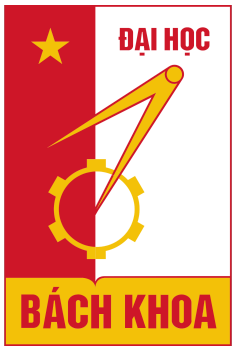
TRƯỜNG ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN TOÁN ỨNG DỤNG VÀ TIN HỌC



Báo cáo môn Phân tích số liệu

Đề tài: Mô hình hồi quy tuyến tính

Giảng viên hướng dẫn: TS. Lê Xuân Lý
Nhóm sinh viên thực hiện: Nhóm 6

Hoàng Phi Long	.....	20185375
Đặng Thị Hồng Nhung	.....	20185391
Phạm Ngọc Thi Thư	.....	20185409
Nguyễn Hải Đăng	.....	20185333
Phạm Thị Linh Chi	.....	20185329
Trần Hải Phong	.....	20185393

Hà Nội, tháng 5/2021

Bảng phân công nhiệm vụ

Họ và tên	Đánh giá công việc
Hoàng Phi Long	1
Đặng Thị Hồng Nhung	1
Phạm Ngọc Thi Thư	1
Nguyễn Hải Đăng	2
Phạm Thị Linh Chi	2
Trần Hải Phong	2

Mục lục

I Giới thiệu bài toán thực tế **4**

II Lý thuyết sử dụng để xử lý bài toán **6**

1 Mô hình hồi quy tuyết tính cổ điển **7**
1.1 Giới thiệu về mô hình hồi quy tuyến tính cổ điển . . . 7
1.2 Ước lượng bình phương cực tiểu . . . 8
1.2.1 Minh họa hình học của ước lượng bình phương cực tiểu . . . 11
1.2.2 Tính chất ước lượng bằng phương pháp bình phương cực tiểu . . . 11
1.2.3 Định lý Gauss về ước lượng bình phương cực tiểu . . . 12
1.2.4 Hệ số xác định R . . . 13
1.3 Mô hình hồi quy tuyến tính . . . 13
1.3.1 Phân phối F, phân phối Student . . . 13
1.3.2 Khoảng tin cậy của các hệ số hồi quy β_j . . . 15
1.3.3 Kiểm định giả thiết về các hệ số hồi quy . . . 18
1.3.4 Ước lượng hàm hồi quy tuyến tính . . . 19
1.4 Kiểm tra sự phù hợp của mô hình . . . 21
1.4.1 Outlier . . . 21
1.4.2 Sự phù hợp của mô hình . . . 22
1.4.3 Kiểm định tính không tương quan của ε_j theo thời gian . . . 25
1.4.4 Q - Q plot . . . 27
1.5 Một số nội dung bổ sung . . . 27
1.5.1 Standardized . . . 27
1.5.2 $t - statistic$ và $p - value$ . . . 27
1.5.3 Xác định các biến quan trọng - Features Selection . . . 28
1.5.4 Các bước tiến hành trong phân tích hồi quy . . . 28

2 Mô hình hồi quy tuyến tính bội **29**
2.1 Mô hình bài toán . . . 29
2.1.1 Mô hình bài toán . . . 29
2.1.2 Tổng kết mô hình . . . 30
2.1.3 Tại sao hồi quy đồng thời? . . . 30

III Ứng dụng vào bài toán thực tế **31**

3 Dữ liệu Advertising **34**
3.1 Có sự cộng tuyến giữa TV, Radio, Newspaper không? . . . 35
3.2 Mối liên hệ này chặt chẽ cỡ nào? . . . 35
3.3 Kênh quảng cáo nào đóng góp vào doanh số? . . . 36

3.4 Liệu có ảnh hưởng cộng năng giữa các kênh quảng cáo hay không? 36

4 Kiểm tra sự phù hợp của mô hình **38**
4.1 Tiêu chuẩn **Student** 38
4.2 Khảo sát đồ thị phần dư + xác định các **Outlier** 38
4.3 Kiểm định tính không tương quan của phần dư theo thời gian 39
4.4 Xác định các điểm **Leverage** 39

5 Xây dựng mô hình cuối cùng **41**
5.1 Ước lượng hệ số hồi quy và khoảng tin cậy của chúng 41
5.2 Xác định hệ số R 41
5.3 Ước lượng hàm HQTT 41

Tài liệu tham khảo **43**

Phần I

Giới thiệu bài toán thực tế

Chúng ta có dữ liệu về doanh số bán một sản phẩm (sales) của một công ty trên 200 thị trường khác nhau, kèm theo ngân sách đầu tư quảng cáo cho sản phẩm đó trên 3 phương tiện truyền thông: TV, Radio và Newspaper. Công ty không thể trực tiếp tăng doanh số bán sản phẩm. Mặt khác, họ có thể kiểm soát chi phí quảng cáo trên từng phương tiện truyền thông. Do đó, nếu xác định được rằng có mối liên hệ giữa quảng cáo và doanh số bán hàng, thì chúng ta có thể hướng dẫn công ty điều chỉnh ngân sách quảng cáo, từ đó gián tiếp tăng doanh số bán hàng. Nói cách khác, mục tiêu của chúng ta là phát triển một mô hình chính xác có thể sử dụng được để dự đoán doanh số bán hàng trên cơ sở ngân sách đầu tư cho quảng cáo.

Phần II

Lý thuyết sử dụng để xử lý bài toán

Chương 1

Mô hình hồi quy tuyết tính cổ điển

1.1 Giới thiệu về mô hình hồi quy tuyến tính cổ điển

Ta tiến hành n quan sát độc lập đồng thời về $k+1$ biến $X_1, ..., X_k, Y$. Giả sử số liệu quan sát tuân theo mô hình sau:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 x_{11} + \cdots + \beta_k x_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 x_{21} + \cdots + \beta_k x_{2k} + \varepsilon_2 \\ \cdots & \\ y_n &= \beta_0 + \beta_1 x_{n1} + \cdots + \beta_k x_{nk} + \varepsilon_n \end{aligned} \tag{1.1}$$

trong đó các sai số $\varepsilon_1, ..., \varepsilon_n$ thỏa mãn 3 điều kiện:

- $E(\varepsilon_j) = 0$ (việc đo đạc không chịu sai lệch hệ thống)
- $D(\varepsilon_j) = \sigma^2$ (phương sai không đổi hay là độ chuẩn xác đo đạc như nhau)
- $cov(\varepsilon_i, \varepsilon_j) = 0$ với mọi $i \neq j = 1 \div n$ (các sai lệch từng bước không ảnh hưởng đến nhau)

Mô hình (1.1) có thể viết dưới dạng ma trận như sau:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots\vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ddots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

hoặc đơn giản hơn:

$$\underbrace{\mathbb{Y}}_{n\times 1} = \underbrace{\mathbb{X}}_{n\times(k+1)} \cdot \underbrace{\beta}_{(k+1)\times 1} + \underbrace{\varepsilon}_{n\times 1} \tag{1.2}$$

với

$$\mathbb{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots\vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

được gọi là ma trận thiết kế

$$\mathbb{Y} = [y_1, ..., y_n]^T; \beta = [\beta_0, ..., \beta_k]^T; [\varepsilon_1, ..., \varepsilon_n]^T$$

và

$$E(\varepsilon) = 0 \tag{1.3}$$
$$cov(\varepsilon) = E(\varepsilon\varepsilon^T) = \sigma^2 I_n$$

* Bản chất của biến phụ thuộc Y

+ Y được giả định là một biến ngẫu nhiên, và có thể được đo lường bằng thang đo khoảng

* Bản chất của biến dự báo X

+ Các biến dự báo được giả định là phi ngẫu nhiên (nonrandom); nghĩa là, các giá trị của biến dự báo được giữ cố định khi lấy mẫu lặp đi lặp lại (repeated sampling).

* Cách xử lý các biến categorical

+ Vận dụng One-Hot-Encoding (Dummy Variables) cách phổ biến nhất để biểu diễn các biến phân loại là sử dụng mã hóa một mã hoặc một mã N, còn được gọi là biến giả. Ý tưởng đằng sau biến giả là thay thế một biến phân loại bằng một hoặc nhiều tính năng mới có thể có các giá trị 0 và 1. Các giá trị 0 và 1 có ý nghĩa trong công thức cho phân loại nhị phân tuyến tính và chúng ta có thể đại diện cho bất kỳ số lượng danh mục nào bằng cách giới thiệu một tính năng mới cho mỗi danh mục

* Ví dụ 1 bảng dữ liệu được xử lí bằng One-Hot-Encoding

CompanyName	Categoricalvalue	Price
VW	1	20000
Acura	2	10011
Honda	3	50000
Honda	3	10000

VW	Acura	Honda	Price
1	0	0	20000
0	1	0	10011
0	0	1	50000
0	0	1	10000

1.2 Ước lượng bình phương cực tiểu

- Bài toán đầu tiên đặt ra là dựa vào bộ số liệu quan sát được $\mathbb{X}, \mathbb{Y}$ hãy ước lượng tham số β, σ^2.
- Nếu ta sử dụng giá trị b là giá trị thử cho β thì giữa các quan sát y_j và $b_0 + b_1 x_{j1} + ... + b_k x_{jk}$ sẽ có độ lệch (sai số):

$$y_j - b_0 - (b_1 x_{j1} + ... + b_k x_{jk})$$

- Phương pháp bình phương tối thiểu là cách chọn giá trị véctơ b sao cho:

$$S(b) = \sum_{j=1}^{n} (y_j - b_0 - b_1 x_{j1} - \cdots - b_k x_{jk})^2$$

$$= (\mathbb{Y} - \mathbb{X}b)^T (\mathbb{Y} - \mathbb{X}b) \to min \tag{2.1}$$

- Đại lượng $\hat{\beta}$ làm cực tiểu hóa phiếm hàm $S(b)$ được gọi là ước lượng bình phương cực tiểu của β,

Ta có:

$$\widehat{\varepsilon_j} = y_j - (\widehat{\beta_0} + \widehat{\beta_1} x_{j1} + \cdots + \widehat{\beta_k} x_{jk}), j = 1 \div n \tag{2.2}$$

gọi là các phần dư của phép hồi quy.
Vì biểu thức theo $X_1, ..., X_k$ là tuyến tính nên phương trình:

$$\widehat{Y} = \widehat{\beta_0} + \widehat{\beta_1} X_1 + \cdots + \widehat{\beta_k} X_k \tag{2.3}$$

được gọi là **phương trình hồi quy tuyến tính mẫu**
Đặt:

$$\begin{aligned} \widehat{y_j} &= \widehat{\beta_0} + \widehat{\beta_1} x_{j1} + \cdots + \widehat{\beta_k} x_{jk} \\ \widehat{\mathbb{Y}} &= (\widehat{y_1}, ..., \widehat{y_n})^T \end{aligned} \tag{2.4}$$

Mệnh đề 1.2.1
Nếu ma trận thiết kế $\mathbb{X}$ không ngẫu nhiên có hạng $k+1 \leq n$ thì ước lượng bình phương cực tiểu có dạng:

$$\widehat{\beta} = (\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \mathbb{Y} \tag{2.5}$$

Khi đó

$$\widehat{Y} = \mathbb{X}\widehat{\beta} = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{Y} = H\mathbb{Y} \tag{2.6}$$

trong đó:

$$H = \mathbb{X}(\mathbb{X}^T \mathbb{X})^{-1} \mathbb{X}^T \text{ cấp } (n \times n) \tag{2.7}$$

$$\widehat{\varepsilon} = \mathbb{Y} - \widehat{Y} = (I_n - H)\mathbb{Y} \tag{2.8}$$

thỏa mãn:

$$\mathbb{X}^T \widehat{\varepsilon} = 0 \text{ và } \widehat{Y}^T \widehat{\varepsilon} = 0, (\widehat{\beta}^T \mathbb{X}^T \widehat{\varepsilon} = 0) \tag{2.9}$$

Tổng các phần dư:

$$\sum_{j=1}^{n} \widehat{\varepsilon_j^2} = \widehat{\varepsilon}^T \widehat{\varepsilon} = \mathbb{Y}^T \mathbb{Y} - \mathbb{Y}^T \mathbb{X} \widehat{\beta} \tag{2.10}$$

Chứng minh
Vì phiếm hàm $S(b) = \sum\limits_{j=1}^{n} (y_j - b_0 - b_1 x_{j1} - \cdots - b_k x_{jk})^2$ là hàm bậc hai theo b nên dễ thấy $\widehat{\beta}$ có thể tìm được từ hệ phương trình sau:

$$\frac{\partial S}{\partial b_i} = 0, i = 0 \div k$$

ta có kết quả:

$$\begin{aligned} \sum_{j=1}^{n} (b_0 + b_1 x_{j1} + \cdots + b_k x_{jk}) &= \sum_{j=1}^{n} y_i \\ b_0 \sum_{j=1}^{n} x_{j1} + b_1 \sum_{j=1}^{n} x_{j1}^2 + \cdots + b_k \sum_{j=1}^{n} x_{jk} x_{j1} &= \sum_{j=1}^{n} y_i x_{j1} \\ b_0 \sum_{j=1}^{n} x_{j1} + b_1 \sum_{j=1}^{n} x_{j1} x_{jk} + \cdots + b_k \sum_{j=1}^{n} x_{jk}^2 &= \sum_{j=1}^{n} y_i x_{jk} \end{aligned}$$

Nếu đặt $x_{j0} = 1, j = 1 \div n$ ta có phương trình sau:

$$\begin{bmatrix} \sum_{j=1}^{n} x_{j0}^2 & \sum_{j=1}^{n} x_{j0}x_{j1} & \cdots & \sum_{j=1}^{n} x_{j0}x_{jk} \\ \sum_{j=1}^{n} x_{j1}x_{j0} & \sum_{j=1}^{n} x_{j1}^2 & \cdots & \sum_{j=1}^{n} x_{j1}x_{jk} \\ \cdots & \cdots & \cdots & \cdots \\ \sum_{j=1}^{n} x_{jk}x_{j0} & \sum_{j=1}^{n} x_{jk}x_{j1} & \cdots & \sum_{j=1}^{n} x_{jk}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} y_j x_{j0} \\ \sum_{j=1}^{n} y_j x_{j1} \\ \vdots \\ \sum_{j=1}^{n} y_j x_{jk} \end{bmatrix}$$

hoặc dưới dạng ma trận:

$$\mathbb{X}^T\mathbb{X}b = \mathbb{X}^T\mathbb{Y} \tag{2.11}$$

Phương trình (2.11) gọi là phương trình chuẩn.
Do $rank(\mathbb{X}) = k + 1$ nên $\mathbb{X}^T\mathbb{X}$ có nghịch đảo, ta suy ra nghiệm:

$$b = \widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$$

Ta thấy $\widehat{\beta}$ là biểu thức tuyến tính theo $\mathbb{Y}$.
Để chứng minh $\widehat{\beta}$ cực tiểu hóa $S(b)$ và thỏa mãn (2.9), (2.10) ta chú ý rằng ma trận H có tính chất sau:

$$(I - H) \text{ là ma trận đối xứng: } (I - H)^T = (I - H)$$

$$(I - H)^2 = (I - H) \text{ tức là } I - H \text{ là ma trận lũy đẳng} \tag{2.12}$$

$$\mathbb{X}(I - H) = \mathbb{X}^T(I - \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T) = \mathbb{X}^T - \mathbb{X}^T = 0 \tag{2.13}$$

Dễ dàng thấy rằng:

$$\begin{aligned} S(b) &= (\mathbb{Y} - \mathbb{X}b)^T(\mathbb{Y} - \mathbb{X}b) = (\mathbb{Y} - \mathbb{X}\widehat{\beta} + \mathbb{X}\widehat{\beta} - \mathbb{X}b)^T(\mathbb{Y} - \mathbb{X}\widehat{\beta} + \mathbb{X}\widehat{\beta} - \mathbb{X}b) \\ &= (\mathbb{Y} - \mathbb{X}\widehat{\beta})^T(\mathbb{Y} - \mathbb{X}\widehat{\beta}) + (\widehat{\beta} - b)^T\mathbb{X}^T\mathbb{X}(\widehat{\beta} - b) \\ &\quad +(\widehat{\beta} - b)^T\mathbb{X}^T(I - H)\mathbb{Y} + \mathbb{Y}^T(I - H)^T\mathbb{X}(\widehat{\beta} - b) \\ &= (\mathbb{Y} - \mathbb{X}\widehat{\beta})^T(\mathbb{Y} - \mathbb{X}\widehat{\beta} + (\widehat{\beta} - b)^T\mathbb{X}^T\mathbb{X}(\widehat{\beta} - b) \\ &\geq (\mathbb{Y} - \mathbb{X}\widehat{\beta})^T(\mathbb{Y} - \mathbb{X}\widehat{\beta}) = S(\widehat{\beta}) \end{aligned}$$

Dấu "=" xảy ra khi $\widehat{\beta} = b$. Hơn nữa:
$\sum_{j=1}^{n} \widehat{\varepsilon_j}^2 = S(\widehat{\beta}) = (\mathbb{Y} - \mathbb{X}\widehat{\beta})^T(\mathbb{Y} - \mathbb{X}\widehat{\beta}) = \mathbb{Y}^T(I - H)(I - H)\mathbb{Y}$
$= \mathbb{Y}^T(I - H)\mathbb{Y}$(tính chất 2) $= \mathbb{Y}^T\mathbb{Y} - \mathbb{Y}^TH\mathbb{Y} = \mathbb{Y}^T\mathbb{Y} - (\mathbb{Y}^T\mathbb{X})\widehat{\beta}$
Đây chính là công thức (2.10).
Từ (2.8),(2.9),(2.10) ta nhận được: $\mathbb{Y}^T\mathbb{Y} = \sum_{j=1}^{n} y_j^2 = \widehat{\mathbb{Y}}^T\widehat{\mathbb{Y}} + \widehat{\varepsilon}^T\widehat{\varepsilon}$
hoặc:

$$\sum_{j=1}^{n} y_j^2 = \sum_{j=1}^{n} \widehat{y}_j^2 + \sum_{j=1}^{n} \widehat{\varepsilon}_j^2 \tag{2.14}$$

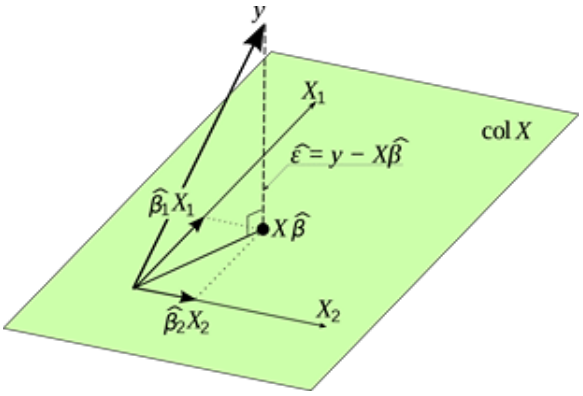
1.2.1 Minh họa hình học của ước lượng bình phương cực tiểu

Dựa theo mô hình hồi quy tuyến tính cổ điển

$$\mathbb{Y} = E(\mathbb{Y}) = \mathbb{X}\beta = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{n1} \end{bmatrix} + \cdots + \beta_k \begin{bmatrix} x_{1k} \\ x_{2k} \\ \vdots \\ x_{nk} \end{bmatrix}$$

Có thể thấy, $E(\mathbb{Y})$ là tổ hợp tuyến tính các cột của ma trận $\mathbb{X}$. Khi β thay đổi, $\mathbb{X}\beta$ thay đổi trên mặt phẳng mẫu chứa các tổ hợp tuyến tính. Vì có sai số ngẫu nhiên ε, vec-tơ quan sát y không nằm trên mặt phẳng mẫu. Vì vậy, y không hẳn là tổ hợp tuyến tính các cột của ma trận $\mathbb{X}$. Đã có:

$$\underbrace{\mathbb{Y}}_{n\times 1} = \underbrace{\mathbb{X}}_{n\times(k+1)} \cdot \underbrace{\beta}_{(k+1)\times 1} + \underbrace{\varepsilon}_{n\times 1} \tag{2.15}$$



Phương pháp bình phương tối thiểu được suy ra từ vec-tơ sai số:

$$\mathbb{Y} - \mathbb{X}b = \text{(vec-tơ quan sát) - (vec-tơ trên mặt phẳng mẫu)}$$

Bình phương độ dài $(\mathbb{Y} - \mathbb{X}b)^T(\mathbb{Y} - \mathbb{X}b)$ là tổng các bình phương $S(b)$. Theo minh họa trên, giá trị của b được chọn sao cho $S(b)$ nhỏ nhất có thể, điểm $\mathbb{X}b$ nằm trên mặt phẳng mẫu gần $\mathbb{Y}$ nhất. Điểm này trùng với chân đường cao hạ từ $\mathbb{Y}$ lên mặt phẳng mẫu.

Với giá trị $b = \widehat{\beta}$, $\widehat{Y} = \mathbb{X}\widehat{\beta}$ là hình chiếu của $\mathbb{Y}$ lên mặt phẳng chứa tổ hợp tuyến tính các cột của $\mathbb{X}$. Vec-tơ $\widehat{\varepsilon} = \mathbb{Y} - \widehat{Y}$ là vec-tơ trực giao của mặt phẳng đó.

1.2.2 Tính chất ước lượng bằng phương pháp bình phương cực tiểu

- Ước lượng $\widehat{\beta}$ là ước lượng không chệch với:

$$E(\widehat{\beta}) = \beta; cov(\widehat{\beta}) = \sigma^2(\mathbb{X}^T\mathbb{X})^{-1} \tag{2.16}$$

- Phần dư $\widehat{\varepsilon}$ có tính chất:

$$E(\widehat{\varepsilon}) = 0; cov(\widehat{\varepsilon}) = \sigma^2(I - H) \tag{2.17}$$

- $\widehat{\sigma}^2 = \dfrac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{n-k-1} = \sum\limits_{j=1}^{n} \dfrac{\widehat{\varepsilon}_j^2}{n-k-1}$ là ước lượng không chệch của σ^2, tức là $E(\widehat{\sigma}^2) = \sigma^2$
- $\widehat{\beta}, \widehat{\varepsilon}$ là không tương quan, tức là:

$$cov(\widehat{\beta}, \widehat{\varepsilon}) = 0; cov(\widehat{\beta}, \widehat{\sigma}^2) = 0 \tag{2.18}$$

Chứng minh

1)

$$\begin{aligned} E\widehat{\beta} &= E(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T E(\mathbb{Y}) \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}\beta = \beta \\ cov(\widehat{\beta}) &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T cov(\mathbb{Y})\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T I\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1} \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1} \end{aligned}$$

2) Do $\widehat{\varepsilon} = (I-H)\mathbb{Y}$ (theo (2.8)) nên:

$$\begin{aligned} E(\widehat{\varepsilon}) &= (I-H)E(\mathbb{Y}) = (I-H)\mathbb{X}\beta = 0, \\ cov(\widehat{\varepsilon}) &= (I-H)I(I-H)\sigma^2 = \sigma^2(I-H). \end{aligned}$$

3) Từ (2) ta suy ra:

$$\begin{aligned} E(\widehat{\varepsilon}^T\widehat{\varepsilon}) &= \sum_{1}^{n} E\widehat{\varepsilon_j^2} = tr(cov(\widehat{\varepsilon})) = \sigma^2 tr(I_n - H) \\ &= \sigma^2(n - tr(H)) \end{aligned}$$

Mặt khác,

$$\begin{aligned} tr(H) &= tr(\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T) = tr((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}) = tr(I_{k+1}) = k+1 \\ \Rightarrow E(\widehat{\varepsilon}^T\widehat{\varepsilon}) &= \sigma^2(n-k-1). \end{aligned}$$

4) Ta có:

$$\begin{aligned} cov(\widehat{\beta}, \widehat{\varepsilon}) &= cov((\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}(I_n - H)\mathbb{Y}) \\ &= (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T cov(\mathbb{Y})(I_n - H) \\ &= \sigma^2(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T(I_n - H) = 0 \end{aligned}$$

1.2.3 Định lý Gauss về ước lượng bình phương cực tiểu

Trong mô hình tuyến tính cổ điển (1.2),(1.3) với hạng đầy đủ $k+1 \leq n$ thì ước lượng:

$$c^T\widehat{\beta} = c_0\widehat{\beta_0} + c_1\widehat{\beta_1} + \cdots + c_k\widehat{\beta_k} \tag{2.19}$$

của $c^T\beta = c_0\beta_0 + c_1\beta_1 + \cdots + c_0\beta_0$ là ước lượng không chệch với phương sai bé nhất so với bất kỳ ước lượng tuyến tính không chệch nào dạng $a^T\mathbb{Y} = a_1y_1 + \cdots + a_ny_n$ Nếu thêm giả thiết rằng ε có phân bố chuẩn $N_n(0, \sigma^2 I_n)$ thì $c^T\widehat{\beta}$ là một ước lượng không chệch với phương sai cực tiểu của $c^T\beta$ so với bất kỳ ước lượng không chệch nào khác.

Chứng minh

1) Do tính chất tuyến tính của kỳ vọng nên rõ ràng $c^T\widehat{\beta}$ là ước lương không chệch của $c^T\beta$. Hơn nữa giả sử $a^T\mathbb{Y}$ là một ước lượng không chệch của $c^T\beta$ thì:

$$E(a^T\mathbb{Y}) = a^T E(\mathbb{Y}) = a^T\mathbb{X}\beta \equiv c^T\beta \Leftrightarrow (a^T\mathbb{X} - c^T)\beta \equiv 0$$

với mọi β, đặc biệt khi $\beta^T = a^T\mathbb{X} - c^T$ ta có:

$$\beta^T\beta = 0 \Leftrightarrow a^T\mathbb{X} - c^T = 0 \Leftrightarrow a^T\mathbb{X} = c^T \tag{2.20}$$

Chú ý rằng

$$c^T\widehat{\beta} = c^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} = a^{*T}\mathbb{Y} \tag{2.21}$$

với $a^{*T} = c^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} \Leftrightarrow a^* = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}c$.

$$\begin{aligned} D(a^T\mathbb{Y}) &= a^T cov(\mathbb{Y})a = \sigma^2 a^T a \\ &= \sigma^2(a - a^* + a^*)^T(a - a^* + a^*) \\ &= \sigma^2(a - a^*)^T(a - a^*) + \sigma^2(a^{*T}a^*) + 2(a - a^*)^T a^*\sigma^2 \\ &= \sigma^2(a - a^*)^T(a - a^*) + \sigma^2 a^{*T}a^* \geq D(a^*\mathbb{Y}) \end{aligned} \tag{2.22}$$

Vì

$$\begin{aligned} (a - a^*)^T a^* &= a^T a^* - a^{*T}a^* \\ &= a^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}c - c^T(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}c \\ &= c^T(\mathbb{X}^T\mathbb{X})^{-1}c - c^T(\mathbb{X}^T\mathbb{X})^{-1}c \\ &= c^T(\mathbb{X}^T\mathbb{X})^{-1}c - c^T(\mathbb{X}^T\mathbb{X})^{-1}c = 0 \end{aligned}$$

Trong (2.22) dấu"=" xảy ra khi và chỉ khi $a = a^*$.

2) Xem **Thống kê toán** - Đào Hữu Hồ, Nguyễn Văn Hữu, Hoàng Hữu Như

1.2.4 Hệ số xác định R

Đại lượng

$$R^2 := \frac{\widehat{Y}^T\widehat{Y} - n(\overline{y})^2}{\mathbb{Y}^T\mathbb{Y} - n(\overline{y})^2} = \frac{\Sigma_1^n\widehat{y}_j^2 - n(\overline{y})^2}{\Sigma_1^n y_j^2 - n(\overline{y})^2} \tag{2.23}$$

gọi là *bình phương của hệ số xác định*, đó là tỷ lệ biến thiên của các biến y_j được giải thích bởi các biến $x_{j1}, ..., x_{jk}$.
Từ (2.14) ta có:

$$\sum_{j=1}^n \widehat{\varepsilon}_j^2 = \left[\sum_{j=1}^n y_j^2 - n(\overline{y})^2\right](1 - R^2) = ns_y^2(1 - R^2) \tag{2.24}$$

ta nhận được phương trình để tính sai số bình phương trung bình.

1.3 Mô hình hồi quy tuyến tính

1.3.1 Phân phối F, phân phối Student

Phân phối F

Định nghĩa:

Nếu gọi U_1 là phân phối khi bình phương với bậc tự do d_1 Và U_2 là phân phối khi bình phương với bậc tự do d_2 thì phân phối F là tỉ số giữa 2 phân phối khi bình phương U_1 và U_2

$$\boxed{F \sim \frac{\frac{U_1}{d_1}}{\frac{U_2}{d_2}}} \tag{3.25}$$

Như vậy phân phối F cũng là phân phối khi bình phương có trị trung bình μ và phương sai σ^2. Phân phối F diễn tả phân phối xác suất liên tục với tần suất xuất hiện tương tự như phân phối rỗng của thống kê kiểm thử. Nó thường hay xảy ra trong quá trình phân tích phương sai và kiểm thử F.

Hàm mật độ:

$$\boxed{f(x; d_1, d_2) = \frac{\sqrt{\frac{(d_1 x)^{d_1} {d_2}^{d_2}}{(d_1 x + d_2)^{d_1 + d_2}}}}{x\beta(\frac{d_1}{2}, \frac{d_2}{2})}}$$

Với:

- d_1: biến dương
- d_2: biến dương
- x: biến ngẫu nhiên

Các tham số đặc trưng:

- $EX = \mu = \frac{d_2}{d_2 - 2}$ (với $d_2 > 2$)
- $VX = \sigma^2 = \frac{2{d_2}^2(d_1 + d_2) - 2}{d_1(d_2 - 2)^2(d_2 - 4)}$ (với $d_2 > 4$)

Ứng dụng trong phân tích phương sai **ANOVA**

Phân phối Student

Định nghĩa: Giả sử $X \sim N(0; 1)$ và $Y \sim X^2(n)$ là hai biến ngẫu nhiên độc lập. Khi đó

$$\boxed{T = \frac{X}{\sqrt{\frac{Y}{n}}}}$$

được gọi là tuân theo phân phối Student với n bậc tự do. Kí hiệu: $T \sim T(n)$. Ứng dụng trong thống kê suy luận phương sai tổng thể khi tổng thể được giả thiết là có phân phối chuẩn, đặc biệt khi cỡ mẫu nhỏ. Ngoài ra ta còn dùng phân phối Student trong kiểm định giả thiết về trung bình khi phương sai tổng thể chưa biết.

Hàm mật độ:

$$\boxed{f(x) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{\pi n}\Gamma(\frac{n}{2})(1 + \frac{x^2}{n})^{\frac{n+1}{2}}}}$$

với $x \in R$

Các tham số đặc trưng:

- $ET = 0$
- $VT = \frac{n}{n-2}$

Chú ý

- Phân phối Student có cùng dạng và tính đối xứng như phân phối chuẩn nhưng nó phản ánh tính biến đổi của phân phối sâu sắc hơn. Phân phối chuẩn không thể dùng để xấp xỉ phân phối mẫu khi có kích thước nhỏ. Trong trường hợp này ta dùng phân phối Student.
- Khi bậc tự do n tăng lên (n > 30) thì phân phối Student tiến nhanh về phân phối chuẩn. Do đó khi n > 30 ta có thể dùng phân phối chuẩn thay thế cho phân phối Student.

1.3.2 Khoảng tin cậy của các hệ số hồi quy β_j

Trong phần này ta xét mô hình hồi quy cổ điển với giả thiết thêm rằng: các ε_j có cùng phân phối chuẩn $N(0, \sigma^2)$ và độc lập, tức là $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ có phân bố chuẩn $N_n(0, \sigma^2 I_n)$

Mệnh đề 1.3.1

1. *$\widehat{\beta}$ có phân bố chuẩn $N_{k+1}(\beta, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$*
2. *$\frac{(n-k-1)\widehat{\sigma}^2}{\sigma^2} = \frac{\Sigma_{j=1}^n \widehat{\varepsilon_j}^2}{\sigma^2}$ có phân bố χ^2 với (n-k-1) bậc tự do.*
3. *$\widehat{\beta}, \widehat{\sigma}^2$ là độc lập.*

Chứng minh

$\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$; $\widehat{\varepsilon} = (I - H)\mathbb{Y}$ là các tổ hợp tuyến tính của vecto Y có phân bố chuẩn $N_n(\mathbb{X}\beta, \sigma^2 I_n)$.
Vì vậy, $\widehat{\beta}$ có phân bố chuẩn $N_{k+1}(\beta, \sigma^2(\mathbb{X}^{\mathbb{T}}\mathbb{X}^{-1}))$, $\widehat{\varepsilon}$ có phân bố chuẩn $N(0, \sigma^2(I - H))$, $cov(\widehat{\beta}, \widehat{\varepsilon}) = 0$ và $(\widehat{\beta}, \widehat{\varepsilon})^T$ có phân bố chuẩn đồng thời chuẩn.
Nên theo tính chất của phân bố chuẩn, ta có :

$$\frac{\widehat{\varepsilon}^T\widehat{\varepsilon}}{\sigma^2} = \sum_{j=1}^{n} \frac{\widehat{\varepsilon_j}^2}{\sigma^2}$$

có phân bố χ^2 với $n - k - 1$ bậc tự do. Thật vậy,

(i) Vì $(I - H)$ là ma trận lũy đẳng nên nếu ta ký hiệu λ và ε là cặp giá trị riêng và vecto riêng của $(I - H)$, ta sẽ có:

$$(I - H)e = \lambda e \Rightarrow (I - H)^2 e = \lambda(I - H)e = \lambda^2 e$$

hoặc $(I - H)e = \lambda^2 e = \lambda e$. Do đó $\lambda = \lambda^2$. Vậy $\lambda = 0$ hoặc 1.
Vì $\text{tr}(I - H) = n - k - 1 = \lambda_1 + ... + \lambda_n$ nên $n - k - 1$ giá trị riêng đầu tiên của $I - H$ là 1 còn$k + 1$ giá trị riêng còn lại bằng 0.

(ii) Giả sử $e_1, ...e_{n-k-1}$ là $n - k - 1$ vecto riêng ứng với giá trị riêng là 1 còn $k + 1$ vecto riêng ứng với giá trị riêng 0 của ma trận $I - H$. Theo công thức khai triển phổ của ma trận ta có:

$$I - H = e_1 e_1^T + ... + e_{n-k-1} e_{n-k-1}^T$$

Đặt

$$V = \begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_{n-k-1} \end{bmatrix} = \begin{bmatrix} e_1^T\varepsilon \\ e_2^T\varepsilon \\ \vdots \\ e_{n-k-1}^T\varepsilon \end{bmatrix}$$

Khi đó V có phân bố chuẩn với $E(V) = 0$, còn

$$cov(V_i, V_j) = e_i^T(\sigma^2 I)e_i = \begin{cases} \sigma^2 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

nên $V_1, ..., V_{n-k-1}$ có phân bố chuẩn độc lập $N(0, 1)$ và V có $N(0, \sigma^2 I_{n-k-1})$.

(ii) Giả sử $e_1, ...e_{n-k-1}$ là $n - k - 1$ vecto riêng ứng với giá trị riêng là 1 còn $k + 1$ vecto riêng ứng với giá trị riêng 0 của ma trận $I - H$. Theo công thức khai triển phổ của ma trận ta có:

$$I - H = e_1e_1^T + ... + e_{n-k-1}e_{n-k-1}^T$$

Mệnh đề 1.3.2
Xét mô hình hồi quy tuyến tính cổ điển $\mathbb{Y} = \mathbb{X}\beta + \varepsilon$ với $\mathbb{X}$ có hạng là $k + 1 \leq n$ và $\varepsilon \sim N(0, \sigma^2 I_n)$. Khi đó miền tin cậy đồng thời mức $(1 - \alpha)$ của β xác định bởi:

$$(\beta - \widehat{\beta})^T\mathbb{X}^T\mathbb{X}(\beta - \widehat{\beta}) \leq (k + 1)\widehat{\sigma}^2F_{k+1,n-k-1}(\alpha) \tag{3.26}$$

trong đó $F_{k+1,n-k-1}(\alpha)$ là phân vị trên mức α của phân bố F với bậc tự do là $k + 1, n - k - 1$. Nói cách khác, với độ tin cậy $(1 - \alpha)$, giá trị chân thực β phải nằm trong Ellipsoid:

$$(x - \widehat{\beta})^T\mathbb{X}^T\mathbb{X}(x - \widehat{\beta}) = (k + 1)\widehat{\sigma}^2F_{k+1,n-k-1}(\alpha)$$

Hơn nữa khoảng tin cậy đồng thời mức $(1 - \alpha)$ của các $\beta_i, i = 0 \div k$ được xác định bởi các mút:

$$\widehat{\beta}_i \pm \sqrt{\widehat{D}(\widehat{\beta}_i)(k + 1)F_{k+1,n-k-1}(\alpha)} \tag{3.27}$$

trong đó $\widehat{D}(\widehat{\beta}_i)$ ký hiệu phần tử thứ i trên đường chéo chính của ma trận $\widehat{\sigma}^2(\mathbb{X}^T\mathbb{X})^{-1}$ và là ước lượng không chệch của $D(\widehat{\beta})$

Chứng minh
Xét ma trận căn bậc hai đối xứng $(X^TX)^{1/2}$ và đặt

$$U = (X^TX)^{1/2}(\widehat{\beta} - \beta).$$

Ta có:

$$\begin{aligned} E(U) &= 0 \\ cov(U) &= (X^TX)^{1/2}cov(\widehat{\beta})(X^TX)^{1/2} \\ &= \sigma^2(X^TX)^{1/2}(X^TX)^{-1}(X^TX)^{1/2} = \sigma^2 I_{k+1} \end{aligned}$$

Vậy U có phân bố chuẩn $N(0, \sigma^2 I_{k+1})$. Do đó $\frac{1}{\sigma^2}U^TU = \frac{1}{\sigma^2}(\widehat{\beta} - \beta)^T(X^TX)(\widehat{\beta} - \beta)$ có phân phối χ^2 với $k + 1$ bậc tự do. Hơn nữa, theo mệnh đề 2.1, $(n - k - 1)\frac{\widehat{\sigma}^2}{\sigma^2}$ có phân bố χ^2 với $n - k - 1$bậc tự do và độc lập với $\widehat{\beta}$, tức là độc lập với U^TU. Vì vậy đại lương

$$F = \frac{(\widehat{\beta} - \beta)^T\mathbb{X}^T\mathbb{X}(\widehat{\beta} - \beta)/(k + 1)}{\widehat{\sigma}^2} = \frac{U^TU/(k + 1)\sigma^2}{(n - k - 1)\widehat{\sigma}^2/(n - k - 1)\sigma^2}$$

có phân bố F với $k+1$ và $n-k-1$ bậc tự do. Từ đó

$$P\{F \leq F_{k+1,n-k-1}(\alpha)\} = 1-\alpha$$

hoặc

$$P(\widehat{\beta}-\beta)^T\mathbb{X}^T\mathbb{X}(\widehat{\beta}-\beta) \leq (k+1)\widehat{\sigma}^2 F_{k+1,n-k-1}(\alpha) = 1-\alpha$$

Mệnh đề 1.3.3
Giả sử $t_{n-k-1}\left(\dfrac{\alpha}{2(k+1)}\right)$ *là phân vị trên mức* $\dfrac{\alpha}{2(k+1)}$ *của phân bố Student với* $n-k-1$ *bậc tự do. Khi đó đồng thời ta có các khoảng tin cậy của* β_i*, với mức tin cậy* $(1-\alpha)$ *cho bởi các đầu mút:*

$$\widehat{\beta_i} \pm t_{n-k-1}\left(\frac{\alpha}{2(k+1)}\right)\sqrt{\widehat{D}(\widehat{\beta_i})} \tag{3.28}$$

Ví dụ 1 Các công ty muốn mua máy tính trước hết phải đánh giá các nhu cầu trong tương lai. Dưới đây là dữ liệu thu thập được từ 7 công ty để đưa ra hàm dự đoán về nhu cầu phần cứng:

STT	x_0	x_1(orders)	x_2(add-delete items)	y(CPU time)
1	1	125.3	2.108	141.5
2	1	146.1	9.213	168.9
3	1	133.9	1.905	154.8
4	1	128.5	0.815	146.5
5	1	151.5	1.061	172.8
6	1	136.2	8.603	160.1
7	1	92	1.125	108.5

Giả sử các thông số này tuân theo mô hình tuyến tính cổ điển, khi đó:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \varepsilon_j, j = 1 \div 7$$

Ta sẽ ước lượng các hệ số hồi quy bằng phương pháp bình phương cực tiểu.

$$(\mathbb{X}^T\mathbb{X})^{-1} = \begin{bmatrix} 8.18996 & -0.06423 & 0.0916 \\ -0.06423 & 0.00052 & -0.00111 \\ 0.09 & -0.001 & 0.015 \end{bmatrix}$$

$$\widehat{\beta} = \begin{bmatrix} \widehat{\beta_0} \\ \widehat{\beta_1} \\ \widehat{\beta_2} \end{bmatrix} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y} = \begin{bmatrix} 7.86 \\ 1.08 \\ 0.45 \end{bmatrix}$$

Vậy phương trình hồi quy tuyến tính mẫu là:

$$\widehat{y} = 7.86 + 1.08x_1 + 0.45x_2$$

Tổng bình phương các phần dư là:

$$\sum_1^n \widehat{\varepsilon_j}^2 = \sum_1^n y_j^2 - \mathbb{Y}^T\mathbb{X}\widehat{\beta} = 13.06195$$

$$\widehat{\sigma}^2 = \frac{1}{n-k-1}\sum_1^n \widehat{\varepsilon_j}^2 = \frac{13.06195}{4} = 3.265486$$

Sau đây là bảng tính các giá trị $\widehat{y_j}, \widehat{\varepsilon_j}$

STT	y_j	$\widehat{y_j}$	$\widehat{\varepsilon_j}$
1	141.5	144.1326	-2.6326
2	168.9	169.7939	-0.89385
3	154.8	155.5053	-0.70525
4	146.5	147	-0.5
5	172.8	171.9575	0.84255
6	160.1	158.8274	1.27265
7	108.5	107.7263	0.77375

Tổng phần dư bằng -1.84275.
Ta có:

$$\widehat{D}(\widehat{\beta_0}) = 3.265486 \times 8.18996 \Rightarrow \sqrt{\widehat{D}(\widehat{\beta_0})} = 5.17148$$
$$\widehat{D}(\widehat{\beta_1}) = 3.265486 \times 0.00052 \Rightarrow \sqrt{\widehat{D}(\widehat{\beta_1})} = 0.0412$$
$$\widehat{D}(\widehat{\beta_2}) = 3.265486 \times 0.015 \Rightarrow \sqrt{\widehat{D}(\widehat{\beta_3})} = 0.22132$$

Khoảng tin cậy của $\beta_0, \beta_1, \beta_2$ mức 0,95:

$$\widehat{\beta_0} \pm t_4 \left(\frac{0,05}{2 \times 3}\right) \sqrt{\widehat{D}(\widehat{\beta_0})} = 7.86 \pm 20.252$$
$$\widehat{\beta_1} \pm t_4 \left(\frac{0,05}{2 \times 3}\right) \sqrt{\widehat{D}(\widehat{\beta_1})} = 1.08 \pm 0.16134$$
$$\widehat{\beta_2} \pm t_4 \left(\frac{0,05}{2 \times 3}\right) \sqrt{\widehat{D}(\widehat{\beta_2})} = 0.45 \pm 0.86671$$

Do đó ta có kết quả:

$$\beta_0 \in (-12.392; 28.112)$$
$$\beta_1 \in (0.91866; 1.24134)$$
$$\beta_2 \in (-0.41671; 1.31671)$$

1.3.3 Kiểm định giả thiết về các hệ số hồi quy

Xét mô hình HQTT cổ điển

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon \tag{3.29}$$

Khi thiết lập phương trình, ta giả sử rằng mọi biến độc lập $X_1, ..., X_k$ đều tham gia phương trình hồi quy. Tuy nhiên, trên thực tế, có một vài biến sẽ không tham gia vào phương trình hồi quy, tức là hệ số β_i của nó bằng 0. Tuy vậy, các hệ số ước lượng có thể khác 0. Bài toán đặt ra là kiểm định xem khi nào hệ số ước lượng được xem là bằng 0 thực sự.

Ta có bài toán kiểm định giả thiết

$$H_0 : \beta_{p+1} = ... = \beta_k = 0 (0 < p < k) \tag{3.30}$$

với đối thiết $K : \exists i \in \{p+1, ..., k\}$ sao cho $\beta_i \neq 0$
Giả thiết H_0 có nghĩa là các biến độc lập không tham gia vào biểu thức tuyến tính, ngược lại đối thiết K nói rằng có ít nhất một trong các biến này có liên quan đên mô hình.

Tổng quát hơn ta xét bài toán kiểm định giả thiết dạng:

$$H_0 : \begin{cases} c_{10}\beta_0 + c_{11}\beta_1 + \cdots + c_{1k}\beta_k = a_1 \\ c_{20}\beta_0 + c_{21}\beta_1 + \cdots + c_{2k}\beta_k = a_2 \\ \cdots \\ c_{k-p,0}\beta_0 + c_{k-p,1}\beta_1 + \cdots + c_{k-p,k}\beta_k = a_{k-p} \end{cases} \Leftrightarrow C\beta = a \tag{3.31}$$

trong đó $C = [c_{ij}]_{k-p,k+1}; a = [a_1, ..., a_{k-p}]^T$
Bài toán đang xét ((3.30)) là trường hợp riêng của ((3.31)) với:

$$C = \begin{bmatrix} 0 & 0 & \cdots & 0 & \vdots & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & \vdots & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \vdots & 0 & 0 & \cdots & 1 \end{bmatrix} = [0 \vdots I_{k-p}]$$

Quy tắc kiểm định:

Bác bỏ giả thuyết $H_0 : C\beta = 0$ nếu:

$$(C\widehat{\beta})(C(\mathbb{X}^T\mathbb{X})^{-1}C^T)^{-1}C\widehat{\beta}/\widehat{\sigma}^2 > (k-p)F_{k-p,n-k-1}(\alpha) \tag{3.32}$$

Nhận xét: Ta có thể sử dụng mệnh đề (2.4) về khoảng tin cậy của $\beta_{p+1}, ..., \beta_k$ với các đầu mút $\widehat{\beta_i} \pm t_{n-k-1}\left(\frac{\alpha}{2(k+1)}\right)\sqrt{\widehat{D}(\widehat{\beta_i})}$ để kiểm định giả thuyết (2.24). Điều đó có nghĩa là nếu tồn tại chỉ số $i \in \{p+1, ..., k\}$ thỏa mãn:

$$|\widehat{\beta_i}| > t_{n-k-1}\left(\frac{\alpha}{2(k-p)}\right)\sqrt{\widehat{D}(\widehat{\beta_i})}$$

thì ta coi $\beta_i \neq 0$ **Ví dụ** Giả thuyết: $\beta_1 = \beta_2 = 0$. Đặt $C = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$
Có: $\widehat{\beta} = \begin{bmatrix} 7.86 \\ 1.08 \\ 0.45 \end{bmatrix}$ $(\mathbb{X}^T\mathbb{X})^{-1} = \begin{bmatrix} 8.18996 & -0.06423 & 0.0916 \\ -0.06423 & 0.00052 & -0.00111 \\ 0.09 & -0.001 & 0.015 \end{bmatrix}$

$\widehat{\sigma}^2 = 3.265486$
$\Rightarrow (C\widehat{\beta})(C(\mathbb{X}^T\mathbb{X})^{-1}C^T)^{-1}C\widehat{\beta}/\widehat{\sigma}^2 = 852.64$
$(k-p)F_{k-p,n-k-1}(\alpha) = 2.F_{2,4}(0.05) = 13.88$
Vì $852.64 > 13.88$ nên ta bác bỏ giả thuyết $\beta_1 = \beta_2 = 0$.

1.3.4 Ước lượng hàm hồi quy tuyến tính

Bài toán đặt ra là ước lượng hàm hồi quy tuyến tính:

$$E(Y|X) = \beta_0 + \beta_1 X_1 + ... + \beta_k X_k$$

tại điểm $X^0 = (1, X_1^0, ..., X_k^0)$ tức là ước lượng tổ hợp tuyến tính sau:

$$E(Y|X) = \beta_0 + \beta_1 X_1^0 + ... + \beta_k X_k^0 = {X^0}^T\beta \tag{3.33}$$

Theo định lý Gauss, ${X^0}^T\widehat{\beta}$ là ước lượng tuyến tính với phương sai cực tiểu

Nếu $\varepsilon \sim N(0, I_n\sigma^2)$ thì $X^{0^T}\widehat{\beta} \sim N(X^{0^T}\beta, \sigma^2 X^{0^T}(\mathbb{X}^T\mathbb{X})^{-1}X^0)$ và do đó khoảng tin cậy mức $(1-\alpha)$ của $X^{0^T}\beta$ chính là:

$$X^{0^T}\widehat{\beta} \pm t_{n-k-1}(\frac{\alpha}{2}) \underbrace{\widehat{\sigma}\sqrt{X^{0^T}(\mathbb{X}^T\mathbb{X})^{-1}X^0}}_{\sqrt{\widehat{D}(X^{0^T}\widehat{\beta})}} \tag{3.34}$$

hoặc

$$X^{0^T}\widehat{\beta} \pm t_{n-k-1}(\frac{\alpha}{2})\sqrt{\widehat{D}(X^{0^T}\widehat{\beta})} \tag{3.35}$$

Ta xét lại ví dụ 1: Các công ty muốn mua máy tính trước hết phải đánh giá các nhu cầu trong tương lai. Dưới đây là dữ liệu thu thập được từ 7 công ty để đưa ra hàm dự đoán về nhu cầu phần cứng:

STT	x_0	x_1(orders)	x_2(add-delete items)	y(CPU time)
1	1	125.3	2.108	141.5
2	1	146.1	9.213	168.9
3	1	133.9	1.905	154.8
4	1	128.5	0.815	146.5
5	1	151.5	1.061	172.8
6	1	136.2	8.603	160.1
7	1	92	1.125	108.5

Giả sử:

$$y_j = \beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2} + \varepsilon_j, j = 1 \div n, n = 7$$

với $\{\varepsilon_j\}$ là dãy độc lập có phân bố chuẩn $N(0, \sigma^2)$. Khi đó ta có thể dùng phương trình hồi quy tuyến tính mẫu:

$$\widehat{y} = 7.86 + 1.08x_1 + 0.45x_2$$

để dự đoán hàm hồi quy $E(Y|X_1 = 130, X_2 = 7.5) = \beta_0 + 1.13\beta_1 + 7.5\beta_2 = X^{0^T}\beta$, tại $X^{0^T} = (1, 130, 7.5)$. Ta có:

$$(\mathbb{X}^T\mathbb{X})^{-1} = \begin{bmatrix} 8.18996 & -0.06423 & 0.0916 \\ -0.06423 & 0.00052 & -0.00111 \\ 0.09 & -0.001 & 0.015 \end{bmatrix}$$

$\widehat{\sigma}^2 = 3.265486$

$$\widehat{\sigma}^2 X^{0^T}(\mathbb{X}^T\mathbb{X})^{-1}X^0 = 1.39325$$

$$\widehat{y_0} = 7.86 + 1.08 \times 130 + 0.45 \times 7.5 = 151.635$$

Vì vậy khoảng tin cậy mức 0,95 của $X^{0^T}\beta$ là:

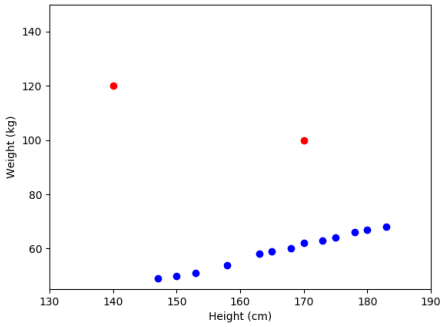
$$\begin{aligned} \widehat{y_0} \pm t_4(0,05/2)\sqrt{\widehat{\sigma}^2 X^{0^T}(\mathbb{X}^T\mathbb{X})^{-1}X^0} &= 151.635 \pm 2.776.1.18 \\ &= 151.635 \pm 3.27568 \end{aligned}$$

hay (148.35932;154.91068)

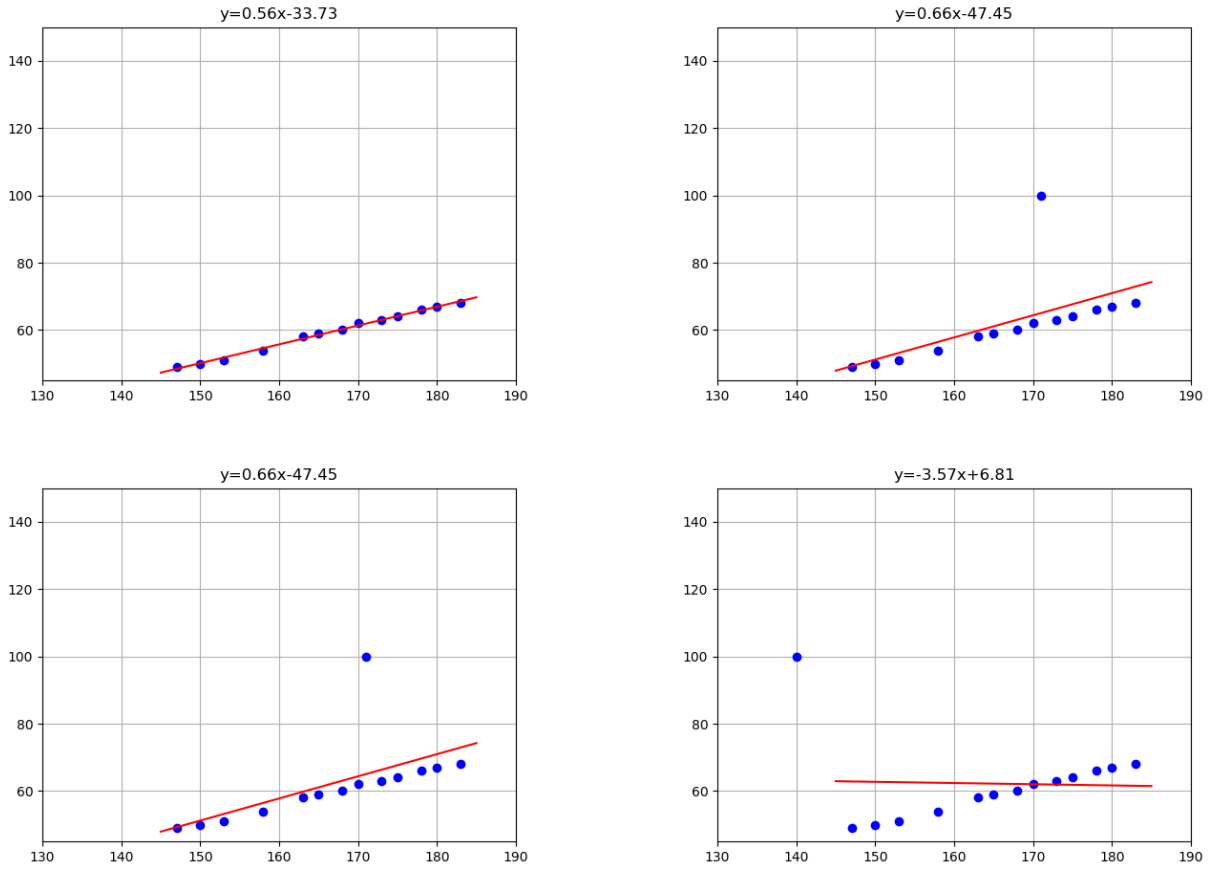
1.4 Kiểm tra sự phù hợp của mô hình

1.4.1 Outlier

Định nghĩa: Outlier là điểm dữ liệu bất thường, khác biệt so với phần còn lại của dữ liệu. [h]



Hình 1.1: Hai điểm nhiễu nằm rất xa so với phần còn lại của bộ dữ liệu



Hình 1.2: Tác động của Outlier đến việc tạo mô hình

Định nghĩa:
- Leverage là độ đo khoảng cách giữa 1 điểm dữ liệu và phần còn lại của bộ dữ liệu dựa trên miền giá trị của bộ dữ liệu.
- Là các phần tử nằm trên đường chéo chính của ma trận H.

Từ 4 hình trên, có thể thấy các điểm outlier có tác động rất lớn đến mô hình, và điểm nào có giá trị của biến dự đoán x nằm càng xa khỏi bộ dữ liệu thì tác động càng lớn. Ta gọi những điểm nằm xa như vậy là high leverage, tương tự nằm gần bộ dữ liệu là low leverage.

Công thức tính Leverage:

$$\boxed{h_{jj} = \frac{1}{n} + \frac{(x_j - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}}$$

Từ công thức trên, có thể thấy rằng mô hình các có nhiều điểm dữ liệu nằm gần nhau thì sức ảnh hưởng hay độ lớn của các điểm high leverage càng giảm.

1.4.2 Sự phù hợp của mô hình

Kiểm tra ý nghĩa của mô hình

Tiêu chuẩn F

Ngay từ đầu ta có cần dùng đến các biến độc lập x_i không hay chỉ cần một giá trị tụ do β_0 (tức là giá trị dự đoán y chỉ đơn giản dao động xung quanh hằng số) là đủ? Để trả lời câu hỏi này ta sẽ sử dụng tiêu chuẩn F (F-test).

Xét đại lượng:

$$F = \frac{(n-k-1)R^2}{k(1-R^2)}$$

với đại lượng R^2 được tính bằng công thức (2.23)

Giả thuyết: $H_0 : \beta_1 = ... = \beta_k = 0$
Đối thuyết: $\exists \beta_j \neq 0$ với $j = 1, .., k$

Các bước tiến hành F-test

Bước 1: Tính đại lượng F.
Bước 2: Tra bảng phân phối Fisher với bậc tự do k và n-k-1, mức ý nghĩa α.
Bước 3: Nếu $F > F_{k,n-k-1}(\alpha)$ thì bác bỏ H_0.

VD: Để nghiên cứu sự phụ thuộc giữa doanh thu Y và chi phí sản xuất X_1,chi phí tiếp thị X_2, người ta điều tra ngẫu nhiên doanh thu của 12 công ty trong 12 thời kỳ ,kết quả ta có bảng sau:

STT	x_1	x_2	y	STT	x_1	x_2	y
1	18	10	127	7	25	14	161
2	25	11	149	8	16	12	128
3	19	6	106	9	17	12	139
4	24	16	163	10	23	12	144
5	15	7	102	11	22	14	159
6	26	17	180	12	15	15	138

Sử dụng công thức ở trên, ta được $R^2 = 0.9756$

$$F = \frac{2.0,9756}{(12-2-1)(1-0,9756)} = 179,6292$$

Tra bảng F với mức ý nghĩa 0.02,ta được:

$$F_{2,9}(0,02) = 6,234$$

Vì $F > F_{2,9}(0,02)$ nên bác bỏ giả thuyết H_0, tức là có sự phụ thuộc vào các biến độc lập.

Kiểm tra tính đa cộng tuyến của $X_1, X_2, .., X_k$

Trong thực tế, các biến X_i có thể không độc lập với nhau, hay nói cách khác là có tương quan, như vậy ta có thể loại bỏ các biến này. Ngoài ra, nếu xảy ra hiện tượng đa cộng tuyến cũng có thể dẫn đến định thức của X^TX gần tới 0, các hệ số trong $\hat{\beta}$ (công thức (2.11)) trở nên rất lớn, gây ra sai số nghiêm trọng (hiện tượng bất ổn của hệ đại số tuyến tính).

Xác định hiện tượng đa cộng tuyến dựa trên các dấu hiệu:
+) Một số phần tử trên đường chéo chính của ma trận $(X^TX)^{-1}$ tỏ ra rất lớn
+) Các hệ số tương quan tuyến tính mẫu $r_{ij} = s_{ij}/\sqrt{s_{ii}s_{jj}}$ tỏ ra rất lớn ($|r_{ij}| > 0,7$).
trong đó $s_{ij} = \frac{1}{n}\sum_{k=1}^{n} X_{ki}X_{kj} - \bar{X}_i\bar{X}_j$

Khắc phục hiện tượng đa cộng tuyến

Để khắc phục hiện tượng đa cộng tuyến, ta làm như sau:

1. Tính các r_{ij}
2. Đặt r_{0i} là hệ số tương quan tuyến tính mẫu giữa Y và X_i cụ thể là:

$$r_{0i} = \frac{s_{0i}}{\sqrt{s_{ii}s_{00}}}$$

   trong đó $s_{00} = s_y^2; s_{0i} = \frac{1}{n}\sum_{j=1}^{n} y_j.x_{ji} - \bar{y} \times \bar{x_i}$
   Nếu $|r_{ij}| > 0.7$ thì:
   loại X_i ra khỏi mô hình nếu $|r_{0i}| < |r_{0j}|$
   loại X_j ra khỏi mô hình nếu $|r_{0i}| > |r_{0j}|$
3. Thực hiện hồi quy sau khi với ma trận X đã loại bỏ X_i hay X_j.

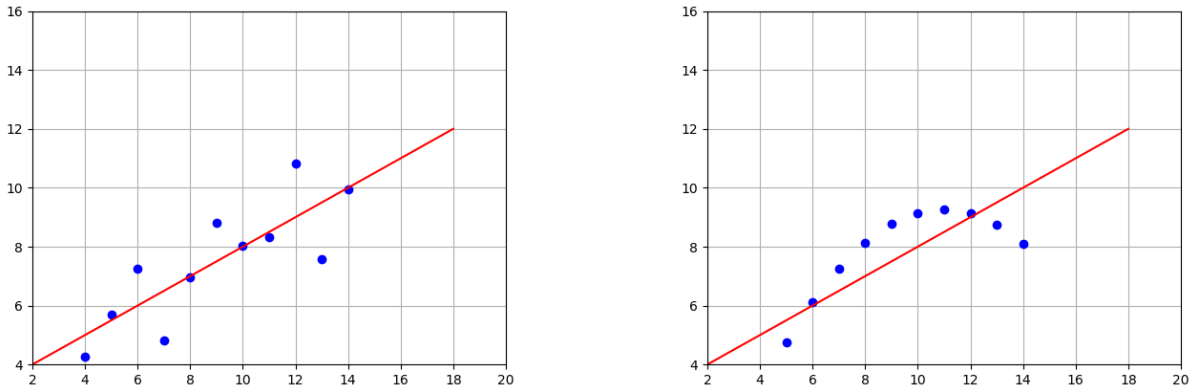
Ở bước 2, ta chọn X_i hoặc X_j để loại bỏ khỏi mô hình dựa trên sự tương quan của các biến này đối với y ,biến nào có hệ số tương quan sát với y hơn (gần 1) thì biến đó sẽ được giữ lại.

Ví dụ:
Ta có ma trận thiết kế và giá trị dự đoán sau:

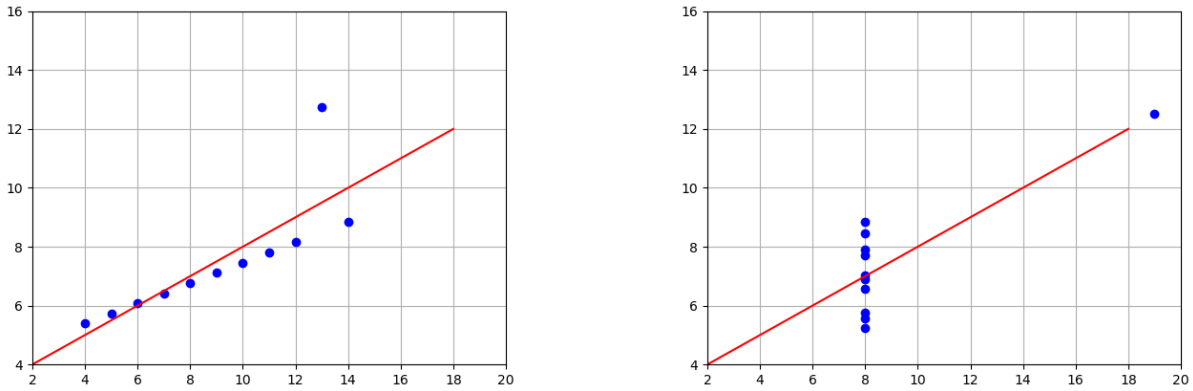
$$\text{X=} \begin{pmatrix} 1 & 1.9999 & 3.0001 \\ 1 & 4.0001 & 4.9999 \\ 1 & 0.0001 & 1.0001 \end{pmatrix} \text{ y} = (8\ ,14\ ,2)$$

Bước 1: Tìm các $|r_{ij}| > 0.7$
$r_{12} \approx 1$
Bước 2: Tính r_{0i}
$r_{01} \approx 0.9999$
$r_{02} \approx 0.9999$
$|r_{02}| \approx |r_{01}$, nên loại một trong 2 biến đều được.
Bước 3: Thực hiện hồi quy sau khi đã loại X_2 (hoặc X_1).

Kiểm tra bằng đồ thị



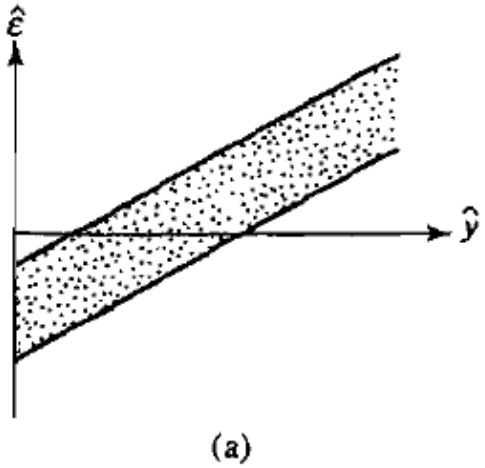
Hình 1.3: Bộ tứ Anscombe



	Giá trị
Mean x	9
Var x	11
Mean y	7.5
Var y	4.125
Cor	0.816

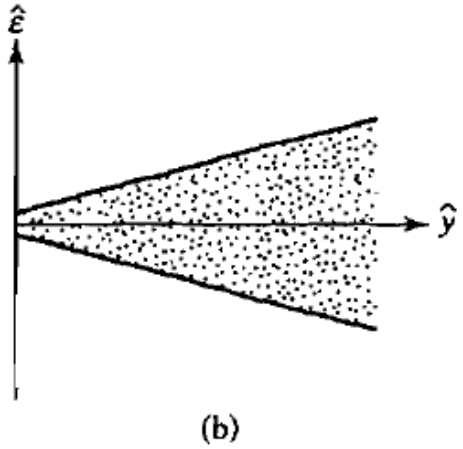
Giả thuyết - Phần dư $\hat{\epsilon_j}$ phụ thuộc vào biến $\hat{y_j}$. (i)
- Phương sai không phải là hằng số. (ii)
- Mô hình dự đoán bỏ sót biến dự đoán z_j. (iii)
- Phần dư $\hat{\epsilon_j}$ không có phân bố chuẩn.(iv)

Đồ thị phần dư $\hat{\epsilon_j}$ và giá trị dự đoán $\hat{y_j}$ 1.Sai số $\hat{\epsilon_j}$ phụ thuộc vào biến $\hat{y_j}$.



Hình 1.4: Tính toán sai hoặc thiếu hệ số tự do β_0

Đồ thị phần dư $\hat{\epsilon_j}$ và một biến dự đoán $\hat{z_j}$ 2.Phương sai không phải là hằng số.



Hình 1.5: Phương sai không phải là hằng số.

1.4.3 Kiểm định tính không tương quan của ε_j theo thời gian

Giả sử y_j được theo dõi theo thời gian $j = 1, 2, ...$ Trường hợp này thường xảy ra khi khảo sát các đại lượng kinh tế. Khi đó rất thường xảy ra trường hợp các ε_j có tương quan với nhau. (Auto correlation)

Để phát hiện tính tự tương quan của các sai số ε ta sử dụng tiêu chuẩn Durbin-Watson như sau:
Đặt:

$$r_1 = \frac{\Sigma_{j=2}^{n} \widehat{\varepsilon_{j-1}} \widehat{\varepsilon_j}}{\Sigma_{j=1}^{n} \widehat{\varepsilon_j}^2} \tag{4.36}$$

Khi đó đại lượng:

$$DW = \sum_{j=2}^{n} (\widehat{\varepsilon_j} - \widehat{\varepsilon_{j-1}})^2 / \sum_{j=2}^{n} \widehat{\varepsilon_j}^2 = 2(1 - r_1) \tag{4.37}$$

sẽ tuân theo phân phối Durbin-Watson Tra bảng Durbin- Watson ứng với mức ý nghĩa α ta tìm được hai số $d_1(k, n, \alpha) < d_2(k, n, \alpha)$, khi đó so sánh DW với d_1, d_2 ta rút ra các kết luận sau:

- Nếu $0 \leq DW < d_1$ thì các ε_j có tự tương quan dương
- Nếu $d_1 \leq DW \leq d_2$ thì không thể nói gì được
- Nếu $d_2 < DW < 4 - d_2$ thì các ε_j không có tự tương quan
- Nếu $4 - d_2 \leq DW \leq 4 - d_1$ thì không thể kết luận được
- Nếu $4 - d_1 < DW \leq 4$ thì các ε_j có tự tương quan âm.

Ta xét lại ví dụ 1:

Các công ty muốn mua máy tính trước hết phải đánh giá các nhu cầu trong tương lai. Dưới đây là dữ liệu thu thập được từ 7 công ty để đưa ra hàm dự đoán về nhu cầu phần cứng:

STT	x_0	x_1(orders)	x_2(add-delete items)	y(CPU time)
1	1	125.3	2.108	141.5
2	1	146.1	9.213	168.9
3	1	133.9	1.905	154.8
4	1	128.5	0.815	146.5
5	1	151.5	1.061	172.8
6	1	136.2	8.603	160.1
7	1	92	1.125	108.5

Ta đã tính được: Tổng bình phương các phần dư $\sum_{1}^{n} \widehat{\varepsilon_j}^2 = 13.06195$

$\Rightarrow r_1 = \frac{\Sigma_{j=2}^{n} \widehat{\varepsilon_{j-1}} \widehat{\varepsilon_j}}{\Sigma_{j=1}^{n} \widehat{\varepsilon_j}^2} = \frac{4.97187}{13.06195} = 0.38064$

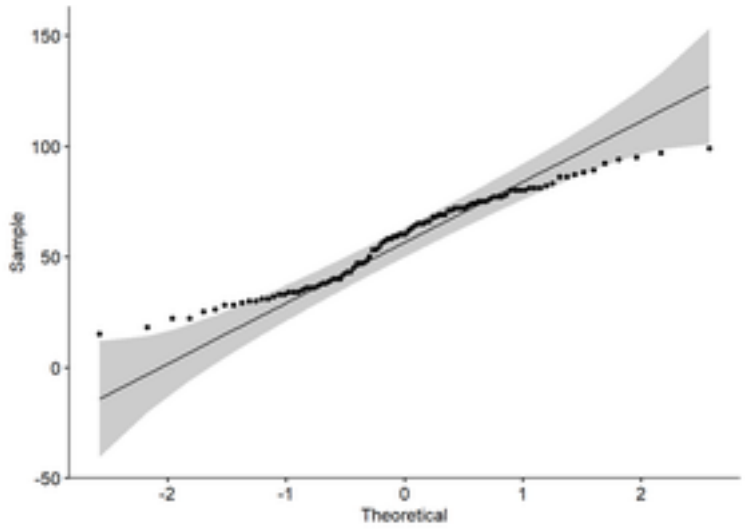
$\Rightarrow DW = 2(1 - r_1) = 1.23872$

Với $\alpha = 0,05; n = 7; k = 2$, tra bảng phân phối Durbin-Watson ta tìm được:

$$d_1 = 0,467; d_2 = 1,896$$

Vậy $d_1 = 0.467 < DW = 1.23872 < d_2 = 1.896$ nên ta không thể kết luận gì được.

1.4.4 Q - Q plot



1.5 Một số nội dung bổ sung

1.5.1 Standardized

$$\boxed{z = \frac{x - \mu}{\sigma}} \tag{5.38}$$

với x là giá trị quan sát, μ là giá trị trung bình mẫu và σ là độ lệch chuẩn mẫu.

- Công thức (5.38) đo lường xem x cách xa μ bao nhiêu độ lệch chuẩn.
- Nếu áp dụng công thức (5.38) với toàn bộ tập X ta sẽ được một mẫu Z mới có trung bình là μ, độ lệch chuẩn là 1.

1.5.2 $t - statistic$ và $p - value$

$$\boxed{t - statistic = \frac{\widehat{\beta_i}}{\sqrt{\widehat{D}(\widehat{\beta_i})}}}$$

với $\widehat{\beta_i}$ là ước lượng hệ số hồi quy cho biến X_i.

- Nếu có mối liên hệ giữa X_i và Y, ta kì vọng $t - statistic$ sẽ có phân phối t với $n - 2$ bậc tự do.
- Gọi $p - value$ là xác suất quan sát được một con số có giá trị tuyệt đối bằng $t - statistic$ hoặc lớn hơn.
- Công thức tính $p - value = 2 * (1 - cdf(n, |t - statistic|))$ với cdf là hàm phân phối tích lũy của phân phối *student*
- Ta bác bỏ giả thuyết $\beta_i = 0$ khi $p - value < 5\%$ hoặc $p - value < 1\%$.
- Khi $n \geq 30$, chúng lần lượt tương ứng với $t - statistic \geq 2$ hoặc $t - statistic \geq 2,75$.

1.5.3 Xác định các biến quan trọng - Features Selection

Có 3 cách tiếp cận kinh điển <Stepwise>

- Chọn tới dần: Chúng ta bắt đầu bằng một mô hình trống - mô hình chứa hệ số chặn mà không chứa biến dự báo. Sau đó chúng ta khớp trùng p hồi quy tuyến tính đơn biến và sau đó thêm vào mô hình trống biến sinh hệ số xác định R lớn nhất. Sau đó lại thêm vào mô hình biến có giá trị RSS nhỏ nhất, thành mô hình hai biến. Tiếp tục thực hiện cách làm này cho đến khi thỏa mãn một quy tắc dừng nào đó.
- Chọn lùi dần: Chúng ta bắt đầu bằng tất cả các biến trong mô hình, và loại bỏ biến nào cho $p-value$ lớn nhất - nghĩa là biến ít ý nghĩa thống kê nhất. Mô hình mới $(p-1)$ biến làm khớp trùng, và biến có $p-value$ lại bị loại bỏ. Chẳng hạn chúng ta có thể cho dừng lại khi tất cả các biến còn lại đều có $p-value$ nhỏ hơn một ngưỡng nào đó.
- Chọn hỗn hợp: Đây là phương pháp kết hợp giữa chọn tới dần và chọn lui dần.

1.5.4 Các bước tiến hành trong phân tích hồi quy

Bước 1: Tiền xử lí dữ liệu:

- Xử lí biến categorial
- Khảo sát tính đơn cộng tuyến tính của $X_1, ..., X_k$
- Sử dụng tiêu chuẩn F, kiểm tra mối liên hệ giữa X và Y

Bước 2: Features Selection - Sử dụng phương pháp Stepwise để xác định các biến quan trọng

- Ước lượng các hệ số hồi quy: $\widehat{\beta} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\mathbb{Y}$
- Xác định hệ số R, các hệ số $t-statistic$, $p-value$ của từng biến X_i
- Lựa chọn các biến quan trọng và loại bỏ những biến không cần thiết
- Lặp lại bước 2 đến khi thỏa mãn điều kiện dừng do người phân tích đặt ra

Bước 3:

- Sử dụng tiêu chuẩn student, kiểm tra các sai số ε
- Khảo sát đồ thị phần dư, xác định các outlier
- Xác định các điểm leverage
- Loại bỏ các outlier và các điểm leverage

Bước 4: Xây dựng mô hình cuối cùng

- Ước lượng hệ số hồi quy và khoảng tin cậy
- Xác định hệ số R
- Ước lượng hàm hồi quy tuyến tính

Chương 2

Mô hình hồi quy tuyến tính bội

2.1 Mô hình bài toán

2.1.1 Mô hình bài toán

Trong phần này, chúng ta sẽ nghiên cứu vấn đề về mô hình hóa quan hệ giữa m biến phụ thuộc $Y_1, Y_2, \cdots, Y_m$ và tập các biến độc lập $x_1, x_2, \cdots, x_k$. Giả sử mỗi biến phụ thuộc tuân theo một mô hình hồi quy, theo đó:

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}x_1 + \cdots + \beta_{k1}x_k + \varepsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}x_1 + \cdots + \beta_{k2}x_k + \varepsilon_2 \\ &\vdots \quad\quad \vdots \\ Y_m &= \beta_{0m} + \beta_{1m}x_1 + \cdots + \beta_{km}x_k + \varepsilon_m \end{aligned}$$

Sai số $\boldsymbol{\varepsilon}^T = [\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n]$ có $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ và $Var(\boldsymbol{\varepsilon}) = \boldsymbol{\Sigma}$.

Để xây dựng khái niệm mô hình mới sao cho phù hợp với mô hình hồi quy tuyến tính cổ điển, ta đặt $[x_{j0}, x_{j1}, \ldots, x_{jk}]$ là là giá trị các biến độc lập ở lần quan sát thứ j, đặt $\mathbf{Y}_j^T = [Y_{j1}, Y_{j2}, \ldots, Y_{jm}]$ là các biến phụ thuộc, và đặt $\boldsymbol{\varepsilon}_j^T = [\varepsilon_{j1}, \boldsymbol{\varepsilon}_{j2}, \ldots, \boldsymbol{\varepsilon}_{jm}]$ là sai số. Về kí hiệu, ma trận:

$$\underset{(n\times(k+1))}{\mathbf{X}} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{bmatrix}$$

cũng tương tự như ma trận X trong mô hình hồi quy tuyến tính đơn biến cổ điển. Các ma trận còn lại:

$$\underset{(n\times m)}{\mathbf{Y}} = \begin{bmatrix} Y_{11} & Y_{12} & \cdots & Y_{1m} \\ Y_{21} & Y_{22} & \cdots & Y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{n1} & Y_{n2} & \cdots & Y_{nm} \end{bmatrix} = \begin{bmatrix} \mathbf{Y}_{(1)} \vdots \mathbf{Y}_{(2)} \vdots \cdots \vdots \mathbf{Y}_{(m)} \end{bmatrix}$$

$$\underset{(k+1)\times m)}{\boldsymbol{\beta}} = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{k1} & \beta_{k2} & \cdots & \beta_{km} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{(1)} \vdots \boldsymbol{\beta}_{(2)} \vdots \cdots \vdots \boldsymbol{\beta}_{(m)} \end{bmatrix}$$

$$\underset{(n\times m)}{\boldsymbol{\varepsilon}} = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = \left[\begin{array}{c:c:c:c} \boldsymbol{\varepsilon}_{(1)} & \boldsymbol{\varepsilon}_{(2)} & \cdots & \boldsymbol{\varepsilon}_{(m)} \end{array} \right]$$

Từ đó, mô hình hồi quy tuyến tính bội được định nghĩa là:

$$\underset{(n\times m)}{\mathbf{Y}} = \underset{(n\times(k+1))}{\mathbf{X}} \underset{((k+1)\times m)}{\boldsymbol{\beta}} + \underset{(n\times m)}{\boldsymbol{\varepsilon}}$$

với

$$E(\boldsymbol{\varepsilon}_{(j)}) = 0 \text{ và } \mathrm{Cov}(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(j)}) = \sigma_{ij}\mathbf{I},\ i, j = 1, 2, \cdots, m$$

2.1.2 Tổng kết mô hình

Như vậy, ta có một số giả thiết ban đầu:

1. $E(\boldsymbol{X}) = \boldsymbol{X}\boldsymbol{\beta}$ hay $E(\boldsymbol{\Sigma}) = \mathbf{0}$.
2. $cov(\boldsymbol{Y}_i) = \boldsymbol{\Sigma}$ với mọi $i = 1, 2, \ldots, n$, trong đó $\widehat{Y_i}$ là hàng thứ i của $\boldsymbol{Y}$.

$$cov(\boldsymbol{Y}_i) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{m1} & \sigma_{m2} & \cdots & \sigma_{mm} \end{pmatrix}$$

3. $cov(\boldsymbol{Y}_i, \boldsymbol{Y}_j) = \mathbf{0}$ với mọi $i \neq j$.

$$\begin{pmatrix} cov(y_{i1}, y_{j1}) & cov(y_{i1}, y_{j2}) & \cdots & cov(y_{i1}, y_{jm}) \\ cov(y_{i2}, y_{j1}) & cov(y_{i2}, y_{j2}) & \cdots & cov(y_{i2}, y_{jm}) \\ \vdots & \vdots & \ddots & \vdots \\ cov(y_{im}, y_{j1}) & cov(y_{im}, y_{j2}) & \cdots & cov(y_{im}, y_{jm}) \end{pmatrix} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}$$

2.1.3 Tại sao hồi quy đồng thời?

Lợi ích của việc thực hiện đồng thời các phép hồi quy thay vì thực hiện tuần tự từng phép hồi quy một là gì? Câu trả lời là để sử dụng được lợi thế của việc có nhiều biến. Ví dụ, có 1 bài toán phân tích phương sai, so sánh các loại thuốc dựa trên một số các biến chỉ số sức khỏe. Có thể xảy ra khả năng là không có các biến riêng lẻ nào có thể thể hiện sự khác nhau giữa các loại thuốc. Nhưng khi kết hợp nhiều biến lại, ta sẽ thấy rõ những sự khác biệt. Sử dụng mô hình toàn thể cũng giúp xử lý khi có nhiều phép so sánh...

Phần III

Ứng dụng vào bài toán thực tế

Dưới đây là những class và function mà bọn em code để xử lý bài toán.

```
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from scipy.stats import f,t
import math
from statsmodels.stats.stattools import durbin_watson as dw
import warnings
warnings.filterwarnings('ignore')

class Linear_Model():
  def __init__(self, X, y):
    def coefficients(X, y):
      XTX_inv = np.linalg.inv(X.T @ X)
      XTy     = X.T @ y
      beta    = XTX_inv @ XTy
      return beta
    self.beta= coefficients(X, y)
    def summary(X, y):
      y_pred = self.predict(X)
      MSE = np.sum(np.square(y_pred - y))

      variance = MSE * (np.linalg.inv(X.T @ X).diagonal())  / (X.shape[0] -X
    .shape[1])

      standard_error = np.sqrt(variance)

      t_statistic = self.beta / standard_error

      p_values = 2*(1 - t.cdf(X.shape[0], np.abs(t_statistic)))

      results = pd.DataFrame({'feature': X.columns,
                              'coefficients': self.beta,
                              'standard_error': standard_error,
                              't-statistic': t_statistic,
                              'P>|t|': p_values})
      t_alpha = t.ppf(1-(0.05/(2*(X.shape[1]))),(X.shape[0]-X.shape[1]))
      results['[0.025'] = results['coefficients'] - results['standard_error'
    ]*np.sqrt(t_alpha*(X.shape[1]))
      results['0.975]'] = results['coefficients'] + results['standard_error'
    ]*np.sqrt(t_alpha*(X.shape[1]))
      results.set_index('feature')
      RSS = np.sum(np.square(y_pred - y))
      TSS = np.sum(np.square(y - np.mean(y)))
      r_squared = 1 -RSS/TSS
      f_stat = (TSS-RSS)/X.shape[1]*(X.shape[0] - X.shape[1] - 1)/RSS
      return results, r_squared, f_stat
    self.summary, self.r_squared, self.f_stat = summary(X,y)
    self.eps_hat = self.predict(X) - y
  def predict(self, X):
    return X @ self.beta
  def f_test(self):
    return bool(self.f_stat > f.ppf(q=1-.01, dfn=X.shape[1], dfd=X.shape[0]-
   X.shape[1]))
  def t_test(self):
    I_H = np.array(X @ np.linalg.inv(X.T @ X)) @ X.T
```

```
    for i in range(len(I_H)):
      I_H[i][i] -= 1
    I_H = -1*I_H
    Lambda, P = np.linalg.eig(I_H)
    for i in range(len(Lambda)):
      Lambda[i] = round(Lambda[i].real)
      for j in range(len(P[0])):
        P[i][j] = P[i][j].real
    Lambda = np.array(Lambda, dtype='int64')
    P = np.array(P, dtype='float64')
    i = 0
    j = len(Lambda)-1
    while i <= j:
      while Lambda[i] != 0:
        i += 1
      while Lambda[j] != 1:
        j -= 1
      if i <= j:
        Lambda[i], Lambda[j] = Lambda[j], Lambda[i]
        P[i], P[j] = P[j], P[i]
        i += 1
        j -= 1
    eps_star = P.T @ self.eps_hat
    eps_star_hat = np.sum(eps_star/(len(Lambda)-len(self.beta)-1))
    t_const = np.sqrt((len(Lambda)-len(self.beta)-1))*eps_star_hat
    t_const = t_const/np.sqrt((np.sum((eps_star-eps_star_hat)**2))/(len(
Lambda)-len(self.beta)-2))
    return t_const > t.ppf(1-0.025, (len(Lambda)-len(self.beta)-2))
```

Chương 3

Dữ liệu Advertising

Tập dữ liệu bao gồm doanh số ở 200 thị trường khác nhau, cùng với ngân sách quảng cáo ở mỗi thị trường đó trên 3 phương tiện truyền thông:

- **TV:** Ngân sách dùng để quảng cáo trên TV.
- **Radio:** Ngân sách dùng để quảng cáo trên Radio.
- **Newspaper:** Ngân sách dùng để quảng cáo trên Báo.
- **Sales:** Doanh số thu được.

Sales là biến Response. Chúng ta sẽ xây dựng mô hình dự đoán **Sales** dựa trên ngân sách đầu tư cho quảng cáo bằng cách trả lời các câu hỏi:

1. Có sự cộng tuyến giữa TV, Radio, Newspaper không?
2. Có mối liên hệ nào giữa doanh số và ngân sách quảng cáo hay không?
3. Mối liên hệ này chặt chẽ cỡ nào?
4. Kênh quảng cáo nào đóng góp vào doanh số?
5. Liệu có ảnh hưởng cộng năng giữa các kênh quảng cáo hay không?

```
data = pd.read_csv("/content/Advertising.csv")
data.drop('Unnamed: 0', axis=1, inplace=True)
data.head()

```

	TV	Radio	Newspaper	Sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

3.1 Có sự cộng tuyến giữa TV, Radio, Newspaper không?

- Xây dựng mô hình hồi quy tuyến tính đa biến sử dụng cả 3 biến **TV**, **Radio**, **Newspaper**, để dự đoán **Sales**.
- Sử dụng tiêu chuẩn F để kiểm tra mối liên hệ giữa **TV**, **Radio**, **Newspaper** và **Sales**.

```
data.corr()
```

	TV	Radio	Newspaper	Sales
TV	1.000000	0.054809	0.056648	0.782224
Radio	0.054809	1.000000	0.354104	0.576223
Newspaper	0.056648	0.354104	1.000000	0.228299
Sales	0.782224	0.576223	0.228299	1.000000

```
X = data[['TV', 'Radio', 'Newspaper']]
X = sm.add_constant(X)
y = data['Sales']
model = Linear_Model(X, y)
model.summary
```

	feature	coefficients	standard_error	t-statistic	P>|t|	[0.025	0.975]
0	const	2.938889	0.311908	9.422288	0.000000	1.948421	3.929358
1	TV	0.045765	0.001395	32.808624	0.000000	0.041335	0.050194
2	Radio	0.188530	0.008611	21.893496	0.000000	0.161185	0.215875
3	Newspaper	-0.001037	0.005871	-0.176715	0.301033	-0.019681	0.017606

```
print("Chi so F_statistic: ",model.f_stat)
```

```
Chỉ số F_statistic:  425.52086943950275
```

```
print("Kiem dinh F:", model.f_test()) #True tuc la beta khac 0
```

```
Kiểm định F: True
```

Như vậy, tồn tại mối liên hệ giữa doanh số và ngân sách quảng cáo.

3.2 Mối liên hệ này chặt chẽ cỡ nào?

- Sử dụng hệ số xác định R để kiểm tra

```
model.r_squared
```

```
0.8972106381789522
```

Hệ số xác định R = 0.897 là con số khá cao, chứng tỏ mối liên hệ này khá chặt chẽ.

3.3 Kênh quảng cáo nào đóng góp vào doanh số?

- Sử dụng phương pháp **stepwise** để lựa chọn những kênh quảng cáo đóng góp vào doanh số.

```
X1 = data[['TV', 'Radio', 'Newspaper']]
X1 = sm.add_constant(X1)
y = data['Sales']
model1 = Linear_Model(X1, y)

```

	feature	coefficients	standard_error	t-statistic	P>|t|	[0.025	0.975]
0	const	2.938889	0.311908	9.422288	0.000000	1.948421	3.929358
1	TV	0.045765	0.001395	32.808624	0.000000	0.041335	0.050194
2	Radio	0.188530	0.008611	21.893496	0.000000	0.161185	0.215875
3	Newspaper	-0.001037	0.005871	-0.176715	0.301033	-0.019681	0.017606

```
model1.r_squared
```

```
0.8972106381789522
```

Vì p-value của Newspaper rất lớn nên ta sẽ loại Newspaper ra khỏi mô hình, vì vậy chỉ có **TV** và **Radio** có đóng góp vào doanh số.

3.4 Liệu có ảnh hưởng cộng năng giữa các kênh quảng cáo hay không?

- Như ta đã thấy chỉ có biến **TV** và **Radio** có mối liên hệ giữa với Sales.
- Ta sẽ xây dựng biến thứ 3 bằng cách lấy **TVxRadio** để thể hiện sự công năng của 2 kênh này.

```
X = data[['TV', 'Radio']]
X['TVxRadio'] = X['TV']*X['Radio']
X = sm.add_constant(X)
y = data['Sales']
model = Linear_Model(X,y)
model.summary
```

	feature	coefficients	standard_error	t-statistic	P>|t|	[0.025	0.975]
0	const	6.750220	0.247871	27.232755	0.000000e+00	5.963102	7.537339
1	TV	0.019101	0.001504	12.698953	0.000000e+00	0.014325	0.023878
2	Radio	0.028860	0.008905	3.240815	9.633934e-08	0.000582	0.057139
3	TVxRadio	0.001086	0.000052	20.726564	0.000000e+00	0.000920	0.001253

```
model.r_squared
```

```
0.9677905498482523
```

Ta có thể thấy, độ chính xác của mô hình đã tăng lên đáng kể, từ 0.897 lên 0.968. Vì vậy, có sự cộng năng giữa **TV** và **Radio**, tức là 2 kênh quảng cáo này bổ trợ cho nhau giúp truyền thông tốt hơn.

Chương 4

Kiểm tra sự phù hợp của mô hình

1. Tiêu chuẩn **Student**.
2. Khảo sát đồ thị phần dư + xác định các**Outlier**.
3. Kiểm định tính không tương quan của phần dư theo thời gian.
4. Xác định các điểm **Leverage**

4.1 Tiêu chuẩn Student

```
model.t_test() #False tuc la khong the bac bo phan du tuan theo phan phoi
    chuan
```

```
False
```

Vậy ta không thể bác bỏ phần dư tuân theo phân phối chuẩn.

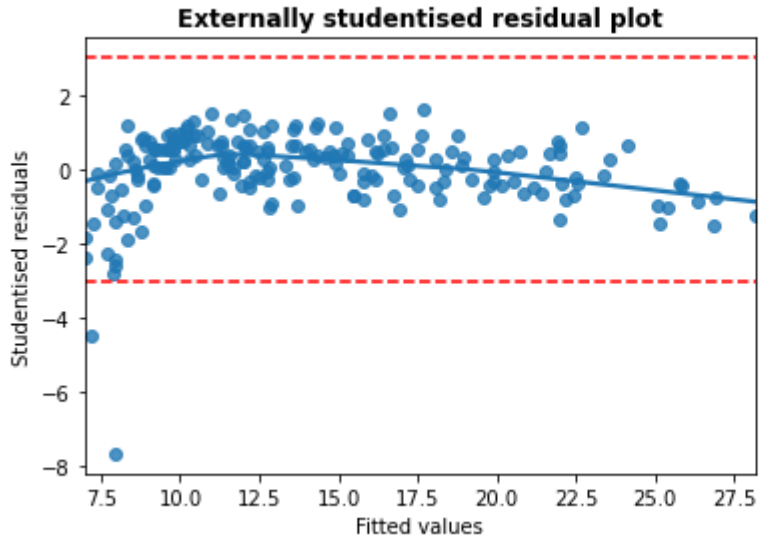
4.2 Khảo sát đồ thị phần dư + xác định các Outlier

- Ta sẽ sử dụng công thức Standardlized để chuẩn hóa phần dư của tất cả các quan sát thành thành mẫu có trung bình là 0 và độ lệch chuẩn là 1. Sau đó loại bỏ những quan sát có trị tuyệt đối phần dư được chuẩn hóa > 3.

```
y_pred = model.predict(X)
residuals = np.array(y - y_pred)
# Estimate variance (externalised)
 i_est   = []
for i in range(X.shape[0]):
    # exclude ith observation from estimation of variance
    external_residuals = np.delete(residuals, i)
      i_est   += [np.sqrt((1 / (X.shape[0] - X.shape[1])) * np.sum(np.square(
    external_residuals)))]
 i_est   = np.array( i_est )
t_stat = residuals /   i_est

```

```
       ax = sns.regplot(x=y_pred, y=t_stat, lowess=True)
plt.xlabel('Fitted values')
plt.ylabel('Studentised residuals')
plt.title('Externally studentised residual plot', fontweight='bold')
ax.axhline(y=3, color='r', linestyle='dashed')
ax.axhline(y=-3, color='r', linestyle='dashed')
```

```
<matplotlib.lines.Line2D at 0x7fe4936a3f10>
```



Ta có thể thấy dữ liệu của chúng ta có 2 điểm **Outlier**. Chúng ta sẽ lưu 2 điểm này vào drop-list để sau này *xóa* 2 điểm này khỏi dữ liệu.

```
drop_list = []
for i in range(len(t_stat)-1):
  if abs(t_stat[i]) > 3:
    drop_list.append(i)
```

4.3 Kiểm định tính không tương quan của phần dư theo thời gian

```
DW = dw(residuals)
print(DW)
```

```
2.2236287420696197
```

- Ta có d1(3, 200, 0.05)= 1.738, d2(3, 200, 0.05)= 1.799
- Vì DW = 2.224 nên DW thuộc khoảng (4-d2, 4-d1) = (2.201, 2.262) nên không thể kết luận được

4.4 Xác định các điểm Leverage

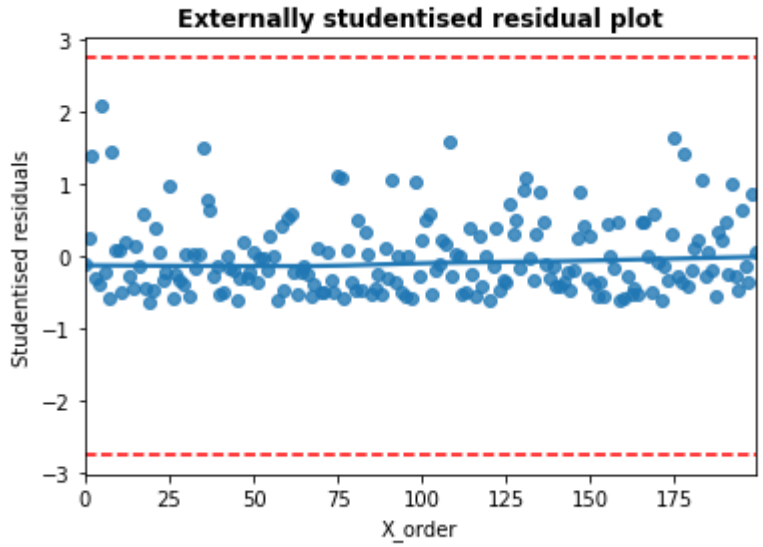
- Vì trung bình leverage của tất cả các quan sát là (p+1)/n với p là số biến dự đoán, n là số lượng quan sát. Ta sẽ sử dụng công thức Standardlized để chuẩn hóa Leverage của tất cả các quan sát thành thành mẫu có trung bình là 0 và độ lệch chuẩn là 1. Sau đó loại bỏ những quan sát có trị tuyệt đối Leverage được chuẩn hóa > 2.75

```
    #leverage
H = np.array(np.array(X @ np.linalg.inv(X.T @ X)) @ X.T)
h_ii = H.diagonal()
l i_est = []
for i in range(X.shape[0]):
    # exclude ith observation from estimation of variance
```

```
    external_leverage = np.delete(h_ii, i)
    l_i_est += [np.sqrt((1 / (X.shape[0] - X.shape[1])) * np.sum(np.square(
    external_leverage)))]
l_i_est = np.array(l_i_est)

# Externally studentised leverage
student_leverage = (h_ii-X.shape[1]/X.shape[0]) / l_i_est * np.sqrt(1 -
    h_ii)

```

```
      ax = sns.regplot(x=np.arange(200), y=student_leverage, lowess=True)
plt.xlabel('X_order')
plt.ylabel('Studentised residuals')
plt.title('Externally studentised residual plot', fontweight='bold')
ax.axhline(y=2.75, color='r', linestyle='dashed')
ax.axhline(y=-2.75, color='r', linestyle='dashed')

```

```
<matplotlib.lines.Line2D at 0x7fe4936a3cd0>
```



- Ta có thể thấy dữ liệu không có điểm nào là điểm Leverage. Ta sẽ xóa 2 outlier nằm trong drop-list.

```
X.drop(drop_list, inplace=True)
y.drop(drop_list, inplace=True)
```

```
(198, 4)
```

Chương 5

Xây dựng mô hình cuối cùng

1. Ước lượng hệ số hồi quy và khoảng tin cậy của chúng
2. Xác định hệ số R
3. Ước lượng hàm HQTT

5.1 Ước lượng hệ số hồi quy và khoảng tin cậy của chúng

```
model = Linear_Model(X,y)
model.summary
```

	feature	coefficients	standard_error	t-statistic	P>|t|	[0.025	0.975]
0	const	6.840642	0.206104	33.190318	0.000000e+00	6.186126	7.495157
1	TV	0.018597	0.001246	14.928898	0.000000e+00	0.014641	0.022552
2	Radio	0.034022	0.007422	4.583877	3.432350e-10	0.010452	0.057591
3	TVxRadio	0.001063	0.000043	24.445089	0.000000e+00	0.000925	0.001201

5.2 Xác định hệ số R

```
model.r_squared
```

```
0.9774918671223143
```

5.3 Ước lượng hàm HQTT

```
#Uoc luong HQTT
interval = []
y_pred = model.predict(X)
XT_X_inv = np.linalg.inv(X.T @ X)
sigma_hat = np.sum(np.square(y_pred-y))/(X.shape[0]-X.shape[1])
for i in range(len(X)):
  X0 = np.array(X.values[i])
  interval.append(t.ppf(1-0.025, (X.shape[0]-X.shape[1]))* np.sqrt(sigma_hat
   *(np.array(X0 @ XT_X_inv) @ X0)))
```

```
predict_interval = pd.DataFrame()
predict_interval['predict'] = y_pred
predict_interval['[0.025'] = y_pred - interval
predict_interval['0.975]'] = y_pred + interval
predict_interval

```

	predict	[0.025	0.975]
0	21.650619	21.449677	21.851560
1	10.864092	10.611038	11.117145
2	9.561230	9.198725	9.923735
3	17.713639	17.542887	17.884392
4	12.645797	12.490294	12.801299
...	...	...	...
195	7.827140	7.534923	8.119357
196	9.249756	9.048570	9.450942
197	12.198275	12.037475	12.359075
198	26.203979	25.892829	26.515129
199	13.571103	13.346151	13.796055

198 rows × 3 columns

Tài liệu tham khảo

1. Applied Multivariate Statistical Analysis - Richard Johnson - 2007
2. Slide bài giảng thầy Lê Xuân Lý
3. An Introduction to Statistical Learning with Applications in R - Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani - 2013