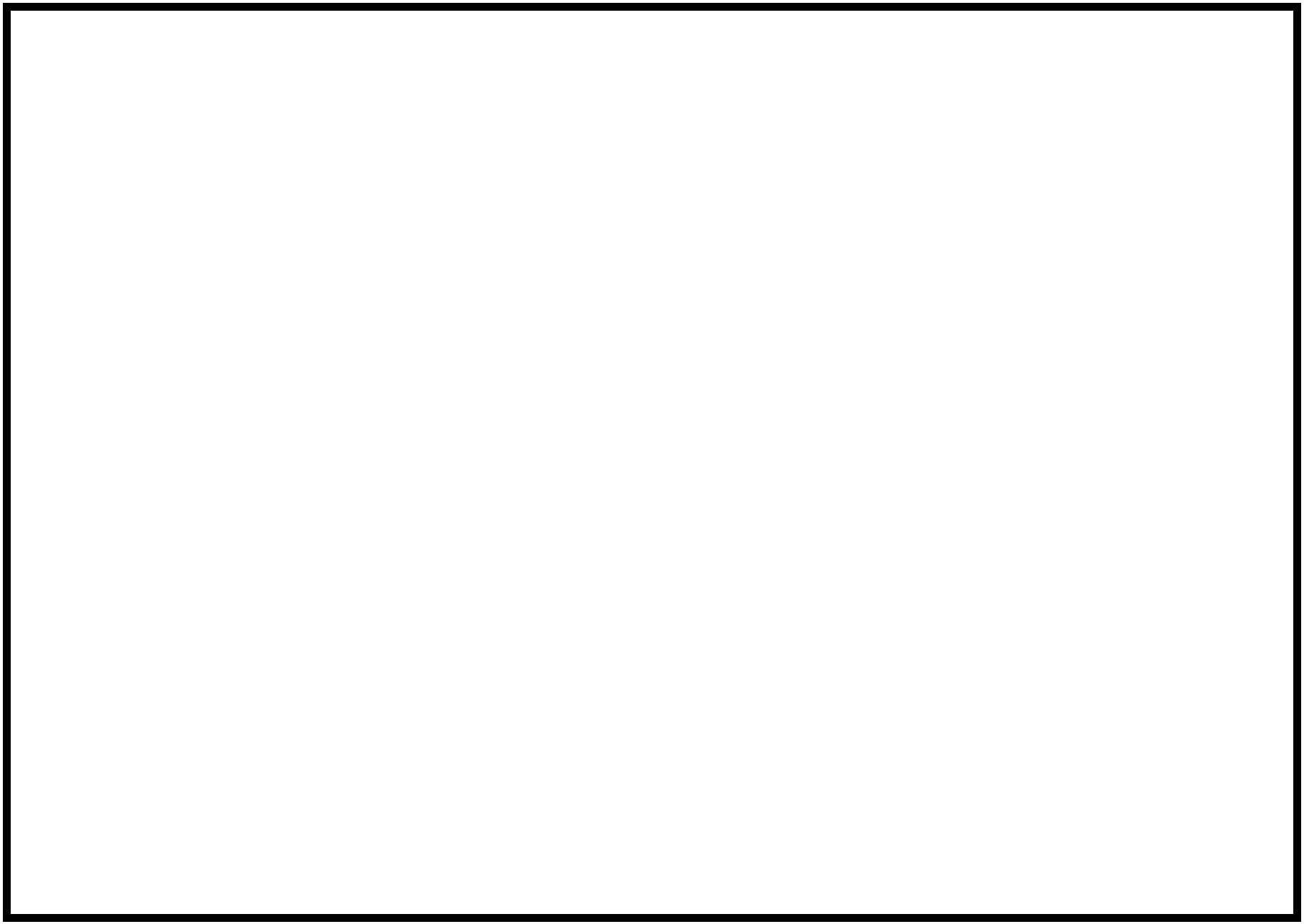
# Metody Inżynierii Wiedzy
# Regresja liniowa - wykład 6

Adam Szmigielski
aszmigie@pjwstk.edu.pl
materiały: $ftp(public) : //aszmigie/MIW$

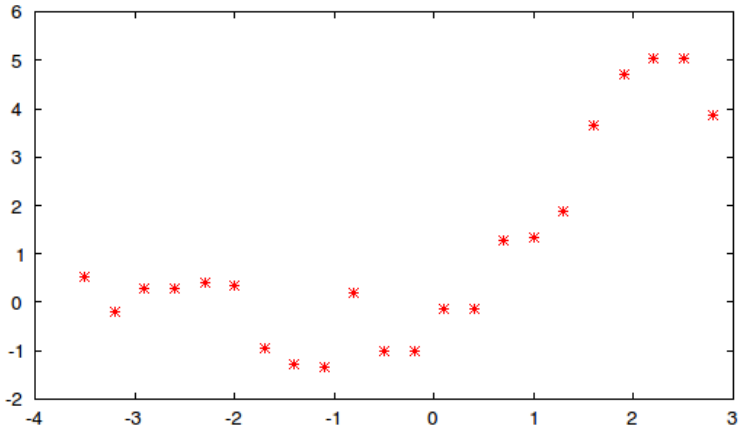
**Rodzaje modeli -test**

- *modele parametryczne* - gdy ilość parametrów określających model jest skończona
  np. możemy traktować parametry równania opisujące w/w krzywą jako model,
- *modele nieparametryczne* - gdy ilość parametrów opisujących model jest nieskończona
  np. możemy traktować krzywą (nieskończony zbiór punktów) jako model.



# "Czarna skrzynka"

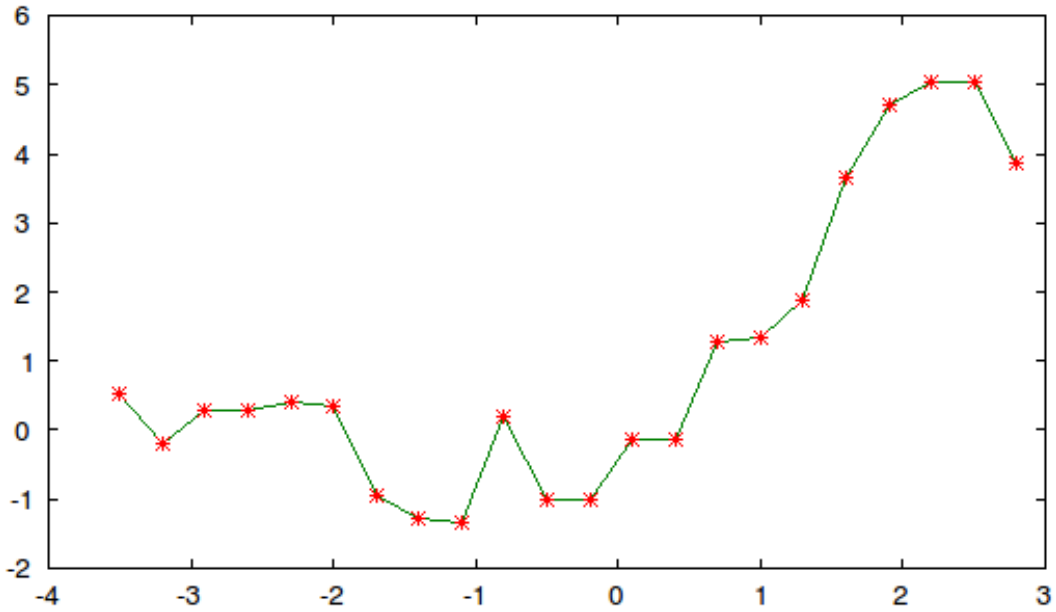
| x | y |
|---|---|
| -3,5000 | 0,5383 |
| -3,2000 | -0,2035 |
| -2,9000 | 0,2970 |
| -2,6000 | 0,2838 |
| -2,3000 | 0,3988 |
| -2,0000 | 0,3455 |
| -1,7000 | -0,9495 |
| -1,4000 | -1,2928 |
| -1,1000 | -1,3353 |
| -0,8000 | 0,1829 |
| -0,5000 | -1,0107 |
| -0,2000 | -1,0122 |
| 0,1000 | -0,1372 |
| 0,4000 | -0,1440 |
| 0,7000 | 1,2802 |
| 1,0000 | 1,3467 |
| 1,3000 | 1,8867 |
| 1,6000 | 3,6428 |
| 1,9000 | 4,6996 |
| 2,2000 | 5,0316 |
| 2,5000 | 5,0429 |
| 2,8000 | 3,8659 |





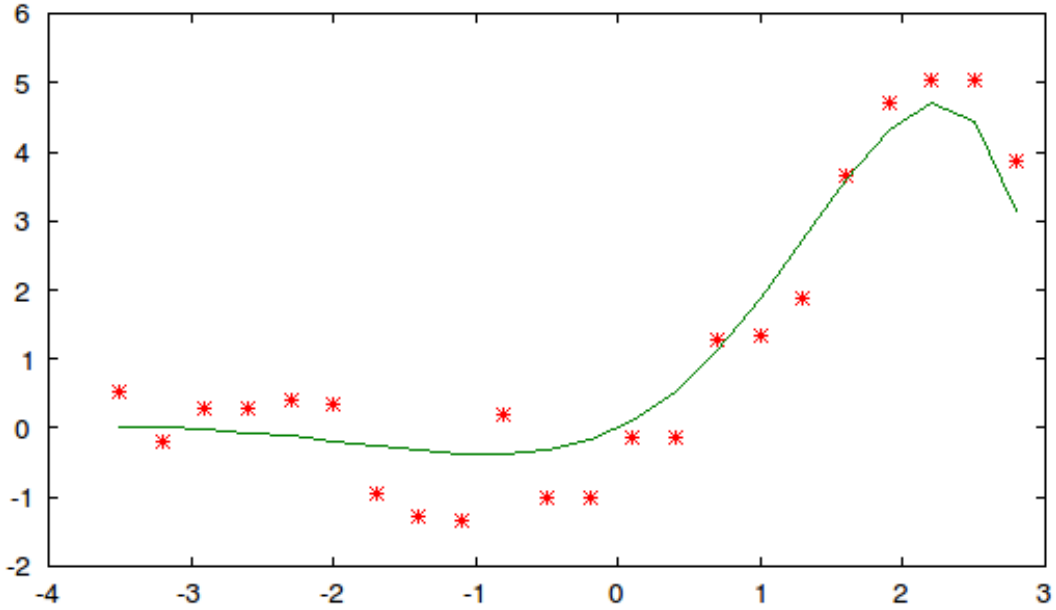
"Czarna skrzynka" mapuje wejście na wyjście.

## Interpolacja




- **Interpolacja** – metoda polegająca na wyznaczaniu w danym przedziale funkcji interpolacyjnej, która przyjmuje w nim z góry zadane wartości, w ustalonych punktach nazywanych węzłami.

## Aproksymacja




- **Aproksymacja** – metoda polegająca na określeniu rozwiązań przybliżonych, które są bliskie rozwiązaniom dokładnym.

**Zależność wejścia od wyjścia - korelacja**

- **Współczynnik korelacji** – liczba określająca w jakim stopniu zmienne są współzależne.
- Istnieje wiele różnych wzorów określanych jako współczynniki korelacji.
- Większość z nich jest znormalizowana: +1 (zupełna korelacja dodatnia), -1 (zupełna korelacja ujemna), 0 (brak korelacji),
- Najczęściej stosowany jest współczynnik korelacji r Pearsona - korelacja liniowa.

**Współczynnik korelacji liniowej Pearsona**

- *Współczynnik korelacji liniowej Pearsona* - współczynnik określający poziom zależności liniowej między zmiennymi losowymi $X$ i $Y$.
- *Współczynnika korelacji liniowej* definiuje się następująco:

$$r_{xy} = \frac{n \cdot \sum_{i=1}^{n}(x_i \cdot y_i) - \sum_{i=1}^{n}(x) \cdot \sum_{i=1}^{n}(y)}{(\sqrt{n \cdot \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \cdot \sqrt{n \cdot \sum_{i=1}^{n} y_i^2 - (\sum_{i=1}^{n} y_i)^2}}$$

- Współczynnik korelacji liniowej dwóch zmiennych jest ilorazem *kowariancji* i iloczynu *odchyleń standardowych* tych zmiennych:

$$r_{xy} = \frac{cov(X, Y)}{\delta_X \cdot \delta_Y}$$

## Model liniowy

- Regresja liniowa zakłada to, że wartość oczekiwana wyjścia, przy danym wejściu $E[y|x]$ jest liniowa - strukturą modelu jest prosta,
- W najprostszym przypadku $y = w \cdot x$,
- Zadanie polega na takim doborze parametru $w$ który najlepiej opisuje danych $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, tj. chcemy maksymalizować prawdopodobieństwo:

$$P(w|x_1, x_2, x_3, \ldots x_n, y_1, y_2 \ldots y_n)$$

- Do maksymalizacji tego prawdopodobieństwa możemy:
  - użyć reguły Bayesa do określenia prawdopodobieństwa warunkowego dla naszych danych,
  - użyć metody największej wiarygodności (ang. *Maximum Likelihood Estimation*).

**Metoda najmniejszych kwadratów (ang. Least Square LS)**

Najbardziej wiarygodny parametr $w$ powinien minimalizować sumę kwadratów błędu

$$E(w) = \sum_{i=1}^{n}(y_i - w \cdot x_i)^2 = \sum_{i=1}^{n} y_i^2 - (2\sum_{i=1}^{n} x_i y_i)w + (\sum_{i=1}^{n} x_i^2)w^2 \longrightarrow min$$

obliczając pochodną błędu po parametrze $w$ i przyrównując wynik do zera otrzymujemy:

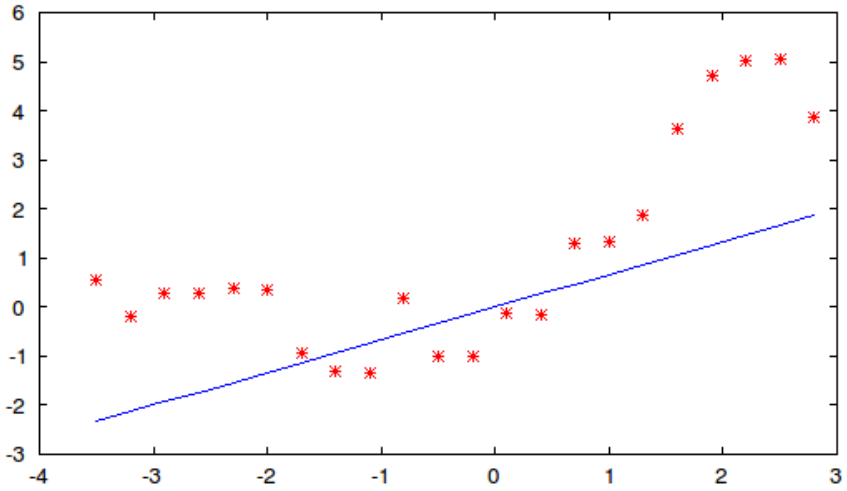
$$\frac{dE(w)}{dw} = -2(\sum_{i=1}^{n} x_i y_i) + 2(\sum_{i=1}^{n} x_i^2)w = 0$$

skąd otrzymujemy:

$$w = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

## Odpowiedź modelu (1 parametr) - przykład

- Nasz model będzie ma postać: $y = w \cdot x$
- obliczone wartości: $w = 0,6663$



**Metoda LS dla modelu** $y = w_1 \cdot x + w_0$

Model będzie najlepiej dopasowany do danych gdy suma błędów kwadratowych osiągnie minimum

$$E(w_1, w_0) = \sum_{i=1}^{n} (y_i - (w_1 \cdot x_i + w_0))^2 \longrightarrow min$$

obliczając pochodne cząstkowe po $w_1$ i $w_0$ i porównując je do zera otrzymujemy

$$\frac{\delta E(w_1, w_0)}{\delta w_1} = 2 \cdot (\sum_{i=1}^{n} (y_i - w_1 \cdot x_i - w_0)) \cdot (-x_i) = 0$$

$$\frac{\delta E(w_1, w_0)}{\delta w_0} = 2 \cdot (\sum_{i=1}^{n} (y_i - w_1 \cdot x_i - w_0)) \cdot (-1) = 0$$

po przekształceniach:

$$w_0 n + w_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$w_0 \sum_{i=1}^{n} x_i + w_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

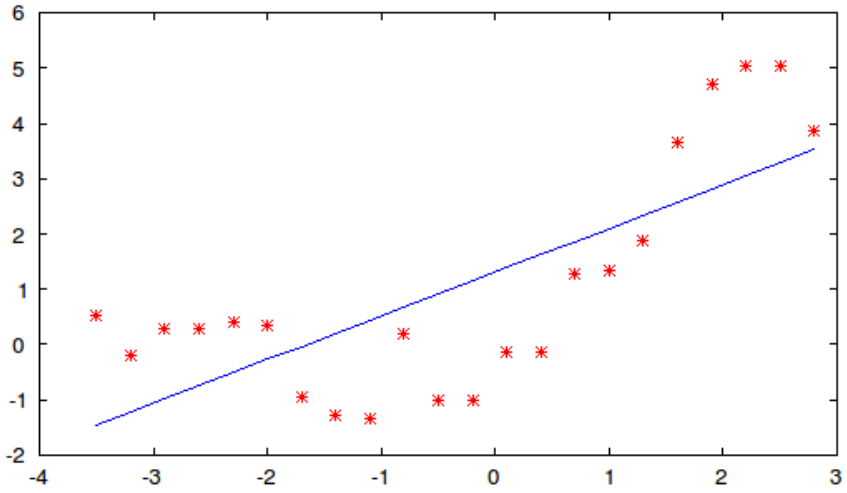
co daje wartości parametrów $w_1$ i $w_0$

$$w_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$w_0 = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

**Odpowiedź modelu $y = w_1 \cdot x + w_0$ - przykład**

- Nasz model będzie ma postać: $y = w_1 \cdot x + w_0$
- obliczone wartości: $w_1 = 0,7888$ i $w_0 = 1,3105$



## Regresja nieliniowa

Załóżmy, że mechanizm mapujący wejście na wyjście jest wielomianem drugiego stopnia:

$$y = w_2 \cdot x^2 + w_1 \cdot x + w_0$$

- Model będzie najlepiej dopasowany do danych gdy suma błędów kwadratowych osiągnie minimum

$$E(w_2, w_1, w_0) = \sum_{i=1}^{n} (y_i - (w_2 \cdot x_i^2 + w_1 \cdot x_i + w_0))^2 \longrightarrow min$$

Podobnie jak w poprzednich przypadkach obliczamy pochodne cząstkowe po $w_2$, $w_1$ i $w_0$ i porównując je do zera:

$$E(w_2, w_1, w_0) = \sum_{i=1}^{n} (y_i - (w_2 \cdot x_i^2 + w_1 \cdot x_i + w_0))^2$$

$$\frac{\delta E(w_2, w_1, w_0)}{\delta w_2} = 2 \cdot \sum_{i=1}^{n} (y_i - w_2 \cdot x_i^2 - w_1 \cdot x_i - w_0) \cdot (-x_i^2) = 0$$

$$\frac{\delta E(w_2, w_1, w_0)}{\delta w_1} = 2 \cdot \sum_{i=1}^{n} (y_i - w_2 \cdot x_i^2 - w_1 \cdot x_i - w_0) \cdot (-x_i) = 0$$

$$\frac{\delta E(w_2, w_1, w_0)}{\delta w_0} = 2 \cdot \sum_{i=1}^{n} (y_i - w_2 \cdot x_i^2 - w_1 \cdot x_i - w_0) \cdot (-1) = 0$$

Po uporządkowaniu otrzymujemy układ 3 równań z 3 niewiadomymi:

$$w_2 \sum_{i=1}^{n} x_i^4 + w_1 \sum_{i=1}^{n} x_i^3 + w_0 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i^2 y_i$$

$$w_2 \sum_{i=1}^{n} x_i^3 + w_1 \sum_{i=1}^{n} x_i^2 + w_0 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i$$

$$w_2 \sum_{i=1}^{n} x_i^2 + w_1 \sum_{i=1}^{n} x_i + w_0 \cdot n = \sum_{i=1}^{n} y_i$$

lub w postaci macierzowej

$$\begin{bmatrix} \sum_{i=1}^{n} x_i^2 y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i & n \end{bmatrix} \cdot \begin{bmatrix} w_2 \\ w_1 \\ w_0 \end{bmatrix}$$

Skąd można wyznaczyć wektor parametrów $[w_2, w_1, w_0]$ jako:

$$\begin{bmatrix} w_2 \\ w_1 \\ w_0 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i^4 & \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 \\ \sum_{i=1}^{n} x_i^3 & \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i & n \end{bmatrix}^{-1} \cdot \begin{bmatrix} \sum_{i=1}^{n} x_i^2 y_i \\ \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}$$

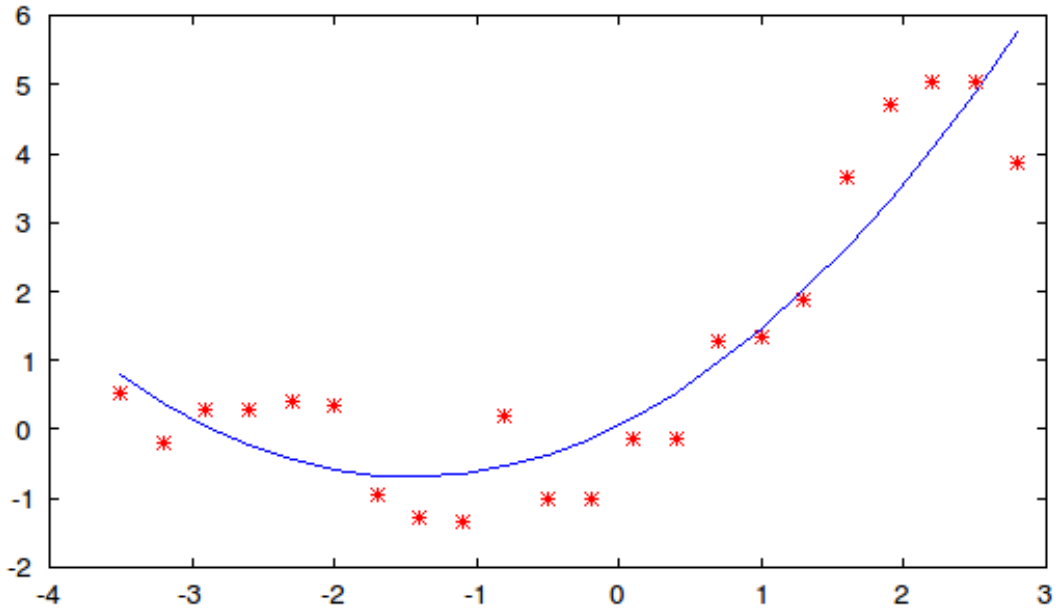
**Odpowiedź modelu** $y = w_2x^2 + w_1x + w_0$

Obliczone parametry:

$w_2 = 0,3561$

$w_1 = 1,0381$

$w_0 = 0,0641$



**Metoda najmniejszych kwadratów - wersja z macierzą obserwacji Fishera**

- Załóżmy, że zależność wyjścia od wejścia opisany jest równaniem

$$y = w_1 \cdot x_1 + w_2 \cdot x_2 + w_3 \cdot x_3 \tag{1}$$

  gdzie $w_1, w_2, w_3$ są parametrami modelu.

- Model posiada trzy wejścia $x_1$, $x_2$ i $x_3$ oraz jedno wyjście $y$.

- Załóżmy, że dysponujemy $n$ liczbą danych. $x_{i1}$ będzie oznaczać wartość 1 wejścia dla i-tej próbki, a $y_i$ wartość wyjścia modelu dla i-tej próbki.

Dla tych danych mamy następujące równanie macierzowe:

$$\underbrace{\begin{bmatrix} y_1 \\ . \\ . \\ y_n \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} x_{11} & x_{12} & x_{13} \\ . & . & . \\ . & . & . \\ x_{n1} & x_{n2} & x_{n3} \end{bmatrix}}_{\Phi} \cdot \underbrace{\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}}_{W}$$

gdzie $Y$ jest wektorem wyjściowym, $W$ wektorem parametrów, a $\phi$ nosi nazwę **macierzy obserwacji Fishera** (ang. *Fisher observation matrix*)

Z równia macierzowego $Y = \Phi \cdot W$ należy wyznaczyć wektor parametrów $W$.

## Rozwiązanie równania

$$Y = \Phi \cdot W \tag{2}$$

W ogólności macierz $\Phi$ nie jest kwadratowa. Równianie (2) mnożymy lewostronnie przez $\Phi^T$

$$\Phi^T \cdot Y = \Phi^T \cdot \Phi \cdot W \tag{3}$$

Ponieważ $\Phi^T \cdot \Phi$ jest macierzą kwadratową możemy obliczyć macierz do nie odwrotną i pomnożyć obie strony równania (3)

$$(\Phi^T \cdot \Phi)^{-1} \cdot \Phi^T \cdot Y = \underbrace{(\Phi^T \cdot \Phi)^{-1} \cdot (\Phi^T \cdot \Phi)}_{I} \cdot W$$

$I$ - macierz jednostkowa

Finalnie rozwiązaniem równia (2) jest wektor W:

$$W = \underbrace{(\phi^T \phi)^{-1} \cdot \phi^T}_{\text{macierz pseudoodrotna}} \cdot Y \tag{4}$$

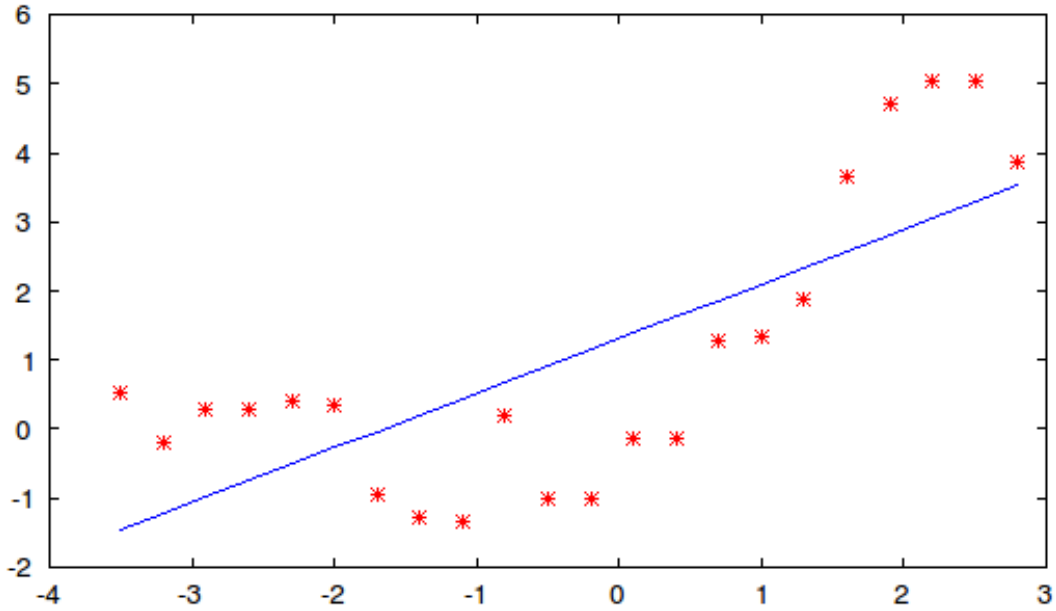
## Regresja liniowa - model $y = w_1 \cdot x + w_0$

$$Y = \begin{bmatrix} 0,5383 \\ -0,2035 \\ 0,2970 \\ 0,2838 \\ 0,3988 \\ 0,3455 \\ -0,9495 \\ -1,2928 \\ -1,3353 \\ 0,1829 \\ -1,0107 \\ -1,0122 \\ -0,1372 \\ -0,1440 \\ 1,2802 \\ 1,3467 \\ 1,8867 \\ 3,6428 \\ 4,6996 \\ 5,0316 \\ 5,0429 \\ 3,8659 \end{bmatrix} \quad X = \begin{bmatrix} -3,5000 \\ -3,2000 \\ -2,9000 \\ -2,6000 \\ -2,3000 \\ -2,0000 \\ -1,7000 \\ -1,4000 \\ -1,1000 \\ -0,8000 \\ -0,5000 \\ -0,2000 \\ 0,1000 \\ 0,4000 \\ 0,7000 \\ 1,0000 \\ 1,3000 \\ 1,6000 \\ 1,9000 \\ 2,2000 \\ 2,5000 \\ 2,8000 \end{bmatrix} \quad \Phi = \begin{bmatrix} -3,5000 & 1 \\ -3,2000 & 1 \\ -2,9000 & 1 \\ -2,6000 & 1 \\ -2,3000 & 1 \\ -2,0000 & 1 \\ -1,7000 & 1 \\ -1,4000 & 1 \\ -1,1000 & 1 \\ -0,8000 & 1 \\ -0,5000 & 1 \\ -0,2000 & 1 \\ 0,1000 & 1 \\ 0,4000 & 1 \\ 0,7000 & 1 \\ 1,0000 & 1 \\ 1,3000 & 1 \\ 1,6000 & 1 \\ 1,9000 & 1 \\ 2,2000 & 1 \\ 2,5000 & 1 \\ 2,8000 & 1 \end{bmatrix} \quad W = \begin{bmatrix} w_1 \\ w_0 \end{bmatrix}$$

**Odpowiedź modelu $y = w_1 x + w_0$ - macierz obserwacji Fishera**

Obliczone parametry:

$w_1 = 0,7888$, $w_0 = 1,3105$

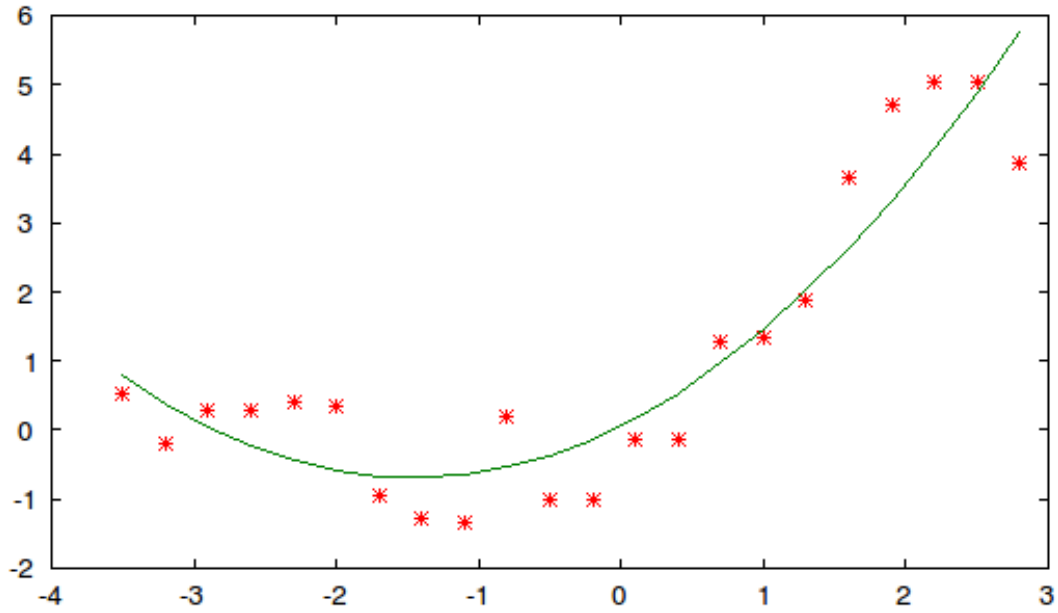


## Regresja liniowa - model $y = w_2 \cdot x^2 + w_1 \cdot x + w_0$

$$Y = \begin{bmatrix} 0,5383 \\ -0,2035 \\ 0,2970 \\ 0,2838 \\ 0,3988 \\ 0,3455 \\ -0,9495 \\ -1,2928 \\ -1,3353 \\ 0,1829 \\ -1,0107 \\ -1,0122 \\ -0,1372 \\ -0,1440 \\ 1,2802 \\ 1,3467 \\ 1,8867 \\ 3,6428 \\ 4,6996 \\ 5,0316 \\ 5,0429 \\ 3,8659 \end{bmatrix} \quad X = \begin{bmatrix} -3,5000 \\ -3,2000 \\ -2,9000 \\ -2,6000 \\ -2,3000 \\ -2,0000 \\ -1,7000 \\ -1,4000 \\ -1,1000 \\ -0,8000 \\ -0,5000 \\ -0,2000 \\ 0,1000 \\ 0,4000 \\ 0,7000 \\ 1,0000 \\ 1,3000 \\ 1,6000 \\ 1,9000 \\ 2,2000 \\ 2,5000 \\ 2,8000 \end{bmatrix} \quad \Phi = \begin{bmatrix} 12,2500 & -3,5000 & 1 \\ 10,2400 & -3,2000 & 1 \\ 8,4100 & -2,9000 & 1 \\ 6,7600 & -2,6000 & 1 \\ 5,2900 & -2,3000 & 1 \\ 4,0000 & -2,0000 & 1 \\ 2,8900 & -1,7000 & 1 \\ 1,9600 & -1,4000 & 1 \\ 1,2100 & -1,1000 & 1 \\ 0,6400 & -0,8000 & 1 \\ 0,2500 & -0,5000 & 1 \\ 0,0400 & -0,2000 & 1 \\ 0,0100 & 0,1000 & 1 \\ 0,1600 & 0,4000 & 1 \\ 0,4900 & 0,7000 & 1 \\ 1,0000 & 1,0000 & 1 \\ 1,6900 & 1,3000 & 1 \\ 2,5600 & 1,6000 & 1 \\ 3,6100 & 1,9000 & 1 \\ 4,8400 & 2,2000 & 1 \\ 6,2500 & 2,5000 & 1 \\ 7,8400 & 2,8000 & 1 \end{bmatrix} \quad W = \begin{bmatrix} w_2 \\ w_1 \\ w_0 \end{bmatrix}$$

**Odpowiedź modelu $y = w_2x^2 + w_1x + w_0$ - macierz obserwacji Fishera**

Obliczone parametry:
$w_2 = 0,3561;\ w_1 = 1,0381,\ w_0 = 0,0641$



**Modele autoregresyjne AR (ang.** *Autoregressive Model*)

- Model parametryczny szeregu czasowego, używany do modelowania i predykcji.
- Model autoregresyjny służy do predykcji liniowej – predykcji wyjścia układu oblicza się w oparciu o wartości wejść z przeszłości.

$$y_{t+1} = \varphi_1 \cdot y_t + \varphi_2 \cdot y_{t-1} + \varphi_3 \cdot y_{t-2} + \epsilon_t, \tag{5}$$

  gdzie $\varphi_1, \varphi_2, \varphi_3$ są parametrami a $\epsilon_t$ jest szumem.
- Wykorzystuje się inne rodzaje modeli do predykcji jak: model ARX, model ARMAX, model ARMA, model ARIMA etc.

**Zastosowanie regresji liniowej do predykcji notowań akcji**

Do opisu przebiegu notowań akcji posłużono się modelem autoregresyjnym

$$y_{t+1} = \varphi_1 \cdot y_t + \varphi_2 \cdot y_{t-1} + \varphi_3 \cdot y_{t-2} + \epsilon_t, \tag{6}$$

gdzie $\varphi_1, \varphi_2, \varphi_3$ są parametrami a $\epsilon_t$ jest szumem. Prognozowana wartość akcji liczona jest w oparciu o trójparametryczny predykcyjny modele autoregresyjny o postaci

$$\hat{y}_{t+1} = \varphi_1 \cdot y_t + \varphi_2 \cdot y_{t-1} + \varphi_3 \cdot y_{t-2}. \tag{7}$$

Parametry te są identyfikowane metodą najmniejszych kwadratów. Do identyfikacji użyto 49 wcześniejszych obserwacji (notowań cen akcji).

Macierz obserwacji $\phi$ ma postać:

$$\phi = \begin{bmatrix} y_t & y_{t-1} & y_{t-2} \\ . & . & . \\ . & . & . \\ y_{t-48} & y_{t-49} & y_{t-50} \end{bmatrix}$$

Wektor parametrów $\varphi = [\varphi_1, \varphi_2, \varphi_3]$ obliczono jako

$$\varphi = (\phi^T \phi)^{-1} \cdot \phi^T \cdot y \tag{8}$$

gdzie $y$ jest wektorem 49 obserwacji

$$y = \begin{bmatrix} y_t \\ . \\ . \\ y_{t-48} \end{bmatrix}$$

Pożądanym efektem prognozy jest to, by prognozowana wartość cen akcji $\hat{y}_{t+1}$ była faktycznie równa przyszłemu notowaniu akcji $y_{t+1}$. Niestety, każda prognoza jest obarczona błędem prognozy. Faktyczna wartość notowań akcji będzie więc równa sumie prognozy oraz błędu prognozy $e_t$

$$y_{t+1} = \hat{y}_{t+1} + \epsilon_t \tag{9}$$

Z równania ( 9 ) można przedstawić błąd prognozy na jeden krok naprzód jako

$$\epsilon_t = y_{t+1} - \hat{y}_{t+1} \tag{10}$$

## Zadania na laboratoria

- Wczytaj dane z pliku *daneXX.txt*[a]. Zaproponuj i zrealizuj podział tych danych na *dane treningowe* i *dane testowe*,
- Zaproponuj liniowy model parametryczny **Model 1**. Określ parametry modelu stosując metodę najmniejszych kwadratów dla danych treningowych,
- Zweryfikuj poprawność **Modelu 1**,
- Zaproponuj bardziej złożony model parametryczny **Model 2**. Określ parametry modelu stosując metodę najmniejszych kwadratów dla danych treningowych,
- Zweryfikuj poprawność **Modelu 2**,
- Porównaj oba modele.

[a]gdzie XX jest numerem zestawu. W każdej linii pliku pierwsza liczba określa wejście a druga wartość wyjścia