Dopasowanie prostej do punktów z niepewnościami dla obydwu współrzędnych. Nowe wyniki dla problemu zapoczątkowanego w XIX wieku

A. Zięba

Wydział Fizyki i Informatyki Stosowanej,
Akademia Górniczo-Hutnicza im. St. Staszica w Krakowie,
al. Mickiewicza 30, 30-059 Kraków

e-mail: andrzej.zieba@fis.agh.edu.pl

Stosując metodę najmniejszych kwadratów, nie zawsze pamiętamy o założeniu, że niepewnościami $u(y_i)$ obarczona jest tylko zmienna zależna. Przypadek niezerowych niepewności dla obydwu współrzędnych (ang. *Errors-in-Variables*) budził mało uwagi. Tematem referatu jest uogólnienia metod: OLS (ang. *Ordinary Least Squares*) oraz WLS (ang. *Weighted Least Squares*) na przypadek $u(x_i) > 0$. Przedstawię spójną prezentację wyników znanych i rezultaty własne.

Model statystyczny problemu. Wariancja efektywna

Dla standardowego dopasowania prostej wartości zmiennej zależnej modeluje wzór

$$y_i = a_0 x_i + b_0 + e_{y,i}, \tag{1}$$

w którym parametry a_0 i b_0 określają „prawdziwą” prostą (nieznaną w eksperymencie), zaś składnik $e_{y,i}$ modeluje błąd przypadkowy zmiennej zależnej.

Gdy wartości x_i są ponadto obarczone błędami $e_{x,i}$, model przybiera postać

$$y_i = a_0(x_{0,i} + e_{x,i}) + b_0 + e_{y,i}. \tag{2}$$

Składnik $a_0 x_{0,i} + b_0$ jest deterministyczny, natomiast $a_0 e_{x,i} + e_{y,i}$ oznacza nową zmienną losową. Modele (1) i (2) mają te same właściwości stochastyczne, o ile zmiennej y_i we wzorze (1) przypiszemy wariancję efektywną

$$\mathrm{Var}(y_i) = a_0^2\ \mathrm{Var}(e_{x,i}) + \mathrm{Var}(e_{y,i}). \tag{3}$$

Punkty o różnych niepewnościach – uogólnienie metody WLS

Wyprowadzenie Reeda [1] równania (4) polegało na uproszczeniu pseudo-kubicznego równania Yorka [2]. Elementarne wyprowadzenie obejmuje następujące kroki:

(i) efektywne wariancje przybliżamy jako niepewności efektywne:

$$u_{\mathrm{eff}}(y_i) = [a^2 u^2(x_i) + u^2(x_i)]^{1/2}$$

i wykorzystujemy do zdefiniowania wag funkcji kryterialnej $M(a, b)$ metody WLS,

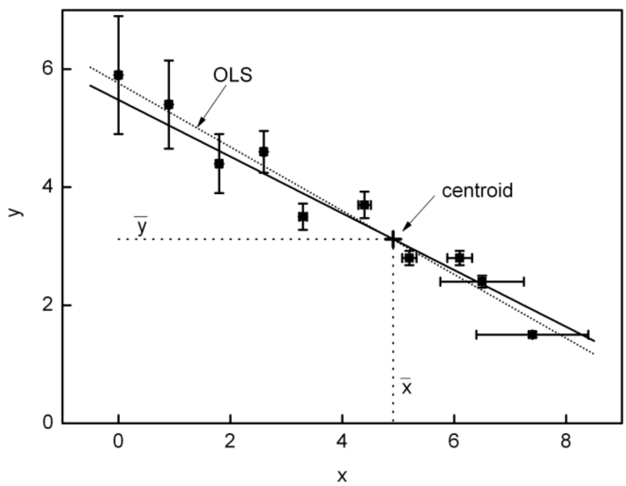
(ii) warunek $\partial M(a,b)/\partial b = 0$ można rozwiązać analitycznie, co redukuje problem do poszukiwania minimum funkcji jednej zmiennej $M(a)$,

(iii) warunek $\partial M(a)/\partial a = 0$ prowadzi do równania pseudo-kwadratowego

$$a = \frac{\sum (y_i - \overline{y})[(x_i - \overline{x}) + aC(y_i - \overline{y})]}{\sum (x_i - \overline{x})[(x_i - \overline{x}) + aC(y_i - \overline{y})]}, \tag{4}$$

w którym $\overline{x}$ oraz $\overline{y}$ oznaczają średnie ważone.

Równanie (4) rozwiązuje się metodą iteracyjną. Dla przykładowych danych (Rysunek 1) punktem startowym jest $a = -0.5396$ z metody OLS; kolejne iteracje dają: -0.4602; -0.4812; -0.4804; i -0.4805.



Rysunek 1: Dopasowanie OLS i dopasowanie prawidłowe (linia ciągła) dla przykładowych danych

Punkty równoważne – uogólnienia metody OLS

Gdy niepewności są te same dla każdego punktu ale $u(x) \neq u(y)$, mamy dopasowanie Deminga (DR – od *Deming Regression*). Najprostszy jest model dopasowania ortogonalnego (OR – od *Orthogonal Regression*) dla którego $u(x) = u(y)$. Wzór na nachylenie prostej dla OR podał Adcock [3],

$$a = D \pm \sqrt{D^2 + 1}, \quad D = \frac{S_{yy} - S_{xx}}{2S_{xy}}, \tag{5}$$

gdzie

$$S_{xx} = \sum (x_i - \overline{x})^2, \quad S_{yy} = \sum (y_i - \overline{y})^2, \quad S_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y}). \tag{6}$$

Natomiast niepewność nachylenia,

$$u^2(a) = \frac{M}{n-2}\frac{(1+a^2)(S_{xx}+S_{yy})}{(S_{xx}-S_{yy})^2+4S_{xy}^2}, \tag{7}$$

wyprowadził Petrolini [4] – ponad sto lat później.

Omówione też zostanie znaczenie formalizmu dla danych o różnym stopniu rozrzutu.

Podziękowania

Dziękuję Cameronowi Reed'owi za użyteczną korespondencję. Temat opracowałem dla przygotowanego podręcznika analizy danych [5].

Bibliografia

[1] Reed B. C., Am. J. Phys. 60, 59 (1992).

[2] York D., Can. J. Phys. 44, 1079 (1966).

[3] Adcock R. J., Analyst 5, 53-54 (1878).

[4] Petrolini A., Am. J. Phys. 82, 1178 (2014).

[5] Zięba A., Analysis of Experimental Data in Science and Technology, Cambridge Scholar Publishing, Newcastle, UK (2023).