Припустимо, що існують дві змінні x і y, де x – незалежна змінна, y – залежна змінна. Співвідношення між x та y є статистичним, а саме

$$y = a + b\cdot x + \varepsilon \qquad (2.1)$$

, де ε є похибка або збурення і має відомий імовірнісний розподіл (тобто є випадковою величиною). Детермінована компонента $a + b\cdot x$ (2.1) містить параметри регресії (a, b), які потрібно оцінити на основі n пар значень (x_j, y_j); $y_j = a + b\cdot x_j + \varepsilon_j$, $j = 1 \ldots n$. Щоб зробити це, вважають, що збурення ε_j задовольняють низці статистичних гіпотез [2]:

1. величини ε_j є незалежними ($E(\varepsilon_k\cdot\varepsilon_j) = 0$, $\forall$ k, j), що означає, з одного боку, відсутність у змінної (x) похибки при вимірюванні, з іншого боку, незалежна змінна (x) є єдиною змінної яка впливає детермінованим способом на поведінку залежної змінної (y);
2. мають нульове математичне сподівання ($E(\varepsilon_j) = 0$, $\forall$ $j = 1 \ldots n$), це припущення є технічним, якщо наприклад $E(\varepsilon_j) = \text{constant}$, $\forall$ $j = 1 \ldots n$, тоді потрібно включити цю константу в величину параметра регресії (a);
3. мають однакові дисперсії ($E(\varepsilon_j^2) = E(\varepsilon_k^2) = \sigma^2$, $\forall$ k, j), що означає, однакову міру ненадійності всіх спостережень (x_j, y_j);
4. розподілені за нормальним законом.

В подальшому, порушення одного або декількох гіпотез 1 – 4 буде розглядатися окремо. На даному етапі, сформулюємо алгоритм отримання характеристик лінійної регресії при умові виконання всіх припущень 1 – 4.

Оцінюємо параметри регресії (a, b) (2.1) використовуючи метод найменших квадратів. Ведемо наступні позначення: $S(\mathbf{B}) = \left\| \mathbf{Y} - \mathbf{X}\cdot\mathbf{B} \right\|^2 \equiv \sum(y_j - \alpha - \beta\cdot x_j)^2$ – сума квадратів залишків, $\mathbf{Y} = (y_1,\ y_2,\ \ldots,\ y_{n-1},\ y_n)^T$, $\mathbf{B} = (\alpha,\ \beta)^T$, $\mathbf{X} = \begin{pmatrix} 1 & 1 & & 1 & 1 \\ x_1 & x_2 & ; \ldots ; & x_{n-1} & x_n \end{pmatrix}^T$. Задача на знаходження мінімуму $S(\mathbf{B})$ має розв'язок [2]:

$$\mathbf{B} = (\mathbf{X}^T\cdot\mathbf{X})^{-1}\cdot\mathbf{Y} \qquad (2.2)$$

Оцінені в такий спосіб параметри (a, b) були позначені (α, β) і дають змогу записати вибіркову функцію регресії (2.3), що представляє усереднене, або закономірне значення залежної змінної (y) при даному значенню незалежної мінної (x):

$$y = \alpha + \beta\cdot x \qquad (2.3)$$

Підраховуємо коефіцієнт детермінації $R^2 = 1 - S(\mathbf{B})/S_1(\mathbf{B})$, що є часткою поясненої ($S_1(\mathbf{B}) - S(\mathbf{B})$) і загальної суми квадратів ($S_1(\mathbf{B})$), де $S_1(\mathbf{B}) = \sum(y_j - E(\mathbf{Y})\ )^2$, $E(\mathbf{Y}) = 1/n\cdot\sum y_j$ – середнє значення, j = 1 ... n. Коефіцієнт детермінації є мірою тісноти лінійного зв'язку між змінними (y) та (x).

Підраховуємо величину довірчих інтервалів, з рівнем довіри $1 - \delta$, для коефіцієнтів вибіркової функції регресії (2.3):

$$\alpha \pm s\cdot Q_1\cdot t(n - 2, \delta/2) \qquad (2.4)$$

$$\beta \pm s\cdot Q_2\cdot t(n - 2, \delta/2) \qquad (2.5)$$

, де $s^2 = S(\mathbf{B})/(n - 2)$, $Q_k^2 = \{(\mathbf{X}^T\cdot\mathbf{X})^{-1}\}_{k,k}$ – $k^й$ діагональний матричний елемент, k = (1, 2), $t(n - 2, \delta/2)$ – інверсна сукупна функція щільності (inverse cumulative density function) для розподілу Стьюдента з $n - 2$ ступенями свободи і рівнем надійності δ [2]. Надалі буде використовуватися лише значення $\delta = 0.05$ що відповідає рівню довіри – 95%. Величина s^2 є незміщеною оцінкою дисперсії для похибки ($s = \sigma$).

Знаходимо рівняння еліпса в перемінних (α, β), що окреслює спільну довірчу область, з рівнем довіри $1 - \delta$, на коефіцієнти регресії (2.11):

$$(\mathbf{B_1} - \mathbf{B})\cdot(\mathbf{X}^T\cdot\mathbf{X})\cdot(\mathbf{B_1} - \mathbf{B}) \le 2\cdot s^2\cdot F(2, n - 2, \delta) \qquad (2.6)$$

, де $\mathbf{B}$ – значення отримані із (2.10), $\mathbf{B_1} = (\alpha, \beta)^T$ – вектор змінних, $F(2, n - 2, \delta)$ – інверсна сукупна функція щільності для розподілу Фішера з 2 та $n - 2$ ступенями свободи і рівнем надійності $\delta = 0.05$ [2].

Знаходимо рівняння кривих, що задають величину довірчого інтервалу, з рівнем довіри $1 - \delta$, для самої вибіркової функції регресії (2.11):

$$\mathbf{z}^T\cdot\mathbf{B} \pm D \qquad (2.7)$$

, де $\mathbf{z}^T = (1, x)$, $D^2 = 2\cdot s^2\cdot\mathbf{z}^T\cdot(\mathbf{X}^T\cdot\mathbf{X})^{-1}\cdot\mathbf{z}\cdot F(2, n - 2, \delta)$, $\mathbf{B}$ – значення отримані по формулі (2.2).

✞✋✍ Перевіряємо справедливість статистичних гіпотез 1 – 4, що до досліджуваних експериментальних даних: (x_j, y_j), $j = 1 \ldots n$. Серед існуючої кількості статистичний критерій [2,1], в даному дослідженні використовується якісний (графічний) критерій. Для цього будуємо графік залежності упорядкованого залишку r_j (studentized residuals), ($r_j < r_k \leftrightarrow j < k$; $j, k = 1 \ldots n$) від квантиля q_j нормального розподілу рівня $j/(n+1)$ (normal quantiles):

$$r_j = (y_j - \alpha - \beta\cdot x_j)/[s\cdot\sqrt{(1 - h_j)}], \qquad (2.8)$$

$$q_j = \Phi(j/(n+1)) \qquad (2.9)$$

, де $h_j = \{\mathbf{X}\cdot(\mathbf{X}^{\mathrm{T}}\cdot\mathbf{X})^{-1}\cdot\mathbf{X}^{\mathrm{T}}\}_{j,j}$ – $j^{й}$ діагональний матричний елемент, $\Phi(x)$ – інверсна сукупна функція щільності для нормального розподілу із нульовим математичним сподіванням та одиничною дисперсією N(0,1), $j = 1 \ldots n$. У випадку справедливості гіпотез 1 – 4, точки на графіку (q_j, r_j) повинні чітко описуватися лінійною залежністю [2].

ЦИТОВАНА ЛІТЕРАТУРА

1. JOHNSON, J. L., F. C. LEONE: Statistics and Experimental Design in Engineering and the Physical Sciences.
2. Bates D. M., Watts D. G. Nonlinear regression analysis and its applications (Wiley, 1988).