EKONOMETRIA

26

Zastosowanie matematyki w ekonomii

Redaktor naukowy **Janusz Łyko**

Wydawnictwo Uniwersytetu Ekonomicznego we Wrocławiu
Wrocław 2009

Spis treści

Summaries

PRACE NAUKOWE UNIWERSYTETU EKONOMICZNEGO WE WROCŁAWIU nr 76
Ekonometria 26 2009

Andrzej Wilkowski
Uniwersytet Ekonomiczny we Wrocławiu

O WSPÓŁCZYNNIKU KORELACJI

Streszczenie: W pracy omówione będą wybrane własności klasycznego współczynnika korelacji oraz jego próbkowego odpowiednika. Podany zostanie także współczynnik zależności prostoliniowej i jego związek ze współczynnikiem korelacji. Przedstawiona zostanie również asymptotyczna normalność próbkowego odpowiednika tego współczynnika.

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

1. Wstęp

Praca poświęcona jest miarom zależności liniowej. Omówiono klasyczny współczynnik korelacji, podano także nowe fakty dotyczące maksymalnego współczynnika korelacji, hipotezy korelacyjnej Gaussa oraz lematu Hoeffdinga. Następnie zdefiniowano współczynnik zależności prostoliniowej (jest to punkt wyjścia do konstrukcji innych miar zależności), przedstawiono jego związek ze współczynnikiem Pearsona. Ostatnia część niniejszego artykułu poświęcona jest asymptotycznej normalności próbkowych odpowiedników obu tych współczynników.

2. Współczynnik korelacji liniowej i maksymalny współczynnik korelacji

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” znaczy zależne liniowo, tzn. 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

$$-1 \le r \le 1,\ r(X,Y)=r(Y,X),\ r(X,Y)=r(mX+n,Y) \text{ o ile } m \ne 0.$$

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:

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

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

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.

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

$$R(X,Y)=\sup_{f,g} r\left(f(X),g(Y)\right),$$

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

Wymieńmy kilka własności współczynnika R:

- jeśli wektor (X, Y) ma rozkład normalny, to $R(X, Y) = |r(X, Y)|$ [Gebelein 1941],
- jeśli niezdegenerowane zmienne losowe $X_1, \ldots, X_n$ są niezależne oraz jednakowo rozłożone, to

$$R(S_m, S_n) = \sqrt{\frac{m}{n}},$$

gdzie $m \leq n$ są naturalne oraz $S_k = \sum_{j=1}^{k} X_j$, $k = 1, \ldots, n$ [Dembo, Kagan, Sheep 2001],

- jeśli niezdegenerowane zmienne losowe $X_1, \ldots, X_n$ są niezależne oraz jednakowo rozłożone, to

$$R\left(\sum_{j=1}^{m} X_j,\ \sum_{j=l+1}^{n} X_j\right) = \frac{m-l}{\sqrt{m(n-l)}},$$

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

Warto wspomnieć również o korelacyjnej nierówności Gaussa. Załóżmy, że A, B są wypukłymi, symetrycznymi podzbiorami przestrzeni $\mathbb{R}^n$. Niech υ będzie gaussowską, centralną miarą na $\mathbb{R}^n$. Wówczas **hipoteza korelacyjna Gaussa** stanowi, że

$$\upsilon(A \cap B) \geq \upsilon(A)\upsilon(B).$$

Dowód tego faktu znajdziemy w pracy [He-Jing, Ze-Chun 2008].

Na zakończenie tego punktu omówione zostanie uogólnienie dobrze znanego lematu Hoeffdinga. Niech $\mathbf{Z}$ będzie losowym wektorem o wartościach w $u + v$-wymiarowym prostokącie $P = [a, b] \subset \mathbb{R}^{u+v}$, $u, v \in \mathbb{N}$. Przypuśćmy, że zbiór P został tak wybrany, aby każda współrzędna Z_i była równa a_i z prawdopodobieństwem zero. Dla niepustego zbioru $K \subset \{1, \ldots, u+v\}$, niech $P_K = \Pi_{k \in K}[a_k, b_k]$ oraz F_K będzie łączną dystrybuantą zmiennych losowych Z_k, dla $k \in K$, uporządkowanych według rosnących indeksów (przyjmujemy, że $F_\varnothing = 1$). Niech $\mathbf{X} = (Z_1, \ldots, Z_u)$, $\mathbf{Y} = (Z_{u+1}, \ldots, Z_{u+v})$. Załóżmy, że f i g są funkcjami o wartościach rzeczywistych określonymi na $P_{\{1, \ldots, u\}}$, $P_{\{u+1, \ldots, u+v\}}$, odpowiednio. **Uogólniony lemat Hoeffdinga** mówi, że jeśli funkcje f i g są lewostronnie ciągłe oraz mają ograniczoną wariację Hardy'ego-Krausa, to

$$Cov(\mathrm{f}(\mathbf{X}), \mathrm{g}(\mathbf{Y})) = \sum_{\varnothing \neq \mathrm{I} \subset \{1,\ldots,\mathrm{u}\}} \ \sum_{\varnothing \neq \mathrm{J} \subset \{\mathrm{u}+1,\ldots,\mathrm{u}+\mathrm{v}\}} (-1)^{|\mathrm{I}|+|\mathrm{J}|} \int_{\mathrm{P_I}} \int_{\mathrm{P_J}} \left(\mathrm{F_{IUJ}} - \mathrm{F_I F_J}\right) \mathrm{df_I dg_J}.$$

Dowód tego faktu znajdziemy w pracy [Beare 2009].

Zdefiniowania wymaga **wariacja Hardy'ego-Krausa.** Niech h oznacza funkcję o wartościach rzeczywistych określoną na n-wymiarowym prostokącie $[a, b] \subset \mathbb{R}^n$.

Wtedy

$$\Delta_R h = \sum_{I \subset \{1,\dots,n\}} (-1)^{|I|} h(x_I),$$

gdzie $R = [c, d] \subset [a, b]$, x_I jest wektorem przestrzeni $\mathbb{R}^n$, w którym i-ta współrzędna jest równa c_i, jeśli $i \in I$, lub d_i, gdy $i \notin I$. Można teraz określić wariację Vitaliego, mianowicie

$$\|h\|_V = \sup \sum_{R \in \wp} |\Delta_R h| ,$$

gdzie supremum dotyczy wszystkich skończonych rodzin n-wymiarowych prostokątów $\wp = \{R_i : 1 \le i \le m\}$ takich, że $\bigcup_{i=1}^m R_i = [a, b]$, oraz wnętrza dowolnej pary tych prostokątów z rodziny $\wp$ są rozłączne. Dla niepustego zbioru $I \subset \{1, \dots, n\}$ niech h_I oznacza funkcję rzeczywistą określoną na $\prod_{i \in I} [a_i, b_i]$ powstałą przez zastąpienie i-tego argumentu funkcji h przez b_i, gdy $i \notin I$ (przyjmujemy, że $h_\varnothing = h(b)$). **Wariacja Hardy'ego-Krausa** dana jest wzorem

$$\|h\|_{HK} = \sum_{\varnothing \ne I \subset \{1,\dots,n\}} \|h_I\|_V .$$

3. Współczynnik zależności prostoliniowej

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.

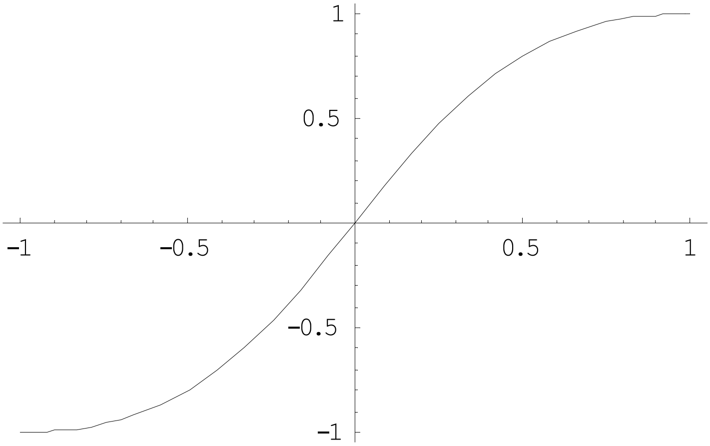
Możemy także napisać:

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

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 (tzn. wariancja równa jeden, wartość oczekiwana zero). Wtedy wzór (1) przyjmie postać:

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

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:

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

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

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

4. 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

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

Chcąc zatem obliczać prawdopodobieństwa, można traktować X_n jako zmienną losową $N(m_{n,}\ s_n^2)$ [Serfling 1991].

Niech $(X_1, Y_1), \ldots, (X_n, Y_n)$ będą niezależnymi obserwacjami, o jednakowym rozkładzie, z pewnego rozkładu dwuwymiarowego (wektor $(\mathbf{X}_1, \mathbf{Y}_1)$ ma taki sam rozkład jak wektor losowy $(\mathbf{X}, \mathbf{Y})$). 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

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

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

Wymieńmy kilka własności próbkowego współczynnika korelacji:

- $\hat{r}_n$ jest $AN(r, n^{-1}\mathbf{dSd}^T)$ [Serfling 1991] , gdzie **S** jest macierzą
- kowariancji wektora (**X**, **Y**, **X**^{2}, **Y**^{2}, **XY**), a wektor

$$\mathbf{d} = \left(\frac{rE(X)}{Var(X)} - \frac{E(Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}, \frac{rE(Y)}{Var(Y)} - \frac{E(X)}{\sqrt{Var(X)}\sqrt{Var(Y)}}, -\frac{r}{2Var(X)}, \right.$$

$$\left. -\frac{r}{2Var(Y)}, \frac{1}{\sqrt{Var(X)}\sqrt{Var(Y)}} \right),$$

- jeżeli zmienna losowa dwuwymiarowa (X, Y) ma rozkład normalny, wówczas $\hat{r}_n$ jest $AN(r, \frac{1-r^2}{n})$, zgodność jest dobra dla $n \geq 500$,

  $\frac{1}{2} ln \frac{1+\hat{r}_n}{1-\hat{r}_n}$ jest $AN\left(\frac{1}{2} ln \frac{1+r}{1-r}, \frac{1}{n-3} \right)$, wystarczającą zgodność mamy dla $n \geq 20$,

  gdy ponadto $r = 0$ (cechy X, Y są wtedy niezależne), to statystyka $\frac{\hat{r}_n}{\sqrt{1-\hat{r}_n^{\,2}}}\sqrt{n-2}$

ma rozkład Studenta z $n-2$ stopniami swobody [Cieciura, Zacharski 2007].

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 ($\mathbf{X}_1$, $\mathbf{Y}_1$) ma taki sam rozkład jak wektor losowy (**X**, **Y**)). Na podstawie wzorów (1) i (2) wnioskujemy, że próbkowy odpowiednik współczynnika k jest postaci

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

Twierdzenie. Niech wektor

$$\mathbf{V} = (\overline{X}, \overline{Y}, \frac{1}{n}\sum_{i=1}^n X_i^2, \frac{1}{n}\sum_{i=1}^n Y_i^2, \frac{1}{n}\sum_{i=1}^n X_i Y_i),$$

funkcja g: $\mathbb{R}^5 \to \mathbb{R}$ 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 $\hat{k}_n$ jest AN(k, n^{-1} $\delta S \delta^T$),
gdzie **S** jest macierzą kowariancji wektora (**X**, **Y**, **X**^{2}, **Y**^{2}, **XY**), a wektor

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

Dowód tego faktu znajduje się w pracy [Wilkowski 2009].

Na zakończenie niniejszego opracowania 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.

Literatura

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

Beare B., *A Generalization of Hoeffding`s Lemma, and a New Class of Covariance Inequalities*, Statistics and Probability Letters 79, Elsevier 2009.

Cieciura M., Zacharski J., *Metody probabilistyczne w ujęciu praktycznym*, 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 no 7.

Gebelein H., *Das Statistische Problem der Korrelation als Variations und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung*, „Z. Angew. Math. Mech." 1941 no 21.

He-Jing H., Ze-Chun H., *Gaussian Correlation Conjecture for Symmetric Convex Sets,* arXiv:0811.0488v1 [math.PR] 4 Nov 2008.

Serfling R.J., *Twierdzenia graniczne statystyki matematycznej*, PWN, Warszawa 1991.

Wilkowski A., *Uwagi o współczynniku korelacji,* Prace Naukowe Uniwersytetu Ekonomicznego we Wrocławiu (w druku), Ekonometria 27 (2009).

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

Yaming Y., *On the Maximal Correlation Coefficient*, „Statistics and Probability Letters" no 78, Elsevier 2008.

ABOUT THE COEFFICIENT OF CORRELATION

Summary: The work discusses some properties of the classical coefficient of correlation and its sample equivalent. Parallel dependence coefficient and its connection with the coefficient of correlation is also given. The author describes asymptotic normalcy of the sample equivalent of this coefficient too.