Metody eksploracji danych
2. Metody regresji

Piotr Szwed
Katedra Informatyki Stosowanej AGH
2016

Zagadnienie regresji

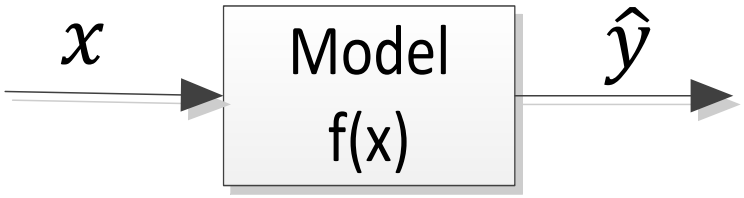
Dane:

- Zbiór uczący: $\mathrm{D} = \{(x_i, y_i)\}_{i=1,m}$
- Obserwacje: (x_i, y_i), wektor cech $x_i \in \boldsymbol{R}^n$
- Wartość wyjściowa jest skalarem $y_i \in \boldsymbol{R}$

Zadanie: dobór funkcji

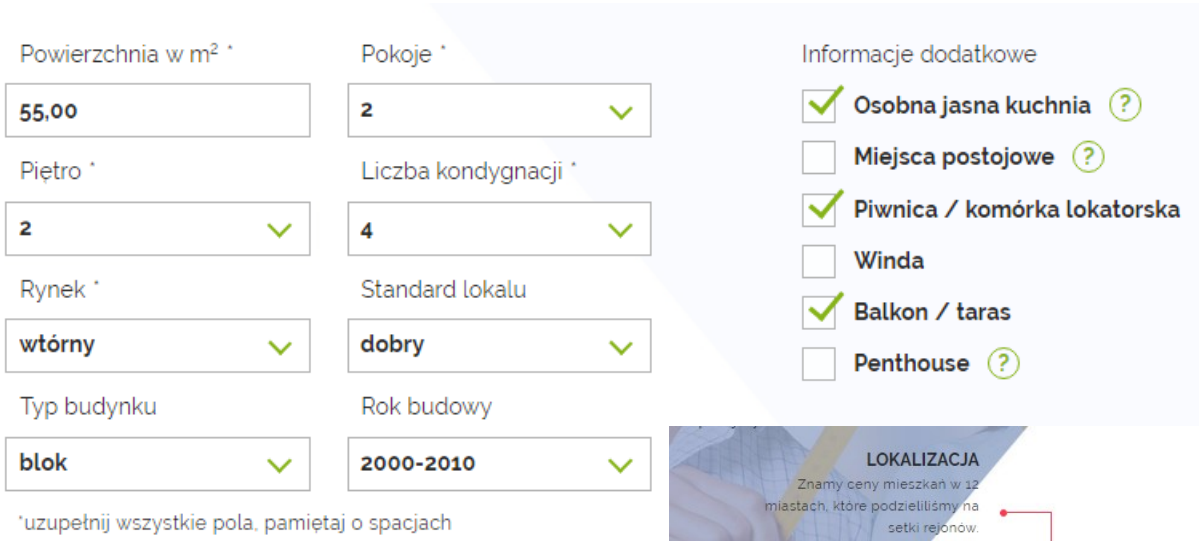
$f(x)$: $\boldsymbol{R}^n \rightarrow \boldsymbol{R}$,

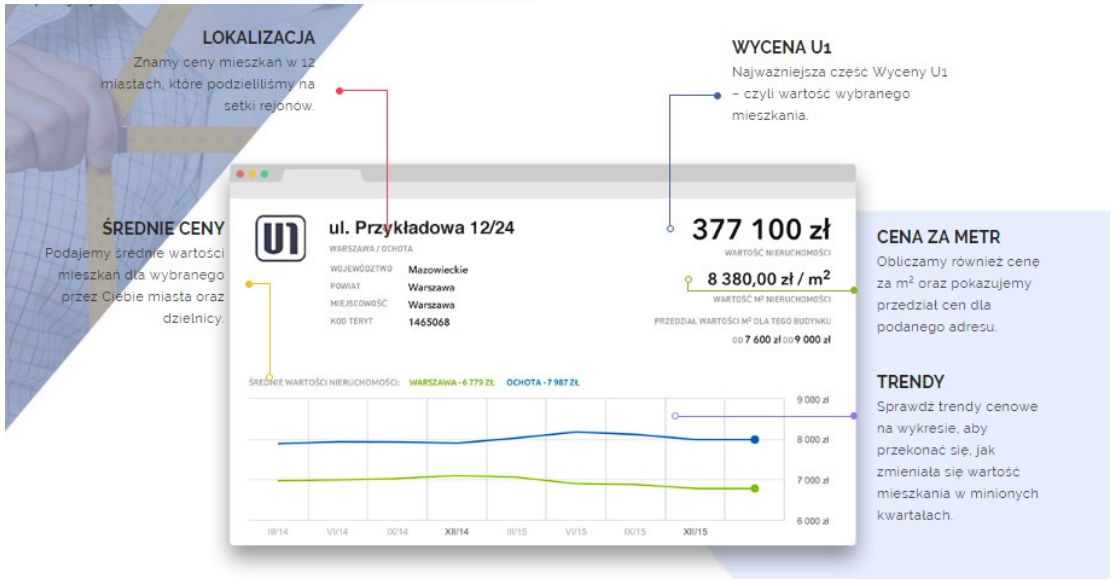
która pozwoli **przewidzieć** wartość wyjściową $\hat{y}$ odpowiadającą x



Przykład 1

Szacowanie wartości nieruchomości na podstawie danych transakcyjnych z uwzględnieniem różnych atrybutów



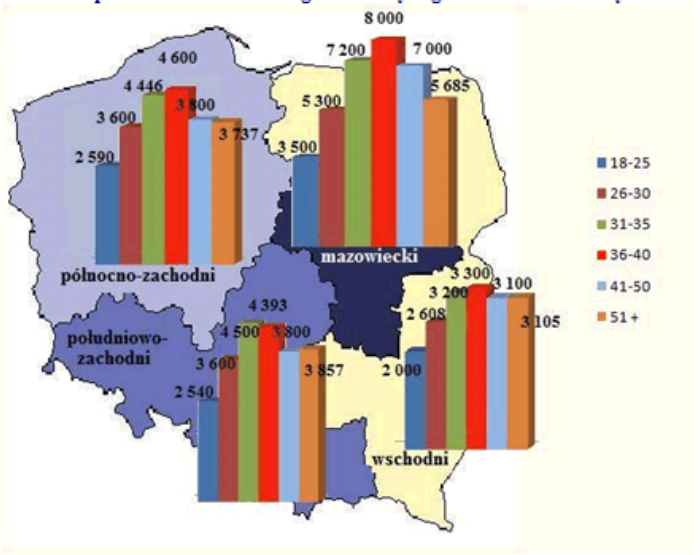


[https://urban.one/]

Przykład 2

Przewidywanie wynagrodzeń w zależności od lokalizacji, wieku, branży, wykształcenia, stanowiska…

Mapa 1. Zróżnicowanie regionalne wynagrodzeń osób w różnym wieku



Źródło: Ogólnopolskie Badanie Wynagrodzeń przeprowadzone przez Sedlak & Sedlak w 2008 roku

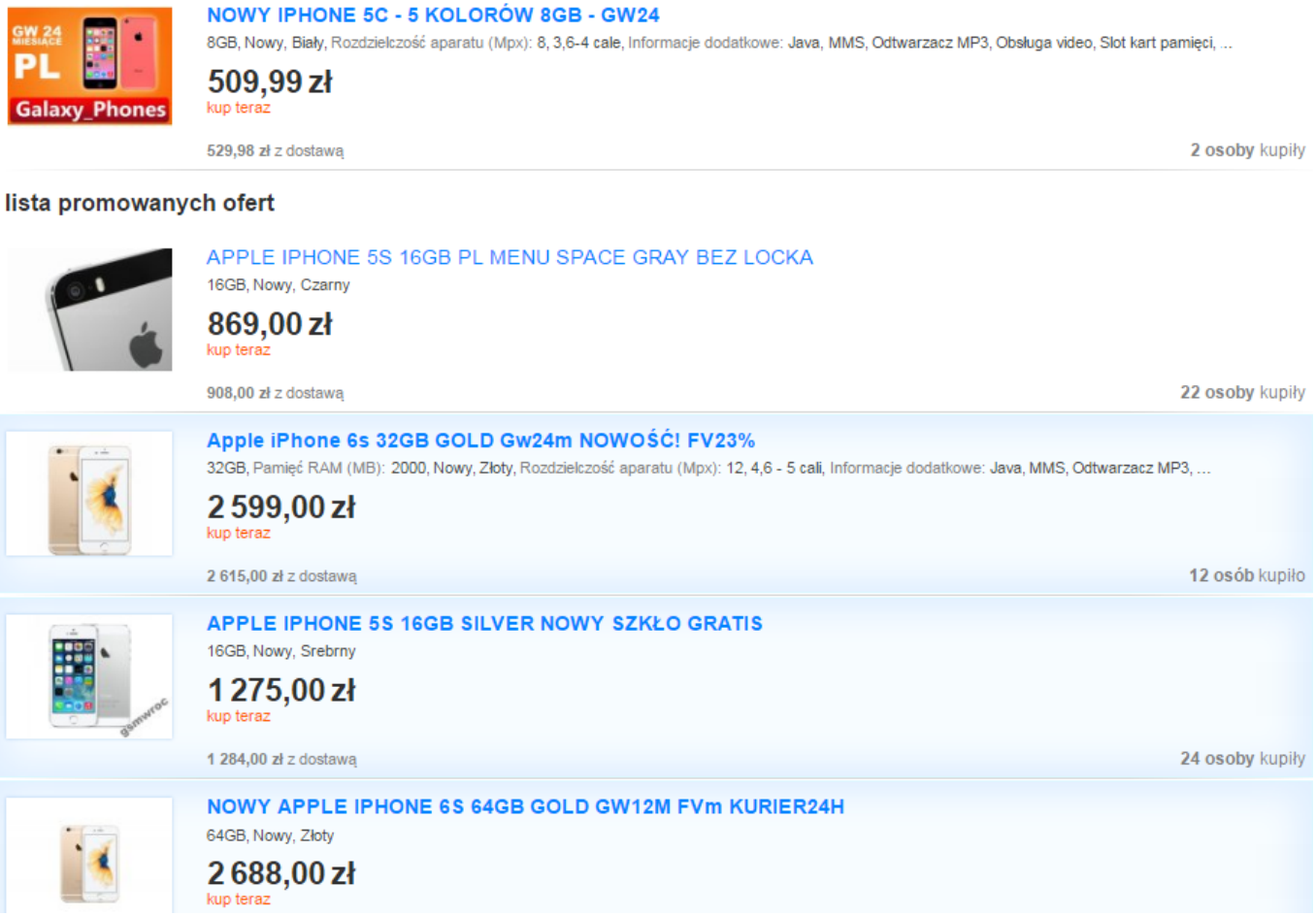
Tabelka 3. Miesięczne wynagrodzenie całkowite osób, które ukończyły różne kierunki studiów

kierunek studiów	próba	25% zarabia mniej	mediana 2015 rok	25% zarabia więcej
informatyka	10 542	4 000	6 250	9 300
elektronika i automatyka	4 100	4 179	5 900	8 500
mechanika i metalurgia	5 000	3 600	5 000	7 400
(...)				
ekonomia, finanse i zarządzanie	27 483	3 099	4 500	7 200
(...)				
pedagogiczne (związane z edukacją)	3 445	2 331	3 013	4 200
sport, rehabilitacja, fizykoterapia	671	2 200	2 800	4 325
pielęgniarstwo	885	2 200	2 670	3 400

ódło: Ogólnopolskie Badanie Wynagrodzeń (OBW) przeprowadzone przez Sedlak & Sedlak w 2015 roku

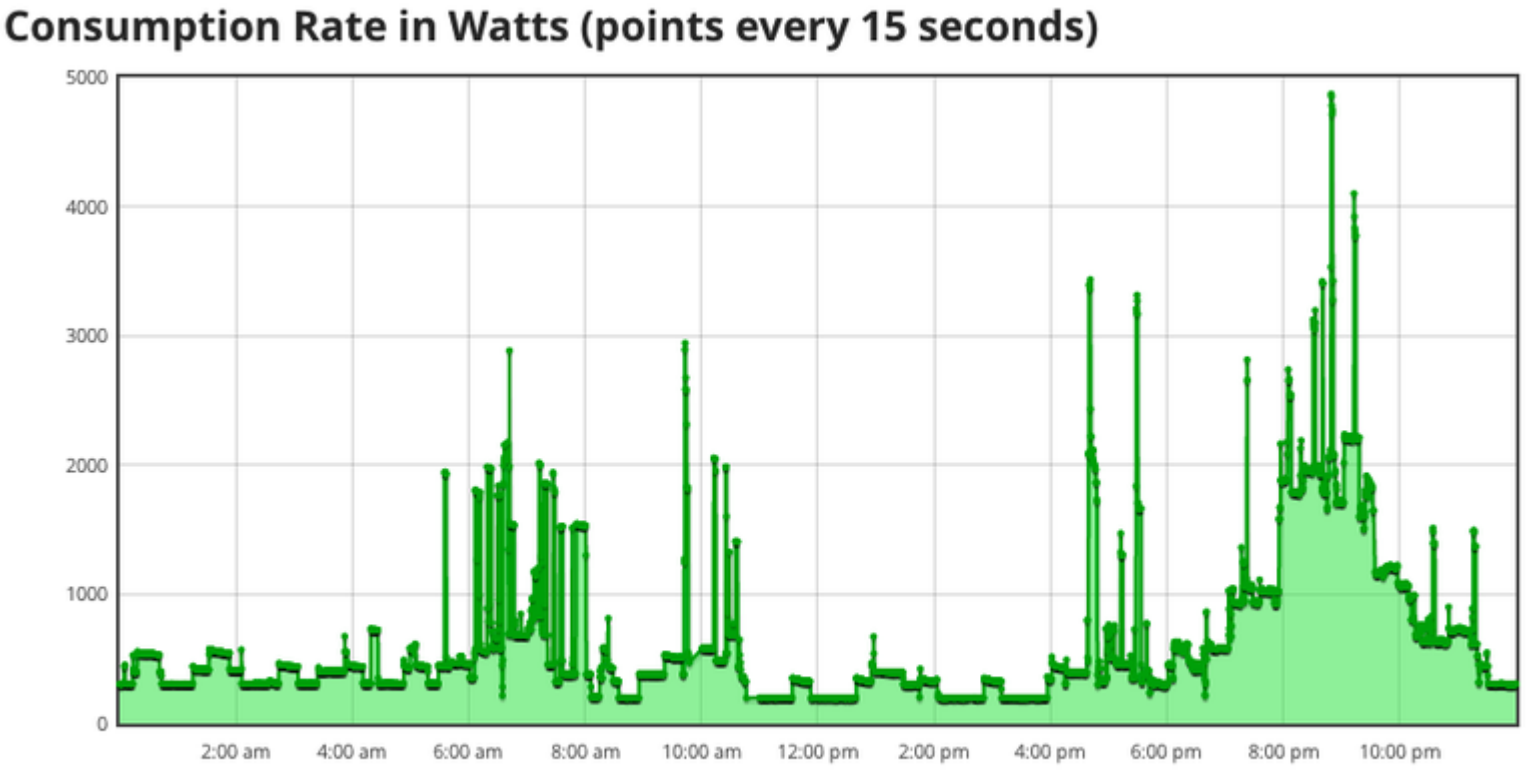
Przykład 3

Przewidywanie wielkości sprzedaży w zależności od typu towaru i ceny



Przykład 4

- Przewidywanie zużycia energii na podstawie danych historycznych



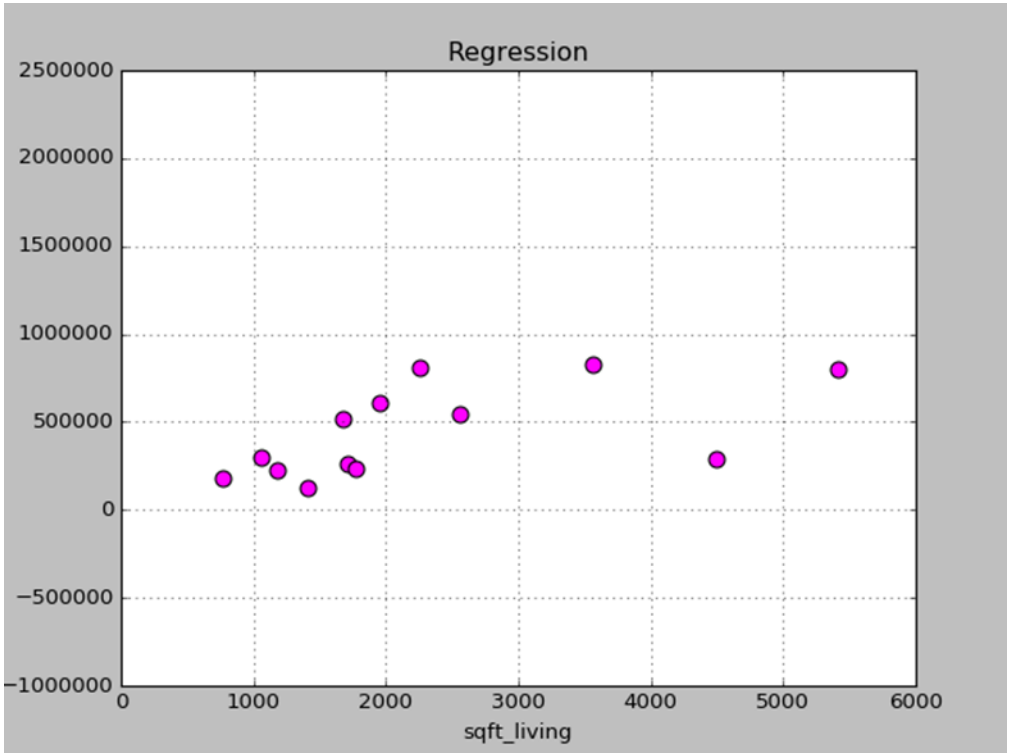
wattvision

[http://wattvision.posthaven.com/?page=5]

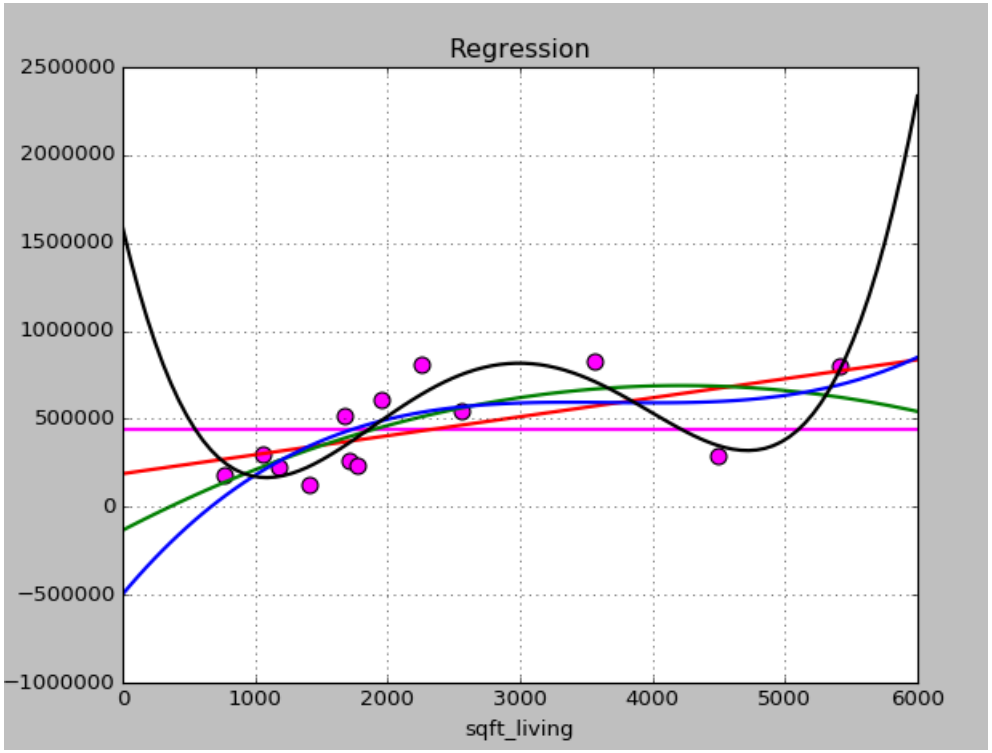
Etapy

- Wybór danych, które zostaną użyte $\mathrm{D} = \{(x_i, y_i)\}_{i=1,m}$
- Wybór postaci modelu
- **Określenie parametrów modelu**
- Walidacja modelu

Przykład: x to powierzchnia domu, y to cena domu



Postać modelu

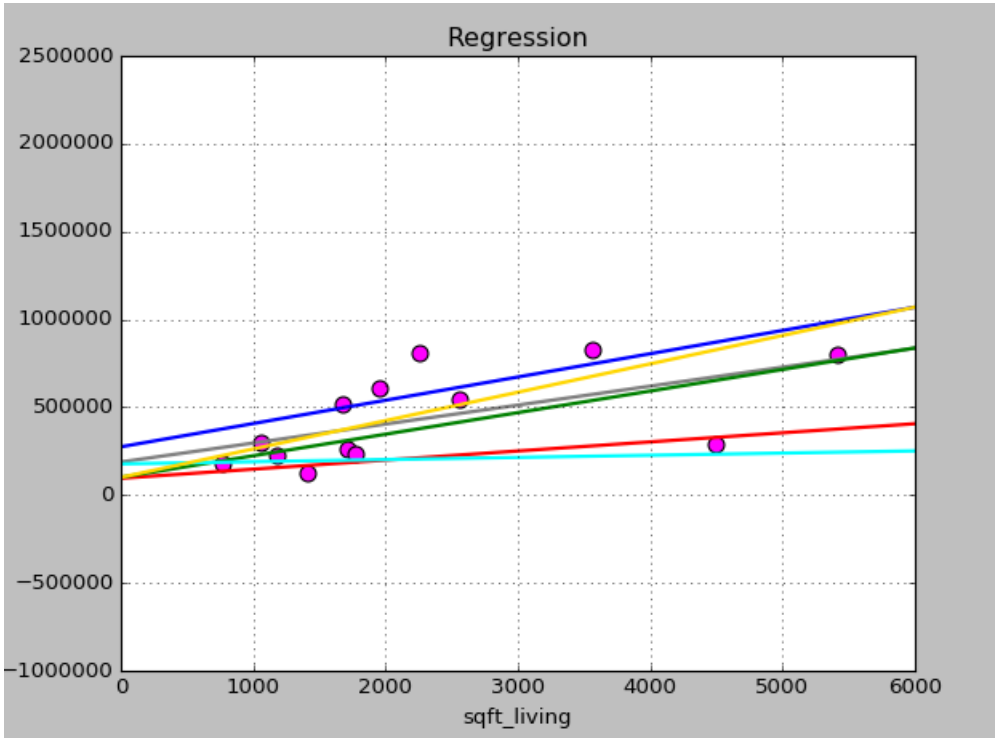


Wybór typowych funkcji (których parametry da się wyznaczyć)

- Funkcja stała?
- Funkcja liniowa?
- Wielomian wyższego rzędu?

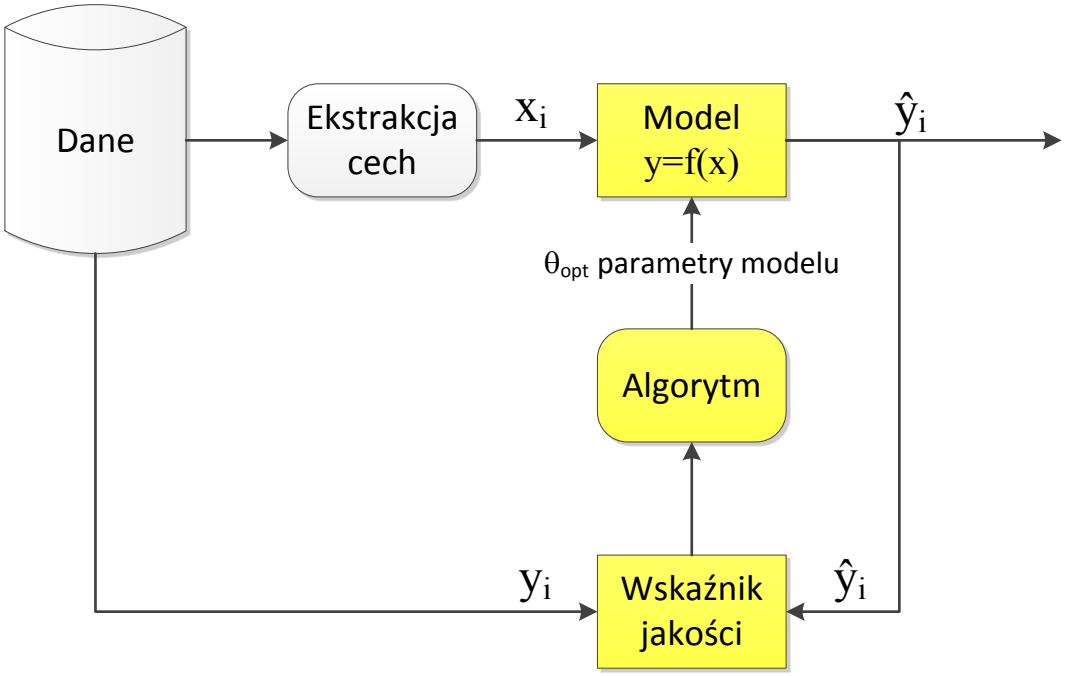
Z reguły bardziej złożona funkcja będzie lepiej przybliżała dane uczące, ale niekoniecznie sprawdzi się przy predykcji.

Parametry modelu



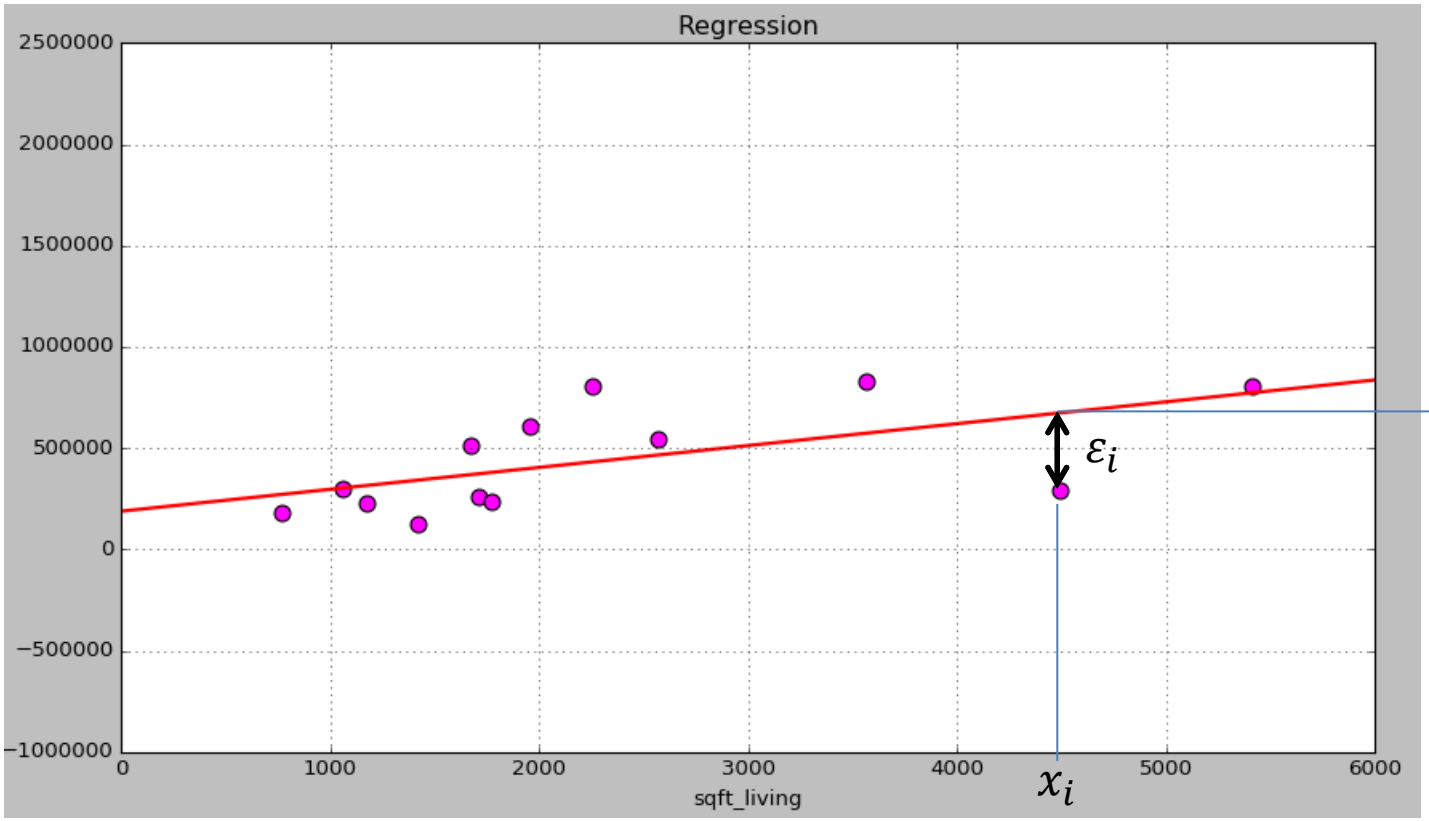
- Zakładamy, że wybraną postacią będzie funkcja liniowa:
  $$y(x) = w_0 + w_1 x$$
- Należy wybrać „najlepszą” lub „najbardziej prawdopodobną” funkcję
- Konieczne są założenia i kryterium pozwalające na ocenę modelu

Przebieg procesu uczenia



- Wskaźnik jakości służy do porównania:
  - y_i- wartości wyjściowej zapisanych w zbiorze uczącym z
  - $\hat{y}_i$- wartości przewidywanej przez model
- Wskaźnik jakości steruje przebiegiem algorytmu lub jest wykorzystywany w jego konstrukcji

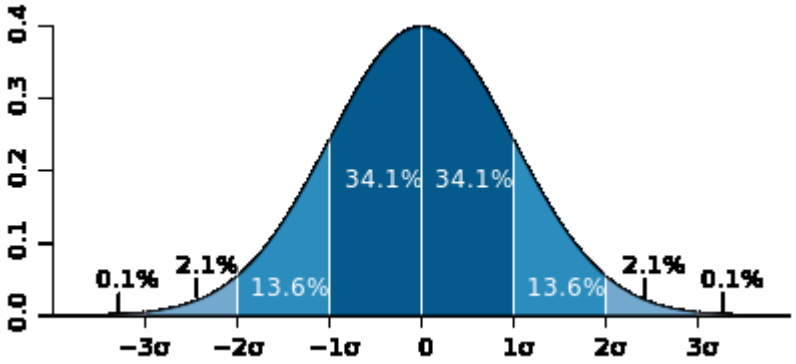
Założenia



Zakłada się, że wartości wyjściowe y_i są obarczone błędem ε

- $y_i = w_0 + w_1 x + \varepsilon_i$
- Błąd ε jest zmienną losową o rozkładzie normalnym (Gaussa): $\varepsilon \sim N(0, \sigma^2)$ - wartość oczekiwana błędu $\mathrm{E}[\varepsilon] = 0$

Założenia



Rozkład normalny

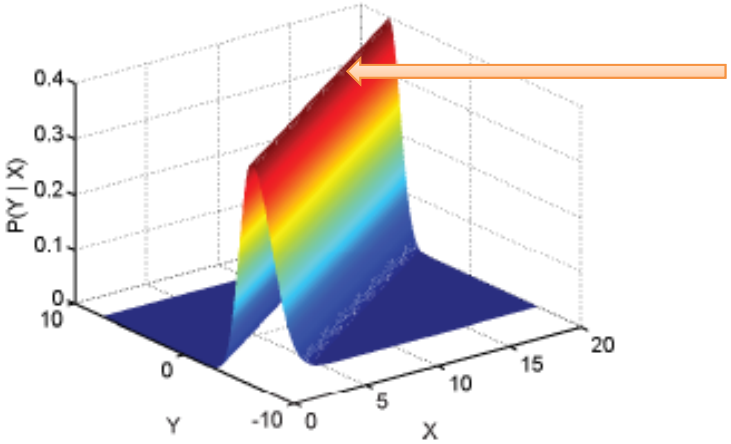
$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$

μ – wartość oczekiwana (średnia wartość)

σ^2 - wariancja

Prawdopodobieństwo, że y przyjmie określoną wartość jest zależne od x oraz parametrów modelu θ:

$$p(y|x, \theta) = N\big(\mu(x), \sigma^2(x)\big)$$



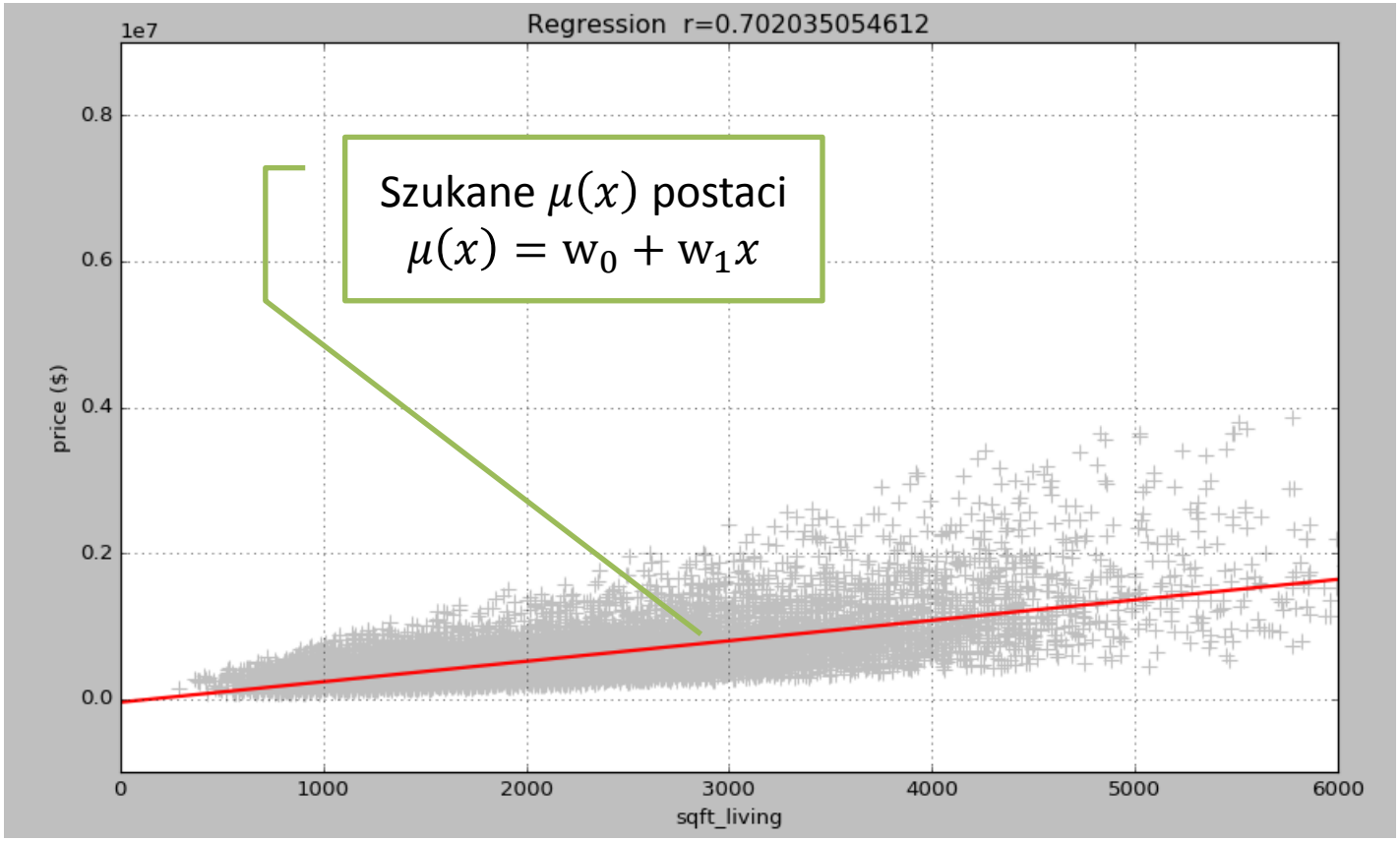
Poszukiwany jest przebieg funkcji po grzbiecie rozkładu

Parametry modelu $\theta = [w_0, w_1]$

[Kevin P. Murphy: Machine Learning A Probabilistic Perspective, MIT Press 2012]

Założenia

- Najczęściej dla uproszczenia zakłada się, że wariancja jest stała $\sigma^2(x) = \sigma^2$
- Poszukiwany jest przebieg $\mu(x)$ – wartości oczekiwanej y zależnej od x



Tu wariancja nie jest stała (cena nie będzie < 0)

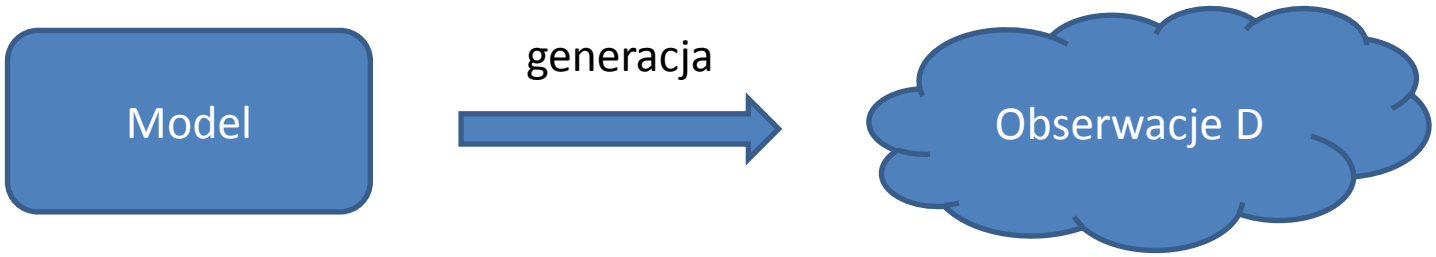
Zbiór: kc_house_data (2700 transakcji sprzedaży)

Maksymalizacja prawdopodobieństwa

W bardzo wielu algorytmach dobór parametrów modelu następuje poprzez maksymalizację lub minimalizację pewnej funkcji celu wyrażającej oczekiwane „dobre” parametry modelu θ.

$$\hat{\theta} = \arg max_{\theta}\, F(D, \theta)$$

Wartość funkcji celu $F(D, \theta)$ zależy od danych D i parametrów modelu θ. Dane są podane na wejściu, więc szukamy optymalnych wartości parametrów modelu $\hat{\theta}$.



Częste założenie:

- Dane pochodzą z pewnego modelu i generowane są z pewnym prawdopodobieństwem, którego rozkład zależy od parametrów modelu.
- Szukany jest model maksymalizujący prawdopodobieństwo D

Maksymalizacja prawdopodobieństwa

Zamierzamy zmaksymalizować prawdopodobieństwo zaobserwowania zbioru uczącego $\mathrm{D} = \{(x_i, y_i)\}_{i=1,m}$

- Prawdopodobieństwo pojedynczej obserwacji (x, y)

$$p(y|x, \theta) = N(\mu(x), \sigma^2)$$

- Prawdopodobieństwo zaobserwowania zbioru uczącego D:

$$p(D|\theta) = p(y_1|x_1, \theta) \cdot p(y_2|x_2, \theta) \cdot \ldots \cdot p(y_m|x_m, \theta)$$

- Szukamy parametrów modelu θ, które **maksymalizują** $p(D|\theta)$

$$\hat{\theta} = \arg max_\theta \, p(D|\theta)$$

- Maksymalizacja iloczynu jest kłopotliwa, więc zamiast wyznaczania miejsca ekstremum funkcji wyznacza się miejsce ekstremum jej logarytmu:

$$\hat{\theta} = \arg max_\theta \, \mathbf{ln}(p(D|\theta))$$

- Zauważmy, że jeśli $f(x) > 0$ to pochodna jej logarytmu wynosi:

$$\ln\big(f(x)\big)' = \frac{1}{f(x)} f'(x),$$

czyli miejsce minimum/maksimum (x_0: $f'(x_0) = 0$) nie ulegnie zmianie

RSS – suma kwadratów błędów

- Zagadnienie **maksymalizacji** zamieniamy na **minimalizację**: $\hat{\theta} = \arg min_\theta \; -\ln(p(D|\theta))$
- Szukamy minimum wyrażenia $NLL(\theta)$ – ang. negative log likelihood:

$$NLL(\theta) = -\ln\big(p(D|\theta)\big) = -\sum_{i=1}^{m} \ln(p(y_i|x_i,\theta))$$

Podstawiając:

$$p(y_i|x_i,\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y_i-\mu(x_i))^2}$$

otrzymujemy:

$$NLL(\theta) = \frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \mu(x_i))^2 + \frac{m}{2} ln(2\pi\sigma^2)$$

Wyrażenie

$$RSS = \sum_{i=1}^{m}(y_i - \mu(x_i))^2 = \sum_{i=1}^{m}(y_i - (w_0 + w_1 x_i))^2$$

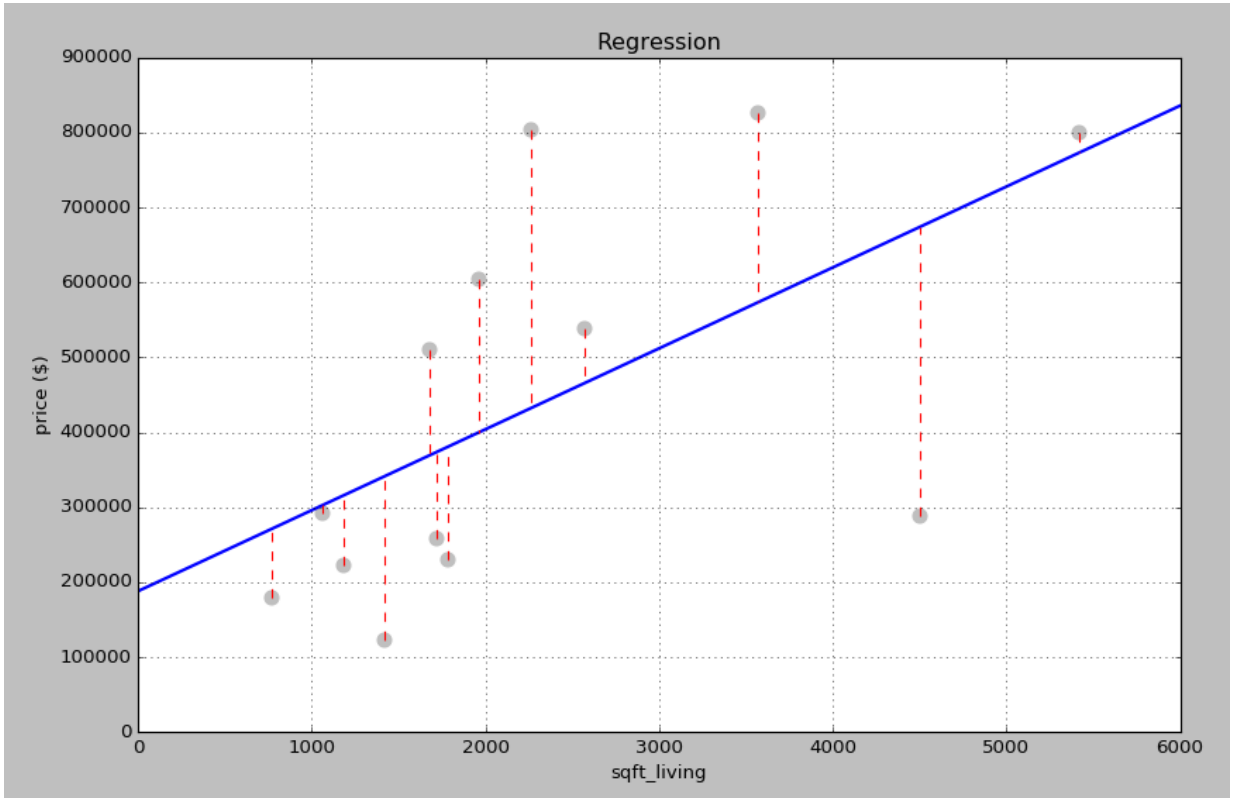
nazywane jest resztową sumą kwadratów błędów (RSS – residual sum of squares).

Celem regresji liniowej jest znalezienie takich parametrów θ, czyli wag: $w_0, w_1, \ldots, w_n$, które zminimalizują RSS.

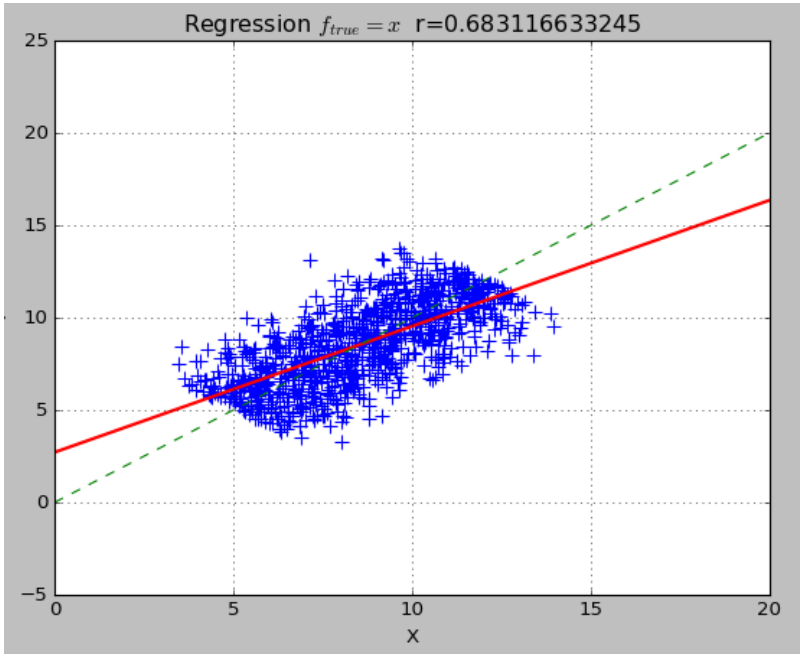
Sposób obliczania RSS

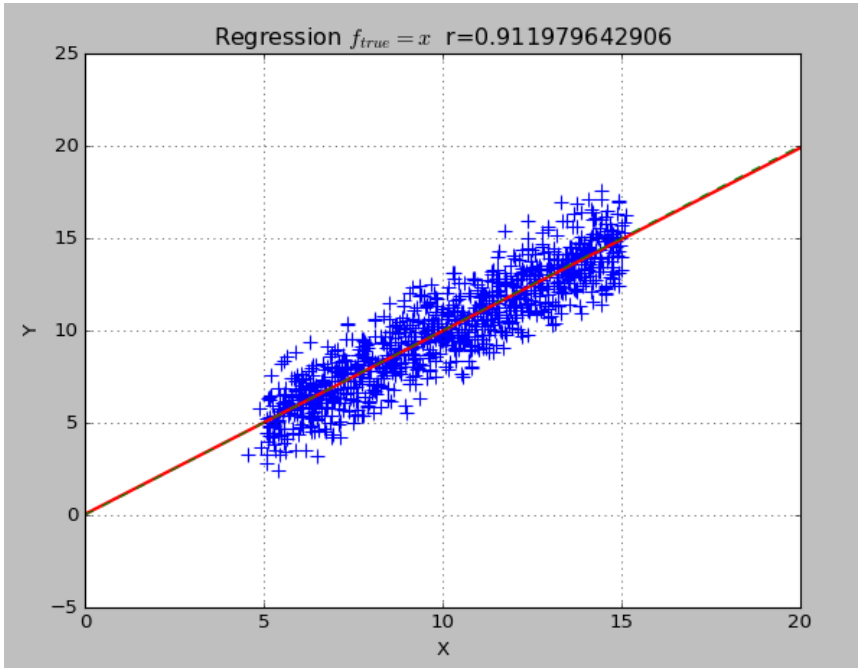
RSS oblicza się jako sumę kwadratów odległości punktów od prostej (hiperpłaszczyzny) w **kierunku pionowym**, a nie sumę odległości (kierunek prostopadły).

$$RSS(w_0, w_1) = \sum_{i=1}^{m} (y_i - (w_0 + w_1 x_i))^2$$



Sposób obliczania RSS





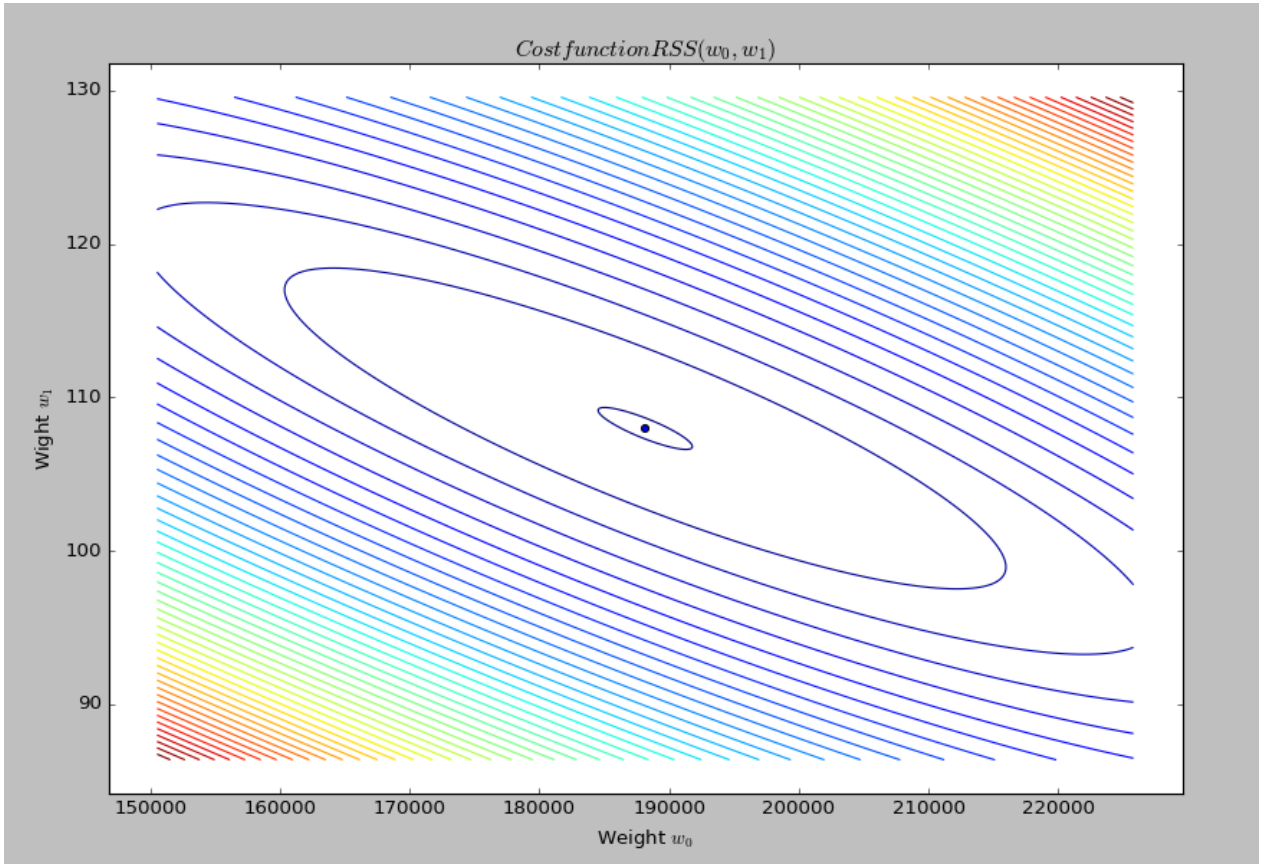
Zaobserwujmy różnicę:

- **po lewej** – wygenerowane losowo punkty tworzą prostokąt ulokowany wzdłuż przerywanej zielonej prostej f_{true}. Czerwona linia regresji jest odległa od f_{true}
- **po prawej** - f_{true} i linia regresji przechodzą przez środek równoległoboku (zgodność)

Minimalizacja RSS

Szukamy optymalnych parametrów:

$$[\widehat{w}_0, \widehat{w}_1] = \arg min_{w_0,w_1} \sum_{i=1}^{m} (y_i - (w_0 + w_1 x_i))^2$$



Wyraz wolny

Jeżeli $x \in \mathrm{R}^n$ równanie regresji ma postać:

$$y(x) = w_0 + \sum_{i=1}^{n} w_i \; x^i$$

Częstym zabiegiem jest pozbycie się wyrazu wolnego (bias) przez zmianę wymiaru na $n + 1$.

$$[x^1, x^2, \dots, x^n] \quad \Rightarrow \quad [1, x^1, x^2, \dots, x^n]$$

Wtedy równanie regresji ma postać:

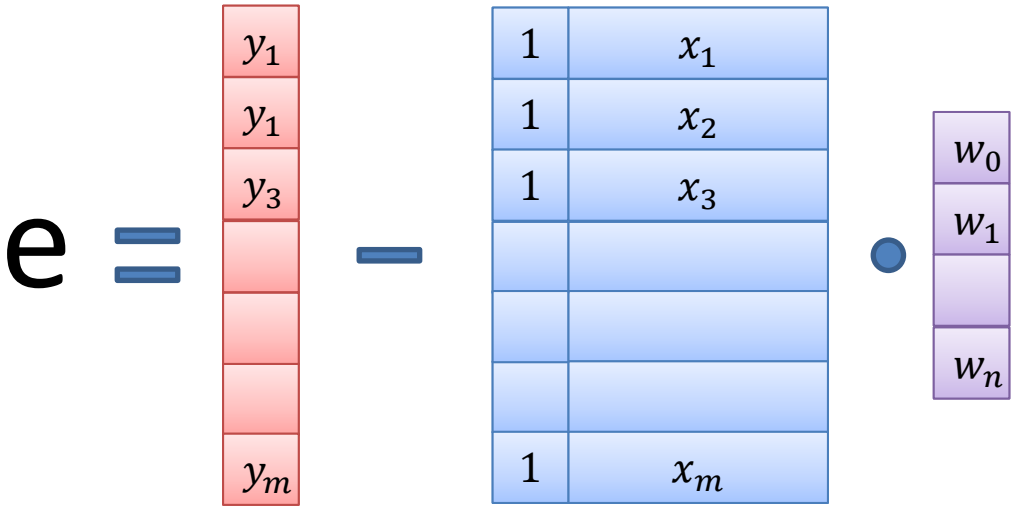
$$y(x) = \; w^T x,$$

gdzie $w^T = \; [w_0, w_1, \dots, w_n]$

Oczywiście także $w^T x = x^{\mathrm{T}} w$

Postać macierzowa

- Elementy zbioru uczącego mogą być przedstawione w postaci:
  - macierzy X o rozmiarach $m \times n + 1$
  - m-wymiarowego wektora y.
- Macierz X ma dodatkowe jedynki w pierwszej kolumnie.
- Niech $w \in \mathbf{R}^{n+1}$ będzie wektorem wag.



Wektor błędu $e \in \mathrm{R}^m$ można więc przedstawić jako: $e = y - \mathrm{X}w$

$$RSS(w) = \sum_{i=1}^{m} e_i \cdot e_i = e^T e = (y - \mathrm{X}w)^T (y - \mathrm{X}w)$$

Minimalizacja RSS

$$RSS(w) = (y - \mathrm{X}w)^T(y - \mathrm{X}w) = y^T y - y^T(Xw) - (Xw)^T y + (Xw)^T(Xw)$$
$$= y^T y - 2w^T X^T y + w^T X^T Xw$$

Dla przypomnienia: $(Xw)^T = w^T X^T$

Aby obliczyć minimum $RSS(w)$ względem należy obliczyć pochodną i przyrównać do 0

$$\frac{d}{dw}(y^T y - 2w^T X^T y + w^T X^T Xw) = -2X^T y + 2X^T Xw$$
$$-2X^T y + 2X^T Xw = 0$$

stąd otrzymujemy **równanie normalne**:

$$X^T Xw = X^T y$$

oraz zależność:

$$w = (X^T X)^{-1} X^T y$$

Równanie normalne jest równaniem liniowym i w może być wyznaczone przez eliminację Gaussa lub obliczenie odwrotności macierzy $(X^T X)^{-1}$.

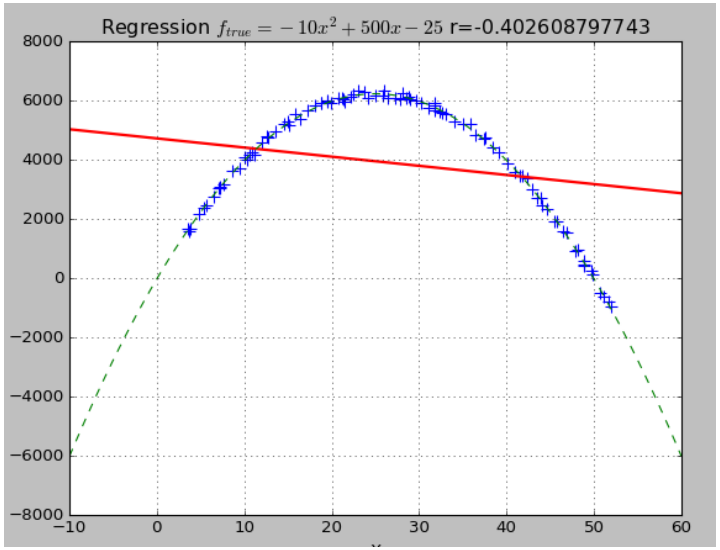
Równanie normalne

$$X^T X w = X^T y$$

- Macierz $X^T X$ ma wymiar: $n + 1 \times \mathrm{n} + 1$
- Czynnik $X^T y$ jest wektorem $n + 1$ wymiarowym
- Problem może zostać sprowadzony do odwracania macierzy $X^T X$.
- Macierz odwrotna istnieje, jeśli macierz $X^T X$ jest nieosobliwa, tzn.
  - liczba niezależnych liniowo obserwacji m musi być większa niż liczba cech ($m > n + 1$)
  - kolumny nie są liniowo zależne, czyli atrybuty nie są silnie skorelowane (np. powierzchnia w stopach i metrach kwadratowych)

Ograniczenia liniowej regresji

- Rzeczywiste relacje pomiędzy X i y mogą być nieliniowe. Stąd generalizacja do modeli nieliniowych.
- Złożoność jest rzędu: $O(m\, n^2 + n^3)$. Czyli np. dla n=30 atrybututów i m=10000 liczba operacji jest b. duża
- W przypadku korelacji pomiędzy zmiennymi macierz $X^T X$ może być źle uwarunkowana (niestabilność numeryczna)



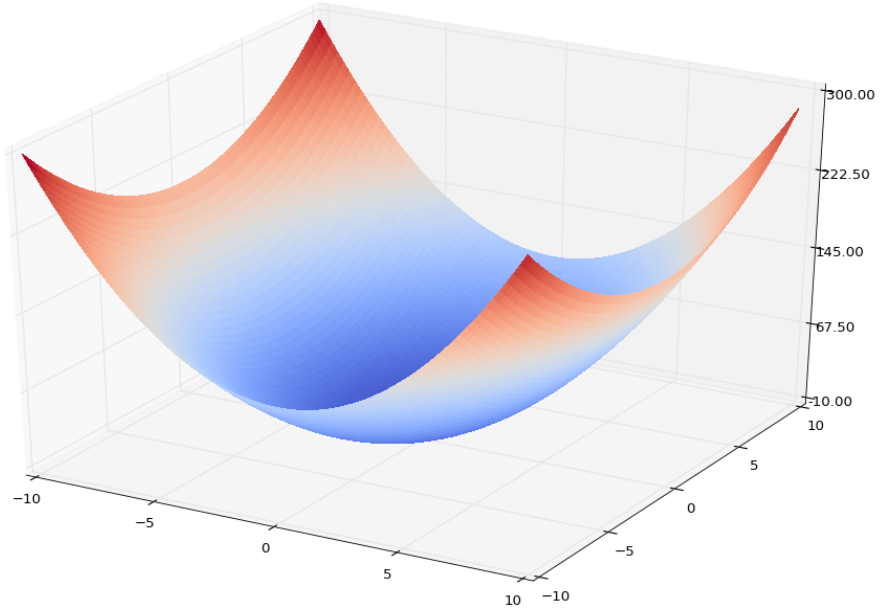
- Wszystkie zmienne (atrybuty) są użyte w modelu, a może zdarzyć się, że jedynie ich podzbiór jest istotny i ma rzeczywisty wpływ na wartość wyjściową.

Metoda gradientu prostego

- W praktyce współczynniki regresji są najczęściej wyznaczane za pomocą **gradientowych** metod optymalizacji.

$$\hat{w} = \arg min_w RSS(w), \text{ gdzie } RSS(w) = \sum_{i=1}^{m} (y_i - w^T x_i)^2$$

- RSS jest wypukłą funkcją wag, stąd globalne minimum istnieje i da się je wyznaczyć za pomocą metod gradientowych.

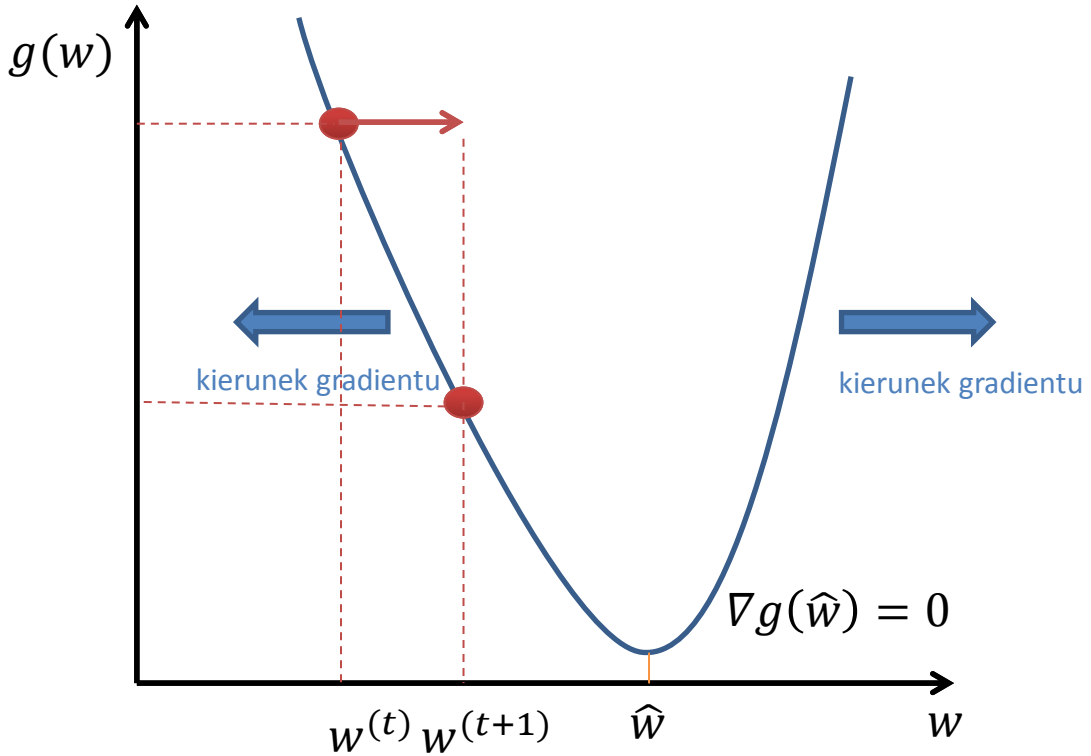


Metoda gradientu prostego

Dla uproszczenia oznaczmy $RSS(w)$ jako $g\ (w)$

- Gradient $\nabla g(w) = [\frac{\partial g}{\partial w_0}, \frac{\partial g}{\partial w_1}, \dots, \frac{\partial g}{\partial w_n}]$ jest $n+1$ wymiarowym wektorem pochodnych cząstkowych w kierunkach $w_0, w_1, \dots, w_n$

- W trakcie optymalizacji osiągnięto punkt $w^{(t)}$
- **Małe** przemieszczenie w kierunku $-\nabla g\left(w^{(t)}\right)$ zazwyczaj przynosi poprawę wartości funkcji celu.



Metoda gradientu prostego

Algorytm

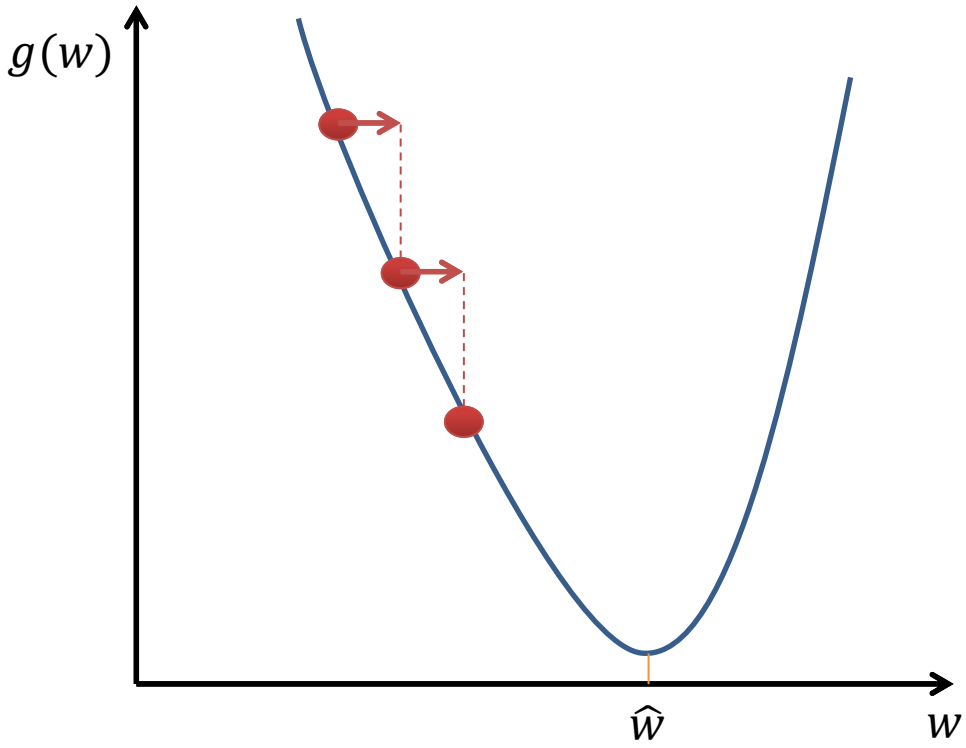
wybierz $w^{(0)}$
$t = 0$
while ! kryterium_stopu:
$w^{(t+1)} = w^{(t)} - \gamma \nabla g(w^{(t)})$
$t = t + 1$

- Współczynnik γ jest małą liczbą
- Typowe kryteria stopu:
  - Maksymalna liczba iteracji: $t = maxiter$
  - Gradient bliski zeru: $\left|\nabla g\left(w^{(t)}\right)\right| < \varepsilon$
  - Brak poprawy funkcji celu: $d(w^{(t+1)} - w^{(t)}) < \varepsilon$

Wielkość kroku

Wielkość kroku γ ma wpływ na zbieżność algorytmu:

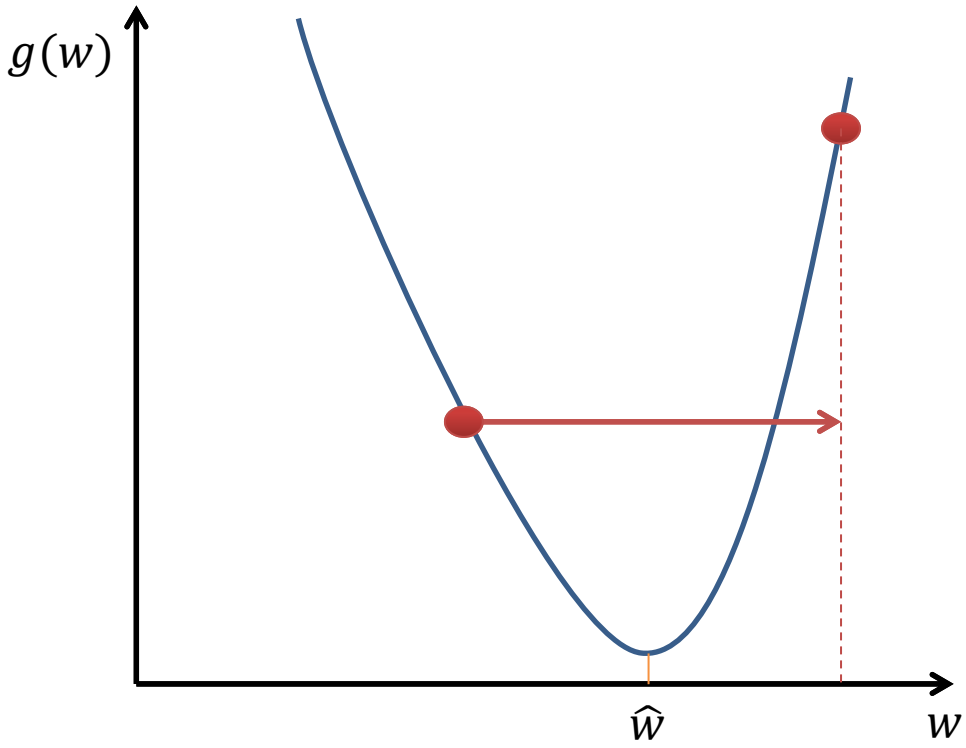
- krok za mały $\rightarrow$ zbyt wolna zbieżność
- krok za duży $\rightarrow$ możliwy wzrost wartości funkcji celu



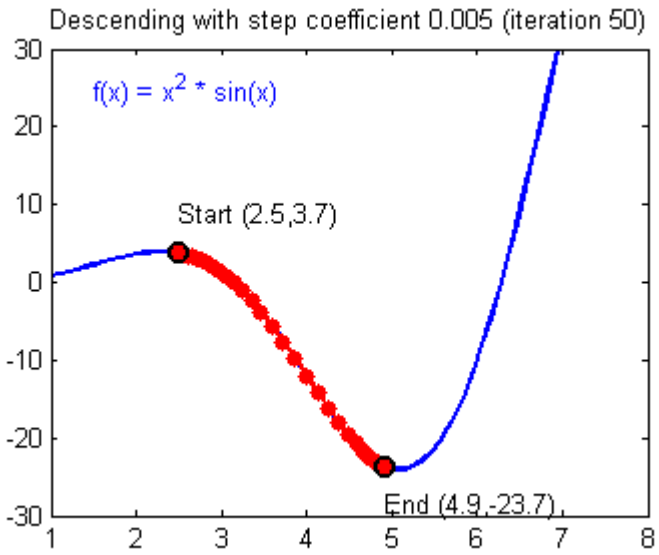
Wielkość kroku

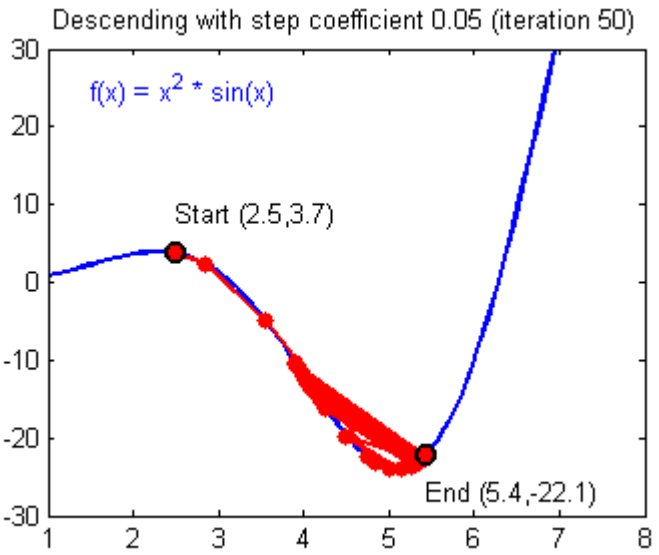
Wielkość kroku γ ma wpływ na zbieżność algorytmu:

- krok za mały $\rightarrow$ zbyt wolna zbieżność
- krok za duży $\rightarrow$ możliwy wzrost wartości funkcji celu



Wielkość kroku - porównanie



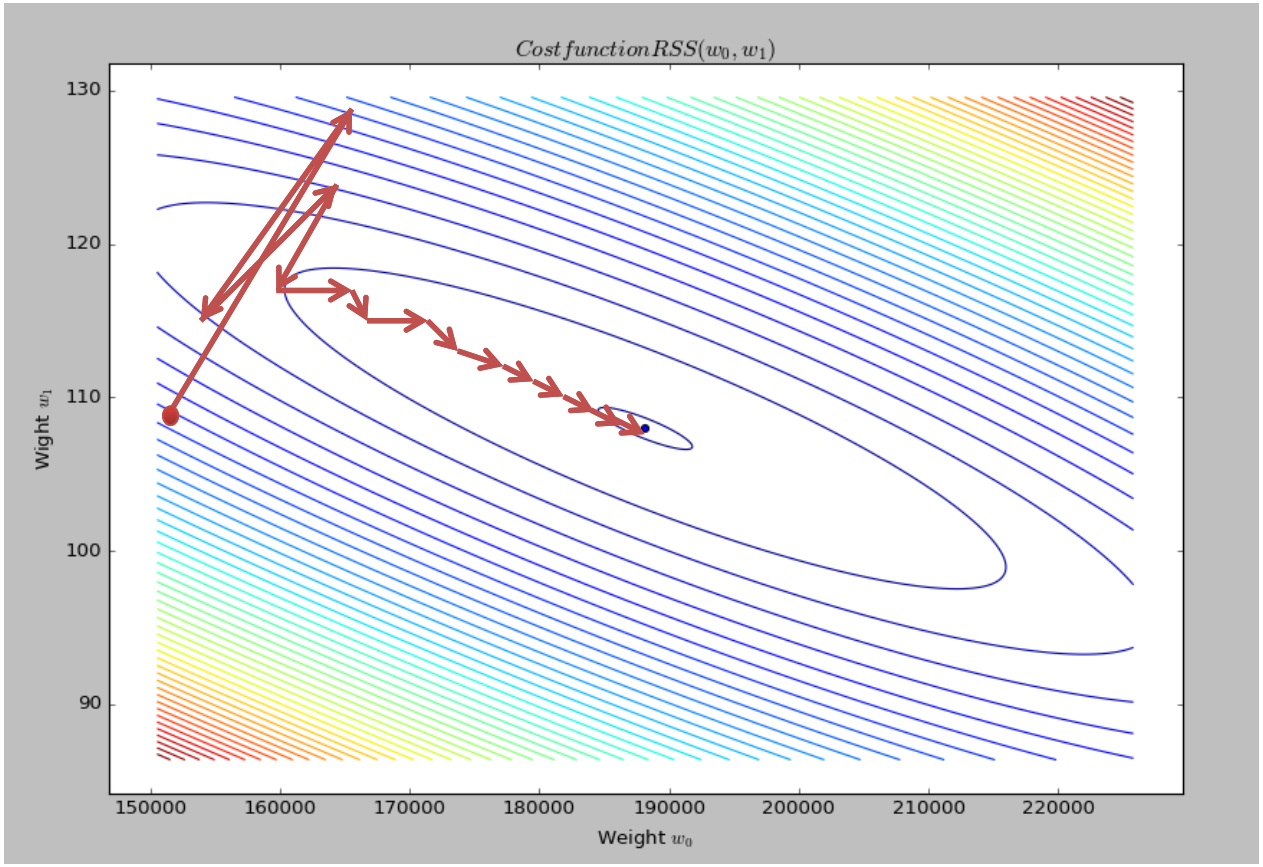


https://www.cs.toronto.edu/~frossard/post/linear_regression/

Zmienny współczynnik kroku

W praktycznych implementacjach wartość współczynnika γ maleje wraz z numerem iteracji, np.:

- $\gamma(t) = \frac{\gamma(0)}{t}$
- $\gamma(t) = \frac{\gamma(0)}{\sqrt{t}}$



Gradient RSS

$$RSS(w) = \sum_{i=1}^{m} (y_i - w^T x_i)^2 = \sum_{i=1}^{m} \left(y_i - \sum_{j=0}^{n} x_i^j w_j \right)^2$$

Uwaga: tu x_i^j oznacza j-ty element i-tego wektora, czyli $x[i][j]$

Pochodna cząstkowa względem w_j:

$$\frac{\partial RSS(w)}{\partial w_j} = \sum_{i=1}^{m} 2(y_i - w^T x_i)(-x_i^j) = -2 \sum_{i=1}^{m} (y_i - w^T x_i) x_i^j$$

- Oznaczmy: $e_i = (y_i - w^T x_i)$. Pochodna $\frac{\partial}{\partial w_j} (e_i)^2 = 2 e_i \frac{\partial}{\partial w_j} e_i$
- Pochodna $\frac{\partial}{\partial w_j} e_i = -x_i^j$
- Pochodna sumy $i = 1, m$ składników jest równa sumie pochodnych
- Gradient:

$$\nabla RSS(w) = [-2 \sum_{i=1}^{m} (y_i - w^T x_i) \cdot 1 \, , -2 \sum_{i=1}^{m} (y_i - w^T x_i) x_i^1 \, , \ldots , -2 \sum_{i=1}^{m} (y_i - w^T x_i) x_i^n]$$

- Wartość $x_i^0 = 1$

Obliczanie gradientu RSS

$$\frac{\partial RSS(w)}{\partial w_j} = -2\sum_{i=1}^{m}(y_i - w^T x_i)x_i^j$$

- Dla i-tej obserwacji błąd mnożony jest przez wartość j-tego atrybutu
- Złożoność obliczeniowa $O(n^2 m)$
- Należy przewidzieć kilkadziesiąt iteracji
- Jeśli obserwacje $(x_i, y_i)_{i=1,m}$ są umieszczone w bazie danych, wielokrotna iteracja może być kosztowna (np. $m = 1000000$).
- Zamiast:

$$\nabla RSS(w) =$$
$$[-2\sum_{i=1}^{m}(y_i - w^T x_i)\cdot 1, -2\sum_{i=1}^{m}(y_i - w^T x_i)x_i^1, \dots, -2\sum_{i=1}^{m}(y_i - w^T x_i)x_i^n]$$

poszczególne składowe mogą być liczone równolegle:

$$\nabla RSS(w) = \sum_{i=1}^{m}[-2(y_i - w^T x_i), -2(y_i - w^T x_i)x_i^1, \dots, -2(y_i - w^T x_i)x_i^n]$$

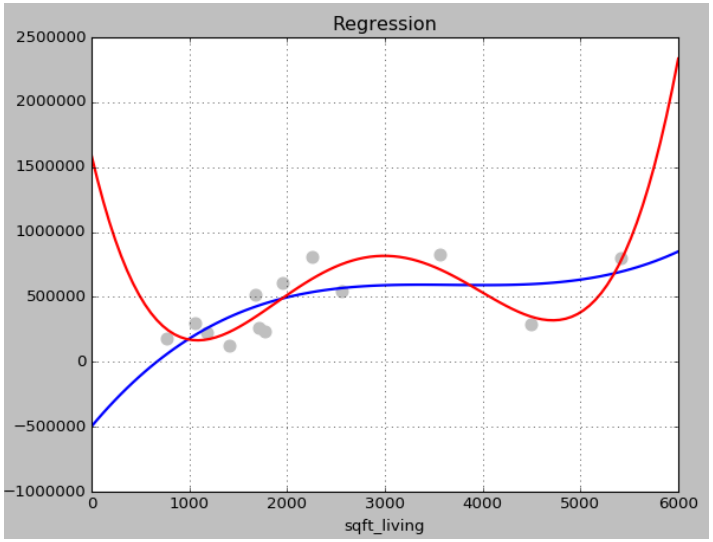
- Warianty algorytmu:
  - **stochastic gradient descent** (wykonanie kroku optymalizacji po odczycie danych)
  - **coordinate descent** (optymalizacja w kierunku jednej ze składowych):

$$\nabla RSS(w) = [const \dots, -2\sum_{i=1}^{m}(y_i - w^T x_i)x_i^j, \dots const]$$

Modele nieliniowe

- Regresja liniowa może być zastosowana do wyznaczenia parametrów modeli będących nieliniowymi funkcjami.
- Często przybliżana funkcja jest wielomianem – **regresja wielomianowa**
- Przykład dla jednej zmiennej:

$$f(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3 \ldots$$



Niebieska krzywa

$$f(x) = -497430.99 + 9.00\text{e}{+02}\, x - 2.45\text{e}{-01}\, x^2 + 2.22\text{e}{-05}\, x^3$$

Czerwona krzywa

$$f(x) = 1589582.0 - 3.21\text{e}{+03}\, x + 2.35\, x^2 - 6.15379704\text{e}{-04}\, x^3 + 5.24\text{e}{-08}\, x^4$$

Modele wielomianowe dla danych wielowymiarowych

Dla n-wymiarowych danych postać wielomianowa może obejmować składniki kolejnych stopni:

- n - pierwszego stopnia
- $n(n+1)/2$ - drugiego stopnia
- $n(n+1)(n+2)/6$ - trzeciego stopnia

$$f(x) = w_0 + \sum w_j x_j + \sum w_{jk} x_j x_k + \sum w_{jkl} x_j x_k \, x_l \ldots$$

- Dla n=30 wielomian stopnia 3 będzie miał potencjalnie 1+30+30*31/2+30*31*32/2=5456 czynników
- Problemem jest:
  - wybór modelu (które kombinacje zmiennych brać pod uwagę)
  - wyznaczenie/przechowywanie/dostęp do macierzy cech X (5456 kolumn?)

Cechy

- W ogólnym przypadku model jest kombinacją liniową dowolnych nieliniowych funkcji danych wejściowych

$$f(x) = w_0 + \sum_{i=1}^{N} w_i h_i(x)$$

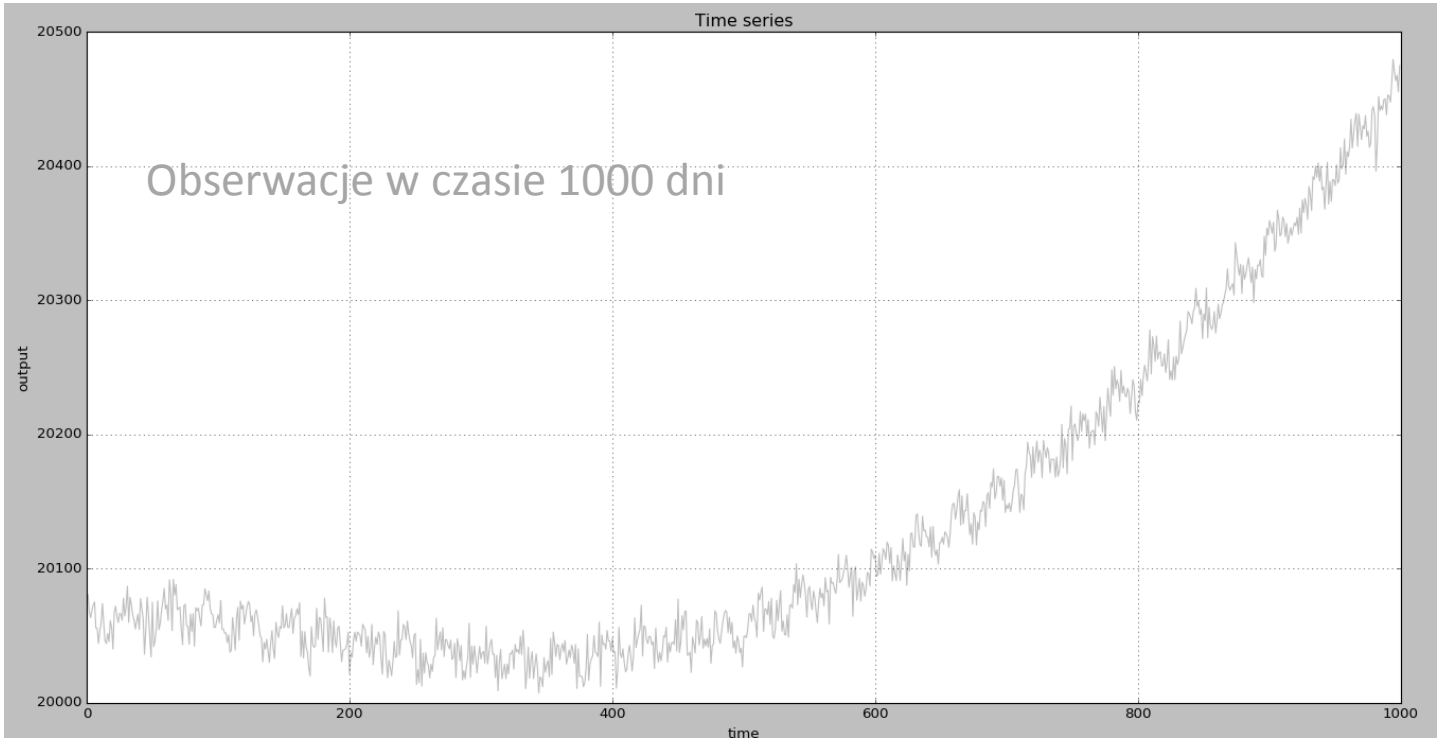
- Funkcja $h_i(x)$: $x \in \mathbf{R}^n \rightarrow \mathbf{R}$ nazywana jest cechą.
- W szczególności może zachodzić $N = n$ oraz $h_i(x) = x[i]$ (i-tą cechą jest i-ty składnik wektora x
- Niezależnie od przebiegu użytych cech – wyznaczane sa współczynniki kombinacji liniowej (wagi w) minimalizujące $RSS(w)$.
- **Model matematyczny i algorytmy pozostają bez zmian**
- Dodając cechę 1 (o indeksie 0) możemy zapisać funkcję, jako:

$$f(x) = w^T h(x)$$

  - w to $N + 1$ wymiarowy wektor wag,
  - $h(x) = [1, h_1(x), \ldots, h_N(x)]$ to wektor cech

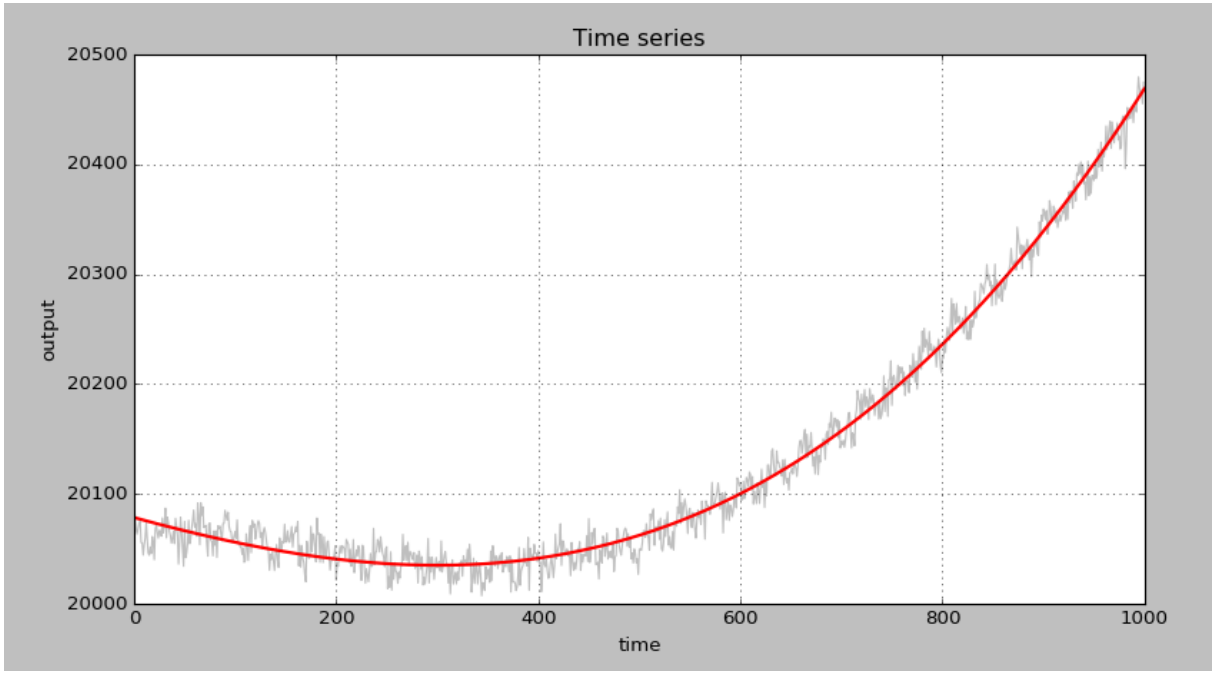
Szeregi czasowe

- Wiele problemów ma charakter przebiegów czasowych, np.
  - zmiana cen w czasie
  - zmiana temperatury
  - zmiana zużycia energii
- Często te przebiegi są kombinacją ogólnych tendencji oraz zmian sezonowych

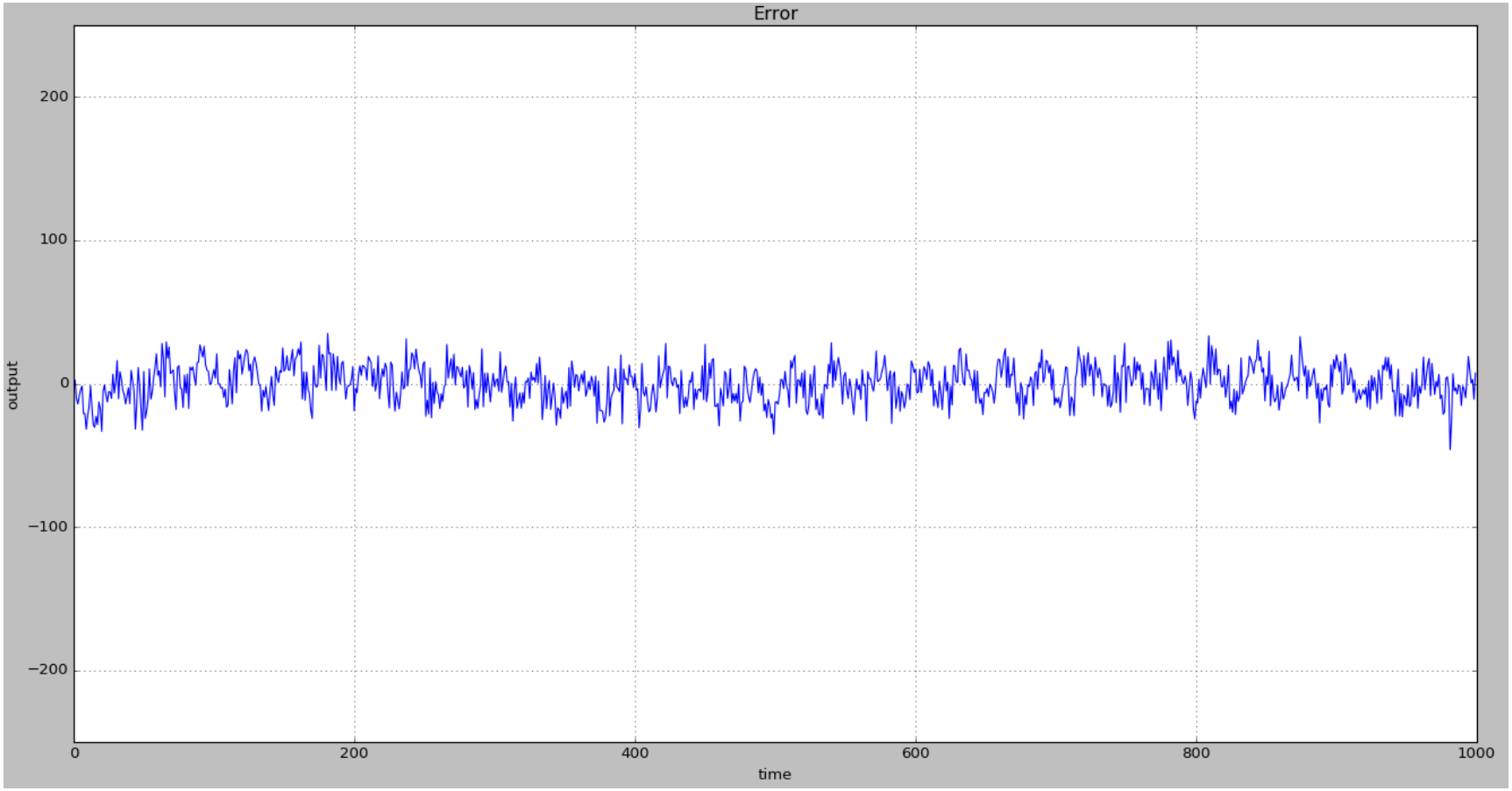


Szeregi czasowe

- Za pomocą regresji można znaleźć krzywą modelującą ogólną tendencję: $y = f(x)$
- Model taki nie uwzględnia przebiegów sezonowych (traktuje je jak błąd): $y_i = f(x_i) + \varepsilon_i$



Wykres błędu



Widoczne 30 dniowe sezonowe fluktuacje

Model

- Model uwzględniający sezonowość ma postać:

$$y = w_0 + w_1 t + w_2 t^2 + w_3 t^3 + w_4 \sin\left(\frac{2\pi}{30} t + \Phi\right)$$

  - Wybrano krzywą 3 stopnia
  - Okres wynosi 30 dni: $\frac{2\pi}{30}$
  - Nieznane przesunięcie fazowe: Φ
- Faza Φ musi zostać przekształcona w cechę. Z zależności: $\sin(a + b) = sin\, a\, cos\, b + \cos a \sin b$ mamy:

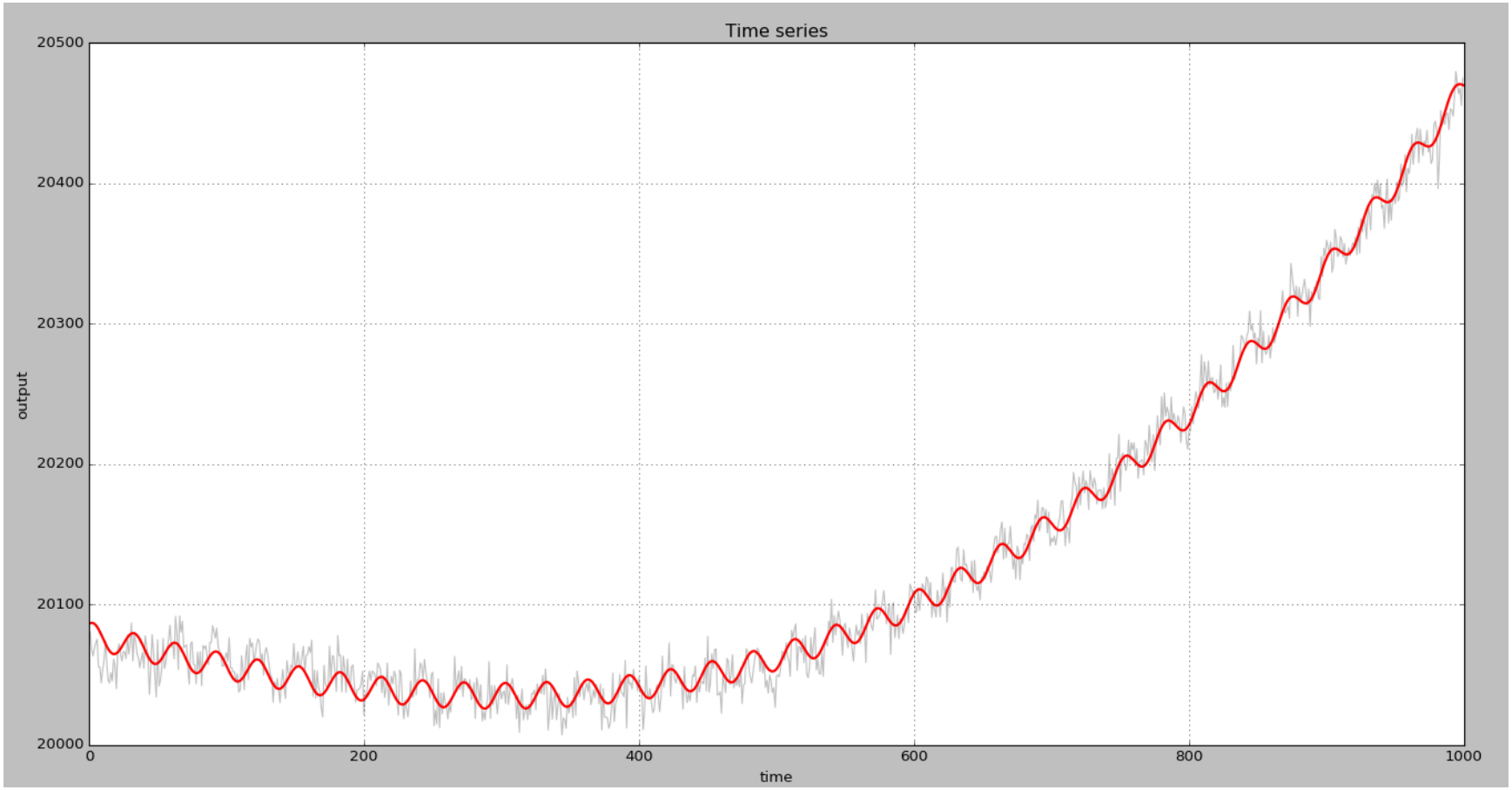
$$\sin\left(\frac{2\pi}{30} t + \Phi\right) = \sin\left(\frac{2\pi}{30} t\right) \cos(\Phi) + \cos\left(\frac{2\pi}{30} t\right) \sin(\Phi)$$

Dla ustalonego Φ czynniki $\cos(\Phi)$ i $\sin(\Phi)$ będą stałe i staną się częścią wag.

Równanie opisujące model musi jednak uwzględnić $\cos\left(\frac{2\pi}{30} t\right)$:

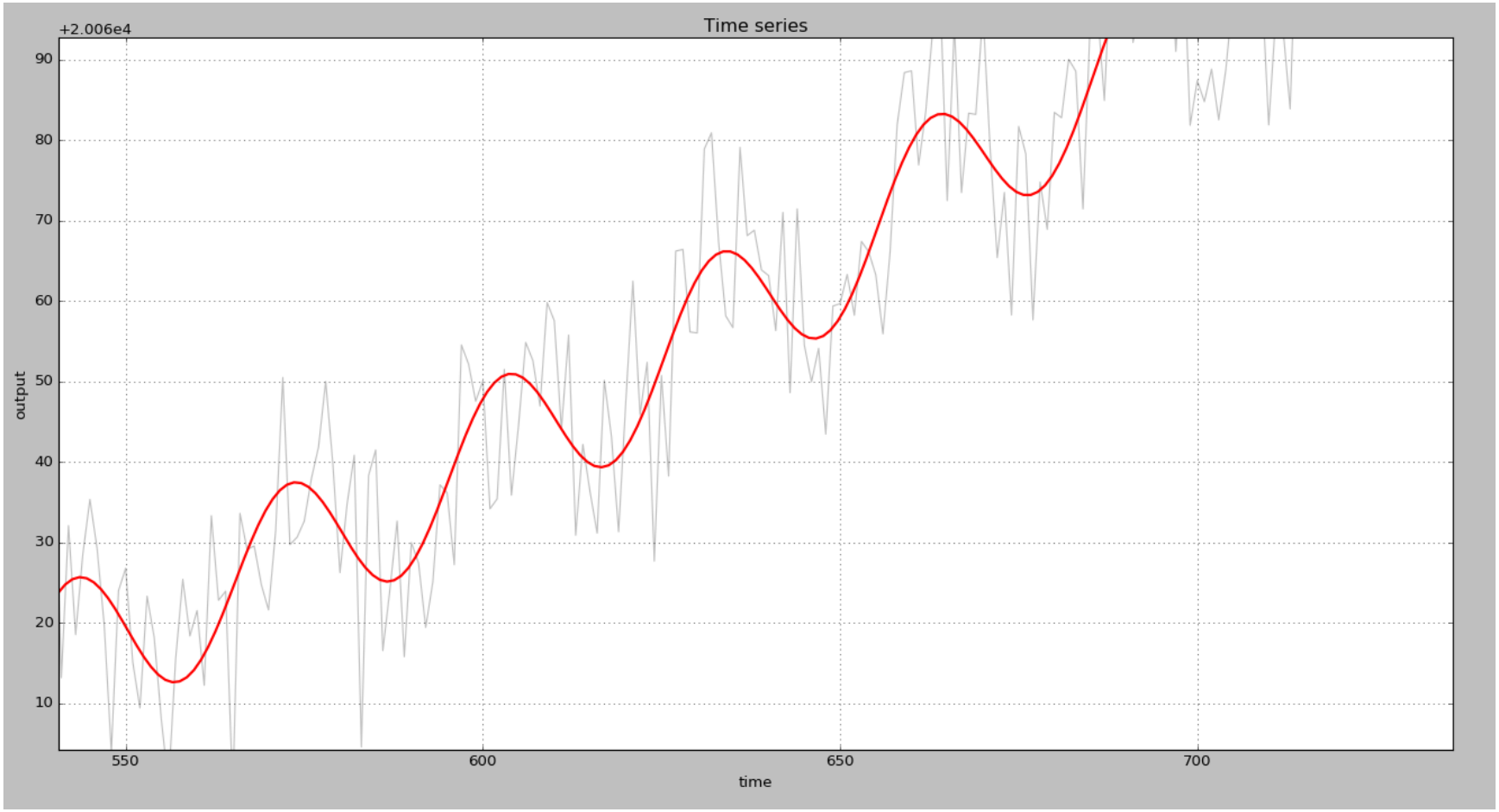
$$y = w_0 + w_1 t + w_2 t^2 + w_3 t^3 + w_4 \sin\left(\frac{2\pi}{30} t\right) + w_5 \cos\left(\frac{2\pi}{30} t\right)$$

Wizualizacja modelu



$$y = 20078.47 - 2.52e-01 \cdot t + 2.39e-04 \cdot t^2 + 4.04e-07 \cdot t^3 + 5.05\sin(\frac{2\pi}{30}t) + 7.67\cos(\frac{2\pi}{30}t)$$

Wizualizacja modelu



Dane przebiegu zostały wygenerowane jako

$$at^2 + bt + c - d\exp(-t) + e\sin\left(\frac{2\pi}{30}t + f\right) + g\,N(0,1)$$