Mô hình hồi quy tuyến tính

MI4020 - Phân tích số liệu

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

Nguyễn Đức Ánh	20204811
Nguyễn Việt Anh	20200039
Nguyễn Bảo Anh	20206110
Nguyễn Sỹ Huân	20200253
Nguyễn Hoàng Nhật	20204772

Mục lục

Lời nói đầu

Qua quá trình học tập và nghiên cứu môn học **Phân tích số liệu** dưới sự hướng dẫn của thầy Lê Xuân Lý, nhóm đã được tìm hiểu về mô hình *Hồi quy tuyến tính* để từ đó thu nạp được những kiến thức đầy thú vị cũng như được biết thêm những ứng dụng thực tiễn của mô hình này. Bài báo cáo này chính là sự kết tinh cô đọng những kiến thức đó, được các thành viên trong nhóm cùng nhau viết nên.

Trong quá trình thực hiện việc viết báo cáo cho chủ đề *Hồi quy tuyến tính* này, các thành viên trong nhóm có gặp phải một số vướng mắc về kiến thức, tuy nhiên nhờ tinh thần đoàn kết của nhóm, các thành viên tích cực cùng nhau tham gia những buổi họp để giải đáp thắc mắc cho nhau. Chính vì thế, bài báo cáo này là công sức công bằng của các thành viên trong nhóm, dựa trên sự công bằng trong việc phân chia các phần việc, cụ thể:

- Phần 2, 3: Nguyễn Sỹ Huân
- Phần 4: Nguyễn Hoàng Nhật
- Phần 5, chương trình minh họa: Nguyễn Bảo Anh
- Phần 6: Nguyễn Đức Ánh
- Phần 7: Nguyễn Việt Anh

Báo cáo này cho dù đã được kiểm tra lại về mặt nội dung, chính tả, ký hiệu,... nhưng cũng không thể tránh khỏi sai sót. Những đóng góp của thầy và bạn đọc sẽ là những đóng góp quý giá giúp nhóm có thể cải thiện kiến thức cũng như cải thiện bản thân mình.

Hà Nội, tháng 2 năm 2023
Nhóm trưởng - *Nguyễn Hoàng Nhật*
Nguyễn Bảo Anh
Nguyễn Việt Anh
Nguyễn Đức Ánh
Nguyễn Sỹ Huân

1 Giới thiệu về hồi quy

Hồi quy là một phương pháp thống kê được sử dụng để kiểm tra mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Mục tiêu của phân tích hồi quy là tìm ra một phương trình toán học mô tả mối quan hệ giữa các biến. Phương pháp này được sử dụng trong nhiều lĩnh vực, bao gồm tài chính, y tế, khoa học xã hội và kỹ thuật. Nó được sử dụng để mô hình hóa và dự đoán một loạt các hiện tượng, chẳng hạn như giá cổ phiếu, hành vi của khách hàng, tiến triển của bệnh tật, giá cả nhà đất và biến đổi khí hậu.

Hồi quy tuyến tính là một công cụ đơn giản nhưng mạnh mẽ đã được áp dụng rộng rãi trong nhiều lĩnh vực vì tính đơn giản và dễ diễn giải của nó. Mục tiêu của hồi quy tuyến tính là tìm một đường thẳng hoặc đường cong phù hợp nhất với các điểm dữ liệu và có thể được sử dụng để đưa ra dự đoán. Đường thẳng hoặc đường cong được gọi là đường hồi quy và quá trình tìm đường phù hợp nhất được gọi là phân tích hồi quy.

Trong báo cáo này, chúng ta sẽ tìm hiểu chi tiết về khái niệm mô hình hồi quy tuyến tính đơn bội, đa bội, sử dụng các giả định làm nền tảng cho kỹ thuật và các bước liên quan đến việc tiến hành phân tích hồi quy.

Một số mô hình hồi quy

Các dạng mô hình hồi quy tuyến tính thường được đề cập tới bao gồm:

- Hồi quy đơn tuyến tính

$$Y = \beta_0 + \beta_1 z + \epsilon$$

- Hồi quy đa tuyến tính

$$Y = \beta_0 + \beta_1 z_1 + \cdots + \beta_n z_n + \varepsilon$$

- Hồi quy tuyến tính đa bội

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

Ngoài ra còn có một số dạng hồi quy tuyến tính khác. Trong bài báo cáo này, ta chủ yếu xét đến hồi quy đa tuyến tính và hồi quy tuyến tính đa bội.

2 Mô hình hồi quy tuyến tính cổ điển

Mô hình hồi quy tuyến tính cổ điển với đơn biến phản hồi có dạng:

$$Y = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r + \varepsilon$$

trong đó:

- Y: biến phản hồi;
- z_i: biến dự đoán;
- β_i: tham số (hệ số hồi quy);
- ε: nhiễu ngẫu nhiên, với giả thiết được đặt ra là $E(\varepsilon) = 0$, $\text{Var}(\varepsilon) = \sigma^2$.

Với n quan sát độc lập Y_i và giá trị của các biến dự đoán tương ứng là z_{ij} $(i = \overline{1,n}, j = \overline{1,r})$:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 z_{11} + \beta_2 z_{12} + \cdots + \beta_r z_{1r} + \varepsilon_1 \\ Y_2 &= \beta_0 + \beta_1 z_{21} + \beta_2 z_{22} + \cdots + \beta_r z_{2r} + \varepsilon_2 \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ Y_n &= \beta_0 + \beta_1 z_{n1} + \beta_2 z_{n2} + \cdots + \beta_r z_{nr} + \varepsilon_n \end{aligned}$$

ta đặt ra giả thiết sau

- $E(\varepsilon_j) = 0$;
- $\text{Var}(\varepsilon_j) = \sigma^2$;
- $\text{Cov}(\varepsilon_j, \varepsilon_k) = 0, j \neq k$.

Ta biểu diễn dưới dạng ma trận như sau

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & \cdots & z_{1r} \\ 1 & z_{21} & z_{22} & \cdots & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & \cdots & z_{nr} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

hay

$$\underset{n\times 1}{\boldsymbol{Y}} = \underset{n\times(r+1)}{\boldsymbol{Z}} \underset{(r+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\varepsilon}},$$

với $\boldsymbol{\varepsilon}$ thỏa mãn:

- $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$;
- $\text{Cov}(\boldsymbol{\varepsilon}) = E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2 \boldsymbol{I}$.

3 Ước lượng bình phương cực tiểu hệ số hồi quy

Mục tiêu của việc phân tích hồi quy là đưa ra công thức biểu diễn mối quan hệ giữa biến phản hồi và các biến dự đoán. Cụ thể trong hồi quy tuyến tính, ta cần tìm mối quan hệ tuyến tính giữa biến phản hồi và biến dự đoán. Để có thể làm được điều này, ta phải ước lượng giá trị của các hệ số hồi quy β_i và tìm phương sai σ^2 của sai số từ bộ dữ liệu. Một cách ước lượng đó là sử dụng ước lượng bình phương cực tiểu. Xét sai số

$$e_i \coloneqq y_i - \beta_0 - \beta_1 z_{i1} - \cdots - \beta_r z_{ir},$$

việc bây giờ cần làm là tìm giá trị thử nghiệm $\widehat{\boldsymbol{\beta}}$ để cực tiểu hóa sai số này. Xét giá trị tổng bình phương sai số:

$$\mathrm{RSS}(\boldsymbol{\beta}) \coloneqq \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 z_{i1} - \cdots - \beta_r z_{ir})^2 = (\boldsymbol{y} - \boldsymbol{Z\beta})'(\boldsymbol{y} - \boldsymbol{Z\beta})$$

Khi đó

$$\widehat{\boldsymbol{\beta}} \coloneqq \underset{\beta}{\operatorname{argmin}} (\boldsymbol{y} - \boldsymbol{Z\beta})'(\boldsymbol{y} - \boldsymbol{Z\beta})$$

được gọi là *ước lượng bình phương cực tiểu* của $\boldsymbol{\beta}$.

3.1 Biểu diễn bằng biến đổi đại số

Xét hàm vector

$$\mathrm{RSS}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{Z\beta})'(\boldsymbol{y} - \boldsymbol{Z\beta})$$

Giá trị $\boldsymbol{\beta}$ để cực tiểu hóa $\mathrm{RSS}(\boldsymbol{\beta})$ là nghiệm của phương trình đạo hàm bằng 0

$$\frac{\partial}{\partial \boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) = \frac{\partial}{\partial \boldsymbol{\beta}} (\boldsymbol{Z\beta} - \boldsymbol{y})'(\boldsymbol{Z\beta} - \boldsymbol{y}) = 2\boldsymbol{Z}'(\boldsymbol{Z\beta} - \boldsymbol{y})$$

Khi đó, bởi vì $\boldsymbol{Z}'$ không là ma trận không nên $\boldsymbol{Z}'\boldsymbol{Z\beta} - \boldsymbol{Z}'\boldsymbol{y} = \boldsymbol{0}$ và ta kết luận được

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$

Lưu ý: Ma trận $\boldsymbol{Z}$ đầy hạng thì ta mới có thể tính nghịch đảo của $\boldsymbol{Z}'\boldsymbol{Z}$. Trường hợp không đầy hạng ta cần dùng giả nghịch đảo (*pseudo inverse*). Trong bài báo cáo này sẽ không xét tới trường hợp cần sử dụng đến giả nghịch đảo, bạn đọc quan tâm có thể tìm hiểu theo từ khóa trên.

3.2 Biểu diễn bằng hình học

Ta có kỳ vọng của vector biến phản hồi:

$$E(\boldsymbol{Y}) = \boldsymbol{Z\beta} = \beta_0 \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} + \beta_1 \begin{bmatrix} z_{11} \\ z_{21} \\ \vdots \\ z_{n1} \end{bmatrix} + \cdots + \beta_r \begin{bmatrix} z_{1r} \\ z_{2r} \\ \vdots \\ z_{nr} \end{bmatrix}$$

hay ta có $E(\boldsymbol{Y})$ là tổ hợp tuyến tính của các vector cột của $\boldsymbol{Z}$.

Kể từ đây ta đặt

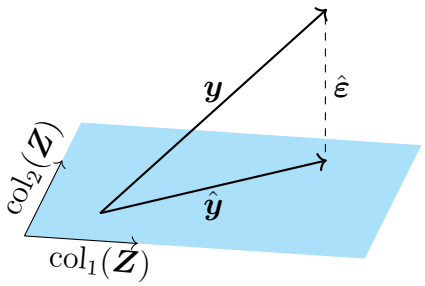
$$\widehat{\boldsymbol{y}} := \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$$
$$\widehat{\boldsymbol{\varepsilon}} := \boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$$
$$(P) := \operatorname{span}\left\{\operatorname{col}_k(\boldsymbol{Z})\right\}_{k=1}^{r+1}$$

Ta sẽ chọn chuẩn Euclide $\|.\|_2$ để biểu diễn các khoảng cách trong không gian. Khi đó $\min \|\varepsilon\|_2^2 = \|\hat{\varepsilon}\|_2^2$ khi và chỉ khi $\widehat{\boldsymbol{y}} = \boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \underset{(P)}{\operatorname{proj}}(\boldsymbol{y})$. Do đó $\widehat{\varepsilon} \perp \operatorname{col}_k(\boldsymbol{Z})$, suy ra

$$\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}) = \boldsymbol{Z}'\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$$

Nếu $\operatorname{rank}(\boldsymbol{Z}) = r + 1$ (đầy hạng) thì

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}$$



Hình 1: Biểu diễn hình học

3.3 Một số kết quả

Định lý 3.1

Xét mô hình hồi quy tuyến tính $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$ với $E(\boldsymbol{\varepsilon}) = 0$ và $\operatorname{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$. Với ma trận $\boldsymbol{Z}$ đầy hạng, $\operatorname{rank}(\boldsymbol{Z}) = r + 1 \leq n$, ước lượng bình phương cực tiểu của $\boldsymbol{\beta}$ là:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y}.$$

Chứng minh:

Đặt $\boldsymbol{H} \coloneqq Z(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ và $\widehat{\boldsymbol{y}} \coloneqq \boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}.$, ta sẽ chứng minh $\boldsymbol{I} - \boldsymbol{H}$ là ma trận đối xứng lũy đẳng.

- Tính đối xứng

$$(\boldsymbol{I} - \boldsymbol{H})' = \left(\boldsymbol{I} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\right)' = \left(\boldsymbol{I} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\right) = \boldsymbol{I} - \boldsymbol{H}$$

- Tính lũy đẳng

$$\begin{aligned}(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}) &= \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{H}^2 \\ &= \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' \\ &= \boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' \\ &= \boldsymbol{I} - \boldsymbol{H}\end{aligned}$$

Ngoài ra, ta cũng có

$$\boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H}) = \boldsymbol{Z}'\left(\boldsymbol{I} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\right) = \boldsymbol{Z}' - \boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}' = \boldsymbol{Z}' - \boldsymbol{Z}' = \boldsymbol{0}$$

$$\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}} = \boldsymbol{y} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

Từ đó suy ra được $\boldsymbol{Z}'\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Z}'(\boldsymbol{y} - \widehat{\boldsymbol{y}}) = \boldsymbol{Z}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{0}$ và $\widehat{\boldsymbol{y}}'\widehat{\boldsymbol{\varepsilon}} = \widehat{\boldsymbol{\beta}}'\boldsymbol{Z}'\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{0}$.

Để chứng minh công thức ước lượng bình phương cực tiểu đã nêu ở trên, ta biến đổi như sau:

$$\begin{aligned}\text{RSS}(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'Z'\boldsymbol{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + 2(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})'\boldsymbol{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\end{aligned}$$

Từ đây có thể thấy rằng $\text{RSS}(\boldsymbol{\beta})$ đạt giá trị nhỏ nhất khi và chỉ khi hạng tử $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ đạt giá trị nhỏ nhất. Trong khi đó

$$(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{Z}'\mathbf{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\mathbf{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}))'\mathbf{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \geq 0$$

và ma trận Z đầy hạng nên $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{Z}'\boldsymbol{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ đạt giá trị nhỏ nhất là 0 khi và chỉ khi $\boldsymbol{Z}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \boldsymbol{0}$ hay $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. □

Như vậy, ở **Định lý 3.1**, chúng ta đã xây dựng được công thức ước lượng bình phương cực tiểu của mô hình hồi quy tuyến tính cổ điển. Tiếp theo ta sẽ tìm hiểu về các tính chất của ước lượng trên.

Định lý 3.2

Xét mô hình hồi quy tuyến tính cổ điển $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$, ước lượng $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'Y}$ thỏa mãn

$$E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$$
$$\text{Cov}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z'Z})^{-1}$$

Sai số $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$ thỏa mãn

$$E(\widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{0}$$
$$\text{Cov}(\widehat{\boldsymbol{\varepsilon}}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$$

Hơn nữa, $\text{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{0}$

Chứng minh:

Trước hết, ta có các biến đổi sau:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'Y} = \left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'}(\boldsymbol{Z}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{\beta} + \left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'\varepsilon}$$
$$\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{Z\beta} + \boldsymbol{\varepsilon}) = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}$$

Kết hợp với một số kết quả đã xuất hiện trong quá trình chứng minh định lý 3.1, ta suy ra được

$$\begin{aligned}
E(\widehat{\boldsymbol{\beta}}) &= \boldsymbol{\beta} + \left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'}E(\boldsymbol{\varepsilon}) = \boldsymbol{\beta} \\
\text{Cov}(\widehat{\boldsymbol{\beta}}) &= \left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'}\,\text{Cov}(\boldsymbol{\varepsilon})\boldsymbol{Z}\left(\boldsymbol{Z'Z}\right)^{-1} \\
&= \boldsymbol{\sigma}^2\left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'Z}\left(\boldsymbol{Z'Z}\right) = \boldsymbol{\sigma}^2\left(\boldsymbol{Z'Z}\right)^{-1} \\
E(\widehat{\boldsymbol{\varepsilon}}) &= (\boldsymbol{I} - \boldsymbol{H})E(\boldsymbol{\varepsilon}) = 0 \\
\text{Cov}(\widehat{\boldsymbol{\varepsilon}}) &= (\boldsymbol{I} - \boldsymbol{H})\,\text{Cov}(\boldsymbol{\varepsilon})(\boldsymbol{I} - \boldsymbol{H})' = \boldsymbol{\sigma}^2(\boldsymbol{I} - \boldsymbol{H}) \\
\text{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\varepsilon}}) &= E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\widehat{\boldsymbol{\varepsilon}}'\right) = \left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'}E\left(\boldsymbol{\varepsilon\varepsilon'}\right)(\boldsymbol{I} - \boldsymbol{H}) \\
&= \boldsymbol{\sigma}^2\left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'}(\boldsymbol{I} - \boldsymbol{H}) = 0
\end{aligned}$$

□

Định lý 3.3 ***(Định lý Gauss về bình phương cực tiểu)***

Xét mô hình hồi quy tuyến tính $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$ với $E(\boldsymbol{\varepsilon}) = \boldsymbol{0}$ và $\text{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}$, ma trận Z đầy hạng. Với ước lượng bình phương cực tiểu $\widehat{\boldsymbol{\beta}}$ của $\boldsymbol{\beta}$ thì $\boldsymbol{c'}\widehat{\boldsymbol{\beta}}$ là ước lượng không chệch của $\boldsymbol{c'\beta}$ và có phương sai nhỏ nhất so với bất kỳ ước lượng tuyến tính không chệch nào khác có dạng

$$\boldsymbol{a'Y} = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n$$

Chứng minh:

Với mỗi giá trị $\boldsymbol{c}$ cố định, xét $\boldsymbol{a}'\boldsymbol{Y}$ là một ước lượng không chệch của $c'\boldsymbol{\beta}$. Khi đó với mọi giá trị $\boldsymbol{\beta}$ thì

$$E\left(\boldsymbol{a}'\boldsymbol{Y}\right) = \boldsymbol{c}'\boldsymbol{\beta}$$

Bên cạnh đó,

$$E\left(\boldsymbol{a}'\boldsymbol{Y}\right) = E\left(\boldsymbol{a}'\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{a}'\boldsymbol{\varepsilon}\right) = \boldsymbol{a}'\boldsymbol{Z}\boldsymbol{\beta}$$

Từ hai công thức trên, ta suy ra được $\boldsymbol{a}'\boldsymbol{Z}\boldsymbol{\beta} = \boldsymbol{c}'\boldsymbol{\beta}$, hoặc có thể viết lại thành $\left(\boldsymbol{c}' - \boldsymbol{a}'\boldsymbol{Z}\right)\boldsymbol{\beta} = 0$ với mọi $\boldsymbol{\beta}$ (Bao gồm cả giá trị $\boldsymbol{\beta} = \left(\boldsymbol{c}' - \boldsymbol{a}'\boldsymbol{Z}\right)'$). Điều này có nghĩa là $\boldsymbol{c}' = \boldsymbol{a}'\boldsymbol{Z}$ với mọi ước lượng không chệch $\boldsymbol{a}'\boldsymbol{Y}$. Đặt $\boldsymbol{u} \coloneqq \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{c}$. Khi đó ta có

$$c'\widehat{\boldsymbol{\beta}} = c'\left(Z'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y} = u'\boldsymbol{Y}$$

Ta có $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ theo định lý 3.2, do đó $\boldsymbol{c}'\widehat{\boldsymbol{\beta}} = \boldsymbol{u}'\boldsymbol{Y}$ là một ước lượng không chệch của $\boldsymbol{c}'\boldsymbol{\beta}$. Như vậy với mọi vector $\boldsymbol{a}$ thỏa mãn điều kiện không chệch $\boldsymbol{c}' = \boldsymbol{a}'Z$ thì

$$\begin{aligned}
\operatorname{Var}\left(\boldsymbol{a}'\boldsymbol{Y}\right) &= \operatorname{Var}\left(\boldsymbol{a}'\boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{a}'\boldsymbol{\varepsilon}\right) \\
&= \operatorname{Var}\left(\boldsymbol{a}'\boldsymbol{\varepsilon}\right) \\
&= \sigma^2\boldsymbol{a}'\boldsymbol{a} \\
&= \sigma^2(\boldsymbol{a} - \boldsymbol{u} + \boldsymbol{u})(\boldsymbol{a} - \boldsymbol{u} + \boldsymbol{u}) \\
&= \sigma^2\left((\boldsymbol{a} - \boldsymbol{u})'(\boldsymbol{a} - \boldsymbol{u}) + \boldsymbol{u}'\boldsymbol{u}\right)
\end{aligned}$$

Vì $\boldsymbol{u}$ cố định nên $(\boldsymbol{a} - \boldsymbol{u})'(\boldsymbol{a} - \boldsymbol{u}) \geq 0$ với mọi $\boldsymbol{a}$. Khi đó $\operatorname{Var}\left(\boldsymbol{a}'\boldsymbol{Y}\right)$ đạt cực tiểu khi $\boldsymbol{a} = \boldsymbol{u}$, hay ước lượng không chệch có phương sai nhỏ nhất là

$$\boldsymbol{u}'\boldsymbol{Y} = \boldsymbol{c}'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y} = \boldsymbol{c}'\widehat{\boldsymbol{\beta}}.$$

□

3.4 Hệ số xác định R^2

Ta có $\boldsymbol{Z}'\widehat{\boldsymbol{\varepsilon}} = 0$ nên suy ra biến đổi sau:

$$\boldsymbol{y}'\boldsymbol{y} = (\widehat{\boldsymbol{y}} + \widehat{\boldsymbol{\varepsilon}})'(\widehat{\boldsymbol{y}} + \widehat{\boldsymbol{\varepsilon}}) = \widehat{\boldsymbol{y}}'\widehat{\boldsymbol{y}} + \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}$$

Với $\overline{y} = \frac{1}{n}\sum_{k=1}^{n} y_k$ và $\overline{\widehat{y}} = \frac{1}{n}\sum_{k=1}^{n} \widehat{y}_k$. Do cột đầu tiên của ma trận $\boldsymbol{Z}$ toàn số 1 nên:

$$0 = \mathbf{1}'\widehat{\boldsymbol{\varepsilon}} = \sum_{k=1}^{n}\widehat{\varepsilon}_k = \sum_{k=1}^{n} y_k - \sum_{k=1}^{n}\widehat{y}_k$$

Suy ra được $\overline{y} = \overline{\widehat{y}}$. Khi đó biểu thức trên có thể biến đổi thành:

$$\sum_{k=1}^{n}(y_k - \overline{y})^2 = \sum_{k=1}^{n}(\widehat{y}_k - \overline{y})^2 + \sum_{k=1}^{n}\widehat{\varepsilon}_k^2$$

Định nghĩa:

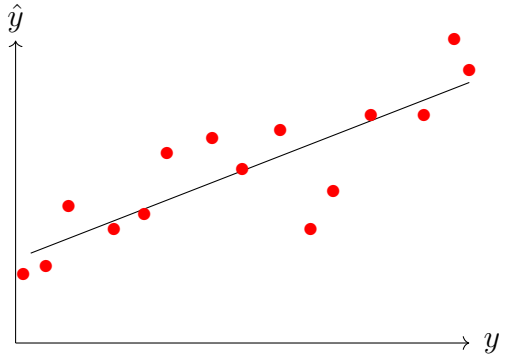
$$R^2 := \frac{\widehat{y}'\widehat{y} - n\overline{y}^2}{y'y - n\overline{y}^2} = 1 - \frac{\sum_{k=1}^{n} \widehat{\varepsilon}_k^2}{\sum_{k=1}^{n}(y_k - \overline{y})^2}$$

R^2 được gọi là hệ số xác định của mô hình hồi quy.

R^2 sẽ tăng khi số lượng biến dự đoán tăng, do đó hệ số này sẽ không phù hợp đối với bộ dữ liệu có nhiều điểm outlier. Để khắc phục hạn chế này, ta sử dụng giá trị $\boldsymbol{R^2}$ **hiệu chỉnh**:

$$R^2_{adj} := 1 - \frac{(1 - R^2)(n - 1)}{n - r - 1}$$

Ta chỉ sử dụng R^2_{adj} với một mẫu chứ không sử dụng đối với một tổng thể.



Hình 2: $R^2 = 0.7$

4 Suy luận về mô hình hồi quy

Phần này trình bày một vài suy luận từ mô hình hồi quy tuyến tính cổ điển được nhắc đến ở phần 3 với giả định rằng các sai số $\boldsymbol{\varepsilon}$ có phân phối chuẩn. Các phương pháp kiểm tra các tính chất của mô hình tổng quát sẽ được trình bày trong phần 6.

4.1 Một vài suy luận liên quan đến tham số hồi quy

Trước khi ta có thể đánh giá vai trò của các biến cụ thể trong hàm hồi quy:

$$E(Y) = \beta_0 + \beta_1 z_1 + \cdots + \beta_r z_r$$

ta cần xác định phân phối mẫu của $\widehat{\boldsymbol{\beta}}$ và thặng dư tổng bình phương $\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}$. Để làm được điều này, ta sẽ giả sử nhiễu $\boldsymbol{\varepsilon}$ có phân phối chuẩn.

Định lý 4.1

Xét $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, với $\boldsymbol{Z}$ có đầy hạng $r+1$ và $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Khi đó:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y} \quad \sim \quad \mathcal{N}_{r+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)$$

độc lập với $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{Y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$. Hơn nữa,

$$n\widehat{\sigma}^2 = \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} \quad \sim \quad \sigma^2\chi^2_{n-r-1}$$

với $\widehat{\sigma}^2$ là ước lượng hợp lý cực đại của σ^2.

Chứng minh:

Từ $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ và $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(0, \sigma^2\boldsymbol{I}\right) \Rightarrow \boldsymbol{Y} \sim \mathcal{N}_n\left(\boldsymbol{Z}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}\right)$.

Mặt khác, $\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{Y}$ và $\widehat{\boldsymbol{\varepsilon}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$ là các tổ hợp tuyến tính của $\boldsymbol{Y}$, do đó, $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{r+1}\left(\boldsymbol{\beta}, \sigma^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)$ và $\widehat{\boldsymbol{\varepsilon}} \sim \mathcal{N}_n\left(0, \sigma^2(\boldsymbol{I} - \boldsymbol{H})\right)$.

Từ định lý 3.2, ta có $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\varepsilon}}) = 0$, suy ra $\widehat{\boldsymbol{\beta}}$ và $\widehat{\boldsymbol{\varepsilon}}$ là hai biến độc lập.

Gọi λ là giá trị riêng ứng với vector riêng $\boldsymbol{v}$ của $\boldsymbol{I} - \boldsymbol{H}$, hay $(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{v} = \lambda\boldsymbol{v}$.

Khi đó

$$(\boldsymbol{I} - \boldsymbol{H})^2\boldsymbol{v} = \lambda(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{v} = \lambda^2\boldsymbol{v}. \tag{1}$$

Lại có, ở chứng minh của định lý 3.1, ta đã chứng minh được rằng $\boldsymbol{I} - \boldsymbol{H}$ là ma trận lũy đẳng, hay $(\boldsymbol{I} - \boldsymbol{H})^2\boldsymbol{v} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{v}$. Từ (1), ta có:

$$\lambda^2\boldsymbol{v} = \lambda\boldsymbol{v} \quad \text{hay} \quad \lambda \in \{0, 1\}. \tag{2}$$

Mặt khác, $\mathrm{tr}(\boldsymbol{I} - \boldsymbol{H}) = n - r - 1 = \lambda_1 + \lambda_2 + \cdots + \lambda_n$, từ (2) suy ra $\boldsymbol{I} - \boldsymbol{H}$ sẽ có $n - r - 1$ giá trị riêng bằng 1 và $r + 1$ giá trị riêng còn lại bằng 0 .

Giả sử $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_{n-r-1}$ là $n - r - 1$ vector riêng tương ứng với các trị riêng $\lambda_1 = \lambda_2 = \cdots = \lambda_{n-r-1} = 1$ của ma trận $\boldsymbol{I} - \boldsymbol{H}$. Khi đó ta có thể biểu diễn ma trận $\boldsymbol{I} - \boldsymbol{H}$ dưới dạng phân tách phổ

$$\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{v}_1\boldsymbol{v}_1' + \boldsymbol{v}_2\boldsymbol{v}_2' + \cdots + \boldsymbol{v}_{n-r-1}\boldsymbol{v}_{n-r-1}'$$

Đặt

$$\boldsymbol{V} \coloneqq \begin{bmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \vdots \\ \boldsymbol{V}_{n-r-1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{v}_1' \\ \boldsymbol{v}_2' \\ \vdots \\ \boldsymbol{v}_{n-r-1}' \end{bmatrix} \boldsymbol{\varepsilon}$$

Khi đó, $\boldsymbol{V}$ sẽ có phân phối chuẩn với $E(\boldsymbol{V}) = 0$ và

$$\mathrm{Cov}\left(\boldsymbol{V}_i, \boldsymbol{V}_j\right) = \boldsymbol{v}_i'\left(\sigma^2\boldsymbol{I}\right)\boldsymbol{v}_j' = \begin{cases} \sigma^2 & \text{nếu } i = j \\ 0 & \text{nếu } i \neq j \end{cases}.$$

Suy ra $\boldsymbol{V}_1, \boldsymbol{V}_2, \ldots, \boldsymbol{V}_{n-r-1}$ độc lập, có cùng phân phối chuẩn $\mathcal{N}\left(0, \sigma^2\right)$ và

$$\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{\varepsilon}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon} = \boldsymbol{V}'\boldsymbol{V} = \boldsymbol{V}_1^2 + \boldsymbol{V}_2^2 + \cdots + \boldsymbol{V}_{n-r-1}^2 \sim \sigma^2\chi^2_{n-r-1}$$

Vậy ta có điều phải chứng minh. □

Chú thích

Ta đặt $s^2 := \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-r-1}$, giá trị này là một ước lượng không chệch của σ^2.

Định lý 4.2

Mô hình hồi quy $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, với $\boldsymbol{Z}$ đầy hạng $r+1$ và $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2\boldsymbol{I}\right)$. Khi đó miền tin cậy mức $100(1-\alpha)\%$ của $\boldsymbol{\beta}$ được cho bởi ellipsolid:

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

trong đó $F_{r+1,n-r-1}(\alpha)$ là phân vị trên mức $100\alpha\%$ của phân phối Fisher với số bậc tự do là $r+1$ và $n-r-1$.
Đồng thời, khoảng tin cậy mức $100(1-\alpha)\%$ của β_i được xác định bởi

$$\widehat{\beta}_i \pm \sqrt{\widehat{\mathrm{Var}}\left(\hat{\beta}_i\right)}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, \quad i = 0, 1, \ldots, r$$

với $\widehat{\mathrm{Var}}\left(\hat{\beta}_i\right)$ là các phần tử thứ i trên đường chéo của ma trận $s^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}$.

Chứng minh:

Xét ma trận căn bậc hai đối xứng $\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{1/2}$. Đặt $\boldsymbol{V} := \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$. Khi đó, $E(\boldsymbol{V}) = 0$ và

$$\mathrm{Cov}(\boldsymbol{V}) = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{1/2}\mathrm{Cov}(\widehat{\boldsymbol{\beta}})\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{1/2} = \sigma^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{1/2}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{1/2} = \sigma^2\boldsymbol{I}_{r+1},$$

và V có phân phối chuẩn $\mathcal{N}_{r+1}\left(0, \sigma^2\boldsymbol{I}\right)$ vì V là tổ hợp tuyến tính của các $\widehat{\beta}_i$.

Do đó,

$$\boldsymbol{V}'\boldsymbol{V} = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{1/2}(\boldsymbol{Z}'\boldsymbol{Z})^{1/2}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{Z}'\boldsymbol{Z})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim \sigma^2\chi^2_{r+1}.$$

Theo Định lý 4.1, ta có $(n-r-1)s^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} \sim \sigma^2\chi^2_{n-r-1}$, độc lập với $\widehat{\boldsymbol{\beta}}$, hay độc lập với $\boldsymbol{V}$. Khi đó,

$$F := \frac{\chi^2_{r+1}/(r+1)}{\chi^2_{n-r-1}/(n-r-1)} = \frac{\boldsymbol{V}'\boldsymbol{V}}{s^2(r+1)} \sim F_{r+1,n-r-1}.$$

Từ đó ta được

$$P\left(F < F_{r+1,n-r-1}(\alpha)\right) = 1 - \alpha.$$

và miền tin cậy mức $100(1-\alpha)\%$ của $\boldsymbol{\beta}$ được xác định bởi ellipsoid

$$(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})' \boldsymbol{Z}' \boldsymbol{Z} (\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}) \leq (r+1) s^2 F_{r+1,n-r-1}(\alpha).$$

Với $i = \overline{0,r}$, từ bất đẳng thức trên ta được

$$\left(\beta_i - \widehat{\beta}_i\right)^2 \leq \widehat{\text{Var}}\left(\hat{\beta}_i\right)(r+1) F_{r+1,n-r-1}(\alpha).$$

Từ đó suy ra khoảng tin cậy đồng thời mức $100(1-\alpha)\%$ của các hệ số hồi quy $\beta_i (i = \overline{0,r})$ được xác định bởi

$$\widehat{\beta}_i \pm \sqrt{\widehat{\text{Var}}\left(\hat{\beta}_i\right)} \sqrt{(r+1) F_{r+1,n-r-1}(\alpha)}.$$

□

Lưu ý

Trong thực tiễn, người ta thường sử dụng công thức đơn giản hơn cho các ước lượng khoảng trong định lý 4.2. Người ta sử dụng khoảng

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

trong đó $t_{n-r-1}(\alpha/2)$ là phân vị trên mức $100(\alpha/2)\%$ của phân phối Student với $n-r-1$ bậc tự do.

Ví dụ 4.1

Dữ liệu dưới bảng 1 được thu thập từ 20 ngôi nhà ở Milwaukee, Wisconsin. Xét xem biến nào có thể loại bỏ trong mô hình hồi quy tuyến tính.

Giả sử mô hình hồi quy tuyến tính cổ điển khớp với dữ liệu giá nhà tại Milwaukee và Wisconsin là

$$Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \varepsilon_j \quad (j = \overline{1,20}),$$

với z_1 là tổng diện tích (ft^2), z_2 là giá ước tính (nghìn đô-la), và Y là giá bán thực tế (nghìn đô-la).

Bây giờ ta khớp dữ liệu trên vào mô hình bằng cách ước lượng hệ số hồi quy như đã trình bày ở phần trước.

Y Giá bán 1000$	z_1 Diện tích 100 ft^2	z_2 Giá ước tính 1000$	Y Giá bán 1000$	z_1 Diện tích 100 ft^2	z_2 Giá ước tính 1000$
74.8	15.31	57.3	71.5	15.18	62.6
74.0	15.20	63.8	71.0	14.44	63.4
72.9	16.25	65.4	78.9	14.87	60.2
70.0	14.33	57.0	86.5	18.63	67.2
74.9	14.57	63.8	68.0	15.20	57.1
76.0	17.33	63.2	102.0	25.76	89.6
72.0	14.48	60.2	84.0	19.05	68.6
73.5	14.91	57.7	69.0	15.37	60.1
74.5	15.25	56.4	88.0	18.06	66.3
73.5	13.89	55.6	76.0	16.35	65.8

Bảng 1: Bảng dữ liệu giá của 20 căn nhà

Thực hiện tính toán, ta có

$$\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1} = \begin{bmatrix} 5.15231 & 0.25443 & -0.14635 \\ 0.25443 & 0.05118 & -0.0172 \\ -0.14635 & -0.0172 & 0.00674 \end{bmatrix}$$

$$\text{và} \quad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'y = \begin{bmatrix} 30.967 \\ 2.634 \\ 0.045 \end{bmatrix}$$

Ta thu được phương trình hồi quy tuyến tính

$$\widehat{y} = 30.967 + 2.634z_1 + 0.045z_2,$$

và giá trị ước lượng không chệch s^2 của σ^2

$$s^2 = \frac{\widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}}}{n-r-1} = \frac{204.995}{17} \approx 12.0585,$$

trong đó, $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}$. Khi đó,

$$\begin{aligned} \left(s^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)_{11} &= 12.0585 \times 5.15231 \\ \left(s^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)_{22} &= 12.0585 \times 0.05118 \\ \left(s^2\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)_{33} &= 12.0585 \times 0.00674 \end{aligned}$$

Với $t_{17}(0.025) = 0.5098$, ta có khoảng tin cậy đồng thời mức 95% của các hệ số hồi quy là:

$$\begin{aligned} \beta_0 &= 30.967 \quad \pm 16.630 \\ \beta_1 &= 2.634 \quad \pm 1.657 \\ \beta_2 &= 0.045 \quad \pm 0.602 \end{aligned}$$

Ta tính được $R^2 \approx 0.834$, cho thấy bộ dữ liệu đã cho có quan hệ hồi quy khá mạnh, các biến được chọn giải thích được tới 83.4% giá bán.

Để ý rằng khoảng tin cậy mức 95% của β_2 ta tính toán được dựa theo công thức đơn giản hơn là

$$\hat{\beta}_2 \pm t_{17}(0.025)\sqrt{\widehat{\text{Var}}\left(\hat{\beta}_2\right)} = 0.045 \pm 2.110(0.285)$$

hay là khoảng

$$(-0.556, 0.647)$$

Do khoảng tin cậy có của β_2 có chứa 0, tức là biến z_2 có thể được loại bỏ khỏi mô hình hồi quy và quá trình phân tích được lặp lại với chỉ một biến dự đoán z_1. Lúc đó, khi đã biết diện tích thì giá trị thẩm định chỉ có ảnh hưởng rất nhỏ đến giá bán thực tế.

Ở mục 4.2 ta sẽ tìm hiểu cách để xác định xem liệu một hoặc nhiều biến dự đoán ra khỏi mô hình hồi quy hay không.

4.2 Kiểm định tỉ số hợp lý cho các tham số hồi quy

Một nhiệm vụ quan trọng khác trong quá trình phân tích hồi quy là đánh giá tác động của các biến dự đoán cụ thể đối với biến phản hồi. Một giả thuyết không có thể được xét đến đó là một vài biến dự đoán z_i có ảnh hưởng không đáng kể đến biến phản hồi Y. Những biến dự đoán này được đánh dấu là $z_{q+1}, z_{q+2}, \ldots, z_r$. Ta xét giả thuyết:

$$H_0 : \beta_{(q+1)} = \beta_{(q+2)} = \ldots = \beta_{(r)} = 0 \text{ hoặc } H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0},$$

với $\boldsymbol{\beta}'_{(2)} = [\beta_{q+1}, \beta_{q+2}, \ldots, \beta_r]$.

Mô hình hồi quy tuyến tính cổ điển $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$, với ma trận $\boldsymbol{Z}$ đầy hạng có thể viết lại dưới dạng

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon} = \left[\begin{array}{c:c} \underset{n\times(q+1)}{\boldsymbol{Z}_{(1)}} & \underset{n\times(r-q)}{\boldsymbol{Z}_{(2)}} \end{array}\right] \left[\begin{array}{c} \underset{(q+1)\times 1}{\boldsymbol{\beta}_{(1)}} \\ \hdashline \underset{(r-q)\times 1}{\boldsymbol{\beta}_{(2)}} \end{array}\right] + \boldsymbol{\varepsilon}$$

$$= \underset{n\times(q+1)}{\boldsymbol{Z}_{(1)}} \underset{(q+1)\times 1}{\boldsymbol{\beta}_{(1)}} + \underset{n\times(r-q)}{\boldsymbol{Z}_{(2)}} \underset{(r-q)\times 1}{\boldsymbol{\beta}_{(2)}} + \boldsymbol{\varepsilon}$$

Dưới giả thuyết không $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}, \boldsymbol{Y} = \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$. Kiểm định tỉ số hợp lý cho H_0 được dựa trên *tổng bình phương mở rộng*, là hiệu của hai tổng bình phương sai số của hai mô hình, cụ thể:

$$\begin{aligned} \textit{Tổng bình phương mở rộng} &= \text{RSS}\left(\boldsymbol{Z}_1\right) - \text{RSS}(\boldsymbol{Z}) \\ &= \left(\boldsymbol{y} - \boldsymbol{Z}_1\hat{\boldsymbol{\beta}}_{(1)}\right)' \left(\boldsymbol{y} - \boldsymbol{Z}_1\hat{\boldsymbol{\beta}}_{(1)}\right) - (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}). \end{aligned}$$

với $\hat{\boldsymbol{\beta}}_{(1)} = \left(\boldsymbol{Z}'_1\boldsymbol{Z}_1\right)^{-1}\boldsymbol{Z}'_1\boldsymbol{y}$.

Định lý 4.3

Xét ma trận $\boldsymbol{Z}$ đầy hạng $r+1$ và $\boldsymbol{\varepsilon} \sim \mathcal{N}_n\left(\boldsymbol{0}, \sigma^2 \boldsymbol{I}\right)$. Xét giả thuyết $H_0 : \boldsymbol{\beta}_{(2)} = \boldsymbol{0}$, khi đó H_0 bị bác bỏ khi:

$$\frac{\left(\text{RSS}\left(\widehat{\boldsymbol{\beta}}_1\right) - \text{RSS}(\widehat{\boldsymbol{\beta}})\right)/(r-q)}{s^2} > F_{r-q,n-r-1}(\alpha).$$

Với $F_{r-q,n-r-1}(\alpha)$ là phân vị trên mức $100\alpha\%$ của phân phối Fisher với $r-q$ và $n-r-1$ bậc tự do.

Chứng minh:

Xét hàm hợp lý của tham số $\boldsymbol{\beta}$ và σ^2

$$\mathcal{L}\left(\boldsymbol{\beta}, \sigma^2\right) = \frac{1}{(2\pi)^{n/2}\sigma^n} \exp\left(-\frac{(\boldsymbol{y}-\boldsymbol{Z\beta})'(\boldsymbol{y}-\boldsymbol{Z\beta})}{2\sigma^2}\right) \leq \frac{1}{(2\pi)^{n/2}\widehat{\sigma}^n} \exp\left(-\frac{n}{2}\right)$$

đạt cực đại đạt được tại

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\boldsymbol{y} \quad \text{và } \widehat{\sigma}^2 = \frac{(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}})'(\boldsymbol{y}-\boldsymbol{Z}\widehat{\boldsymbol{\beta}})}{n}$$

Nếu H_0 đúng thì $\boldsymbol{Y} = \boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{\varepsilon}$, hàm hợp lý của tham số $\boldsymbol{\beta}$ và σ_1^2 là

$$\begin{aligned}\mathcal{L}\left(\boldsymbol{\beta}, \sigma_1^2\right) &= \frac{1}{(2\pi)^{n/2}\sigma_1^n} \exp\left(-\frac{\left(\boldsymbol{y}-\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}\right)'\left(\boldsymbol{y}-\boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)}\right)}{2\sigma_1^2}\right) \\ &\leq \frac{1}{(2\pi)^{n/2}\widehat{\sigma}_1^n} e^{-n/2}.\end{aligned}$$

đạt được tại $\widehat{\boldsymbol{\beta}}_{(1)} = \left(\boldsymbol{Z}'_{(1)}\boldsymbol{Z}_{(1)}\right)^{-1}\boldsymbol{Z}'_{(1)}\boldsymbol{y}$

Hơn nữa

$$\widehat{\sigma}_1^2 = \frac{\left(\boldsymbol{y}-\boldsymbol{Z}_{(1)}\widehat{\boldsymbol{\beta}}_{(1)}\right)'\left(\boldsymbol{y}-\boldsymbol{Z}_{(1)}\widehat{\boldsymbol{\beta}}_{(1)}\right)}{n}$$

Ta bác bỏ giả thuyết $H_0 : \boldsymbol{\beta}_{(2)} = \boldsymbol{0}$ nếu tỉ số sau đủ nhỏ

$$\frac{\max_{\boldsymbol{\beta}_{(1)},\sigma^2} \mathcal{L}\left(\boldsymbol{\beta}_{(1)}, \sigma^2\right)}{\max_{\boldsymbol{\beta},\sigma^2} \mathcal{L}\left(\boldsymbol{\beta}, \sigma^2\right)} = \left(\frac{\widehat{\sigma}_1^2}{\widehat{\sigma}^2}\right)^{-n/2} = \left(\frac{\widehat{\sigma} + \widehat{\sigma}_1^2 - \widehat{\sigma}^2}{\widehat{\sigma}^2}\right)^{-n/2} = \left(1 + \frac{\widehat{\sigma}_1^2 - \widehat{\sigma}^2}{\widehat{\sigma}^2}\right)^{-n/2}$$

tương đương với ta sẽ bác bỏ giả thuyết H_0 nếu giá trị $\left(\widehat{\sigma}_1^2 - \widehat{\sigma}^2\right)/\widehat{\sigma}^2$ đủ lớn hoặc tỉ số sau đủ lớn:

$$\frac{n\left(\widehat{\sigma}_1^2 - \widehat{\sigma}^2\right)/(r-q)}{n\widehat{\sigma}^2/(n-r-1)} = \frac{\text{RSS}(\boldsymbol{\beta}(1)) - \text{RSS}(\boldsymbol{\beta})/(r-q)}{s^2} = F_{r-q,n-r-1}$$

Do đó,

$$P\left(\frac{\text{RSS}\left(\boldsymbol{\beta}_1\right)-\text{RSS}(\boldsymbol{\beta})}{(r-q)s^2} > F_{r-q,n-r-1}(\boldsymbol{\alpha})\right)=\alpha$$

Như vậy với mức ý nghĩa $100\alpha\%$, ta bác bỏ giả thuyết $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$ nếu

$$\frac{\text{RSS}\left(\boldsymbol{\beta}_1\right)-\text{RSS}(\boldsymbol{\beta})}{(r-q)s^2} > F_{r-q,n-r-1}(\alpha).$$

Nhận xét

Quy trình kiểm định tỉ số hợp lý được diễn ra như sau: Để kiểm tra xem một vài hệ số có thể bằng không, kiểm tra mô hình lúc có và không có những hệ số này. Sau đó sử dụng tỉ số F để so sánh sự khác biệt giữa tổng bình phương phần dư được so sánh với tổng bình phương phần dư của toàn bộ mô hình. Quy trình tương tự cũng được áp dụng khi phân tích trường hợp $\boldsymbol{Z}$ không đầy hạng.

Tổng quát hơn, có thể tạo ra các giả thuyết không liên quan đến $r-q$ tổ hợp tuyến tính của $\boldsymbol{\beta}$ có dạng: $H_0 : \boldsymbol{C\beta} = \boldsymbol{A}_0$. Giả sử ma trận $\boldsymbol{C}$ kích cỡ $(r-q)\times(r-1)$ đầy hạng, cho $\boldsymbol{A}_0 = 0$ và xét:

$$H_0 : \boldsymbol{C\beta} = \mathbf{0}.$$

Dưới mô hình đầy đủ, $\boldsymbol{C}\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{r-q}\left(\boldsymbol{C\beta}, \boldsymbol{\sigma}^2\boldsymbol{C}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{C}'\right)$. Ta bác bỏ H_0 ở mức ý nghĩa $100\alpha\%$ nếu

$$\frac{(\boldsymbol{C}\widehat{\boldsymbol{\beta}})'\left(\boldsymbol{C}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{C}'\right)^{-1}(\boldsymbol{C}\widehat{\boldsymbol{\beta}})}{s^2} > (r-q)F_{r-q,n-r-1}(\alpha),$$

Dưới đây, ta xé một ví dụ minh họa cho thấy cách một thí nghiệm không cân bằng có thể được xử lý bằng lý thuyết vừa được nêu ra.

Ví dụ 4.2

Một số người đàn ông và phụ nữ được yêu cầu bình chọn chất lượng dịch vụ ở ba chi nhánh của một chuỗi nhà hàng. Kết quả đánh giá được chuyển thành bảng chỉ số dưới đây (bảng 2), có 18 kết quả được ghi nhận. Cột địa điểm cho biết địa chỉ chi nhánh đang được khảo sát; cột giới tính cho biết người tham gia khảo sát: 0 là nam, 1 là nữ; cột cuối là điểm đánh giá. Có thể thấy, số lượng đánh giá ở mỗi địa điểm là khác nhau và số lượng người tham gia khảo sát ở mỗi giới tính tại mỗi địa điểm cũng khác nhau. Ta sẽ dùng ba biến giả đại diện cho địa điểm và hai biến giả đại diện cho giới tính, ta sẽ xây dựng mô hình hồi quy để chỉ ra mối quan hệ giữa chỉ số dịch vụ Y và địa điểm, giới tính và sự tương tác của họ sử dụng ma trận thiết kế.

Địa điểm	Giới tính	Dịch vụ (Y)	Địa điểm	Giới tính	Dịch vụ (Y)
1	0	15.2	2	0	9.1
1	0	21.2	2	0	18.2
1	0	27.3	2	0	50.0
1	0	21.2	2	1	44.0
1	0	21.2	2	1	63.6
1	1	36.4	3	0	15.2
1	1	92.4	3	0	30.3
2	0	27.3	3	1	36.4
2	0	15.2	3	1	40.9

Bảng 2: Dữ liệu nhà hàng-dịch vụ

Trong ma trận thiết kế dưới đây, mỗi hàng là một phản hồi của khách hàng, cột đầu tiên là hằng số 1, ba cột tiếp theo đại diện cho địa điểm, hai cột kế tiếp đó đại diện cho giới tính, sáu cột cuối là sự tương tác (mỗi cột xác định một cặp tương ứng địa điểm-giới tính).

$$Z = \left[\begin{array}{c|ccc|cc|cccccc}
\overbrace{\text{hằng số}}^{} & \overbrace{\text{địa điểm}}^{} & & & \overbrace{\text{giới tính}}^{} & & \overbrace{\text{tương tác}}^{} & & & & & \\
\hline
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\
\hdashline
1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
\hdashline
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
\hdashline
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\
\hdashline
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
\hdashline
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\
1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1
\end{array}\right]$$

Ta tạo véc tơ hệ số

$$\boldsymbol{\beta}' = [\beta_0, \beta_1, \beta_2, \beta_3, \tau_1, \tau_2, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \gamma_{31}, \gamma_{32}]$$

với các $\beta_i (i > 0)$ và τ_i thể hiện ảnh hưởng của biến vị trí đến điểm phục vụ và các γ_{ik} thể hiện ảnh hưởng của tương tác giữa vị trí và giới tính.

Ma trận $\boldsymbol{Z}$ không đầy hạng và ta xác định được $\text{rank}(\boldsymbol{Z}) = 6$.

Với mô hình đầy đủ, ta tính toán được

$$\text{RSS}(\boldsymbol{Z}) = 2977.4$$

và $n - \text{rank}(\boldsymbol{Z}) = 18 - 6 = 12$.

Mô hình nếu bỏ đi yếu tố tương tác sẽ có ma trận thiết kết là $\boldsymbol{Z}_1$, ma trận này gồm 6 cột đầu tiên của $\boldsymbol{Z}$. Ta tính được

$$\text{RSS}(\boldsymbol{Z}_1) = 3419.1$$

với $n - \text{rank}(\boldsymbol{Z}_1) = 18 - 4 = 14$. Để kiểm tra giả thuyết không

$$H_0 : \gamma_{11} = \gamma_{12} = \gamma_{21} = \gamma_{22} = \gamma_{31} = \gamma_{32} = 0$$

hay là kiểm định xem có bác bỏ được rằng yếu tố tương tác không ảnh hưởng đến mô hình hay không, ta tính toán

$$F = \frac{\left(\text{RSS}\left(\boldsymbol{Z}_1\right) - \text{RSS}(\boldsymbol{Z})\right)/(6-4)}{s^2} = \frac{\left(\text{RSS}\left(\boldsymbol{Z}_1\right) - \text{RSS}(\boldsymbol{Z})\right)/2}{\text{RSS}(\boldsymbol{Z})/12}$$
$$= \frac{(3419.1 - 2977.4)/2}{2977.4/12} = 0.89$$

Tỉ số F này không đủ lớn với mọi mức ý nghĩa α đủ lớn nên ta không thể bác bỏ H_0. Nghĩa là có thể kết luận được rằng điểm dịch vụ không phụ thuộc vào bất cứ tương tác nào giữa vị trí và giới tính, do đó yếu tố này có thể loại ra khỏi mô hình.

Trong các tình huống phân tích phương sai (ANOVA) trong đó số lượng thành phần không bằng nhau, sự khác biệt do các biến dự đoán khác nhau và tương tác của chúng thường không thể được tách thành các lượng độc lập. Để đánh giá ảnh hưởng tương đối của các biến dự đoán đối với biến phản hồi trong trường hợp này, ta cần phải điều chỉnh mô hình có và không có các yếu tố trên và tính toán kiểm định F (xem thêm tại [3]).

5 Vài suy luận từ ước lượng hàm hồi quy

Khi đã có một mô hình hồi quy tuyến tính, ta có thể sử dụng nó để ước lượng các dự đoán mới thông qua các quan sát mới.

Đặt $\boldsymbol{z}_0' = [1, z_{01}, \ldots, z_{0r}]$ là các giá trị của các biến dự đoán. Khi đó $\boldsymbol{z}_0$ và $\widehat{\boldsymbol{\beta}}$ có thể được sử dụng:

1. Để ước lượng hàm hồi quy $\beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r}$, tại $\boldsymbol{z}_0$
2. Để ước lượng giá trị của biến phản hồi Y tại $\boldsymbol{z}_0$

5.1 Ước lượng hàm hồi quy tại z_0

Cho Y_0 là giá trị phản hồi của mô hình hồi quy $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$ khi các biến dự đoán có giá trị $\boldsymbol{z}_0 = [1\ z_{01}\ z_{02}\ \dots\ z_{0r}]'$. Khi đó ta có:

$$E(Y_0 \mid \boldsymbol{z}_0) = \beta_0 + \beta_1 z_{01} + \cdots + \beta_r z_{0r} = \boldsymbol{z}_0'\boldsymbol{\beta}$$

Ước lượng bình phương cực tiểu của $\boldsymbol{z}_0'\boldsymbol{\beta}$ là $\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}}$.

Định lý 5.1

Mô hình hồi quy tuyến tính $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\varepsilon}$ với $\boldsymbol{Z}$ đầy hạng có $\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}}$ là ước lượng không chệch của $E(Y_0 \mid \boldsymbol{z}_0)$ với phương sai nhỏ nhất, là $\mathrm{Var}(\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}}) = \boldsymbol{z}_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_0\sigma^2$. Nếu $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$ thì khoảng tin cậy mức $100(1-\alpha)\%$ của $E(Y_0 \mid \boldsymbol{z}_0) = \boldsymbol{z}_0'\boldsymbol{\beta}$ là

$$\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{(\boldsymbol{z}_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_0)s^2}$$

ở đây $t_{n-r-1}(\alpha/2)$ là phân vị trên mức $100(\alpha/2)\%$ của phân phối Student với $n-r-1$ bậc tự do.

Chứng minh:

Cố định z_0, lúc này $z_0'\boldsymbol{\beta}$ là tổ hợp tuyến tính của các hệ số hồi quy β_i. Ta đã chỉ ra rằng:

$$\mathrm{Var}(\mathbf{z}_0'\boldsymbol{\beta}) = \mathbf{z}_0'\mathrm{Cov}(\boldsymbol{\beta})\mathbf{z}_0 = \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\sigma^2$$

Mặt khác, theo định lý 4.1 ta có $\widehat{\boldsymbol{\beta}} \sim \mathcal{N}_{r+1}(\boldsymbol{\beta}, \sigma^2(\mathbf{Z}'\mathbf{Z})^{-1})$ và $\widehat{\boldsymbol{\beta}}$ độc lập với $s^2/\sigma^2 \sim \chi^2_{n-r-1}/(n-r-1)$. Do đó, tổ hợp tuyến tính $\mathbf{z}_0'\boldsymbol{\beta}$ có phân phối chuẩn $\mathcal{N}(\mathbf{z}_0'\boldsymbol{\beta}, \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\sigma^2)$.

Từ đây suy ra:

$$T := \frac{(\mathbf{z}_0'\widehat{\boldsymbol{\beta}} - \mathbf{z}_0'\boldsymbol{\beta})/\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\sigma^2}}{\sqrt{s^2/\sigma^2}} = \frac{\mathbf{z}_0'\widehat{\boldsymbol{\beta}} - \mathbf{z}_0'\boldsymbol{\beta}}{\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 s^2}} \sim t_{n-r-1}$$

Hay ta có

$$P\left(\mid T \mid = \frac{\mid \mathbf{z}_0'\widehat{\boldsymbol{\beta}} - \mathbf{z}_0'\boldsymbol{\beta} \mid}{\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 s^2}} < t_{n-r-1}\left(\frac{\alpha}{2}\right)\right) = 1-\alpha$$

Và khoảng tin cậy mức $100(1-\alpha)\%$ của $\mathbf{z}_0'\boldsymbol{\beta}$ có thể được xác định bởi:

$$\mathbf{z}_0'\widehat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{\mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0 s^2}$$

□

5.2 Dự đoán quan sát mới tại z_0

Dự đoán quan sát mới Y_0 tại $\boldsymbol{z}_0' = [1, z_{01}, \ldots, z_{0r}]$ theo mô hình hồi quy được:

$$Y_0 = \boldsymbol{z}_0'\boldsymbol{\beta} + \varepsilon_0$$

trong đó $\varepsilon_0 \sim \mathcal{N}(0, \sigma^2)$ và độc lập với các $\boldsymbol{\varepsilon}_i$, do đó cũng độc lập với $\widehat{\boldsymbol{\beta}}$ và s^2. Các sai số ε_i sẽ bị ảnh hưởng từ các ước lượng $\widehat{\boldsymbol{\beta}}$ và s^2 thông qua biến phản hồi Y, nhưng ε_0 thì không. Ta có định lý sau

Định lý 5.2

Xét mô hình hồi quy tuyến tính $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ với $\boldsymbol{Z}$ có hạng đầy đủ và $\boldsymbol{\varepsilon} \sim \mathcal{N}_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Một quan sát mới Y_0 có *dự đoán không chệch*

$$\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}} = \widehat{\beta}_0 + \widehat{\beta}_1 z_{01} + \cdots + \widehat{\beta}_r z_{0r}$$

và phương sai của *sai số dự báo* là

$$\mathrm{Var}(Y_0 - \boldsymbol{z}_0'\widehat{\boldsymbol{\beta}}) = \sigma^2(1 + \boldsymbol{z}_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_0)$$

Chứng minh:
Theo định lý 5.1, kỳ vọng và phương sai của $\mathbf{z}_0'\widehat{\boldsymbol{\beta}}$ là:

$$E(\mathbf{z}_0'\widehat{\boldsymbol{\beta}}) = \mathbf{z}_0'\boldsymbol{\beta} \quad \text{và} \quad \mathrm{Var}(\mathbf{z}_0'\widehat{\boldsymbol{\beta}}) = \mathbf{z}_0'(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{z}_0\sigma^2$$

Mặt khác,

$$\widehat{\varepsilon_0} = Y_0 - z_0'\widehat{\boldsymbol{\beta}} = z_0'\boldsymbol{\beta} + \boldsymbol{\varepsilon_0} - z_0'\widehat{\boldsymbol{\beta}} = \boldsymbol{\varepsilon_0} + z_0'(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})$$

Ta lại có $\widehat{\boldsymbol{\beta}}$ là ước lượng không chệch của $\boldsymbol{\beta}$ nên $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

$$E(\widehat{\varepsilon_0}) = E(\boldsymbol{\varepsilon_0}) + E(z_0'(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})) = 0$$

Do đó ta có $z_0'\boldsymbol{\beta}$ là ước lượng không chệch của Y_0. Hơn nữa, để ý rằng $\boldsymbol{\varepsilon_0}$ và $\widehat{\boldsymbol{\beta}}$ là hai biến độc lập và $\widehat{\boldsymbol{\beta}}$ có phương sai là $\sigma^2(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ nên

$$\mathrm{Var}(\widehat{\varepsilon_0}) = \mathrm{Var}(Y_0) + \mathrm{Var}(z_0'\widehat{\boldsymbol{\beta}}) = \mathrm{Var}(\boldsymbol{\varepsilon_0}) + \mathrm{Var}(z_0'\widehat{\boldsymbol{\beta}}) = \sigma^2(1 + z_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}z_0)$$

□

Hệ quả: Khoảng dự đoán mức $100(1-\alpha)\%$ của Y_0 được cho bởi

$$\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(1 + \boldsymbol{z}_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_0)}$$

ở đây $t_{n-r-1}(\alpha/2)$ là phân vị trên mức $100(\alpha/2)\%$ của phân phối Student với $n-r-1$ bậc tự do.

Giải thích: Ta có hệ quả này là bởi vì $\boldsymbol{\varepsilon}$ có phân phối chuẩn theo giả thiết và $\widehat{\boldsymbol{\beta}}$ cũng có phân phối chuẩn nên:

$$T := \frac{(Y_0 - z_0'\widehat{\boldsymbol{\beta}})/\sqrt{\sigma^2(1 + z_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}z_0)}}{\sqrt{s^2/\sigma^2}} = \frac{Y_0 - z_0'\widehat{\boldsymbol{\beta}}}{\sqrt{s^2(1 + z_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}z_0)}} \sim t_{n-r-1}$$

Do đó,

$$P\left(\mid T \mid = \frac{\mid Y_0 - z_0'\widehat{\boldsymbol{\beta}} \mid}{\sqrt{s^2(1 + z_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}z_0)}} < t_{n-r-1}(\frac{\alpha}{2})\right) = 1 - \alpha$$

.

Sau đây ta xét một ví dụ áp dụng hai kết quả trên.

Ví dụ 5.1

Một nhà nghiên cứu làm việc cho một công ty sản xuất thiết bị máy tính đã thu thập thông tin về 7 loại CPU dựa vào số đơn đặt hàng và số đơn vị vào ra để biết được nhu cầu khách hàng nhằm tạo ra sản phẩm CPU mới phù hợp. Dữ liệu thu thập được cho bởi bảng 3.

Tuổi thọ CPU Y (Giờ)	Số đơn đặt z_1 (Nghìn)	Đơn vị vào-ra z_2 (Nghìn)
141.5	123.5	2.108
168.9	146.1	9.213
154.8	133.9	1.905
146.5	128.5	0.815
172.8	151.5	1.061
160.1	136.2	8.603
108.5	92.0	1.125

Bảng 3: Tuổi thọ của 7 loại CPU

Ta thu được ước lượng hàm hồi quy

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

và cũng tính được giá trị $s = 1.204$. Bây giờ với quan sát mới $\boldsymbol{z}_0 = [1\ 130\ 7.5]'$ ta tiến hành ước lượng hàm hồi quy và dự đoán quan sát mới tại z_0 bằng định lý 5.1 và hệ quả định lý 5.2.

Khoảng tin cậy mức 95% của hàm hồi quy tuyến tính tại $\boldsymbol{z}_0$ là:

$$\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}} \pm t_4(0.025)s\sqrt{\boldsymbol{z}_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_0} = 151.97 \pm 2.776 \times (0.71) = 151.97 \pm 1.97$$

hay là khoảng $[150; 153.94]$. Khoảng dự đoán mức 95% của quan sát mới Y_0 tại $\boldsymbol{z}_0$ là:

$$\boldsymbol{z}_0'\widehat{\boldsymbol{\beta}} \pm t_4(0.025)s\sqrt{1 + \boldsymbol{z}_0'(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{z}_0} = 151.97 \pm 2.776 \times (1.40) = 151.97 \pm 3.89$$

hay là khoảng $[148.08; 155.86]$.

6 Kiểm định mô hình và các yếu tố khác

6.1 Chuẩn hóa dữ liệu

Chuẩn hóa dữ liệu là một phương pháp tiền xử lý phổ biến thường xuyên được sử dụng trong các bài toán phân tích dữ liệu, đặc biệt là khi dữ liệu có yếu tố đơn vị (mét, kilogram, VNĐ,...). Yếu tố này khiến cho dữ liệu trong các khoảng giá trị khác hẳn nhau, chẳng hạn z_1 là biến dự đoán chỉ chiều cao của một người, vậy thì $z_1 \in [0, 3]$ đơn vị mét. Mặt khác, z_2 là biến dự đoán chỉ cân nặng của một người trưởng thành, tính theo đơn vị kilogram, khoảng giá trị cần quan tâm là $[30, 100]$. Sự khác biệt lớn giữa khoảng giá trị của các biến độc lập có thể ảnh hưởng đến việc xây dựng mô hình. Trong trường hợp hai biến "chiều cao" và "cân nặng" như trên, "chiều cao" sẽ có tác động lớn hơn đối với đầu ra (biến phản hồi), nhưng như vậy không khẳng định được rằng biến đó có ý nghĩa quan trọng hơn về mặt thống kê.

Để chuẩn hóa dữ liệu, ta đặt

$$x = \frac{z - \mu}{\sigma},$$

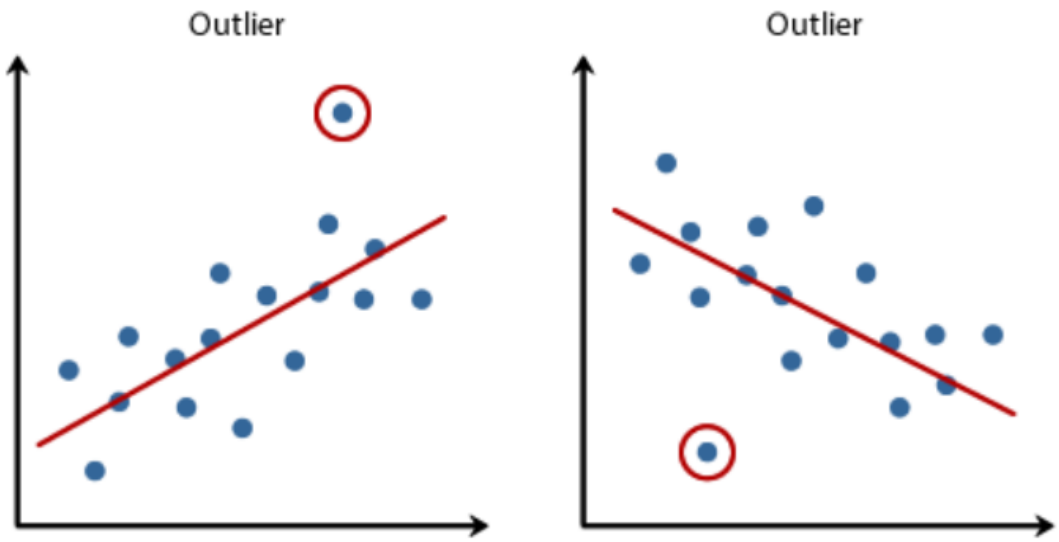
trong đó z là giá trị quan sát, μ là giá trị trung bình mẫu và σ là độ lệch chuẩn mẫu.

Việc chuẩn hóa dữ liệu mang lại cho ta những lợi ích sau

- Dữ liệu sau chuẩn hóa không có thứ nguyên, từ đó giải quyết được vấn đề về việc các khoảng giá trị ứng với từng dữ liệu ở các khoảng cách xa nhau.
- Đo lường xem từng giá trị z có cách xa giá trị trung bình μ của nó hay không. Hơn nữa, việc chuẩn hóa giúp ta thu được một bộ dữ liệu mới có trung bình là 0 và độ lệch chuẩn là 1.

6.2 Tìm hiểu về điểm Outlier

Outlier là điểm dữ liệu bất thường, nằm cách xa so với phần dữ liệu còn lại. Điểm Outlier có vai trò quan trọng trong mô hình hồi quy bởi có ảnh hưởng đến việc đánh giá bình phương tối thiểu và độ dốc của mô hình hồi quy.



Hình 3: Minh họa trực quan về điểm Outlier

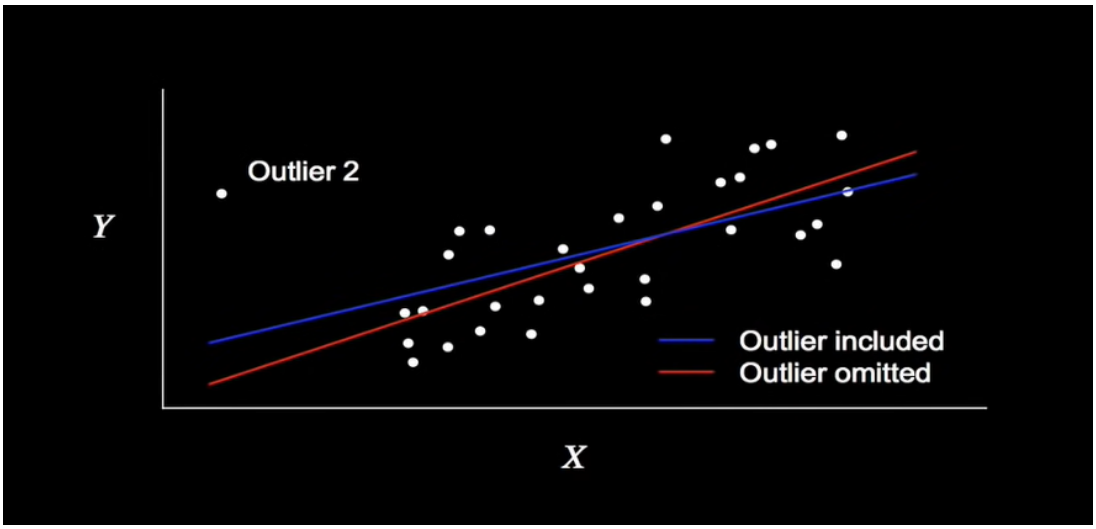
Hình 3 gồm hai bộ dữ liệu khác nhau, tương ứng với đường thẳng hồi quy và các điểm được khoanh tròn là minh họa trực quan về điểm Outlier (ứng với từng bộ dữ liệu).

Ta quan tâm đến ảnh hưởng của điểm Outlier đến độ dốc của mô hình hồi quy và đánh giá bình phương tối thiểu. Ở hình 4, ta thấy rằng có sự khác biệt lớn về độ dốc mô hình hồi quy khi ta xét việc nên giữ điểm Outlier hay không, do vậy việc kiểm soát các điểm Outlier trong bộ dữ liệu là vô cùng quan trọng.

Trong trường hợp có hữu hạn điểm dữ liệu, ta có thể nhận diện và cân nhắc việc giữ/bỏ điểm Outlier, tuy nhiên khi có hàng triệu điểm dữ liệu, để phân loại và lọc điểm Outlier dường như trở nên khó khăn hơn, đòi hỏi các kỹ thuật để xử lý việc loại bỏ các Outlier nhằm đảm bảo về tính chính xác của ước lượng hồi quy. Trong bài báo cáo này, nhóm tác giả sẽ tìm hiểu về hai phương pháp nhằm hỗ trợ giải quyết vấn đề này, bao gồm sử dụng *độ đo Leverage* và *quy tắc 1.5IQR*.

6.3 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 của điểm dữ liệu



Hình 4: Ảnh hưởng của điểm Outlier lên mô hình 2

thứ j trong mô hình hồi quy đơn biến được xác định bởi công thức sau

$$h_{jj} = \frac{1}{n} + \frac{(z_j - \bar{z})^2}{\sum_{j=1}^{n} (z_j - \bar{z})^2}. \tag{6.0}$$

Một cách tiếp cận khác đối với độ đo leverage đó là ta xét ma trận $\boldsymbol{H} := \boldsymbol{Z}\left(\boldsymbol{Z'Z}\right)^{-1}\boldsymbol{Z'}$. Khi đó leverage tại điểm dữ liệu thứ j chính là phần tử h_{jj} trên đường chéo chính của ma trận $\boldsymbol{H}$.

Leverage trung bình cho các quan sát bằng $\dfrac{p+1}{n}$, trong đó p là số biến dự đoán. Từ công thức 6.0, trong trường hợp mô hình gồm p biến dự đoán, ta thấy rằng h_{jj} luôn nằm giữa $\dfrac{1}{n}$ và 1.

Chứng minh: Với p là số biến dự đoán, độ đo leverage trung bình được tính như sau

$$\begin{aligned}
\overline{h} &= \frac{1}{n}\sum_{j=1}^{n} h_{jj} \\
&= \frac{1}{n}\mathrm{tr}(\boldsymbol{H}) \\
&= \frac{1}{n}\mathrm{tr}(\boldsymbol{Z(Z'Z)^{-1}Z'}) \\
&= \frac{1}{n}\mathrm{tr}((\boldsymbol{Z'Z})(\boldsymbol{Z'Z})^{-1}) \\
&= \frac{1}{n}\mathrm{tr}(\boldsymbol{I_{p+1}}) \\
&= \frac{p+1}{n}.
\end{aligned}$$

Tiếp theo, ta xét phương trình

$$\widehat{y} = \boldsymbol{Z}\widehat{\beta} = \boldsymbol{Z(Z'Z)^{-1}Z'}y = \boldsymbol{H}y,$$

ở đó hàng thứ j của $\widehat{\boldsymbol{y}}$ được biểu diễn như sau

$$\widehat{y}_j = h_{jj}y_j + \sum_{k\neq j} h_{jk}y_k.$$

Giả thiết rằng các giá trị y_k cố định, $k \neq j$, ta được

$$\Delta\widehat{y}_j = h_{jj}\Delta y_j.$$

Nếu độ đo Leverage tại điểm dữ liệu thứ j càng lớn thì độ thay đổi của y_j càng có tác động lớn đến độ thay đổi của $\widehat{y}_j$, hay ảnh hưởng đến độ dốc của mô hình hồi quy. Những điểm như vậy được gọi là high leverage và đây là điểm Outlier.

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 lớn đến mô hình, điểm nào có giá trị của biến dự đoán càng xa với bộ dữ liệu thì tác động càng lớn.
- Trong thực tế, ngưỡng chọn Outlier thường là $\dfrac{2p}{n}$ hoặc $\dfrac{3p}{n}$, ở đó p là số biến dự đoán.

6.4 Quy tắc 1.5IQR

Khi làm việc với một bộ dữ liệu, ta cần quan tâm đến một vài đặc trưng nhất định để kiểm soát và có được những thông tin cần thiết về bộ dữ liệu nhằm định vị Outlier, bao gồm những đặc trưng sau:

- Giá trị nhỏ nhất hay thấp nhất của bộ dữ liệu.
- Tứ phân vị thứ nhất $Q1$: phân vị mức 25%.
- Tứ phân vị thứ nhất $Q2$: phân vị mức 50% hay chính là trung vị.
- Tứ phân vị thứ nhất $Q3$: phân vị mức 75%.
- Giá trị lớn nhất hay cao nhất của bộ dữ liệu.

Tiếp đến, ta định nghĩa khoảng cách tứ phân vị (Interquartile range) là giá trị

$$\text{IQR} = Q3 - Q1.$$

Khi đó, theo quy tắc 1.5 IQR, nếu $x \notin [Q1 - 1.5\text{IQR}, Q3 + 1.5\text{IQR}]$ thì x được coi là một điểm Outlier.

Nhận xét

Quy tắc 1.5 IQR để loại điểm Outlier xuất phát từ quy tắc 3σ, được phát biểu rằng với độ tin cậy 0.9973, biến ngẫu nhiên X có phân phối chuẩn $\mathcal{N}(\mu, \sigma^2)$ sẽ lấy giá trị trong khoảng lân cận 3σ của giá trị trung bình μ. Tuy nhiên, theo quy tắc trên chưa lấy được hết lân cận 3σ, do vậy, nếu thực sự cần thiết, ta có thể hiệu chỉnh lên thành quy tắc 1.7 IQR, tại đó nếu $x \notin [Q_1 - 1.7\text{IQR}, Q_3 + 1.7\text{IQR}]$ thì x được gọi là 1 điểm Outlier.

6.5 Kiểm tra tính phụ thuộc vào biến của mô hình

Sau khi đưa ra được một mô hình hoàn chỉnh, điều ta cần quan tâm là kiểm định sự quan trọng của các biến. Điều này bắt nguồn từ việc không phải biến nào cũng có đóng góp lớn đến mô hình, quan hệ chặt chẽ với biến phản hồi. Việc loại bỏ bớt biến đi vừa giúp ta đơn giản hóa mô hình, giúp cho việc tính toán được thực thi nhanh hơn, đồng thời ta có thể hiểu thêm về các nhân tố quyết định đầu ra mà ta mong muốn.

Ta sử dụng tiêu chuẩn F để kiểm tra tính phụ thuộc vào biến $z_i, i = \overline{1,k}$ của mô hình, hay là mô hình chỉ phụ thuộc vào mỗi giá trị tự do của hằng số β_0, trong đó biểu thức F được xác định như sau

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

trong đó

$$R^2 := \frac{\widehat{y}'\widehat{y} - n\overline{y}^2}{y'y - n\overline{y}^2} = 1 - \frac{\sum_{k=1}^{n} \widehat{\varepsilon}_k^2}{\sum_{k=1}^{n}(y_k - \overline{y})^2}.$$

Với ý tưởng tiếp cận như trên, ta đề cập đến các bước nhằm kiểm tra sự phụ thuộc vào biến của mô hình.

Ta xét giả thuyết

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

với k là số biến dự đoán, và đối thuyết

$$H_1 : \exists \beta_j \neq 0, j = \overline{1,k}.$$

Ta thực hiện kiểm định bằng các bước sau:

Các bước thực hiện

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

Ví dụ 6.1

Với mong muốn nghiên cứu sự phụ thuộc giữa doanh thu và chi phí sản xuất và chi phí tiếp thị, người ta điều tra ngẫu nhiên doanh thu của 12 công ty ứng với 12 thời kỳ, ta được bảng 4 sau

Y Doanh thu (Dollar)	z_1 Chi phí sản xuất (Dollar)	z_2 Chi phí tiếp thị (Dollar)
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

Bảng 4: Bảng tương quan giữa doanh thu và chi phí

Dựa theo các bước được đề cập ở trên, trước hết, ta xét cặp giả thuyết đối thuyết

$$H_0 : \beta_1 = \beta_2 = 0 \text{ và } H_1 : \exists \beta_i \neq 0, i = \overline{1,2}.$$

Áp dụng công thức

$$R^2 := \frac{\widehat{y}'\widehat{y} - n\overline{y}^2}{y'y - n\overline{y}^2} = 1 - \frac{\sum_{k=1}^{n} \widehat{\varepsilon}_k^2}{\sum_{k=1}^{n}(y_k - \overline{y})^2},$$

ta được $R^2 = 0.97565653$, từ đó ta được

$$F = \frac{(12-2-1)(0.97565653)}{2(1-0.97565653)} \approx 180.35451558$$

Tra bảng phân phối Fisher, với mức ý nghĩa 5% ta thấy $F_{2,9}(0,05) = 4.26$. Khi đó, ta bác bỏ giả thuyết H_0, tức là có sự ảnh hưởng của các biến dự đoán đối với mô hình.

6.6 Giá trị t-statistic và p-value

Với chọn lọc các biến quan trọng trong mô hình, ta cần tính toán $p-$value cho từng biến, thật vậy, xét bài toán gồm r biến dự đoán, ứng với mỗi $i = \overline{1,r}$, ta quan tâm đến cặp giả thuyết đối thuyết sau

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

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

$$t - \text{statistic} = \frac{\widehat{\beta}_i}{\sqrt{\widehat{D}}},$$

trong đó $\widehat{\beta}_i$ là ước lượng của hệ số hồi quy thứ i và

$$\widehat{D} = s^2 = \frac{1}{n-r-1}\sum_{j=1}^{n}\widehat{\varepsilon}_j^2.$$

Trong trường hợp đối thuyết hai phía, giá trị $p-$value là 2 lần xác suất quan sát được một số có giá trị bằng hoặc lớn hơn giá trị tuyệt đối $t-$statistic, với công thức như sau

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

trong đó CDF là hàm phân phối tích lũy của phân phối Student.

Nhận xét

Giả thuyết H_0 thường bị bác bỏ ở mức ý nghĩa 5% hoặc 1% ứng với khi $p-$value $< 5\%$ hoặc $p-$value $< 1\%$. Với số lượng quan sát đủ lớn, khi $n \geq 30$, các điều kiện bác bỏ sẽ lần lượt tương ứng với $t-$statistic ≥ 2 (mức ý nghĩa 5%) hoặc $t-$statistic ≥ 2.75 (mức ý nghĩa 1%).

6.7 Khảo sát phần dư

Mô hình hồi quy tuyến tính sẽ phù hợp với dãy số liệu đang quan sát nếu các sai số ε_j là các biến ngẫu nhiên độc lập và có cùng phân phối. Mô hình ta đang xét có phân phối chuẩn, do vậy ta cần kiểm tra xem sai số có phân phối chuẩn $\mathcal{N}(0, \sigma^2 \boldsymbol{I}_n)$ hay không?

Nhằm trả lời câu hỏi về giả thuyết nhiễu ngẫu nhiên có tuân theo phân phối chuẩn không, ta mô hình hóa thành bài toán kiểm định.

Ta đưa ra một tiêu chuẩn để chấp nhận hay bác bỏ giả thuyết

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

hay viết theo cách khác, ta có

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

Ta đề cập đến định lý sau

Định lý 6.1

Xét mô hình hồi quy tuyến tính $\boldsymbol{Y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ với ma trận $\boldsymbol{Z}$ đầy hạng. Gọi $v_1, v_2, \ldots, v_n$ là dãy các vector riêng trực chuẩn của ma trận $\boldsymbol{I} - \boldsymbol{H} := \boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$. Khi đó giả thuyết $H_0 : \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2\boldsymbol{I})$ bị bác bỏ với mức ý nghĩa $100\alpha\%$ nếu

$$\left(\sum_{i=1}^{n} v_i\right)' \widehat{\varepsilon} > \frac{t_{n-r-2}(\alpha/2)}{\sqrt{n-r-2}} \sqrt{\left| \left(\left(\sum_{i=1}^{n} v_i \right)' \widehat{\varepsilon} \right)^2 - (n-r-1)\widehat{\varepsilon}'\widehat{\varepsilon} \right|}.$$

Chứng minh. Ta đã chứng minh được ma trận $\boldsymbol{I} - \boldsymbol{H}$ là ma trận đối xứng, lũy đẳng, đồng thời chỉ có các trị riêng là 0 hoặc 1. Thực hiện chéo hóa ma trận, ta được

$$\boldsymbol{I} - \boldsymbol{H} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}',$$

trong đó $\boldsymbol{P} := \left[\begin{array}{c|c|c|c} \boldsymbol{v}_1 & \boldsymbol{v}_2 & \cdots & \boldsymbol{v}_n \end{array} \right]$ là ma trận trực giao gồm các vector riêng trực chuẩn và

$$\boldsymbol{\Lambda} := \operatorname{diag}\{\underbrace{1, \ldots, 1}_{n-r-1}, \underbrace{0, \ldots, 0}_{r+1}\}.$$

Tiếp theo, ta đặt

$$\boldsymbol{e} := \begin{bmatrix} e_1 & e_2 & \ldots & e_n \end{bmatrix}' = \boldsymbol{P}'\hat{\boldsymbol{\varepsilon}}.$$

Nếu giả thuyết H_0 đúng thì

$$E(\boldsymbol{e}) = \boldsymbol{P}'E(\widehat{\boldsymbol{\varepsilon}}) = \boldsymbol{0};$$

$$\operatorname{Cov}(\boldsymbol{e}) = \boldsymbol{P}' \operatorname{Cov}(\widehat{\boldsymbol{\varepsilon}})\, \boldsymbol{P} = \sigma^2 \boldsymbol{P}'\boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'\boldsymbol{P} = \sigma^2\boldsymbol{\Lambda}.$$

Không những vậy, $e_1, e_2, \ldots, e_{n-r-1}$ sẽ là các biến ngẫu nhiên có phân phối chuẩn và $e_{n-r} = e_{n-r+1} = \ldots = e_n = 0$. Khi đó

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n-r-1} e_i \text{ và } \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n-r-1} e_i^2. \tag{6.1}$$

Xét vector ngẫu nhiên

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

Nếu giả thuyết H_0 đúng thì ta sẽ có

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

Hơn nữa, ta còn có biến đổi

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

Xét thống kê

$$T := n\bar{e}\sqrt{\frac{n-r-2}{\left|n^2\bar{e}^2 - (n-r-1)\sum_{i=1}^{n} e_i^2\right|}}. \tag{6.4}$$

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

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

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

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

hay viết lại thành

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

Để ý rằng

$$n\bar{e}=\sum_{i=1}^{n}e_i=\mathbf{1}'\boldsymbol{e}=\mathbf{1}'\boldsymbol{P}'\widehat{\boldsymbol{\varepsilon}}=(\boldsymbol{P}\mathbf{1})'\widehat{\boldsymbol{\varepsilon}}=\left(\sum_{i=1}^{n}\boldsymbol{v}_i\right)'\widehat{\boldsymbol{\varepsilon}};$$

$$\sum_{i=1}^{n}e_i^2=\boldsymbol{e}'\boldsymbol{e}=\widehat{\boldsymbol{\varepsilon}}'\boldsymbol{P}\boldsymbol{P}'\widehat{\boldsymbol{\varepsilon}}=\widehat{\boldsymbol{\varepsilon}}\widehat{\boldsymbol{\varepsilon}}.$$

Thay vào (6.5), ta được

$$\left(\sum_{i=1}^{n}v_i\right)'\widehat{\varepsilon}>\frac{t_{n-r-2}\left(\alpha/2\right)}{\sqrt{n-r-2}}\sqrt{\left|\left(\left(\sum_{i=1}^{n}v_i\right)'\widehat{\varepsilon}\right)^2-(n-r-1)\widehat{\varepsilon}'\widehat{\varepsilon}\right|},$$

hay ta có điều phải chứng minh. □

Ta có một số lưu ý ở quy tắc kiểm định đề ra ở định lý trên

Lưu ý

- Quy tắc kiểm định này không phụ thuộc vào cách chọn hệ vector riêng trực chuẩn của ma trận $\boldsymbol{I}-\boldsymbol{H}$.
- Quy tắc kiểm định này chỉ xác định khi $n > r+2$. Nhưng trong thực tế, thường giá trị n (là số quan sát) sẽ lớn hơn rất nhiều so với r (là số biến dự đoán) nên ràng buộc này hầu như luôn thỏa mãn.

Nhận xét

Khi tiêu chuẩn Student dẫn đến bác bỏ giả thuyết ε không tuân theo phân phối chuẩn $\mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_n)$, đồ thị phần dư thể hiện được những lỗi sau của mô hình.

- Phần dư $\widehat{\varepsilon_j}$ phụ thuộc vào $\widehat{y}_j$.
- Phương sai của các sai số $\widehat{\varepsilon_j}$ không phải là hằng số.
- Mô hình dự đoán bỏ sót biến dự đoán z_j.
- Phần dư $\widehat{\varepsilon_j}$ không có phân phối chuẩn.

6.8 Kiểm tra tính không tương quan của các phần dư theo thời gian

Giả sử Y_j được theo dõi theo thời gian $j = 1, 2, \ldots$ Khi đó, thường xảy các trường hợp ε_j có tương quan với nhau.

Khi đó, ta có thể sử dụng tiêu chuẩn Durbin-Watson để kiểm tra tính tương quan này. Cụ thể, đại lượng

$$DW := \frac{\sum_{j=2}^{n} (\widehat{\varepsilon_j} - \widehat{\varepsilon}_{j-1})^2}{\sum_{j=1}^{n} \widehat{\varepsilon_j}^2},$$

sẽ tuân theo phân phối Durbin-Watson.

Tra bảng Durbin-Watson với mức ý nghĩa α, ta tìm được hai hệ số $d_1(n, p, \alpha) < d_2(n, p, \alpha)$ với n là số điểm của bộ dữ liệu, p là số biến, α là mức ý nghĩa. So sánh với DW, ta có kết luận sau:

- Nếu $0 \leq DW < d_1$ thì các $\widehat{\varepsilon_j}$ 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 $\widehat{\varepsilon_j}$ 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 $\widehat{\varepsilon_j}$ có tự tương quan âm.

Từ tiêu chuẩn trên, ta nhận thấy rằng phương pháp này giúp ta xác nhận các tự tương quan bậc 1, tức là 2 biến liên tiếp. Ta quay lại ví dụ của bảng 4 với

mục tiêu kiểm tra tính không tương quan theo thời gian thông qua tiêu chuẩn Durbin-Watson.

Ví dụ 6.2

Với mong muốn nghiên cứu sự phụ thuộc của doanh thu dựa vào chi phí sản xuất và chi phí tiếp thị, người ta đã tổ chức một cuộc điều tra ngẫu nhiên doanh thu của 12 công ty ứng với 12 thời kỳ, kết quả điều tra thu được trong bảng 5.

Y Doanh thu (Dollar)	z_1 Chi phí sản xuất (Dollar)	z_2 Chi phí tiếp thị (Dollar)
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

Bảng 5: Bảng tương quan giữa doanh thu và chi phí

Từ công thức ở phần 6.8, ta sử dụng tiêu chuẩn Durbin-Watson để kiểm tra tính tương quan. Tính toán đại lượng DW thu được

$$DW = 2.52723823$$

Với các giá trị $\alpha = 0.05, n = 12, k = 2$, ta có được bảng Durbin-Watson như hình 5.

Khi đó, tra bảng Durbin-Watson ta được

$$d_1 = 0.812, d_2 = 1.579,$$

tương đương với $4 - d_2 = 2.421, 4 - d_1 = 3.188$, ta nhận thấy rằng

$$4 - d_2 < DW < 4 - d_1,$$

do vậy ta không kết luận được tính tương quan của các sai số ε_j, $j = \overline{1, n}$.

n\k	1		2		3		4		5		6		7		8		9		10	
6	0.610	1.400																		
7	0.700	1.356	0.467	1.896																
8	0.763	1.332	0.559	1.777	0.367	2.287														
9	0.824	1.320	0.629	1.699	0.455	2.128	0.296	2.588												
10	0.879	1.320	0.697	1.641	0.525	2.016	0.376	2.414	0.243	2.822										
11	0.927	1.324	0.758	1.604	0.595	1.928	0.444	2.283	0.315	2.645	0.203	3.004								
12	0.971	1.331	0.812	1.579	0.658	1.864	0.512	2.177	0.380	2.506	0.268	2.832	0.171	3.149						
13	1.010	1.340	0.861	1.562	0.715	1.816	0.574	2.094	0.444	2.390	0.328	2.692	0.230	2.985	0.147	3.266				
14	1.045	1.350	0.905	1.551	0.767	1.779	0.632	2.030	0.505	2.296	0.389	2.572	0.286	2.848	0.200	3.111	0.127	3.360		
15	1.077	1.361	0.946	1.543	0.814	1.750	0.685	1.977	0.562	2.220	0.447	2.471	0.343	2.727	0.251	2.979	0.175	3.216	0.111	3.438
16	1.106	1.371	0.982	1.539	0.857	1.728	0.734	1.935	0.615	2.157	0.502	2.388	0.398	2.624	0.304	2.860	0.222	3.090	0.155	3.304
17	1.133	1.381	1.015	1.536	0.897	1.710	0.779	1.900	0.664	2.104	0.554	2.318	0.451	2.537	0.356	2.757	0.272	2.975	0.198	3.184
18	1.158	1.391	1.046	1.535	0.933	1.696	0.820	1.872	0.710	2.060	0.603	2.258	0.502	2.461	0.407	2.668	0.321	2.873	0.244	3.073
19	1.180	1.401	1.074	1.536	0.967	1.685	0.859	1.848	0.752	2.023	0.649	2.206	0.549	2.396	0.456	2.589	0.369	2.783	0.290	2.974
20	1.201	1.411	1.100	1.537	0.998	1.676	0.894	1.828	0.792	1.991	0.691	2.162	0.595	2.339	0.502	2.521	0.416	2.704	0.336	2.885
21	1.221	1.420	1.125	1.538	1.026	1.669	0.927	1.812	0.829	1.964	0.731	2.124	0.637	2.290	0.546	2.461	0.461	2.633	0.380	2.806
22	1.239	1.429	1.147	1.541	1.053	1.664	0.958	1.797	0.863	1.940	0.769	2.090	0.677	2.246	0.588	2.407	0.504	2.571	0.424	2.735
23	1.257	1.437	1.168	1.543	1.078	1.660	0.986	1.785	0.895	1.920	0.804	2.061	0.715	2.208	0.628	2.360	0.545	2.514	0.465	2.670
24	1.273	1.446	1.188	1.546	1.101	1.656	1.013	1.775	0.925	1.902	0.837	2.035	0.750	2.174	0.666	2.318	0.584	2.464	0.506	2.613
25	1.288	1.454	1.206	1.550	1.123	1.654	1.038	1.767	0.953	1.886	0.868	2.013	0.784	2.144	0.702	2.280	0.621	2.419	0.544	2.560
26	1.302	1.461	1.224	1.553	1.143	1.652	1.062	1.759	0.979	1.873	0.897	1.992	0.816	2.117	0.735	2.246	0.657	2.379	0.581	2.513
27	1.316	1.469	1.240	1.556	1.162	1.651	1.084	1.753	1.004	1.861	0.925	1.974	0.845	2.093	0.767	2.216	0.691	2.342	0.616	2.470
28	1.328	1.476	1.255	1.560	1.181	1.650	1.104	1.747	1.028	1.850	0.951	1.959	0.874	2.071	0.798	2.188	0.723	2.309	0.649	2.431
29	1.341	1.483	1.270	1.563	1.198	1.650	1.124	1.743	1.050	1.841	0.975	1.944	0.900	2.052	0.826	2.164	0.753	2.278	0.681	2.396
30	1.352	1.489	1.284	1.567	1.214	1.650	1.143	1.739	1.071	1.833	0.998	1.931	0.926	2.034	0.854	2.141	0.782	2.251	0.712	2.363

Hình 5: Bảng Durbin-Watson tại $\alpha = 0.05$

6.9 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.

Định nghĩa

Các biến $\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_k}$ được gọi là đa cộng tuyến nếu tồn tại các hằng số $c_0, c_1, \ldots, c_k$ không đồng thời bằng 0 thỏa mãn $c_0 + \sum_{i=1}^{k} c_i \boldsymbol{Z_i} = 0$. Khi đó, biểu thức tương đương với $c_0 + \sum_{i=1}^{k} c_i \boldsymbol{Z_{ji}} = 0$, $j = \overline{1, n}$.

Nếu ma trận $\boldsymbol{Z}$ không đầy hạng, khi đó giữa các cột của $\boldsymbol{Z}$ có sự đa cộng tuyến, dẫn đến ma trận $\boldsymbol{Z'Z}$ không là ma trận khả nghịch.

Nhận xét

Trên thực tế, khi $|\boldsymbol{Z'Z}| \approx 0$, người ta có thể coi $\boldsymbol{Z_1}, \ldots, \boldsymbol{Z_k}$ có hiện tượng đa cộng tuyến. Khi đó, ước lượng $\widehat{\boldsymbol{\beta}} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'Y}$ thường không ổn định và có phương sai rất lớn, hay các khoảng tin cậy sẽ rất rộng.

Cách nhận biết hiện tượng:

- Một số phần tử trên đường chéo chính của ma trận $(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$ rất lớn.
- Các hệ số tương quan tuyến tính mẫu của các cặp $\boldsymbol{Z_i}$, $\boldsymbol{Z_j}$ là $r_{ij} = s_{ij}/\sqrt{s_{jj}s_{ii}}$ mà $|r_{ij}| > 0.7$ (để hiểu rõ hơn lý do chọn hằng số này, bạn đọc xem thêm tại [1]), trong đó

$$s_{ij} = \frac{1}{n}\sum_{k=1}^{n} z_{ki}z_{kj} - \overline{z}_i\,\overline{z}_j.$$

Để khắc phục, ta thực hiện theo các bước sau.

Các bước thực hiện

- Tính các hệ số tương quan tuyến tính mẫu r_{ij}.
- Đặt r_{0i} là các hệ số tương quan giữa Y và Z_i, với

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

trong đó $s_{00} = s_y^2 = \overline{y^2} - (\overline{y})^2$ và $s_{0j} = \dfrac{1}{n}\sum_{k=1}^{n} y_k z_{kj} - \overline{y}\,\overline{z}_j$.

Nếu $|r_{ij}| > 0.7$ thì $\begin{cases} \text{loại } Z_i \text{ ra khỏi mô hình nếu } |r_{0i}| < |r_{0j}|, \\ \text{loại } Z_j \text{ ra khỏi mô hình nếu } |r_{0i}| > |r_{0j}|. \end{cases}$

- Thực hiện hồi quy tuyến tính sau khi ma trận $\boldsymbol{Z}$ đã loại bỏ biến Z_i hay Z_j.

Cơ sở để thực hiện bước 2 là dựa trên sự tương quan của biến so với y, biến nào thể hiện mức độ tương quan với y lớn hơn, tức có hệ số tương quan gần với 1 hơn, biến đó sẽ được giữ lại.

6.10 Xác định các biến quan trọng

Sau khi đưa ra mô hình hồi quy, ta cần xác định tính quan trọng của các biến, bởi có thể chỉ một tập con số biến dự đoán có quan hệ chặt chẽ đến biến phản hồi. Việc tìm ra một tập con phù hợp và có quan hệ mật thiết với biến phản hồi là nhiệm vụ quan trọng.

Nhóm tác giả đề cập đến một vài thống kê được dùng để đánh giá chất lượng của mô hình, các thông kê phổ biến thường được sử dụng là Mallow's C_p, AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion) và R^2_{adj} hay còn gọi là hệ số xác định hiệu chỉnh. Với p là số biến dự đoán, $\boldsymbol{\beta}_{(2)}$ và $\boldsymbol{Z}_{(1)}$ là vector

và ma trận gồm các giá trị quan sát tương ứng. Ta đề cập đến các công thức ứng với các thông kê được nêu trên như sau:

- (Mallow's C_p) Chọn $\boldsymbol{\beta}_{(2)}$ sao cho $C_p \approx p$

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

- (Akaike information criterion) Chọn $\beta_{(2)}$ để AIC nhỏ nhất

$$AIC = 2p + n\left(1 + \ \ln \frac{2\pi RSS(\boldsymbol{\beta}_{(1)})}{n}\right).$$

- (Bayesian information criterion) Chọn $\beta_{(2)}$ để BIC nhỏ nhất

$$BIC = n\ln \frac{RSS(\boldsymbol{\beta}_{(1)})}{n} + p\ln\, n.$$

Tiếp theo, ta đề xuất ra ba phương pháp nhằm tối ưu hóa số mô hình cần thử, bao gồm *chọn tiến dần*, *chọn lùi dần* và *chọn hỗn hợp* được trình bày dưới đây.

Chọn tiến dần.

1. Gọi M_0 là mô hình *null* (Mô hình không chứa biến dự đoán).
2. Với $k = 0, 1, \ldots, p-1$,
   - Xét tất cả $p - k$ mô hình được tạo bằng cách lấy các biến phản hồi của M_k và thêm vào một biến không có trong M_k.
   - Chọn mô hình **tốt nhất** trong $p - k$ mô hình vừa tạo và gọi mô hình đó là M_{k+1}. **Tốt nhất** ở đây là có giá trị RSS thấp nhất hoặc R^2 cao nhất.
3. Chọn mô hình tốt nhất trong các mô hình $M_0, M_1, \ldots, M_p$ bằng một trong các hệ số Mallow's C_p, AIC, BIC hoặc R^2_{adj}.

Chọn lùi dần.

1. Gọi M_0 là mô hình *full* (Mô hình chứa tất cả các biến dự đoán).
2. Với $k = p, p-1, \ldots, 1$,
   - Xét tất cả k mô hình được tạo bằng cách loại đi một biến trong mô hình M_k.
   - Chọn mô hình **tốt nhất** trong k mô hình vừa tạo và gọi mô hình đó là M_{k-1}. **Tốt nhất** ở đây là có giá trị RSS thấp nhất hoặc R^2 cao nhất.

3. Chọn mô hình tốt nhất trong các mô hình $M_0, M_1, \ldots, M_p$ bằng một trong các hệ số Mallow's C_p, AIC, BIC hoặc R^2_{adj}.

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

1. Gọi M_0 là mô hình *null* (Mô hình không chứa biến dự đoán).
2. Với $k = 0, 1, \ldots, p$,
   - Xét tất cả $p - k$ mô hình được tạo bằng cách lấy các biến phản hồi của M_k và thêm vào một biến không có trong M_k.
   - Chọn mô hình **tốt nhất** trong $p - k$ mô hình vừa tạo và gọi mô hình đó là M_{k+1}. **Tốt nhất** ở đây là có giá trị RSS thấp nhất hoặc R^2 cao nhất.
   - Sau khi chọn được mô hình tốt nhất, dựa vào một trong các hệ số Mallow's C_p, AIC, BIC hoặc R^2_{adj}, 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.

6.11 Tiến hành phân tích hồi quy

6.11.1 Các bước thiết lập mô hình hồi quy tuyến tính

Từ các phương pháp ta đề cập ở trên, để xây dựng một mô hình hồi quy tuyến tính, ta thực hiện tuần tự theo các bước sau

1. Tiền xử lý dữ liệu
   - Chuẩn hóa tập mẫu $x = (z - \overline{\mu})/s$.
   - Khảo sát tính đơn-đa cộng tuyến của các biến dự đoán.
   - Sử dụng quy tắc tứ phân vị để lọc các điểm Outlier.
2. Xác định các biến quan trọng
   - Ước lượng các hệ số hồi quy: $\widehat{\beta} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}$.
   - Sử dụng tiêu chuẩn F, kiểm tra mối liên hệ giữa $\boldsymbol{Z}$ và $\boldsymbol{Y}$.
   - Xác định hệ số $R, t-$statistic, $p-$value của từng biến Z_i.
   - Lựa chọn các biến quan trọng, loại bỏ các biến không cần thiết.
   - Lặp lại quy trình đến khi thỏa mãn yêu cầu đặt ra.
3. Khảo sát phần dư
   - Sử dụng tiêu chuẩn Student, kiểm tra về phân phối của ε.
   - Khảo sát phần dư, xác định Outlier.

- Xác định độ đo Leverage với từng điểm dữ liệu.
- Loại bỏ các Outlier dựa trên độ đo Leverage ứng với từng điểm vừa tìm được.

4. Xây dựng mô hình

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

6.11.2 Thực hiện các bước

Nhóm tác giả sử dụng ngôn ngữ lập trình Python để xây dựng mô hình hồi quy tuyến tính với bộ dữ liệu "Housing.csv" được lấy trên Kaggle.

Trước hết, ta khai báo các thứ viện sử dụng để thiết lập mô hình hồi quy tuyến tính.

```
import matplotlib.pyplot as plt
import statsmodels.api   as sm
import pandas            as pd
import seaborn           as sns
import numpy             as np
import scipy
import math

import warnings
warnings.filterwarnings('ignore')

from google.colab import drive
drive.mount('/content/drive')

np.set_printoptions(suppress=True)
```

Tiếp theo, ta thực hiện đọc dữ liệu từ file data, về tổng quan data gồm biến phản hồi là giá của một ngôi nhà, được đo lường thông qua các biến dự đoán gồm diện tích tổng thể, số lượng phòng tắm, số tầng,... Đường dẫn chứa dữ liệu là https://www.kaggle.com/datasets/ashydv/housing-dataset.

```
path = "/content/drive/MyDrive/Housing.csv"
cols = ['price', 'area', 'bedrooms','bathrooms', 'stories',
        'mainroad', 'guestroom', 'basement', 'hotwaterheating',
        'airconditioning', 'parking', 'prefarea',
        'furnishingstatus']

data = pd.read_csv(path, header = 0, usecols = cols)
data.head()
```

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	furnishingstatus
0	13300000	7420	4	2	3	yes	no	no	no	yes	2	yes	furnished
1	12250000	8960	4	4	4	yes	no	no	no	yes	3	no	furnished
2	12250000	9960	3	2	2	yes	no	yes	no	no	2	yes	semi-furnished
3	12215000	7500	4	2	2	yes	no	yes	no	yes	3	yes	furnished
4	11410000	7420	4	1	2	yes	yes	yes	no	yes	2	no	furnished

Hình 6: 5 hàng đầu tiên của bộ dữ liệu

Tiếp theo, ta ánh xạ dữ liệu, đánh số cho dữ liệu dạng “yes/no” và tạo biến giả (biến dummy - biến độc lập được đưa vào để giải thích các yếu tố định tính).

```
varlist =  ['mainroad', 'guestroom', 'basement', 'hotwaterheating',
            'airconditioning', 'prefarea']

# Applying the mapping function to the housing list
data[varlist] = data[varlist].apply(
                            lambda x: x.map({'yes': 1, 'no': 0}))
```

Tạo biến giả từ các giá trị trong biến gốc là "furnishingstatus".

```
# Get the dummy variables for the feature 'furnishingstatus' and
# store it in a new variable 'status'
status = pd.get_dummies(data['furnishingstatus'],drop_first = True)

# Add the results to the original housing dataframe
data = pd.concat([data, status], axis = 1)

# Drop 'furnishingstatus' as we have created the dummies for it
data.drop(['furnishingstatus'], axis = 1, inplace = True)
```

Bây giờ ta thử in ra một số quan sát đầu tiên để xem thử dữ liệu lúc này như thế nào.

```
data.head()
```

Khi ấy, ta thu được dữ liệu 5 hàng đầu sau khi được ánh xạ và tạo các biến giả như sau:

	price	area	bedrooms	bathrooms	stories	mainroad	guestroom	basement	hotwaterheating	airconditioning	parking	prefarea	semi-furnished	unfurnished
0	13300000	7420	4	2	3	1	0	0	0	1	2	1	0	0
1	12250000	8960	4	4	4	1	0	0	0	1	3	0	0	0
2	12250000	9960	3	2	2	1	0	1	0	0	2	1	1	0
3	12215000	7500	4	2	2	1	0	1	0	1	3	1	0	0
4	11410000	7420	4	1	2	1	1	1	0	1	2	0	0	0

Hình 7: 5 hàng đầu tiên của bộ dữ liệu sau khi ánh xạ

Tiếp đến, ta chia bộ dữ liệu thành 2 tập, 70% huấn luyện, 30% kiểm tra, sau đó chuẩn hóa dữ liệu.

```
from sklearn.model_selection import train_test_split
np.random.seed(0)
df_train, df_test = train_test_split(
    data,
    train_size = 0.7,
    test_size = 0.3,
    random_state = 100
)
```

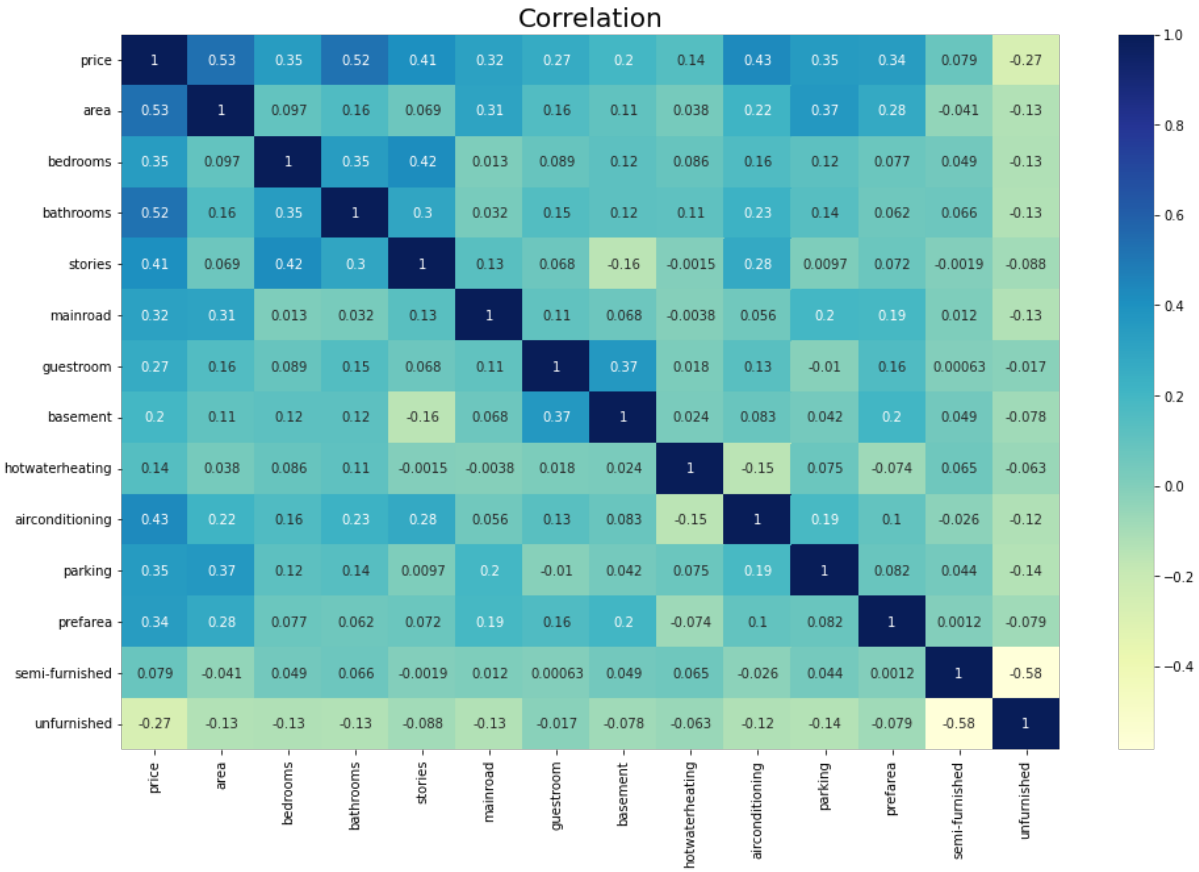
Chuẩn hóa dữ liệu cho các biến ban đầu không có dạng "yes/no" và các biến không phải biến giả.

```
from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()

# Apply scaler() to all the columns except the 'yes-no' and
# 'dummy' variables
num_vars = ['area', 'bedrooms', 'bathrooms',
            'stories', 'parking', 'price']

df_train[num_vars] = scaler.fit_transform(df_train[num_vars])
df_train = df_train.dropna()
```

Tiếp đến, ta kiểm tra tính đa cộng tuyến giữa các biến dự đoán bằng cách quan sát hệ số tương quan giữa các cặp biến trong tập các biến dự đoán xem có giá trị nào quá ngưỡng hay không.

```
# Check the correlation coefficients to see which variables are
# highly correlated
plt.figure(figsize = (16, 10))
plt.title("Correlation", fontsize=20)
corr_matrix = df_train.corr()

sns.heatmap(corr_matrix, annot = True, cmap = "YlGnBu")
plt.show()
```

Khi ấy, ta được biểu đồ như hình 8. Ta nhận thấy rằng không có sự tương quan mạnh nào xảy ra ở đây (không vượt ngưỡng 0.7).



Hình 8: Ma trận hệ số tương quan giữa các biến dự đoán

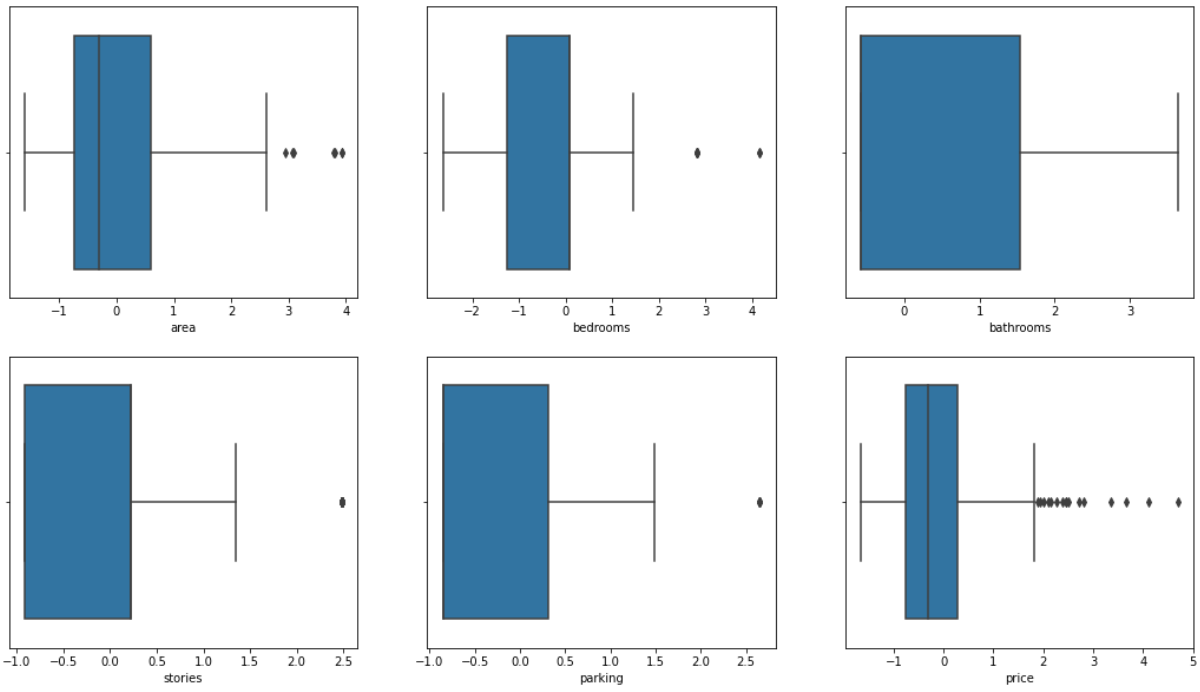
Tiếp theo, ta lọc dữ liệu bằng khoảng cách tứ phân vị, những điểm nào nằm ngoài khoảng $[Q_1 - 1.5IQR; Q_3 + 1.5IQR]$ sẽ bị loại bỏ khỏi tập dữ liệu.

```
def removeOutlierIQR(columns, data, mul=1.5):
    n_rows = math.ceil(len(columns)/3)
    fig, ax = plt.subplots(n_rows, 3, figsize=(18, 5 * n_rows))
    for i, col in enumerate(columns):
        sns.boxplot(x = data[col], ax = ax[i//3, i%3])
        Q1 = data[col].quantile(0.25)
        Q3 = data[col].quantile(0.75)
        IQR = Q3 - Q1
        data = data[(data[col]>=Q1-mul*IQR)&(data[col]<=Q3+mul*IQR)]
```

Áp dụng với các biến trong `columns` và loại bỏ dữ liệu trong `df_train`.

```
columns = ['area', 'bedrooms', 'bathrooms',
           'stories', 'parking', 'price']

removeOutlierIQR(columns, df_train)
```

Khi ấy, ta thu được hình dưới đây, với các hình tròn nhỏ ở ngoài biểu đồ hộp là các điểm Outlier cần loại bỏ.



Hình 9: Biểu đồ hộp của một số cột dữ liệu

Sau khi loại bỏ một số điểm Outlier bằng quy tắc tứ phân vị. Tiếp đến, ta thực hiện thêm hằng số vào mô hình và bắt đầu tiến hành xây dựng mô hình hồi quy tuyến tính.

```
X_train = sm.add_constant(X_train)
lm = sm.OLS(y_train, X_train).fit()
print(lm.summary())
```

Khi chạy mô hình, ta nhận được bảng kết quả thống kê tổng quan của mô hình hồi quy tuyến tính. Nhận thấy rằng vẫn còn những biến độc lập có p-value lớn nên mô hình này sẽ không cần thiết phải sử dụng đầy đủ cả 13 biến dự đoán ban đầu mà ta có thể giảm số lượng biến dự đoán bằng cách chạy hàm chọn lọc biến được đề cập ở trước.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  price   R-squared:                       0.681
Model:                            OLS   Adj. R-squared:                  0.670
Method:                 Least Squares   F-statistic:                     60.40
Date:                Thu, 22 Dec 2022   Prob (F-statistic):           8.83e-83
Time:                        07:59:41   Log-Likelihood:                -322.66
No. Observations:                 381   AIC:                             673.3
Df Residuals:                     367   BIC:                             728.5
Df Model:                          13
Covariance Type:            nonrobust
==================================================================================
                     coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------
const             -0.5374      0.106     -5.054      0.000      -0.746      -0.328
area               0.2701      0.035      7.795      0.000       0.202       0.338
bedrooms           0.0437      0.034      1.267      0.206      -0.024       0.112
bathrooms          0.2873      0.033      8.679      0.000       0.222       0.352
stories            0.2031      0.036      5.661      0.000       0.133       0.274
mainroad           0.3205      0.091      3.520      0.000       0.141       0.499
guestroom          0.1933      0.087      2.233      0.026       0.023       0.364
basement           0.1372      0.071      1.943      0.053      -0.002       0.276
hotwaterheating    0.5392      0.137      3.934      0.000       0.270       0.809
airconditioning    0.4249      0.072      5.899      0.000       0.283       0.567
parking            0.1102      0.033      3.365      0.001       0.046       0.175
prefarea           0.3776      0.075      5.040      0.000       0.230       0.525
semi-furnished     0.0058      0.075      0.078      0.938      -0.142       0.153
unfurnished       -0.1970      0.081     -2.440      0.015      -0.356      -0.038
==============================================================================
Omnibus:                       93.687   Durbin-Watson:                   2.093
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              304.917
Skew:                           1.091   Prob(JB):                     6.14e-67
Kurtosis:                       6.801   Cond. No.                         7.82
==============================================================================
```

Ta chọn lọc biến dự đoán thông qua phương thức chọn hỗn hợp.

```
def mixed_selection(curr_preds, potential_preds,
                        predictors, response, tol=.05):
    lm = sm.OLS(response, predictors[curr_preds]).fit()
    while potential_preds != []:
        best_pred = None
        best_r_squared = lm.rsquared_adj
        # loop to determine if any of the predictors can better
        # the r-squared
        for pred in potential_preds:
            preds = curr_preds[:] + [pred]
            new_r_squared = sm.OLS(response,
                        predictors[preds]).fit().rsquared_adj
            if new_r_squared > best_r_squared:
                best_r_squared = new_r_squared
                best_pred = pred
        # a potential predictor improved the r-squared; remove
        # it from potential_preds and add it to curr_preds
        if best_pred != None:
            curr_preds.append(best_pred)
            potential_preds.remove(best_pred)
        else:
            # none of the remaining potential predictors
            # improved the adjust r-squared; exit loop
            break
        # fit a new lm using the new predictors, look at the
        # p-values two-tailed p values for t-stats of the params
        pvals = sm.OLS(response,
                    predictors[curr_preds]).fit().pvalues
        # remove the feature from curr_preds that have a p-value
        # that is too large
        for feat in pvals.index:
            if pvals[feat] > tol and feat != 'const':
                curr_preds.remove(feat)
    return curr_preds
curr_preds = ['const']
potential_preds = list(lm.params.index)
potential_preds.remove('const')
mixed_selection(curr_preds,potential_preds,predictors=X_train,
                response=y_train, tol=.05)
```

Kết quả của quá trình chọn lọc này cho ra 10 biến như sau:

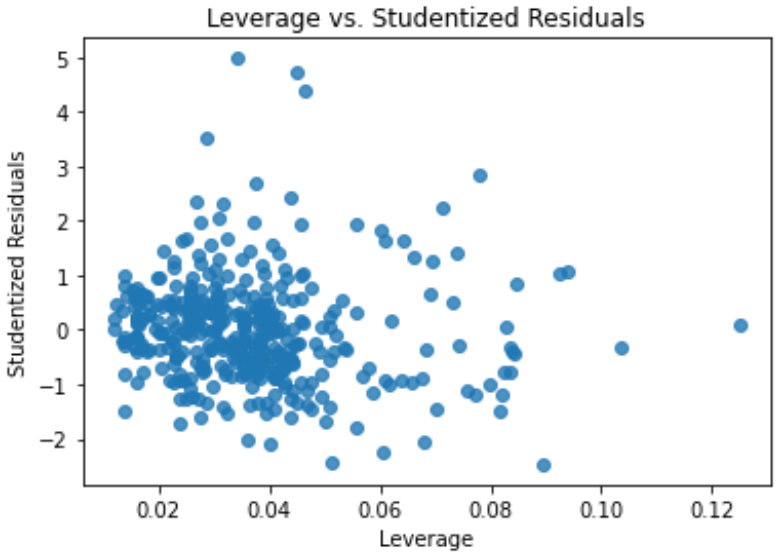
```
['const',
 'area',
 'bathrooms',
 'stories',
 'airconditioning',
 'prefarea',
 'hotwaterheating',
 'mainroad',
 'unfurnished',
 'parking',
 'guestroom']
```

Xử lý Leverage.

```
influence = lm.get_influence()
inf_sum = influence.summary_frame()
print(inf_sum.head())

student_resid = influence.resid_studentized_external
(cooks, p) = influence.cooks_distance
(dffits, p) = influence.dffits
leverage = influence.hat_matrix_diag

X_train['leverage'] = leverage
X_train['student_resid'] = student_resid

print('Leverage vs. Studentized Residuals')
sns.regplot(x = leverage, y = lm.resid_pearson, fit_reg = False)
plt.title('Leverage vs. Studentized Residuals')
plt.xlabel('Leverage')
plt.ylabel('Studentized Residuals')
plt.show()
```

Chạy mô hình sau khi chọn lọc biến và loại bỏ outlier.

```
# Model after selecting a mixture of predictors and removing
# outliers
lm_1 = sm.OLS(y_train,X_train[curr_preds]).fit()
print(lm_1.summary())
```



Hình 10: Đồ thị mối liên hệ giữa Leverage và thặng dư Student

OLS Regression Results

Dep. Variable:	price	R-squared:	0.674
Model:	OLS	Adj. R-squared:	0.665
Method:	Least Squares	F-statistic:	74.94
Date:	Thu, 22 Dec 2022	Prob (F-statistic):	5.28e-82
Time:	07:59:42	Log-Likelihood:	-317.08
No. Observations:	374	AIC:	656.2
Df Residuals:	363	BIC:	699.3
Df Model:	10		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.4700	0.094	-5.026	0.000	-0.654	-0.286
area	0.2632	0.035	7.422	0.000	0.193	0.333
bathrooms	0.3102	0.034	9.219	0.000	0.244	0.376
stories	0.2025	0.032	6.299	0.000	0.139	0.266
airconditioning	0.4288	0.072	5.975	0.000	0.288	0.570
prefarea	0.4102	0.074	5.542	0.000	0.265	0.556
hotwaterheating	0.4617	0.155	2.981	0.003	0.157	0.766
mainroad	0.2809	0.092	3.038	0.003	0.099	0.463
unfurnished	-0.2098	0.065	-3.208	0.001	-0.338	-0.081
parking	0.1111	0.033	3.390	0.001	0.047	0.175
guestroom	0.2793	0.082	3.393	0.001	0.117	0.441

Omnibus:	106.873	Durbin-Watson:	2.094
Prob(Omnibus):	0.000	Jarque-Bera (JB):	367.027
Skew:	1.253	Prob(JB):	2.00e-80
Kurtosis:	7.156	Cond. No.	7.69

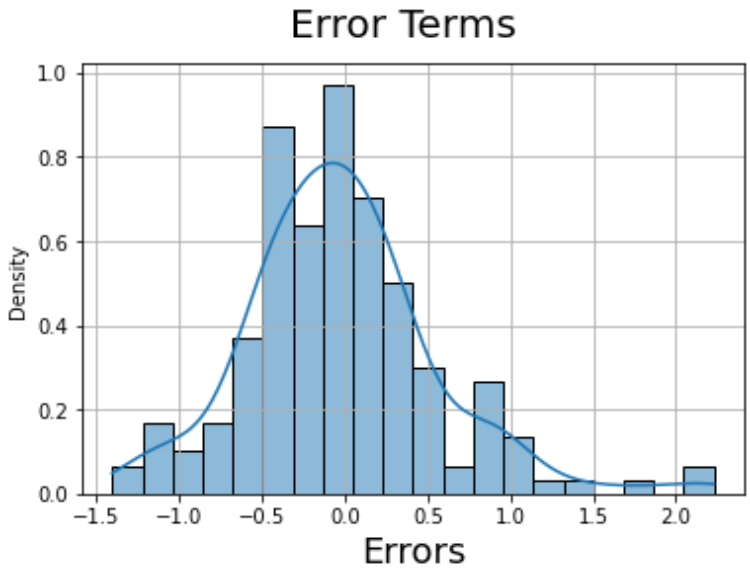
Hình 11: Kết quả chạy mô hình với 10 biến được chọn

Công đoạn tiếp theo là phân tích thặng dư. Đầu tiên, đưa ra dự đoán từ mô hình và tính toán sai số của dự đoán này từ giá trị quan sát được của biến phản hồi Y.

```
# Forecast of new observations
y_pred = lm_1.predict(X_test[curr_preds])   # yhat
error_term = y_test - y_pred # epsilonhat
```

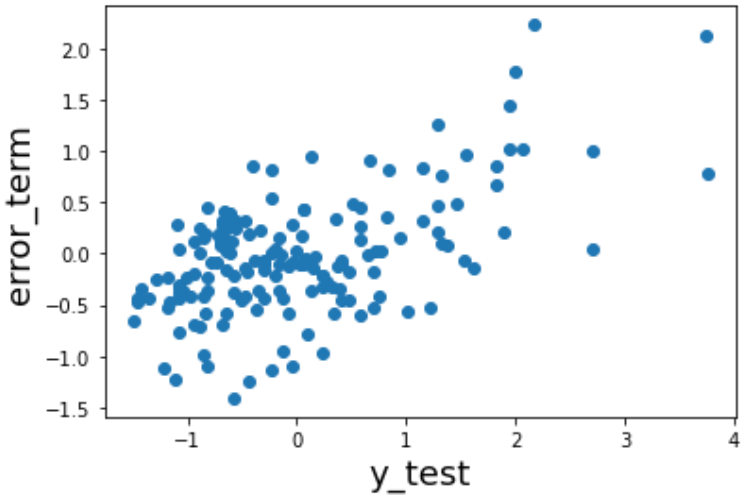
Sau đó ta vẽ đồ thị thể hiện mối quan hệ giữa giá trị dự đoán và giá trị quan sát được của biến phản hồi.

```
fig = plt.figure()
plt.grid()
sns.histplot(error_term, bins = 20, kde=True, stat="density")
fig.suptitle('Error Terms', fontsize = 20) # Plot heading
plt.xlabel('Errors', fontsize = 18)  # X-label
plt.show()
```

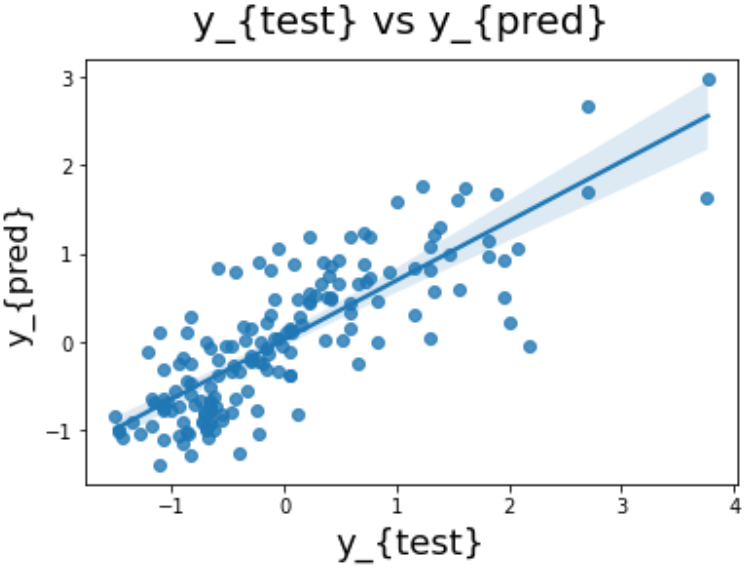


Ta lại vẽ thêm đồ thị để xem xét mối quan hệ giữa y và $\hat{\varepsilon}$, giữa y và $\hat{y}$.

```
plt.scatter(y_test, error_term)
plt.xlabel('y_test',  fontsize = 18)
plt.ylabel('error_term',  fontsize = 18)
plt.show()
```




```
fig = plt.figure()
sns.regplot(y_test, y_pred, fit_reg=True)
fig.suptitle(r'y_{test} vs y_{pred}', fontsize=20)
plt.xlabel('y_{test}', fontsize=18)
plt.ylabel('y_{pred}', fontsize=16)
plt.show()
```



Nhìn vào các đồ thị trên thì thấy rằng mô hình đã khá phù hợp, áp dụng tiêu chuẩn Student để kiểm định về tính chuẩn của phần dư ta kết luận được giá trị của các hệ số hồi quy với mô hình hồi quy tuyến tính ứng với bộ dữ liệu là:

```
const             -0.469996
area               0.263179
bathrooms          0.310166
stories            0.202548
airconditioning    0.428813
prefarea           0.410150
hotwaterheating    0.461710
mainroad           0.280866
unfurnished       -0.209827
parking            0.111054
guestroom          0.279283
dtype: float64
```

7 Mô hình hồi quy tuyến tính đa bội

7.1 Giới thiệu mô hình

Tiếp theo, dựa theo mô hình hồi quy tuyến tính đơn bội được đề cập ở trên, ta đề cập đến mô hình hồi quy tuyến tính đa bội, ở đó trả về m thành phần phản hồi tương ứng với m biến phản hồi thay vì một biến phản hồi như ở mô hình hồi quy tuyến tính đơn bội.

Khi đó ta có biến phản hồi được biểu diễn như sau

$$\boldsymbol{Y} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{nm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Y}_{(1)} & \boldsymbol{Y}_{(2)} & \cdots & \boldsymbol{Y}_{(m)} \end{bmatrix},$$

với $\boldsymbol{Y}_{(i)}$ là cột chứa giá trị n quan sát của thành phần phản hồi thứ i.

Ngoài ra ta có ma trận hệ số và ma trận nhiễu được biểu diễn lần lượt như sau

$$[\boldsymbol{\beta}] = \begin{bmatrix} \beta_{01} & \beta_{02} & \cdots & \beta_{0m} \\ \beta_{11} & \beta_{12} & \cdots & \beta_{1m} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{r1} & \beta_{r2} & \cdots & \beta_{rm} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{(1)} & \boldsymbol{\beta}_{(2)} & \cdots & \boldsymbol{\beta}_{(m)} \end{bmatrix},$$

với $\boldsymbol{\beta_{(i)}} = [\beta_{0i},\ \beta_{1i},\ \ldots\ , \beta_{ri}]'$ và

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

với $\boldsymbol{\varepsilon_{(i)}} = [\varepsilon_{0i},\ \varepsilon_{1i},\ \ldots\ ,\varepsilon_{ri}]'$. Tương tự như ở mô hình hồi quy tuyến tính đơn bội, ta giả sử $E(\boldsymbol{\varepsilon}) = 0$. Từ đây ta rút ra biểu diễn của mô hình hồi quy tuyến tính đa bội dưới dạng hệ phương trình như sau:

$$\begin{cases} \boldsymbol{Y}_{(1)} = \beta_{01} + \beta_{11}Z_1 + \cdots + \beta_{r1}Z_r + \varepsilon_{(1)}, \\ \boldsymbol{Y}_{(2)} = \beta_{02} + \beta_{12}Z_1 + \cdots + \beta_{r2}Z_r + \varepsilon_{(2)}, \\ \ldots \\ \boldsymbol{Y}_{(m)} = \beta_{0m} + \beta_{1m}Z_1 + \cdots + \beta_{rm}Z_r + \varepsilon_{(m)}. \end{cases}$$

Khác với mô hình hồi quy tuyến tính đơn bội, các nhiễu trong cùng một quan sát của mô hình hồi quy tuyến tính đa bội có thể liên quan đến nhau. Cụ thể, nhiễu ε_{ij} và ε_{ik} với $j \neq k$ có thể có hệ số tương quan khác 0,

$$\sigma_{kj} = \mathrm{Cov}(\varepsilon_{ij}, \varepsilon_{ik}) \neq 0.$$

Từ đây ta định nghĩa ma trận $\boldsymbol{\Sigma}$ như sau

$$\boldsymbol{\Sigma} = (\sigma_{ij})_{m\times m},$$

với $\sigma_{ij}\boldsymbol{I}_n = \mathrm{Cov}(\varepsilon_{(i)}, \varepsilon_{(j)}), \quad i, j = \overline{1, m}.$

Ngoài ra, ma trận giá trị $\boldsymbol{Z}$ có kích thước và biểu diễn giống như khi xét mô hình tuyến tính đơn bội

$$\boldsymbol{Z} = \begin{bmatrix} 1 & z_{11} & \ldots & z_{1r} \\ 1 & z_{21} & \ldots & z_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & \ldots & z_{nr} \end{bmatrix}.$$

Từ đây ta có phương trình ma trận tổng quát cho mô hình hồi quy tuyến tính đa bội với n biến dự đoán và m biến phản hồi.

$$\underset{(n\times m)}{\boldsymbol{Y}} = \underset{(n\times(r+1))}{\boldsymbol{Z}} \times \underset{((r+1)\times m)}{[\boldsymbol{\beta}]} + \underset{(n\times m)}{[\boldsymbol{\varepsilon}]}.$$

7.2 Ước lượng tham số

7.2.1 Ước lượng ma trận hệ số β

Trong mô hình hồi quy tuyến tính đơn bội, ước lượng $\widehat{\boldsymbol{\beta}}_{(i)}$, cột thứ i của ma trận $[\widehat{\boldsymbol{\beta}}]$, có giá trị

$$\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y}_{(i)}.$$

Khi đó, giá trị của ước lượng $\widehat{\boldsymbol{\beta}}$ là

$$\begin{aligned}[\widehat{\boldsymbol{\beta}}] &= \left[\begin{array}{cccc} \widehat{\boldsymbol{\beta}}_{(1)} & \widehat{\boldsymbol{\beta}}_{(2)} & \dots & \widehat{\boldsymbol{\beta}}_{(m)} \end{array}\right] \\ &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\left[\begin{array}{cccc} \boldsymbol{Y}_{(1)} & \boldsymbol{Y}_{(2)} & \dots & \boldsymbol{Y}_{(m)} \end{array}\right] \\ &= (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \end{aligned} \tag{7.1}$$

Với ước lượng này của $[\boldsymbol{\beta}]$, ta rút ra một số tính chất quan trọng sau

Định lý 7.1

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

Chứng minh.

Thật vậy, do kỳ vọng của ma trận sai số ngẫu nhiên $\boldsymbol{\varepsilon}$ là 0, ta có

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

Định lý 7.2

$[\widehat{\boldsymbol{\beta}}]$ là ước lượng hợp lý cực đại của $[\boldsymbol{\beta}]$ khi xét hàm hợp lý cho n quan sát

$$\mathcal{L}([\boldsymbol{\beta}], \boldsymbol{\Sigma}, [\boldsymbol{Y}]) = \frac{1}{(2\pi)^{\frac{mn}{2}} \mid \boldsymbol{\Sigma} \mid^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{Z}_i[\boldsymbol{\beta}])\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{Z}_i[\boldsymbol{\beta}])'\right),$$

với giả thiết sai số có phân phối chuẩn kỳ vọng bằng 0.

Chứng minh.

Theo giả thiết, chúng 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 $\mathcal{N}_m(\mathbf{0}, \boldsymbol{\Sigma})$ với hàm mật độ xác suất đồng thời như sau

$$f(\boldsymbol{\varepsilon}_i) = \frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} \exp\left(-\frac{\boldsymbol{\varepsilon}_i'\boldsymbol{\Sigma}^{-1}\boldsymbol{\varepsilon}_i}{2}\right).$$

Ta viết lại thành,

$$f([\boldsymbol{\beta}], \boldsymbol{\Sigma}, \boldsymbol{Y}_i) = \frac{1}{\sqrt{(2\pi)^m|\boldsymbol{\Sigma}|}} \exp\left(-\frac{S_i}{2}\right),$$

trong đó

$$S_i := \sum_{a=1}^{m}\sum_{b=1}^{m}(Y_{ia} - (Z\beta)_{ia})'\,\Sigma_{ab}^{-1}\,(Y_{ib} - (Z\beta)_{ib}).$$

Từ đây, ta xây dựng hàm hợp lý cho n quan sát như sau

$$\mathcal{L}([\boldsymbol{\beta}], \boldsymbol{\Sigma}, [\boldsymbol{Y}]) = \frac{1}{\sqrt{(2\pi)^{mn}|\boldsymbol{\Sigma}|^n}} \exp\left(-\frac{S}{2}\right),$$

trong đó

$$S := \sum_{i=1}^{n}\sum_{a=1}^{m}\sum_{b=1}^{m} (Y_{ia} - (Z\beta)_{ia})' \Sigma_{ab}^{-1} (Y_{ib} - (Z\beta)_{ib}).$$

Ta muốn tìm ước lượng $[\widehat{\boldsymbol{\beta}}]$ để cực đại hóa $\mathcal{L}$. Do đại lượng $\frac{1}{\sqrt{(2\pi)^{mn}|\boldsymbol{\Sigma}|^n}}$ không phụ thuộc vào $[\boldsymbol{\beta}]$, ta cần cực tiểu hóa S. Khi đó, ta cần tính $\frac{\partial S}{\partial \beta_{kc}}$ với $k = 1, 2, \ldots, r+1$ và $c = 1, 2, \ldots, m$.

Trước hết, ta cố định i, với $a \neq c$, khi đó ta được

$$S_i = (Y_{ia} - (Z\beta)_{ia})' \sum_{b=1}^{m} \Sigma_{ab}^{-1} (Y_{ib} - (Z\beta)_{ib}).$$

Vì vậy,

$$\begin{aligned}
\frac{\partial S_i}{\partial \beta_{kc}} &= \frac{\partial}{\partial \beta_{kc}} (Y_{ia} - (Z\beta)_{ia})' [\Sigma_{a1}^{-1} (Y_{i1} - (Z\beta)_{i1}) + \Sigma_{a2}^{-1} (Y_{i2} - (Z\beta)_{i2}) \\
&\qquad + \cdots + \Sigma_{ac}^{-1} (Y_{ic} - (Z\beta)_{ic}) + \ldots] \\
&= (Y_{ia} - (Z\beta)_{ia})' \Sigma_{ac}^{-1} \frac{\partial}{\partial \beta_{kc}} (Y_{ic} - (Z\beta)_{ic}) \\
&= (Y_{ia} - (Z\beta)_{ia})' \Sigma_{ac}^{-1} \frac{\partial}{\partial \beta_{kc}} \left(Y_{ic} - \sum_{j=1}^{m} Z_{ij}\beta_{jc}\right) \\
&= -(Y_{ia} - (Z\beta)_{ia})' \Sigma_{ac}^{-1} Z_{ik}
\end{aligned}$$

Với $a = c$ ta có:

$$\begin{aligned}
S_i &= (Y_{ic} - (Z\beta)_{ic})' \sum_{b=1}^{m} \Sigma_{cb}^{-1} (Y_{ib} - (Z\beta)_{ib}) \\
&= (Y_{ic} - (Z\beta)_{ic})' [\Sigma_{c1}^{-1} (Y_{i1} - (Z\beta)_{i1}) + \Sigma_{c2}^{-1} (Y_{i2} - (Z\beta)_{i2}) \\
&\qquad + \cdots + \Sigma_{cc}^{-1} (Y_{ic} - (Z\beta)_{ic}) + \ldots]
\end{aligned}$$

Vậy nên,

$$\begin{aligned}
\frac{\partial S_i}{\partial \beta_{kc}} &= \frac{\partial}{\partial \beta_{kc}} (Y_{ic} - (Z\beta)_{ic})' [\Sigma_{c1}^{-1} (Y_{i1} - (Z\beta)_{i1}) + \Sigma_{c2}^{-1} (Y_{i2} - (Z\beta)_{i2}) \\
&\qquad + \cdots + \Sigma_{cc}^{-1} (Y_{ic} - (Z\beta)_{ic}) + \ldots] \\
&= -Z_{ik} \sum_{b=1}^{m} \Sigma_{cb}^{-1} (Y_{ib} - (Z\beta)_{ib}) - (Y_{ic} - (Z\beta)_{ic}) \Sigma_{cc}^{-1} Z_{ik}
\end{aligned}$$

Từ 2 điều trên ta suy ra:

$$\frac{\partial S_i}{\partial \beta_{kc}} = -2Z_{ik}\sum_{b=1}^{m}\Sigma_{cb}^{-1}\left(Y_{ib} - (Z\beta)_{ib}\right).$$

Tổng hợp n quan sát ta được:

$$\begin{aligned}\frac{\partial S}{\partial \beta_{kc}} &= -2\sum_{i=1}^{n}Z_{ik}\sum_{b=1}^{m}\Sigma_{cb}^{-1}\left(Y_{ib} - (Z\beta)_{ib}\right)\\ &= -2\sum_{i=1}^{n}\sum_{b=1}^{m}\Sigma_{cb}^{-1}\left(Z_{ik}Y_{ib} - Z_{ik}(Z\beta)_{ib}\right)\\ &= -2\sum_{b=1}^{m}\Sigma_{cb}^{-1}\left(Z'Y - Z'Z\beta\right)_{kb}.\end{aligned}$$

Điều này đúng $\forall k = 1, 2, \ldots, r+1$ và $c = 1, 2, \ldots, m$ hay là ta phải có $\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{Z}'[\boldsymbol{Y}] - \boldsymbol{Z}'\boldsymbol{Z}[\boldsymbol{\beta}]\right)' = \boldsymbol{0}$.

Như vậy, $\boldsymbol{Z}'[\boldsymbol{Y}] - \boldsymbol{Z}'\boldsymbol{Z}[\boldsymbol{\beta}] = \boldsymbol{0} \Rightarrow [\widehat{\boldsymbol{\beta}}] = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'[\boldsymbol{Y}]$.

Định lý 7.3

$E([\widehat{\boldsymbol{\varepsilon}}]) = \boldsymbol{0}$ và $\mathrm{Cov}(\boldsymbol{\beta}_{(i)}, \boldsymbol{\beta}_{(j)}) = \sigma_{ij}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}$

Chứng minh.

Dễ thấy $E([\bar{\varepsilon}]) = \boldsymbol{0}$ vì với mọi $j = \overline{1, m}$, ta có

$$E\left(\widehat{\boldsymbol{\varepsilon}}_{(j)}\right) = E\left(\boldsymbol{Y}_{(j)} - \boldsymbol{Z}\widehat{\boldsymbol{\beta}}_{(j)}\right) = E\left(\boldsymbol{Z}\left(\boldsymbol{\beta}_{(j)} - \widehat{\boldsymbol{\beta}}_{(j)}\right) + \boldsymbol{\varepsilon}_{(j)}\right) = \boldsymbol{0}$$

Mặt khác, với phép đặt $\boldsymbol{C} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$, ta có

$$\begin{aligned}\mathrm{Cov}\left(\widehat{\boldsymbol{\beta}}_{(i)}, \widehat{\boldsymbol{\beta}}_{(j)}\right) &= \mathrm{Cov}\left(\boldsymbol{C}\boldsymbol{Y}_{(i)}, \boldsymbol{C}\boldsymbol{Y}_{(j)}\right)\\ &= E\left[\left(\boldsymbol{C}\boldsymbol{Y}_{(i)} - E\left(\boldsymbol{C}\boldsymbol{Y}_{(i)}\right)\right)\left(\boldsymbol{C}\boldsymbol{Y}_{(j)} - E\left(\boldsymbol{C}\boldsymbol{Y}_{(j)}\right)\right)'\right]\\ &= \boldsymbol{C}E\left[\left(\boldsymbol{Y}_{(i)} - \boldsymbol{Z}\boldsymbol{\beta}_{(i)}\right)\left(\boldsymbol{Y}_{(j)} - \boldsymbol{Z}\boldsymbol{\beta}_{(j)}\right)'\right]\boldsymbol{C}'\\ &= \boldsymbol{C}E\left[\boldsymbol{\varepsilon}_{(i)}\boldsymbol{\varepsilon}'_{(j)}\right]\boldsymbol{C}'\\ &= \boldsymbol{C}\,\mathrm{Cov}\left(\boldsymbol{\varepsilon}_{(i)}, \boldsymbol{\varepsilon}_{(j)}\right)\boldsymbol{C}'\\ &= \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\left(\sigma_{ij}\boldsymbol{I}\right)\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\\ &= \sigma_{ij}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\end{aligned}$$

Định lý 7.4

$[\widehat{\boldsymbol{\beta}}]$ có phân phối chuẩn.

Chứng minh.

Với giả thiết $\boldsymbol{\varepsilon}_{(i)} \sim \mathcal{N}_n\left(0, \sigma_{ii}\boldsymbol{I}\right)$ ta được $\boldsymbol{Y}_{(i)} \sim \mathcal{N}_n\left(\boldsymbol{Z}\boldsymbol{\beta}, \sigma_{ii}\boldsymbol{I}\right)$. Do đó

$$\widehat{\boldsymbol{\beta}}_{(i)} = \left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}\boldsymbol{Y}_{(i)} \sim \mathcal{N}_{r+1}\left(\boldsymbol{\beta}_{ii,}, \sigma_{ii}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\right)$$

Như vậy $[\widehat{\boldsymbol{\beta}}]$ có phân phối chuẩn.

7.2.2 Ước lượng ma trận biến phản hồi Y

Sử dụng ước lượng bình phương cực tiểu $[\widehat{\boldsymbol{\beta}}]$ vừa tìm được ở công thức (7.1), ta dễ dàng xây dựng các ma trận ước lượng của $[\boldsymbol{Y}]$ và $[\boldsymbol{\varepsilon}]$ như sau:

$$[\widehat{\boldsymbol{Y}}] = \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}] = \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'[\boldsymbol{Y}],$$
$$[\widehat{\boldsymbol{\varepsilon}}] = [\boldsymbol{Y}] - [\widehat{\boldsymbol{Y}}] = \left(\boldsymbol{I} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\right)[\boldsymbol{Y}].$$

Hơn nữa, ta có

$$\boldsymbol{Z}'[\widehat{\boldsymbol{\varepsilon}}] = \boldsymbol{Z}'\left(\boldsymbol{I} - \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\right)[\boldsymbol{Y}] = \left(\boldsymbol{Z}' - \boldsymbol{Z}'\boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'\right)[\boldsymbol{Y}] = \boldsymbol{0},$$
$$[\widehat{\boldsymbol{Y}}]'[\widehat{\boldsymbol{\varepsilon}}] = (\boldsymbol{Z}[\widehat{\boldsymbol{\beta}}])'[\widehat{\boldsymbol{\varepsilon}}] = [\widehat{\boldsymbol{\beta}}]'\boldsymbol{Z}'[\widehat{\boldsymbol{\varepsilon}}] = \boldsymbol{0}.$$

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

$$[\boldsymbol{Y}]'[\boldsymbol{Y}] = ([\widehat{\boldsymbol{Y}}] + [\widehat{\boldsymbol{\varepsilon}}])'([\widehat{\boldsymbol{Y}}] + [\widehat{\boldsymbol{\varepsilon}}]) = [\widehat{\boldsymbol{Y}}]'[\widehat{\boldsymbol{Y}}] + [\widehat{\boldsymbol{\varepsilon}}]'[\widehat{\boldsymbol{\varepsilon}}].$$

7.2.3 Ước lượng ma trận Σ

Định lý 7.5

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

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}[\widehat{\boldsymbol{\varepsilon}}]'[\widehat{\boldsymbol{\varepsilon}}] = \frac{1}{n}([\boldsymbol{Y}] - \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}])'([\boldsymbol{Y}] - \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}]),$$

là ước lượng hợp lý cực đại của $\boldsymbol{\Sigma}$.

Chứng minh.

Trước hết ta viết lại ma trận sai số $\boldsymbol{\varepsilon}$ như sau

$$[\boldsymbol{\varepsilon}] = \begin{bmatrix} \varepsilon_{11} & \varepsilon_{12} & \cdots & \varepsilon_{1m} \\ \varepsilon_{21} & \varepsilon_{22} & \cdots & \varepsilon_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \varepsilon_{n2} & \cdots & \varepsilon_{nm} \end{bmatrix} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Khi đó hàm mật độ đồng thời của mỗi ε_i là

$$f(\varepsilon_i) = \frac{1}{(2\pi)^{m/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}\varepsilon_i \boldsymbol{\Sigma}^{-1}\varepsilon_i'\right)$$
$$= \frac{1}{(2\pi)^{m/2}\det(\boldsymbol{\Sigma})^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{Y}_i - \boldsymbol{Z}_i[\boldsymbol{\beta}])\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{Z}_i[\boldsymbol{\beta}])'\right),$$

với ma trận $\boldsymbol{\Sigma}$ được định nghĩa như trên.

Từ đó với n quan sát độc lập, ta có hàm hợp lý cực đại cho n quan sát

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{Y}) = \frac{1}{(2\pi)^{\frac{mn}{2}}\det(\boldsymbol{\Sigma})^{\frac{n}{2}}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{Z}_i\boldsymbol{\beta})\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{Z}_i\boldsymbol{\beta})'\right),$$

$$\boldsymbol{Q} = [q_{ij}]_{n\times n} := [\widehat{\boldsymbol{\varepsilon}}]'[\boldsymbol{\varepsilon}] = ([\boldsymbol{Y}] - \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}])'([\boldsymbol{Y}] - \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}]).$$

Khi đó

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

trong đó y_{ij} và $\widehat{y}_{ij}$ lần lượt là phần tử ở hàng i, cột j của ma trận $[\boldsymbol{Y}]$ và $[\widehat{\boldsymbol{Y}}] = \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}]$. Với $\left(\boldsymbol{\Sigma}^{-1}\right)_{ij}$ là phần tử ở hàng i, cột j của ma trận Σ^{-1}, ta có biến đổi

$$\begin{aligned}
\sum_{i=1}^{n}(\boldsymbol{Y}_i - \boldsymbol{Z}\boldsymbol{\beta}_i)\boldsymbol{\Sigma}^{-1}(\boldsymbol{Y}_i - \boldsymbol{Z}\boldsymbol{\beta}_i)' &= \sum_{i=1}^{n}\left(\sum_{j=1}^{m}\sum_{k=1}^{m}(y_{ij} - \widehat{y}_{ij})\left(\boldsymbol{\Sigma}^{-1}\right)_{jk}(y_{ik} - \widehat{y}_{ik})\right) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m}\left(\left(\boldsymbol{\Sigma}^{-1}\right)_{jk}\sum_{i=1}^{n}(y_{ij} - \widehat{y}_{ij})(y_{ik} - \widehat{y}_{ik})\right) \\
&= \sum_{j=1}^{m}\sum_{k=1}^{m}\left(\boldsymbol{\Sigma}^{-1}\right)_{jk} q_{jk} \\
&= \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\right).
\end{aligned}$$

Như vậy hàm hợp lý trên có thể được viết lại thành

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

Tới đây ta sẽ tìm ma trận $\boldsymbol{\Sigma}$ để cực đại hóa log-hàm hợp lý

$$\log\mathcal{L}([\boldsymbol{\beta}], \boldsymbol{\Sigma}, [\boldsymbol{Y}]) = -\frac{mn}{2}\log 2\pi + \frac{n}{2}\log\det\left(\boldsymbol{\Sigma}^{-1}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\right).$$

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

Ta viết lại log-hàm hợp lý, khi đó ta được

$$\begin{aligned}\log \mathcal{L}([\boldsymbol{\beta}], \boldsymbol{\Sigma}, [\boldsymbol{Y}]) &= -\frac{mn}{2}\log 2\pi + \frac{n}{2}\log\det\left(\boldsymbol{Q}^{-1/2}\boldsymbol{A}\boldsymbol{Q}^{-1/2}\right) - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{Q}^{-1/2}\boldsymbol{A}\boldsymbol{Q}^{-1/2}\boldsymbol{Q}\right)\\ &= -\frac{mn}{2}\log 2\pi + \frac{n}{2}\log\det\left(\boldsymbol{Q}^{-1}\right) + \frac{n}{2}\log\det\boldsymbol{A} - \frac{1}{2}\operatorname{tr}\left(\boldsymbol{Q}^{-1/2}\boldsymbol{A}\boldsymbol{Q}^{1/2}\right)\\ &= -\frac{mn}{2}\log 2\pi - \frac{n}{2}\log\det\boldsymbol{Q} + \frac{n}{2}\log\det\boldsymbol{A} - \frac{1}{2}\operatorname{tr}\boldsymbol{A}\end{aligned}$$

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

$$\det(\boldsymbol{A}) = \prod_i \lambda_i, \quad \text{và} \quad \operatorname{tr}(\boldsymbol{A}) = \sum_i \lambda_i.$$

Từ các kết quả này, ta được

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

Như vậy log-hàm hợp lý 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 mỗi $i = \overline{1,m}$. Mặt khác, do $\boldsymbol{A}$ là ma trận đối xứng nên ta có thể viết $\boldsymbol{A}$ dưới dạng chéo hóa

$$\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}',$$

trong đó $\boldsymbol{\Lambda} = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ và $\boldsymbol{P}$ là ma trận trực giao. Do các trị riêng của $\boldsymbol{A}$ đều bằng n nên

$$\boldsymbol{A} = \boldsymbol{P}(n\boldsymbol{I})\boldsymbol{P}' = n\boldsymbol{P}\boldsymbol{P}' = n\boldsymbol{I}.$$

Từ đây suy ra ước lượng hợp lý cực đại của $\boldsymbol{\Sigma}$ là

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

Định lý 7.6

$\widehat{\boldsymbol{\Sigma}}$ là ước lượng chệch của $\boldsymbol{\Sigma}$.

Chứng minh.

Trước hết ta chứng minh $E\left[\widehat{\boldsymbol{\varepsilon}}'_{(i)}\widehat{\boldsymbol{\varepsilon}}_{(j)}\right] = \sigma_{ij}(n-r-1)$, với mọi $i, j = \overline{1,m}$. Thật vậy, với các ma trận đối xứng lũy đẳng $\boldsymbol{H} := \boldsymbol{Z}\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}'$ và $\boldsymbol{I} - \boldsymbol{H}$, ta có biến

đổi sau

$$\begin{aligned}
E\left[\widehat{\boldsymbol{\varepsilon}}'_{(j)}\widehat{\boldsymbol{\varepsilon}}_{(k)}\right] &= E\left[\left(\boldsymbol{Y}_{(j)} - \widehat{\boldsymbol{Y}}_{(j)}\right)'\left(\boldsymbol{Y}_{(k)} - \widehat{\boldsymbol{Y}}_{(k)}\right)\right] \\
&= E\left[\left(\boldsymbol{Y}_{(j)} - \boldsymbol{H}\boldsymbol{Y}_{(j)}\right)'\left(\boldsymbol{Y}_{(k)} - \boldsymbol{H}\boldsymbol{Y}_{(k)}\right)\right] \\
&= E\left[\boldsymbol{Y}'_{(j)}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}_{(k)}\right] \\
&= E\left[\left(\boldsymbol{Y}_{(j)} - \boldsymbol{Z}\boldsymbol{\beta}_{(j)}\right)'(\boldsymbol{I} - \boldsymbol{H})\left(\boldsymbol{Y}_{(k)} - \boldsymbol{Z}\boldsymbol{\beta}_{(k)}\right)\right] \\
&= E\left[\boldsymbol{\varepsilon}'_{(j)}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}_{(k)}\right] \\
&= E\left[\operatorname{tr}\left((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}_{(k)}\boldsymbol{\varepsilon}'_{(j)}\right)\right] \\
&= \operatorname{tr}\left((\boldsymbol{I} - \boldsymbol{H})E\left[\boldsymbol{\varepsilon}_{(k)}\boldsymbol{\varepsilon}'_{(j)}\right]\right) \\
&= \sigma_{jk}\operatorname{tr}(\boldsymbol{I} - \boldsymbol{H}) \\
&= \sigma_{jk}(n - r - 1).
\end{aligned}$$

Như vậy, $E(\widehat{\boldsymbol{\Sigma}}) = (n - r - 1)\boldsymbol{\Sigma}/n$ và do đó $\widehat{\boldsymbol{\Sigma}}$ là ước lượng chệch của $\boldsymbol{\Sigma}$. Từ chứng minh trên, ta có thể suy ra được ước lượng không chệch của $\boldsymbol{\Sigma}$.

7.3 Kiểm định tham số của mô hình

7.3.1 Phương pháp Wilk's Lambda

Mục đích của bài toán kiểm định này là lược bỏ những thuộc tính có mức ảnh hưởng không đáng kể đến kết quả của mô hình hồi quy tuyến tính, để có thể giảm bớt thời gian tính toán và không quá ảnh hưởng đến độ chính xác của mô hình.

Cặp giả thuyết - đối thuyết ta cần kiểm định như sau

$$H_0 \ : \ [\boldsymbol{\beta}]_{(2)} = \boldsymbol{0}; \quad H_1 \ : \ [\boldsymbol{\beta}]_{(2)} \neq \boldsymbol{0} \quad \text{ở đây} \quad [\boldsymbol{\beta}] = \begin{bmatrix} [\boldsymbol{\beta}]_{(1)} \\ {\scriptstyle (q+1)\times m} \\ [\boldsymbol{\beta}]_{(2)} \\ {\scriptstyle (r-q)\times m} \end{bmatrix}$$

Khi đó nếu ta viết $\boldsymbol{Z}$ dưới dạng

$$\boldsymbol{Z} = \left[\begin{array}{c:c} \underset{n\times(q+1)}{\boldsymbol{Z}_{(1)}} & \underset{n\times(r-q)}{\boldsymbol{Z}_{(2)}} \end{array}\right]$$

Dưới tác động của giả thuyết H_0 thì mô hình trở thành $\boldsymbol{Y} = \boldsymbol{Z}_{(1)}[\boldsymbol{\beta}]_{(1)} + [\boldsymbol{\varepsilon}]_{(1)}$. Độ lệch giữa các giá trị sai số trung bình bình phương của hai mô hình khi đó là

$$\left(\widehat{Y}_{(1)} - \boldsymbol{Z}_{(1)}[\widehat{\boldsymbol{\beta}}]_{(1)}\right)'\left(\widehat{Y}_{(1)} - \boldsymbol{Z}_{(1)}[\widehat{\boldsymbol{\beta}}]_{(1)}\right) - (\widehat{Y} - \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}])'(\widehat{Y} - \boldsymbol{Z}[\widehat{\boldsymbol{\beta}}]) = n\left(\widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}}\right),$$

trong đó

$$[\widehat{\boldsymbol{\beta}}]_{(1)} = \left(\boldsymbol{Z}'_{(1)}\boldsymbol{Z}_{(1)}\right)^{-1}\boldsymbol{Z}'_{(1)}[\boldsymbol{Y}] \ \text{ và } \ \widehat{\boldsymbol{\Sigma}}_1 = \frac{1}{n}\left(\widehat{Y}_{(1)} - \boldsymbol{Z}_{(1)}[\widehat{\boldsymbol{\beta}}]_{(1)}\right)'\left(\widehat{Y}_{(1)} - \boldsymbol{Z}_{(1)}[\widehat{\boldsymbol{\beta}}]_{(1)}\right).$$

Xét tỷ số kiểm định Λ được định nghĩa bởi

$$\Lambda := \frac{\max\limits_{[\boldsymbol{\beta}]_{(1)}\boldsymbol{\Sigma}} \mathcal{L}\left([\boldsymbol{\beta}]_{(1)}, \boldsymbol{\Sigma}\right)}{\max\limits_{[\boldsymbol{\beta}],\boldsymbol{\Sigma}} \mathcal{L}([\boldsymbol{\beta}], \boldsymbol{\Sigma})} = \frac{\mathcal{L}\left([\widehat{\boldsymbol{\beta}}]_{(1)}, \widehat{\boldsymbol{\Sigma}}_1\right)}{\mathcal{L}([\widehat{\boldsymbol{\beta}}], \widehat{\boldsymbol{\Sigma}})} = \left(\frac{\det \widehat{\boldsymbol{\Sigma}}}{\det \widehat{\boldsymbol{\Sigma}}_1}\right)^{\frac{n}{2}}.$$

Khi đó thống kê Wilk's lambda được xác định bởi

$$\Lambda^{\frac{2}{n}} = \frac{\det \widehat{\boldsymbol{\Sigma}}}{\det \widehat{\boldsymbol{\Sigma}}_1}.$$

Từ thống kê này, ta đưa ra tiêu chuẩn để kiểm định tham số như sau.

Định lý 7.7

Xét mô hình hồi quy tuyến tính đa bội $[\boldsymbol{Y}] = \boldsymbol{Z}[\boldsymbol{\beta}] + [\boldsymbol{\varepsilon}]$ với rank $\boldsymbol{Z} = r+1 \leq n-m$ và nhiễu $[\boldsymbol{\varepsilon}]$ có phân phối chuẩn. Khi đó ta bác bỏ $H_0 : [\boldsymbol{\beta}]_{(2)} = \mathbf{0}$ nếu như đại lượng

$$-2\ln\Lambda = -n\ln\frac{\det \widehat{\boldsymbol{\Sigma}}}{\det \widehat{\boldsymbol{\Sigma}}_1}$$

nhận giá trị đủ lớn. Hơn nữa khi n lớn, ta sử dụng thống kê hiệu chỉnh

$$T = -\left(n - r - 1 - \frac{1}{2}(m - r + q - 1)\right)\ln\frac{\det \widehat{\boldsymbol{\Sigma}}}{\det \widehat{\Sigma}_1},$$

thống kê này sẽ xấp xỉ phân phối Chi-bình phương $\chi^2_{m(r-q)}$.

Chứng minh của định lý trên xem tại [2].

7.3.2 Một số phương pháp kiểm định khác

Ngoài việc sử dụng tỷ số hợp lý (the likelihood ratio), các phép thử khác đã được đề xuất để kiểm định giả thuyết $H_0 : \boldsymbol{\beta}_{(2)} = \mathbf{0}$. Chẳng hạn, đặt $\boldsymbol{E}$ là ma trận tổng bình phương và tích vô hướng các nhiễu

$$\boldsymbol{E} = n\boldsymbol{\Sigma}.$$

Tiếp theo, ta có $\boldsymbol{H}$ là ma trận tổng bình phương và tích vô hướng của phần dư

$$\boldsymbol{H} = n\left(\widehat{\boldsymbol{\Sigma}}_1 - \widehat{\boldsymbol{\Sigma}}\right).$$

Ta sẽ tìm các trị riêng $\eta_1 \geq \eta_2 \geq \cdots \geq \eta_s$ của ma trận $\boldsymbol{H}\boldsymbol{E}^{-1}$, trong đó $s = \min\{p, r-q\}$, tương đương với việc tìm các nghiệm của phương trình

$$\det\left(\widehat{\boldsymbol{\Sigma}}_1 - (\eta+1)\widehat{\boldsymbol{\Sigma}}\right) = 0.$$

Từ đây ta có các thống kê

$$\begin{aligned}
\text{Wilk's lambda} &= \prod_{i=1}^{s} \frac{1}{1+\eta_i} = \frac{\det \boldsymbol{E}}{\det(\boldsymbol{E}+\boldsymbol{H})}.\\
\text{Pillai's trace} &= \sum_{i=1}^{s} \frac{\eta_i}{1+\eta_i} = \operatorname{tr}\left(\boldsymbol{H}(\boldsymbol{H}+\boldsymbol{E})^{-1}\right).\\
\text{Hotelling-Lawley trace} &= \sum_{i=1}^{s} \eta_i = \operatorname{tr}\left(\boldsymbol{H}\boldsymbol{E}^{-1}\right).\\
\text{Roy's greatest root} &= \frac{\eta_1}{1+\eta_1}.
\end{aligned}$$

Trong các phép thử trên, phép thử Roy sẽ cho kết quả tệ so với 3 phép thử còn lại nếu như số lượng quan sát quá ít hoặc ma trận $\boldsymbol{H}\boldsymbol{E}^{-1}$ có nhiều trị riêng lớn xấp xỉ nhau, khi đó, phép thử Roy sẽ cho kết quả đối ngược với các phép thử khác. Vì vậy, ta chỉ nên dùng phép thử Roy khi có đủ số lượng quan sát ($n > 30$) và ma trận $\boldsymbol{H}\boldsymbol{E}^{-1}$ chỉ có duy nhất một trị riêng trội.

7.4 Dự đoán từ ước lượng mô hình

Xét mô hình hồi quy tuyến tính đa bội $[\boldsymbol{Y}] = \boldsymbol{Z}[\boldsymbol{\beta}] + [\boldsymbol{\varepsilon}]$ với rank $\boldsymbol{Z} = r+1 \leq n-m$ và nhiễu $[\boldsymbol{\varepsilon}]$ có phân phối chuẩn, với các tham số đã được ước lượng và tinh chỉnh. Khi đó, ta có thể tiến hành sử dụng mô hình trong việc dự đoán với các dữ liệu sẵn có. Vấn đề được đặt ra là ta phải dự đoán trung bình của các biến phản hồi dựa trên một quan sát $\boldsymbol{z}_0$ mà ta vừa thu thập được. Suy luận về trung bình của các biến phản hồi có thể được đưa ra dựa vào các phân phối mà ta đã kết luận được từ việc ước lượng các tham số hồi quy. Trước hết ta có

$$\widehat{\boldsymbol{\beta}}'\boldsymbol{z}_0 \sim \mathcal{N}_m\left(\boldsymbol{\beta}'\boldsymbol{z}_0, \boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0\boldsymbol{\Sigma}\right) \quad \text{và} \quad n\widehat{\boldsymbol{\Sigma}} \sim \boldsymbol{W}_{n-r-1}(\boldsymbol{\Sigma}).$$

Giá trị cần tìm của hàm hồi quy là $\boldsymbol{\beta}'\boldsymbol{z}_0$, dựa vào kiến thức về thống kê T^2 ta có thể viết

$$T^2 = \left(\frac{[\widehat{\boldsymbol{\beta}}]'\boldsymbol{z}_0 - [\boldsymbol{\beta}]'\boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0'\left(\boldsymbol{Z}'\mathbf{Z}\right)^{-1}\boldsymbol{z}_0}}\right)' \left(\frac{n}{n-r-1}\widehat{\boldsymbol{\Sigma}}\right)^{-1} \left(\frac{[\widehat{\boldsymbol{\beta}}]'\boldsymbol{z}_0 - [\boldsymbol{\beta}]'\boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0}}\right),$$

và miền ellipsoid với độ tin cậy $100(1-\alpha)\%$ của $\widehat{\boldsymbol{\beta}}'\boldsymbol{z}_0$ được cho bởi bất đẳng thức

$$\begin{aligned}
&\left(\frac{[\widehat{\boldsymbol{\beta}}]'\boldsymbol{z}_0 - [\boldsymbol{\beta}]'\boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0}}\right)' \left(\frac{n}{n-r-1}\widehat{\boldsymbol{\Sigma}}\right)^{-1} \left(\frac{[\widehat{\boldsymbol{\beta}}]'\boldsymbol{z}_0 - [\boldsymbol{\beta}]'\boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0}}\right)\\
&\qquad \leq \boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0\left[\frac{m(n-r-1)}{n-r-m}F_{m,n-r-m}(\alpha)\right],
\end{aligned}$$

mà ở đó, $F_{m,n-r-m}(\alpha)$ là phân vị trên mức $100\alpha\%$ của phân phối Fisher với các bậc tự do là m và $n-r-m$. Từ đây ta có khoảng tin cậy đồng thời với độ tin cậy $100(1-\alpha)\%$ của $E\left(\boldsymbol{Y}_i \mid \boldsymbol{Z}=\boldsymbol{z}_0\right)=\boldsymbol{z}_0'\boldsymbol{\beta}_{(i)}$ là

$$\boldsymbol{z}_0\widehat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m}F_{m,n-r-m}(\alpha)}\sqrt{\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0\left(\frac{n}{n-r-1}\widehat{\sigma}_{ii}\right)},$$

với $\widehat{\boldsymbol{\beta}}_{(i)}$ là cột thứ i của ma trận $\widehat{\boldsymbol{\beta}}$ và $\widehat{\sigma}_{ii}$ là phần tử ở hàng thứ i trên đường chéo chính của ma trận $\widehat{\boldsymbol{\Sigma}}$. Một vấn đề khác là việc đưa ra dự đoán về các giá trị của $\boldsymbol{Y}_0=[\boldsymbol{\beta}]'\boldsymbol{z}_0+\boldsymbol{\varepsilon}_0$, trong đó $\boldsymbol{\varepsilon}_0$ độc lập với các vector trong ma trận $[\boldsymbol{\varepsilon}]$. Bây giờ ta lại có

$$\boldsymbol{Y}_0-[\widehat{\boldsymbol{\beta}}]'\boldsymbol{z}_0 \sim \mathcal{N}_m\left(\boldsymbol{0},\left(1+\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0\right)\boldsymbol{\Sigma}\right).$$

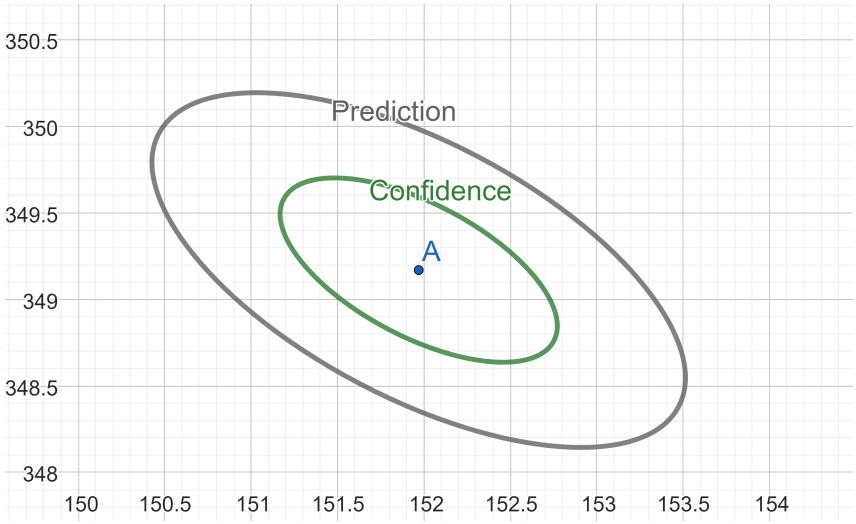
Khi đó miền ellipsoid dự đoán với độ tin cậy $100(1-\alpha)$ % là:

$$\left(\frac{[\widehat{\boldsymbol{\beta}}]'\boldsymbol{z}_0-[\boldsymbol{\beta}]'\boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0}}\right)'\left(\frac{n}{n-r-1}\widehat{\boldsymbol{\Sigma}}\right)^{-1}\left(\frac{[\widehat{\boldsymbol{\beta}}]'\boldsymbol{z}_0-[\boldsymbol{\beta}]'\boldsymbol{z}_0}{\sqrt{\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0}}\right)$$
$$\leq\left(1+\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0\right)\left[\frac{m(n-r-1)}{n-r-m}F_{m,n-r-m}(\alpha)\right].$$

Khi đó, khoảng dự đoán đồng thời với độ tin cậy $100(1-\alpha)$ % là:

$$\boldsymbol{z}_0\widehat{\boldsymbol{\beta}}_{(i)} \pm \sqrt{\frac{m(n-r-1)}{n-r-m}F_{m,n-r-m}(\alpha)}\sqrt{\left(1+\boldsymbol{z}_0'\left(\boldsymbol{Z}'\boldsymbol{Z}\right)^{-1}\boldsymbol{z}_0\right)\left(\frac{n}{n-r-1}\widehat{\sigma}_{ii}\right)}.$$

So sánh hai công thức trên, ta nhận xét rằng miền elip dự đoán sẽ có kích thước lớn hơn miền elip tin cậy, tuy nhiên hai elip này sẽ có cùng tâm, như hình vẽ dưới đây.



8 Tổng kết

Trong bài báo cáo này, nhóm tác giả đã trình bày những tìm hiểu về chủ đề *Mô hình hồi quy tuyến tính*, bao gồm: phần giới thiệu về hồi quy (phần 1, 2), ước lượng bình phương cực tiểu của hệ số hồi quy (phần 3), một vài suy luận về mô hình hồi quy cùng các yếu tố liên quan (phần 4, 6), ước lượng hàm hồi quy (phần 5), và nghiên cứu về mô hình hồi quy tuyến tính đơn bội và đa bội (phần 7).

Với mong muốn người đọc được dễ hình dung hơn về mô hình, hiểu rõ hơn về những lý thuyết cần thiết, có được một tài liệu tham khảo tốt khi muốn tìm hiểu về các mô hình hồi quy tuyến tính, mang mô hình ứng dụng vào thực tế. Các định lý, kết quả trong mỗi phần đều được nhóm tác giả chứng minh và nhóm tác giả cũng đã thử xây dựng mô hình hồi quy tuyến tính đa biến dựa trên một bộ dữ liệu thực tế được trình bày trong phần 6.11.

Lời cảm ơn

Báo cáo này được thực hiện và hoàn thành tại Đại học Bách Khoa Hà Nội, nằm trong nội dung học phần *Phân tích số liệu* kỳ học 20221.

Thông qua bài báo cáo này, nhóm tác giả đã đúc kết được cho mình những kinh nghiệm và tích lũy thêm kiến thức đối với học phần Phân tích số liệu. Nhóm tác giả xin gửi lời cảm ơn sâu sắc đến ThS. Lê Xuân Lý, giảng viên giảng dạy học phần. Trong suốt quá trình học tập học phần này, thầy đã có những góp ý và đề xuất cũng như tận tình chỉ dạy để các nhóm nói chung và nhóm 6 nói riêng có thể hoàn thiện báo cáo một cách tốt nhất. Nhóm tác giả xin kính chúc thầy sức khỏe và thành công trên con đường sắp tới.

Tài liệu

[1] Dormann, C. F., J. Elith, S. Bacher, et al. (2013). *Collinearity: a review of methods to deal with it and a simulation study evaluating their performance*. Ecography, **36**(1), 27-46.

[2] Box, G. E. P (1949). *A General Distribution Theory for a Class of Likelihood Criteria*. Biometrika, **36**, 317-346.

[3] Richard Johnson, Dean Wichern (2014). *Applied Multivariate Statistical Analysis*. Pearson New International Edition, **7**, 360-401.