NHÓM 8:
• Chu Thị Bảo Ngọc - 18110009
• Nguyễn Thị Minh Mỹ - 18110152
• Nguyễn Hoàng Thông - 1611383

MÔ HÌNH HỒI QUY TUYẾN TÍNH NHIỀU CHIỀU

NỘI DUNG TÌM HIỂU :

PHẦN 1: GIỚI THIỆU

Xét ví dụ sau:

-Một căn nhà rộng $x_1(m^2)$ có x_2 phòng ngủ và cách trung tâm thành phố $x_3(km)$. Giả sử chúng ta đã có số liệu thống kê từ 1000 căn nhà trong thành phố đó, liệu rằng khi có một căn nhà mới với các thông số về diện tích, số phòng ngủ và khoảng cách tới trung tâm, chúng ta có thể dự đoán được giá của căn nhà đó không? Nếu có thì hàm dự đoán $\mathbf{y} = \mathbf{f}(\mathbf{x})$ sẽ có dạng như nào?. Ở đây $\mathbf{x} = [x_1, x_2, x_3]$ là vector hàng chứa thông tin input,$\mathbf{y}$ là một số vô hướng (scalar) biểu diễn output (tức là giá của) căn nhà trong ví dụ này.

-Một cách đơn giản nhất, chúng ta có thể thấy rằng:
(i) diện tích nhà càng lớn thì giá nhà càng cao.
(ii) số lượng phòng ngủ càng lớn thì giá nhà càng cao.
(iii) càng xa trung tâm thì giá nhà càng giảm.

-Một hàm số đơn giản nhất có thể mô tả mối quan hệ giữa giá nhà và 3 đại lượng đầu vào này là gì?

-Phân tích hồi quy là một phương pháp thống kê nhằm dự báo giá trị của một hay nhiều biến phụ thuộc dựa vào tập hợp các giá trị của một số biến độc lập. Đây cũng là phương pháp để xác định xem các biến độc lập ảnh hưởng lên các biến phụ thuộc như thế nào.

PHẦN 2: MÔ HÌNH HỒI QUY TUYẾN TÍNH CỔ ĐIỂN

-Cho $z_1, z_2, ..., z_r$ là r biến độc lập (independent variables, predictor variables, explanatory variables) có liên quan đến biến phụ thuộc Y (dependent variable, response variable). Với r=4, ta có:

$$\begin{aligned} \text{Y} &= \text{Giá trị thị trường của nhà ở} \\ z_1 &= \text{Diện tích khu vực sinh sống} \\ z_2 &= \text{Vị trí} \\ z_3 &= \text{Giá trị thẩm định năm ngoái} \\ z_4 &= \text{Chất lượng xây dựng} \end{aligned}$$

-Để mô hình hóa quan hệ tuyến tính trong đó diễn tả sự thay đổi của biến Y theo biến z cho trước người ta sử dụng mô hình hồi quy tuyến tính cổ điển. Mô hình hồi quy tuyến tính cổ điển có dạng như sau:

$$\text{Y} = \beta_0 + \beta_1 z_1 + ... + \beta_r z_r + \epsilon$$

$$[\text{Biến phụ thuộc}] = [\text{Trung bình(phụ thuộc vào } z_1, z_2, ..., z_r)] + [\text{Sai số}]$$

-Từ đây ta có thể xây dựng với n quan sát độc lập trên Y và các giá trị liên quan z_i thì mô hình hoàn chỉnh có dạng như sau:

$$\begin{aligned} Y_1 &= \beta_0 + \beta_1 z_{11} + \beta_2 z_{12} + ... + \beta_r z_{1r} + \epsilon_1 \\ Y_2 &= \beta_0 + \beta_1 z_{21} + \beta_2 z_{22} + ... + \beta_r z_{2r} + \epsilon_2 \\ &\vdots \\ Y_n &= \beta_0 + \beta_1 z_{n1} + \beta_2 z_{n2} + ... + \beta_r z_{nr} + \epsilon_n \end{aligned} \tag{7-1}$$

Trong đó:
Y_n: Giá trị biến phụ thuộc Y trong lần quan sát thứ n.
z_{nr} : Giá trị của biến độc lập z trong lần quan sát thứ n.
β_r : Các tham số chưa biết.
ϵ_n : Sai số trong lần quan sát thứ n.

-Và sai số có các tính chất sau:

$$\begin{aligned} &(i): && E(\epsilon_j) = 0 \\ &(ii): && Var(\epsilon_j) = \sigma^2 (\text{hằng số}) \\ &(iii): && \text{Cov}(\epsilon_j, \epsilon_k) = 0, j \neq k. \end{aligned} \tag{7-2}$$

-Ở dạng ma trận, (7-1) trở thành:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & z_{11} & z_{12} & ... & z_{1r} \\ 1 & z_{21} & z_{22} & ... & z_{2r} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{n1} & z_{n2} & ... & z_{nr} \end{bmatrix} . \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_r \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

hoặc

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

-Và lúc này ta có các tính chất sau:

$$\begin{aligned} &(i): && E(\boldsymbol{\epsilon}) = \mathbf{0} \\ &(ii): && Cov(\boldsymbol{\epsilon}) = E(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2 \boldsymbol{I} \end{aligned}$$

-Mỗi cột của **Z** bao gồm n giá trị biến độc lập tương ứng trong khi hàng thứ j của **Z** chứa các giá trị cho tất cả các biến dự báo trong lần thử thứ j.

$$\boxed{\begin{gathered} \underset{(n\times 1)}{\boldsymbol{Y}} = \underset{(n\times(r+1))}{\boldsymbol{Z}} \underset{((r+1)\times 1)}{\boldsymbol{\beta}} + \underset{(n\times 1)}{\boldsymbol{\epsilon}} \\ E(\boldsymbol{\epsilon}) = \underset{(n\times 1)}{\mathbf{0}}; Cov(\boldsymbol{\epsilon}) = \underset{(n\times n)}{\sigma^2 \boldsymbol{I}} \end{gathered}} \tag{7-3}$$

Trong đó, β và σ^2 là những tham số chưa biết và ma trận hồi quy (ma trận mô hình) **Z** có dòng thứ j $[z_{j0}, z_{j1}, ..., z_{jr}]$.

Ví dụ 7.1: (Dự báo mô hình hồi quy đường thẳng). Xác định mô hình hồi quy tuyến tính theo dạng đường thẳng.

$$E(Y) = \beta_0 + \beta_1 z_1$$

và dữ liệu sau:

z_1	0	1	2	3	4
y	1	4	3	8	9

Giải:

-Đặt **Y'**= $[Y_1, Y_2, Y_3, Y_4, Y_5]$ là các biến phụ thuộc (biến kết cục) được quan sát và sai số ngẫu nhiên $\boldsymbol{\epsilon}' = [\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4, \epsilon_5]$ và ta có thể viết như sau:$\boldsymbol{Y = Z\beta + \epsilon}$ với:

$$\mathbf{Y}=\begin{bmatrix}Y_1\\Y_2\\\vdots\\Y_5\end{bmatrix},\mathbf{Z}=\begin{bmatrix}1&z_{11}\\1&z_{21}\\\vdots&\vdots\\1&z_{51}\end{bmatrix},\boldsymbol{\beta}=\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix},\boldsymbol{\epsilon}=\begin{bmatrix}\epsilon_1\\\epsilon_2\\\vdots\\\epsilon_5\end{bmatrix}$$

-Dữ liệu bài toán này được chứa trong vetor quan sát phụ thuộc $\boldsymbol{y}$ và ma trận hồi quy $\boldsymbol{Z}$ với:

$$\mathbf{y}=\begin{bmatrix}1\\4\\3\\8\\9\end{bmatrix},\mathbf{Z}=\begin{bmatrix}1&0\\1&1\\1&2\\1&3\\1&4\end{bmatrix}\text{ Lúc đó: }\begin{bmatrix}1\\4\\3\\8\\9\end{bmatrix}=\begin{bmatrix}1&0\\1&1\\1&2\\1&3\\1&4\end{bmatrix}\begin{bmatrix}\beta_0\\\beta_1\end{bmatrix}+\begin{bmatrix}\epsilon_1\\\epsilon_2\\\vdots\\\epsilon_5\end{bmatrix}$$

Ví dụ 7.2 (Ma trận hồi quy cho ANOVA một chiều dưới dạng mô hình hồi quy). Xác định ma trận nếu mô hình hồi quy tuyến tính được áp dụng cho tình huống ANOVA một chiều trong ví dụ 6.7.

Population 1: 9,6,9
Population 2: 0,2
Population 3: 3,1,2

Giải:

-Ta đặt các biến giả như sau:

$$z_1=\begin{cases}1 & \text{nếu quan sát từ quần thể 1}\\0 & \text{nếu khác}\end{cases},z_2=\begin{cases}1 & \text{nếu quan sát từ quần thể 2}\\0 & \text{nếu khác}\end{cases},z_3=\begin{cases}1 & \text{nếu quan sát từ quần thể 3}\\0 & \text{nếu khác}\end{cases}$$

và $\beta_0=\mu,\beta_1=\tau_1,\beta_2=\tau_2,\beta_3=\tau_3$. Lúc đó:

$$Y_j=\beta_0+\beta_1z_{j1}+\beta_2z_{j2}+\beta_3z_{j3}+\epsilon_j;i,j=1,2,...,8$$

-Trong đó ta sắp xếp các quan sát từ 3 quần thể theo thứ tự. Do đó, ta thu hồi được vector phụ thuộc Y và ma trận hồi quy (ma trận mô hình) như sau:

$$\underset{(8\times1)}{\boldsymbol{Y}}=\begin{bmatrix}9\\6\\9\\0\\2\\3\\1\\2\end{bmatrix};\underset{(8\times4)}{\boldsymbol{Z}}=\begin{bmatrix}1&1&0&0\\1&1&0&0\\1&1&0&0\\1&0&1&0\\1&0&1&0\\1&0&0&1\\1&0&0&1\\1&0&0&1\end{bmatrix}$$

PHẦN 3: ƯỚC LƯỢNG BÌNH PHƯƠNG TỐI THIỂU

-Phương pháp bình phương tối thiểu (nhỏ nhất) được đưa ra bởi nhà toán học người Đức-Carl F. Gauss, tên gọi là OLS (ordinary least squares). Mục tiêu của phương pháp này là tìm cực tiểu cho bình phương sai số.

-Theo phương pháp bình phương tối thiểu, ta cần tìm **b** sao cho tổng bình phương sai số

$$S(\mathbf{b})=\sum_{j=1}^{n}(y_j-b_0-b_1z_{j1}-b_2z_{j2}-...-b_rz_{jr})^2=(\mathbf{y}-\mathbf{Zb})'(\mathbf{y}-\mathbf{Zb}),\quad j=1,2,..,n$$

đạt giá trị nhỏ nhất.

Trong đó:
$\mathbf{y}'=[y_1,...,y_n]$
$\mathbf{Z}=[1,z_{j1},...,z_{jr}]$
$\mathbf{b}'=[b_0,...,b_r]$

-Các hệ số **b** tìm được từ phương pháp OLS được gọi là các ước lượng bình phương tối thiểu của các tham số hồi quy $\boldsymbol{\beta}$, kí hiệu là $\hat{\boldsymbol{\beta}}$.

-Giá trị $\hat{y_j} = \hat{\beta}_0 + \hat{\beta}_1 z_{j1} + ... + \hat{\beta}_r z_{jr}$ được gọi là ước lượng hay giá trị dự báo của y_j

-Các sai số $\hat{\epsilon_j} = y_j - \hat{\beta}_0 - \hat{\beta}_0 z_{j1} - ... - \hat{\beta}_r z_{jr}, (j = \overline{1,n})$ được gọi là những phần dư (residuals).

Kết quả 7.1: **Z** có hạng đầy đủ là $r+1 \leq n$. Lúc này ta có ước lượng bình phương tối thiểu của β được xác định như sau:

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

Đặt $\hat{\mathbf{y}} = \mathbf{Z}\hat{\boldsymbol{\beta}} = \mathbf{Hy}$ biểu thị giá trị ước lượng của biến phụ thuộc (fitted values) của **y** với $\mathbf{H} = \mathbf{Z}(\mathbf{Z}'.\mathbf{Z})^{-\mathbf{1}}.\mathbf{Z}'$ và **H** được gọi là ma trận "mũ" ("hat" matrix). Lúc này phần dư (sai số) là:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = [\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z'Z})^{-\mathbf{1}}\boldsymbol{Z'}]\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$$

thỏa mãn $\boldsymbol{Z'\hat{\epsilon}} = \mathbf{0}$ và $\hat{\boldsymbol{y}}'\hat{\boldsymbol{\epsilon}} = 0$. Ngoài ra:

$$\begin{aligned}\text{Tổng bình phương (phần dư)} &= \sum_{j=1}^{n}(y_j - \hat{\beta}_0 - \hat{\beta}_1 z_{j1} - \hat{\beta}_2 z_{j2} - ... - \hat{\beta}_r z_{jr})^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \\ &= \boldsymbol{y'}[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z'Z})^{-\mathbf{1}}\boldsymbol{Z'}]\boldsymbol{y} = \boldsymbol{y'y} - \boldsymbol{y'Z}\hat{\boldsymbol{\beta}}\end{aligned}$$

Chứng minh:

-Phương pháp OLS sẽ lựa chọn các hệ số hồi quy trong vetor sao cho bình phương sai số của mô hình ước lượng là nhỏ nhất. Đặt

$$\begin{aligned}S(\boldsymbol{\beta}) = \sum_{j=1}^{n}(y_j - \beta_0 - \beta_1 z_{j1} - \beta_2 z_{j2} - ... - \beta_r z_{jr})^2 &= \|\boldsymbol{y} - \boldsymbol{Z\beta}\|^2 \\ &= \langle \boldsymbol{y} - \boldsymbol{Z\beta}, \boldsymbol{y} - \boldsymbol{Z\beta} \rangle \\ &= (\boldsymbol{y} - \boldsymbol{Z\beta})^T (\boldsymbol{y} - \boldsymbol{Z\beta}) \\ &= \left(\boldsymbol{y^T} - \boldsymbol{\beta^T Z^T}\right)(\boldsymbol{y} - \boldsymbol{Z\beta}) \\ &= \boldsymbol{y^T y} - \boldsymbol{y^T Z\beta} - \boldsymbol{\beta^T Z^T y} + \boldsymbol{\beta^T Z^T Z\beta}\end{aligned}$$

-Vì $\boldsymbol{y^T Z\beta}$ là ma trận đối xứng nên $\boldsymbol{y^T Z\beta} = \left(\boldsymbol{y^T Z\beta}\right)^T = \boldsymbol{\beta^T Z^T y} \Rightarrow S(\boldsymbol{\beta}) = \boldsymbol{y^T y} - \mathbf{2}\boldsymbol{y^T Z\beta} + \boldsymbol{\beta^T Z^T Z\beta}$

-Ta đi tìm $\bigtriangledown S(\boldsymbol{\beta}) = \bigtriangledown \left(\boldsymbol{y^T y} - \mathbf{2}\boldsymbol{y^T Z\beta} + \boldsymbol{\beta^T Z^T Z\beta}\right) = \bigtriangledown \boldsymbol{\beta^T Z^T Z\beta} - 2 \bigtriangledown \boldsymbol{y^T Z\beta}$

• $\bigtriangledown \boldsymbol{y^T Z\beta}$

-Đặt $\varphi(\boldsymbol{\beta}) = \boldsymbol{yZ\beta^T} = \boldsymbol{a}^T\boldsymbol{\beta}$ với $\boldsymbol{y^T Z} = \boldsymbol{a^T} = (a_0, a_1, ..., a_r), \boldsymbol{a} \in \Re^{r+1}$

$\Rightarrow \varphi(\boldsymbol{\beta}) = a_0\beta_0 + a_1\beta_1 + ... + a_r\beta_r \Rightarrow \dfrac{\partial\varphi(\boldsymbol{\beta})}{\partial\beta_i} = a_i$ với i=0,...,r

$$\Rightarrow \bigtriangledown\varphi(\boldsymbol{\beta}) = \begin{bmatrix} \dfrac{\partial\varphi}{\varphi\beta_0} \\ \vdots \\ \dfrac{\partial\varphi}{\varphi\beta_r} \end{bmatrix} = \begin{bmatrix} a_0 \\ \vdots \\ a_r \end{bmatrix} = \mathbf{a} \Rightarrow \bigtriangledown\varphi(\boldsymbol{\beta}) = \bigtriangledown(\boldsymbol{y^T Z\beta}) = a = \boldsymbol{Z^T y} \text{ và } \bigtriangledown^2(\boldsymbol{y^T Z\beta}) = 0 \quad (1)$$

• $\bigtriangledown \boldsymbol{\beta^T Z^T Z\beta}$

-Đặt $\boldsymbol{Z^T Z} = \boldsymbol{A}$ với $\boldsymbol{A}$ là ma trận (r+1) dòng và (r+1) cột.

$\Rightarrow \boldsymbol{\beta^T Z^T Z\beta} = \boldsymbol{\beta^T A\beta} = \sum_{i=1}^{r+1}\sum_{j=1}^{r+1} a_{ij}\beta_i\beta_j = a_{11}\beta_1^2 + a_{22}\beta_2^2 + ... + a_{12}\beta_1\beta_2 + a_{21}\beta_2\beta_1$

$$\begin{aligned}
\varphi(x_1,...,x_n) =& \sum_{i=1}^{r+1}[x_i \sum_{j=1}^{r+1} a_{ij}x_j] \\
\frac{\partial \varphi}{\partial x_k}(\mathbf{X}) =& \sum_{i=1}^{r+1}\left(\frac{\partial}{\partial x_k}(x_i)\sum_{j=1}^{r+1} a_{ij}x_j + x_i \frac{\partial}{\partial x_k}(\sum_{j=1}^{r+1} a_{ij}x_j)\right) \text{ với k=1,...,r+1} \\
=& \sum_{j=1}^{r+1} a_{kj}x_j + \sum_{i=1}^{r+1} x_i a_{ik} \\
=& 2\sum_{j=1}^{r+1} a_{kj}x_j = 2[\mathbf{A}]_k\mathbf{X} \\
\Rightarrow & \bigtriangledown \varphi(x) = 2\mathbf{A}\text{x}
\end{aligned}$$

-Vậy $\bigtriangledown(\boldsymbol{\beta^T Z^T Z \beta}) = 2(\boldsymbol{Z^T Z})\boldsymbol{\beta}$ và $\bigtriangledown^2(\boldsymbol{\beta^T Z^T Z \beta}) = 2(\boldsymbol{Z^T Z})$ (2)

-Từ (1) và (2) ta có $\bigtriangledown S(\boldsymbol{\beta}) = 2(\mathbf{Z^T Z})\boldsymbol{\beta} - 2\boldsymbol{Z^T y} = 0 \Rightarrow \boldsymbol{\beta} = (\boldsymbol{Z^T Z})^{-1}\boldsymbol{Z^T y}$

-Như vậy bằng phương pháp OLS ta đã tìm được $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z^T Z})^{-1}\boldsymbol{Z^T y}$ ■

Ví dụ 7.3 (Tính ước lượng bình phương tối thiểu, sai số, tổng bình phương sai số)
Tính ước lượng bình phương tối thiểu $\hat{\boldsymbol{\beta}}$, sai số $\boldsymbol{\epsilon}$ và tổng bình phương sai số cho mô hình hồi quy có dạng đường thẳng (bậc nhất)

$$Y_j = \beta_0 + \beta_1 z_{j1} + \epsilon_j$$

phù hợp với bộ dữ liệu sau

z_1	0	1	2	3	4
y	1	4	3	8	9

-Ta có:

$$\mathbf{Z}=\begin{bmatrix}1 & 0\\ 1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 4\end{bmatrix} \Rightarrow \mathbf{Z'} = \begin{bmatrix}1 & 1 & 1 & 1 & 1\\ 0 & 1 & 2 & 3 & 4\end{bmatrix}, \mathbf{y}=\begin{bmatrix}1\\4\\3\\8\\9\end{bmatrix}$$

$$\mathbf{Z'Z}=\begin{bmatrix}1 & 1 & 1 & 1 & 1\\ 0 & 1 & 2 & 3 & 4\end{bmatrix}\cdot\begin{bmatrix}1 & 0\\ 1 & 1\\ 1 & 2\\ 1 & 3\\ 1 & 4\end{bmatrix} = \begin{bmatrix}5 & 10\\ 10 & 30\end{bmatrix}, (\mathbf{Z'Z})^{-1} = \begin{bmatrix}0.6 & -0.2\\ -0.2 & 0.1\end{bmatrix}$$

$$\mathbf{Z'y}=\begin{bmatrix}1 & 1 & 1 & 1 & 1\\ 0 & 1 & 2 & 3 & 4\end{bmatrix}\cdot\begin{bmatrix}1\\4\\3\\8\\9\end{bmatrix} = \begin{bmatrix}25\\70\end{bmatrix}$$

-Từ các tính toán trên ta có được $\hat{\boldsymbol{\beta}} = \begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = \boldsymbol{(Z'Z)^{-1}Z'y} = \begin{bmatrix}0.6 & -0.2\\ -0.2 & 0.1\end{bmatrix}\cdot\begin{bmatrix}25\\70\end{bmatrix} = \begin{bmatrix}1\\2\end{bmatrix}$

Vậy phương trình phù hợp là $\hat{y} = 1 + 2z$

-Vector giá trị dự báo là: $\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}} \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 9 \end{bmatrix}$. Từ đó ta có sai số là $\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \begin{bmatrix} 1 \\ 4 \\ 3 \\ 8 \\ 9 \end{bmatrix} - \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \\ 9 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{bmatrix}$

-Tổng bình phương sai số là: $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} = \begin{bmatrix} 0 & 1 & -2 & 1 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 1 \\ -2 \\ 1 \\ 0 \end{bmatrix} = 6$

3.1 Phân rã tổng bình phương:

-Phân rã tổng bình phương của trung bình có công thức như sau:

$$\boxed{\sum_{j=1}^{n}(y_j - \bar{y})^2 = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 + \sum_{j=1}^{n}\hat{\epsilon}_j^2}$$

-Ý nghĩa của các thành phần:

$\sum_{j=1}^{n}(y_j - \bar{y})^2$: Total sum of squares - TSS: là tổng bình phương của tất cả các sai lệch giữa các giá trị quan sát y_j và giá trị trung bình.

$\sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2$: Regression sum of squares - RSS : là tổng bình phương tất cả các sai lệch giữa các giá trị của biến phụ thuộc $\boldsymbol{y}$ nhận được từ hàm hồi quy mẫu và giá trị trung bình của chúng. Phần này đo độ chính xác của hàm hồi quy.

$\sum_{j=1}^{n}\hat{\epsilon}_j^2$:Error sum of squares - ESS: là tổng bình phương của tất cả các sai lệch giữa các quan sát $\boldsymbol{y}$ và các giá trị nhận được từ hàm hồi quy.

-Ta có thể viết thành: TSS = RSS + ESS

Chứng minh:

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(y_i - \overline{y} + \hat{y}_i - \hat{y}_i)^2 = \sum_{i=1}^{n}((\hat{y}_i - \bar{y}) + \underbrace{(y_i - \hat{y}_i)}_{\hat{\varepsilon}_i})^2 \\
&= \sum_{i=1}^{n}((\hat{y}_i - \bar{y})^2 + 2\hat{\varepsilon}_i(\hat{y}_i - \bar{y}) + \hat{\varepsilon}_i^2) \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{y}_i - \bar{y}) \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n}\hat{\varepsilon}_i(\hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \cdots + \hat{\beta}_r z_{ir} - \bar{y}) \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 + 2(\hat{\beta}_0 - \bar{y})\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i}_{0} + 2\hat{\beta}_1\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i z_{i1}}_{0} + \cdots + 2\hat{\beta}_r\underbrace{\sum_{i=1}^{n}\hat{\varepsilon}_i z_{ir}}_{0} \\
&= \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\hat{\varepsilon}_i^2 = \text{ESS} + \text{RSS}
\end{aligned}$$

-Nếu đem chia 2 vế của TSS = RSS + ESS cho TSS ta được:

$$1 = \frac{RSS}{TSS} + \frac{ESS}{TSS} = \frac{\sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2} + \frac{\sum_{j=1}^{n}\hat{\epsilon}_j^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$\Rightarrow R^2 = \frac{RSS}{TSS} = 1 - \frac{ESS}{TSS}$$

$$= 1 - \frac{\sum_{j=1}^{n}\hat{\epsilon}_j^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

-Trong đó $\mathbf{R^2}$ được gọi là hệ số xác định (coefficient of determination) một con số thống kê tổng hợp khả năng giải thích của một phương trình. Nó biểu thị tỷ lệ biến thiên của biến phụ thuộc do tổng mức biến thiên của các biến giải thích gây ra. Như vậy, $\mathbf{R^2}$ phải nằm giữa 0 và 1. Khi $\mathbf{R^2}$ càng gần 0, khả năng giải thích càng kém và điều ngược lại sẽ đúng khi các giá trị của nó tiến dần tới 1.

3.2 Khía cạnh hình học của bình phương tối thiểu:

-Vector trung bình của biến phụ thuộc $\mathbf{Y}=\boldsymbol{E(Y)} = \boldsymbol{Z\beta} = \beta_0 \begin{bmatrix}1\\1\\\vdots\\1\end{bmatrix} + \beta_1 \begin{bmatrix}z_{11}\\z_{21}\\\vdots\\z_{n1}\end{bmatrix} + ... + \beta_r \begin{bmatrix}z_{r1}\\z_{r1}\\\vdots\\z_{r1}\end{bmatrix}$

-Như ta thấy,$\boldsymbol{E(Y)}$ là một tổ hợp tuyến tính của các cột của $\boldsymbol{Z}$
-Khi $\boldsymbol{\beta}$ thay đổi, $\boldsymbol{Z\beta}$ mở rộng trên mặt phẳng mô hình của tất cả các tổ hợp tuyến tính. Thông thường, vector phụ thuộc $\boldsymbol{y}$ sẽ không nằm trong mặt phẳng mô hình vì sai số ngẫu nhiên $\boldsymbol{\epsilon}$, nghĩa là $\boldsymbol{y}$ không (chính xác) là tổ hợp tuyến tính của các cột của $\boldsymbol{Z}$

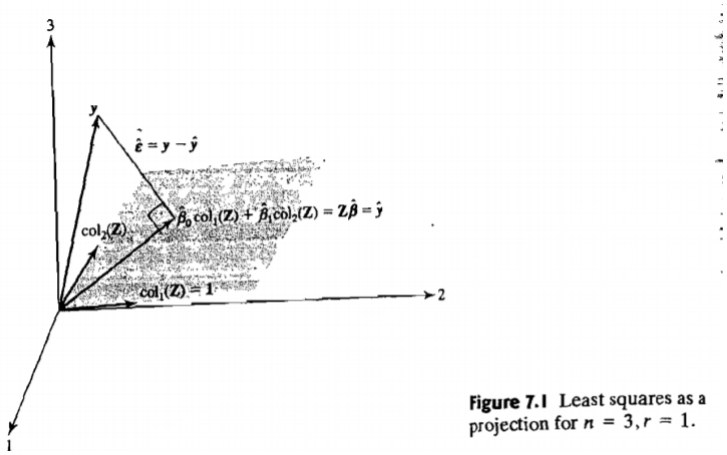


Figure 7.1 Least squares as a projection for $n = 3, r = 1$.

3.3 Thuộc tính lấy mẫu của ước lượng bình phương tối thiểu cổ điển:

Kết quả 7.2: Theo mô hình hồi quy tuyến tính (7-3), ước lượng bình phương tối thiểu $\hat{\boldsymbol{\beta}} = \boldsymbol{(Z.Z')^{-1}Z'Y}$ có

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

Các sai số $\hat{\boldsymbol{\epsilon}}$ có tính chất sau:

$$E(\hat{\boldsymbol{\epsilon}}) = \mathbf{0} \text{ và } \mathrm{Cov}(\hat{\boldsymbol{\epsilon}}) = \sigma^2[\boldsymbol{I - Z(Z'Z)^{-1}Z'}] = \sigma^2[\boldsymbol{I - H}]$$

Đồng thời, $E(\hat{\boldsymbol{\epsilon}}'\boldsymbol{\epsilon}) = (n - r - 1)\sigma^2$. Do đó, khi định nghĩa:

$$s^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-(r+1)} = \frac{\mathbf{Y}'[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}']\mathbf{Y}}{n-r-1} = \frac{\mathbf{Y}'[\mathbf{I} - \mathbf{H}]\mathbf{Y}}{n-r-1}$$

thì ta có $E(s^2) = \sigma^2$

Hơn nữa, $\hat{\boldsymbol{\beta}}$ và $\hat{\boldsymbol{\epsilon}}$ không tương quan.

Chứng minh: *Tham khảo tại link http://www.prenhall.com/statistics*

-Ước lượng bình phương tối thiểu $\hat{\boldsymbol{\beta}}$ có tính chất là phương sai tối thiểu lần đầu tiên được thiết lập bởi Gauss. Kết quả 3 sau đây liên quan đến các ước lượng "tốt nhất" của các hàm tham số tuyến tính có dạng:
$c_0\beta_0 + c_1\beta_1 + ... + c_r\beta_r$ với $\mathbf{c}$ bất kì.

***Kết quả 7.3*(Định lý Gauss-Markov về bình phương tối thiểu):**

Cho $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}$ với $E(\boldsymbol{\epsilon}) = 0, Cov(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$ và $\boldsymbol{Z}$ là ma trận có thứ hạng đầy đủ r+1. Công cụ ước lượng

$$\boldsymbol{c'\hat{\beta}} = c_0\hat{\beta}_0 + c_1\hat{\beta}_1 + ... + c_r\hat{\beta}_r$$

là ước lượng tuyến tính không chệch có phương sai nhỏ nhất trong số các ước lượng tuyến tính có dạng

$$\mathbf{a'Y} = a_1Y_1 + a_2Y_2 + ... + a_nY_n$$

không chệch cho $\boldsymbol{c'\beta}$.

Chứng minh:

-Với bất kì $\mathbf{c}$, đặt $\mathbf{a'Y}$ là ước lượng không chệch cho $\boldsymbol{c'\beta}$ tức là

$$\begin{aligned} E(\mathbf{a'Y}) &= \boldsymbol{c'\beta} \\ E(\boldsymbol{a'Z\beta + a'\epsilon}) &= \boldsymbol{c'\beta} \\ \boldsymbol{a'Z\beta} &= \boldsymbol{c'\beta} \\ (\boldsymbol{c' - a'Z})\boldsymbol{\beta} &= 0 \\ \Rightarrow \mathbf{c'} &= \mathbf{a'Z} \end{aligned}$$

-Ta xét $\boldsymbol{c'\hat{\beta}} = \boldsymbol{c'}(\boldsymbol{Z'Z})^{-1}\mathbf{Z'Y} = \mathbf{a^{*\prime}Y}$ với $\mathbf{a^*} = \mathbf{Z(Z'Z)^{-1}c} \Rightarrow E(\boldsymbol{c'\hat{\beta}}) = \boldsymbol{c'}E(\hat{\boldsymbol{\beta}}) = \boldsymbol{c'\beta}$ (Do kết quả 7.2: $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$))

-Vậy $\boldsymbol{c'\hat{\beta}} = \mathbf{a^{*\prime}Y}$ là ước lượng không chệch $\boldsymbol{c'\beta}$. Do đó, với bất kỳ c thỏa mãn yêu cầu ước lượng không chệch

$$\begin{aligned} Var(\mathbf{a'Y}) &= Var(\boldsymbol{a'Z\beta + a'\epsilon}) = Var(\boldsymbol{a'\epsilon}) = \boldsymbol{a'I}\sigma^2\boldsymbol{a} \\ &= \sigma^2(a - a^* + a^*)'(a - a^* + a^*) \\ &= \sigma^2[(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*) + \mathbf{a}^{*\prime}\mathbf{a}^*] \end{aligned}$$

-Vì $(\mathbf{a} - \mathbf{a}^*)'\mathbf{a}^* = (\mathbf{a} - \mathbf{a}^*)'\mathbf{Z}(\mathbf{ZZ^{-1}})^{-1}\mathbf{c} = 0$ từ điều kiện $(\mathbf{a} - \mathbf{a}^*)'\mathbf{Z} = \mathbf{a'Z} - \mathbf{a}^{*\prime}\mathbf{Z} = \mathbf{c'} - \mathbf{c'} = \mathbf{0}$

-Mà $\mathbf{a}^*$ cố định, nên $(\mathbf{a} - \mathbf{a}^*)'(\mathbf{a} - \mathbf{a}^*)$ dương trừ khi $\mathbf{a} = \mathbf{a}^*$. Var$(\mathbf{a'Y})$ nhỏ nhất khi $\boldsymbol{a^{*\prime}Y} = \boldsymbol{c'(Z'Z)^{-1}Z'Y} = \boldsymbol{c'\hat{\beta}}$ ■

PHẦN 4: CÁC SUY LUẬN VỀ MÔ HÌNH HỒI QUY

4.1 Các suy luận liên quan đến các tham số hồi quy:

Kết quả 7.4: Cho $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}$, trong đó $\boldsymbol{Z}$ có hạng đầy đủ bằng r+1 và $\boldsymbol{\epsilon}$ tuân theo phân phối $N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$. Khi đó, ước lượng hợp lý cực đại của $\boldsymbol{\beta}$ cũng chính là ước lượng bình phương tối thiểu $\hat{\boldsymbol{\beta}}$.
Hơn nữa, ta có $\hat{\boldsymbol{\beta}} = \boldsymbol{(Z'Z)^{-1}Z'Y}$ tuân theo phân phối $N_{r+1}(\boldsymbol{\beta}, \sigma^2\boldsymbol{(Z'Z)}^{-1})$ và độc lập với các phần dư $\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y - Z\hat{\beta}}$.
Đồng thời, $n\hat{\sigma}^2 = \hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}$ tuân theo phân phối $\sigma^2\chi^2_{n-r-1}$, trong đó $\hat{\sigma}^2$ là ước lượng hợp lý cực đại của σ^2.

Chứng minh: *Tham khảo tại link http://www.prenhall.com/statistics*

- Miền tin cậy của $\boldsymbol{\beta}$ là một hình ellipsoid, có thể được biểu diễn bằng ước lượng $s^2\boldsymbol{(Z'Z)}^{-1}$ của ma trận hiệp phương sai, trong đó $s^2 = \dfrac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n - r - 1}$

Kết quả 7.5: Cho $\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}$, trong đó $\boldsymbol{Z}$ có hạng đầy đủ bằng r+1 và $\boldsymbol{\epsilon}$ tuân theo phân phối $N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$. Khi đó, miền tin cậy $100(1-\alpha)\%$ cho $\boldsymbol{\beta}$ là:

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

Trong đó: $F_{r+1,n-r-1}(\alpha)$ là phân vị trên thứ 100α của phân phối Fisher với bậc tự do là r+1 và n-r-1.
Ta cũng có: Khoảng tin cậy đồng thời với độ tin cậy $100(1-\alpha)\%$ cho các giá trị β_i là:

$$\hat{\beta}_i \pm \sqrt{\widehat{Var}(\hat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, \qquad i = 0, 1, ..., r$$

Trong đó, $\widehat{Var}(\hat{\beta}_i)$ là phần tử thuộc đường chéo chính của $s^2\boldsymbol{(Z'Z)}^{-1}$ tương ứng với $\hat{\beta}_i$.

Chứng minh:

- Đặt $\boldsymbol{V} = (\boldsymbol{Z'Z})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$

- Theo Kết quả 7.2, $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$

$$\begin{aligned}\Rightarrow E(\boldsymbol{V}) &= E[(\boldsymbol{Z'Z})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\ &= (\boldsymbol{Z'Z})^{\frac{1}{2}}E(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\boldsymbol{Z'Z})^{\frac{1}{2}}(\boldsymbol{\beta} - \boldsymbol{\beta}) \\ &= 0\end{aligned}$$

- Theo Kết quả 7.2, $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z'Z})^{-1}$

$$\begin{aligned}\Rightarrow Cov(\boldsymbol{V}) &= Cov[(\boldsymbol{Z'Z})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\ &= (\boldsymbol{Z'Z})^{\frac{1}{2}}Cov(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})[(\boldsymbol{Z'Z})^{\frac{1}{2}}]' \\ &= (\boldsymbol{Z'Z})^{\frac{1}{2}}Cov(\hat{\boldsymbol{\beta}})(\boldsymbol{Z'Z})^{\frac{1}{2}} \\ &= (\boldsymbol{Z'Z})^{\frac{1}{2}}\sigma^2(\boldsymbol{Z'Z})^{-1}(\boldsymbol{Z'Z})^{\frac{1}{2}} \\ &= \sigma^2(\boldsymbol{Z'Z})^{\frac{1}{2}}(\boldsymbol{Z'Z})^{-1}(\boldsymbol{Z'Z})^{\frac{1}{2}} \\ &= \sigma^2(\boldsymbol{Z'Z})^{\frac{1}{2}}(\boldsymbol{Z'Z})^{\frac{-1}{2}}(\boldsymbol{Z'Z})^{\frac{-1}{2}}(\boldsymbol{Z'Z})^{\frac{1}{2}} \\ &= \sigma^2\boldsymbol{I}\end{aligned}$$

- Theo Kết quả 7.4, ta có $\hat{\boldsymbol{\beta}} \sim N_{r+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z'Z})^{-1})$

$\Rightarrow \boldsymbol{V} = (\boldsymbol{Z'Z})^{\frac{1}{2}}\hat{\boldsymbol{\beta}} - (\boldsymbol{Z'Z})^{\frac{1}{2}}\boldsymbol{\beta}$ tuân theo phân phối chuẩn (theo Kết quả 4.3) (1)

- Đồng thời: $$\begin{aligned}\boldsymbol{V'V} &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'[(\boldsymbol{Z'Z})^{\frac{1}{2}}]'(\boldsymbol{Z'Z})^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{Z'Z})^{\frac{1}{2}}(Z'Z)^{\frac{1}{2}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &= (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\boldsymbol{Z'Z})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \qquad (2)\end{aligned}$$

(1),(2) $\Rightarrow \boldsymbol{V'V} \sim \sigma^2\chi^2(r+1)$

- Theo Kết quả 7.4, $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \sim \sigma^2\chi^2(n-r-1)$ và $\hat{\boldsymbol{\epsilon}}$ độc lập với $\hat{\boldsymbol{\beta}}$

$\Rightarrow$ Nếu đặt $s^2 = \dfrac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-r-1}$ thì:

$(n-r-1)s^2 \sim \sigma^2\chi^2(n-r-1)$ và $(n-r-1)s^2$ độc lập với $\hat{\boldsymbol{\beta}}$, do đó cũng độc lập với V.

Vì vậy,

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

Từ đó suy ra miền tin cậy $100(1-\alpha)\%$ cho $\boldsymbol{\beta}$ là:

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

- Đặt $\boldsymbol{A}^{-1} = \dfrac{\boldsymbol{Z'Z}}{s^2}$, $\quad c^2 = (r+1)F_{r+1,n-r-1}(\alpha)$ và $\boldsymbol{u_i}' = (0, ..., 0, 1, 0, ..., 0)$

Khi đó: $\boldsymbol{z} = \boldsymbol{\beta} - \hat{\boldsymbol{\beta}}$ là một vector nằm trong miền ellipsoid tin cậy của $\boldsymbol{\beta}$. (vì $\hat{\boldsymbol{\beta}}$ là tâm của hình ellipsoid này)

- Ta có:

$\boldsymbol{z'u_i} = \beta_i - \hat{\beta}_i$ là phần tử thứ i của vector $\boldsymbol{z}$.

$$\begin{aligned}\boldsymbol{u_i'Au_i} &= \boldsymbol{u_i}'\left(\frac{\boldsymbol{Z'Z}}{s^2}\right)^{-1}\boldsymbol{u_i} \\ &= \boldsymbol{u_i}'s^2(\boldsymbol{Z'Z})^{-1}\boldsymbol{u_i} \\ &= \widehat{Var}(\hat{\beta}_i)\end{aligned}$$ là phần tử thứ i trên đường chéo chính của $s^2(\boldsymbol{Z'Z})^{-1}$

- Hơn nữa, khoảng tin cậy đồng thời $100(1-\alpha)\%$ cho các β_i là hình chiếu của miền ellipsoid tin cậy $100(1-\alpha)\%$ cho $\boldsymbol{\beta}$ xuống các trục tọa độ.

- Theo Kết quả 5A.1 (trang 258), ta có:

$$|\boldsymbol{z'u_i}| \le c\sqrt{\boldsymbol{u_i'Au_i}}$$
$$\Leftrightarrow |\beta_i - \hat{\beta}_i| \le \sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}\sqrt{\widehat{Var}(\hat{\beta}_i)}$$
$$\Leftrightarrow \hat{\beta}_i - \sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}\sqrt{\widehat{Var}(\hat{\beta}_i)} \le \beta_i \le \hat{\beta}_i + \sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}\sqrt{\widehat{Var}(\hat{\beta}_i)}$$

Vậy: Khoảng tin cậy đồng thời $100(1-\alpha)\%$ cho các β_i là:

$$\hat{\beta}_i \pm \sqrt{\widehat{Var}(\hat{\beta}_i)}\sqrt{(r+1)F_{r+1,n-r-1}(\alpha)}, \qquad i = 0, 1, ..., r \qquad \blacksquare$$

- Trong ứng dụng, người ta thường không quan tâm đến tính chất khoảng tin cậy "đồng thời" của ước lượng khoảng trong Kết quả 7.5. Khi cần tìm các biến độc lập quan trọng, họ có thể thay thế $(r+1)F_{r+1,n-r-1}(\alpha)$ bằng giá trị $t_{n-r-1}\left(\frac{\alpha}{2}\right)$ và sử dụng khoảng:

$$\hat{\beta}_i \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{\widehat{Var}(\hat{\beta}_i)}, \qquad i = 0, 1, ..., r \tag{7-11}$$

Ví dụ 7.4: Các dữ liệu trong bảng 7.1 được thu thập từ 20 căn nhà ở khu Milwaukee, Wisconsin.

Đặt z_1 = Tổng diện tích căn nhà (đơn vị: $100ft^2$)
z_2 = Giá trị thẩm định (đơn vị: \$1000)
Y = Giá bán (đơn vị: \$1000)

Table 7.1 Real-Estate Data

z_1 Total dwelling size ($100\,ft^2$)	z_2 Assessed value (\$1000)	Y Selling price (\$1000)
15.31	57.3	74.8
15.20	63.8	74.0
16.25	65.4	72.9
14.33	57.0	70.0
14.57	63.8	74.9
17.33	63.2	76.0
14.48	60.2	72.0
14.91	57.7	73.5
15.25	56.4	74.5
13.89	55.6	73.5
15.18	62.6	71.5
14.44	63.4	71.0
14.87	60.2	78.9
18.63	67.2	86.5
15.20	57.1	68.0
25.76	89.6	102.0
19.05	68.6	84.0
15.37	60.1	69.0
18.06	66.3	88.0
16.35	65.8	76.0

Dựa vào bộ dữ liệu, hãy ước lượng hàm hồi quy tuyến tính $Y_j = \beta_0 + \beta_1 z_{j1} + \beta_2 z_{j2} + \epsilon_j$ bằng cách dùng phương pháp bình phương tối thiểu.

Giải: (Xem file code)

4.2 Kiểm định tỉ số hợp lí cực đại cho tham số hồi quy:

- Một giả thuyết không (H_0) cho rằng một số biến z_i không làm ảnh hưởng đến biến Y. Ta ký hiệu các biến độc lập z_i này là $z_{q+1}, z_{q+2}, ..., z_r$. Khi đó, câu khẳng định rằng $z_{q+1}, z_{q+2}, ..., z_r$ không ảnh hưởng đến Y có thể được viết lại thành giả thuyết thống kê

$$H_0 : \beta_{q+1} = \beta_{q+2} = ... = \beta_r = 0 \qquad \text{hoặc} \qquad H_0 : \boldsymbol{\beta}_{(2)} = 0 \tag{7-12}$$

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

- Đặt

$$\boldsymbol{Z} = \left[\begin{array}{c:c} \underset{n\times(q+1)}{\boldsymbol{Z}_1} & \underset{n\times(r-q)}{\boldsymbol{Z}_2} \end{array} \right], \qquad \boldsymbol{\beta} = \left[\begin{array}{c} \underset{(q+1)\times 1}{\boldsymbol{\beta}_{(1)}} \\ \hdashline \underset{(r-q)\times 1}{\boldsymbol{\beta}_{(2)}} \end{array} \right]$$

Ta có thể biểu diễn mô hình hồi quy tuyến tính tổng quát thành:

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon} = \left[\begin{array}{c:c} \boldsymbol{Z}_1 & \boldsymbol{Z}_2 \end{array} \right] . \left[\begin{array}{c} \boldsymbol{\beta}_{(1)} \\ \hdashline \boldsymbol{\beta}_{(2)} \end{array} \right] = \boldsymbol{Z}_{(1)}\boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_{(2)}\boldsymbol{\beta}_{(2)} + \boldsymbol{\epsilon}$$

- Đối với giả thuyết không $H_0 : \boldsymbol{\beta}_{(2)} = 0, \boldsymbol{Y} = \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{\epsilon}$, kiểm định tỉ số hợp lý cho H_0 được dựa vào

$$\begin{aligned} \text{Extra sum of squares (ESS)} =& SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z}) \\ =& \text{Tổng bình phương phần dư của } \boldsymbol{Z}_1 \text{ - Tổng bình phương phần dư của } \boldsymbol{Z} \\ =& (\boldsymbol{y} - \boldsymbol{Z}_1\hat{\boldsymbol{\beta}}_{(1)})'(\boldsymbol{y} - \boldsymbol{Z}_1\hat{\boldsymbol{\beta}}_{(1)}) - (\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}}) \end{aligned} \tag{7-13}$$

Trong đó: $\hat{\beta}_{(1)} = (\boldsymbol{Z}_1'\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1'\boldsymbol{y}$

- Ta định nghĩa tỉ số F là:

$$F = \frac{(SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z}))/(r-q)}{s^2} = \frac{n(\hat{\sigma_1}^2 - \hat{\sigma}^2)/(r-q)}{n\hat{\sigma}^2/(n-r-1)}$$

Tỉ số F tuân theo phân phối Fisher với bậc tự do là r-q và n-r-1.

Kết quả 7.6: Cho $\boldsymbol{Z}$ có hạng đầy đủ là r+1 và $\boldsymbol{\epsilon}$ tuân theo phân phối $N_n(\boldsymbol{0}, \sigma^2\boldsymbol{I})$. Kiểm định tỉ số hợp lý cho $H_0 : \boldsymbol{\beta}_{(2)} = 0$ tương đương với một kiểm định cho H_0 dựa vào extra sum of squares ở (7-13) và $s^2 = \dfrac{(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})}{n-r-1}$. Đặc biệt, kiểm định tỉ số hợp lý sẽ bác bỏ H_0 nếu:

$$F = \frac{SS_{res}(\boldsymbol{Z}_1) - SS_{res}(\boldsymbol{Z})}{s^2(r-q)} > F_{r-q,n-r-1}(\alpha)$$

Trong đó: $F_{r-q,n-r-1}(\alpha)$ là phân vị trên thứ 100α của phân phối Fisher với bậc tự do là r-q và n-r-1.

Chứng minh:

- Xét hàm hợp lý ứng với $\boldsymbol{\beta}$ và σ^2 là:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \frac{1}{(2\pi)^{\frac{n}{2}}\sigma^n} e^{\frac{-(\boldsymbol{y}-\boldsymbol{Z\beta})'(\boldsymbol{y}-\boldsymbol{Z\beta})}{2\sigma^2}} \\ &\leq \frac{1}{(2\pi)^{\frac{n}{2}}\hat{\sigma}^n} e^{\frac{-n}{2}} \end{aligned}$$

Dấu "=" xảy ra khi $\sigma^2 = \dfrac{(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})}{n}$ và $\boldsymbol{\beta} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'y}$

Do đó:

$\hat{\sigma}^2 = \dfrac{(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - \boldsymbol{Z}\hat{\boldsymbol{\beta}})}{n}$ là ước lượng hợp lý cực đại của σ^2

$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z'Z})^{-1}\boldsymbol{Z'y}$ là ước lượng hợp lý cực đại của $\boldsymbol{\beta}$

- Xét giả thuyết $H_0 : \boldsymbol{Y} = \boldsymbol{Z_1\beta_{(1)}} + \boldsymbol{\epsilon}$ và:

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

Trong đó: $\hat{\boldsymbol{\beta}}_{\boldsymbol{(1)}} = (\boldsymbol{Z_1'Z})^{-1}\boldsymbol{Z_1'y}$

Đồng thời: $\hat{\sigma}_1^2 = \dfrac{(\boldsymbol{y} - \boldsymbol{Z_1}\hat{\boldsymbol{\beta}}_{\boldsymbol{(1)}})'(\boldsymbol{y} - \boldsymbol{Z_1}\hat{\boldsymbol{\beta}}_{\boldsymbol{(1)}})}{n}$

- Việc bác bỏ $H_0 : \boldsymbol{\beta_{(2)}} = 0$ đối với các giá trị nhỏ của tỉ số hợp lý

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

cũng tương đương với việc bác bỏ H_0 đối với các giá trị lớn của $\dfrac{\hat{\sigma}_1^2 - \hat{\sigma}^2}{\hat{\sigma}^2}$ hay tỉ số

$$F = \frac{n(\hat{\sigma}_1^2 - \hat{\sigma}^2)/(r-q)}{n\hat{\sigma}^2/(n-r-1)} = \frac{(SS_{res}(\boldsymbol{Z_1}) - SS_{res}(\boldsymbol{Z}))/(r-q)}{s^2}$$

Vậy: Kiểm định tỉ số hợp lý sẽ bác bỏ H_0 nếu:

$$F = \frac{(SS_{res}(\boldsymbol{Z_1}) - SS_{res}(\boldsymbol{Z}))/(r-q)}{s^2} > F_{r-q,n-r-1}(\alpha) \qquad \blacksquare$$

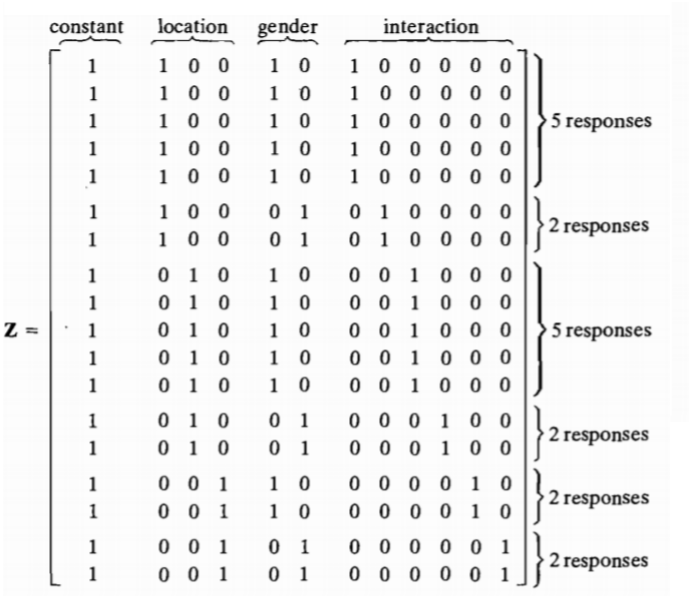
Ví dụ 7.5: (Kiểm tra mức quan trọng của các biến độc lập thêm vào bằng cách sử dụng extra sum-of-squares)
Các khách hàng nam và nữ tham gia đánh giá dịch vụ tại ba cơ sở của một chuỗi nhà hàng lớn. Bảng 7.2 chứa các dữ liệu của n = 18 khách hàng. Mỗi điểm dữ liệu trong bảng được phân loại dựa vào địa điểm (1,2 hoặc 3) và giới tính (nam = 0 và nữ = 1).

Table 7.2 Restaurant-Service Data

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

Ta thấy bảng dữ liệu trên có số lượng các quan sát trong các nhóm không bằng nhau. Ví dụ: nhóm thuộc địa điểm 1 và giới tính nam có 5 quan sát, trong khi nhóm thuộc địa điểm 2 và giới tính nữ chỉ có 2 quan sát.

Sử dụng 3 biến giả cho địa điểm (Địa điểm 1, Địa điểm 2, Địa điểm 3) và 2 biến giả cho giới tính (Nam, Nữ), ta có thể lập được một mô hình hồi quy tuyến tính cho thấy quan hệ giữa kết quả đánh giá Y với địa điểm, giới tính và mối liên hệ giữa địa điểm và giới tính thông qua ma trận Z như sau:



Hỏi rằng mối liên hệ giữa địa điểm và giới tính có tác động đến kết quả đánh giá Y hay không? Mức độ ảnh hưởng của từng địa điểm lên kết quả đánh giá Y có như nhau hay không? (nghĩa là yếu tố địa điểm có tác động lên kết quả đánh giá Y hay không)

Giải: (Xem file code)

-Tỉ số F nhỏ hơn $F_{2,12}(\alpha)$ với mọi giá trị α có ý nghĩa. Từ đó, ta kết luận rằng kết quả đánh giá Y không phụ thuộc vào mối liên hệ giữa địa điểm và giới tính. Vì vậy, các biến biểu diễn mối liên hệ này có thể được bỏ ra khỏi mô hình.

-Sử dụng Extra sum of squares, ta có thể xác định được rằng ba địa điểm có cùng mức độ ảnh hưởng lên kết quả đánh giá Y, nghĩa là các địa điểm khác nhau sẽ không dẫn đến kết quả đánh giá Y khác nhau (Lúc này ta nói yếu tố địa điểm không tác động lên kết quả đánh giá Y):

Xét $\boldsymbol{Z} = \begin{bmatrix} 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 & 1 \end{bmatrix}$

-Đặt $\boldsymbol{Z}_2$ = [Cột Địa điểm 1], $\boldsymbol{Z}_1$ = [Các cột còn lại của $\boldsymbol{Z}$]. Ta sẽ tính được Extra sum of squares ứng với Địa điểm 1.
-Đặt $\boldsymbol{Z}_2$ = [Cột Địa điểm 2], $\boldsymbol{Z}_1$ = [Các cột còn lại của $\boldsymbol{Z}$]. Ta sẽ tính được Extra sum of squares ứng với Địa điểm 2.
-Đặt $\boldsymbol{Z}_2$ = [Cột Địa điểm 3], $\boldsymbol{Z}_1$ = [Các cột còn lại của $\boldsymbol{Z}$]. Ta sẽ tính được Extra sum of squares ứng với Địa điểm 3.

-Vì giá trị Extra sum of squares ứng với ba địa điểm này là bằng nhau nên các tỉ số F tương ứng cũng bằng nhau. Nghĩa là, ba địa điểm này có cùng mức độ ảnh hưởng lên kết quả đánh giá Y.

-Tương tự, ta cũng có thể kiểm chứng được rằng hai giới tính nam và nữ có mức độ ảnh hưởng khác nhau lên kết quả đánh giá Y, nghĩa là nam và nữ sẽ cho kết quả đánh giá Y khác nhau (Lúc này ta nói yếu tố giới tính có tác động lên kết quả đánh giá Y).

PHẦN 5: CÁC SUY LUẬN TỪ HÀM HỒI QUY ƯỚC LƯỢNG

- Giả sử ta đã tìm được mô hình hồi quy phù hợp với bộ dữ liệu. Cho $\boldsymbol{z}_0' = (1, z_{01}, ..., z_{0r})$ là các giá trị của các biến độc lập. Khi đó, $\boldsymbol{z}_0$ và $\hat{\boldsymbol{\beta}}$ có thể được sử dụng để:
(a): Ước lượng hàm hồi quy $\beta_0 + \beta_1 z_{01} + ... + \beta_r z_{0r}$ tại $\boldsymbol{z}_0$.
(b): Ước lượng giá trị của biến phụ thuộc Y tại $\boldsymbol{z}_0$.

5.1 Ước lượng hàm hồi quy tại z_0:

- Cho Y_0 là giá trị của biến phụ thuộc khi các biến độc lập có giá trị là $\boldsymbol{z}_0' = (1, z_{01}, ..., z_{0r})$. Theo mô hình hồi quy tuyến tính cổ điển (7-3), giá trị kỳ vọng của Y_0 là:

$$E(Y_0|z_0) = \beta_0 + \beta_1 z_{01} + ... + \beta_r z_{0r} = \boldsymbol{z}_0'\boldsymbol{\beta} \tag{7-15}$$

Ước lượng bình phương tối thiểu của $E(Y_0|\boldsymbol{z}_0)$ là $\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}$.

Kết quả 7.7: Đối với mô hình hồi quy tuyến tính cổ điển (7-3), $\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}$ là ước lượng tuyến tính không chệch của $E(Y_0|\boldsymbol{z}_0)$ với phương sai nhỏ nhất, $Var(\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}) = \boldsymbol{z}_0'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z}_0\sigma^2$.
Nếu các sai số $\boldsymbol{\epsilon}$ tuân theo phân phối chuẩn thì khoảng tin cậy $100(1-\alpha)\%$ cho $E(Y_0|\boldsymbol{z}_0) = \boldsymbol{z}_0'\boldsymbol{\beta}$ là:

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

Trong đó: $t_{n-r-1}\left(\frac{\alpha}{2}\right)$ là phân vị trên thứ $100\left(\frac{\alpha}{2}\right)$ của phân phối Student với bậc tự do là n-r-1.

Chứng minh:

- Với $\boldsymbol{z}_0$ cho trước, $\boldsymbol{z}_0'\boldsymbol{\beta}$ là một tổ hợp tuyến tính của $\beta_0, ..., \beta_r$.

- Theo Kết quả 7.3, $\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}$ là một ước lượng tuyến tính không chệch có phương sai nhỏ nhất của $\boldsymbol{z}_0'\boldsymbol{\beta} = E(\boldsymbol{Y}_0|\boldsymbol{z}_0)$

- Theo Kết quả 7.2, $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{Z'Z})^{-1}$. Do đó:

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

$$E(\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}) = \boldsymbol{z}_0'E(\hat{\boldsymbol{\beta}}) = \boldsymbol{z}_0'\boldsymbol{\beta} \tag{2}$$

- Theo Kết quả 7.4, $\hat{\boldsymbol{\beta}} \sim N_{r+1}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{Z'Z})^{-1})$ và $\hat{\boldsymbol{\beta}}$ độc lập với $\hat{\boldsymbol{\epsilon}}$

$\Rightarrow \hat{\boldsymbol{\beta}}$ độc lập với $\dfrac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\sigma^2(n-r-1)} = \dfrac{s^2}{\sigma^2}$ với $s^2 = \dfrac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-r-1}$

- Đồng thời:

Vì $\dfrac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\sigma^2(n-r-1)} \sim \dfrac{\sigma^2\chi^2(n-r-1)}{\sigma^2(n-r-1)}$ (theo Kết quả 7.4)
nên $\dfrac{s^2}{\sigma^2} \sim \dfrac{\chi^2(n-r-1)}{n-r-1}$

- Vì $\hat{\boldsymbol{\beta}}$ tuân theo phân phối chuẩn nên $\boldsymbol{z}_0'\hat{\boldsymbol{\beta}}$ tuân theo phân phối chuẩn (3)

(1),(2),(3) $\Rightarrow \boldsymbol{z}_0'\hat{\boldsymbol{\beta}} \sim N_{r+1}(\boldsymbol{z}_0'\boldsymbol{\beta}, \sigma^2\boldsymbol{z}_0'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z}_0)$ (theo Kết quả 4.2)

- Ta có:

$$\frac{(\boldsymbol{z_0'}\hat{\boldsymbol{\beta}} - \boldsymbol{z_0'}\boldsymbol{\beta})/\sqrt{\sigma^2 \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0}}}{\sqrt{s^2/\sigma^2}} = \frac{\boldsymbol{z_0'}\hat{\boldsymbol{\beta}} - \boldsymbol{z_0'}\boldsymbol{\beta}}{\sqrt{s^2(\boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}} \sim St(n-r-1)$$

- Khoảng tin cậy $100(1-\alpha)\%$ cho $E(\boldsymbol{Y_0}|\boldsymbol{z_0}) = \boldsymbol{z_0'}\boldsymbol{\beta}$ là tập hợp các giá trị $\boldsymbol{z_0'}\boldsymbol{\beta}$ thỏa:

$$\left| \frac{\boldsymbol{z_0'}\hat{\boldsymbol{\beta}} - \boldsymbol{z_0'}\boldsymbol{\beta}}{\sqrt{s^2(\boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}} \right| \leq t_{n-r-1}\left(\frac{\alpha}{2}\right)$$

$$\Leftrightarrow \boldsymbol{z_0'}\hat{\boldsymbol{\beta}} - t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(\boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})} \leq \boldsymbol{z_0'}\boldsymbol{\beta} \leq \boldsymbol{z_0'}\hat{\boldsymbol{\beta}} + t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(\boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}$$

Vậy: Khoảng tin cậy $100(1-\alpha)\%$ cho $E(\boldsymbol{Y_0}|\boldsymbol{z_0}) = \boldsymbol{z_0'}\boldsymbol{\beta}$ là:

$$\boldsymbol{z_0'}\hat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(\boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})} \qquad \blacksquare$$

5.2 Dự báo một giá trị quan sát mới tại z_0

- Dựa vào mô hình hồi quy tuyến tính cổ điển ở (7-3), ta có:

$$\boldsymbol{Y}_0 = \boldsymbol{z_0'}\boldsymbol{\beta} + \boldsymbol{\epsilon}_0$$

hoặc

(Giá trị quan sát mới $\boldsymbol{Y}_0$) = (Giá trị kỳ vọng của $\boldsymbol{Y}_0$ tại $\boldsymbol{z}_0$) + (Sai số mới)

Trong đó: $\boldsymbol{\epsilon}_0$ tuân theo phân phối $N(0, \sigma^2)$ và độc lập với $\boldsymbol{\epsilon}$. Do đó, $\boldsymbol{\epsilon}_0$ cũng độc lập với $\hat{\boldsymbol{\beta}}$ và s^2.

Kết quả 7.8: Dựa vào mô hình hồi quy tuyến tính cổ điển ở (7-3), giá trị quan sát mới Y_0 có biến độc lập không chệch:

$$\boldsymbol{z_0'}\hat{\boldsymbol{\beta}} = \hat{\beta}_0 + \hat{\beta}_1 z_{01} + ... + \hat{\beta}_r z_{0r}$$

Phương sai của sai số dự báo $Y_0 - \boldsymbol{z_0'}\hat{\boldsymbol{\beta}}$ là:

$$Var(Y_0 - \boldsymbol{z_0'}\hat{\boldsymbol{\beta}}) = \sigma^2(1 + \boldsymbol{z_0'}(\boldsymbol{Z'Z})^{-1}\boldsymbol{z}_0)$$

Nếu các sai số $\boldsymbol{\epsilon}$ tuân theo phân phối chuẩn thì khoảng dự đoán $100(1-\alpha)\%$ cho Y_0 là:

$$\boldsymbol{z_0'}\hat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(1 + \boldsymbol{z_0'}(\boldsymbol{Z'Z})^{-1}\boldsymbol{z}_0)}$$

Trong đó: $t_{n-r-1}\left(\frac{\alpha}{2}\right)$ là phân vị trên thứ 100 $\left(\frac{\alpha}{2}\right)$ của phân phối Student với bậc tự do là n-r-1.

Chứng minh:

- Từ mô hình hồi quy, ta có:

$$\boldsymbol{Y_0} = \boldsymbol{z_0'}\boldsymbol{\beta} + \boldsymbol{\epsilon_0}$$

- Theo Kết quả 7.7, $\boldsymbol{z_0'}\hat{\boldsymbol{\beta}}$ là ước lượng tuyến tính không chệch của $E(\boldsymbol{Y_0}|\boldsymbol{z_0}) = \boldsymbol{z_0'}\boldsymbol{\beta}$, với $E(\boldsymbol{z_0'}\hat{\boldsymbol{\beta}}) = \boldsymbol{z_0'}\boldsymbol{\beta}$ và

$Var(\boldsymbol{z_0'}\hat{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0}$

- Sai số dự báo là:

$\hat{\boldsymbol{\epsilon}}_0 = \boldsymbol{Y_0} - \boldsymbol{z_0}'\hat{\boldsymbol{\beta}} = \boldsymbol{z_0}'\boldsymbol{\beta} + \boldsymbol{\epsilon_0} - \boldsymbol{z_0}'\hat{\boldsymbol{\beta}} = \boldsymbol{\epsilon_0}'(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$

$\Rightarrow E(\hat{\boldsymbol{\epsilon}}_0) = E(\boldsymbol{\epsilon}_0) + \boldsymbol{z_0}'E(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})$

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

$= 0 \qquad (1)$

$\Rightarrow$ Biến độc lập $\boldsymbol{z_0'}\hat{\boldsymbol{\beta}}$ là không chệch.

- Vì $\boldsymbol{\epsilon_0}$ độc lập với $\hat{\boldsymbol{\beta}}$ nên:

$$\begin{aligned} Var(\hat{\boldsymbol{\epsilon}}_0) &= Var(\boldsymbol{Y_0} - \boldsymbol{z_0}'\hat{\boldsymbol{\beta}}) \\ &= Var(\boldsymbol{z_0'}\boldsymbol{\beta} + \boldsymbol{\epsilon_0} - \boldsymbol{z_0'}\hat{\boldsymbol{\beta}}) \\ &= Var(\boldsymbol{\epsilon_0} - \boldsymbol{z_0}'\hat{\boldsymbol{\beta}}) \\ &= Var(\boldsymbol{\epsilon_0}) + Var(\boldsymbol{z_0}'\hat{\boldsymbol{\beta}}) \\ &= \sigma^2 + \sigma^2 \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0} \\ &= \sigma^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0}) \qquad (2) \end{aligned}$$

- Nếu ta giả sử $\boldsymbol{\epsilon}$ tuân theo phân phối chuẩn thì:

$\hat{\boldsymbol{\beta}}$ tuân theo phân phối chuẩn. (theo Kết quả 7.4)

$\Rightarrow \hat{\boldsymbol{\epsilon}}_0 = \boldsymbol{Y_0} - \boldsymbol{z_0}'\hat{\boldsymbol{\beta}} = -\boldsymbol{z_0}'\hat{\boldsymbol{\beta}} + (\boldsymbol{z_0}'\boldsymbol{\beta} + \boldsymbol{\epsilon_0})$ tuân theo phân phối chuẩn. (theo Kết quả 4.2) (3)

(1),(2),(3) $\Rightarrow \hat{\boldsymbol{\epsilon}}_0 \sim N(0, \sigma^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0}))$

$$\Rightarrow \frac{\hat{\boldsymbol{\epsilon}}_0}{\sqrt{\sigma^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}} \sim N(0,1)$$

- Ta lại có:

$\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}} \sim \sigma^2\chi^2(n-r-1)$ (theo Kết quả 7.4)

$$\Rightarrow \frac{s^2}{\sigma^2} = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\sigma^2(n-r-1)} \sim \frac{\sigma^2\chi^2(n-r-1)}{\sigma^2(n-r-1)} \text{ với } s^2 = \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{n-r-1}$$

$$\Rightarrow \sqrt{\frac{s^2}{\sigma^2}} \sim \sqrt{\frac{\chi^2(n-r-1)}{n-r-1}}$$

$$\Rightarrow \frac{\dfrac{\hat{\boldsymbol{\epsilon}}_0}{\sqrt{\sigma^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}}}{\sqrt{\dfrac{s^2}{\sigma^2}}} \sim St(n-r-1)$$

$$\Rightarrow \frac{\boldsymbol{Y_0} - \boldsymbol{z_0}'\hat{\boldsymbol{\beta}}}{\sqrt{s^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}} \sim St(n-r-1)$$

- Khoảng dự đoán $100(1-\alpha)\%$ cho $\boldsymbol{Y_0}$ là tập hợp các giá trị $\boldsymbol{Y_0}$ thỏa:

$$\left| \frac{\boldsymbol{Y_0} - \boldsymbol{z_0}'\hat{\boldsymbol{\beta}}}{\sqrt{s^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}} \right| \leq t_{n-r-1}\left(\frac{\alpha}{2}\right)$$

$$\Leftrightarrow \boldsymbol{z_0}'\hat{\boldsymbol{\beta}} - t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})} \leq \boldsymbol{Y_0} \leq \boldsymbol{z_0}'\hat{\boldsymbol{\beta}} + t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})}$$

Vậy: Khoảng dự đoán $100(1-\alpha)\%$ cho $\boldsymbol{Y_0}$ là:

$$\boldsymbol{z_0}'\hat{\boldsymbol{\beta}} \pm t_{n-r-1}\left(\frac{\alpha}{2}\right)\sqrt{s^2(1 + \boldsymbol{z_0}'(\boldsymbol{Z'Z})^{-1}\boldsymbol{z_0})} \qquad \blacksquare$$

- Ta thấy rằng khoảng dự đoán cho Y_0 thì rộng hơn khoảng tin cậy để ước lượng giá trị của hàm hồi quy $E(Y_0|\boldsymbol{z}_0) = \boldsymbol{z}_0'\boldsymbol{\beta}$.

- Vì $Y_0 = \boldsymbol{z}_0'\boldsymbol{\beta} + \epsilon_0$ nên việc dự báo một giá trị quan sát mới Y_0 tại $\boldsymbol{z}_0$ không chắc chắn bằng việc ước lượng giá trị $E(Y_0|\boldsymbol{z}_0) = \boldsymbol{z}_0'\boldsymbol{\beta}$.

Ví dụ 7.6: Những công ty nào muốn xem xét việc mua máy tính thì trước tiên phải đánh giá các nhu cầu trong tương lai để xác định loại thiết bị phù hợp. Một chuyên gia máy tính đã thu thập dữ liệu từ bảy công ty tương tự nhau để xây dựng một hàm dự báo cho các yêu cầu về phần cứng máy tính để quản lý kho hàng. Các dữ liệu được trình bày trong bảng 7.3.

Đặt z_1 = Số lượng đơn đặt hàng (đơn vị: nghìn)
z_2 = Số lượng thiết bị thêm - bớt (đơn vị: nghìn)

Y = Thời gian CPU xử lý (đơn vị: giờ)

Table 7.3 Computer Data

z_1 (Orders)	z_2 (Add–delete items)	Y (CPU time)
123.5	2.108	141.5
146.1	9.213	168.9
133.9	1.905	154.8
128.5	.815	146.5
151.5	1.061	172.8
136.2	8.603	160.1
92.0	1.125	108.5

Source: Data taken from H. P. Artis, *Forecasting Computer Requirements: A Forecaster's Dilemma* (Piscataway, NJ: Bell Laboratories, 1979).

Hãy xây dựng khoảng tin cậy 95% cho thời gian trung bình mà CPU xử lý, $E(Y_0|\boldsymbol{z}_0) = \beta_0 + \beta_1 z_{01} + \beta_2 z_{02}$ tại $\boldsymbol{z}_0^{'} = (1.130, 7.5)$.
Đồng thời, hãy tìm khoảng dự đoán 95% cho thời gian CPU xử lý ở một cơ sở mới tương ứng với $\boldsymbol{z}_0$.

Giải: (Xem file code)

PHẦN 6: KIỂM TRA MÔ HÌNH VÀ CÁC KHÍA CẠNH KHÁC CỦA PHƯƠNG PHÁP HỒI QUY

6.1 Mô hình có hiệu quả không?

- Dĩ nhiên, chúng ta bắt buộc phải kiểm tra mức độ phù hợp của mô hình trước khi sử dụng hàm hồi quy ước lượng để suy luận.

- Tất cả thông tin về mức độ thiếu phù hợp của mô hình được biểu diễn bằng các phần dư:

$$\begin{aligned}\hat{\epsilon}_1 &= y_1 - \hat{\beta}_0 - \hat{\beta}_1 z_{11} - ... - \hat{\beta}_r z_{1r}\\ \hat{\epsilon}_2 &= y_2 - \hat{\beta}_0 - \hat{\beta}_1 z_{21} - ... - \hat{\beta}_r z_{2r}\\ &\cdots\\ \hat{\epsilon}_n &= y_n - \hat{\beta}_0 - \hat{\beta}_1 z_{n1} - ... - \hat{\beta}_r z_{nr}\end{aligned}$$

hoặc

$$\hat{\boldsymbol{\epsilon}} = [\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\boldsymbol{y} = [\boldsymbol{I} - \boldsymbol{H}]\boldsymbol{y} \qquad (7\text{-}16)$$

- Nếu mô hình được xem là đúng thì mỗi phần dư $\hat{\epsilon}_j$ sẽ là ước lượng của sai số ϵ_j, trong đó ϵ_j được giả sử tuân theo phân phối chuẩn $N(0, \sigma^2)$. Mặc dù $E(\hat{\boldsymbol{\epsilon}}) = \boldsymbol{0}$ nhưng ma trận hiệp phương sai $\sigma^2[\boldsymbol{I} - \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}']\boldsymbol{y} = \sigma^2[\boldsymbol{I} - \boldsymbol{H}]$ không là ma trận chéo. Các phần dư có phương sai không cân bằng và hệ số tương quan khác 0. Tuy nhiên, thật may mắn khi các hệ số tương quan thường nhỏ và các phương sai cũng gần bằng nhau.

- Vì các phần dư $\hat{\boldsymbol{\epsilon}}$ có ma trận hiệp phương sai $\sigma^2[\boldsymbol{I} - \boldsymbol{H}]]$ nên các phương sai của ϵ_j có thể khác nhau rất nhiều nếu các phần tử trên đường chéo của $\boldsymbol{H}$, các giá trị đòn bẩy h_{jj}, có sự khác nhau đáng kể. Vì vậy, nhiều nhà thống kê thích phân tích hình ảnh dựa vào các phần dư đã được student hóa. Sử dụng bình phương trung bình phần dư s^2 làm ước lượng của σ^2, ta có:

$$\widehat{Var}(\hat{\epsilon}_j) = s^2(1 - h_{jj}), \qquad j = 1, 2, ..., n \qquad (7\text{-}17)$$

và các phần dư được student hóa là:

$$\hat{\epsilon}_j^* = \frac{\hat{\epsilon_j}}{\sqrt{s^2(1 - h_{jj})}}, \qquad j = 1, 2, ..., n \qquad (7\text{-}18)$$

- Ta kỳ vọng rằng các phần dư đã được student hóa sẽ có đồ thị nhìn gần giống với các hình vẽ độc lập của phân phối chuẩn $N(0, 1)$. Một số gói phần mềm sử dụng các phương sai ước lượng "delete-one" $s^2(j)$ để có thể student hóa các $\hat{\epsilon_j}$. Trong đó, $s^2(j)$ chính là bình phương trung bình phần dư khi xóa giá trị quan sát thứ j khỏi bộ dữ liệu.

- Ta nên vẽ các phần dư theo nhiều cách khác nhau để có thể dễ dàng nhận thấy được những giá trị bất thường. Sau đây là một số cách vẽ đồ thị để phục vụ cho mục đích phân tích tổng quát:

1 - Vẽ đồ thị của các phần dư $\hat{\epsilon}_j$ theo các giá trị dự báo $\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 z_{j1} + ... + \hat{\beta}_r z_{jr}$:
Sự sai lệch so với các giả thiết của mô hình thường được biểu thị thông qua hai dạng hiện tượng:

(a) : *Các phần dư bị phụ thuộc vào những giá trị dự báo.* Xét Hình 7.2(a). Ta thấy các phép tính số học đều bị sai, hoặc là một tham số β_0 đang bị thiếu trong mô hình.

(b) : *Phương sai không là hằng số.* Đồ thị của phần dư có thể có dạng hình phễu như trong Hình 7.2(b), khi đó sẽ có sự biến thiên lớn giữa các giá trị lớn của $\hat{y}$ và sự biến thiên nhỏ giữa các giá trị nhỏ của $\hat{y}$. Nếu trường hợp này xảy ra, phương sai của sai số sẽ không là hằng, và ta sẽ cần dùng đến các phép biến đổi hoặc phương pháp bình phương tối thiểu có trọng số (hoặc cả hai) để ước lượng cho mô hình hồi quy.
Trong Hình 7.2(d), các phần dư tạo thành một dải băng nằm ngang. Đây là một điều lý tưởng vì nó cho thấy các phương sai là bằng nhau và không phụ thuộc vào $\hat{\boldsymbol{y}}$.

2 - Vẽ đồ thị của các phần dư $\hat{\epsilon}_j$ theo một biến độc lập (ví dụ như z_1), hoặc theo tích của một số biến độc lập (ví dụ như z_1^2 hoặc $z_1 z_2$):
Nếu đồ thị này có hình dạng mang tính hệ thống thì có nghĩa rằng mô hình hồi quy còn đang thiếu một vài tham số. Trường hợp này được biểu diễn ở Hình 7.2(c).

3 - Vẽ đồ thị Q-Q và histogram:
Để xác định xem các sai số có tuân theo phân phối chuẩn hay không, ta sẽ đi kiểm tra các phần dư $\hat{\epsilon}_j$ hoặc $\hat{\epsilon}_j^*$ bằng các kỹ thuật như trong mục 4.6 ở chương 4. Các đồ thị Q-Q, histogram và dot diagram có thể giúp chúng ta nhận diện được các giá trị bất thường hoặc những sự sai lệch lớn mà ta cần chú ý đến khi phân tích. Nếu n lớn, những sự sai lệch nhỏ sẽ không ảnh hưởng nhiều đến các suy luận về tham số β.

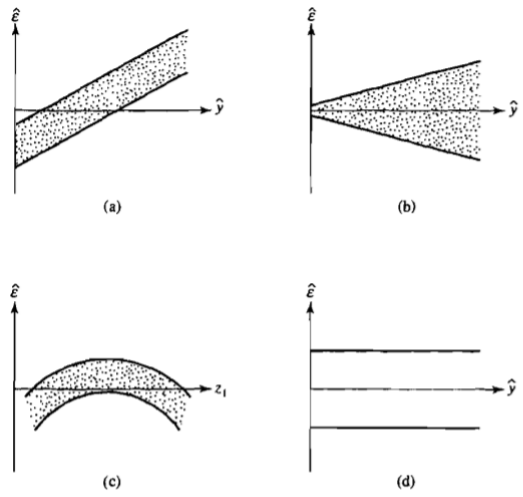


Figure 7.2 Residual plots.

4 - Vẽ đồ thị của các phần dư $\hat{\epsilon}_j$ theo thời gian :
Giả thiết về sự độc lập là một giả thiết quan trọng nhưng khó có thể kiểm tra được. Nếu các dữ liệu tuân theo trình tự thời gian một cách tự nhiên thì một đồ thị của các phần dư theo thời gian sẽ có hình dạng mang tính hệ thống.

Đặt
$$r_1 = \frac{\sum_{j=2}^{n} \hat{\epsilon}_j \hat{\epsilon}_{j-1}}{\sum_{j=1}^{n} \hat{\epsilon}_j^2} \tag{7-19}$$

Ta có r_1 là hệ số tự tương quan bậc nhất của các phần dư tương ứng với các thời điểm kế tiếp nhau. Một kiểm định về tính độc lập dựa vào thống kê

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

được gọi là Kiểm định Durbin - Watson.

Ví dụ 7.7: Xét các dữ liệu máy tính trong Ví dụ 7.6. Ba đồ thị phần dư được cho trong Hình 7.3. Cỡ mẫu n = 7 là quá nhỏ để có thể đánh giá chính xác, nhưng theo đồ thị thì có vẻ như các giả thiết hồi quy là hợp lý.

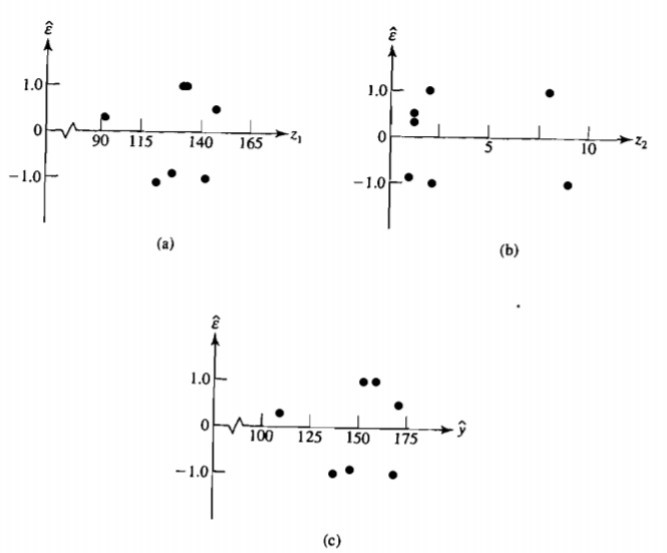


Figure 7.3 Residual plots for the computer data of Example 7.6.

- Nếu một bộ giá trị của các biến độc lập tương ứng với nhiều giá trị quan sát của biến phụ thuộc thì ta có thể tiến hành một kiểm định về sự thiếu phù hợp.

6.2 Giá trị đòn bẩy và mức độ ảnh hưởng:

- Mặc dù việc phân tích các phần dư là hữu ích trong việc đánh giá mức độ phù hợp của mô hình, nhưng những sự sai lệch trong mô hình hồi quy thường bị che khuất bởi quá trình đánh giá mức độ phù hợp. Ví dụ, những giá trị ngoại lai trong biến phụ thuộc hoặc biến độc lập có thể gây ra ảnh hưởng rất lớn đối với quá trình phân tích, tuy nhiên các giá trị này lại khó có thể được phát hiện từ các đồ thị phần dư.

- Ta định nghĩa giá trị đòn bẩy h_{jj} là phần tử thứ j trên đường chéo chính của $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}$. Giá trị đòn bẩy h_{jj} liên quan đến khoảng cách từ giá trị quan sát thứ j đến n - 1 giá trị quan sát còn lại. Trong hồi quy tuyến tính với một biến độc lập z thì:

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

Giá trị đòn bẩy trung bình là $\dfrac{r+1}{n}$.
Đồng thời, giá trị đòn bẩy h_{jj} còn là mức độ ảnh hưởng của một trường hợp lên sự phù hợp của mô hình.

- Vector các giá trị dự báo là:

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

Trong đó, hàng thứ j thể hiện giá trị $\hat{y}_j = h_{jj}y_j + \sum_{k\neq j} h_{jk}y_k$

- Giả sử ta cố định các giá trị khác y_j của y. Khi đó:

$$(\text{Sự thay đổi của giá trị } \hat{y}_j) = h_{jj}\ (\text{Sự thay đổi của giá trị } y_j)$$

Nếu giá trị đòn bẩy h_{jj} lớn so với các giá trị h_{jk} thì y_j sẽ là yếu tố chính ảnh hưởng đến giá trị dự báo $\hat{y}_j$.

- Các giá trị quan sát nào có ảnh hưởng lớn đến kết quả suy luận từ dữ liệu thì được gọi là các quan sát có ảnh hưởng (influential observations). Những phương pháp đánh giá mức độ ảnh hưởng của các quan sát thường được dựa vào sự thay đổi của vector tham số ước lượng $\hat{\boldsymbol{\beta}}$ khi xóa đi các quan sát đó.

- Nếu sau quá trình kiểm tra, ta thấy các giả thiết hồi quy đều được thỏa (một cách tương đối) thì ta có thể yên tâm đưa ra các suy luận về $\boldsymbol{\beta}$ và các giá trị trong tương lai của biến $\boldsymbol{Y}$.

6.3 Một số vấn đề khác trong hồi quy:

1 - Lựa chọn các biến độc lập từ một tập hợp lớn:
- Trên thực tế, rất khó để có thể lập được một hàm hồi quy phù hợp ngay lập tức. Khi số lượng các biến độc lập là

quá lớn thì ta không thể đưa hết chúng vào mô hình. Hiện tại đã có nhiều phương pháp để có thể tìm được tập hợp các biến độc lập "tốt nhất". Một trong những phương pháp đó là sử dụng ***thống kê Mallow's*** C_p:

$$C_p = \left(\frac{\begin{array}{c}\text{Tổng bình phương phần dư cho mô hình con} \\ \text{chứa p tham số, trong đó bao gồm hệ số chặn}\end{array}}{\text{Phương sai của phần dư cho mô hình đầy đủ}} \right) - (n - 2p)$$

Trong đó: Hệ số chặn là tham số β_0 trong mô hình hồi quy tuyến tính cổ điển.

- Xét đồ thị của C_p theo p, trong đó mỗi điểm (p, C_p) tương ứng với một tập hợp các biến độc lập. Các mô hình tốt sẽ có điểm (p, C_p) nằm gần với đường thẳng 45^o (nghĩa là C_p gần bằng p).

- Hình 7.4 là đồ thị C_p cho các dữ liệu máy tính trong Ví dụ 7.6 với ba biến độc lập z_1, z_2, z_3:

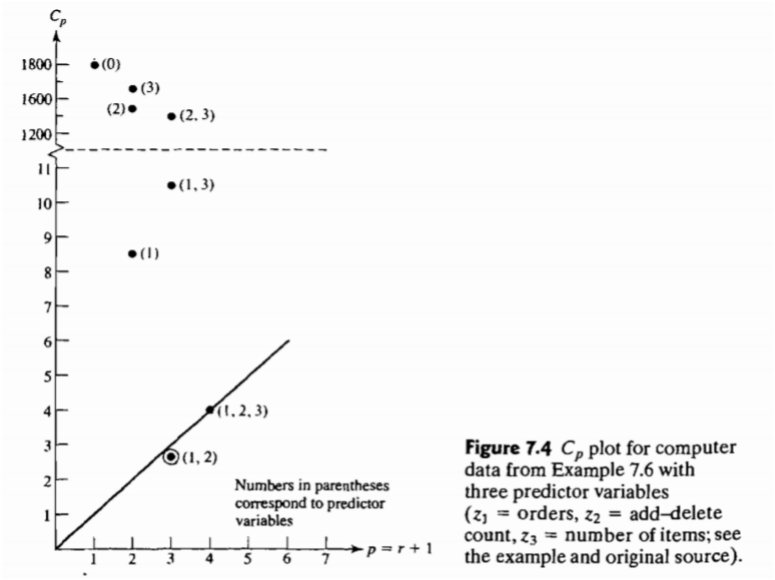


Figure 7.4 C_p plot for computer data from Example 7.6 with three predictor variables (z_1 = orders, z_2 = add–delete count, z_3 = number of items; see the example and original source).

Theo Hình 7.4, ta thấy điểm $(3, C_3)$ tương ứng với tập hợp z_1, z_2 nằm gần đường thẳng 45^o nhất, nên tập z_1, z_2 là tập hợp các biến độc lập "tốt nhất" để đưa vào mô hình hồi quy cho Ví dụ 7.6.

- Nếu có quá nhiều biến độc lập thì ràng buộc về chi phí sẽ làm hạn chế số lượng mô hình có thể xét đến được. Một cách tiếp cận khác để chọn lựa các biến độc lập cho mô hình là ***Phương pháp hồi quy từng bước***. Phương pháp này sẽ giúp chọn ra các biến độc lập quan trọng mà không cần phải xét đến từng trường hợp.

- Các bước thuật toán của phương pháp hồi quy từng bước có thể được diễn giải như sau:

Bước 1: Biến độc lập đầu tiên được đưa vào hàm hồi quy là biến có ảnh hưởng lớn nhất đến sự thay đổi của biến phụ thuộc Y (nghĩa là hệ số tương quan giữa biến đó với Y là lớn nhất).

Bước 2: Trong số các biến độc lập chưa được chọn, biến tiếp theo được đưa vào hàm hồi quy là biến có ảnh hưởng lớn nhất đến tổng bình phương hồi quy. Độ ảnh hưởng đó được xác định bằng kiểm định F ở Kết quả 7.6. Nếu giá trị của thống kê F lớn hơn giá trị *F to enter* thì biến độc lập được xem là quan trọng và sẽ được đưa vào mô hình. Trong đó, *F to enter* là một giá trị do người dùng nhập vào.

Bước 3: Sau khi một biến mới được thêm vào hàm hồi quy, độ ảnh hưởng đến tổng bình phương hồi quy của các biến còn lại trong hàm hồi quy sẽ được kiểm tra bằng kiểm định F. Nếu giá trị của thống kê F nhỏ hơn giá trị *F to remove* tương ứng với mức ý nghĩa cho trước thì biến đó sẽ bị đưa ra khỏi mô hình. Trong đó, *F to remove* là một giá trị do người dùng nhập vào.

Bước 4: Ta lặp lại Bước 2 và Bước 3 đến khi việc thêm các biến vào hàm hồi quy không còn có ảnh hưởng lớn và việc xóa đi các biến trong hàm hồi quy có ảnh hưởng lớn. Khi đó, thuật toán sẽ kết thúc.

- Tuy nhiên, ta không thể chắc chắn rằng phương pháp này sẽ chọn ra được các biến độc lập tốt nhất cho việc dự báo. Đồng thời, phương pháp này cũng không có khả năng nhận diện được khi nào thì nên sử dụng những phép biến đổi đối với các biến độc lập.

- Một phương pháp khác để tìm ra mô hình phù hợp là sử dụng ***Tiêu chí thông tin Akaike (AIC)***:

$$AIC = n\ ln\left(\frac{\text{Tổng bình phương phần dư cho mô hình con chứa p tham số, trong đó bao gồm hệ số chặn}}{n}\right) + 2p$$

- Ta cần chọn mô hình nào có giá trị AIC nhỏ.

2 - Hiện tượng cộng tuyến:

- *Cộng tuyến là hiện tượng các biến độc lập trong mô hình có mối quan hệ tuyến tính với nhau.*

- Nếu $\boldsymbol{Z}$ không có hạng đầy đủ thì một số tổ hợp tuyến tính, ví dụ như $\boldsymbol{Za}$, phải bằng $\boldsymbol{0}$. Trong trường hợp này, ta nói các cột của $\boldsymbol{Z}$ cộng tuyến với nhau. Khi đó, $\boldsymbol{Z'Z}$ sẽ không khả nghịch. Trên thực tế, trong phân tích hồi quy, hiếm khi nào $\boldsymbol{Za} = \boldsymbol{0}$ hoàn toàn. Tuy nhiên, nếu các tổ hợp tuyến tính của các cột của $\boldsymbol{Z}$ gần bằng $\boldsymbol{0}$ thì kết quả của phép tính $(\boldsymbol{Z'Z})^{-1}$ sẽ không ổn định.

- Thông thường, các phần tử trên đường chéo chính của $(\boldsymbol{Z'Z})^{-1}$ là khá lớn. Điều này dẫn đến phương sai ước lượng của các $\hat{\beta}_i$ cũng sẽ lớn và lúc đó ta khó có thể nhận biết được các hệ số hồi quy $\hat{\beta}_i$ "quan trọng".

- Để khắc phục hậu quả của hiện tượng cộng tuyến, ta có các cách sau:

(a) : Trong cặp biến độc lập mà có hệ số tương quan lớn, ta bỏ bớt đi một biến.

(b) : Liên hệ biến phụ thuộc Y với các thành phần chính của các biến độc lập. Nghĩa là, các hàng $z_j^{'}$ của $\boldsymbol{Z}$ được xem như một mẫu, và một số thành phần chính đầu tiên của các biến độc lập được tính theo cách ở Mục 8.3. Khi đó, biến Y sẽ được đưa vào phân tích hồi quy với các biến độc lập mới này.

3 - Hiện tượng chệch xảy ra do sử dụng sai mô hình:

- Giả sử mô hình hồi quy đúng phải có $\boldsymbol{Z} = \left[\begin{array}{c:c} \boldsymbol{Z}_1 & \boldsymbol{Z}_2 \end{array}\right]$ với hạng r + 1 và:

$$\underset{(n\times 1)}{\boldsymbol{Y}} = \left[\begin{array}{c:c} \underset{(n\times(q+1))}{\boldsymbol{Z}_1} & \underset{(n\times(r-q))}{\boldsymbol{Z}_2} \end{array}\right] \left[\begin{array}{c} \underset{((q+1)\times 1)}{\boldsymbol{\beta}_{(1)}} \\ \hdashline \underset{((r-q)\times 1)}{\boldsymbol{\beta}_{(2)}} \end{array}\right] + \underset{(n\times 1)}{\boldsymbol{\epsilon}} \tag{7-20}$$

$$= \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_2\boldsymbol{\beta}_{(2)} + \boldsymbol{\epsilon}$$

Trong đó: $E(\boldsymbol{\epsilon}) = \boldsymbol{0}$ và $Var(\boldsymbol{\epsilon}) = \sigma^2\boldsymbol{I}$.

- Tuy nhiên, nhà phân tích đã sơ suất khi ước lượng mô hình hồi quy mà chỉ sử dụng q biến độc lập đầu tiên để tìm giá trị nhỏ nhất của tổng bình phương phần dư $(\boldsymbol{Y} - \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)})'(\boldsymbol{Y} - \boldsymbol{Z}_1\boldsymbol{\beta}_{(1)})$. Ta đã biết ước lượng bình phương tối thiểu của $\boldsymbol{\beta}_{(1)}$ là $\hat{\boldsymbol{\beta}}_{(1)} = (\boldsymbol{Z}_1^{'}\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^{'}\boldsymbol{Y}$. Khi đó:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{(1)}) &= (\boldsymbol{Z}_1^{'}\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^{'}E(\boldsymbol{Y}) \\ &= (\boldsymbol{Z}_1^{'}\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^{'}(\boldsymbol{Z}_1\boldsymbol{\beta}_{(1)} + \boldsymbol{Z}_2\boldsymbol{\beta}_{(2)} + E(\boldsymbol{\epsilon})) \\ &= \boldsymbol{\beta}_{(1)} + (\boldsymbol{Z}_1^{'}\boldsymbol{Z}_1)^{-1}\boldsymbol{Z}_1^{'}\boldsymbol{Z}_2\boldsymbol{\beta}_{(2)} \end{aligned} \tag{7-21}$$

Ta thấy: $\hat{\boldsymbol{\beta}}_{(1)}$ là một ước lượng chệch của $\boldsymbol{\beta}_{(1)}$ trừ phi các cột của $\boldsymbol{Z}_1$ vuông góc với các cột của $\boldsymbol{Z}_2$ (vì lúc đó $\boldsymbol{Z}_1^{'}\boldsymbol{Z}_2 = \boldsymbol{0}$). Nếu các biến độc lập quan trọng đang bị thiếu trong mô hình thì các ước lượng bình phương tối thiểu $\hat{\boldsymbol{\beta}}_{(1)}$ có thể sai.