



Báo cáo cuối kỳ

PHÂN TÍCH SỐ LIỆU

Chủ đề 7

HỒI QUY TUYẾN TÍNH

Giảng viên hướng dẫn: ThS. Lê Xuân Lý

Nhóm sinh viên thực hiện: Nhóm 7

Phan Thanh Tùng	20210913
Bùi Minh Châu	20210110
Quách Thái Dương	20210253
Nguyễn Minh Dương	20216917
Lê Thanh Hùng	20216935
Hoàng Kim Khánh	20216839
Nguyễn Thị Nhã Linh	20216940
Đào Mai Sơn	20216958
Khổng Nguyên Thiêm	20210814
Doãn Chí Thường	20210831
Lê Minh Tiến	20216893

Mục lục

Chương 1 Tổng quan về mô hình hồi quy tuyến tính	5
1.1 Giới thiệu	5
1.2 Sự cần thiết của mô hình	5
1.3 Dữ liệu cho phân tích	6
1.3.1 Dữ liệu chéo	6
1.3.2 Dữ liệu chuỗi thời gian	6
1.3.3 Dữ liệu hỗn hợp	7
1.4 Tính tuyến tính trong mô hình hồi quy	8
Chương 2 Mô hình hồi quy tuyến tính	9
2.1 Mô hình hồi quy tuyến tính cổ điển	9
2.2 Giới thiệu một số mô hình hồi quy tuyến tính khác	10
2.2.1 Mô hình dạng log - log	11
2.2.2 Mô hình dạng bán loga	11
2.2.3 Mô hình dạng đa thức	12
Chương 3 Phương pháp bình phương tối thiểu	15
3.1 Ước lượng bình phương cực tiểu	15
3.2 Tiếp cận bằng hình học	18
3.3 Phân tách tổng bình phương - Hệ số xác định R^2	19
3.4 Tính chất ước lượng bằng phương pháp bình phương cực tiểu	20
3.5 Định lý Gauss về ước lượng bình phương cực tiểu	22
Chương 4 Ước lượng khoảng của mô hình hồi quy	25
4.1 Khoảng tin cậy của các hệ số hồi quy β_j	25
4.2 Kiểm định tỷ số hợp lý của các hệ số hồi quy	30
Chương 5 Ước lượng hàm hồi quy tuyến tính	35
5.1 Ước lượng hàm hồi quy tại Z_0	35
5.2 Dự đoán một quan sát mới tại Z_0	36
Chương 6 Kiểm tra sự phù hợp của mô hình	41
6.1 Một số khái niệm	41
6.1.1 Chuẩn hóa tập mẫu	41
6.1.2 Outlier	41
6.1.3 Độ đo Leverage	44
6.1.4 Quy tắc 1.5IQR	45
6.1.5 Giá trị t -statistic và p -value	45
6.2 Kiểm định tính phụ thuộc vào biến của mô hình	46
6.3 Kiểm tra tính đa cộng tuyến của các biến dự đoán	47
6.4 Khảo sát phần dư	49
6.4.1 Khảo sát đồ thị của các phần dư	52
6.4.2 Khảo sát đồ thị phần dư Q-Q	53
6.5 Kiểm định tính không tương quan của phần dư theo thời gian	54
6.5.1 Kiểm tra bằng đồ thị	54
6.5.2 Kiểm tra bằng phương pháp Durbin - Watson	55
6.6 Xác định các biến quan trọng	56
Chương 7 Phần code chương trình thực hành với dữ liệu cụ thể	59
7.1 Mô tả dữ liệu	59
7.2 Các bước xây dựng một mô hình hồi quy hoàn chỉnh	60
7.2.1 Tiền xử lý dữ liệu	60

7.2.2	Chuẩn hóa tập mẫu	62
7.2.3	Lọc bỏ outlier bằng leverage	62
7.2.4	Khảo sát tính đa cộng tuyến	64
7.3	Xác định tiêu chuẩn và các biến quan trọng	65
7.3.1	Ước lượng tham số hồi quy	65
7.3.2	Xác định các giá trị t-statistic , p-value	66
7.3.3	Đánh giá mức độ quan trọng của biến đối với mô hình	68
7.3.4	Khảo sát phần dư	68
7.4	Kiểm tra và đánh giá mô hình	69
7.4.1	Thực hiện ước lượng lại tham số và R^2 của mô hình	70
7.4.2	Ước lượng khoảng tham số	70
7.4.3	Test mô hình và đánh giá	71
7.4.4	So sánh 2 mô hình hồi quy	71
7.5	Một số nhận xét cho mô hình dữ liệu thực tế	72
Chương 8	Mô hình hồi quy tuyến tính bội	73
8.1	Mô hình bài toán	73
8.2	Ước lượng tham số	74
8.3	Các tính chất quan trọng	78
8.4	Kiểm định tỉ số hợp lý cho tham số hồi quy	80
8.5	Các thống kê nhiều chiều khác	83
8.6	Đưa ra dự đoán từ mô hình hồi quy tuyến tính đa bội	84
Kết luận		87
Tài liệu tham khảo		87

FaMI
1956

Lời mở đầu

Phân tích dữ liệu là một học phần quan trọng trong ngành công nghệ thông tin. Chúng ta đang bước vào kỷ nguyên số với các nguồn dữ liệu lớn (Big Data), việc cấp thiết là cần phải xử lý và phân tích dữ liệu. Ngành phân tích dữ liệu sẽ là một ngành triển vọng trong tương lai, các công ty doanh nghiệp từ nhỏ đến lớn đều rất cần các nhóm phụ trách các công việc liên quan đến phân tích, xử lý, đánh giá dữ liệu, từ đó đưa ra các quyết định, hướng phát triển cho công ty. Sinh viên theo đuổi ngành phân tích dữ liệu sẽ có rất nhiều các vị trí việc làm tiềm năng ví dụ như Chuyên gia phân tích trí tuệ doanh nghiệp (Business Intelligence Analyst), Phân tích dữ liệu kinh doanh (Business Data Analytics), Kỹ sư khoa học dữ liệu (Data Scientist), Kỹ sư dữ liệu (Data Engineer), Chuyên gia phân tích định lượng (Quantitative Analyst), ...

Phân tích dữ liệu sẽ tập trung vào việc thu thập, khai thác, quản lý và xử lý bộ dữ liệu, từ đó đưa ra các nhận định, dự đoán xu hướng hoạt động của tương lai. Một quá trình phân tích dữ liệu sẽ gồm nhiều giai đoạn, đầu tiên là kiểm tra dữ liệu - dữ liệu dành cho phân tích có tốt hay không, dữ liệu có được thu thập một cách khách quan hay không; tiếp theo là làm sạch dữ liệu - loại bỏ những dữ liệu trùng lặp, dữ liệu nhiễu cũng như nội suy ra các dữ liệu bị mất mát; sau đó là phân tích khám phá dữ liệu - trực quan hóa các dữ liệu để có được thông tin chi tiết, liên quan đến dữ liệu; và cuối cùng là sử dụng các mô hình thuật toán để xác định mối quan hệ giữa các biến, đánh giá đưa ra dự báo cho một số biến phản hồi đại diện cho các dữ liệu. Nhận thấy tầm quan trọng của việc phân tích dữ liệu chúng em đã quyết định lựa chọn tìm hiểu, nghiên cứu một giai đoạn quan trọng trong quá trình phân tích dữ liệu đó là giai đoạn thiết lập mô hình hóa và thuật toán cho dữ liệu. Có rất nhiều mô hình cũng như thuật toán trong phân tích dữ liệu và trong phạm vi của bài báo cáo này chúng em xin được trình bày các mô hình hồi quy tuyến tính đa biến dùng để đánh giá tác động của các biến độc lập lên các biến phụ thuộc. Nội dung của bài báo cáo gồm có 10 phần: Tổng quan về mô hình hồi quy tuyến tính, Mô hình hồi quy tuyến tính, Ước lượng bình phương cực tiểu, Ước lượng khoảng của mô hình hồi quy, Ước lượng hàm hồi quy tuyến tính, Kiểm tra sự phù hợp của mô hình, Mô hình hồi quy tuyến tính bội, Một cách tiếp cận khác của mô hình hồi quy tuyến tính, So sánh hai cách tiếp cận của mô hình hồi quy tuyến tính, Phần code chương trình thực hành với dữ liệu cụ thể.

Để có thể hoàn thành bài báo cáo này, chúng em xin được gửi lời cảm ơn chân thành và sâu sắc đến thầy **ThS. Lê Xuân Lý**, thầy đã tận tình giảng dạy và hướng dẫn chúng em trong suốt quá trình học tập và làm bài báo cáo.



FaMI

1956

Bảng đánh giá thành viên

Tên thành viên	MSSV	Đánh giá giữa kỳ	Đánh giá cuối kỳ
Phan Thanh Tùng (Nhóm trưởng)	20210913	+1.5	+ 1.5
Bùi Minh Châu	20210110	+1.5	+ 1.5
Quách Thái Dương	20210253	+1.5	+1.5
Nguyễn Minh Dương	20216917	+1.5	+1.5
Lê Thanh Hùng	20216934	+1.5	+1.5
Hoàng Kim Khánh	20216839	+1.5	+1.5
Nguyễn Thị Nhã Linh	20210526	+1.5	+1.5
Đào Mai Sơn	20216979	+1.5	+1.5
Khổng Nguyên Thiêm	20210814	+1.5	+1.5
Doãn Chí Thường	20210831	+1.5	+1.5
Lê Minh Tiến	20216893	+1.5	+1.5



FaMI

1956

Tổng quan về mô hình hồi quy tuyến tính

1.1 Giới thiệu

Phân tích hồi quy là một kỹ thuật thống kê để điều tra và mô hình hóa mối quan hệ giữa các biến. Các ứng dụng của hồi quy rất nhiều và xảy ra trong hầu hết các lĩnh vực, bao gồm kỹ thuật, khoa học, vật lý và hóa học, kinh tế, quản lý, khoa học đời sống, sinh học và khoa học xã hội. Trên thực tế, phân tích hồi quy có thể là kỹ thuật thống kê được sử dụng rộng rãi nhất. Ví dụ: dự đoán giá nhà, tiền lương, ...

Hồi quy tuyến tính có lẽ là công cụ tiêu chuẩn đơn giản và phổ biến nhất được sử dụng cho bài toán hồi quy. Xuất hiện từ đầu thế kỷ 19, hồi quy tuyến tính được phát triển từ một vài giả thuyết đơn giản. Ta có thể chia mô hình hồi quy tuyến tính làm 3 dạng

- Hồi quy tuyến tính đơn

$$Y = \beta_0 + \beta_1 Z + \varepsilon$$

- Hồi quy tuyến tính bội

$$Y = \beta_0 + \beta_1 Z_1 + \dots + \beta_n Z_n + \varepsilon$$

- Hồi quy tuyến tính đa bội

$$\begin{pmatrix} Y_{11} & Y_{12} \\ Y_{21} & Y_{22} \end{pmatrix} = \begin{pmatrix} 1 & Z_{11} & Z_{12} \\ 1 & Z_{21} & Z_{22} \end{pmatrix} \begin{pmatrix} \beta_{01} & \beta_{02} \\ \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix} + \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} \\ \varepsilon_{21} & \varepsilon_{22} \end{pmatrix}$$

Ta chủ yếu sẽ xét mô hình hồi quy tuyến tính bội và mô hình hồi quy tuyến tính đa bội trong bài báo cáo này.

1.2 Sự cần thiết của mô hình

Mô hình hồi quy tuyến tính bội cho phép đánh giá tác động riêng phần của một biến độc lập lên biến phụ thuộc khi biến độc lập khác trong mô hình không đổi.

Ngoài ra, việc đưa thêm các biến độc lập thích hợp vào mô hình đồng nghĩa với việc sử dụng thêm thông tin trong việc giải thích sự thay đổi của biến phụ thuộc, do đó cải thiện chất lượng dự báo của mô hình.

1.3 Dữ liệu cho phân tích

Trong bất kì phân tích mô hình nào thì dữ liệu cũng đều là quan trọng nhất, dữ liệu có tốt thì kết quả chạy mô hình mới tốt được. Dữ liệu "tốt" ở đây được hiểu là các dữ liệu được người khảo sát lấy khách quan (lấy ngẫu nhiên từ một tổng thể nghiên cứu và đảm bảo tính đo đạc chính xác). Dựa vào không gian và thời gian của dữ liệu thì người ta chia thành 3 loại dữ liệu: Chuỗi thời gian, dữ liệu chéo và dữ liệu hỗn hợp. Trong phần báo cáo về chương mô hình hồi quy tuyến tính bội này thì từ phần 7.2 - 7.9 dữ liệu được chúng em sử dụng để phân tích và đánh giá là dữ liệu chéo còn phần 7.10 dữ liệu được sử dụng là dữ liệu chuỗi thời gian.

1.3.1 Dữ liệu chéo

Dữ liệu chéo là các dữ liệu về một hay nhiều biến được thu thập tại cùng một thời điểm (thời kỳ) nhưng ở các không gian (địa phương, đơn vị, ...) khác nhau và được lấy một cách ngẫu nhiên từ tổng thể nghiên cứu. Phân tích dữ liệu chéo thường bao gồm so sánh sự khác biệt giữa các đối tượng.

Ví dụ về dữ liệu chéo: Dữ liệu về tỷ lệ sẵn sàng mua xe trong năm tới theo các nhóm tuổi và giới tính.

Will you buy a car next year?							
	Total	Gen Z: 18-24	Millennial: 25-39	Gen X: 40-55	Boomer: 56-75	Male	Female
No	73%	80%	81%	72%	62%	71%	75%
Yes	27%	20%	19%	28%	38%	29%	25%

Hình 1.1: Ví dụ về dữ liệu chéo

1.3.2 Dữ liệu chuỗi thời gian

Dữ liệu chuỗi thời gian là chuỗi các số liệu được thu thập trong một thời kỳ hay một khoảng thời gian lặp lại như nhau trên cùng một không gian địa điểm. Khoảng lặp lại như nhau có thể là giá (giá chứng khoán), tuần, tháng, ... Khoảng thời gian này được gọi là tần số của dữ liệu, tần số này có thể là giờ, ngày, tuần, tháng, quý, năm, ...

Ví dụ dữ liệu chuỗi thời gian: Dữ liệu về giá vàng ở Việt Nam qua các năm từ 2010 - 2020.

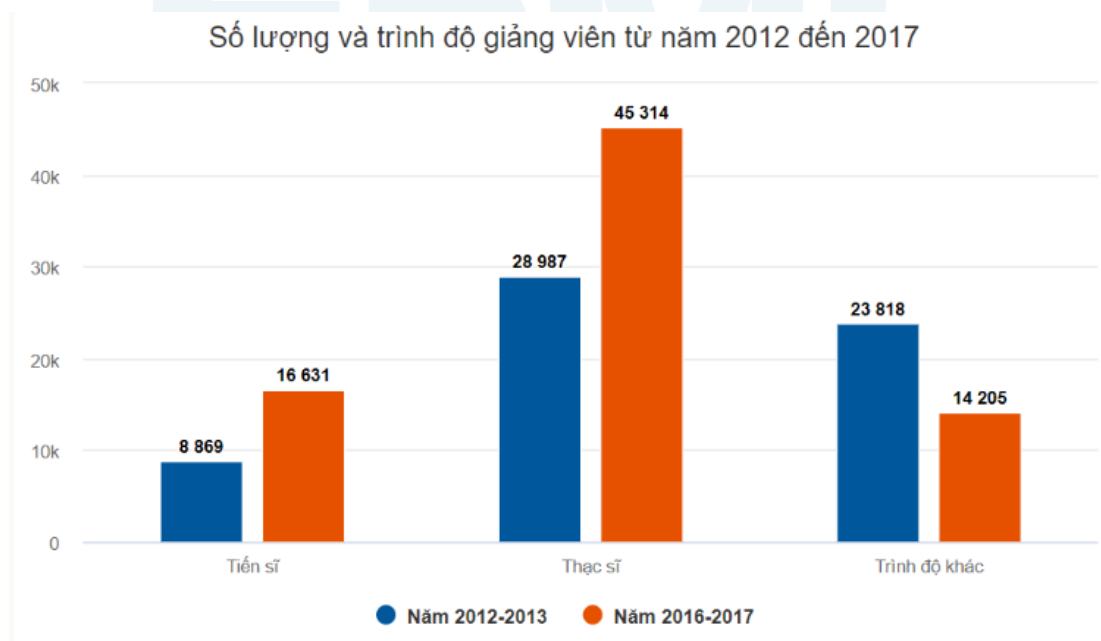


Hình 1.2: Ví dụ về dữ liệu chuỗi thời gian

1.3.3 Dữ liệu hỗn hợp

Dữ liệu hỗn hợp là các dữ liệu được thu thập theo thời gian và không gian (là sự kết hợp giữa dữ liệu chéo và dữ liệu chuỗi thời gian)

Ví dụ: Dữ liệu về số lượng và trình độ giảng viên từ năm 2012 đến 2017.



Hình 1.3: Ví dụ về dữ liệu hỗn hợp

1.4 Tính tuyến tính trong mô hình hồi quy

Tính tuyến tính của hàm hồi quy được hiểu là tuyến tính theo tham số, có nghĩa là tuyến tính ở các hệ số hồi quy và nó có thể tuyến tính hoặc phi tuyến ở các biến Z và Y .

Ví dụ: Mô hình hồi quy sau đây:

1. $Y = \beta_1 + \beta_2 Z^2 + \varepsilon$

2. $\log(Y) = \beta_1 + \beta_2 \log(Z) + \varepsilon$

đều được hiểu là các mô hình hồi quy tuyến tính, vì chúng đều tuyến tính ở các hệ số hồi quy.

FaMI
1956

Mô hình hồi quy tuyến tính

2.1 Mô hình hồi quy tuyến tính cổ điển

Mở đầu về chương hồi quy tuyến tính bội (hay còn được gọi là mô hình hồi quy tuyến tính nhiều biến), ta xét mô hình hồi quy tuyến tính cổ điển.

Giả sử Z_1, Z_2, \dots, Z_k là k biến độc lập dùng để dự báo và Y là biến phụ thuộc cần dự báo. Ví dụ Y là giá nhà hiện nay, khi đó Y phụ thuộc vào các yếu tố:

- Z_1 là diện tích ngôi nhà
- Z_2 là vị trí vùng (khoảng cách đến trung tâm thành phố)
- Z_3 là số lượng phòng ngủ
- Z_4 là số lượng phòng khách
- ...

Mô hình hồi quy tuyến tính cổ điển khẳng định rằng Y phụ thuộc tuyến tính vào các yếu tố chính Z_i theo như phương trình:

$$Y = \beta_0 + \beta_1 Z_1 + \dots + \beta_k Z_k + \varepsilon \quad (2.1)$$

Trong đó β_i với $i = 0, 1, 2, \dots, k$ là các hệ số hồi quy chưa biết, ε là sai số ngẫu nhiên.

Tiến hành n quan sát độc lập đồng thời về $k + 1$ biến Z_1, \dots, Z_k, Y . Mô hình hoàn chỉnh trở thành:

$$\begin{aligned} y_1 &= \beta_0 + \beta_1 z_{11} + \dots + \beta_k z_{1k} + \varepsilon_1 \\ y_2 &= \beta_0 + \beta_1 z_{21} + \dots + \beta_k z_{2k} + \varepsilon_2 \\ &\vdots \\ y_n &= \beta_0 + \beta_1 z_{n1} + \dots + \beta_k z_{nk} + \varepsilon_n \end{aligned} \quad (2.2)$$

Trong đó:

- β_i là các hệ số hồi quy
- Y là biến phụ thuộc
- Z_i là các biến giải thích (các yếu tố chính)

- ε là phân nhiễu - sai số (đại diện cho tất cả các biến không được đưa vào mô hình do các lý do như không có sẵn dữ liệu, khó đo lường và các biến này ảnh hưởng lên Y là không đáng kể)

Các sai số $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$ thỏa mãn 3 điều kiện:

1. $E(\varepsilon_j) = 0$
2. $\text{Var}(\varepsilon_j) = \sigma^2$ (hằng số không đổi)
3. $\text{Cov}(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$ (các sai số ε là không tương quan với nhau)

Mô hình trên có thể được viết dưới dạng ma trận như sau:

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

Hoặc viết dưới dạng tổng quát:

$$\underbrace{Y}_{n \times 1} = \underbrace{Z}_{n \times (k+1)} \cdot \underbrace{\beta}_{(k+1) \times 1} + \underbrace{\varepsilon}_{n \times 1} \quad (2.3)$$

Với ma trận Z được gọi là ma trận thiết kế:

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

Và

$$Y = [y_1, y_2, \dots, y_n]^T, \beta = [\beta_0, \beta_1, \dots, \beta_k]^T, \varepsilon = [\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n]^T$$

$$E(\varepsilon) = \underbrace{0}_{n \times 1} \text{ và } \text{Cov}(\varepsilon) = \underbrace{\sigma^2 I}_{n \times n}$$

Ở đây: ma trận β và σ^2 là chưa biết và nhiệm vụ là cần đi tìm β và σ^2 .

2.2 Giới thiệu một số mô hình hồi quy tuyến tính khác

Một số dạng biến thể của mô hình hồi quy trên, phù hợp với các tình huống và các yêu cầu khác nhau của mô hình kinh tế cũng như với bản chất của số liệu.

2.2.1 Mô hình dạng log - log

a) Khái niệm

Trong kinh tế thường gặp nhiều hàm số phi tuyến, chẳng hạn như hàm sản xuất có dạng Cobb-Douglas như nhau:

$$Q = e^{\beta_0} K^{\beta_1} L^{\beta_2}$$

Trong đó Q, K và L lần lượt là sản lượng, vốn và lao động. Khi đưa thêm yếu tố nhiễu ngẫu nhiên vào, ta có:

$$Q = e^{\beta_0} K^{\beta_1} L^{\beta_2} e^{\varepsilon}$$

Hàm này phi tuyến theo cả tham số và biến số K, L. Ta sẽ đưa hàm trên về dạng hàm tuyến tính theo hệ số hồi quy như sau:

$$\ln(Q) = \beta_0 + \beta_1 \ln(K) + \beta_2 \ln(L) + \varepsilon$$

Một cách tổng quát, mô hình hồi quy dạng log-log có thể được viết như sau:

$$\ln(Y) = \beta_0 + \beta_1 \ln(Z_1) + \dots + \beta_k \ln(Z_k) + \varepsilon \quad (2.4)$$

b) Tính chất

Các hệ số β_i còn được gọi là các hệ số co giãn riêng phần của từng Z_i , nghĩa là nếu Z_i gia tăng (giảm) 1% và các yếu tố khác không đổi thì trung bình Y tăng (giảm) là β_i %.

c) Ví dụ

Dạng hàm chi tiêu được thiết lập như sau:

$$\ln(CT) = -0.1 + 0.92\ln(TN) + 0.06\ln(TS) + \varepsilon$$

Trong đó: CT, TN, TS lần lượt là chi tiêu, thu nhập và tài sản. Khi thu nhập (TN) tăng 1% và tài sản không đổi thì trung bình chi tiêu gia đình tăng 0.92%.

d) Nhận xét

Các mô hình dạng log-log được sử dụng khá rộng rãi nhất là các mô hình nghiên cứu về hàm cung cầu.

2.2.2 Mô hình dạng bán loga

a) Khái niệm

Trong một số trường hợp, mô hình log-log nói trên là không thực sự phù hợp, chẳng hạn quan hệ giữa tiền lương và số năm kinh nghiệm của người lao động, hoặc tiền lương và trình độ học vấn (thường được đo bởi số năm học ở trường). Khi đó mô hình dạng bán loga dưới đây có thể là phù hợp:

$$\ln(Y) = \beta_0 + \beta_1 Z + \varepsilon$$

Tổng quát mô hình dạng bán loga có thể được viết dưới dạng như sau:

$$\ln(Y) = \beta_0 + \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_k Z_k + \varepsilon \quad (2.5)$$

Hoặc cũng có thể gặp mô hình dạng:

$$Y = \beta_0 + \beta_1 \ln(Z_1) + \beta_2 \ln(Z_2) + \cdots + \beta_k \ln(Z_k) + \varepsilon \quad (2.6)$$

b) Tính chất

Tính chất của dạng hàm bán loga:

+ Dạng (2.5): nếu Z_i gia tăng (giảm) 1 đơn vị và các yếu tố khác không đổi thì trung bình Y tăng (giảm) là $100\beta_i\%$.

+ Dạng (2.6): nếu Z_i gia tăng (giảm) 1% và các yếu tố khác không đổi thì trung bình Y tăng (giảm) là $\frac{\beta_i}{100}$ đơn vị.

c) Ví dụ

Giả sử quan hệ giữa thu nhập (TN) và trình độ học vấn (ED - đo bằng số năm học ở trường) là như sau:

$$\ln(\text{TN}) = 2.5 + 5.6\text{ED} + \varepsilon$$

Khi đó có thể nói rằng cứ mỗi năm đi học giúp mức thu nhập trung bình tăng 5.6%.

d) Nhận xét:

- Dạng hàm logarit thường được sử dụng khi các biến số đều nhận giá trị dương như dân số, GDP, số lao động, ... hoặc các biến số mà phân phối có đuôi lệch như thu nhập, mức lương, ... Việc lấy logarit trong trường hợp này thường giúp làm cho phân phối của sai số ngẫu nhiên gần với phân phối chuẩn hơn, do nó nhận cả các giá trị âm và làm giúp gia tăng tính đối xứng của phân phối xác suất.
- Việc sử dụng mô hình dạng logarit có một ưu thế là các kết quả ước lượng không phụ thuộc vào đơn vị của các biến số.

2.2.3 Mô hình dạng đa thức

a) Khái niệm

Trong phân tích hồi quy cũng thường gặp các dạng hàm có chứa các số mũ bậc cao của biến độc lập - là các hàm dạng đa thức. Dạng hàm tổng quát

$$Q = \beta_0 + \beta_{11}Z_1 + \beta_{12}Z_1^2 + \cdots + \beta_{k1}Z_k + \beta_{k2}Z_k^2 + \cdots + \varepsilon \quad (2.7)$$

b) Ví dụ

Trong kinh tế lao động thường gặp bài toán về quan hệ tiền lương W và tuổi (Age) của người lao động, ta xét mô hình hồi quy sau:

$$W = \beta_1 + \beta_2 \text{Age} + \beta_3 \text{Age}^2 + \varepsilon$$

Mô hình hồi quy dạng đa thức thường được dùng để nghiên cứu hàm chi phí hoặc tiền lương.

Kết luận: Tất cả các mô hình nêu trên thì đều có thể đưa về dạng mô hình hồi quy tuyến tính theo cả tham số và biến số như mô hình hồi quy cổ điển bằng cách đặt biến mới. Và khi nghiên cứu ta cần lựa chọn mô hình hồi quy tuyến tính phù hợp với vấn đề và dữ liệu mình đang nghiên cứu để đạt kết quả tốt.





FaMI

1956

Phương pháp bình phương tối thiểu

3.1 Ước lượng bình phương cực tiểu

Với các mô hình hồi quy tuyến tính, phương pháp ước lượng thông dụng nhất là phương pháp bình phương cực tiểu (least square). Đây là một phương pháp tối ưu hóa để lựa chọn một đường phù hợp nhất cho một dải dữ liệu ứng với cực trị của tổng các bình phương sai số thống kê giữa đường phù hợp nhất và dữ liệu. Phương pháp này được giới thiệu bởi Gauss vào những năm cuối thế kỷ XVIII.

Bài toán đặt ra là hãy dựa trên ma trận \mathbb{Z} và vector \mathbb{Y} của các giá trị quan sát được hãy ước lượng tham số β và σ^2 .

Nếu ta sử dụng giá trị b là giá trị thử cho β thì giữa các quan sát y_j và $b_0 + b_1 z_{j1} + \dots + b_k z_{jk}$ sẽ có độ lệch (sai số) khác 0:

$$y_j - b_0 - (b_1 z_{j1} + \dots + b_k z_{jk})$$

Phương pháp bình phương tối thiểu là cách chọn giá trị vectơ b sao cho cực tiểu hóa hàm $S(b)$:

$$\begin{aligned} S(b) &= \sum_{j=1}^n (y_j - b_0 - b_1 z_{j1} - \dots - b_k z_{jk})^2 \\ &= (\mathbb{Y} - \mathbb{Z}b)^T (\mathbb{Y} - \mathbb{Z}b) \rightarrow \min \end{aligned} \quad (3.1)$$

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:

$$\hat{\varepsilon}_j = y_j - (\hat{\beta}_0 + \hat{\beta}_1 z_{j1} + \dots + \hat{\beta}_k z_{jk}), j = 1 \div n \quad (3.2)$$

gọi là các **phần dư** của phép hồi quy.

Vì biểu thức theo Z_1, \dots, Z_k là tuyến tính nên phương trình:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 Z_1 + \dots + \hat{\beta}_k Z_k \quad (3.3)$$

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

Đặt:

$$\begin{aligned} \hat{y}_j &= \hat{\beta}_0 + \hat{\beta}_1 z_{j1} + \dots + \hat{\beta}_k z_{jk} \\ \hat{\mathbf{Y}} &= (\hat{y}_1, \dots, \hat{y}_n)^T \end{aligned} \quad (3.4)$$

Định lý 3.1

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

$$\hat{\beta} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} \quad (3.5)$$

Khi đó

$$\hat{Y} = \mathbf{Z} \hat{\beta} = \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} = H \mathbf{Y} \quad (3.6)$$

trong đó:

$$H = \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \text{ cấp } (n \times n) \quad (3.7)$$

$$\hat{\epsilon} = \mathbf{Y} - \hat{Y} = (\mathbf{I}_n - H) \mathbf{Y} \quad (3.8)$$

thỏa mãn:

$$\mathbf{Z}^T \hat{\epsilon} = 0 \text{ và } \hat{Y}^T \hat{\epsilon} = 0, (\hat{\beta}^T \mathbf{Z}^T \hat{\epsilon} = 0) \quad (3.9)$$

Tổng các phần dư:

$$\sum_{j=1}^n \hat{\epsilon}_j^2 = \hat{\epsilon}^T \hat{\epsilon} = \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T \mathbf{Z} \hat{\beta} \quad (3.10)$$

Chứng minh. Vì phiếm hàm $S(b) = \sum_{j=1}^n (y_j - b_0 - b_1 z_{j1} - \dots - b_k z_{jk})^2$ là hàm bậc hai theo b và b làm cực tiểu phiếm hàm này nên dễ thấy $\hat{\beta}$ có thể tìm được từ hệ phương trình sau:

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

ta có kết quả:

$$\begin{aligned} \sum_{j=1}^n (b_0 + b_1 z_{j1} + \dots + b_k z_{jk}) &= \sum_{j=1}^n y_j \\ b_0 \sum_{j=1}^n z_{j1} + b_1 \sum_{j=1}^n z_{j1}^2 + \dots + b_k \sum_{j=1}^n z_{jk} z_{j1} &= \sum_{j=1}^n y_j z_{j1} \\ b_0 \sum_{j=1}^n z_{j1} + b_1 \sum_{j=1}^n z_{j1} z_{jk} + \dots + b_k \sum_{j=1}^n z_{jk}^2 &= \sum_{j=1}^n y_j z_{jk} \end{aligned}$$

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

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

hoặc dưới dạng ma trận:

$$\mathbf{Z}^T \mathbf{Z} \mathbf{b} = \mathbf{Z}^T \mathbf{Y} \quad (3.11)$$

Phương trình (3.11) gọi là phương trình chuẩn.

Do $\text{rank}(\mathbf{Z}) = k + 1$ nên $\mathbf{Z}^T \mathbf{Z}$ có nghịch đảo, ta suy ra nghiệm:

$$\mathbf{b} = \hat{\beta} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y}$$

Ta thấy $\hat{\beta}$ là biểu thức tuyến tính theo \mathbf{Y} .

Để chứng minh $\hat{\beta}$ cực tiểu hóa $S(\mathbf{b})$ và thỏa mãn (3.9), (3.10) ta chú ý rằng ma trận $I - H$ là ma trận đối xứng lũy đẳng. Thật vậy:

(i) Tính đối xứng

$$(I - H)^T = (I - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T)^T = (I - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T) = (I - H)$$

(ii) Tính lũy đẳng

$$\begin{aligned} (I - H)^2 &= (I - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T)^2 \\ &= I - 2\mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T + \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \\ &= I - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T = I - H \end{aligned}$$

(iii) Ta có:

$$\mathbf{Z}^T (I - H) = \mathbf{Z}^T (I - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T) = \mathbf{Z}^T - \mathbf{Z}^T = 0$$

Dễ dàng thấy rằng:

$$\begin{aligned} S(\mathbf{b}) &= (\mathbf{Y} - \mathbf{Z}\mathbf{b})^T (\mathbf{Y} - \mathbf{Z}\mathbf{b}) = (\mathbf{Y} - \mathbf{Z}\hat{\beta} + \mathbf{Z}\hat{\beta} - \mathbf{Z}\mathbf{b})^T (\mathbf{Y} - \mathbf{Z}\hat{\beta} + \mathbf{Z}\hat{\beta} - \mathbf{Z}\mathbf{b}) \\ &= (\mathbf{Y} - \mathbf{Z}\hat{\beta})^T (\mathbf{Y} - \mathbf{Z}\hat{\beta}) + (\hat{\beta} - \mathbf{b})^T \mathbf{Z}^T \mathbf{Z} (\hat{\beta} - \mathbf{b}) \\ &\quad + (\hat{\beta} - \mathbf{b})^T \mathbf{Z}^T (I - H) \mathbf{Y} + \mathbf{Y}^T (I - H)^T \mathbf{Z} (\hat{\beta} - \mathbf{b}) \end{aligned}$$

Vì $(\hat{\beta} - b)^T \mathbf{Z}^T (I - H) \mathbf{Y} = 0$ và $\mathbf{Y}^T (I - H)^T \mathbf{Z} (\hat{\beta} - b) = 0$ do $\mathbf{Z}^T (I - H) = 0$ theo tính chất **iii** nên:

$$S(b) = (\mathbf{Y} - \mathbf{Z}\hat{\beta})^T (\mathbf{Y} - \mathbf{Z}\hat{\beta}) + (\hat{\beta} - b)^T \mathbf{Z}^T \mathbf{Z} (\hat{\beta} - b) \geq (\mathbf{Y} - \mathbf{Z}\hat{\beta})^T (\mathbf{Y} - \mathbf{Z}\hat{\beta}) = S(\hat{\beta})$$

Dấu "=" xảy ra khi $\hat{\beta} = b$. Hơn nữa:

$$\begin{aligned} \sum_{j=1}^n \hat{\varepsilon}_j^2 &= S(\hat{\beta}) = (\mathbf{Y} - \mathbf{Z}\hat{\beta})^T (\mathbf{Y} - \mathbf{Z}\hat{\beta}) = \mathbf{Y}^T (I - H) (I - H) \mathbf{Y} \\ &= \mathbf{Y}^T (I - H) \mathbf{Y} \text{ (theo tính chất ii)} = \mathbf{Y}^T \mathbf{Y} - \mathbf{Y}^T H \mathbf{Y} = \mathbf{Y}^T \mathbf{Y} - (\mathbf{Y}^T \mathbf{Z}) \hat{\beta} \end{aligned}$$

Đây chính là công thức (3.10).

Từ (3.8), (3.9), (3.10) ta nhận được: $\mathbf{Y}^T \mathbf{Y} = \sum_{j=1}^n y_j^2 = \hat{\mathbf{Y}}^T \hat{\mathbf{Y}} + \hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}$

hoặc:

$$\sum_{j=1}^n y_j^2 = \sum_{j=1}^n \hat{y}_j^2 + \sum_{j=1}^n \hat{\varepsilon}_j^2 \quad (3.12)$$

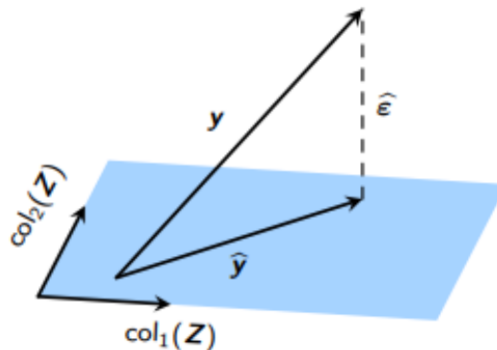
3.2 Tiếp cận bằng hình học

Trước đó, ta đã tìm hiểu rằng đại lượng $\hat{\beta}$ làm cực tiểu hóa phiếm hàm $S(b)$ được gọi là ước lượng bình phương cực tiểu của β . Bây giờ, ta sẽ nhìn dưới một góc độ hình học trực quan hơn.

Đặt $\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{Z}\beta$, suy ra $S(b) = (\mathbf{Y} - \mathbf{Z}\beta)^T (\mathbf{Y} - \mathbf{Z}\beta) = \boldsymbol{\varepsilon}^T \boldsymbol{\varepsilon}$

Yêu cầu tìm đại lượng $\hat{\beta}$ để $S(b)$ min tương đương với việc tìm đại lượng để $\hat{\beta}$ cực tiểu hóa độ dài vector sai số $\boldsymbol{\varepsilon}$.

Đặt $\hat{\mathbf{Y}} = \mathbf{Z}\hat{\beta}$, $\hat{\boldsymbol{\varepsilon}} = \mathbf{Y} - \mathbf{Z}\hat{\beta}$. Khi $b = \hat{\beta}$ làm cho tổng bình phương sai số đạt giá trị nhỏ nhất, hay $\hat{\boldsymbol{\varepsilon}}$ sẽ phải vuông góc với mặt phẳng (P) chứa tổ hợp tuyến tính các vector của \mathbf{Z} . Điều này có được là do tính chất hình học: đường vuông góc có độ dài nhỏ hơn đường xiên. Ta có thể minh họa hình học qua hình dưới:



Hình 3.1: Minh họa cho ước lượng bình phương cực tiểu

Bằng biểu diễn hình học như trên, ta có thể thấy ngay $\hat{\mathbf{Y}}$ là hình chiếu của \mathbf{Y} lên (P). Câu hỏi đặt ra là tìm ra hình chiếu $\hat{\mathbf{Y}}$ này như thế nào? Do $\hat{\mathbf{e}}$ vuông góc với mặt phẳng (P) nên sẽ vuông góc với tất cả các vector của Z hay $\mathbf{Z}^T \hat{\mathbf{e}} = 0$. Từ đây suy ra được rằng nếu ma trận Z có $\text{rank} = r + 1$ thì:

$$\hat{\beta} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y}$$

Và từ đó ta tìm được $\hat{\mathbf{Y}} = \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y}$

3.3 Phân tách tổng bình phương - Hệ số xác định R^2

Ở công thức (3.12) ta đã có:

$$\sum_{j=1}^n y_j^2 = \sum_{j=1}^n \hat{y}_j^2 + \sum_{j=1}^n \hat{\epsilon}_j^2$$

Đặt $\bar{y} = \sum_{k=1}^n \frac{y_k}{n}$ và $\bar{\hat{y}} = \sum_{k=1}^n \frac{\hat{y}_k}{n}$. Cũng từ tính chất $\mathbf{Z}^T \hat{\mathbf{e}} = 0$ ta được $1^T \hat{\mathbf{e}} = 0$ do cột đầu tiên của ma trận Z gồm toàn số 1. Khi đó:

$$0 = 1^T \hat{\mathbf{e}} = \sum_{k=1}^n \hat{\epsilon}_k = \sum_{k=1}^n y_k - \sum_{k=1}^n \hat{y}_k$$

Hay $\bar{y} = \bar{\hat{y}}$. Trừ hai vế của đẳng thức (3.12) cho $n\bar{y}^2 = n\bar{\hat{y}}^2$ ta được:

$$y^T y - n\bar{y}^2 = \hat{y}^T \hat{y} - n\bar{\hat{y}}^2 + \hat{\epsilon}^T \hat{\epsilon}$$

hoặc viết lại thành:

$$\sum_{k=1}^n (y_k - \bar{y})^2 = \sum_{k=1}^n (\hat{y}_k - \bar{\hat{y}})^2 + \sum_{k=1}^n \hat{\epsilon}_k^2 \quad (3.13)$$

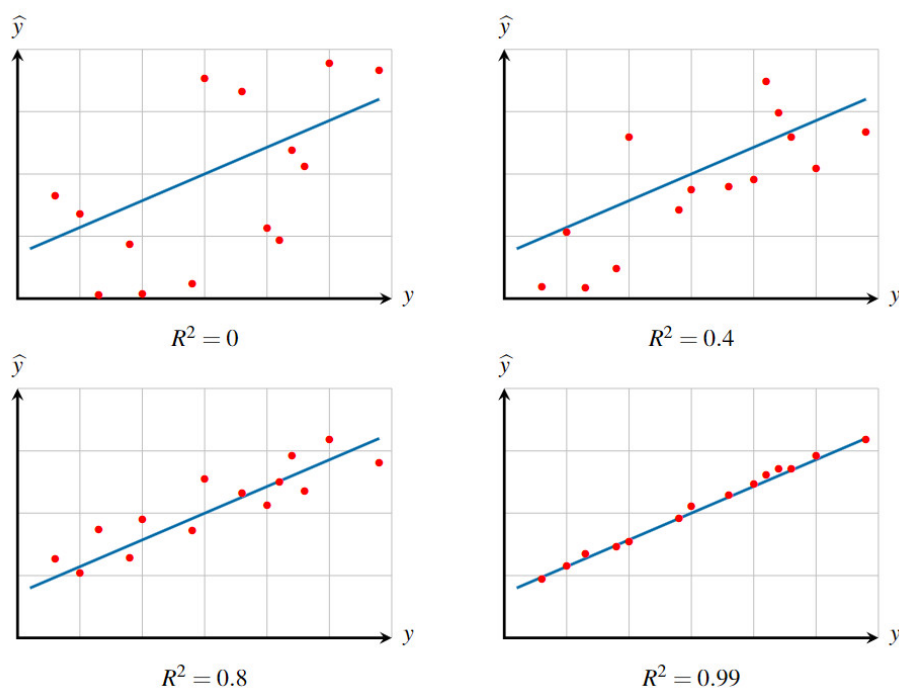
Đẳng thức (3.13) được gọi là phân tách tổng bình phương. Từ đẳng thức này mà người ta xây dựng được một tiêu chuẩn để đánh giá mức độ phù hợp của bộ dữ liệu với mô hình hồi quy tuyến tính. Ta định nghĩa

$$R^2 := \frac{\hat{y}^T \hat{y} - n\bar{\hat{y}}^2}{y^T y - n\bar{y}^2} = \frac{\sum_{k=1}^n (\hat{y}_k - \bar{\hat{y}})^2}{\sum_{k=1}^n (y_k - \bar{y})^2} = 1 - \frac{\sum_{k=1}^n \hat{\epsilon}_k^2}{\sum_{k=1}^n (y_k - \bar{y})^2}$$

Giá trị của R^2 gọi là bình phương của hệ số xác định, đó là tỷ lệ biến thiên của các biến được giải thích bởi các biến z_{j1}, \dots, z_{jk} .

Chẳng hạn, nếu như $R^2 = 0.5$ thì có nghĩa là khoảng một nửa số quan sát trong bộ dữ liệu có thể được biểu diễn bằng mô hình. Nếu như $R^2 = 1$ có nghĩa là mô hình đã cho đi qua tất cả các điểm hay $\hat{\epsilon}_j = 0$ với mọi j . Khi $\hat{\beta}_0 = \bar{y}$ hoặc $\hat{\beta}_1 = \hat{\beta}_2 = \dots = \hat{\beta}_k = 0$ thì $R^2 = 0$. Trong trường hợp này, các biến dự đoán z_{j1}, \dots, z_{jk} không có mối liên hệ với nhau.

Người ta đã nghiên cứu và chỉ ra rằng với giá trị $R^2 \geq 0.5$ thì mô hình sẽ được đánh giá là phù hợp.



Hình 3.2: Ví dụ về một giá trị R^2

Thay vì sử dụng hệ số xác định R^2 thì ta sẽ sử dụng hệ số xác định \bar{R}^2 hiệu chỉnh.

- Kí hiệu: \bar{R}^2
- Định nghĩa:

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

Giá trị \bar{R}^2 thường được sử dụng thay cho giá trị R^2 thông thường khi so sánh các mô hình hồi quy có số lượng biến số khác nhau..

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

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

$$E(\hat{\beta}) = \beta; cov(\hat{\beta}) = \sigma^2(\mathbf{Z}^T \mathbf{Z})^{-1} \quad (3.14)$$

- Phần dư $\hat{\epsilon}$ có tính chất:

$$E(\hat{\epsilon}) = 0; cov(\hat{\epsilon}) = \sigma^2(I - H) \quad (3.15)$$

- $\hat{\sigma}^2 = \frac{\hat{\epsilon}^T \hat{\epsilon}}{n-k-1} = \sum_{j=1}^n \frac{\hat{\epsilon}_j^2}{n-k-1}$ là ước lượng không chệch của σ^2 , tức là $E(\hat{\sigma}^2) = \sigma^2$

- $\hat{\beta}, \hat{\epsilon}$ là không tương quan, tức là:

$$cov(\hat{\beta}, \hat{\epsilon}) = 0; cov(\hat{\beta}, \hat{\sigma}^2) = 0 \quad (3.16)$$

Chứng minh. Trước tiên ta xét các phép biến đổi sau đây:

$$\begin{aligned}
 \hat{\beta} &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} \\
 &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (\mathbf{Z}\beta + \varepsilon) \\
 &= \beta + (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \varepsilon \\
 \hat{\varepsilon} &= \mathbf{Y} - \mathbf{Z}\hat{\beta} \\
 &= \mathbf{Y} - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} \\
 &= (\mathbf{I} - \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T) \mathbf{Y} \\
 &= (\mathbf{I} - \mathbf{H}) \mathbf{Y}
 \end{aligned}$$

$$\text{Với } \mathbf{H} = \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T$$

1) Ta có phép biến đổi sau:

$$\begin{aligned}
 E(\hat{\beta}) &= E(\beta + (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \varepsilon) \\
 &= E(\beta) + E((\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \varepsilon) \\
 &= \beta + (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T E(\varepsilon) \\
 &= \beta \\
 \text{cov}(\hat{\beta}) &= \text{cov}(\beta + (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \varepsilon) \\
 &= \text{cov}((\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \varepsilon) \\
 &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \text{cov}(\varepsilon) \mathbf{Z}(\mathbf{Z}^T \mathbf{Z})^{-1} \\
 &= \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \\
 &= \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1}
 \end{aligned}$$

2) Do $\hat{\varepsilon} = (\mathbf{I} - \mathbf{H}) \mathbf{Y}$ nên:

$$\begin{aligned}
 E(\hat{\varepsilon}) &= (\mathbf{I} - \mathbf{H}) E(\mathbf{Y}) = (\mathbf{I} - \mathbf{H}) \mathbf{Z}\beta = \mathbf{0} \\
 \text{cov}(\hat{\varepsilon}) &= (\mathbf{I} - \mathbf{H})^T \mathbf{I} (\mathbf{I} - \mathbf{H}) \sigma^2 = \sigma^2 (\mathbf{I} - \mathbf{H})
 \end{aligned}$$

3) Từ (2) ta suy ra:

$$\begin{aligned}
 E(\hat{\varepsilon}^T \hat{\varepsilon}) &= \sum_{j=1}^n E \hat{\varepsilon}_j^2 = \text{tr}(\text{cov}(\hat{\varepsilon})) = \sigma^2 \text{tr}(\mathbf{I}_n - \mathbf{H}) \\
 &= \sigma^2 (n - \text{tr}(\mathbf{H}))
 \end{aligned}$$

Mặt khác,

$$\begin{aligned}\text{tr}(H) &= \text{tr} \left(\mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \right) = \text{tr} \left((\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Z} \right) = \text{tr} (I_{k+1}) = k+1 \\ \Rightarrow E(\hat{\boldsymbol{\varepsilon}}^T \hat{\boldsymbol{\varepsilon}}) &= \sigma^2 (n - k - 1)\end{aligned}$$

4) Ta có:

$$\begin{aligned}\text{cov}(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\varepsilon}}) &= \text{cov} \left((\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} (I_n - H) \mathbf{Y} \right) \\ &= (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \text{cov}(\mathbf{Y}) (I_n - H) \\ &= \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T (I_n - H) = 0\end{aligned}$$

□

3.5 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:

$$c^T \hat{\boldsymbol{\beta}} = c_0 \hat{\beta}_0 + c_1 \hat{\beta}_1 + \cdots + c_k \hat{\beta}_k$$

của $c^T \boldsymbol{\beta} = c_0 \beta_0 + c_1 \beta_1 + \cdots + c_k \beta_k$ là ước lượng không chệch với phương sai bé nhất so với bất kỳ ước lượng tuyến tính không chệch nào dạng $a^T \mathbf{Y} = a_1 y_1 + \cdots + a_n y_n$. Nếu thêm giả thiết rằng $\boldsymbol{\varepsilon}$ có phân bố chuẩn $N_n(0, \sigma^2 I_n)$ thì $c^T \hat{\boldsymbol{\beta}}$ là một ước lượng không chệch với phương sai cực tiểu của $c^T \boldsymbol{\beta}$ so với bất kỳ ước lượng không chệch nào khác.

Chứng minh. Do tính chất tuyến tính của kỳ vọng nên rõ ràng $c^T \hat{\boldsymbol{\beta}}$ là ước lượng không chệch của $c^T \boldsymbol{\beta}$.

1) Giả sử $a^T \mathbf{Y}$ là một ước lượng không chệch của $c^T \boldsymbol{\beta}$ thì:

$$E(a^T \mathbf{Y}) = a^T E(\mathbf{Y}) = a^T \mathbf{Z} \boldsymbol{\beta} \equiv c^T \boldsymbol{\beta} \Leftrightarrow (a^T \mathbf{Z} - c^T) \boldsymbol{\beta} \equiv 0$$

Với mọi $\boldsymbol{\beta}$, đặc biệt khi $\boldsymbol{\beta}^T = a^T \mathbf{Z} - c^T$, ta có:

$$\boldsymbol{\beta}^T \boldsymbol{\beta} = 0 \Leftrightarrow a^T \mathbf{Z} - c^T = 0 \Leftrightarrow a^T \mathbf{Z} = c^T$$

Chú ý rằng:

$$c^T \hat{\boldsymbol{\beta}} = c^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} = a^{*T} \mathbf{Y}$$

với $a^{*T} = c^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \Leftrightarrow a^* = \mathbf{Z} (\mathbf{Z}^T \mathbf{Z})^{-1} c$.

$$\begin{aligned} \text{Var}(a^T \mathbf{Y}) &= a^T \text{cov}(\mathbf{Y}) a = \sigma^2 a^T a \\ &= \sigma^2 (a - a^* + a^*)^T (a - a^* + a^*) \\ &= \sigma^2 (a - a^*)^T (a - a^*) + \sigma^2 (a^*)^T a^* + 2(a - a^*)^T a^* \sigma^2 \\ &= \sigma^2 (a - a^*)^T (a - a^*) + \sigma^2 a^{*T} a^* \geq \text{Var}(a^* \mathbf{Y}) \end{aligned}$$

Vì:

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

Trong (3.21) dấu "=" xảy ra $\Leftrightarrow a = a^*$

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

□

FaMI
1956



FaMI

1956

Ước lượng khoảng của mô hình hồi quy

4.1 Khoảng tin cậy của các hệ số hồi quy β_j

Giả sử ta có một mô hình hồi quy tuyến tính cổ điển như sau:

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \beta_4 z_4 + \varepsilon, \text{ trong đó:}$$

- Y là lượng cholesterol trong máu
- z_1 là tuổi
- z_2 là trọng lượng cơ thể
- z_3 là mức độ tuân thủ tiêu chuẩn dinh dưỡng
- z_4 là chiều cao
- ε là sai số hồi quy
- $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ là các tham số hồi quy chưa biết

Ta sẽ muốn đánh giá sự quan trọng của các biến dự đoán z_1, z_2, z_3, z_4 . Ta làm điều này bằng cách tìm ước lượng khoảng của các tham số $\beta_1, \beta_2, \beta_3, \beta_4$.

Ví dụ, với ước lượng khoảng của β_4 là $(0.255, 0.361)$ có chứa $\beta_4 = 0$, ta có thể kết luận chiều cao đóng góp không nhiều trong việc dự đoán lượng cholesterol trong máu và do đó, có thể được cân nhắc loại bỏ khỏi mô hình hồi quy khi cần thiết. Quá trình suy luận này được gọi là suy luận liên quan đến tham số hồi quy.

Trước khi có thể đánh giá sự quan trọng của các biến cụ thể trong hàm hồi quy $Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3 + \dots + \beta_k z_k + \varepsilon$ (2.1), ta phải tìm ra phân phối mẫu của ước lượng bình phương cực tiểu $\hat{\beta}$ và phân phối mẫu của tổng bình phương sai số $\hat{\varepsilon}^T \hat{\varepsilon}$. Để làm được điều đó, ta sẽ giả sử rằng sai số ε có phân phối chuẩn.

Hệ quả 4.1

Cho $Y = Z\beta + \varepsilon$ với Z là ma trận đầy hạng $r + 1$ và $\varepsilon \sim N_n(0, \sigma^2 I)$. Khi đó, ước lượng hợp lý cực đại của β bằng với ước lượng bình phương cực tiểu $\hat{\beta}$. Hơn nữa:

$$\bullet \hat{\beta} = (Z^T Z)^{-1} Z^T Y \sim N_{r+1}(\beta, \sigma^2 (Z^T Z)^{-1}) \quad (4.1.1)$$

$$\bullet \hat{\beta} \text{ độc lập với sai số } \hat{\varepsilon} = Y - Z\hat{\beta} \quad (4.1.2)$$

$$\bullet n\hat{\sigma}^2 = \hat{\varepsilon}^T \hat{\varepsilon} \sim \sigma^2 X_{n-r-1}^2, \text{ trong đó } \hat{\sigma}^2 \text{ là ước lượng hợp lý cực đại của } \sigma^2 \quad (4.1.3)$$

Chứng minh hệ quả 4.1.1.

Vì $Y = Z\beta + \varepsilon$ là tổ hợp tuyến tính của ε phân phối chuẩn nên Y phân phối chuẩn. Ta có:

- $E[Y] = E[Z\beta + \varepsilon] = Z\beta + E[\varepsilon] = Z\beta$
- $cov[Y] = cov[Z\beta + \varepsilon] = cov[\varepsilon] = \sigma^2 I$

Vậy $Y \sim N_n(Z\beta, \sigma^2 I)$.

Ở (3.5), ta đã có $\hat{\beta} = (Z^T Z)^{-1} Z^T Y$, là tổ hợp tuyến tính của Y phân phối chuẩn nên $\hat{\beta}$ phân phối chuẩn. Sử dụng kết quả (3.16), ta có:

- $E[\hat{\beta}] = \beta$
- $cov[\hat{\beta}] = \sigma^2 (Z^T Z)^{-1}$

Vậy $\hat{\beta} \sim N_{r+1}(\beta, \sigma^2 (Z^T Z)^{-1})$.

Chứng minh hệ quả 4.1.2.

Thật vậy, ở (3.8) ta đã có $\hat{\varepsilon} = (I - H)Y$, là tổ hợp tuyến tính của Y phân phối chuẩn nên $\hat{\varepsilon}$ phân phối chuẩn. Sử dụng kết quả (3.17), ta có:

- $E[\hat{\varepsilon}] = 0$
- $cov[\hat{\varepsilon}] = \sigma^2 (I - H)$

Vậy $\hat{\varepsilon} \sim N_n(0, \sigma^2 (I - H))$.

Kết hợp kết luận này với $\hat{\beta} \sim N_{r+1}(\beta, \sigma^2 (Z^T Z)^{-1})$ vừa chứng minh và $cov(\hat{\beta}, \hat{\varepsilon}) = 0$ ở (3.18), ta được $\hat{\beta}$ và $\hat{\varepsilon}$ độc lập.

Chứng minh hệ quả 4.1.3.

Bước 1: Biểu diễn ma trận $I - H$ dưới dạng phân tách phổ

Vì Z là ma trận đầy hạng $r + 1$ nên

$$rank(Z) = rank(Z(Z^T Z)^{-1} Z^T) = rank(H) = r + 1$$

Gọi $\lambda_i, e_i, i = \overline{1, n}$ là các giá trị riêng và vector riêng trực chuẩn tương ứng của ma trận lũy đẳng $I - H$ (3.12), ta có:

$$\lambda_i e_i = (I - H)e_i = (I - H)(I - H)e_i = (I - H)\lambda_i e_i = \lambda_i (I - H)e_i = \lambda_i^2 e_i$$

Suy ra $\lambda_i = \lambda_i^2$ hay $\lambda_i \in \{0, 1\} \forall i = \overline{1, n}$. Đặt $\Lambda = diag(\lambda_1, \lambda_2, \dots, \lambda_n)$ và $U = [e_1 e_2 \dots e_n]$. Do $I - H$ là ma trận lũy đẳng nên $U^T = U^{-1}$. Khi đó:

$$H = I - (I - H) = U^{-1}U - U^{-1}\Lambda U = U^T U - U^T \Lambda U = U^T (I - \Lambda) U$$

Bước 2: Tạo phân phối chi-bình phương $\sigma^2 \chi_{n-r-1}^2$ từ $e_1, e_2, \dots, e_{n-r-1}$ và ε

Đặt $E = \begin{bmatrix} E_1 \\ E_2 \\ \dots \\ E_{n-r-1} \end{bmatrix} = \begin{bmatrix} e_1^T \\ e_2^T \\ \dots \\ e_{n-r-1}^T \end{bmatrix} \varepsilon = M\varepsilon$. Vì E tuyến tính theo ε phân phối chuẩn nên E phân phối chuẩn. Ta có:

- $E[E_i] = E[e_i^T \varepsilon] = e_i^T E[\varepsilon] = 0$
- $cov[E_i] = cov[e_i^T \varepsilon] = e_i^T cov[\varepsilon] e_i = \sigma^2 |e_i|^2 = \sigma^2$
- $cov[E_i, E_j] = cov[e_i^T \varepsilon, e_j^T \varepsilon] = e_i^T cov[\varepsilon] e_j = \sigma^2 e_i^T e_j = 0 \forall i \neq j$

Vậy $E_1, E_2, \dots, E_{n-r-1}$ độc lập và $E_i \sim N(0, \sigma^2)$ hay $\frac{E_i}{\sigma} \sim N(0, 1) \forall i = \overline{1, n-r-1}$. Do đó:

$$\begin{aligned} \left(\frac{E_1}{\sigma}\right)^2 + \left(\frac{E_2}{\sigma}\right)^2 + \dots + \left(\frac{E_{n-r-1}}{\sigma}\right)^2 &\sim \chi_{n-r-1}^2 \\ \implies E_1^2 + E_2^2 + \dots + E_{n-r-1}^2 &\sim \sigma^2 \chi_{n-r-1}^2 \end{aligned}$$

Bước 3: Chứng minh $\widehat{\varepsilon}^T \widehat{\varepsilon} \sim \sigma^2 \chi_{n-r-1}^2$

Ta đã có $I - H$ là ma trận đối xứng (3.11), lũy đẳng (3.12) và $Z^T(I - H) = 0$ (3.13) nên $(Z^T(I - H))^T = (I - H)Z = 0$. Khi đó:

$$\begin{aligned} \widehat{\varepsilon}^T \widehat{\varepsilon} &= ((I - H)Y)^T (I - H)Y = Y^T (I - H)^T (I - H)Y = (Z\beta + \varepsilon)^T (I - H)(Z\beta + \varepsilon) \\ &= \beta^T Z^T (I - H)\beta Z + \beta^T Z^T (I - H)\varepsilon + \varepsilon^T (I - H)Z\beta + \varepsilon^T (I - H)\varepsilon = \varepsilon^T (I - H)\varepsilon \\ &= \varepsilon^T (e_1 e_1^T + e_2 e_2^T + \dots + e_{n-r-1} e_{n-r-1}^T) \varepsilon = \varepsilon^T M^T M \varepsilon = E^T E = E_1^2 + E_2^2 + \dots + E_{n-r-1}^2 \sim \sigma^2 \chi_{n-r-1}^2 \end{aligned}$$

Hệ quả 4.2: Cho $Y = Z\beta + \varepsilon$ với Z là ma trận đầy hạng $r + 1$ và $\varepsilon \sim N_n(0, \sigma^2 I)$. Khi đó:

i) Miền tin cậy mức $100(1 - \alpha)\%$ của β được cho bởi:

$$(\beta - \widehat{\beta})^T Z^T Z (\beta - \widehat{\beta}) \leq (r + 1) s^2 F_{r+1, n-r-1}(\alpha) \quad (4.2.1)$$

ii) Khoảng tin cậy đồng thời mức $100(1 - \alpha)\%$ của β_i được cho bởi:

$$\widehat{\beta}_i \pm \sqrt{\widehat{Var}(\widehat{\beta}_i)} \sqrt{(r + 1) F_{r+1, n-r-1}(\alpha)}, i = \overline{0, r} \quad (4.2.2)$$

trong đó $\widehat{Var}(\widehat{\beta}_i)$ là phần tử đường chéo chính của ma trận $s^2(Z^T Z)^{-1}$ tương ứng với β_i .

Chứng minh hệ quả 4.2.1.

Bước 1: Chứng minh $V^T V = (\beta - \widehat{\beta})^T (Z^T Z) (\beta - \widehat{\beta}) \sim \sigma^2 \chi_{r+1}^2$

Xét ma trận căn bậc 2 đối xứng $(Z^T Z)^{\frac{1}{2}}$. Đặt $V = (Z^T Z)^{\frac{1}{2}} (\beta - \widehat{\beta})$, là tổ hợp tuyến tính của $\widehat{\beta}$ phân phối chuẩn nên V phân phối chuẩn. Ta có:

- $E[V] = E[(Z^T Z)^{\frac{1}{2}} (\beta - \widehat{\beta})] = (Z^T Z)^{\frac{1}{2}} E[\beta - \widehat{\beta}] = 0$
- $cov[V] = (Z^T Z)^{\frac{1}{2}} cov(\beta - \widehat{\beta}) ((Z^T Z)^{\frac{1}{2}})^T = (Z^T Z)^{\frac{1}{2}} cov[\widehat{\beta}] (Z^T Z)^{\frac{1}{2}} = (Z^T Z)^{\frac{1}{2}} \sigma^2 (Z^T Z)^{-1} (Z^T Z)^{\frac{1}{2}} = \sigma^2 I$

Vậy $V \sim N_{r+1}(0, \sigma^2 I)$. khi đó, $V_1^2 + V_2^2 + \dots + V_{r+1}^2 = V^T V = (\beta - \hat{\beta})^T (Z^T Z)^{\frac{1}{2}} (Z^T Z)^{\frac{1}{2}} (\beta - \hat{\beta}) = (\beta - \hat{\beta})^T (Z^T Z) (\beta - \hat{\beta}) \sim \sigma^2 \chi_{r+1}^2$

Bước 2: Tạo phân phối fisher từ $V^T V$ và $\hat{\varepsilon}^T \hat{\varepsilon}$ độc lập, có phân phối chi-bình phương

Ở (4.1.3), ta đã có $\hat{\varepsilon}^T \hat{\varepsilon} \sim \sigma^2 \chi_{n-r-1}^2$, độc lập với $\hat{\beta}$ do đó độc lập với $V^T V$.

$$\Rightarrow F = \frac{\frac{V^T V}{\sigma^2(r+1)}}{\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{\sigma^2(n-r-1)}} = \frac{V^T V}{(r+1)s^2} \sim F_{r+1, n-r-1}$$

$$\Rightarrow (\beta - \hat{\beta})^T (Z^T Z) (\beta - \hat{\beta}) = V^T V = (r+1)s^2 F$$

Do $P(F \leq F_{r+1, n-r-1}(\alpha)) = 1 - \alpha$ nên ta được miền tin cậy mức $100(1 - \alpha)\%$ của β là

$$(\beta - \hat{\beta})^T Z^T Z (\beta - \hat{\beta}) \leq (r+1)s^2 F_{r+1, n-r-1}(\alpha)$$

Đây chính là elip độ tin cậy của β . Tâm của elip nằm ở ước lượng hợp lý cực đại $\hat{\beta}$, hướng và kích thước của nó được xác định bởi các giá trị riêng và các vector riêng của ma trận $Z^T Z$.

Chứng minh hệ quả 4.2.2.

Do Z là ma trận đầy hạng $r+1$, ta chọn hệ vector cơ sở của Z là hệ vector cơ sở chính tắc của R^{r+1} . Theo (5A.1), hình chiếu của elip độ tin cậy lên vector u_i là:

$$\begin{aligned} |(\beta - \hat{\beta})u_i| &\leq \sqrt{(r+1)F_{r+1, n-r-1}(\alpha)} \sqrt{u_i^T s^2 (Z^T Z)^{-1} u_i} \\ &\Leftrightarrow |\beta_i - \hat{\beta}_i| \leq \sqrt{(r+1)F_{r+1, n-r-1}(\alpha)} \sqrt{\widehat{Var}(\hat{\beta}_i)} \\ &\Rightarrow \beta_i \leq \hat{\beta}_i \pm \sqrt{(r+1)F_{r+1, n-r-1}(\alpha)} \sqrt{\widehat{Var}(\hat{\beta}_i)} \end{aligned}$$

trong đó $\widehat{Var}(\hat{\beta}_i)$ là phần tử đường chéo chính của ma trận $s^2 (Z^T Z)^{-1}$ tương ứng với β_i .

Trong thực tế, ta thường bỏ qua mức tin cậy “đồng thời” của các ước lượng khoảng trong hệ quả này, thay thế $\sqrt{(r+1)F_{r+1, n-r-1}(\alpha)}$ bởi $t_{n-r-1}(\frac{\alpha}{2})$ và sử dụng khoảng:

$$\hat{\beta}_i \pm t_{n-r-1}(\frac{\alpha}{2}) \sqrt{\widehat{Var}(\hat{\beta}_i)} \quad (4.2.3)$$

khi nghiên cứu các biến dự đoán quan trọng.

Ví dụ 4.1. (Xây dựng mô hình hồi quy cho dữ liệu bất động sản)

Từ dữ liệu thẩm định trong bảng được thu thập từ 20 ngôi nhà ở Milwaukee, Wisconsin và khu vực lân cận. Ta có mô hình hồi quy:

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \varepsilon_j$$

Ta tính được các ước lượng bình phương cực tiểu $\hat{\beta}$ của hệ số hồi quy là:

$$(Z^T Z)^{-1} = \begin{bmatrix} 5.1523 & 0.2544 & -0.1463 \\ 0.2544 & 0.0512 & -0.0172 \\ -0.1463 & -0.0172 & 0.0067 \end{bmatrix}$$

$$\hat{\beta} = (Z^T Z)^{-1} Z^T Y = \begin{bmatrix} 30.967 \\ 2.634 \\ 0.045 \end{bmatrix}$$

Tiếp theo, ta đi tính độ lệch chuẩn ước tính của các ước lượng này:

STT	z_1 Tổng diện tích nhà (100/t ²)	z_2 Giá thẩm định (\$1000)	Y Giá bán (\$1000)
1	15.31	57.3	74.8
2	15.20	63.8	74.0
3	16.25	65.4	72.9
4	14.33	57.0	70.0
5	14.57	63.8	74.9
6	17.33	63.2	76.0
7	14.48	60.2	72.0
8	14.91	57.7	73.5
9	15.25	56.4	74.5
10	13.89	55.6	73.5
11	15.18	62.6	71.5
12	14.44	63.4	71.0
13	14.87	60.2	78.9
14	18.63	67.2	86.5
15	15.20	57.1	68.0
16	25.76	89.6	102.0
17	19.05	68.6	84.0
18	15.37	60.1	69.0
19	18.06	66.3	88.0
20	16.35	65.8	76.0

$$\hat{\varepsilon} = Y - Z\hat{\beta} = [0.9280 \ 0.1252 \ \dots \ -0.9939]^T$$

$$s = \sqrt{\frac{\hat{\varepsilon}^T \hat{\varepsilon}}{n-r-1}} = \sqrt{\frac{205.0013}{17}} = 3.473$$

$$\sqrt{\widehat{Var}(\hat{B}_1)} = s\sqrt{(Z^T Z)^{-1}_{11}} = 3.473\sqrt{5.1523} = 7.88$$

$$\sqrt{\widehat{Var}(\hat{B}_2)} = s\sqrt{(Z^T Z)^{-1}_{22}} = 3.473\sqrt{0.0512} = 0.785$$

$$\sqrt{\widehat{Var}(\hat{B}_3)} = s\sqrt{(Z^T Z)^{-1}_{33}} = 3.473\sqrt{0.0067} = 0.285$$

Cuối cùng, ta tính hệ số xác định R^2 :

$$\hat{Y} = Z\hat{\beta} = [73.8720 \ 73.8748 \ \dots \ 76.9939]^T$$

$$\bar{Y} = \sum_{j=1}^{20} Y_j = 76.55$$

$$R^2 = \frac{\sum_{j=1}^{20} (\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^{20} (Y_j - \bar{Y})^2} = \frac{1032.23}{1237.87} = 0.834$$

Do đó, phương trình phù hợp là:

$$\hat{Y} = 30.967 + 2.634z_1 + 0.045z_2$$

(7.88) (0.785) (0.285)

Các con số trong ngoặc đơn là độ lệch chuẩn ước tính của các hệ số bình phương tối thiểu. Hơn nữa, $R^2 = 0.834$ cho thấy rằng dữ liệu thể hiện một mối quan hệ hồi quy mạnh. Nếu các sai số ε vượt qua các kiểm tra chẩn đoán sẽ được mô tả trong phần 6 thì phương trình này có thể được sử dụng để dự đoán giá bán của một căn nhà khác trong khu vực dựa trên kích thước và giá thẩm định của nó.

Khoảng tin cậy mức 95% của β_2 là:

$$\hat{\beta}_2 \pm t_{17}(0.025) \sqrt{\widehat{Var}(\hat{\beta}_2)} = 0.045 \pm 2.110 * 0.285 = (-0.556, 0.647)$$

Do khoảng tin cậy có chứa $\beta_2 = 0$, biến z_2 có thể được loại bỏ khỏi mô hình hồi quy và việc phân tích lặp lại với biến dự đoán duy nhất z_1 . Vậy là khi biết được diện tích nhà, giá trị thẩm định đường như chỉ góp một phần nhỏ trong việc dự đoán giá bán.

4.2 Kiểm định tỷ số hợp lý của các hệ số hồi quy

Xét mô hình HQTTC cổ điển

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{Z}_1 + \beta_2 \mathbf{Z}_2 + \dots + \beta_k \mathbf{Z}_k + \varepsilon$$

Khi thiết lập phương trình, ta giả sử rằng mọi biến độc lập $\mathbf{Z}_1, \dots, \mathbf{Z}_k$ 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, hoặc có ảnh hưởng rất ít đến biến phản hồi và ta có thể lược bỏ đi.

Đặt

$$\beta = \begin{bmatrix} \beta_{(1)} \\ (q+1) \times 1 \\ \dots \\ \beta_{(2)} \\ (k-q) \times 1 \end{bmatrix}$$

Ta có bài toán kiểm định giả thuyết:

$$H_0 : \beta_{q+1} = \dots = \beta_k (0 < q < k) \quad \text{Hay} \quad H_0 : \beta_{(2)} = 0 \quad (4.3.1)$$

với đối thuyết $K : \exists i \in \{q+1, \dots, k\}$ sao cho $\beta_i \neq 0$

Giả thuyết H_0 có nghĩa là các biến độc lập không tham gia vào biểu thức tuyến tính (2.1), ngược lại đối thuyết K nói rằng có ít nhất một trong các biến này có liên quan đến mô hình.

Xét mô hình hồi quy tuyến tính cổ điển

$$\begin{aligned} Y = \mathbf{Z}\beta + \varepsilon &= \begin{bmatrix} \mathbf{Z}_{(1)} & \vdots & \mathbf{Z}_{(2)} \\ n \times (q+1) & & n \times (k-q) \end{bmatrix} \begin{bmatrix} \beta_{(1)} \\ (q+1) \times 1 \\ \dots \\ \beta_{(2)} \\ (k-q) \times 1 \end{bmatrix} + \varepsilon \\ &= \mathbf{Z}_{(1)}\beta_{(1)} + \mathbf{Z}_{(2)}\beta_{(2)} + \varepsilon \end{aligned}$$

Hệ quả 4.2

Xét mô hình hồi quy tuyến tính $\mathbf{Y} = \mathbf{Z}\beta + \varepsilon$ với ma trận \mathbf{Z} đầy hạng và $\varepsilon \sim N_n(0, \sigma^2 \mathbf{I})$. Khi đó ta bác bỏ giả thuyết $H_0 : \beta_{(2)} = 0$ với độ tin cậy $1 - \alpha$ nếu :

$$SS_{res}(\mathbb{Z}_{(1)}) - SS_{res}(\mathbb{Z}) > \hat{\sigma}^2(k-q)F_{k-q, n-k-1}(\alpha) \quad (4.4.1)$$

Chứng minh hệ quả 4.4.1

Hàm hợp lý tham số β và σ^2 là :

$$L(\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2} \sigma^n} e^{-\frac{(y - Z\beta)^T (y - Z\beta)}{2\sigma^2}} \leq \frac{1}{(2\pi)^{n/2} \hat{\sigma}^n} e^{-\frac{n}{2}}$$

Max của $L(\beta, \sigma^2)$ xảy ra khi $\beta = \hat{\beta} = (Z^T Z)^{-1} Z^T Y$ và $\hat{\sigma}^2 = \frac{(y - Z\hat{\beta})^T (y - Z\hat{\beta})}{n}$ ($\hat{\sigma}^2$ là ước lượng hợp lý cực đại của σ^2)

Tương tự với $Y = Z_{(1)}\beta_{(1)} + \varepsilon$ thì:

$$\max_{\beta_{(1)}, \sigma^2} L(\beta_{(1)}, \sigma^2) = \frac{1}{(2\pi)^{n/2} \hat{\sigma}_1^n} e^{-\frac{n}{2}}$$

tại $\beta_{(1)} = \hat{\beta}_{(1)} = (Z_{(1)}^T Z_{(1)})^{-1} Z_{(1)}^T Y$ và $\hat{\sigma}_1^2 = \frac{(y - Z_{(1)}\hat{\beta}_{(1)})^T (y - Z_{(1)}\hat{\beta}_{(1)})}{n}$ ($\hat{\sigma}_1^2$ là ước lượng hợp lý cực đại của σ_1^2)

Ta bác bỏ giả thuyết $H_0 : \beta_{(2)} = 0$ nếu

$$\frac{\max L(\beta_{(1)}, \hat{\sigma}_1^2)}{\max L(\beta, \hat{\sigma}^2)} = \left(\frac{\hat{\sigma}_1^2}{\hat{\sigma}^2} \right)^{-n/2} = \left(1 + \frac{\hat{\sigma}_1^2 - \hat{\sigma}^2}{\hat{\sigma}^2} \right)^{-n/2}$$

nhỏ hay giá trị $\frac{\hat{\sigma}_1^2 - \hat{\sigma}^2}{\hat{\sigma}^2}$ lớn

$$F = \frac{n(\hat{\sigma}_1^2 - \hat{\sigma}^2)/(k - q)}{n\hat{\sigma}^2/(n - k - 1)} = \frac{(SS_{res}(Z_{(1)}) - SS_{res}(Z))/(k - q)}{\hat{\sigma}^2}$$

$F \sim F_{k-q, n-k-1} \Rightarrow P(F > F_{k-q, n-k-1}(\alpha)) = \alpha$

Suy ra ta bác bỏ giả thuyết $H_0 : \beta_{(2)} = 0$ với độ tin cậy $1 - \alpha$ nếu :

$$SS_{res}(Z_{(1)}) - SS_{res}(Z) > \hat{\sigma}^2(k - q)F_{k-q, n-k-1}(\alpha)$$

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

$$H_0 : \begin{cases} c_{10}\beta_0 + c_{11}\beta_1 + \dots + c_{1k}\beta_k = a_1 \\ c_{20}\beta_0 + c_{21}\beta_1 + \dots + c_{2k}\beta_k = a_2 \\ \dots \\ c_{k-q,0}\beta_0 + c_{k-q,1}\beta_1 + \dots + c_{k-q,k}\beta_k = a_{k-q} \end{cases} \Leftrightarrow C\beta = a \quad (4.4.2)$$

trong đó $C = [c_{ij}]_{k-q, k+1}; a = [a_1, \dots, a_{k-q}]^T$

Bài toán đang xét (4.4.2) là trường hợp riêng của (4.4.1) với:

$$C = \begin{bmatrix} 0 & 0 & \dots & 0 & \vdots & 1 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & \vdots & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 0 & \vdots & 0 & 0 & \dots & 1 \end{bmatrix} = [0; I_{k-q}]$$

Định lý 4.1

Quy tắc kiểm định:

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

$$((C\hat{\beta})^T (C(\mathbf{Z}^T \mathbf{Z})^{-1} C^T)^{-1} C\hat{\beta}) / \hat{\sigma}^2 > (k-q) F_{k-q, n-k-1}(\alpha) \quad (4.4.3)$$

Định lý 4.2

Nhận xét

Ta có thể sử dụng mệnh đề (4.2.3) về khoảng tin cậy của $\beta_{q+1}, \dots, \beta_k$ với các đầu mút

$\hat{\beta}_i \pm t_{n-k-1} \left(\frac{\alpha}{2} \right) \sqrt{\widehat{D}(\hat{\beta}_i)}$ để kiểm định giả thuyết (4.6). Điều đó có nghĩa là nếu tồn tại chỉ số $i \in \{q+1, \dots, k\}$ thỏa mãn:

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

thì ta coi $\beta_i \neq 0$

Ví dụ 4.1

Để biết sự phụ thuộc giữa khối lượng và các kích thước của loài cá Smelt người ta đo khối lượng của 14 con cá và thu được kết quả như sau :

STT	Y Weight (g)	z_1 Length1 (cm)	z_2 Length2 (cm)	z_3 Length3 (cm)	z_4 Height (cm)	z_5 Width (cm)
1	6.7	9.3	9.8	10.8	1.7388	1.0476
2	7.5	10	10.5	11.6	1.972	1.16
3	7	10.1	10.6	11.6	1.7284	1.1484
4	9.7	10.4	11	12	2.196	1.38
5	9.8	10.7	11.2	12.4	2.0832	1.2772
6	8.7	10.8	11.3	12.6	1.9782	1.2852
7	10	11.3	11.8	13.1	2.2139	1.2838
8	9.9	11.3	11.8	13.1	2.2139	1.1659
9	9.8	11.4	12	13.2	2.2044	1.1484
10	12.2	11.5	12.2	13.4	2.0904	1.3936
11	13.4	11.7	12.4	13.5	2.43	1.269
12	12.2	12.1	13	13.8	2.277	1.2558
13	19.7	13.2	14.3	15.2	2.8728	2.0672
14	19.9	13.8	15	16.2	2.9322	1.8792

Bài giải :

Ta xét mô hình hồi quy tuyến tính cổ điển cho Weight (Cân nặng) của mỗi con cá

$$\mathbf{Y} = \beta_0 + \beta_1 \mathbf{z}_{j1} + \beta_2 \mathbf{z}_{j2} + \beta_3 \mathbf{z}_{j3} + \beta_4 \mathbf{z}_{j4} + \beta_5 \mathbf{z}_{j5} \quad (j = \overline{1, 14})$$

Ta có

$$(\mathbf{Z}^T \mathbf{Z})^{-1} = \begin{bmatrix} 21.18 & -16.47 & 10.12 & 2.97 & 3.88 & -2.72 \\ -16.47 & 29.83 & -13.46 & -12.47 & -1.77 & 5.71 \\ 10.12 & -13.46 & 9.66 & 2.34 & 0.17 & -3.46 \\ 2.97 & -12.47 & 2.34 & 8.54 & 0.18 & -1.63 \\ 3.88 & -1.77 & 0.17 & 0.18 & 6.91 & -2.66 \\ -2.72 & 5.71 & -3.46 & -1.63 & -2.66 & 5.10 \end{bmatrix}$$

$$\hat{\beta} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y} = [-13.190 \quad -4.804 \quad 4.490 \quad 0.944 \quad 3.576 \quad 3.519]^T$$

Tổng bình phương phần dư :

$$\sum_{j=1}^{14} \hat{\varepsilon}_j^2 = \sum_{j=1}^{14} y_j^2 - \mathbf{Y}^T \mathbf{Z} \hat{\beta} = 4.1087$$

Suy ra

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{j=1}^{14} \hat{\varepsilon}_j^2 = \frac{4.1087}{8} = 0.5136$$

Xét giả thuyết: $\beta_3 = \beta_4 = 0$. Với bài toán đang xét và giả thuyết trên ta có $n = 14, k = 5, q = 2$

$$\Rightarrow C = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

$$\Rightarrow ((C\hat{\beta})^T (C(\mathbf{Z}^T \mathbf{Z})^{-1} C^T)^{-1} C\hat{\beta}) / \hat{\sigma}^2 = 16.8705$$

Với độ tin cậy 95% ta có: $(k-q)F_{k-q, n-k-1}(\alpha) = 3.F_{3,8}(0.05) = 12.1985$

Vì $16.8705 > 12.1985$ nên ta bác bỏ giả thuyết $\beta_3 = \beta_4 = 0$ với độ tin cậy 95%

Vậy khối lượng loài cá phụ thuộc ít nhất vào 1 trong 2 yếu tố Length3, Height.

FaMI
1956



FaMI

1956

Ước lượng hàm hồi quy tuyến tính

5.1 Ước lượng hàm hồi quy tại \mathbf{Z}_0

Đặt $\mathbf{Z}_0^T = [1 \ z_{01} \ z_{02} \ \cdots \ z_{0k}]$ là biến dự đoán, ta có Y_0 biểu thị giá trị phản hồi của mô hình hồi quy tại \mathbf{Z}_0 . Khi đó:

$$E(Y_0 | \mathbf{Z}_0) = \beta_0 + \beta_1 z_{01} + \cdots + \beta_k z_{0k} = \mathbf{Z}_0^T \boldsymbol{\beta} \quad (5.1)$$

Theo định lý Gauss, ước lượng bình phương cực tiểu của $E(Y_0 | \mathbf{Z}_0)$ là $\mathbf{Z}_0^T \hat{\boldsymbol{\beta}}$, hơn nữa còn là ước lượng hiệu quả của $E(Y_0 | \mathbf{Z}_0)$ với phương sai cực tiểu:

$$\text{Var}(\mathbf{Z}_0^T \hat{\boldsymbol{\beta}}) = \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 \sigma^2 \quad (5.2)$$

Định lý 5.1

Xét mô hình hồi quy tuyến tính $\mathbf{Y} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ với ma trận \mathbf{Z} đầy hạng, nếu $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2 \mathbf{I})$, khi đó khoảng tin cậy mức $100(1 - \alpha)\%$ của $\mathbf{Z}_0^T \boldsymbol{\beta}$ là

$$\mathbf{Z}_0^T \hat{\boldsymbol{\beta}} \pm t_{n-k-1} \left(\frac{\alpha}{2} \right) \sqrt{\mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 s^2}$$

Chứng minh định lý 5.1

Với \mathbf{Z}_0 cố định, $\mathbf{Z}_0^T \boldsymbol{\beta}$ là một tổ hợp tuyến tính của β_i , nên theo định lý Gauss, ta có:

$$\text{Var}(\mathbf{Z}_0^T \hat{\boldsymbol{\beta}}) = \mathbf{Z}_0^T \text{Cov}(\hat{\boldsymbol{\beta}}) \mathbf{Z}_0 = \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 \sigma^2$$

Khi $\text{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1}$ (bởi công thức 3.14)

Giả định rằng $\boldsymbol{\varepsilon}$ có phân phối chuẩn.

Mệnh đề 4.1 khẳng định $\hat{\boldsymbol{\beta}}$ có phân phối $N_{k+1}(\boldsymbol{\beta}, \sigma^2 (\mathbf{Z}^T \mathbf{Z})^{-1})$ độc lập với s^2/σ^2 và s^2/σ^2 có phân phối $\chi_{n-k-1}^2/(n-k-1)$.

Kết quả, sự kết hợp tuyến tính $\mathbf{Z}_0^T \hat{\boldsymbol{\beta}}$ có phân phối $N(\mathbf{Z}_0^T \boldsymbol{\beta}, \sigma^2 \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0)$

Suy ra $(\mathbf{Z}_0^T \hat{\boldsymbol{\beta}} - \mathbf{Z}_0^T \boldsymbol{\beta}) / \sqrt{\sigma^2 \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0}$ có phân phối $N(0; 1)$.

Chia tỷ lệ này cho $\sqrt{s^2/\sigma^2}$ có phân phối $\sqrt{\chi_{n-k-1}^2/(n-k-1)}$, chúng ta được:

$$\frac{(\mathbf{Z}_0^T \hat{\beta} - \mathbf{Z}_0^T \beta) / \sqrt{\sigma^2 \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0}}{\sqrt{s^2/\sigma^2}} = \frac{(\mathbf{Z}_0^T \hat{\beta} - \mathbf{Z}_0^T \beta)}{\sqrt{s^2 \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0}} \sim t_{n-k-1}$$

Hay:

$$P\left(\frac{|\mathbf{Z}_0^T \hat{\beta} - \mathbf{Z}_0^T \beta|}{\sqrt{s^2 \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0}} < t_{n-k-1} \left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

Nên khoảng tin cậy của $E(Y_0 | \mathbf{Z}_0) = \mathbf{Z}_0^T \beta$ có các mút xác định bởi:

$$\mathbf{Z}_0^T \hat{\beta} \pm t_{n-k-1} \left(\frac{\alpha}{2}\right) \sqrt{\mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 s^2}$$

5.2 Dự đoán một quan sát mới tại \mathbf{Z}_0

Với \mathbf{Z}_0 là vector cho trước, thay vào mô hình hồi quy ta được :

$$Y_0 = \mathbf{Z}_0^T \beta + \varepsilon_0$$

Trong đó $\varepsilon_0 \sim N(0, \sigma^2)$ và độc lập với ε_i , do đó cũng độc lập với $\hat{\beta}$ và s^2 . Các sai số ε_i sẽ bị ảnh hưởng từ các ước lượng $\hat{\beta}$ và s^2 thông qua biến phản hồi Y , nhưng ε_0 thì không.

Định lý 5.2

Xét mô hình hồi quy tuyến tính $\mathbf{Y} = \mathbf{Z}\beta + \varepsilon$ với ma trận \mathbf{Z} đầy hạng, một dự đoán mới có dự đoán không chệch

$$\mathbf{Z}_0^T \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 z_{01} + \dots + \hat{\beta}_k z_{0k}$$

với phương sai của sai số dự báo $Y_0 - \mathbf{Z}_0^T \hat{\beta}$ là

$$\text{Var}(Y_0 - \mathbf{Z}_0^T \hat{\beta}) = \sigma^2 \left(1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0\right)$$

Với ε có phân phối chuẩn, khoảng tin cậy mức $100(1 - \alpha)\%$ của Y_0 được đưa ra là:

$$\mathbf{Z}_0^T \hat{\beta} \pm t_{n-k-1} \left(\frac{\alpha}{2}\right) \sqrt{s^2 \left(1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0\right)} \quad (5.3)$$

Chứng minh định lý 5.2

Chúng ta dự đoán Y_0 bởi $\mathbf{Z}_0^T \hat{\beta}$, ước lượng $E(Y_0 | \mathbf{Z}_0)$.

Theo định lý 5.1, $\mathbf{Z}_0^T \hat{\beta}$ có:

$$E(\mathbf{Z}_0^T \hat{\beta}) = \mathbf{Z}_0^T \beta \quad \text{và} \quad \text{Var}(\mathbf{Z}_0^T \hat{\beta}) = \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 \sigma^2$$

Dự báo sai số là $Y_0 - \mathbf{Z}_0^T \hat{\beta} = \mathbf{Z}_0^T \beta + \varepsilon_0 - \mathbf{Z}_0^T \hat{\beta} = \varepsilon_0 + \mathbf{Z}_0^T (\beta - \hat{\beta})$

Vì vậy, $E(Y_0 - \mathbf{Z}_0^T \hat{\beta}) = E(\varepsilon_0) + E(\mathbf{Z}_0^T (\beta - \hat{\beta})) = 0$

Từ ε_0 và $\hat{\beta}$ độc lập

$$\begin{aligned} \text{Var}(Y_0 - \mathbf{Z}_0^T \hat{\beta}) &= \text{Var}(\varepsilon_0) + \text{Var}(\mathbf{Z}_0^T \hat{\beta}) = \sigma^2 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 \sigma^2 \\ &= \sigma^2 \left(1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0\right) \end{aligned}$$

Nếu giả định thêm rằng ε có phân phối chuẩn, sau đó $\hat{\beta}$ có phân phối chuẩn, và sự kết hợp tuyến tính $Y_0 - \mathbf{Z}_0^T \hat{\beta}$ cũng vậy.

Kết quả là: $(Y_0 - \mathbf{Z}_0^T \hat{\beta}) / \sqrt{\sigma^2 (1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0)}$ có phân phối $N(0, 1)$.

Chia tỷ lệ này cho $\sqrt{s^2 / \sigma^2}$ có phân phối $\sqrt{\chi_{n-k-1}^2 / (n-k-1)}$.

Chúng ta đạt được: $\frac{(Y_0 - \mathbf{Z}_0^T \hat{\beta})}{\sqrt{s^2 (1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0)}}$ có phân phối t_{n-k-1}

Hay:

$$P\left(\frac{|Y_0 - \mathbf{Z}_0^T \hat{\beta}|}{\sqrt{s^2 (1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0)}} < t_{n-k-1} \left(\frac{\alpha}{2}\right)\right) = 1 - \alpha$$

nên ta có khoảng tin cậy của Y_0 được xác định bởi

$$\mathbf{Z}_0^T \hat{\beta} \pm t_{n-k-1} \left(\frac{\alpha}{2}\right) \sqrt{s^2 (1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0)}$$

Ví dụ 5.1

Để nghiên cứu sự phụ thuộc giữa yêu cầu phần cứng máy tính đối với nhu cầu sử dụng của khách hàng từ những đơn đặt hàng đã có, từ đó đưa ra dự báo về các yêu cầu phần cứng máy tính trong tương lai đồng thời có biện pháp quản lí hàng tồn kho. Một nhà khoa học máy tính đưa ra dữ liệu từ 7 trang web khác nhau được thể hiện ở bảng sau:

Bảng 5.1: Tuổi thọ của 7 loại CPU		
Z_1 Số đơn đặt (nghìn)	Z_2 Lượng đơn vị vào-ra (nghìn)	Y Tuổi thọ CPU (giờ)
123.5	2.108	141.5
146.1	9.213	168.9
133.9	1.905	154.8
128.5	0.815	146.5
151.5	1.061	172.8
136.2	8.603	160.1
92.0	1.125	108.5

Xây dựng khoảng tin cậy 95% cho tuổi thọ CPU trung bình và tìm khoảng thời gian dự đoán 95% cho tuổi thọ CPU tại cơ sở mới ứng với $\mathbf{Z}_0 = [1 \ 130 \ 7.5]^T$

Giải ví dụ

Ta xét mô hình hồi quy tuyến tính cổ điển cho tuổi thọ CPU

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \varepsilon_j$$

Ta tính được phương trình hồi quy tuyến tính mẫu:

$$\hat{y} = 8.42 + 1.08z_1 + 0.42z_2$$

để dự đoán hàm hồi quy và dự đoán quan sát mới tại $\mathbf{Z}_0 = [1 \ 130 \ 7.5]^T$

Ta có:

$$\bullet \mathbf{Z} = \begin{bmatrix} 1 & 123.5 & 2.108 \\ 1 & 146.1 & 9.213 \\ 1 & 133.9 & 1.905 \\ 1 & 128.5 & 0.815 \\ 1 & 151.5 & 1.061 \\ 1 & 136.2 & 8.603 \\ 1 & 92.0 & 1.125 \end{bmatrix}$$

$$\bullet (\mathbf{Z}^T \mathbf{Z})^{-1} = \begin{bmatrix} 8.1796 & -0.0641 & 0.0883 \\ -0.0641 & 0.0005 & -0.0010 \\ 0.0883 & -0.0010 & 0.0144 \end{bmatrix}$$

$$\bullet \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 = 0.3699$$

$$\bullet \mathbf{Z}_0^T \hat{\beta} = 151.97$$

$$\bullet s^2 = 1.449$$

- $t_{n-k-1} \left(\frac{\alpha}{2} \right) = t_4(0.025) = 2.776$

Khoảng tin cậy mức 95% cho tuổi thọ CPU trung bình là:

$$\mathbf{Z}_0^T \hat{\beta} \pm t_{n-k-1} \left(\frac{\alpha}{2} \right) \sqrt{\mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 s^2} = 151.97 \pm 2.033$$

Khoảng thời gian dự đoán mức 95% cho tuổi thọ CPU tại cơ sở mới với điều kiện \mathbf{Z}_0 là:

$$\mathbf{Z}_0^T \hat{\beta} \pm t_{n-k-1} \left(\frac{\alpha}{2} \right) \sqrt{s^2 \left(1 + \mathbf{Z}_0^T (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}_0 \right)} = 151.97 \pm 3.912$$

FaMI
1956



FaMI

1956

Kiểm tra sự phù hợp của mô hình

6.1 Một số khái niệm

6.1.1 Chuẩn hóa tập mẫu

Trong các bài toán phân tích dữ liệu thì chuẩn hóa dữ liệu là một phương pháp thường xuyên được sử dụng, đặc biệt là khi dữ liệu của chúng ta có những yếu tố đơn vị như là mét, kilogram,... Các yếu tố này khiến cho dữ liệu mà ta xử lý có những khoảng giá trị khác hẳn nhau, chẳng hạn như giữa biến dự đoán chiều cao $[0,2]$ và biến dự đoán thu nhập $[0,20000000]$ của một người. Để chuẩn hóa dữ liệu, ta đặt:

$$z^* = \frac{z - \mu}{s}$$

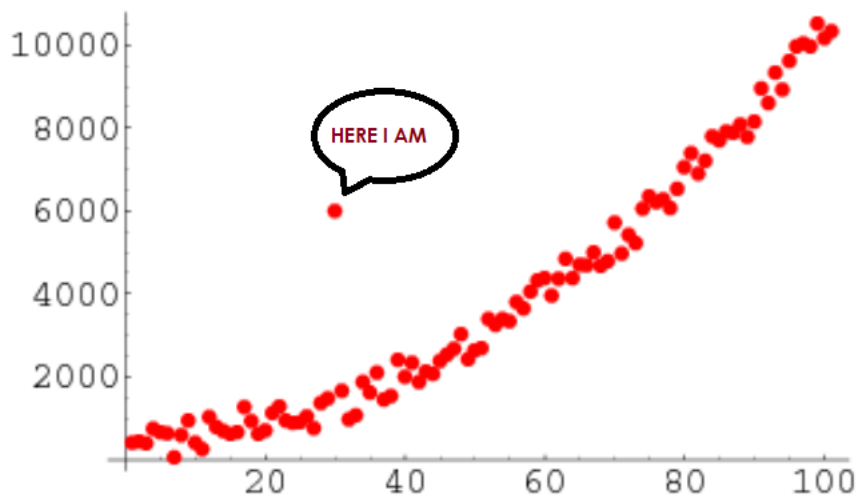
trong đó z là giá trị quan sát, μ là giá trị trung bình mẫu và s là độ lệch chuẩn mẫu.

Việc đặt chuẩn hóa như vậy có tác dụng như sau:

- Đo lường xem từng giá trị z cách xa giá trị trung bình μ là bao nhiêu.
- Khi ta áp dụng công thức này với toàn bộ dữ liệu Z , ta sẽ nhận được mẫu Z^* mới có trung bình là 0 và độ lệch chuẩn là 1.

6.1.2 Outlier

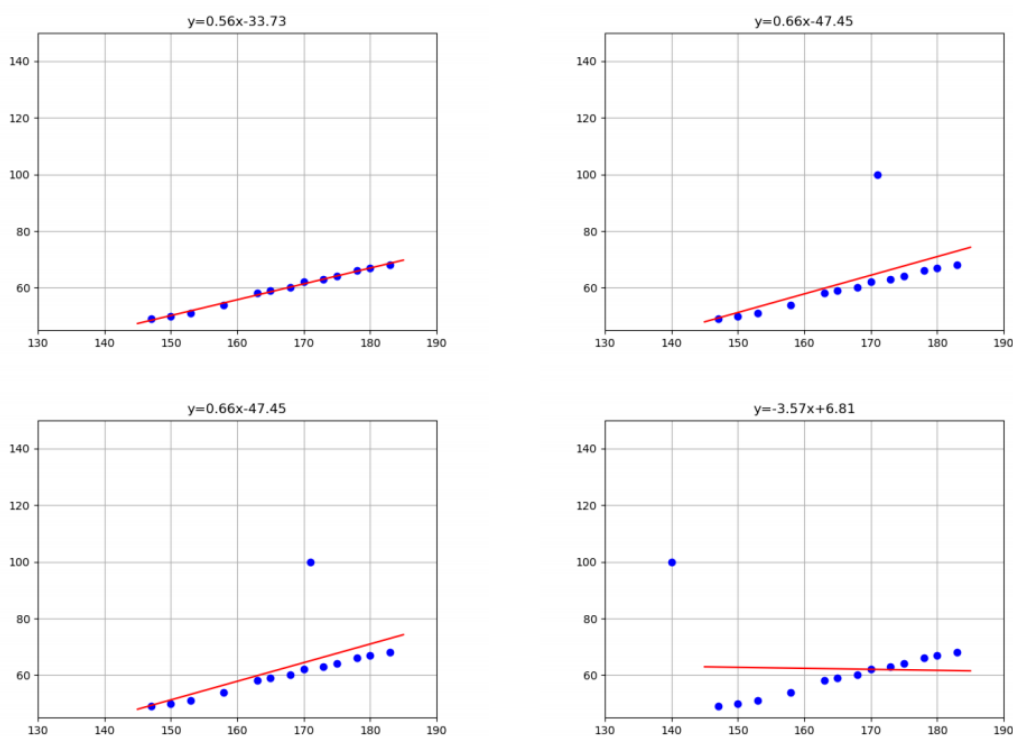
Outlier là điểm dữ liệu bất thường, khác biệt so với phần còn lại của bộ dữ liệu. Như ở ảnh bên dưới, có 1 điểm dữ liệu khác với phần còn lại nên điểm này gọi là điểm outlier.



Hình 6.1: Hình ảnh về điểm outlier

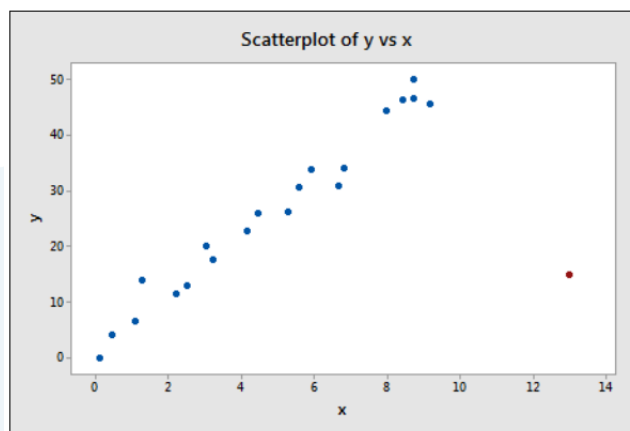
Tác động của điểm outlier đến mô hình hồi quy:

Điểm nào càng xa khỏi bộ dữ liệu thì tác động càng lớn. Trong hình vẽ dưới đây, ở hình 1 và hình 2 ta thấy ở hình 2 điểm outlier sẽ kéo đường hồi quy nhích lên một chút. So sánh hình 3 và 4, cả 2 hình đều có điểm outlier nhưng điểm ở hình 4 cách xa hơn nên nó kéo đường hồi quy xa với điểm dữ liệu hơn nên làm đường hồi quy không tốt.



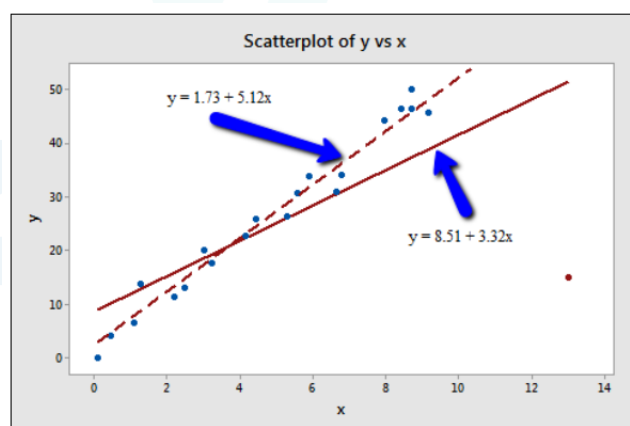
Hình 6.2: Tác động của điểm outlier đến mô hình

Để hiểu rõ hơn về tác động của điểm outlier đối với mô hình thì ta xét mô hình sau:



Hình 6.3: Biểu đồ dữ liệu ban đầu

Theo như hình vẽ trên thì điểm màu đỏ là điểm outlier. Vậy ta sẽ tìm đường hồi quy 2 lần: 1 lần bao gồm điểm outlier và một lần loại trừ điểm outlier.



Hình 6.4: Mô hình chứa điểm outlier và không chứa điểm outlier

Ở hình trên thì đường liền nét biểu thị phương trình hồi quy bao gồm điểm outlier, trong khi đường đứt nét biểu thị phương trình hồi quy loại trừ điểm outlier.

⇒ Ta thấy rằng sự tồn tại của điểm outlier làm giảm đáng kể độ dốc của đường hồi quy, làm ảnh hưởng đến mô hình nên ta cần loại bỏ điểm này.

Ở những hình vẽ trên, điểm outlier được thể hiện một cách trực quan, tuy nhiên trong việc thực nghiệm nhận biết điểm outlier lại không đơn giản như vậy. Đặc biệt là khi ta có hàng trăm nghìn thậm chí hàng triệu điểm dữ liệu thì việc nhận dạng Outlier bằng mắt thường gần như là điều không thể. Trong bài báo cáo này, ta sẽ tìm hiểu và nghiên cứu về 2 phương pháp hỗ trợ công việc này là “Độ đo Leverage” và “Quy tắc 1.5IQR”.

6.1.3 Độ đo Leverage

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. Leverage được xác định bởi công thức như sau:

$$h_{kj} = \frac{1}{n} + \frac{(z_{kj} - \bar{z})^2}{\sum_{j=1}^n (z_{kj} - \bar{z})^2}$$

Những điểm nằm xa so với bộ dữ liệu gọi là những điểm high leverage, ngược lại thì những điểm nằm gần là lower leverage.

Ta thấy rằng, h_{jj} luôn nằm giữa $1/n$ và 1 . Hơn nữa, Leverage trung bình cho các quan sát luôn bằng $(p+1)/n$, p là số biến dự đoán. Do vậy nếu một quan sát mà có độ đo Leverage quá lớn so với $(p+1)/n$, ta có thể nghi ngờ đó là một điểm high leverage.

Nhận xét

- Càng có nhiều điểm dữ liệu nằm gần nhau thì sức ảnh hưởng hay độ lớn của điểm high leverage càng giảm.
- Các điểm high leverage có tác động rất lớn đến mô hình, điểm nào có giá trị của biến dự đoán nằm càng xa khỏi bộ dữ liệu thì tác động càng lớn.
- Những điểm $h_{kj} > 3 * \bar{h}$ thì được coi là những điểm outlier. Với $\bar{h} = \frac{1}{n} \sum_{j=1}^n h_{kj}$

Ví dụ 6.1

Để nghiên cứu sự phụ thuộc giữa doanh thu Y và chi phí sản xuất Z_1 , chi phí tiếp thị Z_2 (đơn vị triệu Dollar) người ta điều tra ngẫu nhiên doanh thu của 12 công ty trong 12 thời kỳ, kết quả ta có bảng sau:

Y (Doanh thu)	Z_1 (Chi phí sản xuất)	Z_2 (Chi phí tiếp thị)
127	18	10
149	25	11
106	19	6
163	24	16
102	15	7
180	26	17
161	25	14
128	16	12
139	17	12
144	23	12
159	22	14
138	15	15

Xét ví dụ 6.1, ta có bảng sau :

j	z_1	h_{1j}	$h_{1j} > 3 * \frac{1}{n} \sum_{j=1}^n h_{1j}$	j	z_2	h_{2j}	$h_{2j} > 3 * \frac{1}{n} \sum_{j=1}^n h_{2j}$
1	18	0,114	0	1	10	0,121	0
2	25	0,192	0	2	11	0,094	0
3	19	0,094	0	3	6	0,391	0
4	24	0,15	0	4	16	0,202	0
5	15	0,235	0	5	7	0,299	0
6	26	0,245	0	6	17	0,272	0
7	25	0,192	0	7	14	0,111	0
8	16	0,184	0	8	12	0,084	0
9	17	0,144	0	9	12	0,084	0
10	23	0,118	0	10	12	0,084	0
11	22	0,096	0	11	14	0,111	0
12	15	0,235	0	12	15	0,148	0

Như vậy trong bộ dữ liệu ở ví dụ 6.1 không có điểm outlier nào.

6.1.4 Quy tắc 1.5IQR

Bất kỳ tập dữ liệu nào cũng có thể mô tả bằng năm giá trị đặc trưng, năm giá trị này cung cấp cho chúng ta thông tin cần thiết để định vị outlier bao gồm (lần lượt theo thứ tự tăng dần). Xét các tứ phân vị:

- Giá trị nhỏ nhất hoặc thấp nhất của tập dữ liệu.
- Tứ phân vị thứ nhất Q_1 : Phân vị mức 25%
- Tứ phân vị thứ hai Q_2 : Phân vị mức 50% (trung vị)
- Tứ phân vị thứ ba Q_3 : Phân vị mức 75%
- Giá trị lớn nhất hoặc cao nhất của tập dữ liệu.

Định nghĩa **khoảng cách tứ phân vị** (Interquartile range) là giá trị:

$$IQR = Q_3 - Q_1$$

Khi đó nếu x không thuộc $[Q_1 - 1.5IQR; Q_3 + 1.5IQR]$ thì x được coi là một điểm outlier.

6.1.5 Giá trị t -statistic và p -value

Để tính được p -value này, ta cần xét cặp giả thuyết và đối thuyết sau với mỗi $i = 1, r$

$$H_0 : \beta_i = 0 \text{ và } H_1 : \beta_i \neq 0$$

Việc bác bỏ hay chấp nhận giả thuyết H_0 được thực hiện thông qua giá trị t -statistic được xác định bởi công thức:

$$t\text{-statistic}(Z_i) = \frac{\hat{\beta}_i}{\sqrt{\widehat{D}(\hat{\beta}_i)}}$$

trong đó β_i là ước lượng hệ số hồi quy và $\widehat{D}(\widehat{\beta}_i)$ là phần tử thứ $i + 1$ trên đường chéo chính của ma trận $\widehat{\sigma}^2(\mathbf{Z}^T \mathbf{Z})^{-1}$, trong đó:

$$\widehat{\sigma}^2 = \frac{1}{n - k - 1} \sum_{j=1}^n \widehat{\varepsilon}_j^2$$

Giá trị p -value là xác suất xảy ra sai lầm loại 1 (bác bỏ H_0 khi H_0 đúng với $H_0: \beta_i = 0$)

$$p\text{-value}(Z_i) = 2(1 - \text{CDF}(n, |t\text{-statistic}(Z_i)|))$$

trong đó CDF là hàm phân phối xác suất của phân phối Student.

Ta rút ra được các hệ quả:

- Nếu có mối liên hệ giữa Z_i và Y , ta kì vọng t -statistic có phân phối Student với $n - k - 1$ bậc tự do.
- Giả thuyết $H_0: \beta_i = 0$ bị bác bỏ khi $p\text{-value} < 5\%$ hoặc $p\text{-value} < 1\%$ tùy mức ý nghĩa mà ta chọn (thường là 5%).

6.2 Kiểm định tính phụ thuộc vào biến của mô hình

Sau khi đã đưa ra được mô hình hoàn chỉnh, một bước nữa không thể thiếu đó là kiểm định sự "quan trọng" của các biến. Điều này bắt nguồn từ việc trong thực tế không phải biến dự đoán nào cũng có đóng góp lớn đến đầu ra của mô hình, không phải yếu tố nào cũng có quan hệ chặt chẽ đối với biến phản hồi.

Vì thế, việc loại bỏ bớt biến là vô cùng cần thiết, một mặt điều này làm đơn giản hóa mô hình, giúp cho các tính toán được thực thi nhanh hơn, mặt khác, nó giúp ta hiểu thêm về các nhân tố quyết định đến đầu ra mà ta mong muốn.

Để kiểm định xem liệu trong số các biến Z_1, Z_2, \dots, Z_k có biến nào là có thể loại khỏi mô hình hồi quy tuyến tính bội, ta tiến hành F -test. Việc này được tiến hành bằng cách giả sử rằng một vài hệ số β_k của chúng ta bằng 0 (đồng nghĩa với việc loại bỏ Z_k) hay ta xét cặp giả thiết và đối thuyết như sau:

- Giả thuyết $H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$
- Đối thuyết $H_1: \nexists \beta_j \neq 0$ với $j = \overline{1, k}$

Xét đại lượng:

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

trong đó:

$$R^2 := \frac{\widehat{Y}^T \widehat{Y} - n(\bar{y})^2}{Y^T Y - n(\bar{y})^2} = \frac{\sum_1^n \widehat{y}_j^2 - n(\bar{y})^2}{\sum_1^n y_j^2 - n(\bar{y})^2}$$

hay còn được biểu diễn theo công thức:

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

Các bước kiểm tra như sau:

1. Tính đại lượng F-score.
2. Tra bảng phân phối Fisher với bậc tự do là k và $n - k - 1$, mức ý nghĩa α .
3. Nếu $F > F_{k, n-k-1}(\alpha)$ thì ta bác bỏ H_0 và chấp nhận H_1 .

Xét lại ví dụ 6.1:

Ta đi sẽ kiểm tra xem mô hình có phụ thuộc vào biến hay không với $\alpha = 0,01$.

Bảng tính các giá trị $\hat{y}_j, \hat{\varepsilon}_j$:

STT	y_j	\hat{y}_j	$\hat{\varepsilon}_j$	STT	y_j	\hat{y}_j	$\hat{\varepsilon}_j$
1	127	124,9666	2,033	7	161	161,5420	-0,542
2	149	147,2659	1,734	8	128	129,4733	-1,473
3	106	108,4382	-2,438	9	139	131,979	7,021
4	163	168,5537	-5,554	10	144	147,0132	-3,013
5	102	103,1741	-1,174	11	159	154,0249	4,975
6	180	178,3238	1,676	12	138	141,2437	-3,244

Ta có:

$$\hat{\varepsilon}_j^T \hat{\varepsilon}_j = \sum_{j=1}^n \hat{\varepsilon}_j^2 = 144,3734; \quad s_y^2 = \bar{y}^2 - (\bar{y})^2 = \frac{245626}{12} - \left(\frac{1696}{12} \right)^2 = 493,7222$$

$$\text{Vậy } 144,3734 = 12 \times 493,7222(1 - R^2) \rightarrow (1 - R^2) = 0,0224 \rightarrow R = 0,9756$$

$$F = \frac{0,9756 \times (12 - 2 - 1)}{2 \times 0,0244} = 179,6292$$

Tra bảng Fisher ta được:

$$F_{2,9}(0,01) = 8,02$$

Ta thấy $F > F_{2,9}(0,01)$, do đó ta cần bác bỏ giả thiết rằng $\beta_1 = \dots = \beta_k = 0$, tức là có sự phụ thuộc tuyến tính vào các biến độc lập.

6.3 Kiểm tra tính đa cộng tuyến của các biến dự đoán

Sự đa cộng tuyến giữa các biến dự đoán ám chỉ trường hợp mà ở đó, hai hay nhiều biến dự đoán có sự ràng buộc với nhau. Hệ quả điển hình của hiện tượng này đó là các phần tử trên đường chéo chính của ma trận $\mathbf{Z}^T \mathbf{Z}$ sẽ rất lớn, dẫn đến các ước lượng $\hat{\beta}_i$ lớn theo và gây cản trở trong việc xác định độ "quan trọng" (significant) của các biến dự đoán.

Cách nhận biết hiện tượng này:

- Một số phần tử trên đường chéo chính của ma trận $\mathbf{Z}^T \mathbf{Z}$ tỏ ra rất lớn.

- Các hệ số tương quan tuyến tính mẫu của các cặp Z_i, Z_j là $r_{ij} = s_{ij} / \sqrt{s_{jj}s_{ii}}$ tỏ ra lớn ($|r_{ij}| \geq 0,7$), trong đó:

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^n (z_{ki} - \bar{z}_i) \times (z_{kj} - \bar{z}_j)$$

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

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

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

$$\text{trong đó: } s_{00} = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y}) \times (y_j - \bar{y}); s_{0i} = \frac{1}{n-1} \sum_{j=1}^n (y_j - \bar{y}) \times (z_{ji} - \bar{z}_i)$$

Khi đó, nếu thấy $|r_{ij}| \geq 0,7$ thì:

- Ta loại biến Z_i ra khỏi mô hình nếu $|r_{0i}| < |r_{0j}|$.
- Ta loại biến Z_j ra khỏi mô hình nếu $|r_{0i}| > |r_{0j}|$.

Bản chất của bước này chính là ta giữ lại biến có độ tương quan với biến phụ thuộc Y cao hơn và loại bỏ biến có độ tương quan với Y thấp hơn.

3. Thực hiện hồi quy sau khi ma trận Z đã loại bỏ biến Z_i hay Z_j .

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

$$\mathbf{Z} = \begin{bmatrix} 1 & 1.9999 & 2.8889 \\ 1 & 4.0001 & 4.9994 \\ 1 & 0.0009 & 2.0003 \end{bmatrix}, \mathbf{Y} = [8, 14, 2]^T$$

Ta tiến hành làm như sau:

- Bước 1: Tìm các $|r_{ij}| \geq 0,7$. Ta tính được ma trận hiệp phương sai và ma trận hệ số tương quan mẫu là:

$$\mathbf{S} = \begin{bmatrix} 3.9984 & 2.9986 \\ 2.9986 & 2.3731 \end{bmatrix}, \mathbf{R} = \begin{bmatrix} 1 & 0.9734 \\ 0.9734 & 1 \end{bmatrix}$$

- Bước 2: Tính các r_{0i}

$$s_{00} = 36$$

$$s_{01} = 11.9976, \quad s_{02} = 8.9973$$

$$\Rightarrow r_{01} = s_{01} / \sqrt{s_{11}s_{00}} = 0.9999 \text{ và } r_{02} = s_{02} / \sqrt{s_{22}s_{00}} = 0.9734$$

Ta thấy: $|r_{01}| > |r_{02}|$ nên ta sẽ loại biến Z_2 ra khỏi mô hình.

- Bước 3: Ta thực hiện hồi quy sau khi đã loại bỏ biến Z_2 .

6.4 Khảo sát phần dư

Một giả thiết quan trọng của mô hình hồi quy tuyến tính đó là nhiễu ε có phân phối chuẩn. Trên thực tế, ta không thể biết được sai số của phép đo có phân phối chuẩn hay không nên ta phải đưa vào bài toán kiểm định.

Ta sẽ đưa ra tiêu chuẩn Student để chấp nhận hay bác bỏ giả thuyết:

$$H_0 : \varepsilon \sim \mathcal{N}_n(0, \sigma^2 I)$$

hay

$$H_0 : \hat{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2(I - H))$$

Định lý 6.1

Xét mô hình hồi quy tuyến tính $\mathbf{Y} = \mathbf{Z}\beta + \varepsilon$ với ma trận \mathbf{Z} đầy hạng. Khi đó giả thuyết $H_0 : \varepsilon \sim \mathcal{N}_n(0, \sigma^2 \mathbf{I})$ bị bác bỏ với mức ý nghĩa $100(1 - \alpha)\%$ nếu

$$\sqrt{n-r-2} \left(\sum_{i=1}^n \mathbf{v}_i \right)^T \hat{\varepsilon} > t_{n-r-2} \left(\frac{\alpha}{2} \right) \sqrt{\left| n \left(\sum_{i=1}^n \mathbf{v}_i \right)^T \hat{\varepsilon} - (n-r-1) \hat{\varepsilon}^T \hat{\varepsilon} \right|}$$

trong đó $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ là các vector riêng trực chuẩn của ma trận $\mathbf{I} - \mathbf{H}$

Chứng minh:

Do ma trận $\mathbf{I} - \mathbf{H}$ là ma trận lũy đẳng nên nó chỉ có các trị riêng là 0 hoặc 1. Viết ma trận $\mathbf{I} - \mathbf{H}$ dưới dạng chéo hóa, ta được:

$$\mathbf{I} - \mathbf{H} = \mathbf{P}\Lambda\mathbf{P}^T$$

trong đó $\mathbf{P} = [\mathbf{v}_1 : \mathbf{v}_2 : \dots : \mathbf{v}_n]$ là ma trận trực giao gồm các vector riêng trực chuẩn, với:

$$\Lambda = \text{diag}\{\underbrace{1, \dots, 1}_{n-r-1}, \underbrace{0, \dots, 0}_{r+1}\}$$

Đặt

$$\mathbf{e} = [e_1 \ e_2 \ \dots \ e_n]^T = \mathbf{P}^T \hat{\varepsilon}$$

Nếu giả thuyết H_0 đúng thì:

$$E(\mathbf{e}) = \mathbf{P}^T E(\hat{\varepsilon}) = 0$$

$$\text{Cov}(\mathbf{e}) = \mathbf{P}^T \text{Cov}(\sigma^2(\mathbf{I} - \mathbf{H}))\mathbf{P} = \sigma^2 \mathbf{P}^T \mathbf{P} \Lambda \mathbf{P}^T \mathbf{P} = \sigma^2 \Lambda$$

Không những vậy, $e_1, e_2, \dots, e_{n-r-1}$ sẽ là các biến ngẫu nhiên có phân phối chuẩn và $e_{n-r} = e_{n-r+1} =$

$\dots = e_n = 0$. Khi đó:

$$\sum_{i=1}^n e_i = \sum_{i=1}^{n-r-1} e_i \quad \text{và} \quad \sum_{i=1}^n e_i^2 = \sum_{i=1}^{n-r-1} e_i^2 \quad (6.1)$$

Xét các vector ngẫu nhiên

$$\bar{e} = \frac{1}{n} \sum_{i=1}^n e_i \quad \text{và} \quad \tilde{e} = \frac{1}{n-r-1} \sum_{i=1}^{n-r-1} e_i$$

Nếu giả thuyết H_0 đúng thì ta sẽ có

$$n\bar{e} = (n-r-1)\tilde{e} \quad (6.2)$$

Hơn nữa, ta còn có biến đổi

$$\begin{aligned} \sum_{i=1}^{n-r-1} (e_i - \tilde{e})^2 &= \sum_{i=1}^{n-r-1} e_i^2 + (n-r-1)\tilde{e}^2 - 2 \sum_{i=1}^{n-r-1} e_i \tilde{e} \\ &= \sum_{i=1}^{n-r-1} e_i^2 + (n-r-1)\tilde{e}^2 - 2 \sum_{i=1}^n e_i \tilde{e} \\ &= \sum_{i=1}^{n-r-1} e_i^2 + (n-r-1)\tilde{e}^2 - 2\tilde{e}(n\bar{e}) \\ &= \sum_{i=1}^{n-r-1} e_i^2 - (n-r-1)\tilde{e}^2 \end{aligned} \quad (6.3)$$

Xét thống kê

$$T = n\bar{e} \sqrt{\frac{n-r-2}{n^2\bar{e}^2 - (n-r-1)\sum_{i=1}^n e_i^2}} \quad (6.4)$$

Nếu giả thiết H_0 đúng thì theo các công thức (6.1), (6.2) và (6.3), ta có

$$\begin{aligned}
 T &= \left(\sum_{i=1}^n e_i \right) \sqrt{\frac{n-r-2}{\left| (n-r-1)^2 \bar{e}^2 - (n-r-1) \sum_{i=1}^n e_i^2 \right|}} \\
 &= \left(\sum_{i=1}^{n-r-1} e_i \right) \frac{1}{\sqrt{n-r-1}} \sqrt{\frac{n-r-2}{\sum_{i=1}^n e_i^2 - (n-r-1) \bar{e}^2}} \\
 &= \left(\sum_{i=1}^{n-r-1} e_i \right) \frac{1}{\sqrt{n-r-1}} \sqrt{\frac{n-r-2}{\sum_{i=1}^{n-r-1} (e_i - \bar{e})^2}} \\
 &= \frac{\frac{1}{n-r-1} \sum_{i=1}^{n-r-1} e_i}{\sqrt{\frac{1}{n-r-2} \sum_{i=1}^{n-r-1} (e_i - \bar{e})^2}} \sqrt{n-r-1} \sim t_{n-r-2}
 \end{aligned}$$

Như vậy tiêu chuẩn để bác bỏ giả thuyết H_0 với mức ý nghĩa $100(1-\alpha)\%$ là

$$|T| > t_{n-r-2} \left(\frac{\alpha}{2} \right)$$

hay viết lại thành

$$n\bar{e}\sqrt{n-r-2} > t_{n-r-2} \left(\frac{\alpha}{2} \right) \sqrt{\left| n^2 \bar{e}^2 - (n-r-1) \sum_{i=1}^n e_i^2 \right|} \quad (6.5)$$

Chú ý rằng

$$\begin{aligned}
 n\bar{e} &= \sum_{i=1}^n e_i = \mathbf{1}^T \mathbf{e} = \mathbf{1}^T \mathbf{P}^T \hat{\mathbf{e}} = (\mathbf{P}\mathbf{1})^T \hat{\mathbf{e}} = \left(\sum_{i=1}^n \mathbf{v}_i \right)^T \hat{\mathbf{e}} \\
 \sum_{i=1}^n e_i^2 &= \mathbf{e}^T \mathbf{e} = \hat{\mathbf{e}}^T \mathbf{P} \mathbf{P}^T \hat{\mathbf{e}} = \hat{\mathbf{e}}^T \hat{\mathbf{e}}
 \end{aligned}$$

Thay vào tiêu chuẩn ở công thức (6.5), ta được điều phải chứng minh.

Ta có một số lưu ý ở định lý 6.1:

1. Quy tắc kiểm định đề ra ở Định lý 6.1 không phụ thuộc vào cách chọn hệ vector riêng trực chuẩn của ma trận $\mathbf{I} - \mathbf{H}$ (lưu ý rằng ma trận $\mathbf{I} - \mathbf{H}$ có trị riêng bội, cụ thể là trị riêng bằng 1 với bội $n-r-1$ và trị riêng bằng 0 với bội $r+1$ nên vector riêng trực chuẩn ứng với từng trị riêng sẽ không duy nhất)

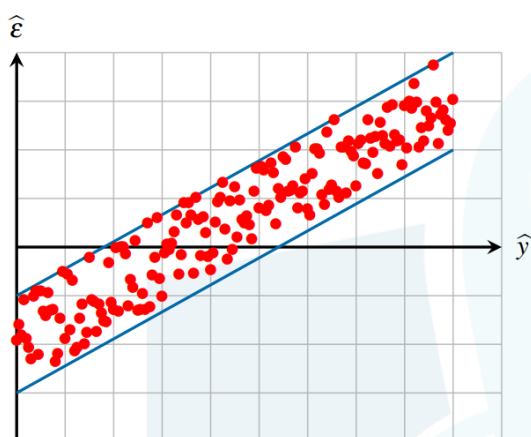
2. Thống kê T ở công thức (6.4) chỉ xác định khi $n \geq r + 2$. Nhưng trong thực tế, thường giá trị n (là số quan sát) sẽ lớn hơn rất nhiều so với $r + 1$ (là số biến dự đoán) nên ràng buộc này hầu như luôn thỏa mãn.

6.4.1 Khảo sát đồ thị của các phần dư

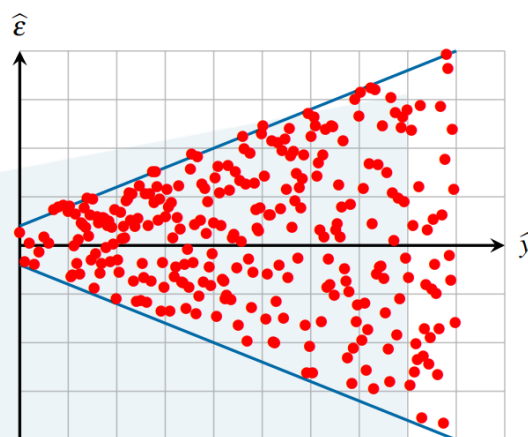
Các mô hình hồi quy tuyến tính đều được xây dựng trên giả thuyết giữa biến dự đoán và biến phản hồi tồn tại một quan hệ tuyến tính. Tuy nhiên, nếu độ chính xác trong dự đoán mô hình giảm đi nghiêm trọng thì điều ở trên không xảy ra dẫn đến các suy luận của chúng ta đều vô nghĩa. Vì thế, đồ thị phần dư là một công cụ hữu ích cho việc phát hiện mối quan hệ phi tuyến tính. Đối với những mô hình hồi quy tuyến tính đơn, ta vẽ đồ thị phân tán giữa sai số ε và biến dự đoán z còn đối với mô hình hồi quy tuyến tính bội, khi số lượng biến dự đoán quá nhiều, ta sẽ vẽ đồ thị phân tán giữa phần dư $\hat{\varepsilon}_j$ và giá trị ước lượng \hat{y} .

Khi tiêu chuẩn Student đưa đến việc bác bỏ giả thiết ε tuân theo phân bố chuẩn $\mathcal{N}_n(0, \sigma^2 \mathbf{I})$, đồ thị phần dư có thể xảy ra các khả năng sau:

- **Hình (a): Đồ thị phần dư $\hat{\varepsilon}_j$ phụ thuộc vào biến \hat{y}_j .** Nguyên nhân có thể là do có sự sai sót trong bước tính toán hoặc thành phần β_0 đã bị loại khỏi mô hình.
- **Hình (b): Đồ thị phần dư với phương sai biến thiên.** Đồ thị có dạng hình phễu (nón), đồ thị này có thể xảy ra do phương sai lớn với \hat{y} lớn và ngược lại. Trong trường hợp này, ta có thể áp dụng một vài biến đổi với đầu ra hoặc sử dụng phương pháp bình phương cực tiểu có trọng số đã được đề cập ở Chương 3.

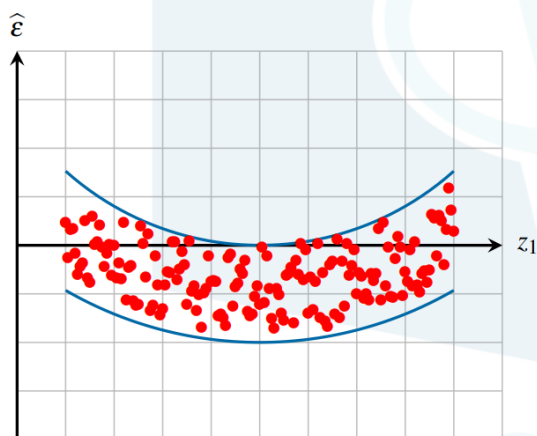


(a) Đồ thị phần dư phụ thuộc vào biến \hat{y}_i

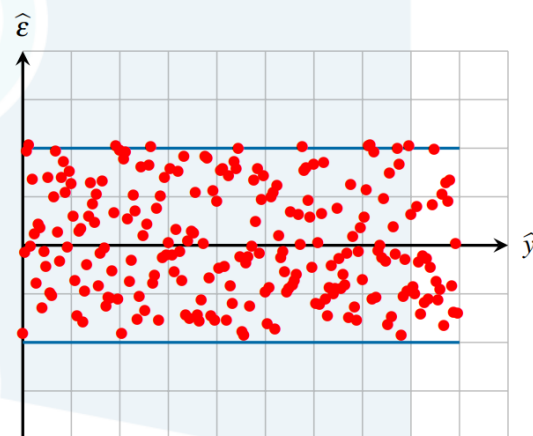


(b) Đồ thị phần dư với phương sai biến thiên

- **Hình (c): Đồ thị phần dư với tích các biến dự đoán.** Mô hình dự đoán bỏ sót biến dự đoán z_j . Đồ thị này biểu diễn lũy thừa của một biến dự đoán hoặc tích của nhiều biến dự đoán với nhau như z_1^2 hay $z_1 z_2$.
- **Hình (d): Đồ thị phần dư lý tưởng.** Mô hình hồi quy của ta sẽ tốt nhất khi đồ thị phần dư ở dạng này, vì phương sai không đổi (do đồ thị có dạng một dải nằm ngang).



(c) Đồ thị phần dư với tích các biến dự đoán



(d) Đồ thị phần dư lý tưởng

6.4.2 Khảo sát đồ thị phần dư Q-Q

Ta có thể dùng đồ thị Q-Q để kiểm tra tính phân phối chuẩn của phần dư. Phần dư tuân theo phân phối chuẩn khi đồ thị Q-Q có quan hệ tuyến tính (đường thẳng). Tuy nhiên, việc chỉ đánh giá dựa trên đồ thị chỉ cho ta cái nhìn trực quan, chưa mang tính lý thuyết. Vậy nên ta có thể dựa vào hệ số tương quan sau:

$$r = \frac{\sum_{j=1}^n (q_j - \bar{q})(\hat{\varepsilon}_j - \bar{\hat{\varepsilon}})}{\left\{ \sum_{j=1}^n (q_j - \bar{q})^2 \sum_{j=1}^n (\hat{\varepsilon}_j - \bar{\hat{\varepsilon}})^2 \right\}^{1/2}}$$

Lưu ý:

- q_j là nghiệm của phương trình $\Phi(q_j) = \frac{(j - 0,5)}{n}$.
- $\hat{\varepsilon}_j$ phải được sắp xếp theo thứ tự tăng dần.
- Đặt ngưỡng cần kiểm tra với mức ý nghĩa α là k .
- Nếu $r > k$ thì phần dư tuân theo phân phối chuẩn. Ngược lại thì phần dư không tuân theo phân phối chuẩn.

Ví dụ:

Xét lại ví dụ 6.1, ta có bảng sau:

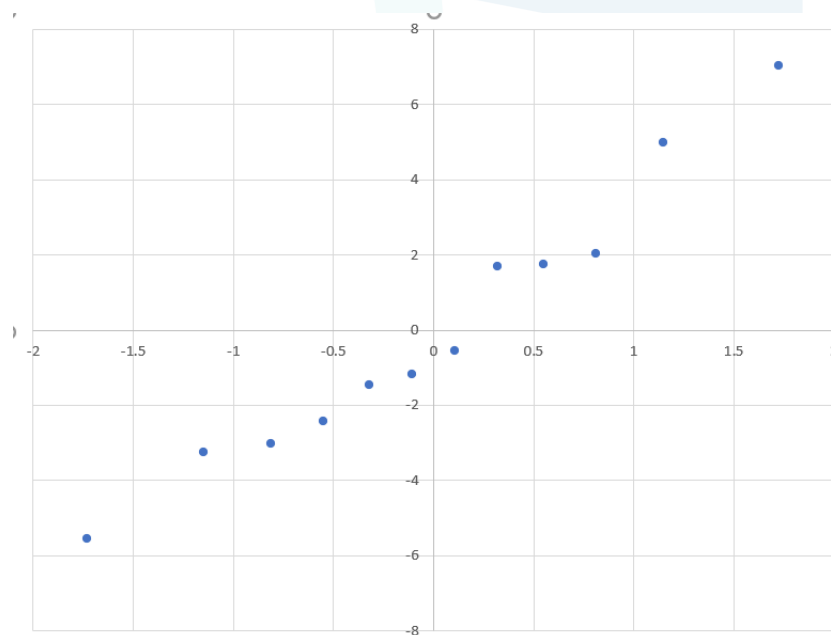
j	$(j-1/2)/n$	q_j	$\hat{\varepsilon}_j$	j	$(j-1/2)/n$	q_j	$\hat{\varepsilon}_j$
1	0,042	-1,728	-5,554	7	0,542	0,105	-0,542
2	0,125	-1,15	-3,244	8	0,625	0,319	1,676
3	0,208	-0,813	-3,013	9	0,708	0,548	1,734
4	0,292	-0,548	-2,438	10	0,792	0,813	2,033
5	0,375	-0,319	-1,473	11	0,875	1,15	4,975
6	0,458	-0,105	-1,174	12	0,958	1,728	7,021

Ta tính được $r = \frac{\sum_{j=1}^n (q_j - \bar{q})(\hat{\varepsilon}_j - \bar{\varepsilon})}{\left\{ \sum_{j=1}^n (q_j - \bar{q})^2 \sum_{j=1}^n (\hat{\varepsilon}_j - \bar{\varepsilon})^2 \right\}^{1/2}} = 0,981$

Với $\alpha = 0,01$, $n = 12$, tra bảng phân vị chuẩn tắc, ta có: $k = 0,708$.

Vì $r > k \Rightarrow$ Phần dư tuân theo phân phối chuẩn.

Đồ thị Q-Q của bài toán trên được biểu diễn ở hình dưới đây. Có thể thấy các điểm dữ liệu có quan hệ tuyến tính nên phần dư tuân theo phân phối chuẩn.

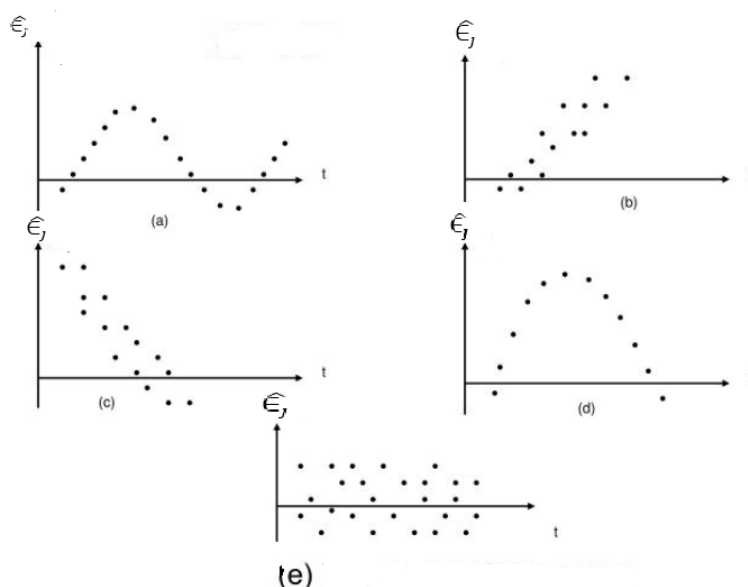


Hình 6.5: Đồ thị Q - Q ứng với bài toán trên

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

6.5.1 Kiểm tra bằng đồ thị

Ta thấy 4 hình a, b, c, d các giá trị $\hat{\varepsilon}_j$ đều có phân phối theo 1 trình tự, phân phối đồ thị nào đó nên nó tương quan với nhau. Còn ở hình e không có hiện tượng tự tương quan vì các giá trị $\hat{\varepsilon}_j$ xáo trộn 1 cách ngẫu nhiên không theo 1 trình tự nào.



Hình 6.6: Đồ thị kiểm tra tính tương quan theo thời gian

Nhận xét: Đối với bộ dữ liệu lớn, việc nhìn trực quan bằng đồ thị sẽ khá khó khăn để áp dụng với bộ dữ liệu bất kỳ.

6.5.2 Kiểm tra bằng phương pháp Durbin - Watson

Giả sử Y_j được theo dõi theo thời gian $j = 1, 2, \dots$. Khi đó, thường xảy ra các trường hợp ε_j có tương quan với nhau. Để khắc phục tình trạng này, ta sử dụng **tiêu chuẩn Durbin - Watson**. Đây là tiêu chuẩn thống kê được đưa ra để đánh giá, kiểm định sự tự tương quan trong các phần dư của mô hình hồi quy.

Trị số thống kê Durbin - Watson (kí hiệu là DW) nhận giá trị từ 0 đến 4. Giá trị $DW = 2$ cho biết trong mô hình không có sự tự tương quan giữa các phần dư. Giá trị $0 < DW < 2$ cho biết có sự tự tương quan dương. Giá trị $2 < DW < 4$ cho biết có sự tự tương quan âm giữa các phần dư.

Để phát hiện tính tự tương quan của các phần dư $\hat{\varepsilon}_j$ ta có thể sử dụng tiêu chuẩn Durbin - Watson. Trước hết, đặt:

$$r_1 = \frac{\sum_{j=2}^n \hat{\varepsilon}_j \hat{\varepsilon}_{j-1}}{\sum_{j=1}^n \hat{\varepsilon}_j^2}$$

Xét đại lượng:

$$DW = \frac{\sum_{j=2}^n (\hat{\varepsilon}_j - \hat{\varepsilon}_{j-1})^2}{\sum_{j=1}^n \hat{\varepsilon}_j^2} = 2(1 - r_1)$$

Khi đó giá trị DW sẽ tuân theo **phân phối Durbin - Watson**. Tra bảng Durbin - Watson ứng với mức ý nghĩa α ta tìm được 2 chỉ số $d_1(k, n, \alpha) < d_2(k, n, \alpha)$, với n là số điểm của bộ dữ liệu, k là số

biến, α là mức ý nghĩa.

Khi đó so sánh DW với d_1, d_2 ta có các kết luận sau:

- Nếu $0 \leq DW < d_1$ thì các $\hat{\varepsilon}_j$ có tự tương quan dương.
- Nếu $d_1 \leq DW \leq d_2$ thì không thể kết luận được.
- Nếu $d_2 < DW < 4 - d_2$ thì các $\hat{\varepsilon}_j$ không có tự tương quan bậc nhất.
- 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 $\hat{\varepsilon}_j$ có tự tương quan âm.

Lưu ý: Phương pháp Durbin - Watson chỉ giúp xác nhận các tự tương quan bậc 1, tức là sự tự tương quan giữa 2 biến liên tiếp. Đây cũng là một mặt hạn chế khá lớn của phương pháp này.

Ví dụ:

Xét lại ví dụ 6.1, ta có bảng sau:

STT	y_j	\hat{y}_j	$\hat{\varepsilon}_j$	STT	y_j	\hat{y}_j	$\hat{\varepsilon}_j$
1	127	124,9666	2,033	7	161	161,5420	-0,542
2	149	147,2659	1,734	8	128	129,4733	-1,473
3	106	108,4382	-2,438	9	139	131,979	7,021
4	163	168,5537	-5,554	10	144	147,0132	-3,013
5	102	103,1741	-1,174	11	159	154,0249	4,975
6	180	178,3238	1,676	12	138	141,2437	-3,244

- Ta có $r_1 = \frac{\sum_{j=2}^n \hat{\varepsilon}_j \hat{\varepsilon}_{j-1}}{\sum_{j=1}^n \hat{\varepsilon}_j^2} = \frac{-45,3437}{144,2298} = -0,3144$ và $DW = 2(1 + 0,3144) = 2,6288$
- Với $\alpha = 0,05; n = 12; k = 2$, ta tra bảng phân phối Durbin - Watson:
 $d_1 = 0,812; d_2 = 1,579$
- Như vậy, ta có: $4 - d_2 < DW < 4 - d_1$

=> Ta chưa thể đưa ra kết luận về tính tự tương quan của các phần dư.

6.6 Xác định các biến quan trọng

Như đã nói ở trên, một bước không thể thiếu sau khi đưa ra được mô hình hồi quy đó là xác định tính quan trọng của các biến dự đoán thông qua F -test để đưa ra được p -value cho từng hệ số tương ứng với biến. Sau đó ta dựa vào mức ý nghĩa (thường là 5%) để loại bỏ các biến (p -value > 5%). Tất nhiên, có thể sẽ xảy ra tình huống mà ở đó tất cả biến dự đoán đều quan trọng đối với biến phản hồi nhưng trong thực tế điều này hiếm khi xảy ra nếu không muốn nói là không thể. Thay vào đó, biến phản hồi thường chỉ có tương quan chặt chẽ với một tập con các biến dự đoán. Do vậy, cần tìm được một tập con phù hợp nhất trong các biến dự đoán sao cho vẫn đảm bảo được độ tin cậy của mô hình.

Lý tưởng mà nói, ta luôn muốn chọn được tập con phù hợp bằng cách thử tất cả các mô hình với các tập con có thể có trong số p biến dự đoán. Vậy làm cách nào để quyết định xem mô hình con nào

tốt hơn?

Một vài thống kê có thể được đưa ra để đánh giá chất lượng của mô hình, các thống kê phổ biến thường được sử dụng đó là Mallows's C_p , AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) và giá trị R^2 hiệu chỉnh (\bar{R}^2) đã được đề cập ở mục trên. Với p là số biến được sử dụng, xét $\beta_{(2)}$ và $\mathbf{Z}_{(1)}$ là vector và ma trận gồm các giá trị quan sát của tương ứng. Dưới đây là công thức cho các thống kê được nêu trên:

- Giá trị R^2 hiệu chỉnh

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

- (Mallows's C_p) Chọn $\beta_{(2)}$ sao cho $C_p \approx p$

$$C_p = \frac{\text{RSS}(\mathbf{Z}_{(1)})}{\text{RSS}(\mathbf{Z})} + 2p - n$$

- (Akaike Information Criterion) Chọn $\beta_{(2)}$ để AIC nhỏ nhất

$$\text{AIC} = 2p + n \left(1 + \ln \frac{2\pi \text{RSS}(\mathbf{Z}_{(1)})}{n} \right)$$

- (Bayesian Information Criterion) Chọn $\beta_{(2)}$ để BIC nhỏ nhất

$$\text{BIC} = n \ln \frac{\text{RSS}(\mathbf{Z}_{(1)})}{n} + p \ln n$$

Tổng quát, với mô hình cho trước p biến dự đoán, ta tiến hành các phương pháp kiểm tra và chọn ra mô hình tốt nhất. Ta đưa ra 3 phương pháp như sau:

Chọn tiến dần.

1. Gọi \mathcal{M}_0 là mô hình *null* (mô hình không chứa biến dự đoán).
2. Với $k = 0, 1, \dots, p - 1$
 - Xét tất cả $p - k$ mô hình được tạo bằng cách lấy các biến phản hồi của \mathcal{M}_k và thêm vào một biến không có trong \mathcal{M}_k .
 - Chọn mô hình tốt nhất trong $p - k$ mô hình vừa tạo và gọi mô hình đó là \mathcal{M}_{k+1} . Tốt nhất ở đây là có giá trị RSS cao nhất hoặc R^2 cao nhất.
3. Chọn mô hình tốt nhất trong các mô hình $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$.

Chọn lùi dần.

1. Gọi \mathcal{M}_p là mô hình *full* (mô hình chứa tất cả các biến dự đoán).
2. Với $k = p, p - 1, \dots, 1$
 - Xét tất cả k mô hình được tạo bằng cách loại đi một biến trong mô hình \mathcal{M}_k .

- Chọn mô hình "tốt nhất" trong k mô hình vừa tạo và gọi mô hình đó là \mathcal{M}_{k-1} . Tốt nhất ở đây là có giá trị RSS cao nhất hoặc R^2 cao nhất.

3. Chọn mô hình tốt nhất trong các mô hình $\mathcal{M}_0, \mathcal{M}_1, \dots, \mathcal{M}_p$.

Chọn hỗn hợp. (Kết hợp giữa chọn tiến dần và chọn lùi dần)

1. Gọi \mathcal{M}_0 là mô hình *null* (mô hình không chứa biến dự đoán).

2. Với $k = 0, 1, \dots, p$

- Xét tất cả $p - k$ mô hình được tạo bằng cách lấy các biến phản hồi của \mathcal{M}_k và thêm vào một biến không có trong \mathcal{M}_k .
- Chọn mô hình tốt nhất trong $p - k$ mô hình vừa tạo và gọi mô hình đó là \mathcal{M}_{k+1} . Tốt nhất ở đây là có giá trị RSS cao nhất hoặc R^2 cao nhất.
- Sau khi chọn được mô hình tốt nhất, ta tiến hành kiểm tra xem có thể loại được biến nào trong mô hình đó hay không. Biến loại đi sẽ không được thêm lại vào mô hình ở các vòng lặp tiếp theo.

FaMI
1956

Phần code chương trình thực hành với dữ liệu cụ thể

Từ cơ sở kiến thức lý thuyết các phần ở trên, nhóm em áp dụng cho dữ liệu là giá nhà thực tế để áp dụng dự đoán cho tương lai. Dữ liệu được lấy trên Kaggle và đã được tiền xử lý để làm sạch data.

7.1 Mô tả dữ liệu

Tên file gốc: `kc_house_data.csv`

Tên file data định dạng: `Data_House_Main.xlsx`

Với 10 cột dữ liệu và 10000 dòng, trong đó bao gồm các cột:

Date: Thời gian (Từ ngày 2/5/2014 đến ngày 2/10/2014) (object)

Dữ liệu được mô tả với 9 cột như sau:

- **Price:** Giá nhà (float64)
- **Bedrooms:** số phòng ngủ (int64)
- **Bathrooms:** số phòng tắm (float64)
- **Area:** diện tích (float64)
- **Floors:** Tầng (int64)
- **Grade:** Hạng mục nhà (int64)
- **yr_built:** Năm ngôi nhà được xây (int64)
- **latitude:** vĩ độ (float64)
- **longitude:** kinh độ (float64)

Thực hiện chọn mô hình hồi quy cổ điển:

$$Y = \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \dots + \beta_r z_r + \varepsilon$$

Trong đó:

- Biến phản hồi Y: Price (giá nhà)
- Các cột còn lại sẽ là biến dự đoán.

7.2 Các bước xây dựng một mô hình hồi quy hoàn chỉnh

Thông qua những phương pháp mà ta sử dụng ở bên trên, để xây dựng một mô hình hồi quy tuyến tính hoàn chỉnh, ta làm theo các bước sau:

- **B1: Tiền xử lý dữ liệu**

- ◇ Đọc file excel bằng thư viện pandas sau đó thực hiện asarray để dữ liệu trở thành các ma trận 2 chiều.
- ◇ Vẽ đồ thị kiểm tra sự phân bố giá trị của từng biến Z so với giá trị Y
- ◇ Chuẩn hóa tập mẫu theo chuẩn Z-score $Z = \frac{x - \mu}{\sigma}$
- ◇ Lọc bỏ outlier bằng Leverage (với mức đánh giá $3p/n$ trong đó p là số biến tham gia dự đoán)
- ◇ Khảo sát đa cộng tuyến tính

- **B2: Xác định các tiêu chuẩn và biến quan trọng**

- ◇ Ước lượng tham số hồi quy $\hat{\beta} = (Z^T Z)^{-1} Z^T Y$
- ◇ Xác định hệ số R^2 , t -statistic, p -value của từng biến Z_i
- ◇ Đánh giá mức độ quan trọng của biến đối với mô hình Chọn tiến dần (Xuất phát từ việc đánh giá một mô hình Null) sau đó lần lượt bổ sung các biến chưa có vào để đánh giá
- ◇ Tiến hành lặp để lựa chọn mô hình (Chọn mô hình cho ta giá trị R^2 lớn nhất)

- **B3: Khảo sát phần dư**

- ◇ Sử dụng tiêu chuẩn Student, kiểm tra sai số ε
- ◇ Xác định Outlier và lọc bỏ chúng dựa vào giá trị Leverage

- **B4: Xây dựng mô hình**

- ◇ Ước lượng lại hệ số hồi quy và khoảng tin cậy tham số
- ◇ Xác định hệ số R -squared (hoặc R)
- ◇ Ước lượng hàm hồi quy tuyến tính

- **B5: Kiểm tra mức độ ổn định của mô hình và đánh giá**

- ◇ Chuẩn hóa tập test theo chuẩn đã thực hiện từ tập train
- ◇ Đánh giá độ chính xác của mô hình dựa trên chỉ số MAPE (sai số tuyệt đối trên tỷ lệ)

7.2.1 Tiền xử lý dữ liệu

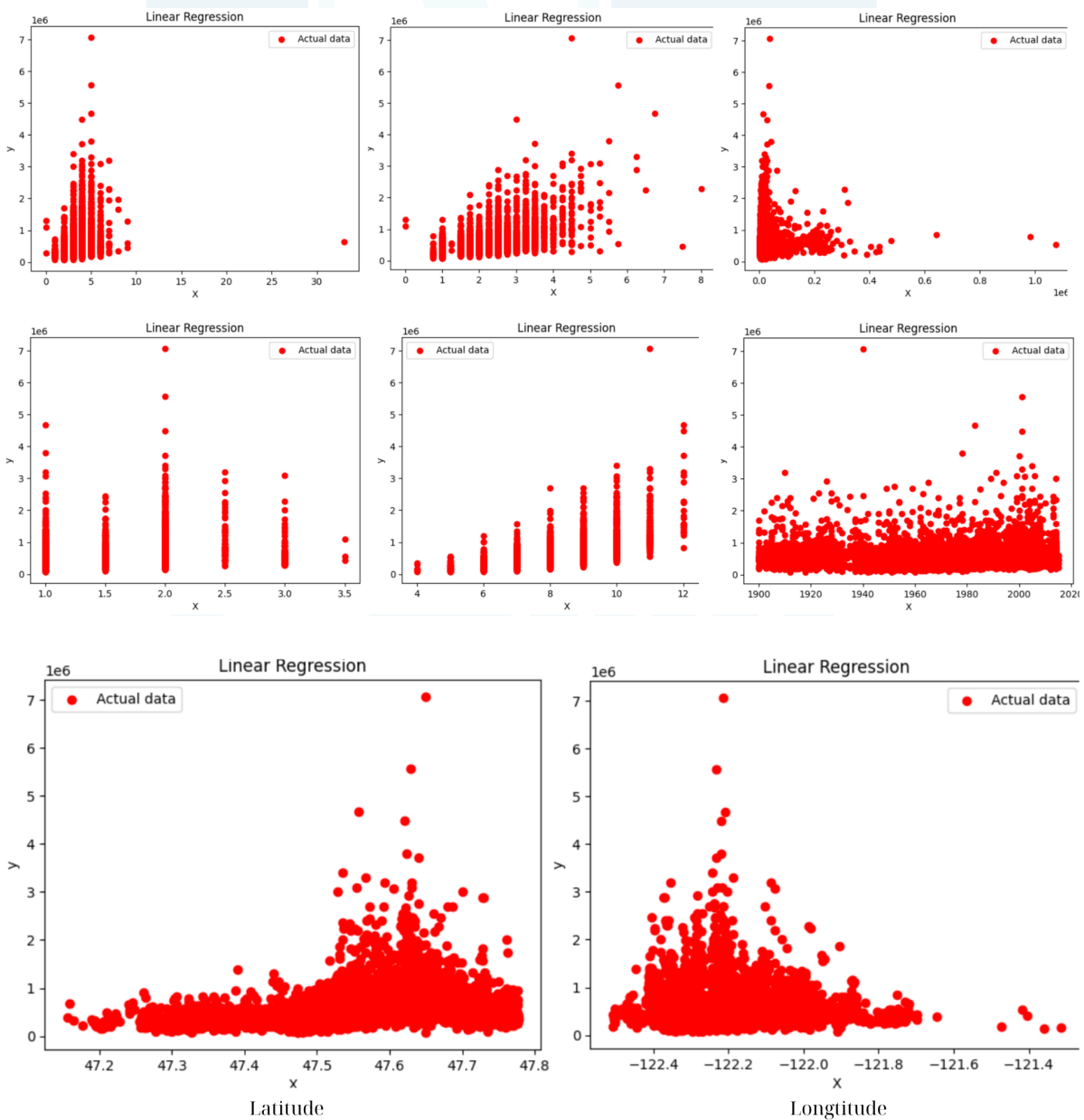
Dữ liệu được đọc từ excel thông qua thư viện pandas sau đó định dạng lại chúng trong 1 array 2 chiều (định dạng ma trận). Dữ liệu truyền vào có 9 cột. khi đó, cột đầu tiên của ma trận chính là Y và 8 cột còn lại sẽ là ma trận Z.

```
data_frame = pd.read_excel("Data_Train_DONE.xlsx")
matrix_data = np.asarray(data_frame.astype(np.float64))
Y = matrix_data.T[0]

Z = np.zeros((len(matrix_data.T[0]), len(matrix_data.T)))
for i in range(len(Z)):
    Z[i][0] = 1

for row in range(len(Z)):
    for col in range(len(Z[0])-1):
        Z[row][col+1] = matrix_data[row][col+1]
```

Sau khi đưa được dữ liệu vào thành các ma trận ta sẽ thực hiện quan sát sự phân bố điểm của từng biến vào phản hồi Y. Ta thu được kết quả như sau:



7.2.2 Chuẩn hóa tập mẫu

Công thức Z-score cho một điểm dữ liệu x trong một tập dữ liệu có trung bình μ và độ lệch chuẩn σ là:

$$Z = \frac{x - \mu}{\sigma}$$

Trong đó:

- x là giá trị của điểm dữ liệu
- μ là trung bình của tập dữ liệu

- σ là độ lệch chuẩn của tập dữ liệu: $\sigma^2 = \frac{\sum_{i=1}^N (X_i - \mu)^2}{N}$

Vì dữ liệu có nhiều cột nên ta chuẩn hóa cho từng cột một. Ta chuẩn hóa với tập train sau đó lưu lại bộ μ và σ của mỗi cột để về sau chuẩn hóa cho tập test theo đúng chuẩn ở trên. Sở dĩ chuẩn hóa tập test theo tập train để đảm bảo tính đồng nhất về phạm vi dữ liệu, tránh trường hợp data leakage (hiện tượng rò rỉ thông tin làm mô hình học thông tin không nên được biết từ tập test) và tạo ra được dự đoán chính xác.

Dưới đây là thuật toán chuẩn hóa tập mẫu

```
def CHTM(Z):
    arr_u = np.zeros(len(Z[0])-1)
    arr_sigma = np.zeros(len(Z[0])-1)
    for k in range(1, len(Z[0])):
        u = sum(Z.T[k])/len(Z)
        sig = 0
        for leg in range(len(Z)):
            sig += (Z.T[k][leg] - u)**2
        sigma_BP = sig/len(Z)
        sigma = math.sqrt(sigma_BP)

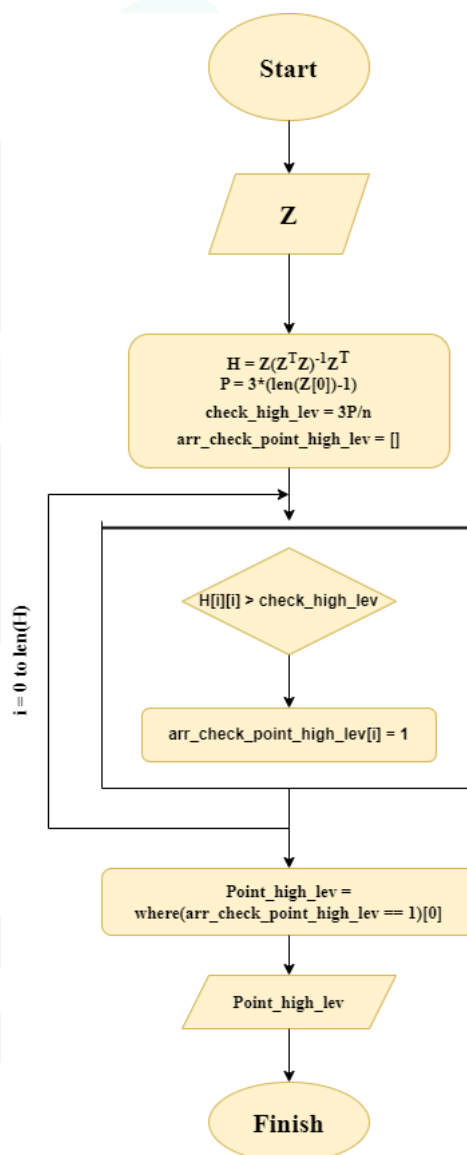
        for ele in range(len(Z)):
            Z[ele][k] = (Z[ele][k]-u)/sigma
        arr_u[k-1] = u
        arr_sigma[k-1] = sigma
    return Z, arr_u, arr_sigma
Z, arr_u, arr_sigma = CHTM(Z)
```

7.2.3 Loại bỏ outlier bằng leverage

Quay trở lại khái niệm về độ đo Leverage. Leverage là độ đo khoảng cách giữa một điểm dữ liệu và phần còn lại của bộ dữ liệu dựa trên miền giá trị của bộ dữ liệu đó. Độ đo Leverage được xác định bởi công thức:

$$h_{jj} = \frac{1}{n} + \frac{(z_j - \bar{z})^2}{\sum_{j=1}^n (z_j - \bar{z})^2}$$

Một cách định nghĩa khác cho độ đo leverage đó là ta sẽ xét ma trận $H := Z(Z^T Z)^{-1}Z$. Khi đó leverage tại điểm dữ liệu thứ j chính là phần tử h_{jj} trên đường chéo chính của ma trận H . Trên thực tế thì ngưỡng đánh giá high leverage thường là $2p/n$ hoặc $3p/n$. Đối với bộ dữ liệu này ta sẽ chọn mức đánh giá là $3p/n$ với p là số biến dự đoán của mô hình.



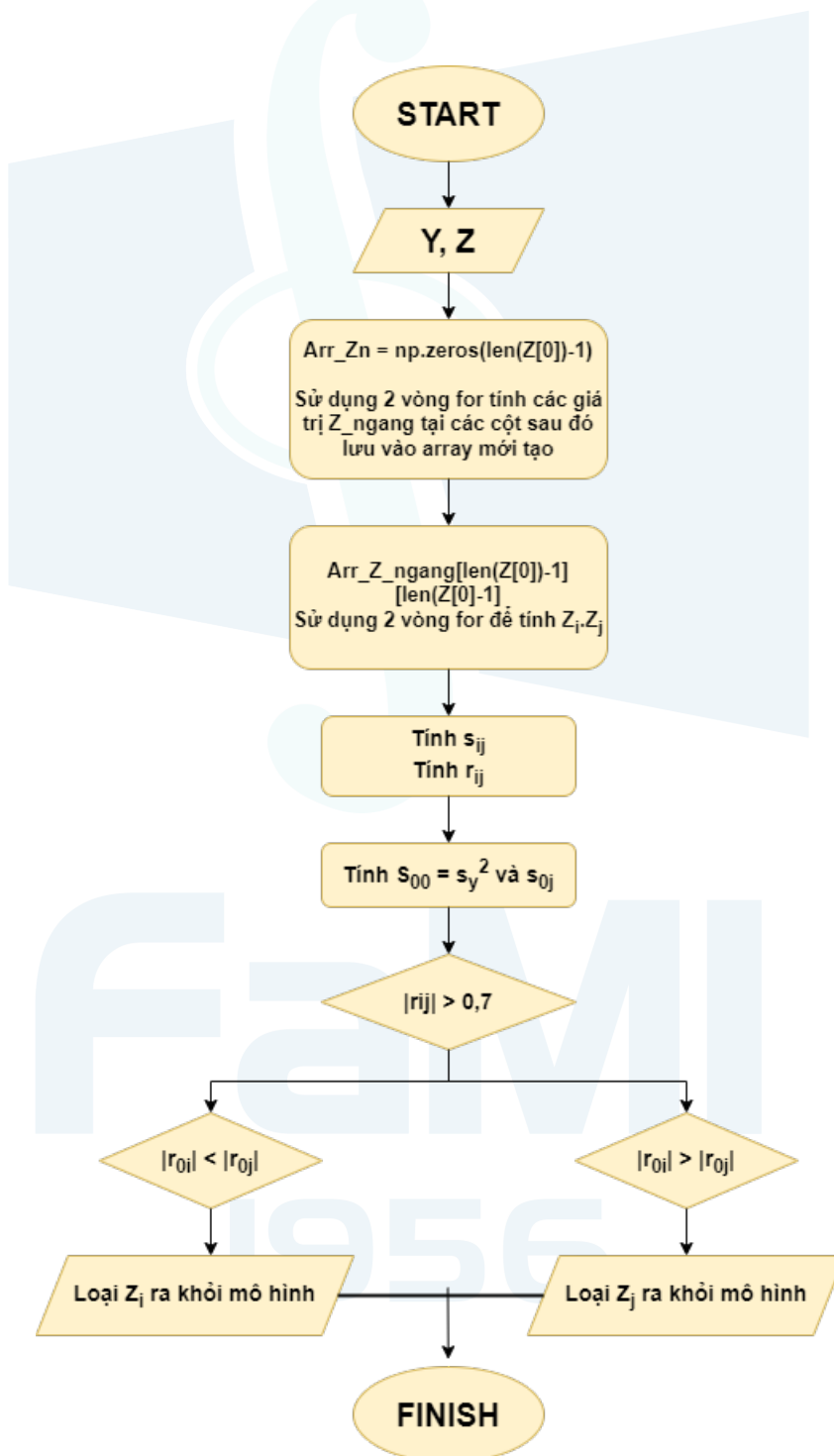
Sau khi chạy chương trình thu được 187 điểm outlier tại các vị trí dưới đây:

```

Các điểm outline với ngưỡng đánh giá 3p/n = [ 82  97 101 114 130 179 254 255 266 286 299 306 348 359
380 401 459 496 505 513 519 530 589 609 620 679 725 793
797 851 938 939 1037 1085 1100 1133 1147 1173 1261 1290 1300 1313
1319 1344 1348 1355 1359 1444 1480 1496 1498 1514 1575 1598 1637 1686
1783 1835 1873 1906 1926 1958 2039 2074 2093 2141 2297 2301 2303 2345
2432 2433 2454 2457 2484 2521 2523 2648 2668 2721 2771 2828 2898 2913
3027 3044 3061 3080 3167 3168 3204 3239 3258 3302 3321 3327 3379 3396
3419 3480 3542 3572 3595 3596 3630 3678 3701 3723 3732 3749 3762 3840
3850 3851 3907 3908 3948 3967 3979 4023 4034 4124 4134 4176 4240 4249
4267 4366 4420 4468 4490 4493 4520 4529 4532 4552 4571 4574 4607 4699
4723 4726 4758 4804 4854 4910 4919 4973 5187 5215 5348 5391 5420 5425
5434 5444 5463 5510 5573 5576 5909 5946 5972 5988 6017 6039 6114 6261
6275 6340 6353 6361 6389 6413 6466 6598 6610 6614 6616 6665 6689 6697
6715 6795 6908 6940 6960]
Số điểm outline là: 187
  
```

7.2.4 Khảo sát tính đa cộng tuyến

Đa cộng tuyến giữa các biến dự đoán ám chỉ trường hợp mà ở đó, hai hay nhiều biến dự đoán có sự ràng buộc với nhau. Việc xác định sự đa cộng tuyến giữa các biến với nhau và giữa các biến với phản hồi đều mang ý nghĩa lớn cho mô hình dữ liệu. Dưới đây là sơ đồ thuật toán để khảo sát tính đa cộng tuyến:



Sơ đồ thuật toán trên sẽ đánh giá sự phụ thuộc của các biến dự đoán với mô hình. Và đưa ra có thể loại bỏ biến nào ra khỏi mô hình.

Ma trận tương quan của các biến dự đoán

1	0.5187	0.1030	0.1736	0.3699	0.1760	-0.0172	0.1706
0.5187	1	0.0994	0.5106	0.6676	0.5142	0.0334	0.2564
0.1030	0.0994	1	-0.0596	0.1749	0.0534	-0.0943	0.2816
0.1736	0.5106	-0.0596	1	0.4630	0.4968	0.0541	0.1452
0.3699	0.6676	0.1749	0.4630	1	0.4436	0.1168	0.2293
0.1760	0.5142	0.0534	0.4968	0.4436	1	-0.1452	0.4512
-0.0172	0.0334	-0.0943	0.0541	0.1168	-0.1452	1	-0.1160
0.1706	0.2564	0.2816	0.1452	0.2293	0.4512	-0.1160	1

Vecto tương quan giữa Y và Z_i

0.3220827	0.51196767	0.11873617	0.26454373	0.66679007	0.04349533	0.31354426	0.03184803
-----------	------------	------------	------------	------------	------------	------------	------------

Nhận xét: Quan sát từ ma trận tương quan giữa các biến Z_i và tương quan giữa Y và từng biến Z ta nhận thấy rằng sự đa cộng tuyến tính giữa các biến chưa thể loại bỏ được biến nào theo điều kiện $|r_{ij}| > 0.7$

7.3 Xác định tiêu chuẩn và các biến quan trọng

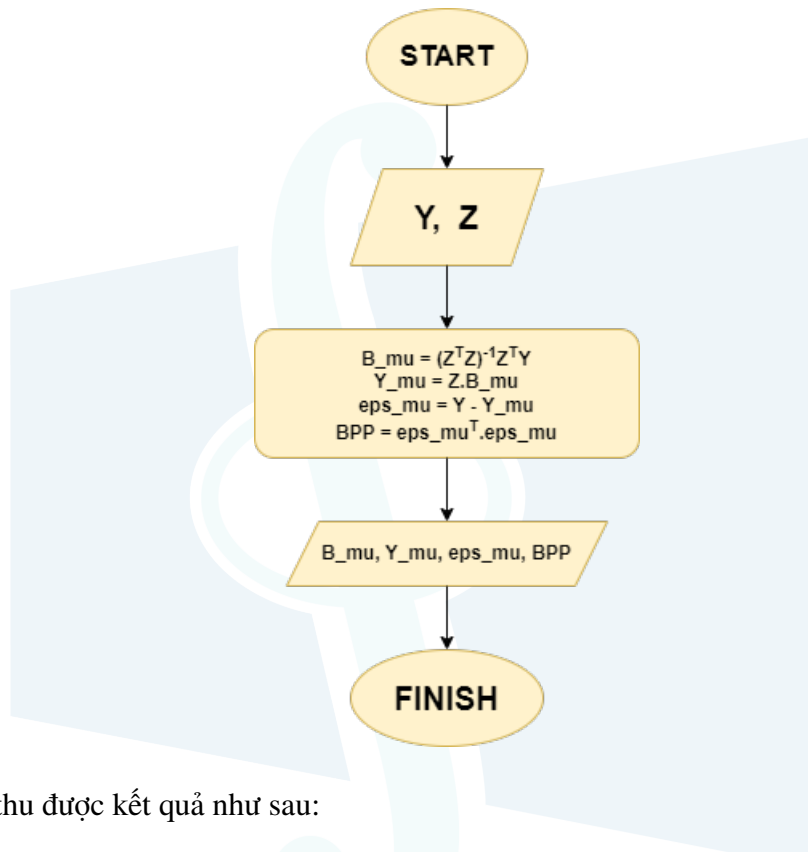
7.3.1 Ước lượng tham số hồi quy

Ta thực hiện ước lượng tham số hồi quy: $\hat{\beta} = (Z^T Z)^{-1} Z^T Y$

Xác định hệ số R^2 : còn được gọi là R-squared, là một độ đo thường được sử dụng để đánh giá mức độ phù hợp của mô hình hồi quy tuyến tính. Nó thể hiện phần trăm phương sai của biến phụ thuộc đã được mô hình hóa bởi mô hình hồi quy. R-squared có giá trị từ 0 đến 1 và được tính bằng cách so sánh phương sai của giá trị dự đoán từ mô hình với phương sai của biến phụ thuộc. Giá trị càng gần 1 thì mô hình càng tốt, tức là nó mô tả phần lớn sự biến động của dữ liệu.

$$\text{Hệ số } R^2 \text{ được xác định bởi công thức: } R^2 = \frac{\sum_{j=1}^n (\hat{Y}_j - \bar{Y})^2}{\sum_{j=1}^n (Y_j - \bar{Y})^2}$$

Xây dựng sơ đồ thuật toán bình phương cực tiểu



Chạy code thu được kết quả như sau:

```

Hệ số ước lượng hồi quy: [ 548301.47126472  9541.30168672  92757.09935123  20576.00575253
2438.02854392  215527.84565516 -115673.86326345  64386.84842327
-4503.30057737]
Độ lệch chuẩn ước lượng: [2747.02490925 3559.2000007 4501.88319244 7248.05102809 3385.2770652
3884.95676687 3719.87659011 2799.94294073 3225.74700734]
  
```

7.3.2 Xác định các giá trị t-statistic , p-value

Bắt đầu xuất phát với cặp giả thuyết $H_0 : \beta_i = 0$ và $H_1 : \beta_i \neq 0$

Giá trị t-statistic được định nghĩa là:

$$t - statistic(Z_i) = \frac{\hat{\beta}_i}{\sqrt{\hat{D}(\hat{\beta}_i)}}$$

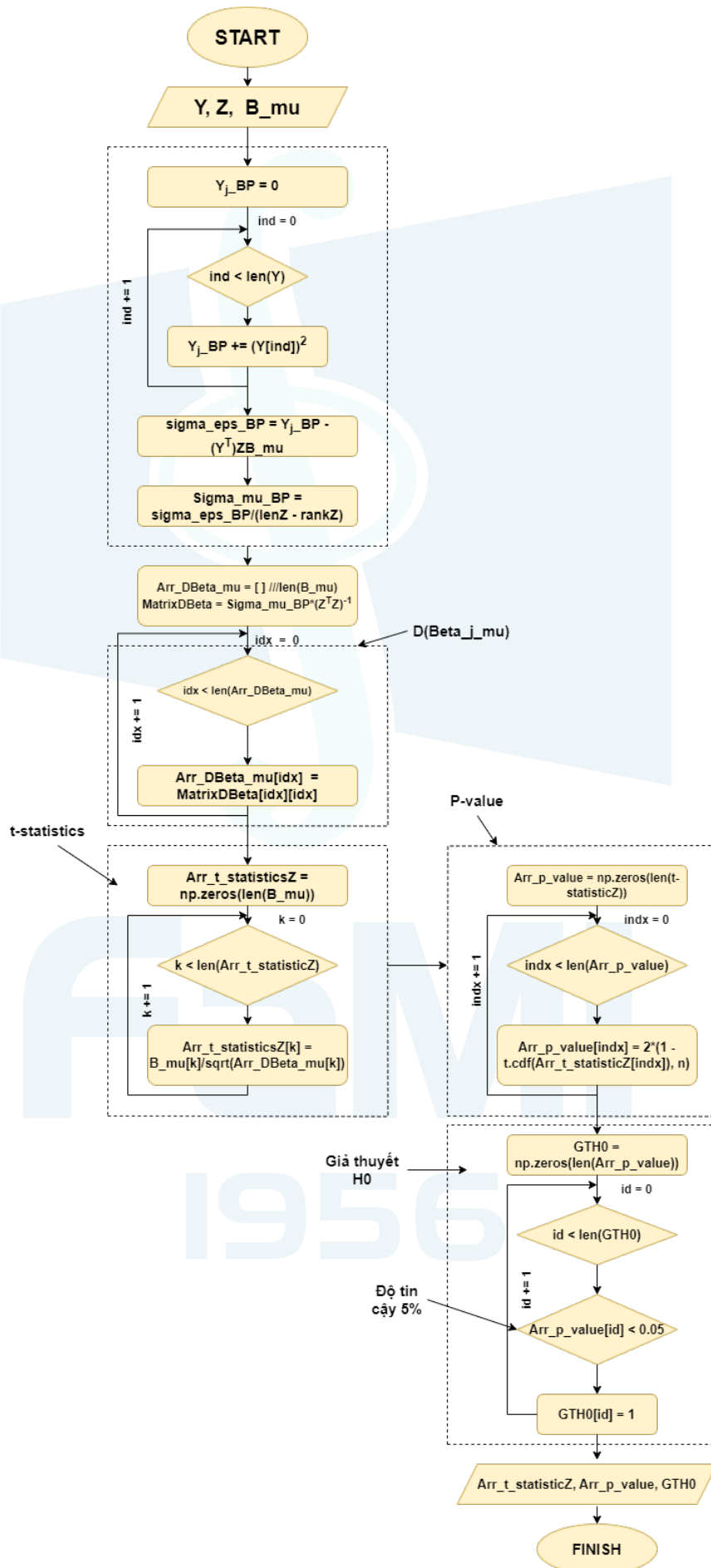
Trong đó $\hat{\beta}_i$ là ước lượng hệ số hồi quy và $\hat{D}(\hat{\beta}_i)$ là phần tử thứ i+1 trên đường chéo chính của ma trận $\hat{\sigma}^2(Z^T Z)^{-1}$, với:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{j=1}^n \hat{\varepsilon}_j^2$$

Ta chọn mức độ bác bỏ giả thuyết H_0 khi p-value < 5%, trong đó, giá trị p-value được xác định như sau:

$$p - value = 2(1 - CDF(n, |t - statistic|))$$

Ta có sơ đồ thuật toán xác định 2 giá trị t - statistic và p - value như sau:



Kết quả khi thực hiện thuật toán thu được như sau:

```
t-statistic tại Z: [183.56740215 2.46543742 18.94923513 2.6108299 0.66234356
51.02181994 -28.59864497 21.14884893 -1.28392419]
p-Value: [0. 0.01370842 0. 0.00905154 0.50777287 0.
2. 0. 1.80078902]
Bác bỏ giả thuyết H0: [1. 1. 1. 1. 0. 1. 0. 1. 0.]
```

Quan sát kết quả thu được:

- Nhận thấy rằng các biến Z_4, Z_6, Z_8 không phụ thuộc tuyến tính vào biến phản hồi Y , từ đó ta có thể đánh giá được, các hệ số B_4, B_6, B_8 có thể bằng 0. Từ đó ta có thể bác bỏ 3 biến trên.

- Lưu ý rút ra: Việc bỏ đi các biến như vậy có thể làm độ chính xác của mô hình dự đoán bị giảm đi 1 chút nhưng không đáng kể do nó không phụ thuộc tuyến tính vào Y với mức tin cậy cho trước. Chính vì thế tùy trường hợp mà ta có xác định nên loại bỏ biến hay không. Việc loại bỏ biến như vậy có thể làm giảm 1 chút về độ chính xác của mô hình nhưng nó sẽ đánh đổi lại về tốc độ, thời gian xử lý và tài nguyên cung cấp. Việc loại bỏ nên được kiểm thử chi tiết để tránh những mất mát không đáng có.

- Đối với bài toán này em sẽ thực hiện kiểm tra với 2 trường hợp:

- Thứ nhất là xử lý cho Data đầy đủ
- Thứ hai là loại bỏ 3 biến trên (Z_4, Z_6, Z_8)

7.3.3 Đánh giá mức độ quan trọng của biến đối với mô hình

Ở đây ta chọn xét mô hình M_k tiến xuất phát với bộ biến Null (trống biến) sau đó bổ sung lần lượt các biến chưa có trong mô hình thành biến M_{k+1} nếu R^2 của mô hình mới lớn hơn mô hình cũ thì chọn mô hình mới để tiếp tục lặp cho đến hết và trả ra các biến góp phần quan trọng nhất cho mô hình (làm cho mô hình có R^2 cao nhất)

Kết quả đánh giá mô hình:

```
[0.59307593 0.59264614 0.56768638 0.59259395 0.59304491 0.40900573
0.55836564 0.59215719 0.59263104 0.40618972 0.56717175 0.56584782
0.26716128 0.59258581 0.39882785 0.40305633 0.55780396 0.55639816
0.18948414 0.59215595 0.39577524 0.40077788 0.56554573 0.24876033
0.22119607 0.39426238 0.55606618 0.16459757 0.12766425 0.39171801
0.20668167 0. ]
```

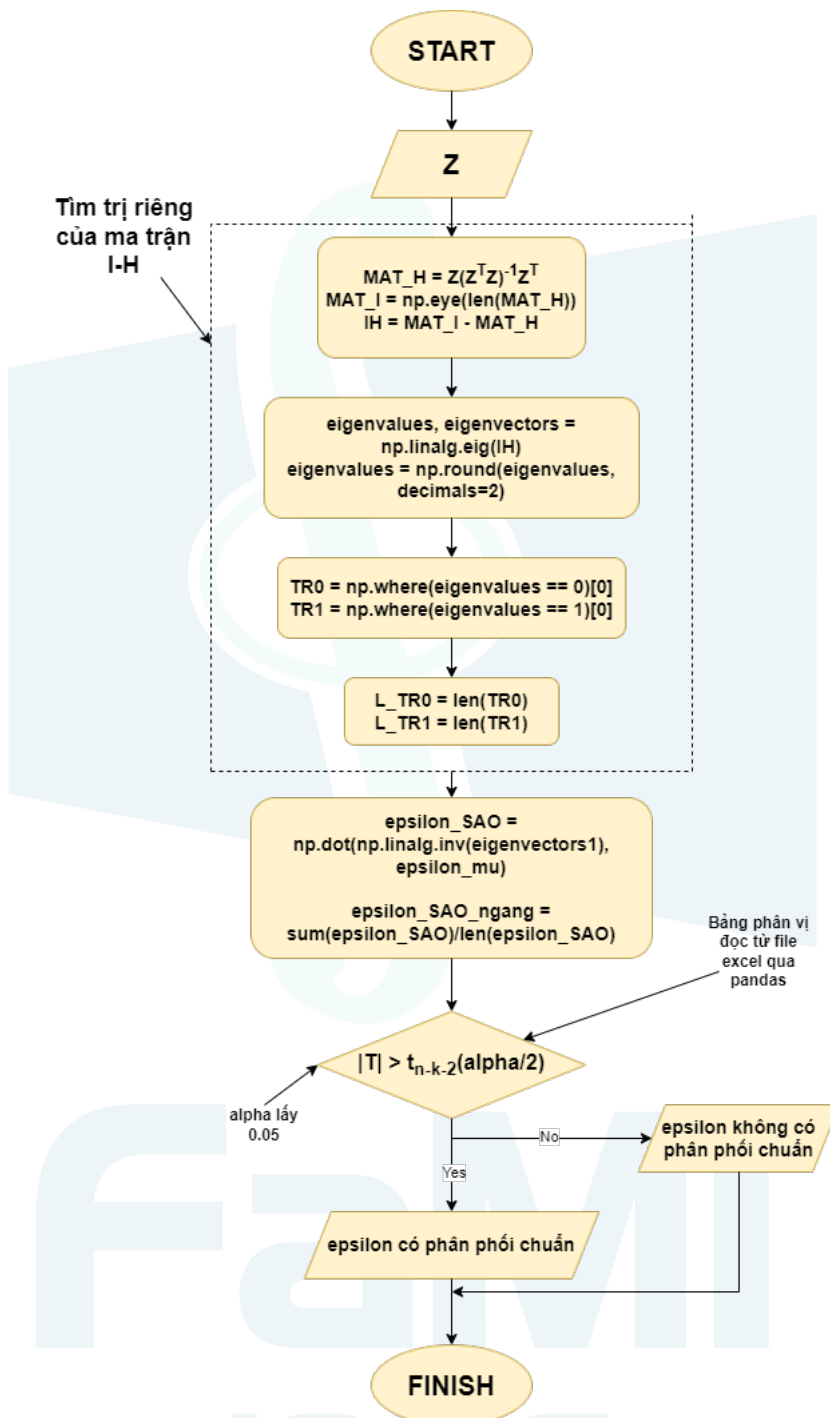
Từ bảng giá trị thu được ở trên có thể nhận thấy rằng mô hình gồm 8 biến có R^2 lớn nhất, cho nên ta sẽ chọn thực hiện xây dựng với mô hình này.

Quan sát bảng dữ liệu cho thấy cũng có nhiều mô hình có R^2 gần với mô hình 8 biến tức là những mô hình đó có các biến ít ảnh hưởng đến giá trị phản hồi. Ta cũng có thể chọn nó để giảm chi phí train mô hình.

Ở đây em chọn tiến hành với mô hình có R^2 lớn nhất tức là mô hình có 8 biến dự đoán Z_i .

7.3.4 Khảo sát phần dư

Như ở trên đã trình bày chúng ta cần phải khảo sát xem phần dư của mô hình của đạt phân phối chuẩn hay không. Bỏ lẽ đối với một mô hình hồi quy thì phần dư đạt phân phối chuẩn thì mô hình chúng ta đang xét mới thực sự có ý nghĩa. Vì thế mà chúng ta sẽ đặt ra giả thuyết $H_0: \varepsilon \sim$ phân phối chuẩn. Áp dụng ta xây dựng được thuật toán:



Kết quả trả ra cho thấy mô hình mà chúng ta đang xét ở đây có phần dư tuân theo phân phối chuẩn

```

eps_SAO = [ 5.68794906e+22 -1.12953711e+24  7.16484727e+23 ... -3.21523573e+22
 3.34500489e+22 -4.86961172e+22]
eps_SAO_ngang = 8.797720803061088e+22
TC_STUDENT = 6.720978047918398e-15
Epsilon tuân theo phân phối chuẩn!
  
```

7.4 Kiểm tra và đánh giá mô hình

Thực hiện lọc bỏ outlier thêm một lần nữa ta thu được 155 điểm với ngưỡng đánh giá high leverage là $3p/n$.


```

Các điểm outline với ngưỡng đánh giá 3p/n = [ 8 29 58 77 116 241 259 267 268 322 342 349 355 449
452 460 461 466 478 682 695 754 802 838 903 962 1000 1034
1063 1085 1090 1156 1193 1198 1265 1278 1302 1339 1399 1429 1499 1512
1624 1672 1696 1807 1822 1872 1915 1984 2017 2115 2131 2182 2187 2243
2367 2440 2536 2545 2632 2665 2681 2696 2731 2859 2876 2984 3023 3108
3156 3287 3289 3313 3322 3405 3477 3585 3596 3622 3687 3698 3741 3773
3803 3853 3856 3926 3973 4013 4029 4035 4057 4081 4092 4165 4179 4209
4210 4289 4413 4458 4552 4575 4629 4630 4734 4817 4890 4945 4964 4965
4984 5017 5044 5148 5177 5216 5225 5355 5367 5390 5501 5570 5599 5685
5695 5740 5763 5764 5817 5819 5872 5922 5995 6045 6093 6096 6105 6135
6141 6248 6273 6280 6311 6328 6457 6519 6579 6646 6671 6678 6718 6742
6803]
Số điểm outline là: 155

```

7.4.1 Thực hiện ước lượng lại tham số và R^2 của mô hình

```

Hệ số ước lượng hồi quy: [ 545873.00188899 10426.8781372 86864.92383146 17174.32496493
2330.59019748 212557.88335641 -109766.86525783 64381.02043107
-4892.70206688]
Độ lệch chuẩn ước lượng: [ 2890.68379077 3431.51000965 4349.82814032 10778.26881108
3260.45526118 3771.21170049 3592.05340364 2662.89828028
3128.59060164]
Dữ liệu thể hiện mối quan hệ hồi quy yếu , R^2 = 0.6020617028813046

```

Nhận thấy sau khi lọc bỏ các điểm high-leverage đã làm cho chỉ số R^2 tăng lên 1 chút từ 0.59307 lên thành 0.60206. Tuy chỉ tăng thêm được một chút nhưng chúng ta cũng có thể thấy được tác động của điểm outlier có ảnh hưởng tới độ chính xác của mô hình.

7.4.2 Ước lượng khoảng tham số

Với chứng minh ở trên, chúng ta tiến hành xây dựng thuật toán. Ở đây thì có sử dụng đến phân vị student cho nên ta thực hiện import file excel vào sau đó cho code đọc excel để tìm phân vị student. Với n đủ lớn thì ta xấp xỉ nó với phân vị laplace qua công thức:

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$$

Kết quả thu được cho khoảng tin cậy của các tham số

```

Khoảng tin cậy của các tham số:
[[ 540207.26165909 551538.7421189 ]
 [ 3701.11851829 17152.6377561 ]
 [ 78339.26067643 95390.58698649]
 [ -3951.08190479 38299.73183465]
 [ -4059.90211442 8721.08250939]
 [ 205166.30842346 219949.45828937]
 [-116807.28992897 -102726.44058669]
 [ 59161.73980171 69600.30106042]
 [ -11024.7396461 1239.33551233]]

```

7.4.3 Test mô hình và đánh giá

Chúng ta chia bộ dữ liệu là 2 phần với tỷ lệ 7:3, trong đó 70% là dành cho tập train, 30% là dành cho tập test. Đầu tiên tiến hành chuẩn hóa tập test với chuẩn đã xác định ở tập train. Sở dĩ chuẩn hóa tập test theo tập train để đảm bảo tính đồng nhất về phạm vi dữ liệu, tránh trường hợp data leakage (hiện tượng rò rỉ thông tin làm mô hình học thông tin không nên được biết từ tập test) và tạo ra được dự đoán chính xác. Thực hiện dự đoán và đánh giá độ chính xác của mô hình dựa trên chỉ số MAPE (sai số tuyệt đối trên tỷ lệ).

- Đo lường độ chính xác của mô hình dự đoán theo tỷ lệ tuyệt đối giữa giá trị thực tế và giá trị dự đoán
- $MAPE = \frac{1}{n} \sum_{i=1}^n \left(\frac{|T_i - T_d|}{T_i} \right) \times 100$ với T_i là giá trị thực tế và T_d là giá trị dự đoán

```
[ [ 0.61850822 -0.51986267 -0.23909922 ... -1.80592345 0.87069099
-0.6942589 ]
[ 1.65107621 0.12759551 -0.17318897 ... -0.28298596 0.09973179
0.94788727]
[ 0.61850822 0.12759551 -0.12896494 ... 0.08928765 -0.03780674
0.48471784]
...
[-0.41405976 -1.49104994 -0.1254168 ... 0.02160154 -1.65630064
0.85665693]
[ 0.61850822 0.4513246 -0.29473406 ... -1.50133595 0.48335276
-1.06619799]
[ 0.61850822 0.12759551 -0.19669185 ... 0.66461959 -1.46300326
0.22506225]]
26.984737877148927
```

Chúng ta có thể quan sát được ma trận Z sau khi đã định chuẩn và giá trị MAPE xấp xỉ 27 % điều đó cho thấy mô hình có khả năng hồi quy nhưng chưa thực sự tốt. Trên thực tế giá trị MAPE đánh giá một mô hình có độ chính xác tốt hay không thường đạt mức từ 10-20%. Như vậy với mức chỉ số MAPE ở ngưỡng 27% ta có thể nói rằng mô hình có thể dùng để dự đoán cho tương lai nhưng chưa thực sự tốt.

7.4.4 So sánh 2 mô hình hồi quy

	Mô hình gốc (R^2 lớn nhất)	Mô hình loại bỏ 3 biến Z_4, Z_6, Z_8
R^2	Đầu: 0.59307592865 Loại outlier: 0.602061702881	Đầu: 0.516019416011 Loại outlier: 0.534916648135
Epsilon	Tuân theo phân phối chuẩn	Tuân theo phân phối chuẩn
Time	9p38s/train	8p11s/train
MAPE	26.984737877148927	29.27214638492003%

Quan sát so sánh trên, chúng ta có thể nhận thấy rằng, việc loại bỏ các biến sẽ dẫn đến mức độ phù hợp của mô hình đang giảm rất nhiều (đặc trưng là thể hiện qua hệ số R^2). Đồng thời chỉ số MAPE cũng đã tăng lên sau khi loại bỏ biến, điều đó có nghĩa là sai số đã tăng lên sau khi loại bỏ biến. Có một điều mà chúng ta nhận thấy được rất rõ, đó là thời gian train đã giảm xuống. Điều này có ý nghĩa cho thực tế rất nhiều bởi trong một số trường hợp nhất định thì việc đánh đổi hiệu suất để lấy thời gian cũng là một điều cần làm.

7.5 Một số nhận xét cho mô hình dữ liệu thực tế

Mô hình hồi quy tuyến tính được áp dụng rộng rãi trên nhiều lĩnh vực thực tế. Nó có ứng dụng khá cao cho việc dự đoán nhiều thành phần. Nhưng tất nhiên không phải lúc nào cũng có thể dùng được phương pháp này. Chúng ta có thể rút ra một vài nhận định như sau:

- Mô hình chỉ có ý nghĩa khi phần dư đạt phân phối chuẩn
- Mô hình không áp dụng cho trường hợp đa cộng tuyến nhiều
- Không áp dụng cho những lĩnh vực dự báo ngắn hạn, dữ liệu biến động nhanh chóng
- Không dùng cho những bộ dữ liệu có yếu tố phi tuyến lớn. Nếu có yếu tố phi tuyến chúng ta nên sử dụng các mạng neural để thực hiện xác định một cách chính xác hơn

FaMI
1956

Mô hình hồi quy tuyến tính bội

8.1 Mô hình bài toán

Mô hình hóa quan hệ giữa m biến phản hồi Y_1, Y_2, \dots, Y_m và tập các biến dự đoán z_1, z_2, \dots, z_k . Giả sử mỗi biến phản hồi tuân theo một mô hình hồi quy, theo đó:

$$\begin{aligned} Y_1 &= \beta_{01} + \beta_{11}z_1 + \dots + \beta_{k1}z_k + \varepsilon_1 \\ Y_2 &= \beta_{02} + \beta_{12}z_1 + \dots + \beta_{k2}z_k + \varepsilon_2 \\ &\vdots \\ Y_m &= \beta_{0m} + \beta_{1m}z_1 + \dots + \beta_{km}z_k + \varepsilon_m \end{aligned}$$

Với sai số $\varepsilon^T = [\varepsilon_1 \ \varepsilon_2 \ \dots \ \varepsilon_m]$ được giả sử tuân theo phân phối chuẩn m chiều với $E(\varepsilon) = 0$ và $\text{Var}(\varepsilon) = \Sigma$ là một ma trận đối xứng xác định dương.

Để xây dựng khái niệm mô hình mới sao cho phù hợp với mô hình hồi quy tuyến tính cổ điển, ta ký hiệu $\mathbf{Z}_j = [z_{j0} \ z_{j1} \ \dots \ z_{jk}]$ là giá trị các biến dự đoán ở lần quan sát thứ j , $\mathbf{Y}_j^T = [Y_{j1} \ Y_{j2} \ \dots \ Y_{jm}]$ là các biến phản hồi và $\varepsilon_j^T = [\varepsilon_{j1} \ \varepsilon_{j2} \ \dots \ \varepsilon_{jm}]$ là sai số.

Về kí hiệu ma trận:

$$\mathbf{Z}_{(n \times (k+1))} = \begin{bmatrix} 1 & z_{11} & \dots & z_{1k} \\ 1 & z_{21} & \dots & z_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \dots & z_{nk} \end{bmatrix}$$

cũng tương tự trong mô hình hồi quy tuyến tính đơn biến phản hồi cổ điển.

Các ma trận còn lại:

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

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

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

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

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

với các giả thiết

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

Tương tự như mô hình hồi quy tuyến tính cổ điển, sai số tại các lần thử nghiệm khác nhau sẽ không có sự tương quan, nghĩa là $\text{Cov}(\varepsilon_{ij}, \varepsilon_{kq}) = 0$ với $j \neq q$. Tuy nhiên, các sai số trong cùng một lần thử nghiệm thì có thể tương quan với nhau.

8.2 Ước lượng tham số

Vì biến phản hồi thứ j tuân theo mô hình hồi quy cổ điển

$$Y_{(j)} = \mathbf{Z} \boldsymbol{\beta}_{(j)} + \boldsymbol{\varepsilon}_{(j)}$$

Tổng bình phương sai số $RSS(\boldsymbol{\beta}_{(j)})$ được cực tiểu hóa bởi $\hat{\boldsymbol{\beta}}_{(j)} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y}_{(j)}$. Do đó:

$$RSS([\boldsymbol{\beta}]) = \sum_{j=1}^n RSS(\boldsymbol{\beta}_{(j)}) = \sum_{j=1}^n (\mathbf{Y}_{(j)} - \mathbf{Z} \boldsymbol{\beta}_{(j)})^T (\mathbf{Y}_{(j)} - \mathbf{Z} \boldsymbol{\beta}_{(j)})$$

Vì mỗi số hạng trong tổng trên đều không âm nên $RSS([\boldsymbol{\beta}])$ đạt cực tiểu khi từng $RSS(\boldsymbol{\beta}_{(j)})$ đạt cực tiểu, tức là $\hat{\boldsymbol{\beta}}_{(j)} = \widehat{\boldsymbol{\beta}}_{(j)} = (\mathbf{Z}^T \mathbf{Z})^{-1} \mathbf{Z}^T \mathbf{Y}_{(j)}$.

Vậy

$$[\hat{\beta}] = [\hat{\beta}_{(1)} : \hat{\beta}_{(2)} : \dots : \hat{\beta}_{(m)}] = (Z^T Z)^{-1} Z^T [Y_{(1)} : Y_{(2)} : \dots : Y_{(m)}] = (Z^T Z)^{-1} Z^T [Y] \quad (8.1)$$

là một ước lượng bình phương cực tiểu cho $[\beta]$.

Sử dụng ước lượng bình phương cực tiểu $[\hat{\beta}]$ vừa tìm được ở công thức (8.1), ta có thể xây dựng các ma trận ước lượng của $[Y]$ và $[\varepsilon]$ như sau:

$$[\hat{Y}] = Z\hat{\beta} = Z(Z^T Z)^{-1} Z^T [Y]$$

$$[\hat{\varepsilon}] = [Y] - [\hat{Y}] = (I - Z(Z^T Z)^{-1} Z^T) [Y]$$

Hơn nữa, ta có:

$$Z^T [\hat{\varepsilon}] = Z^T (I - Z(Z^T Z)^{-1} Z^T) [Y] = (Z^T - Z^T Z(Z^T Z)^{-1} Z^T) [Y] = 0$$

$$[\hat{Y}]^T [\hat{\varepsilon}] = (Z\hat{\beta})^T [\hat{\varepsilon}] = [\hat{\beta}]^T Z^T [\hat{\varepsilon}] = 0$$

Điều này kéo theo mọi cột của Z và $[\hat{Y}]$ trực giao với mọi cột của $[\hat{\varepsilon}]$. Mặt khác, do $[Y] = [\hat{Y}] + [\hat{\varepsilon}]$ nên $[Y]^T [Y] = ([\hat{Y}] + [\hat{\varepsilon}])^T ([\hat{Y}] + [\hat{\varepsilon}])$

Như vậy:

$$\begin{aligned} [Y]^T [Y] &= [\hat{Y}]^T [\hat{Y}] + [\hat{\varepsilon}]^T [\hat{\varepsilon}] \\ \left(\begin{array}{c} \text{Tích vô hướng của} \\ \text{các vector phản hồi} \end{array} \right) &= \left(\begin{array}{c} \text{Tích vô hướng của} \\ \text{các vector dự đoán} \end{array} \right) + \left(\begin{array}{c} \text{Tích vô hướng của} \\ \text{các vector nhiễu} \end{array} \right) \end{aligned}$$

Định lý 8.1

Xét mô hình hồi quy tuyến tính bội $[Y] = Z[\beta] + [\varepsilon]$, với $\text{rank} Z = k + 1 \leq n - m$, và nhiễu $[\varepsilon]$ có phân phối chuẩn. Khi đó:

$$\hat{\Sigma} = \frac{1}{n} [\hat{\varepsilon}]^T [\hat{\varepsilon}] = \frac{1}{n} \left([Y] - Z[\hat{\beta}] \right)^T \left([Y] - Z[\hat{\beta}] \right)$$

là ước lượng hợp lý cực đại của Σ

Chứng minh: Theo giả thiết, ta có n quan sát và tại quan sát thứ i thì vector $\varepsilon_i = [\varepsilon_{i1} \ \varepsilon_{i2} \ \dots \ \varepsilon_{im}]$ tuân theo phân phối chuẩn m chiều $N_m(0, \Sigma)$ với hàm mật độ xác suất đồng thời:

$$\begin{aligned} f(\varepsilon_i) &= \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} \exp \left(-\frac{1}{2} \varepsilon_i \Sigma^{-1} \varepsilon_i^T \right) \\ &= \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} \exp \left(-\frac{1}{2} (Y_i - Z\beta_i) \Sigma^{-1} (Y_i - Z\beta_i)^T \right) \end{aligned}$$

Từ đây ta có được hàm hợp lý cho n quan sát:

$$L([\beta], \Sigma, [Y]) = \prod_{i=1}^n f(\epsilon_i) = \frac{1}{\sqrt{(2\pi)^{mn}(\det \Sigma)^n}} \exp \left(-\frac{1}{2} \sum_{i=1}^n (Y_i - Z\beta_i) \Sigma^{-1} (Y_i - Z\beta_i)^T \right) \quad (8.2)$$

Đặt:

$$Q = [q_{ij}]_{m \times m} := [\hat{\epsilon}]^T [\hat{\epsilon}] = ([Y] - Z[\hat{\beta}])^T ([Y] - Z[\hat{\beta}])$$

Khi đó:

$$q_{ij} = \sum_{k=1}^n (y_{ki} - \hat{y}_{ki})(y_{kj} - \hat{y}_{kj})$$

trong đó y_{ij} và \hat{y}_{ij} lần lượt là phần tử ở hàng i, cột j của ma trận $[Y]$ và $[\hat{Y}] = Z[\hat{\beta}]$.

Với $(\Sigma^{-1})_{ij}$ là phần tử ở hàng i, cột j của ma trận Σ^{-1} , ta có biến đổi:

$$\begin{aligned} \sum_{i=1}^n (Y_i - Z\beta_i) \Sigma^{-1} (Y_i - Z\beta_i)^T &= \sum_{i=1}^n \left(\sum_{j=1}^m \sum_{k=1}^m (y_{ij} - \hat{y}_{ij}) (\Sigma^{-1})_{jk} (y_{ik} - \hat{y}_{ik}) \right) \\ &= \sum_{j=1}^m \sum_{k=1}^m \left((\Sigma^{-1})_{jk} \sum_{i=1}^n (y_{ij} - \hat{y}_{ij})(y_{ik} - \hat{y}_{ik}) \right) \\ &= \sum_{j=1}^m \sum_{k=1}^m (\Sigma^{-1})_{jk} q_{jk} \\ &= tr(\Sigma^{-1} Q) \end{aligned}$$

Như vậy hàm hợp lý ở công thức (8.2) có thể viết lại thành:

$$L([\beta], \Sigma, [Y]) = \frac{1}{\sqrt{(2\pi)^{mn}(\det \Sigma)^n}} \exp \left(-\frac{1}{2} tr(\Sigma^{-1} Q) \right)$$

Tới đây ta sẽ tìm ma trận Σ để cực đại hóa log-hàm hợp lý:

$$\log L([\beta], \Sigma, [Y]) = -mn \log 2\pi + n \log \det(\Sigma^{-1}) - tr(\Sigma^{-1} Q)$$

Do Q là ma trận đối xứng xác định dương nên tồn tại ma trận nghịch đảo Q^{-1} và ma trận căn bậc hai $Q^{1/2}$. Đặt $A = Q^{1/2} \Sigma^{-1} Q^{1/2}$, khi đó $\Sigma^{-1} = Q^{-1/2} A Q^{-1/2}$.

Ta viết lại log-hàm hợp lý thành:

$$\begin{aligned}
 \log L([\beta], \Sigma, [Y]) &= -mn \log 2\pi + n \log \det(Q^{-1/2} A Q^{-1/2}) - \text{tr}(Q^{-1/2} A Q^{-1/2} Q) \\
 &= -mn \log 2\pi + n \log \det(Q^{-1}) + n \log \det A - \text{tr}(Q^{-1/2} A Q^{1/2}) \\
 &= -mn \log 2\pi - n \log \det Q + n \log \det A - \text{tr} A
 \end{aligned} \tag{8.3}$$

Gọi λ_i ($i = \overline{1, m}$) là các trị riêng của ma trận A. Ta có các kết quả đã có trong đại số tuyến tính như sau:

$$\det A = \prod_{i=1}^m \lambda_i \text{ và } \text{tr} A = \sum_{i=1}^m \lambda_i$$

Thay các kết quả này vào (8.3) ta được:

$$\log L([\beta], \Sigma, [Y]) = -mn \log 2\pi - n \log \det Q + \sum_{i=1}^m (-\lambda_i + n \log \lambda_i)$$

Như vậy log-hàm hợp lý sẽ đạt cực đại khi và chỉ khi các giá trị $(-\lambda_i + n \log \lambda_i)$ đạt cực đại, tương đương với $\lambda_i = n$ với $i = \overline{1, m}$. Mặt khác, do A là ma trận đối xứng nên ta có thể viết A dưới dạng chéo hóa:

$$A = P \Lambda P^T$$

trong đó $\Lambda = \text{diag}\{\lambda_1, \lambda_2, \dots, \lambda_m\}$ và P là ma trận trực giao. Do các trị riêng của Λ đều bằng n nên:

$$A = P(nI)P^T = nPP^T = nI$$

Từ đây suy ra ước lượng hợp lý cực đại của Σ là:

$$\hat{\Sigma} = Q^{1/2} A^{-1} Q^{1/2} = Q^{1/2} \left(\frac{1}{n} I \right) Q^{1/2} = \frac{1}{n} Q = \frac{1}{n} ([Y] - Z[\hat{\beta}])^T ([Y] - Z[\hat{\beta}])$$

Giá trị lớn nhất của hàm hợp lý cực đại đạt tại $[\hat{\beta}]$ và $\hat{\Sigma}$ là:

$$\begin{aligned}
 L([\hat{\beta}], \hat{\Sigma}, [Y]) &= \frac{1}{\sqrt{(2\pi)^{mn} (\det \hat{\Sigma})^n}} \exp \left(-\frac{1}{2} \text{tr}(\hat{\Sigma}^{-1} Q) \right) \\
 &= \frac{1}{\sqrt{(2\pi)^{mn} (\det \hat{\Sigma})^n}} \exp \left(-\frac{1}{2} \text{tr}(n([\epsilon]^T [\epsilon])^{-1} ([\epsilon]^T [\epsilon])) \right) \\
 &= \frac{1}{\sqrt{(2\pi)^{mn} (\det \hat{\Sigma})^n}} \exp \left(-\frac{1}{2} \text{tr}(nI) \right) \\
 &= (\det \hat{\Sigma})^{-\frac{n}{2}} (2\pi e)^{-\frac{mn}{2}}
 \end{aligned}$$

8.3 Các tính chất quan trọng

Tính chất 8.1

$[\hat{\beta}]$ là ước lượng không chệch của $[\beta]$.

Chứng minh: Yêu cầu tương đương với chứng minh $E([\hat{\beta}]) = [\beta]$. Thật vậy, ta có:

$$\begin{aligned} E([\hat{\beta}]) &= E((Z^T Z)^{-1} Z^T [Y]) \\ &= E((Z^T Z)^{-1} Z^T (Z[\beta] + [\varepsilon])) \\ &= E((Z^T Z)^{-1} Z^T Z[\beta]) + E((Z^T Z)^{-1} Z^T [\varepsilon]) \\ &= [\beta] + (Z^T Z)^{-1} Z^T E([\varepsilon]) = [\beta]. \end{aligned}$$

Tính chất 8.2

$[\hat{\beta}]$ là ước lượng hợp lý cực đại của $[\beta]$ (Khi $[\varepsilon]$ có phân phối chuẩn).

Chứng minh: Vì $[\varepsilon]$ có phân phối chuẩn nên hàm hợp lý cho n quan sát sẽ có dạng như công thức (8.2):

$$L([\beta], \Sigma, [Y]) = \frac{1}{\sqrt{(2\pi)^{mn} (\det \Sigma)^n}} \exp \left(-\frac{1}{2} \sum_{i=1}^n (Y_i - Z\beta_i) \Sigma^{-1} (Y_i - Z\beta_i)^T \right)$$

Đặt:

$$\begin{aligned} S_i &:= (Y_i - Z\beta_i) \Sigma^{-1} (Y_i - Z\beta_i)^T \\ S &:= \sum_{i=1}^n (Y_i - Z\beta_i) \Sigma^{-1} (Y_i - Z\beta_i)^T = \sum_{i=1}^n S_i \end{aligned}$$

Ta sẽ tìm giá trị $[\hat{\beta}]$ để cực đại hóa hàm hợp lý trên. Do đại lượng $1/\sqrt{(2\pi)^{mn} (\det \Sigma)^n}$ không phụ thuộc vào $[\beta]$ nên L đạt cực đại khi và chỉ khi giá trị S đạt cực tiểu. Mặt khác, bởi vì S là tổng của dạng toàn phương xác định không âm S_i nên S đạt cực tiểu khi và chỉ khi các giá trị S_i đạt cực tiểu.

Ta có:

$$\frac{\partial S_i}{\partial [\beta]^T} = (\Sigma^{-1} + (\Sigma^{-1})^T) (Y_i^T - [\beta]^T Z_i^T) (-Z_i) = -2\Sigma^{-1} (Y_i^T Z_i - [\beta]^T Z_i^T Z_i)$$

Do ma trận Σ^{-1} là ma trận đối xứng xác định dương nên $\partial S_i / \partial [\beta]^T = 0$ khi và chỉ khi $Y_i^T Z_i = [\beta]^T Z_i^T Z_i$. Lấy tổng với i từ 1 đến n , ta được:

$$\sum_{i=1}^n Y_i^T Z_i = \sum_{i=1}^n [\beta]^T (Z_i^T Z_i) \Leftrightarrow [Y]^T Z = [\beta]^T (Z^T Z) \Leftrightarrow [\beta] = (Z^T Z)^{-1} Z^T [Y]$$

Chứng minh hoàn tất.

Tính chất 8.3

$$E([\hat{\varepsilon}]) = 0 \text{ và } Cov(\hat{\beta}_{(i)}, \hat{\beta}_{(j)}) = \sigma_{ij}(Z^T Z)^{-1}.$$

Chứng minh: Với mọi $j = \overline{1, m}$ thì ta có:

$$\begin{aligned} E(\hat{\varepsilon}_{(j)}) &= E(Y_{(j)} - \hat{Y}_{(j)}) \\ &= E([I - Z(Z^T Z)^{-1} Z^T] Y_{(j)}) \\ &= E([I - Z(Z^T Z)^{-1} Z^T] \varepsilon_{(j)}) \quad (\forall i \ Y_{(j)} = Z\beta_{(j)} + \varepsilon_{(j)}) \\ &= 0 \end{aligned}$$

Đặt $C = (Z^T Z)^{-1} Z^T$. Khi đó:

$$\begin{aligned} Cov(\hat{\beta}_{(i)}, \hat{\beta}_{(j)}) &= Cov(CY_{(i)}, CY_{(j)}) \\ &= E[(CY_{(i)} - E(CY_{(i)}))(CY_{(j)} - E(CY_{(j)}))^T] \\ &= CE[(Y_{(i)} - Z\beta_{(i)})(Y_{(j)} - Z\beta_{(j)})^T]C^T \\ &= CE(\varepsilon_{(i)}\varepsilon_{(j)}^T)C^T \\ &= (Z^T Z)^{-1} Z^T (\sigma_{ij} I) ((Z^T Z)^{-1} Z^T)^T \\ &= \sigma_{ij} (Z^T Z)^{-1} Z^T ((Z^T Z)^{-1} Z^T)^T \\ &= \sigma_{ij} (Z^T Z)^{-1} ((Z^T Z)^{-1} Z^T Z)^T \\ &= \sigma_{ij} (Z^T Z)^{-1} \end{aligned}$$

Chứng minh hoàn tất.

Tính chất 8.4

Với $Y_{n \times m} = Z_{n \times (r+1)} \beta_{(r+1) \times m} + \varepsilon_{n \times m}$, $\text{rank}(Z) = r + 1$, $n \geq (r + 1) + m$ và $\varepsilon_{(i)}$ có phân phối chuẩn, $i = \overline{1, m}$ thì $[\hat{\beta}]$ có phân phối chuẩn.

Chứng minh: Ta có $\varepsilon_{(i)}$ có phân phối chuẩn. Khi đó: $\varepsilon_{(i)} \sim N_n(0, \sigma_{ii} I)$

$$\Rightarrow Y_{(i)} \sim N_n(Z\beta_{(i)}, \sigma_{ii} I)$$

$$\Rightarrow \hat{\beta}_{(i)} = (Z^T Z)^{-1} Z^T Y_{(i)} \sim N_{r+1}(\beta_{(i)}, \sigma_{ii} (Z^T Z)^{-1})$$

Vậy $[\hat{\beta}]$ có phân phối chuẩn.

(Tính chất này sẽ được thể hiện trong phần *Đưa ra dự đoán từ mô hình hồi quy tuyến tính đa bội* phía sau).

Tính chất 8.5

$\hat{\Sigma}$ là ước lượng chệch của Σ , với $\hat{\Sigma} = \hat{\varepsilon}^T \hat{\varepsilon} / n$.

Chứng minh: Ta chứng minh $E(\hat{\varepsilon}_{(j)}^T \hat{\varepsilon}_{(k)}) = \sigma_{jk}(n - r - 1)$, với mọi $j, k = \overline{1, m}$.

Ta gọi $H = Z(Z^T Z)^{-1} Z^T$. Khi đó:

$$\begin{aligned}
 E(\widehat{\varepsilon}_{(j)}^T \widehat{\varepsilon}_{(k)}) &= E\left[\left(Y_{(j)} - \widehat{Y}_{(j)}\right)^T \left(Y_{(k)} - \widehat{Y}_{(k)}\right)\right] \\
 &= E\left[\left(Y_{(j)} - HY_{(j)}\right)^T \left(Y_{(k)} - HY_{(k)}\right)\right] \\
 &= E\left[\left((I - H)Y_{(j)}\right)^T (I - H)Y_{(k)}\right] \\
 &= E\left[\left((I - H)(Z\beta_{(j)} + \varepsilon_{(j)})\right)^T (I - H)(Z\beta_{(k)} + \varepsilon_{(k)})\right] \\
 &= E\left[\left((I - H)\varepsilon_{(j)}\right)^T (I - H)\varepsilon_{(k)}\right] \\
 &= E\left(\varepsilon_{(j)}^T (I - H)^T (I - H)\varepsilon_{(k)}\right) \\
 &= E\left(\varepsilon_{(j)}^T (I - H)\varepsilon_{(k)}\right) \\
 &= E\left(\text{tr}\left((I - H)\varepsilon_{(k)}\varepsilon_{(j)}^T\right)\right) \quad (*) \\
 &= \text{tr}\left((I - H)E\left(\varepsilon_{(k)}\varepsilon_{(j)}^T\right)\right) \quad (**) \\
 &= \sigma_{jk} \text{tr}(I - H) = \sigma_{jk}(n - r - 1) \quad (***)
 \end{aligned}$$

Như vậy, $E(\widehat{\Sigma}) = (n - r - 1)\Sigma/n$ và do đó $\widehat{\Sigma}$ là ước lượng chệch của Σ .

(Các phép biến đổi (*), (**) và (***) có được từ áp dụng các kết quả 2A.12, 4.9 và 7.1 trong sách *Applied Multivariate Statistical Analysis*).

Từ kết quả trên, ta có thể suy ra được ước lượng không chệch của Σ là:

$$\widehat{\Sigma}^* = n\widehat{\Sigma}/(n - r - 1)$$

8.4 Kiểm định tỉ số hợp lý cho tham số hồi quy

Tương tự như mô hình cổ điển, giả thiết các biến phụ thuộc không phụ thuộc vào $z_{q+1}, z_{q+2}, \dots, z_r$ trở thành:

$$H_0: \beta_{(2)} = 0 \text{ với } \beta = \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix}$$

$\begin{matrix} ((q+1) \times m) \\ ((r-q) \times m) \end{matrix}$

Đặt $Z = \left[\begin{array}{c|c} Z_1 & Z_2 \\ \hline (n \times (q+1)) & (n \times (r-q)) \end{array} \right]$, viết lại mô hình thành:

$$E(Y) = Z\beta = [Z_1 | Z_2] \begin{bmatrix} \beta_{(1)} \\ \beta_{(2)} \end{bmatrix} = Z_1\beta_{(1)} + Z_2\beta_{(2)}$$

Khi giả thuyết $H_0 : \beta_{(2)} = 0$ đúng thì $Y = Z_1 \beta_{(1)} + \varepsilon$ và kiểm định tỉ số hợp lý cho H_0 được xây dựng dựa trên đại lượng phần tổng bình phương và tích vô hướng sai lệch:

$$\left(Y - Z_1 \hat{\beta}_{(1)}\right)^T \left(Y - Z_1 \hat{\beta}_{(1)}\right) - \left(Y - Z \hat{\beta}\right)^T \left(Y - Z \hat{\beta}\right) = n \left(\hat{\Sigma}_1 - \hat{\Sigma}\right)$$

$$\text{với } \hat{\beta}_{(1)} = (Z_1^T Z_1)^{-1} Z_1^T Y \text{ và } \hat{\Sigma}_1 = \frac{1}{n} \left(Y - Z_1 \hat{\beta}_{(1)}\right)^T \left(Y - Z_1 \hat{\beta}_{(1)}\right).$$

Xét tỷ số kiểm định Λ được định nghĩa:

$$\Lambda = \frac{\max_{\beta_{(1)}, \Sigma} L(\beta_{(1)}, \Sigma)}{\max_{\beta, \Sigma} L(\beta, \Sigma)} = \frac{L(\hat{\beta}_{(1)}, \hat{\Sigma}_1)}{L(\hat{\beta}, \hat{\Sigma})} = \left(\frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}\right)^{\frac{n}{2}}$$

$$\text{với } L(\hat{\beta}, \hat{\Sigma}) = (2\pi)^{-mn/2} |\hat{\Sigma}|^{-n/2} e^{-mn/2}.$$

Do đó, ta hoàn toàn có thể sử dụng thống kê Wilks' lambda:

$$\Lambda^{\frac{2}{n}} = \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}$$

Định lý 8.2

Mô hình hồi quy bội với Z có hạng đủ $\text{rank } Z = k + 1$, $n \geq (k + 1) + m$ và sai số ε có phân phối chuẩn.

Nếu $H_0 : \beta_{(2)} = 0$ đúng thì thống kê $n\hat{\Sigma}$ có phân phối $W_{p, n-k-1}(\Sigma)$ độc lập với thống kê $n(\hat{\Sigma}_1 - \hat{\Sigma})$ có phân phối $W_{p, k-q}(\Sigma)$. Khi đó, ta bác bỏ H_0 nếu biểu thức dưới đây đủ lớn:

$$-2 \ln \Lambda = -n \ln \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|} = -n \ln \frac{|n\hat{\Sigma}|}{|n\hat{\Sigma} + n(\hat{\Sigma}_1 - \hat{\Sigma})|}$$

Diễn giải: Ta hình dung điểm làm cực đại hàm hợp lý, gọi là ước lượng hợp lý cực đại, là điểm mà tại đó hiểu rằng ước lượng là hợp lý nhất, theo nghĩa là nếu tham số nhận giá trị ước lượng đó thì mô hình trên lý thuyết và dữ liệu ta có là phù hợp với nhau nhất. Do đó, ở đây muốn xác định $\beta_{(2)}$ có phải một giá trị hợp lý cho β hay không, ta so sánh cực đại của hàm $(\beta_{(1)}, \Sigma)$ với cực đại của hàm (β, Σ) , là trường hợp không bị giới hạn bởi H_0 .

Nếu H_0 đúng, thì rõ ràng sai lệch $n(\hat{\Sigma}_1 - \hat{\Sigma})$ phải đủ nhỏ, nghĩa là sự thay đổi này không ảnh hưởng đến mô hình. Do đó, ta bác bỏ H_0 nếu tỷ số trên là đủ lớn.

Khi n đủ lớn, thống kê hiệu chỉnh:

$$- \left[n - r - 1 - \frac{1}{2}(m - r + q + 1) \right] \ln \frac{|\hat{\Sigma}|}{|\hat{\Sigma}_1|}$$

có phân phối xấp xỉ phân phối Chi - bình phương $\chi_{m(r-q)}^2$.

Ví dụ 8.1

Kiểm định độ quan trọng của các biến dự báo bổ sung với nhiều biến phụ thuộc

Hệ thống phục vụ ở 3 địa điểm trong chuỗi nhà hàng lớn được đánh giá theo hai tiêu chí đánh giá chất lượng phục vụ, bởi khách hàng thân thiết nam và nữ. Giả sử, xét mô hình hồi quy cho phép các yếu tố là vị trí địa lí, giới tính, và sự tác động qua lại giữa 2 yếu tố này, ảnh hưởng đến cả hai bằng chất lượng phục vụ. Ma trận thiết kế ở ví dụ 7.1 không đổi với trường hợp 2 biến phụ thuộc. Ta sẽ kiểm định giả thiết sự tác động qua lại giữa 2 yếu tố này không ảnh hưởng đến 2 biến phụ thuộc. Chương trình máy tính cung cấp dữ liệu sau:

$$\begin{pmatrix} \text{Sai số mô hình} \\ \text{ban đầu} \end{pmatrix} = n\hat{\Sigma} = \begin{bmatrix} 2977,39 & 1021,72 \\ 1021,72 & 2050,95 \end{bmatrix}$$

$$\begin{pmatrix} \text{Sai số} \\ \text{chênh lệch} \end{pmatrix} = n(\hat{\Sigma}_1 - \hat{\Sigma}) = \begin{bmatrix} 441,76 & 246,16 \\ 246,16 & 366,12 \end{bmatrix}$$

Table 7.2 Restaurant-Service Data

Location	Gender	Service (Y)
1	0	15.2
1	0	21.2
1	0	27.3
1	0	21.2
1	0	21.2
1	1	36.4
1	1	92.4
2	0	27.3
2	0	15.2
2	0	9.1
2	0	18.2
2	0	50.0
2	1	44.0
2	1	63.6
3	0	15.2
3	0	30.3
3	1	36.4
3	1	40.9

$$\mathbf{Z} = \begin{array}{c} \begin{array}{cccccc} \text{constant} & \text{location} & \text{gender} & \text{interaction} & & \end{array} \\ \left[\begin{array}{cccccc} 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right] \end{array} \begin{array}{l} \left. \begin{array}{l} \\ \\ \\ \\ \end{array} \right\} 5 \text{ responses} \\ \left. \begin{array}{l} \\ \end{array} \right\} 2 \text{ responses} \\ \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} 5 \text{ responses} \\ \left. \begin{array}{l} \\ \end{array} \right\} 2 \text{ responses} \\ \left. \begin{array}{l} \\ \end{array} \right\} 2 \text{ responses} \\ \left. \begin{array}{l} \end{array} \right\} 2 \text{ responses} \end{array}$$

Đặt $\beta_{(2)}$ là ma trận sự tác động qua lại giữa vị trí địa lý và giới tính đối với 2 biến phụ thuộc. Ta sử dụng định lý trên để kiểm định $H_0 : \beta_{(2)} = 0$.

Lấy $\alpha = 0.05$, ta có:

$$\begin{aligned} & - \left[n - r_1 - 1 - \frac{1}{2}(m - r_1 + q_1 + 1) \right] \ln \frac{|n\hat{\Sigma}|}{|n\hat{\Sigma} + n(\hat{\Sigma}_1 - \hat{\Sigma})|} \\ & = - \left[18 - 5 - 1 - \frac{1}{2}(2 - 5 + 3 + 1) \right] \ln(0,7605) = 3,28 \end{aligned}$$

Để thấy vì $3,28 < \chi^2_{2(5-3)}(0,05) = 9,49$ nên không bác bỏ H_0 với mức ý nghĩa 5%. Suy ra vị trí địa lý và giới tính không cần phải có tác động lẫn nhau.

8.5 Các thống kê nhiều chiều khác

Đặt E là ma trận tổng bình phương và tích vô hướng các sai số cỡ $p \times p$

$$E = n\hat{\Sigma}$$

Ta có H là ma trận tổng bình phương và tích vô hướng của phần dư

$$H = n(\hat{\Sigma}_1 - \hat{\Sigma})$$

Các thống kê có thể được định nghĩa trực tiếp theo E và H , hoặc các trị riêng khác 0 của HE^{-1} là $\eta_1 \geq \eta_2 \geq \dots \geq \eta_s$, với $s = \min(p, k - q)$, chính là nghiệm của phương trình $|(\hat{\Sigma}_1 - \hat{\Sigma}) - \eta\hat{\Sigma}| = 0$. Từ

đây ta có các định nghĩa:

- Wilks' lambda $= \prod_{i=1}^s \frac{1}{1 + \eta_i} = \frac{|E|}{|E + H|}$
- Pillai's trace $= \sum_{i=1}^s \frac{\eta_i}{1 + \eta_i} = \text{tr} [H(H + E)^{-1}]$
- Hotelling-Lawley trace $= \sum_{i=1}^s \eta_i = \text{tr}(HE^{-1})$
- Roy's greatest root $= \frac{\eta_1}{1 + \eta_1}$

8.6 Đưa ra dự đoán từ mô hình hồi quy tuyến tính đa bội

Xét mô hình $Y = Z\beta + \varepsilon$, ε có phân phối chuẩn, các tham số đã được ước lượng và tính chỉnh. Bài toán là dự đoán kỳ vọng của các biến phụ thuộc tương ứng với giá trị $z_0^T = [1, z_{01}, \dots, z_{0k}]$ cố định của các biến độc lập.

Ta có $\hat{\beta}^T z_0$ có phân phối $N_m(\beta^T z_0, z_0^T (Z^T Z)^{-1} z_0 \Sigma)$ và $n\hat{\Sigma}$ độc lập với $\hat{\beta}^T z_0$ và có phân phối $W_{n-r-1}(\Sigma)$. Giá trị chưa biết của hàm hồi quy tại z_0 là $\beta^T z_0$. Do đó, từ mục 5.2 (sách *Applied Multivariate Statistical Analysis*) về thống kê T^2 , ta có thể viết:

$$T^2 = \left(\frac{\hat{\beta}^T z_0 - \beta^T z_0}{\sqrt{z_0^T (Z^T Z)^{-1} z_0}} \right)^T \left(\frac{n}{n-r-1} \hat{\Sigma} \right)^{-1} \left(\frac{\hat{\beta}^T z_0 - \beta^T z_0}{\sqrt{z_0^T (Z^T Z)^{-1} z_0}} \right)$$

và miền ellipsoid với độ tin cậy $100(1 - \alpha)\%$ của $\beta^T z_0$ bởi bất phương trình:

$$\begin{aligned} & \left(\beta^T z_0 - \hat{\beta}^T z_0 \right)^T \left(\frac{n}{n-r-1} \hat{\Sigma} \right)^{-1} \left(\beta^T z_0 - \hat{\beta}^T z_0 \right) \\ & \leq z_0^T (Z^T Z)^{-1} z_0 \left[\frac{m(n-r-1)}{n-r-m} F_{m, n-r-m}(\alpha) \right] \end{aligned}$$

với $F_{m, n-r-m}(\alpha)$ là phân vị trên mức ý nghĩa 100α của phân phối Fisher với m và $n-r-m$ bậc tự do.

Ta có khoảng tin cậy đồng thời với độ tin cậy $100(1 - \alpha)\%$ cho $E(Y_i) = z_0^T \beta_{(i)}$ là:

$$z_0^T \hat{\beta}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m} F_{m, n-r-m}(\alpha)} \sqrt{z_0^T (Z^T Z)^{-1} z_0 \left(\frac{n}{n-r-1} \hat{\sigma}_{ii} \right)} \text{ với } i = \overline{1, m}$$

với $\hat{\beta}_{(i)}$ là cột thứ i của $\hat{\beta}$ và $\hat{\sigma}_{ii}$ là thành phần thứ i trên đường chéo của $\hat{\Sigma}$.

Bài toán tiếp theo là dự đoán giá trị vectơ quan sát mới $Y_0 = \beta^T z_0 + \varepsilon_0$ tại z_0 trong đó ε_0 độc lập với các vector trong ε .

Ta có $Y_0 - \hat{\beta}^T z_0 = (\beta - \hat{\beta})^T z_0 + \varepsilon_0$ có phân phối $N_m(0, (1 + z_0^T (Z^T Z)^{-1} z_0) \Sigma)$ độc lập đối với $n\hat{\Sigma}$, do đó miền ellipsoid dự đoán với độ tin cậy $100(1 - \alpha)\%$ cho Y_0 trở thành:

$$\begin{aligned} & \left(Y_0 - \hat{\beta}^T z_0 \right)^T \left(\frac{n}{n-r-1} \hat{\Sigma} \right)^{-1} \left(Y_0 - \hat{\beta}^T z_0 \right) \\ & \leq (1 + z_0^T (Z^T Z)^{-1} z_0) \left[\frac{m(n-r-1)}{n-r-m} F_{m, n-r-m}(\alpha) \right] \end{aligned}$$

Khoảng tin cậy đồng thời với độ tin cậy $100(1 - \alpha)\%$ cho từng Y_{0i} là:

$$z_0^T \hat{\beta}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m} F_{m, n-r-m}(\alpha)} \sqrt{(1 + z_0^T (Z^T Z)^{-1} z_0) \left(\frac{n}{n-r-1} \hat{\sigma}_{ii} \right)} \text{ với } i = 1, 2, \dots, m$$

Nhận xét: So sánh hai công thức, ta thấy khoảng dự đoán giá trị chính xác rộng hơn khoảng dự đoán giá trị kì vọng, bởi sự xuất hiện của sai số ngẫu nhiên ε_{0i} .

FaMI
1956



FaMI
1956

Kết luận

Qua quá trình làm việc nhóm tìm hiểu và nghiên cứu về chương "**Mô hình hồi quy tuyến tính bội**" chúng em đã rút ra được một số nhận xét sau:

- Nội dung của bài báo cáo đã trình bày được các khái niệm và kiến thức cơ bản về mô hình hồi quy tuyến tính, cách xây dựng mô hình và các ví dụ minh họa đi kèm theo từng phần.
- Bài báo cáo đã giới thiệu về mô hình hồi quy tuyến tính đa biến, có chương trình sử dụng code viết bằng mã Python để phân tích.
- Do khả năng về kiến thức còn hạn chế nên nhóm vẫn chưa đi tìm hiểu được sâu, được rộng về tất cả các mô hình hồi quy. Trong tương lai, nhóm sẽ cố gắng tìm hiểu, nghiên cứu kỹ hơn về các mô hình hồi quy nói riêng và các kỹ thuật phân tích dữ liệu nói chung.

Nhóm 7 kính mong sẽ nhận được sự góp ý từ thầy để có thể hoàn thiện chủ đề cũng như hoàn thiện kiến thức của mình hơn nữa.

Chúng em xin chân thành cảm ơn!

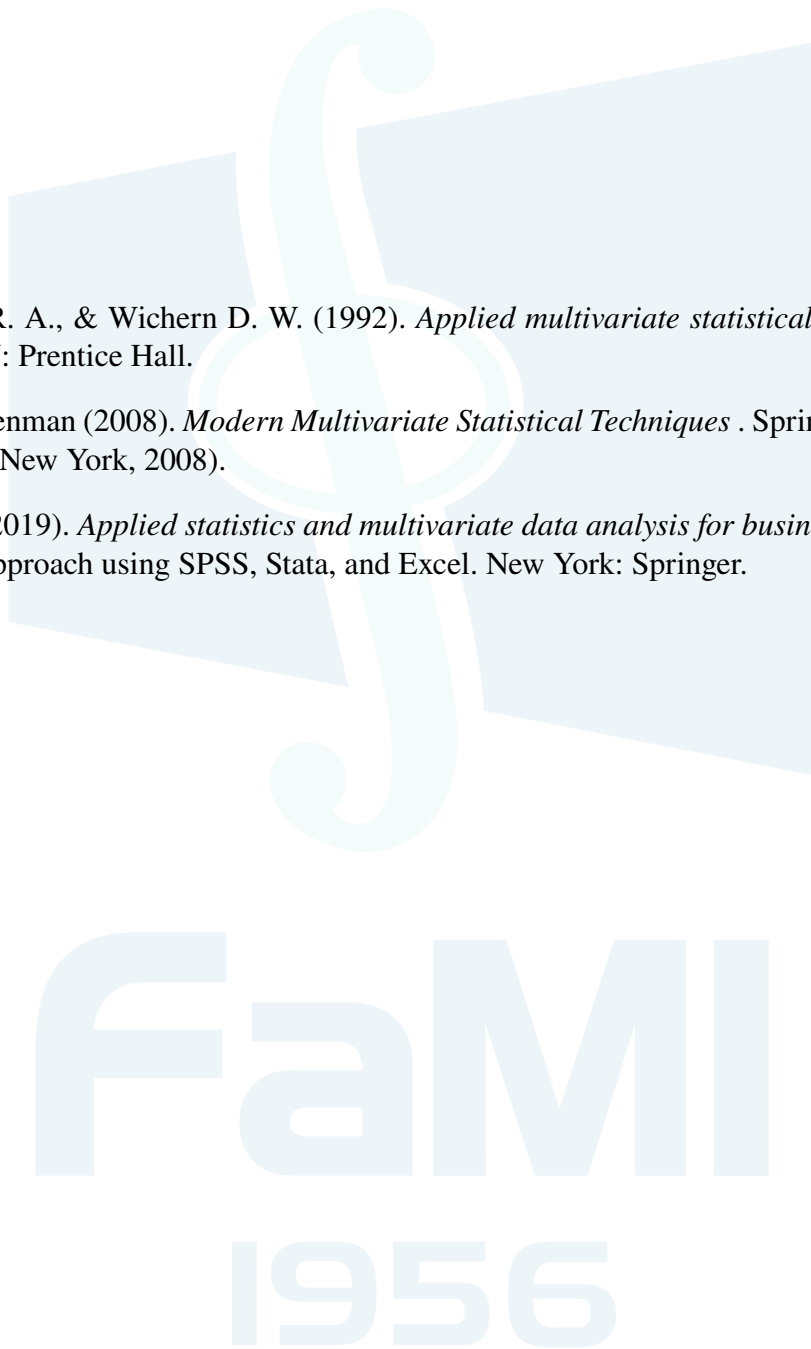
FaMI
1956



FaMI

1956

Tài liệu tham khảo

- 
- [1] Johnson R. A., & Wichern D. W. (1992). *Applied multivariate statistical analysis*. Englewood Cliffs, N.J: Prentice Hall.
- [2] Alan J. Izenman (2008). *Modern Multivariate Statistical Techniques* . Springer Texts in Statistics (Springer New York, 2008).
- [3] Cleff T. (2019). *Applied statistics and multivariate data analysis for business and economics: A modern approach using SPSS, Stata, and Excel*. New York: Springer.