Z. HELLWIG (Wrocław)

WYZNACZANIE PARAMETRÓW REGRESJI LINIOWEJ METODĄ DWÓCH PUNKTÓW

Niech (ξ, η) będzie zmienną losową typu ciągłego. Zakładamy, że w rozkładzie tej zmiennej krzywe regresji typu pierwszego są liniami prostymi.

Równania linii regresji można więc napisać w postaci

$$y_x - \nu = a_y(x-\mu), \tag{1}$$

$$x_y - \mu = a_x(y-\nu), \tag{2}$$

gdzie μ, ν są współrzędnymi środka ciężkości populacji, a_x i a_y zaś są współczynnikami regresji. Jeżeli $\mu = 0$ i $\nu = 0$, to równania (1) i (2) przybierają prostszą postać

$$y_x = a_y x,$$

$$x_y = a_x y.$$

Linie (1) i (2) są liniami regresji typu pierwszego, tzn.

$$y_x = \mathrm{E}(\eta|\xi = x),$$

$$x_y = \mathrm{E}(\xi|\eta = y).$$

Estymując na podstawie próbki parametry występujące w (1) i (2), korzysta się na ogół z metody najmniejszej sumy kwadratów. Ujemną stroną tej metody jest to, że wymaga ona wielu żmudnych operacji rachunkowych, jak podnoszenie do kwadratu, obliczanie iloczynów, sumowanie długich kolumn liczb oraz wyciąganie pierwiastków. Poniżej proponuję inną, prostszą rachunkowo metodę wyznaczania parametrów linii regresji.

Metoda ta, zwana dalej *metodą dwóch punktów*, daje, jak zobaczymy, estymatory zgodne i nieobciążone. Efektywność ich jest mniejsza od efektywności estymatorów uzyskanych metodą klasyczną, lecz za to rachunki związane z ich wyznaczaniem są łatwiejsze. Rozkład proponowanych estymatorów jest zbieżny do rozkładu normalnego.

Przypuśćmy, że w celu oszacowania parametrów a_x i a_y, z populacji generalnej pobrano próbkę liczącą n punktów o współrzędnych (x_i, y_i) $(i = 1, 2, \ldots, n)$. Obliczamy średnie

$$\bar{x} = (x_1+x_2+\ldots+x_n)/n, \quad \bar{y} = (y_1+y_2+\ldots+y_n)/n,$$

a następnie dzielimy zbiór wszystkich punktów na dwa podzbiory, które oznaczymy liczbami rzymskimi I i II. Do podzbioru I zaliczamy punkty, których współrzędne x są większe od $\bar{x}$, do podzbioru II pozostałe punkty. Środki ciężkości tych podzbiorów wyznaczają prostą, na której leży środek ciężkości wszystkich punktów, tzn. punkt o współrzędnych $(\bar{x}, \bar{y})$.

Otóż za estymator parametru a_y proponuję wziąć współczynnik kierunkowy tej prostej. Oznaczymy go symbolem a_y, przy czym

$$a_y = (\bar{y}_{\mathrm{I}}-\bar{y})/(\bar{x}_{\mathrm{I}}-\bar{x}).$$

We wzorze tym mamy

$$\bar{y}_{\mathrm{I}} = (y_{i_1}+y_{i_2}+\ldots+y_{i_k})/k, \quad \bar{x}_{\mathrm{I}} = (x_{i_1}+x_{i_2}+\ldots+x_{i_k})/k,$$

gdzie $(x_{i_1}, y_{i_1}), \ldots, (x_{i_k}, y_{i_k})$ są to te spośród punktów (x_i, y_i), których współrzędne x_i spełniają warunek $x_i > \bar{x}$. Litera k oznacza tu liczbę punktów o odciętych większych od $\bar{x}$. Zwróćmy uwagę, że k jest zmienną losową, która może przybierać wartości $1, 2, \ldots, n-1$.

Udowodnimy, że estymator a_y jest zgodny i nieobciążony, oraz zbadamy jego efektywność.

TWIERDZENIE 1. *Współczynnik a_y jest estymatorem zgodnym współczynnika regresji α_y.*

Dowód. Mamy udowodnić, że dla dowolnego $\varepsilon > 0$

$$\mathrm{P}\{|a_y-\alpha_y| > \varepsilon\} \underset{n\to\infty}{\to} 0.$$

W dowodzie wykorzystamy następujący

LEMAT 1. *Jeśli* $\mathrm{E}(\eta|\xi = x) = \alpha x$, *to dla każdego* h

$$\mathrm{E}(\eta|x-h < \xi < x+h) = \alpha\mathrm{E}(\xi|x-h < \xi < x+h).$$

Istotnie, z definicji mamy

$$\mathrm{E}(\eta|x-h < \xi < x+h) = \int_{x-h}^{x+h}\int_{-\infty}^{\infty} yf(x, y)\Big[\int_{x-h}^{x+h}\Big(\int_{-\infty}^{+\infty} f(x, y)\,dy\Big)dx\Big]^{-1}dydx,$$

gdzie $f(x, y)$ jest gęstością dwuwymiarowego rozkładu prawdopodobieństwa.

Oznaczmy

$$C = \int_{x-h}^{x+h}\Big[\int_{-\infty}^{\infty} f(x, y)\,dy\Big]dx.$$

W takim razie jest

$$\mathrm{E}(\eta|x-h<\xi<x+h) = C^{-1}\int_{x-h}^{x+h}\int_{-\infty}^{\infty} yf(x,y)dxdy.$$

Ponieważ zachodzi równość

$$f(xy) = f(y|x)f_1(x),$$

w której $f(y|x)$ jest gęstością warunkową prawdopodobieństwa zmiennej η, a $f_1(x)$ gęstością brzegową zmiennej ξ, więc

$$\begin{aligned}
\mathrm{E}(\eta|x-h<\xi<x+h) &= C^{-1}\int_{x-h}^{x+h}\int_{-\infty}^{\infty} yf(y|x)f_1(x)dy\,dx = \\
&= C^{-1}\int_{x-h}^{x+h} f_1(x)\Big[\int_{-\infty}^{\infty} yf(y|x)dy\Big]dx = \\
&= C^{-1}\int_{x-h}^{x+h} f_1(x)\,\mathrm{E}(\eta|\xi=x)dx = \\
&= C^{-1}\int_{x-h}^{x+h} f_1(x)\alpha x\,dx = \\
&= \alpha\Big(\int_{x-h}^{x+h} xf_1(x)dx\Big)\Big/\int_{x-h}^{x+h} f_1(x)dx = \\
&= \alpha\mathrm{E}(\xi|x-h<\xi<x+h).
\end{aligned}$$

Otrzymaliśmy zatem równość

$$\mathrm{E}(\eta|x-h<\xi<x+h) = \alpha\mathrm{E}(\xi|x-h<\xi<x+h), \qquad \text{c. n. d.}$$

Z udowodnionego lematu wynika następujący

WNIOSEK.

$$\mathrm{E}(\bar{y}_{\mathrm{I}}) = \alpha\mathrm{E}(\bar{x}_{\mathrm{I}}).$$

Mamy bowiem

$$\mathrm{E}(y_1|x>\bar{x}) = \mathrm{E}(y_2|x_2>\bar{x}) = \ldots = \mathrm{E}(y_n|x_n>\bar{x}), \tag{3}$$

z definicji zaś otrzymujemy

$$\mathrm{E}(\bar{y}_{\mathrm{I}}) = \mathrm{E}\big((y_{i_1}+\ldots+y_{i_k})/k\big).$$

Wobec tego, przy ustalonym k jest

$$\begin{aligned}
&\mathrm{E}\big((y_{i_1}+\ldots+y_{i_k})/k\big) = \\
&= \mathrm{E}\Big(\frac{1}{k}(y_1+\ldots+y_k)\Big|x_i>\bar{x},\ i=1,2,\ldots,k;\bar{x}\geqslant x_i,\ i=k+1,k+2,\ldots,n\Big).
\end{aligned}$$

W takim razie

$$\mathrm{E}(\bar{y}_{\mathrm{I}}) = \mathrm{E}(y_1|x_1 > \bar{x}).$$

Analogicznie, łatwo wykazać, że

$$\mathrm{E}(\bar{x}_{\mathrm{I}}) = \mathrm{E}(x_1|x_1 > \bar{x}).$$

Dla uzasadnienia słuszności wniosku pozostaje więc do udowodnienia równość

$$\mathrm{E}(y_1|x_1 > \bar{x}) = a\mathrm{E}(x_1|x_1 > \bar{x}).$$

Równość powyższą można również napisać inaczej, a mianowicie

$$\mathrm{E}(y_1|x_1 > z) = a\mathrm{E}(x_1|x_1 > z),$$

gdzie x_1 i z są zmiennymi losowymi niezależnymi. Wystarczy bowiem warunek $x_1 > \bar{x}$ napisać w postaci $x_1 > (x_2+\ldots+x_n)/(n-1)$ i przyjąć $z = (x_2+\ldots+x_n)/(n-1)$.

Dla dowolnego rzeczywistego u mamy na podstawie lematu

$$\mathrm{E}(y_1|x_1 > u) = a\mathrm{E}(x_1|x_1 > u).$$

Oznaczając przez $g(u)$ gęstość prawdopodobieństwa zmiennej losowej z widzimy, że

$$\begin{aligned}\mathrm{E}(y_1|x_1 > \bar{x}) &= \int_{-\infty}^{\infty} \mathrm{E}(y_1|x_1 > u)g(u)du = \\ &= \int_{-\infty}^{\infty} a\mathrm{E}(x_1|x_1 > u)g(u)du = \\ &= a\int_{-\infty}^{\infty} \mathrm{E}(x_1|x_1 > u)g(u)du = \\ &= a\mathrm{E}(x_1|x_1 > z), \quad \text{c. n. d.}\end{aligned}$$

Wiadomo, że $\bar{x}$ zmierza według prawdopodobieństwa do $\mathrm{E}(x)$ i że $\bar{y}$ zmierza według prawdopodobieństwa do $\mathrm{E}(\bar{y})$. Można również pokazać, że $\bar{x}_{\mathrm{I}}$ zmierza według prawdopodobieństwa do $\mathrm{E}(\bar{y}_{\mathrm{I}})$ oraz $\bar{y}_{\mathrm{I}}$ zmierza według prawdopodobieństwa do $\mathrm{E}(\bar{y}_{\mathrm{I}})$. Możemy więc korzystać z twierdzenia Słuckiego (zob. [3], str. 282), które głosi, że jeśli zmienne losowe $\xi_n, \eta_n, \ldots, \varrho_n$ są zbieżne według prawdopodobieństwa odpowiednio do stałych $x, y, \ldots, r$, to każda wymierna funkcja $R(\xi_n, \eta_n, \ldots, \varrho_n)$ zdąża według prawdopodobieństwa do stałej $R(x, y, \ldots, r)$, jeśli tylko wielkość $R(x, y, \ldots, r)$ jest skończona.

Z twierdzenia tego wynika, że $a_y = (\bar{y}_{\mathrm{I}}-\bar{y})/(\bar{x}_{\mathrm{I}}-\bar{x})$ zdąża według prawdopodobieństwa do $\big(\mathrm{E}(\bar{y}_{\mathrm{I}})-\mathrm{E}(\bar{y})\big) / \big(\mathrm{E}(\bar{x}_{\mathrm{I}})-\mathrm{E}(\bar{x})\big)$. Na podstawie wniosku mamy

$$\frac{\mathrm{E}(y_{\mathrm{I}})-\mathrm{E}(\bar{y})}{\mathrm{E}(\bar{x}_{\mathrm{I}})-\mathrm{E}(\bar{x})} = \frac{a\mathrm{E}(\bar{x}_{\mathrm{I}})-a\mathrm{E}(\bar{x})}{\mathrm{E}(\bar{x}_{\mathrm{I}})-\mathrm{E}(\bar{x})} = a_y,$$

a więc a_y zdąża według prawdopodobieństwa do α_y, czego należało dowieść.

Obecnie wykażemy, że a_y jest estymatorem nieobciążonym.

TWIERDZENIE 2.

$$\mathrm{E}(a_y) = \alpha_y.$$

Dowód. Niech $u_1, u_2, \ldots, u_n$ będą dowolnymi liczbami rzeczywistymi. Wobec tego, jeśli $x_1 = u_1, x_2 = u_2, \ldots, x_n = u_n$, to

$$\mathrm{E}(a_y | x_1 = u_1, \ldots, x_n = u_n) = \big(\mathrm{E}(\bar{y}_\mathrm{I}) - \mathrm{E}(\bar{y})\big)/(\bar{x}_\mathrm{I} - \bar{x}).$$

Ale

$$\frac{\mathrm{E}(\bar{y}_\mathrm{I}) - \mathrm{E}(\bar{y})}{\bar{x}_\mathrm{I} - \bar{x}} =$$

$$= \frac{[\mathrm{E}(\eta|\xi = x_1) + \ldots + \mathrm{E}(\eta|\xi = x_k)]/k - [\mathrm{E}(\eta|\xi = x_1) + \ldots + \mathrm{E}(\eta|\xi = x_n)]/n}{\bar{x}_\mathrm{I} - \bar{x}},$$

a ponieważ z założenia mamy $\mathrm{E}(\eta|\xi = x) = \alpha x$, więc

$$\frac{\mathrm{E}(\bar{y}_\mathrm{I}) - \mathrm{E}(\bar{y})}{\bar{x}_\mathrm{I} - \bar{x}} = \frac{(\alpha x_1 + \ldots + \alpha x_k)/k - (\alpha x_1 + \ldots + \alpha x_n)/n}{\bar{x}_\mathrm{I} - \bar{x}} = \alpha.$$

Oczywiście jest

$$\mathrm{E}(a_y) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \mathrm{E}(a_y | x_1 = u_1, \ldots, x_n = u_n) f_1(u_1) \ldots f_1(u_n)\, du_1 \ldots du_n =$$

$$= \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} \alpha f_1(u_1) \ldots f_1(u_n)\, du_1 \ldots du_n =$$

$$= \alpha \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_1(u_1) \ldots f_1(u_n)\, du_1 \ldots du_n,$$

przy czym $f_1(u)$ oznacza gęstość brzegową zmiennej ξ.

Ponieważ mamy

$$\int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_1(u_1) \ldots f_1(u_n)\, du_1 \ldots du_n = 1,$$

więc

$$\mathrm{E}(a_y) = \alpha_y, \qquad \text{c. n. d.}$$

Przystąpimy obecnie do porównania wariancji estymatora a_y z wariancją estymatora A_y (symbol A_y oznacza współczynnik regresji z próbki, otrzymany metodą klasyczną).

Niech będzie $u_n \leqslant u_{n-1} \leqslant \ldots \leqslant u_1$. Wprowadźmy oznaczenia [1]

$$V(A_y) = \mathcal{V}(A_y | x_1 = u_1, \ldots, x_n = u_n),$$

$$V(a_y) = \mathcal{V}(a_y | x_1 = u_1, \ldots, x_n = u_n)$$

oraz

$$s_u = \sqrt{\Big(\sum_1^n (u-\overline{u})^2\Big)/n}, \qquad d_u = \Big(\sum_1^n |u-\overline{u}|\Big)/n.$$

TWIERDZENIE 3.

$$e(a_y | x_1 = u_1, \ldots, x_1 = u_n) = \frac{V(A_y)}{V(a_y)} \geqslant \frac{d^2 u}{s^2 u} \cdot \frac{n-1}{n}.$$

Zbadajmy, czemu jest równe $V(a_y)$. Mamy

$$V(a_y) = \mathcal{V}(\overline{y}_{\mathrm{I}} - \overline{y})/(\overline{x}_{\mathrm{I}} - \overline{x})^2. \tag{4}$$

LEMAT 2.

$$d_x = \Big(\sum_{i=1}^n |x_i - \overline{x}|\Big)/n = 2k(\overline{x}_{\mathrm{I}} - \overline{x})/n, \tag{5}$$

gdzie k jest liczbą x-ów większych od $\overline{x}$.

Dowód. Oznaczając $z_i = x_i - \overline{x}$ możemy napisać równość następującą:

$$\sum_{i=1}^k z_i = \sum_{j=1}^l (-z_j), \tag{6}$$

gdzie każde $z_i > 0$, a każde $z_j \leqslant 0$, przy czym $k+1 = n$.

Odchylenie przeciętne d_x jest równe wyrażeniu

$$\frac{1}{n}\sum_{i=1}^n |x_i - \overline{x}| = \frac{1}{n}\sum_{i=1}^n |z_i| = \frac{1}{n}\Big(\sum_{i=1}^k |z_i| + \sum_{i=1}^l |z_i|\Big).$$

Ze względu na (6) mamy

$$d_x = \frac{2}{n}\sum_{i=1}^k z_i = \frac{2}{n}\sum_{j=1}^l (-z_j). \tag{7}$$

Zbadajmy obecnie, czemu jest równa różnica $\overline{x}_{\mathrm{I}} - \overline{x}$. Mamy

$$\overline{x}_{\mathrm{I}} - \overline{x} = \frac{1}{k}\sum_{i=1}^k (\overline{x} + z_i) - \overline{x} = \frac{1}{k}\Big(k\overline{x} + \sum_{i=1}^k z_i\Big) - \overline{x} =$$

$$= \frac{1}{k}\Big(k\overline{x} + \sum_{i=1}^k z_i - k\overline{x}\Big) = \frac{1}{k}\sum_{i=1}^k z_i.$$

---

[1] Korzystamy tutaj z wariancji warunkowych, gdyż jak wiadomo ([2], str. 156), nie zwęża to ogólności rozważań.

Lemat zostanie udowodniony, gdy otrzymany wynik podstawimy do (7). Mamy bowiem

$$\sum_{i=1}^{k} z_i = k(\bar{x}_{\mathrm{I}}-\bar{x}),$$

skąd

$$d_x = 2k(\bar{x}_{\mathrm{I}}-\bar{x})/n, \qquad \text{c. n. d.}$$

Na mocy lematu 2 wzór (4) można również napisać w innej postaci, mianowicie

$$V(a_y) = \frac{4k^2}{n^2 d_x^2}\,\mathcal{V}(\bar{y}_{\mathrm{I}}-\bar{y}). \tag{8}$$

We wzorze tym

$$\mathcal{V}(\bar{y}_{\mathrm{I}}-\bar{y}) = \mathcal{V}\left(\frac{y_1+y_2+\ldots+y_k}{k} - \frac{y_1+y_2+\ldots+y_k+y_{k+1}+\ldots+y_n}{n}\right) =$$

$$= \mathcal{V}\left(\frac{n\sum_{r=1}^{k} y_r}{kn} - \frac{k\sum_{r=1}^{k} y_r}{kn} - \frac{k\sum_{r=k+1}^{n} y_r}{kn}\right) =$$

$$= \mathcal{V}\left((n-k)\frac{\sum_{r=1}^{r} y_r}{kn} - \frac{k\sum_{r=k+1}^{n} y_r}{kn}\right) =$$

$$= \frac{(n-k)^2}{k^2n^2}\sum_{r=1}^{k}\mathcal{V}(y_r) + \frac{k^2}{k^2n^2}\sum_{r=k+1}^{n}\mathcal{V}(y_r).$$

Ponieważ jednak mamy

$$\mathcal{V}(y_r) = \mathcal{V}(\eta|\xi = x_r) = \mathcal{V}(\alpha x+\beta+\Delta), \quad \text{gdzie} \quad \Delta = y-\alpha x-\beta, \ \beta = \nu-\alpha\mu,$$

więc

$$\mathcal{V}(y_r) = \mathcal{V}(\Delta).$$

W takim razie

$$\mathcal{V}(y_{\mathrm{I}}-\bar{y}) = \frac{(n-k)^2}{k^2n^2}\,k\mathcal{V}(\Delta) + \frac{k^2}{k^2n^2}\,(n-k)\mathcal{V}(\Delta) =$$

$$= \mathcal{V}(\Delta)\left(\frac{n^2k-2nk^2+k^3+nk^2-k^3}{k^2n^2}\right) =$$

$$= \mathcal{V}(\Delta)\left(\frac{n^2k-nk^2)}{k^2n^2}\right) = \mathcal{V}(\Delta)\left(\frac{1}{k}-\frac{1}{n}\right).$$

Wobec tego wzór (8) przybierze postać

$$V(a_y) = \frac{4k^2}{n^2 d_x^2}\,\mathcal{V}(\Delta)\left(\frac{1}{k} - \frac{1}{n}\right),$$

czyli

$$V(a_y) = \frac{\mathcal{V}(\Delta)}{nd_x^2}\left(\frac{4k}{n} - \frac{4k^2}{n^2}\right). \tag{9}$$

Zajmijmy się wariancją $V(A_y)$; otrzymujemy

$$V(A_y) = \mathcal{V}\Big(\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\Big)\Big/\Big[\sum_{i=1}^{n}(x_i-\bar{x})^2\Big]^2 =$$

$$= \mathcal{V}[(x_1-\bar{x})(y_1-\bar{y})+\ldots+(x_n-\bar{x})(y_n-\bar{y})]\Big/\Big[\sum_{i=1}^{n}(x_i-\bar{x})^2\Big]^2 =$$

$$= \frac{n-1}{n}\Big[\sum_{i=1}^{n}(x_i-\bar{x})\mathcal{V}(\Delta)\Big]\Big/\Big[\sum_{i=1}^{n}(x_i-\bar{x})^2\Big]^2 =$$

$$= \frac{n-1}{n}\,\mathcal{V}(\Delta)\Big/\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

czyli

$$V(A_y) = \frac{\mathcal{V}(\Delta)}{ns_x^2}\cdot\frac{n-1}{n}.$$

Stąd mamy

$$e(a_y|x_1 = u_1, \ldots, x_n = u_n) = \frac{\big(\mathcal{V}(\Delta)/ns_x^2\big)\big((n-1)/n\big)}{\big(\mathcal{V}(\Delta)/nd_x^2\big)(4k/n - 4k^2/n^2)} =$$

$$= \frac{d_x^2}{s_x^2}\cdot\frac{n^2}{4k(n-k)}\cdot\frac{n-1}{n}.$$

Wyrażenie

$$n^2/4k(n-k) \tag{10}$$

osiąga minimum ze względu na k, gdy $k = \frac{1}{2}n$. Podstawiając $k = \frac{1}{2}n$ do (10) otrzymujemy

$$n^2/\big(4\cdot\tfrac{1}{2}n(n-\tfrac{1}{2}n)\big) = 1,$$

a wobec tego

$$e(a_y|x_1 = u_1, \ldots, x_n = u_n) \geqslant (d_x^2/s_x^2)\big((n-1)/n\big), \quad \text{c. n. d.} \tag{11}$$

Stosunek $e = e(a_y|x_1 = u_1, \ldots, x_n = u_n)$ wariancji $V(A_y)$ i $V(a_y)$ możemy traktować jako zmienną losową ze względu na $x_1, \ldots, x_n$.

Stosując cytowane wyżej twierdzenie Słuckiego do prawej strony wzoru (11) otrzymujemy, że e zmierza według prawdopodobieństwa do D_x^2/σ_x^2, gdzie

$$D_x = \int_{-\infty}^{\infty} |x| f_1(x)\,dx, \qquad \sigma_x^2 = \int_{-\infty}^{\infty} x^2 f_1(x)\,dx.$$

Jeżeli populacja ma rozkład normalny, to e zmierza według prawdopodobieństwa do $2/\pi$.

Zajmiemy się obecnie zbadaniem rozkładu estymatora a_y. Udowodnimy następujące

TWIERDZENIE 4. *Rozkład zmiennej losowej*

$$t = \frac{(a_y - a_y)}{\sqrt{V(a_y)}}$$

jest dla $n \to \infty$ *zbieżny do rozkładu normalnego* $N(0,1)$.

Dowód. Zauważmy najpierw, że na podstawie twierdzenia 2 jest $\mathrm{E}(t) = 0$, a na podstawie wzoru (9), $V(t) = 1$. Po prostych przekształceniach, w których wykorzystuje się wzór (5), otrzymujemy

$$t = \frac{\sqrt{n}\,[(\bar{y}_1 - \bar{y})\,k/n - \frac{1}{2} a_y d_x]}{\sqrt{\mathcal{V}(\Delta)}\,\sqrt{(1-k/n)\,k/n}}. \tag{12}$$

LEMAT 3. *Jeżeli* $n \to \infty$, *to*

$$\mathrm{P}\{|k/n - [1 - F(\mu+0)]| \to 0\} = 1,$$

gdzie $F(x)$ *oznacza dystrybuantę brzegową zmiennej* ξ.

Dowód. Na podstawie twierdzenia Gliwienki ([1], str. 282) mamy

$$\mathrm{P}\{\sup_{-\infty<x<\infty} |F_n(x) - F(x)| \to 0\} = 1,$$

gdy $n \to \infty$; we wzorze tym $F_n(x)$ oznacza empiryczną dystrybuantę próbki, a $F(x)$ oznacza dystrybuantę populacji.

Wobec tego jest

$$\mathrm{P}\{\sup_{-\infty<x<\infty} |F_n(\mu) - F(\mu)| \underset{n\to\infty}{\to} 0\} = 1.$$

Ponieważ z drugiej strony

$$P\{|\bar{x}_n - \mu| \underset{n\to\infty}{\to} 0\} = 1,$$

więc ze względu na ciągłość dystrybuanty $F(x)$, otrzymujemy, że

$$\mathrm{P}\{\sup_{-\infty<x<\infty} |F_n(\bar{x}_n) - F(\mu)| \underset{n\to\infty}{\to} 0\} = 1.$$

Ale

$$k/n = 1-F_n(\bar{x}+0),$$

skąd

$$\mathrm{P}\Big\{\sup_{-\infty<x<\infty} |k/n-[1-F(\mu+0)]| \underset{n\to\infty}{\to} 0\Big\} = 1.$$

Lemat został więc udowodniony.

We wzorze (12) wartości $\mathcal{V}(\Delta)$ i a_y są stałe, a k/n i d_x zmierzają według prawdopodobieństwa do stałych.

Niech $H(t)$ będzie dystrybuantą zmiennej t; wtedy

$$H(t) = \mathrm{P}(k=1)H(t|k=1)+\ldots+\mathrm{P}(k=n-1)H(t|k=n-1),$$

gdzie $H(t|k=k_0)$ oznacza warunkową dystrybuantę zmiennej t pod warunkiem, że $k=k_0$. Niech m będzie taką liczbą naturalną, że dla każdego k, spełniającego nierówność $m<k<n-m$, jest

$$\sup|\Phi(t)-F(t|k)| < \varepsilon. \tag{13}$$

We wzorze tym mamy $\Phi(t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t} e^{-x^2/2}\,dx$, ε zaś jest dowolną liczbą rzeczywistą, większą od zera.

Nierówność (13) możemy napisać dlatego, że z definicji zmiennej t wynika, iż rozkład warunkowy $H(t|k)$, gdy n, k i $(n-k)$ rosną nieograniczenie, zmierza do $\Phi(t)$.

Istotnie, występująca w liczniku prawej strony wzoru (12) różnica średnich

$$\bar{y}_{\mathrm{I}}-\bar{y} = \frac{(n-k)\sum_{1}^{k}(y|x>\bar{x}) - k\sum_{k+1}^{n}(y|x\leqslant\bar{x})}{nk} =$$

$$= \frac{n-k}{n}\left[\frac{1}{k}\sum_{1}^{k}(y|x>\bar{x}) - \frac{1}{n-k}\sum_{k+1}^{n}(y|x\leqslant\bar{x})\right] =$$

$$= \frac{n-k}{n}[\bar{Y}_{\mathrm{I}}-\bar{Y}_{\mathrm{II}}],$$

gdzie $\bar{Y}_{\mathrm{I}}$ i $\bar{Y}_{\mathrm{II}}$ są to niezależne od siebie średnie arytmetyczne niezależnych zmiennych losowych. Jeżeli $n\to\infty$, $k\to\infty$, $n-k\to\infty$, to na podstawie centralnego twierdzenia granicznego rozkład różnicy $\bar{Y}_{\mathrm{I}}-\bar{Y}_{\mathrm{II}}$ zmierza do rozkładu normalnego, a na mocy lematu 3 wyrażenie $(n-k)/n$ zdąża według prawdopodobieństwa do stałej. Wobec tego na podstawie twierdzenia o zbieżności ([3], str. 281) mamy $H(t|k)\to\Phi(t)$.

LEMAT 4. *Dla każdego m jest*

$$[\mathrm{P}(k=1)+\ldots+\mathrm{P}(k=m)+\mathrm{P}(k=n-m)+\ldots+\mathrm{P}(k=n-1)] \to 0,$$

gdy $n \to \infty$.

Dowód. Na podstawie lematu 3 istnieje takie n_0, że dla $n > n_0$ mamy

$$\mathrm{P}\{|k/n-[1-F(\mu+0)]| < \varepsilon\} > 1-\varepsilon.$$

Dla wygody oznaczmy

$$1-F(\mu+0) = q, \quad F(\mu+0) = p;$$

będzie wtedy

$$\mathrm{P}\{k < n(q-\varepsilon) \text{ lub } k > n(q+\varepsilon)\} < \varepsilon.$$

Widać stąd, że dla spełnienia nierówności

$$\mathrm{P}\{k < m \text{ lub } k > n-m\} < \varepsilon,$$

trzeba, żeby było $n(q-\varepsilon) > m$ oraz $n(q+\varepsilon) < n-m$, czyli żeby n było większe od większej z dwóch liczb $m/(q-\varepsilon)$, $m/(p-\varepsilon)$. Wynika stąd, że zawsze można dobrać takie n, że będzie spełniona nierówność

$$\mathrm{P}\{k < m \text{ lub } k > n-m\} < \varepsilon.$$

Lemat został więc udowodniony. Niech

$$S(t) = \sum_{k=m+1}^{n-m-1} p(k)H(t|k).$$

W takim razie na mocy lematu 4 istnieje zawsze takie n_0, zależne od m i ε, że dla $n > n_0$ spełniona jest nierówność

$$|H(t)-S(t)| < \varepsilon. \tag{14}$$

Z wzorów (13) i (14) wynika nierówność

$$|\Phi(t)-H(t)| < 2\varepsilon,$$

co dowodzi twierdzenia.

Na zakończenie przytoczymy przykład liczbowy. Pozwoli on zaznajomić się z techniką znajdowania parametrów regresji metodą dwóch punktów.

Kolumna x-ów w podanej tablicy (str. 78 i 79) przedstawia produkcję prądu elektrycznego mierzoną na stykach generatorów, natomiast kolumna y-ów zawiera dane liczbowe dotyczące miału węglowego zużytego w związku z tą produkcją.

Materiał pochodzi z elektrowni hutniczej, mającej dwa turbogeneratory systemu Škoda o mocy 3100 kW każdy oraz dwa kotły wodnorurkowe systemu Duquesne, opalane pyłem węglowym.

Zużycie węgla (miału) podane jest brutto, tak jak wykazała waga kontrolna. Produkcja prądu jest również podana brutto, gdyż pomiarów prądu dokonano na zaciskach generatorów.

Używając jakichkolwiek znaków wyróżniających (w niniejszym tekście użyto znaku +) łatwo podzielić zbiory liczb zawartych w kolumnie x-ów i y-ów na podzbiory. Jeżeli po podziale okaże się, że liczność podzbiorów jest niejednakowa — należy w obliczeniach oprzeć się na podzbiorach mniej licznych, gdyż skraca to rachunki.

Tablica i związane z nią obliczenia są proste i nie wymagają dalszych wyjaśnień.

Współczynniki regresji, znalezione metodą dwóch punktów, wyrażają się liczbami

$$a_y = 0{,}694, \quad a_x = 1{,}178.$$

Przy użyciu metody klasycznej otrzymalibyśmy następujące rezultaty:

$$A_y = 0{,}716, \quad A_x = 1{,}288.$$

Wprowadzimy następującą miarę rozbieżności między współczynnikami regresji, otrzymanymi obu metodami:

$$\Delta a_y/A_y = (a_y - A_y)/A_y;$$

nazywać ją będziemy *rozbieżnością względną współczynników regresji.*

W przytoczonym przykładzie jest

$$\frac{\Delta a_y}{A_y} = \frac{0{,}022}{0{,}716} = 0{,}031; \qquad \frac{\Delta a_x}{A_x} = \frac{0{,}110}{1{,}288} = 0{,}085.$$

Wyniki uzyskane za pomocą obu metod mało się różnią między sobą (2), natomiast obliczenia wymagane przez metodę dwóch punktów są znacznie mniej uciążliwe od obliczeń, związanych ze stosowaniem metody klasycznej.

Właśnie dlatego metoda dwóch punktów może oddać znaczne usługi praktyczne, zwłaszcza w tych przypadkach, gdy badanie związku korelacyjnego oparte jest na próbkach o dużej liczności.

---

(2) Warto podkreślić, że w praktyce na ogół większe znaczenie od rozbieżności między współczynnikami regresji ma rozbieżność między rzędnymi linii regresji otrzymanymi metodą klasyczną i metodą dwóch punktów.

Oto względna miara tej rozbieżności, dotycząca linii regresji y względem x:

$$\Delta y_x/\mathcal{Y}_x = |y_x - \mathcal{Y}_x|/\mathcal{Y}_x,$$

gdzie

$$y_x = a_y x + \bar{y} - a_y \bar{x}; \qquad \mathcal{Y}_x = A_y x + \bar{y} - A_y \bar{x}.$$

Łatwo wykazać, że jest $\Delta y_x/\mathcal{Y}_x \leqslant \Delta a_y/A_y$.

TABLICA 1

L. p.	x dziesiątków ton miesięcznie	y dziesiątków tys. kWh miesięcznie	$x > \bar{x}$	$y\|x > \bar{x}$	$y > \bar{y}$	$x\|y > \bar{y}$
1	183	175				
2	184	172				
3	180	168				
4	164	156				
5	177	190+			190	177
6	159	160				
7	147	142				
8	151	153				
9	164	149				
10	122	128				
11	167	167				
12	188	172				
13	180	162				
14	156	160				
15	163	154				
16	175	160				
17	173	179				
18	158	144				
19	190	193+			193	190
20	180	169				
21	196+	177	196	177		
22	206+	190+	206	190	190	206
23	199+	186+	199	186	186	199
24	201+	180+	201	180	180	201
25	207+	182+	207	182	182	207
26	209+	190+	209	190	190	209
27	184	164				
28	165	164				
29	142	149				
30	116	133				
31	147	164				
32	175	168				
33	197+	176	197	176		
34	202+	186+	202	186	186	202
35	189	183+			183	189
36	190	176				
37	180	191+			191	180
38	170	167				
39	182	161				
40	189	180+			180	189
41	213+	191+	213	191	191	213
42	301+	264+	301	264	264	301
43	225+	202+	225	202	202	225
44	234+	214+	234	214	214	234

TABLICA 1 (cd.)

L. p.	x dziesiątków ton miesięcznie	y dziesiątków tys. kWh miesięcznie	$x > \bar{x}$	$y\|x > \bar{x}$	$y > \bar{y}$	$x\|y > \bar{y}$
45	203+	184+	203	184	184	203
46	192+	189+	192	189	189	192
47	191+	179	191	179		
48	146	155				
49	193+	173	193	173		
50	187	166				
51	187	182+			182	187
52	212+	205+	212	205	205	212
53	251+	216+	251	216	216	251
54	220+	201+	220	201	201	220
55	180	164				
56	207+	191+	207	191	191	207
57	190+	173	190	173		
58	185	174				
59	186	170				
60	181	166				
61	192+	179	192	179		
62	203+	191+	203	191	191	203
63	277+	247+	277	247	247	277
64	299+	257+	299	257	257	299
65	215+	206+	215	206	206	215
66	200+	188+	200	188	188	200
67	192+	167	192	167		
68	187	183+			183	187
69	194+	190+	194	190	190	194
70	194+	182+	194	182	182	194
71	190	178				
72	278+	241+	278	241	241	278
Razem	13712	12888	6693	6097	6175	6641
	$\bar{x} = 190{,}4$	$\bar{y} = 179$	$\bar{x}_{\mathrm{I}} = 215{,}9$	$\bar{y}_{\mathrm{I}} = 196{,}7$		

Otrzymujemy

$$a_y = \frac{196{,}7 - 179}{215{,}8 - 190{,}4} = \frac{17{,}7}{25{,}4} = 0{,}694;$$

podobnie

$$a_x = \frac{214{,}2 - 190{,}4}{199{,}2 - 179} = \frac{23{,}8}{20{,}2} = 1{,}178,$$

gdzie $214{,}2 = 6641/31$ oraz $199{,}2 = 6175/31$.

Prace cytowane

[1] Б. В. Гнеденко, *Курс теории вероятностей*, Москва-Ленинград 1950.
[2] M. G. Kendall, *The advanced theory of statistics*, II, London 1946.
[3] К. Крамер, *Математические методы статистики*, Москва 1948.

Praca wpłynęła 26. 3. 1955

З. ГЕЛЛВИГ (Вроцлав)

ОПРЕДЕЛЕНИЕ ЛИНЕЙНОЙ РЕГРЕСИИ МЕТОДОМ ДВУХ ТОЧЕК

РЕЗЮМЕ

Чтобы определить расположение линии регрессии η относительно ξ методом двух точек, требуется найти центр тяжести точек, абсциссы которых больше (или меньше) $\bar{x}$.

Расположение линии регрессии η относительно ξ определяется при помощи центра тяхести всех точек и центра тяжести точек, абсциссы которых больше (или меньше) $\bar{x}$. Подобным образом определяется расположение линии регрессии ξ относительно η.

Этим методом получены следующие коэффициенты регрессии:

$$a_y = (\bar{y}_I - \bar{y})/(\bar{x}_I - \bar{x}), \qquad a_x = (\bar{x}_I - \bar{x})/(\bar{y}_I - \bar{y}).$$

В статье доказано, что эти эстиматоры состоятельны и несмещены. Их относительная эффективность, при предположении, что распределение (ξ, η) нормальное, стремится по вероятности к $2/\pi$.

Распределение эстиматоров при возрастании выборки стремится к нормальному распределению.

Доказано автором, что

$$d_x = \Big(\sum_{i=1}^{n} |x_i - \bar{x}|\Big)/n = 2k|\bar{x}_I - \bar{x}|/n.$$

Полученная формула значительно упрощает исчисления связанные с вычислением среднего отклонения, особенно тогда, когда при вычислении отклонения используэтся упорядоченный ряд.

Определение параметров линии регрессии методом двух точек, благодаря своей простоте, может иметь большое практическое значение, главным образом в тех случаях, когда оцена коэффициентов регрессии основывается на большой выборке.

Z. HELLWIG (Wrocław)

DETERMINING LINEAR REGRESSION PARAMETERS BY MEANS OF THE TWO POINT METHOD

SUMMARY

In order to determine the position of the line of regression of η with respect to ξ by the two point method we determine the centre of gravity of those points whose abscissae are greater (or less) than $\bar{x}$.

The position of the line of regression of η with respect to ξ is determined by means of the centre of gravity of all points and the centre of gravity of those points whose abscissae are greater (less) than $\bar{x}$. In the same way we determine the position of the line of regression of ξ with respect to η.

The regression coefficients obtained by this method are expressed by the formulae

$$a_y - (\bar{y}_I - \bar{y})/(\bar{x}_I - \bar{x}), \qquad a_x = (\bar{x}_I - \bar{x})/(\bar{y}_I - \bar{y}).$$

The author shows that there estimators are consistent and unbiased. Under the assumption that the distribution (ξ, η) is normal, their relative efficiency is convergent in probability, to $2/\pi$. With the growth of the sample the distribution of the estimators tends to a normal distribution.

In examining the efficiency of the estimators the author proves the following lemma:

$$d_x = \Big(\sum_{i=1}^{n} |x_i - \bar{x}|\Big)/n = 2k|\bar{x}_I - \bar{x}|/n.$$

The formula deduced greatly simplifies calculations connected with computing the mean deviation, particularly if we make use of an ordered series in calculating the deviation.

Owing to its simplicity, the method of determining regression line parameters may be of considerable practical value particular in those cases where the estimation of regression coefficients is made of the basis of a large sample.