Colloquium Biometricum 38
2008, 33–49

METODY BADANIA OBSERWACJI WPŁYWOWYCH W ROZPOZNANIU WARUNKÓW WODNYCH W GLEBIE

Agnieszka Kamińska[1)], Mirosława Wesołowska-Janczarek[1)], Antoni Grzywna[2)]

[1)] Katedra Zastosowań Matematyki i Informatyki
[2)]Katedra Melioracji i Budownictwa Rolniczego
Uniwersytet Przyrodniczy w Lublinie
Akademicka 13, 20-950 Lublin
e-mail: agnieszka.kaminska@up.lublin.pl

Streszczenie

W pracy przedstawiono zastosowanie technik diagnostyki regresji liniowej wielu zmiennych do analizy zależności zapasu wody w warstwie korzenienia się roślin od rzędnej położenia zwierciadła wody gruntowej. Wykorzystane metody diagnostyczne pozwoliły dokonać kompleksowej analizy wpływu poszczególnych obserwacji i ich struktury na wyniki estymacji, wskazać obserwacje naruszające jednorodność danych oraz określić przyczynę ich wystąpienia.

Słowa kluczowe: diagnostyka modelu regresji, stosunki wodne w glebie

Klasyfikacja AMS 2000: 62J05

1. Wstęp

Podczas rozwiązywania problemów, w których stosuje się analizę regresji, można natrafić na szereg trudności. Analizowane dane empiryczne wykazują często różny stopień nietypowości, co powoduje znaczne pogorszenie efektywności szacowanego równania regresji. Ważne bowiem jest to, aby model nie był

nadmiernie uwarunkowany przez pojedyncze obserwacje o wartościach mocno różniących się od typowych dla próby, które mogą istotnie zakłócić wyniki obliczeń i prowadzić do błędnych wniosków. Powszechnie stosowaną metodą szacowania parametrów równania regresji jest metoda najmniejszych kwadratów, która jest bardzo wrażliwa na wszelkie odstępstwa od przyjętych założeń. Stąd też wynika potrzeba opracowania metod odpowiedniej diagnostyki danych wykorzystywanych w procesie modelowania. Dokładna analiza struktury danych wykorzystywanych do estymacji modelu regresji, połączona z analizą reszt oraz analizą wrażliwości, pozwala często dokonać istotnych, korzystnych zmian w specyfikacji modelu oraz wykryć błędy w danych statystycznych.

W pracy przedstawiono zastosowanie technik diagnostyki regresji liniowej wielu zmiennych do analizy zależności zapasu wody w warstwie korzeniowej gleby od rzędnej położenia zwierciadła wody gruntowej. Wykorzystane metody diagnostyczne pozwoliły dokonać kompleksowej analizy wpływu poszczególnych obserwacji i ich struktury na wyniki estymacji, wskazać obserwacje naruszające jednorodność danych oraz określić przyczynę ich wystąpienia.

2. Materiał i metody badań

Do badań wybrano obiekt melioracyjny Ochoża położony na Pojezierzu Łęczyńsko-Włodawskim w dolinie rzeki Tyśmienica. Badania stosunków wodnych prowadzono w latach 1999-2002 w 5 przekrojach hydrometrycznych łącznie w 59 punktach. Badania obejmowały analizę właściwości fizyko-chemicznych gleb, waloryzację użytkowania terenu, inwentaryzację stanu urządzeń melioracyjnych i pomiary hydrologiczne, prowadzono je w różnych punktach położonych na środku odwadnianej kwatery na użytkach zielonych. W celu prowadzenia w nich pomiarów stanów wody gruntowej w m założono studzienki pomiarowe. W tych samych punktach z warstwy 0–30 cm pobierano próbki gleby dla których oznaczano uwilgotnienie metodą suszarkowo-wagową. Następnie uwilgotnienie gleby przeliczano na wielkość zapasu wody w mm w badanej warstwie gleby. W celu odniesienia otrzymanych zależności do danych warunków siedliskowych wykonano odkrywki glebowe, z których pobrano próbki do badań. Na podstawie przeprowadzonych badań terenowych określono budowę morfologiczną profilu glebowego, a następnie ustalono rodzaj gleby.

Jednym z analizowanych elementów było określenie wpływu rzędnej położenia zwierciadła wody gruntowej [m n. p. p.] na wielkość zapasu wody w warstwie korzeniowej gleby [mm], gdzie n. p. p. oznacza nad poziomem porównawczym przyjętym na poziomie zera wodowskazu w rzece. W niniejszej pracy analizie poddano wyniki uzyskane na kwaterze nr 4, w piezometrze 44, przeprowadzone łącznie w 32 terminach. Zestawienie wykorzystanych w pracy danych wraz ze szczegółowym opisem można znaleźć w pracy Grzywny (2003).

3. Model regresji liniowej wielu zmiennych

W pracy rozważany jest ogólny model liniowej regresji wielokrotnej postaci

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_{p-1} X_{p-1} + \varepsilon,$$

opisujący zależność między zmienną objaśnianą Y, a zmiennymi objaśniającymi $X_1, X_2, ..., X_{p-1}$.

W notacji macierzowej model ten przyjmuje postać

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

gdzie $\mathbf{y}$ jest wektorem obserwacji o wymiarach $(n \times 1)$, $\mathbf{X}$ jest niestochastyczną macierzą pełnego rzędu, o znanej postaci (ustalonych elementach) i wymiarach $(n \times p)$, gdzie $n > p$, $\boldsymbol{\beta}$ jest wektorem parametrów o wymiarach $(p \times 1)$, a $\boldsymbol{\varepsilon}$ jest wektorem błędów losowych o wymiarach $(n \times 1)$. Dodatkowo $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I})$, co oznacza, że błędy losowe ε_i ($i = 1,2...,n$), są niezależnymi zmiennymi losowymi o rozkładzie normalnym, ze średnią 0 i wariancją σ^2 dla każdej z n obserwacji.

Przyjęte założenia dotyczące wektora losowego $\boldsymbol{\varepsilon}$ implikują, iż wartość oczekiwana wektora losowego $\mathbf{y}$, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$, a jego macierz wariancji $\mathbf{D}(\mathbf{y}) = \sigma^2 \mathbf{I}$.

Estymatory współczynników regresji otrzymane metodą najmniejszych kwadratów mają postać

$$\mathbf{b} = (\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y},$$

Wektor oszacowanych wartości zmiennej objaśnianej ma postać

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} = \mathbf{H}\mathbf{y},$$

gdzie

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{T}}.$$

Wektor reszt jest postaci

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y},$$

przy czym $E(\mathbf{e}) = \mathbf{0}$, $\mathbf{D}(\mathbf{e}) = \sigma^2(\mathbf{I} - \mathbf{H})$, $\mathbf{e} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$ oraz $\frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{\sigma^2} \sim \chi^2_{n-p}$,

gdzie χ^2_{n-p} oznacza rozkład χ^2 z $n - p$ stopniami swobody.

Nieobciążonym estymatorem nieznanej wariancji σ^2 jest

$$s^2 = \frac{\mathbf{e}^{\mathbf{T}}\mathbf{e}}{n-p} = \frac{\mathbf{y}^T(\mathbf{I} - \mathbf{H})\mathbf{y}}{n-p}.$$

4. Reszty i ich modyfikacje

Przy szacowaniu parametrów modelu czyni się wiele założeń dotyczących błędów losowych. Czy przyjęcie tych założeń było słuszne, rozstrzyga się na etapie weryfikacji modelu. Analiza błędów losowych pozwala także na ocenę trafności doboru postaci modelu oraz zestawu zmiennych objaśniających. Narzędziami służącymi do sprawdzania przyjętych założeń o błędach losowych są testy statystyczne. Weryfikacja hipotez jest przeprowadzana na podstawie ciągu

reszt będących oszacowaniami odchyleń losowych. Bardzo przydatne są również metody graficzne i wykresy reszt.

W diagnostyce regresji proponuje się zastosowanie prostych transformacji reszt uzyskanych po zastosowaniu MNK dla celów wnioskowania statystycznego postaci

$$r_i = \frac{e_i}{\mathrm{s}\sqrt{1-h_{ii}}}, \; i = 1,2,...,n,$$

oraz

$$r_i^* = \frac{e_i}{s_{(i)}\sqrt{1-h_{ii}}}, \; i = 1,2,...,n,$$

gdzie $s_{(i)}$ jest oceną odchylenia standardowego składnika losowego σ, wyznaczonego po usunięciu i-tej obserwacji (Chatterjee i Hadi, 1988).

5. Wskaźniki wpływu

Ważną rolę w diagnostyce liniowych modeli regresji odgrywa macierz $\mathbf{H}$. Wartości diagonalnych tej macierzy h_{ii} nazywane są wskaźnikami wpływu (*ang. leverage*) i mieszczą się w przedziale $\langle 1/n;1)$. W literaturze znaleźć można różne wartości progowe dla tego wskaźnika, na przykład wg Belsleya i in. (1980) wynosi ona $2p/n$, natomiast Velleman i Welsch (1981) podają $3p/n$. Obserwacje dla których wartości h_{ii} przekraczają wartość graniczną tej miary zwykło nazywać się obserwacjami o wysokim wskaźniku wpływu (*ang. high-leverage point*).

6. Miary wrażliwości

W literaturze dotyczącej diagnostyki regresji można znaleźć propozycje statystyk służących ocenie stopnia, w jakim usunięcie pojedynczej obserwacji wpływa na wartość oszacowanych parametrów i wartości przewidywanych $\hat{y}_i$,

pozostałych po usunięciu tej obserwacji, oraz macierz wariancji estymatorów parametrów. Trzy najczęściej stosowane miary ogólnego wpływu obserwacji na szacowany model regresji to tzw. odległość Cooka (C_i), tzw. odległość Welscha- Kuha (*DFFITS*) oraz tzw. zmodyfikowana wartość Cooka (C_i^*). Miary te mogą być wyrażone jako funkcje wskaźników wpływu h_{ii} oraz reszt modyfikowanych (r_i lub r_i^*).

Cook (1977) zaproponował miarę

$$C_i = \frac{1}{p}\frac{h_{ii}}{1-h_{ii}} r_i^2 ,$$

sugerując, aby wartości C_i porównywać z percentylami centralnego rozkładu F z p i $n-p$ stopniami swobody (chociaż statystyka ta w rzeczywistości nie ma rozkładu F). Obserwacje, dla których wartości statystyki Cooke'a przekraczają wartości graniczne tej miary, uznaje się za istotnie wpływające na wielkość zmian obserwowanych w wektorze ocen $\mathbf{b}$.

W pracy Belsley i in. (1980) zaproponowano miarę

$$DFFITS_i = \left|r_i^*\right| \sqrt{\frac{h_{ii}}{1-h_{ii}}} .$$

Chociaż $DFFITS_i$ nie ma dokładnie rozkładu t, niemniej jednak zachowuje się podobnie do statystyki t. Przy spełnionych założeniach modelowych $r_i^* \sim t_{(n-p-1)}$, wówczas za graniczną wartość dla $DFFITS_i$ można uznać $DF\breve{F}ITS_\alpha = t_\alpha \sqrt{p/(n-p)}$, gdzie t_α jest odczytane z tablic rozkładu $t-$Studenta, przy ustalonym poziomie istotności α i $n-p-1$ stopniami swobody (Belsley i in., 1980).

Welsch i Kuh (1977) zaproponowali miarę

$$C_i^* = \left|r_i^*\right| \sqrt{\frac{h_{ii}}{1-h_{ii}}\frac{n-p}{p}} ,$$

Chatterjee i Hadi (1986) podają dla tej statystyki wartość graniczną $\breve{C}_i^* = 2\sqrt{(n-p)/p}$.

Belsley i in. (1980) zaproponowali statystykę do badania wpływu i-tej obserwacji na macierz wariancji estymatorów parametrów modelu jako

$$COVRATIO_i = \left(\frac{s_{(i)}^2}{s^2} \right)^p \frac{1}{1-h_{ii}}.$$

Przybliżone wartości graniczne $COVRATIO_i$ spełniają nierówność $\left|COVRATIO_i - 1\right| \geq 3p/n$, a ich przekroczenie oznacza silną wpływowość danej obserwacji (Belsley i in., 1980).

Obserwacja może wykazywać wpływowość w przestrzeni o mniejszym wymiarze. W pracy Belsley i in. (1980) zaproponowano miarę będącą standaryzowaną różnicą między j-tym współczynnikiem regresji otrzymanym na podstawie wszystkich n obserwacji (b_j), a j-tym współczynnikiem regresji otrzymanym po usunięciu i-tej obserwacji ($b_{j(i)}$)

$$DFBETAS_{ij} = \frac{b_j - b_{j(i)}}{\sqrt{\sigma_{b_j}^2}} = \frac{b_j - b_{j(i)}}{s_{(i)}\sqrt{(\mathbf{X}^T\mathbf{X})_{jj}^{-1}}}, \quad j = 0,1,\ldots,p-1.$$

Dla zbioru danych o małej i umiarkowanej liczbie obserwacji za graniczną wartość uznaje się $\left| \breve{DFBETAS}_{ij} \right| = 1$, natomiast dla dużych zbiorów danych $\left| \breve{DFBETAS}_{ij} \right| = 2/\sqrt{n}$.

7. Analiza wyników badań

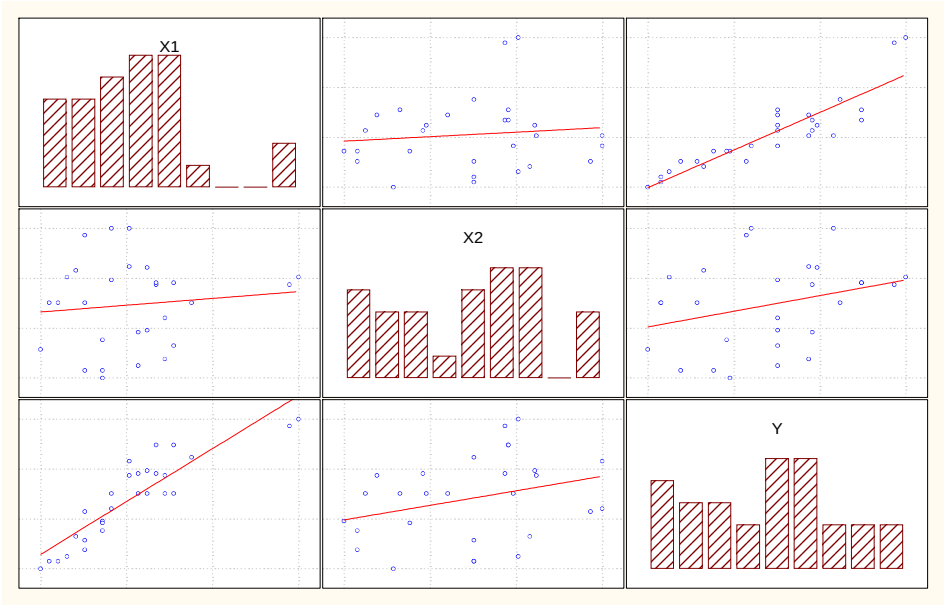
Dla rozważanego punktu pomiarowego zaproponowano następującą postać analityczną modelu opisującego zależność zapasu wody w warstwie korzeniowej gleby (Y) od rzędnej zwierciadła wody gruntowej (X_1) oraz zapasu wody zgromadzonego w tym samym miejscu w poprzednim miesiącu (X_2):

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon .$$

Otrzymano równanie regresji postaci

$$Y = 40.491 + 55.963 X_1 + 0.198 X_2 \quad R^2 = 0.842 , \; R_p^2 = 0.8295 \qquad (7.1)$$

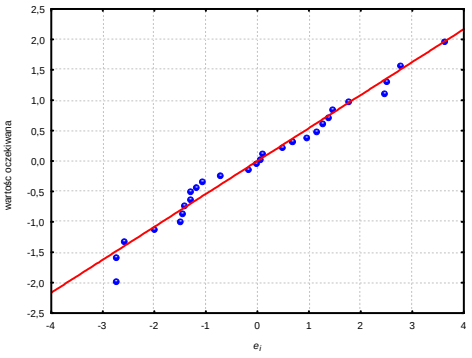
Na rys.1 przedstawiono macierzowy wykres rozrzutu danych wraz z histogramami poszczególnych zmiennych. Potwierdza on wysoką liniową zależność między zapasem wody w warstwie korzeniowej gleby (Y), a rzędną zwierciadła wody gruntowej (X_1), dodatkowo pozwala wniknąć w strukturę danych. Mianowicie, dwie obserwacje, o numerach 1 oraz 20, na wykresie rozrzutu Y względem X_1 wydają się być znacznie oddalone od pozostałych obserwacji ze zbioru danych.



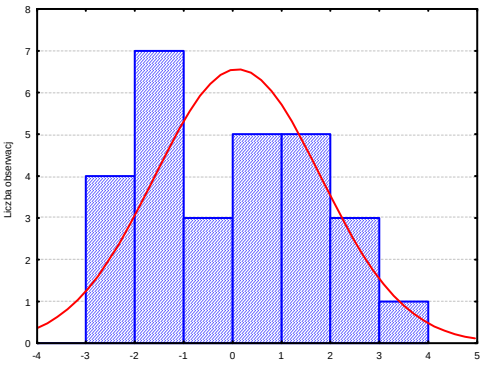
Rys. 1. Macierzowy wykres rozrzutu dla modelu (7.1)

Kolejnym krokiem było przeprowadzenie szczegółowej analizy reszt. W tym celu wykorzystane zostały zarówno testy statystyczne jak i metody graficzne (wykresy reszt). Analizę losowego charakteru reszt przeprowadzono testem Walda-Wolfowitza, stwierdzając iż nie ma podstaw do odrzucenia hipo-

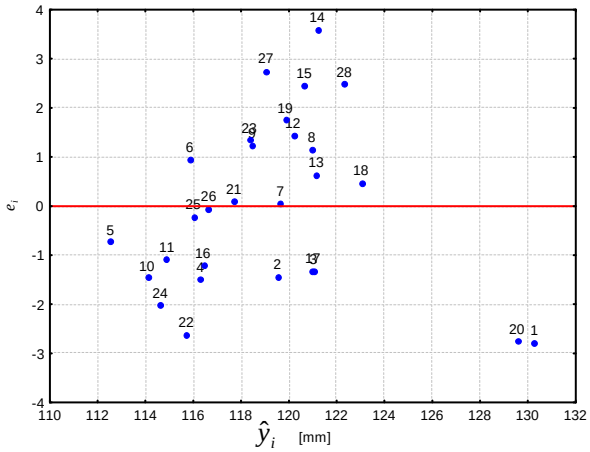
tezy o losowym rozkładzie reszt. Do zweryfikowania hipotezy o całkowitym braku zjawiska autokorelacji składników losowych rzędu pierwszego zastosowano test dwustronny Durbina-Watsona, na podstawie którego stwierdzono, iż nie ma podstaw do przyjęcia, iż zachodzi autokorelacja reszt, na deklarowanym poziomie istotności $\alpha = 0.02$. Do wstępnej oceny normalności rozkładu wykorzystano normalny wykres prawdopodobieństwa reszt (rys. *2*) oraz histogram reszt (rys. 3). Analizując oba wykresy stwierdzono, że rozkład reszt umiarkowanie odstaje od rozkładu normalnego. Wizualna ocena normalności może wydawać się mało dokładna. Formalną weryfikację hipotezy o normalności roz-



Rys. 2. Normalny wykres prawdopodobieństwa reszt dla modelu (7.1)



Rys. 3. Histogram reszt dla modelu (7.1)



Rys. 4. Wykres rozrzutu reszt względem wartości przewidywanych zmiennej zależnej dla modelu (7.1)

kładu reszt zrealizowano testem Shapiro-Wilka nie odrzucając hipotezy o normalnym rozkładzie reszt, przy poziomie istotności $\alpha = 0.05$. Do oceny stabilności wariancji reszt posłużono się wykresem rozrzutu reszt względem przewidywanych wartości zmiennej zależnej (rys. 4), na którym widoczne jest poważne naruszenie tego założenia. Wykres powinien przedstawiać równomierny pas punktów bez wyraźnej tendencji wzrostu lub spadku wariancji reszt przy wzroście wartości przewidywanych zmiennej objaśnianej.

Analiza wrażliwości pozwoliła na ocenę ex-post wpływu poszczególnych obserwacji na wyniki estymacji. W tabelach 1-2 zestawiono wartości reszt i ich modyfikacji, wartości wskaźników wpływu oraz miar wrażliwości dla wszystkich obserwacji ze zbioru danych.

Tabela 1. Wartości reszt, ich modyfikacji i wartości wskaźników wpływu dla modelu (7.1)

Numer obserwacji	e_i	r_i	r_i^*	h_{ii}
1	-2,767	-1.830	-1.927	0.315
2	-1.427	-0.851	-0.847	0.158
3	-1.319	-0.741	-0.734	0.051
4	-1.500	-0.845	-0.840	0.057
5	-0.729	-0.431	-0.424	0.144
6	0.948	0.560	0.553	0.142
7	0.058	0.032	0.032	0.052
8	1.147	0.649	0.641	0.063
9	1.244	0.700	0.692	0.053
10	-1.461	-0.844	-0.839	0.103
11	-1.090	-0.640	-0.633	0.132
12	1.434	0.832	0.827	0.110
13	0.645	0.362	0.355	0.048
14	3.605	2.024	2.169	0.049
15	2.453	1.457	1.492	0.150
16	-1.215	-0.699	-0.692	0.095
17	-1.303	-0.748	-0.741	0.091
18	0.484	0.275	0.269	0.071
19	1.747	0.988	0.988	0.064
20	-2.748	-1.777	-1.863	0.283
21	0.090	0.054	0.052	0.162
22	-2.616	-1.507	-1.549	0.097
23	1.352	0.784	0.778	0.110
24	-2.021	-1.159	-1.167	0.089
25	-0.209	-0.122	-0.120	0.122
26	-0.045	-0.025	-0.025	0.064
27	2.757	1.549	1.597	0.051
28	2.486	1.404	1.433	0.060

Tabela 2. Zestawienie miar wrażliwości dla poszczególnych obserwacji w modelu (7.1)

Nr	C_i	$DFFITS_i$	C_i^*	$COVRATIO_i$	$DFBETAS(b_0)$	$DFBETAS(b_1)$	$DFBETAS(b_2)$
1	0.513	1.307	3.772	1.071	0.726	-1.197	-0.150
2	0.045	0.368	1.062	1.229	0.241	0.092	-0.319
3	0.010	0.172	0.495	1.115	-0.011	-0.085	0.053
4	0.014	0.208	0.599	1.099	-0.050	0.128	-0.022
5	0.010	0.174	0.503	1.291	-0.134	0.122	0.076
6	0.017	0.225	0.650	1.268	0.193	-0.033	-0.188
7	0.0001	0.008	0.021	1.192	0.003	0.002	-0.004
8	0.009	0.167	0.483	1.147	-0.104	0.021	0.106
9	0.009	0.164	0.474	1.125	-0.046	-0.053	0.084
10	0.027	0.285	0.823	1.155	-0.091	0.231	-0.034
11	0.020	0.247	0.713	1.239	-0.215	0.078	0.186
12	0.028	0.291	0.840	1.167	0.150	0.117	-0.221
13	0.002	0.080	0.231	1.169	-0.037	0.025	0.029
14	0.071	0.496	1.432	0.694	-0.244	0.150	0.198
15	0.125	0.629	1.814	1.019	-0.459	-0.071	0.549
16	0.017	0.225	0.649	1.177	0.040	0.139	-0.125
17	0.018	0.236	0.680	1.162	-0.085	-0.120	0.152
18	0.001	0.075	0.216	1.206	-0.022	0.053	-0.003
19	0.022	0.260	0.750	1.072	-0.141	-0.025	0.174
20	0.416	1.172	3.383	1.051	0.598	-1.078	-0.075
21	0.0002	0.023	0.067	1.319	-0.012	-0.011	0.019
22	0.081	0.509	1.470	0.940	0.030	0.352	-0.238
23	0.025	0.274	0.790	1.178	0.190	0.043	-0.224
24	0.044	0.366	1.057	1.051	-0.112	0.284	-0.043
25	0.0006	0.045	0.129	1.285	-0.038	0.008	0.036
26	0.0001	0.007	0.019	1.208	-0.005	0.002	0.004
27	0.043	0.373	1.075	0.880	0.169	0.058	-0.204
28	0.042	0.364	1.049	0.940	-0.195	0.183	0.123

Analizując te tabele można zauważyć że:

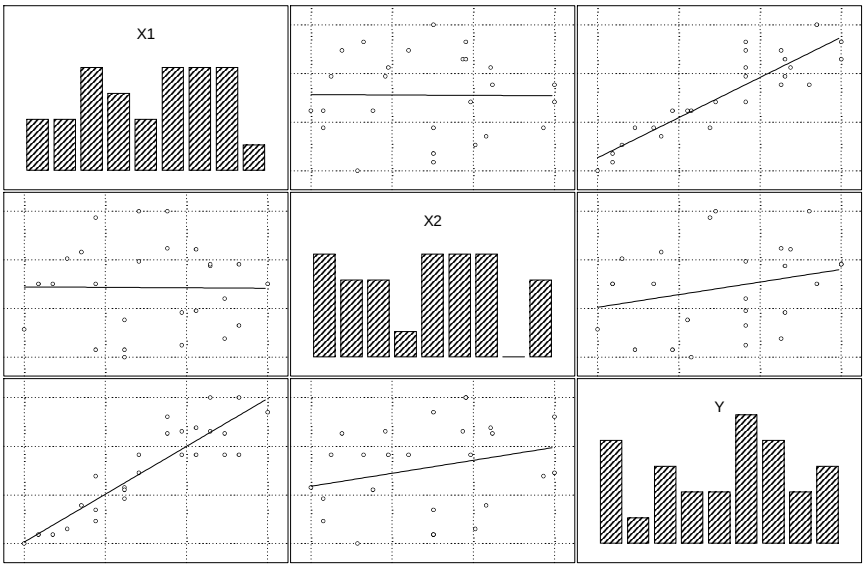
1. Dwie największe wartości wskaźnika wpływu h_{ii} otrzymano dla obserwacji o numerach 1 i 20. Przekroczyły one wartość graniczną wg Belsleya.
2. Największą wartość reszty studentyzowanej można zaobserwować dla obserwacji 14. Obserwacje 1 i 20 również osiągnęły duże wartości, jednak żadna z nich nie przekroczyła poziomu krytycznego. Ponadto, otrzymano bardzo małą wartość wskaźnika wpływu dla obserwacji 14.
3. Miary wrażliwości C_i, $DFFITS_i$, i C_i^* osiągnęły dla obserwacji 1 i 20 największe wartości, zdecydowanie odbiegające od pozostałych obser-

wacji. Oznacza to, iż usunięcie każdej z wymienionych obserwacji ma istotny wpływ na zmianę wartości oszacowań parametrów modelu. Tak wysokie wartości tych miar dla tych obserwacji są konsekwencją wystąpienia wysokich wartości h_{ii} oraz umiarkowanych wartości studentyzowanych reszt r_i.

4. Analiza wartości $DFBETAS_{ij}$ dostarczyła dodatkowych informacji. Mianowicie usunięcie zarówno obserwacji o numerze 1 jaki i 20 wywarło istotny wpływ na zmianę wartości oszacowań parametru przy zmiennej X_1.

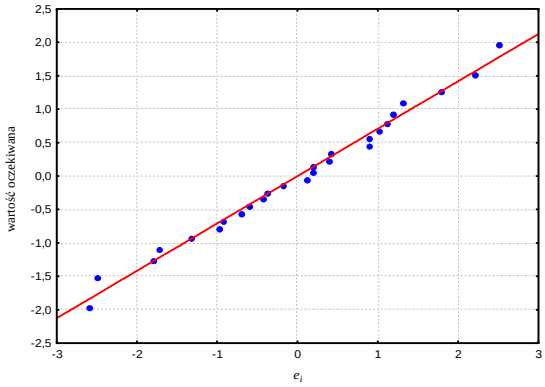
Ponieważ zostały wykryte dwie obserwacje wpływowe dokonano estymacji oraz diagnostyki modelu otrzymanego po wykluczeniu ich z analizy. Równanie regresji otrzymane po usunięciu obserwacji o numerze 1 i 20 przyjęło postać

$$Y = 18.546 + 75.317X_1 + 0.227X_2,\ R^2 = 0.884,\ R_p^2 = 0.873 \quad (7.2)$$

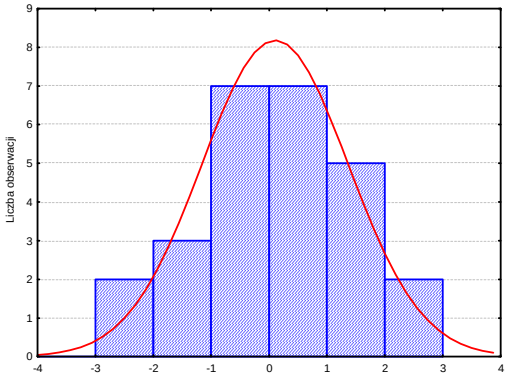


Rys. 5. Macierzowy wykres rozrzutu z histogramami dla modelu (7.2)

Macierzowy wykres rozrzutu danych (rys. 5), podobnie jak dla modelu z kompletną liczbą obserwacji (7.1), potwierdził wysoką liniową zależność między zapasem wody (Y) a rzędną zwierciadła wody (X_1), nie wskazując jednocześnie na występowanie obserwacji mających szczególny wpływ na oszacowanie modelu (7.2).



Rys. 6. Wykres normalny prawdopodobieństwa reszt dla modelu (7.2)



Rys. 7. Histogram reszt dla modelu (7.2)

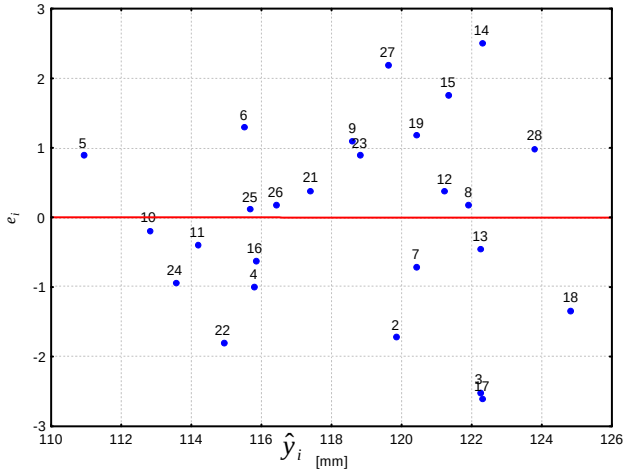
Analiza reszt nie wykazała istnienia znaczących odstępstw od założeń modelowych (rys. 6-8). Analizę regresji oraz analizę wariancji dla tego modelu przedstawiono w tabeli 3 oraz tabeli 4. Stwierdzono, iż łączny efekt oddziaływania obu zmiennych objaśniających na zmienną objaśnianą jest statystycznie istotny. Wszystkie współczynniki w równaniu (7.2) są istotne i zarówno wartość współczynnika determinacji R^2, jak i poprawionego współczynnika determinacji R^2 wzrosła w stosunku do modelu (7.1). Wysokie wartości tych współczynników należy tłumaczyć dużą jednorodnością danych.

Tabela 3. Wyniki analizy regresji dla modelu (7.2)

X_j	b_j	b.s.(b_j)	$t(b_i)$	$P\left[\lvert t\rvert > \lvert t(b_j)\rvert\right]$
w.wolny	18.546	9.204	2.015	0.0557
X_1	75.317	5.939	12.682	$7 \cdot 10^{-12}$
X_2	0.227	0.060	3.783	0.00096
$R = 0.940$	$R^2 = 0.884$	$R_p^2 = 0.873$	$s = 1.408$	

Tabela 4. Wyniki analizy wariancji dla modelu regresji (7.2)

Źródła zmienności	Sumy kwadratów	Stopnie swobody	Statystyka testowa F	*P*
Regresja Błąd	345.789 45.593	2 23	87.219	$2 \cdot 10^{-11}$
Całość	391.382			



Rys. 8. Wykres rozrzutu reszt względem wartości przewidywanych zmiennej zależnej dla modelu (7.2)

Analiza wrażliwości dla modelu (7.2) wykazała jednakowy wpływ obserwacji na wyniki estymacji (tabela 5-6). Zarówno wartości reszt studentyzowanych jak i wskaźników wpływu dla wszystkich obserwacji nie przekroczyły poziomu krytycznego. Dla obserwacji o numerze 14, uznanej za nietypową dla modelu pełnego, po usunięciu obserwacji wpływowych nie zanotowano dużych wartości reszt. Oznacza to, iż wystąpienie obserwacji o dużej reszcie w modelu (7.1) było wynikiem wpływowości obserwacji o numerach 1 i 20. Żadna z miar wrażliwości nie osiągnęła wartości granicznej potwierdzając tezę o stosunkowo równej sile oddziaływania wszystkich obserwacji na wyniki estymacji modelu (7.2).

Tabela 5. Wartości reszt, ich modyfikacji i wartości wskaźników wpływu dla modelu (7.2)

Numer obserwacji	e_i	r_i	r_i^*	h_{ii}
2	-1.719	-1.333	-1.357	0.161
3	-2.502	-1.862	-1.976	0.089
4	-0.987	-0.725	-0.717	0.065
5	0.892	0.715	0.707	0.204
6	1.301	1.000	1.000	0.146
7	-0.701	-0.515	-0.507	0.068
8	0.200	0.149	0.145	0.087
9	1.108	0.809	0.803	0.054
10	-0.175	-0.134	-0.131	0.147
11	-0.374	-0.287	-0.282	0.146
12	0.375	0.287	0.281	0.140
13	-0.444	-0.329	-0.322	0.080
14	2.510	1.860	1.974	0.081
15	1.775	1.378	1.407	0.163
16	-0.607	-0.455	-0.447	0.105
17	-2.596	-1.984	-2.131	0.136
18	-1.325	-1.026	-1.027	0.158
19	1.184	0.873	0.868	0.073
21	0.400	0.311	0.304	0.165
22	-1.794	-1.355	-1.381	0.115
23	0.893	0.675	0.667	0.116
24	-0.928	-0.703	-0.695	0.121
25	0.120	0.091	0.089	0.125
26	0.193	0.142	0.138	0.066
27	2.198	1.610	1.672	0.060
28	1.004	0.760	0.752	0.119

Tabela 6. Zestawienie miar wrażliwości dla poszczególnych obserwacji w modelu (7.2)

Nr	C_i	$DFFITS_i$	C_i^*	$COVRATIO_i$	$DFBETAS\ (b_0)$	$DFBETAS(b_1)$	$DFBETAS\ (b_2)$
2	0.114	-0.595	-1.646	1.070	0,363	0,046	-0,516
3	0.113	-0.617	-1.708	0.768	0,192	-0,452	0,101
4	0.012	-0.188	-0.522	1.140	-0,069	0,119	-0,012
5	0.046	0.369	1.021	1.359	0,31	-0,297	-0,154
6	0.057	0.413	1.144	1.171	0,327	-0,085	-0,345
7	0.006	-0.136	-0.378	1.183	-0,002	-0,07	0,055
8	0.001	0.045	0.125	1.249	-0,033	0,021	0,027
9	0.012	0.191	0.530	1.107	-0,054	-0,027	0,098
10	0.001	-0.055	-0.151	1.337	-0,028	0,047	-0,002
11	0.005	-0.116	-0.322	1.323	-0,1	0,05	0,087

12	0.004	0.113	0.314	1.314	0,015	0,066	-0,07
13	0.003	-0.095	-0.263	1.224	0,061	-0,06	-0,033
14	0.102	0.588	1.627	0.763	-0,389	0,366	0,223
15	0.123	0.621	1.718	1.054	-0,458	0,082	0,537
16	0.008	-0.153	-0.425	1.243	-0,004	0,095	-0,076
17	0.207	-0.846	-2.342	0.753	0,054	-0,595	0,393
18	0.066	-0.445	-1.233	1.180	0,253	-0,387	-0,023
19	0.020	0.244	0.676	1.114	-0,149	0,047	0,162
21	0.006	0.135	0.375	1.352	-0,047	-0,053	0,106
22	0.080	-0.499	-1.381	1.006	-0,084	0,358	-0,191
23	0.020	0.241	0.667	1.217	0,108	0,064	-0,186
24	0.023	-0.258	-0.714	1.218	-0,129	0,213	-0,011
25	0.000	0.034	0.093	1.305	0,026	-0,007	-0,027
26	0.000	0.037	0.102	1.221	0,024	-0,011	-0,021
27	0.055	0.422	1.168	0.849	0,063	0,158	-0,195
28	0.026	0.276	0.764	1.201	-0,194	0,21	0,087

8. Podsumowanie

Z przeprowadzonych rozważań wynika, że:
Równanie regresji liniowej postaci

$$Y = 18.546 + 75.317X_1 + 0.227X_2$$

w zadawalającym stopniu oddaje charakter zależności zapasu wody w warstwie korzeniowej gleby od rzędnej położenia zwierciadła wody gruntowej i może być stosowane do prognozowania.
Wykryte w trakcie analizy diagnostycznej obserwacje wpływowe wskazały na warunki, w których oszacowany związek między zmiennymi może nie znajdować potwierdzenia. Wpływowość obserwacji numer 1 spowodowana była czynnikami przyrodniczymi (opady, temperatura), natomiast obserwacji o numerze 20 zmianą kierunku przepływu gruntowego (piętrzenie wody w rzece) (Grzywna, 2003). Zebranie większej liczby danych przy obecności tych czynników pozwoli wykryć istniejące związki regresyjne między badanymi zmiennymi w tych odmiennych warunkach.

Literatura cytowana

Belsley D. A., Kuh E., Welsch R. E. (1980). *Regression diagnostics: identifying influential data and sources of collinearity*. Wiley, New York.

Chatterjee S., Hadi A. S. (1986). Influential observation, high leverage points, and outliers in linear regression. *Statistical Science* 1, 379-416.

Cook R. D. (1977). Detection of influential observations in linear regression. *Technometrics* 19, 15-18.

Cook R. D. (1993). Exploring partial residual plots. *Technometrics* 35, 351-62.

Cook R. D., Weisberg S. (1982). *Residuals and influence in regression*. New York and London, Chapman and Hall.

Chatterjee S., Hadi A. S. (1988). *Sensitivity analysis in linear regression*. New York, John Wiley & Sons.

Grzywna A. (2003). *Analiza stosunków wodno-glebowych wybranego fragmentu doliny rzeki Tyśmienicy. Praca* doktorska, AR w Lublinie.

Neter J., Kutner M. H., Nachtsheim C. J., Wasserman W. (1996). *Applied linear statistical models*. Chicago: Richard D. Irwin, Inc. and Times Mirror Higher Education Group, Inc.

Velleman P. F., Welsch R. E. (1981). Efficient computing of regression diagnostics. *The American Statistician* 35, 234-242.

Welsch R. E., Kuh E. (1977). *Linear regression diagnostics*. Technical Report 923-977, Sloan School of Management, Masachusetts Institute of Technology.

THE APPLICATION OF REGRESSION DIAGNOSTICS IN DIAGNOSIS OF WATER CONDITIONS IN SOIL

Summary

The aim of the paper is to present methods of linear regression diagnostics to determine the relations between hydrological elements on a grassland situated in river valley. The used technics allowed to make complex analysis of dependence of particular observations on estimation results, identification of observations having the influence on the linear regression results and to determine the reason of their appearance.

Key words and phrases: regression diagnostics, water conditions in soil

Classification AMS 2000: 62J05