[wg Jobson, 227]

Niech $y = X\beta + u,\ \ Eu = 0, Vu = \sigma^2 I,\ \ X \in M(n, p+1)$. Uwaga! W macierzy planu X pierwsza kolumna składa się z samych jedynek, a $p+1$ jest liczbą współczynników regresji dla p zmiennych. Warunek najmniejszych kwadratów $\|y - X\beta\|^2$daje estymator b parametru β postaci

$$b = \left(X^T X\right)^{-1} X^T y$$

Wtedy

$$\widehat{y} = Xb = X\left(X^T X\right)^{-1} X^T y$$

Oznaczmy $H = X\left(X^T X\right)^{-1} X^T$, a przez $e = y - \widehat{y}$ wektor reszt z regresji próbkowej. Wtedy

$$\begin{aligned} y &= X\beta + u, \\ y &= Xb + e \\ \widehat{y} &= Hy, e = (I-H)\, y \end{aligned}$$

Własność 1

$$\begin{aligned} HX &= X, \quad H1_n = 1_n, \quad (I-H)\, X = 0 \\ H &= H^T,\ H^2 = H,\ (I-H)^2 = I - H \end{aligned}$$

Dowód

$$\begin{aligned} H^T &= \left(X\left(X^T X\right)^{-1} X^T\right)^T = X\left(X^T X\right)^{-1} X^T = H, \\ H^2 &= X\left(X^T X\right)^{-1}\left(X^T X\right)\left(X^T X\right)^{-1} X^T = X\left(X^T X\right)^{-1} X^T = H, \\ (I-H)^2 &= I - H - H + H^2 = I - H \end{aligned}$$

H jest więc operatorem rzutu prostopadłego na płaszczyznę regresji próbkowej, $I - H$ operatorem rzutu prostopadłego na płaszczyznę błędu próbkowego.

Własność 2

Reszty u od regresji populacji są niezależne ($Vu = \sigma^2 I$), natomiast reszty e od regresji próbkowej są zależne, $Ve = \sigma^2 (I-H)$.

Dowód

$$\begin{aligned} Ee &= E(I-H)y = (I-H)Ey = (I-H)X\beta \\ (I-H)X &= 0 \Rightarrow Ee = 0 \end{aligned}$$

$$\begin{aligned} Ve &= Eee^T = E(I-H)yy^T(I-H) = (I-H)Eyy^T(I-H), \\ Eyy^T &= E(X\beta+u)(X\beta+u)^T = X\beta\beta^TX^T + Euu^T = X\beta\beta^TX^T + \sigma^2 I, \\ (I-H)Eyy^T &= ((I-H)X)\beta\beta^TX^T + \sigma^2(I-H) = \sigma^2(I-H), \\ Ve &= \sigma^2(I-H)(I-H) = \sigma^2(I-H) \end{aligned}$$

Będzie dowód, że $\frac{1}{n} \leqslant h_{ii} \leqslant 1$. Gdy $h_{ii} = 1$ to $Ve_i = 0$ a więc z prawdopodobieństwem 1 reszta $e_i = 0$.

Definicja

Resztą standaryzowaną nazywamy wielkość

$$r_i = \frac{e_i}{s\sqrt{1-h_{ii}}},$$

gdzie s jest estymatorem σ.

Z faktu, że $HX = X$ wynika, że $H1_n = 1_n$.

Wtedy

$$\overline{y} = \frac{1}{n}1_n^Ty = \frac{1}{n}1_n^THy = \frac{1}{n}1_n^T\widehat{y} = \overline{\widehat{y}}$$

i trójkąt o bokach y, $\widehat{y}, \overline{y}1_n$ jest trójkątem prostokątnym.

Stąd dla każdego y zachodzą nierówności

$$0 \leqslant \|\widehat{y} - \overline{y}1_n\|^2 \leqslant \|y - \overline{y}1_n\|^2$$

Mamy

$$\begin{aligned} \|\widehat{y} - \overline{y}1_n\|^2 &= \|\widehat{y}\|^2 - 2\overline{y}1_n^T\widehat{y} + \|\overline{y}1_n\|^2 \\ &= \|Hy\|^2 - 2n\overline{y}\overline{\widehat{y}} + n(\overline{y})^2 \\ &= y^THy - n(\overline{y})^2 \end{aligned}$$

$$\begin{aligned} \|y - \overline{y}1_n\|^2 &= \|y\|^2 - 2\overline{y}1_n^Ty + \|\overline{y}1_n\|^2 \\ &= \|y\|^2 - n(\overline{y})^2 \end{aligned}$$

A więc, dla każdego y zachodzą nierówności

$$\|y\|^2 \geqslant y^T H y \geqslant n\left(\overline{y}\right)^2$$

Kładąc $y^T = [0, 0, ..., 0, 1, 0, ..., 0]$ z jedynką na i-tym miejscu otrzymamy

$$\begin{aligned}
\frac{1}{n} &\leqslant h_{ii} \leqslant 1, \\
\sum_i h_{ii} &= Tr\left(H\right) = Tr\left(X\left(X^T X\right)^{-1} X^T\right) \\
&= Tr\left(\left(X^T X\right)^{-1} X^T X\right) = Tr\left(I\right) \\
&= p+1
\end{aligned}$$

Współczynnik h_{ii} opisuje wpływ i-tej oberwacji na regresję, gdyż

$$\widehat{y_i} = \sum_j h_{ij} y_j = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

$$\begin{aligned}
\sum_{j=1}^{n} h_{ij}^2 &= \left(HH^T\right)_{ii} = \left(H^2\right)_{ii} = h_{ii} \\
\sum_{j \neq i} h_{ij}^2 &= h_{ii}\left(1 - h_{ii}\right), \\
\sum_{j \neq i} h_{ij}^2 &\leqslant \frac{1}{4}
\end{aligned}$$

Dla $h_{ii} \geqslant \frac{1}{2}$ duże wartości h_{ii} oznaczają małe wartości h_{ij}
Przyjjmiemy, że wartość h_{ii} jest duża, gdy

$$h_{ii} > 2\frac{1}{n} \sum_i h_{ii} = \frac{2\left(p+1\right)}{n}$$

Ocenimy teraz jakie modyfikacje w oszacowaniu równania regresji spowoduje usunięcie jednej, i-tej obserwacji.

Oznaczmy przez $X_{(i)}$ macierz X po usunięciu i-tego wiersza ($X_{(i)} \in M\left(n-1, p+1\right)$), zaś przez $y_{(i)}$ $n-1$ - wymiarowy wektor y po usunięciu i-tego wiersza.

Niech $b_{(i)}$ oznacza $p+1$- wymiarowy wektor estymatorów współczynników regresji, zaś $\widehat{y_{(i)}}$ oznacza n - wymiarowy wektor wartości regresji, gdy z danych usunięto i-tą obserwację.

Wtedy

$$\begin{aligned} \widehat{y_{(i)}} &= Xb_{(i)}, \\ b_{(i)} &= \left(X_{(i)}^T X_{(i)}\right)^{-1} X_{(i)}^T y_{(i)} \end{aligned}$$

Zauważmy, że $X_{(i)}^T y_{(i)} = X^T y - X_i^T y_i$, gdzie X_i jest i-tym wierszem macierzy X.

$$b_{(i)} = \left(X_{(i)}^T X_{(i)}\right)^{-1} X_{(i)}^T y_{(i)} = \left(X_{(i)}^T X_{(i)}\right)^{-1} \left(X^T y - X_i^T y_i\right)$$

Skorzystamy ze wzoru [Madansky, 140]

$$\left(X_{(i)}^T X_{(i)}\right)^{-1} = \left(X^T X\right)^{-1} + \frac{\left(X^T X\right)^{-1} X_i^T X_i \left(X^T X\right)^{-1}}{1 - h_{ii}}$$

$$\begin{aligned} \left(X_{(i)}^T X_{(i)}\right)^{-1} X^T y &= \left(X^T X\right)^{-1} X^T y + \frac{\left(X^T X\right)^{-1} X_i^T X_i \left(X^T X\right)^{-1} X^T y}{1 - h_{ii}} = \\ &= b + \frac{\left(X^T X\right)^{-1} X_i^T \widehat{y_i}}{1 - h_{ii}} \end{aligned}$$

gdyż $\widehat{y_i} = X_i \left(X^T X\right)^{-1} X^T y$

$$\begin{aligned} \left(X_{(i)}^T X_{(i)}\right)^{-1} X_i^T y_i &= \left(X^T X\right)^{-1} X_i^T y_i + \frac{\left(X^T X\right)^{-1} X_i^T X_i \left(X^T X\right)^{-1} X_i^T y_i}{1 - h_{ii}} = \\ &= \left(X^T X\right)^{-1} X_i^T y_i + \frac{\left(X^T X\right)^{-1} X_i^T h_{ii} y_i}{1 - h_{ii}} = \\ &= \left(X^T X\right)^{-1} X_i^T y_i \left(1 + \frac{h_{ii}}{1 - h_{ii}}\right) = \frac{\left(X^T X\right)^{-1} X_i^T y_i}{1 - h_{ii}} \end{aligned}$$

gdyż $h_{ii} = X_i \left(X^T X\right)^{-1} X_i^T$
Stąd

$$\begin{aligned} b_{(i)} &= b + \frac{\left(X^T X\right)^{-1} X_i^T \widehat{y_i}}{1 - h_{ii}} - \frac{\left(X^T X\right)^{-1} X_i^T y_i}{1 - h_{ii}} = \\ &= b - \frac{e_i \left(X^T X\right)^{-1} X_i^T}{1 - h_{ii}} \end{aligned}$$

$$\widehat{y_{(i)}} = Xb_{(i)} = X\left(b - \frac{e_i\left(X^TX\right)^{-1}X_i^T}{1-h_{ii}}\right) =$$
$$= \widehat{y} - \frac{e_iX\left(X^TX\right)^{-1}X_i^T}{1-h_{ii}} = \widehat{y} - \frac{sr_iX\left(X^TX\right)^{-1}X_i^T}{\sqrt{1-h_{ii}}}$$

Odległością Cooka nazywamy liczbę

$$D_i = \frac{\left\|\widehat{y} - \widehat{y_{(i)}}\right\|^2}{s^2\left(p+1\right)}$$

Odległość Cooka ocenia skutki usunięcia i-tej obserwacji. Jest standaryzowaną średnią odległością pomiędzy wartościami regresji przed i po usunięciu i-tej obserwacji.

$$D_i = \frac{\left\|\widehat{y} - \widehat{y_{(i)}}\right\|^2}{s^2\left(p+1\right)} = \frac{\left\|\frac{sr_iX\left(X^TX\right)^{-1}X_i^T}{\sqrt{1-h_{ii}}}\right\|^2}{s^2\left(p+1\right)} = \frac{s^2r_i^2\left\|X\left(X^TX\right)^{-1}X_i^T\right\|^2}{s^2\left(p+1\right)\left(1-h_{ii}\right)},$$
$$\left\|X\left(X^TX\right)^{-1}X_i^T\right\|^2 = X_i\left(X^TX\right)^{-1}X^TX\left(X^TX\right)^{-1}X_i^T = X_i\left(X^TX\right)^{-1}X_i^T = h_{ii},$$
$$D_i = \frac{s^2r_i^2h_{ii}}{s^2\left(p+1\right)\left(1-h_{ii}\right)} = r_i^2\frac{h_{ii}}{\left(p+1\right)\left(1-h_{ii}\right)}$$

Pierwszy składnik iloczynu odpowiada za odchylenia od regresji, zaś drugi - za wielkość wpływu