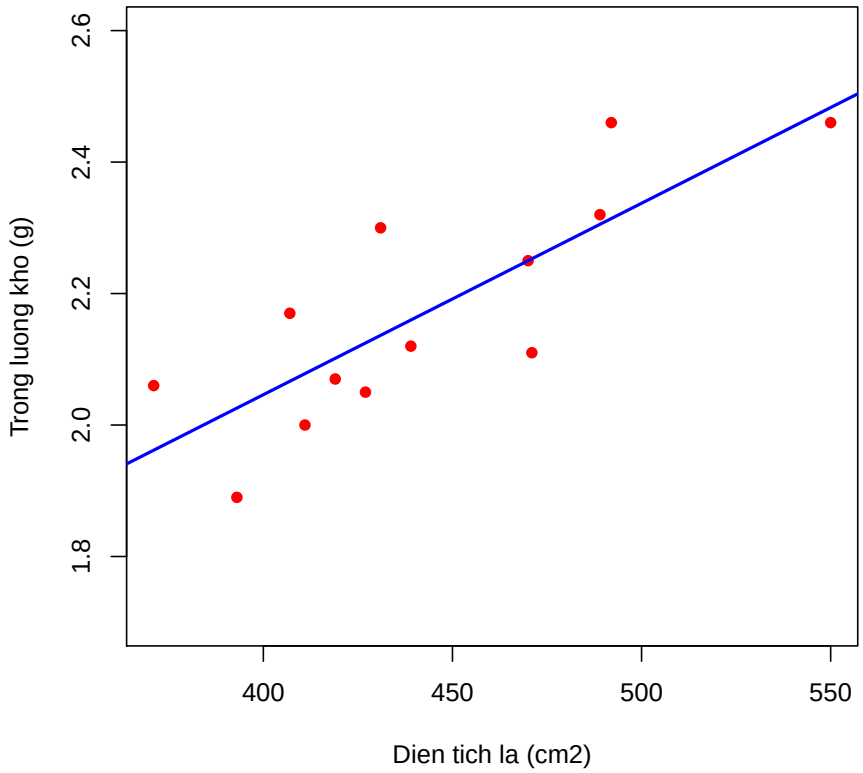
$$S_{xx} = 28465.69, \quad S_{xy} = 82.8977.$$

Ta tính được

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{82.8977}{28465.69} = 0.002912,$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_0 \bar{x} = 2.1738 - 0.002912 \times 443.8462 = 0.8813.$$

Hồi quy tuyến tính đơn: ví dụ

(b) Vẽ đường thẳng hồi quy ước lượng trên đồ thị phân tán:



ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

3 Tính chất của các ước lượng hệ số hồi quy

Tính chất của các ước lượng bình phương bé nhất

Định lý 1

Xét $Y = \beta_0 + \beta_1 X + \epsilon$ là một mô hình hồi quy tuyến tính đơn với $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Với n cặp giá trị quan trắc $(x_i, y_i), i = 1, \ldots, n$ ta có

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i.$$

Gọi $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng của β_0 và β_1 tìm được từ phương pháp bình phương bé nhất, khi đó

1. *$\hat{\beta}_0$ và $\hat{\beta}_1$ tuân theo luật phân phối chuẩn.*
2. *Kỳ vọng và phương sai của $\hat{\beta}_0$ và $\hat{\beta}_1$ lần lượt là*

$$\mathbb{E}(\hat{\beta}_0) = \beta_0, \ \mathbb{V}ar(\hat{\beta}_0) = \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\sigma^2, \tag{14}$$

$$\mathbb{E}(\hat{\beta}_1) = \beta_1, \ \mathbb{V}ar(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}}. \tag{15}$$

Ta có $\hat{\beta}_0$ và $\hat{\beta}_1$ lần lượt là các ước lượng không chệch cho β_0 và β_1.

Tính chất của các ước lượng bình phương bé nhất

Mệnh đề 1

Ước lượng của phương sai σ^2 của sai số của mô hình được cho bởi

$$\hat{\sigma}^2 = \frac{\mathrm{SSE}}{n-2}. \tag{16}$$

Ta cũng có $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$ tức là $\hat{\sigma}^2$ là một ước lượng không chệch cho σ^2.

Tính chất của các ước lượng bình phương bé nhất

Mệnh đề 1

Ước lượng của phương sai σ^2 của sai số của mô hình được cho bởi

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-2}. \tag{16}$$

Ta cũng có $\mathbb{E}[\hat{\sigma}^2] = \sigma^2$ tức là $\hat{\sigma}^2$ là một ước lượng không chệch cho σ^2.

Định nghĩa 2

Trong mô hình hồi quy tuyến tính đơn, sai số chuẩn (SE) của các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ là

$$\text{SE}(\hat{\beta}_0) = \sqrt{\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\hat{\sigma}^2}, \tag{17}$$

$$\text{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}, \tag{18}$$

Tính chất của các ước lượng bình phương bé nhất

Định lý 2 (Gauss - Markov)

Xét mô hình hồi quy tuyến tính đơn

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

có $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng bình phương bé nhất cho β_0 và β_1, khi đó $\hat{\beta}_0$ và $\hat{\beta}_1$ là các ước lượng không chệch tốt nhất.

ĐẠI HỌC
QUỐC GIA
TP. HỒ CHÍ MINH

4 Hệ số xác định R^2

Độ đo sự biến thiên của dữ liệu

Gọi:

- SST: Tổng bình phương toàn phần (Total Sum of Squares),

$$\mathrm{SST} = \sum_{i=1}^{n} (y_i - \bar{y})^2.$$

  SST còn được ký hiệu là S_{yy}.

- SSR: Tổng bình phương hồi quy (Regression Sum of Squares),

$$\mathrm{SSR} = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$$

- SSE: Tổng bình phương sai số (Error Sum of Squares),

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

Độ đo sự biến thiên của dữ liệu

- SST: đo sự biến thiên của các giá trị y_i xung quanh giá trị trung tâm của dữ liệu $\bar{y}$,
- SSR: giải thích sự biến thiên liên quan đến mối quan hệ tuyến tính của X và Y,
- SSE: giải thích sự biến thiên của các yếu tố khác (không liên quan đến mối quan hệ tuyến tính của X và Y).

Ta chứng tỏ được:

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y}_i)^2, \tag{19}$$

$$\mathrm{SST}=\mathrm{SSR}+\mathrm{SSE}\,.$$

Hệ số xác định

Định nghĩa 3

Hệ số xác định (Coefficient of Determination) là tỷ lệ của tổng sự biến thiên trong biến phụ thuộc gây ra bởi sự biến thiên của các biến độc lập (biến giải thích) so với tổng sự biến thiên toàn phần.
Hệ số xác định thường được gọi là R - bình phương (R-squared), ký hiệu là R^2.
Công thức tính:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSR}}{\text{SSR} + \text{SSE}}. \tag{20}$$

Chú ý: $0 \leq R^2 \leq 1$.

Hệ số xác định

Định nghĩa 3

Hệ số xác định (Coefficient of Determination) là tỷ lệ của tổng sự biến thiên trong biến phụ thuộc gây ra bởi sự biến thiên của các biến độc lập (biến giải thích) so với tổng sự biến thiên toàn phần.
Hệ số xác định thường được gọi là R - bình phương (R-squared), ký hiệu là R^2.
Công thức tính:

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSR}}{\text{SSR} + \text{SSE}}. \tag{20}$$

Chú ý: $0 \leq R^2 \leq 1$.

Hệ số xác định của một mô hình hồi quy cho phép ta đánh giá mô hình tìm được có giải thích tốt cho mối liên hệ giữa biến phụ thuộc Y và biến phụ thuộc X hay không.

Hệ số xác định

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSR}}{\text{SSR} + \text{SSE}}$$

Hệ số xác định

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSR}}{\text{SSR} + \text{SSE}}$$

- Tính SSR:

$$\begin{aligned}
\text{SSR} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \hat{\beta}_1^2 S_{xx} = \hat{\beta}_1 \hat{\beta}_1 S_{xx} = \hat{\beta}_1 \frac{S_{xy}}{S_{xx}} S_{xx} \\
&= \hat{\beta}_1 S_{xy}.
\end{aligned}$$

Hệ số xác định

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{\text{SSR}}{\text{SSR} + \text{SSE}}$$

- Tính SSR:

$$\begin{aligned}
\text{SSR} &= \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 = \sum_{i=1}^{n} (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\
&= \sum_{i=1}^{n} (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2 = \hat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 \\
&= \hat{\beta}_1^2 S_{xx} = \hat{\beta}_1 \hat{\beta}_1 S_{xx} = \hat{\beta}_1 \frac{S_{xy}}{S_{xx}} S_{xx} \\
&= \hat{\beta}_1 S_{xy}.
\end{aligned}$$

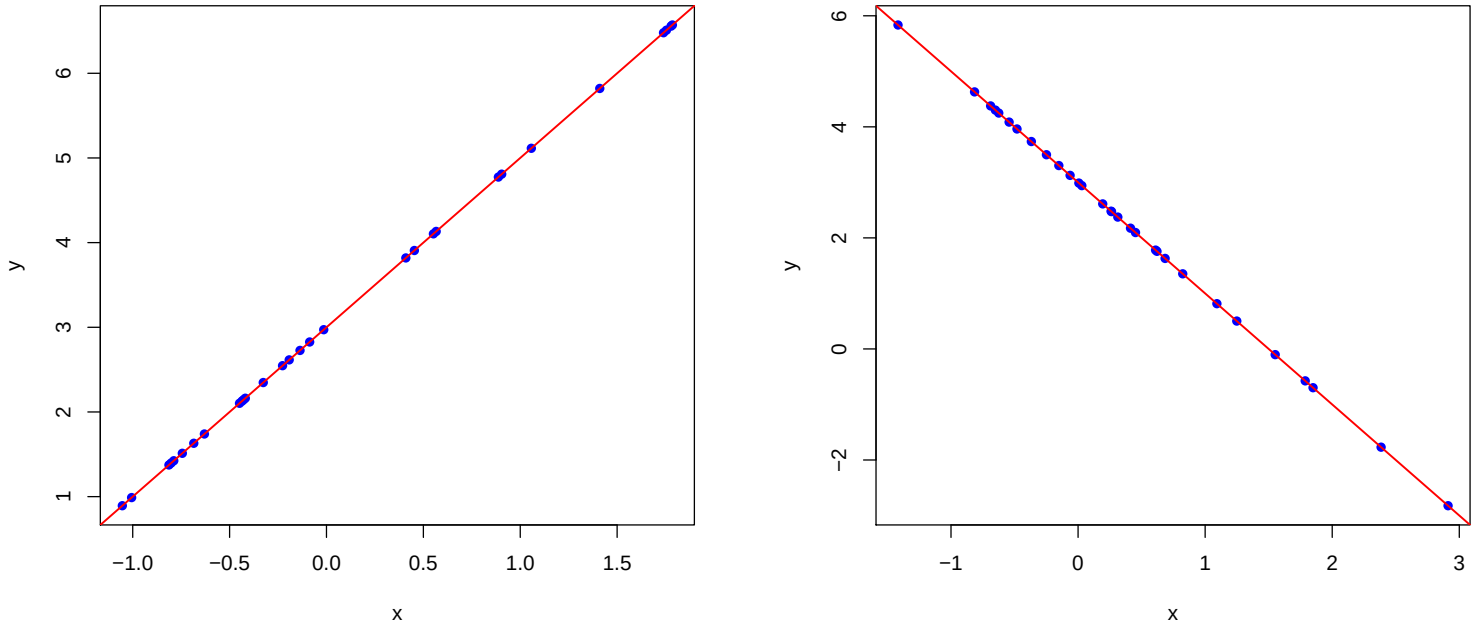
- Tính SSE:

$$\text{SSE} = \text{SST} - \text{SSR} = \text{SST} - \hat{\beta}_1 S_{xy},$$

với

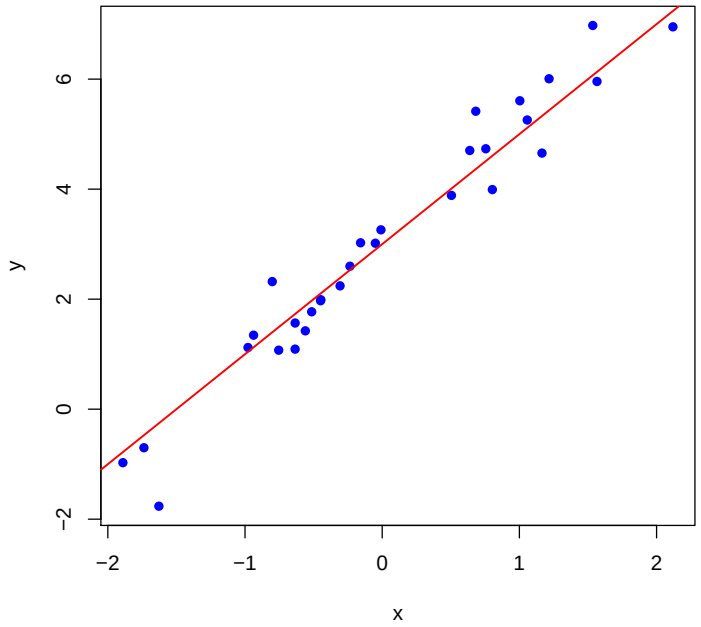
$$\text{SST} = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}.$$

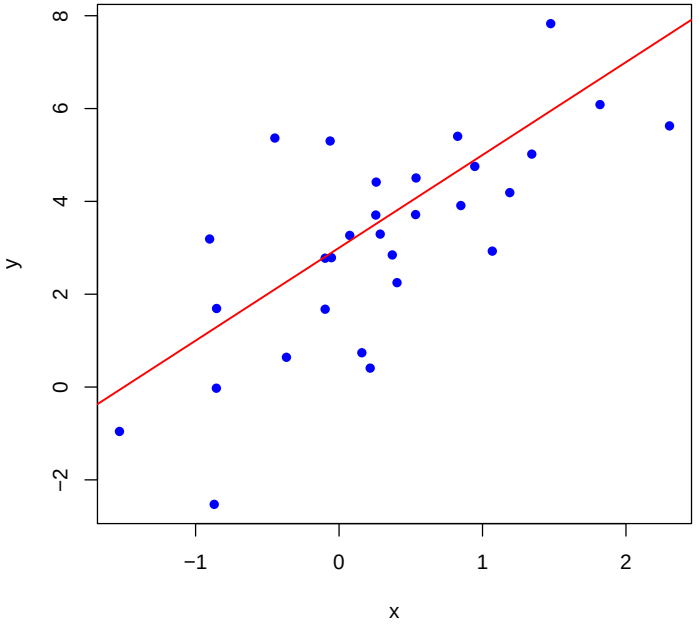
R^2 và mối liên hệ giữa X và Y



- $R^2 = 1$: X và Y có mối liên hệ tuyến tính hoàn hảo. 100% sự biến thiên của Y được giải thích bởi X.

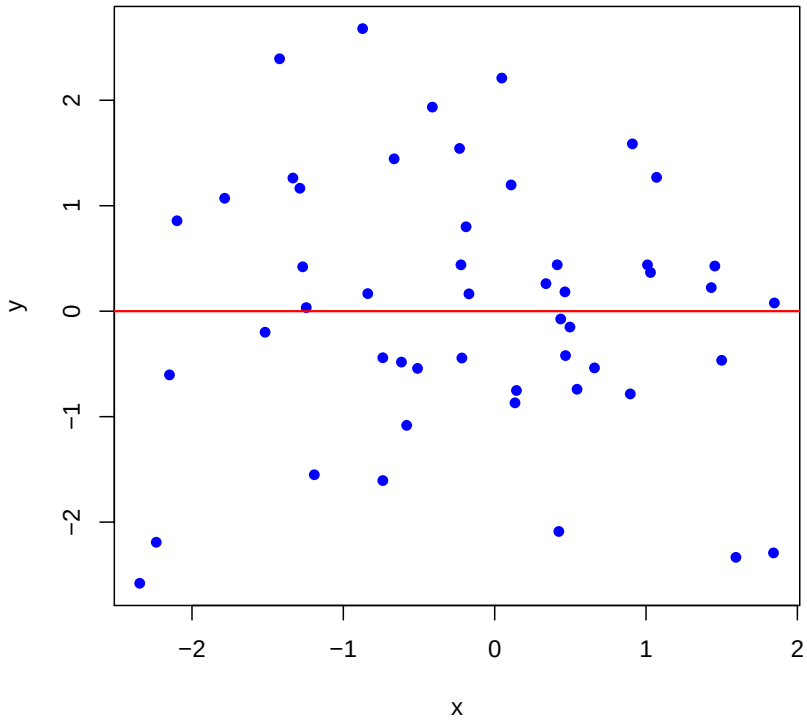
R^2 và mối liên hệ giữa X và Y





- R^2 càng gần 1 thì mối quan hệ tuyến tính giữa X và Y càng mạnh. Đa số sự biến thiên của Y được giải thích bởi X.
- R^2 càng gần 0 thì mối quan hệ tuyến tính giữa X và Y càng yếu. Sự biến thiên của Y càng ít được giải thích bởi X.

R^2 và mối liên hệ giữa X và Y



- $R^2 = 0$: không có mối liên hệ tuyến tính giữa X và Y. Không có sự biến thiên nào của Y được giải thích bởi X.

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

4 Hồi quy tuyến tính đơn: ví dụ

Hồi quy tuyến tính đơn: ví dụ

Ví dụ 1

Một nhà thực vật học khảo sát mối liên hệ giữa tổng diện tích bề mặt (đv: cm^2) của các lá cây đậu nành và trọng lượng khô (đv: g) của các cây này. Nhà thực vật học trồng 13 cây trong nhà kính và đo tổng diện tích lá và trọng lượng của các cây này sau 16 ngày trồng, kết quả cho bởi bảng sau:

x_i	411	550	471	393	427	431	492	371	470	419	407	489	439
y_i	2.00	2.46	2.11	1.89	2.05	2.30	2.46	2.06	2.25	2.07	2.17	2.32	2.12

(a) *Vẽ biểu đồ phân tán biểu diễn diện tích lá X và trọng lượng khô Y của cây đậu nành với mẫu quan sát đã cho.*

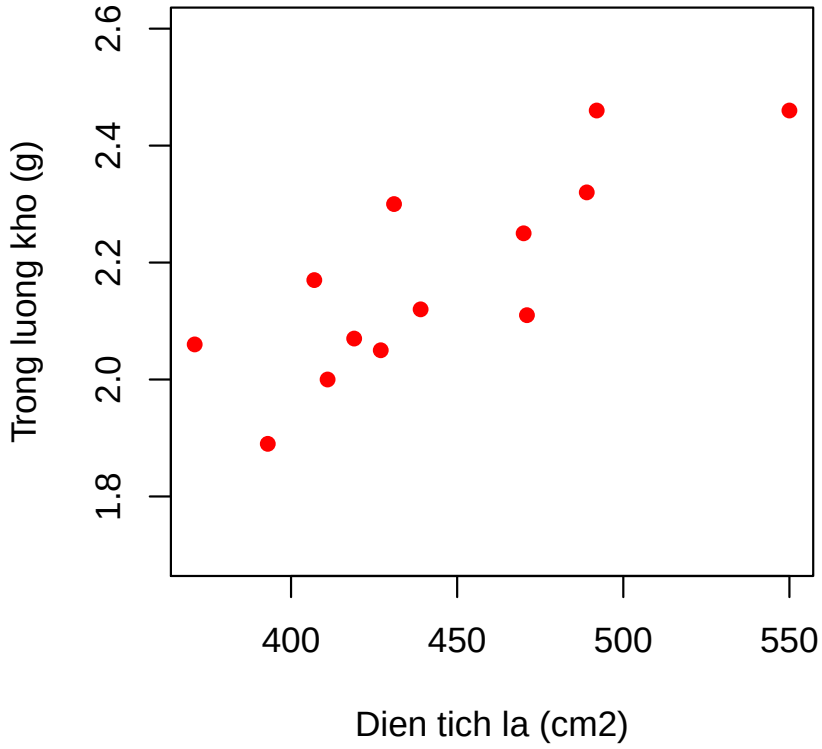
(b) *Tìm đường thẳng hồi quy biểu diễn mối liên hệ giữa trọng lượng cây Y theo diện tích lá X. Vẽ đường thẳng hồi quy tìm được trên đồ thị phân tán.*

(c) *Tính hệ số R^2 và nhận xét về mô hình.*

Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(a) Vẽ đồ thị phân tán:



Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(b) Tìm đường thẳng hồi quy ước lượng

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1.$$

Nhắc lại, các ước lượng $\hat{\beta}_0$ và $\hat{\beta}_1$ được tính bởi:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}},$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

với

$$S_{xx} = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n},$$
$$S_{xy} = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}.$$

Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Suy ra,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 443.8462, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.1738,$$

Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Suy ra,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 443.8462, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.1738,$$

và

$$S_{xx} = 28465.69, \quad S_{xy} = 82.8977.$$

Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(b) Từ bảng số liệu ta tính được

$$\sum_{i=1}^{n} x_i = 5770, \quad \sum_{i=1}^{n} x_i^2 = 2589458,$$
$$\sum_{i=1}^{n} y_i = 28.26, \quad \sum_{i=1}^{n} x_i y_i = 12625.99.$$

Suy ra,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = 443.8462, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 2.1738,$$

và

$$S_{xx} = 28465.69, \quad S_{xy} = 82.8977.$$

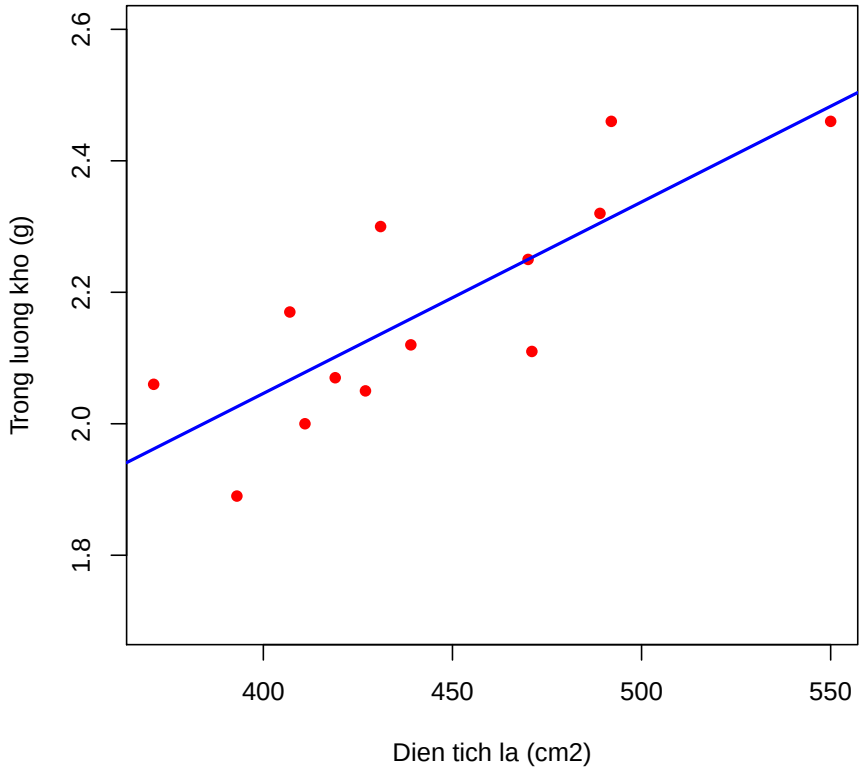
Ta tính được

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{82.8977}{28465.69} = 0.002912,$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_0 \bar{x} = 2.1738 - 0.002912 \times 443.8462 = 0.8813.$$

Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(b) Vẽ đường thẳng hồi quy ước lượng trên đồ thị phân tán:



Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(c) Tính hệ số xác định R^2: nhắc lại công thức tính hệ số xác định

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

Hồi quy tuyến tính đơn: ví dụ

Giải ví dụ 1:

(c) Tính hệ số xác định R^2: nhắc lại công thức tính hệ số xác định

$$R^2 = \frac{\text{SSR}}{\text{SST}},$$

với

$$\text{SST} = S_{yy} = \sum_{i=1}^{n} y_i^2 - \frac{\left(\Sigma_{i=1}^{n} y_i\right)^2}{n} = 0.3637,$$

và

$$\text{SSR} = \hat{\beta}_1 S_{xy} = 0.002912 \times 82.8977 = 0.2414.$$

Vậy:

$$R^2 = \frac{SSR}{SST} = \frac{0.2414}{0.3637} = 0.6637.$$

ĐẠI HỌC
QUỐC GIA
TP. HỒ CHÍ MINH

5 Phân tích thặng dư

Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.

Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:

Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:

1. Tuyến tính: mối quan hệ giữa X và Y là tuyến tính tính, tức là

$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$

Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:

1. Tuyến tính: mối quan hệ giữa X và Y là tuyến tính tính, tức là
$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$
2. Phương sai bằng nhau: phương sai của biến đáp ứng (biến phụ thuộc) Y là hằng số với mọi giá trị của biến độc lập X, tức là $\mathbb{V}\mathrm{ar}(Y|X = x) = \sigma^2$.

Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:

1. Tuyến tính: mối quan hệ giữa X và Y là tuyến tính tính, tức là
$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$
2. Phương sai bằng nhau: phương sai của biến đáp ứng (biến phụ thuộc) Y là hằng số với mọi giá trị của biến độc lập X, tức là $\mathbb{V}\mathrm{ar}(Y|X = x) = \sigma^2$.
3. Độc lập: các quan trắc của biến đáp ứng Y độc lập với nhau.

Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:

1. Tuyến tính: mối quan hệ giữa X và Y là tuyến tính tính, tức là
$$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$
2. Phương sai bằng nhau: phương sai của biến đáp ứng (biến phụ thuộc) Y là hằng số với mọi giá trị của biến độc lập X, tức là $\mathbb{V}\mathrm{ar}(Y|X = x) = \sigma^2$.
3. Độc lập: các quan trắc của biến đáp ứng Y độc lập với nhau.
4. Phân phối chuẩn: với mỗi giá trị của biến độc lập, phân phối có điều kiện (cho trước giá trị x) của biến đáp ứng là phân phối chuẩn, $Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.

Phân tích thặng dư

- Phân tích thặng dư (Residual Analysis) được sử dụng để kiểm tra các giả định của mô hình hồi quy tuyến tính.
- Các giả định của mô hình:
  1. Tuyến tính: mối quan hệ giữa X và Y là tuyến tính tính, tức là
  $$\mathbb{E}[Y|X = x] = \beta_0 + \beta_1 x.$$
  2. Phương sai bằng nhau: phương sai của biến đáp ứng (biến phụ thuộc) Y là hằng số với mọi giá trị của biến độc lập X, tức là $\mathbb{V}\mathrm{ar}(Y|X = x) = \sigma^2$.
  3. Độc lập: các quan trắc của biến đáp ứng Y độc lập với nhau.
  4. Phân phối chuẩn: với mỗi giá trị của biến độc lập, phân phối có điều kiện (cho trước giá trị x) của biến đáp ứng là phân phối chuẩn, $Y|X = x \sim \mathcal{N}(\beta_0 + \beta_1 x, \sigma^2)$.
- Việc kiểm tra các giả định trên thông thường sẽ được thực hiện thông qua các giá trị thặng dư, cho bởi
$$e_i = y_i - \hat{y}_i, \quad i = 1, \ldots, n,$$
với $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$.

Phân tích thặng dư

- Đồ thị các giá trị thặng dư: các cặp $(\hat{y}_i, e_i)$, $i = 1, \ldots, n$. (Hoặc ta vẽ các giá trị e_i tương ứng với các giá trị của biến độc lập x_i).

Phân tích thặng dư

- Đồ thị các giá trị thặng dư: các cặp $(\hat{y}_i, e_i)$, $i = 1, \ldots, n$. (Hoặc ta vẽ các giá trị e_i tương ứng với các giá trị của biến độc lập x_i).
- Nếu các giả định về 1, 2 và 3 thỏa thì ta sẽ nhận thấy đồ thị thặng dư gồm các điểm phân tán đều trên mặt phẳng Oxy và phân tán đề xung quanh đường thẳng $y = 0$.

Phân tích thặng dư

- Đồ thị các giá trị thặng dư: các cặp $(\hat{y}_i, e_i)$, $i = 1, \ldots, n$. (Hoặc ta vẽ các giá trị e_i tương ứng với các giá trị của biến độc lập x_i).
- Nếu các giả định về 1, 2 và 3 thỏa thì ta sẽ nhận thấy đồ thị thặng dư gồm các điểm phân tán đều trên mặt phẳng Oxy và phân tán đề xung quanh đường thẳng $y = 0$.
- Trường hợp một trong các giả định trên bị vi phạm, chẳng hạn như phương sai thay đổi, mối quan hệ giữa các biến không tuyến tính, ta sẽ thấy các điểm trên đồ thị thặng dư sẽ phân bố theo một hình dạng cụ thể nào đó.

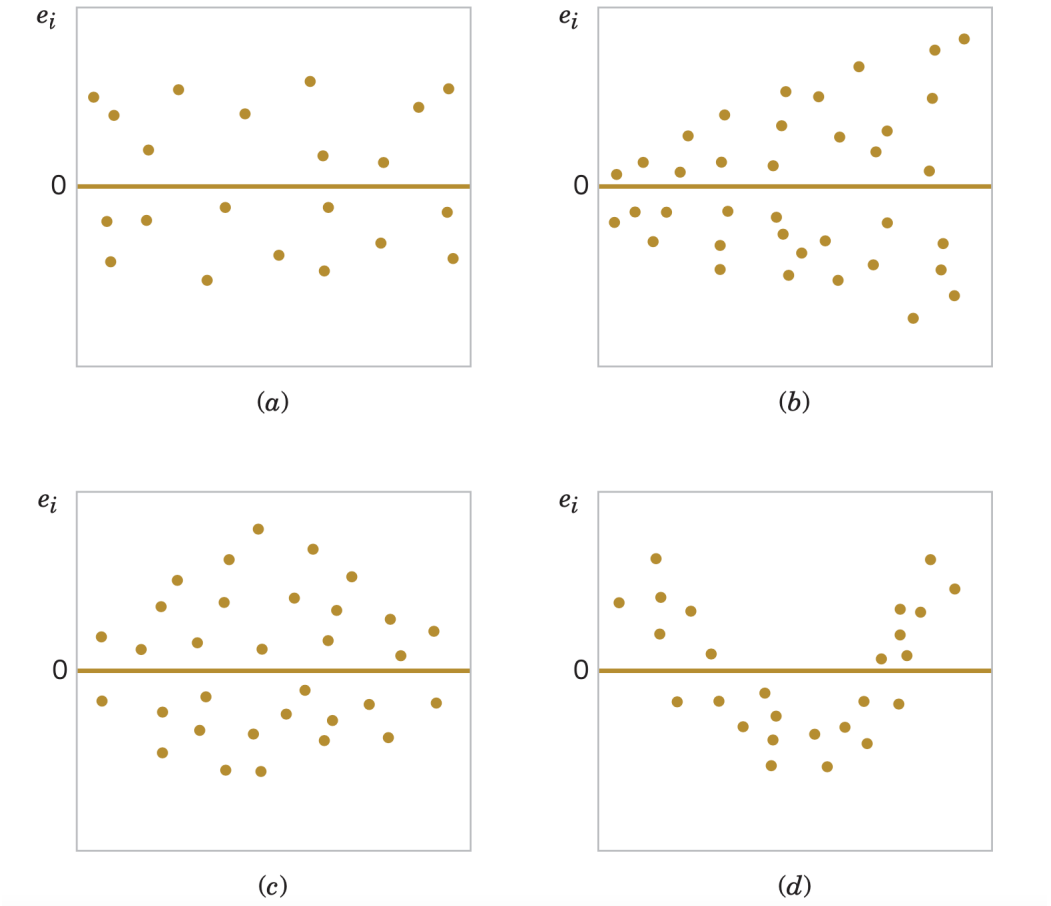
Phân tích thặng dư

- Đồ thị các giá trị thặng dư: các cặp $(\hat{y}_i, e_i)$, $i = 1, \ldots, n$. (Hoặc ta vẽ các giá trị e_i tương ứng với các giá trị của biến độc lập x_i).
- Nếu các giả định về 1, 2 và 3 thỏa thì ta sẽ nhận thấy đồ thị thặng dư gồm các điểm phân tán đều trên mặt phẳng Oxy và phân tán đề xung quanh đường thẳng $y = 0$.
- Trường hợp một trong các giả định trên bị vi phạm, chẳng hạn như phương sai thay đổi, mối quan hệ giữa các biến không tuyến tính, ta sẽ thấy các điểm trên đồ thị thặng dư sẽ phân bố theo một hình dạng cụ thể nào đó.
- Đồ thị thặng dư cũng giúp cho ta xác định được sự tồn tại của các điểm outlier.

Phân tích thặng dư

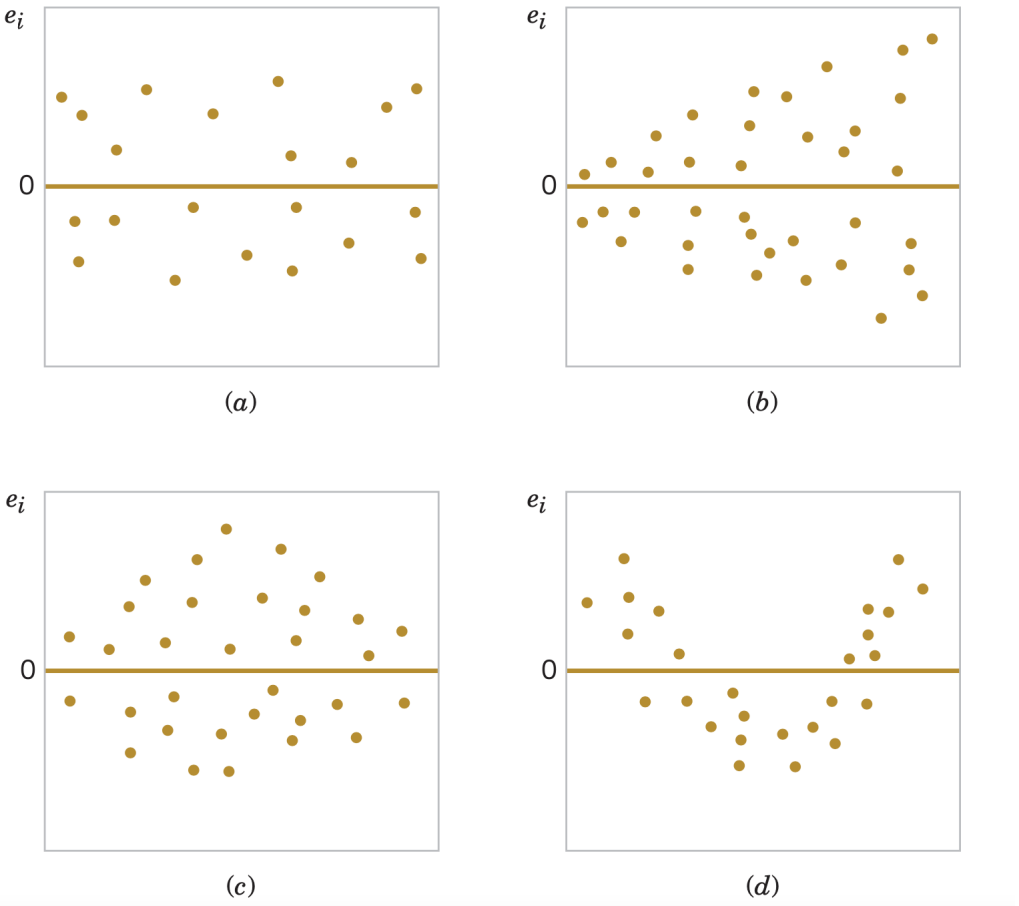
- Đồ thị các giá trị thặng dư: các cặp $(\hat{y}_i, e_i)$, $i = 1, \ldots, n$. (Hoặc ta vẽ các giá trị e_i tương ứng với các giá trị của biến độc lập x_i).
- Nếu các giả định về 1, 2 và 3 thỏa thì ta sẽ nhận thấy đồ thị thặng dư gồm các điểm phân tán đều trên mặt phẳng Oxy và phân tán đề xung quanh đường thẳng $y = 0$.
- Trường hợp một trong các giả định trên bị vi phạm, chẳng hạn như phương sai thay đổi, mối quan hệ giữa các biến không tuyến tính, ta sẽ thấy các điểm trên đồ thị thặng dư sẽ phân bố theo một hình dạng cụ thể nào đó.
- Đồ thị thặng dư cũng giúp cho ta xác định được sự tồn tại của các điểm outlier.
- Để kiểm tra giả định về phân phối chuẩn (giả định 4), ta thường dùng đồ thị Normal Q-Q Plot.

Phân tích thặng dư



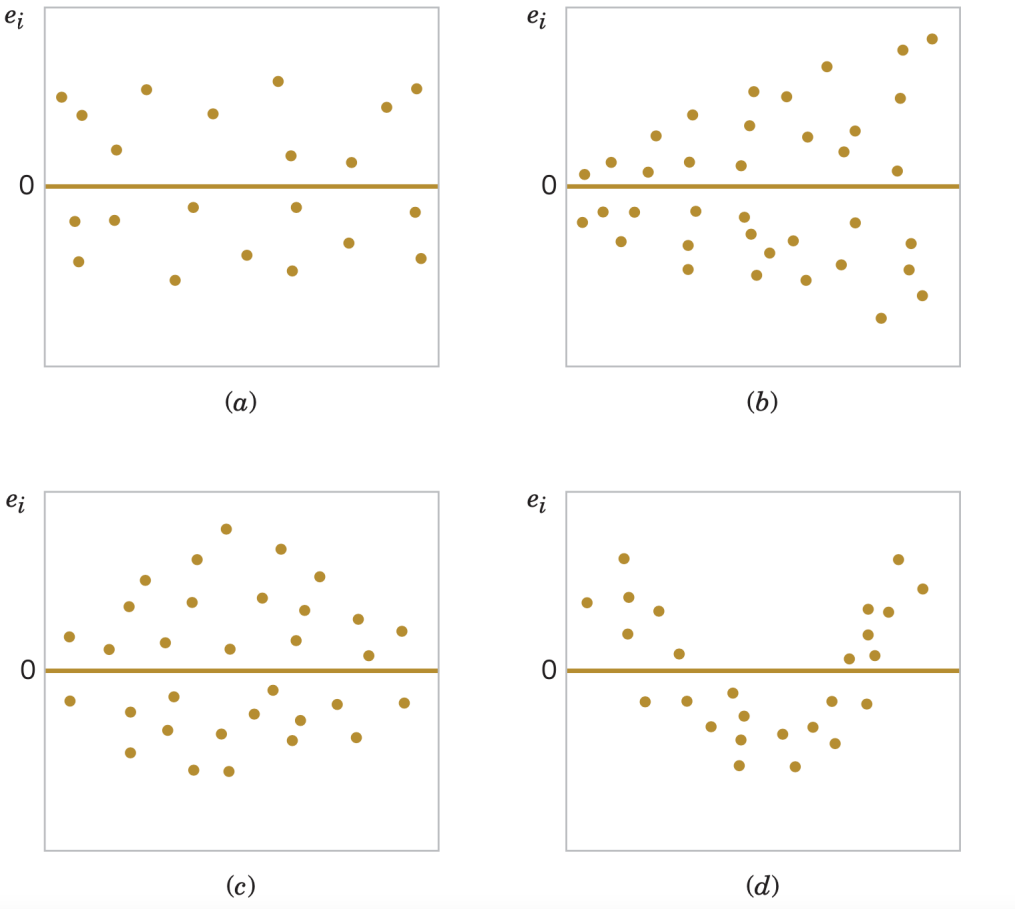
- (a): các giả định của mô hình được thỏa mãn.

Phân tích thặng dư



- (a): các giả định của mô hình được thỏa mãn.
- (b): phương sai tăng dần theo thời gian hoặc theo biên độ của x_i hay y_i. (c): phương sai không bằng nhau.

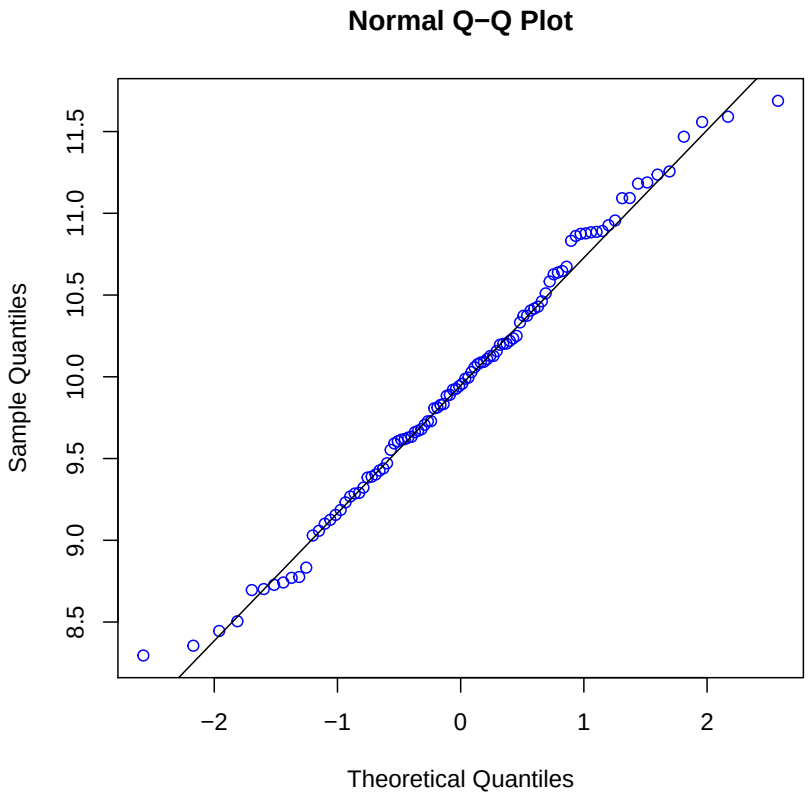
Phân tích thặng dư



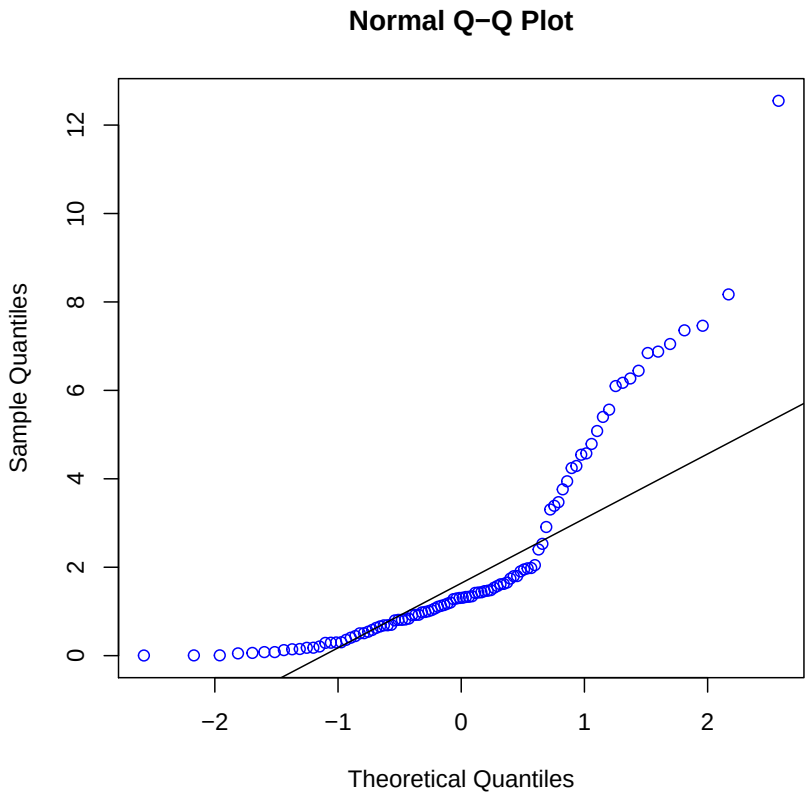
- (a): các giả định của mô hình được thỏa mãn.
- (b): phương sai tăng dần theo thời gian hoặc theo biên độ của x_i hay y_i. (c): phương sai không bằng nhau.
- (d): mối quan hệ giữa X và Y là phi tuyến tính.

Phân tích thặng dư

- Kiểm tra phân phối chuẩn sử dụng đồ thị Normal Q-Q Plot.



Dữ liệu tuân theo phân phối chuẩn



Dữ liệu không tuân theo phân phối chuẩn

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

6 Kiểm định giả thuyết cho các hệ số hồi quy

Kiểm định giả thuyết cho các hệ số hồi quy

Bài toán:

- Giả sử ta cần xây dựng một mô hình hồi quy với biến phụ thuộc Y và một tập các biến giải thích $X_1, X_2, \ldots, X_p$.

Kiểm định giả thuyết cho các hệ số hồi quy

Bài toán:

- Giả sử ta cần xây dựng một mô hình hồi quy với biến phụ thuộc Y và một tập các biến giải thích $X_1, X_2, \ldots, X_p$.
- Trong tập hợp các biến $X_1, X_2, \ldots, X_p$ này, có những biến giải thích tốt cho Y, cũng có thể có những biến không liên quan hoặc có mối liên hệ rất nhỏ với Y.

Kiểm định giả thuyết cho các hệ số hồi quy

Bài toán:

- Giả sử ta cần xây dựng một mô hình hồi quy với biến phụ thuộc Y và một tập các biến giải thích $X_1, X_2, \ldots, X_p$.
- Trong tập hợp các biến $X_1, X_2, \ldots, X_p$ này, có những biến giải thích tốt cho Y, cũng có thể có những biến không liên quan hoặc có mối liên hệ rất nhỏ với Y.
- Ta có thể xét mô hình hồi quy tuyến tính tổng quát (hồi quy bội):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

Kiểm định giả thuyết cho các hệ số hồi quy

Bài toán:

- Giả sử ta cần xây dựng một mô hình hồi quy với biến phụ thuộc Y và một tập các biến giải thích $X_1, X_2, \ldots, X_p$.
- Trong tập hợp các biến $X_1, X_2, \ldots, X_p$ này, có những biến giải thích tốt cho Y, cũng có thể có những biến không liên quan hoặc có mối liên hệ rất nhỏ với Y.
- Ta có thể xét mô hình hồi quy tuyến tính tổng quát (hồi quy bội):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

- Để xác định biến nào có ý nghĩa đối với mô hình, ta có thể thực hiện kiểm định giả thuyết đối với các hệ số hồi quy tương ứng, cụ thể,

$$H_0 : \beta_j = 0 \quad \text{với} \quad H_1 : \beta_j \neq 0,$$

với $j = 0, \ldots, p$.

Kiểm định giả thuyết cho các hệ số hồi quy

Bài toán:

- Giả sử ta cần xây dựng một mô hình hồi quy với biến phụ thuộc Y và một tập các biến giải thích $X_1, X_2, \ldots, X_p$.
- Trong tập hợp các biến $X_1, X_2, \ldots, X_p$ này, có những biến giải thích tốt cho Y, cũng có thể có những biến không liên quan hoặc có mối liên hệ rất nhỏ với Y.
- Ta có thể xét mô hình hồi quy tuyến tính tổng quát (hồi quy bội):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon.$$

- Để xác định biến nào có ý nghĩa đối với mô hình, ta có thể thực hiện kiểm định giả thuyết đối với các hệ số hồi quy tương ứng, cụ thể,

$$H_0 : \beta_j = 0 \quad \text{với} \quad H_1 : \beta_j \neq 0,$$

  với $j = 0, \ldots, p$.
- Trong nội dung chương trình học, ta đang khảo sát mô hình hồi quy tuyến tính đơn $Y = \beta_0 + \beta_1 X + \epsilon$, nên ta sẽ xét bài toán kiểm định giả thuyết cho β_0 và β_1.

Kiểm định giả thuyết cho β_0

- Bài toán kiểm định giả thuyết cho hệ số chặn β_0 trong mô hình hồi quy tuyến tính đơn như sau:

$$\begin{cases} H_0 : \beta_0 = b_0 \\ H_1 : \beta_0 \neq b_0 \end{cases}$$

với giá trị b_0 và mức ý nghĩa α cho trước. Thông thường $b_0 = 0$.

Kiểm định giả thuyết cho β_0

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,

Kiểm định giả thuyết cho β_0

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,

Kiểm định giả thuyết cho β_0

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{\mathrm{SE}(\hat{\beta}_0)}, \qquad \text{với } \mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

Kiểm định giả thuyết cho β_0

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{\mathrm{SE}(\hat{\beta}_0)}, \qquad \text{với } \mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_0}| > t^{n-2}_{1-\alpha/2}$.

Kiểm định giả thuyết cho β_0

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{\mathrm{SE}(\hat{\beta}_0)}, \qquad \text{với } \mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_0}| > t^{n-2}_{1-\alpha/2}$.
5. Kết luận: Bác bỏ H_0 / Chưa đủ cơ sở để bác bỏ H_0.

Kiểm định giả thuyết cho β_0

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_0} = \frac{\hat{\beta}_0 - b_0}{\mathrm{SE}(\hat{\beta}_0)}, \quad \text{với } \mathrm{SE}(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left(1 + \frac{\bar{x}^2}{S_{xx}}\right)}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_0}| > t^{n-2}_{1-\alpha/2}$.
5. Kết luận: Bác bỏ H_0 / Chưa đủ cơ sở để bác bỏ H_0.
6. Hoặc ta có thể sử dụng p-giá tri tính bởi

$$p = 2\mathbb{P}(T_{n-2} \geq |t_{\beta_0}|),$$

và bác bỏ H_0 khi $p \leq \alpha$.

Kiểm định giả thuyết cho β_0

- Bài toán kiểm định giả thuyết cho hệ số góc β_1 trong mô hình hồi quy tuyến tính đơn như sau:

$$\begin{cases} H_0 : \beta_1 = b_1 \\ H_1 : \beta_1 \neq b_1 \end{cases}$$

với giá trị b_1 và mức ý nghĩa α cho trước. Thông thường $b_1 = 0$.

Kiểm định giả thuyết cho β_1

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,

Kiểm định giả thuyết cho β_1

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,

Kiểm định giả thuyết cho β_1

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}, \quad \text{với } \mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

Kiểm định giả thuyết cho β_1

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}, \quad \text{với } \mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_1}| > t^{n-2}_{1-\alpha/2}$.

Kiểm định giả thuyết cho β_1

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}, \quad \text{với } \mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_1}| > t^{n-2}_{1-\alpha/2}$.
5. Kết luận: Bác bỏ H_0 / Chưa đủ cơ sở để bác bỏ H_0.

Kiểm định giả thuyết cho β_1

Các bước kiểm định

1. Phát biểu giả thuyết H_0 và đối thuyết H_1,
2. Xác định mức ý nghĩa α,
3. Tính giá trị thống kê kiểm định:

$$t_{\beta_1} = \frac{\hat{\beta}_1 - b_1}{SE(\hat{\beta}_1)}, \quad \text{với } \mathrm{SE}(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}.$$

4. Bác bỏ H_0 khi: $|t_{\beta_1}| > t^{n-2}_{1-\alpha/2}$.
5. Kết luận: Bác bỏ H_0 / Chưa đủ cơ sở để bác bỏ H_0.
6. Hoặc ta có thể sử dụng p-giá trị tính bởi

$$p = 2\mathbb{P}(T_{n-2} \geq |t_{\beta_1}|),$$

và bác bỏ H_0 khi $p \leq \alpha$.

ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH

3 Hệ số tương quan

Hệ số tương quan mẫu

Định nghĩa

Với mẫu cỡ n: $(x_i, y_i), i = 1, \ldots, n$, hệ số tương quan mẫu, ký hiệu r_{XY}, được xác định như sau

$$r_{XY} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \, \Sigma_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{21}$$

Hệ số tương quan mẫu

Định nghĩa

Với mẫu cỡ n: $(x_i, y_i), i = 1, \ldots, n$, hệ số tương quan mẫu, ký hiệu r_{XY}, được xác định như sau

$$r_{XY} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \, \Sigma_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{21}$$

- Ta có: $-1 \leq r_{XY} \leq 1$.

Hệ số tương quan mẫu

Định nghĩa

Với mẫu cỡ n: $(x_i, y_i), i = 1, \ldots, n$, hệ số tương quan mẫu, ký hiệu r_{XY}, được xác định như sau

$$r_{XY} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \, \Sigma_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}. \tag{21}$$

- Ta có: $-1 \leq r_{XY} \leq 1$.
- $-1 \leq r_{XY} < 0$: tương quan âm. r_{XY} càng gần -1 biểu thị mối liên hệ tuyến tính nghịch giữa X và Y càng mạnh.

Hệ số tương quan mẫu

Định nghĩa

Với mẫu cỡ n: $(x_i, y_i), i = 1, \ldots, n$, hệ số tương quan mẫu, ký hiệu r_{XY}, được xác định như sau

$$r_{XY} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \, \Sigma_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{21}$$

- Ta có: $-1 \leq r_{XY} \leq 1$.
- $-1 \leq r_{XY} < 0$: tương quan âm. r_{XY} càng gần -1 biểu thị mối liên hệ tuyến tính nghịch giữa X và Y càng mạnh.
- $0 < r_{XY} \leq 1$: tương quan dương. r_{XY} càng gần 1 biểu thị mối liên hệ tuyến tính thuận giữa X và Y càng mạnh.

Hệ số tương quan mẫu

Định nghĩa

Với mẫu cỡ n: $(x_i, y_i), i = 1, \ldots, n$, hệ số tương quan mẫu, ký hiệu r_{XY}, được xác định như sau

$$r_{XY} = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma_{i=1}^{n}(x_i - \bar{x})^2 \, \Sigma_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}. \tag{21}$$

- Ta có: $-1 \leq r_{XY} \leq 1$.
- $-1 \leq r_{XY} < 0$: tương quan âm. r_{XY} càng gần -1 biểu thị mối liên hệ tuyến tính nghịch giữa X và Y càng mạnh.
- $0 < r_{XY} \leq 1$: tương quan dương. r_{XY} càng gần 1 biểu thị mối liên hệ tuyến tính thuận giữa X và Y càng mạnh.
- r_{XY} càng gần 0, biểu thị mối liên hệ tuyến tính yếu. $r_{XY} = 0$: không có mối liên hệ tuyến tính giữa X và Y.

Mối quan hệ giữa hệ số tương quan mẫu và hệ số xác định

- Chú ý rằng,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \sqrt{\frac{S_{yy}}{S_{xx}}}\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} = \sqrt{\frac{S_{yy}}{S_{xx}}}r_{XY} = \sqrt{\frac{SST}{S_{xx}}}r_{XY},$$

  vì S_{yy} chính là SST.

- Suy ra,

$$r_{XY}^2 = \hat{\beta}_1{}^2\frac{S_{xx}}{\mathrm{SST}} = \hat{\beta}_1\frac{S_{xy}}{S_{xx}}\frac{S_{xx}}{\mathrm{SST}} = \frac{\hat{\beta}_1 S_{xy}}{\mathrm{SST}} = \frac{\mathrm{SSR}}{\mathrm{SST}} = R^2.$$

- Vậy, hệ số xác định R^2 của mô hình hồi quy tuyến tính đơn bằng với bình phương của hệ số tương quan mẫu

$$R^2 = r_{XY}^2.$$

ĐẠI HỌC
QUỐC GIA
TP. HỒ CHÍ MINH

8 Bài tập

Bài tập

Bài tập 1

Trong một bài báo về Nghiên cứu Bê tông: "Đặc tính bề mặt gần bê tông: tính thấm nội tại" trình bày dữ liệu về cường độ nén (X) và độ thấm nội tại (Y) của các hỗn hợp bê tông và phương pháp xử lý khác nhau. Số liệu được tóm tắt như sau:

$$n = 14,\ \sum_{i=1}^{n} x_i = 43,\ \sum_{i=1}^{n} x_i^2 = 157.42,\ \sum_{i=1}^{n} y_i = 572,$$
$$\sum_{i=1}^{n} y_i^2 = 23530,\ \sum_{i=1}^{n} x_i y_i = 1697.80.$$

(a) Xác định đường thẳng hồi quy ước lượng mô tả mối quan hệ tuyến tính giữa cường độ nén và độ thấm nội tại của bê tông.

(b) Ước lượng phương sai σ^2 của sai số.

(c) Sử dụng đường thẳng hồi quy ước lượng, hãy tiên đoán độ thấm nội tại của bê tông khi cường độ nén $x_0 = 4.3$?

(d) Tính hệ số xác định R^2 và cho nhận xét về mối liên hệ giữa X và Y.

Bài tập

Bài tập 2

Xét mẫu gồm 10 cặp giá trị (x_i, y_i) cho bởi bảng

x_i	-1	0	2	-2	5	6	8	11	12	-3
y_i	-5	-4	2	-7	6	9	13	21	20	-9

(a) Vẽ biểu đồ phân tán cho dữ liệu, tìm đường thẳng hồi quy ước lượng.
(b) Tìm ước lượng $\hat{\sigma}^2$ cho phương sai σ^2 của sai số ngẫu nhiên.
(c) Tính hệ số xác đinh R^2 và hệ số tương quan mẫu r_{XY}.
(d) Thực hiện kiểm định giả thuyết cho hệ số β_1.

Bài tập

Bài tập 3

Một nghiên cứu ảnh hưởng việc gia tăng liều dùng X (mg/kg) của một loại thuốc ngủ trên thời gian ngủ Y (giờ). Kết quả thực nghiệm ghi nhận được như sau:

x_i	1	1	2	2	3	4	5	5
y_i	1	1.2	1.5	1.7	2	2.2	2.5	2.2

(a) Tìm phương trình hồi quy của Y theo X.

(b) Tìm $\hat{\sigma}^2$ và hệ số xác định R^2.

(c) Nếu liều dùng thuốc ngủ là $x_0 = 4$ (mg/kg), thì thời gian ngủ dự đoán bằng bao nhiêu?

(d) Có tài liệu cho biết phương trình hồi quy của Y theo X là $y = 0.29x + 0.93$. Hỏi kết quả quan sát có phù hợp với phương trình cho biết không? $\alpha = 0.05$.