Techniki Optymalizacji: Metody regresji

Wojciech Kotłowski

Instytut Informatyki Politechniki Poznańskiej
email: imię.nazwisko@cs.put.poznan.pl

pok. 2 (CW) tel. (61)665-2936 konsultacje: piątek 15:10-16:40
Slajdy dostępne pod adresem: `http://www.cs.put.poznan.pl/wkotlowski/to/`

6.11.2018

Spis treści

Spis treści

Problem regresji

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

Problem regresji

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

Przykłady

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

Regresja liniowa

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

$$\hat{Y} = w_1 X + w_0$$

  - dla wielu zmiennych:

$$\hat{Y} = w_1 X_1 + \ldots + w_m X_m + w_0 = \boldsymbol{w}^\top \boldsymbol{X} + w_0$$

Regresja liniowa

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

$$\hat{Y} = w_1 X + w_0$$

  - dla wielu zmiennych:

$$\hat{Y} = w_1 X_1 + \ldots + w_m X_m + w_0 = \boldsymbol{w}^\top \boldsymbol{X} + w_0$$

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

$$\hat{Y} = w_1 X + w_2 X^2 + \ldots + w_0 \quad \Longrightarrow \quad \hat{Y} = w_1 X_1 + w_2 X_2 + \ldots + w_0$$

Jak to się zwykle robi

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

$$(x_{11}, x_{12}, \ldots, x_{1m}, y_1) \qquad (\boldsymbol{x}_1, y_1)$$
$$(x_{21}, x_{22}, \ldots, x_{2m}, y_2) \qquad (\boldsymbol{x}_2, y_2)$$
$$\ldots \qquad \text{lub w skrócie} \qquad \ldots$$
$$(x_{n1}, x_{n2}, \ldots, x_{nm}, y_n) \qquad (\boldsymbol{x}_n, y_n)$$

Jak to się zwykle robi

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

$$(x_{11}, x_{12}, \ldots, x_{1m}, y_1) \qquad (\boldsymbol{x}_1, y_1)$$
$$(x_{21}, x_{22}, \ldots, x_{2m}, y_2) \qquad (\boldsymbol{x}_2, y_2)$$
$$\ldots \qquad \text{lub w skrócie} \qquad \ldots$$
$$(x_{n1}, x_{n2}, \ldots, x_{nm}, y_n) \qquad (\boldsymbol{x}_n, y_n)$$

- Wyznaczamy współczynniki $w_0, w_1, \ldots, w_m$ na danych.

Jak to się zwykle robi

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

$$(x_{11}, x_{12}, \ldots, x_{1m}, y_1) \qquad (\boldsymbol{x}_1, y_1)$$
$$(x_{21}, x_{22}, \ldots, x_{2m}, y_2) \qquad (\boldsymbol{x}_2, y_2)$$
$$\ldots \qquad \text{lub w skrócie} \qquad \ldots$$
$$(x_{n1}, x_{n2}, \ldots, x_{nm}, y_n) \qquad (\boldsymbol{x}_n, y_n)$$

- Wyznaczamy współczynniki $w_0, w_1, \ldots, w_m$ 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:

$$(x_{11}, x_{12}, \ldots, x_{1m}, y_1) \qquad (\boldsymbol{x}_1, y_1)$$
$$(x_{21}, x_{22}, \ldots, x_{2m}, y_2) \qquad (\boldsymbol{x}_2, y_2)$$
$$\ldots \qquad \text{lub w skrócie} \qquad \ldots$$
$$(x_{n1}, x_{n2}, \ldots, x_{nm}, y_n) \qquad (\boldsymbol{x}_n, y_n)$$

- Wyznaczamy współczynniki $w_0, w_1, \ldots, w_m$ na danych.
- Testujemy nasz model na osobnym zbiorze testowym (również ze znanymi wartościami Y)

Przykład: szacowanie czasu pracy programistów

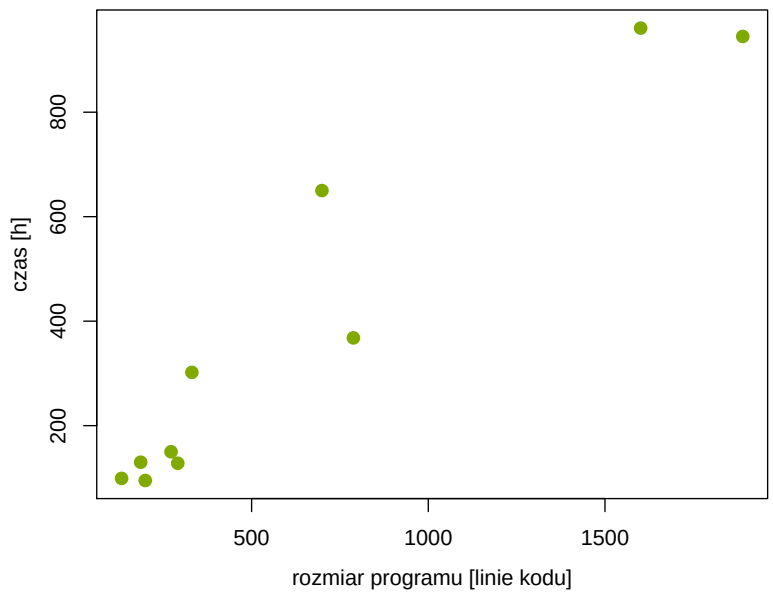
X	Y
Rozmiar programu	Oszacowany czas
186	130
699	650
132	99
272	150
291	128
331	302
199	95
1890	945
788	368
1601	961

Przykład: szacowanie czasu pracy programistów

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

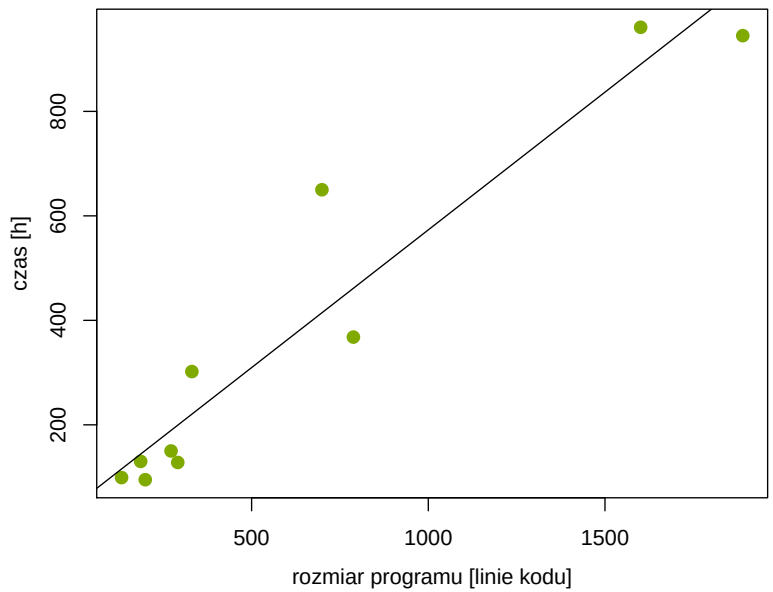
Przykład: szacowanie czasu pracy programistów

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



$$w_0 = 45.93, w_1 = 0.5273$$

Spis treści

Jak wyznaczyć współczynniki?

Problem optymalizacji

Mając zbiór danych $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, wyznacz współczynniki $w_0, w_1, \ldots, w_m$ tak, aby wartości przewidywane przez model:

$$\hat{y}_i = w_1 x_{i1} + \ldots + w_m x_{im} + w_0 = \boldsymbol{w}^\top \boldsymbol{x}_i + w_0$$

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

Jak wyznaczyć współczynniki?

Problem optymalizacji

Mając zbiór danych $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$, wyznacz współczynniki $w_0, w_1, \ldots, w_m$ tak, aby wartości przewidywane przez model:

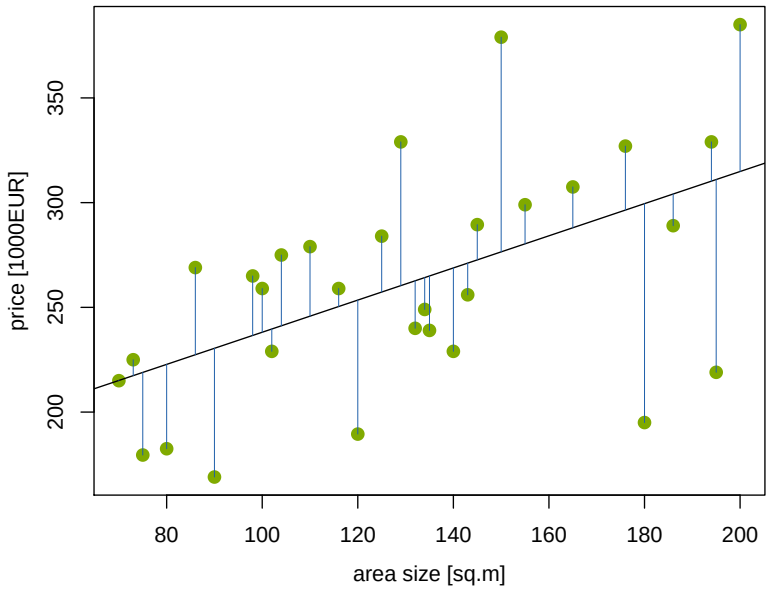
$$\hat{y}_i = w_1 x_{i1} + \ldots + w_m x_{im} + w_0 = \boldsymbol{w}^\top \boldsymbol{x}_i + w_0$$

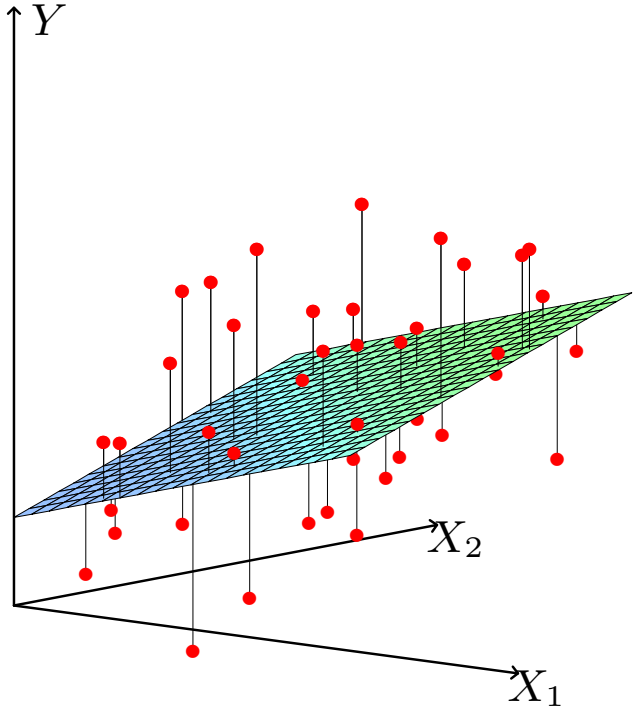
na wszystkich danych $(i = 1, \ldots, n)$ były jak najbliżej prawdziwych wartości y_i.

- Uwaga: zwykle dodajemy jeszcze jedną zmienną wejściową X_0 stale równą 1 i chowamy współczynnik w_0 do wektora $\boldsymbol{w}$, otrzymując:

$$\hat{y}_i = w_0 x_{i0} + w_1 x_{i1} + \ldots + w_m x_{im} = \boldsymbol{w}^\top \boldsymbol{x}_i$$

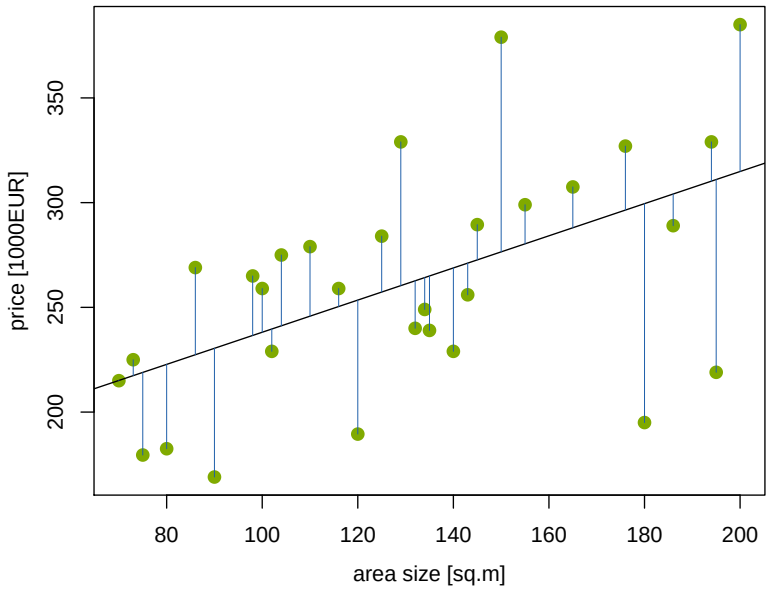
Odchylenia (błędy) na danych

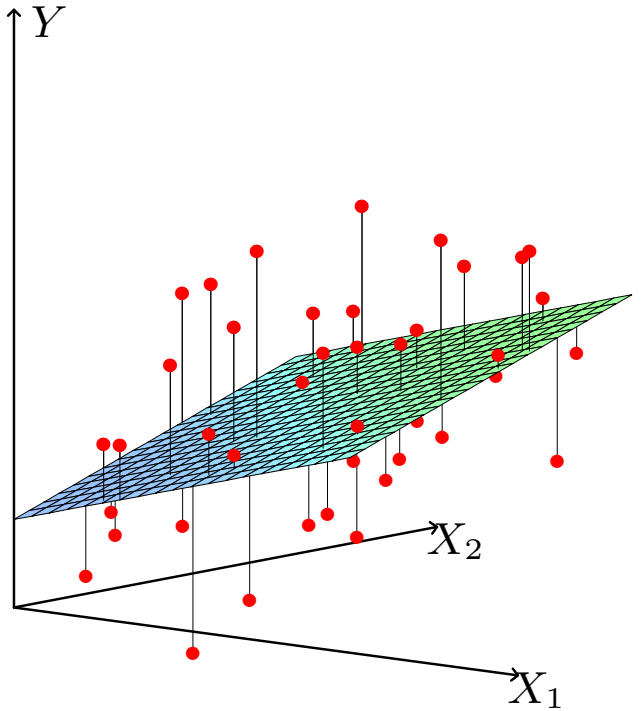




Źródło: Hastie, Tibshirani, Friedman, *Elements of statistical learning*.

Odchylenia (błędy) na danych



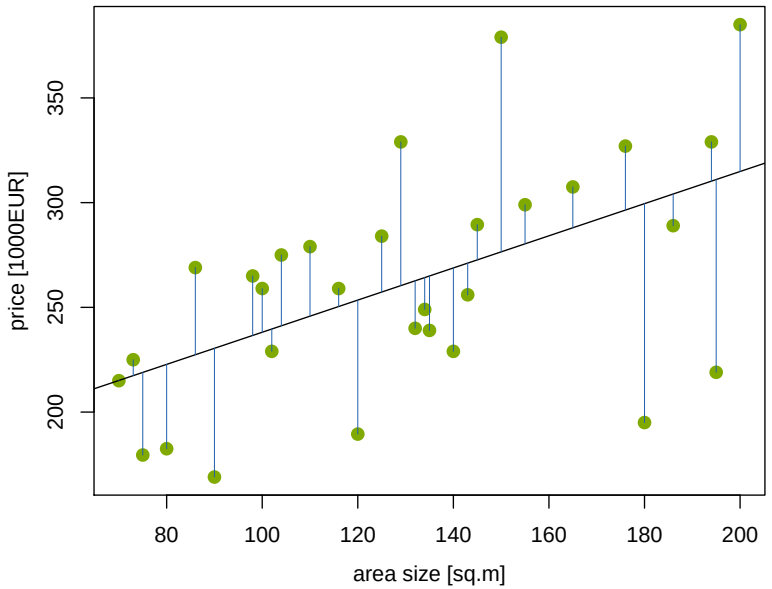


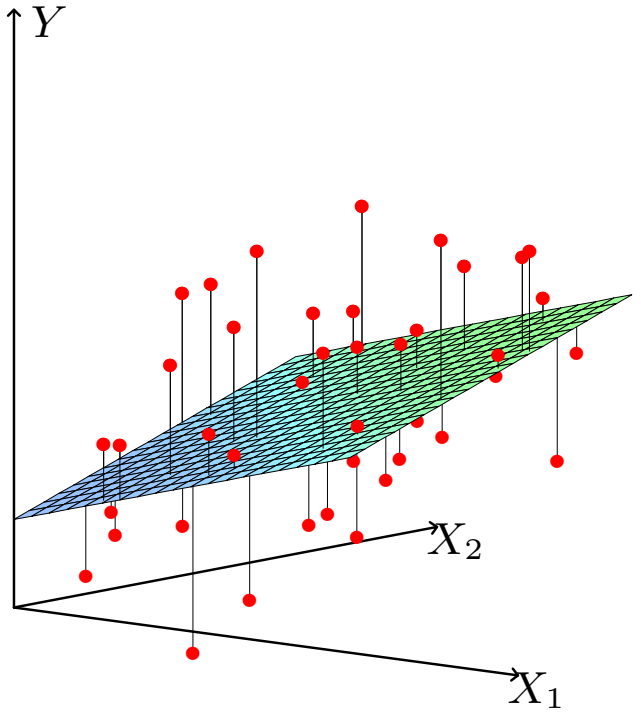
Źródło: Hastie, Tibshirani, Friedman, *Elements of statistical learning*.

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

$$\delta_i = y_i - \hat{y}_i = y_i - \boldsymbol{w}^\top \boldsymbol{x}_i$$

Odchylenia (błędy) na danych





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

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

$$\delta_i = y_i - \hat{y}_i = y_i - \boldsymbol{w}^\top \boldsymbol{x}_i$$

- Jak zmierzyć sumaryczny błąd?

Trzy kryteria

Trzy kryteria

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

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

Trzy kryteria

- Minimalizacja sumy kwadratów błędów/odchyleń (*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/odchyleń (*least absolute deviations* – LAD)

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

Trzy kryteria

- Minimalizacja sumy kwadratów błędów/odchyleń (*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/odchyleń (*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)

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

Zalety i wady

	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 wyjatkowymi przypadkami).

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

- Minimalizacja kwadratów błędów (LS):
  $\min_{w_0} \left\{ 99 \cdot (w_0 - 1)^2 + (w_0 - 1000)^2 \right\}$

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

- Minimalizacja kwadratów błędów (LS):

$$\min_{w_0} \left\{ 99 \cdot (w_0 - 1)^2 + (w_0 - 1000)^2 \right\} \quad \Longrightarrow w_0 = \text{avg}(\boldsymbol{y}) = 10.99$$

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

- Minimalizacja kwadratów błędów (LS):
  $\min_{w_0} \left\{ 99 \cdot (w_0 - 1)^2 + (w_0 - 1000)^2 \right\} \quad \Longrightarrow w_0 = \text{avg}(\boldsymbol{y}) = 10.99$
- Minimalizacja wartości bezwzględnych (błędów) (LAD):
  $\min_{w_0} \left\{ 99 \cdot |w_0 - 1| + |w_0 - 1000| \right\}$

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

- Minimalizacja kwadratów błędów (LS):
  $\min_{w_0} \left\{ 99 \cdot (w_0 - 1)^2 + (w_0 - 1000)^2 \right\} \quad \Longrightarrow w_0 = \text{avg}(\boldsymbol{y}) = 10.99$
- Minimalizacja wartości bezwzględnych (błędów) (LAD):
  $\min_{w_0} \left\{ 99 \cdot |w_0 - 1| + |w_0 - 1000| \right\} \quad \Longrightarrow w_0 = \text{median}(\boldsymbol{y}) = 1$

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

- Minimalizacja kwadratów błędów (LS):
  $\min_{w_0} \left\{ 99 \cdot (w_0 - 1)^2 + (w_0 - 1000)^2 \right\} \quad \Longrightarrow w_0 = \text{avg}(\boldsymbol{y}) = 10.99$
- Minimalizacja wartości bezwzględnych (błędów) (LAD):
  $\min_{w_0} \left\{ 99 \cdot |w_0 - 1| + |w_0 - 1000| \right\} \quad \Longrightarrow w_0 = \text{median}(\boldsymbol{y}) = 1$
- Minimalizacja największego błędu (MM):
  $\min_{w_0} \max \left\{ |w_0 - 1|, |w_0 - 1000| \right\}$

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

- Minimalizacja kwadratów błędów (LS):
  $\min_{w_0} \left\{ 99 \cdot (w_0 - 1)^2 + (w_0 - 1000)^2 \right\} \quad \Longrightarrow w_0 = \text{avg}(\boldsymbol{y}) = 10.99$
- Minimalizacja wartości bezwzględnych (błędów) (LAD):
  $\min_{w_0} \left\{ 99 \cdot |w_0 - 1| + |w_0 - 1000| \right\} \quad \Longrightarrow w_0 = \text{median}(\boldsymbol{y}) = 1$
- Minimalizacja największego błędu (MM):
  $\min_{w_0} \max \left\{ |w_0 - 1|, |w_0 - 1000| \right\} \quad \Longrightarrow w_0 = \text{middle}(\boldsymbol{y}) = 500.5$

Odporność na wartości odstające

Najprostszy przypadek: Brak $\boldsymbol{X}$, tylko Y.
Model zawiera tylko stałą:

$$\hat{y} = w_0.$$

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

- Minimalizacja kwadratów błędów (LS):
  $\min_{w_0} \left\{ 99 \cdot (w_0 - 1)^2 + (w_0 - 1000)^2 \right\} \quad \Longrightarrow w_0 = \text{avg}(\boldsymbol{y}) = 10.99$
- Minimalizacja wartości bezwzględnych (błędów) (LAD):
  $\min_{w_0} \left\{ 99 \cdot |w_0 - 1| + |w_0 - 1000| \right\} \quad \Longrightarrow w_0 = \text{median}(\boldsymbol{y}) = 1$
- Minimalizacja największego błędu (MM):
  $\min_{w_0} \max \left\{ |w_0 - 1|, |w_0 - 1000| \right\} \quad \Longrightarrow w_0 = \text{middle}(\boldsymbol{y}) = 500.5$

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

Spis treści

Optymalizacja

„Metoda najmniejszych kwadratów”

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

Optymalizacja

„Metoda najmniejszych kwadratów”

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

- Funkcja wypukła, kwadratowa.

Optymalizacja

„Metoda najmniejszych kwadratów”

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

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

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

Metoda najmniejszych kwadratów

Metoda najmniejszych kwadratów

- Funkcja celu:

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

Metoda najmniejszych kwadratów

- Funkcja celu:

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

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij}$$

Metoda najmniejszych kwadratów

- Funkcja celu:

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

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij}$$

- Przyrównanie do zera:

$$\sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \sum_{k=1}^{m} w_k x_{ik} x_{ij}$$

Metoda najmniejszych kwadratów

- Funkcja celu:

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

- Pochodne:

$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij}$$

- Przyrównanie do zera:

$$\sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij} = 0 \quad \Longrightarrow \quad \sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \sum_{k=1}^{m} w_k x_{ik} x_{ij}$$

- Wektorowo:

$$\sum_{i=1}^{n} y_i \boldsymbol{x}_i = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \boldsymbol{w} \quad \Longrightarrow \quad \boldsymbol{w} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \right).$$

MNK jako metoda Newtona-Raphsona

MNK jako metoda Newtona-Raphsona

- Funkcja celu: $$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$

MNK jako metoda Newtona-Raphsona

- Funkcja celu: $$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$
- Pochodne: $$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij}$$

MNK jako metoda Newtona-Raphsona

- Funkcja celu: $$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$
- Pochodne: $$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij}$$
- Gradient: $$\nabla_L(\boldsymbol{w}) = -2 \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) \boldsymbol{x}_i$$

MNK jako metoda Newtona-Raphsona

- Funkcja celu: $$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$
- Pochodne: $$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij}$$
- Gradient: $$\nabla_L(\boldsymbol{w}) = -2 \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) \boldsymbol{x}_i$$
- Drugie pochodne: $$\frac{\partial^2 L(\boldsymbol{w})}{\partial w_j w_k} = \sum_{i=1}^{n} 2 x_{ik} x_{ij}$$

MNK jako metoda Newtona-Raphsona

- Funkcja celu: $$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2$$
- Pochodne: $$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij}$$
- Gradient: $$\nabla_L(\boldsymbol{w}) = -2 \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) \boldsymbol{x}_i$$
- Drugie pochodne: $$\frac{\partial^2 L(\boldsymbol{w})}{\partial w_j w_k} = \sum_{i=1}^{n} 2 x_{ik} x_{ij}$$
- Hesjan: $$\boldsymbol{H}_L(\boldsymbol{w}) = \sum_{i=1}^{n} 2 \boldsymbol{x}_i \boldsymbol{x}_i^\top$$

MNK jako metoda Newtona-Raphsona

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

MNK jako metoda Newtona-Raphsona

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

- Rozwiązanie początkowe: $\boldsymbol{w}_0 = 0$.

$$\nabla_L(\boldsymbol{w}_0) = -2\sum_{i=1}^{n} y_i\boldsymbol{x}_i, \qquad \boldsymbol{H}_L(\boldsymbol{w}_0) = \sum_{i=1}^{n} 2\boldsymbol{x}_i\boldsymbol{x}_i^\top$$

MNK jako metoda Newtona-Raphsona

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

- Rozwiązanie początkowe: $\boldsymbol{w}_0 = 0$.

$$\nabla_L(\boldsymbol{w}_0) = -2\sum_{i=1}^{n} y_i\boldsymbol{x}_i, \qquad \boldsymbol{H}_L(\boldsymbol{w}_0) = \sum_{i=1}^{n} 2\boldsymbol{x}_i\boldsymbol{x}_i^\top$$

- Krok metodą Newtona-Rapshona:

$$\boldsymbol{w}_1 = \boldsymbol{w}_0 - \boldsymbol{H}_L^{-1}(\boldsymbol{w}_0)\nabla_L(\boldsymbol{w}_0)$$

MNK jako metoda Newtona-Raphsona

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

- Rozwiązanie początkowe: $\boldsymbol{w}_0 = 0$.

$$\nabla_L(\boldsymbol{w}_0) = -2\sum_{i=1}^{n} y_i\boldsymbol{x}_i, \qquad \boldsymbol{H}_L(\boldsymbol{w}_0) = \sum_{i=1}^{n} 2\boldsymbol{x}_i\boldsymbol{x}_i^\top$$

- Krok metodą Newtona-Rapshona:

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

MNK jako metoda Newtona-Raphsona

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

- Rozwiązanie początkowe: $\boldsymbol{w}_0 = 0$.

$$\nabla_L(\boldsymbol{w}_0) = -2\sum_{i=1}^{n} y_i\boldsymbol{x}_i, \qquad \boldsymbol{H}_L(\boldsymbol{w}_0) = \sum_{i=1}^{n} 2\boldsymbol{x}_i\boldsymbol{x}_i^\top$$

- Krok metodą Newtona-Rapshona:

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

- Newton-Rapshon rozwiązuje MNK w jednym roku!

Metoda Levenberga-Marquada

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

Metoda Levenberga-Marquada

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

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

Metoda Levenberga-Marquada

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

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

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

Metoda Levenberga-Marquada

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

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

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

Regresja grzbietowa

- Funkcja celu:

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

Regresja grzbietowa

- Funkcja celu:

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

- Pochodne:

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

Regresja grzbietowa

- Funkcja celu:
$$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|^2.$$
- Pochodne:
$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij} + 2\lambda w_j$$
- Przyrównanie pochodnych do zera daje:
$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \sum_{k=1}^{m} w_k x_{ik} x_{ij} + \lambda w_j$$

Regresja grzbietowa

- Funkcja celu:
$$L(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^\top \boldsymbol{x}_i)^2 + \lambda \|\boldsymbol{w}\|^2.$$
- Pochodne:
$$\frac{\partial L(\boldsymbol{w})}{\partial w_j} = \sum_{i=1}^{n} -2(y_i - \boldsymbol{w}^\top \boldsymbol{x}_i) x_{ij} + 2\lambda w_j$$
- Przyrównanie pochodnych do zera daje:
$$\sum_{i=1}^{n} y_i x_{ij} = \sum_{i=1}^{n} \sum_{k=1}^{m} w_k x_{ik} x_{ij} + \lambda w_j$$
- Wektorowo:
$$\sum_{i=1}^{n} y_i \boldsymbol{x}_i = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top \right) \boldsymbol{w} + \lambda \boldsymbol{w} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \lambda \boldsymbol{I} \right) \boldsymbol{w}$$
$$\implies \boldsymbol{w} = \left(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^\top + \lambda \boldsymbol{I} \right)^{-1} \left(\sum_{i=1}^{n} y_i \boldsymbol{x}_i \right).$$

Spis treści

Optymalizacja

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

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

Optymalizacja

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

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

Sprowadzamy do problemu programowania liniowego:

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

$$y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^-.$$

Optymalizacja

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

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

Sprowadzamy do problemu programowania liniowego:

- Dla każdego $i = 1, \ldots, n$ wprowadzamy dwie zmienne $\sigma_i^+, \sigma_i^- \geq 0$ takie, że
$$y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^-.$$
- Zauważmy, że wtedy:
$$|y_i - \boldsymbol{w}^\top \boldsymbol{x}_i| \leq \sigma_i^+ + \sigma_i^-,$$

Optymalizacja

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

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

Sprowadzamy do problemu programowania liniowego:

- Dla każdego $i = 1, \ldots, n$ wprowadzamy dwie zmienne $\sigma_i^+, \sigma_i^- \geq 0$ takie, że
$$y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^-.$$

- Zauważmy, że wtedy:
$$|y_i - \boldsymbol{w}^\top \boldsymbol{x}_i| \leq \sigma_i^+ + \sigma_i^-,$$
czyli
$$L(\boldsymbol{w}) \leq \sum_{i=1}^{n} \sigma_i^+ + \sigma_i^-,$$

Optymalizacja

Minimalizacja sumy wartości bezwzględnych błędów (LAD):

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

Sprowadzamy do problemu programowania liniowego:

- Dla każdego $i = 1, \ldots, n$ wprowadzamy dwie zmienne $\sigma_i^+, \sigma_i^- \geq 0$ takie, że
$$y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^-.$$

- Zauważmy, że wtedy:
$$|y_i - \boldsymbol{w}^\top \boldsymbol{x}_i| \leq \sigma_i^+ + \sigma_i^-,$$
czyli
$$L(\boldsymbol{w}) \leq \sum_{i=1}^{n} \sigma_i^+ + \sigma_i^-,$$
i równość zachodzi dokładnie gdy dla każdego $i = 1, \ldots, n$, jedno z σ_i^+, σ_i^- jest równe 0.

Optymalizacja

Rozwiązujemy problem:

$$\begin{aligned} \min \quad & L'(\boldsymbol{w}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^{n} \sigma_i^+ + \sigma_i^- \\ \text{p.o.} \quad & y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^- && i = 1, \ldots, n \\ & \sigma_i^+, \sigma_i^- \geq 0 && i = 1, \ldots, n. \end{aligned}$$

Optymalizacja

Rozwiązujemy problem:

$$\begin{aligned} \min \quad & L'(\boldsymbol{w}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^{n} \sigma_i^+ + \sigma_i^- \\ \text{p.o.} \quad & y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^- && i = 1, \ldots, n \\ & \sigma_i^+, \sigma_i^- \geq 0 && i = 1, \ldots, n. \end{aligned}$$

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

Optymalizacja

Rozwiązujemy problem:

$$\begin{aligned} \min \quad & L'(\boldsymbol{w}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^{n} \sigma_i^+ + \sigma_i^- \\ \text{p.o.} \quad & y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^- && i = 1, \ldots, n \\ & \sigma_i^+, \sigma_i^- \geq 0 && i = 1, \ldots, n. \end{aligned}$$

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

Optymalizacja

Rozwiązujemy problem:

$$\begin{aligned} \min \quad & L'(\boldsymbol{w}, \boldsymbol{\sigma}^+, \boldsymbol{\sigma}^-) = \sum_{i=1}^{n} \sigma_i^+ + \sigma_i^- \\ \text{p.o.} \quad & y_i - \boldsymbol{w}^\top \boldsymbol{x}_i = \sigma_i^+ - \sigma_i^- && i = 1, \ldots, n \\ & \sigma_i^+, \sigma_i^- \geq 0 && i = 1, \ldots, n. \end{aligned}$$

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

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

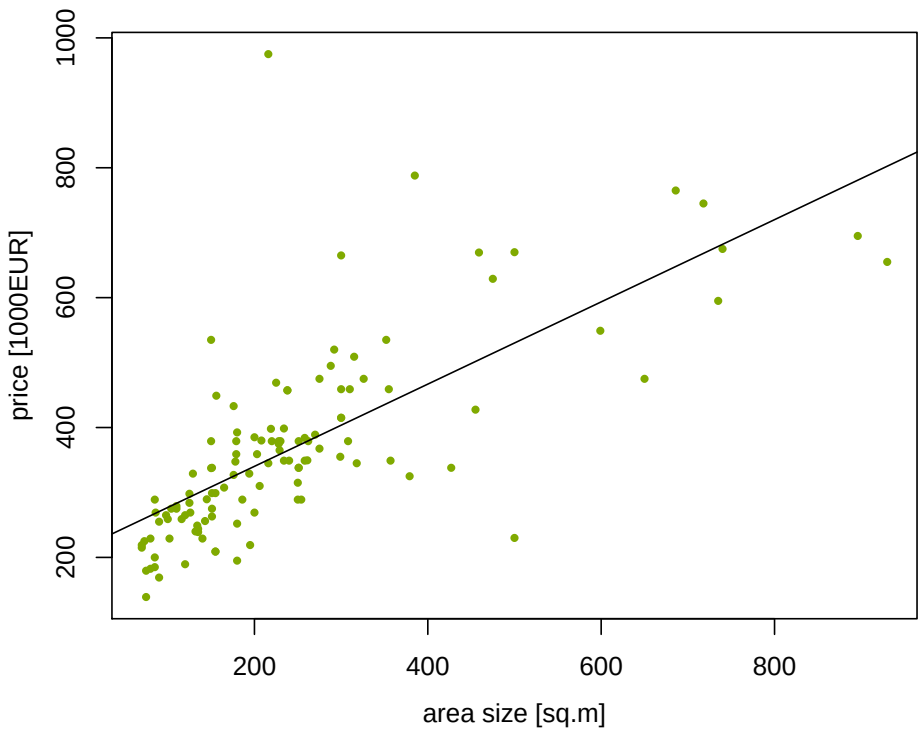
więc rozwiązaliśmy problem LAD.

Spis treści

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
...	...	



Wycena domów

- 119 domów, opisanych 9 cechami wejściowymi ($\boldsymbol{X}$) 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

Wycena domów

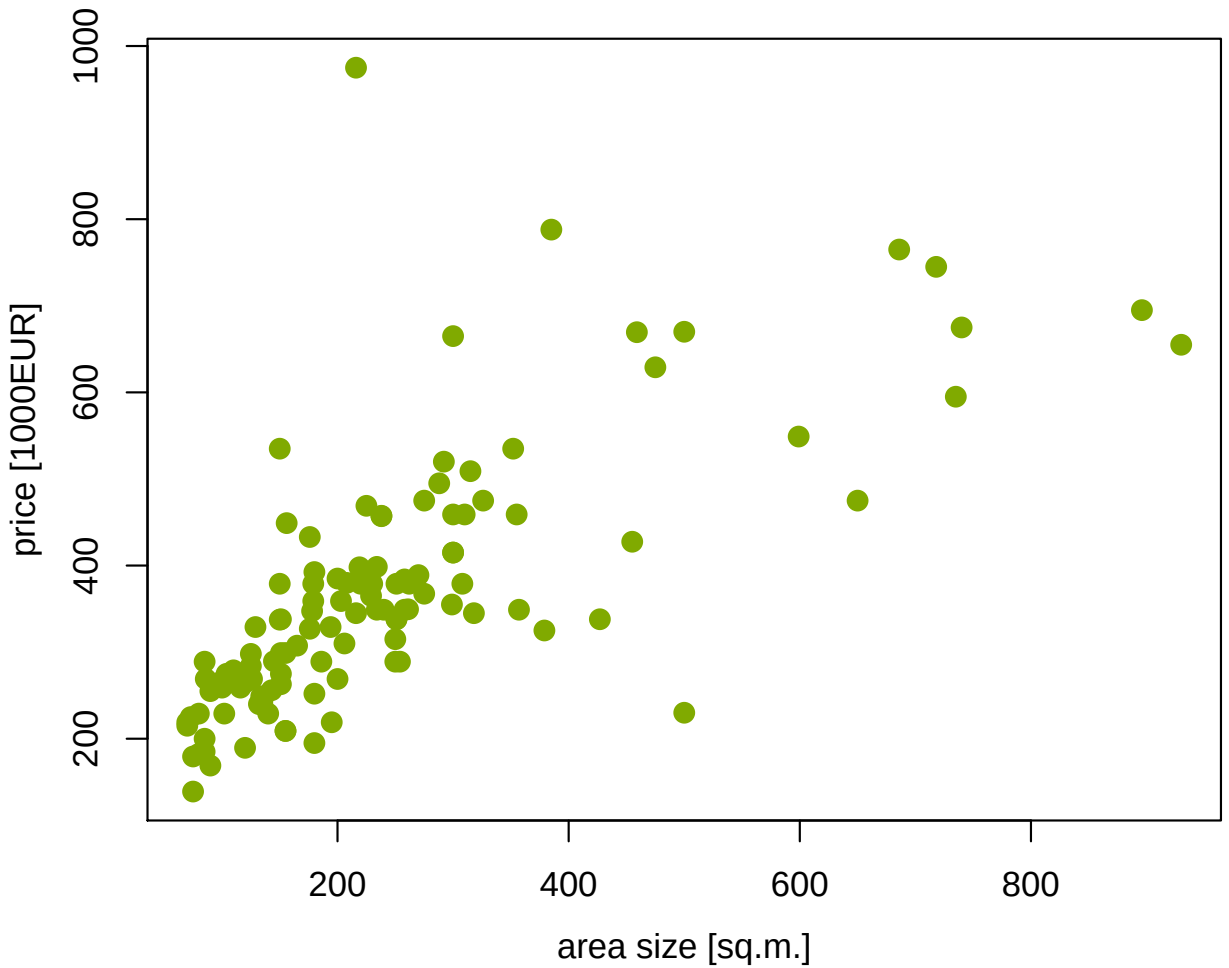
Zbiór danych:

	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
$\boldsymbol{x}_1$	4	385	5	6	775	3	3	3	1934	788
$\boldsymbol{x}_2$	3	156	2	1	350	1	1	1	1996	449
$\boldsymbol{x}_3$	4	90	3	1	200	1	1	1	1950	169
$\boldsymbol{x}_4$	3	86	3	2	410	3	2	1	1966	269
$\boldsymbol{x}_5$	1	73	3	2	330	3	3	1	1950	225
$\boldsymbol{x}_6$	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.

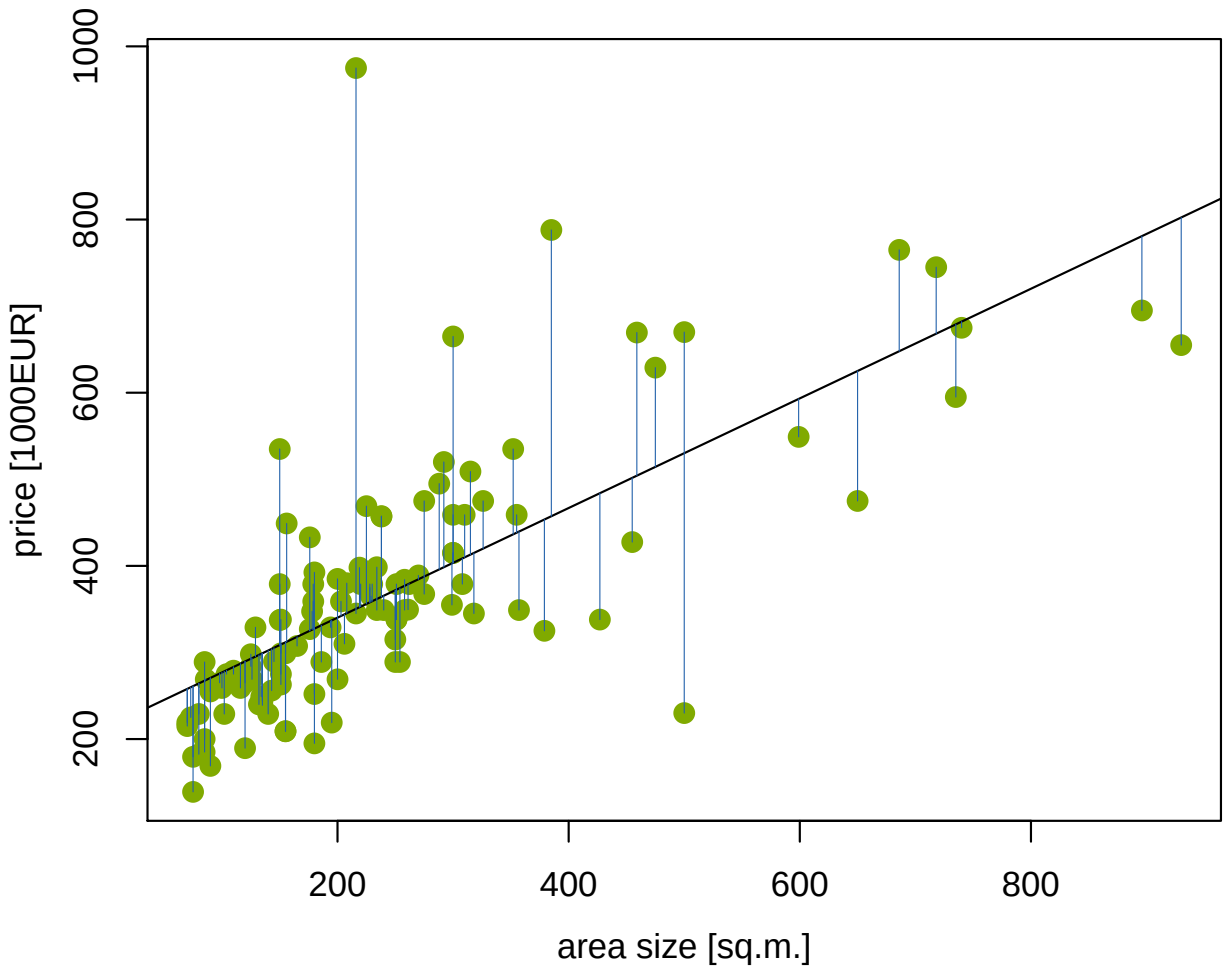
Wykres Y =PRICE w funkcji X =AREA:



Ilustracyjny przykład

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

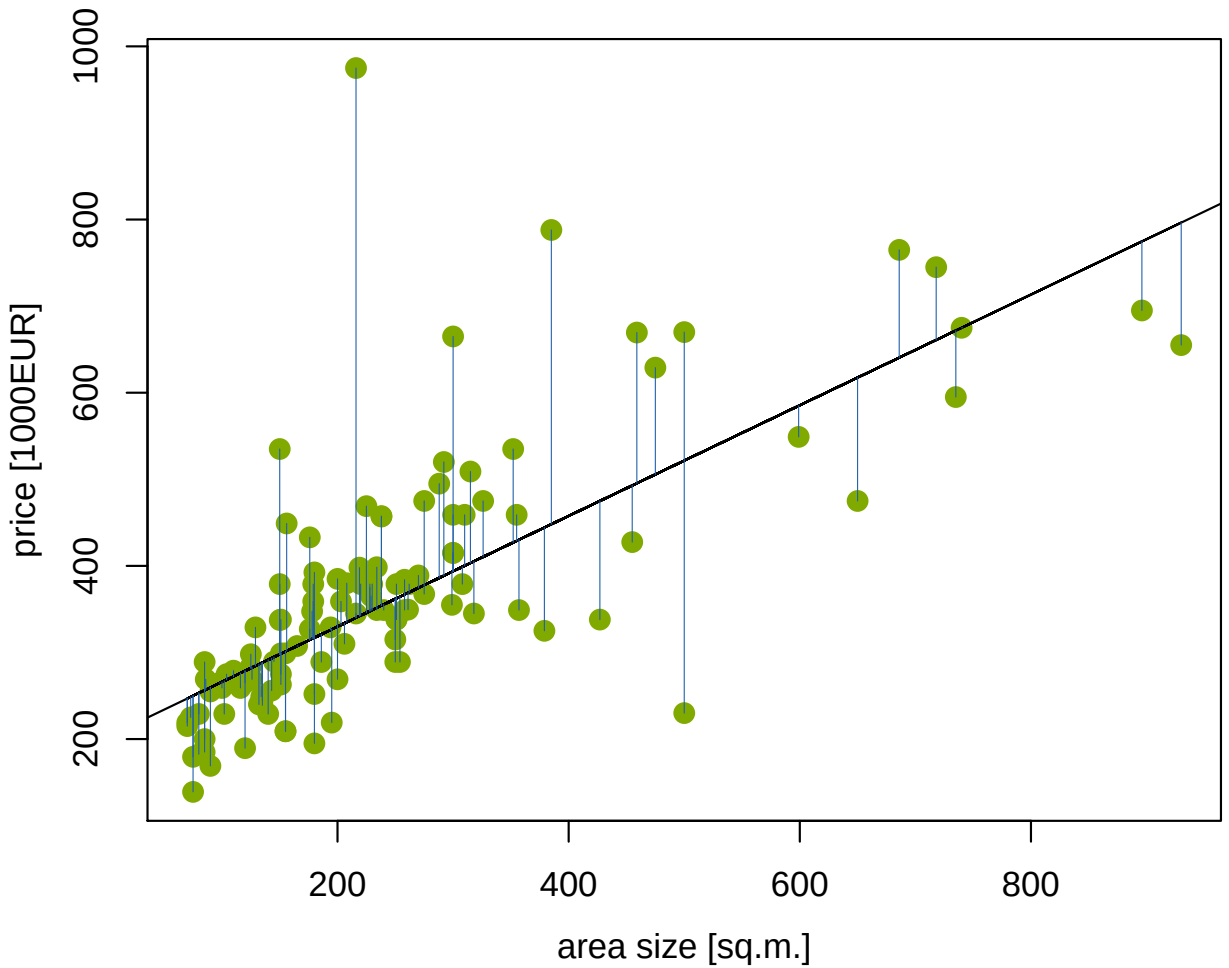
Metoda najmniejszych kwadratów (LS):



Ilustracyjny przykład

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

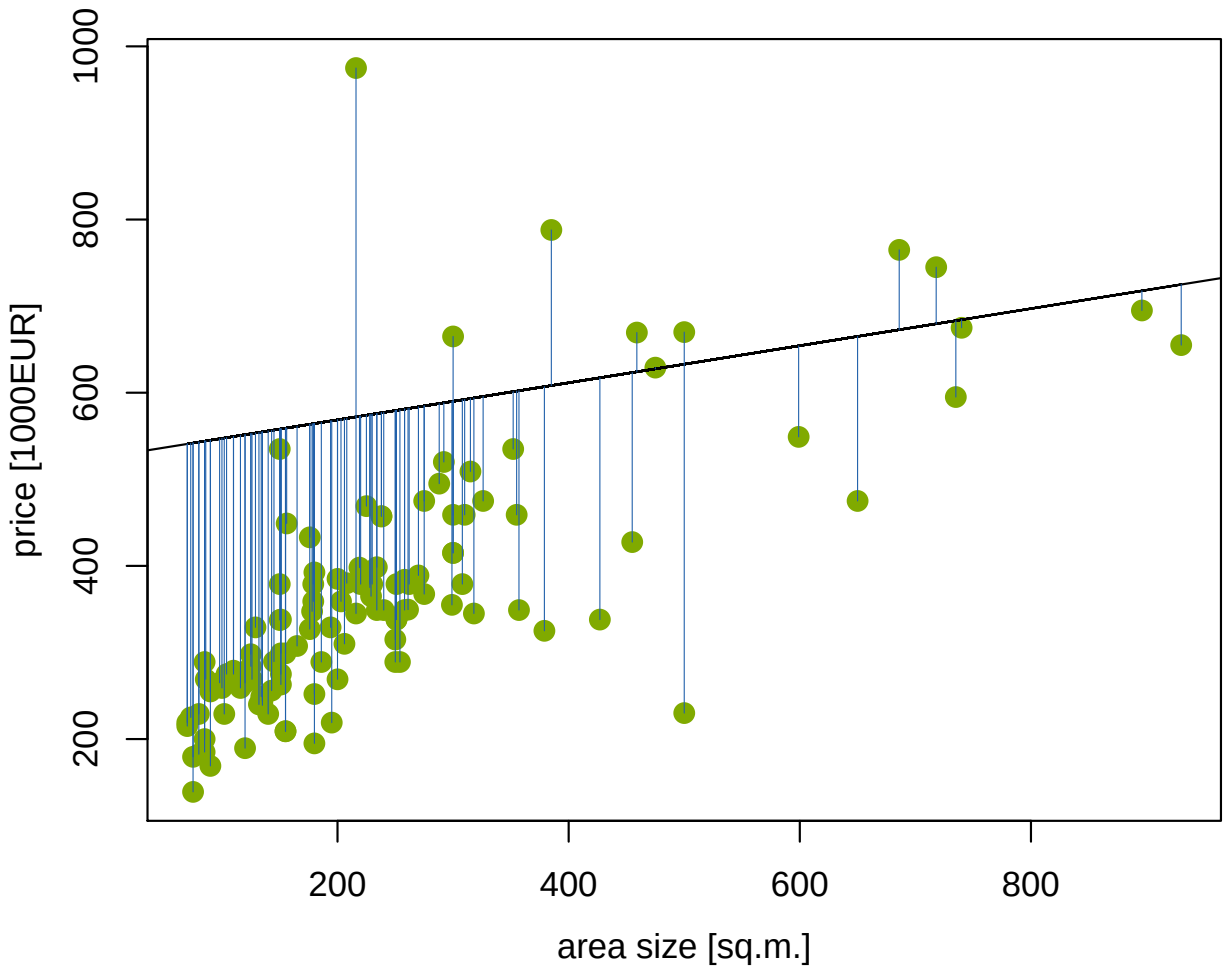
Minimalizacja sumy wartości bewzględnych (LAD):



Ilustracyjny przykład

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

Minimalizacja największego błędu (MM):



Koniec na dzisiaj :)