

# Techniki Optymalizacji: Metody regresji

Wojciech Kotłowski

Instytut Informatyki Politechniki Poznańskiej  
email: [imię.nazwisko@cs.put.poznan.pl](mailto:imię.nazwisko@cs.put.poznan.pl)

pok. 2 (CW) tel. (61)665-2936 konsultacje: wtorek 15:00-16:30  
Slajdy dostępne pod adresem: <http://www.cs.put.poznan.pl/wkotlowski/>

18.11.2013

- 1 Problem regresji
- 2 Metody regresji liniowej
- 3 Minimalizacja kwadratów błędów
- 4 Minimalizacja sumy wartości bezwzględnych błędów
- 5 Przykład: wycena domów

- 1 Problem regresji
- 2 Metody regresji liniowej
- 3 Minimalizacja kwadratów błędów
- 4 Minimalizacja sumy wartości bezwzględnych błędów
- 5 Przykład: wycena domów

# Problem regresji

- Przewidywania/wyjaśnienie zmian jednej zmiennej ( $Y$ ) pod wpływem zmian innych zmiennych ( $X$ ).
- Powód: zmienne  $X$  zwykle łatwe do pozyskania,  $Y$  – trudne lub niemożliwe do pozyskania

# Problem regresji

- Przewidywania/wyjaśnienie zmian jednej zmiennej ( $Y$ ) pod wpływem zmian innych zmiennych ( $X$ ).
- Powód: zmienne  $X$  zwykle łatwe do pozyskania,  $Y$  – trudne lub niemożliwe do pozyskania

## Przykłady

- $X$  – ceny akcji w ostatnim tygodniu,  $Y$  – cena akcji jutro.
- $X$  – wyniki testów medycznych,  $Y$  – poziom zaawansowania choroby.
- $X$  – wielkość programu,  $Y$  – czas pisania programu.
- $X$  – warunki na drodze, czas, lokalizacja,  $Y$  – średnia prędkość samochodów.
- $X$  – cechy domu  $Y$  – cena domu.

- Modelujemy zmienną  $Y$  jako funkcję liniową  $\mathbf{X}$ .
  - dla jednej zmiennej:

$$\hat{Y} = a_1 X + a_0$$

- dla wielu zmiennych:

$$\hat{Y} = a_1 X_1 + \dots + a_m X_m + a_0 = \mathbf{a}^\top \mathbf{X} + a_0$$

- Modelujemy zmienną  $Y$  jako funkcję liniową  $\mathbf{X}$ .
  - dla jednej zmiennej:

$$\hat{Y} = a_1 X + a_0$$

- dla wielu zmiennych:

$$\hat{Y} = a_1 X_1 + \dots + a_m X_m + a_0 = \mathbf{a}^\top \mathbf{X} + a_0$$

- Model liniowy jest ogólniejszy niż myślicie!
  - **Przykład:** regresja wielomianowa to regresja liniowa!
  - Mając  $X$ , wprowadzamy zmienne:  
 $X_1 = X, X_2 = X^2, X_3 = X^3, \dots$

$$\hat{Y} = a_1 X + a_2 X^2 + \dots + a_0 \implies \hat{Y} = a_1 X_1 + a_2 X_2 + \dots + a_0$$

# Jak to się zwykle robi

- Otrzymujemy zbiór danych historycznych, na którym znane są wartości  $Y$ :

$$\begin{array}{ll} (x_{11}, x_{12}, \dots, x_{1m}, y_1) & (\boldsymbol{x}_1, y_1) \\ (x_{21}, x_{22}, \dots, x_{2m}, y_2) & (\boldsymbol{x}_2, y_2) \\ \dots & \text{lub w skrócie } \dots \\ (x_{n1}, x_{n2}, \dots, x_{nm}, y_n) & (\boldsymbol{x}_n, y_n) \end{array}$$



# Jak to się zwykle robi

- Otrzymujemy zbiór danych historycznych, na którym znane są wartości  $Y$ :

$$\begin{array}{ll} (x_{11}, x_{12}, \dots, x_{1m}, y_1) & (\boldsymbol{x}_1, y_1) \\ (x_{21}, x_{22}, \dots, x_{2m}, y_2) & (\boldsymbol{x}_2, y_2) \\ \dots & \text{lub w skrócie } \dots \\ (x_{n1}, x_{n2}, \dots, x_{nm}, y_n) & (\boldsymbol{x}_n, y_n) \end{array}$$

- Wyznaczamy współczynniki  $a_0, a_1, \dots, a_m$  na danych.

# Jak to się zwykle robi

- Otrzymujemy zbiór danych historycznych, na którym znane są wartości  $Y$ :

$$\begin{array}{ll} (x_{11}, x_{12}, \dots, x_{1m}, y_1) & (\mathbf{x}_1, y_1) \\ (x_{21}, x_{22}, \dots, x_{2m}, y_2) & (\mathbf{x}_2, y_2) \\ \dots & \text{lub w skrócie } \dots \\ (x_{n1}, x_{n2}, \dots, x_{nm}, y_n) & (\mathbf{x}_n, y_n) \end{array}$$

- Wyznaczamy współczynniki  $a_0, a_1, \dots, a_m$  na danych.
- Testujemy nasz model na osobnym zbiorze testowym (również ze znanymi wartościami  $Y$ )

# Jak to się zwykle robi

- Otrzymujemy zbiór danych historycznych, na którym znane są wartości  $Y$ :

$$\begin{array}{ll} (x_{11}, x_{12}, \dots, x_{1m}, y_1) & (\mathbf{x}_1, y_1) \\ (x_{21}, x_{22}, \dots, x_{2m}, y_2) & (\mathbf{x}_2, y_2) \\ \dots & \text{lub w skrócie } \dots \\ (x_{n1}, x_{n2}, \dots, x_{nm}, y_n) & (\mathbf{x}_n, y_n) \end{array}$$

- Wyznaczamy współczynniki  $a_0, a_1, \dots, a_m$  na danych.
- Testujemy nasz model na osobnym zbiorze testowym (również ze znanymi wartościami  $Y$ )

## Przykład: szacowanie czasu pracy programistów

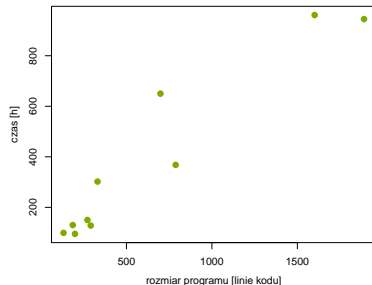
$X$	$Y$
Rozmiar programu	Oszacowany czas
186	130
699	650
132	99
272	150
291	128
331	302
199	95
1890	945
788	368
1601	961

# Przykład: szacowanie czasu pracy programistów

$X$	$Y$
Rozmiar programu	Oszacowany czas
$x_1 = 186$	$y_1 = 130$
$x_2 = 699$	$y_2 = 650$
$x_3 = 132$	$y_3 = 99$
$x_4 = 272$	$y_4 = 150$
$x_5 = 291$	$y_5 = 128$
$x_6 = 331$	$y_6 = 302$
$x_7 = 199$	$y_7 = 95$
$x_8 = 1890$	$y_8 = 945$
$x_9 = 788$	$y_9 = 368$
$x_{10} = 1601$	$y_{10} = 961$

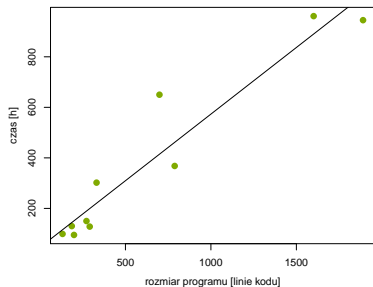
# Przykład: szacowanie czasu pracy programistów

$X$	$Y$
Rozmiar programu	Oszacowany czas
$x_1 = 186$	$y_1 = 130$
$x_2 = 699$	$y_2 = 650$
$x_3 = 132$	$y_3 = 99$
$x_4 = 272$	$y_4 = 150$
$x_5 = 291$	$y_5 = 128$
$x_6 = 331$	$y_6 = 302$
$x_7 = 199$	$y_7 = 95$
$x_8 = 1890$	$y_8 = 945$
$x_9 = 788$	$y_9 = 368$
$x_{10} = 1601$	$y_{10} = 961$



# Przykład: szacowanie czasu pracy programistów

$X$	$Y$
Rozmiar programu	Oszacowany czas
$x_1 = 186$	$y_1 = 130$
$x_2 = 699$	$y_2 = 650$
$x_3 = 132$	$y_3 = 99$
$x_4 = 272$	$y_4 = 150$
$x_5 = 291$	$y_5 = 128$
$x_6 = 331$	$y_6 = 302$
$x_7 = 199$	$y_7 = 95$
$x_8 = 1890$	$y_8 = 945$
$x_9 = 788$	$y_9 = 368$
$x_{10} = 1601$	$y_{10} = 961$



$$a_0 = 45.93, a_1 = 0.5273$$

- 1 Problem regresji
- 2 Metody regresji liniowej
- 3 Minimalizacja kwadratów błędów
- 4 Minimalizacja sumy wartości bezwzględnych błędów
- 5 Przykład: wycena domów



# Jak wyznaczyć współczynniki?

## Problem optymalizacji

Mając zbiór danych  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ , wyznacz współczynniki  $a_0, a_1, \dots, a_n$  tak, aby wartości przewidywane przez model:

$$\hat{y}_i = a_1 x_{i1} + \dots + a_n x_{in} + a_0 = \mathbf{a}^\top \mathbf{x}_i + a_0$$

na wszystkich danych ( $i = 1, \dots, n$ ) były jak najbliżej prawdziwych wartości  $y_i$ .

# Jak wyznaczyć współczynniki?

## Problem optymalizacji

Mając zbiór danych  $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)$ , wyznacz współczynniki  $a_0, a_1, \dots, a_n$  tak, aby wartości przewidywane przez model:

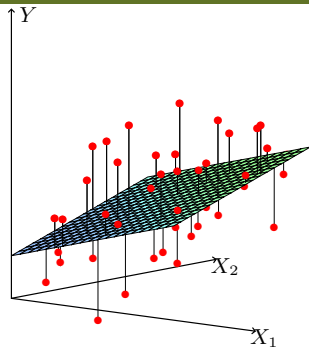
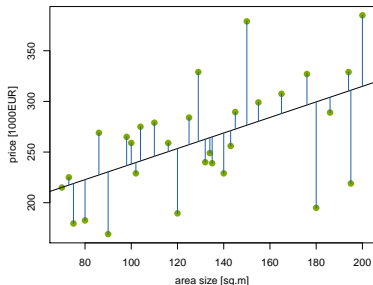
$$\hat{y}_i = a_1 x_{i1} + \dots + a_n x_{in} + a_0 = \mathbf{a}^\top \mathbf{x}_i + a_0$$

na wszystkich danych ( $i = 1, \dots, n$ ) były jak najbliżej prawdziwych wartości  $y_i$ .

- Uwaga: zwykle dodajemy jeszcze jedną zmienną wejściową  $X_0$  stale równą 1 i chowamy współczynnik  $a_0$  do wektora  $\mathbf{a}$ , otrzymując:

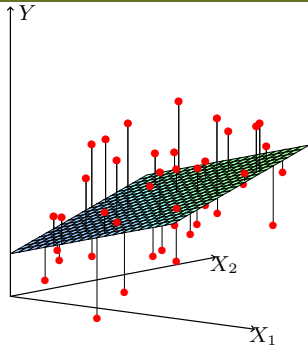
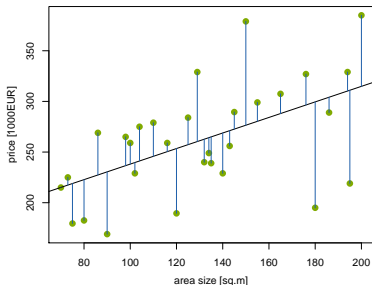
$$\hat{y}_i = a_0 x_{i0} + a_1 x_{i1} + \dots + a_n x_{in} = \mathbf{a}^\top \mathbf{x}_i$$

# Odchylenia (błędy) na danych



Źródło: Hastie, Tibshirani, Friedman, *Elements of statistical learning*.

# Odchylenia (błędy) na danych

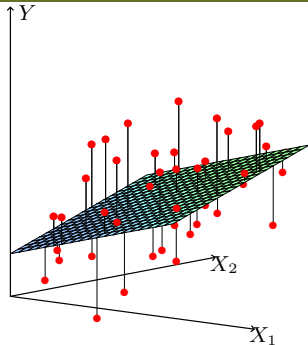
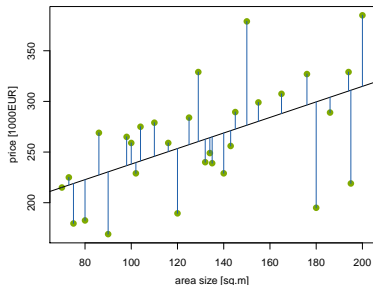


Źródło: Hastie, Tibshirani, Friedman, *Elements of statistical learning*.

- Odchylenie (błąd) na danym  $x_i$  to różnica między prawdziwą wartością  $y_i$ , a wartością przewidywaną przez model  $\hat{y}_i$  (punkt na prostej):

$$\delta_i = y_i - \hat{y}_i = y_i - \mathbf{a}^\top \mathbf{x}_i$$

# Odchylenia (błędy) na danych



Źródło: Hastie, Tibshirani, Friedman, *Elements of statistical learning*.

- Odchylenie (błąd) na danym  $x_i$  to różnica między prawdziwą wartością  $y_i$ , a wartością przewidywaną przez model  $\hat{y}_i$  (punkt na prostej):

$$\delta_i = y_i - \hat{y}_i = y_i - \mathbf{a}^\top \mathbf{x}_i$$

- Jak zmierzyć sumaryczny błąd?

# Trzy kryteria

- Minimalizacja sumy kwadratów błędów/odchyłeń (*least squares* – LS)

$$\min : \quad z = \sum_{i=1}^n \delta_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Minimalizacja sumy kwadratów błędów/odchyłeń  
(*least squares* – LS)

$$\min : \quad z = \sum_{i=1}^n \delta_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Minimalizacja sumy wartości bezwzględnych błędów/odchyłeń  
(*least absolute deviations* – LAD)

$$\min : \quad z = \sum_{i=1}^n |\delta_i| = \sum_{i=1}^n |y_i - \hat{y}_i|$$



- Minimalizacja sumy kwadratów błędów/odchyłeń  
(*least squares* – LS)

$$\min : \quad z = \sum_{i=1}^n \delta_i^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Minimalizacja sumy wartości bezwzględnych błędów/odchyłeń  
(*least absolute deviations* – LAD)

$$\min : \quad z = \sum_{i=1}^n |\delta_i| = \sum_{i=1}^n |y_i - \hat{y}_i|$$

- Minimalizacja największego z błędów  
(*minimax* – MM)

$$\min : \quad z = \max_{i=1, \dots, n} |\delta_i| = \max_{i=1, \dots, n} |y_i - \hat{y}_i|$$

	LS	LAD	MM
optymalizacja	analityczny wzór	progr. liniowe	progr. liniowe
stabilność rozwiązania	stabilne	niestabilne	niestabilne
wartości odstające	nieodporna	odporna	bardzo nieodporna

- Zwykle wybór między LS a LAD.
- MM nie nadaje się do stosowania w regresji! (poza wyjątkowymi przypadkami).

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $\mathbf{X}$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $\mathbf{X}$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_{100} = 1000$ .

Współczynniki wyznaczone na danych:

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $\mathbf{X}$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_6 = 1000$ .

Współczynniki wyznaczone na danych:

- Minimalizacja kwadratów błędów (LS):

$$\min_{a_0} \left\{ 99 \cdot (a_0 - 1)^2 + (a_0 - 1000)^2 \right\}$$

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $\mathbf{X}$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_{100} = 1000$ .

Współczynniki wyznaczone na danych:

- Minimalizacja kwadratów błędów (LS):

$$\min_{a_0} \left\{ 99 \cdot (a_0 - 1)^2 + (a_0 - 1000)^2 \right\} \implies a_0 = \text{avg}(\mathbf{y}) = 10.99$$

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $\mathbf{X}$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_{100} = 1000$ .

Współczynniki wyznaczone na danych:

- Minimalizacja kwadratów błędów (LS):

$$\min_{a_0} \left\{ 99 \cdot (a_0 - 1)^2 + (a_0 - 1000)^2 \right\} \implies a_0 = \text{avg}(\mathbf{y}) = 10.99$$

- Minimalizacja wartości bezwzględnych (błędów) (LAD):

$$\min_{a_0} \left\{ 99 \cdot |a_0 - 1| + |a_0 - 1000| \right\}$$

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $X$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_{100} = 1000$ .

Współczynniki wyznaczone na danych:

- Minimalizacja kwadratów błędów (LS):

$$\min_{a_0} \left\{ 99 \cdot (a_0 - 1)^2 + (a_0 - 1000)^2 \right\} \implies a_0 = \text{avg}(\mathbf{y}) = 10.99$$

- Minimalizacja wartości bezwzględnych (błędów) (LAD):

$$\min_{a_0} \left\{ 99 \cdot |a_0 - 1| + |a_0 - 1000| \right\} \implies a_0 = \text{median}(\mathbf{y}) = 1$$



# Odporność na wartości odstające

Najprostszy przypadek: Brak  $X$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_{100} = 1000$ .

Współczynniki wyznaczone na danych:

- Minimalizacja kwadratów błędów (LS):

$$\min_{a_0} \left\{ 99 \cdot (a_0 - 1)^2 + (a_0 - 1000)^2 \right\} \implies a_0 = \text{avg}(\mathbf{y}) = 10.99$$

- Minimalizacja wartości bezwzględnych (błędów) (LAD):

$$\min_{a_0} \left\{ 99 \cdot |a_0 - 1| + |a_0 - 1000| \right\} \implies a_0 = \text{median}(\mathbf{y}) = 1$$

- Minimalizacja największego błędu (MM):

$$\min_{a_0} \max \left\{ |a_0 - 1|, |a_0 - 1000| \right\}$$

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $X$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_{100} = 1000$ .

Współczynniki wyznaczone na danych:

- Minimalizacja kwadratów błędów (LS):

$$\min_{a_0} \left\{ 99 \cdot (a_0 - 1)^2 + (a_0 - 1000)^2 \right\} \implies a_0 = \text{avg}(\mathbf{y}) = 10.99$$

- Minimalizacja wartości bezwzględnych (błędów) (LAD):

$$\min_{a_0} \left\{ 99 \cdot |a_0 - 1| + |a_0 - 1000| \right\} \implies a_0 = \text{median}(\mathbf{y}) = 1$$

- Minimalizacja największego błędu (MM):

$$\min_{a_0} \max \left\{ |a_0 - 1|, |a_0 - 1000| \right\} \implies a_0 = \text{middle}(\mathbf{y}) = 500.5$$

# Odporność na wartości odstające

Najprostszy przypadek: Brak  $\mathbf{X}$ , tylko  $Y$ .

Model zawiera tylko stałą:

$$\hat{y} = a_0.$$

Dane: 100 punktów,  $y_1 = y_2 = \dots = y_{99} = 1$  oraz jedna przypadkowo źle wpisana wartość  $y_{100} = 1000$ .

Współczynniki wyznaczone na danych:

- Minimalizacja kwadratów błędów (LS):

$$\min_{a_0} \left\{ 99 \cdot (a_0 - 1)^2 + (a_0 - 1000)^2 \right\} \implies a_0 = \text{avg}(\mathbf{y}) = 10.99$$

- Minimalizacja wartości bezwzględnych (błędów) (LAD):

$$\min_{a_0} \left\{ 99 \cdot |a_0 - 1| + |a_0 - 1000| \right\} \implies a_0 = \text{median}(\mathbf{y}) = 1$$

- Minimalizacja największego błędu (MM):

$$\min_{a_0} \max \left\{ |a_0 - 1|, |a_0 - 1000| \right\} \implies a_0 = \text{middle}(\mathbf{y}) = 500.5$$

Gdy są  $\mathbf{X}$ , prosta regresji dla LS i MM będzie przyciągana zbyt mocno do wartości odstających! (szczególnie MM: tragedia!)

- 1 Problem regresji
- 2 Metody regresji liniowej
- 3 Minimalizacja kwadratów błędów**
- 4 Minimalizacja sumy wartości bezwzględnych błędów
- 5 Przykład: wycena domów

„Metoda najmniejszych kwadratów”

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

„Metoda najmniejszych kwadratów”

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Funkcja wypukła, kwadratowa.

„Metoda najmniejszych kwadratów”

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Funkcja wypukła, kwadratowa.
- Rozwiązanie poprzez przyrównanie pochodnych po wszystkich  $a_j$  do 0:

$$\mathbf{a}_{LS} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right),$$

# Metoda najmniejszych kwadratów



# Metoda najmniejszych kwadratów

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

# Metoda najmniejszych kwadratów

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij}$$

# Metoda najmniejszych kwadratów

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij}$$

- Przyprowadzenie do zera:

$$\sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij} = 0 \quad \implies \quad \sum_{i=1}^n y_i x_{ij} = \sum_{i=1}^n \sum_{k=1}^m a_k x_{ik} x_{ij}$$

# Metoda najmniejszych kwadratów

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij}$$

- Przyprowadzenie do zera:

$$\sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij} = 0 \quad \implies \quad \sum_{i=1}^n y_i x_{ij} = \sum_{i=1}^n \sum_{k=1}^m a_k x_{ik} x_{ij}$$

- Wektorowo:

$$\sum_{i=1}^n y_i \mathbf{x}_i = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{a} \quad \implies \quad \mathbf{a} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right).$$

# MNK jako metoda Newtona-Raphsona

# MNK jako metoda Newtona-Raphsona

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

# MNK jako metoda Newtona-Raphsona

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij}$$

# MNK jako metoda Newtona-Raphsona

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij}$$

- Gradient:

$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$



# MNK jako metoda Newtona-Raphsona

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij}$$

- Gradient:

$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$

- Drugie pochodne:

$$\frac{\partial^2 L(\mathbf{a})}{\partial a_j \partial a_k} = \sum_{i=1}^n 2x_{ik}x_{ij}$$

# MNK jako metoda Newtona-Raphsona

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij}$$

- Gradient:

$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$

- Drugie pochodne:

$$\frac{\partial^2 L(\mathbf{a})}{\partial a_j \partial a_k} = \sum_{i=1}^n 2x_{ik}x_{ij}$$

- Hesjan:

$$\mathbf{H}_L(\mathbf{a}) = \sum_{i=1}^n 2\mathbf{x}_i \mathbf{x}_i^\top$$

# MNK jako metoda Newtona-Raphsona

- Gradient: 
$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$
- Hesjan: 
$$\mathbf{H}_L(\mathbf{a}) = \sum_{i=1}^n 2 \mathbf{x}_i \mathbf{x}_i^\top$$

# MNK jako metoda Newtona-Raphsona

■ Gradient: 
$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$

■ Hesjan: 
$$\mathbf{H}_L(\mathbf{a}) = \sum_{i=1}^n 2 \mathbf{x}_i \mathbf{x}_i^\top$$

■ Rozwiązanie początkowe:  $\mathbf{a}_0 = 0$ .

$$\nabla_L(\mathbf{a}_0) = -2 \sum_{i=1}^n y_i \mathbf{x}_i, \quad \mathbf{H}_L(\mathbf{a}_0) = \sum_{i=1}^n 2 \mathbf{x}_i \mathbf{x}_i^\top$$

# MNK jako metoda Newtona-Raphsona

■ Gradient: 
$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$

■ Hesjan: 
$$\mathbf{H}_L(\mathbf{a}) = \sum_{i=1}^n 2\mathbf{x}_i \mathbf{x}_i^\top$$

■ Rozwiązanie początkowe:  $\mathbf{a}_0 = 0$ .

$$\nabla_L(\mathbf{a}_0) = -2 \sum_{i=1}^n y_i \mathbf{x}_i, \quad \mathbf{H}_L(\mathbf{a}_0) = \sum_{i=1}^n 2\mathbf{x}_i \mathbf{x}_i^\top$$

■ Krok metodą Newtona-Raphsona:

$$\mathbf{a}_1 = \mathbf{a}_0 - \mathbf{H}_L^{-1}(\mathbf{a}_0) \nabla_L(\mathbf{a}_0)$$

# MNK jako metoda Newtona-Raphsona

■ Gradient: 
$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$

■ Hesjan: 
$$\mathbf{H}_L(\mathbf{a}) = \sum_{i=1}^n 2 \mathbf{x}_i \mathbf{x}_i^\top$$

■ Rozwiązanie początkowe:  $\mathbf{a}_0 = \mathbf{0}$ .

$$\nabla_L(\mathbf{a}_0) = -2 \sum_{i=1}^n y_i \mathbf{x}_i, \quad \mathbf{H}_L(\mathbf{a}_0) = \sum_{i=1}^n 2 \mathbf{x}_i \mathbf{x}_i^\top$$

■ Krok metodą Newtona-Raphsona:

$$\mathbf{a}_1 = \mathbf{a}_0 - \mathbf{H}_L^{-1}(\mathbf{a}_0) \nabla_L(\mathbf{a}_0) = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right)$$

# MNK jako metoda Newtona-Raphsona

- Gradient: 
$$\nabla_L(\mathbf{a}) = -2 \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i) \mathbf{x}_i$$

- Hesjan: 
$$\mathbf{H}_L(\mathbf{a}) = \sum_{i=1}^n 2 \mathbf{x}_i \mathbf{x}_i^\top$$

- Rozwiązanie początkowe:  $\mathbf{a}_0 = \mathbf{0}$ .

$$\nabla_L(\mathbf{a}_0) = -2 \sum_{i=1}^n y_i \mathbf{x}_i, \quad \mathbf{H}_L(\mathbf{a}_0) = \sum_{i=1}^n 2 \mathbf{x}_i \mathbf{x}_i^\top$$

- Krok metodą Newtona-Raphsona:

$$\mathbf{a}_1 = \mathbf{a}_0 - \mathbf{H}_L^{-1}(\mathbf{a}_0) \nabla_L(\mathbf{a}_0) = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right)$$

- Newton-Rapshon rozwiązuje MNK w jednym roku!

- Co jeśli hesjan  $\mathbf{H}_L$  jest osobliwy (nieodwracalny)?



- Co jeśli hesjan  $\mathbf{H}_L$  jest osobliwy (nieodwracalny)?
- Dodajemy do hesjanu macierz jednostkową  $\mathbf{I}$  przemnożoną przez (małą) stałą  $\lambda$ :

$$\mathbf{a} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\top} + \lambda \mathbf{I} \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right).$$

- Co jeśli hesjan  $\mathbf{H}_L$  jest osobliwy (nieodwracalny)?
- Dodajemy do hesjanu macierz jednostkową  $\mathbf{I}$  przemnożoną przez (małą) stałą  $\lambda$ :

$$\mathbf{a} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\top} + \lambda \mathbf{I} \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right).$$

- Czy istnieje modyfikacja problemu regresji, które rozwiązaniem jest jeden krok Levenberga-Marquada?

- Co jeśli hesjan  $\mathbf{H}_L$  jest osobliwy (nieodwracalny)?
- Dodajemy do hesjanu macierz jednostkową  $\mathbf{I}$  przemnożoną przez (małą) stałą  $\lambda$ :

$$\mathbf{a} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^{\top} + \lambda \mathbf{I} \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right).$$

- Czy istnieje modyfikacja problemu regresji, które rozwiązaniem jest jeden krok Levenberga-Marquada?  
 $\implies$  Regresja grzbietowa.

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2 + \lambda \|\mathbf{a}\|^2.$$

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2 + \lambda \|\mathbf{a}\|^2.$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij} + 2\lambda a_j$$

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2 + \lambda \|\mathbf{a}\|^2.$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij} + 2\lambda a_j$$

- Przyporównanie pochodnych do zera daje:

$$\sum_{i=1}^n y_i x_{ij} = \sum_{i=1}^n \sum_{k=1}^m a_k x_{ik} x_{ij} + \lambda a_j$$

- Funkcja celu:

$$L(\mathbf{a}) = \sum_{i=1}^n (y_i - \mathbf{a}^\top \mathbf{x}_i)^2 + \lambda \|\mathbf{a}\|^2.$$

- Pochodne:

$$\frac{\partial L(\mathbf{a})}{\partial a_j} = \sum_{i=1}^n -2(y_i - \mathbf{a}^\top \mathbf{x}_i)x_{ij} + 2\lambda a_j$$

- Przyporównanie pochodnych do zera daje:

$$\sum_{i=1}^n y_i x_{ij} = \sum_{i=1}^n \sum_{k=1}^m a_k x_{ik} x_{ij} + \lambda a_j$$

- Wektorowo:

$$\begin{aligned} \sum_{i=1}^n y_i \mathbf{x}_i &= \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{a} + \lambda \mathbf{a} = \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top + \lambda \mathbf{I} \right) \mathbf{a} \\ \implies \mathbf{a} &= \left( \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i^\top + \lambda \mathbf{I} \right)^{-1} \left( \sum_{i=1}^n y_i \mathbf{x}_i \right). \end{aligned}$$

- 1 Problem regresji
- 2 Metody regresji liniowej
- 3 Minimalizacja kwadratów błędów
- 4 Minimalizacja sumy wartości bezwzględnych błędów
- 5 Przykład: wycena domów



Minimalizacja sumy wartości bezwzględnych błędów (LAD):

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n |y_i - \mathbf{a}^\top \mathbf{x}_i|.$$

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n |y_i - \mathbf{a}^\top \mathbf{x}_i|.$$

Sprowadzamy do problemu programowania liniowego:

- Dla każdego  $i = 1, \dots, n$  wprowadzamy dwie zmienne  $\sigma_i^+, \sigma_i^- \geq 0$  takie, że

$$y_i - \mathbf{a}^\top \mathbf{x}_i = \sigma_i^+ - \sigma_i^-.$$

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n |y_i - \mathbf{a}^\top \mathbf{x}_i|.$$

Sprowadzamy do problemu programowania liniowego:

- Dla każdego  $i = 1, \dots, n$  wprowadzamy dwie zmienne  $\sigma_i^+, \sigma_i^- \geq 0$  takie, że

$$y_i - \mathbf{a}^\top \mathbf{x}_i = \sigma_i^+ - \sigma_i^-.$$

- Zauważmy, że wtedy:

$$|y_i - \mathbf{a}^\top \mathbf{x}_i| \leq \sigma_i^+ + \sigma_i^-,$$

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n |y_i - \mathbf{a}^\top \mathbf{x}_i|.$$

Sprowadzamy do problemu programowania liniowego:

- Dla każdego  $i = 1, \dots, n$  wprowadzamy dwie zmienne  $\sigma_i^+, \sigma_i^- \geq 0$  takie, że

$$y_i - \mathbf{a}^\top \mathbf{x}_i = \sigma_i^+ - \sigma_i^-.$$

- Zauważmy, że wtedy:

$$|y_i - \mathbf{a}^\top \mathbf{x}_i| \leq \sigma_i^+ + \sigma_i^-,$$

czyli

$$L(\mathbf{a}) \leq \sum_{i=1}^n \sigma_i^+ + \sigma_i^-,$$

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

$$\min_{\mathbf{a}} : L(\mathbf{a}) = \sum_{i=1}^n |y_i - \mathbf{a}^\top \mathbf{x}_i|.$$

Sprowadzamy do problemu programowania liniowego:

- Dla każdego  $i = 1, \dots, n$  wprowadzamy dwie zmienne  $\sigma_i^+, \sigma_i^- \geq 0$  takie, że

$$y_i - \mathbf{a}^\top \mathbf{x}_i = \sigma_i^+ - \sigma_i^-.$$

- Zauważmy, że wtedy:

$$|y_i - \mathbf{a}^\top \mathbf{x}_i| \leq \sigma_i^+ + \sigma_i^-,$$

czyli

$$L(\mathbf{a}) \leq \sum_{i=1}^n \sigma_i^+ + \sigma_i^-,$$

i równość zachodzi dokładnie gdy dla każdego  $i = 1, \dots, n$ , jedno z  $\sigma_i^+, \sigma_i^-$  jest równe 0.

Rozwiązujemy problem:

$$\min \quad L'(\mathbf{a}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^n \sigma_i^+ + \sigma_i^-$$

$$\begin{aligned} \text{p.o.} \quad y_i - \mathbf{a}^\top \mathbf{x}_i &= \sigma_i^+ - \sigma_i^- & i = 1, \dots, n \\ \sigma_i^+, \sigma_i^- &\geq 0 & i = 1, \dots, n. \end{aligned}$$

Rozwiązujemy problem:

$$\min \quad L'(\mathbf{a}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^n \sigma_i^+ + \sigma_i^-$$

$$\begin{aligned} \text{p.o.} \quad y_i - \mathbf{a}^\top \mathbf{x}_i &= \sigma_i^+ - \sigma_i^- & i = 1, \dots, n \\ \sigma_i^+, \sigma_i^- &\geq 0 & i = 1, \dots, n. \end{aligned}$$

- Minimalizujemy górne ograniczenie funkcji  $L(\mathbf{a})$ .
- Wiemy, że w optimum dokładnie jedno z  $\sigma_i^+, \sigma_i^-$  jest równe 0.

Rozwiązujemy problem:

$$\min \quad L'(\mathbf{a}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^n \sigma_i^+ + \sigma_i^-$$

$$\begin{aligned} \text{p.o.} \quad y_i - \mathbf{a}^\top \mathbf{x}_i &= \sigma_i^+ - \sigma_i^- & i = 1, \dots, n \\ \sigma_i^+, \sigma_i^- &\geq 0 & i = 1, \dots, n. \end{aligned}$$

- Minimalizujemy górne ograniczenie funkcji  $L(\mathbf{a})$ .
- Wiemy, że w optimum dokładnie jedno z  $\sigma_i^+, \sigma_i^-$  jest równe 0.
- **Dowód:** jeśli oba  $\sigma_i^+, \sigma_i^- > 0$ , to możemy oba zmniejszyć o  $\delta$ , zachowując ograniczenia, a zmniejszając funkcję celu o  $2\delta$  – sprzeczność!



Rozwiązujemy problem:

$$\min \quad L'(\mathbf{a}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^n \sigma_i^+ + \sigma_i^-$$

$$\begin{aligned} \text{p.o.} \quad y_i - \mathbf{a}^\top \mathbf{x}_i &= \sigma_i^+ - \sigma_i^- & i = 1, \dots, n \\ \sigma_i^+, \sigma_i^- &\geq 0 & i = 1, \dots, n. \end{aligned}$$

- Minimalizujemy górne ograniczenie funkcji  $L(\mathbf{a})$ .
- Wiemy, że w optimum dokładnie jedno z  $\sigma_i^+, \sigma_i^-$  jest równe 0.
- **Dowód:** jeśli oba  $\sigma_i^+, \sigma_i^- > 0$ , to możemy oba zmniejszyć o  $\delta$ , zachowując ograniczenia, a zmniejszając funkcję celu o  $2\delta$  – sprzeczność!
- **Wniosek:** W optimum

$$L'(\mathbf{a}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = L(\mathbf{a}),$$

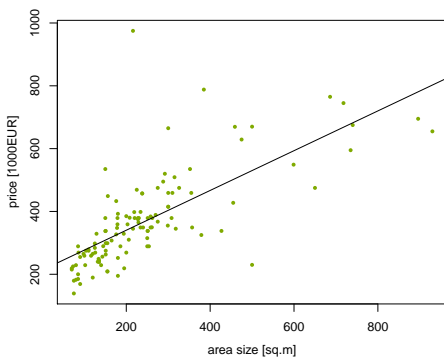
więc rozwiązaliśmy problem LAD.

- 1 Problem regresji
- 2 Metody regresji liniowej
- 3 Minimalizacja kwadratów błędów
- 4 Minimalizacja sumy wartości bezwzględnych błędów
- 5 Przykład: wycena domów

# Wycena domów

- Den Bosch ('s-Hertogenbosch), Holandia
- 119 domów.

	$X$	$Y$
	living area	sale price
$x_1$	385	788
$x_2$	156	449
$x_3$	90	169
$x_4$	86	269
$x_5$	73	225
$x_6$	125	298
...	...	



- 119 domów, opisanych 9 cechami wejściowymi ( $X$ ) i 1 wyjściową ( $Y$ ).

---

DISTR	type of district, four categories ranked from bad (1) to good (4)
AREA	total area including garden
BEDR	number of bedrooms
TYPE	apartment (1), row house (2), corner house (3), semidetached (4), detached (5), villa (6)
VOL	volume of the house
STOR	number of storeys
GARD	type of garden, four categories ranked from bad to good
GARG	no garage (1), normal garage (2), large garage (3)
YEAR	build year
PRICE	selling price

---

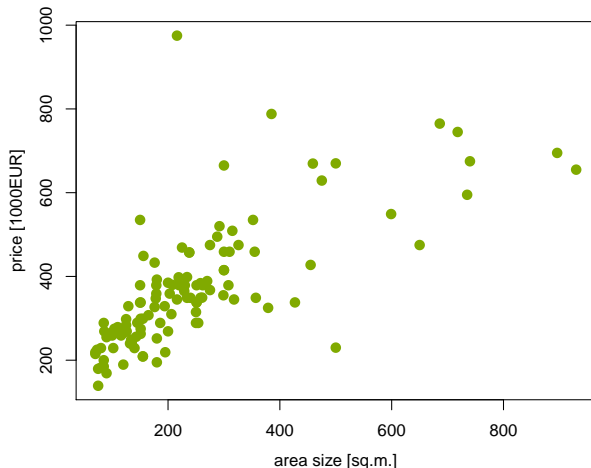
## Zbiór danych:

	$X_1$	$X_2$	$X_3$	$X_4$	$X_5$	$X_6$	$X_7$	$X_8$	$X_9$	$Y$
	DISTR	AREA	BEDR	TYPE	VOL	STOR	GARD	GARG	YEAR	PRICE
$x_1$	4	385	5	6	775	3	3	3	1934	788
$x_2$	3	156	2	1	350	1	1	1	1996	449
$x_3$	4	90	3	1	200	1	1	1	1950	169
$x_4$	3	86	3	2	410	3	2	1	1966	269
$x_5$	1	73	3	2	330	3	3	1	1950	225
$x_6$	3	125	2	1	300	2	1	2	1950	298
...	...	...	...	...	...	...	...	...	...	...

# Ilustracyjny przykład

Weźmy tylko jedną zmienną  $X = X_2$  (AREA) dla zilustrowania wyników na płaszczyźnie.

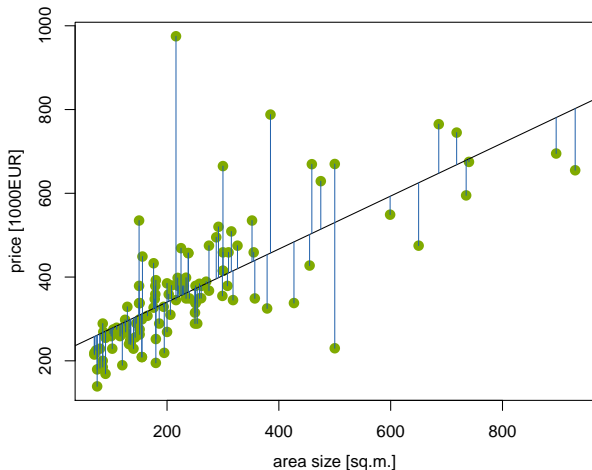
Wykres  $Y = \text{PRICE}$  w funkcji  $X = \text{AREA}$ :



# Ilustracyjny przykład

Weźmy tylko jedną zmienną  $X = X_2$  (AREA) dla zilustrowania wyników na płaszczyźnie.

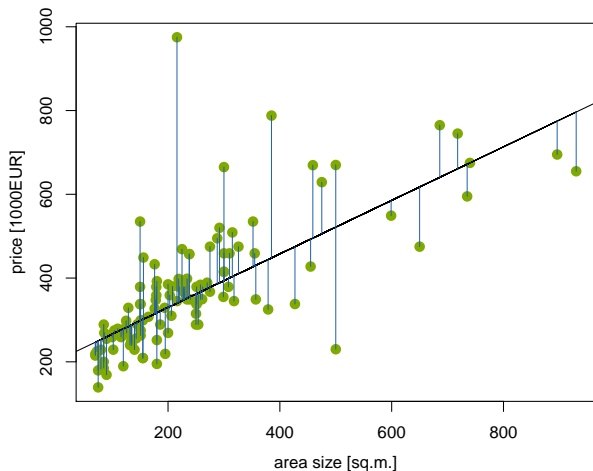
Metoda najmniejszych kwadratów (LS):



# Ilustracyjny przykład

Weźmy tylko jedną zmienną  $X = X_2$  (AREA) dla zilustrowania wyników na płaszczyźnie.

Minimalizacja sumy wartości bezwzględnych (LAD):

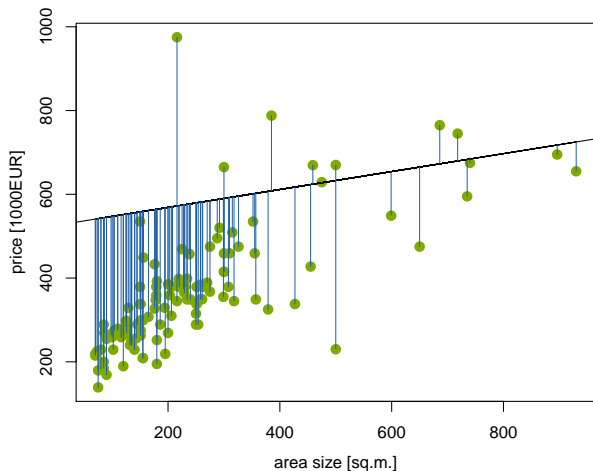




# Ilustracyjny przykład

Weźmy tylko jedną zmienną  $X = X_2$  (AREA) dla zilustrowania wyników na płaszczyźnie.

Minimalizacja największego błędu (MM):



Koniec na dzisiaj :)