Testowanie jakości dopasowania.

Test χ^2 jakości dopasowania

Metoda najmniejszych kwadratów opiera się na założeniu, że najlepszą funkcją opisującą zależność między wielkościami jest taka, która minimalizuje ważoną sumę kwadratów odchyleń wartości y_i od dopasowywanej funkcji $y(x_i)$. Tę sumę można scharakteryzować wielkością wariancji dopasowania s^2, która jest estymatorem wariancji danych σ^2. Dla funkcji $y(x_i)$, liniowo zależnej od m parametrów i dopasowanej do n punktów, mamy:

$$s^2 = \frac{1}{n-m} \frac{\sum_{i=1}^{n} \left\{ \left(1/\sigma_i^2\right) \left[y_i - y(x_i)\right]^2 \right\}}{(1/n)\sum_{i=1}^{n} \left(1/\sigma_i^2\right)} = \frac{1}{\nu} \sum_{i=1}^{n} w_i \left[y_i - y(x_i)\right]^2$$

gdzie czynnik $\nu = n - m$ jest liczbą stopni swobody dopasowania funkcji o m parametrach do n punktów, a czynniki wagowe dla każdego punktu wynoszą

$$w_i = \frac{1/\sigma_i^2}{\frac{1}{n}\sum_{i=1}^{n} \left(1/\sigma_i^2\right)}$$

i są równe odwrotnościom wariancji $1/\sigma_i^2$ opisującym niepewności pomiarowe dla tego punktu unormowanych do średniej z wszystkich czynników wagowych $\left(\sum w_i = n\right)$.

Wariancja dopasowania jest również scharakteryzowana przez samą wartość χ^2:

$$\chi^2 \equiv \sum_{i=1}^{n} \left\{ \frac{1}{\sigma_i^2} \left[y_i - y(x_i)\right]^2 \right\}$$

gdzie

$$y(x_i) = \sum_{j=1}^{m} a_j f_j(x_i)$$

Związek między s^2 a χ^2 najwyraźniej widać, jeżeli porównać s^2 ze zredukowana χ_ν^2:

$$\chi^2_\nu = \frac{\chi^2}{\nu} = \frac{s^2}{\left\langle \sigma_i^2 \right\rangle}$$

albo

$$\chi^2 = \nu \frac{s^2}{\left\langle \sigma_i^2 \right\rangle}$$

gdzie $\left\langle \sigma_i^2 \right\rangle$ jest ważoną średnią indywidualnych wariancji:

$$\left\langle \sigma_i^2 \right\rangle = \frac{\frac{1}{n}\sum\left(\left(\frac{1}{\sigma_i^2}\right)\sigma_i^2\right)}{\frac{1}{n}\sum\left(\frac{1}{\sigma_i^2}\right)} = \left[\frac{1}{n}\sum\frac{1}{\sigma_i^2}\right]^{-1}$$

i jest równe σ^2 w przypadku gdy wszystkie niepewności są jednakowe $\sigma_i = \sigma$.

Wariancja σ^2 charakteryzuje rozkład jakiemu podlegają wartości wielkości mierzonej – jest miarą rozrzutu wartości mierzonych – i nie może być miarą jakości dopasowania. Z drugiej strony estymator wariancji dopasowania s^2 względem dopasowanej funkcji jest miarą rozrzutu zarówno samych danych jak i jakości dopasowania. Zatem określenie χ^2 jako stosunku wariancji dopasowania s^2 do wariancji samych danych σ^2 pomnożonego przez liczbę stopni swobody robi z niej wygodną miarę jakości dopasowania.

Jeżeli dopasowana funkcja jest dobrym przybliżeniem rzeczywistej zależności, to wartość s^2 powinna zgadzać się z wartością σ^2, a wartość zredukowana χ^2_ν powinna być około jedności, $\chi^2_\nu \approx 1$. Jeżeli dopasowana funkcja nie jest właściwa dla danych punktów, to różnice $y_i - y(x_i)$ będą większe i większa będzie wariancja dopasowania dając wartość χ^2_ν większa od jedności. Wartość χ^2_ν mniejsza od 1 nie oznacza koniecznie lepszego dopasowania – jest prostym odzwierciedleniem faktu, że wartości s^2 i χ^2_ν są też zmiennymi losowymi i fluktuują od jednej serii pomiarowej do drugiej. Bardzo mała wartość χ^2_ν może oznaczać pomyłkę przy ustalaniu niepewności wartości wielkości mierzonej.

W tablicach statystycznych można znaleźć wartości dystrybuanty rozkładu χ^2 i obliczyć prawdopodobieństwo:

$$P_\chi(\chi^2;\nu) = \int_{\chi^2}^{\infty} p_\chi(x^2;\nu)\, dx^2,$$

że przypadkowy zestaw danych wylosowanych z rozkładu wyjściowego da wartość χ^2 równą lub większą od danej.

W przypadku właściwego doboru funkcji i dobrego dopasowania doświadczalna wartość χ^2_ν powinna być bliska oczekiwanej i prawdopodobieństwo $P_\chi(\chi^2;\nu)$ powinno wynosić około 0,5. Gorsze dopasowanie da powiększoną wartość χ^2_ν, a odpowiednie prawdopodobieństwo będzie mniejsze.

Trzeba pamiętać o pewnej dwuznaczności χ^2_ν, która jest zależna zarówno od danych pomiarowych i od wybranej funkcji, a zatem nawet właściwie dobrana funkcja może dać czasami dużą wartość χ^2_ν.

Współczynnik korelacji liniowej

Dane pomiarowe składają się z par zmierzonych wartości wielkości fizycznych $\{x_i, y_i\}$. Zanim dopasujemy do nich funkcje liniową (lub jakąkolwiek inną), powinniśmy zapytać, czy między mierzonymi wielkościami w ogóle występuje jakaś zależność fizyczna.

Jeżeli założymy, że wielkość Y jest wielkością zależną, to chcielibyśmy wiedzieć, czy dane dają się przedstawić przy pomocy funkcji liniowej

$$y = a\,x + b$$

Poprzednio otrzymaliśmy analityczne rozwiązanie dla najlepszej (w sensie metody minimalizacji χ^2) parametru a, który jest współczynnikiem kierunkowym dopasowanej funkcji

$$a = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - \left(\sum x_i\right)^2}$$

(czynniki wagowe zostały opuszczone dla lepszej przejrzystości wzoru). Jeżeli wielkości X i Y są niezależne od siebie, to również niezależne i nieskorelowane są wyniki pomiarów. Nie powinniśmy obserwować żadnej tendencji wzrostu (lub zmniejszania się) wartości y wraz ze wzrostem x, a współczynnik kierunkowy a wyniesie 0.

Ponieważ interesuje nas wzajemna relacja między wielkościami X i Y, to równie dobrze możemy zapytać o zależność

$$x = a'\,y + b'.$$

W tym wypadku parametry a' i b' będą miały inne wartości (i wymiary), ale jeżeli dane są skorelowane, to powinien między nimi zachodzić jakiś związek. Dla parametru a' można otrzymać rozwiązanie w postaci

$$a' = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum y_i^2 - \left(\sum y_i\right)^2}$$

i jeśli dane nie są skorelowane, to znowu współczynnik kierunkowy odwróconej zależności powinien wynosić $a' = 0$.
Jeżeli dane są zależne w sposób całkowicie jednoznaczny (całkowicie skorelowane), to powinien zachodzić związek

$$y = \frac{1}{a'}x - \frac{b'}{a'} = a\,x + b$$

oraz równość współczynników

$$\frac{1}{a'} = a \qquad -\frac{b'}{a'} = b\,.$$

W przypadku całkowitej korelacji $a\,a' = 1$. Jeżeli nie ma żadnej korelacji, to oba współczynniki są zerami i związek powyżej w ogóle nie zachodzi. Jeżeli zdefiniujemy, jako miarę korelacji liniowej, wielkość r

$$r^2 \equiv a\,a'$$

albo

$$r \equiv \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\,\sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}\,.$$

Współczynnik korelacji r przyjmuje wartości od 0, w przypadku braku korelacji, do ±1 przy całkowitej korelacji. Znak nie jest istotny dla istnienia korelacji, ważna jest natomiast wartość bezwzględna współczynnika.

Najczęściej istnienie korelacji testujemy porównując otrzymaną wartość r z rozkładem prawdopodobieństwa dla populacji, która jest całkowicie nieskorelowana. Porównanie daje nam informację, czy jest prawdopodobne, że analizowane dane mogły zostać wylosowane z populacji nieskorelowanej. Jeżeli prawdopodobieństwo przypadkowego otrzymania wartości równej lub większej od $|r|$ (lub równej lub mniejszej od $-|r|$) jest niewielkie, to mamy prawo sądzić, że nasze dane są skorelowane.

Współczynnik korelacji liniowej (w przypadku braku korelacji między zmiennymi) ma następujący symetryczny rozkład prawdopodobieństwa:

$$p_r(x;\nu) = \frac{1}{\sqrt{\pi}}\,\frac{\Gamma\left[(\nu+1)/2\right]}{\Gamma(\nu/2)}\left(1-x^2\right)^{(\nu-2)/2}$$

Tablice statystyczne podają wartości prawdopodobieństwa dla n nieskorelowanych par wartości

$$P_c(r;n) = P\left[\left(x > |r|\right) \cup \left(x < -|r|\right)\right] = 2\int_{|r|}^{1} p_r(x;n-2)\,dx$$

(W przypadku zależności liniowej liczba stopni swobody $\nu = n-2$.)

Przykład

1.
Dla danych liczbowych z przykładu pomiarów spadku napięcia wzdłuż drutu oporowego otrzymujemy

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - \left(\sum x_i\right)^2}\ \sqrt{n\sum y_i^2 - \left(\sum y_i\right)^2}}$$

$$= \frac{9\times 779{,}3 - 450{,}0\times 12{,}44}{\sqrt{9\times 28500 - (450{,}0)^2}\ \sqrt{9\times 21{,}32 - (12{,}44)^2}}$$

$$= 0{,}9994$$

W tablicach znajdujemy dla $n = 9$ wartość $P_c(0{,}898;\ 9) = 0{,}001$.
Oznacza to, że. $P_c(0{,}9994;\ 9) < 0{,}001$

2.
Dla danych liczbowych z pomiarów liczby impulsów licznika G-M w funkcji odległości preparatu otrzymujemy:

$$r = \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{\sqrt{\sum w_i \sum w_i x_i^2 - \left(\sum w_i x_i\right)^2} \times \sqrt{\sum w_i \sum w_i y_i^2 - \left(\sum w_i y_i\right)^2}}$$

$$= \frac{S_w S_{xy} - S_x S_y}{\sqrt{\left(S_w S_{xx} - (S_x)^2\right)\times\left(S_w S_{yy} - (S_y)^2\right)}}$$

$$= \frac{0{,}03570\times 81{,}02 - 0{,}1868\times 10{,}0}{\sqrt{\left(0{,}03570\times 1{,}912 - (0{,}1868)^2\right)\times\left(0{,}03570\times 3693{,}0 - (10{,}0)^2\right)}}$$

$$= 0{,}9938$$

Dla $n = 10$ w tablicach znajdujemy

$$P_c(0{,}9938;10) < P_c(0{,}872;10) = 0{,}001.$$

W obu przykładach odpowiednie prawdopodobieństwa są na tyle małe, że z dużą pewnością możemy uznać istnienie korelacji między mierzonymi wartościami.

Współczynniki korelacji liniowej między wieloma zmiennymi

Jeżeli zmienna zależna jest liniową funkcją więcej niż jednej zmiennej niezależnej,

$$y_i = a_0 + a_1 x_{i1} + a_2 x_{i2} + a_3 x_{i3} + \ldots$$

to możemy sprawdzać korelacje między $\{y_i\}$ a każdą ze zmiennych niezależnych $\{x_{ij}\}$ (pierwszy indeks oznacza numer pomiaru, a drugi zmiennej niezależnej). Nie ma znaczenia, czy x_{ij} są oddzielnymi zmiennymi, potęgami x_i, czy dowolnymi funkcjami $f_j(x_i)$.

Wprowadzimy pojęcie kowariancji z próby s_{jk}:

$$s_{jk} \equiv \frac{1}{n-1} \sum_{i=1}^{n} \left[(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k) \right]$$

gdzie odpowiednie średnie wynoszą oczywiście:

$$\bar{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij} \qquad \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}$$

(wagi są pominięte, żeby nie komplikować formy wzorów).
Przy takim podejściu estymatorem wariancji z próby j-tej zmiennej jest

$$s_j^2 \equiv s_{jj} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)^2$$

Trzeba zwrócić uwagę, że wariancje z prób są miarą szerokości przedziałów zmienności odpowiednich zmiennych i nie maja nic wspólnego z niepewnościami, z jakimi mierzymy ich wartości.

Zauważmy, że

$$s_{jk} \equiv \frac{1}{n-1}\sum_{i=1}^{n}\left[\left(x_{ij}-\bar{x}_j\right)\left(x_{ik}-\bar{x}_k\right)\right]$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}x_{ik}-\bar{x}_j x_{ik}-x_{ij}\bar{x}_k+\bar{x}_j\bar{x}_k\right)$$

$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(x_{ij}x_{ik}-\bar{x}_j x_{ik}-x_{ij}\bar{x}_k+\bar{x}_j\bar{x}_k\right)$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_{ij}x_{ik}-\sum_{i=1}^{n}\bar{x}_j x_{ik}-\sum_{i=1}^{n}x_{ij}\bar{x}_k+\sum_{i=1}^{n}\bar{x}_j\bar{x}_k\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_{ij}x_{ik}-\bar{x}_j\sum_{i=1}^{n}x_{ik}-\bar{x}_k\sum_{i=1}^{n}x_{ij}+\bar{x}_j\bar{x}_k\sum_{i=1}^{n}1\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_{ij}x_{ik}-\bar{x}_j\ n\,\bar{x}_k-\bar{x}_k\ n\,\bar{x}_j+\bar{x}_j\bar{x}_k n\right]$$

$$= \frac{1}{n-1}\left[\sum_{i=1}^{n}x_{ij}x_{ik}-n\,\bar{x}_j\bar{x}_k\right] = \frac{1}{n-1}\left[\sum_{i=1}^{n}x_{ij}x_{ik}-\frac{1}{n}\sum_{i=1}^{n}x_{ij}\sum_{i=1}^{n}x_{ik}\right]$$

Porównując to z wzorem definiującym współczynnik korelacji

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{\left(n\sum x_i^2-\left(\sum x_i\right)^2\right)\left(n\sum y_i^2-\left(\sum y_i\right)^2\right)}},$$

który po podzieleniu licznika i mianownika przez n przyjmuje postać

$$r = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left(\sum x_i^2-\frac{1}{n}\left(\sum x_i\right)^2\right)\left(\sum y_i^2-\frac{1}{n}\left(\sum y_i\right)^2\right)}}$$

możemy przez analogię zapisać

$$r_{jk} = \frac{s_{jk}}{s_j \cdot s_k}.$$

r_{jk} jest współczynnikiem korelacji liniowej z próby między dwoma dowolnymi zmiennymi x_j i x_k. Podobnie współczynnikiem korelacji między j-tą zmienną x_j a zmienną zależną y jest

$$r_{jy} = \frac{s_{jy}}{s_j \cdot s_y}.$$

W szczególnym przypadku dopasowania wielomianu $y(x) = \sum_{k=0}^{m} a_k x^k$,

kolejne zmienne x_j są potęgami zmiennej niezależnej x

$$x_{ij} = x_i^j$$

i współczynnik korelacji między zmienną zależną i j-tym składnikiem wielomianu wynosi

$$r_{jy} = \frac{s_{jy}}{s_j \cdot s_y}$$

gdzie

$$s_j^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^{2j} - \frac{1}{n}\left(\sum_{i=1}^{n} x_i^j\right)^2\right]$$

$$s_y^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2\right]$$

$$s_{jy} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^j y_i - \frac{1}{n}\sum_{i=1}^{n} x_i^j \sum_{i=1}^{n} y_i\right]$$

Jeżeli niepewności punktów pomiarowych nie są wszystkie jednakowe, to musimy uwzględnić odpowiednie wagi statystyczne w definicjach wariancji, kowariancji i współczynnika korelacji z próby. Wzory na wartości współczynników korelacji w formie

$$r_{jk} = \frac{s_{jk}}{s_j \cdot s_k}$$

pozostają niezmienione. Wzory na wariancje i kowariancje z próby muszą 9natomiast być zmodyfikowane:

$$s_{jk} \equiv \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{1}{\sigma_i^2}\left(x_{ij} - \bar{x}_j\right)\left(x_{ik} - \bar{x}_k\right)\right]}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}$$

$$s_j^2 \equiv s_{jj} \frac{\frac{1}{n-1}\sum_{i=1}^{n}\left[\frac{1}{\sigma_i^2}\left(x_{ij}-\bar{x}_j\right)^2\right]}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}.$$

Średnie $\bar{x}_j$ i $\bar{x}_k$ są też ważone

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij} w_i = \frac{\sum_{i=1}^{n}\frac{x_{ij}}{\sigma_i^2}}{\sum_{i=1}^{n}\frac{1}{\sigma_i^2}}$$

unormowanymi czynnikami wagowymi równymi

$$w_i = \frac{1/\sigma_i^2}{\frac{1}{n}\sum_{i=1}^{n}\left(1/\sigma_i^2\right)}$$