Estymacja i prognozowanie

Maciej Kostrzewski
AGH
Kraków

1 luty 2010

1 Regresja Wieloraka

Motywacja: ceny mieszkań, a ...?

Rozwiązanie: Opis związku między Y a $X_1, ..., X_k$. Tablica danych.

$$\begin{matrix} y_1 & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ y_n & x_{n1} & \dots & x_{nk} \end{matrix}$$

Poszukujemy hiperpłaszczyzny najlepiej dopasowanej do tego zbioru. Model regresji z k zmiennymi objaśniającymi:

$$y_i = \beta_0 x_{i0} + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \varepsilon_i,\ i = 1, ..., n$$

gdzie $x_{i0} \equiv 1$.

Uwaga 1 *β_0 jest odpowiednikiem wyrazu wolnego w regresji liniowej*

Postać macierzowa:

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \dots & x_{1k} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \dots & x_{nk} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y = X\beta + \varepsilon$$

y - wektor obserwacji zmiennej objaśniającej $(nx1)$ (realizacja zmiennej Y)
X - macierz obserwacji zmiennych objaśniających $(nx\,(1+k))$
β - wektor współczynników $((1+k)\,x1)$
ε -wektor reszt $(nx1)$.

Dodatkowe założenia

$$X = \begin{bmatrix} x_{10} & \dots & x_{1k} \\ \vdots & & \vdots \\ x_{n0} & \dots & x_{nk} \end{bmatrix}$$

ma ustalone elementy (brak losowości)

$$\begin{aligned} E\varepsilon &= 0 \\ E\varepsilon\varepsilon^T &= \sigma^2\mathbf{I} \\ rz\left(X\right) &= 1+k\leq n \end{aligned}$$

Obserwacja 1 *:*

1. $rz\left(X\right)=1+k$
2. $E\varepsilon_i^2=\sigma^2$
3. $E\varepsilon_i\varepsilon_j=0$ dla $i\neq j$.
4. X jest deterministyczna $\Rightarrow X\perp\varepsilon$

Estymacja:

$$\widehat{Y}=X\widehat{\beta}+e$$

Metoda najmniejszych kwadratów:

$$SSE=e^Te=\left(y-X\widehat{\beta}\right)^T\left(y-X\widehat{\beta}\right)=y^Ty-2\beta^TX^Ty+\widehat{\beta}^TX^TX\widehat{\beta}$$

$$\min_{\widehat{\beta}}SSE\left(\widehat{\beta}\right)=\min_{\widehat{\beta}}y^Ty-2\widehat{\beta}^TX^Ty+\widehat{\beta}^TX^TX\widehat{\beta}$$

Różniczkujemy

$$\begin{aligned} \frac{\partial SSE}{\partial\widehat{\beta}} &= -2X^Ty+2X^TX\widehat{\beta} \\ \frac{\partial SSE}{\partial\widehat{\beta}} &= 0 \\ X^TX\widehat{\beta} &= X^Ty \text{ (układ równań normalnych} \end{aligned}$$

$$\widehat{\beta}=\left(X^TX\right)^{-1}X^Ty$$

Gdyby $rzX<1+k$ wówczas estymator nie jest określony jednoznacznie. Wartości teoretyczne:

$$\widehat{y}=X\widehat{\beta}$$

$$e=\widehat{\varepsilon}=y-\widehat{y}$$

Lemat 1 *Niech* $\widehat{\beta}=\left(X^TX\right)^{-1}X^Ty$ *wówczas*

1. $E\left(\widehat{\beta}\right)=\beta$
2. $\sum_{\widehat{\beta}\widehat{\beta}}=\sigma^2\left(X^TX\right)^{-1}$

Definicja 1 *Estymator $\widehat{\theta}$ jest najlepszym nieobciążonym estymatorem wektora θ (inaczej najefektywniejszym), gdy $\widehat{\theta}$ jest estymatorem nieobciążonym oraz macierz*

$$E\left(\widetilde{\theta}-\theta\right)\left(\widetilde{\theta}-\theta\right)^{T}-E\left(\widehat{\theta}-\theta\right)\left(\widehat{\theta}-\theta\right)^{T}$$

jest nieujemnie określona, gdzie $\widetilde{\theta}$ jest dowolnym innym nieobciążonym estymatorem θ.

Twierdzenie 1 (Gaussa i Markowa) *W modelu regresji wielorakiej najlepszym nieobciążonym estymatorem liniowym wektora β jest wektor wyznaczony metodą najmniejszych kwadratów*

Twierdzenie 2 *Nieobciążonym estymatorem wariancji σ^2 składnika losowego jest*[1]

$$\widehat{\sigma}^2=\frac{SSE}{n-k-1}$$

gdzie $SSE=e^Te=\sum_{i=1}^{n}e_i^2=\sum_{i=1}^{n}\left(y_i-\widehat{y}_i\right)^2$.

Definicja 2 *Klasyczny model normalnej regresji wielorakiej to model regresji wielorakiej z dodatkowym założeniem $\varepsilon\sim N\left(0,\sigma^2I\right)$ tzn. $E\varepsilon=0$ oraz $E\varepsilon\varepsilon^T=\sigma^2I$*

Obserwacja 2 *Z braku korelacji między ε_i i ε_j wynika, że ε_i dla $i=1,...,n$ są niezależnymi ziennymi losowymi o rozkładach normalnych t. że $E\varepsilon_i=0$ oraz $Var\left(\varepsilon_i\right)=\sigma^2$.*

Wniosek 1 *Niech $i\in\{0,...,k\}$*

$$H0:\beta_i=c_i$$

$$t_{statystyka}=\frac{\widehat{\beta}_i-c_i}{SE\left(\widehat{\beta}_i\right)}\sim t_{n-k-1}$$

gdzie $SE\left(\widehat{\beta}_i\right)=\sqrt{\left(\sigma^2\left(X^TX\right)^{-1}\right)_{ii}}$. Jest to test dwustronny.

Wniosek 2 *W praktyce interesuje nas wpływ zmiennych niezależnych (bez wyrazu wolnego) na Y.*

$$\begin{aligned}H0 &: \beta_1=0,...,\beta_k=0\\ H1 &: \exists i\in\{0,...,k\}:\beta_i\neq 0\end{aligned}$$

$$F_{statystyka}=\frac{MSR}{MSE}\sim F_{k,n-k-1}$$

gdzie $MSR=\frac{SSR}{k}$ (k bo tyle zmiennych niezaleznych X), $MSE=\frac{SSE}{n-k-1}$ (bo n punktów ale $k+1$ parametrów do oszacowania), $SSR=\sum_{i=1}^{n}\left(\widehat{y}_i-\overline{y}\right)^2$. Jest to test prawostronny.

[1] Pierwiastek z niego nie jest estymatorem nieobciążonym odchylenia σ.

Uwaga 2 *Jeśli nie można odrzucić hipotezy $H0$, to analiza regresji się kończy. W przeciwnym wypadku wiemy, że są statystyczne podstawy, by przypuszczać, że zachodzi związek liniowy pomiędzy zmienną objaśnianą i co najmniej jedną zmienną niezależną.*

Uwaga 3 *W praktyce przeprowadzamy najpierw test F, a dopiero potem testy t.*

Przedziały ufności:
Idea: estymator±wartość krytyczna*odchylenie standardowe (estymatora)

Twierdzenie 3 $(1-\alpha)\,100\%$ *przedział ufności dla parametru β_i jest postaci:*

$$\widehat{\beta}_i \pm t_{n-k-1,\alpha/2} SE\left(\widehat{\beta}_i\right)$$

Ćwiczenie 1 *Wykonać analizę regresji wielorakiej dla danych SMSA. Przyjąć, że zmienna Crime jest zmienną objaśnianą.*

Rozwiązanie 1 *1. Dokonać wyboru zmiennych kierując się intuicją*

2. *Zaznacz obszar 5x(1+k), naciśnij* F2 *wpisz* =REGLINP(Zakres_Y;Zakres_X;1;1) ; CTRL+SHIFT+ENTER:

$\widehat{\beta}_k$	$\widehat{\beta}_{k-1}$	$\widehat{\beta}_{k-2}$	...	$\widehat{\beta}_0$
$SE\left(\widehat{\beta}_k\right)$	$SE\left(\widehat{\beta}_{k-1}\right)$	$SE\left(\widehat{\beta}_{k-2}\right)$	...	$SE\left(\widehat{\beta}_0\right)$
r^2	$\widehat{\sigma}$	–	...	–
F	df	–	...	–
SSR	SSE	–	...	–

3. *Interpretacja współczynników*

4. *Wykonaj analizę regresji korzystając z* Narzędzia>Analiza Danych>Regresja *(zaznaczyć tytuły)*

5. *Porównaj wyniki uzyskane w punktach* 2 *i* 4.

6. *Wyjaśnij co oznacza "Istotność F". Czy "Istotność F "to p-value?[TAK] Rozkład.F(x)=???, a co piszą w Pomocy?*

7. *Dokonaj analizy istotności wpływu poszczególnych zmiennych objaśniających, korzystając z testu t, na zmienną objaśniana Y.*

8. *Stworzyć macierz korelacji:* Narzędzia>Analiza Danych>Korelacja. *Przeprowadzić dyskusję nad doborem zmiennych objaśniających.*

Współliniowość:

Ćwiczenie 2 *Wykonać próbę analizy regresji dla dowolnych zmiennych: Y, X_1, X_2, X_3, gdzie $X_3 = 0,2X_1 - 0,4X_2$.*

Załóżmy, że zmienna X_{k+1} wywołuje współliniowość po dołączeniu do modelu w którym znajdują się $X_1, ..., X_k$. Niech ta współliniowość wynika ze współzależności tej zmiennej ze zbiorem zmiennych niezależnych. Podstawowym skutkiem współliniowości jest zbyt wysoka wariancja estymatorów współczynników regresji. Aby zmierzyć ten skutek współlinowości oblicza się VIF

Definicja 3 *Wskaźnik nadmiaru wariancji VIF (variance inflation factor) związany ze zmienną X_{k+1}:*

$$VIF\left(X_{k+1}\right) = \frac{1}{1 - R_{k+1}^2}$$

gdzie R_{k+1}^2 jest wartością współczynnika R^2 dla regresji gdzie zmienną zależną jest X_{k+1} a zbiorem zmiennych niezależnych jest $X_1, ..., X_k$.

Uwaga 4 *Można wykazać, że VIF jest ilorazem wariancji estymatora β_{k+1} do wariancji tego współczynnika, gdyby zmienna X_{k+1} była nieskorelowana z pozostałymi, stąd nazwa miary jako wskaźnik nadmiaru wariancji estymatora.*

Uwaga 5 *VIF jest kolejnym wskaźnikiem, obok macierzy korelacji, na istnienie współliniowości.*

Ćwiczenie 3 *Zastosować VIF jako miarę współliniowości dla danych SMSA*

Dobór zmiennych

Ćwiczenie 4

1. *Dobór zmiennych do modelu na podstawie zmian współczynnika R_p^2.*
2. *Przeprowadzić dobór zmiennych dla SMSA (ograniczyć zbiór zmiennych objaśniających do silnie sokrelowanych z CRIME - powyżej 0,9) metodą w przód.*

Częściowy test F

Wychodzimy od modelu w którym znajduje się już $k-l$ zmiennych. Chcemy sprawdzić istotność związku Y oraz pewnego l elementowego podzbioru zmiennych objaśniających, przy założeniu, że w modelu znajduje się już $k-l$ zmiennych -jest to tzw. względna istotność, bo względem $k-l$ zmiennych.

Model

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 X_1 + ... + \boldsymbol{\beta}_k X_k + \varepsilon$$

nazywamy modelem pełnym.

Model zredukowany to model zawierający $k-l$ zmiennych.

Statystyka:

$$F_{l,n-(1+k)} = \frac{\left(SSE_R - SSE_F\right)/l}{MSE_F}$$

gdzie SSE_R to SSE dla modelu zredukowanego, SSE_F i $MSE_F = \frac{SSE_F}{n-(1+k)}$ to odpowiednio SSE i MSE obie obliczone dla pełnego modelu.

Dobór w przód

Punktem wyjścia jest model bez zmiennych. W kolejnych krokach dołącza się zmienną wg. kryterium: najwyższa wartość testu F (równoważnemu t) przy założeniu, że F jest powyżej z góry ustalonego (przez użytkownika lub program) progu. Dobór drugiej i ewentualnie kolejnych odbywa się za pomoca testu częściowego F. Procedura kończy się, gdy nie ma już zmiennej dla której wartość statystyki (częściowego F) spełniałaby kryterium progu.

Algorithm 1 *1. Do modelu dołączamy zmienną X_j, gdy*

$$\begin{aligned} j \quad &: \quad F_j = \max\{F_i : i = 1, ..., k\} \Leftrightarrow |t_j| = \max\{|t_i| : i = 1, ..., k\} \\ &\quad \textit{oraz} \\ p_{in} \quad &\geq \quad p - value\,(j) \end{aligned}$$

2. W modelu znajduje się wektor zmiennych,gdzie $J \subset \{1, ..., k\}$ to ich zbiór indeksów. Dołącza się do modelu X_j, gdy

$$\begin{aligned} j \quad &: \quad F_j = \max\{F_i : i = \{1, ..., k\} \setminus J\} \\ &\quad \textit{gdzie } F \textit{ to częściowy test } F \textit{ oraz} \\ p_{in} \quad &\geq \quad p - value\,(j) \end{aligned}$$

Uwaga 6 *Którą ze zmiennych współliniowych usunąć z modelu? Wyłączamy tę zmienną, której usunięcie najmniej zmniejszy R^2*

Regresja krokowa

Algorithm 2 *1. Do modelu dołączamy zmienną X_j, gdy*

$$\begin{aligned} j \quad &: \quad F_j = \max\{F_i : i = 1, ..., k\} \Leftrightarrow |t_j| = \max\{|t_i| : i = 1, ..., k\} \\ &\quad \textit{oraz} \\ p_{in} \quad &\geq \quad p - value\,(j) \end{aligned}$$

2. Jeśli $p - value\,(j) > p_{out}$ to dołączoną zmienną wykluczamy - zapętlenie algorytmu; zwrócić uwagę na ustalenie relacji między p_{in} a p_{out}

3. W modelu znajduje się wektor zmiennych,gdzie $J \subset \{1, ..., k\}$ to ich zbiór indeksów. Dołącza się do modelu X_j, gdy

$$\begin{aligned} j \quad &: \quad F_j = \max\{F_i : i = \{1, ..., k\} \setminus J\} \\ &\quad \textit{gdzie } F \textit{ to częściowy test } F \textit{ oraz} \\ p_{in} \quad &\geq \quad p - value\,(j) \end{aligned}$$

Z modelu wykluczamy X_l, gdy

$$\begin{aligned} l \quad &: \quad F_l = \min\{F_i : i \in J\} \\ &\quad \textit{gdzie } F \textit{ to częściowy test } F \textit{ oraz} \\ p_{out} \quad &\leq \quad p - value\,(j) \end{aligned}$$

Koniec algorytmu, gdy nie ma zmiennych spełniających kryteria dołaczenia i nie ma zmiennych spełniających zpełniających warunki wykluczenia.

Uwaga 7 *1. Zwykle $p_{in} = 0,05$ i $p_{out} = 0,05$*

2. Gdyby $p_{in} > p_{out}$ to procedura może okazać sie rozbieżna, tj. w kolejnych krokach zmienna będzie dołączana, po to by w kolejnym zostać wykluczoną.

Regresja jakościowa:
dane Zarobki

1. Statystyki opsiowe dla zarobków mężczyzn oraz zarobków kobiet; analiza
2. Korelacja między wykształceniem, a zarobkami
3. Przeprowadzić analizę regresji wielorakiej - Y=zarobki, X_1=wykształcenie, X_2=płeć
4. Czy na podstawie analizy regresji można wyciągnąć wniosek o dyskryminacji kobiet?
5. Czy uzyskane dwie linie regresji są do siebie równoległe?

Przykład 1 *dane PasyBezpieczeństwa.*

Regresja nieliniowa:

Definicja 4 *Model regresji*

$$Y = F(X_1, ..., X_k),$$

gdzie F jest dowolną funkcją.
Modele regresji dzielimy na:

1. modele liniowe (było)
2. modele nieliniowe linearyzowane (takie, które możemy sprowadzić do modeli liniowych)
3. modele nieliniowe nielinearyzowalne

Modele linearyzowalne

Uwaga 8 *Transformację możemy odgadnąć obserwując wykresy rozrzutu zmiennej zależnej i zmiennych niezależnych.*

Model	Transformacja	Model po transformacji	
$Y = aX^p\varepsilon$ (f.potęgowa)	$\ln(Y)$	$\ln(Y) = \ln(a) + p\ln(X) + \ln(\varepsilon)$	$Y >$
$Y = ab^X\varepsilon$ (f.wykładnicza)	$\ln(Y)$	$\ln(Y) = \ln(a) + \ln(b)\,X + \ln(\varepsilon)$	$a >$
$Y = a_0 + a_1X + ... + a_pX^p + \varepsilon$	$X_1 = X, X_2 = X^2, ..., X_p = X^p$	$Y = a_0 + a_1X_1 + ... + a_pX_p + \varepsilon$	
$Y = a + b\frac{1}{X} + \varepsilon$ (f. hiperboliczna)	$X_1 = \frac{1}{X}$	$Y = a + bX_1 + \varepsilon$	
$Z = \frac{a}{1+be^{-X}+\varepsilon}$ (f. logistyczna)	$Y = \frac{1}{Z}$, $X_1 = e^{-X}$	$Y = \frac{1}{a} + \frac{b}{a}X_1 + \frac{1}{a}\varepsilon$	

Uwaga 9 *Często rezygnujemy z lepszego dopasowania na rzecz gorszego, jeśli to drugie ma dobre (lepsze) merytoryczne uzasadnienie (interpretację)*

Modele nieliniowe nielinearyzowalne

Modele które nie da się przekształcić do modeli liniowych np: $Y = ab^X + \varepsilon$

Przykład 2 *plik kombajn*

Regresja pozorna (ang. spurious regression)

Regresja pozorna ma miejsce, gdy trend zmiennej objaśniającej i trend zmiennej objaśnianej są podobne. Współczynniki regresji przy zmiennych objaśniających mogą być statystycznie istotnie różne od zera, wartość współczynnika deterinacji R^2 może być wysoka, jednakże zależność ma charakter złudny, przypadkowy, pozorny - nie ma bowiem rozsądnego uzasadnienia związku między zmiennymi!

Przykład 3 *Produkcja czekolady i produkcja energii - plik Komajn*

Przykład 4 *Inne...*

Przykład 5 ***Interpretacja*** *w modelu regresji w oparciu o Rosen (1982) "The Impact of Proposition 13 on Housing Prices in Northern California: A Test of the Interjurisdictional Capitalization Hypothesis".*

Problem założeń modelu regresji

Heteroskedastyczność

Wstępne badanie przeprowadzamy analizując wykresy: $(\widehat{y}_i, e_i)$, $\left(\widehat{y}_i, e_i^2\right)$, (numer obserwacji,$e_i$).

Jeśli reszty rosną lub maleją wraz ze wzrostem wartości teoretycznych y to mamy przesłannkę za heteroskedastycznością.

Przy heteroskedastyczności estymatory MNK mogą nie być efektywne.

Heteroskedastyczności można spróbować pozbyć się stosując transformacje Boxa i Coxa:

$$\begin{cases} y^\lambda & \lambda \neq 0 \\ \ln|y| & \lambda = 0 \end{cases}$$

1. $\lambda = -1$, to $\frac{1}{Y}$

2. $\lambda = 0$, to $\ln(Y)$ (stosujemy, gdy e_i^2 rośnie)

3. $\lambda = \frac{1}{2}$, to $\sqrt{Y}$

4. $\lambda = 2$, to Y^2

Przykład 6 *Plik Farmakologia.*

Normalność.

Normalność nie jest wymagana na etapie estymacji, ale jest potrzebna do weryfikacji istotności parametrów. Testy t i F są odporne na „niewielkieódchylenia od normalności. Rola założenia o normalności zmniejsza się przy wzrastającej próbie.

Uwaga 10 *W praktyce jeśli $n < 15(1 + k)$ to przyjmuje się, że zbiór danych jest mały. Wówczas ważne jest testowanie założenia normalności.*

Niezależność ε_i.

Definicja 5 *Autokorelacja zaburzenia losowego z opóźnieniem rzędu l to korelacja między ε_i i ε_{i-l}; oznaczenie ρ_l.*

Uwaga 11 *W praktyce najczęściej występuje autokorelacja pierwszego rzędu*

Mowa tu o korelacji między składnikami losowymi, pojawiająca się w szeregach czasowych. Korelacja ta wynika z wzajemnego skorelowania pominiętych zmiennych objasniających, które są reprezentowane przez składniki losowe (błędy).

Twierdzenie 4 *Test Durbina-Watsona:*

Założenia:

1. Modelu musi uwzględniać wyraz wolny
2. Składniki resztowe mają rozkład normalny
3. W modelu nie występuje zmienna opóźniona (np. nie można stosować testu dla $X_n = \beta_0 + \beta X_{n-1} + \varepsilon_n$)

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

$$\begin{aligned} H0 &: \rho_1 = 0 \\ H1 &: \rho_1 \neq 0 \text{ brak autokorelacji} \\ H1 &: \rho_1 > 0 \text{ dodatnia autokorelacja, gdy } d < 2 \\ H1 &: \rho_1 < 0 \text{ ujemna autokorelacja, gdy } d > 2 \end{aligned}$$

Dla $H1 : \rho_1 \neq 0$ **oraz poziomu istotności** 2α :

$d < d_L$	$d_L \leq d \leq d_U$	$d_U < d < 4 - d_U$	$4 - d_U \leq d \leq 4 - d_L$	$4 - d_L < d$
$H1$	Test nie roztrzyga	$H0$	Test nie roztrzyga	$H1$

Dla $H1 : \rho_1 > 0$ **oraz poziomu istotności** α :

$d < d_L$	$d_L \leq d \leq d_U$	$d_U < d < 4 - d_U$	$4 - d_U \leq d \leq 4 - d_L$	$4 - d_L < d$
$H1$	Test nie roztrzyga		$H0$	

Dla $H1 : \rho_1 < 0$ **oraz poziomu istotności** α :

$d < d_L$	$d_L \leq d \leq d_U$	$d_U < d < 4 - d_U$	$4 - d_U \leq d \leq 4 - d_L$	$4 - d_L < d$
	$H0$		Test nie roztrzyga	$H1$

Uwaga 12 *Mankamentem testu jest, że nie dla każdej wartości statystyki test wskazuje na hipotezę*

Uwaga 13 *W przypadku statystycznego udowodnienia istnienia autokorelacji wówczas wyniki analizy regresji są niewiarygodne, rozwiązaniem jest zastosowanie uogólnionej metody najmniejszych kwadratów.*

Przykład 7 *plik TestDW*

Obserwacje nietypowe

Przykład 8 *Plik outliers*

Obserwacje nietypowe inaczej obserwacje skrajne (outliers) to obserwacje istotnie różniące się od pozostałych.

Dla regresji prostej wykrycie obserwacji odstających umożliwia analiza wykresu. W przypadku regresji wielorakiej czasami analiza reszt oszacowanego modelu umożliwia wykrycie obserwacji nietypowych.

Uwaga 14 *Idealny model nie dopuszcza do obserwacji odstających. W idealnym modelu każda z obserwacji jest typowa.*

Uwaga 15 *Przyczyny występowania danych nietypowych*

1. *błędy w trakcie zapisu danych*
2. *słaby model, który nie uwzględnia istotnej zmiennej objaśniającej - mówimy o danych nietypowych dla modelu.*
3. *Nietypowe zjawisko/warunki w okresie badanym np. okres trwania wojny*

Uwaga 16 *Jedną z konsekwencji występowania nietypowych danych jest zmiana wartości estymatorów.*

Uwaga 17 *Wśród danych nietypowych wyróżniamy wpływowe i te nie mające wpływu na estymację parametrów. Pierwsze z nich mogą być groźne.*

Algorithm 3 1. *Identyfikacja obserwacji odstających.*

2. *Wyznaczenie wpływu obserwacji odstających na analizę regresji.*

3. *Decyzja o wykluczeniu lub pozostawieniu w bazie danych przypadków odstających i wpływowych.*

Uwaga 18 *Najprostrze jest usunięcie zmiennej odstającej i ponowne wykonanie analizy regresji. Takie działanie może prowadzić do błędów.*

Uwaga 19 *Dane nietypowe mogą przyciągać płaszczyznę regresji do siebie, a wówczas, gdy występuja np. 2 nietypowe obserwacje obok siebie to ich identyfikacja jest trudna - obserwacje te nawzajem tuszują się.*

Uwaga 20 *Postępowanie*

1. *Analiza najmniejszych i największych wartości każdej ze zmiennych objaśniających i objaśnianej.*

2. *Obserwacja reszt - duże reszty mogą wskazywać na obserwacje odstające.*

3. *Wewnętrznie studentyzowane reszty:*

$$e_i^{st} = \frac{e_i}{\sqrt{[1-h_{ii}]\frac{1}{n-1-k}\sum_{i=1}^{n} e_i^2}} = \frac{e_i}{\sqrt{MSE\,[1-h_{ii}]}},$$

gdzie k to liczba zmiennych niezależnych. Jeśli $|e_i^{st}| > 3$ to należy się przyjrzeć danemu przypadkowi.

4. $\widehat{Y} = X\widehat{\beta} = HY$, $H = X\left(X^TX\right)^{-1}X^T$, $e = (I-H)Y$, *macierz kowariancji* $\sum_{e_ie_j} = \sigma^2(I-H)$, $Var(e_i) = \sigma^2(1-h_{ii})$, *gdzie* h_{ii} *to element z diagonalii* H.

   Można pokazać, że h_{ii} *jest odległością i tego przypadku od „środka ciężkości danych"(średniego przypadku)*X.

   Im większe jest h_{ii} *tym mniejsza jest wariancja błędu, gdyż* $Var(e_i) = \sigma^2(1-h_{ii})$. *W przypadku* $h_{ii} = 1$, *to* $Var(e_i) = 0$, *co oznacza, że wartość teoretyczna pokrywa się z prawdziwą. Przypadki z dużą wielkością* h_{ii} *mają małą wariancję reszt, zatem ich wykrycie na podstawie wyłacznie analizy dużych wartości reszt jest niemożliwe - jakimś rozwiązaniem jest analiza przypadków o małych i dużych wartościach reszt.*

   h_{ii} *to wskaźnik czy dany przypadek jest odstający od pozostałych (w kontekście zmiennych objaśniającyh) oraz czy jest wpływowy (na model tj.y).* h_{ii} *nazywany jest dźwignią.*

   Im większa jest wartość h_{ii} *tym jest jego wpływ na analizę regresji jest większy, gdyż* $\widehat{y}_i$ *jest liniową kombinacją* Y *z wagą* h_{ii} *(*$\widehat{Y} = HY$ *).*

   Im większa jest wartość h_{ii} *tym bardziej odstający jest i ty przypadek, a jednocześnie większy jest jego wpływ na analizę regresji .*

   Wskazówka praktyczna: obserwacje uznaje się za odstające, gdy $h_{ii} > \frac{2(1+k)}{n}$. *Gdy* $h_{ii} \geq 0,5$ *to mówimy, że przypadek ma bardzo dużą dźwignię (wpływ). Gdy* $0,2 < h_{ii} < 0,5$ *to mówimy o średnim wpływie czy średniej wielkości dźwignii.*

5.

$$d_i = y_i - \widehat{y}_{i(i)}$$

$$d_i = \frac{e_i}{1 - h_{ii}}$$

Zauważmy, że, gdy h_{ii} rośnie to d_i rośnie.

6. *Studentized deleted Residuals.*

$$d_i^* = \frac{d_i}{\sqrt{Var\left(d_i\right)}} = \frac{d_i}{\sqrt{MSE_{(i)}\left(1 - h_{ii}\right)}}$$

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

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

7. *Identyfikacja przypadków wpływowych - miary DFFITS, DFBETAS i odległość Cook'a.*

   - *DF=różnica (difference); FIT=dopasowanie; S=studentyzowalne*

$$\left(DFFITS\right)_i = \frac{\widehat{y}_i - \widehat{y}_{i(i)}}{\sqrt{MSE_{(i)} h_{ii}}}$$

Miara określa wpływ i-tego przypadku na teoretyczną wartość $\widehat{y}_i$ (tj. w sytuacji, gdy każdy z przypadków jest brany pod uwagę w analizie).

$$DFFITS_i = d_i^* \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

gdy h_{ii} rosnie to $|DFFITS_i|$ też rośnie.
Wskazówki praktyczne: dla małej próby luż średnio licznej próby jeśli $|DFFITS_i| > 1$ to mówimy, ze przypadek jest wpływowy; dla licznej próby jełśi $|DFFITS_i| > 2\sqrt{\frac{1+k}{n}}$ to mówimy, że przypadek jest wpływowy

   - *Odległość Cooka*

$$D_i = \frac{\left(\widehat{\beta} - \widehat{\beta}_{(i)}\right)^T X^T X \left(\widehat{\beta} - \widehat{\beta}_{(i)}\right)}{\left(1 + k\right) MSE} \sim F(1 + k, n - 1 - k)$$

$\widehat{\beta}_{(i)}$ jest wektorem parametrów wyestymowanych bez uwzględnienia i-tego przypadku. Miarę tę można obliczyć dla pełnego modelu, gdyż

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

Wraz ze wzrostem e_i lub h_{ii} rośnie D_i.
Wskazówka praktyczna : $P(F < D_i) = p$, gdzie $F \sim F(1+k, n-1-k)$ to jeśli $p < 0,2$ to i-ty przypadek ma mały wpływ na model, jeśli $p > 0,5$ to i-ty przypadek ma duży wpływ na model.

Uwaga 21 *Jesli obserwacja jest nietypowa i nie ma uzasadnienia, że jest wynikiem błedu gromadzenia danych oraz nie ma sensownej interpretacji jej wystepowania. Wówczas o wiele lepszym posunieciem niż jej eliminacja jest zmniejszenie jej wpływu. Jeśli obserw. odst. dotyczy jednej ze zmiennych niezale nych wówczas należy zastosowac transformacje zmiennych tj logarytm, pierw. kwadratowy i inne. Oczywiscie to mam sens jesli transformacja nie zrodzi innych problemów.*

Uwaga 22 *Excel w kolumnie „Std. składniki resztowe"podaje:*

$$\frac{e_i}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n} e_i^2}}$$