DỰ BÁO: MÔ HÌNH HỒI QUY (PREDICTION: REGRESSION MODELS)

Nguyễn Văn Hạnh

AI Academy Vietnam

Tháng 8 năm 2024

Nội dung

1. Giới thiệu mô hình hồi quy
2. Ước lượng tham số mô hình: Phương pháp bình phương nhỏ nhất
3. Kiểm định trong mô hình hồi quy tuyến tính bội.
4. Đánh giá và lựa chọn mô hình

Giới thiệu

- Trong nhiều lĩnh vực chúng ta cần nghiên cứu về quan hệ giữa các biến với nhau.

Giới thiệu

- Trong nhiều lĩnh vực chúng ta cần nghiên cứu về quan hệ giữa các biến với nhau.
- Phương pháp phân tích tương quan và hồi quy được áp dụng.

Giới thiệu

- Trong nhiều lĩnh vực chúng ta cần nghiên cứu về quan hệ giữa các biến với nhau.
- Phương pháp phân tích tương quan và hồi quy được áp dụng.
- Phân tích tương quan (correlation) là phương pháp nghiên cứu mối quan hệ tuyến tính giữa 2 biến dựa trên đo lường mức độ quan hệ hay cường độ quan hệ tuyến tính.

Giới thiệu

- Trong nhiều lĩnh vực chúng ta cần nghiên cứu về quan hệ giữa các biến với nhau.
- Phương pháp phân tích tương quan và hồi quy được áp dụng.
- Phân tích tương quan (correlation) là phương pháp nghiên cứu mối quan hệ tuyến tính giữa 2 biến dựa trên đo lường mức độ quan hệ hay cường độ quan hệ tuyến tính.
- Phân tích hồi quy (Regression) là phương pháp nghiên cứu mối quan hệ giữa 2 hay nhiều biến, mà cụ thể một hay nhiều biến sẽ là biến độc lập (ảnh hưởng đến biến mục tiêu), và biến còn lại sẽ là biến mục tiêu (bị ảnh hưởng bởi biến độc lập), mô hình hoá, định lượng hoá mối quan hệ này để qua đó xác định được giá trị của biến mục tiêu nếu các biến độc lập thay đổi như thế nào.

Phân tích hồi quy

- Mô hình hồi quy để định lượng hoá mối quan hệ giữa biến mục tiêu Y và các biến độc lập $(X_1, X_2, \ldots, X_k)$ được cho bởi phương trình:

$$Y = f(X_1, X_2, \ldots, X_k) + \epsilon = f(X) + \epsilon$$

Phân tích hồi quy

- Mô hình hồi quy để định lượng hoá mối quan hệ giữa biến mục tiêu Y và các biến độc lập $(X_1, X_2, \ldots, X_k)$ được cho bởi phương trình:

$$Y = f(X_1, X_2, \ldots, X_k) + \epsilon = f(X) + \epsilon$$

Trong đó:

  - f là một hàm toán học định lượng hoá mối quan hệ

Phân tích hồi quy

- Mô hình hồi quy để định lượng hoá mối quan hệ giữa biến mục tiêu Y và các biến độc lập $(X_1, X_2, \ldots, X_k)$ được cho bởi phương trình:

$$Y = f(X_1, X_2, \ldots, X_k) + \epsilon = f(X) + \epsilon$$

Trong đó:

- f là một hàm toán học định lượng hoá mối quan hệ
- Y là biến mục tiêu (biến phụ thuộc, biến đầu ra)

Phân tích hồi quy

- Mô hình hồi quy để định lượng hoá mối quan hệ giữa biến mục tiêu Y và các biến độc lập $(X_1, X_2, \ldots, X_k)$ được cho bởi phương trình:

$$Y = f(X_1, X_2, \ldots, X_k) + \epsilon = f(X) + \epsilon$$

Trong đó:

- f là một hàm toán học định lượng hoá mối quan hệ
- Y là biến mục tiêu (biến phụ thuộc, biến đầu ra)
- $X = (X_1, X_2, \ldots, X_k)$ là các biến độc lập (biến giải thích, biến đầu vào)

Phân tích hồi quy

- Mô hình hồi quy để định lượng hoá mối quan hệ giữa biến mục tiêu Y và các biến độc lập $(X_1, X_2, \ldots, X_k)$ được cho bởi phương trình:

$$Y = f(X_1, X_2, \ldots, X_k) + \epsilon = f(X) + \epsilon$$

Trong đó:

- f là một hàm toán học định lượng hoá mối quan hệ
- Y là biến mục tiêu (biến phụ thuộc, biến đầu ra)
- $X = (X_1, X_2, \ldots, X_k)$ là các biến độc lập (biến giải thích, biến đầu vào)
- ϵ là sai số nhẫu nhiên.

Một số dạng mô hình hồi quy phổ biến

- Mô hình hồi quy tuyến tính: hàm f là hàm tuyến tính,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

hay

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Trong đó:

  - Biến mục tiêu (biến phụ thuộc, biến đầu ra) Y là biến định lượng liên tục
  - $X = (X_1, X_2, \ldots, X_k)$ là các biến độc lập (biến giải thích, biến đầu vào)
  - Sai số nhẫu nhiên ϵ được giả sử có phân bố chuẩn.

Một số dạng mô hình hồi quy phổ biến

- Mô hình hồi quy logistic: Biến phụ thuộc Y là biến nhị phân (Y nhận 2 giá trị 0 hoặc 1, "có" hoặc "không"):

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

hay

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon)}}$$

Trong đó:

  - $p = \mathbb{P}(Y = 1)$ là xác suất xảy ra giá trị $Y = 1$.
  - $X = (X_1, X_2, \ldots, X_k)$ là các biến độc lập (biến giải thích, biến đầu vào)
  - ϵ là sai số nhẫu nhiên

Một số dạng mô hình hồi quy phổ biến

- Mô hình hồi quy đa thức: hàm f là hàm đa thức bậc k,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k + \epsilon$$

hay

$$\mathbb{E}(Y|X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_k X^k$$

Trong đó:
  - Biến mục tiêu (biến phụ thuộc, biến đầu ra) Y là biến định lượng liên tục
  - $X = (X_1, X_2, \ldots, X_k)$ là các biến độc lập (biến giải thích, biến đầu vào)
  - ϵ là sai số nhẫu nhiên

Một số dạng mô hình hồi quy phổ biến khác

- Mô hình hồi quy Ridge
- Mô hình hồi quy LASSO
- Mô hình hồi quy Elastic Net
- Mô hình hồi quy phân vị
- Mô hình hồi quy Poisson
- Mô hình hồi quy Cox

Dữ liệu thực nghiệm

- Các giá trị quan sát của 2 biến X và Y là kết quả của một phép thử, một thí nghiệm ngoài thực địa, trong phòng thí nghiệm

Dữ liệu thực nghiệm

- Các giá trị quan sát của 2 biến X và Y là kết quả của một phép thử, một thí nghiệm ngoài thực địa, trong phòng thí nghiệm
- Giá trị của biến X có thể kiểm soát, và chúng ta quan sát giá trị kết quả của biến Y:
  - $X =$ liều lượng thuốc, $Y =$ thay đổi lưu lượng máu trong bệnh nhân.
  - $X =$ lượng phân bón, $Y =$ năng suất lúa.

Dữ liệu thực nghiệm

- Các giá trị quan sát của 2 biến X và Y là kết quả của một phép thử, một thí nghiệm ngoài thực địa, trong phòng thí nghiệm
- Giá trị của biến X có thể kiểm soát, và chúng ta quan sát giá trị kết quả của biến Y:
  - X = liều lượng thuốc, Y = thay đổi lưu lượng máu trong bệnh nhân.
  - X = lượng phân bón, Y = năng suất lúa.
- Chúng ta quan sát giá trị cả 2 biến X và Y, không kiểm soát được biến nào:
  - X = cân nặng, Y = chiều cao của một người.
  - X = chiều cao của giống cây, Y = năng suất của giống cây đó.

Ví dụ

Ví dụ: Một kĩ sư tại một công ty chất bán dẫn muốn mô hình hoá mối quan hệ giữa hệ số khuếch đại của các thiết bị transistor HFE (y) với ba thông số: cực phát – RS (x_1), cực gốc – RS (x_2) và cực góp – RS (x_3). Số liệu được cho bởi bảng ở dưới đây

x_1	x_2	x_3	y
14.62	226	7	128.4
15.63	220	3.375	52.62
14.62	217.4	6.375	113.9
15	220	6	98.01
14.5	226.5	7.625	139.9
15.25	224.1	6	102.6
16.12	220.5	3.375	48.14
15.13	223.5	6.125	109.6
15.5	217.6	5	82.68
15.13	228.5	6.625	112.6
15.5	230.2	5.75	97.52
16.12	226.5	3.75	59.06
15.13	226.6	6.125	118.8
15.63	225.6	5.375	89.09
15.38	229.7	5.875	101
14.38	234	8.875	171.9
15.5	230	4	66.8
14.25	224.3	8	157.1
14.5	240.5	10.87	208.4
14.62	223.7	7.375	113.4

Mô hình hồi quy tuyến tính cho ví dụ

Mô hình hồi quy tuyến tính cho ví dụ này là:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

Trong đó

- Y (hệ số khuếch đại) là biến đầu ra hay biến phụ thuộc
- $X = (X_1, X_2, X_3)$ (thông số của 3 cực) là biến độc lập
- ϵ gọi là sai số ngẫu nhiên (sai số thực nghiệm)

Mô hình hồi quy tuyến tính

Xem xét một bộ dữ liệu thực nghiệm gồm biến phụ thuộc Y và biến giải thích $X = (X_1, X_2, \ldots, X_k)$. Giả sử mối quan hệ giữa X và Y là tuyến tính:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

hay

$$E(Y|X = x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Trong đó:

- ϵ gọi là sai số ngẫu nhiên (sai số thực nghiệm), được giả sử có phân bố chuẩn $\mathcal{N}(0, \sigma^2)$.
- β_0 gọi là hệ số chặn (giá trị trung bình của Y khi $X = 0$).
- $(\beta_1, \ldots, \beta_k)$ gọi là hệ số góc: mô tả tốc độ biến thiên của biến Y (β_i là sự thay đổi của Y khi X_i thay đổi 1 đơn vị).

Mô hình hồi quy tuyến tính

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Các tham số β_0 và β_i đặc trưng cho mối tương quan này gọi là các hệ số hồi quy.

Mô hình hồi quy tuyến tính

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Các tham số β_0 và β_i đặc trưng cho mối tương quan này gọi là các hệ số hồi quy.

Vấn đề: Chúng ta không biết giá trị của $\beta_0, \beta_1, \ldots, \beta_k$:

- Bộ dữ liệu gồm n cặp quan sát $(x^{(i)}, y_i), i = 1, 2, \ldots, n$ với $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})$

Mô hình hồi quy tuyến tính

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Các tham số β_0 và β_i đặc trưng cho mối tương quan này gọi là các hệ số hồi quy.

Vấn đề: Chúng ta không biết giá trị của $\beta_0, \beta_1, \ldots, \beta_k$:

- Bộ dữ liệu gồm n cặp quan sát $(x^{(i)}, y_i), i = 1, 2, \ldots, n$ với $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})$
- Dữ liệu được sử dụng để ước lượng tham số $\beta_0, \beta_1, \ldots, \beta_k$, tức là để phù hợp mô hình với dữ liệu nhằm:
  - xác định mối tương quan giữa Y và X

Mô hình hồi quy tuyến tính

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Các tham số β_0 và β_i đặc trưng cho mối tương quan này gọi là các hệ số hồi quy.

Vấn đề: Chúng ta không biết giá trị của $\beta_0, \beta_1, \ldots, \beta_k$:

- Bộ dữ liệu gồm n cặp quan sát $(x^{(i)}, y_i), i = 1, 2, \ldots, n$ với $x^{(i)} = (x_1^{(i)}, x_2^{(i)}, \ldots, x_k^{(i)})$
- Dữ liệu được sử dụng để ước lượng tham số $\beta_0, \beta_1, \ldots, \beta_k$, tức là để phù hợp mô hình với dữ liệu nhằm:
    - xác định mối tương quan giữa Y và X
    - Sử dụng mối tương quan này để dự báo giá trị của biến phụ thuộc Y quan sát được khi biết giá trị đầu vào của biến độc lập $X = x$.

Ước lượng $\beta_0, \beta_1, \ldots, \beta_k$: Phương pháp bình phương nhỏ nhất

- Ta tìm ước lượng $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ để dự báo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$

Ước lượng $\beta_0, \beta_1, \ldots, \beta_k$: Phương pháp bình phương nhỏ nhất

- Ta tìm ước lượng $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ để dự báo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$
- Với các quan sát $(x^{(i)}, y_i), i = 1, 2, \ldots, n$, áp dụng cho mô hình HQTT: $y_i = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_k x_k^{(i)} + \epsilon_i = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)} + e_i$

VINBIGDATA VINGROUP AI Academy Vietnam

Ước lượng $\beta_0, \beta_1, \ldots, \beta_k$: Phương pháp bình phương nhỏ nhất

- Ta tìm ước lượng $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ để dự báo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$
- Với các quan sát $(x^{(i)}, y_i), i = 1, 2, \ldots, n$, áp dụng cho mô hình HQTT: $y_i = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_k x_k^{(i)} + \epsilon_i = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)} + e_i$
- Chúng ta làm phù hợp mô hình với dữ liệu bằng cách tìm ước lượng hệ chặn $\hat{\beta}_0$ và hệ số góc $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$.

Ước lượng $\beta_0, \beta_1, \ldots, \beta_k$: Phương pháp bình phương nhỏ nhất

- Ta tìm ước lượng $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ để dự báo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$
- Với các quan sát $(x^{(i)}, y_i), i = 1, 2, \ldots, n$, áp dụng cho mô hình HQTT: $y_i = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_k x_k^{(i)} + \epsilon_i = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)} + e_i$
- Chúng ta làm phù hợp mô hình với dữ liệu bằng cách tìm ước lượng hệ chặn $\hat{\beta}_0$ và hệ số góc $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$.
- Với mỗi quan sát (x_i, y_i), ta có $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)}) = e_i$

Ước lượng $\beta_0, \beta_1, \ldots, \beta_k$: Phương pháp bình phương nhỏ nhất

- Ta tìm ước lượng $(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$ để dự báo $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k$
- Với các quan sát $(x^{(i)}, y_i), i = 1, 2, \ldots, n$, áp dụng cho mô hình HQTT: $y_i = \beta_0 + \beta_1 x_1^{(i)} + \beta_2 x_2^{(i)} + \ldots + \beta_k x_k^{(i)} + \epsilon_i = \hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)} + e_i$
- Chúng ta làm phù hợp mô hình với dữ liệu bằng cách tìm ước lượng hệ chặn $\hat{\beta}_0$ và hệ số góc $(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$.
- Với mỗi quan sát (x_i, y_i), ta có $y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)}) = e_i$
- Tổng bình phương sai lệch cho tập dữ liệu (x_i, y_i):

$$f(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)})]^2$$

Phương pháp bình phương nhỏ nhất

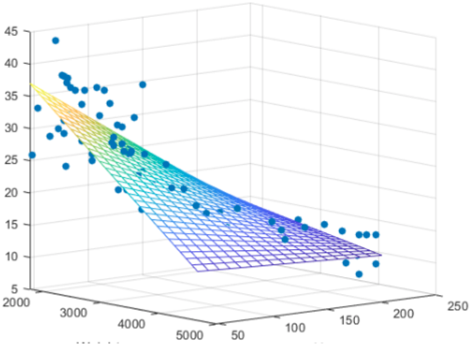
- Phương pháp bình phương nhỏ nhất: Tìm các ước lượng $\hat{\beta}_0, \ldots, \hat{\beta}_k$ làm cực tiểu hoá hàm số:

$$f(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)})]^2$$

Phương pháp bình phương nhỏ nhất

- Phương pháp bình phương nhỏ nhất: Tìm các ước lượng $\hat{\beta}_0, \ldots, \hat{\beta}_k$ làm cực tiểu hoá hàm số:

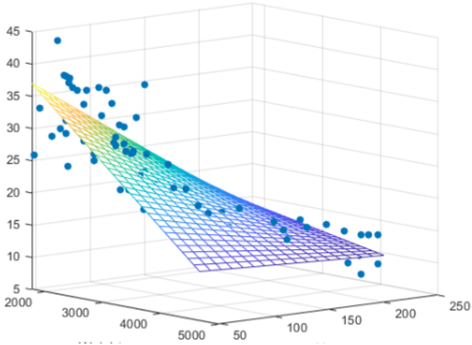
$$f(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)})]^2$$



Phương pháp bình phương nhỏ nhất

- Phương pháp bình phương nhỏ nhất: Tìm các ước lượng $\hat{\beta}_0, \ldots, \hat{\beta}_k$ làm cực tiểu hoá hàm số:

$$f(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)})]^2$$



Phương pháp bình phương nhỏ nhất

- Phương pháp bình phương nhỏ nhất: Tìm các ước lượng $\hat{\beta}_0, \ldots, \hat{\beta}_k$ làm cực tiểu hoá hàm số:

$$f(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)})]^2$$

Phương pháp bình phương nhỏ nhất

- Phương pháp bình phương nhỏ nhất: Tìm các ước lượng $\hat{\beta}_0, \ldots, \hat{\beta}_k$ làm cực tiểu hoá hàm số:

$$f(\hat{\beta}_0, \ldots, \hat{\beta}_k) = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_1^{(i)} + \hat{\beta}_2 x_2^{(i)} + \ldots + \hat{\beta}_k x_k^{(i)})]^2$$

- Giải hệ phương trình sau để tìm ước lượng $\hat{\beta}_0, \ldots, \hat{\beta}_k$:

$$\frac{\partial f(\hat{\beta}_0, \ldots, \hat{\beta}_p)}{\partial \hat{\beta}_i} = 0; i = 0, 2, \ldots, k$$

Biểu diễn dưới dạng ma trận

- Dữ liệu gồm *n* quan sát cho bởi bảng:

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{n1}	x_{n2}	...	x_{nk}

Biểu diễn dưới dạng ma trận

- Đặt

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix},$$

và

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \text{ and } \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

- Khi đó ta có mô hình dạng ma trận:

$$Y = X\beta + \epsilon.$$

Biểu diễn dưới dạng ma trận

Phương pháp bình phương nhỏ nhất:

- Tính tổng bình phương sai lệch:

$$L = \sum_{i=1}^{n} e_i^2 = e^t e = (y - X\hat{\beta})^t y - X\hat{\beta} t$$

- Tìm $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta_k})$ sao cho L đạt giá trị nhỏ nhất.
- Khi đó

$$\left.\frac{\partial L}{\partial \boldsymbol{\beta}}\right|_{\hat{\beta}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}$$

- Ta được nghiệm:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Ước lượng trong mô hình hồi quy tuyến tính

- Mô hình hồi quy dự báo:

$$\hat{y} = X\hat{\beta}$$

hay

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \ldots + \hat{\beta}_k X_{ik}$$

Ví dụ

Ví dụ: Một kĩ sư tại một công ty chất bán dẫn muốn mô hình hoá mối quan hệ giữa hệ số khuếch đại của các thiết bị transistor HFE (y) với ba thông số: cực phát – RS (x_1), cực gốc – RS (x_2) và cực góp – RS (x_3). Số liệu được cho bởi bảng ở dưới đây

x_1	x_2	x_3	y
14.62	226	7	128.4
15.63	220	3.375	52.62
14.62	217.4	6.375	113.9
15	220	6	98.01
14.5	226.5	7.625	139.9
15.25	224.1	6	102.6
16.12	220.5	3.375	48.14
15.13	223.5	6.125	109.6
15.5	217.6	5	82.68
15.13	228.5	6.625	112.6
15.5	230.2	5.75	97.52
16.12	226.5	3.75	59.06
15.13	226.6	6.125	118.8
15.63	225.6	5.375	89.09
15.38	229.7	5.875	101
14.38	234	8.875	171.9
15.5	230	4	66.8
14.25	224.3	8	157.1
14.5	240.5	10.87	208.4
14.62	223.7	7.375	113.4

Ví dụ: hệ số khuếch đại của các thiết bị transistor

Ví dụ: Hệ số khuếch đại của các thiết bị transistor

- Hãy lập mô hình hồi quy tuyến tính cho bài toán trên
- Hãy tính ước lượng của các tham số trong mô hình

x_1	x_2	x_3	y
14.62	226	7	128.4
15.63	220	3.375	52.62
14.62	217.4	6.375	113.9
15	220	6	98.01
14.5	226.5	7.625	139.9
15.25	224.1	6	102.6
16.12	220.5	3.375	48.14
15.13	223.5	6.125	109.6
15.5	217.6	5	82.68
15.13	228.5	6.625	112.6
15.5	230.2	5.75	97.52
16.12	226.5	3.75	59.06
15.13	226.6	6.125	118.8
15.63	225.6	5.375	89.09
15.38	229.7	5.875	101
14.38	234	8.875	171.9
15.5	230	4	66.8
14.25	224.3	8	157.1
14.5	240.5	10.87	208.4
14.62	223.7	7.375	113.4

Ví dụ: hệ số khuếch đại của các thiết bị transistor

Trả lời:

- Mô hình hồi quy tuyến tính cho bài toán trên:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

trong đó:

- Biến phụ thuộc Y là hệ số khuếch đại của các thiết bị transistor HFE
- Các biến độc lập gồm thông số của cực phát – RS (x_1), của cực gốc – RS (x_2) và của cực góp – RS (x_3)
- Sai số ngẫu nhiên ϵ có phân bố chuẩn $N(0; \sigma^2)$
- Các tham số của mô hình gồm: $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$ và σ^2.

Ví dụ: hệ số khuếch đại của các thiết bị transistor

- Kết quả chạy mô hình:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8251    95.5182  -0.051    0.960
x1           -9.2272     7.1119  -1.297    0.213
x2            0.6383     0.4314   1.479    0.158
x3           17.6320     2.5272   6.977 3.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.703 on 16 degrees of freedom
Multiple R-squared:  0.9762,
  Adjusted R-squared:  0.9718
F-statistic: 218.9 on 3 and 16 DF,  p-value: 3.378e-13
```

- Ước lượng của hệ số hồi quy β là

$$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = (-4.8251; -9.2272; 0.6383; 17.632)$$

- Ước lượng của σ^2 là $\hat{\sigma}^2 = 6.703^2 = 44.93$

Kiểm định ý nghĩa của mô hình hồi quy

Ta kiểm định cặp giả thuyết đối thuyết sau:
H_0 : Biến phụ thuộc Y không có mối quan hệ tuyến tính với các biến giải thích $X_1, X_2, \ldots, X_k$
H_1 : Biến phụ thuộc Y có mối quan hệ tuyến tính với các biến giải thích $X_1, X_2, \ldots, X_k$
Tương đương với:

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$$

$$H_1 : \exists \beta_i \neq 0$$

Kiểm định ý nghĩa của mô hình hồi quy

Bảng phân tích phương sai cho mô hình hồi quy:

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Trong đó

- Tổng bình phương tổng thể: $SS_T = \sum_{i=1}(y_i - \bar{y})^2$

Kiểm định ý nghĩa của mô hình hồi quy

Bảng phân tích phương sai cho mô hình hồi quy:

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Trong đó

- Tổng bình phương tổng thể: $SS_T = \sum_{i=1}(y_i - \bar{y})^2$
- Tổng bình phương cho mô hình hồi quy: $SS_R = \sum_{i=1}(\hat{y}_i - \bar{y})^2$

Kiểm định ý nghĩa của mô hình hồi quy

Bảng phân tích phương sai cho mô hình hồi quy:

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Trong đó

- Tổng bình phương tổng thể: $SS_T = \sum_{i=1}(y_i - \bar{y})^2$
- Tổng bình phương cho mô hình hồi quy: $SS_R = \sum_{i=1}(\hat{y}_i - \bar{y})^2$
- Tổng bình phương sai số: $SS_E = \sum_{i=1}(y_i - \hat{y}_i)^2 = SS_T - SS_R$

Kiểm định ý nghĩa của mô hình hồi quy

Bảng phân tích phương sai cho mô hình hồi quy:

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Trong đó

- Phương sai cho mô hình hồi quy: $MS_R = SS_R/k$

Kiểm định ý nghĩa của mô hình hồi quy

Bảng phân tích phương sai cho mô hình hồi quy:

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Trong đó

- Phương sai cho mô hình hồi quy: $MS_R = SS_R/k$
- Phương sai sai số: $MS_E = SS_E/(n-k-1)$

Kiểm định ý nghĩa của mô hình hồi quy

Bảng phân tích phương sai cho mô hình hồi quy:

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Trong đó

- Phương sai cho mô hình hồi quy: $MS_R = SS_R/k$
- Phương sai sai số: $MS_E = SS_E/(n-k-1)$
- Giá trị F thực nghiệm: $F_{obs} = MS_R/MS_E$

Kiểm định ý nghĩa của mô hình hồi quy

Bảng phân tích phương sai cho mô hình hồi quy:

Analysis of Variance for Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Quy tắc kiểm định: Ta bác bỏ H_0 (tức là biến phụ thuộc Y có mối quan hệ tuyến tính với các biến độc lập $X_1, X_2, \ldots, X_k$) nếu:

$$F_{obs} > F_{k;n-k-1;\alpha} \Leftrightarrow p = P(F_{k;n-k-1} > F_{obs}) < \alpha$$

Ví dụ

Ví dụ: Hãy kiểm định xem giữa hệ số khuếch đại của các thiết bị transistor HFE (y) có mối quan hệ tuyến tính với ba thông số: cực phát – RS (x_1), cực gốc – RS (x_2) và cực góp – RS (x_3) ở mức ý nghĩa $\alpha = 5\%$ hay không?
Ta kiểm định cặp giả thuyết đối thuyết sau:

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

$$H_1 : \exists \beta_i \neq 0$$

ở mức ý nghĩa $\alpha = 0.05$

Ví dụ

Kết quả chạy mô hình hồi quy:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8251    95.5182  -0.051    0.960
x1           -9.2272     7.1119  -1.297    0.213
x2            0.6383     0.4314   1.479    0.158
x3           17.6320     2.5272   6.977 3.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.703 on 16 degrees of freedom
Multiple R-squared:  0.9762,
  Adjusted R-squared:  0.9718
F-statistic: 218.9 on 3 and 16 DF,  p-value: 3.378e-13
```

Vì $p = 3.378 * 10^{-13} < 0.05 = \alpha$ nên ta bác bỏ H_0. Vậy biến phụ thuộc y có mối quan hệ tuyến tính với ba biến độc lập x_1, x_2, x_3 ở mức ý nghĩa $\alpha = 5\%$

Kiểm định các hệ số hồi quy

Ta kiểm định cặp giả thuyết đối thuyết sau:
H_0 : Biến phụ thuộc Y không có mối quan hệ tuyến tính với biến giải thích X_j (Trong mô hình không xét biến độc lập X_j)
H_1 : Biến phụ thuộc Y có mối quan hệ tuyến tính với các biến giải thích X_j (Trong mô hình có xét biến độc lập X_j)
Tương đương với:

$$H_0 : \beta_j = 0$$

$$H_1 : \beta_j \neq 0$$

Kiểm định các hệ số hồi quy

Quy tắc kiểm định:

- Tính tiêu chuẩn kiểm định: giá trị T thực nghiệm

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

trong đó: C_{jj} là phần tử trên đường chéo chính của ma trận $(X^t X)^{-1}$ tương ứng với $\hat{\beta}_j$, $\hat{\sigma}^2 = SS_E/(n-k-1)$ là ước lượng của phương sai σ^2 của sai số ngẫu nhiên.

Kiểm định các hệ số hồi quy

Quy tắc kiểm định:

- Tính tiêu chuẩn kiểm định: giá trị T thực nghiệm

$$T = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{jj}}}$$

  trong đó: C_{jj} là phần tử trên đường chéo chính của ma trận $(X^t X)^{-1}$ tương ứng với $\hat{\beta}_j$, $\hat{\sigma}^2 = SS_E/(n-k-1)$ là ước lượng của phương sai σ^2 của sai số ngẫu nhiên.

- Ta bác bỏ H_0 nếu $|T| > t(n-k-1; \alpha/2)$
  $\Leftrightarrow p - value = 2P(T_{n-k-1} > |T|) < \alpha$

Ví dụ

Ví dụ: Một kĩ sư tại một công ty chất bán dẫn muốn mô hình hoá mối quan hệ giữa hệ số khuếch đại của các thiết bị transistor HFE (y) với ba thông số: cực phát – RS (x_1), cực gốc – RS (x_2) và cực góp – RS (x_3). Với mức ý nghĩa $\alpha = 0.05$, hãy kiểm định xem hệ số khuếch đại của các thiết bị transistor HFE (y) có quan hệ tuyến tính với thông số của cực phát – RS (x_1) hay không?
Ta kiểm định cặp GT- ĐT sau:

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_1 \neq 0$$

ở mức ý nghĩa 0.05

Ví dụ: hệ số khuếch đại của các thiết bị transistor

- Kết quả chạy mô hình:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8251    95.5182  -0.051    0.960
x1           -9.2272     7.1119  -1.297    0.213
x2            0.6383     0.4314   1.479    0.158
x3           17.6320     2.5272   6.977 3.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.703 on 16 degrees of freedom
Multiple R-squared:  0.9762,
  Adjusted R-squared:  0.9718
F-statistic: 218.9 on 3 and 16 DF,  p-value: 3.378e-13
```

- Ta có p-giá trị $= 0.213 > \alpha = 0.05$ nên ta chấp nhận H_0. Vậy hệ số khuếch đại của các thiết bị transistor HFE (y) không có quan hệ tuyến tính với thông số của cực phát – RS x_1 ở mức ý nghĩa 0.05

Hệ số xác định mô hình

- Hệ số xác định của mô hình:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Hệ số xác định mô hình

- Hệ số xác định của mô hình:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

- Hệ số xác định đã hiệu chỉnh của mô hình:

$$R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)}$$

Hệ số xác định mô hình

- Hệ số xác định của mô hình:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

- Hệ số xác định đã hiệu chỉnh của mô hình:

$$R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)}$$

- Ý nghĩa của hệ số xác định: cho biết tỷ lệ % các biến độc lập giải thích cho sự biến thiên của biến phụ thuộc thông qua mô hình tuyến tính.

Ví dụ: hệ số khuếch đại của các thiết bị transistor

- Kết quả chạy mô hình đầy đủ:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8251    95.5182  -0.051    0.960
x1           -9.2272     7.1119  -1.297    0.213
x2            0.6383     0.4314   1.479    0.158
x3           17.6320     2.5272   6.977 3.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.703 on 16 degrees of freedom
Multiple R-squared:  0.9762,
  Adjusted R-squared:  0.9718
F-statistic: 218.9 on 3 and 16 DF,  p-value: 3.378e-13
```

- Hệ số xác định của mô hình là $R^2 = 0.9762 = 97.62\%$, có nghĩa là các biến độc lập x_1, x_2, x_3 giải thích được 97.62% sự biến thiên của biến phụ thuộc y.

Lựa chọn mô hình

Thuật toán lựa chọn mô hình (phương pháp loại bỏ lùi):

- Bước 1: Chạy mô hình đầy đủ với tất cả các biến độc lập

Lựa chọn mô hình

Thuật toán lựa chọn mô hình (phương pháp loại bỏ lùi):

- Bước 1: Chạy mô hình đầy đủ với tất cả các biến độc lập
- Bước 2: Loại bỏ biến độc lập nào có p-giá trị lớn nhất và lớn hơn α_{crit} (cho trước, thường chọn α_{crit} trong khoảng $0.15 - 0.2$)

Lựa chọn mô hình

Thuật toán lựa chọn mô hình (phương pháp loại bỏ lùi):

- Bước 1: Chạy mô hình đầy đủ với tất cả các biến độc lập
- Bước 2: Loại bỏ biến độc lập nào có p-giá trị lớn nhất và lớn hơn α_{crit} (cho trước, thường chọn α_{crit} trong khoảng $0.15 - 0.2$)
- Bước 3: Chạy mô hình mới sau đó quay lại bước 2

Lựa chọn mô hình

Thuật toán lựa chọn mô hình (phương pháp loại bỏ lùi):

- Bước 1: Chạy mô hình đầy đủ với tất cả các biến độc lập
- Bước 2: Loại bỏ biến độc lập nào có p-giá trị lớn nhất và lớn hơn α_{crit} (cho trước, thường chọn α_{crit} trong khoảng $0.15 - 0.2$)
- Bước 3: Chạy mô hình mới sau đó quay lại bước 2
- Bước 4: Dừng đến khi nào tất cả các p-giá trị của các biến độc lập nhỏ hơn α_{crit} và ta thu được mô hình cuối cùng.

Ví dụ: hệ số khuếch đại của các thiết bị transistor

- Kết quả chạy mô hình đầy đủ:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8251    95.5182  -0.051    0.960
x1           -9.2272     7.1119  -1.297    0.213
x2            0.6383     0.4314   1.479    0.158
x3           17.6320     2.5272   6.977 3.12e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.703 on 16 degrees of freedom
Multiple R-squared:  0.9762,
  Adjusted R-squared:  0.9718
F-statistic: 218.9 on 3 and 16 DF,  p-value: 3.378e-13
```

- Loại bỏ biến x_1 vì có p-giá trị lớn nhất và nhỏ hơn $\alpha_{crit} = 0.15$

Ví dụ: hệ số khuếch đại của các thiết bị transistor

- Kết quả chạy mô hình sau khi loại bỏ biến x_1:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -84.7057    74.4811  -1.137    0.271
x2            0.2918     0.3456   0.844    0.410
x3           20.6337     1.0371  19.895 3.26e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.837 on 17 degrees of freedom
Multiple R-squared:  0.9737,
  Adjusted R-squared:  0.9706
F-statistic: 314.9 on 2 and 17 DF,  p-value: 3.69e-14
```

- Loại bỏ biến x_2 vì có p-giá trị lớn nhất và nhỏ hơn $\alpha_{crit} = 0.15$

Ví dụ: hệ số khuếch đại của các thiết bị transistor

- Kết quả chạy mô hình sau khi loại bỏ biến x_1 và x_2:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.9820     5.3816  -4.085 0.000696 ***
x3           21.1439     0.8362  25.285 1.63e-15 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1

Residual standard error: 6.782 on 18 degrees of freedom
Multiple R-squared:  0.9726,
  Adjusted R-squared:  0.9711
F-statistic: 639.3 on 1 and 18 DF,  p-value: 1.626e-15
```

- Mô hình cuối cùng được lựa chọn là $Y = \beta_0 + \beta_3 X_3 + \epsilon$

Bài tập

Điện năng tiêu thụ trong tháng (y) của một nhà máy hoá chất được cho là chịu liên quan bởi các yếu tố: nhiệt độ trung bình xung quanh (x_1), số ngày nhà máy hoạt động trong tháng (x_2), độ tinh khiết trung bình của sản phẩm (x_3), tổng sản lượng sản phẩm sản suất trong tháng (x_4). Số liệu điện năng tiêu thụ trong năm 2019 của nhà máy trên được cho bởi bảng ở dưới đây.

y	x_1 (°F)	x_2 (ngày)	x_3 (%)	x_4 (tấn)
240	25	24	91	100
236	31	21	90	95
270	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

Bài tập

Xét mô hình hồi quy tuyến tính sau:
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon$ với giả thiết $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

1) Với mức ý nghĩa 0,05 hãy kiểm định xem có hay không mối quan hệ tuyến tính giữa biến phụ thuộc y với các biến giải thích x_1, x_2, x_3, x_4, tức là hãy kiểm định cặp giả thuyết – đối thuyết sau:
$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0 \text{ với } H_1: \exists\, \beta_j \neq 0;\ j = 1,2,3,4.$$
2) Hãy tính ước lượng của các hệ số hồi quy $\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)$. Từ đó hãy dự báo lượng tiêu thụ điện trong tháng của nhà máy trên nếu $x_1 = 75°\text{F}$, $x_2 = 24$ ngày, $x_3 = 90\%$ và $x_4 = 98$ tấn.
3) Với mức ý nghĩa 0,05 hãy kiểm định cặp giả thuyết – đối thuyết sau:
$$H_0: \beta_1 = 0 \text{ với } H_1: \beta_1 \neq 0$$
4) Với mức ý nghĩa 0,05 hãy kiểm định cặp giả thuyết – đối thuyết sau:
$$H_0: \beta_3 = 0 \text{ với } H_1: \beta_3 \neq 0$$
5) Hãy tính hệ số xác định đã hiệu chỉnh của mô hình.
6) Hãy chạy thuật toán loại bỏ lùi để tìm mô hình tối ưu cho bài toán trên.