

Korelacje nieparametryczne

Zadanie 1

Skorzystaj z opcji **Korelacje (Spearmana, tau Kendalla, gamma)** dostępnej w oknie **Statystyki nieparametryczne**

Oceń za pomocą współczynnika Spearmana współzależność między opiniami (postawionymi w punktach) wydanymi przez 2 lekarzy o zdrowiu 10 pacjentów. Dlaczego nie możemy zastosować współczynnika Pearsona?

$$r_s = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)}$$

gdzie:

d_i – różnice między rangami.

Zadanie 2

Badano związek między czasem leczenia chorych na zaburzenia układu krążenia, a aktywnością pewnego enzymu w organizmie chorych. Czy korelacja aktywności enzymu względem czasu leczenia jest istotna na poziomie istotności 0.05? Ocen na wykresie rozrzutu czy występują skrajne wartości. Jeżeli występują to dokonaj ponownych obliczeń po odrzuceniu wartości skrajnych. Napisz równanie regresji. Sprawdź jak wygląda badana zależność w podgrupach kobiety, mężczyźni.

Analiza regresji prostej

Zagadnienia:

- Klasyczny model regresji liniowej
- Metoda najmniejszych kwadratów
- Założenia regresji prostej
- Analiza regresji prostej w programie Statistica

Klasyczny model regresji liniowej

Przyjmijmy, że zbiorowość jest badana ze względu na dwie zmienne X i Y . Narzędziem badania mechanizmu powiązań między tymi zmiennymi jest funkcja regresji. Funkcja regresji to funkcja matematyczna o określonej postaci (w tym przypadku jest to funkcja liniowa), która jest przybliżeniem faktycznej zależności między zmiennymi. Szukana zmienna Y jest tradycyjnie nazywana *zmienną objaśnianą, lub zależną*. Zmienna X nazywa się *zmienną objaśniającą lub niezależną*. Jeżeli powiązanie ma postać liniową to oznacza to, że jednakowym przyrostom zmiennej niezależnej odpowiadają jednakowe (co do kierunku i siły) zmiany zmiennej zależnej. Model regresji liniowej możemy zapisać następująco:

$$Y = E(Y|X = x) = \beta_0 + \beta_1 x + \varepsilon \quad (1)$$

$E(Y | X=x)$ -wartość zmiennej Y oczekiwana przy warunku, że zmienna przyjmuje wartość x

ε - składnik losowy, wyraża wpływ wszystkich czynników, które obok X mogą wpływać na Y

β_0, β_1 szacujemy na podstawie n –elementowej próby, składającej się z par obserwacji (x_i, y_i) dla $i = 1, 2, \dots, n$. Oszacowana funkcja regresji przyjmuje postać

$$y_i = b_0 + b_1 x_i + e_i = \hat{y}_i + e_i \quad (2)$$

$i=1, 2, \dots, n$ -kolejny numer obserwacji,

$e_i = y_i - \hat{y}_i$ - tzw. reszty

Metoda najmniejszych kwadratów (MNK)

MNK polega na oszacowaniu parametrów funkcji (1) tak aby suma kwadratów reszt, opisująca rozbieżność między wartościami empirycznymi zmiennej zależnej, a jej wartościami teoretycznymi osiągnęła minimum:

$$\text{Wyrażenie } \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_1 x_i - b_0)^2 \text{ ma osiągnąć minimum}$$

y_i - wartości empiryczne zmiennej Y

$\hat{y}_i$ - wartości teoretyczne wyznaczone z (2)

b_0 i b_1 są współczynnikami regresji. b_0 to wyraz wolny (intercept) równania regresji, b_1 nazywany jest oceną współczynnika regresji liniowej (slope), mówi on o ile jednostek średnio wzrośnie lub zmaleje wartość zmiennej zależnej, gdy wartość zmiennej niezależnej zmieni się o jedną jednostkę.

Uzyskane metodą MNK oceny b_0 i b_1 (estymatory MNK) mają postać opisaną wzorami:

$$b_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}; \quad b_0 = \bar{y} - b_1 \bar{x}$$

$\bar{x}, \bar{y}$ - wartości średnie odpowiednio X, Y

Założenia regresji prostej

Estymatory otrzymane MNK powinny być efektywne i nieobciążone, w tym celu muszą być spełnione następujące warunki

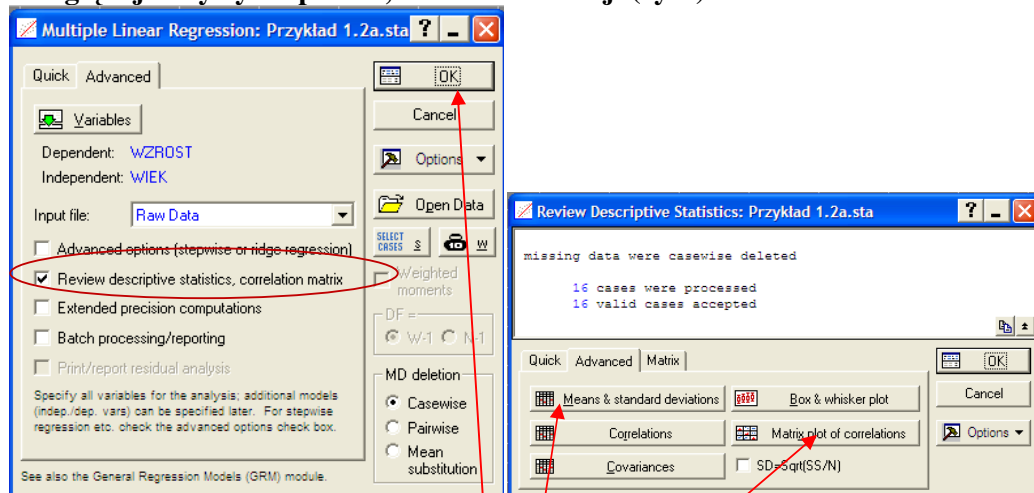
1. Model jest liniowy względem parametrów $y_i = \beta_0 + \beta_1 x_i$
W tym celu rozpoczynamy analizę regresji od przestudiowania pod kątem liniowości wykres rozrzutu
2. Liczba obserwacji n musi być większa lub równa liczbie oszacowanych parametrów b_0 i b_1 tj. $n \geq 2$
Im większa liczba obserwacji tym większa precyzja oszacowania parametrów
3. Składowik losowy e_i ma wartość oczekiwaną równą 0 dla wszystkich $i = 1, 2, \dots, n$
Tzn. czynniki nieuwzględnione w modelu nie oddziałują w istotny sposób na średnią wartość zmiennej Y, w przeciwnym wypadku parametry obciążone są błędem systematycznym
4. Wariancja reszt jest taka sama dla wszystkich obserwacji (homoscedastyczność).

- Tzn. czynniki nie ujęte w modelu mają taką samą zmienność (rozrzut) niezależnie od numeru obserwacji
5. Reszty są nieskorelowane, czyli e_i oraz e_j są od siebie niezależne dla wszystkich par i i j , gdzie $i, j = 1, 2, \dots, n$ oraz $i \neq j$
Jeżeli obserwowana wielkość zmiennej zależnej w momencie t silnie zależy od wielkości w momentach wcześniejszych to możemy mieć do czynienia z autokorelacją reszt
 6. Reszty mają rozkład normalny

Analizy regresji w pakiecie Statistica

Zadanie 3

W pliku zebrano dane dot. wieku i wzrostu 16-osobowej grupy dzieci. Wczytaj dane do pakietu Statistica. Z Menu **Statystyka** wybierz opcję **Regresja Wieloraka**. Zaznacz opcję **Przeglądaj statystyki opisowe, macierz korelacji** (rys1).



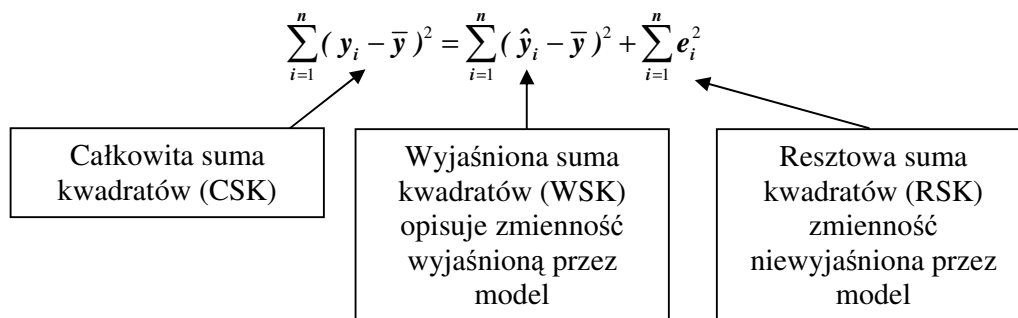
Rys. 1 Okno z opcjami statystyk opisowych

- policz średnie i odchylenia standardowe
- wykreśl macierzowy wykres rozrzutu (opcja macierzowy wykres korelacji). Uwaga MNK jest bardzo czuła na obserwacje odstające, sprawdź czy występują.
- dokonaj oceny parametrów modelu b_0 i b_1 (w tym celu odznacz opcję **Przeglądaj statystyki opisowe, macierz korelacji**) i naciśnij **OK**, a następnie wybierz **Podsumowanie: wyniki regresji**

$$\text{Wzrost} = b_1 \text{Wiek} + b_0$$

Zapisz uzyskaną postać modelu i zinterpretuj go.

- ile wynosi błąd standardowy estymacji S_e ?
- jakim błędem obarczone są wyliczone współczynniki regresji b_0 i b_1 (S_{b_0} , S_{b_1})?
- na podstawie współczynnika determinacji R^2 oceń dopasowanie modelu.



$$R^2 = \frac{WSK}{CSK}$$

CSK, WSK, RSK możesz odczytać z arkusza analizy wariancji ANOVA (**Wyniki regresji wielorakiej zakładka Zaawansowane**).

Uzupełnij zdanie

Model wyjaśnia% zaobserwowanej zmienności.

d) Ile wynosi znormalizowany współczynnik regresji Beta? Zinterpretuj uzyskany wynik. Współczynniki regresji nie możemy porównywać ze względu na różne miary, dlatego wprowadzono znormalizowane parametry równania regresji (przez odjęcie średniej arytmetycznej i podzielenie przez odchylenie standardowe). Wyraz wolny równa się 0.

W programie Statistica znormalizowane współczynniki regresji nazywają się Beta i ich interpretacja jest następująca: zmiana zmiennej niezależnej o jedno odchylenie standardowe powoduje wzrost wartości zmiennej zależnej o Beta jej odchylenia standardowego. Zaletą jest niezależność od jednostek miary zmiennej X.

e) oceń istotność parametrów modelu (test t w arkuszu wyników **Podsumowanie regresji zmiennej zależnej**)

f) Wykorzystując zbudowany model wylicz prognozę dla osoby mającej 14.5 roku. W tym celu wybierz opcję **Predykcja zmiennej zależnej** na zakładce **Reszty, Założenia, Predykcja**. Wprowadź wartość 14.5 w oknie **Określ wartość zmiennych niezależnych**. Ile wynosi prognoza wzrostu i 95% granice ufności dla warunkowej wartości oczekiwanej wzrostu?.

Zadanie 4

Dla 30 pacjentów zebrano informacje dotyczące wieku i ciśnienia skurczowego krwi. Czy jest powiązanie między ciśnieniem, a wiekiem. Przeprowadź analizę regresji prostej jak powyżej.

Analiza regresji wielorakiej

W badaniach medycznych i biologicznych obserwujemy zazwyczaj nie jedną (analiza regresji prostej), a wiele cech i ich oddziaływanie na zmienną objaśnianą. Metoda, która służy do rozwiązywania takich zagadnień nazywa się regresją wieloraką.

Założmy, że dysponujemy n obserwacjami zmiennych $x_1, x_2, \dots, x_k$, o których przypuszczamy, że wywierają wpływ na zmienną y . Liniowy model regresji wielorakiej przyjmie postać:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k + \varepsilon$$

β_k – parametry modelu (współczynniki regresji),

ε – składnik losowy,

Założenia regresji wielorakiej

1. Model jest liniowy względem parametrów $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$
dla $i = 1, 2, \dots, n$
W tym celu rozpoczynamy analizę regresji od przestudiowania pod kątem liniowości wykres rozrzutu
2. Liczba obserwacji n musi być większa lub równa liczbie oszacowanych parametrów tj. $n \geq k+2$
Im większa liczba obserwacji tym większa precyzja oszacowania parametrów
3. Żadna ze zmiennych niezależnych nie jest kombinacją liniową innych zmiennych niezależnych – założenie o braku współliniowości
4. Składnik losowy e_i ma wartość oczekiwaną równą 0 dla wszystkich $i = 1, 2, \dots, n$
Tzn. czynniki nieuwzględnione w modelu nie oddziałują w istotny sposób na średnią wartość zmiennej Y , w przeciwnym wypadku parametry obciążone są błędem systematycznym
5. Wariancja reszt jest taka sama dla wszystkich obserwacji (homoscedastyczność).
Tzn. czynniki nie ujęte w modelu mają taką samą zmienność (rozrzut) niezależnie od numeru obserwacji
6. Reszty są nieskorelowane, czyli e_i oraz e_j są od siebie niezależne dla wszystkich par i i j , gdzie $i, j = 1, 2, \dots, n$ oraz $i \neq j$
Jeżeli obserwowana wielkość zmiennej zależnej w momencie t silnie zależy od wielkości w momentach wcześniejszych to możemy mieć do czynienia z autokorelacją reszt
7. Reszty mają rozkład normalny

Zadanie 5

W grupie 12 dzieci zebrano dane dotyczące wieku, wzrostu (cale) i wagi (funty). Za pomocą analizy regresji wielorakiej przeprowadź analizę wpływu wieku i wzrostu na wagę.

$$Waga = \beta_0 + \beta_1 \text{Wiek} + \beta_2 \text{Wzrost} + \varepsilon$$

Analiza Reszt

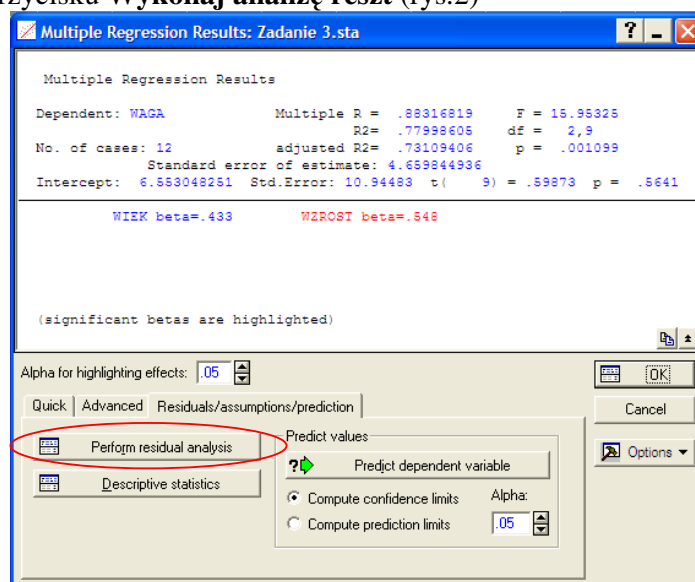
Zagadnienia

- Metody weryfikacji założeń analizy regresji
- Statystyka Durbin-Watsona
- Statystyki związane z resztami

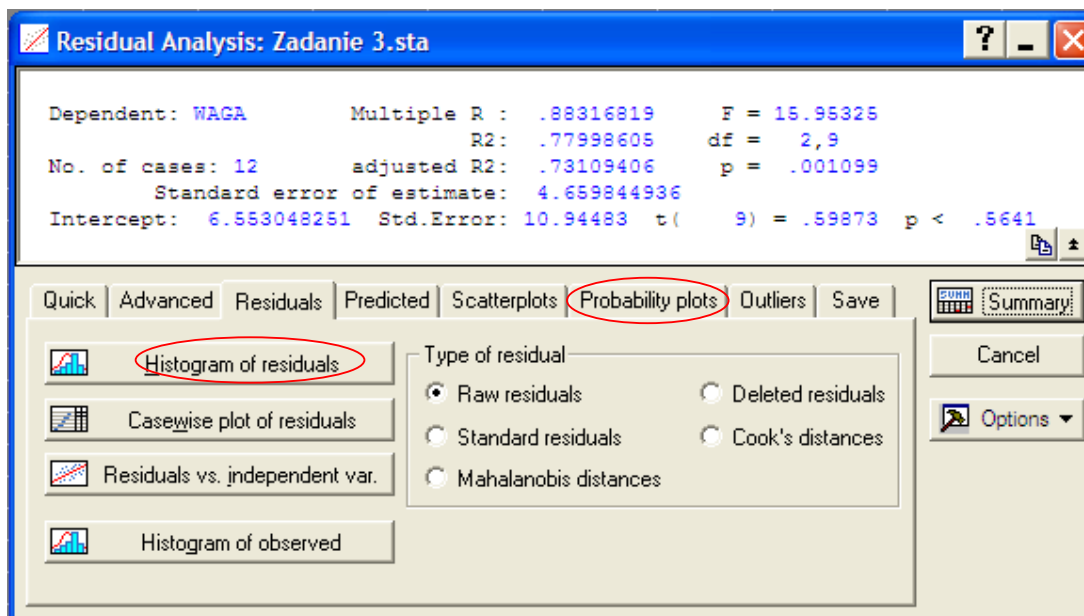
Większość założeń analizy regresji dotyczy reszt. Reszta losowa jest to różnica między wartościami obserwowanymi a teoretycznymi zmiennej losowej Y . Poprawnie skonstruowany model to taki, który posiada pożądane własności reszt (patrz założenia analizy regresji powyżej):

- a) normalność rozkładu reszt
- b) brak autokorelacji reszt
- c) stałość wariancji reszt

W programie Statistica zestaw statystyk i wykresów potrzebnych do analizy reszt jest dostępnych na karcie **Reszty, założenia, predykcja** w oknie **Wyniki regresji wielorakiej** po naciśnięciu przycisku **Wykonaj analizę reszt** (rys.2)



Rys.2 Okno **Wyniki regresji wielorakiej**, karta **Reszty, założenia, predykcja**

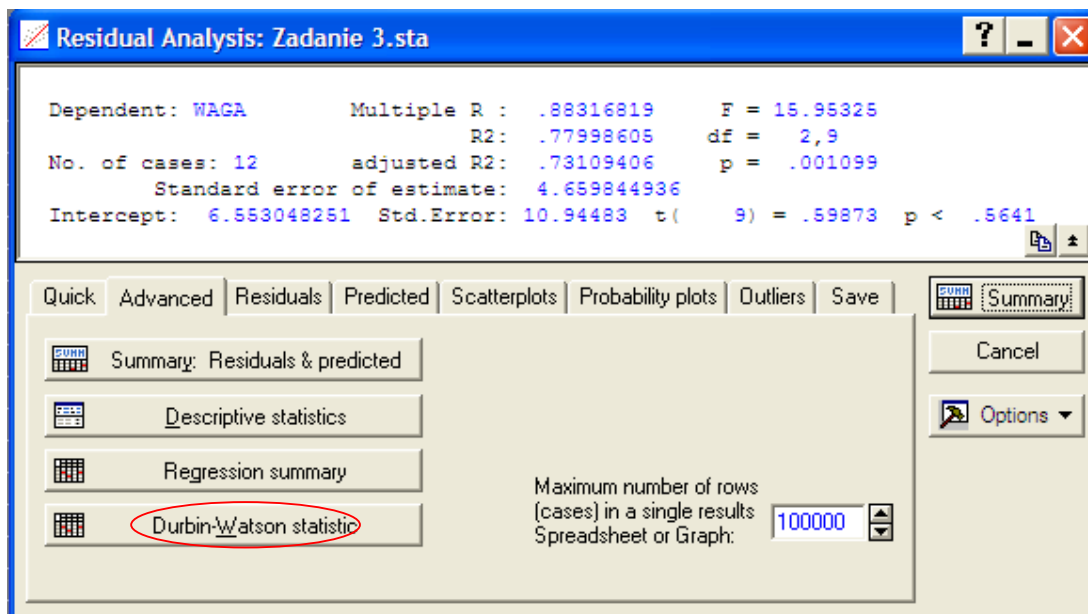


Rys.3 Okno Analiza reszt

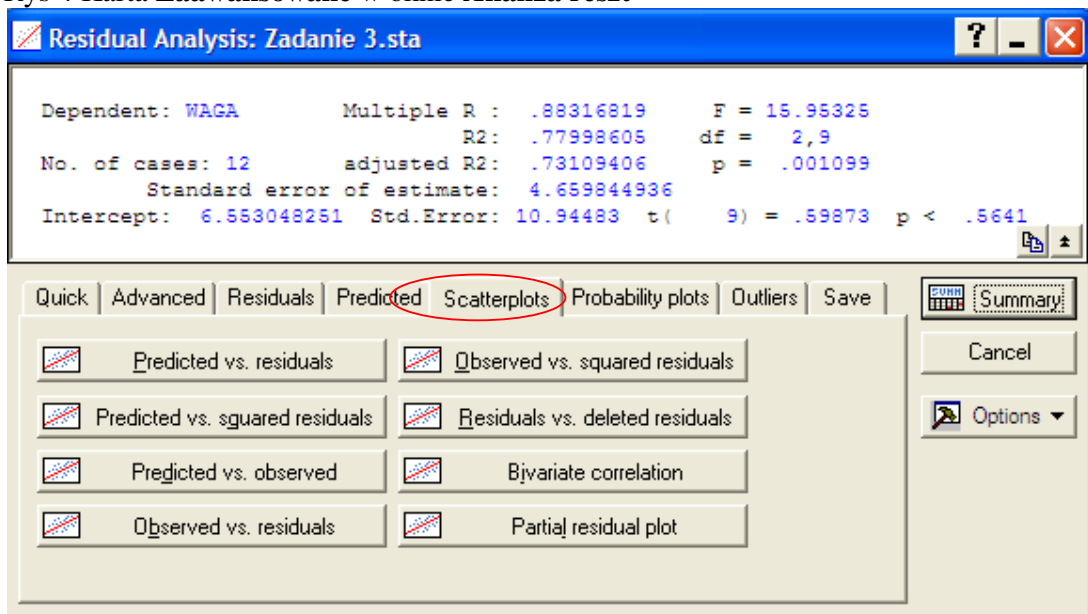
a) Zaczniemy od założenia: **reszty modelu mają rozkład normalny**. Założenie jest istotne do weryfikacji istotności otrzymanych parametrów modelu. Na karcie zatytułowanej **Wykresy prawdopodobieństwa** w oknie **Analiza reszt** (rys 3) możesz dokonać graficznego sprawdzenia normalności (wykres normalności reszt). Jeżeli reszty nie mają rozkładu normalnego to punkty będą się odchyłać od linii prostej. Wykres normalności reszt pozwala na wykrycie punktów odstających. Możesz też wykreślić histogram reszt na karcie **Reszty** (rys3.). Jeżeli nie masz pewności co do interpretacji graficznej zastosuj test normalności Shapiro-Wilka, dostępny w module **Statystyki podstawowe i tabele**.

b) Następne założenie mówi o **braku autokorelacji reszt**. Autoregulacja reszt jest zjawiskiem niepożądanym i dotyczy najczęściej szeregów czasowych lub przestrzennych i jest definiowana jako korelacja pomiędzy elementami tego samego ciągu obserwacji, uporządkowanymi czasowo lub przestrzennie. Do weryfikacji tego założenia służy test Durбина-Watsona (DW), który jest dostępny na karcie **Zaawansowane** w oknie **Analiza reszt** (rys.4). Otrzymujemy wartość testu d i estymator współczynnika autokorelacji ($d=2(1-\text{Seryjna Kor.})$). Wartość statystyki d należy do przedziału $<0,4>$, Wartość d bliska 0 wskazuje na autokorelację dodatnią, a bliska 4 na autokorelację ujemną. Wartość d bliska 2 na brak autokorelacji.

c) **stałość wariancji reszt** (homoscedastyczność) można sprawdzić za pomocą wykresu rozrzutu reszt. Wykres rozrzutu reszt względem wartości przewidywanych lub rozrzutu wartości przewidywanych względem kwadratów reszt można znaleźć na zakładce **Wykresy rozrzutu** w oknie **Analiza reszt** (rys 5). Reszty powinny być równomiernie rozproszone, nie wykazywać różnic w rozproszeniu wyników dla różnych wartości przewidywanych. W przypadku stwierdzenia niespełnienia tego założenia można dokonać takiej transformacji modelu, by otrzymać model homoscedastyczny lub zastosować ważoną metodę najmniejszych kwadratów



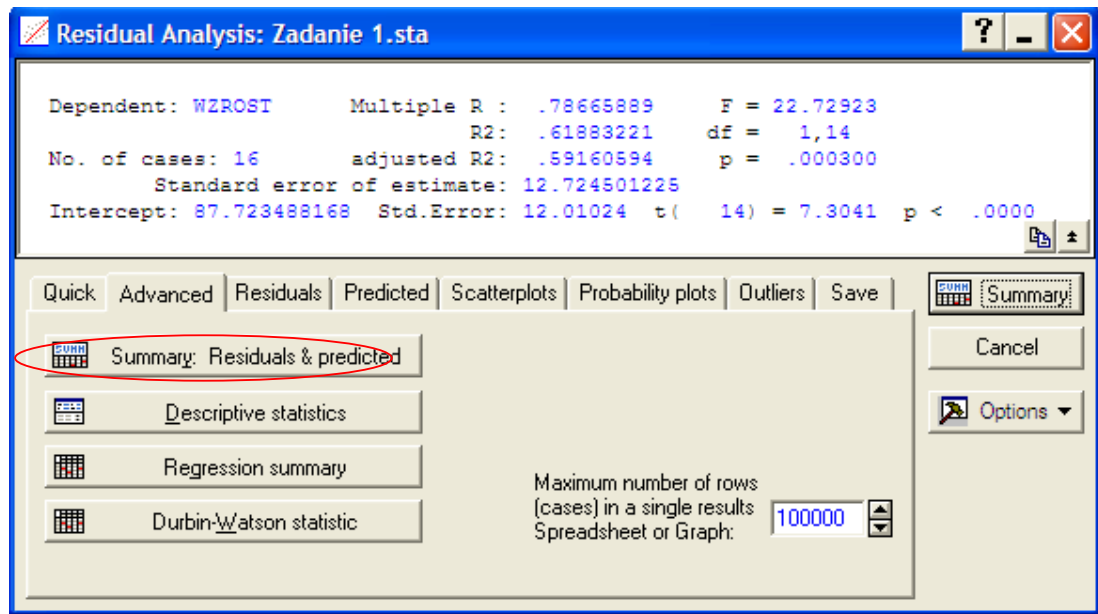
Rys 4 Karta **Zaawansowane** w oknie **Analiza reszt**



Rys 5. Karta **Wykresy rozrzutu** w oknie **Analiza reszt**

Obserwacje nietypowe w analizie regresji

W pakiecie STATISTICA istnieje wiele narzędzi (statystyk i wykresów) do wykrywania odstających punktów. Przede wszystkim wykorzystamy w tym celu analizę reszty modelu. W oknie **Analiza reszt** na karcie **Zaawansowane** po naciśnięciu przycisku **Podsumowanie: reszty i przewidywane** otwiera się okno oferujące różne narzędzia, za pomocą których będziemy szukać przypadków o szczególnie dużych resztach (Rys6).



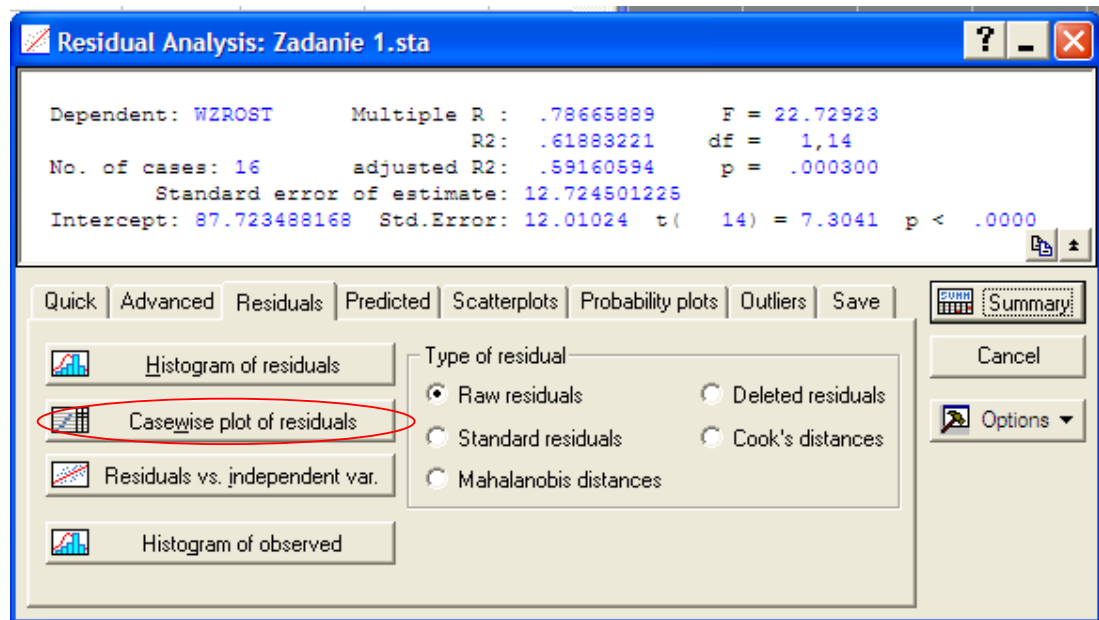
1	2	3	4	5	6	7	8	9
Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance

Rys. 6 Analiza wartości przewidywanych i reszt

1. Wartości obserwowanej zmiennej zależnej
2. Wartości przewidywane wyliczone z równania regresji
3. Wartości reszt
4. Standaryzowana wartość przewidywana
5. Standaryzowana wartość resztowa (wartość resztowa podzielona przez pierwiastek kwadratowy z średniego kwadratu reszt)
6. Błędy standardowe niestandaryzowanej wartości przewidywanej
7. **Odległość Mahalanobisa.** Jest to uogólnienie odległości euklidesowej uwzględniające skorelowanie zmiennych. Jeżeli bowiem zmienne niezależne nie są powiązane, to odległość Mahalanobisa jest równa odległości euklidesowej. Miara ta pozwala ustalić, czy daną obserwację można zaliczyć do odstających. Dzięki niej możemy "wyłapać" obserwacje potencjalnie groźne ze względu na bardzo dużą odległość od pozostałych. Nie zawsze możemy to stwierdzić, przeglądając wartości reszt lub reszt standaryzowanych, gdyż obserwacje takie silnie "przyciągają" linię regresji i przez to mogą mieć małe reszty.
8. **Usunięte wartości resztowe.** Jest to reszta dla danej obserwacji, obliczana tak, jakby ten przypadek nie był włączony do analizy regresji (tzn. równanie regresji jest wyliczone na podstawie danych z próby z wyjątkiem tej obserwacji). Jeśli reszta usunięta różni się znacznie od reszty standaryzowanej, oznacza to, że dany przypadek jest przypadkiem odstającym, ponieważ jego usunięcie istotnie zmieniło równanie regresji.
9. **Odległość Cooka.** Jest to miara stopnia zmiany współczynników regresji, gdyby dany przypadek pominąć w obliczeniach współczynników. O ile odległości Mahalanobisa mierzą odległość przypadku od środka ciężkości wyznaczonego przez zmienne

niezależne, a reszty standaryzowane - od linii regresji, o tyle odległości Cooka łączą te dwie odległości i przez to są łączną miarą wpływu poszczególnych obserwacji na linię regresji.

Graficznie wartości reszt i statystki można przedstawić na **Wykresie reszt wg przypadków** z **okna Analiza reszt** (Rys. 7) i na wykresie rozrzutu reszt względem reszt usuniętych wybierając ten wykres na karcie **Wykresy rozrzutu** (rys. 5).



Rys 7. Wykres reszt według przypadków

Zadanie 6

Przeprowadź analizę reszt i analizę obserwacji nietypowych dla danych z zadania 3 (przed i po usunięciu obserwacji nietypowych)

Zadanie 7

W pewnym eksperymencie farmakologicznym analizowano oddziaływanie pewnego preparatu na zmianę średniej objętości krwinki czerwonej (MCV). Przeprowadź analizę obserwacji odstających na podstawie analizy reszt.