Analiza regresji – sprawdzanie założeń

W analizie regresji, podobnie jak w analizie wariancji, w celu sprawdzenia założeń wykonywana jest ***analiza reszt***. W analizie badane są:

- *normalność*,
- *homoscedastyczność* (stałość wariancji błędu dla poszczególnych wartości zmiennej niezależnej),
- *niezależność*

wartości resztowych.

W analizie reszt analizowane mogą być:

- reszty,
- standaryzowane reszty,
- studentyzowane reszty,
- reszty usunięte,
- studentyzowane reszty usunięte,
- ...

Wektor reszt $\hat{\mathbf{e}}$

Reszty reprezentują różnice pomiędzy wartościami obserwowanymi y_i a wartościami otrzymywanymi z wykorzystywanego modelu $\hat{y}_i$, tzn. wyznaczane są jako:

$$\hat{e}_i = y_i - \hat{y}_i.$$

W postaci macierzowej wektor reszt $\hat{\mathbf{e}}$ zapisuje się jako:

$$\hat{\mathbf{e}} = \mathbf{y} - \hat{\mathbf{y}}.$$

Zapisując aproksymowane funkcją regresji wartości wyjść obiektu $\hat{\mathbf{y}}$ w postaci:

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\mathbf{b}} \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \\ &= \mathbf{H}\mathbf{y}, \end{aligned}$$

można wektor reszt opisać jako:

$$\begin{aligned} \hat{\mathbf{e}} &= \mathbf{y} - \hat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

gdzie: $\mathbf{H} = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T$ to tzw. *macierz kapeluszowa* (od *ang. for putting hat on*, macierz nakłada kapelusz na $\mathbf{y}$)

***Własności macierzy* H**

Symetria ($\mathbf{H} = \mathbf{H}^T$)

$$\begin{aligned}\mathbf{H}^T &= \left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)^T \\ &= \mathbf{X}\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^T \\ &= \mathbf{X}\left(\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right)^T\mathbf{X}^T \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ &= \mathbf{H}.\end{aligned}$$

Idempodentność ($\mathbf{H}\,\mathbf{H} = \mathbf{H}$)

$$\begin{aligned}\mathbf{H}\cdot\mathbf{H} &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\cdot\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ &= \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T \\ &= \mathbf{H}.\end{aligned}$$

***Własności macierzy* I – H**

Symetria

macierz **H** jest symetryczna, macierz **I – H** ma poza główną przekątną wartości przeciwne do macierzy **H**, też jest więc symetryczna

Idempodentność

$$\begin{aligned}(\mathbf{I}-\mathbf{H})\cdot(\mathbf{I}-\mathbf{H}) &= \\ &= \mathbf{I}-2\mathbf{H}-\mathbf{H}\cdot\mathbf{H} \\ &= \mathbf{I}-2\mathbf{H}-\mathbf{H} \\ &= \mathbf{I}-\mathbf{H}.\end{aligned}$$

***Własności wektora reszt* $\hat{\mathbf{e}}$**

Wartość oczekiwana

$$E[\hat{\mathbf{e}}] = E[(\mathbf{I}-\mathbf{H})\mathbf{y}] = (\mathbf{I}-\mathbf{H})\cdot E[\mathbf{y}] = (\mathbf{I}-\mathbf{H})\cdot \mathbf{Xb} = \mathbf{Xb} - \mathbf{HXb} =$$

$$= \mathbf{Xb} - \mathbf{X}\underbrace{\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}}_{=\mathbf{I}}\mathbf{b} = \mathbf{Xb} - \mathbf{Xb} = \mathbf{0}.$$

Wariancja

$$D^2[\hat{\mathbf{e}}] = D^2[(\mathbf{I}-\mathbf{H})\mathbf{y}] = (\mathbf{I}-\mathbf{H})\cdot D^2[\mathbf{y}]\cdot(\mathbf{I}-\mathbf{H})^T = (\mathbf{I}-\mathbf{H})\cdot\sigma^2\mathbf{I}\cdot(\mathbf{I}-\mathbf{H})^T =$$

$$= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H})^T = \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \qquad \textit{(z symetrii macierzy } \mathbf{I}-\mathbf{H}\textit{)}$$

$$= \sigma^2(\mathbf{I}-\mathbf{H})(\mathbf{I}-\mathbf{H}) = \sigma^2(\mathbf{I}-\mathbf{H}). \qquad \textit{(z idempodentności } \mathbf{I}-\mathbf{H}\textit{)}$$

Własności macierzy H cd.

h_{ii} – element z głównej przekątnej macierzy **H**, tzw. ***przełożenie*** lub ***współczynnik dźwignięcia*** (*ang. leverage*), jest miarą odległości *i–tej* obserwacji od centrum zdefiniowanego przez macierz wejść **X**.

Interpretację współczynnika h_{ii} można łatwo pokazać dla funkcji regresji zawierającej wyraz wolny i jedną zmienną niezależną, niech:

$$SS_x = \sum (x_i - \bar{x})^2 = \sum x_i^2 - 2\sum \bar{x} x_i + \sum \bar{x}^2 = \sum x_i^2 - 2\bar{x}\sum x_i + n\bar{x}^2 = \sum x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 = \sum x_i^2 - n\bar{x}^2,$$

wtedy:

$$\left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \left(\begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}\begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}\right)^{-1} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}^{-1} = \begin{bmatrix} n & n\bar{x} \\ n\bar{x} & \sum x_i^2 \end{bmatrix}^{-1} = \frac{1}{SS_x}\begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix},$$

więc:

$$h_{ii} = \begin{bmatrix} 1 & x_i \end{bmatrix}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\begin{bmatrix} 1 \\ x_i \end{bmatrix} = \frac{1}{SS_x}\begin{bmatrix} 1 & x_i \end{bmatrix}\begin{bmatrix} \frac{1}{n}\sum x_i^2 & -\bar{x} \\ -\bar{x} & 1 \end{bmatrix}\begin{bmatrix} 1 \\ x_i \end{bmatrix} = \frac{1}{SS_x}\left(\frac{1}{n}\sum x_i^2 + x_i^2 - 2x_i\bar{x} + \bar{x}^2 - \bar{x}^2\right) =$$

$$= \frac{1}{SS_x}\left(\frac{1}{n}\sum x_i^2 - \bar{x}^2 + (x_i - \bar{x})^2\right) = \frac{\frac{1}{n}\sum x_i^2 - \bar{x}^2}{SS_x} + \frac{(x_i - \bar{x})^2}{SS_x} = \frac{\frac{1}{n}SS_x}{SS_x} + \frac{1}{SS_x}(x_i - \bar{x})^2 = \frac{1}{n} + \frac{1}{SS_x}\underbrace{(x_i - \bar{x})^2}$$

odległość i-tej obserwacji od centrum określonego przez macierz wejść

Własności współczynnika dźwignięcia

własność 1:

$$0 \le h_{ii} \le 1.$$

Wcześniej pokazano, że macierz **H** jest idempodentna (**H = H H**), współczynnik h_{ii} spełnia więc zależność:

$$h_{ii} = \sum_{j=1}^{n} h_{ij}\, h_{ji},$$

ze względu na symetrię macierzy (**H = H**^{T}), zależność tą można przepisać jako:

$$h_{ii} = \sum_{j=1}^{n} h_{ij}\, h_{ij} = \sum_{j=1}^{n} h_{ij}^2 = h_{ii}^2 + \underbrace{\sum_{j \ne i}^{n} h_{ij}^2}_{\ge 0}$$

więc:

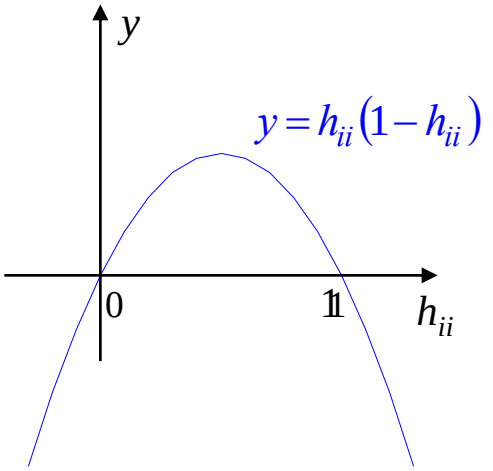
$$h_{ii} - h_{ii}^2 \ge 0,$$

czyli:

$$h_{ii}\left(1 - h_{ii}\right) \ge 0,$$

ostatecznie:

$$0 \le h_{ii} \le 1.$$



Własności współczynnika dźwignięcia cd.

własność 2:

$$\sum h_{ii} = p+1,$$

gdzie: $p+1$ – liczba współczynników równania regresji.

Suma elementów z głównej przekątnej macierzy kwadratowej to jej ślad:

$$\sum h_{ii} = tr(\mathbf{H}) = tr\Big(\underbrace{\mathbf{X}}_{\mathbf{A}}\underbrace{\big(\mathbf{X}^T\mathbf{X}\big)^{-1}}_{\mathbf{B}}\underbrace{\mathbf{X}^T}_{\mathbf{C}}\Big),$$

z własności śladu: $tr(\mathbf{ABC}) = tr(\mathbf{CAB}),$

$$\sum h_{ii} = tr\Big(\mathbf{X}^T\mathbf{X}\big(\mathbf{X}^T\mathbf{X}\big)^{-1}\Big) = tr\underbrace{(\mathbf{I})}_{\text{macierz jednostkowa o wymiarze } (p+1)\times(p+1)} = p+1.$$

Uwagi!

1. Średnia wartość współczynnika dźwignięcia wynosi: $\bar{h} = \frac{p+1}{n}$.

2. Obserwację dla której współczynnik dźwignięcia ma wartość większą niż $2\bar{h}$ *uznaje się za odległą.*

W ***analizie reszt*** analizowane mogą być:

- ***reszty*** *(ang. residuals), reszty surowe*

$$\hat{e}_i = y_i - \hat{y}_i,$$

- ***reszty standaryzowane*** *(ang. standarized residuals)*

$$d_i = \frac{\hat{e}_i}{\hat{\sigma}},$$

gdzie: $\hat{\sigma} = \sqrt{MS_e}$.

Uwagi!

1. Standaryzacja jest próbą doprowadzenia rozkładu reszt do rozkładu o wariancji równej 1, *przy założeniu, że estymatorem wariancji reszt jest tzw. błąd standardowy estymacji* MS_e, *(pokazuje się, że* $E[MS_e] = \sigma^2$*).*

2. Ze slajdu 6. wynika jednak, że wariancja reszty $\hat{e}_i$ *wynosi* $\sigma^2 (1 - h_{ii})$. *Dodatkowo, ze względu na to, że:* $h_{ii} \leq 1$, *reszty standaryzowane są przeszacowane, zaleca się analizę reszt studentyzowanych.*

- ***reszty studentyzowane***, *reszty wewnętrznie studentyzowane*
  (ang. studentized residuals lub internally studentized residuals)

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}},$$

gdzie: $\hat{\sigma} = \sqrt{MS_e}$, h_{ii} – *współczynnik dźwignięcia.*

- ***studentyzowane reszty usunięte***, *reszty zewnętrznie studentyzowane*
  (ang. extrenally studentized residuals lub R–Student)

Standaryzacja wykonywana w resztach studentyzowanych wykorzystuje oszacowaną wartość wariancji błędów losowych $\hat{\sigma}$ wyznaczoną ze wszystkich obserwacji, w celu uniezależnienia szacowania $\hat{\sigma}$ od *i–tej* obserwacji obserwację tą wyłącza się z szacowania i wyznacza się $\hat{\sigma}_{(-i)}$, ostatecznie:

$$t_i = \frac{\hat{e}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}},$$

gdzie: $\hat{\sigma}^2_{(-i)} = \dfrac{(n-p-1)MS_e - \hat{e}_i^2/(1-h_{ii})}{n-p-2}.$

- ***reszty usunięte*** *(ang. deleted residuals, PRESS residuals)*

  Podczas wyznaczania ***reszt usuniętych***, resztę dla każdej obserwacji wyznacza się wyłączając obserwację z analizy regresji a wartość reszty oblicza się na podstawie przewidywanej (na podstawie znalezionej nowej funkcji regresji) wartości zmiennej wyjściowej:

$$\hat{e}_{(-i)} = y_i - \hat{y}_{(-i)},$$

gdzie: $\hat{e}_{(-i)}$ *– reszta usunięta,* $\hat{y}_{(-i)}$ *– prognozowana wartość zmiennej wyjściowej wyznaczona z równania regresji otrzymanego bez uwzględniania* i*–tej obserwacji,*

Pokazuje się, że wartość reszty usuniętej można otrzymać bez wyznaczania nowego równania regresji:

$$\hat{e}_{(-i)} = \frac{\hat{e}_i}{1 - h_{ii}},$$

Uwagi!

1. Reszty usunięte wykorzystywane są do obliczenia prognozowanej sumy kwadratów błędów (ang. prediction error sum of squares): $PRESS = \Sigma e^2_{(-i)}$.

2. Standaryzacja reszt $\hat{e}_{(-i)}$ *, wariancja reszty* $\hat{e}_{(-i)}$ *wynosi* $\sigma^2/(1 - h_{ii})$*, daje reszty studentyzowane:*

$$\frac{\hat{e}_{(-i)}}{\hat{\sigma}/\sqrt{1-h_{ii}}} = \frac{\hat{e}_i/(1-h_{ii})}{\hat{\sigma}/\sqrt{1-h_{ii}}} = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}} = r_i.$$

Odległość Cook-a

Odległość Cook-a dla *i–tej* obserwacji jest miarą określającą wielkość zmian współczynników równania regresji w przypadku, gdy *i–ta* obserwacja jest usuwana.

Idea, odległość można wyznaczyć licząc:

$$D_i \propto \left(\hat{b}_0 - \hat{b}_{0(-i)}\right)^2 + \left(\hat{b}_1 - \hat{b}_{1(-i)}\right)^2 + \ldots + \left(\hat{b}_p - \hat{b}_{p(-i)}\right)^2 = \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(-i)}\right)^T \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(-i)}\right)$$

gdzie: $\hat{b}_{0(-i)}, \hat{b}_{1(-i)}, \ldots, \hat{b}_{p(-i)}$ – współczynniki równania regresji po usunięciu *i–tej* obserwacji.

Cook zaproponował ważenie sumowanych składników przy pomocy wyrażenia:

$$\frac{\mathbf{X}^T\mathbf{X}}{(p+1)\,\hat{\sigma}^2}.$$

Ostatecznie odległość Cook-a jest definiowana jako:

$$D_i = \frac{\left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(-i)}\right)^T \mathbf{X}^T\mathbf{X} \left(\hat{\mathbf{b}} - \hat{\mathbf{b}}_{(-i)}\right)}{(p+1)\,\hat{\sigma}^2}.$$

Pokazuje się również, że odległość tą można wyznaczyć z wzoru:

$$D_i = \frac{r_i^2}{(p+1)} \frac{h_{ii}}{(1-h_{ii})}.$$

Uwaga! D_i *może być duże ze względu na dużą wartość: a) reszty* r_i *i b) współczynnika dźwignięcia* h_{ii}.

Analiza reszt – podsumowanie

Wskaźniki	*Definicja*	*Uwagi*
reszty surowe	$\hat{e}_i = y_i - \hat{y}_i$	$\mathcal{N}\left(0, \sigma\sqrt{1-h_{ii}}\right)$
reszty standaryzowane	$d_i = \frac{\hat{e}_i}{\hat{\sigma}},$	$\approx t_{n-p-1}$
reszty usunięte	$\hat{e}_{(-i)} = y_i - \hat{y}_{(-i)}$	$\mathcal{N}\left(0, \sigma/\sqrt{1-h_{ii}}\right)$
studentyzowane wewnętrznie	$r_i = \hat{e}_i / \left(\hat{\sigma}\sqrt{1-h_{ii}}\right)$	$\approx t_{n-p-1}$
studentyzowane zewnętrznie	$t_i = \hat{e}_i / \left(\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}\right)$	t_{n-p-2}
współczynnik dźwignięcia	h_{ii} – i-ty element przekątnej macierzy **H**	
odległość Cook-a	$D_i = \frac{r_i^2}{(p+1)} \frac{h_{ii}}{(1-h_{ii})}$	

W *analizie reszt* szczególną uwagę poświęca się:

- ***obserwacjom odstającym*** *(ang. outliers)*

  za odstającą uznawana jest obserwacja, która ma dużą resztę, tzn. obserwowana wartość zmiennej zależnej mocno odbiega od wartości prognozowanej z modelu regresji,

  przyjmuje się, że reszta jest duża jeśli jej wartość przekracza 3 odchylenia standardowe,

- ***obserwacjom o dużej dźwigni*** *(ang. leverage points)*

  obserwacje takie mają dużą wartość *współczynnika dźwignięci*a h_{ii},

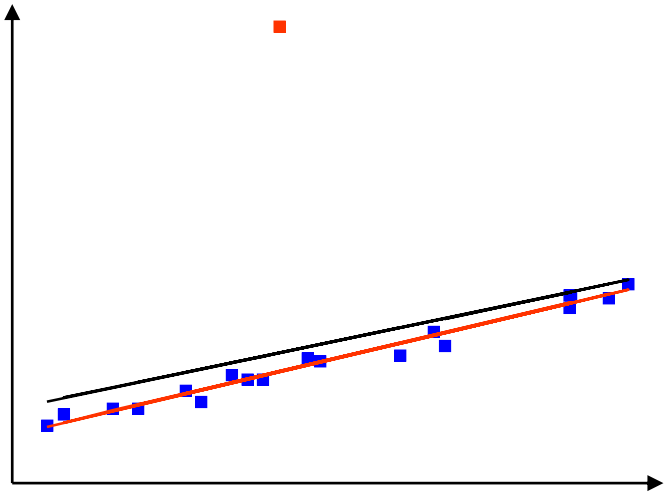
  przyjmuje się, że współczynnik dźwignięcia jest duży jest przekracza wartość $2\bar{h} = 2\frac{p+1}{n}$,

- ***obserwacjom wpływowym*** *(ang. influential observations)*

  obserwacje, które mocno wpływają na postać funkcji regresji, poszukiwanie obserwacji wpływowych sprowadza się do wyznaczania modelu z obserwacją i bez obserwacji,

  obserwacje wpływowe mają dużą wartość reszty usuniętej i dużą odległość Cook-a (np. > 1).

Uwaga! Nawet niezbyt duża reszta obserwacji o dużej dźwigni może powodować, że obserwacja taka staje się obserwacją wpływową.

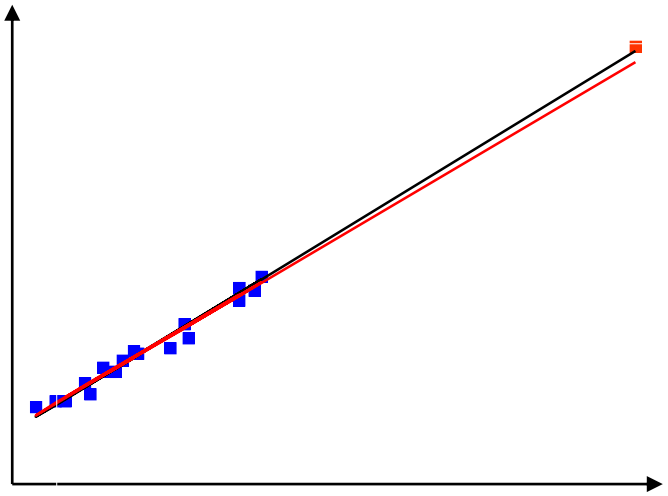


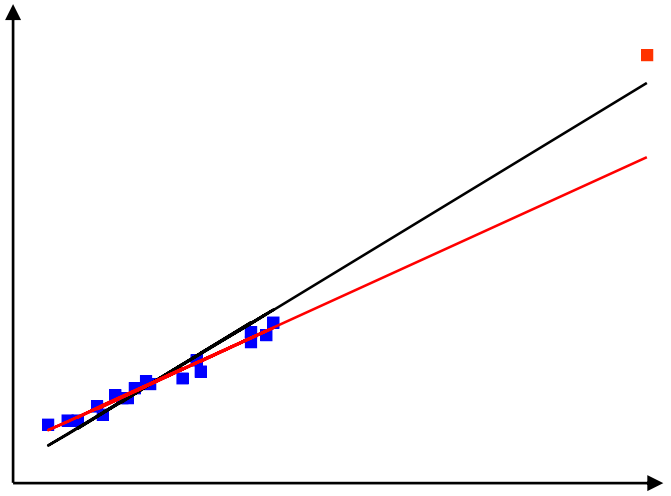
wyróżniona obserwacja

funkcja regresji

funkcja regresji po usunięciu obserwacji

Obserwacja o małej dźwigni i dużej reszcie nie wpływa mocno na funkcję regresji





Obserwacja o dużej dźwigni i niewielkiej reszcie wpływa mocno na funkcję regresji

Analiza reszt pozwala na weryfikację założeń *analizy regresji*.

Normalność rozkładu reszt

- analiza graficzna:
  - histogramy reszt
  - wykresy normalności
- weryfikacja testów zgodności

Niezależność reszt

- test Durbin-Watsona (testowana jest hipoteza zerowa o zerowej korelacji pomiędzy kolejnymi resztami), *rozkład wykorzystywany w teście jest nietypowy i wartości krytyczne muszą być odczytywane z tablic*

Homoscedastyczność reszt (*stałość wariancji reszt*)

- *analiza graficzna:*
  - wykres rozrzutu reszt względem wartości przewidywanych
  - wykres rozrzutu reszt względem zmiennej niezależnej

Przykład 1. Badając zależność zmiennej zależnej od jednej zmiennej niezależnej, przeprowadzono 5 doświadczeń, których wyniki zapisano w tabeli.

Lp.	*x*	*y*
1	-2	2
2	-1	3
3	0	5
4	1	6
5	7	14,5

Przeprowadź analizę regresji, zakładając, że zależność zmiennych jest liniowa, tzn.:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x.$$

Macierz wejść i wektor obserwowanych wyjść przyjmują w tym przypadku postać:

$$\mathbf{X} = \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 7 \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \\ 14{,}5 \end{bmatrix}.$$

W kolejnych krokach wyznaczono:

$$\mathbf{X}^T\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ -2 & -1 & 0 & 1 & 7 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ 1 & -1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 7 \end{bmatrix} = \begin{bmatrix} 5 & 5 \\ 5 & 55 \end{bmatrix}, \qquad \mathbf{C} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} = \begin{bmatrix} 0{,}22 & -0{,}02 \\ -0{,}02 & 0{,}02 \end{bmatrix}, \qquad \hat{\mathbf{b}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \begin{bmatrix} 4{,}7 \\ 1{,}4 \end{bmatrix}$$

Znaleziona zależność ma więc postać: $\hat{y} = 4{,}7 + 1{,}4x.$

(oznacza to, że każda zmiana wartości zmiennej niezależnej o 1 przyczynia się do zmiany wartości zmiennej zależnej o 1,4).

Przykład 1.cd. W kolejnym kroku zbadano jakość dopasowania znalezionej funkcji. Obliczono błędy:

$$SS_e = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}\hat{e}_i^2 = 0{,}2, \qquad MS_e = \frac{SS_e}{n-p-1} = \frac{0{,}2}{5-1-1} = 0{,}0667,$$

$$SS_r = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = (2-6{,}1)^2 + \ldots + (14{,}5-6{,}1)^2 = 98, \qquad MS_r = \frac{SS_r}{p} = \frac{98}{1} = 98,$$

$$SS_T = SS_r + SS_e = 0{,}2 + 98 = 98{,}2, \qquad MS_T = \frac{SS_T}{n-1} = \frac{98{,}2}{1} = 98{,}2.$$

Współczynniki determinacji wynoszą w tym przypadku:

$$R^2 = 1 - \frac{SS_e}{SS_T} = 1 - \frac{0{,}2}{98{,}2} = 0{,}998, \qquad \bar{R}^2 = 1 - \frac{n-1}{n-p-1}\frac{SS_e}{SS_T} = 1 - \frac{5-1}{5-1-1}\,\frac{98}{98{,}2} = 0{,}997.$$

świadczą więc o ***dobrym dopasowaniu*** modelu.

Następnie zbadano istotność funkcji i jej współczynników:

$$F_n = \frac{MS_r}{MS_e} = \frac{98}{0{,}0667} \approx 1470, \qquad p-value = 1 - F_{F(p,n-p-1)}(F_n) = 1 - F_{F(1,5-1-1)}(1470) \approx 0{,}000039,$$

$$t_n = \frac{\hat{b}_0}{s\sqrt{c_{00}}} = \frac{4{,}7}{\sqrt{0{,}0667 \cdot 0{,}22}} \approx 38{,}809, \qquad p-value = 2F_{t(5-1-1)}(-38{,}809) \approx 0{,}000038,$$

$$t_n = \frac{\hat{b}_1}{s\sqrt{c_{11}}} = \frac{1{,}4}{\sqrt{0{,}0667 \cdot 0{,}02}} \approx 38{,}341, \qquad p-value = 2F_{t(5-1-1)}(-38{,}341) \approx 0{,}000039,$$

Funkcja i jej współczynniki okazały się ***istotne***.

Przykład 1.cd. Na koniec wyznaczone zostały reszty (właściwa analiza nie została wykonana ze względu na zbyt małą liczbę obserwacji). Wyznaczono *macierz kapeluszową* i prognozowane wartości zmiennej zależnej:

$$\mathbf{H}=\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T=\begin{bmatrix} 0{,}38 & 0{,}32 & 0{,}26 & 0{,}2 & -0{,}16 \\ 0{,}32 & 0{,}28 & 0{,}24 & 0{,}2 & -0{,}04 \\ 0{,}26 & 0{,}24 & 0{,}22 & 0{,}2 & 0{,}08 \\ 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 & 0{,}2 \\ -0{,}16 & -0{,}04 & 0{,}08 & 0{,}2 & 0{,}92 \end{bmatrix},\qquad \mathbf{y}=\begin{bmatrix} 2 \\ 3 \\ 5 \\ 6 \\ 14{,}5 \end{bmatrix},\qquad \hat{\mathbf{y}}=\mathbf{H}\mathbf{y}=\begin{bmatrix} 1{,}9 \\ 3{,}3 \\ 4{,}7 \\ 6{,}1 \\ 14{,}5 \end{bmatrix}.$$

Z zależności:

$$1)\ \hat{e}_i=y_i-\hat{y}_i,\qquad 2)\ d_i=\frac{\hat{e}_i}{\hat{\sigma}},\qquad 3)\ \hat{e}_{(-i)}=\frac{\hat{e}_i}{1-h_{ii}},\qquad 4)\ r_i=\frac{\hat{e}_i}{\hat{\sigma}\sqrt{1-h_{ii}}},\qquad 5)\ t_i=\frac{\hat{e}_i}{\hat{\sigma}_{(-i)}\sqrt{1-h_{ii}}},\qquad 6)\ D_i=\frac{r_i^2}{p+1}\frac{h_{ii}}{1-h_{ii}}$$

wyznaczone zostały: 1. *reszty*, 2. *reszty standaryzowane*, 3. *reszty usunięte*,
reszty studentyzowane: 4. *wewnętrznie* i 5. *zewnętrznie*, 6. *odległość Cook'a*.

$\hat{\sigma}=\sqrt{MS_e}\approx 0{,}26,$

$2\bar{h}=2\frac{p+1}{n}=0{,}8,$

Lp.	h_{ii}	$\hat{e}_i$	d_i	$\hat{e}_{(-i)}$	r_i	t_i	D_i
1	0,38	0,1	0,39	0,16	0,49	0,42	0,07
2	0,28	-0,3	-1,16	-0,42	-1,37	-1,83	0,36
3	0,22	0,3	1,16	0,38	1,32	1,65	0,24
4	0,20	-0,1	-0,39	-0,13	-0,43	-0,37	0,02
5	0,92	0,0	0,00	0,00	0,00	0,00	0,00

obserwacja o dużej dźwigni — *reszty są zbliżone i niezbyt duże* — $D_i < 1$

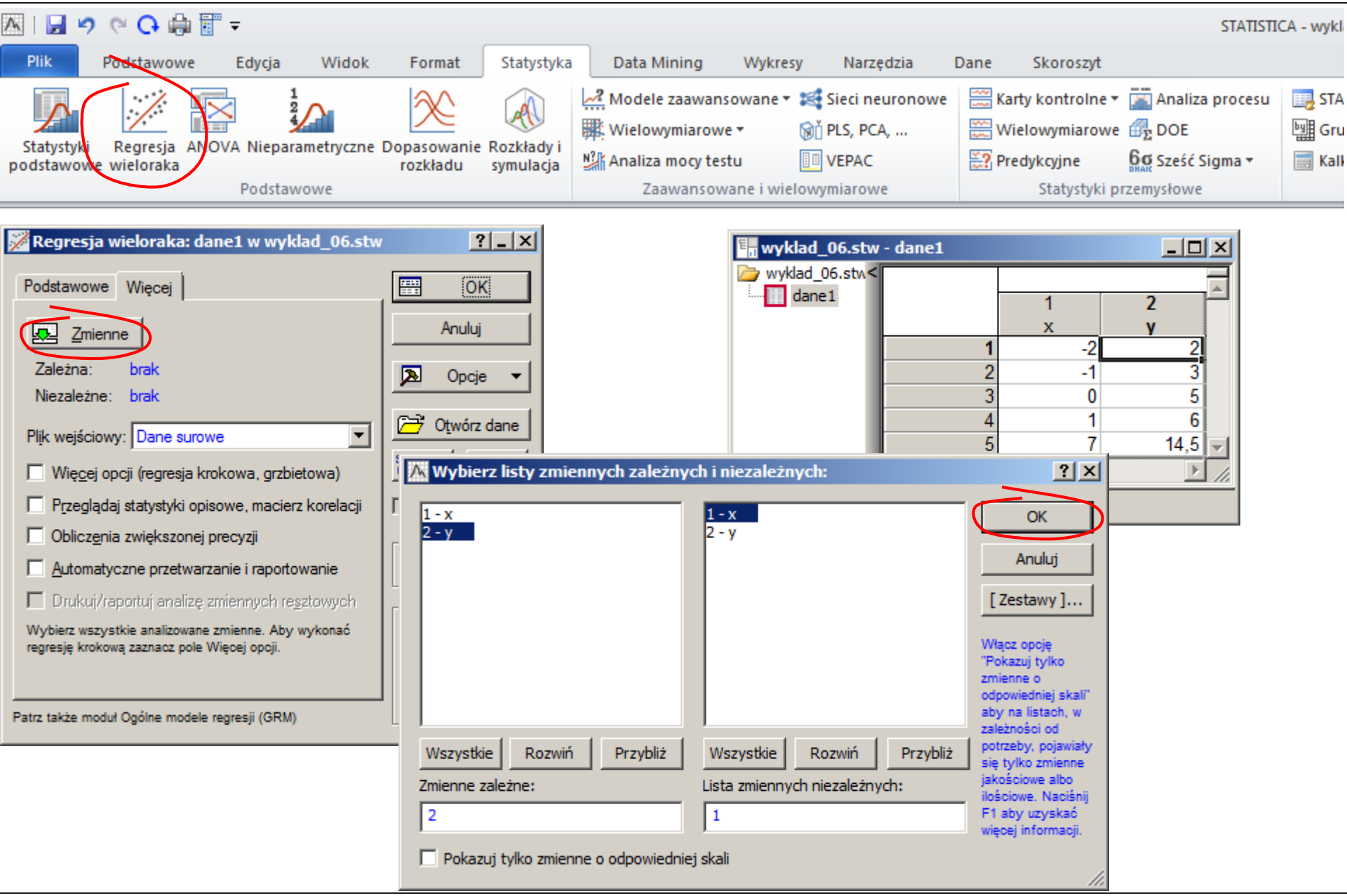
Otrzymane wyniki wskazują, że w analizowanym zbiorze danych nie ma obserwacji:

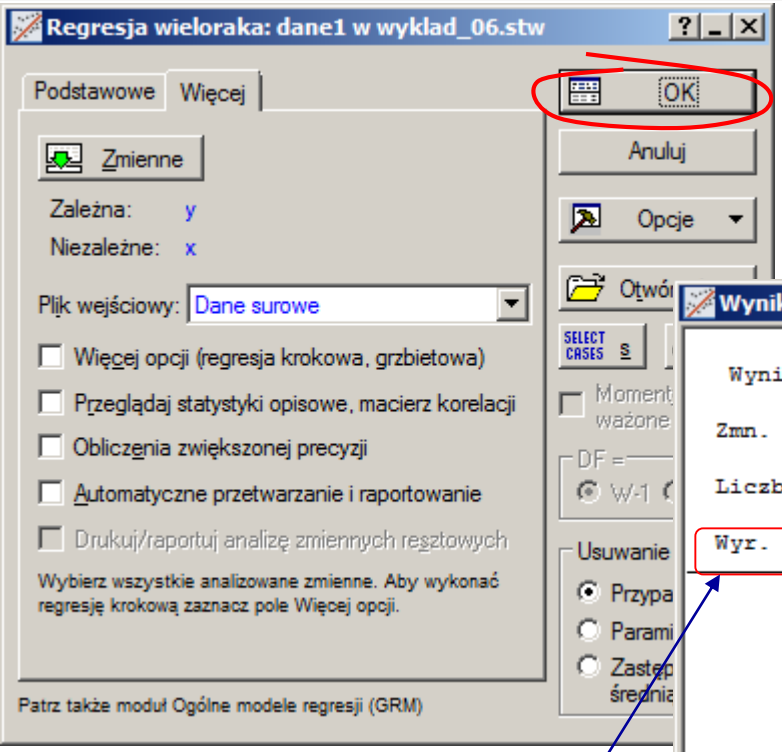
- *odstających,*
- *wpływowych.*

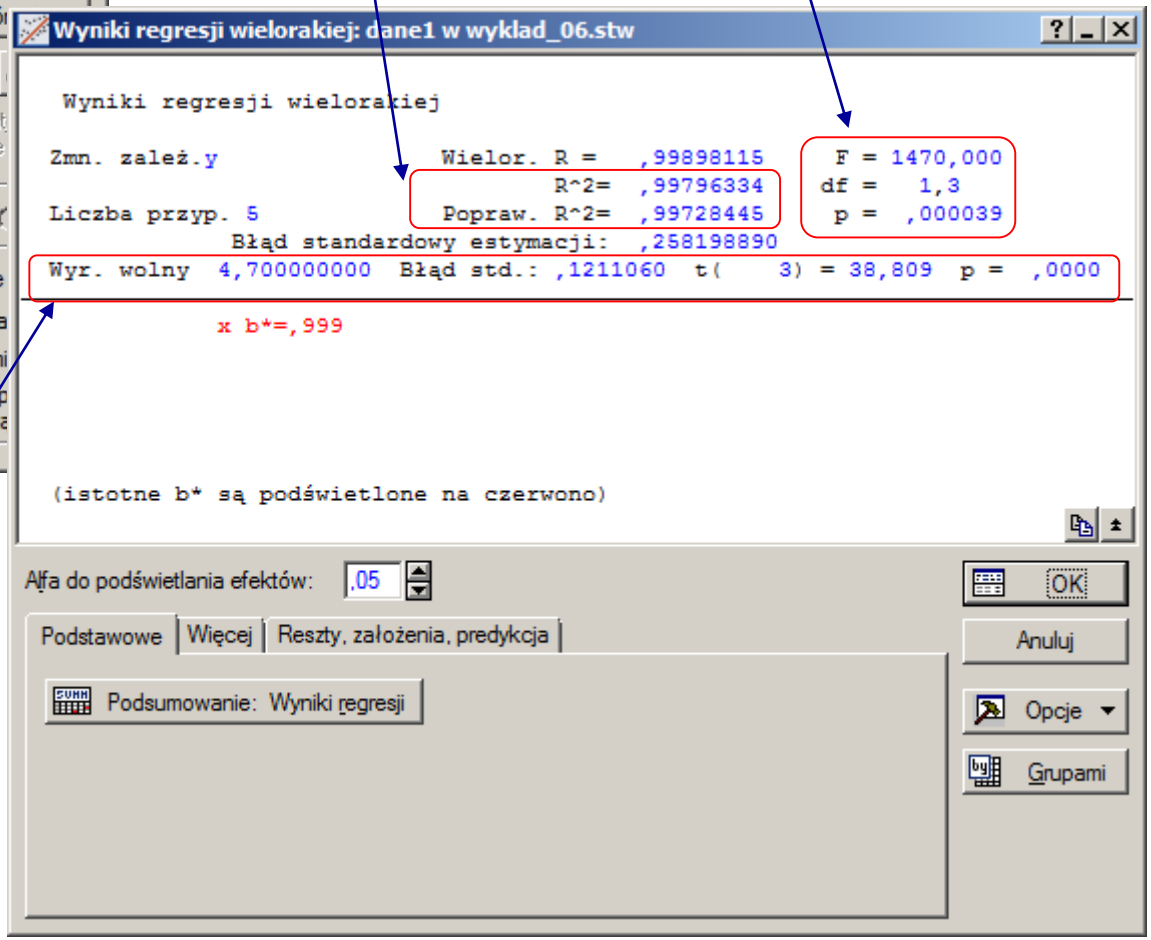
Obserwacja o dużej dźwigni nie ma wpływu na wynik analizy.

W tym przypadku nie ma więc konieczności rozważania usuwania żadnej z obserwacji.

STATISTICA – analiza regresji – regresja wieloraka







test istotności funkcji regresji

F – *statystyka testowa,*

df – *liczba stopni swobody F,*

p – *p–value,*

R^2, $\overline{R}^2$

test istotności dla wyrazu wolnego

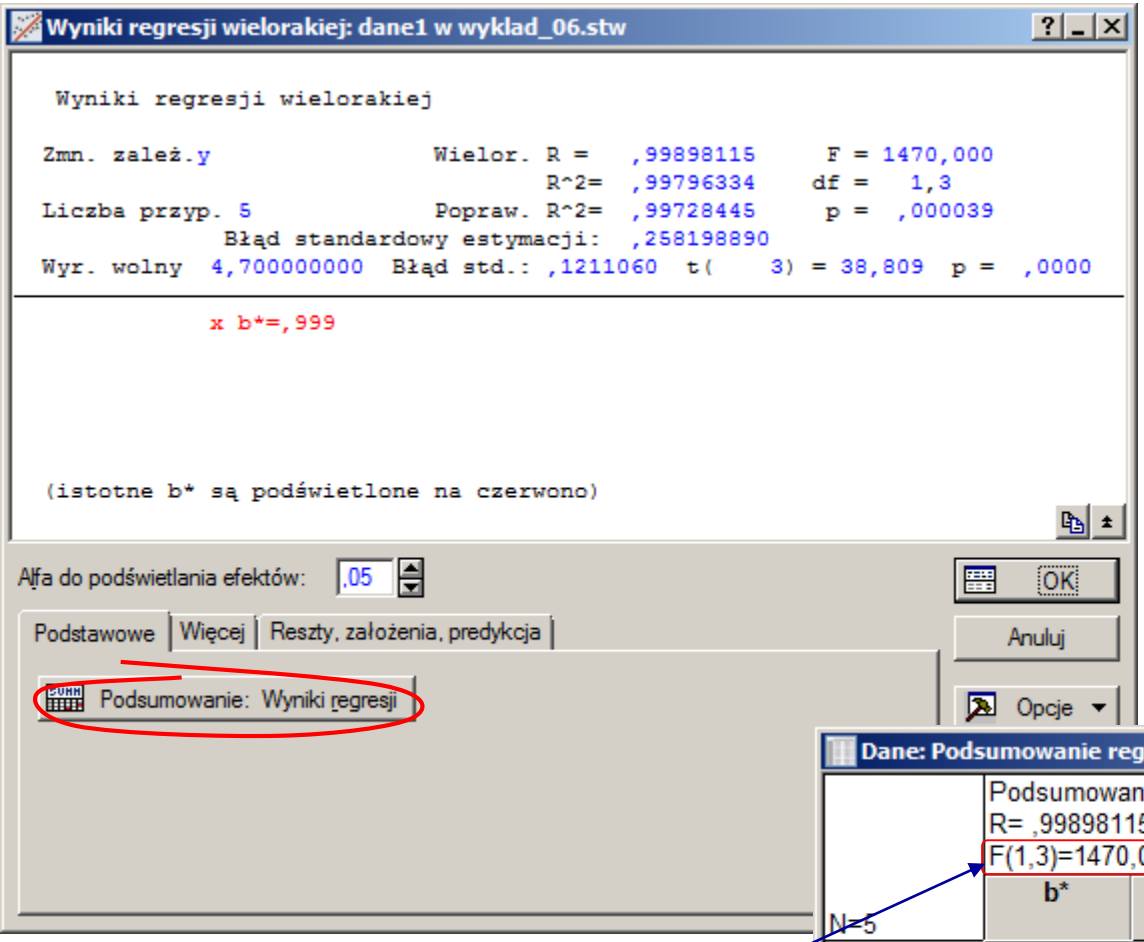
Wyr. wolny – $\hat{b}_0$

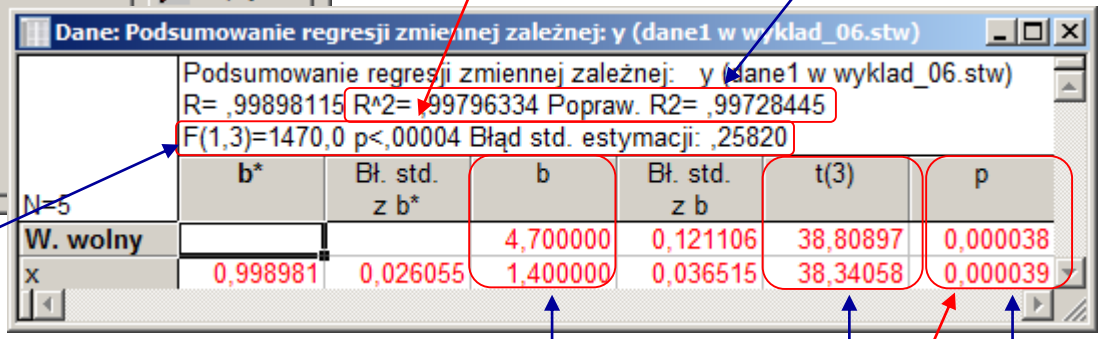
t – *statystyka testowa*

(w nawiasie liczba stopni swobody)

p – *p–value,*

STATISTICA – analiza regresji – regresja wieloraka





funkcja regresji istotna

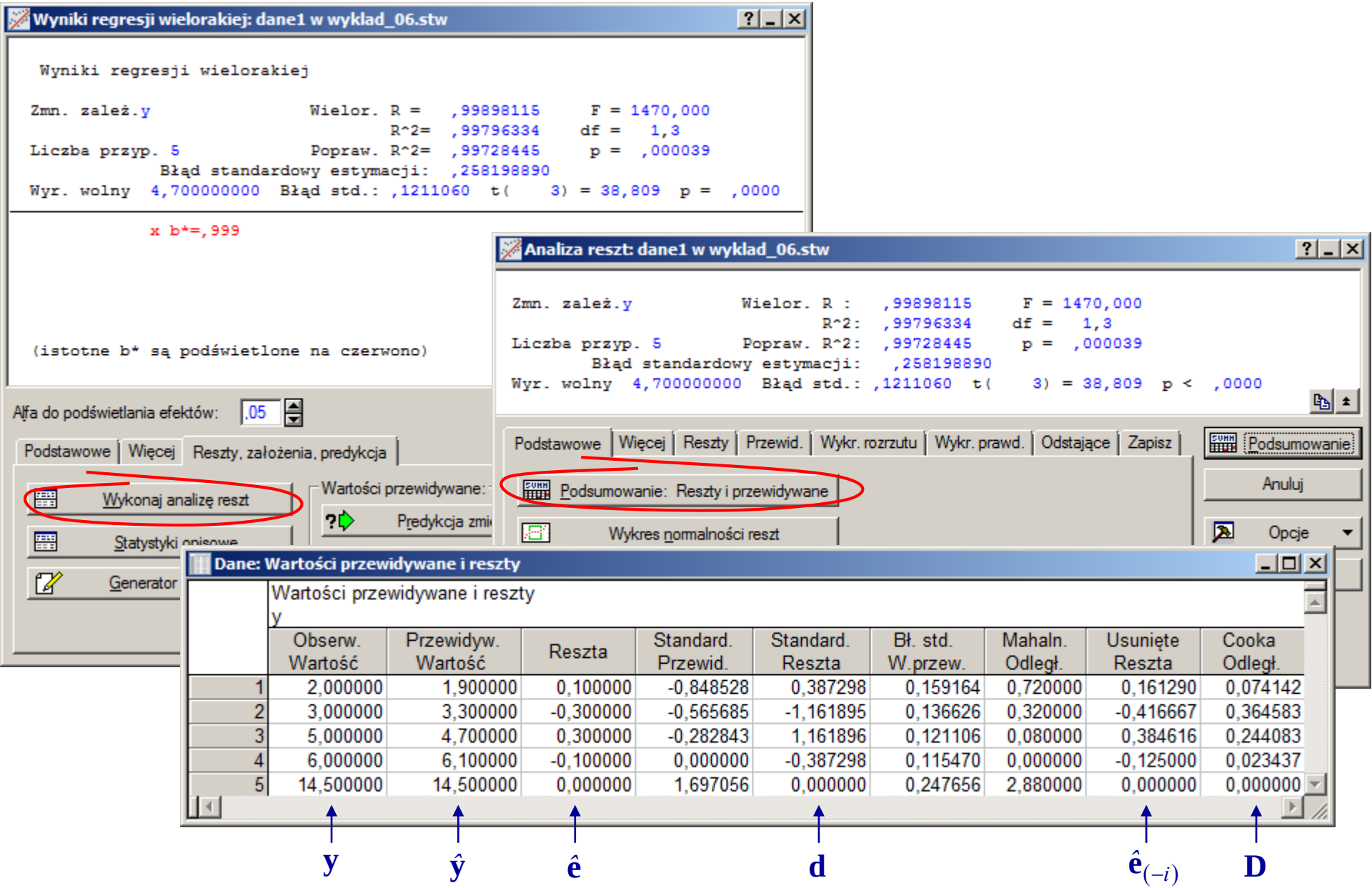
R^2, $\overline{R}^2$

test istotności funkcji regresji

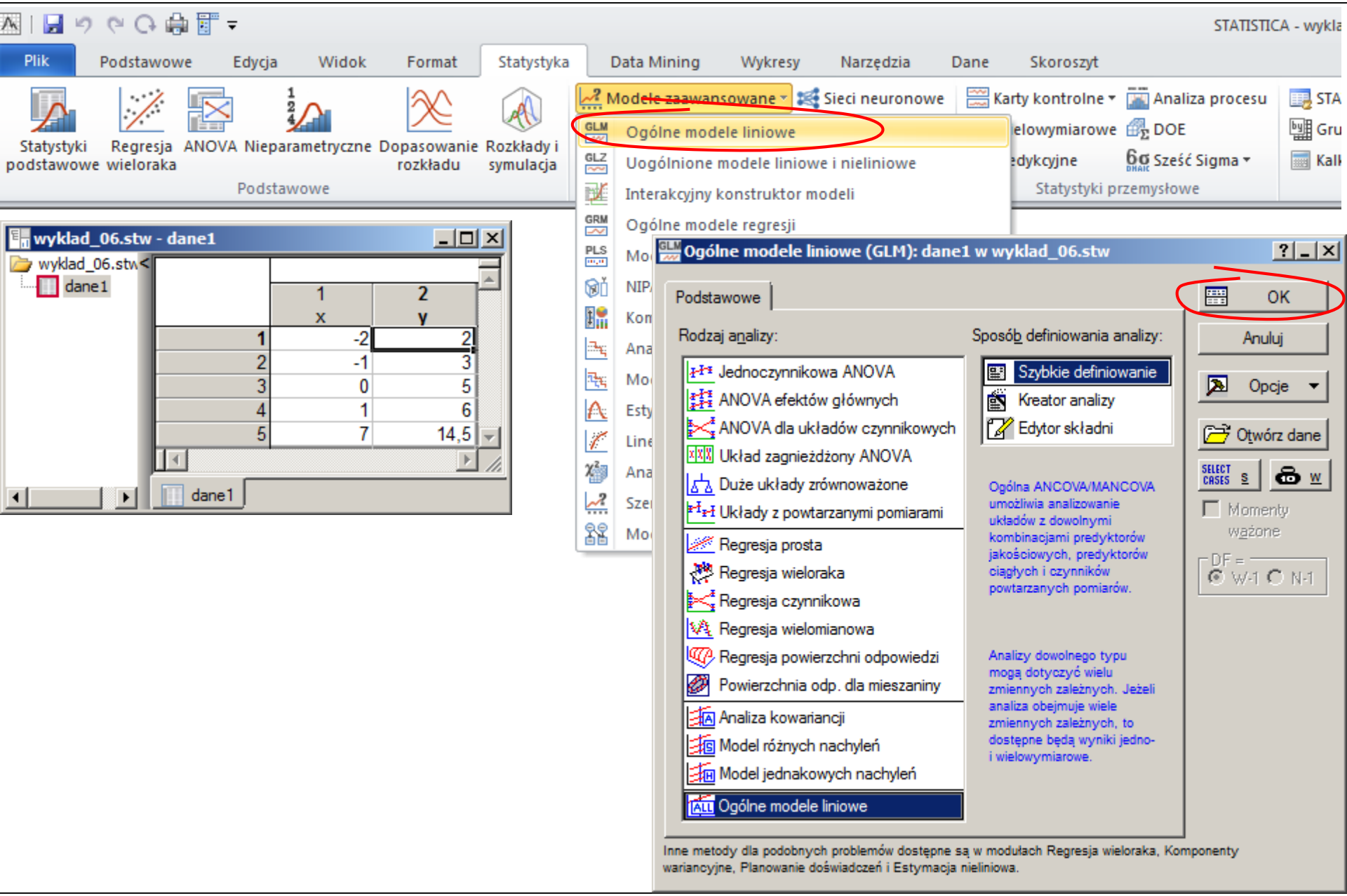
test istotności współczynników funkcji regresji: $\hat{b}_i$ t_n $p-value$

współczynniki funkcji regresji istotne

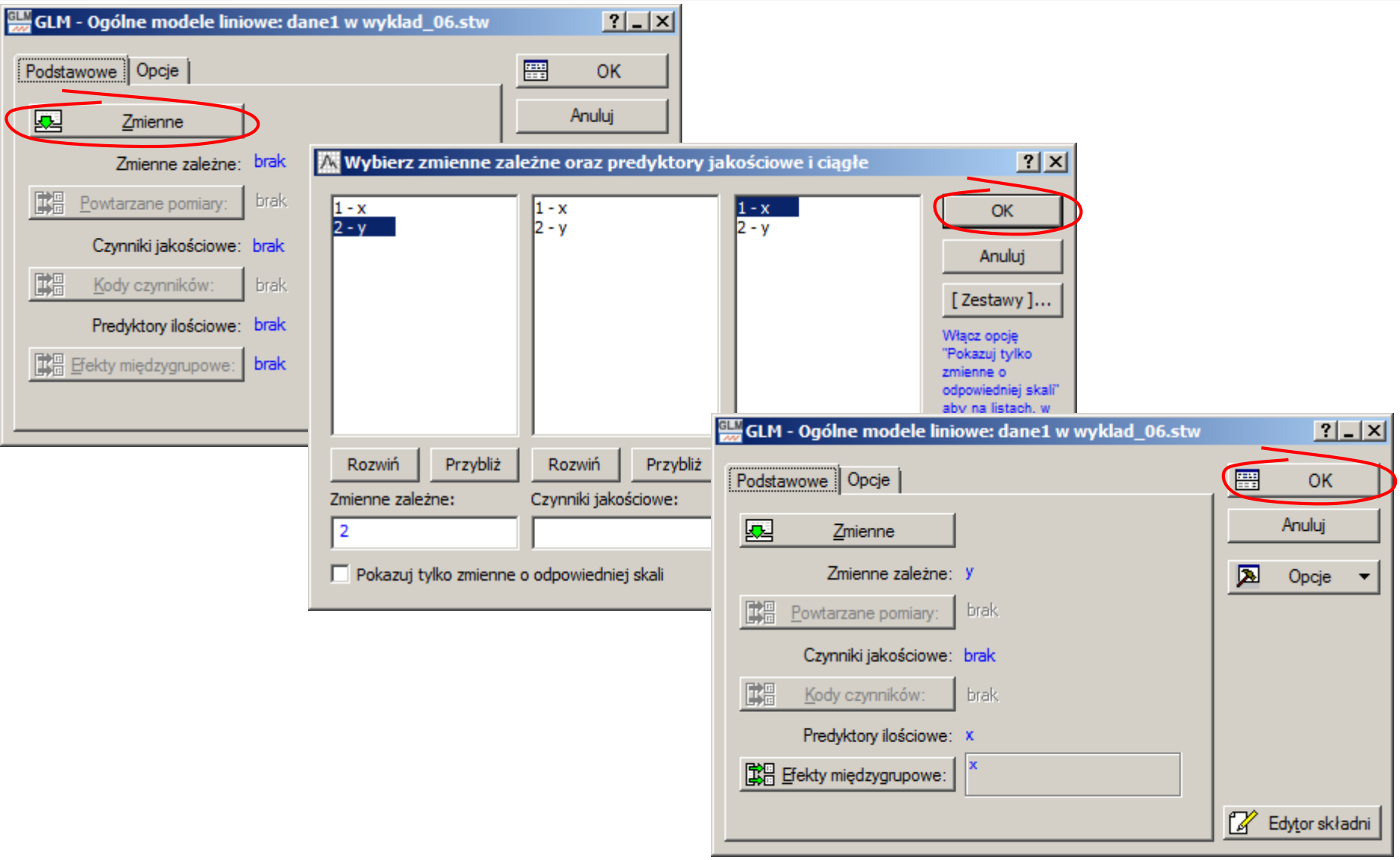
STATISTICA – analiza regresji – regresja wieloraka



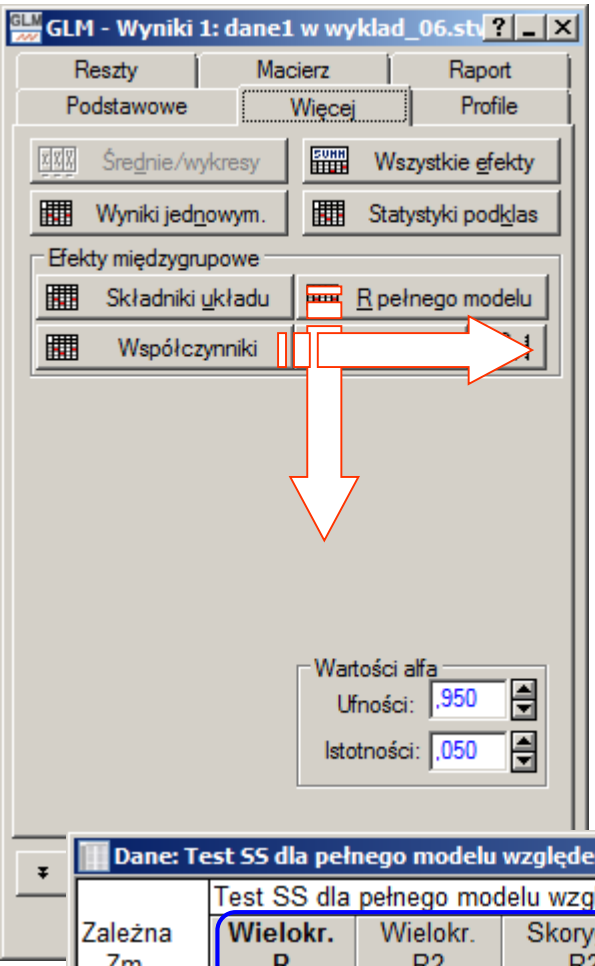
STATISTICA – analiza regresji – GLM



STATISTICA – analiza regresji – GLM



STATISTICA – analiza regresji – GLM



Dane: Oceny parametrów (dane1 w wyklad_06.stw)

Oceny parametrów (dane1 w wyklad_06.stw)
Parametryzacja z sigma-ograniczeniami

Efekt	y Param.	y Bł. std.	y t	y p
Wyraz wolny	4,700000	0,121106	38,80897	0,000038
x	1,400000	0,036515	38,34058	0,000039

test istotności współczynników funkcji regresji

$\hat{b}_i$

$p-value$

współczynniki funkcji regresji istotne

Dane: Test SS dla pełnego modelu względem SS dla reszt (dane1 w wyklad_06.stw)

Test SS dla pełnego modelu względem SS dla reszt (dane1 w wyklad_06.stw)

Zależna Zm.	Wielokr. R	Wielokr. R2	Skorygow R2	SS Model	df Model	MS Model	SS Reszta	df Reszta	MS Reszta	F	p
y	0,998981	0,997963	0,997284	98,00000	1	98,00000	0,200000	3	0,066667	1470,000	0,000039

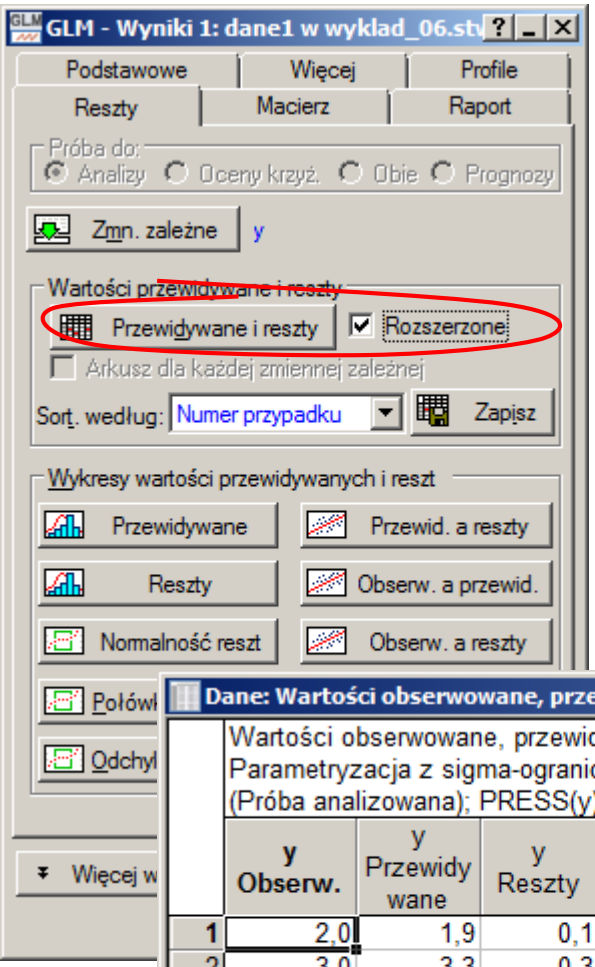
jakość dopasowania funkcji regresji

test istotności funkcji regresji

$p-value$

funkcja regresji istotna

STATISTICA – analiza regresji – GLM



Dane: Wartości obserwowane, przewidywane i reszty*

Wartości obserwowane, przewidywane i reszty
Parametryzacja z sigma-ograniczeniami
(Próba analizowana); PRESS(y) = 0,363180

	y Obserw.	y Przewidywane	y Reszty	y Bł.std.p.	y z przew.	y z reszt	y Stu.res.	y W.wpływ.	y R. usun.	y Std.r.us.	y Odl.Mah.	y Odl.Cooka	y Dffitsa	y z Dffitsa
1	2,0	1,9	0,1	0,16	-0,85	0,39	0,49	0,38	0,16	0,42	0,72	0,07	0,06	0,33
2	3,0	3,3	-0,3	0,14	-0,57	-1,16	-1,37	0,28	-0,42	-1,83	0,32	0,36	-0,12	-1,14
3	5,0	4,7	0,3	0,12	-0,28	1,16	1,32	0,22	0,38	1,65	0,08	0,24	0,08	0,88
4	6,0	6,1	-0,1	0,12	0,00	-0,39	-0,43	0,20	-0,13	-0,37	0,00	0,02	-0,03	-0,18
5	14,5	14,5	-0,0	0,25	1,70	-0,00	-0,00	0,92	-0,00	-0,00	2,88	0,00	-0,00	-0,00

y $\hat{y}$ $\hat{e}$ d r h $\hat{e}_{(-i)}$ t D

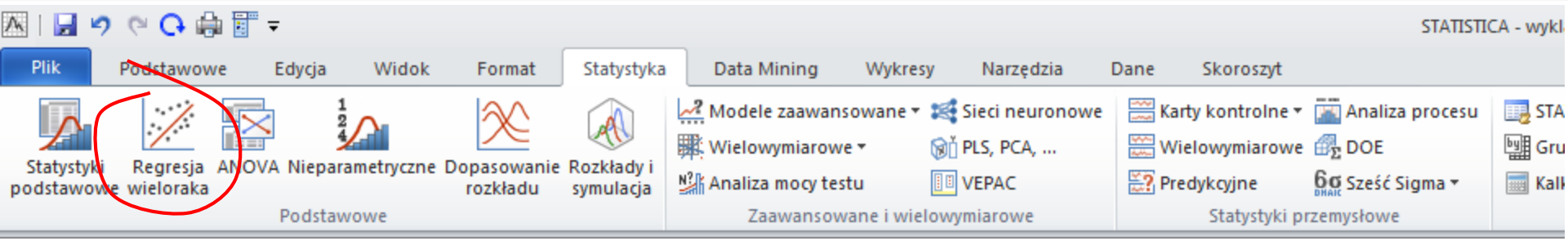
Przykład 2.

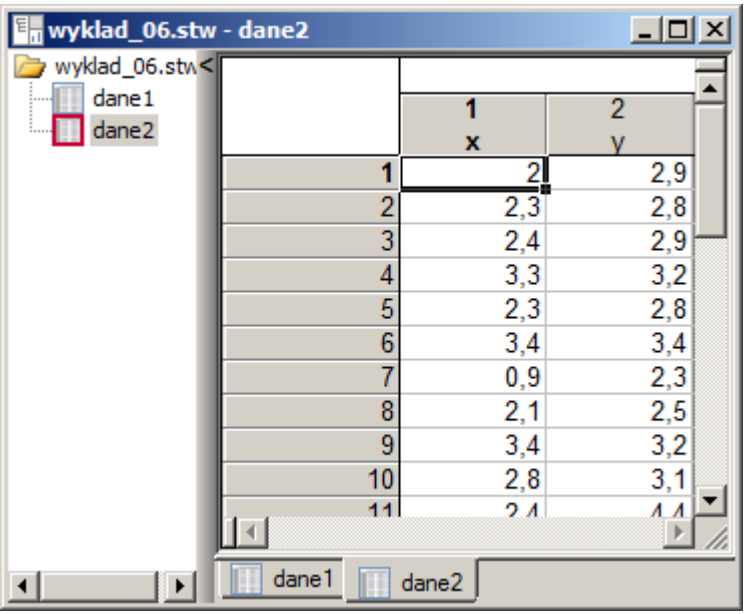
W celu zbadania zależności pomiędzy kosztami produkcji pewnego artykułu (*cecha Y*, w milionach zł) a wielkością produkcji (*cecha X*, w milionach sztuk) zebrano dane z 30 zakładów produkujących ten artykuł.

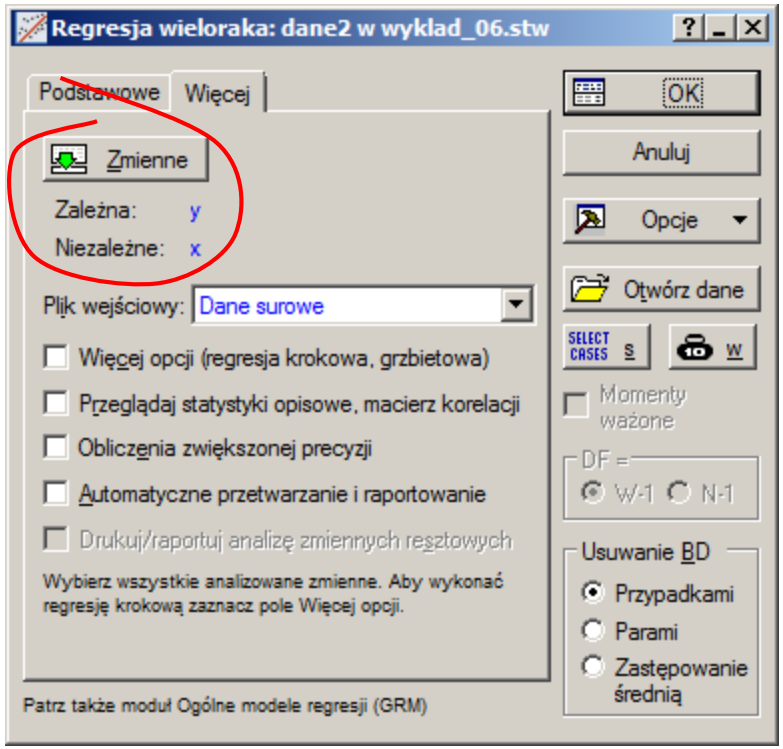
nr	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
x_i	2	2,3	2,4	3,3	2,3	3,4	0,9	2,1	3,4	2,8	2,4	3,2	2,8	3	1,7
y_i	2,9	2,8	2,9	3,2	2,8	3,4	2,3	2,5	3,2	3,1	4,4	3,2	3	2,8	2,2

nr	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
x_i	2,1	2,4	2,3	1	1,8	2,5	2	1,1	2,2	2,6	1	3,1	1,5	1,3	3,5
y_i	2,6	2,6	2,7	2,3	2,7	2,8	2,4	2,4	3,2	3,2	2,5	2,9	2,6	2,5	3,5

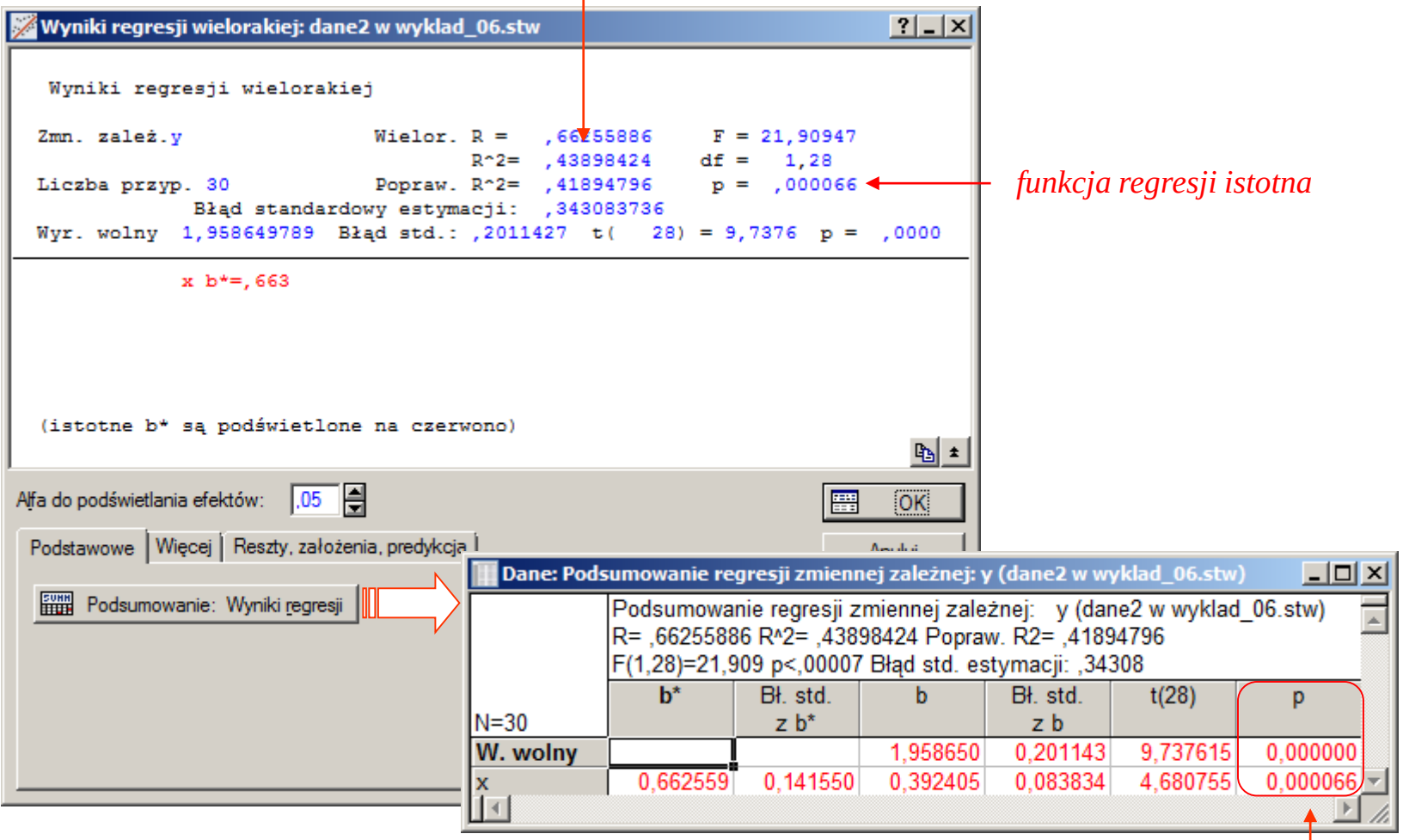
STATISTICA – analiza regresji – regresja wieloraka







dopasowanie funkcji regresji nie jest duże



funkcja regresji istotna

Otrzymane równanie regresji $\hat{y} = 1{,}96 + 0{,}35\,x$ *oznacza, że:*

zwiększanie produkcji o 1 *mln sztuk powoduje wzrost kosztów produkcji o* 0,35 *mln zł.*

współczynniki funkcji regresji istotne

STATISTICA – analiza regresji – regresja wieloraka

Wyniki regresji wielorakiej: dane2 w wyklad_06.stw

Wyniki regresji wielorakiej

Zmn. zależ.y	Wielor. R =	,66255886	F = 21,90947
	R^2=	,43898424	df = 1,28
Liczba przyp. 30	Popraw. R^2=	,41894796	p = ,000066
Błąd standardowy estymacji:		,343083736	
Wyr. wolny 1,958649789	Błąd std.:	,2011427	t(28) = 9,7376 p = ,0000

x b*=,663

(istotne b* są podświetlone na czerwono)

Alfa do podświetlania efektów: ,05

Podstawowe | Więcej | Reszty, założenia, predykcja

Wykonaj analizę reszt

Statystyki opisowe

Generator kodów

Wartości przew

Prec

Oblicz grani

Oblicz grani

Analiza reszt: dane2 w wyklad_06.stw

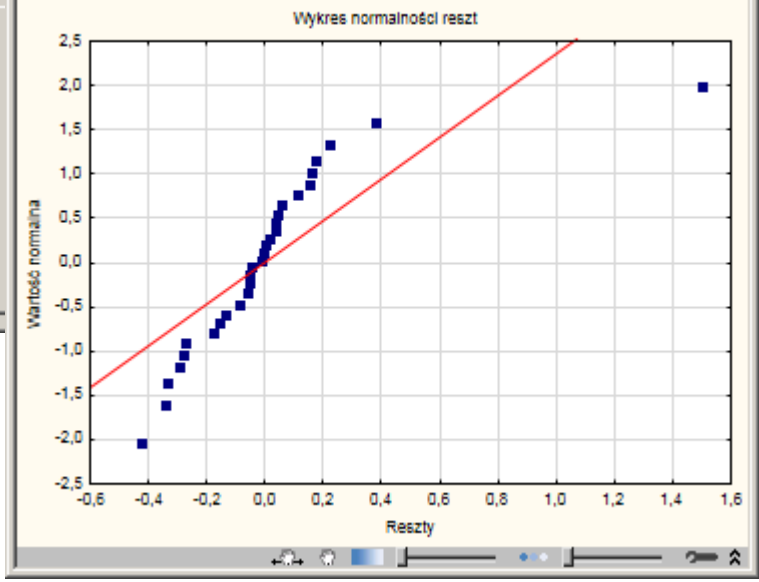
Zmn. zależ.y	Wielor. R :	,66255886	F = 21,90947
	R^2:	,43898424	df = 1,28
Liczba przyp. 30	Popraw. R^2:	,41894796	p = ,000066
Błąd standardowy estymacji:		,343083736	
Wyr. wolny 1,958649789	Błąd std.:	,2011427	t(28) = 9,7376 p < ,0000

Podstawowe | Więcej | Reszty | Przewid. | Wykr. rozrzu

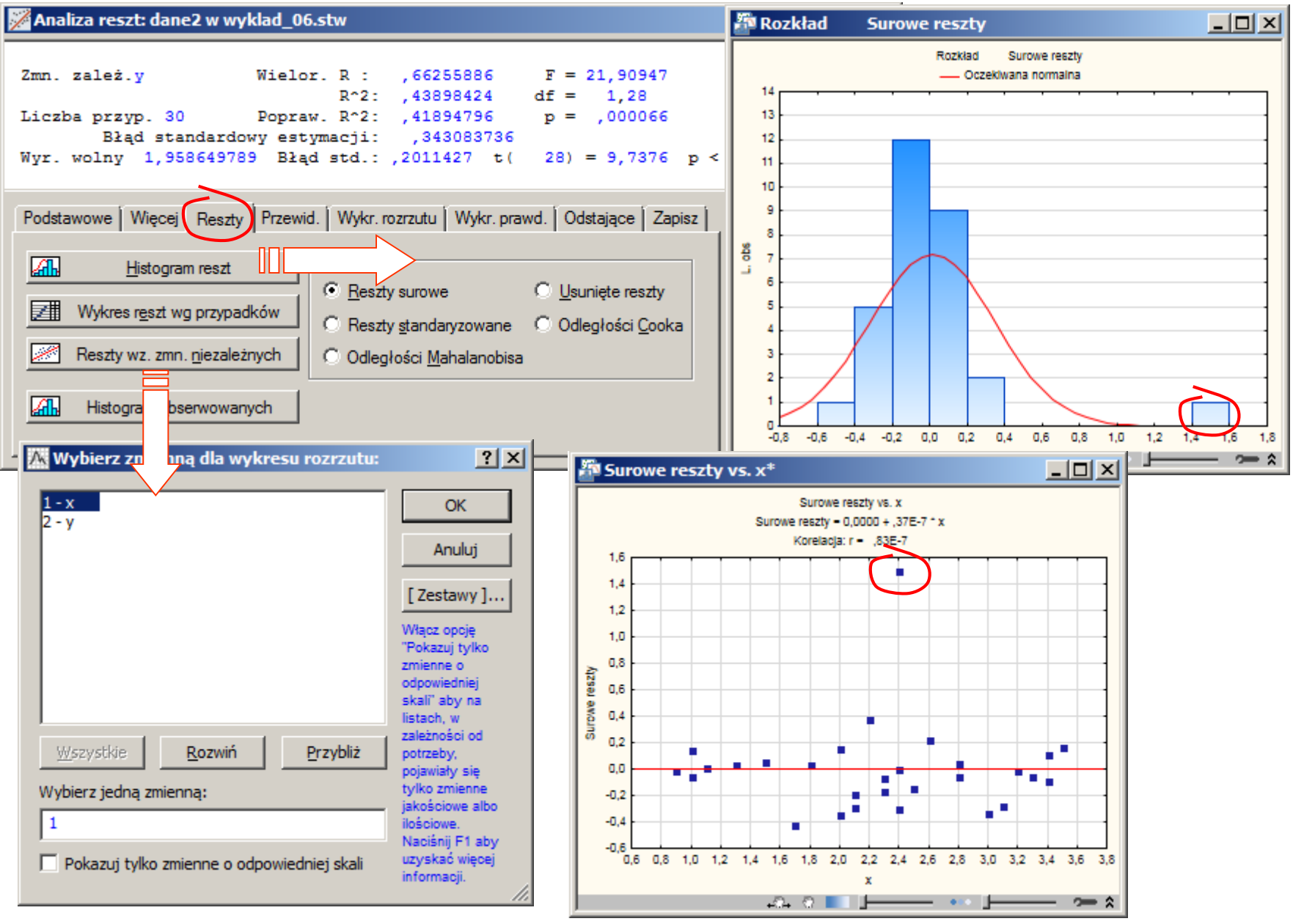
Podsumowanie: Reszty i przewidywane

Wykres normalności reszt

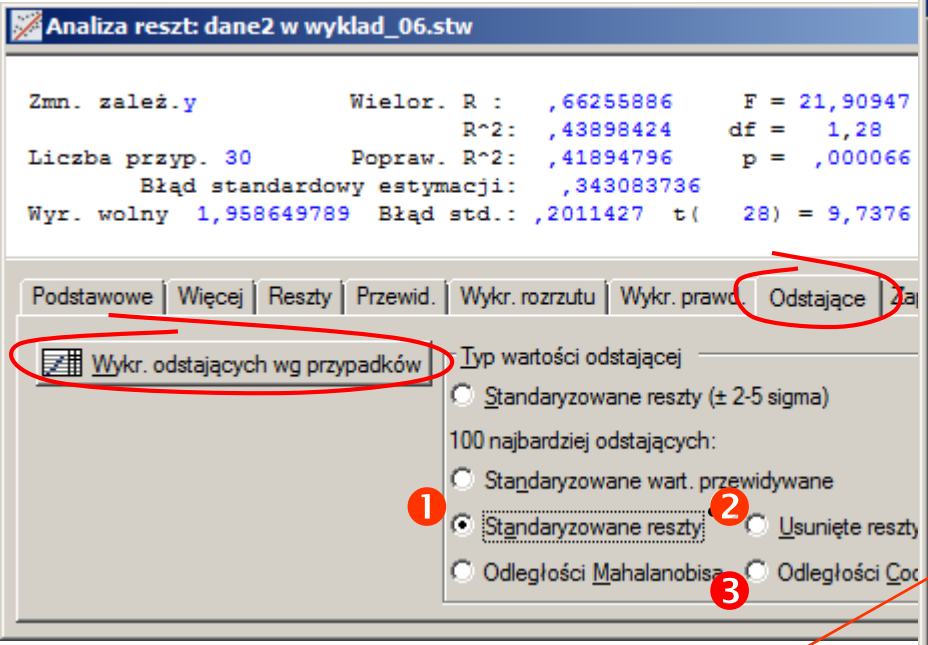
Wykres normalności reszt*

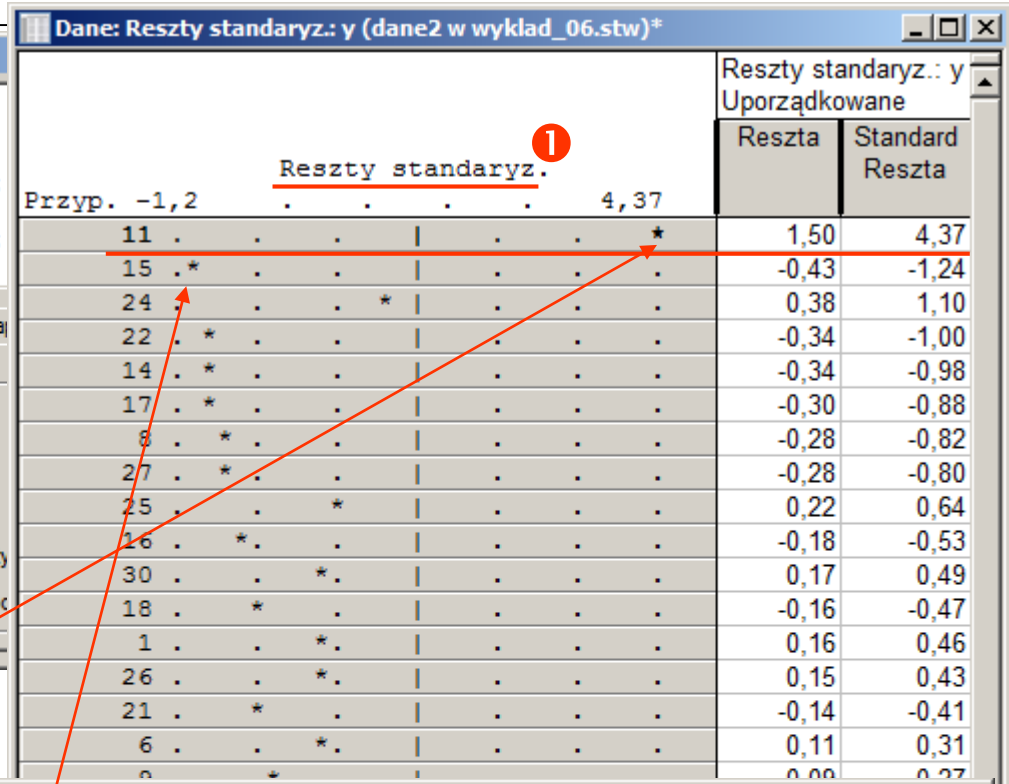


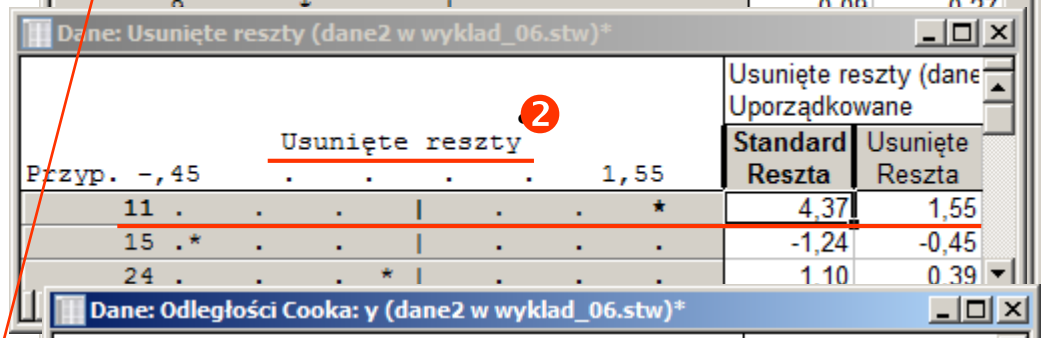
STATISTICA – analiza regresji – regresja wieloraka

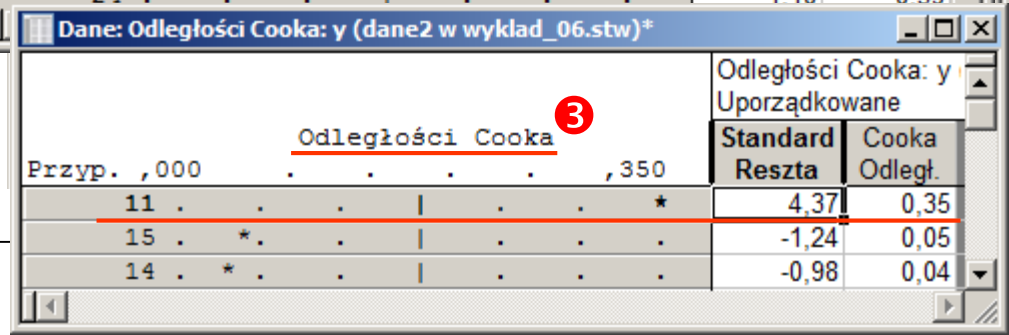


STATISTICA – analiza regresji – regresja wieloraka



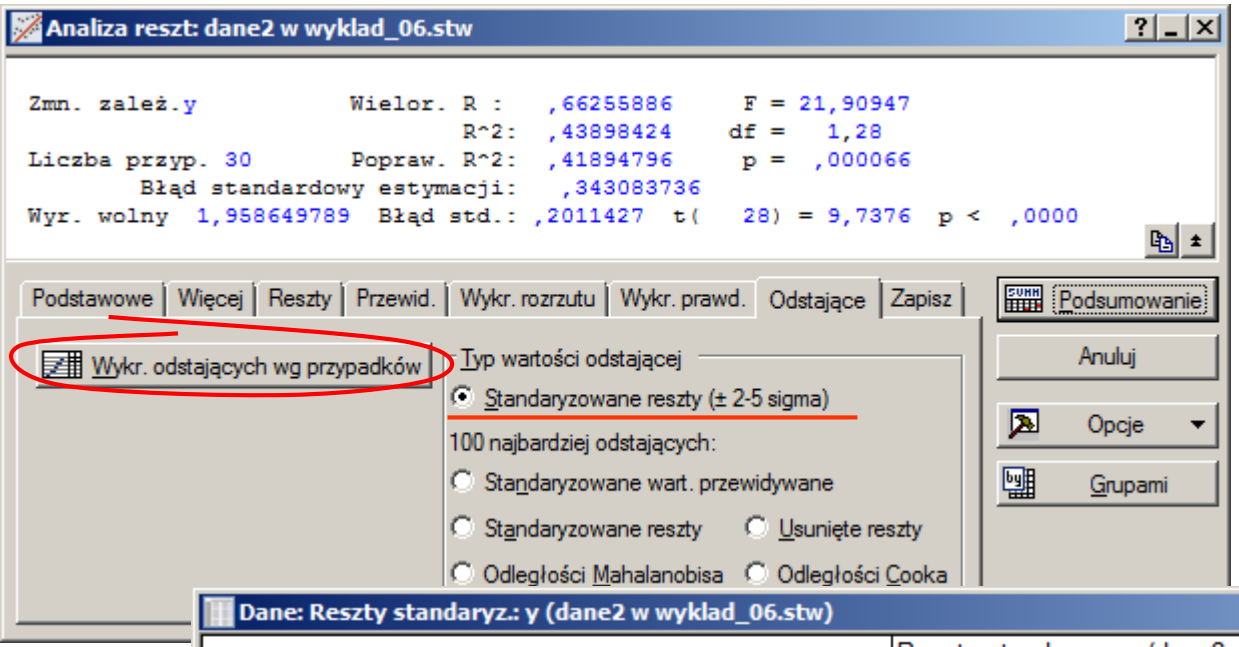






wybrane reszty są porządkowane w kolejności malejącej

- *największa reszta standaryzowana wystąpiła w 11 obserwacji, d_{11}=4,37,*
  reszta przekraczająca 3 wskazuje, że obserwacja 11 jest obserwacją odstającą,
  11 obserwacja ma również największą wartość:
  reszty usuniętej: $\hat{e}_{(-11)}$=1,55 i
  odległości Cook-a: D_{11} = 0,35,
  wartości te nie wskazują jednak na to, że obserwacja 11 jest obserwacją wpływową
- *kolejną resztę otrzymano w obserwacji 15, d_{15}= –1,24*



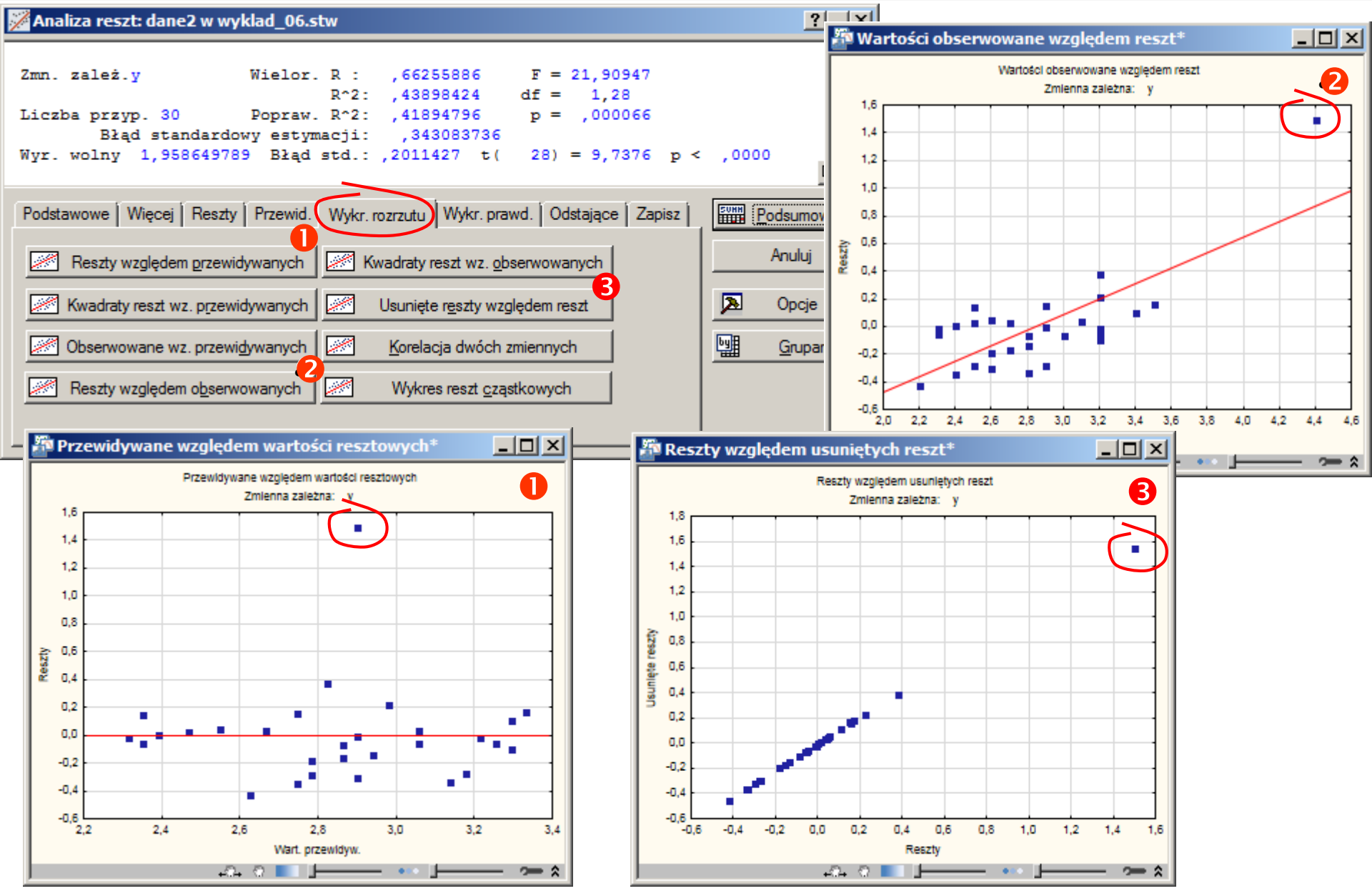
Dane: Reszty standaryz.: y (dane2 w wyklad_06.stw)

Reszty standaryz.: y (dane2 w wyklad_06.stw)
Odstające

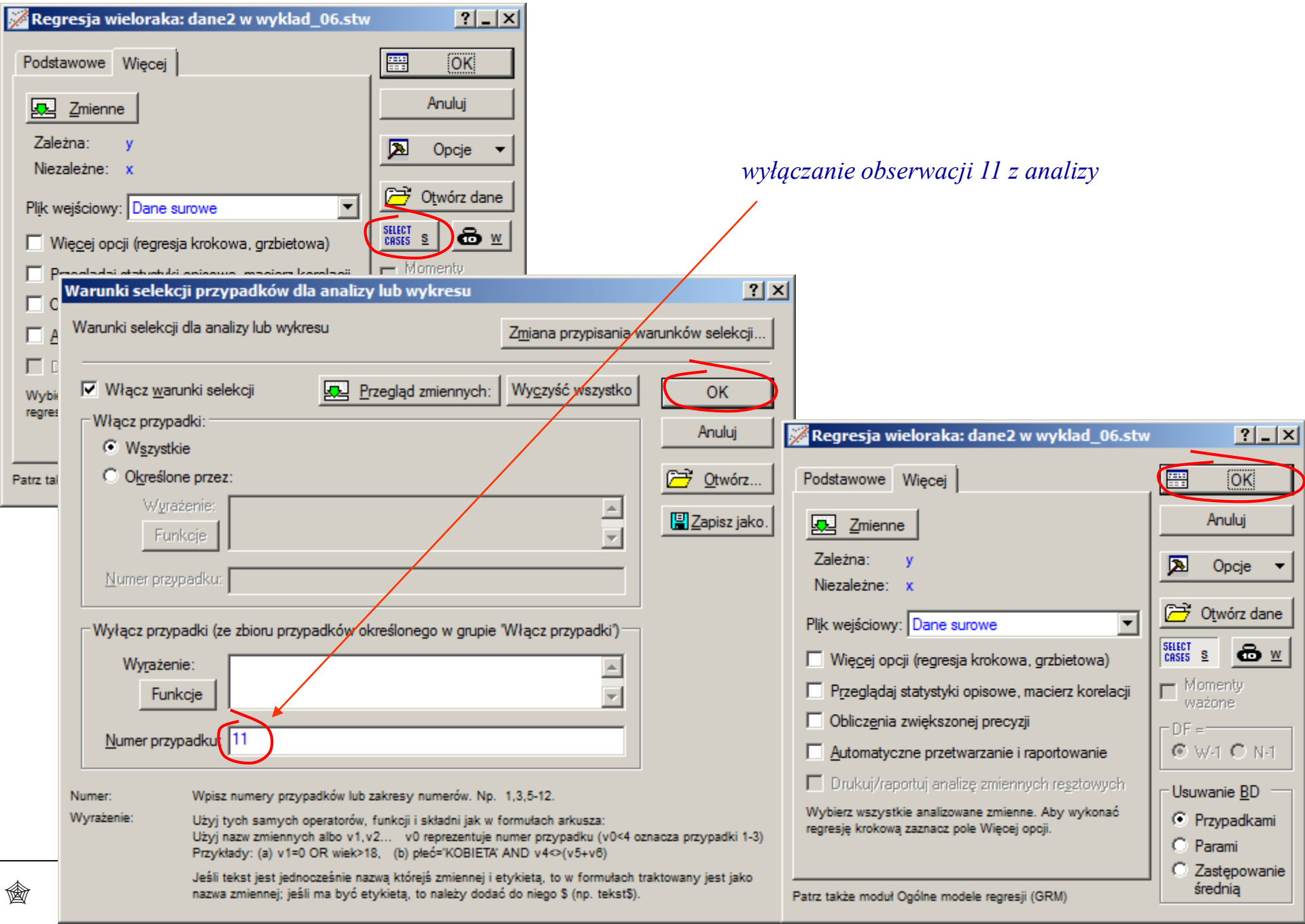
Przyp.	Reszty standaryz. -5. -4. -3. ±2. 3. 4. 5.	Obserw. Wartość	Przewidyw. Wartość	Reszta	Standard Przewid.	Standard Reszta	Bł. std. W.przew.
11	. . . | . .* .	4,400000	2,900422	1,499578	0,157906	4,370881	0,063441
Minimum	. . . | . .* .	4,400000	2,900422	1,499578	0,157906	4,370881	0,063441
Maksim.	. . . | . .* .	4,400000	2,900422	1,499578	0,157906	4,370881	0,063441
Średnia	. . . | . .* .	4,400000	2,900422	1,499578	0,157906	4,370881	0,063441
Mediana	. . . | . .* .	4,400000	2,900422	1,499578	0,157906	4,370881	0,063441

Obserwacja 11 jest wskazywana przez STATISTIC-ę jako obserwacja odstająca.

STATISTICA – analiza regresji – regresja wieloraka

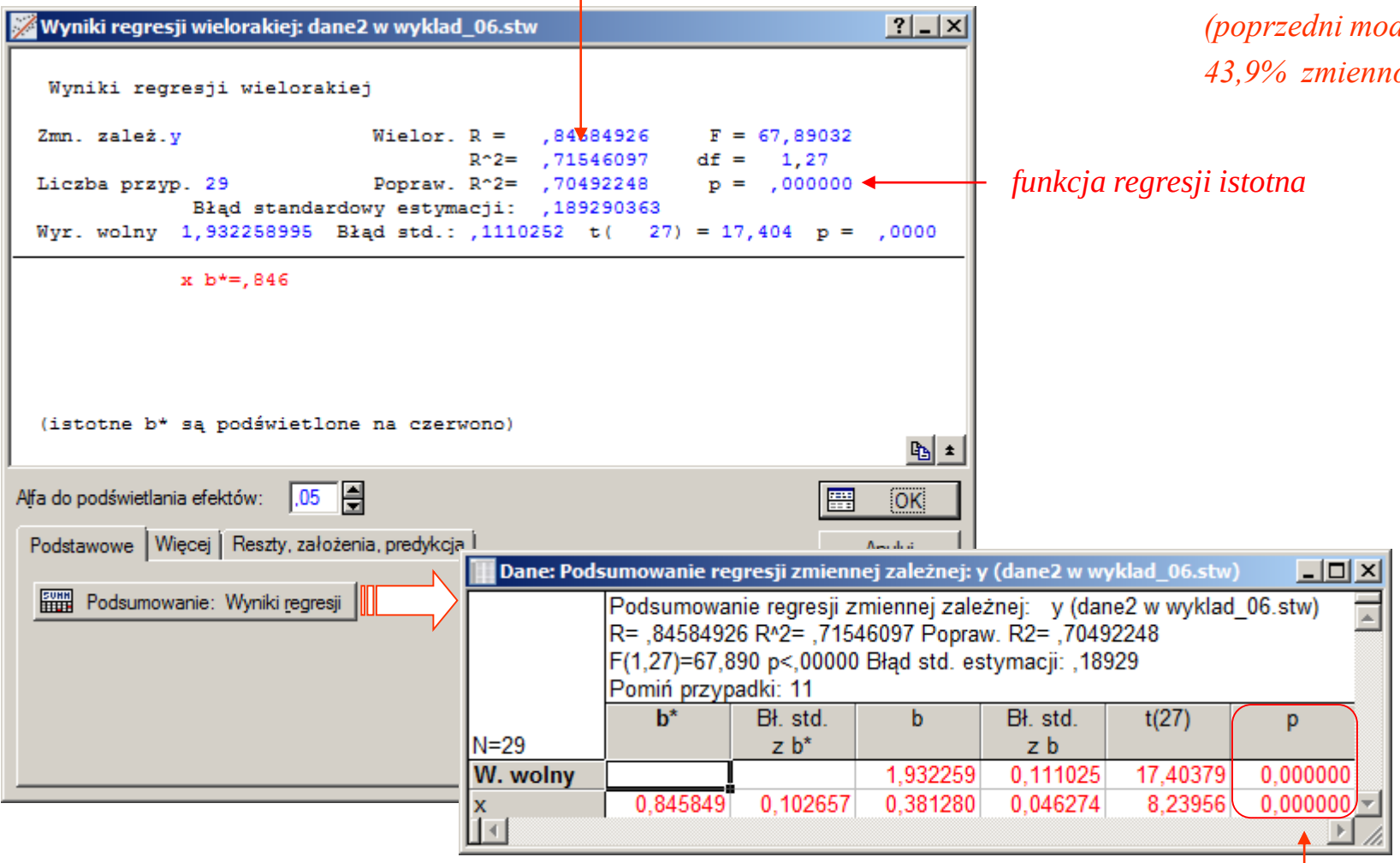


STATISTICA – analiza regresji – regresja wieloraka



dopasowanie funkcji regresji wzrosło, model wyjaśnia 71,5% zmienności kosztów produkcji

(poprzedni model wyjaśniał tylko 43,9% zmienności)



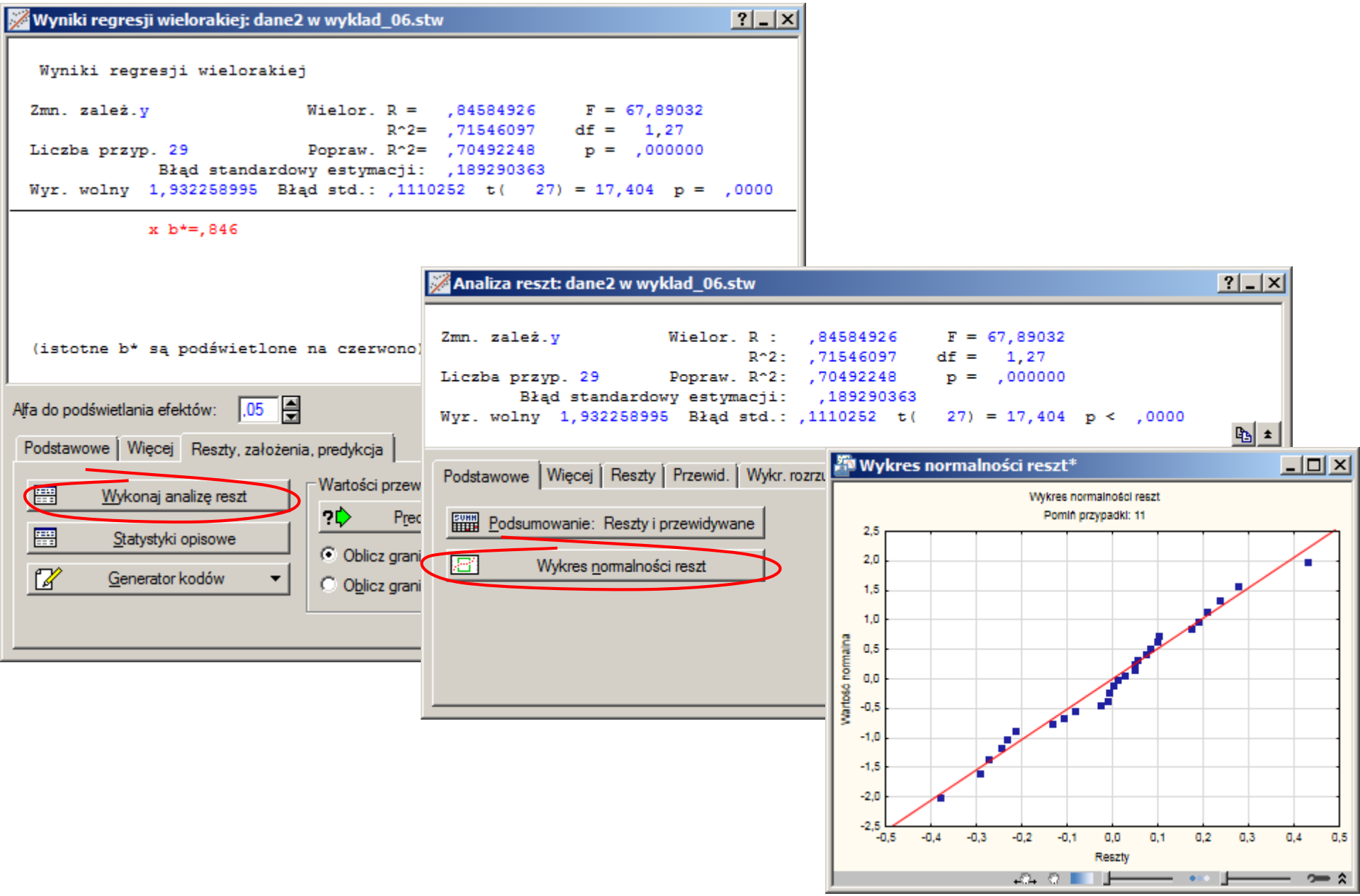
funkcja regresji istotna

Otrzymane model: $\hat{y} = 1{,}93 + 0{,}38\,x$

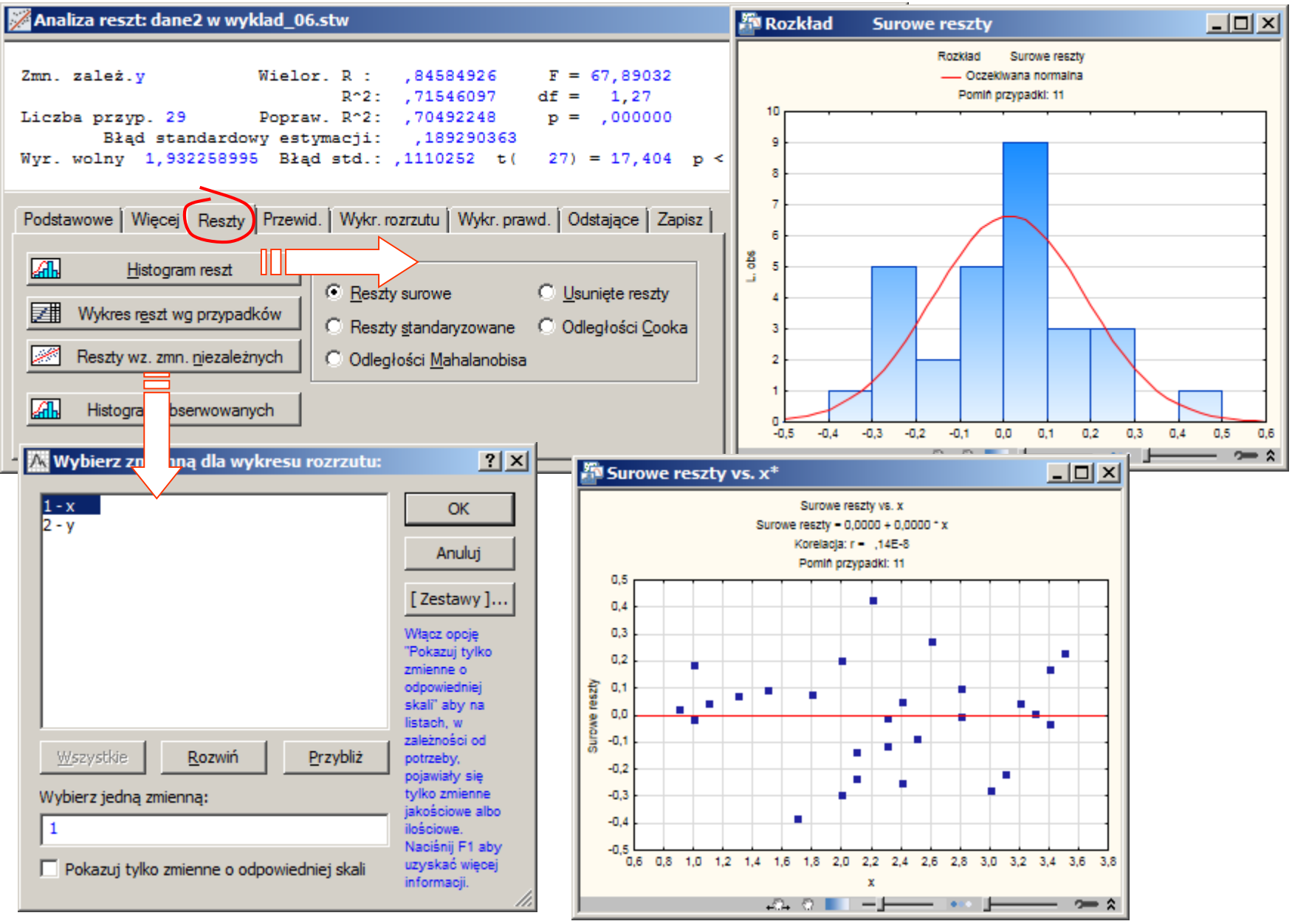
nie różni się istotnie od poprzedniego: $\hat{y} = 1{,}96 + 0{,}35\,x$

współczynniki funkcji regresji istotne

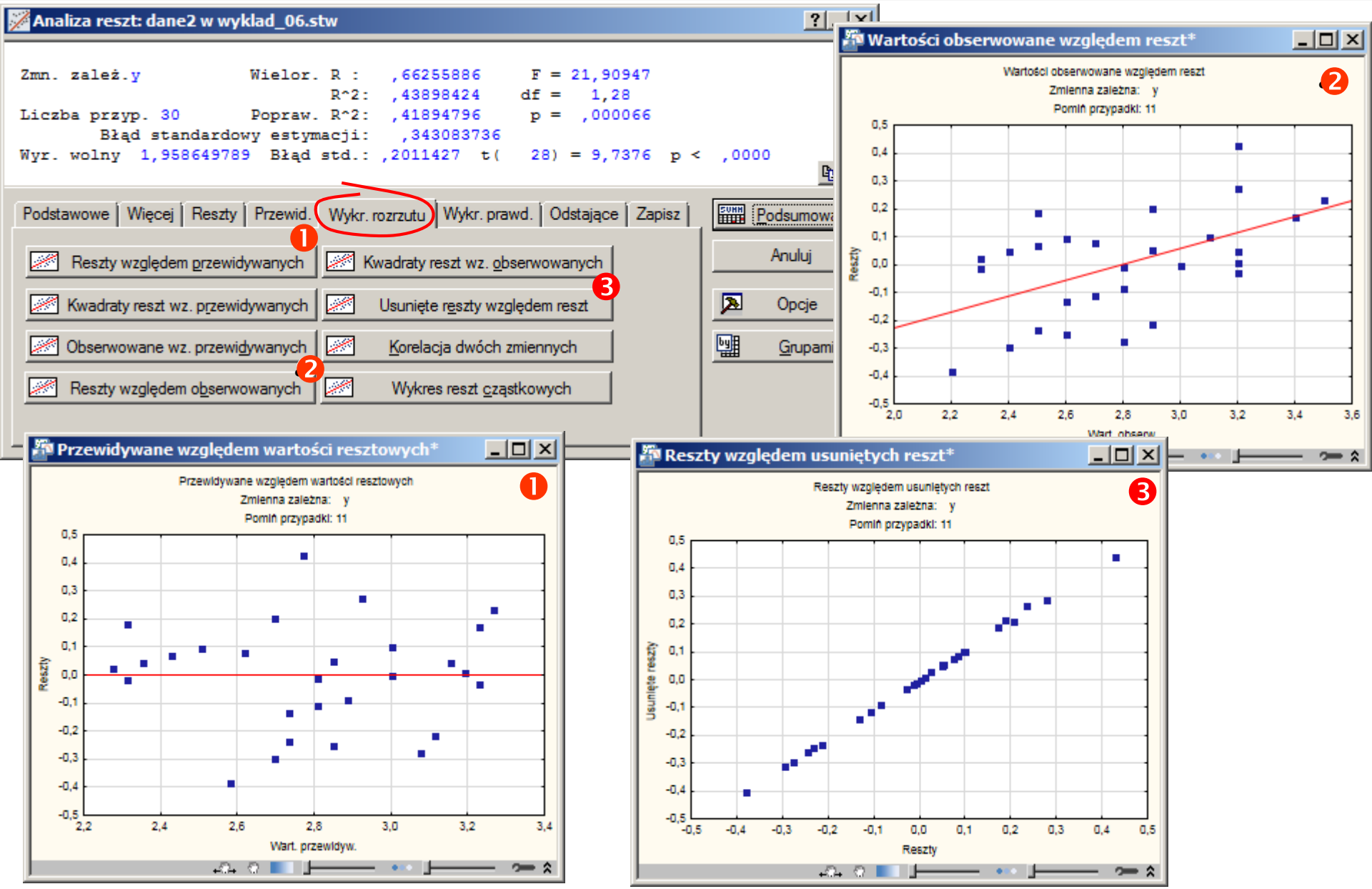
STATISTICA – analiza regresji – regresja wieloraka



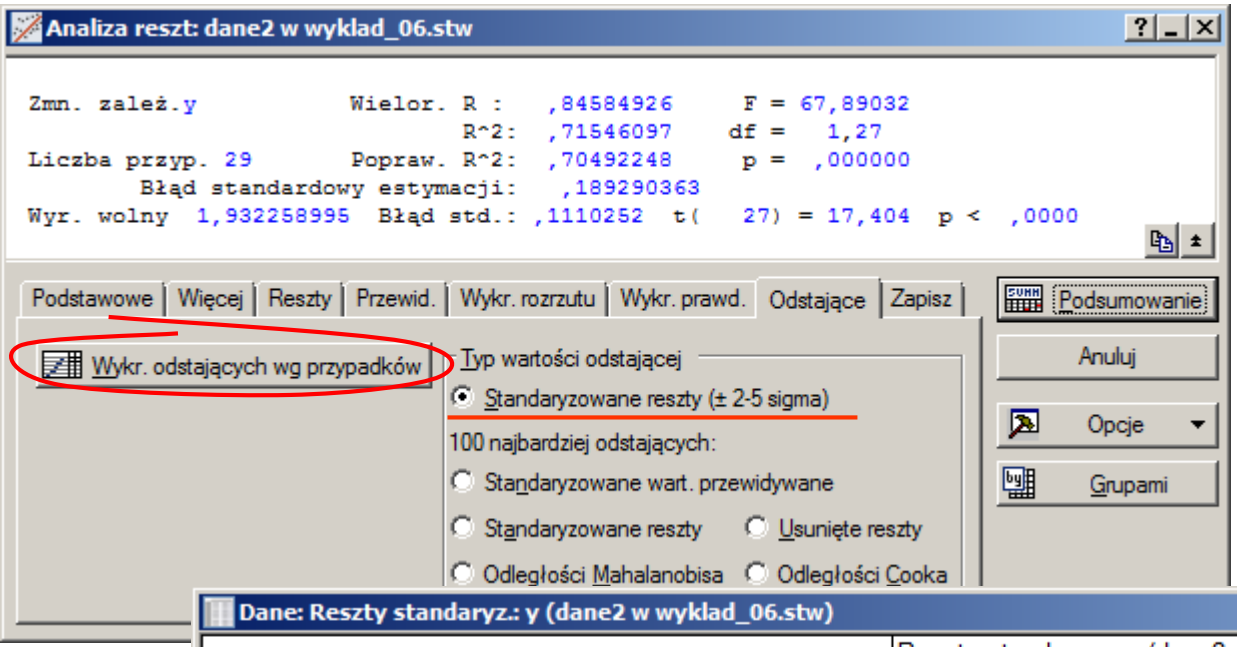
STATISTICA – analiza regresji – regresja wieloraka



STATISTICA – analiza regresji – regresja wieloraka



STATISTICA – analiza regresji – regresja wieloraka



Dane: Reszty standaryz.: y (dane2 w wyklad_06.stw)

Reszty standaryz.: y (dane2 w wyklad_06.stw)
Odstające
Pomiń przypadki: 11

Przyp.	Reszty standaryz. -5. -4. -3. ±2. 3. 4. 5.	Obserw. Wartość	Przewidyw. Wartość	Reszta	Standard Przewid.	Standard Reszta	Bł. std. W.przew.
15	. . . * . . .	2,200000	2,580435	-0,380435	-0,744919	-2,00980	0,044109
24	. . . |* . . .	3,200000	2,771075	0,428925	-0,098133	2,26596	0,035325
Minimum	. . . * . . .	2,200000	2,580435	-0,380435	-0,744919	-2,00980	0,035325
Maksim.	. . . |* . . .	3,200000	2,771075	0,428925	-0,098133	2,26596	0,044109
Średnia	. . . |* . . .	2,700000	2,675755	0,024245	-0,421526	0,12808	0,039717
Mediana	. . . |* . . .	2,700000	2,675755	0,024245	-0,421526	0,12808	0,039717

Obserwacja 15 i 24 są wskazywane jako obserwacje odstające (wielkość reszt standaryzowanych tych obserwacji nie przekracza jednak 3). Analiza reszt wskazuje więc, że otrzymany model (bez uwzględnienia obserwacji 11) jest akceptowalny.