PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 84

Ekonometria 27 2009

Andrzej Wilkowski
Uniwersytet Ekonomiczny we Wrocławiu

UWAGI O WSPÓŁCZYNNIKU KORELACJI

Streszczenie: W pracy omówiono wybrane własności klasycznego współczynnika korelacji i maksymalnego współczynnika korelacji. Wprowadzono także współczynnik zależności prostoliniowej oraz podano dowód asymptotycznej normalności jego próbkowego odpowiednika.

Słowa kluczowe: współczynnik korelacji, maksymalny współczynnik korelacji, współczynnik zależności prostoliniowej, asymptotyczna normalność.

Jednym z celów niniejszej pracy jest omówienie wybranych własności współczynnika korelacji (i jego estymatora), a także maksymalnego współczynnika korelacji (niektóre z nich zostały udowodnione w roku 2008, są zatem dosyć aktualne). W dalszej części artykułu podano współczynnik zależności prostoliniowej i jego związki ze współczynnikiem korelacji (są to wyniki własne, m.in. autora). Współczynnik ten jest teoretyczną podstawą konstrukcji innych miar zależności nieliniowej między zmiennymi losowymi.

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

$$rX,Y = CovX,\ YVarX\ VarY\ . \tag{1}$$

Oczywiście $-1 \le r \le 1$, $rX,Y = rY,X$, $rX,Y = rmX + n,Y$, o ile $m \neq 0$.

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

$$y - E(Y)VarY = rx - E(X)VarX, \tag{2}$$

$$y - E(Y)VarY = 1\ r\ x - E(\mathrm{X})VarX, \tag{3}$$

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 następujące wyrażenia:

$$\mathrm{mina},\ b \in RE(Y - b - aX)^2 = VarY_1 - r^2\ , \tag{4}$$

$$\mathrm{mina},\ b \in RE(X - b - aY)^2 = VarX_1 - r^2\ . \tag{5}$$

Widać z tego, że każda zmienna ma wariancję zmniejszoną w stosunku $(1 - r^2)$: 1, na skutek 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.

Maksymalny współczynnik korelacji *R(X, Y)* między zmiennymi losowymi X oraz Y został wprowadzony przez Gebeleina [Gebelein 1941]. Definiuje go wyrażenie:

$$RX,Y = supf,grfX,gY \tag{6}$$

gdzie supremum dotyczy wszystkich funkcji f, g, takich że $0 < Var(fX,Var(gY)) < \infty$.

Wymieńmy kilka własności współczynnika R:

- jeśli wektor (X, Y) ma rozkład normalny, to $R(X, Y) = r(X,Y)$ [Lancaster 1957],
- jeśli niezdegenerowane zmienne losowe $X_1, \ldots, X_n$ są niezależne, o jednakowym rozkładzie, to

$$RSm,Sn = mn, \tag{7}$$

gdzie $m \leq n$ są naturalne oraz $Sk = j = 1kXj$, $j = 1, \ldots, n$ [Dembo, Kagan, Sheep 2001],

- jeśli niezdegenerowane zmienne losowe $X_1, \ldots, X_n$ są niezależne, o jednakowym rozkładzie, to

$$R(j = 1mXj,\ j = 1 + 1nXj) = m - lmn - 1, \tag{8}$$

gdzie liczby naturalne l, m, n, spełniają warunek: $1 \leq 1 + 1 \leq m \leq n$ [Yaming 2008].

Warto wspomnieć o korelacyjnej nierówności Gaussa. Załóżmy, że A, B są wypukłymi, symetrycznymi podzbiorami przestrzeni Rn. Niech υ będzie gaussowską, centralną miarą na Rn. Wówczas **hipoteza korelacyjna Gaussa** stanowi, że:

$$\upsilon A \cap B \geq \upsilon A \upsilon B. \tag{9}$$

Dowód tego znajduje się w pracy [He-Jing, Ze-Chun 2008].

Zauważmy, że jeśli mamy proste regresji zmiennych losowych X oraz Y:

$$y = a_1 x + b_1, \tag{10}$$

$$x = a_2 y + b_2, \tag{11}$$

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

$$r2X,Y = a1a2. \tag{12}$$

Zdefiniujemy **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:

$$kX,Y = \cos\alpha = a1 + a2a12 + 1\ a22 + 1, \tag{13}$$

gdzie α jest kątem przecięcia prostych regresji.

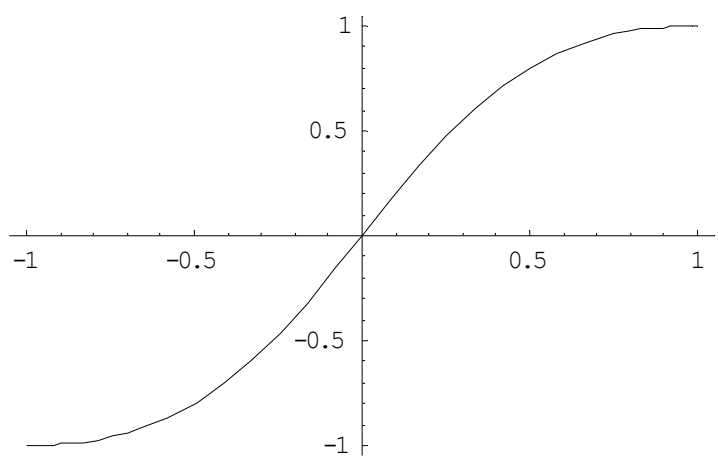
Możemy także napisać:

$$kVarX,VarY,r = VarX + VarYrVarX + r2VarY\ VarY + r2VarX. \tag{14}$$

Z tego wynika, 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 (tzn. wariancja równa jeden, wartość oczekiwana równa zero). Wtedy wzór (14) przyjmie postać:

$$kr = 2r1 + r2\ ,\ r \in -1, 1. \tag{15}$$

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:

$$maxVarX,\ VarY > 0,\ r \in [-1,1]\ kVarX,\ VarY,\ r - r = 105 - 22\ 2. \tag{16}$$

Maksimum jest osiągane dla $r = \pm 5 - 2$ oraz $Var(X) = Var(Y)$. Dowód znajduje się w pracy [Wilkowski 1994].

Zwróćmy jeszcze uwagę na fakt, że współczynnik korelacji liniowej r jest również kosinusem kąta, ale między innymi wektorami.

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 $sn2$**, jeżeli $sn2 > 0$ dla dostatecznie dużych n oraz

$$Xn - mn\ sn\ d\ N0,1. \tag{17}$$

Zapisujemy to jako: **X_n jest $AN(mn, sn2)$.** 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 $AN(mn,sn2)$, to nie wynika z tego, że ciąg (X_n) w ogóle zbiega według rozkładu. Mamy jednak zawsze

$$suptPXn \le t - P(N(mn,sn2) \le t)0,\ n\infty\ . \tag{18}$$

Chcąc zatem obliczać prawdopodobieństwa, można traktować X_n jako zmienną losową $N(mn,sn2)$ [Serfling 1999].

Niech $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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

$$rX,Y = CovX,\ YVarX\ VarY\ . \tag{19}$$

Jego próbkowy odpowiednik ma postać:

$$rn = i = 1nXi - X(Yi-Y)\ i = 1n(Xi - X\ )2\ i = 1n(Yi - Y\ )2, \tag{20}$$

gdzie $X = 1ni = 1nXi$, $Y = 1ni = 1nYi$.

Okazuje się, że ***rn*** jest $AN(r, n^{-1}\mathbf{dSd}^{\mathrm{T}})$ [Serfling 1999], gdzie **S** jest macierzą kowariancji wektora (X, Y, X^2, Y^2, XY), a wektor $\mathbf{d} = (rE(X)Var(X) - E(Y)VarX\ VarY,\ rE(Y)Var(Y) - E(X)VarX\ VarY, -r2Var(X), -r2Var(Y), 1VarX\ VarY)$.

W rozważaniach dotyczących asymptotycznej normalności korzysta się z następującego twierdzenia.

Twierdzenie A. Przypuśćmy, że $Xn = Xn1, \ldots, Xnk$ jest $AN\mu, bn2\Sigma$, gdzie Σ jest pewną macierzą kowariancji oraz $bn \to 0$. Niech $g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_m(\mathbf{x}))$, $\mathbf{x} = (x_1, \ldots, x_k)$, będzie taką funkcją wektorową, że każda jej składowa g_i jest funkcją rzeczywistą, a różniczka w punkcie $\mathbf{x} = \mu$ jest niezerowa (wystarczy, aby co najmniej jedna pochodna cząstkowa była ciągła i różna od zera w tym punkcie). Niech

$$\mathrm{D} = \partial gi\partial xj\ x = \mu, \tag{21}$$

oczywiście **D** jest macierzą wymiaru $m \times k$.

Wówczas

$$g(\mathbf{X}_n) \text{ jest } AN(g(\mu), bn2D\Sigma DT). \tag{22}$$

Dowód tego twierdzenia znajduje się w książce [Serfling 1999].

Poprzednio 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 $(X_1, Y_1), \ldots, (X_n, Y_n)$ 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 (14) i (20) wnioskujemy, że próbkowy odpowiednik współczynnika k ma postać:

$$kn = (i = 1n(Xi - X)2 + i = 1n(Yi - Y)2)\, rn\, i = 1n(Xi - X)2 + rn2i =$$

$$= 1n(Yi - Y)2\, rn2i = 1n(Xi - X)2 + i = 1n(Yi - Y)2\,. \tag{23}$$

Twierdzenie B. Niech wektor **V = X, Y, 1ni = 1nXi2, 1ni = 1nYi2, 1ni = = 1nXiYi**, funkcja g: $R5 \to R$ będzie określona wzorem

$$gz1, z2, z3, z4, z5 = z5 - z1z2(z3 - z12\, z4 - z22 + z4 - z22z3 - z12)z3 -$$

$$z12 + z5 - z1z22z3 - z12\, z4 - z22 + z5 - z1z22z4 - z22. \tag{24}$$

Wówczas kn jest $AN(k, n^{-1\delta S\delta T})$, gdzie **S** jest macierzą kowariancji wektora (X, Y, X^2, Y^2, XY), a wektor

$$\delta = \partial g \partial z1z = EV, \ldots, \partial g \partial z5z = EV\,. \tag{25}$$

Dowód. Zauważmy, że wektor **V** jest $AN(E(\mathbf{V}), n^{-1}\mathbf{S})$ [Serfling 1999]. Widać, że $kn = g(V)$, funkcja g spełnia założenia twierdzenia A (w naszym przypadku $b_n = n^{-1/2}$). Mamy zatem tezę twierdzenia B.

Na zakończenie zauważmy, że idee związane ze współczynnikiem zależności prostoliniowej mogą być wykorzystane przy konstrukcji innych miar zależności. Analogicznie jak w wypadku prostych regresji można definiować krzywe regresji (wprowadzając np. zaburzenie w części liniowej równania krzywej, raz przy jednej zmiennej, a raz przy drugiej). Następnie trzeba wyznaczyć punkt przecięcia krzywych regresji, który leży bliżej środka ciężkości zbioru danych. Kosinus kąta, pod jakim przecinają się krzywe w tym punkcie, będzie wtedy współczynnikiem zależności nieliniowej.

Literatura

Antoniewicz R., *Metoda najmniejszych kwadratów dla zależności niejawnych i jej zastosowania w ekonomii,* Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 445, AE, Wrocław1988.

Brandt S., *Data Analysis. Statistical and Computational Methods for Statistics and Engineers,* ed. 3, Springer Verlag, New York 1999.

Cieciura M., Zacharski J., *Metody probabilistyczne w ujęciu praktycznym*, Wydawnictwo Naukowe PWN, Warszawa 2007.
Cramer H., *Metody matematyczne w statystyce*, PWN, Warszawa 1958.
Dembo A., Kagan A., Sheep L.A., *Remarks on the maximum correlation coefficient*, „Bernoulli" 2001, 7.
Gebelein H., *Das statistische Problem der Korrelation als Variations und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung*, Z. Angew. Math. Mech. 1941, 21.
He-Jing H., Ze-Chun H., *Gaussian Correlation Conjecture for Symmetric Convex Sets*, arXiv:0811.0488v1 [math.PR], 4 November 2008.
Lancaster H.O., *Some properties of the bivariate normal distribution considered in the form of a contingency table*, „Biometrica" 44, s. 289-292.
Renyi A., *On measures of dependence,* Acta Math. Hungar. 1959, 10.
Serfling R.J., *Twierdzenia graniczne statystyki matematycznej*, PWN, Warszawa 1999.
Yaming Y., *On the maximal correlation coefficient,* „Statistics and Probability Letters" 2008, 78.
Wilkowski A., *Współczynnik zależności prostoliniowej a współczynnik korelacji*, Prace Naukowe Akademii Ekonomicznej we Wrocławiu nr 667, AE, Wrocław 1994.

NOTE ON CORRELATION COEFFICIENT

Summary: The paper discusses correlation coefficient and the maximal coefficient of correlation. It also introduces the line dependent coefficient and gives the proof of the normal asymptotic distribution of its sample equivalent.