# Analiza korelacji i regresji

# NIEZALEŻNOŚĆ ZMIENNYCH

**Zależność stochastyczna** dwóch zmiennych losowych polega na tym, że zmiana jednej z nich zmienia rozkład prawdopodobieństwa drugiej zmiennej

Cecha *X* jest stochastycznie niezależna od cechy *Y* gdy:

$\bar{x}_1 = \bar{x}_2 = \ldots \bar{x}_k$ oraz $s_1^2(x) = s_2^2(x) = \ldots = s_k^2(x)$

Cecha *Y* jest stochastycznie niezależna od cechy *X* gdy:

$\bar{y}_1 = \bar{y}_2 = \ldots \bar{y}_r$ oraz $s_1^2(y) = s_2^2(y) = \ldots = s_r^2(y)$

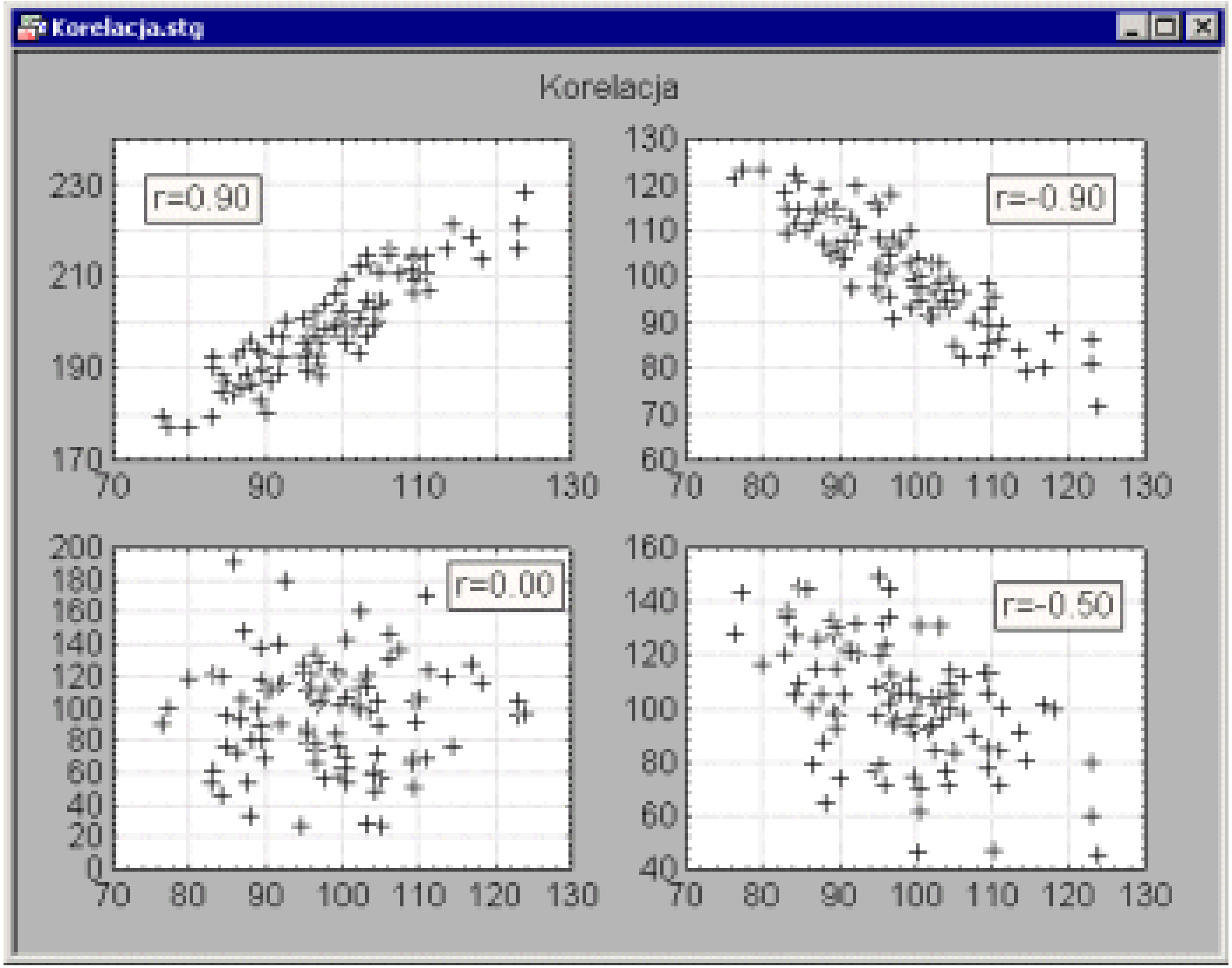
**Zależność korelacyjna** dwóch zmiennych losowych polega na tym, że zmiana jednej z nich pociąga za sobą zmianę średnich wartości drugiej cechy

Cecha *X* jest korelacyjnie niezależna od cechy *Y* gdy: $\bar{x}_1 = \bar{x}_2 = \ldots \bar{x}_k$

Cecha *Y* jest korelacyjnie niezależna od cechy *X* gdy: $\bar{y}_1 = \bar{y}_2 = \ldots \bar{y}_r$

**Zależność funkcyjna** dwóch zmiennych losowych polega na tym, że zmiana wartości jednej z nich pociąga za sobą ściśle określoną zmianę wartości drugiej cechy

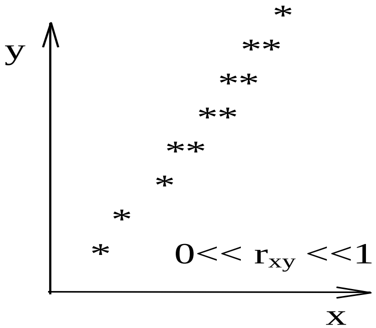
# Korelacyjny diagram rozrzutu



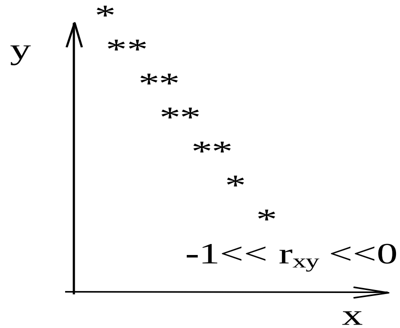
# Korelacyjny diagram rozrzutu

## Korelacja liniowa

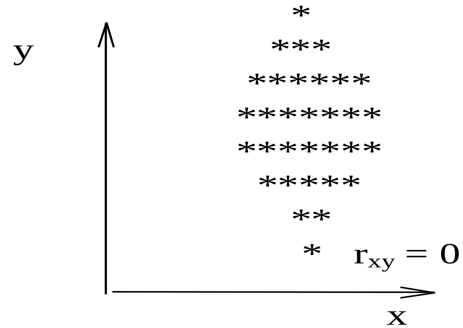
Silna dodatnia




Silna ujemna

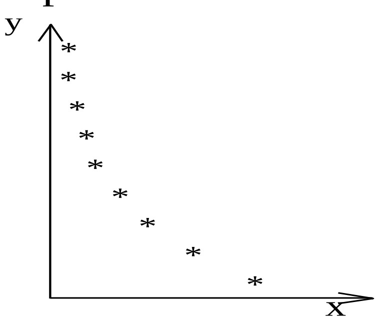



## Brak korelacji

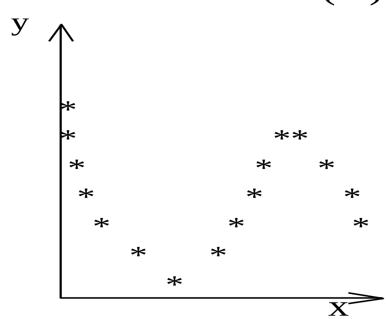



## Korelacja nieliniowa

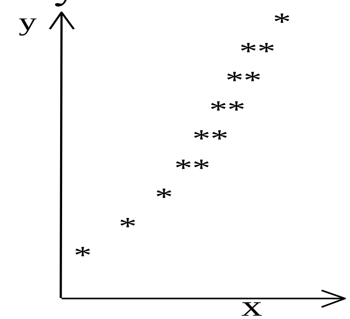
hiperbola



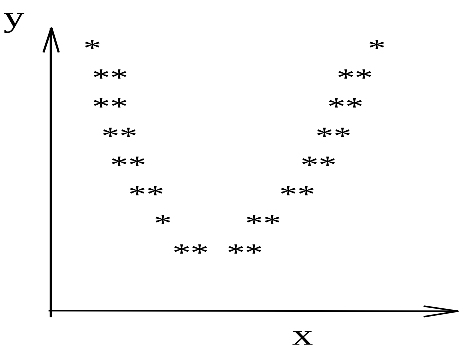

wielomian (3$^{o}$)



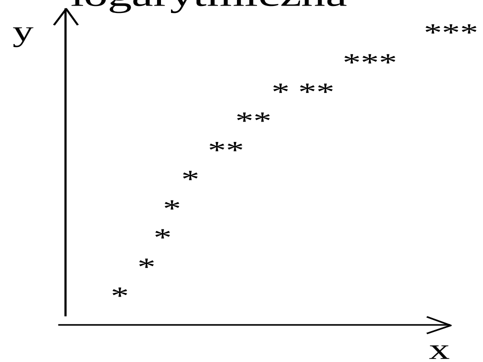

wykładnicza




parabola (2$^{o}$)




logarytmiczna



# Współczynnik korelacji liniowej Pearsona

- **$r_{xy}= r_{yx}$ - mierzy siłę i kierunek związku prosto-linowego pomiędzy dwoma cechami**
- ***współczynnik korelacji*** **- standaryzacja kowariancji**
- ***kowariancja*** **- średnia arytmetyczna iloczynu odchyleń zmiennych X i Y od wartości średnich arytmetycznych**
- ***związek liniowy*** **- jednostkowym przyrostom przyczyny (zmiennej niezależnej) towarzyszy stały przyrost skutku (zmiennej zależnej)**
- **kowariancja informuje o:**

  **cov(X,Y)=0 braku zależności korelacyjnej**

  **cov(X,Y)<0 ujemnej zależności korelacyjnej**

  **cov(X,Y)>0 dodatniej zależności korelacyjnej**
- **Unormowanym miernikiem natężenia siły związku dwóch zmiennych jest *współczynnik korelacji liniowej Pearsona***

$$r_{xy} = r_{yx} = \frac{\operatorname{cov}(X,Y)}{s(x)s(y)}$$

$$r \in \langle -1;1 \rangle$$

$$r = \frac{\operatorname{cov}(X,Y)}{s(x)\cdot s(y)} = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{n \cdot s(x)\cdot s(y)} = \frac{\frac{1}{n}\sum_j x_j y_j - \bar{x}\cdot\bar{y}}{s(x)\cdot s(y)}$$

# Współczynnik korelacji liniowej Pearsona

| | |
|---|---|
| **$r = 1$** | **- korelacja dodatnia - związek funkcyjny** |
| **$0 < r < 1$** | **- korelacja dodatnia niedoskonała** |
| **$r = 0$** | **- brak korelacji prostoliniowej** |
| **$-1 < r < 0$** | **- korelacja ujemna niedoskonała** |
| **$r = -1$** | **- korelacja ujemna - związek funkcyjny** |

$$r_{xy} = r_{yx} = \frac{\operatorname{cov}(X, Y)}{s(x)s(y)}$$

$$r \in \langle -1 ; 1 \rangle$$

# Funkcja regresji

- Analityczny wyraz przyporządkowania średnich wartości zmiennej objaśnianej konkretnym wartościom zmiennych objaśniających.

- Modele regresji I-szego rodzaju:

$$Y = \beta_{1y} \cdot X + \beta_{0y} + \xi$$

$$X = \beta_{1x} \cdot Y + \beta_{0x} + Z$$

- **Składnik losowy ξ** *(Z)* ma swoje źródło w mechanizmach losowych
  – badana zależność ma charakter losowy,
  – nie jest możliwe uwzględnienie wszystkich czynników wpływających na poziom badanego zjawiska,
  – nie znamy dokładnej postaci analitycznej funkcji
  – pomiar danych liczbowych nie jest dokładny

# Funkcja regresji

- Estymatorem funkcji regresji I rodzaju jest funkcja regresji II rodzaju :

$$\hat{y} = b_{1y} \cdot x + b_{0y}$$

$$\hat{x} = b_{1x} \cdot y + b_{0x}$$

- która spełnia warunek (funkcja kryterium KMNK) :

$$E\{[Y - \hat{Y}]^2\} = E\{[Y - (b_{1y} \cdot x + b_{0y})]^2\} \rightarrow \min$$

$$E\{[X - \hat{X}]^2\} = E\{[X - (b_{1x} \cdot y + b_{0x})]^2\} \rightarrow \min$$

- Warunek ten oznacza, iż
  - średnie odchylenie zmiennej losowej od prostej regresji jest równe zeru,
  - prosta regresji przechodzi przez punkt o współrzędnych (x = E(X), y = E(Y))
- Model regresji II rodzaju zapiszemy jako :

$$y = b_{1y} \cdot x + b_{0y} + e$$

$$x = b_{1x} \cdot y + b_{0x} + z$$

- Wektory reszt określone są następująco :

$$e_j = y_j - \hat{y}_j$$

$$z_j = x_j - \hat{x}_j$$

# Etapy estymacji parametrów modelu regresji

1. Dobór zmiennych do modelu:

2. Określenie postaci analitycznej modelu

3. Wybór metody szacunku parametrów strukturalnych modelu

4. Oszacowanie parametrów strukturalnych modelu.

5. Oszacowanie parametrów struktury stochastycznej modelu

6. Ocena jakości modelu.

# Założenia KMNK

1. Postać modelu jest liniowa
2. Zmienne objaśniające są nielosowe
3. Składnik losowy ma nadzieję matematyczną równą zero $E(\xi) = 0$ i stałą wariancję $D^2(\xi)$ = const.

Przez stałość wariancji rozumie się, że nie zależy ona od kolejnych realizacji zmiennych objaśniających modelu.

**Homoscedastyczność - heteroscedastyczność**

**Scedastyczność** - poważny problem, gdyż błędy losowe szacowane na podstawie różnych wartości zmiennej niezależnej X mogą się zmieniać. Gdy warunek ten nie jest spełniony należy zrezygnować z KMNK na rzecz metod alternatywnych.

4. Realizacje zmiennych objaśniających są niezależne, co sprawia, że ciąg $\{\xi_j\}$ jest ciągiem niezależnych zmiennych losowych
5. Składnik losowy ξ nie jest skorelowany ze zmiennymi objaśniającymi
6. Błędy losowe charakteryzują się rozkładem normalnym $N[0, D^2(\xi)]$

# Szacowanie parametrów strukturalnych modelu

$$\hat{y}_j = b_{1y} \cdot x_j + b_{0y}$$

$$\begin{cases} \sum_{j=1}^{n} Y_j = b_{1y} \sum_{j=1}^{n} X_j + n b_{0y} \\ \sum_{j=1}^{n} X_j Y_j = b_{1y} \sum_{j=1}^{n} X_j^2 + b_{0y} \sum_{j=1}^{n} X_j \end{cases}$$

$$b_{1y} = \frac{\sum_{j=1}^{n} (X_j - \overline{X})(Y_j - \overline{Y})}{\sum_{j=1}^{n} (X_j - \overline{X})^2} = \frac{\sum_{j=1}^{n} X_j Y_j - n\overline{X}\,\overline{Y}}{\sum_{j=1}^{n} X_j{}^2 - n\overline{X}^2}$$

$$b_{0y} = \overline{Y} - b_{1y} \overline{X}$$

# Szacowanie parametrów strukturalnych modelu

$$\hat{x}_j = b_{1x} \cdot y_j + b_{0x}$$

$$\begin{cases} \sum_{j=1}^{n} X_j = b_{1x} \sum_{j=1}^{n} Y_j + n b_{0x} \\ \sum_{j=1}^{n} X_j Y_j = b_{1x} \sum_{j=1}^{n} Y_j^2 + b_{0x} \sum_{j=1}^{n} Y_j \end{cases}$$

$$b_{1x} = \frac{\sum_{j=1}^{n} (X_j - \overline{X})(Y_j - \overline{Y})}{\sum_{j=1}^{n} (Y_j - \overline{Y})^2} = \frac{\sum_{j=1}^{n} X_j Y_j - n\overline{X}\,\overline{Y}}{\sum_{j=1}^{n} {Y_j}^2 - n\overline{Y}^2}$$

$$b_{0x} = \overline{X} - b_{1y}\overline{Y}$$

# Szacowanie parametrów strukturalnych modelu

- $b_{1y}$ - współczynnik regresji informuje o ile przeciętnie zmieni się wartość zmiennej zależnej gdy zmienna niezależna wzrośnie o jednostkę

- $b_{1y}$ - tangens kąta nachylenia linii regresji do osi *OX*

- $b_{0y}$ - wyraz wolny - rzędna punktu przecięcia linii regresji z osią *OY*

## Relacje między współczynnikami regresji i współczynnikiem korelacji

$$r = \pm\sqrt{b_{1y} \cdot b_{1x}}$$

$$b_{1x} = r\frac{s(x)}{s(y)}$$

$$b_{1y} = r\frac{s(y)}{s(x)}$$

# Miary jakości dopasowania funkcji regresji

1. **Wariancja składnika resztowego (nieobciążony estymator wariancji składnika losowego $D^2(\xi)$ )**

$$s^2(e_j) = s^2_{y(x)} = \frac{\sum_{j=1}^{n} e_j^2}{n-k} = \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{n-k}$$

dla funkcji liniowej dwóch zmiennych :

$$s^2(e_j) = s^2_{y(x)} = \frac{\sum_{j=1}^{n} (y_j - (b_{1y} x_j + b_{oy}))^2}{n-k} \approx s^2(y) \cdot (1 - r^2)$$

# Miary jakości dopasowania funkcji regresji

2. Zgodnie z równością wariancyjną - **wariancja ogólna *s²(y)*** równa jest sumie **wariancji resztowej *niewyjaśnionej regresją*** i **wariancji teoretycznej *wyjaśnionej regresją*** :

$$s^2(y) = s^2(\hat{y}) + s^2(e)$$

$$\sum_{j=1}^{n}(y_j - \overline{y})^2 = \sum_{j=1}^{n}(\hat{y}_j - \overline{y})^2 + \sum_{j=1}^{n}(y_j - \hat{y})^2$$

# Miary jakości dopasowania funkcji regresji

3. **Współczynnik zbieżności resztowej -** ***indeterminacji, alienacji, nieokreslonoości*** $\varphi^2$

$$\varphi^2 = \frac{s^2(e)}{s^2(y)} = \frac{\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}{\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

$$0 \le \varphi^2 \le 1$$

# Miary jakości dopasowania funkcji regresji

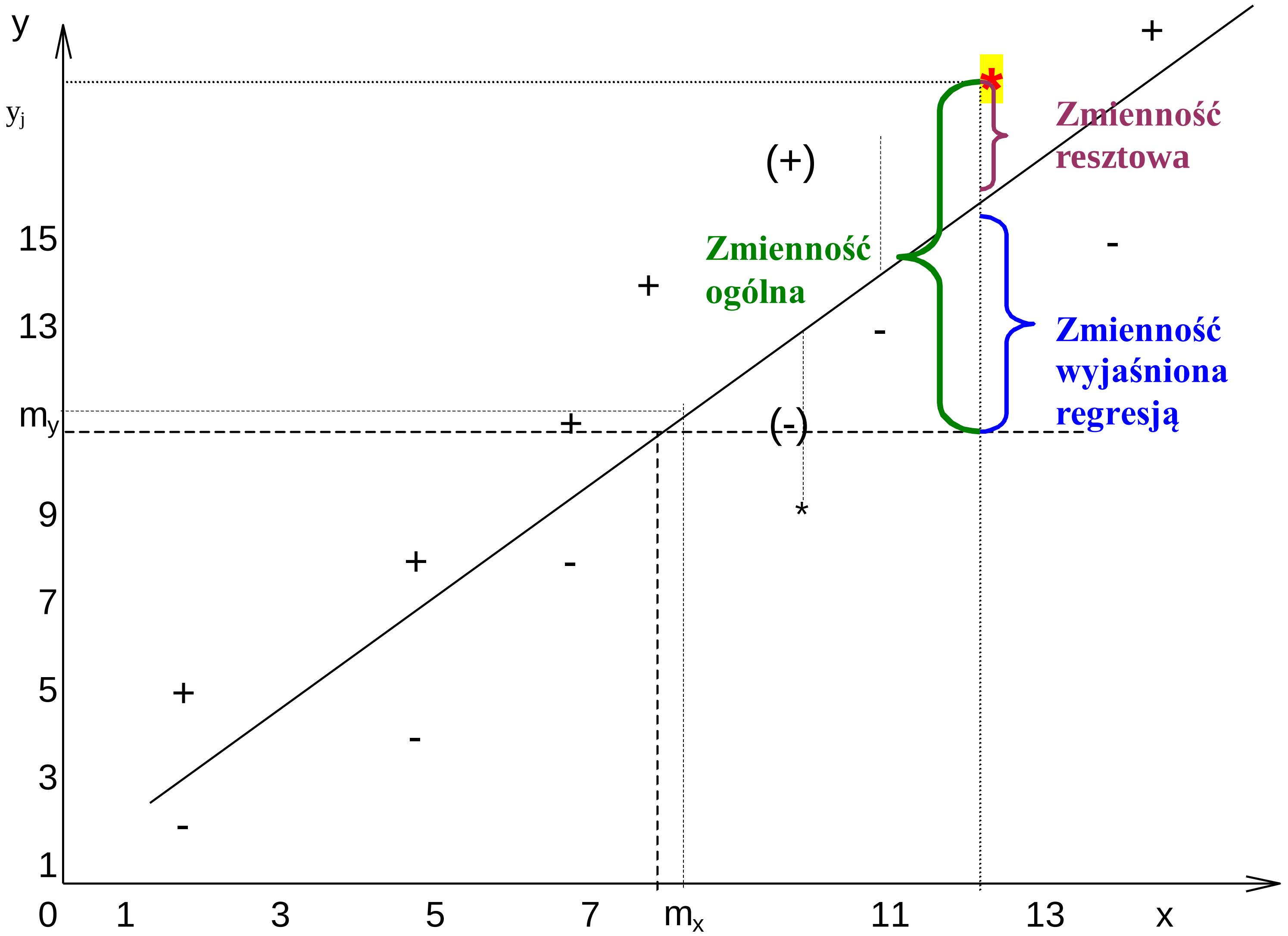
**4. Współczynnik determinacji -** ***określoności***

$$R^2 = \frac{s^2(\hat{y})}{s^2(y)} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \overline{y})^2}{\sum_{j=1}^{n} (y_j - \overline{y})^2} = 1 - \frac{s^2(e)}{s^2(y)} = 1 - \varphi^2$$

$$0 \le R^2 \le 1$$

$$R^2 + \varphi^2 = 1$$

# Dekompozycja wariancji zmiennej zależnej



# Błędy standardowe szacunku parametrów funkcji regresji

Współczynników regresji:

Wyrazów wolnych:

$$s(b_{1y}) = \frac{s_{y(x)}}{\sqrt{\sum x_j^2 - n\bar{x}^2}} = \frac{s_{y(x)}}{\sqrt{\sum (x_j - \bar{x})^2}}$$

$$s(b_{0y}) = \sqrt{\frac{s_{y(x)}^2 \sum x_j^2}{n\sum (x_j - \bar{x})^2}} = \sqrt{\frac{s_{y(x)}^2 \sum x_j^2}{n(\sum x_j^2 - n\bar{x}^2)}}$$

$$s(b_{1x}) = \frac{s_{x(y)}}{\sqrt{\sum y_j^2 - n\bar{y}^2}} = \frac{s_{x(y)}}{\sqrt{\sum (y_j - \bar{y})^2}}$$

$$s(b_{0x}) = \sqrt{\frac{s_{x(y)}^2 \sum y_j^2}{n\sum (y_j - \bar{y})^2}} = \sqrt{\frac{s_{x(y)}^2 \sum y_j^2}{n(\sum y_j^2 - n\bar{y}^2)}}$$

# Pełny zapis modelu regresji

$$y_j = b_{1y} x + b_{0y} + e_j$$

$$(s_{(b_{1y})})(s_{(b_{0y})})(s_e)$$

- parametry strukturalne

- parametry struktury stochastycznej

$$x_j = b_{1x} y_i + b_{0x} + z_j$$

$$(s_{(b_{1x})})(s_{(b_{0x})})(s_z)$$

# Analiza regresji w *Excelu*

1) *Funkcja statystyczna* Excel ➔ REGLINP

**Podaje wartości:**

| | |
|---|---|
| $b_{yi}$ | - ocen parametrów strukturalnych modelu regresji, |
| $s(b_{yi})$ | - błędy standardowe ocen parametrów strukturalnych, |
| $R^2$ | - wartość współczynnika determinacji, |
| *Sey* | - standardowy błąd oceny *y* - błąd standardowy szacunku funkcji regresji - odchylenie standardowe składnika resztowego, |
| *F* | - statystykę F, |
| *df* | - liczbę stopni swobody, |
| *ssreg* | - sumę kwadratów wyjaśnioną regresją, |
| *Ssresid* | - sumę kwadratów nie wyjaśnioną regresją. |

**W wyniku zastosowania funkcji REGLINP otrzymujemy:**

| | |
|---|---|
| 1294.053321 | 25855.31826 |
| 37.75274933 | 2503.665761 |
| 0.646298476 | 21711.11342 |
| 1174.916964 | 643 |
| 5.53823E+11 | 3.03092E+11 |

**co należy zapisać następująco:**

$$y_i = 1294{,}05 \cdot x_i + 25855{,}52 + e_i$$

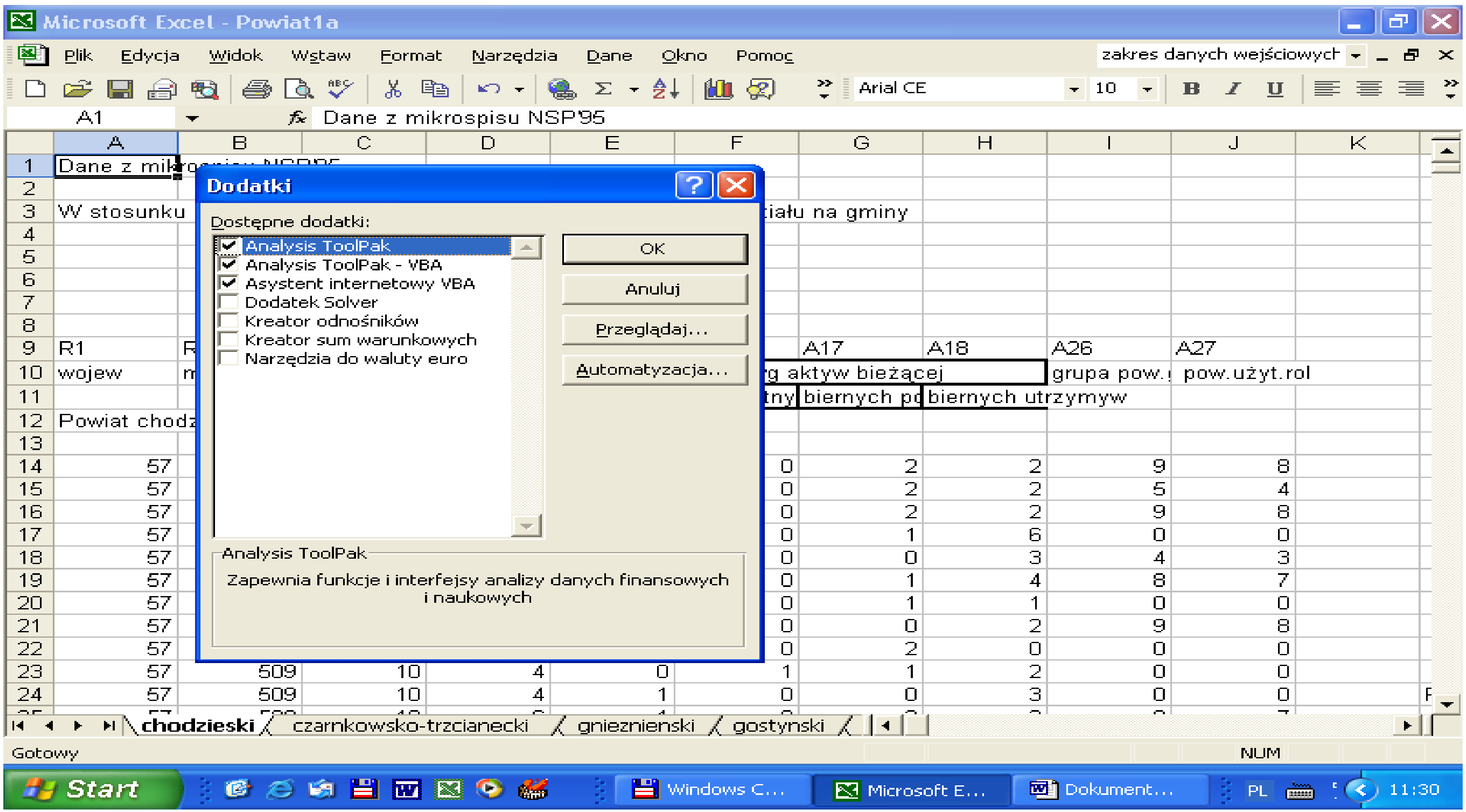
$$(37{,}75) \qquad (2503{,}67) \quad (21711{,}11)$$

$$R^2 = 0{,}646$$

# Analiza regresji w *Excelu*

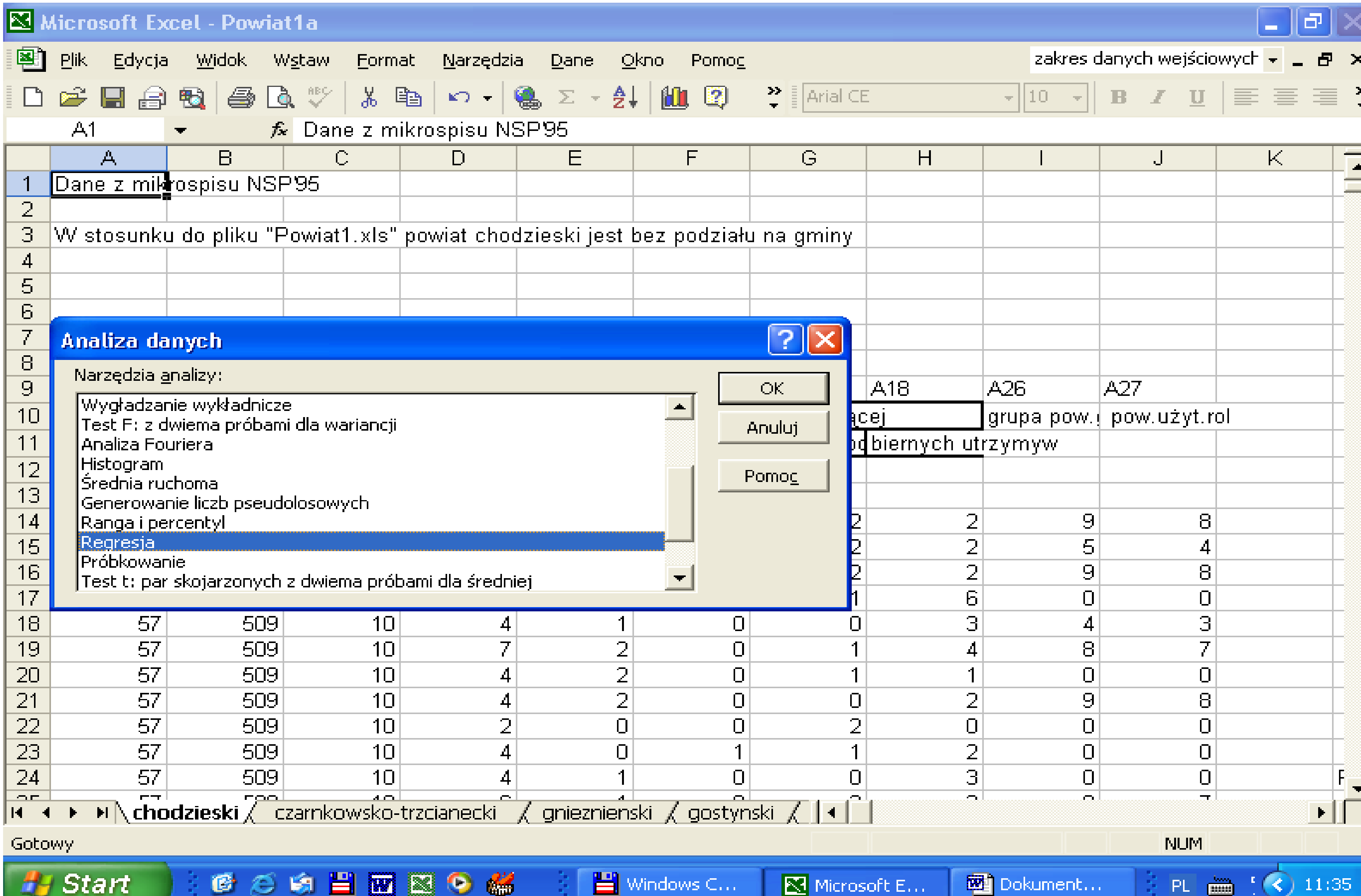
**2) Wybór metody:** ***Narzędzia*** **➔*Analiza danych***

**Przygotowanie Dodatków**



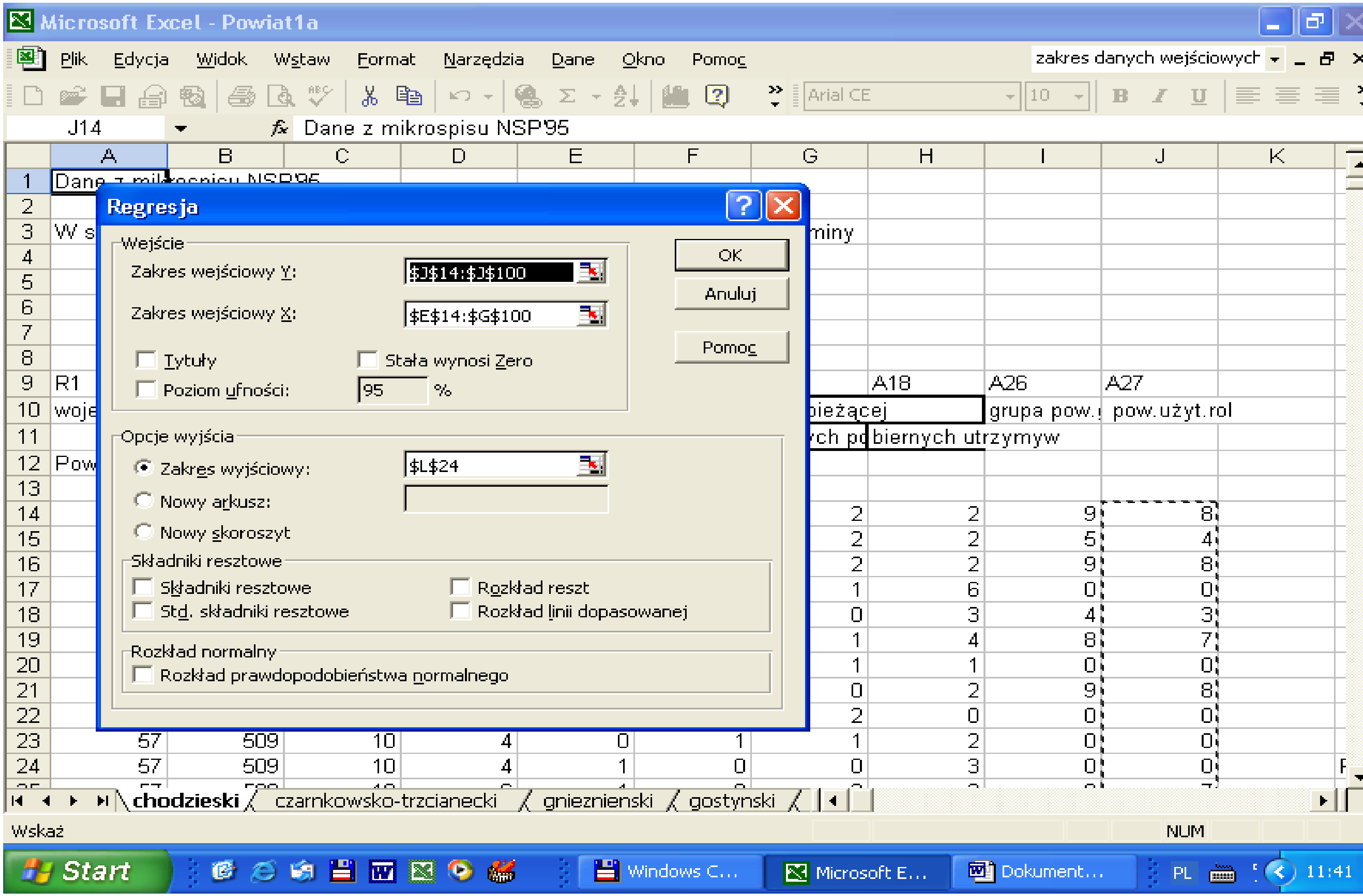
# Analiza regresji w *Excelu*

*2) Narzędzia ➔Analiza danych*



# Analiza regresji w *Excelu*

## *2) Okno dialogowe Regresja — wypełnienie panelu*



# Analiza regresji w *Excelu*

***2)*** **Okno dialogowe Regresja — wypełnienie panelu**

**Dane wejściowe -** Dane wejściowe powinny być numeryczne w postaci wektorów kolumnowych o tej samej liczbie wierszy.

**Zakres wejściowy Y -** Wprowadź odwołanie do zakresu wejściowego zmiennej zależnej. Zakres musi składać się z pojedynczej kolumny danych.

**Zakres wejściowy X -** Wprowadź odwołanie do zakresu wejściowego zmiennych niezależnych. Program Microsoft Excel porządkuje zmienne niezależne z tego zakresu rosnąco od lewej do prawej. Maksymalna liczba zmiennych niezależnych jest równa 16.

**Tytuły -** Zaznacz to pole wyboru, jeżeli pierwszy wiersz albo pierwsza kolumna zakresów wejściowych zawiera etykiety. Wyczyść je, jeżeli zakres wejściowy nie zawiera etykiet; program *Excel* generuje odpowiednie etykiety danych w tabeli wyjściowej.

**Poziom ufności -** Zaznacz to pole, aby uwzględnić dodatkowy poziom w podsumowującej tabeli wyjściowej. W polu wprowadź wartość poziomu ufności, który będzie stosowany oprócz domyślnego poziomu 95 procent.

**Stała wynosi zero -** Zaznacz to pole, jeżeli chcesz wymusić, aby linia regresji przechodziła przez początek układu współrzędnych.

**Zakres wyjściowy -** Wprowadź odwołanie do lewej górnej komórki tabeli wyników.

Jeśli spełnione są założenia (3-6), to $b_i$ są najlepszymi estymatorami parametrów $\beta_i$

**Nowy arkusz** - Kliknij, aby wstawić w bieżącym skoroszycie nowy arkusz i wkleić do niego wyniki, rozpoczynając od komórki A1. Nazwę nowego arkusza wpisz w polu.

**Nowy skoroszyt -** Kliknij, aby utworzyć nowy skoroszyt i wkleić wyniki do nowego arkusza w nowym skoroszycie.

**Składniki resztkowe -** Zaznacz to pole, aby uwzględnić składniki resztkowe w tabeli wyjściowej składników resztkowych.

**Standaryzowane składniki resztkowe -** Zaznacz to pole, aby uwzględnić standaryzowane składniki resztkowe w tabeli wyjściowej składników resztkowych.

**Rozkład reszt** - Zaznacz to pole wyboru, aby wygenerować wykres każdej zmiennej niezależnej w funkcji składnika resztkowego.

**Rozkład linii dopasowanej** - Zaznacz to pole, aby wygenerować wykres wartości teoretycznych (prognozowanych)

**Rozkład prawdopodobieństwa normalnego -** Zaznacz to pole, aby wygenerować wykres rozkładu prawdopodobieństwa normalnego.

# Analiza regresji w *Excelu*

## *2)* Wyniki analizy regresji

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| PODSUMOWANIE - WYJŚCIE | | | | | | | | |
| | | | | | | | | |
| *Statystyki regresji* | | | | | | | | |
| Wielokrotność R | 0,735786196 | | | | | | | |
| R kwadrat | 0,541381326 | | | | | | | |
| Dopasowany R kwadrat | 0,524804747 | | | | | | | |
| Błąd standardowy | 2,275715012 | | | | | | | |
| Obserwacje | 87 | | | | | | | |
| | | | | | | | | |
| ANALIZA WARIANCJI | | | | | | | | |
| | *df* | *SS* | *MS* | *F* | *Istotność F* | | | |
| Regresja | 3 | 507,417426 | 169,1391 | 32,659413 | 4,86E-14 | | | |
| Resztkowy | 83 | 429,8469418 | 5,178879 | | | | | |
| Razem | 86 | 937,2643678 | | | | | | |
| | | | | | | | | |
| | *Współczynniki* | *Błąd standardowy* | *t Stat* | *Wartość-p* | *Dolne 95%* | *Górne 95%* | *Dolne 95,0%* | *Górne 95,0%* |
| Przecięcie | -1,123984475 | 0,55911616 | -2,01029 | 0,0476482 | -2,236044 | -0,011925 | -2,236044331 | -0,01192462 |
| Zmienna X 1 | 2,282432125 | 0,25559222 | 8,929975 | 9,017E-14 | 1,7740693 | 2,7907949 | 1,774069342 | 2,790794909 |
| Zmienna X 2 | -0,443186869 | 0,726899877 | -0,60969 | 0,5437297 | -1,888962 | 1,0025881 | -1,888961882 | 1,002588145 |
| Zmienna X 3 | 1,186514967 | 0,320388406 | 3,703364 | 0,0003821 | 0,5492751 | 1,8237548 | 0,549275142 | 1,823754793 |