Regresja liniowa

Weronika Piotrowska,

Golem 2022

Agenda

1. Wprowadzenie
2. Regresja liniowa od A do Z
    1. Jak zminimalizować MSE?
    2. Spadek gradientu
    3. Więcej zmiennych
    4. Regularyzacja
3. Inne rodzaje regresjii
4. Kodowańsko

Wprowadzenie

Na podstawie tego

Chcemy przewidzieć to

No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
4	2013.5	13.3	561.9845	5	24.98746	121.54391	54.8
5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
9	2013.5	31.7	5512.038	1	24.95095	121.48458	18.8
10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1
11	2013.083	34.8	405.2134	1	24.97349	121.53372	41.4
12	2013.333	6.3	90.45606	9	24.97433	121.5431	58.1
13	2012.917	13	492.2313	5	24.96515	121.53737	39.3
14	2012.667	20.4	2469.645	4	24.96108	121.51046	23.8
15	2013.5	13.2	1164.838	4	24.99156	121.53406	34.3

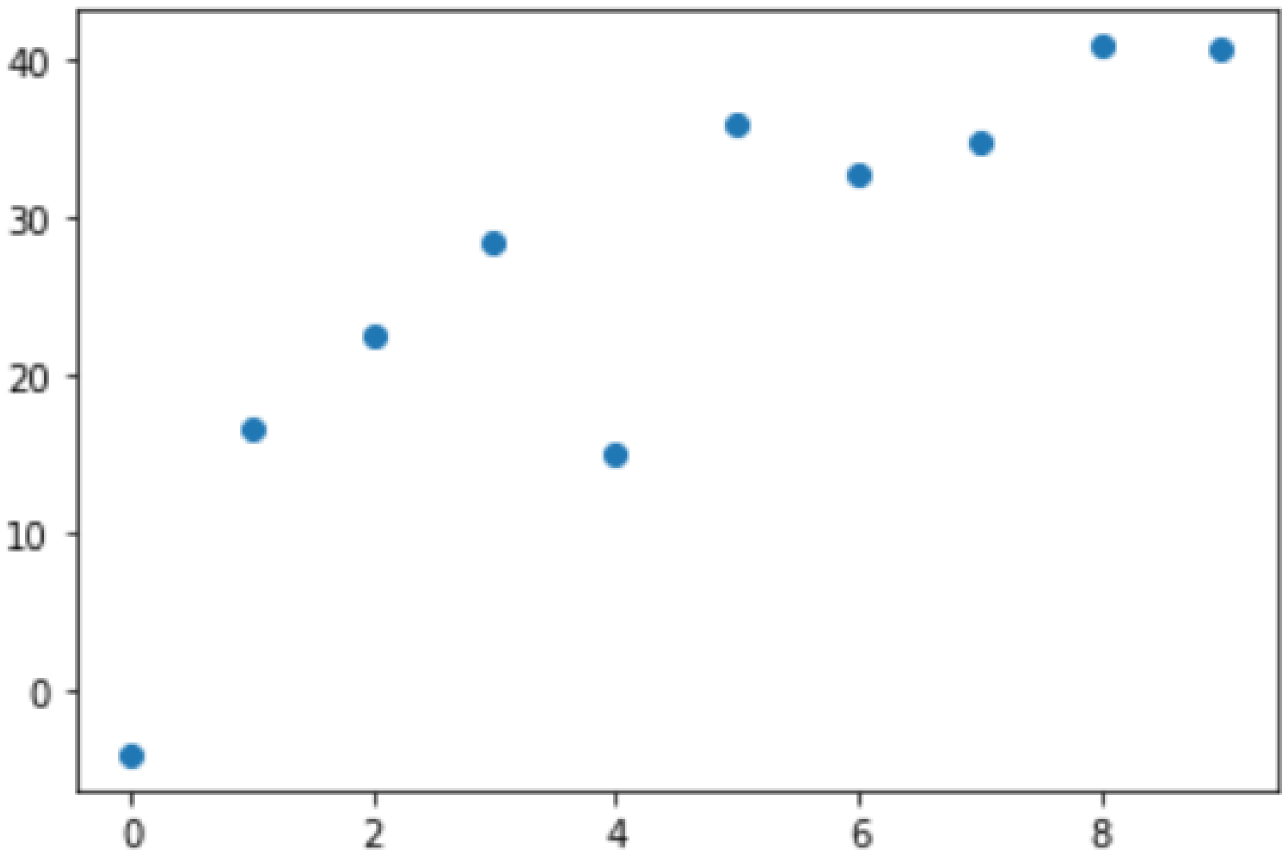
Wprowadzenie

features

target / label

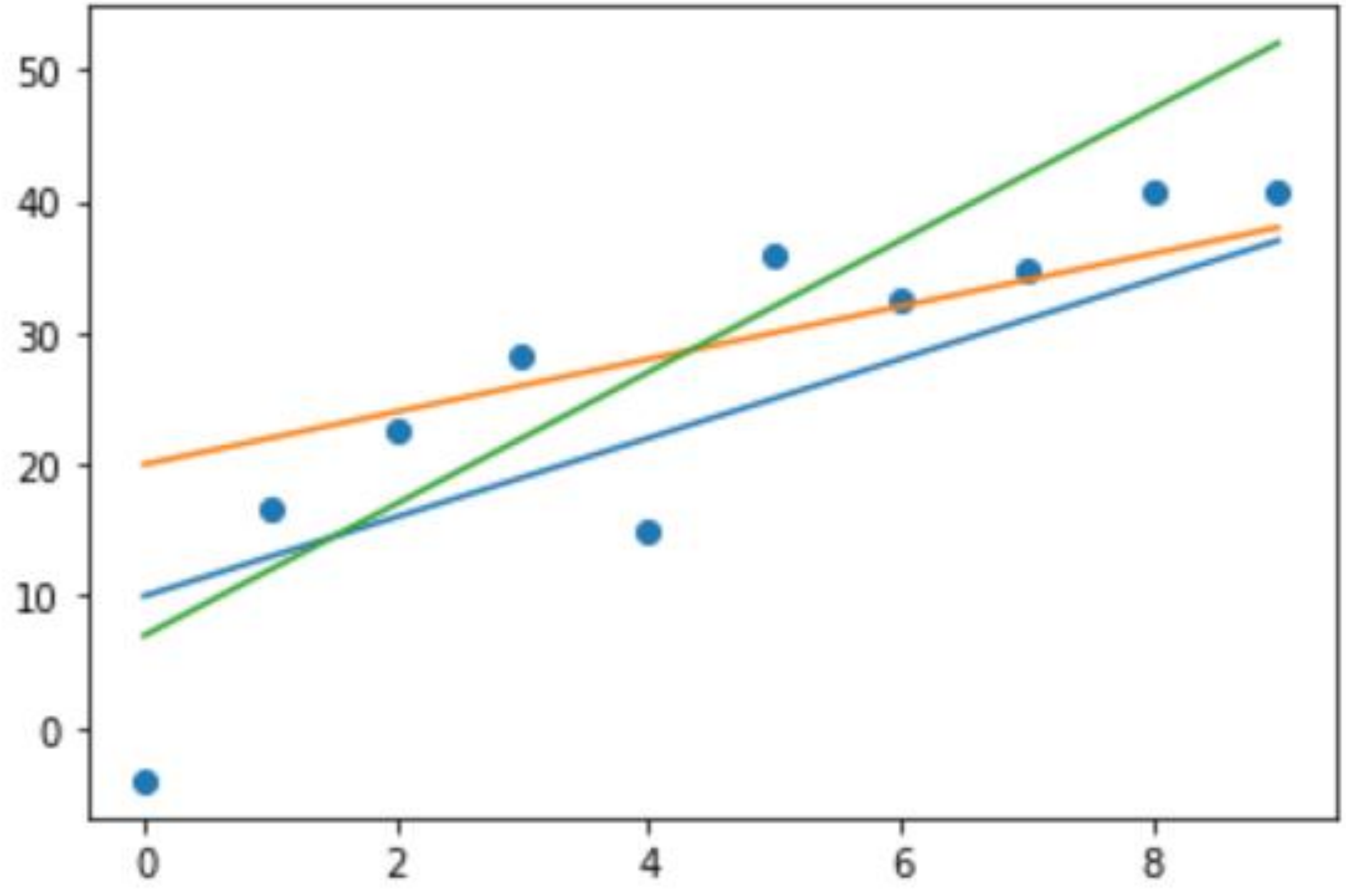
No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
4	2013.5	13.3	561.9845	5	24.98746	121.54391	54.8
5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
9	2013.5	31.7	5512.038	1	24.95095	121.48458	18.8
10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1
11	2013.083	34.8	405.2134	1	24.97349	121.53372	41.4
12	2013.333	6.3	90.45606	9	24.97433	121.5431	58.1
13	2012.917	13	492.2313	5	24.96515	121.53737	39.3
14	2012.667	20.4	2469.645	4	24.96108	121.51046	23.8
15	2013.5	13.2	1164.838	4	24.99156	121.53406	34.3

	x	y
0	0	-4.0
1	1	16.6
2	2	22.6
3	3	28.4
4	4	15.0
5	5	35.8
6	6	32.6
7	7	34.7
8	8	40.8
9	9	40.7



$$y = \beta_0 + \beta_1 x$$

Jak znaleźć najlepszą wartość β_0 i β_1?



MSE

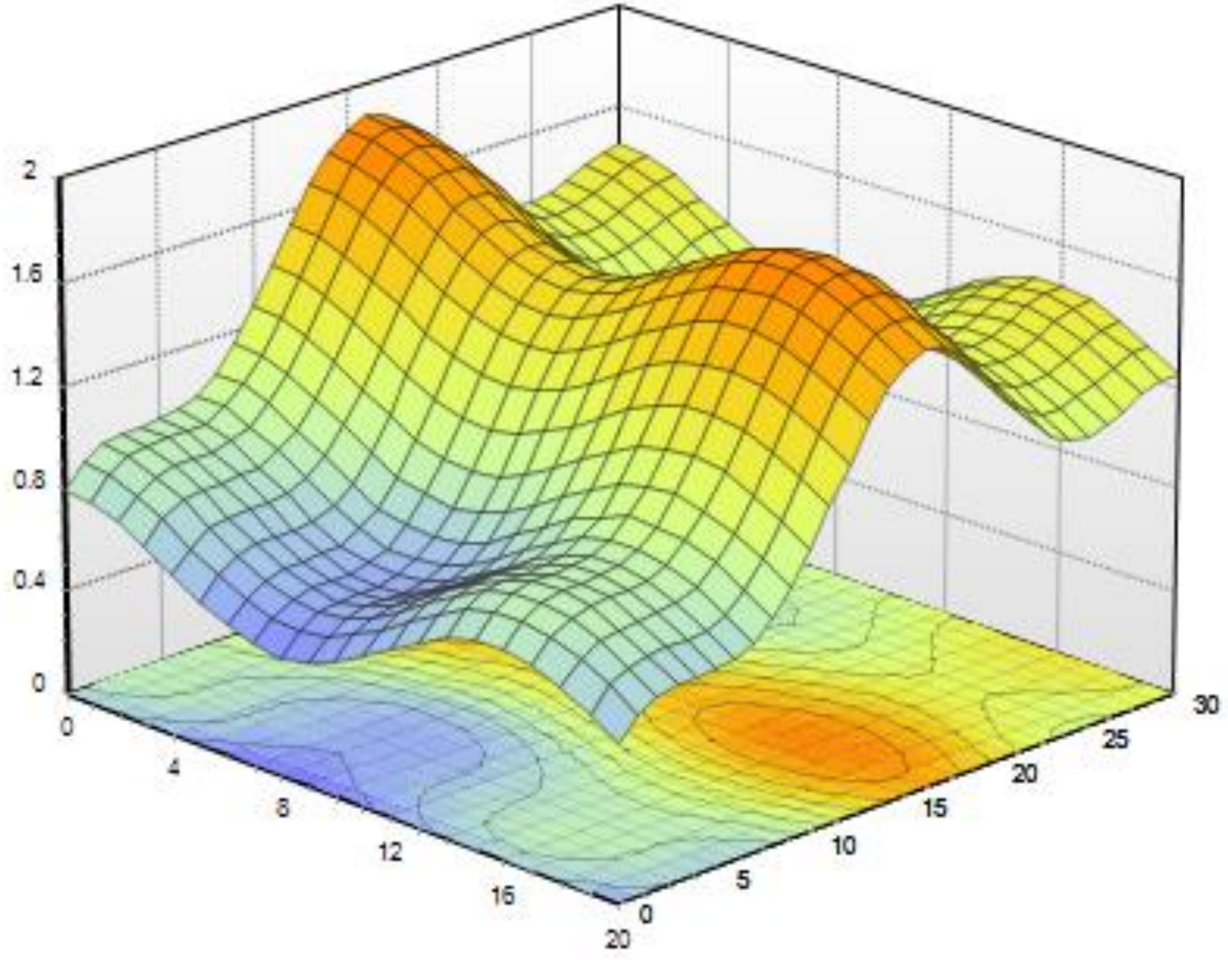
$$MSE = \frac{1}{n}\sum_{i=0}^{n} ||y_i - f(x_i)||_2^2$$

W tym przypadku możemy to uprościć (Dlaczego?)

$$MSE = \frac{1}{n}\sum_{i=0}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2$$

MSE jako funkcja od β_0 i β_1

- Oś pionowa: MSE
- Osie poziome: β_0 i β_1



$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - (\beta_0 + \beta_1 x_i))^2$$

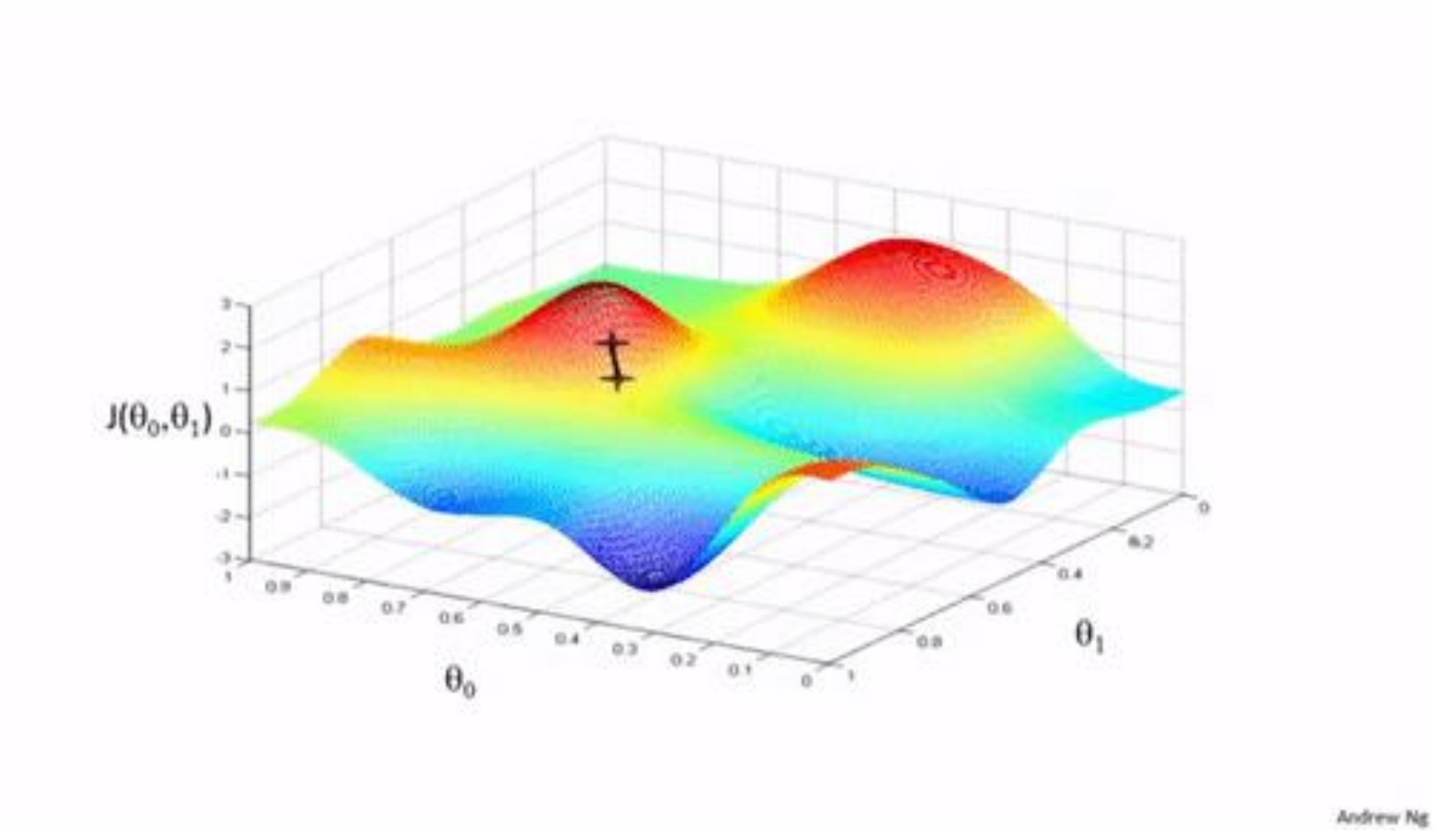
Znaleźć taką wartość β_0 i β_1 żeby MSE była najmniejsza

Pochodne!

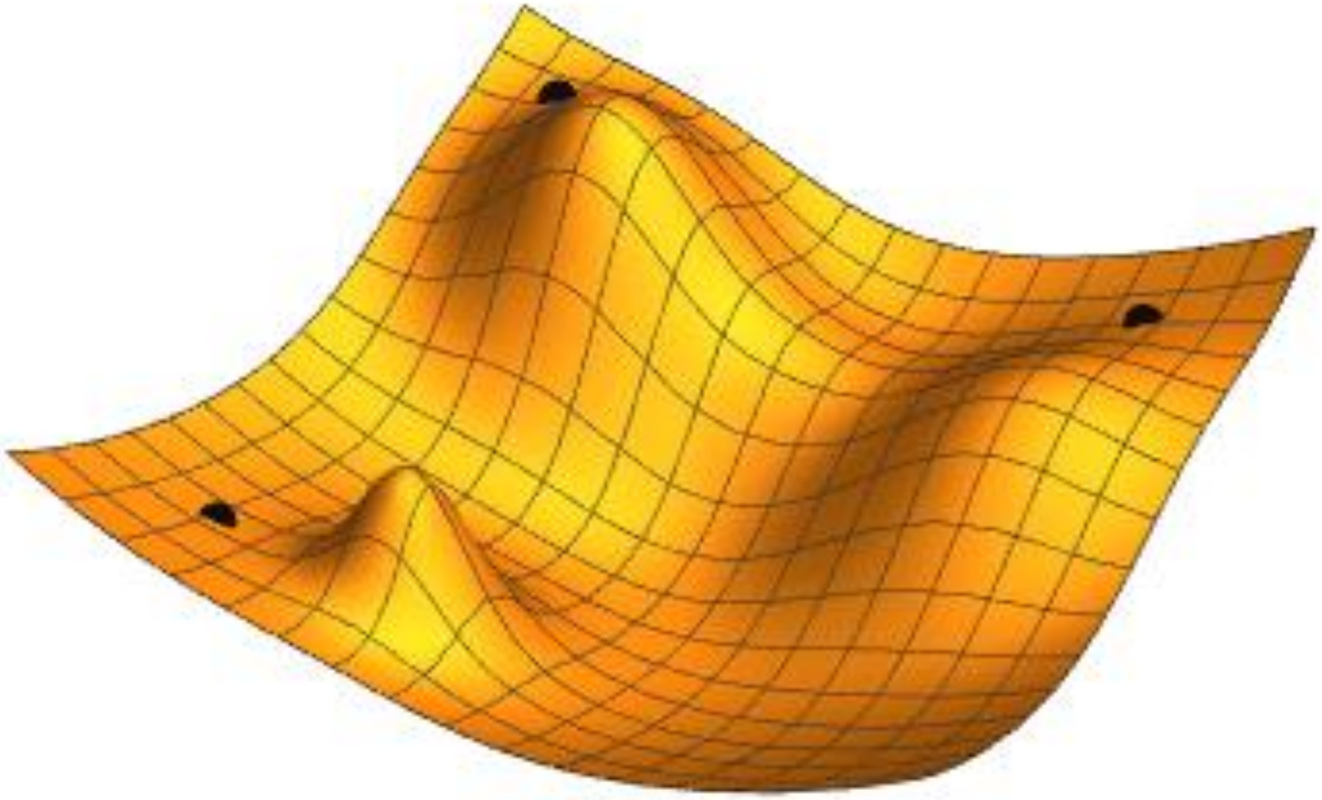
$$\frac{\partial MSE}{\partial \beta_0} = \frac{-2}{n}\sum_{i=0}^{n}(y_i - (\beta_0 + \beta_1 x_i))$$

$$\frac{\partial MSE}{\partial \beta_1} = \frac{-2}{n}\sum_{i=0}^{n} x_i(y_i - (\beta_0 + \beta_1 x_i))$$

Spadek gradientu



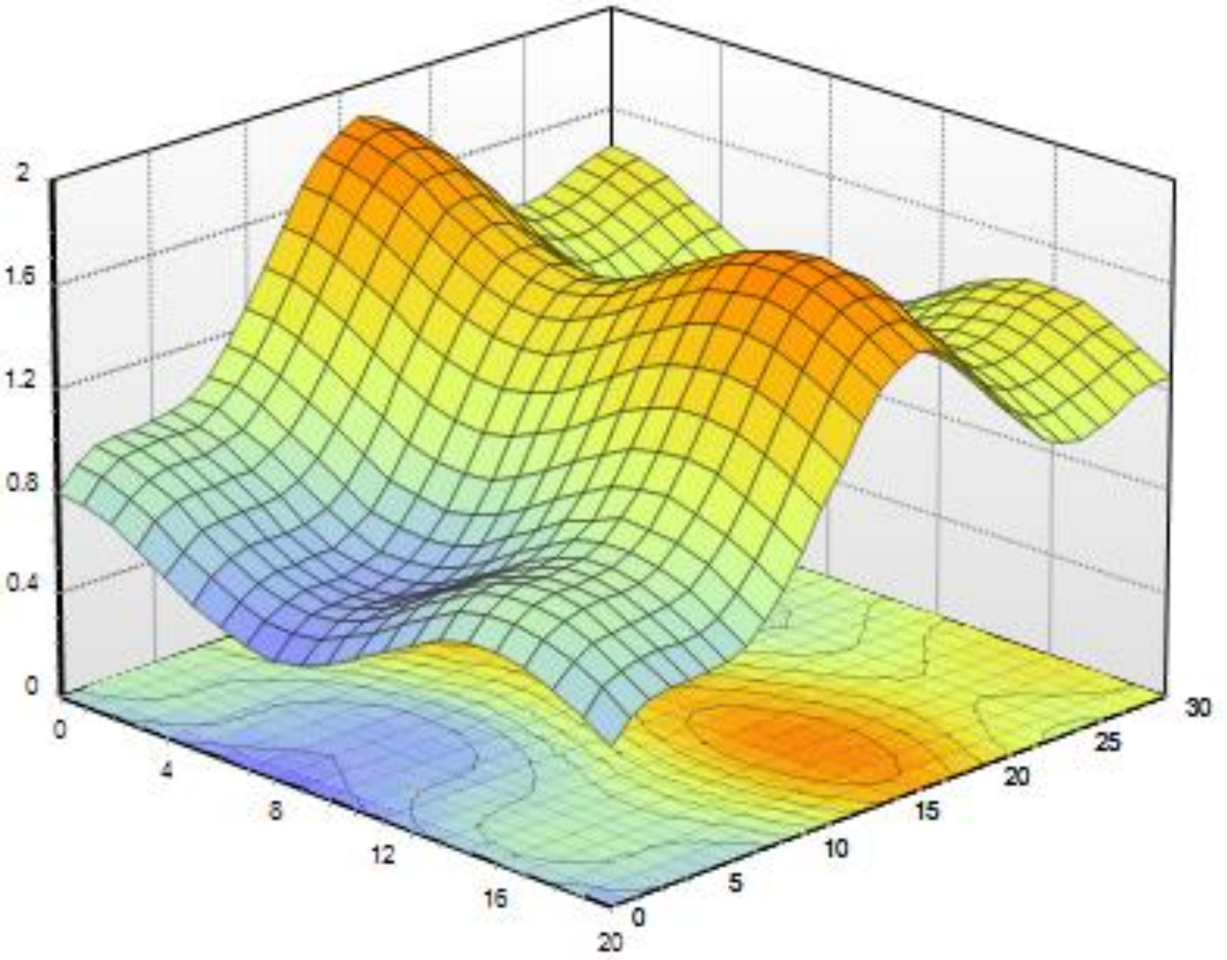
Spadek gradientu



Learning rate

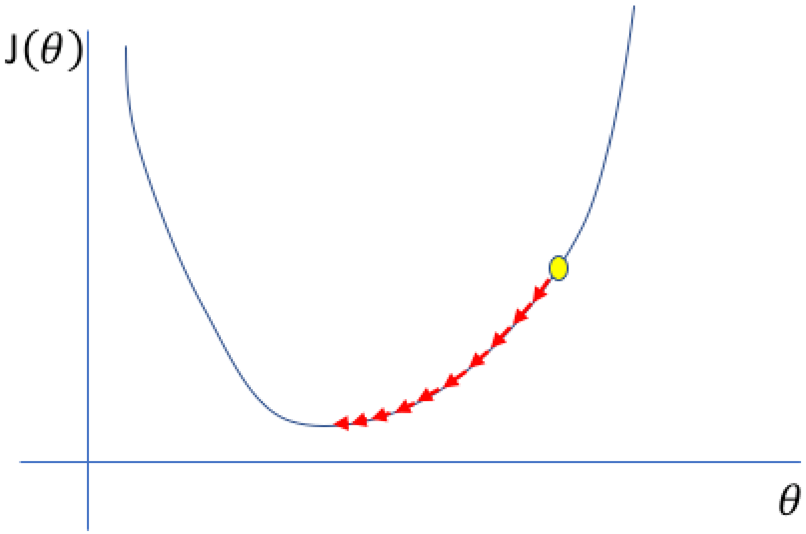
$$\beta_0 = \beta_0 - \frac{\partial MSE}{\partial \beta_0} \cdot \alpha$$

$$\beta_1 = \beta_1 - \frac{\partial MSE}{\partial \beta_1} \cdot \alpha$$



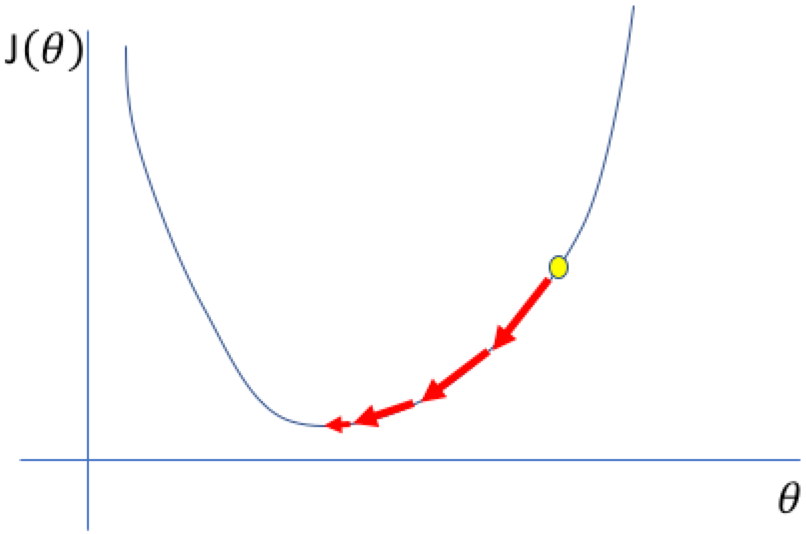
Learning rate

Too low



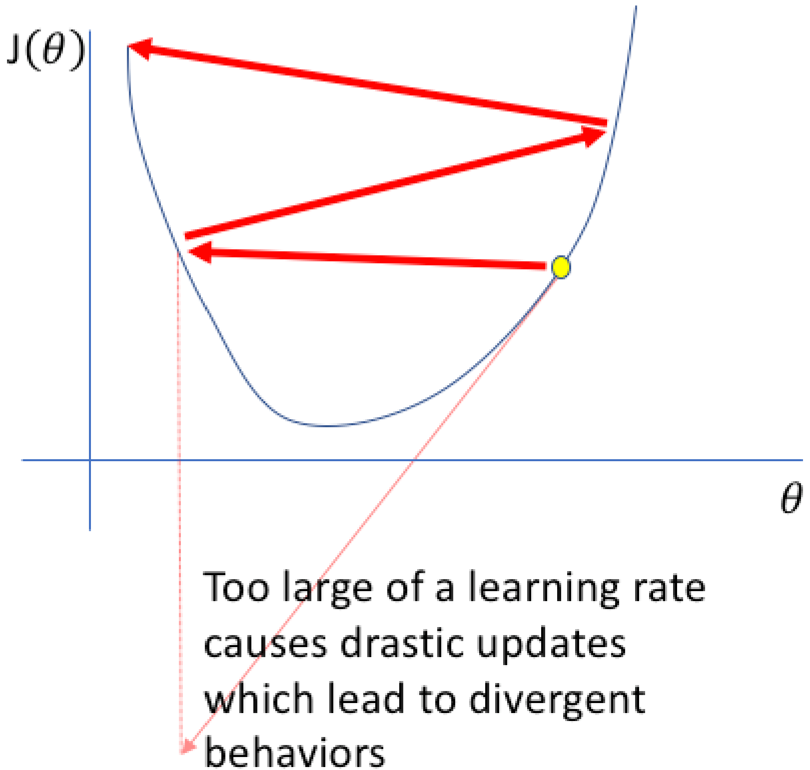
A small learning rate requires many updates before reaching the minimum point

Just right



The optimal learning rate swiftly reaches the minimum point

Too high



Too large of a learning rate causes drastic updates which lead to divergent behaviors

Więcej zmiennych

No	X1 transaction date	X2 house age	X3 distance to the nearest MRT station	X4 number of convenience stores	X5 latitude	X6 longitude	Y house price of unit area
1	2012.917	32	84.87882	10	24.98298	121.54024	37.9
2	2012.917	19.5	306.5947	9	24.98034	121.53951	42.2
3	2013.583	13.3	561.9845	5	24.98746	121.54391	47.3
4	2013.5	13.3	561.9845	5	24.98746	121.54391	54.8
5	2012.833	5	390.5684	5	24.97937	121.54245	43.1
6	2012.667	7.1	2175.03	3	24.96305	121.51254	32.1
7	2012.667	34.5	623.4731	7	24.97933	121.53642	40.3
8	2013.417	20.3	287.6025	6	24.98042	121.54228	46.7
9	2013.5	31.7	5512.038	1	24.95095	121.48458	18.8
10	2013.417	17.9	1783.18	3	24.96731	121.51486	22.1
11	2013.083	34.8	405.2134	1	24.97349	121.53372	41.4
12	2013.333	6.3	90.45606	9	24.97433	121.5431	58.1
13	2012.917	13	492.2313	5	24.96515	121.53737	39.3
14	2012.667	20.4	2469.645	4	24.96108	121.51046	23.8
15	2013.5	13.2	1164.838	4	24.99156	121.53406	34.3

Więcej zmiennych

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

- Wszystko robimy tak samo tylko dla większej ilości β
- Nie da się zwizualizować ☹

Regularyzacja

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n$$

Jeśli na przykład β_0 = 0.05, β_1=2 , β_2=500, to wynik zależy tylko od x_2

Regularyzacja

Zmieniamy troche funkcję straty

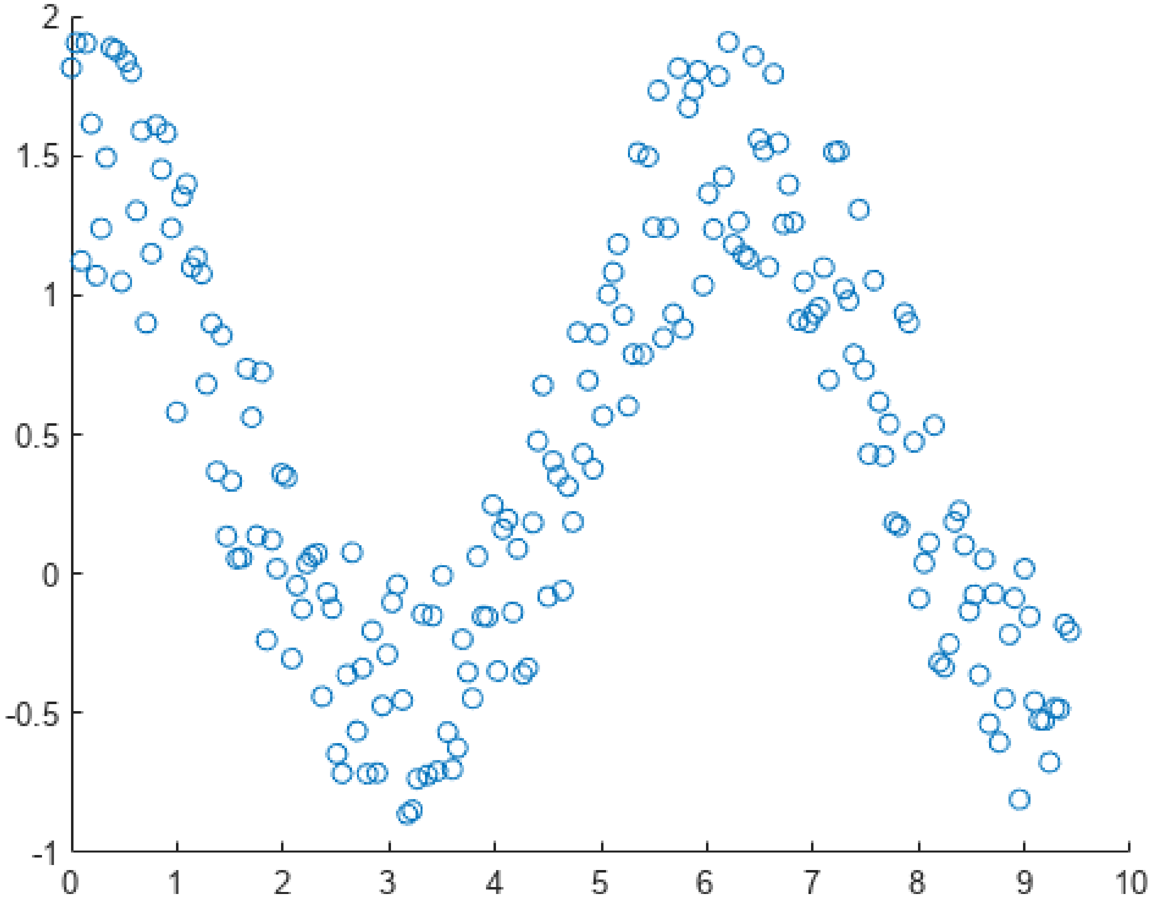
Ridge

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - pred_i)^2 + \lambda\sum_{j=0}^{m}\beta_j^2$$

Lasso

$$MSE = \frac{1}{n}\sum_{i=0}^{n}(y_i - pred_i)^2 + \lambda\sum_{j=0}^{m}|\beta_j|$$

Inne modele regresyjne



Regresja wielomianowa

$$y = \beta_0 + \beta_1 x_1 \longrightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \longrightarrow \text{???}$$

Regresja wielomianowa

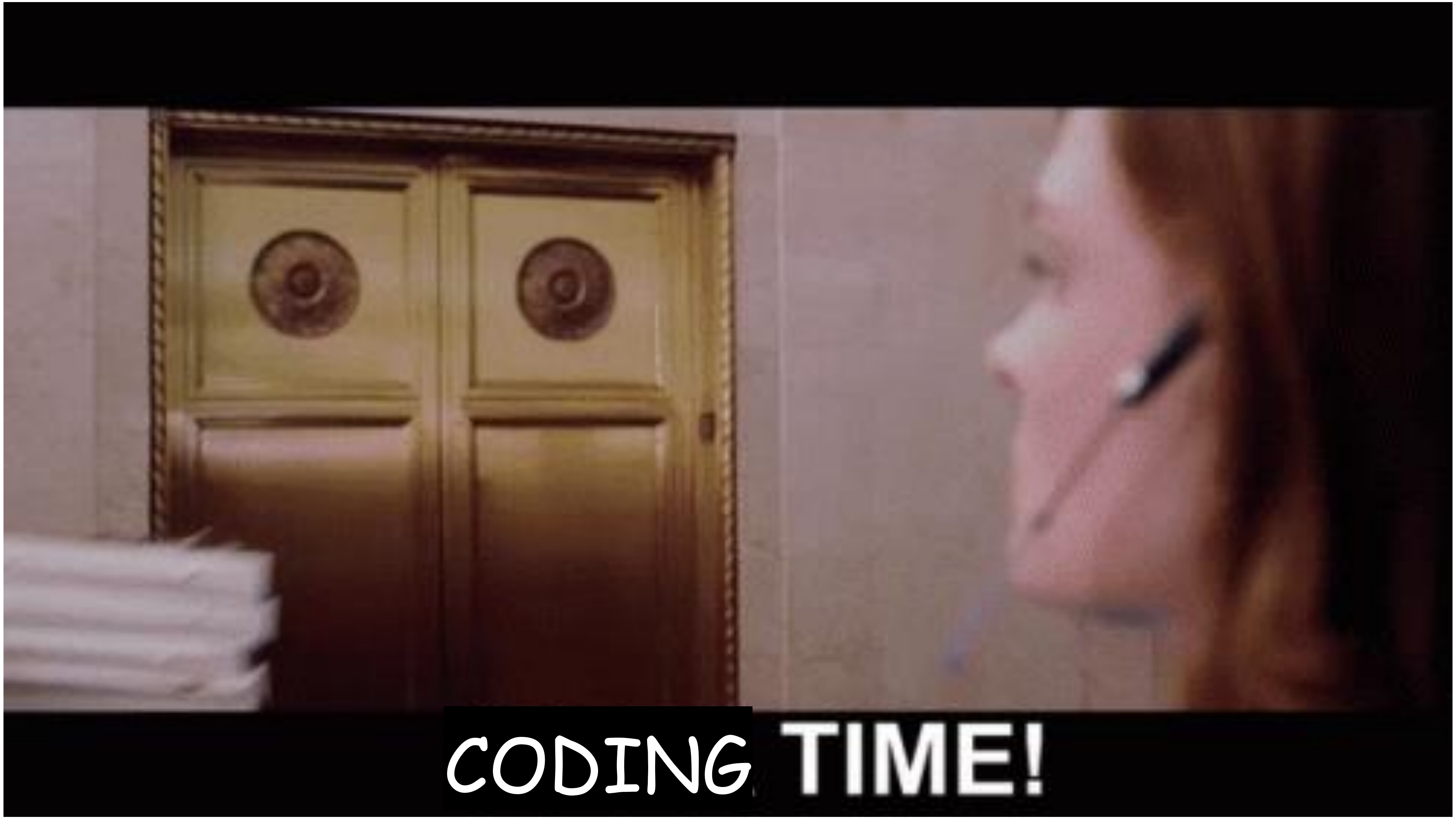
$$y = \beta_0 + \beta_1 x_1 \longrightarrow y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 \, x_2^2 + \beta_5 x_1 x_2$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Regresja wielomianowa

- Bierzemy pod uwagę wszystkie możliwe iloczyny wszystkich x-ów do określonego stopnia wielomianu
- Każdy iloczyn dostaje swój współczynnik β
- Dalej robimy jak w przypadku zwykłej regresji liniowej


CODING TIME!