EKONOMETRIA ECONOMETRICS 33 • 2011

ISSN 1507-3866

Andrzej Wilkowski
Uniwersytet Ekonomiczny we Wrocławiu

ZASTOSOWANIE WSPÓŁCZYNNIKA ZALEŻNOŚCI PROSTOLINIOWEJ I WIELOŚREDNIEJ W ANALIZIE DANYCH

Streszczenie: Zdefiniowano współczynnik zależności prostoliniowej rozumiany jako kosinus kąta, pod jakim przecinają się proste regresji. Podobnie jak klasyczny współczynnik korelacji współczynnik zależności prostoliniowej jest asymptotycznie normalny. Tak jak w przypadku prostych regresji można zdefiniować pojęcie stożkowych regresji. Jest to przykład współczynnika zależności nieliniowej, który można określić, wychodząc od współczynnika zależności prostoliniowej. Dalej przedstawiono wielośrednią, uogólnienie klasycznego pojęcia wartości oczekiwanej zmiennej losowej. Średnia może być uważana za aproksymację średniokwadratową zmiennej losowej jednym punktem. Wielośrednia jest aproksymacją zmiennej więcej niż jednym punktem jednocześnie. Przy definiowaniu wielośredniej korzysta się ze standardowej metody momentów oraz faktów z teorii wielomianów ortogonalnych.

Słowa kluczowe: współczynnik zależności prostoliniowej, asymptotyczna normalność, wielośrednia.

1. Współczynnik zależności prostoliniowej

Najczęściej używanym typem współczynnika korelacji jest tzw. współczynnik korelacji ***r* Pearsona**, nazywany również **współczynnikiem korelacji liniowej**. Współczynnik korelacji liniowej Pearsona (dalej nazywany po prostu współczynnikiem korelacji) wymaga, aby dwie zmienne zostały zmierzone co najmniej na skali przedziałowej. Określa on stopień „proporcjonalnych" powiązań wartości dwóch zmiennych. Wartość korelacji (współczynnik korelacji) nie zależy od jednostek miary, w jakich wyrażamy badane zmienne, np. korelacja pomiędzy wzrostem i ciężarem będzie taka sama bez względu na to, w jakich jednostkach (cale i funty czy centymetry i kilogramy) wyrazimy badane wielkości. Określenie „proporcjonalne" oznacza zależne liniowo, tzn. że korelacja jest silna, jeśli może być „opisana" za pomocą linii prostej (nachylonej do góry lub na dół).

Przypomnijmy, że **współczynnikiem korelacji liniowej *r*** zmiennych losowych X i Y nazywamy wielkość

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}.$$

Oczywiście $-1 \leq r \leq 1$, $r(X,Y) = r(Y,X), r(X,Y) = r(mX+n,Y), o\ ile\ m \neq 0$.

Przedstawia on ważną charakterystykę rozkładu wektora losowego (*X*, *Y*). Główne jego własności są ściśle związane z dwiema prostymi regresji

$$\frac{y - E(Y)}{\sqrt{Var(Y)}} = r\frac{x - E(X)}{\sqrt{Var(X)}},$$
$$\frac{y - E(Y)}{\sqrt{Var(Y)}} = \frac{1}{r}\frac{x - E(X)}{\sqrt{Var(X)}},$$

które są prostymi najlepszej zgodności, w sensie metody najmniejszych kwadratów, z masą prawdopodobieństwa w rozkładzie zmiennej (*X*, *Y*) [Cramer 1958]. Miarami zgodności tych prostych są poniższe wyrażenia:

$$\min_{a,b\in\mathbb{R}} E(Y - b - aX)^2 = Var(Y)(1 - r^2),$$

$$\min_{a,b\in\mathbb{R}} E(X - b - aY)^2 = Var(X)(1 - r^2).$$

Widać z tego, że każda zmienna ma wariancję zmniejszoną w stosunku $(1 - r^2) : 1$ wskutek odjęcia od niej jej najlepszej średniokwadratowej liniowej oceny wyrażonej w zależności od drugiej zmiennej. Współczynnik *r* można zatem uważać za miarę stopnia liniowości wykazywanej przez rozkład wektora losowego (*X*, *Y*). Stopień ten osiąga wartość największą, gdy $|r| = 1$, a cała masa prawdopodobieństwa jest rozpostarta na prostej. Przypadek przeciwny zachodzi, gdy $r = 0$, wtedy nie można zmniejszyć wariancji jakiejkolwiek zmiennej losowej przez odjęcie funkcji liniowej drugiej zmiennej.

Zauważmy, że mając proste regresji zmiennych losowych *X* oraz *Y*:

$$y = a_1 x + b_1,$$
$$x = a_2 y + b_2,$$

możemy także wyznaczyć **współczynnik korelacji liniowej *r*** ; mianowicie:

$$r^2(X,Y) = |a_1 a_2|.$$

Zdefiniujemy obecnie **współczynnik zależności prostoliniowej *k*** zmiennych *X*, *Y* [Antoniewicz 1988]. Będziemy go rozumieli jako kosinus kąta, pod jakim przecinają się proste regresji. Po łatwych przekształceniach otrzymujemy:

$$k(X,Y) = \cos\alpha = \frac{a_1 + a_2}{\sqrt{a_1^2 + 1}\sqrt{a_2^2 + 1}},$$

gdzie α jest kątem przecięcia prostych regresji.

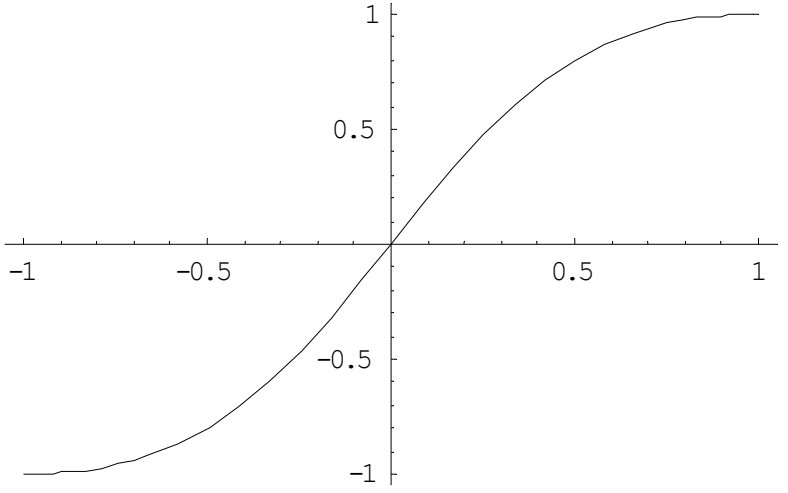
Możemy także napisać:

$$k(Var(X), Var(Y), r) = \frac{\big(Var(X) + Var(Y)\big)r}{\sqrt{Var(X) + r^2 Var(Y)}\,\sqrt{Var(Y) + r^2 Var(X)}}. \quad (1)$$

Z powyższego widać, że **współczynnik zależności prostoliniowej *k*** jest równy jeden, gdy między zmiennymi jest dokładna zależność liniowa, jeśli zaś $k = 0$, to takiej zależności nie ma. Oczywiście $k^2 = 1$ tylko wtedy, gdy $r^2 = 1$, oraz $k = 0$, gdy $r = 0$. Wartości pośrednie nie są jednak przyjmowane jednoznacznie. Może się zdarzyć, że przy ustalonej wielkości współczynnika r otrzymamy różne wartości współczynnika k (w zależności od wariancji). Rozpatrzmy teraz unormowane zmienne losowe. Wtedy wzór (1) przyjmie postać:

$$k(r) = \frac{2r}{1 + r^2}\ , r \in [-1\,, 1]\,.$$

Rysunek 1 przedstawia wykres tej funkcji.



Rys. 1. Wykres funkcji $k(r)$

Źródło: opracowanie własne.

Można wyznaczyć maksymalną różnicę między współczynnikami k oraz r. Okazuje się, że:

$$\max_{Var(X),Var(Y)>0,r\in[-1,1]} |k(Var(X), Var(Y), r) - r| = \frac{\sqrt{10\sqrt{5} - 22}}{2}.$$

Maksimum jest osiągane dla $r = \pm\sqrt{\sqrt{5} - 2}$ oraz Var(X) = Var(Y). Dowód tego faktu można znaleźć w pracy [Wilkowski 1994].

Na zakończenie tego punktu zwróćmy uwagę na fakt, że współczynnik korelacji liniowej *r* jest również kosinusem kąta, ale między innymi wektorami.

2. Asymptotyczna normalność miar zależności liniowej

Jednym z ważniejszych rodzajów zbieżności według rozkładu jest zbieżność do rozkładu normalnego. Ciąg zmiennych losowych (X_n) zbiega według rozkładu do $N(m, s^2)$, $s > 0$, jeżeli równoważnie ciąg $((X_n - m)/s)$ zbiega według rozkładu do $N(0,1)$. Ogólniej mówimy, że **ciąg zmiennych losowych (X_n) jest asymptotycznie normalny o średniej m_n i wariancji s_n^2**, jeżeli $s_n^2 > 0$ dla dostatecznie dużych *n* oraz

$$\frac{X_n - m_n}{s_n} \xrightarrow{d} N(0,1)\,.$$

Zapisujemy to jako: X_n jest $AN(m_n, s_n^2)$. Oczywiście ciągi (m_n) oraz (s_n) są ciągami stałych. Liczby te nie muszą być jednak średnią i odchyleniem standardowym zmiennej losowej X_n; zmienna ta nie musi mieć ani średniej, ani odchylenia standardowego. Zauważmy, że jeżeli X_n jest $AN(m_n, s_n^2)$, to nie wynika stąd, że ciąg (X_n) w ogóle zbiega według rozkładu. Mamy jednak zawsze

$$\sup_t |P(X_n \le t) - P(N(m_n, s_n^2) \le t)| \to 0\,, n \to \infty\,.$$

Chcąc zatem obliczać prawdopodobieństwa, można traktować X_n jako zmienną losową $N(m_n, s_n^2)$ [Serfling 1999].

Niech (X_1, Y_1), …, (X_n, Y_n) będą niezależnymi obserwacjami, o jednakowym rozkładzie, z pewnego rozkładu dwuwymiarowego (wektor (X_1, Y_1) ma taki sam rozkład jak wektor losowy (X, Y)). Jak pamiętamy, współczynnikiem korelacji liniowej zmiennych losowych *X* i *Y* jest wielkość

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}\,.$$

Jego próbkowy odpowiednik ma postać

$$\hat{r}_n = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}\sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}\,, \tag{2}$$

gdzie $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$, $\overline{Y} = \frac{1}{n}\sum_{i=1}^n Y_i$.

W poprzednim punkcie został wprowadzony współczynnik zależności prostoliniowej *k* zmiennych losowych *X* oraz *Y*, rozumiany jako kosinus kąta, pod jakim przecinają się proste regresji tych zmiennych. W dalszym ciągu (X_1, Y_1), …, (X_n, Y_n) będą niezależnymi obserwacjami o jednakowym rozkładzie, z pewnego rozkładu dwuwymiarowego (wektor (X_1, Y_1) ma taki sam rozkład jak wektor losowy (X, Y)).

Na podstawie wzorów (1) i (2) wnioskujemy, że próbkowy odpowiednik współczynnika *k* jest postaci:

$$\hat{k}_n = \frac{(\sum_{i=1}^n (X_i - \overline{X})^2 + \sum_{i=1}^n (Y_i - \overline{Y})^2)\,\hat{r}_n}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2 + \hat{r}_n^2 \sum_{i=1}^n (Y_i - \overline{Y})^2}\sqrt{\hat{r}_n^2 \sum_{i=1}^n (X_i - \overline{X})^2 + \sum_{i=1}^n (Y_i - \overline{Y})^2}}.$$

Twierdzenie. Niech wektor $\mathbf{V} = \left(\overline{X}, \overline{Y}, \frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n Y_i^2, \frac{1}{n}\sum_{i=1}^n X_i\ Y_i\right)$, funkcja $g\colon \mathbb{R}^5 \to \mathbb{R}$ będzie określona wzorem

$$g(z_1, z_2, z_3, z_4, z_5) = \frac{(z_5 - z_1 z_2)\left(\sqrt{\frac{z_3 - z_1^2}{z_4 - z_2^2}} + \sqrt{\frac{z_4 - z_2^2}{z_3 - z_1^2}}\right)}{\sqrt{z_3 - z_1^2 + \frac{(z_5 - z_1 z_2)^2}{z_3 - z_1^2}}\sqrt{z_4 - z_2^2 + \frac{(z_5 - z_1 z_2)^2}{z_4 - z_2^2}}}.$$

Wówczas

$$\hat{k}_n \text{ jest } AN(k, n^{-1}\boldsymbol{\delta S \delta^T}),$$

gdzie **S** jest macierzą kowariancji wektora (X, Y, X^2, Y^2, XY), a wektor

$$\boldsymbol{\delta} = \left(\frac{\partial g}{\partial z_1}|\boldsymbol{z} = E(\boldsymbol{V}), \ldots, \frac{\partial g}{\partial z_5}|\boldsymbol{z} = E(\boldsymbol{V})\right).$$

Dowód tego faktu znajdziemy w pracy [Wilkowski 2008].

Na zakończenie tego punktu zauważmy, że współczynnik zależności prostoliniowej może być punktem wyjścia do konstrukcji innych miar zależności. W tym celu wystarczy zdefiniować krzywe regresji, a kosinus kąta, pod jakim się one przecinają, traktować jako współczynnik zależności względem tej klasy krzywych.

3. Współczynnik zależności stożkowej

W poprzednim punkcie zauważyliśmy, że bliska zeru wartość współczynnika *k* bądź *r* oznacza tylko słabą zależność liniową, albo jej brak, między zmiennymi. Mogą one natomiast zależeć od siebie nieliniowo. Naturalnym rozszerzeniem klasy funkcji liniowych są krzywe stożkowe, czyli zbiory punktów powstałych na przecięciu stożka (ściślej powierzchni stożkowej) i płaszczyzny. W zależności od kąta przecięcia należą do nich: elipsa (w szczególnym przypadku okrąg), parabola (szczególnie półprosta, w zasadzie niebędąca stożkową), hiperbola (szczególnym przypadkiem jest para półprostych o wspólnym początku, także niezaliczana do stożkowych). Do jednoznacznego wyznaczenia stożkowej wystarczy także pięć punktów płaszczyzny, z których żadne trzy nie leżą na jednej prostej. Krzywe stożkowe są nazywane inaczej **krzywymi drugiego stopnia**, gdyż można je w kartezjańskim układzie współ-

rzędnych opisać równaniem algebraicznym drugiego stopnia względem obu zmiennych x i y:

$$ax^2 + bxy + cy^2 + dx + ey + f = 0\ ,$$

gdzie liczby rzeczywiste a, b, c nie są równocześnie równe zeru.

Analogicznie jak w przypadku prostych regresji wprowadzimy teraz pojęcie stożkowych regresji. Załóżmy, że zmienne losowe X oraz Y mają dodatni i skończony czwarty moment ($0 < E(X^4), E(Y^4) < \infty$). Niech liczby rzeczywiste a_1, b_1, c_1, d_1, f_1, a_2, b_2, c_2, e_2, f_2 spełniają poniższe równości (realizują odpowiednie minima średniokwadratowe):

$$\min_{a,b,c,d,f\in R} E(aX^2 + bXY + cY^2 + dX - Y + f\ )^2$$
$$= E(a_1X^2 + b_1XY + c_1Y^2 + d_1X - Y + f_1)^2\ ,$$
$$\min_{a,b,c,e,f\in R} E(aX^2 + bXY + cY^2 - X + eY + f\ )^2$$
$$= E(a_2X^2 + b_2XY + c_2Y^2 - X + e_2Y + f_2)^2\ .$$

Stożkową regresji względem zmiennej *x* [Wilkowski 1993] nazywamy krzywą określoną wzorem:

$$a_1x^2 + b_1xy + c_1y^2 + d_1x - y + f_1 = 0. \tag{3}$$

Stożkową regresji względem zmiennej y [Wilkowski 1993] nazywamy krzywą określoną wzorem:

$$a_2x^2 + b_2xy + c_2y^2 - x + e_2y + f_2 = 0. \tag{4}$$

Zauważmy, że powyższe równania różnią się tylko częścią liniową. Można to uzasadnić analogią do prostych regresji (zaburzenie w jednym z nich części liniowej, a w drugim nieliniowej wydaje się niecelowe, natomiast możliwość zaburzenia w obu równaniach składowej kwadratowej jest do rozważenia).

Parametry stożkowych regresji spełniają poniższe układy równań [Wilkowski 1993]:

$$\begin{gathered}
E(X^4)\,a_1 + E(X^3Y)\,b_1 + E(X^2Y^2)\,c_1 + E(X^3)\,d_1 + E(X^2)\,f_1 = E(X^2Y)\\
E(X^3Y)\,a_1 + E(X^2Y^2)\,b_1 + E(XY^3)\,c_1 + E(X^2Y)\,d_1 + E(XY)\,f_1 = E(XY^2)\\
E(X^2Y^2)\,a_1 + E(XY^3)\,b_1 + E(Y^4)\,c_1 + E(XY^2)\,d_1 + E(Y^2)\,f_1 = E(Y^3)\\
E(X^3)\,a_1 + E(X^2Y)\,b_1 + E(XY^2)\,c_1 + E(X^2)\,d_1 + E(X)\,f_1 = E(XY)\\
E(X^2)\,a_1 + E(XY)\,b_1 + E(Y^2)c_1 + E(X)d_1 + f_1 = E(Y)\ ,
\end{gathered} \tag{5}$$

$$\begin{gathered}
E(X^4)\,a_2 + E(X^3Y)\,b_2 + E(X^2Y^2)\,c_2 + E(X^2Y)\,e_2 + E(X^2)\,f_2 = E(X^3)\\
E(X^3Y)\,a_2 + E(X^2Y^2)\,b_2 + E(XY^3)\,c_2 + E(XY^2)\,e_2 + E(XY)\,f_2 = E(X^2Y)\\
E(X^2Y^2)\,a_2 + E(XY^3)\,b_2 + E(Y^4)\,c_2 + E(Y^3)\,e_2 + E(Y^2)\,f_2 = E(XY^2)\\
E(X^2Y)\,a_2 + E(XY^2)\,b_2 + E(Y^3)\,c_2 + E(Y^2)\,e_2 + E(Y)\,f_2 = E(XY)\\
E(X^2)\,a_2 + E(XY)\,b_2 + E(Y^2)\,c_2 + E(Y)\,e_2 + f_2 = E(X)\ .
\end{gathered} \tag{6}$$

Współczynnikiem zależności stożkowej k_s [Wilkowski 1993] zmiennych losowych X i Y nazywamy wielkość

$$k_s = cos\ \alpha\ , \tag{7}$$

gdzie α jest kątem przecięcia stożkowych regresji (3), (4), liczonym w punkcie przecięcia leżącym najbliżej, w sensie odległości euklidesowej, punktu ($E(X)$, $E(Y)$).

Oczywiście wzór (7) można także zapisać w postaci:

$$k_s = \frac{1+m_1 m_2}{\sqrt{1+m_1^2}\sqrt{1+m_2^2}}, \tag{8}$$

gdzie m_1, m_2 są współczynnikami kierunkowymi stycznych w punkcie przecięcia, będącym najbliżej punktu ($E(X)$, $E(Y)$), krzywych regresji (3), (4).

W przypadku gdy sto*żkowe regresji się nie przecinają lub $k_s = 0$, nie ma zależności stożkowej między zmiennymi X, Y*. Jeżeli $k_s^2 = 1$, mamy do czynienia z dokładną zależnością stożkową. Jak wiadomo, stożkowe regresji (3), (4) mogą przecinać się w kilku punktach. Do obliczeń wykorzystujemy punkt przecięcia będący najbliżej punktu ($E(X)$, $E(Y)$). Wydaje się, że kosinus liczony właśnie w tym punkcie daje najwięcej informacji (pozostałe punkty przecięcia mogą posłużyć do wyznaczania lokalnych współczynników zależności).

Przykład 1. Współczynnik zależności parabolicznej

Jednym z ważniejszych pojęć stosowanych w ekonometrycznej analizie kosztów jest tzw. funkcja kosztów. Po odrzuceniu składnika losowego funkcja ta często jest parabolą. Załóżmy zatem, że stożkowe regresji (3), (4) są parabolami o osiach symetrii równoległych do prostej OY. Wtedy ich równania redukują się do postaci

$$y = a_1x^2 + d_1x + f_1 \text{ oraz } x = a_2x^2 + e_2y + f_2,$$

a parametry spełniają zależność

$$\min_{a,d,f\in R} E(aX^2 + dX - Y + f)^2 = E(a_1X^2 + d_1X - Y + f_1)^2\ ,$$

$$\min_{a,e,f\in R} E(aX^2 - X + eY + f)^2 = E(a_2X^2 - X + e_2Y + f_2)^2\ .$$

Układy równań (5), (6) mają wtedy postać:

$$E(X^4)\ a_1 + E(X^3)\ d_1 + E(X^2)\ f_1 = E(X^2Y)$$
$$E(X^3)\ a_1 + E(X^2)\ d_1 + E(X)\ f_1 = E(XY)$$
$$E(X^2)\ a_1 + E(X)d_1 + f_1 = E(Y)\ ,$$

$$E(X^4)\ a_2 + E(X^2\ Y)\ e_2 + E(X^2)\ f_2 = E(X^3)$$
$$E(X^2Y)\ a_2 + E(Y^2)\ e_2 + E(Y)\ f_2 = E(XY)$$
$$E(X^2)\ a_2 + E(Y)\ e_2 + f_2 = E(X).$$

Parabole regresji przecinają się w punktach (x_1, y_1) , (x_2, y_2) , gdzie

$$x_{1,2} = \frac{(1 - d_1 e_2) \pm \sqrt{(d_1 e_2 - 1)^2 - 4(a_1 e_2 + a_2)(f_1 e_2 + f_2)}}{2(a_1 e_2 + a_2)}.$$

Punkt leżący bliżej punktu $(E(X), E(Y))$ oznaczmy jako (x_0, y_0). Na podstawie wzoru (8) możemy napisać

$$k_s = \frac{1 + (2a_1 x_0 + d_1)(\frac{-2a_2 x_0 + 1}{e_2})}{\sqrt{1 + (2a_1 x_0 + d_1)^2}\sqrt{1 + (\frac{-2a_2 x_0 + 1}{e_2})^2}}.$$

Przypuśćmy, że mamy dane fikcyjne jak w poniższej tabeli.

X	0	1	2	3	4	5	6	7	8	9	10	11	12
Y	0	11	26	29	32	35	36	35	32	26	18	11	0

Parabole regresji, względem x oraz względem y, mają postać:

$$y = -0{,}9958\, x^2 + 11{,}7424\, x + 1{,}6392,$$

$$y = -1{,}0033\, x^2 + 10{,}913\, x + 7{,}223.$$

4. Wielośrednia

Ważnymi charakterystykami liczbowymi, wykorzystywanymi w badaniach statystycznych i w teorii prawdopodobieństwa, są momenty zmiennej losowej. Zmienną losową X rozumiemy jako funkcję mierzalną, określoną na standardowej przestrzeni probabilistycznej (Ω, F, P), o wartościach rzeczywistych. Wtedy **momenty** $m_1, m_2, \ldots$ można wyznaczyć za pomocą wzorów :

$$m_k = E(X^k) = \int_{\Omega} X^k(\omega) P(d\omega),$$

gdzie $k = 1, 2, \ldots$, a kolejne całki są bezwzględnie zbieżne.

Pierwszy moment $m_1 = E(X)$ nazywamy **wartością oczekiwaną, średnią, lub przeciętną** zmiennej losowej. Kombinacja pierwszego i drugiego momentu, zdefiniowana następująco

$$V_1(X) = E(X - E(X))^2 = m_2 - m_1^2$$

jest **wariancją** zmiennej losowej. Wielość ta mierzy odchylenie funkcji X od wartości średniej. Wielomian $X - E(X)$ ma własność minimalizowania średniokwadratowej normy:

$$\min_{a\in R} E(X-a)^2 = E(X-E(X))^2 = V_1(X).$$

Można zatem powiedzieć, że wartość oczekiwana jest najlepszym, w sensie normy kwadratowej, przybliżeniem jednopunktowym danej zmiennej losowej.

Najstarszą ogólną metodą tworzenia ocen parametrów rozkładu za pomocą zbioru wartości w próbie jest właśnie metoda momentów wprowadzona przez K. Pearsona i szeroko stosowana przez niego i jego szkołę [Cramer 1958]. Polega ona na przyrównywaniu pewnej liczby momentów w próbie do odpowiednich momentów rozkładu, będących funkcjami nieznanych parametrów. Rozwiązując uzyskane równania względem parametrów, otrzymujemy szukane ich oceny. Metoda ta w praktyce prowadzi często do stosunkowo prostych rachunków.

W dalszym ciągu zakładamy, że zmienna losowa X ma gęstość f_X. Wtedy kolejne momenty można obliczyć na podstawie wzorów:

$$m_k = \int_{-\infty}^{+\infty} x^k f_X(x)dx,$$

$k = 1, 2, \ldots$, gdy oczywiście całki te są bezwzględnie zbieżne.

Istotne znaczenie mają maksima funkcji gęstości f_X. W statystyce charakterystyki te nazywamy **modalnymi** zmiennej losowej X. Wyznaczają one punkty koncentracji masy prawdopodobieństwa. W przypadku gęstości jednomodalnej dobrym przybliżeniem mody jest wartość oczekiwana $E(X)$ (gdy funkcja f_X jest symetryczna, obie te wielkości się pokrywają).

Niech zmienna losowa X ma teraz skończone momenty zwykłe rzędu $2n - 1$:

$$E(X^k) = m_k < \infty, \quad k = 1, 2, \ldots, 2n-1.$$

Wtedy wielomian p_n, minimalizujący normę:

$$\min_{a,b,\ldots,c\in R} E(X^n + aX^{n-1} + bX^{n-2} + \ldots + c)^2 = E(p_n(X))^2,$$

ma postać [Cramer 1958]

$$p_n(x) = K\begin{vmatrix} 1 & m_1 & \ldots & m_n \\ \ldots & \ldots & \ldots & \ldots \\ m_{n-1} & m_n & \ldots & m_{2n-1} \\ 1 & x & \ldots & x^n \end{vmatrix}, \text{ gdzie } K \neq 0. \tag{9}$$

Ponieważ p_n jest wielomianem ortogonalnym stopnia n [Szego 1975], więc:

$$p_n(x) = (x - s_1) \ldots (x - s_n), \text{ gdzie } s_1 < \ldots < s_n .$$

Uporządkowaną *n*-kę (s_1, …, s_n) nazywamy *n*-średnią (wielośrednią) zmiennej losowej X [Antoniewicz 2005]. Wektor ten jest aproksymacją zmiennej losowej *n* punktami. Analogonami wariancji i odchylenia standardowego będą wyrażenia:

$$V_n(X) = E((X - s_1)...(X - s_n))^2,$$

$$\sqrt[2n]{\mathbf{V_n(X)}} = \sqrt[2n]{\mathbf{E((X - s_1) ... (X - s_n))^2}}\ . \tag{10}$$

Charakterystyki te mierzą jednoczesne odchylenie zmiennej losowej X od *n* miejsc koncentracji prawdopodobieństwa. Inne charakterystyki związane z wielośrednią można znaleźć w pracach: [Antoniewicz, Wilkowski 2004; Antoniewicz 2005].

Przy analizie danych liczbowych standardowo wyznacza się wartość średnią, wariancję i odchylenie standardowe zmiennej losowej *X*. Można jednak pójść krok dalej. W tym celu należy najpierw obliczyć momenty zwykłe wyższych rzędów, następnie kolejno wyznaczyć 2-średnią (s_1, s_2), 3-średnią (s_1, s_2, s_3) itd. Na koniec, mając odpowiednie wielośrednie, znajdujemy charakterystyki mierzące odchylenie zmiennej X od punktów koncentracji prawdopodobieństwa: $V_1(X) = E(X - E(X))^2, V_2(X) = E((X - s_1)(X - s_2))^2$ itd. Następnie wyznaczamy pierwiastki odpowiedniego stopnia każdej z nich (będą to analogony odchylenia standardowego). Przypuśćmy, że najmniejszy pierwiastek jest na *k*-tym miejscu. Oznacza to, że w naszym przypadku mamy *k* punktów koncentracji w sensie średniokwadratowym. Wniosek ten pozwala na bardziej precyzyjną analizę danych liczbowych.

Przykład 2. Dwumodalny rozkład Webera

Propozycją rozkładu dwumodalnego, o ciągłej funkcji gęstości, określonego na prostej, niebędącego mieszanką, jest rozkład Webera [Antoniewicz, Wilkowski 2004; Wilkowski 2009]. Zmienna losowa X o tym rozkładzie ($X \sim W(\alpha, \beta, \gamma)$) ma gęstość postaci

$$g_{\alpha,\beta,\gamma}(x) = \frac{1}{z(\alpha,\beta)} e^{\alpha(x-\gamma)^2 - \beta(x-\gamma)^4}; \quad x, \gamma \in R;\ \alpha, \beta > 0\ .$$

Komentarza wymaga stała normalizacyjna $z(\alpha,\beta)$. Okazuje się, że powyższą całkę można wyrazić przez funkcje specjalne Webera [Bateman, Erdelyi 1953], które są stablicowane. Mianowicie:

$$z(\alpha,\beta) = \int_{\Re} e^{\alpha x^2 - \beta x^4} dx = \exp\left(\frac{\alpha^2}{8\beta}\right) \frac{\sqrt{\Pi}}{\sqrt[4]{2\beta}} D_{-\frac{1}{2}}\left(-\frac{\alpha}{\sqrt{2\beta}}\right),$$

gdzie D jest funkcją Webera.

Z definicji widać, że funkcja g ma dwie mody (maksima) w punktach

$$Mo_1 = -\sqrt{\frac{\alpha}{2\beta}} + \gamma,\ \ Mo_2 = \sqrt{\frac{\alpha}{2\beta}} + \gamma,$$

a wartość oczekiwana zmiennej losowej X wynosi

$$E(X) = \gamma.$$

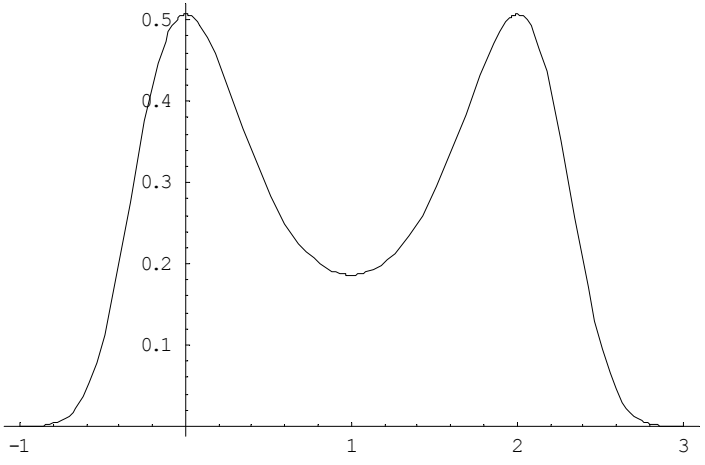
Momenty wyższych rzędów wyrażone są przez konfluentne funkcje hipergeometryczne [Wilkowski 2009]. Przypuśćmy, że $\boldsymbol{E(X) = 0}$, wtedy

$$M(X^{2n-1}) = 0,$$

$$M(X^{2n}) = \frac{1}{2z(\alpha,\beta)} \beta^{\frac{1}{4}(3-2n)} \left(\sqrt{\beta}\Gamma\left(\frac{1}{4}+\frac{n}{2}\right) H\left(\frac{1}{4}+\frac{n}{2},\frac{1}{2},\frac{\alpha^2}{4\beta}\right) + \right.$$
$$\left. +\alpha\Gamma\left(\frac{3}{4}+\frac{n}{2}\right) H\left(\frac{3}{4}+\frac{n}{2},\frac{3}{2},\frac{\alpha^2}{4\beta}\right)\right),$$

gdzie $n = 1, 2, ...,$ a H oznacza konfluentną funkcję hipergeometryczną [Bateman, Erdelyi 1953]. Możemy ją traktować jako uogólnienie innych funkcji specjalnych: funkcji Webera, funkcji Bessela, wielomianów Laguerre'a i Hermite'a itd. Wartości jej są stablicowane.

Na rysunku 2 przedstawiamy funkcję gęstości rozkładu $W(2,1,1)$:



Rys. 2. Gęstość rozkładu $W(2,1,1)$

Źródło: opracowanie własne.

Wartości modalne są w punktach: $Mo_1 = 0, Mo_2 = 2,$ a wartość oczekiwana $E(X) = 1$ staje się w tym przypadku wartością „nieoczekiwaną", gdyż masa prawdopodobieństwa skupiona jest wokół modalnych. Wielomian p_2 ze wzoru (9) ma w tym wypadku postać:

$$p_2(x) = x^2 + ax + b = (x - s_1)(x - s_2), \text{ gdzie}$$

$$a = -\frac{E(X^3) - E(X^2)E(X)}{V_1(X)},\ b = -\frac{E^2(X^2) - E(X^3)E(X)}{V_1(X)}, \tag{11}$$

$$s_1 = -0{,}08746,\ s_2 = 1{,}91254.$$

Dwuśrednia zmiennej losowej X o rozkładzie Webera W(2,1,1) jest zatem parą uporządkowaną $(s_1, s_2) = (-0{,}08746 ; 1{,}91253)$. Biśrednia lepiej aproksymuje modalne tego rozkładu niż wartość oczekiwana $E(X) = 1$. Wariancja tej zmiennej oraz jej analogon dla dwuśredniej wynoszą:

$$V_1(X) = E(X - E(X))^2 = 0{,}83274,,$$

$$V_2(X) = E((X - s_1)(X - s_2))^2 = 0{,}38928.$$

Okazuje się zatem, że odchylenie mierzone jednocześnie od obu miejsc koncentracji prawdopodobieństwa jest mniejsze niż rozrzut zmiennej losowej wokół średniej, czego także należało się spodziewać. Wyznaczanie kolejnych n-średnich nie polepsza jakości aproksymacji, w sensie średniokwadratowym, zmiennej losowej X o rozkładzie W(2,1,1). Mamy np.:

$$(s_1, s_2, s_3) = (-0{,}76371 ; 0{,}82532 ; 2{,}69894)$$

oraz

$$V_3(X) = E((X - s_1)(X - s_2)(X - s_3))^2 = 6{,}66794.$$

Przykład 3. PKB *per capita* to jeden z najczęściej stosowanych na świecie mierników zamożności państwa przez dochody jego obywateli. Jest on niewątpliwie związany z jakością życia w danym kraju. Relacja ta jest jednak trudna do zdefiniowania, bo jak wiadomo, większość ludzi woli zarabiać 50 000 dolarów rocznie, pod warunkiem, że ich sąsiedzi będą mieli 40 000, niż by sami mieli 100 000 rocznie, ale ich sąsiedzi 120 000. Poniżej przedstawiamy wysokość PKB w dolarach amerykańskich w 30 krajach w roku 2006 (dane pochodzą z Wikipedii):

1. Australia	30 897.
2. Austria	33 432.
3. Belgia	31 244.
4. Brazylia	8 561.
5. Bułgaria	9 223.
6. Dania	34 740.
7. Finlandia	31 208.
8. Grenada	8 198.
9. Holandia	30 862.
10. Hongkong	33 479.
11. Iran	7 980.
12. Irlandia	40 610.
13. Islandia	35 115.

14.Kanada	34 273.
15. Katar	31 397.
16. Kazachstan	8 318.
17. Kostaryka	10 434.
18. Luksemburg	69 800.
19. Malezja	11 201.
20. Meksyk	10 186.
21. Norwegia	42 364.
22. Rosja	11 041.
23. Rumunia	8785.
24. Stany Zjednoczone	41 399.
25. Szwajcaria	32 571.
26. Tajlandia	8 368.
27. Tunezja	8 255.
28. Turcja	7 950.
29. Turkmenistan	8 098.
30. Urugwaj	10 720

Na podstawie powyższych danych wyznaczamy wartość oczekiwaną, wariancję i odchylenie standardowe. Wynoszą one odpowiednio:

$$\boldsymbol{E(X) = 23024,}$$

$$\boldsymbol{V_1(X) = 23870350},$$

$$\boldsymbol{\sqrt{V_1(X)} = 15450}.$$

Zauważmy, że przeciętna staje się tutaj raczej „wartością nieoczekiwaną" (w żadnym państwie PKB nawet nie zbliża się do niej), a stosunkowo duże odchylenie standardowe sugeruje tylko znaczny rozrzut danych.

Obliczmy teraz dwuśrednią $\boldsymbol{(s_1, s_2)}$. Korzystając ze wzorów (11), dostajemy:

$$\boldsymbol{a = -58753,8,}$$

$$\boldsymbol{b = 5,83939 \times 10^8}.$$

Ponieważ dwuśrednia to pierwiastki wielomianu

$$\boldsymbol{x^2 + ax + b},$$

ostatecznie otrzymujemy

$$\boldsymbol{(s_1, s_2) = (12671,7\ ;\ 46082,1)}.$$

Analogonem wariancji w tej sytuacji będzie wielkość

$$\boldsymbol{V_2(X) = E((X - s_1)(X - s_2))^2 = 9,86242 \times 10^{15}}\ ,$$

a odchylenia standardowego – jej czwarty pierwiastek

$$\sqrt[4]{V_2(X)} = 9965,43 .$$

W tym przypadku dwuśrednia bardziej odpowiada rzeczywistości niż zwykła wartość średnia. PKB jednych państw są w pobliżu pierwszej współrzędnej, a drugiej grupy państw znajdują się blisko drugiego składnika dwuśredniej. Również analogon odchylenia standardowego dla dwuśredniej jest mniejszy niż zwykłe odchylenie standardowe. W tej sytuacji nie ma potrzeby wyznaczania kolejnych wielośrednich. Okazało się, że dzięki wielośredniej możemy analizować próbki niejednorodne i wychwycić te niejednorodności, a także dzięki pomiarom analogonów odchyleń standardowych weryfikować nasze przypuszczenia co do natury tych niejednorodności. Jest to *novum* w teorii i praktyce statystyki i probabilistyki.

Literatura

Antoniewicz R. (1988), *Metoda najmniejszych kwadratów dla zależności niejawnych i jej zastosowania w ekonomii,* Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 445, AE, Wrocław.

Antoniewicz R. (2005), *O średnich i przeciętnych,* AE, Wrocław.

Antoniewicz R., Wilkowski A. (2004), *O pewnym rozkładzie dwumodalnym*, „Przegląd Statystyczny", R. LI-Zeszyt 1.

Bateman H., Erdelyi A. (1953), *Higher Transcendental Functions*, Mc Graw-Hill. Book Company, New York.

Cramer H. (1958), *Metody matematyczne w statystyce,* PWN, Warszawa.

Laurent P.-J. (1975), *Aproksymacja i optymalizacja* (tłumaczenie z francuskiego), Wydawnictwo „Mir", Moskwa.

Serfling R.J. (1999), *Twierdzenia graniczne statystyki matematycznej,* PWN, Warszawa.

Szego G. (1975), *Orthogonal Polynomials,* Coll. Publ., XXIII, Amer. Math. Soc., Providence.

Wilkowski A. (1993), *O współczynniku zależności parabolicznej,* „Badania Operacyjne i Decyzje", nr 2.

Wilkowski A. (1994), *Współczynnik zależności prostoliniowej a współczynnik korelacji,* Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 667, AE, Wrocław.

Wilkowski A. (2008), *Uwagi o współczynniku korelacji,* Ekonometria 27, red. J. Dziechciarz, Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu nr 84, Wrocław.

Wilkowski A. (2009), *Uwagi o wielośredniej,* [w:] *Zastosowania ekonometrii*, red. A. Barczak, Prace Naukowe Akademii Ekonomicznej w Katowicach, Katowice.

LINE DEPENDENT COEFFICIENT AND MULTIAVERAGE IN DATA ANALYSIS

Summary: In this paper we talk about new statistic tools which enable more precise economic data analysis. Firstly, we define line dependent coefficient as a cosine of angle made of the cross of regression lines. It is the base, thanks to which we can define other nonlinear relation coefficients. Just like the classic correlation coefficient, line dependent coefficient is also asymptotically normal. Natural expansion of the line function class are conics. As in

case of regression lines, we can define regression conics (their equations vary in lineal parts). Conic dependent coefficient is a cosine of cross angle of regression conics (we choose cross point, which is the nearest from the set of points barycentre). It is the example of non-linear dependent coefficient, which can be defined on the basis of line dependent coefficient. The second part of this article is about multiaverage, generalization of the classic expected value of the random variable idea. The average may be considered as root-mean-square average approximation of the random variable with one point. Multiaverage is approximation of the variable with more than just one point at the same time. While defining multiaverage, we use standard moments method and some facts from the orthogonal polynomial theory.

Key words: line dependent coefficient, conic dependent coefficient, multiaverage.