Maciej Kostrzewski

Regresja wielokrotna

Przygotowano w oparciu o „Applied Linear Regression Models" Neter, Wasserman, Kutner

Model regresji:

p-1 zmiennych niezależnych$X_1,...,X_{p-1}$

$Y_i=\beta_0+\beta_1\mathbf{X_{i,1}}+...+\beta_{p-1}X_{i,p-1}+\varepsilon_i$

Jeśli założymy, że $X_{i0}=1$ wówczas postać równoważna, jest postaci:

$$Y_i=\sum_{k=0}^{p-1}\beta_k X_{ik}+\varepsilon_i$$

Zakładając $E(\varepsilon_i)=0$ wówczas

$E(Y)=\beta_0+\beta_1X_1+...+\beta_{p-1}X_{p-1}$ (tzw. Funkcja reakcji/odpowiedzi response function)

Jest to hiperpłaszczyzna.

Interpretacja parametru β_k: wskazuje on na zmiany Y ze względu na jednostkowy wzrost niezależnej zmiennej X_k, przy ustalonym poziomie pozostałych zmiennych. Warto zauważyć, że na model każda ze zmiennych ma jednakowy wpływ (demokracja zmiennych niezależnych).

Ogólny model regresji liniowej

W ogólnej sytuacji zmienne $\mathbf{X_{i,1},...,}X_{i,p-1}$ nie muszą reprezentować różnych niezależnych zmiennych[1]. Zdefiniujmy Ogólny model regresji liniowej o błędach normalnych w postaci:

$Y_i=\beta_0+\beta_1\mathbf{X_{i,1}}+...+\beta_{p-1}X_{i,p-1}+\varepsilon_i$, gdzie

$\beta_0,\beta_1,...,\beta_{p-1}$ są parametrami

$\mathbf{X_{i,1},...,}X_{i,p-1}$ znane wielkości

ε_i - $iidN(0,\sigma^2)$

Wniosek z postaci Y:

$Y\text{-}N(\beta_0+\beta_1\mathbf{X_{i,1}}+...+\beta_{p-1}X_{i,p-1};\sigma^2)$

Przykład:

$Y_i=\beta_0+\beta_1\mathbf{X_1}+\beta_2X^2{}_i+\varepsilon_i$

Przykład:

Teorię regresji można wykorzystać w przypadku

$Y_i=1/(\beta_0+\beta_1\mathbf{X_{i,1}}+...+\beta_{p-1}X_{i,p-1}+\varepsilon_i)$,

definiując nową zmienną zależną postaci

$Zi=1/Yi=\beta_0+\beta_1\mathbf{X_{i,1}}+...+\beta_{p-1}X_{i,p-1}+\varepsilon_i$

Zapis macierzowy

Równanie regresji można zapisać w postaci macierzowej:

$Y=X\beta+\varepsilon$ (wymiary: nx1=nxp px1+nx1)

[1] Dopuszczamy sytuację, że zmienna jest np.: postaci X_1*X_2

$$Y=\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \qquad X=\begin{bmatrix} 1 & X_{11} & \cdots & X_{1,p-1} \\ 1 & X_{21} & \cdots & X_{2,p-1} \\ \vdots & \vdots & & \vdots \\ 1 & X_{n1} & \cdots & Xn,p-1 \end{bmatrix} \qquad \beta=\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \quad \varepsilon=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Y-wektor reakcji (zmienna zależna)
β- wektor parametrów
X –macierz stałych (zmienne niezależne)
ε-wektor niezależnych zmiennych o rozkładzie normalnym o E(ε)=0 i macierzy kowariancji $\sigma^2\{\varepsilon\}=\sigma^2 I$.

Łatwo widać, ze E(Y)=Xβ, natomiast macierz kowariancji dla Y wynosi $\sigma^2\{Y\}=\sigma^2 I$ (nxn).

Estymatory najmniejszych kwadratów

Oznaczmy wektor estymowanych parametrów:

$$b=\begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}.$$

Równanie normalne ma postać X'Xb=X'Y
Estymatory są postaci; $b=(X'X)^{-1}X'Y$ (px1 = pxp px1)

Wartości teoretyczne i reszty

Oznaczmy wektor wartości teoretycznych (dopasowanych, wynikających z modelu) $\hat{Y}_i$ przez $\hat{Y}$, a wektor reszt $e_i=Y_i-\hat{Y}_i$ przez e

$$\hat{Y}=\begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \\ \hat{Y}_n \end{bmatrix} \qquad e=\begin{bmatrix} e_1 \\ e_2 \\ \\ e_n \end{bmatrix}.$$

Wielkości teoretyczne definiuje się jako $\hat{Y}=Xb$, a reszty jako $e=Y-\hat{Y}=Y-Xb$.
Ponadto przyjmijmy oznaczenie:

$$\hat{Y}=HY, \quad H=X\left(X'X\right)^{-1}X'.$$

Stad e=(I-H)Y, wówczas macierz kowariancji ma postać $\sigma^2\{e\}=\sigma^2(I-H)$, która jest estymowana przez $s^2\{e\}=MSE(I-H)$. W szczególności $\sigma^2\{e_i\}=\sigma^2(1-h_{ii})$, gdzie h_{ii} jest elementem z diagonali macierzy H. $h_{ii}=X_i'\left(X'X\right)^{-1}X_i'$, gdzie $X_i=\begin{bmatrix} 1 \\ X_{i,1} \\ \vdots \\ X_{i,p-1} \end{bmatrix}$ (i-ty wiersz macierzy X)

(przypomnienie: $s^2=\dfrac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2}{n-1}$, MSE=SSE/(n-p) –szczegóły poniżej)

Analiza wariancji

Przypomnienie:

$SSE = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$ (error sum of squares),

$SSR = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2$ (regression sum of squares, Z definicji e_i wynika, że $\overline{\hat{Y}} = \overline{Y}$),

SSTO=SSR+ SSE= $\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2$. (total sum of squares)

Uzasadnienie dekompozycji:

$Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + Y_i - \hat{Y}_i$ (całkowite odchylenie (SSTO)=odchylenie dopasowania modelu regresji wokół średniej (SSR)+odchylenie wokół dopasowanej lini regresji (SSE))

Z powyższego związku wynika:

$$\sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n} \left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2$$

Im większy jest SSTO tym większa jest wariancja pomiędzy obserwacjami Y.
Im większe jest SSE tym większe jest odchylenie obserwacji od dopasowanej prostej regresji.
Im większe SSR w stosunku do SSTO tym większe jest znaczenie regresji w opisie (wyjaśnianiu) całkowitej wariancji obserwacji Y.

SSTO=Y'Y-(1/n)Y'JY=Y'[I-(1/n)J]Y
SSE=e'e=(Y-Xb)'(Y-Xb)=Y'Y-b'X'Y=Y'(I-H)Y
SSR=b'X'Y-(1/n)Y'JY=Y'[H-(1/n)J]Y,gdzie

$$J = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix}$$ (nxn)

SSTO n-1 st. swobody
SSE n-p (ponieważ estymujemy p parametrów)
SSR p-1 (odpowiada liczbie zmiennych niezależnych)

Źródło wariancji	SS	Df	MS
Zmienność wewnątrz modelu	SSR=b'X'Y-(1/n)Y'JY	p-1	MSR=SSR/(p-1)
Zmienność reszt	SSE=Y'Y-b'X'Y	n-p	MSE=SSE/(n-p)
Zmienność całkowita	SSTO=Y'Y-(1/n)Y'JY	n-1	

E(MSE)= σ^2 , E(MSR)= σ^2 +nieujemna wielkość,
np.:dla p-1=2:

E(MSR)= $\sigma^2 + \left[\beta_1^2 \sum \left(X_{i1} - \overline{X}_1\right)^2 + \beta_2^2 \sum \left(X_{i2} - \overline{X}_2\right)^2 + 2\beta_1\beta_2 \sum \left(X_{i1} - \overline{X}_1\right)\left(X_{i2} - \overline{X}_2\right)\right] / 2$

Test F dla modelu regresji

Testujemy czy zachodzi związek pomiędzy zmienną zależną i układem zmiennych niezależnych zadanym modelem regresji tzn.

H0: $\beta_1=\beta_2=...=\beta_{p-1}=0$

H1:nie wszystkie β_k są zerami
Statystyka testowa:
F*=MSR/MSE
Poziom istotności α (błąd Igo rodzaju)
Jeśli F*<=F(1-α ;p-1,n-p) przyjmuję H0
Jeśli F*>F(1-α ;p-1,n-p) przyjmuję H1

Współczynnik dopasowania
R^2=SSR/SSTO=1-SSE/SSTO (*100%) określa w ilu procentach zmienność Y jest wyjaśniana przez zmienność modelu. Mierzy proporcjonalną redukcję zmienności Y wynikającą z zastosowania modelu regresji.

Własność:

- $0 \le R^2 \le 1$
- Jeśli wszystkie $\beta_k=0$, to $R^2=0$
- Jeśli $R^2=1$, to musi zachodzić $Y_i = \hat{Y}_i$
- Duże wielkości R^2 nie oznaczają, że dopasowany model jest użyteczny np: równocześnie mogą występować duże wielkości MSE, które uniemożliwiają właściwe wnioskowanie.
- Dodanie zmiennej niezależnej do modelu może zwiększyć wielkość tego współczynnika (nigdy zmniejszyć). Dzieje się tak ponieważ SSE nigdy się nie zwiększy przy większej liczbie zmiennych niezależnych, a SSTO jest zawsze takie samo. Model ekonometryczny powinien być oszczędny. Potrzebna jest miara która będzie preferować modele prostsze. Poprawiony współczynnik determinacji : $R^2{}_p = 1 - \dfrac{\dfrac{SSE}{n-p}}{\dfrac{SSTO}{n-1}} = 1 - \left(\dfrac{n-1}{n-p}\right)\dfrac{SSE}{SSTO}$

Współczynnik korelacji jest określany jako $R = \sqrt{R^2}$

Wnioskowanie o parametrach
Estymatory najmniejszych kwadratów są nieobciążone E(b)=β
Macierz kowariancji:

$$\sigma^2\{b\} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0,b_1\} & \sigma\{b_0,b_{p-1}\} \\ \sigma\{b_1,b_0\} & \sigma^2\{b_1\} & \sigma\{b_1,b_{p-1}\} \\ \sigma\{b_{p-1},b_0\} & \sigma\{b_{p-1},b_1\} & \sigma^2\{b_{p-1}\} \end{bmatrix}$$, wynosi $\sigma^2\{b\}= \sigma^2(X'X)^{-1}$,

Estymowana macierz kowariancji:

$$s^2\{b\} = \begin{bmatrix} s^2\{b_0\} & s\{b_0,b_1\} & s\{b_0,b_{p-1}\} \\ s\{b_1,b_0\} & s^2\{b_1\} & s\{b_1,b_{p-1}\} \\ s\{b_{p-1},b_0\} & s\{b_{p-1},b_1\} & s^2\{b_{p-1}\} \end{bmatrix}$$, wynosi $s^2\{b\}= MSE(X'X)^{-1}$,

Przedziałowa estymacja parametrów
Dla regresji o błędach normalnych:
$\dfrac{b_k - \beta_k}{s\{b_k\}} \approx t(n-p), \quad k = 0,1,\ldots, p-1$, zatem przedziały ufności dla β_k z przedziałem ufności 1-α są postaci: $b_k \pm t(1-\alpha/2;n-p)s\{b_k\}$

Testowanie współczynników

H0: $\beta_k=0$
H1: $\beta_k\neq 0$
Statystyka $t^*=b_k/s\{b_k\}$
Reguła decyzyjna:
Jeśli $|t^*|\leq t(1-\alpha/2;n-p)$ przyjmujemy H0, w innym przypadku przyjmujemy H1

Prognoza
Przedział prognozy dla jednego kroku z prawdopodobieństwem 1-α dla nowych obserwacji $Y_{h(nowy)}$ odpowiadającej wielkości zmiennej zależnej X_h wynosi:

$$\hat{Y}_h \pm t(1-\alpha/2;n-p)s\{Y_{h(nowy)}\}$$

gdzie $s^2\{Y_{h(nowy)}\}=MSE+s^2\{\hat{Y}_h\}=MSE+X'_h s^2\{b\}X_h=MSE(1+X'_h(X'X)^{-1}X_h)$

Dobór zmiennych niezależnych do modelu

Wzrost zmienność modelu regresji po dołączenia zmiennej X_2 do modelu zawierającego już zmienną X_1:

$$SSR(X_2|X_1)=SSE(X_1)-SSE(X_1,X_2)= SSR(X_1,X_2)- SSR(X_1)$$

(ostatnia równość wynika z SSTO=SSR+SSE)

Analogiczne związki gdy dołączymy X_3 do modelu w którym rozważano już X_1, X_2.
$SSR(X_3|X_1,X_2)=SSE(X_1,X_2)-SSE(X_1,X_2,X3)= SSR(X_1,X_2,X3)- SSR(X_1,X_2)$

Podobnie zbadamy wpływ dołączenia kilku zmiennych np.:
$SSR(X_2,X_3|X_1)=SSE(X_1)-SSE(X_1,X_2,X3)= SSR(X_1,X_2,X_3)- SSR(X_1)$

Dekompozycja SSR
$SSR(X_1,X_2)= SSR(X_1) + SSR(X_2|X_1)$,
a ponieważ kolejność indeksów w lewej stronie równości nie ma znaczenia, zatem zachodzi:
$SSR(X_1,X_2)= SSR(X_2) + SSR(X_1|X_2)$

W analogiczny sposób można zapisać związek:
$SSR(X_1,X_2, X_3)= SSR(X_1) + SSR(X_2|X_1)+ SSR(X_3|X_1,X_2)= SSR(X_1) + SSR(X_2, X_3|X_1)$

Źródło zmienności	SS	df	MS
Regresja	$SSR(X_1,X_2,X_3)$	3	$MSR(X_1,X_2,X_3)$
X_1	$SSR(X_1)$	1	$MSR(X_1)$
$X_2\|X_1$	$SSR(X_2\| X_1)$	1	$MSR(X_2\| X_1)$
$X_3\|X_1,X_2$	$SSR(X_3\|X_1,X_2)$	1	$MSR(X_3\|X_1,X_2)$
Błąd	$SSE(X_1,X_2, X_3)$	n-4	$MSE(X_1,X_2, X_3)$
Całkowita	SSTO	n-1	

Uzasadnienie:
Liczba st. swobody (degree of freedom, df) dla $SSR(X_1,X_2,X_3)$ musi być 3, gdyż $SSR(X_1,X_2,X_3)= SSR(X_1)+ SSR(X_2| X_1)+ SSR(X_3|X_1,X_2)$, a każda ze składowych ma df=1 bo 1 dodatkowy parametr do wyestymowania.

Testowanie istotności współczynników regresji

Rozważmy $Y_i=\beta_0+\beta_1 X_{i,1}+\beta_2 X_{i,2}+\beta_3 X_{i,3}+\varepsilon_i$
(niniejszy test jest równoważny z testem powyżej, bo $(t^*)^2=F^*$)
H0: $\beta_3=0$

H1: $\beta_3 \neq 0$

Oznaczenie
F – oznacza model pełny, w tej sytuacji ozn. że rozważamy X_1, X_2, X_3
R – model zredukowany tzn X_1, X_2
Wówczas:
$SSE(F)=SSE(X_1,X_2,X_3)$, $df_F=n-4$ (bo 4 parametry)
Jeśli H0, to mamy $Y_i=\beta_0+\beta_1 X_{i,1}+\beta_2 X_{i,2}+\varepsilon_i$ (oczywiście tutaj błędy są inne, to nie jest to samo co powyżej).
Wówczas
$SSE(R)=SSE(X_1,X_2)$, $df_R=n-3$ (bo 3 parametry)

Statystyka $F^*=\frac{SSE(R)-SSE(F)}{df_R-df_F} \Big/ \frac{SSE(F)}{df_F}$ (przypadek ogólny).

Reguła decyzyjna:
Jeśli $F^* \leq F(1-\alpha; df_R - df_F, df_F)$ przyjmujemy H0, w innym przypadku przyjmujemy H1
Warto zauważyć, że $df_R - df_F=1$, $df_F=n-p$ (gdzie p to liczba parametrów regresji)

Współczynniki determinacji cząstkowej
Idea
R^2 mierzy proporcjonalną redukcję zmienności Y uzyskaną poprzez rozważanie całego zbioru zmiennych niezależnych.
Współczynniki korelacji cząstkowej mierzą brzegowy wpływ jednej ze zmiennych na wyjaśnianie zmienności zmiennej Y, gdy pozostałe zmienne niezależne są już włączone do modelu.

Rozważmy:
$Y_i=\beta_0+\beta_1 X_{i,1}+\beta_2 X_{i,2}+\varepsilon_i$

$SSE(X_2)$ mierzy rozproszenie Y, gdy do modelu jest włączona zmienna X_2
$SSE(X_1,X_2)$ mierzy rozproszenie Y, gdy do modelu włączone są zmienne X_1 i X_2
Redukcja zmienności Y związana ze zmienną X_1 , gdy X_2 jest już w modelu wyraża się poprzez współczynnik korelacji cząstkowej:

$$r^2_{Y1.2}=\frac{SSE(X_2)-SSE(X_1,X_2)}{SSE(X_2)}=\frac{SSR(X_1 \mid X_2)}{SSE(X_2)}$$

Ogólniejszy przypadek:

$$r^2_{Y1.23}=\frac{SSR(X_1 \mid X_2,X_3)}{SSE(X_2,X_3)}$$

$$r^2_{Y4.123}=\frac{SSR(X_4 \mid X_1,X_2,X_3)}{SSE(X_1,X_2,X_3)}$$

Współczynniki korelacji cząstkowej

Współczynnik korelacji cząstkowej np:
$r_{Yk.2}=sgn(\beta_k)(r^2_{Yk.2})^{0.5}$
Współczynniki te stosowane są do identyfikacji zmiennych niezależnych, na podstawie tych informacji włącza się lub nie zmienne do modelu.

Przykład

$$r^2_{Y2.1} = \frac{(r_{Y2} - r_{12}r_{Y1})^2}{(1-r^2_{12})(1-r^2_{Y1})}$$

$$r^2_{Y2.13} = \frac{(r_{Y2.3} - r_{12.3}r_{Y1.3})^2}{(1-r^2_{12.3})(1-r^2_{Y1.3})}$$

, gdzie r_{12} to korelacja pomiędzy X_1 i X_2

Standaryzowany model regresji wielokrotnej

- Umożliwia porównanie wyestymowanych parametrów w tych samych jednostkach

Przykład:

$\hat{Y} = 200 + 20000X_1 + 0.2X_2$

na pierwszy rzut oka X_2 ma istotny wpływ na Y , a X_2 znikomy. Wniosek może być błędny bo nie podano jednostek, przypuśćmy zatem, że
Y w dolarach, X_1 w tysiącach dolarów, X_2 w centach
Wpływ wzrostu o 1000 dolarów zmiennej X_1 na Y jest taki sam jak wzrost zmiennej X_2 o 1000 dolarów

- Umożliwia kontrolę błędów zaokrągleń w procesie estymacji parametrów[2].
- Zagadnienie mało istotne dla modeli o 3ch zmiennych niezależnych lub mniej – tutaj błędy zaokrągleń nie są tak istotne zwłaszcza, jeśli maszyna licząca używa podwójnej precyzji tj 16 cyfr po przecinku)

Uniknięcie błędów zaokrągleń odbywa się poprzez zastosowanie tzw. transformacji korelacyjnej

$$Y_i^{'} = \frac{1}{\sqrt{n-1}}\left(\frac{Y_i - \overline{Y}}{s_Y}\right)$$

$$X_{ik}^{'} = \frac{1}{\sqrt{n-1}}\left(\frac{X_{ik} - \overline{X}_k}{s_k}\right) \qquad k=1,...,p-1,$$

gdzie

$$s_Y = \sqrt{\frac{\sum_i (Y_i - \overline{Y})^2}{n-1}} \qquad s_k = \sqrt{\frac{\sum_i (X_{ik} - \overline{X}_k)^2}{n-1}}$$

Model regresji dla nowych zmiennych $Y^{'}$ i $X^{'}_k$ nazywany jest modelem standardowej regresji i wyraża się poprzez
$Y^{'}_i = \beta^{'1}\mathbf{X}^{'}_{i,1} + ... + \beta^{'}_{p-1}X^{'}_{i,p-1} + \varepsilon^{'}_i$ (nie ma wyrazu wolnego! Jego uwzględnienie i tak okaże się bezcelowe bo estymatory najmniejszych kwadratów dają mu wartość równą 0)
Istnieje związek pomiędzy nowymi i starymi parametrami:

$$\beta_k = (s_Y/s_k)\,\beta^{'}_k \qquad (k=1,...,p-1)$$

$$\beta_0 = \overline{Y} - \beta_1\overline{X}_1 - ... - \beta_{p-1}\overline{X}_{p-1}$$

Uwaga:
Dla standaryzowanego modelu regresji macierz $X'X=r_{XX}$ tzn.jest macierzą korelacji. W konsekwencji jej elementy należą do przedziału [-1,1], co powoduje, że nie występują tak liczne problemy przy odwracaniu macierzy jak w standardowy modelu. Ponadto $X'Y=r_{YX}$.

[2] Problem w obliczeniach wynika przede wszystkim przy obliczaniu $(X'X)^{-1}$

Uzasadnienie:

$$\sum X_{i1}^{'} X_{i2}^{'} = \sum \left(\frac{X_{i1} - \bar{X}_1}{s_1 \sqrt{n-1}} \right) \left(\frac{X_{i2} - \bar{X}_2}{s_2 \sqrt{n-1}} \right) = \frac{1}{n-1} \frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{s_1 s_2} =$$

$$\frac{\sum (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)}{\left[\sum (X_{i1} - \bar{X}_1)^2 \sum (X_{i2} - \bar{X}_2)^2 \right]^{1/2}} = r_{12}$$

Przykład:
Dla p=2:

$$b_1^{'} = \frac{r_{Y1} - r_{12} r_{Y2}}{1 - r_{12}^2}, \quad b_2^{'} = \frac{r_{Y2} - r_{12} r_{Y1}}{1 - r_{12}^2}.$$

Interpretacja
Z porównania wielkości współczynników wyciągany jest wniosek o tym która zmienna ma większy wpływ na zmienną zależną. Z taką interpretacją należy być jednak ostrożny zwłaszcza w przypadku zmienych niezależnych między którymi występuje zależność.

Uwaga:
Interpretacja współczynników regresji tj. jak zmienia się zmienna zależna jeśli pewną zmienną niezależną zmienimy o jednostkę a pozostałe ustalimy, jest nieprawidłowa, gdy wśród zmiennych niezależnych występuje silna korelacja, gdyż zmieniając jedną ze zmiennych niezależnych zmieniamy równocześnie skorelowaną z nią inną zmienną niezależną.

Analiza odstających wartości zmiennej X

Na początku była definicja h_{ii}. Można pokazać, że zachodzą dla niego:

$$0 \le h_{ii} \le 1, \quad \sum_{i=1}^{n} h_{ii} = p \qquad \text{(gdzie p to liczba parametrów).}$$

h_{ii} jest wskaźnikiem czy dany i-ty przypadek jest odstający (w odniesieniu do zmiennych niezależnych), nazywany jest dźwignią.

Uzasadnienie: można pokazać, że jest on odległością pomiędzy i-tym przypadkiem a „środkiem" danych (przypadek średni).

Obserwacja

1. Czym większa wartość h_{ii} tym bardziej odstający jest i-ty przypadek a jednocześnie tym większy jest jego wpływ na obliczenie wartości teoretycznych $\hat{Y}_i$. Jest tak dlatego, że $\hat{Y}_i$. jest liniową kombinacją Y gdzie h_{ii} jest wagą Y_i ($\hat{Y} = HY$).
2. Czym większe h_{ii} tym mniejsza wariancja błędu, gdyż $\sigma^2\{e_i\} = \sigma^2(1 - h_{ii})$. W przypadku skrajnym tj. gdy $h_{ii}=1$ to $\sigma^2\{e_i\} = 0$, zatem wartość teoretyczna pokrywa się z prawdziwą. Zatem przypadki z dużą wielkością h_{ii} mają niezbyt dużą wariancję reszt, zatem wykrycie wartości odstającej na podstawie obserwacji wyłącznie reszt może być niemożliwe.

Wskazówka praktyczna:
Obserwację uznaje się za odstającą jeśli $h_{ii} > 2p/n$.
$h_{ii} > 0{,}5$ – mówi się bardzo duża dźwignia
$0{,}2 < h_{ii} < 0{,}5$ – średnia wielkość dźwigni

Uwaga porządkująca:
Idea:
Najpierw należy zidentyfikować elementy odstające, potem przeprowadzić dla nich analizę ich wpływu na analizę model. Następnie podjąć decyzję czy wyeliminować przypadki z bazy danych.

Analiza odstających wartości zmiennej Y

Eliminacja reszt studentyzowanych
Reszty:

$$e_i = Y_i - \hat{Y}_i$$

Standaryzowane reszty:

$$\frac{e_i}{\sqrt{MSE}}$$

Wewnętrznie studentyzowane reszty

W sytuacji gdy e_i mają znacząco różne wariancje $\sigma^2\{e_i\}$ $(=\sigma^2(1-h_{ii}))$ należy rozważyć e_i w stosunku do $\sigma\{e_i\}$, które umożliwią nam rozpoznanie różnic w przypadku błędów. Wówczas wewnętrznie studentyzowane reszty definiuje się jako:

$e_i^* = \frac{e_i}{s\{e_i\}}$, gdzie $s^2\{e_i\}=MSE(1-h_{ii})$ jest nieobciążonym estymatorem wariancji

Reszty e_i mają różne próbkowe wariancje o ile wartości dźwigni hii zmieniają się znacząco, podczas, gdy wewnętrznie studentyzowane reszty mają stałą wariancję (o ile model jest właściwy).

Usunięte reszty (deleted residuals)
Motywacja:
Załóżmy, że dla pewnego i Y_i jest obserwacją odstającą. Na dopasowanie współczynników regresji może mieć duży wpływ taka obserwacja. W konsekwencji wartość teoretyczna $\hat{Y}_i$ może być bardzo blisko Y_i tzn. reszta $e_i = Y_i - \hat{Y}_i$ jest mała, a to nie świadczy o tym że obserwacja jest odstająca. Inna sytuacja będzie jeśli przed dokonaniem analizy usunięto odstający przypadek. Wówczas reszta będzie duża odzwierciedlając tym samym nietypowość Y_i.

Idea: eliminujemy obserw. odstającą, przeprowadzamy analizę regresji, następnie do wyestymowanego modelu regresji podstawiamy wyeliminowany przypadek i obliczamy tzw. usunięte reszty:

$d_i = Y_i - \hat{Y}_{i(i)}$, gdzie ost.symbol oznacza, ze i –ty przypadek został wyeliminowany z analizy regresji.

Można pokazać, że $d_i = \frac{e_i}{1-h_{ii}}$ (czynniki wyliczone w oparciu o wszystkie przypadki).

Obserwacja:
hii rośnie, to rośnie di.
Reszty te w odróżnieniu od pierwotnych identyfikują obserwacje odstające.

Fakt:

1) $s^2(d_i)=MSE_{(i)}/(1-h_{ii})$, gdzie $MSE_{(i)}$ zostało obliczone, gdy i-ty przypadek został ominięty.
2) $\frac{d_i}{s(d_i)} \quad - \quad t(n-p-1)$

Studentyzowane usunięte reszty (Studentized Deleted Residuals)

Kombinacja dwóch powyżej zdefiniowanych reszt prowadzi do pojęcia **Studentized Deleted Residuals:**

$$d_i^* = \frac{d_i}{s\{d_i\}} = \frac{e_i}{\sqrt{MSE_{(i)}(1-h_{ii})}}.$$

Wyznaczanie takich reszt nie musi łączyć się z dokonywaniem analizy regresji przy każdorazowym wyeliminowaniu kolejnego przypadku bo:

$$d_i^* = e_i\left[\frac{n-p-1}{SSE(1-h_{ii})-e_i^2}\right]^{1/2}$$

Fakt

d_i^* - t(n-p-1)

Duże wartości d_i^* przemawiają za tym, że i-ta obserwacja jest odstająca

Identyfikacja przypadków wpływowych – miary DFFITS, DFBETAS i odległość Cook'a

Umiejąc zidentyfikować obserwacje odstające należy zbadać czy są to przypadki wpływowe.

$$(DFFITS)i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$$ (DF – difference)

Jest to mara określająca wpływ i-tego przypadku na teoretyczną wartość $\hat{Y}_i$ (tj. gdy wszystkie przypadki są uwzględnione w analizie).

Interpretacja:

DFFITSi dla i-tego przypadku przedstawia „z grubsza" liczbę wyestymowanych odchyleń standardowych wpływu i-tego przypadku.

DFFITSi można liczyć na podstawie wszystkich przypadków:

$$DFFITS_i = d_i^*\left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

Jeśli i-ty przypadek X jest odstający i ma dużą wartość h_{ii}, to miara DFFITS rośnie (na wartość bezwzględną).

Zasada praktyczna:

1) dla małej lub średnio licznej próby: przypadek jest wpływowy jeśli |DFFITSi|>1
2) dla licznej próby |DFFITSi|>2(p/n)^0.5

DFBETAS

Mierzy wpływ i-tego przypadku na każdy ze współczynników regresji b_k (k=0,1,...p-1)

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}},$$ k=0,1,...,p-1, gdzie

$b_{k(i)}$ jest oszacowaniem k-tego współczynnika, gdy usuniemy i-ty przypadek.

c_{kk} jest k-tym elementem z diagonalii macierzy $(X'X)^{-1}$

Zasada praktyczna:

1) dla małej lub średnio licznej próby: przypadek jest wpływowy jeśli |DFBETASi|>1
2) dla licznej próby |DFFITSi|>2/n^0.5

Odległość Cook'a

Mierzy wpływ i-tego przypadku na cały wektor estymowanych parametrów modelu regresji.

$$D_i = \frac{\left(b - b_{(i)}\right)' X' X \left(b - b_{(i)}\right)}{pMSE},$$

gdzie $b_{(i)}$ jest wektorem parametrów regresji wyestymowanym na podstawie danych z pominięciem i-tego przypadku, natomiast b, gdy do analizy dopuszczono wszystkie wielkości.
Miarę tę można liczyć na podstawie pełnego modelu, gdyż:

$$D_i = \frac{e_i^2}{pMSE}\left[\frac{h_{ii}}{\left(1 - h_{ii}\right)^2}\right]$$

Czym większe jest e_i lub h_{ii}, tym większe jest D_i.

Przyjmuje się, że rozkład Di przybliża się poprzez F(p,n-p). Wyznaczając wielkość percentyla wnioskuje się o roli i-tego przypadku:

a) percentyl<=0,2, to i-ty przypadek ma mały wpływ na współczynniki regresji
b) percentyl=>0,5, to i-ty przypadek ma duży wpływ na dopasowanie współczynników regresji

Uwaga krytyczna:
Prezentowane miary wpływu danego przypadku zwykle właściwie diagnozują. Jednak nietrudno wyobrazić sobie sytuację, gdy mamy 2 obserwacje nietypowe blisko siebie. Wówczas usunięcie jednej z nich nie ma istotnego odzwierciedlenia we współczynnikach bo ta druga dana ,,tuszuje" jej brak wpływu.

Wnioskowanie o wpływie przypadku
Jeśli nie ma wyraźnego wpływu obserwacji nietypowej na analizę regresji, to nie ma sensu wprowadzania specjalnej diagnostyki dla takiego przypadku. Sytuacja zmienia się gdy wpływ jest istotny.
Załóżmy, że mamy daną odstającą o istotnym wpływie na analizę regresji. Musimy zdecydować co zrobić z taką obserwacją.

1) jeśli istnieją przesłanki wskazujące na nietypową sytuację, której wynikiem jest nietypowy przypadek i taka sytuacja nie jest istotna, bo bardzo sporadyczna (np. zawał operatora skrawarki) to usuwamy przypadek
2) Jeśli błąd pomiaru nie miał miejsca i nie było to przejawem nietypowego zdarzenia wówczas to raczej model jest nieodpowiedni (na wskutek nieuwzględnienia dodatkowej zmiennej niezależnej)
3) Jeśli obserwacja jest nietypowa i nie ma uzasadnienia, że jest wynikiem błędu gromadzenia danych oraz nie ma sensownej interpretacji jej występowanie. Wówczas o wiele lepszym posunięciem niż jej eliminacja jest zmniejszenie jej wpływu. Jeśli obserw. odst. dotyczy jednej ze zmiennych niezależnych wówczas należy zastosować transformację zmiennych tj logarytm, pierw. kwadratowy i inne. Oczywiście to mam sens jeśli transformacja nie zrodzi innych problemów.