**УДК 330.46:53**

# НЕЛІНІЙНА РЕГРЕСІЯ ПРИ ТЕОРЕТИЧНІЙ ОБРОБЦІ ЕКСПЕРИМЕНТАЛЬНИХ ДАНИХ

Афанасьєва Л.М. к. техн.н., доцент кафедри вищої математики й ЕММ,
Харківський національний економічний університет імені Семена Кузнеця, м. Харків, Україна

***Анотація*** **–прпонується алгоритм визначення параметрів теоретичного рівняння регресії у випадку нелінійної моделі .**

***Ключові слова*** —теоретична функція, нелінійна регресія, лінеарізація моделі, ряд Тейлора, формфактор, крітерій згоди.

Нехай на основі експерименту отримано ряд значень випадкової величини В. Треба знайти таку функцію *F(a,x)*,яка найкращим способом відповідає експериментальним даним. У загальному вигляді теоретична функція *F(a,x)* може містити нелінійні параметри, і традиційний метод найменших квадратів як метод розв'язання системи лінійних рівнянь виявляється непридатним. У низці випадків, наприклад, таких, як визначення сталої радіоактивного розпаду або параметрів гауссового розподілу, можна прологарифмувати *F*, і зробивши заміну змінних, здобути систему лінійних рівнянь відносно шуканих параметрів. Але часто заміною змінних лінеаризувати задачу не вдається. Саме така ситуація й виникає при аналізі формфакторів, що їх вимірюють у широкому діапазоні переданих імпульсів. У борнівському наближенні формфактор інтерпретується як Фур'є-образ розподілу заряду і є нелінійною функцією параметрів, що підлягають оцінці. У цьому випадку для лінеаризації задачі функцію *F* розкладають у ряд Тейлора.

Спочатку з яких-небудь міркувань знаходять початкові наближені значення коефіцієнтів $a_k^0$, а потім розкладають підгінну функцію *F(a,x)* у ряд Тейлора в околі точки $F(a^0,x)$ по степенях похибки нульового наближення $\Delta a_k = a_k - a_k^0$. Якщо величини $a_k^0$ визначені достатньо точно, так, що $\Delta a_k / a_k << 1$, то в розвиненні можна залишити тільки члени першого порядку малості:

$$F(a_k,x) = F(a_k^0,x) + \sum_{k=0}^{m}\left(\frac{\partial F(a_k,x)}{\partial a_k}\right)_{a_k=a_k^0}\Delta a_k \qquad (1)$$

Якщо розвинення (1) є справедливим, то тим самим задачу зведено до лінійної, для якої матриця **Ф** являє собою матрицю частинних похідних першого порядку від вихідних функцій за відповідними параметрами, а вектор **В** спостережуваних величин задається відхиленнями вимірюваних значень від значень функцій у точках розвинення. Здобута внаслідок розвинення функція є лінійною відносно приростів $\Delta a_l$, і для того, щоб знайти їх, потрібно розв'язати звичайну систему лінійних рівнянь. У підсумку дістанемо "уточнені" значення оцінок параметрів: $a_l = a_l^o + \Delta a_l$, які знову використовуються як наближені значення для обчислення нових величин $a_l$. У процесі ітерацій щоразу дістаємо нові значення $a_l$, і отже, похідні та матриця **Ф** змінюватимуться. Ця ітераційна процедура триватиме доти, доки не буде виконано умову

$$\left|S_i^2 - S_{i-1}^2\right| \le \varepsilon,$$

де $\varepsilon$ – це наперед задана мала величина. У процесі подібного уточнення кожна ітерація виконується так само, як й у лінійному випадку.

По закінченні процесу здобудемо параметри $\breve{a}_k$, які відповідають мінімуму зваженої суми квадратів відхилень експериментальних даних від лінії регресії

$$S_{\min}^2 = \sum_i w_i (F(\breve{a},x_i) - y_i)^2 \ . \qquad (2)$$

Величина $S^2_{min}$ обумовлена двома факторами: розкидом (дисперсією) вихідних даних і тим, наскільки близько лінія регресії проходить до експериментальних точок. Тобто ця величина містить у собі і якість експериментальних даних, і якість підгонки. Частка $S^2_{min}$ до середньозваженої суми квадратів відхилень вихідних даних від їхнього середнього значення

$$R^2 = \frac{\sum_i w_i (F(\breve{a}, x_i) - y_i)^2}{\sum_i w_i (\bar{y} - y_i)^2} \qquad (3)$$

показує, яка частина повного розкиду вихідного масиву даних обумовлена наявністю лінійної регресії між ними. $R^2$ називають *коефіцієнтом детермінації* (coefficient of determination), а $\sqrt{R^2}$ – *кореляційною часткою*.

Для оцінки адекватності застосованої математичної моделі та масиву даних часто використовується середнє значення кореня квадратного з $S^2_{min}$, що його називають стандартним відхилом (standard deviation)

$$SD = \sqrt{\frac{\sum_i w_i (F(\breve{a}, x_i) - y_i)^2}{n-m}}\,. \qquad (4)$$

Визначимо тепер похибки отриманих коефіцієнтів $\breve{a}_k$. У математичній статистиці показано, що середньоквадратичні похибки оцінок є діагональними елементами оберненої матриці системи нормальних рівнянь, помноженими на величину $S^2_{min}$, яка припадає на ступінь вільності. Кількість ступенів вільності визначається як кількість експериментальних точок $n$ мінус кількість накладених на них лінійних зв'язків $m$, і отже, похибки $\Delta a_k$ обчислюються за формулою:

$$\Delta a_k = \sqrt{\frac{S^2_{\min}}{n-m}\left(\mathbf{\Phi}^{-1}\right)_{k,k}}\,. \qquad (5)$$

Тому проблему зіставлення результатів експерименту з теорією формулюють так: як підібрати параметри моделі, щоб теоретична крива, що її проведено по експериментальних точках, задовольняла умові мінімуму функціонала

$$S^2(R,g) = \sum_i \frac{(F(R,g,x_i) - Y_i)^2}{(\Delta Y_i)^2} = \min\,. \qquad (6)$$

Оскільки $F(R,g,x_i)$ є нелінійною функцією параметрів $R$ і $g$, які підлягають оцінці, то для лінеаризації задачі функцію $F$ розвивають у ряд Тейлора в околі точки $F(R_0,g_0,x)$ по степенях похибок нульового наближення $\Delta R = R - R_0$ і $\Delta g = g - g_0$, утримуючи лише члени першого порядку малості

$$F(R,g,x) = F(R_0,g_0,x) + \left(\frac{\partial F(R,g,x)}{\partial R}\right)_{\substack{R=R_0,\\ g=g_0}} \cdot \Delta R + \left(\frac{\partial F(R,g,x)}{\partial g}\right)_{\substack{R=R_0,\\ g=g_0}} \cdot \Delta g\,.$$

Позначемо

$$F_0(x) = F(R_0,g_0,x);\; F'_{R_0} = \left(\frac{\partial F(R,g,x)}{\partial R}\right)_{\substack{R=R_0,\\ g=g_0}};\quad F'_{g_0} = \left(\frac{\partial F(R,g,x)}{\partial g}\right)_{\substack{R=R_0,\\ g=g_0}}$$

і запишемо систему рівнянь, що відповідає мінімуму функціонала (6), у матричному вигляді

$$\begin{pmatrix} \sum_{i=1}^{n} F'^{\,2}_{R_0}(x_i) & \sum_{i=1}^{n} F'_{R_0}(x_i) F'_{a_1}(x_i) \\ \sum_{i=1}^{n} F'_{g_0}(x_i) F'_{a_0}(x_i) & \sum_{i=1}^{n} F'^{\,2}_{g_0}(x_i) \end{pmatrix} \cdot \begin{pmatrix} \Delta R \\ \Delta g \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{n} [Y_i - F_0(x_i)] F'_{R_0}(x_i) \\ \sum_{i=1}^{n} [Y_i - F_0(x_i)] F'_{g_0}(x_i) \end{pmatrix}. \qquad (7)$$

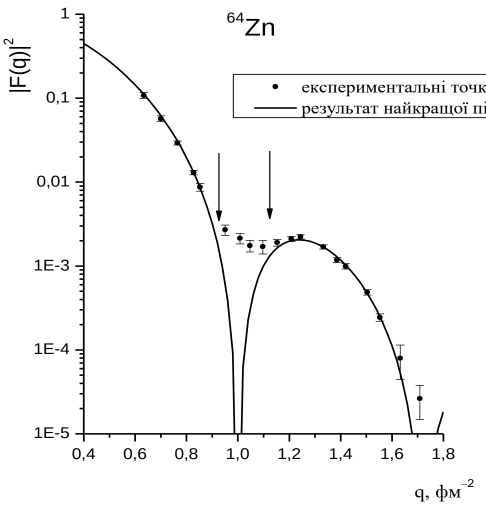


*Рис.1. Формфактор пружного розсіяння електронів на ядрі $^{64}$Zn*

Ця система рівнянь є лінійною відносно приростів $\Delta R$ і $\Delta g$. Розв'язавши систему, здобувають "уточнені" значення оцінок параметрів: $R=R_0+\Delta R$ і $g=g_0+\Delta g$, які знову використовують як початкове наближення для обчислення нових значень $R$ і $g$. Ітераційна процедура триває доти, доки на $p$-му кроці не буде досягнута умова $\left|\chi_p^2-\chi_{p-1}^2\right|\leq\varepsilon,$ де $\varepsilon$ – це наперед задана мала величина. Рис.1 демонструє результат підгонки пружного формфактора в моделі Хелма до експериментальних даних для $^{64}$Zn, що його виконано за допомогою ітераційної процедури. Оскільки борнівське наближення не працює в діапазоні дифракційного мінімуму, то точки, які лежать у цьому діапазоні, слід виключити з підгонки. Стрілками зазначено діапазон точок, що їх виключено з підгонки.

Після того, як методом найменших квадратів здобуто оцінки параметрів та їхніх похибок, розглядають питання про ступінь згоди між теорією та результатами експерименту. У кожному конкретному випадку застосовують найбільш прийнятні методи порівняння та критерії згоди. експерименту в порівнянні з теорією використовується не лише для ілюстрації результатів, але і як аргумент при їхньому обговоренні: практично в кожній науковій публікації можна зустріти слова авторів "як

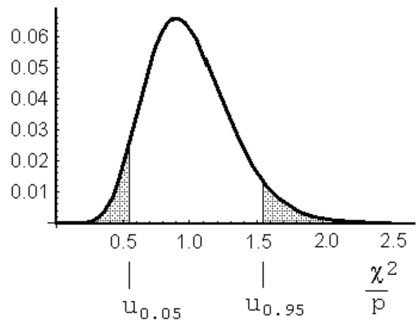


*Рис. 2.Розподіл $\chi^2$ для p = 20*

При вирішенні питання про згоду в ядерній фізиці прийнято розглядати як критерій згоди залишкову суму

$$S_{\min}^2=\sum_{i=1}^{n}\left[\frac{F(\hat{a},x_i)-Y_i}{\Delta Y_i}\right]^2$$

(тут $\hat{a}$ – це набір параметрів, що їх здобуто методом найменших квадратів). Щоб установити кількісний критерій, який би визначав, за яких значень $S^2_{min}$ згоду можна вважати гарною, задовільною або поганою, потрібно знати закон розподілу цієї величини. Виявляється, що вона задовольняє $\chi^2$-розподілу з $n-m$ ступенями вільності і середнім значенням $\left\langle\chi^2\right\rangle=n-m-1$. Звідси дістаємо важливий висновок про те, що якщо

$$\frac{S_{\min}^2}{n-m-1}\approx 1,$$

то згоду між теорією та експериментом слід вважати гарною. Більш точні висновки про ступінь згоди можна зробити, якщо скористатись розподілом $\chi^2(p,x)$ при заданій кількості ступенів вільності $p$. На рис.2 показаний розподіл $\chi^2(p=20,x)$ як функція від $x/p$. Заштриховано діапазони $x$, імовірність влучення в які складає менше п'яти відсотків. Нижня та верхня межі інтервалу, відповідно, $u_{0.05}$ і $u_{0.95}$, визначаються інтегралами

$$\int_0^{u_{0.05}} \chi^2(p=20,x)dx = 0.05 \quad \text{и}$$

$$\int_0^{u_{0.95}} \chi^2(p=20,x)dx = 0.95 .$$

Ці межі називають *квантилями* розподілу $\chi_p^2(p,x)$. Якщо здобута з експериментальних даних частка $\chi^2/p$ потрапляє до інтервалу 1 ÷ $u_{0.95}$ або $u_{0.05}$ ÷ 1, то з імовірністю 95% можна стверджувати, що передбачення теоретичної моделі збігаються з експериментом. Якщо ж $\chi^2/p$ >>1, то це означає, що теорія не узгоджується з експериментом. З іншого боку, малі значення $\chi^2/p$ настільки ж малоймовірні, як і великі.

Часто параметри, що їх здобуто методом найменших квадратів, використовують для обчислення інших величин. Наприклад, визначають середньоквадратичний радіус розподілу заряду з параметрами *R* і *g*:

$$< r^2 >^{\frac{1}{2}} = Q = \sqrt{\frac{3}{5}}(\mathrm{R}^2 + 5\mathrm{g}^2)^{1/2} . \qquad (8)$$

Оскільки нам відомі лише оцінки *R, g*, то й величина *Q* визначається з похибкою. Середньоквадратичну похибку цієї оцінки можна здобути, розглядаючи *Q* як випадкову величину, що представляє собою функцію випадкових величин *R* і *g* з дисперсіями $\sigma_R^2 = (\Delta R)^2$, $\sigma_g^2 = (\Delta g)^2$ і середніми значеннями $\bar{R}, \bar{g}$. Для цього розвинемо функцію *Q* у ряд Тейлора в околі $\bar{R}, \bar{g}$, залишаючи лише перші члени ряду. Тоді для дисперсії $\sigma_Q^2 = \left[Q(R,g) - Q(\bar{R},\bar{g})\right]^2$ є справедливою наближена формула:

$$\sigma_Q^2 \approx \left[ \left.\frac{\partial Q}{\partial R}\right|_{\bar{R},\bar{g}} \Delta R + \left.\frac{\partial Q}{\partial g}\right|_{\bar{R},\bar{g}} \Delta g \right]^2 = \left(\frac{\partial Q}{\partial R}\right)^2 \sigma_R^2 + \left(\frac{\partial Q}{\partial g}\right)^2 \sigma_g^2 + 2\frac{\partial Q}{\partial R}\frac{\partial Q}{\partial g}\sigma_{Rg} \qquad (9)$$

Величину $\sigma_{Rg} = \overline{\left(R-\bar{R}\right)\left(g-\bar{g}\right)}$ називають "коваріацією" або кореляційним моментом *R* і *g*. Вона визначається недіагональними елементами оберненої матриці $\mathbf{\Phi}^{-1}$ системи нормальних рівнянь з невідомими параметрами *R* і *g*. Якщо *R* і *g* є незалежними, то $\sigma_{Rg} = 0$, і остаточний вираз для стандартного (середньоквадратичного) відхилу $\sigma_Q$ набуває вигляду:

$$\sigma_Q^2 \approx \left(\frac{\partial Q}{\partial R}\right)^2 \sigma_R^2 + \left(\frac{\partial Q}{\partial g}\right)^2 \sigma_g^2 . \qquad (10)$$

Це співвідношення називають "правилом перенесення похибок (Error Propagation)" для обчислення середньоквадратичної похибки величини, що залежить від випадкових величин. Із цього правила випливають наступні два відомі прості співвідношення для обчислення похибки *ΔU*.

А. $U = x \pm y.\ \frac{\partial U}{\partial x} = 1,\ \frac{\partial U}{\partial y} = 1, \Rightarrow \Delta U = \sqrt{\Delta x^2 + \Delta y^2}$ .

Тут стандартний відхил *ΔU* є середньоквадратичною сумою відхилень *Δx* і *Δy*.

Б. $U = x \cdot y$ (чи $\frac{\mathrm{x}}{\mathrm{y}}$). $\qquad \frac{\Delta U}{U} = \sqrt{\left(\frac{\Delta x}{x}\right)^2 + \left(\frac{\Delta y}{y}\right)^2}$ .

В цьому разі ми знаходимо не абсолютну величину *ΔU*, а відносну похибку *ΔU/U* як середньоквадратичну суму відносних похибок *x* і *y*.

## Список використаної літератури

1. Д. Худсон. Статистика для физиков. − М.: Мир, 1970 – 296 с.

2.Ферстер Е., Ренц Б. Методы корреляционного и регрессионного анализа.-М.:Финансы и статистика,1983.-394 с.

3.КобзарьА.И. Прикладная математическая статистика. Для инженеров и научных работников-М.:ФИЗМАТЛИТ,2012.-812 с.

Автор

**Афанасьєва Л.М,** к. техн.н., доцент, Харківський національний економічний університет імені Семена Кузнеця (lydia.afa@gmail.com)

Тези доповіді надійшли 9 лютого 2018 року.

Опубліковано в авторській редакції.