Dopasowanie dowolnej funkcji do danych pomiarowych. Część 2.

Metoda mieszana Levenberga-Marquardta

Levenberg zaproponował, żeby skuteczny algorytm znajdowania minimum χ^2 stosował metodę gradientu z dala od minimum i metodę rozwinięcia w jego pobliżu.
Rozwinięcie funkcji χ^2 w szereg potęgowy prowadziło do równania macierzowego:

$$\boldsymbol{\beta} = \boldsymbol{\delta a}\,\boldsymbol{\alpha}\,,\ \left(\beta_i = \sum_j \alpha_{ij}\delta a_j\right)$$

którego rozwiązanie możemy zapisać jako:

$$\boldsymbol{\delta a} = \boldsymbol{\beta}\,\boldsymbol{\alpha}^{-1}.$$

Biorąc pod uwagę definicje wektorów $\boldsymbol{\delta a}$ i $\boldsymbol{\beta}$

$$\boldsymbol{\delta a} = \mathbf{a}_{k+1} - \mathbf{a}_k$$

$$\beta_i = -\frac{1}{2}\frac{\partial \chi_0^2}{\partial a_i}$$

możemy zapisać

$$\mathbf{a}_{k+1} = \mathbf{a}_k - \left[\frac{1}{2}\vec{\nabla}\chi^2(\mathbf{a}_k)\right]\boldsymbol{\alpha}^{-1}$$

Ten związek podaje wartość wektora parametrów $\mathbf{a}_{k+1}$ w następnym $(k+1)$ kroku.

Metoda największego spadku (gradientu) prowadziła natomiast do następującej wartości $\mathbf{a}_{k+1}$:

$$\mathbf{a}_{k+1} = \mathbf{a}_k - const\cdot\vec{\nabla}\chi^2(\mathbf{a}_k)$$

Jednym z problemów przy stosowaniu tej metody jest odpowiedni dobór stałej, tzn. jak duży powinien być krok w kierunku największego spadku. Marquardt zauważył po pierwsze, że macierz krzywizny powinna zawierać w sobie tę informacje w jakiejś postaci. Wymiary poszczególnych elementów wektora gradientu (albo $\boldsymbol{\beta}$) są takie jak wymiary odwrotności odpowiednich parametrów $1/a_i$. W przypadku jednowymiarowym, stała powinna mieć zatem wymiar taki jak a_i^2. W macierzy $\boldsymbol{\alpha}$ nadającym się kandydatem jest odwrotność elementu diagonalnego α_{ii}^{-1}. Zatem odwrotność elementu diagonalnego może

spełniać rolę czynnika skalującego wielkość kroku. Jednak sam czynnik skalujący może mieć wartość za dużą, więc podzielmy go przez bezwymiarowy czynnik λ, któremu w razie potrzeby można nadać dużą wartość $\lambda >> 1$, żeby skrócić kolejny krok. W tej sytuacji związek między wielkością kroku w kierunku a_i a składową wektora $\boldsymbol{\beta}$ ($\boldsymbol{\beta} = -1/2\, \vec{\nabla}\chi^2$) można zapisać:

$$\delta a_i = \frac{1}{\lambda \alpha_{ii}} \beta_i \text{ albo } \beta_i = \lambda \alpha_{ii} \delta a_i$$

Kolejną rzeczą zauważoną przez Marquardta było, że oba podejścia da się praktycznie połączyć zmieniając definicję elementów diagonalnych macierzy $\boldsymbol{\alpha}$:

$$\alpha'_{ii} \equiv \alpha_{ii}(1+\lambda)$$
$$\alpha'_{ij} \equiv \alpha_{ij} \qquad (i \neq j)$$

i zapisując odpowiednie równanie w postaci

$$\boldsymbol{\beta} = \boldsymbol{\delta a}\, \boldsymbol{\alpha}', \left(\beta_i = \sum_j \alpha'_{ij}\, \delta a_j \right)$$

Zmieniając wartość czynnika λ od wartości bardzo dużych do zera przechodzimy w sposób ciągły od metody największego spadku (gradientu) do metody rozwinięcia funkcji χ^2.

Algorytm Marquardta przedstawia się następująco:

- Wybierz początkowe wartości parametrów $\mathbf{a}$ i oblicz wartość $\chi^2(\mathbf{a})$
- Wybierz jakąś rozsądną wartość λ, np. $\lambda = 0{,}001$
- (*) Rozwiąż odpowiednie równanie macierzowe (układ równań liniowych) znajdując $\boldsymbol{\delta a} = \boldsymbol{\beta}\, \boldsymbol{\alpha}'^{-1}$ i oblicz $\chi^2(\mathbf{a} + \boldsymbol{\delta a})$
- Jeżeli $\chi^2(\mathbf{a} + \boldsymbol{\delta a}) \geq \chi^2(\mathbf{a})$, to zwiększ λ o czynnik 10 i wróć do (*)
- Jeżeli $\chi^2(\mathbf{a} + \boldsymbol{\delta a}) < \chi^2(\mathbf{a})$, to zmniejsz λ o czynnik 10, uaktualnij wartości parametrów $\mathbf{a} \leftarrow \mathbf{a} + \boldsymbol{\delta a}$ i wróć do (*)
- Jeżeli w kolejnych krokach wartość χ^2 zmniejsza się w sposób mało znaczący (np. $\Delta\chi^2 < 0{,}1$), to kończymy proces iteracji, ustalamy $\lambda = 0$ i obliczamy macierz kowariancji $\boldsymbol{\varepsilon} = \boldsymbol{\alpha}'^{-1}$ znalezionych parametrów $\mathbf{a}$.

Praktyczna uwaga dotycząca obliczania wartości elementów macierzy krzywizny $\boldsymbol{\alpha}$.

$$\alpha_{kl} \equiv \frac{1}{2}\frac{\partial^2 \chi^2}{\partial a_k \partial a_l}$$

Z uwagi na definicję funkcji χ^2

$$\frac{\partial^2 \chi^2}{\partial a_k \partial a_l} = 2\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[\frac{\partial y(x_i;\mathbf{a})}{\partial a_k}\frac{\partial y(x_i;\mathbf{a})}{\partial a_l} - [y_i - y(x_i;\mathbf{a})]\frac{\partial^2 y(x_i;\mathbf{a})}{\partial a_k \partial a_l}\right]$$

elementy macierzy $\boldsymbol{\alpha}$ powinny być równe:

$$\alpha_{kl} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[\frac{\partial y(x_i;\mathbf{a})}{\partial a_k}\frac{\partial y(x_i;\mathbf{a})}{\partial a_l} - [y_i - y(x_i;\mathbf{a})]\frac{\partial^2 y(x_i;\mathbf{a})}{\partial a_k \partial a_l}\right].$$

W dużej liczbie zastosowań praktycznych wartości te oblicza się z pominięciem drugich pochodnych:

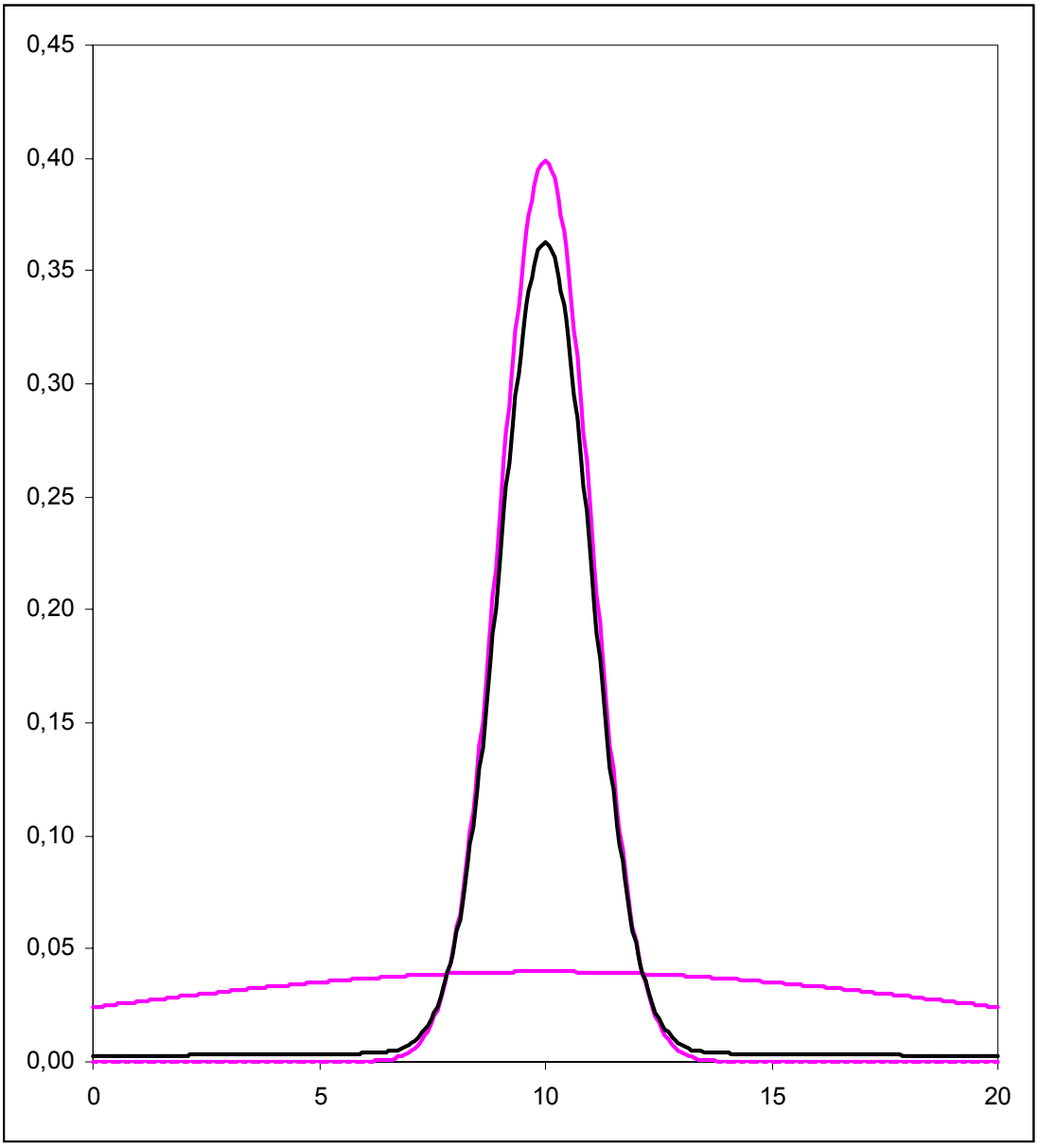
$$\alpha_{kl} = \sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[\frac{\partial y(x_i;\mathbf{a})}{\partial a_k}\frac{\partial y(x_i;\mathbf{a})}{\partial a_l}\right]$$

Za takim rozwiązaniem przemawia kilka względów. Pominięcie drugich pochodnych w definicji upraszcza obliczenia i skraca czas realizacji algorytmu. Zwykle krzywizna funkcji zmienia się wolniej niż sama funkcja, co oznacza, że wartości drugich pochodnych są mniejsze od wartości pierwszych pochodnych. Można się spodziewać, że (dla dobrych $\mathbf{a}$) wartości $\left[\frac{y_i - y(x_i;\mathbf{a})}{\sigma_i^2}\right]$ mają rozkład normalny $N(0,1)$, co oznacza, że spodziewana wartość sumy $\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[[y_i - y(x_i;\mathbf{a})]\frac{\partial^2 y(x_i;\mathbf{a})}{\partial a_k \partial a_l}\right]$ powinna być bliska zeru, a w każdym razie mała w porównaniu z pierwszym składnikiem $\sum_{i=1}^{n} \frac{1}{\sigma_i^2}\left[\frac{\partial y(x_i;\mathbf{a})}{\partial a_k}\frac{\partial y(x_i;\mathbf{a})}{\partial a_l}\right]$.

Okazuje się dodatkowo, że jeśli model (wartości $\mathbf{a}$) jest źle dobrany albo dane zawierają wartości odstające, to uwzględnienie drugich pochodnych może prowadzić do destabilizacji procesu iteracji (braku zbieżności).

Metody estymacji parametrów odporne na odstępstwa od rozkładu normalnego.

Załóżmy, że błędy pomiarowe podlegają pewnemu rozkładowi, który nie jest normalny.



Niech

$$p\left(y_i, y(x_i;\mathbf{a})\right)$$

jest rozkładem gęstości prawdopodobieństwa mierzonej wartości y_i.
Dobierzmy dodatkowo pewną funkcję $\rho\left(y_i, y(x_i;\mathbf{a})\right)$ taką, że

$$p\left(y_i, y(x_i;\mathbf{a})\right) = \exp\left[-\rho\left(y_i, y(x_i;\mathbf{a})\right)\right]$$

Na przykład, dla rozkładu normalnego byłaby to funkcja

$$\rho\left(y_i, y(x_i;\mathbf{a})\right) = \frac{1}{2}\left(\frac{y_i - y(x_i;\mathbf{a})}{\sigma_i}\right)^2 + \ln\left(\sigma_i\sqrt{2\pi}\right).$$

W takim przypadku, stosując metodę największej wiarygodności, funkcję wiarygodności moglibyśmy zapisać w następującej postaci:

$$P = \prod_{i=1}^{n}\left\{\exp\left[-\rho\left(y_i, y(x_i;\mathbf{a})\right)\right]\delta y\right\}.$$

Postępując dalej tak jak przy wprowadzaniu MNK zauważamy, że maksymalizacji funkcji wiarygodności odpowiada minimalizacja wartości następującej sumy:

$$\sum_{i=1}^{n}\rho\left(y_i, y(x_i;\mathbf{a})\right).$$

Często jest tak, że wartość funkcji ρ nie zależy niezależnie od obu swoich argumentów – wartości zmierzonej y_i i oczekiwanej $y(x_i)$ – ale raczej od ich różnicy (przynajmniej po przeskalowaniu pewnym czynnikiem wagowym σ_i przypisanym każdemu punktowi). W takim przypadku o wyznaczanych estymatorach mówimy, że są lokalne, a problem sprowadza się do minimalizacji sumy:

$$\sum_{i=1}^{n}\rho\left(\frac{y_i - y(x_i;\mathbf{a})}{\sigma_i}\right),$$

gdzie $\rho(z)$ jest funkcją tylko jednej zmiennej $z \equiv \left[y_i - y(x_i)\right]/\sigma_i$.
Wprowadźmy jeszcze oznaczenie

$$\psi(z) \equiv \frac{d\rho(z)}{dz}.$$

Wtedy odpowiednikiem układu m równań dla wyznaczenia parametrów $\{a_k\}$ staje się:

$$0 = \sum_{i=1}^{n} \frac{1}{\sigma_i} \psi\left(\frac{y_i - y(x_i)}{\sigma_i} \right) \left(\frac{\partial y(x_i; \mathbf{a})}{\partial a_k} \right) \qquad k = 1, \ldots, m .$$

Dla specjalnego przypadku rozkładu normalnego $\psi(z) = z$ i układ równań jest taki sam jak dla MNK.

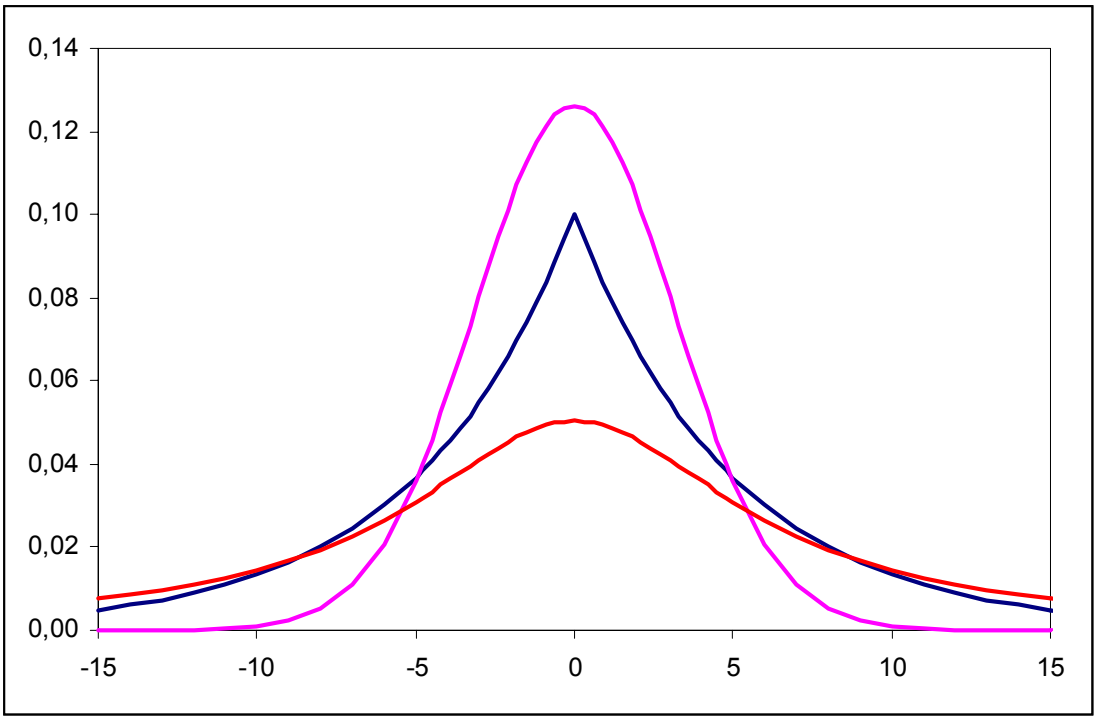
Jeżeli błędy mają np., rozkład podwójny wykładniczy

$$P(y_i - y(x_i)) \propto \exp\left(- \left| \frac{y_i - y(x_i)}{\sigma_i} \right| \right),$$

wtedy

$$\rho(z) = |z| \qquad \psi(z) = \operatorname{sgn}(z)$$

Dla rozkładu podwójnego wykładniczego estymator największej wiarygodności jest otrzymywany przez minimalizację sumy *bezwzględnych odchyleń* (ważonych). Ogony rozkładu, chociaż maleją wykładniczo, to jednak dla każdej wartości odchylenia dają wartość większą niż funkcja rozkładu normalnego.



Jeszcze bardziej rozległym rozkładem, który w niektórych przypadkach może być bardziej realistyczny, jest rozkład Lorentza (Cauchy'ego):

$$P(y_i - y(x_i)) \propto \frac{1}{1+\frac{1}{2}\left(\frac{y_i - y(x_i)}{\sigma_i}\right)^2}$$

Dla rozkładu Lorentza

$$\rho(z) = \ln\left(1+\tfrac{1}{2}z^2\right) \qquad \psi(z) = \frac{z}{1+\frac{1}{2}z^2}$$

W równaniu

$$0 = \sum_{i=1}^{n} \frac{1}{\sigma_i} \psi\left(\frac{y_i - y(x_i)}{\sigma_i}\right)\left(\frac{\partial y(x_i;\mathbf{a})}{\partial a_k}\right)$$

wartość funkcji ψ pełni rolę swego rodzaju czynnika wagowego, którego wartość zależy od odchyłki $\left[\frac{y_i - y(x_i;\mathbf{a})}{\sigma_i}\right]$. Dla błędów o rozkładzie normalnym ta waga jest tym większa im większa odchyłka. Dla rozkładu podwójnego wykładniczego wagi są względnie jednakowe i pod uwagę brany jest tylko znak odchyłki. Wreszcie dla rozkładu Lorentza, z najbardziej wyrazistymi ogonami, waga najpierw rośnie, a następnie maleje ze wzrostem odchyłki tak, że punkty bardzo odchylone (wyniki naprawdę odstające) praktycznie nie są brane pod uwagę.

Dopasowanie linii prostej przez minimalizację bezwzględnych odchyłek

Przyjmujemy model w postaci:

$$y(x;a,b) = ax + b$$

i zakładamy dalej, że niepewności wszystkich pomiarów są jednakowe, to znaczy wszystkie $\sigma_i = \sigma$.

Parametry a i b wyznaczamy minimalizując sumę odchyłek bezwzględnych:

$$\sum_{i=1}^{n} |y_i - ax_i - b|.$$

Wykorzystamy fakt, że dla zbioru punktów c_i ich mediana c_M jest wielkością minimalizującą sumę odchyleń bezwzględnych:

$$\sum_{i=1}^{n} |c_i - c_M|.$$

Dowód dla mediany:

$$\frac{d}{dc_M}\left(\sum_{i=1}^{n}|c_i - c_M|\right) = \sum_{i=1}^{n}\frac{d|c_i - c_M|}{dc_M} = \sum_{i=1}^{n}\left(-\operatorname{sgn}(c_i - c_M)\right)$$

$$\sum_{i=1}^{n}\operatorname{sgn}(c_i - c_M) = 0$$

Wynika stąd, że dla ustalonego a wartością b, która minimalizuję sumę bezwzględnych odchyłek jest

$$b = mediana\{y_i - ax_i\}$$

Różniczkując sumę ze względu na parametr a otrzymujemy warunek

$$\frac{\partial}{\partial a}\left(\sum_{i=1}^{n}|y_i - ax_i - b|\right) = \sum_{i=1}^{n}\left(-x_i \operatorname{sgn}(y_i - ax_i - b)\right) = 0$$

Podstawiając wynik otrzymany dla parametru b otrzymujemy równanie z jedną niewiadomą a:

$$\sum_{i=1}^{n} x_i \operatorname{sgn}\left(y_i - ax_i - mediana\{y_i - ax_i\}\right) = 0,$$

które można rozwiązać metodami numerycznymi.

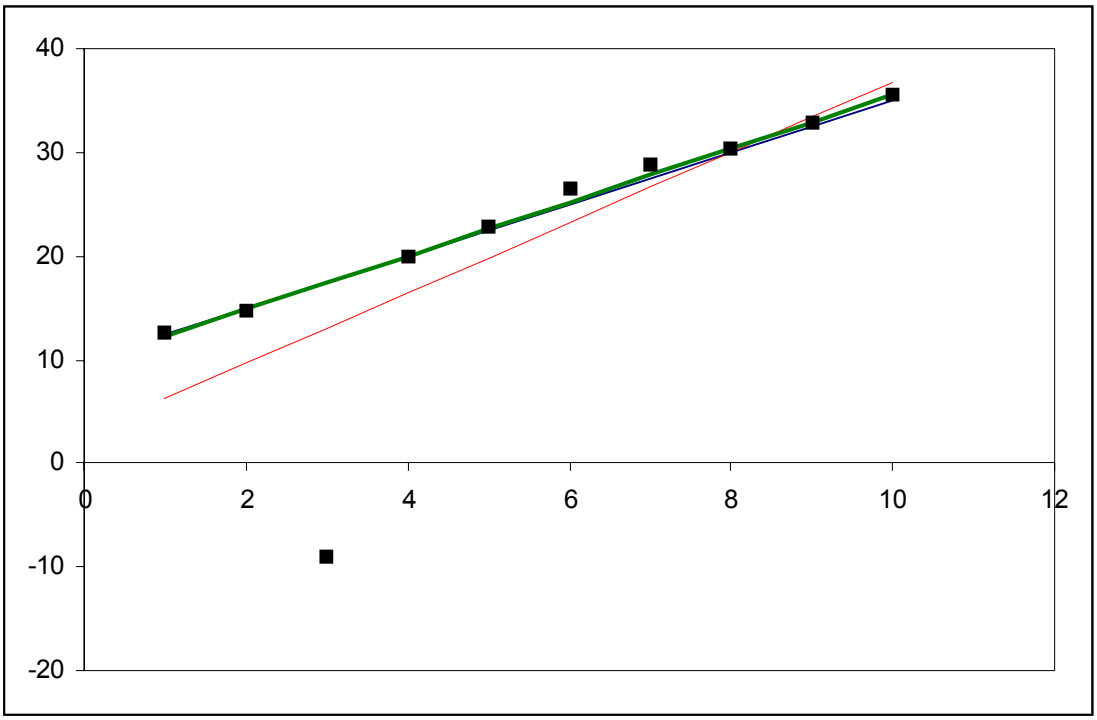
Przykład

Model

$$y = 10 + 2{,}5x$$

normalny N(0;1) 90% ×2
równomierny $P_R(0,1)$ 10% ×20

i	$N(0,1)$	$P_R(0,1)$	ε_i	X_i	Y_i
1	0,1000	0,5519	0,200	1	12,700
2	-0,1010	0,3616	-0,202	2	14,798
3	-1,3275	0,9921	-26,551	3	-9,051
4	-0,0445	0,2084	-0,089	4	19,911
5	0,1433	0,8761	0,287	5	22,787
6	0,7632	0,5023	1,526	6	26,526
7	0,6594	0,4973	1,319	7	28,819
8	0,1705	0,1077	0,341	8	30,341
9	0,2115	0,3413	0,423	9	32,923
10	0,2529	0,3927	0,506	10	35,506



MNK

$a = 3{,}3912 \quad b = 2{,}8743$

Minimalizacja sumy bezwzględnych odchyłek

$a = 2{,}5824 \quad b = 9{,}6814$

Numeryczne rozwiązanie równania

$$f(a)=\sum_{i=1}^{n} x_i \operatorname{sgn}\left(y_i - a x_i - mediana\{y_i - a x_i\}\right)=0$$