

Dobór funkcji regresji

Funkcja regresji powinna w możliwie jak największym stopniu wyjaśniać zmienność zmiennej zależnej, jednocześnie jednak powinna mieć możliwie najprostszą strukturę. Dobór modelu funkcji regresji nie jest sprawą prostą, stosowane są różne strategie.

Regresja krokowa

Pozwala na sekwencyjny dobór zmiennych podczas konstruowania funkcji regresji, stosowane są dwa podejścia:

- selekcja postępująca
konstruowanie modelu rozpoczyna się od jednej zmiennej niezależnej, w kolejnych krokach dodawane są kolejne zmienne,
- eliminacja wsteczna
poszukiwanie optymalnego modelu rozpoczyna się od modelu maksymalnego a w kolejnych krokach kolejno usuwane są zmienne o najmniejszym wpływie na zmienną zależną, proces eliminacji kończy się gdy w modelu występują wyłącznie zmienne, które w istotny sposób wyjaśniają zmienność zmiennej zależnej.

W regresji krokowej wprowadzanie lub usuwanie zmiennych odbywa się w oparciu o progowe wartości poziomów istotności współczynników modelu lub w oparciu o wartości progowe tzw. **cząstkowego testu F** – proces uzupełniania modelu kończy się gdy nie można znaleźć zmiennej spełniającej nałożone warunki. Jeżeli istnieje kilka zmiennych które można dodać (usunąć) z modelu wybierana jest zmienna dająca największy poziom istotności współczynnika lub testu F (przy usuwaniu: zmienna o najmniejszym poziomie istotności lub najmniejszej wartości testu F).

Funkcję regresji można zapisać w postaci:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}, \quad (1)$$

W celu zbadania czy wkład określonego zbioru zmiennych jest istotny, wektor współczynników modelu i macierz wejść są rozdzielane w sposób następujący:

$$\mathbf{y} = [\mathbf{X}_1 \quad \mathbf{X}_2] \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix} + \mathbf{e}, \quad (2)$$

gdzie: \mathbf{X}_1 i \mathbf{X}_2 to kolumny macierzy \mathbf{X} odpowiadające współczynnikom \mathbf{b}_1 i \mathbf{b}_2 .

Zmienne uznaje się za istotne jeżeli związane z nimi współczynniki są istotnie różne od zera. Istotność współczynników bada się testując hipotezę o ich braku istotności:

$$H_0: \mathbf{b}_1 = \mathbf{0}$$

wobec hipotezy alternatywnej:

$$H_1: \mathbf{b}_1 \neq \mathbf{0}.$$

Jeżeli współczynniki \mathbf{b}_1 nie są istotne, model (1) redukuje się do postaci:

$$\mathbf{y} = \mathbf{X}_2 \mathbf{b}_2 + \mathbf{e}, \quad (3)$$

Modele (1) i (2) to tzw. *modele pełne*, model (3) to *model zredukowany* lub *częściowy*.



Jakość dopasowania funkcji regresji: model pełny i zredukowany

Dopasowanie danych empirycznych do modelu można mierzyć wielkością błędu SS_e – im mniejszy błąd tym lepsze dopasowanie. Błędy dla modeli pełnego $SS_e(\mathbf{b})$ i zredukowanego $SS_e(\mathbf{b}_2)$ spełniają zależność:

$$SS_e(\mathbf{b}) \leq SS_e(\mathbf{b}_2)^*.$$

**Własność ta jest konsekwencją metody wyznaczania współczynników regresji – współczynniki te wyznaczone są w taki sposób aby minimalizować błąd SS_e , minimalizacja błędu na większej dziedzinie $SS_e(\mathbf{b})$ nie może dać gorszego wyniku niż minimalizacja na błędzie na podzbiorze tej dziedziny $SS_e(\mathbf{b}_2)$ – znalezione zostanie rozwiązanie dające błąd taki sam lub mniejszy.*

Wnioski

Uzupełnianie modelu o kolejne zmienne prowadzi do:

- zmniejszania błędu SS_e ,
- zwiększania wartości SS_r ** (zmienności wyjaśnionej modelem)

***Zmienność całkowita zmiennej zależnej SS_T jest niezależna od przyjętego modelu, więc skoro zachodzi zależność $SS_T = SS_r + SS_e$ to zmniejszanie wartości błędu SS_e prowadzi do zwiększania wartości SS_r .*

Dekompozycja całkowitej zmienności zmiennej zależnej SS_T dla modelu pełnego oznacza, że:

$$SS_T = SS_r(\mathbf{b}) + SS_e(\mathbf{b}),$$

a dla modelu zredukowanego:

$$SS_T = SS_r(\mathbf{b}_2) + SS_e(\mathbf{b}_2),$$

Zmienność całkowita w przypadku redukcji modelu nie ulega zmianie, więc:

$$SS_T = SS_r(\mathbf{b}) + SS_e(\mathbf{b}) = SS_r(\mathbf{b}_2) + SS_e(\mathbf{b}_2).$$

Ocenę wpływu dodatkowych zmiennych umożliwia tzw. *dodatkowa suma kwadratów* (ang. *extra sum of squares*) definiowana jako:

$$SS_r(\mathbf{b}_1|\mathbf{b}_2) = SS_r(\mathbf{b}) - SS_r(\mathbf{b}_2) = SS_e(\mathbf{b}_2) - SS_e(\mathbf{b}),$$

gdzie: $SS_r(\mathbf{b}_1|\mathbf{b}_2)$ to zmienność wyjaśniona równaniem regresji po wprowadzeniu współczynników \mathbf{b}_1 pod warunkiem, że model zawierał już współczynniki \mathbf{b}_2 , zmienna $SS_r(\mathbf{b}_1|\mathbf{b}_2)$ ma k stopni swobody (k to liczba współczynników w wektorze \mathbf{b}_1).

Do przeprowadzenia testu istotności współczynników \mathbf{b}_1 wykorzystuje się, statystykę F_p :

$$F_p = \frac{SS_r(\mathbf{b}_1|\mathbf{b}_2)}{k} \bigg/ \frac{SS_e(\mathbf{b})}{n-p-1} = \frac{SS_r(\mathbf{b}_1|\mathbf{b}_2)/k}{MS_e(\mathbf{b})}.$$

Przykład 1. Celem badań doświadczalnych jest określenie funkcji obiektu badań

$\hat{y} = f(x_1, x_2)$ przy założeniu, że:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + b_3x_1x_2 + b_4x_1^2 + b_5x_2^2, \quad 1 \leq x_1 \leq 3, \quad 0 \leq x_2 \leq 4.$$

Planując eksperyment zdecydowano o przeprowadzeniu 9 doświadczeń: po jednym na poziomie minimalnym maksymalnym i pośrednim dla każdej zmiennej wejściowej.

Funkcja obiektu będzie skonstruowana z wykorzystaniem regresji krokowej postępującej.

Lp.	x_1	x_2	y
1	1	0	2
2	1	2	8
3	1	4	18
4	2	0	4
5	2	2	8
6	2	4	20
7	3	0	5
8	3	2	8
9	3	4	22

Krok 0. model: $\hat{y} = b_0$

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 2 \\ 8 \\ 18 \\ 4 \\ 8 \\ 20 \\ 5 \\ 8 \\ 22 \end{bmatrix}$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx 10,56,$$

$$SS_e(\hat{b}_0) \approx 442,22,$$

$$MS_e(\hat{b}_0) \approx 55,28,$$

Krok 1. Do modelu można wprowadzić zmienną:

a) x_1 , (tzn.: model: $\hat{y} = b_0 + b_1 x_1$)

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3 \\ 1 & 3 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 8,22 \\ 1,17 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_1) \approx 434,06, \quad MS_e(\hat{b}_0, \hat{b}_1) \approx 62,01,$$

$$F_p = \frac{SS_r(\hat{b}_1 | \hat{b}_0)}{MS_e(\hat{b}_0, \hat{b}_1)} = \frac{SS_e(\hat{b}_0) - SS_e(\hat{b}_0, \hat{b}_1)}{MS_e(\hat{b}_0, \hat{b}_1)} \approx \frac{442,22 - 434,06}{62,01} \approx 0,13.$$

b) x_2 , (tzn.: model: $\hat{y} = b_0 + b_2 x_2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 0 \\ 1 & 2 \\ 1 & 4 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 8,22 \\ 1,17 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_2) \approx 42,06, \quad MS_e(\hat{b}_0, \hat{b}_2) \approx 6,01,$$

$$F_p = \frac{SS_r(\hat{b}_2 | \hat{b}_0)}{MS_e(\hat{b}_0, \hat{b}_2)} = \frac{SS_e(\hat{b}_0) - SS_e(\hat{b}_0, \hat{b}_2)}{MS_e(\hat{b}_0, \hat{b}_2)} \approx \frac{442,22 - 42,06}{6,01} \approx 66,58.$$

Krok 1. Do modelu można wprowadzić:

c) interakcję zmiennych x_1 i x_2 , (tzn.: model: $\hat{y} = b_0 + b_3 x_1 x_2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 2 \\ 1 & 4 \\ 1 & 0 \\ 1 & 4 \\ 1 & 8 \\ 1 & 0 \\ 1 & 6 \\ 1 & 12 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 4,32 \\ 1,56 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_3) \approx 111,75, \quad MS_e(\hat{b}_0, \hat{b}_3) \approx 15,96,$$

$$F_p = \frac{SS_r(\hat{b}_3 | \hat{b}_0)}{MS_e(\hat{b}_0, \hat{b}_3)} = \frac{SS_e(\hat{b}_0) - SS_e(\hat{b}_0, \hat{b}_3)}{MS_e(\hat{b}_0, \hat{b}_3)} \approx \frac{442,22 - 111,75}{15,96} \approx 20,71.$$

d) kwadrat zmiennej x_1 , (tzn.: model: $\hat{y} = b_0 + b_4 x_1^2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 4 \\ 1 & 4 \\ 1 & 4 \\ 1 & 9 \\ 1 & 9 \\ 1 & 9 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 9,24 \\ 0,28 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_4) \approx 434,41, \quad MS_e(\hat{b}_0, \hat{b}_4) \approx 62,06,$$

$$F_p = \frac{SS_r(\hat{b}_4 | \hat{b}_0)}{MS_e(\hat{b}_0, \hat{b}_4)} = \frac{SS_e(\hat{b}_0) - SS_e(\hat{b}_0, \hat{b}_4)}{MS_e(\hat{b}_0, \hat{b}_4)} \approx \frac{442,22 - 434,41}{62,06} \approx 0,13.$$

Krok 1. Do modelu można wprowadzić:

e) kwadrat zmiennej x_2 , (tzn.: model: $\hat{y} = b_0 + b_5 x_2^2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 1 & 16 \\ 1 & 0 \\ 1 & 4 \\ 1 & 16 \\ 1 & 0 \\ 1 & 4 \\ 1 & 16 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 3,78 \\ 1,02 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_5) \approx 12,78, \quad MS_e(\hat{b}_0, \hat{b}_5) \approx 1,83,$$
$$F_p = \frac{SS_r(\hat{b}_5 | \hat{b}_0)}{MS_e(\hat{b}_0, \hat{b}_5)} = \frac{SS_e(\hat{b}_0) - SS_e(\hat{b}_0, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_5)} \approx \frac{442,22 - 12,78}{1,83} \approx 235,18.$$

Podsumowanie: wprowadzanie dodatkowych zmiennych do modelu dało następujące wartości częściowego testu F :

- zmienna x_1 : $F_p \approx 0,13$,
- zmienna x_2 : $F_p \approx 66,58$,
- interakcja zmiennych: $F_p \approx 20,71$,
- kwadrat zmiennej x_1 : $F_p \approx 0,13$,
- kwadrat zmiennej x_2 : $F_p \approx 235,18$.

Największą wartość testu otrzymano w wyniku wprowadzenia do modelu kwadratu zmiennej x_2 , model zostanie więc uzupełniony o tę zmienną.



Krok 2. Do modelu można wprowadzić zmienną:

a) x_1 , (tzn.: model: $\hat{y} = b_0 + b_5x_2^2 + b_1x_1$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 4 & 1 \\ 1 & 16 & 1 \\ 1 & 0 & 2 \\ 1 & 4 & 2 \\ 1 & 16 & 2 \\ 1 & 0 & 3 \\ 1 & 4 & 3 \\ 1 & 16 & 3 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 1,45 \\ 1,02 \\ 1,17 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5) \approx 4,62, \quad MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5) \approx 0,77,$$

$$F_p = \frac{SS_r(\hat{b}_1 | \hat{b}_0, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5)} = \frac{SS_e(\hat{b}_0, \hat{b}_5) - SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5)} \approx \frac{12,78 - 4,62}{0,77} \approx 10,6.$$

b) x_2 , (tzn.: model: $\hat{y} = b_0 + b_5x_2^2 + b_2x_2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 4 & 2 \\ 1 & 16 & 4 \\ 1 & 0 & 0 \\ 1 & 4 & 2 \\ 1 & 16 & 4 \\ 1 & 0 & 0 \\ 1 & 4 & 2 \\ 1 & 16 & 4 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 3,67 \\ 0,96 \\ 0,25 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_2, \hat{b}_5) \approx 12,67, \quad MS_e(\hat{b}_0, \hat{b}_2, \hat{b}_5) \approx 2,11,$$

$$F_p = \frac{SS_r(\hat{b}_2 | \hat{b}_0, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_2, \hat{b}_5)} = \frac{SS_e(\hat{b}_0, \hat{b}_5) - SS_e(\hat{b}_0, \hat{b}_2, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_2, \hat{b}_5)} \approx \frac{12,78 - 12,67}{2,11} \approx 0,05.$$

Krok 2. Do modelu można wprowadzić:

c) interakcję zmiennych x_1 i x_2 , (tzn.: model: $\hat{y} = b_0 + b_5x_2^2 + b_3x_1x_2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 4 & 2 \\ 1 & 16 & 4 \\ 1 & 0 & 0 \\ 1 & 4 & 4 \\ 1 & 16 & 8 \\ 1 & 0 & 0 \\ 1 & 4 & 6 \\ 1 & 16 & 12 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 1,45 \\ 1,02 \\ 1,17 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_3, \hat{b}_5) \approx 6,74, \quad MS_e(\hat{b}_0, \hat{b}_3, \hat{b}_5) \approx 1,12,$$

$$F_p = \frac{SS_r(\hat{b}_3 | \hat{b}_0, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_3, \hat{b}_5)} = \frac{SS_e(\hat{b}_0, \hat{b}_5) - SS_e(\hat{b}_0, \hat{b}_3, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_3, \hat{b}_5)} \approx \frac{12,78 - 6,74}{1,12} \approx 5,39$$

d) kwadrat zmiennej x_1 , (tzn.: model: $\hat{y} = b_0 + b_5x_2^2 + b_4x_1^2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 4 & 1 \\ 1 & 16 & 1 \\ 1 & 0 & 4 \\ 1 & 4 & 4 \\ 1 & 16 & 4 \\ 1 & 0 & 9 \\ 1 & 4 & 9 \\ 1 & 16 & 9 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 2,46 \\ 1,02 \\ 0,28 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_4, \hat{b}_5) \approx 4,97, \quad MS_e(\hat{b}_0, \hat{b}_4, \hat{b}_5) \approx 0,83,$$

$$F_p = \frac{SS_r(\hat{b}_4 | \hat{b}_0, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_4, \hat{b}_5)} = \frac{SS_e(\hat{b}_0, \hat{b}_5) - SS_e(\hat{b}_0, \hat{b}_4, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_4, \hat{b}_5)} \approx \frac{12,78 - 4,97}{0,83} \approx 9,43.$$

Krok 2.

Podsumowanie: wprowadzanie dodatkowych zmiennych do modelu dało następujące wartości cząstkowego testu F :

- zmienna x_1 : $F_p \approx 10,6$,
- zmienna x_2 : $F_p \approx 0,05$,
- interakcja zmiennych: $F_p \approx 5,39$,
- kwadrat zmiennej x_1 : $F_p \approx 9,43$.

Największą wartość testu otrzymano w wyniku wprowadzenia do modelu zmiennej x_1 , model zostanie więc uzupełniony o tą zmienną.

Krok 3. Do modelu można wprowadzić zmienną:

a) x_2 , (tzn.: model: $\hat{y} = b_0 + b_5 x_2^2 + b_1 x_1 + b_2 x_2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 4 & 1 & 2 \\ 1 & 16 & 1 & 4 \\ 1 & 0 & 2 & 0 \\ 1 & 4 & 2 & 2 \\ 1 & 16 & 2 & 4 \\ 1 & 0 & 3 & 0 \\ 1 & 4 & 3 & 2 \\ 1 & 16 & 3 & 4 \end{bmatrix} \quad \hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 1,33 \\ 0,96 \\ 1,17 \\ 0,25 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_5) \approx 4,5, \quad MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_5) \approx 0,9,$$

$$F_p = \frac{SS_r(\hat{b}_2 | \hat{b}_0, \hat{b}_1, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_5)} = \frac{SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5) - SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_2, \hat{b}_5)} \approx \frac{4,62 - 4,5}{0,9} \approx 0,13.$$

Krok 3. Do modelu można wprowadzić zmienną:

b) interakcję zmiennych x_1 i x_2 , (tzn.: model: $\hat{y} = b_0 + b_5x_2^2 + b_1x_1 + b_3x_1x_2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 4 & 1 & 2 \\ 1 & 16 & 1 & 4 \\ 1 & 0 & 2 & 0 \\ 1 & 4 & 2 & 4 \\ 1 & 16 & 2 & 8 \\ 1 & 0 & 3 & 0 \\ 1 & 4 & 3 & 6 \\ 1 & 16 & 3 & 12 \end{bmatrix}$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 1,83 \\ 0,96 \\ 0,92 \\ 0,13 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_3, \hat{b}_5) \approx 4,25, \quad MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_3, \hat{b}_5) \approx 0,85,$$

$$F_p = \frac{SS_r(\hat{b}_3 | \hat{b}_0, \hat{b}_1, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_3, \hat{b}_5)} = \frac{SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5) - SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_3, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_3, \hat{b}_5)} \approx \frac{4,62 - 4,25}{0,85} \approx 0,43.$$

c) kwadrat zmiennej x_1 , (tzn.: model: $\hat{y} = b_0 + b_5x_2^2 + b_1x_1 + b_4x_1^2$)

$$\mathbf{X} = \begin{bmatrix} 1 & 0 & 1 & 1 \\ 1 & 4 & 1 & 1 \\ 1 & 16 & 1 & 1 \\ 1 & 0 & 2 & 4 \\ 1 & 4 & 2 & 4 \\ 1 & 16 & 2 & 4 \\ 1 & 0 & 3 & 9 \\ 1 & 4 & 3 & 9 \\ 1 & 16 & 3 & 9 \end{bmatrix}$$

$$\hat{\mathbf{b}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \approx \begin{bmatrix} 0,89 \\ 1,02 \\ 1,83 \\ -0,17 \end{bmatrix}, \quad SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_4, \hat{b}_5) \approx 4,56, \quad MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_4, \hat{b}_5) \approx 0,91,$$

$$F_p = \frac{SS_r(\hat{b}_4 | \hat{b}_0, \hat{b}_1, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_4, \hat{b}_5)} = \frac{SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5) - SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_4, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_4, \hat{b}_5)} \approx \frac{4,62 - 4,56}{0,91} \approx 0,06.$$

Krok 3. Podsumowanie: wprowadzanie zmiennych do modelu dało następujące wartości testu F :

- zmienna x_2 : $F_p \approx 0,13$,
- interakcja zmiennych: $F_p \approx 0,43$,
- kwadrat zmiennej x_1 : $F_p \approx 0,06$.

Największą wartość testu otrzymano w wyniku wprowadzenia do modelu interakcji zmiennych.

Zakładając jednak, że ustalono graniczną (równą 1) wartość, którą powinien osiągać test – żadna ze zmiennych nie zostanie dodana do modelu a proces konstruowania modelu został zakończony.

Ostatecznie wyznaczony został model:

$$\hat{y} = \hat{b}_0 + \hat{b}_1 x_1 + \hat{b}_5 x_2^2 = 1,45 + 1,17 x_1 + 1,02 x_2^2.$$

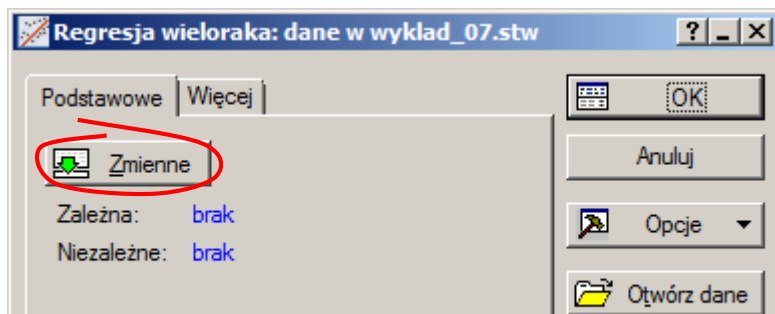
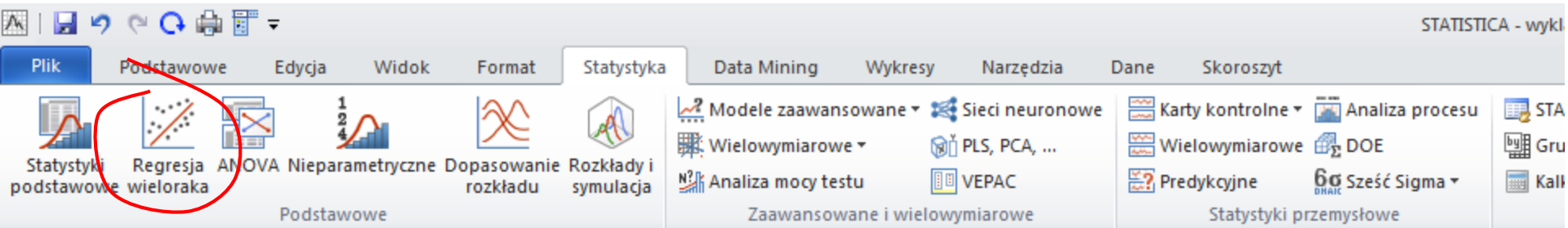
Znaleziona funkcja jest statystycznie istotna;

$$SS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5) \approx 4,62, \quad MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5) \approx 0,77, \quad SS_r(\hat{b}_0, \hat{b}_1, \hat{b}_5) \approx 437,6, \quad MS_r(\hat{b}_0, \hat{b}_1, \hat{b}_5) \approx 218,8,$$

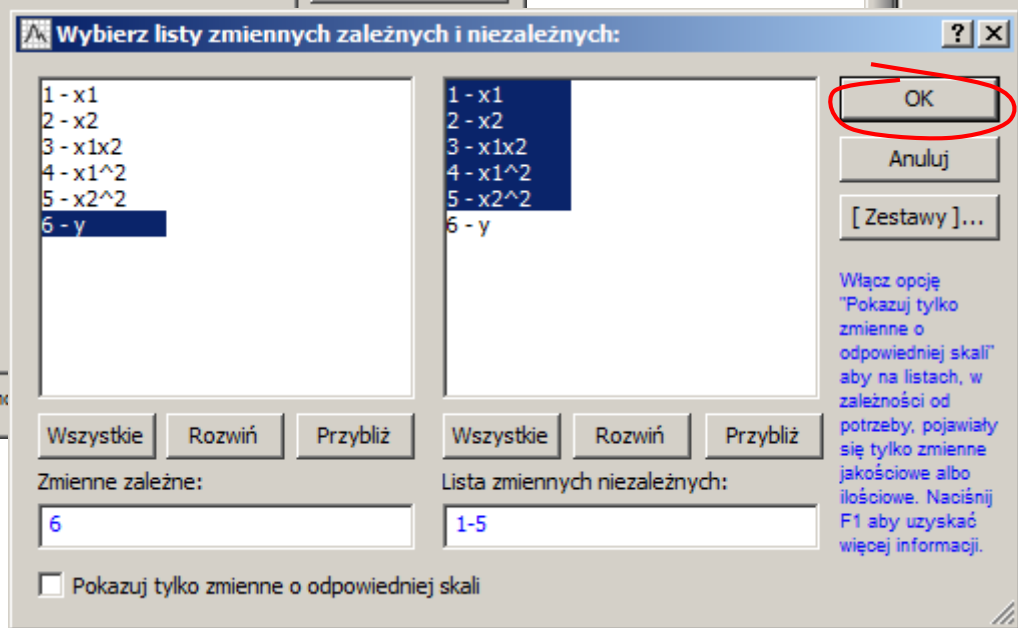
$$F_n = \frac{MS_r(\hat{b}_0, \hat{b}_1, \hat{b}_5)}{MS_e(\hat{b}_0, \hat{b}_1, \hat{b}_5)} \approx \frac{218,8}{0,77} \approx 284,4.$$

$$p\text{-value} = 1 - F_{F(p, n-p-1)}(F_n) = 1 - F_{F(2, 9-2-1)}(218,8) \approx 0,000001.$$

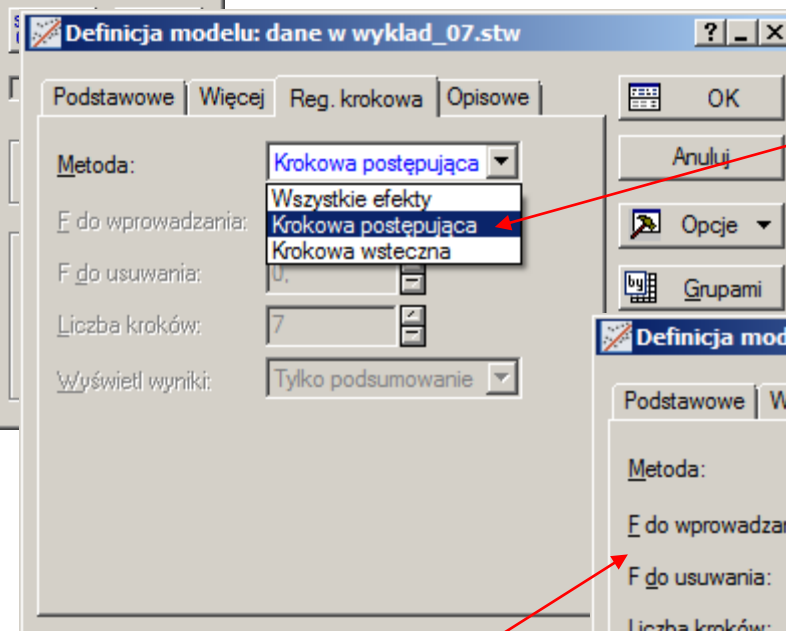
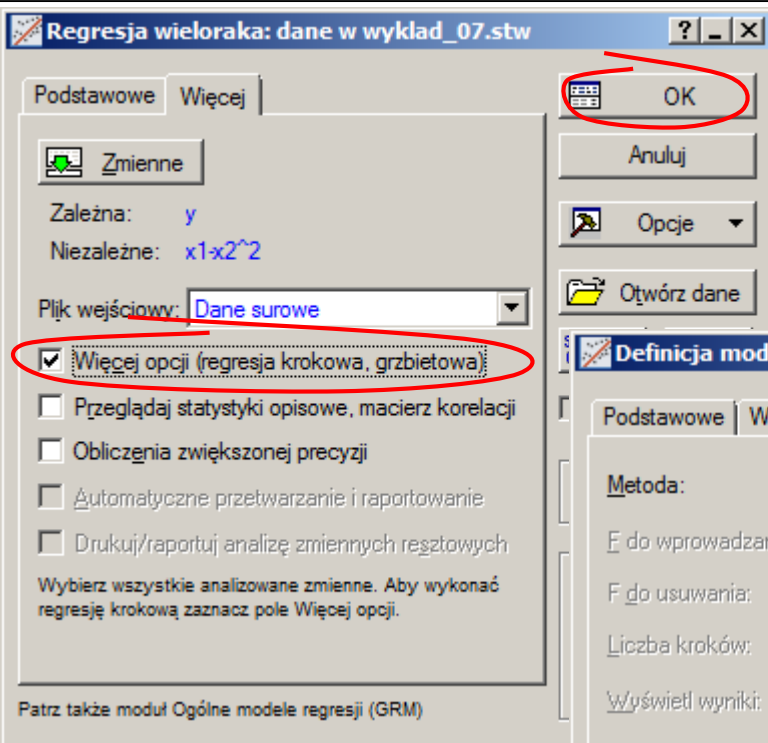
STATISTICA – regresja krokowa



	1	2	3	4	5	6
	x1	x2	x1x2	x1^2	x2^2	y
1	1	0	0	1	0	2
2	1	2	2	1	4	8
3	1	4	4	1	16	18
4	2	0	0	4	0	4
5	2	2	4	4	4	8
6	2	4	8	4	16	20
7	3	0	0	9	0	5
8	3	2	6	9	4	8
9	3	4	12	9	16	22



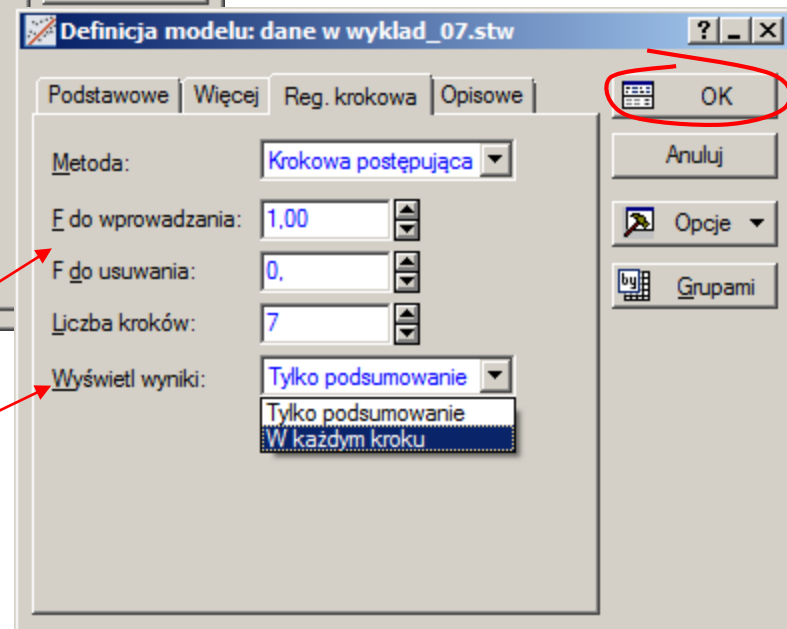
STATYSTICA – regresja krokowa



sposób konstruowania modelu

wartości progowe cząstkowego testu F

sposób wyświetlania wyników



Wyniki regresji wielorakiej: dane w wyklad_07.stw

Wyniki regresji wielorakiej (Krok 0)

Zmn. zależ.y	Wielor. R = 0,0000000	F = 0,00000
	R ² = 0,0000000	df = 0,8
Liczba przyp. 9	Popraw. R ² = 0,0000000	p = -0,00000
Błąd standardowy estymacji: 7,434902674		

Wyniki regresji wielorakiej: dane w wyklad_07.stw

Wyniki regresji wielorakiej (Krok 1)

Zmn. zależ.y	Wielor. R = ,98544196	F = 235,1799
	R ² = ,97109586	df = 1,7
Liczba przyp. 9	Popraw. R ² = ,96696670	p = ,000001
Błąd standardowy estymacji: 1,351298385		
Wyr. wolny 3,782051282	Błąd std.: ,6308534	t(7) = 5,9951 p = ,0005

$x^2 \cdot b^* = ,985$

(istotne b* są podświetlone na czerwono)

Alfa do podświetlania efektów: .05

Podstawowe Więcej Reszty, założenia, predykcja

Podsumowanie: Wyniki regresji	Korelacje cząstkowe
ANOVA (sum. dobroć dopasow.)	Nadmiarowość
Kowariancja współczynników	Podsumowanie r. krokowej
Aktualna macierz wymiany	ANOVA poprawiona na średnią

nnych

Następny

Anuluj

Opcje

Grupami

stkowe

wość

krokowej

a na średnią

STATYSTICA – regresja krokowa

Wyniki regresji wielorakiej: dane w wyklad_07.stw

Wyniki regresji wielorakiej (krok 2 rozwiązanie)
 żadne inne F do wpraw. nie przekracza progu

Zmn. zależ. y Wielokr. R = ,99476791 F = 284,4445
 R^2= ,98956320 df = 2,6

Liczba przyp. 9 Popraw. R^2= ,98608427 p = ,000001
 Błąd standardowy estymacji: ,877058019

Wyr. wolny 1,448717949 Błąd std.: ,8249081 t(6) = 1,7562 p = ,1296

x2^2 b* = ,985

x1 b* = ,136

(istotne b* są podświetlone na czerwono)

Alfa do podświetlania efektów: .05

Podstawowe Więcej Reszty, założenia, predykcja

Podsumowanie: Wyniki regresji 1

Korelacje częściowe

ANOVA (sum. dobroć dopasow.) Nadmiarowość

Kowariancja współczynników Podsumowanie r. krokowej 2

Aktualna macierz wymiany ANOVA poprawiona na średnią

Dane: Podsumowanie regresji zmiennej zależnej: y (dane w wyklad_07.stw)

Podsumowanie regresji zmiennej zależnej: y (dane w wyklad_07.stw)
 R= ,99476791 R^2= ,98956320 Popraw. R2= ,98608427
 F(2,6)=284,44 p<,00000 Błąd std. estymacji: ,87706

	b*	Bł. std. z b*	b	Bł. std. z b	t(6)	p
N=9						
W. wolny			1,448718	0,824908	1,75622	0,129571
x2^2	0,985442	0,041707	1,016026	0,043001	23,62778	0,000000
x1	0,135895	0,041707	1,166667	0,358057	3,25832	0,017285

Dane: Podsumowanie regresji krokowej; DV: y (dane w wyklad_07.stw)

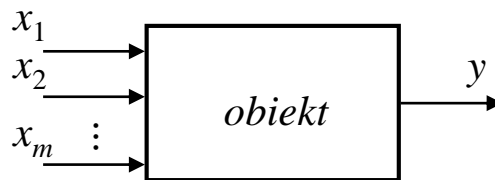
Podsumowanie regresji krokowej; DV: y (dane w wyklad_07.stw)

Zmienna	Krok +do/-z	Wielokr. Spearman	Wielokr. R-kwadr.	R-kwadr. zmiana	F - do wpraw/us	p	Zmienne zawarte
x2^2	1	0,985442	0,971096	0,971096	235,1799	0,000001	1
x1	2	0,994768	0,989563	0,018467	10,6167	0,017285	2

Zmienne losowe mogą być:

- *mieralne (ilościowe)* – przyjmują wartości ze zbioru liczbowego,
- *niemierzalne (jakościowe)* – cechy których nie można wyrazić ilościowo, są opisywane słownie lub wyrażane przy pomocy wybranej skali.

Analizy wariancji i regresji wykorzystywane są do konstruowania modeli *zmiennej objaśnianej* w funkcji *zmiennych objaśniających*:



Analiza wariancji

- *zmienna objaśniana*: zmienna ilościowa,
- *zmiennne objaśniające*: zmiennne jakościowe

Analiza regresji

- *zmienna objaśniana*: zmienna ilościowa,
- *zmiennne objaśniające*: zmiennne ilościowe

Jednoczynnikowa analiza wariancji – podejście regresyjne

W eksperymencie badana jest istotność wpływu *jednej zmiennej niezależnej* na *jedną zmienną zależną* w przypadku, gdy zmienna niezależna może przyjmować *wartości na kilku poziomach*. Dla wyników otrzymanych w eksperymencie tworzony jest model w postaci:

$$y_{ij} = \mu_i + e_{ij}, \quad (*)$$

gdzie y_{ij} to wynik j -tej powtórki doświadczenia przeprowadzonego na i -tym poziomie, μ_i to średnia wartość zmiennej wyjściowej dla i -tego poziomu, e_{ij} – błąd losowy,

po wprowadzeniu ogólnej średniej μ i efektu τ_i i -tego poziomu czynnika (spełniającego założenie $\sum \tau_i = 0$) model ten może być zapisany w postaci:

$$y_{ij} = \mu + \tau_i + e_{ij}. \quad (**)$$

W modelu (*) występuje a parametrów: μ_1, \dots, μ_a , w modelu (**): $(a+1)$ parametrów: $\mu, \tau_1, \dots, \tau_a$ stąd założenie: $\sum \tau_i = 0$.

poziom zmiennej niezależnej	numer doświadczenia			
	1	2	...	r
1				
⋮				
a				

Jednoczynnikowa analiza wariancji – podejście regresyjne (kodowanie zmiennych jakościowych)

Wprowadzając $(a-1)$ fikcyjnych zmiennych ilościowych (tzw. zmienne zero-jedynkowe):

$$x_i = \begin{cases} 1, & \text{gdy zmienna jakościowa jest na poziomie } i, \\ 0, & \text{w przeciwnym przypadku} \end{cases}$$

tworzona jest funkcja regresji: $y = b_0 + b_1x_1 + \dots + b_{a-1}x_{a-1} + e$.

Wyniki eksperymentu można zapisać w postaci macierzy wejść i wektora obserwacji:

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & \vdots & 0 \\ 1 & 0 & 1 & \vdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \vdots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & \vdots & 0 \\ 1 & 0 & 1 & \vdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & \vdots & 0 \end{bmatrix} \begin{cases} 1 \text{ powtórka doświadczeń} \\ \\ \\ r\text{-ta powtórka doświadczeń} \end{cases} = \mathbf{y} = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{a1} \\ \dots \\ y_{1r} \\ y_{2r} \\ \vdots \\ y_{ar} \end{bmatrix}$$

zmienna x_0 odpowiadająca wyrazowi wolnemu
zmiennne x_1, \dots, x_{a-1}

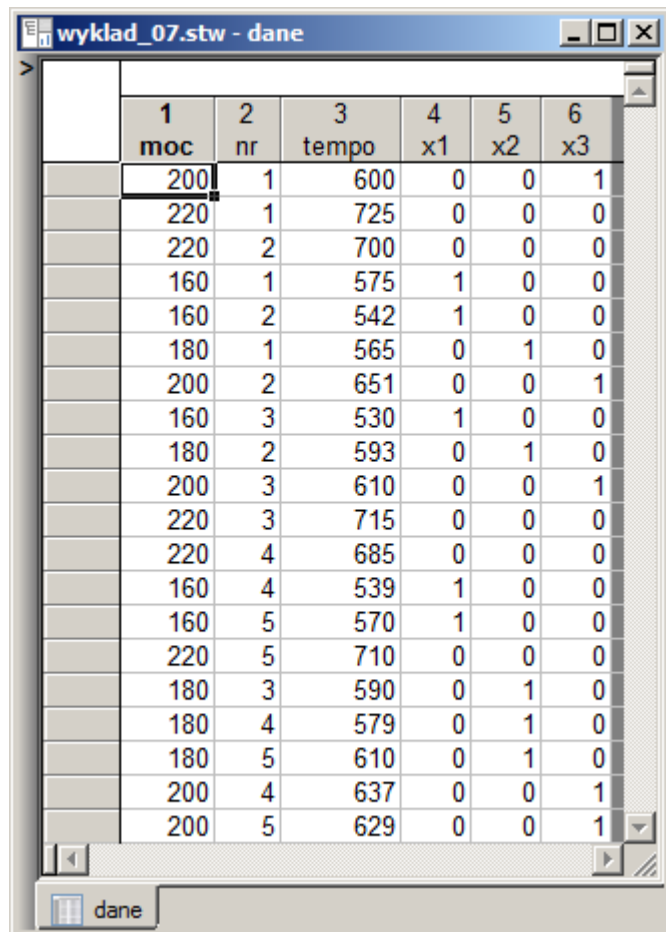
	numer doświadczenia			
	1	2	...	r
1				
⋮				
a				



Badanie istotności wpływu – przykład 1. (wykład 3)

Przykład 1.* Należy zbadać wpływ mocy reaktora plazmowego na szybkość trawienia płytek krzemowych. Planując eksperyment zdecydowano o wyborze 4 poziomów mocy: 160, 180, 200 i 220W i 5 doświadczeń dla każdego z ustalonych poziomów mocy.

Po zaplanowaniu eksperymentu i ustaleniu kolejności prowadzenia poszczególnych doświadczeń wyniki uzyskanych szybkości trawienia w [$\text{\AA}/\text{min}$] zapisano w arkuszu:



	1	2	3	4	5	6
	moc	nr	tempo	x1	x2	x3
	200	1	600	0	0	1
	220	1	725	0	0	0
	220	2	700	0	0	0
	160	1	575	1	0	0
	160	2	542	1	0	0
	180	1	565	0	1	0
	200	2	651	0	0	1
	160	3	530	1	0	0
	180	2	593	0	1	0
	200	3	610	0	0	1
	220	3	715	0	0	0
	220	4	685	0	0	0
	160	4	539	1	0	0
	160	5	570	1	0	0
	220	5	710	0	0	0
	180	3	590	0	1	0
	180	4	579	0	1	0
	180	5	610	0	1	0
	200	4	637	0	0	1
	200	5	629	0	0	1

Znaczenie zmiennych:

nr

numer kolejny doświadczenia na podanym poziomie mocy,

tempo

zmierzona szybkość trawienia,

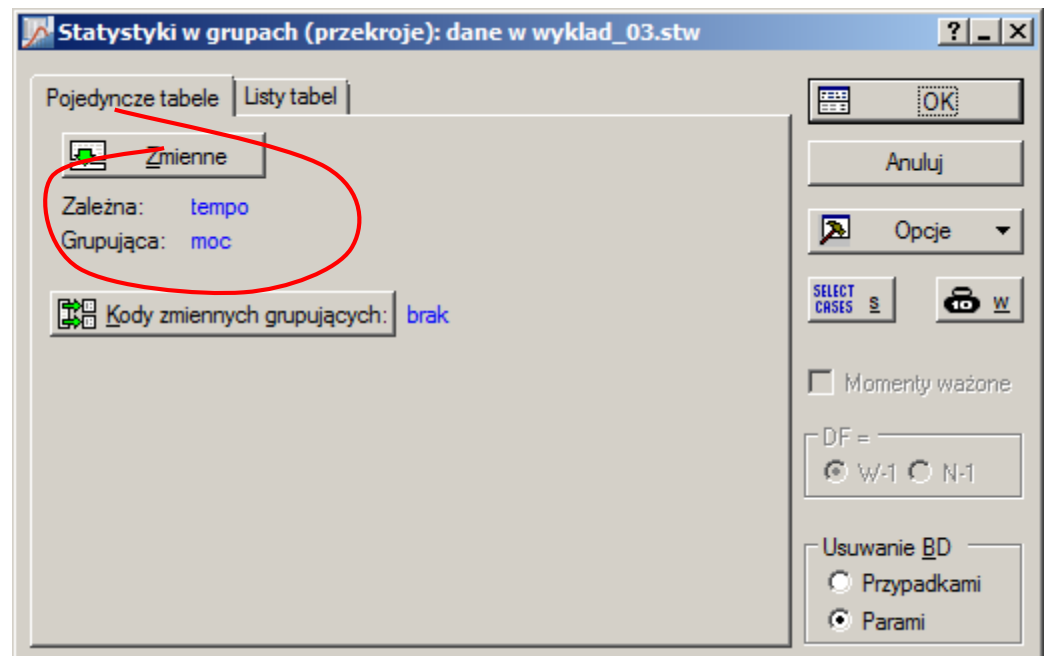
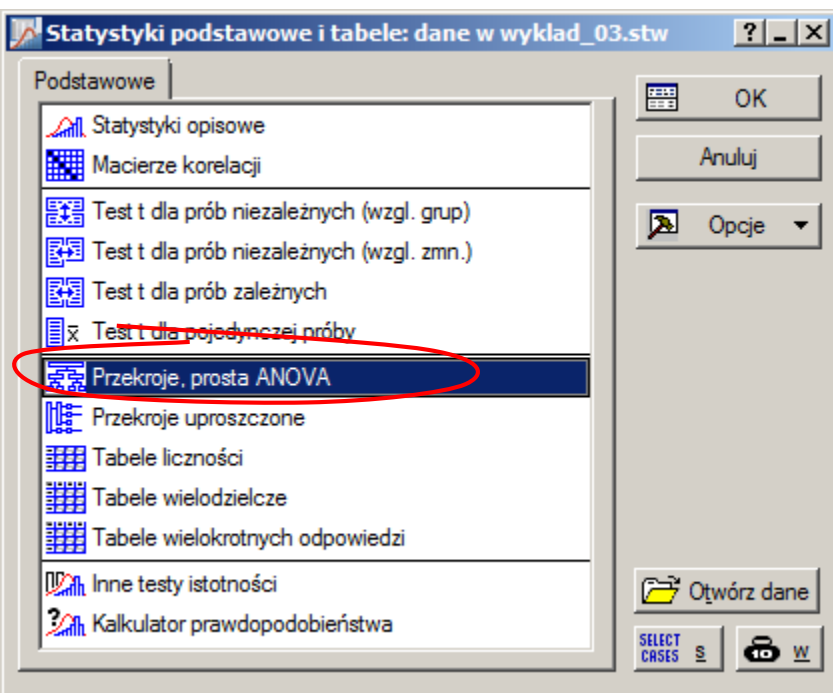
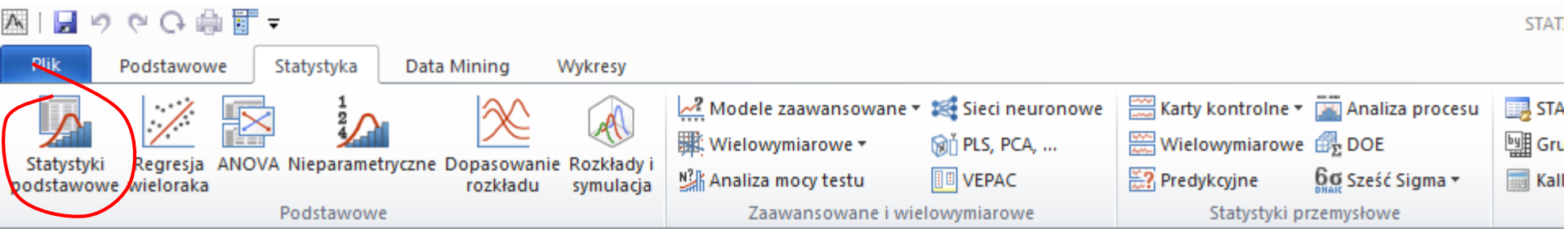
x1, x2, x3

fikcyjne zmienne zero-jedynkowe wykorzystywane w podejściu regresyjnym, dla mocy:

- 160: $x1 = 1, x2 = 0, x3 = 0,$
- 180: $x1 = 0, x2 = 1, x3 = 0,$
- 200: $x1 = 0, x2 = 0, x3 = 1,$
- 220: $x1 = 0, x2 = 0, x3 = 0.$

* Montgomery D. C., *Design and Analysis of Experiments*, Wiley, 2012

STATISTICA – jednoczynnikowa analiza wariancji (wykład 3)



STATYSTICA – jednoczynnikowa analiza wariancji (metoda 1.)

Statystyki w grupach - wyniki: dane w wyklad_03.stw

ZALEŻNA:1 zmienna: tempo

Podstawowe | Statystyki opisowe | Testy ANOVA | Post-hoc

Podsum.: tabela statystyk **3** Wykresy interakcji

Dokładne tabele dwudzielcze Skategoryzow. wykresy ramka-wąsy

Analiza wariancji **1**

Statystyki w grupach - wyniki: Spreadsheet w wyklad_03b.stw

Zmienna: tempo

Test NIR lub porównanie zaplanowane **2**

Test Scheffé

Test Newman-Keulsa i rozstępy krytyczne poziom alfa dla rozstępów krytycznych: .050

Test Duncana wiel. rozstępów i rozstępy kryt.

Test rozsądnej istotnej różnicy (RIR) Tukeya

Test RIR Tukeya dla nierównych licznosci

Dodatkowe testy post hoc (Dunnette'a, Bonferroniego, układów złożonych) dostępne są w module GLM.

Dane: Analiza wariancji (dane w wyklad_03.stw)

Analiza wariancji (dane w wyklad_03.stw)
Zaznaczone efekty są istotne z $p < ,05000$ **1**

Zmienna	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F	p
tempo	66870,55	3	22290,18	5339,200	16	333,7000	66,79707	0,000000

Dane: Tabe...

Tabela przekrońw st:
N=20 (Zmienna: Zależ

moc	tempo Średnie
160	551,2000
180	587,4000
200	625,4000
220	707,0000
Ogół	617,7500

poziomy zmiennej wejściowej w istotny sposób wpływają na wartość zmiennej zależnej

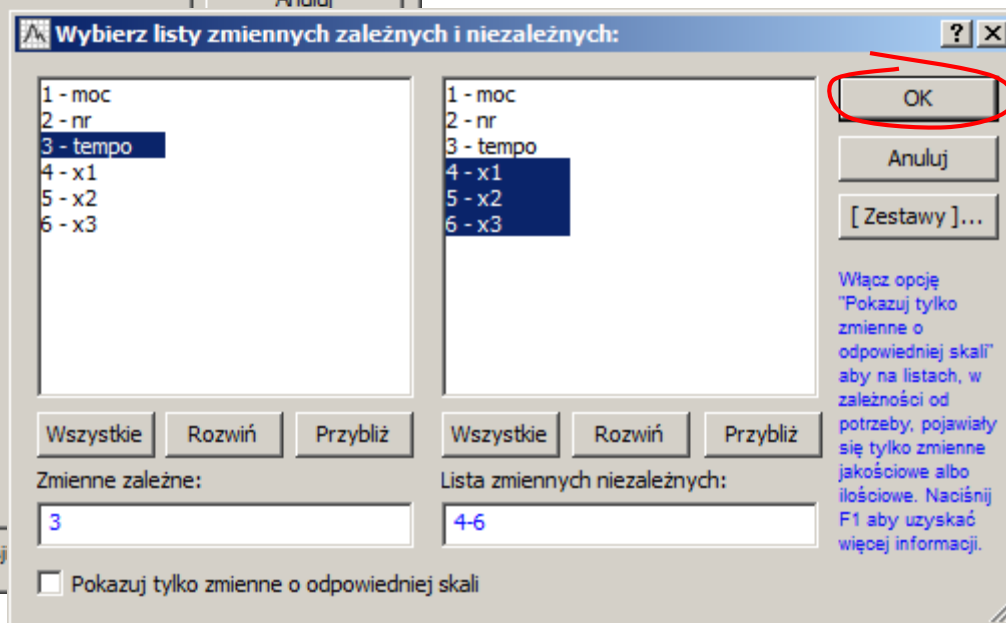
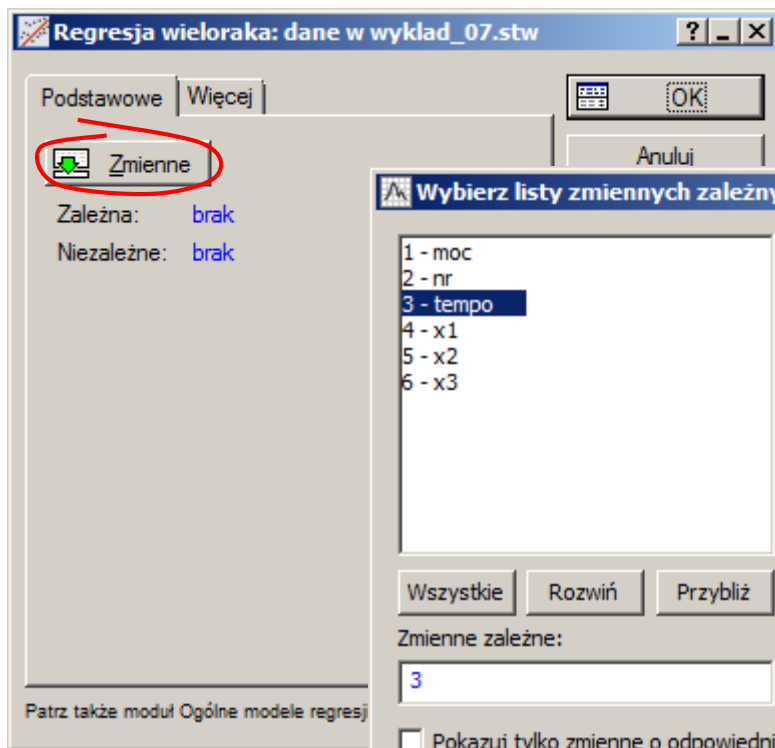
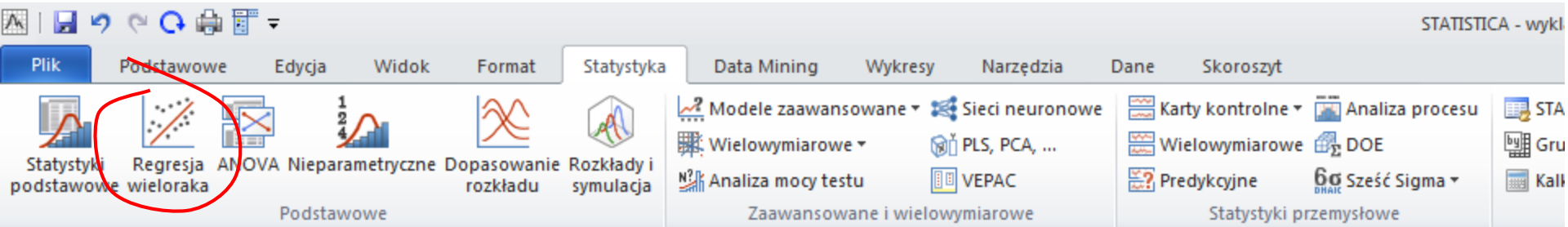
wszystkie poziomy zmiennej wejściowej wpływają na zmienną wyjściową w sposób istotny

Dane: Test NIR; Zmienna: (Spreadsheet w wyklad_03b.stw)

Test NIR; Zmienna: (Spreadsheet w wyklad_03b.stw)
Zaznaczone różnice są istotne z $p < ,05000$ **2**

		{1}	{2}	{3}	{4}
		M=551,20	M=587,40	M=625,40	M=707,00
160	{1}		0,006416	0,000008	0,000000
180	{2}	0,006416		0,004624	0,000000
200	{3}	0,000008	0,004624		0,000003
220	{4}	0,000000	0,000000	0,000003	

STATISTICA – analiza regresji



STATYSTICA – analiza regresji

Wyniki regresji wielorakiej: dane w wyklad_07.stw

Wyniki regresji wielorakiej

Zmn. zależ. tempo Wielor. R = ,96232003 F = 66,79707
 R^2= ,92605985 df = 3,16
 Liczba przyp. 20 Popraw. R^2= ,91219607 p = ,000000
 Błąd standardowy estymacji: 18,267457404
 Wyr. wolny 707,0000000 Błąd std.: 8,169456 t(16) = 86,542 p = 0,0000

x1 b*=-1,1 x2 b*=-,86 x3 b*=-,59

(istotne b* są podświetlone na czerwono)

funkcja regresji istotna

Dane: Analiza wariancji ; DV: tempo (dane w wyklad_07.stw)

Efekt	Suma kwadrat.	df	Średnia kwadrat.	F	p
Regres.	66870,55	3	22290,18	66,79707	0,000000
Reszta	5339,20	16	333,70		
Razem	72209,75				

Alfa do podświetlania efektów: .05

Podstawowe Więcej Reszty, założenia, predykcja

Podsumowanie: Wyniki regresji 1 Korelacje cząstkowe

ANOVA (sum. dobroć dopasow.) 2 Nadmiarowość

Kowariancja współczynników Podsumowanie r. krokowej

Anuluj Opcje Grupami

Dane: Podsumowanie regresji zmiennej zależnej: tempo (dane w wyklad_07.stw)

Podsumowanie regresji zmiennej zależnej: tempo (dane w wyklad_07.stw)
 R= ,96232003 R^2= ,92605985 Popraw. R^2= ,91219607
 F(3,16)=66,797 p<,00000 Błąd std. estymacji: 18,267

N=20	b*	Bł. std. z b*	b	Bł. std. z b	t(16)	p
W. wolny			707,000	8,16946	86,5419	0,000000
x1	-1,12276	0,083258	-155,800	11,55335	-13,4853	0,000000
x2	-0,86188	0,083258	-119,600	11,55335	-10,3520	0,000000
x3	-0,58804	0,083258	-81,600	11,55335	-7,0629	0,000003

współczynniki funkcji regresji istotne

Analiza wariancji

Dane: Analiza wariancji (dane w wyklad_03.stw)

Analiza wariancji (dane w wyklad_03.stw)
Zaznaczone efekty są istotne z $p < ,05000$

Zmienna	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F	p
tempo	66870,55	3	22290,18	5339,200	16	333,7000	66,79707	0,000000

Zmienna: (Spreadsheet w wyklad_03b.stw)

st NIR; Zmienna: (Spreadsheet w wyklad_03b.stw)
Znaczone różnice są istotne z $p < ,05000$

	{1}	{2}	{3}	{4}
moc	M=551,20	M=587,40	M=625,40	M=707,00
160 {1}		0,006416	0,000008	0,000000
180 {2}	0,006416		0,004624	0,000000
200 {3}	0,000008	0,004624		0,000003
220 {4}	0,000000	0,000000	0,000003	

Dane: Tabe...

Tabela przekrojów st:
N=20 (Zmienne zależne)

moc	tempo Średnie
160	551,2000
180	587,4000
200	625,4000
220	707,0000
Ogół	617,7500

- ① $\mu_1 = 551,2 = 707,0 - 155,8 = b_0 + b_1$
- ② $\mu_2 = 587,4 = 707,0 - 119,6 = b_0 + b_2$
- ③ $\mu_3 = 625,4 = 707,0 - 81,6 = b_0 + b_3$
- ④ $\mu_4 = 707,0 = b_0$

Analiza regresji

Dane: Analiza wariancji ; DV: tempo (dane w wyklad_07.stw)

Analiza wariancji ; DV: tempo (dane w wyklad_07.stw)

Efekt	Suma kwadrat.	df	Średnia kwadrat.	F	p
Regres.	66870,55	3	22290,18	66,79707	0,000000
Reszta	5339,20	16	333,70		
Razem	72209,75				

Dane: Podsumowanie regresji zmiennej zależnej: tempo (dane w wyklad_07.stw)

Podsumowanie regresji zmiennej zależnej: tempo (dane w wyklad_07.stw)
R= ,96232003 R^2= ,92605985 Popraw. R2= ,91219607
F(3,16)=66,797 p<,00000 Błąd std. estymacji: 18,267

	b*	Bł. std. z b*	b	Bł. std. z b	t(16)	p
N=20						
W. wolny			707,000	8,16946	86,5419	0,000000
x1	-1,12276	0,083258	-155,800	11,55335	-13,4853	0,000000
x2	-0,86188	0,083258	-119,600	11,55335	-10,3520	0,000000
x3	-0,58804	0,083258	-81,600	11,55335	-7,0629	0,000003



Analiza wariancji

Dane: Analiza wariancji (dane w wyklad_03.stw)

Analiza wariancji (dane w wyklad_03.stw)
Zaznaczone efekty są istotne z $p < ,05000$

Zmienna	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F	p
tempo	66870,55	3	22290,18	5339,200	16	333,7000	66,79707	0,000000

Zmienna: (Spreadsheet w wyklad_03b.stw)

st NIR; Zmienna: (Spreadsheet w wyklad_03b.stw)
Znaczone różnice są istotne z $p < ,05000$

	{1}	{2}	{3}	{4}
M=551,20	M=587,40	M=625,40	M=707,00	

Dane: Tabe...

Tabela przekrojów st:
N=20 (Zmienne zależne)

moc	tempo Średnie
160	551,2000
180	587,4000
200	625,4000
220	707,0000
Ogół	617,7500

moc	{1}	{2}	{3}	{4}
160	M=551,20	0,006416	0,000008	0,000000
180	0,006416	M=587,40	0,004624	0,000000
200	0,000008	0,004624	M=625,40	0,000003
220	0,000000	0,000000	0,000003	M=707,00

- ① $\mu_1 = 551,2 = b_0$
- ② $\mu_2 = 587,4 = 551,2 + 36,2 = b_0 + b_1$
- ③ $\mu_3 = 625,4 = 551,2 + 74,2 = b_0 + b_2$
- ④ $\mu_4 = 707,0 = 551,2 + 155,8 = b_0 + b_3$

kodowanie, dla mocy:

- 160: $x_1 = 0, x_2 = 0, x_3 = 0,$
- 180: $x_1 = 1, x_2 = 0, x_3 = 0,$
- 200: $x_1 = 0, x_2 = 1, x_3 = 1,$
- 220: $x_1 = 0, x_2 = 0, x_3 = 1.$

Analiza regresji (inne kodowanie)

Dane: Analiza wariancji ; DV: tempo (dane w wyklad_07.stw)

Analiza wariancji ; DV: tempo (dane w wyklad_07.stw)

Efekt	Suma kwadrat.	df	Średnia kwadrat.	F	p
Regres.	66870,55	3	22290,18	66,79707	0,000000
Reszta	5339,20	16	333,70		
Razem	72209,75				

Dane: Podsumowanie regresji zmiennej zależnej: tempo (dane w wyklad_07.stw)

Podsumowanie regresji zmiennej zależnej: tempo (dane w wyklad_07.stw)
R= ,96232003 R²= ,92605985 Popraw. R²= ,91219607
F(3,16)=66,797 p<,00000 Błąd std. estymacji: 18,267

	b*	Bł. std. z b*	b	Bł. std. z b	t(16)	p
N=20						
W. wolny			551,2000	8,16946	67,47084	0,000000
x1	0,260871	0,083258	36,2000	11,55335	3,13329	0,006416
x2	0,534714	0,083258	74,2000	11,55335	6,42238	0,000008
x3	1,122755	0,083258	155,8000	11,55335	13,48526	0,000000



Eksperymenty jednoczynnikowe: związki pomiędzy analizą wariancji i regresji

Jednoczynnikowa analiza wariancji – podejście regresyjne (kodowanie zmiennych jakościowych)

Wykorzystując kodowanie z sigma ograniczeniami (każda ze zmiennych fikcyjnych sumuje się do 0):

$$x_i = \begin{cases} 1 & \text{gdy zmienna jakościowa jest na poziomie } i, \\ -1 & \text{gdy zmienna jakościowa jest na poziomie } a, \\ 0 & \text{w przeciwnym przypadku} \end{cases}$$

Po wprowadzeniu $(a-1)$ zmiennych fikcyjnych, tworzona jest funkcja regresji:

$$y = b_0 + b_1x_1 + \dots + b_{a-1}x_{a-1} + e,$$

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & \vdots & 0 \\ 1 & 0 & 1 & \vdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & \vdots & -1 \\ \dots & \dots & \dots & \dots & \dots \\ 1 & 1 & 0 & \vdots & 0 \\ 1 & 0 & 1 & \vdots & 0 \\ 1 & \vdots & \vdots & \vdots & \vdots \\ 1 & -1 & -1 & \vdots & -1 \end{bmatrix} \begin{cases} 1 \text{ powtórka doświadczeń} \\ \\ \\ r\text{-ta powtórka doświadczeń} \end{cases} = \mathbf{y} = \begin{bmatrix} y_{11} \\ y_{21} \\ \vdots \\ y_{a1} \\ \dots \\ y_{1r} \\ y_{2r} \\ \vdots \\ y_{ar} \end{bmatrix}$$

zmienna x_0 odpowiadająca wyrazowi wolnemu
zmiennne x_1, \dots, x_{a-1}

	numer doświadczenia			
	1	2	...	r
1				
⋮				
a				



Jednoczynnikowa analiza wariancji – podejście regresyjne (kodowanie zmiennych jakościowych)

Kodowanie z sigma ograniczeniami prowadzi do układu zależności:

$$\begin{aligned} \hat{y}_1 &= \hat{b}_0 + \hat{b}_1 \cdot 1 + \hat{b}_2 \cdot 0 + \dots + \hat{b}_{a-1} \cdot 0 && = \hat{b}_0 + \hat{b}_1 \\ \hat{y}_2 &= \hat{b}_0 + \hat{b}_1 \cdot 0 + \hat{b}_2 \cdot 1 + \dots + \hat{b}_{a-1} \cdot 0 && = \hat{b}_0 + \hat{b}_2 \\ &\dots\dots\dots && \dots\dots\dots \\ \hat{y}_{a-1} &= \hat{b}_0 + \hat{b}_1 \cdot 0 + \hat{b}_2 \cdot 0 + \dots + \hat{b}_{a-1} \cdot 1 && = \hat{b}_0 + \hat{b}_{a-1} \\ \hat{y}_a &= \hat{b}_0 + \hat{b}_1 \cdot (-1) + \hat{b}_2 \cdot (-1) + \dots + \hat{b}_{a-1} \cdot (-1) && = \hat{b}_0 - \hat{b}_1 - \hat{b}_2 - \dots - \hat{b}_{a-1} \end{aligned},$$

Wprowadzając oznaczenie: $\hat{b}_a = -\sum_{i=1}^{a-1} \hat{b}_i$ dla każdego poziomu zmiennej niezależnej można zapisać:

$$\hat{y}_i = \hat{b}_0 + \hat{b}_i,$$

Porównując powyższy model z modelem wykorzystywanym w analizie wariancji: $\hat{y}_i = \mu + \tau_i$, otrzymuje się następującą interpretację współczynników funkcji regresji:

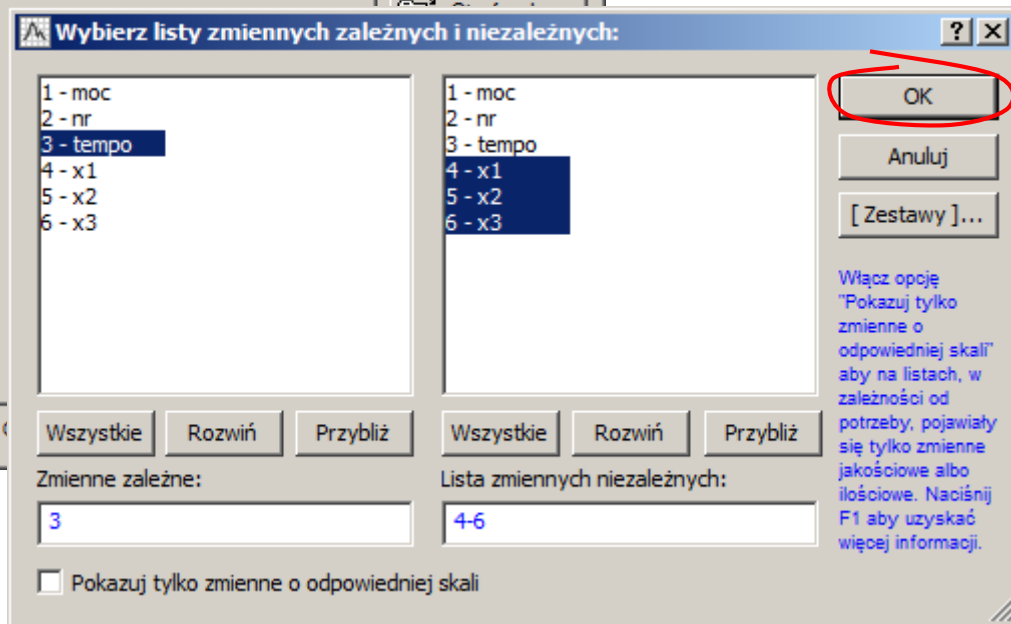
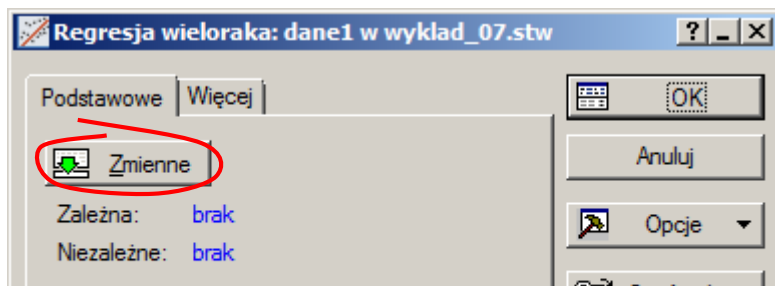
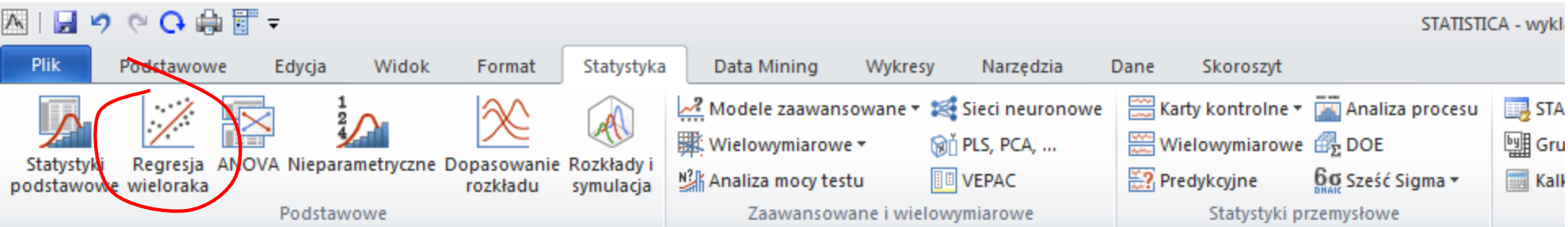
$$\hat{b}_0 = \mu,$$

$$\hat{b}_i = \tau_i.$$

oznacza, to że:

- wyraz wolny odpowiada średniej policzonej dla wszystkich poziomów zmiennej objaśniającej,
- pozostałe współczynniki f. regresji opisują przesunięcia średnich względem poziomu referencyjnego opisanego przez wyraz wolny

STATISTICA – analiza regresji



	1	2	3	4	5	6
	moc	nr	tempo	x1	x2	x3
	200	1	600	0	0	1
	220	1	725	-1	-1	-1
	220	2	700	-1	-1	-1
	160	1	575	1	0	0
	160	2	542	1	0	0
	180	1	565	0	1	0
	200	2	651	0	0	1
	160	3	530	1	0	0
	180	2	593	0	1	0
	200	3	610	0	0	1
	220	3	715	-1	-1	-1
	220	4	685	-1	-1	-1
	160	4	539	1	0	0
	160	5	570	1	0	0
	220	5	710	-1	-1	-1
	180	3	590	0	1	0
	180	4	579	0	1	0
	180	5	610	0	1	0
	200	4	637	0	0	1
	200	5	629	0	0	1

STATYSTICA – analiza regresji

Wyniki regresji wielorakiej: dane1 w wyklad_07.stw

Wyniki regresji wielorakiej

Zmn. zależ. tempo Wielor. R = ,96232003 F = 66,79707
 R^2= ,92605985 df = 3,16
 Liczba przyp. 20 Popraw. R^2= ,91219607 p = ,000000
 Błąd standardowy estymacji: 18,267457404
 Wyr. wolny 617,75000000 Błąd std.: 4,084728 t(16) = 151,23 p = 0,0000

x1 b*=-,78 x2 b*=-,36 x3 b*=-,090

(istotne b* są podświetlone na czerwono)

funkcja regresji istotna

Dane: Analiza wariancji ; DV: tempo (dane1 w wyklad_07.stw)

Efekt	Suma kwadrat.	df	Średnia kwadrat.	F	p
Regres.	66870,55	3	22290,18	66,79707	0,000000
Reszta	5339,20	16	333,70		
Razem	72209,75				

Alfa do podświetlania efektów: .05

Podstawowe Więcej Reszty, założenia, predykcja

- Podsumowanie: Wyniki regresji
- ANOVA (sum. dobroć dopasow.)
- Kowariancja współczynników
- Korelacje częściowe
- Nadmiarowość
- Podsumowanie r. krokowej

Dane: Podsumowanie regresji zmiennej zależnej: tempo (dane1 w wyklad_07.stw)

Podsumowanie regresji zmiennej zależnej: tempo (dane1 w wyklad_07.stw)
 R= ,96232003 R^2= ,92605985 Popraw. R^2= ,91219607
 F(3,16)=66,797 p<,00000 Błąd std. estymacji: 18,267

N=20	b*	Bł. std. z b*	b	Bł. std. z b	t(16)	p
W. wolny			617,7500	4,084728	151,2341	0,000000
x1	-0,783159	0,083258	-66,5500	7,074956	-9,4064	0,000000
x2	-0,357158	0,083258	-30,3500	7,074956	-4,2898	0,000563
x3	0,090025	0,083258	7,6500	7,074956	1,0813	0,295602

istotne współczynniki funkcji regresji

nieistotny współczynnik funkcji regresji

Analiza wariancji

Dane: Analiza wariancji (dane w wyklad_03.stw)

Analiza wariancji (dane w wyklad_03.stw)
Zaznaczone efekty są istotne z $p < ,05000$

Zmienna	SS Efekt	df Efekt	MS Efekt	SS Błąd	df Błąd	MS Błąd	F	p
tempo	66870,55	3	22290,18	5339,200	16	333,7000	66,79707	0,000000

Zmienna: (Spreadsheet w wyklad_03b.stw)

st NIR; Zmienna: (Spreadsheet w wyklad_03b.stw)
Znaczone różnice są istotne z $p < ,05000$

	{1}	{2}	{3}	{4}
moc	M=551,20	M=587,40	M=625,40	M=707,00
160 {1}		0,006416	0,000008	0,000000
180 {2}	0,006416		0,004624	0,000000
200 {3}	0,000008	0,004624		0,000003
220 {4}	0,000000	0,000000	0,000003	

Dane: Tabe...

Tabela przekrojów st:
N=20 (Zmienne zależne)

moc	tempo Średnie
160	551,2000
180	587,4000
200	625,4000
220	707,0000
Ogół	617,7500

- ① $\mu_1 = 551,2 = 617,75 - 66,55 = b_0 + b_1$
- ② $\mu_2 = 587,4 = 617,75 - 30,35 = b_0 + b_2$
- ③ $\mu_3 = 625,4 = 617,75 + 7,65 = b_0 + b_3$
- ④ $\mu_4 = 707,0 = 617,75 - (-66,55 - 30,35 + 7,65) = b_0 - (b_1 + b_2 + b_3)$

Analiza regresji

Dane: Analiza wariancji ; DV: tempo (dane1 w wyklad_07.stw)

Analiza wariancji ; DV: tempo (dane1 w wyklad_07.stw)

Efekt	Suma kwadrat.	df	Średnia kwadrat.	F	p
Regres.	66870,55	3	22290,18	66,79707	0,000000
Reszta	5339,20	16	333,70		
Razem	72209,75				

Dane: Podsumowanie regresji zmiennej zależnej: tempo (dane1 w wyklad_07.stw)

Podsumowanie regresji zmiennej zależnej: tempo (dane1 w wyklad_07.stw)
R= ,96232003 R^2= ,92605985 Popraw. R^2= ,91219607
F(3,16)=66,797 p<,00000 Błąd std. estymacji: 18,267

	b*	Bł. std. z b*	b	Bł. std. z b	t(16)	p
N=20						
W. wolny			617,7500	4,084728	151,2341	0,000000
x1	-0,783159	0,083258	-66,5500	7,074956	-9,4064	0,000000
x2	-0,357158	0,083258	-30,3500	7,074956	-4,2898	0,000563
x3	0,090025	0,083258	7,6500	7,074956	1,0813	0,295602

STATISTICA – analiza regresji – GLM

The screenshot shows the STATISTICA software interface. The 'Modely zaawansowane' menu is open, and 'Ogólne modele liniowe' is highlighted with a red circle. The main menu bar includes 'Plik', 'Podstawowe', 'Edycja', 'Widok', 'Format', 'Statystyka', 'Data Mining', 'Wykresy', 'Narzędzia', 'Dane', and 'Skoroszyt'. The toolbar contains icons for 'Statystyki podstawowe', 'Regresja wieloraka', 'ANOVA', 'Nieparametryczne', 'Dopasowanie rozkładu', and 'Rozkłady i symulacja'.

wykład_07.stw* - dane1

	1 moc	2 nr	3 tempo
	200	1	600
	220	1	725
	220	2	700
	160	1	575
	160	2	542
	180	1	565
	200	2	651
	160	3	530
	180	2	593
	200	3	610
	220	3	715
	220	4	685
	160	4	539
	160	5	570
	220	5	710
	180	3	590
	180	4	579
	180	5	610
	200	4	637
	200	5	629

dane1 dane1 dar

Ogólne modele liniowe (GLM): dane1 w wykład_07.stw

Podstawowe

Rodzaj analizy:

- Jednoczynnikowa ANOVA
- ANOVA efektów głównych
- ANOVA dla układów czynnikowych
- Układ zagnieżdżony ANOVA
- Duże układy zrównoważone
- Układy z powtarzaniem pomiarów
- Regresja prosta
- Regresja wieloraka
- Regresja czynnikowa
- Regresja wielomianowa
- Regresja powierzchni odpowiedzi
- Powierzchnia odp. dla mieszania
- Analiza kowariancji
- Model różnych nachyleń
- Model jednakowych nachyleń
- Ogólne modele liniowe**

Sposób definiowania analizy:

- Szybkie definiowanie
- Kreator analizy
- Edytor składowi

Ogólna ANCOVA/MANCOVA umożliwia analizowanie układów z dowolnymi kombinacjami predyktorów jakościowych, predyktorów ciągłych i czynników powtarzanych pomiarów.

Analizy dowolnego typu mogą dotyczyć wielu zmiennych zależnych. Jeżeli analiza obejmuje wiele zmiennych zależnych, to dostępne będą wyniki jedno- i wielowymiarowe.

Inne metody dla podobnych problemów dostępne są w modułach Regresja wieloraka, Komponenty wariacyjne, Planowanie doświadczeń i Estymacja nieliniowa.

OK

Anuluj

Opcje

Otwórz dane

SELECT CASES

Momenty ważone

DF =

W-1 N-1

STATYSTICA – analiza regresji – GLM

GLM - Ogólne modele liniowe: dane1 w wyklad_07.stw

Podstawowe Opcje

Zmienne

Zmienne zależne: brak

Powtarzane pomiary: brak

Czynniki jakościowe: brak

Kody czynników: brak

Predyktory ilościowe: brak

Efekty międzygrupowe: brak

Wybierz zmienne zależne oraz predyktory jakościowe i ciągłe

1 - moc
2 - nr
3 - tempo

1 - moc
2 - nr
3 - tempo

1 - moc
2 - nr
3 - tempo

OK

Anuluj

[Zestawy]...

Włącz opcje

GLM - Ogólne modele liniowe: dane1 w wyklad_07.stw

Podstawowe Opcje

Zmienne

Zmienne zależne: tempo

Powtarzane pomiary: brak

Czynniki jakościowe: moc

Kody czynników: brak

Predyktory ilościowe: brak

Efekty międzygrupowe: moc

OK

Anuluj

Opcje

GLM - Ogólne modele liniowe: dane1 w wyklad_07.stw

Podstawowe Opcje

Czynniki losowe: brak

Delta wymiatania: 1.E- 7

Delta odwracania: 1.E- 12

Parametryzacja

Sigma-ograniczenia

Bez wyrazu wolnego

Brak dopasowania

Opcje

Sumy kwadratów

Typ I (sekwencyjne)

Typ II (cząstkowe)

Typ III (ortogonalne)

Typ IV (estymowalne)

Typ V (pełny rząd)

Typ VI

Opcja krzyżowa: wyłączona

OK

Anuluj

Opcje

Edytor składni



STATYSTICA – analiza regresji – GLM

GLM - Wyniki 1: dane1 w wyklad_07.stw

Porównania | Profile | Reszty | Macierz | Raport

Podstawowe | Więcej | Średnie

Średnie/wykresy | Wszystkie efekty

Wyniki jednowym. | Statystyki podklas

Efekty międzygrupowe

Składniki układu | R pełnego modelu

Współczynniki | Estymacja

Wartości alfa

Ufności: .950

Istotności: .050

Dane: Oceny parametrów (dane1 w wyklad_07.stw)

Oceny parametrów (dane1 w wyklad_07.stw)
Parametryzacja z sigma-ograniczeniami

Efekt	Poziom Efekt	Kolumna	tempo Param.	tempo Bł. std.	tempo t	tempo p
Wyraz wolny		1	617,7500	4,084728	151,2341	0,000000
moc	160	2	-66,5500	7,074956	-9,4064	0,000000
moc	180	3	-30,3500	7,074956	-4,2898	0,000563
moc	200	4	7,6500	7,074956	1,0813	0,295602

wyniki identyczne jak na slajdzie 16.

Dane: Test SS dla pełnego modelu względem SS dla reszt (dane1 w wyklad_07.stw)

Test SS dla pełnego modelu względem SS dla reszt (dane1 w wyklad_07.stw)

Zależna Zm.	Wielokr. R	Wielokr. R2	Skorygow R2	SS Model	df Model	MS Model	SS Reszta	df Reszta	MS Reszta	F	p
tempo	0,962320	0,926060	0,912196	66870,55	3	22290,18	5339,200	16	333,7000	66,79707	0,000000

p-value
funkcja regresji istotna



Badanie istotności wpływu – przykład 2. (wykład 4)

Przykład 2.* Należy wybrać jeden z trzech rodzajów materiałów, który ma być zastosowany w baterii zasilającej urządzenie wystawione na działanie dużych różnic temperatur. Zaplanowano eksperyment, którego celem miało być ustalenie, który z materiałów jest bardziej odporny na wahania temperatury. Zdecydowano o wyborze 3 temperatur zgodnych z warunkami w których będzie pracowało urządzenie 15, 70 i 125°F (-9,44°C, 21,11°C, 51,67°C) i zaplanowano po 4 doświadczenia dla każdej kombinacji: (materiał, temperatura). Po zaplanowaniu eksperymentu i ustaleniu kolejności prowadzenia poszczególnych doświadczeń wyniki uzyskanych czasów pracy baterii (w [godz.]) zapisano w arkuszu:

	1 materiał	2 temperatura	3 czas pracy	4 x1	5 x2	6 x3	7 x4	8 x5	9 x6	10 x7	11 x8
	1	15	130	1	0	1	0	1	0	0	0
	1	15	74	1	0	1	0	1	0	0	0
	1	15	155	1	0	1	0	1	0	0	0
	1	15	180	1	0	1	0	1	0	0	0
	1	70	34	1	0	0	1	0	1	0	0
	1	70	80	1	0	0	1	0	1	0	0
	1	70	40	1	0	0	1	0	1	0	0
	1	70	75	1	0	0	1	0	1	0	0
	1	125	20	1	0	-1	-1	-1	-1	0	0
	1	125	82	1	0	-1	-1	-1	-1	0	0
	1	125	70	1	0	-1	-1	-1	-1	0	0
	1	125	58	1	0	-1	-1	-1	-1	0	0
	2	15	150	0	1	1	0	0	0	1	0
	2	15	159	0	1	1	0	0	0	1	0
	2	15	188	0	1	1	0	0	0	1	0
	2	15	126	0	1	1	0	0	0	1	0
	2	70	136	0	1	0	1	0	0	0	1
	2	70	106	0	1	0	1	0	0	0	1
	2	70	122	0	1	0	1	0	0	0	1
	2	70	115	0	1	0	1	0	0	0	1
	2	125	25	0	1	-1	-1	0	0	-1	-1

Znaczenie zmiennych:

materiał i temperatura: symbol materiału z którego wykonana została bateria i temperatura, w której przeprowadzono doświadczenie,

x1, x2, x3, x4, x5, x6, x7, x8

fikcyjne zmienne wykorzystywane do kodowania z sigma ograniczeniami,

zmienne x1 i x2 kodują materiał:

mat.1: x1=1, x2=0, mat.2: x1=0, x2=1,

mat.3: x1=-1, x2=-1,

zmienne x3 i x4 kodują temperaturę:

15°F: x3=1, x4=0, 70°F: x3=0, x4=1,

125°F: x3=-1, x4=-1,

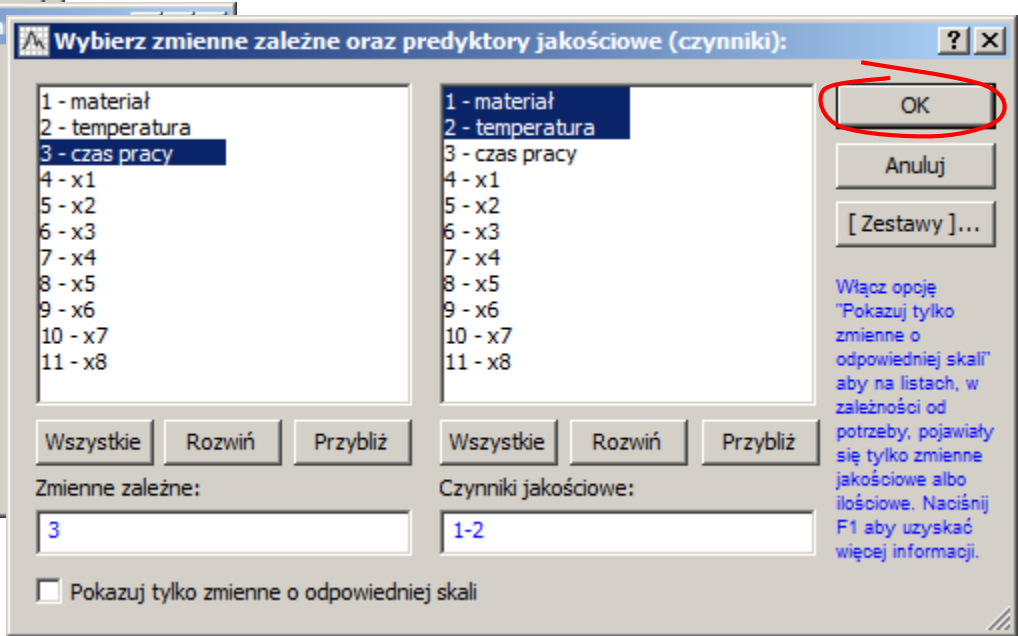
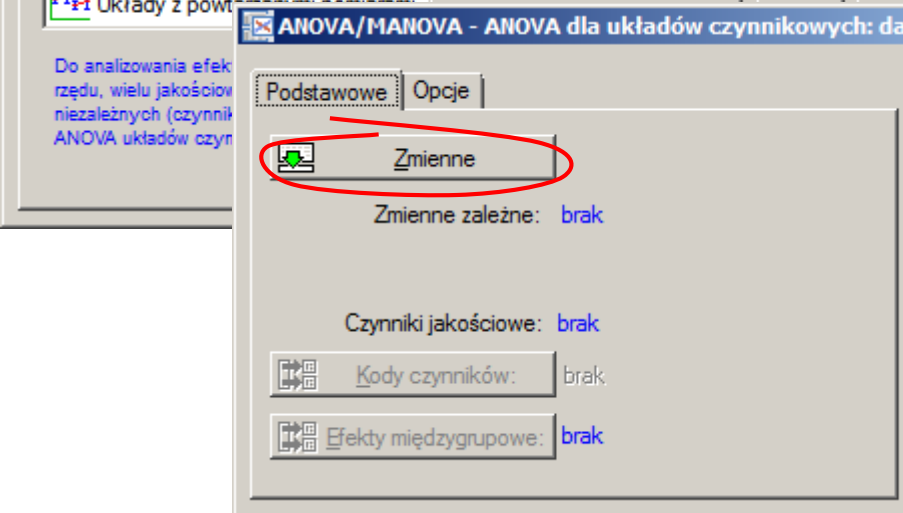
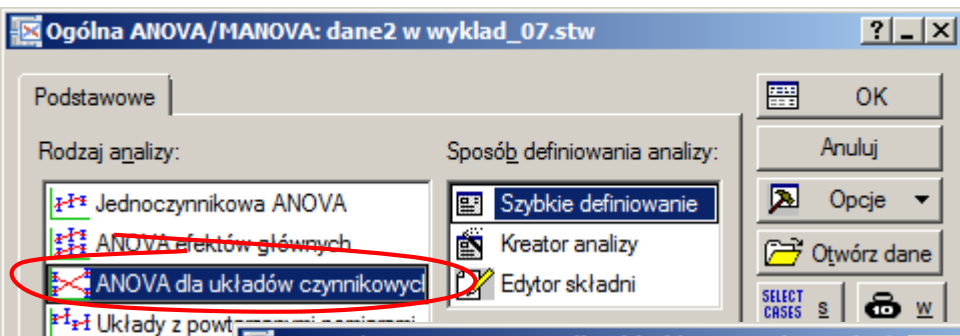
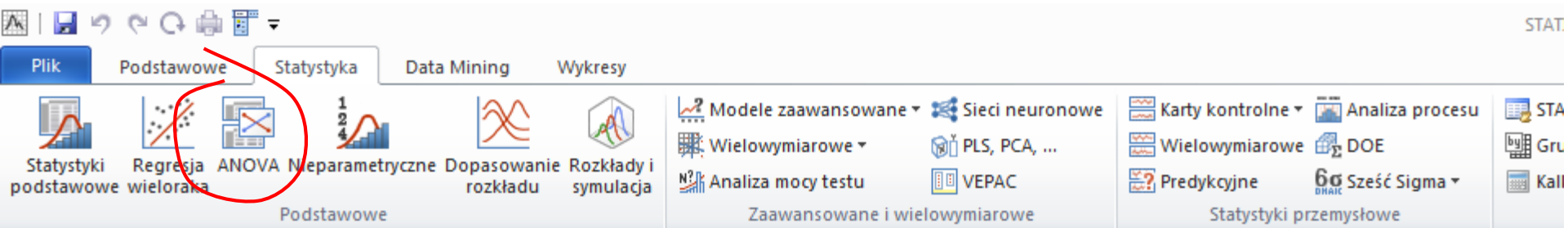
zmienne x5-x8 kodują interakcje:

x5 = x1·x3, x6 = x1·x4,

x7 = x2·x3, x8 = x2·x4.



STATISTICA – dwuczynnikowa analiza wariancji (wykład 4)



STATYSTICA – dwuczynnikowa analiza wariancji (wykład 4)

ANOVA - Wyniki 1: dane2 w wyklad_07. ? _ X

Porównania | Profile | Reszty | Macierz | Raport

Podstawowe | Więcej | Średnie

Średnie/wykresy | Wszystkie efekty 1

Wyniki jednowym. | Statystyki podklas

Efekty międzygrupowe

Składniki układu | R pełnego modelu

Współczynniki 2 | Estymacja

Wartości alfa

Ufności: .950

Istotności: .050

Więcej wyników | Zmień | Zamknij

Grupami | Opcje

Dane: Jednowymiarowe testy istotności dla czas pracy (dane2 w wyklad_07.stw)*

Jednowymiarowe testy istotności dla czas pracy (dane2 w wyk
Parametryzacja z sigma-ograniczeniami
Dekompozycja efektywnych hipotez

Efekt	SS	Stopnie swobody	MS	F	p
Wyraz wolny	400900,0	1	400900,0	593,7386	0,000000
materiał	10683,7	2	5341,9	7,9114	0,001976
temperatura	39118,7	2	19559,4	28,9677	0,000000
materiał*temperatura	9613,8	4	2403,4	3,5595	0,018611
Błąd	18230,7	27	675,2		

Czas pracy baterii zależy od obydwu testowanych czynników oraz od ich interakcji.

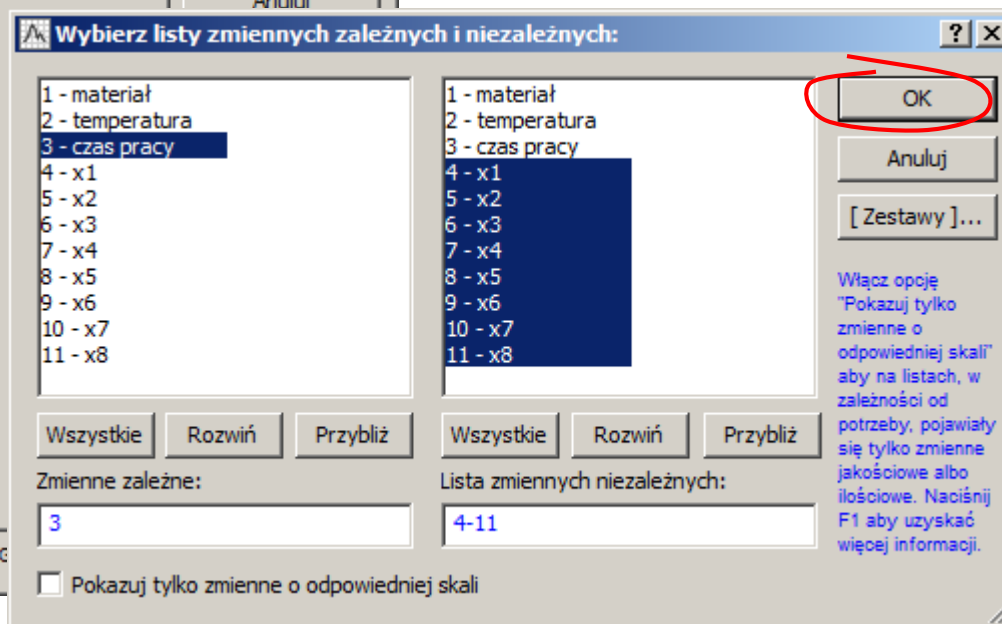
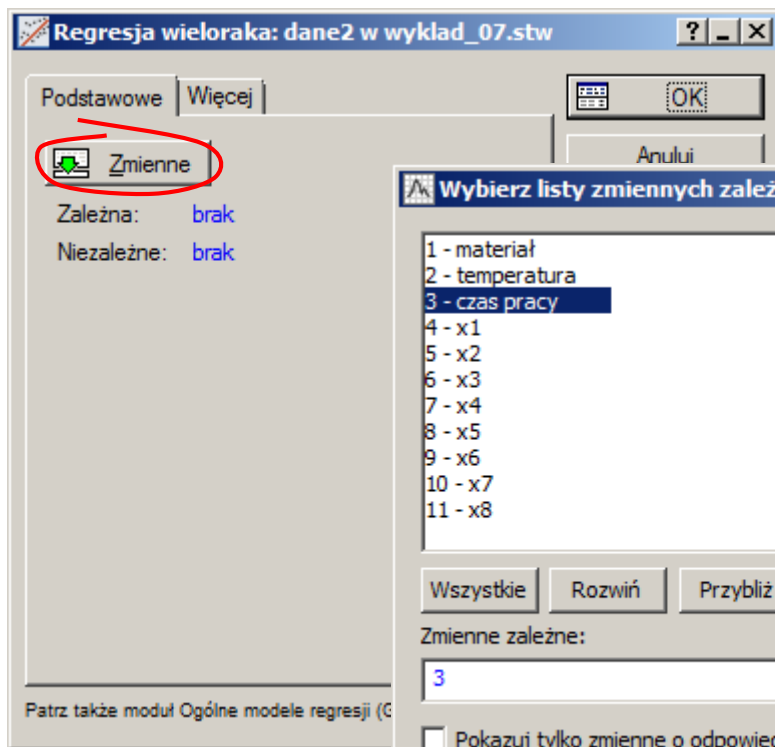
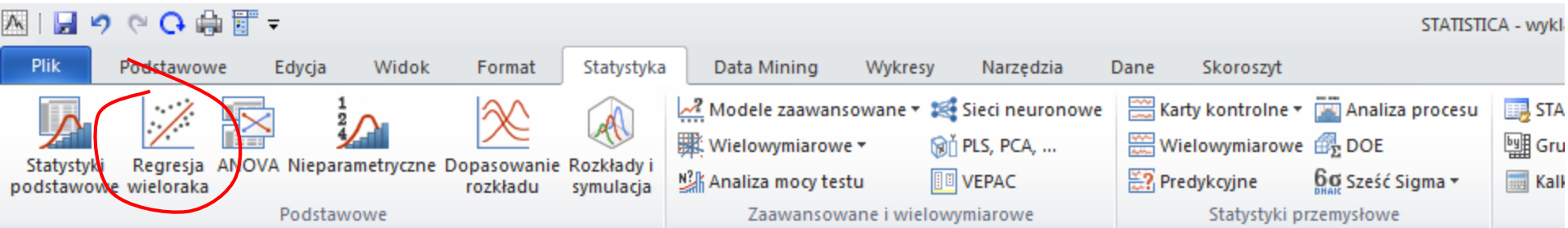
Dane: Oceny parametrów (dane2 w wyklad_07.stw)

Oceny parametrów (dane2 w wyklad_07.stw)
Parametryzacja z sigma-ograniczeniami

Efekt	Poziom Efekt	Kolumna	czas pracy Param.	czas pracy Bł. std.	czas pracy t	czas pracy p
Wyraz wolny		1	105,5278	4,330810	24,36675	0,000000
materiał	1	2	-22,3611	6,124690	-3,65098	0,001106
materiał	2	3	2,8056	6,124690	0,45807	0,650565
temperatura	15	4	39,3056	6,124690	6,41756	0,000001
temperatura	70	5	2,0556	6,124690	0,33562	0,739753
materiał*temperatura	1	6	12,2778	8,661620	1,41749	0,167777
materiał*temperatura	2	7	-27,9722	8,661620	-3,22944	0,003250
materiał*temperatura	3	8	8,1111	8,661620	0,93644	0,357346
materiał*temperatura	4	9	9,3611	8,661620	1,08076	0,289364



STATISTICA – analiza regresji



STATYSTICA – analiza regresji

Wyniki regresji wielorakiej: dane2 w wyklad_07.stw

Wyniki regresji wielorakiej

Zmn. zależ. **czas pracy** Wielor. R = ,87476270 F = 10,99953
R²= ,76520978 df = 8,27
Liczba przyp. 36 Popraw. R²= ,69564230 p = ,000001 ← *funkcja regresji istotna*
Błąd standardowy estymacji: 25,984860264
Wyr. wolny 105,5277778 Błąd std.: 4,330810 t(27) = 24,367 p = 0,0000

x1 b*=-,39 x2 b*=-,049 x3 b*=-,691
x4 b*=-,036 x5 b*=-,176 x6 b*=-,40
x7 b*=-,116 x8 b*=-,134

(istotne b* są podświetlone na czerwono)

Dane: Analiza wariancji ; DV: czas pracy (dane2 w wyklad_07...

Efekt	Suma kwadrat.	df	Średnia kwadrat.	F	p
Regres.	59416,22	8	7427,028	10,99953	0,000001
Reszta	18230,75	27	675,213		
Razem	77646,97				

Alfa do podświetlania efektów: .05

Podstawowe Więcej Reszty, założenia, predykcja

Podsumowanie: Wyniki regresji 1

ANOVA (sum. dobroć dopasow.) 2

Kowariancja współczynników

Aktualna macierz wymiany

Dane: Podsumowanie regresji zmiennej zależnej: czas pracy (dane2 w wyklad_07...

Podsumowanie regresji zmiennej zależnej: czas pracy (dane2 w wyklad_07...
R= ,87476270 R²= ,76520978 Popraw. R²= ,69564230
F(8,27)=11,000 p<,00000 Błąd std. estymacji: 25,985

	b*	Bł. std. z b*	b	Bł. std. z b	t(27)	p
N=36						
W. wolny			105,5278	4,330810	24,36675	0,000000
x1	-0,393130	0,107678	-22,3611	6,124690	-3,65098	0,001106
x2	0,049324	0,107678	2,8056	6,124690	0,45807	0,650565
x3	0,691031	0,107678	39,3056	6,124690	6,41756	0,000001
x4	0,036139	0,107678	2,0556	6,124690	0,33562	0,739753
x5	0,176245	0,124336	12,2778	8,661620	1,41749	0,167777
x6	-0,401536	0,124336	-27,9722	8,661620	-3,22944	0,003250
x7	0,116434	0,124336	8,1111	8,661620	0,93644	0,357346
x8	0,134377	0,124336	9,3611	8,661620	1,08076	0,289364

wyniki są identyczne jak w analizie wariancji (slajd 22)