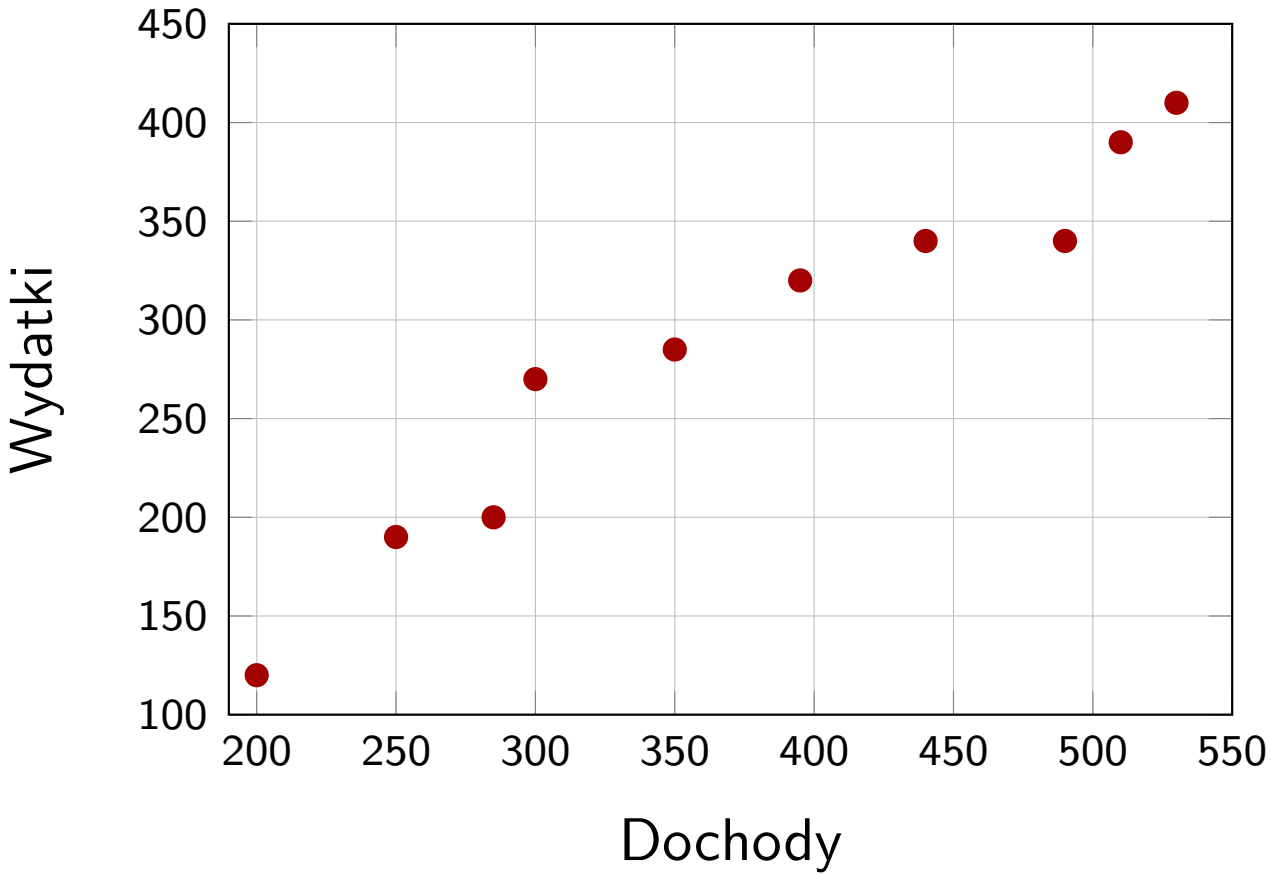
Regresja Liniowa

Zachariasz Jażdżewski, Bartosz Guzowzski

18 maja 2024

Wprowadzenie

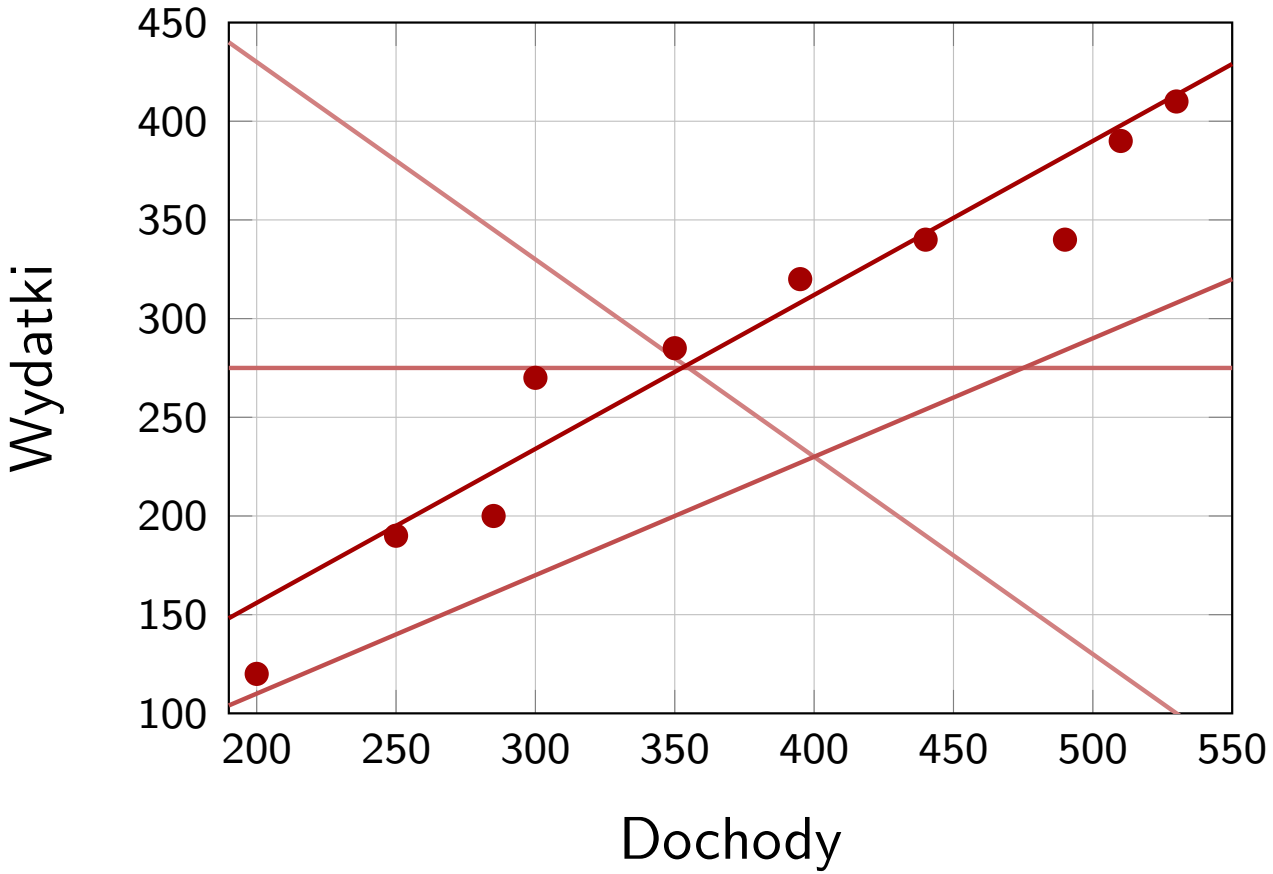
Indeks	Dochody	Wydatki
1	200	120
2	250	190
3	285	200
4	300	270
5	350	285
6	395	320
7	440	340
8	490	340
9	510	390
10	530	410


Wydatki
450
400
350
300
250
200
150
100
200
250
300
350
400
450
500
550
Dochody

Mamy $n = 10$ obserwacji, które przedstawiamy na płaszczyźnie jako punkty

$$(x_i, y_i) \in \{(200, 120), (250, 190), ..., (530, 410)\}, \quad i = 1, 2, ..., n$$

- Pozytywna korelacja między x a y
- Chcemy dopasować krzywą do danych
- Punkty układają się wzdłuż jakiejś prostej
- Chcemy wybrać najlepszą prostą
- Czym jest ta "najlepsza prosta"?



- Która prosta jest lepsza od innych?
- Dlaczego?
- Potrzebujemy obiektywnej metody!

Metoda najmniejszych kwadratów

Rysunek: Carl Friedrich Gauss

- Minimalizacja błędu
- Chcemy znaleźć parametry a i b
- Różnica między punktami a szacowaną prostą jak najmniejsza

$$y_i - (\hat{a}x_i + \hat{b}) \leftarrow \text{reszta}$$

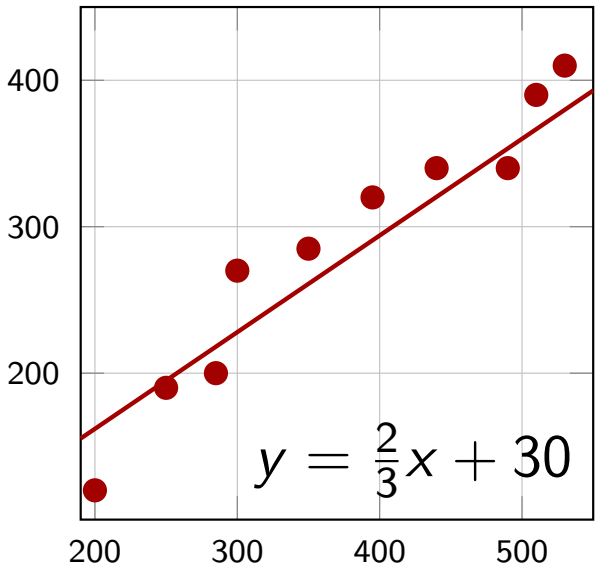
$$\hat{a}, \hat{b} \leftarrow \text{estymatory}$$

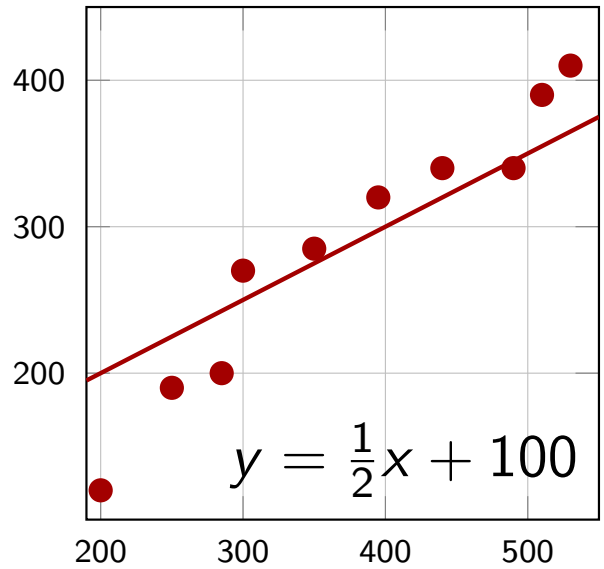
- Im bliżej estymatory są bliskie prawdziwym parametrom, tym reszta jest mniejsza

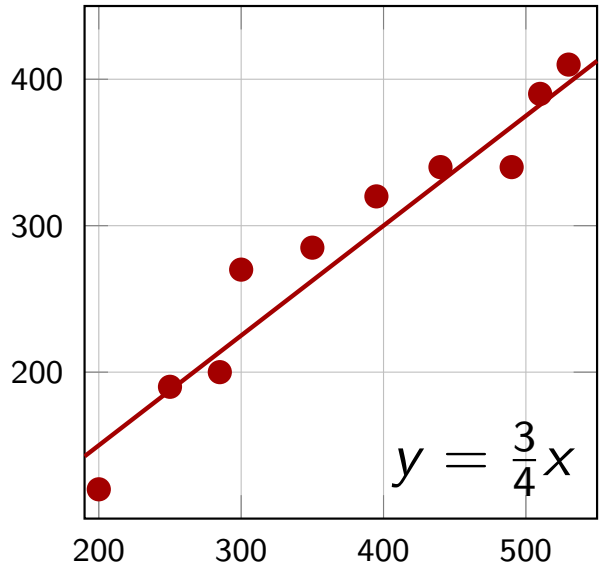
- Zkwadratujmy reszty
- Pozbywamy się problemu ujemnych i dodatnich reszt

$$(y_1 - \hat{a}x_1 - \hat{b})^2 + (y_2 - \hat{a}x_2 - \hat{b})^2 + ... + (y_n - \hat{a}x_n - \hat{b})^2 =$$
$$= \sum_{i=1}^{n} (y_i - \hat{a}x_i - b)^2 \leftarrow \text{suma kwadratów różnic}$$

- Suma najmniejsza $\iff$ każda różnica do kwadratu najmniejsza


400
300
200
200
300
400
500
$y = \frac{2}{3}x + 30$


400
300
200
200
300
400
500
$y = \frac{1}{2}x + 100$


400
300
200
200
300
400
500
$y = \frac{3}{4}x$

- Dla $y = \frac{2}{3} + 30$ suma kwadratów reszt wynosi $6\,769.4$
- Dla $y = \frac{1}{2} + 100$ suma kwadratów reszt wynosi $14\,112.5$
- Dla $y = \frac{3}{4}$ suma kwadratów reszt wynosi $5\,259.4$

Zatem trzecia linia najlepiej jest dopasowana do danych spośród tej trójki

Wzory na parametry najlepszej prostej

$$a = \frac{\sum_{i=1}^{n} y_i(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad b = \bar{y} - \bar{x}a$$

gdzie $\overline{x}$ i $\overline{y}$ to średnie arytmetyczne zmiennych x i y

Metoda najmniejszych kwadratów w praktyce

Podstawmy nasze dane do wzorów!

- $\bar{x} = \frac{1}{10}(200 + 250 + \cdots + 530) = 375$
- $\bar{y} = \frac{1}{10}(120 + 190 + \cdots + 200) = 286.5$

Następnie obliczymy odpowiednie sumy do wzoru na współczynnik a

$$\sum_{i=1}^{10} y_i(x_i - \bar{x}) = 120 \cdot (200 - 375) + ... + 410 \cdot (530 - 375) = 93\,675$$

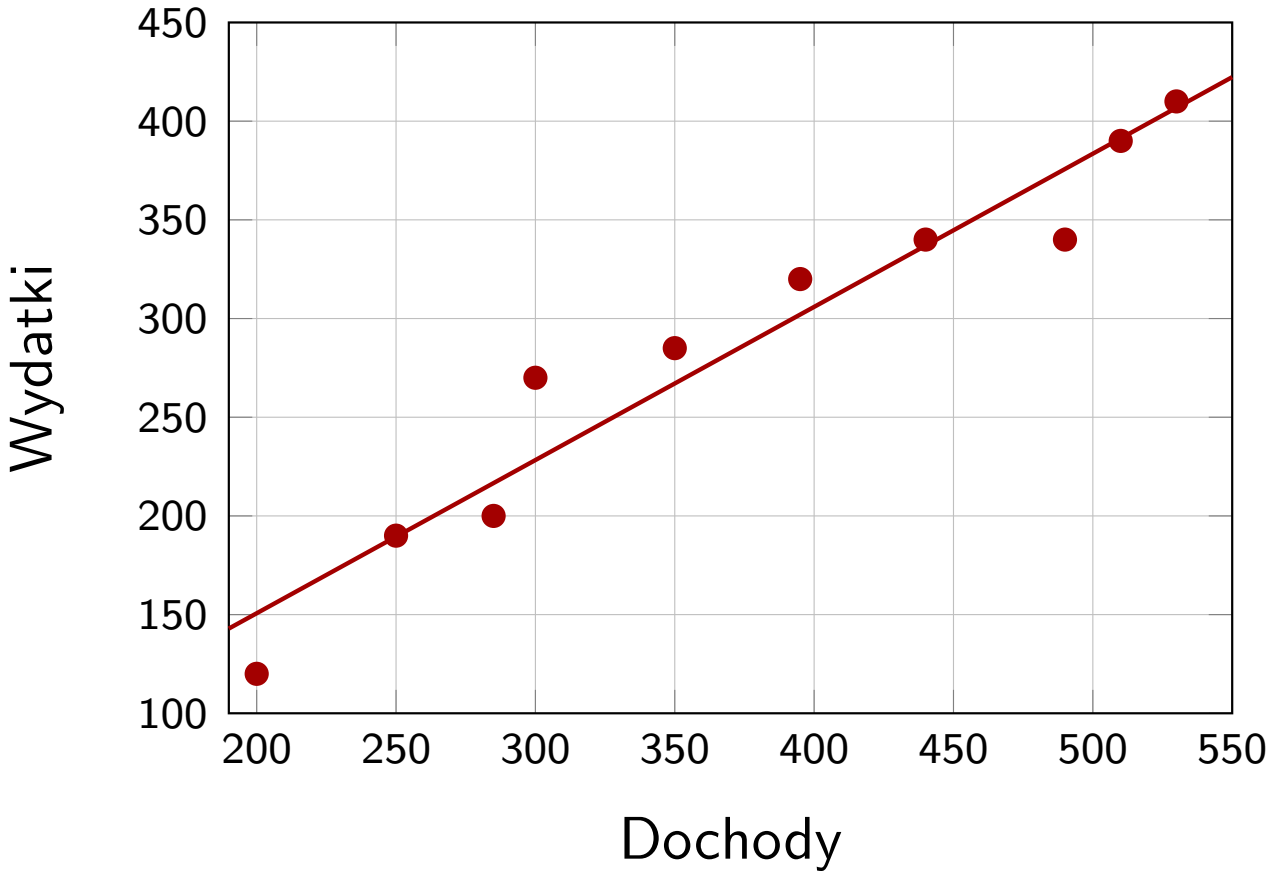
$$\sum_{i=1}^{10} (x_i - \bar{x})^2 = (200 - 375)^2 + ... + (530 - 375)^2 = 120\,700$$

Zostało nam już tylko podstawienie liczb do wzorów na parametry a i b

- $a = \frac{93\,675}{120\,700} \approx 0.78$
- $b = 286.5 - 375 \cdot 0.78 \approx -4.54$

Zatem szukana funkcja jest postaci

$$y = 0.7761x - 4.5367$$



$$\sum_{i=1}^{n} (y_i - \hat{a}x_i - b)^2 = 4901.5$$

Podsumowanie

- Regresja liniowa to opis korelacji między zmiennymi za pomocą krzywej
- Dopasowanie prostej to znalezienie takich parametrów a i b, że funkcja odległości punktów pomiarowych od prostej przyjmuje wartość minimalną
- Metoda najmniejszych kwadratów to takie dopasowanie prostej, aby suma kwadratów różnic między punktami pomiarowymi, a prostą była jak najmniejsza

Dla zainteresowanych tematem będących ciekawych skąd wzięły się wzory na parametry a i b, wykorzystujemy tutaj optymalizację funkcji dwóch zmiennych. Znajdujemy ekstremum lokalne funkcji sumy kwadratów różnic.