M A T H E M A T I C A L E C O N O M I C S

No. 7(14) 2011

NOTES ON LINE DEPENDENT COEFFICIENT AND MULTIAVERAGE

Andrzej Wilkowski

Abstract. In this paper we discuss new statistic tools which enable more precise economics data analysis. Firstly, we define line dependent coefficientas a cosine of the angle made of the cross of regression lines. This is the basis thanks to which we can define other nonlinear relation coefficients such as conic dependent coefficient. Just like the classic correlation coefficient, line dependent coefficient is also asymptotically normal. The second part of this article is about multiaverage, a generalization of the classic expected value of the random variable idea. The average may be considered as the root-mean-square average approximation of the random variable with one point. Multiaverage is an approximation of the random variable with more than just one point at the same time (which is important when we talk about random variables, whose distributions are mixtures, or about multimodal densities). While defining multiaverage, we use the standard moments method and some facts from the orthogonal polynomial theory. In this paper we give some numerical examples in which we use the aforementioned tools.

Keywords: correlation coefficient, line dependent coefficient, conic dependent coefficient, multiaverage.

JEL Classification: C19.

1. Introduction

A line dependent coefficient was defined by Antoniewicz (1988) as a cosine of an angle made of the cross of regression lines. On the basis of this concept, we can define other nonlinear relation coefficients (see (Antoniewicz, 2005; Wilkowski, 2009)). Just like the classic correlation coefficient, line dependent coefficient is also asymptotically normal (Wilkowski, 2009). Next we present multiaverage, a 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 value. Multiaverage is an approximation of the variable

Andrzej Wilkowski
Department of Mathematics, Wrocław University of Economics, Komandorska Street 118/120, 53-345 Wrocław, Poland.
E-mail: andrzej.wilkowski@ue.wroc.pl

with more than just one point at a time (which is important when we talk about random variables, which distributions are mixtures, or about multimodal densities) (McLachlan, Peel, 2004). While defining multiaverage, we use standard moments method (Cramer, 1958) and orthogonal polynomial theory (Brandt, 1999; Szego, 1975).

2. Line dependent coefficient

Let X, Y be random variables on the same probability space. Classic correlation coefficient is then given by the equation (1):

$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}}. \tag{1}$$

Obviously, $-1 \le r \le 1$, $r(X, Y) = r(Y, X)$, $r(X, Y) = r(mX + n, Y)$ if only $m \neq 0$. It is an important characteristic of random vector (X, Y).

For regression lines of random variables X and Y,

$$y = a_1 x + b_1,$$

$$x = a_2 y + b_2,$$

we can easily find correlation coefficient, namely:

$$r^2(X,Y) = |a_1 a_2|. \tag{2}$$

Line dependent coefficient k, is defined as a cosine of angle between crossing regression lines. It is not very hard to see that k is given by the formula:

$$k(X,Y) = \cos\alpha = \frac{a_1 + a_2}{\sqrt{a_1^2 + 1}\sqrt{a_2^2 + 1}}, \tag{3}$$

where α is an angle made of the cross of regression lines. We have also:

$$k\left(Var(X), Var(Y), r\right) = \frac{\left(Var(X) + Var(Y)\right) r}{\sqrt{Var(X) + r^2 Var(Y)}\sqrt{Var(Y) + r^2 Var(X)}}. \tag{4}$$

One of the most important kinds of convergence in distribution is convergence to normal distribution. Sequence of random variables (X_n) converges in distribution to $N(m, s^2)$, $s > 0$ if equivalently sequence $((X_n - m)/s)$ con-

verges in distribution to $N(0, 1)$. Most generally, we say that the sequence of random variables (X_n) is asymptotically normal if $s_n^2 > 0$ for n high enough and

$$\frac{X_n - m_n}{s_n} \xrightarrow{d} N(0,1), \quad n \to \infty. \tag{5}$$

We write it as X_n is $AN(m_n, s_n^2)$.

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent, identically distributed observations of two-dimensional random vector (X, Y). The sample correlation coefficient is then:

$$\hat{r}_n = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}, \tag{6}$$

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

According to (4) and (6), we can see that sample line dependent coefficient is given by:

$$\hat{k}_n = \frac{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2\right)\hat{r}_n}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \hat{r}_n^2\sum_{i=1}^{n}(Y_i - \bar{Y})^2}\sqrt{\hat{r}_n^2\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}. \tag{7}$$

Theorem. *Let us assume that vector*

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

and function $g: \mathbb{R}^5 \to \mathbb{R}$ *is given by the formula*:

$$g(z_1, z_2, z_3, z_4, z_5) = \frac{(z_5 - z_1 z_2)\left(\sqrt{\frac{z_3 - z_1^2}{z_4 - z_2^2}} + \sqrt{\frac{z_4 - z_2^2}{z_3 - z_1^2}}\right)}{\sqrt{z_3 - z_1^2 + \frac{(z_5 - z_1 z_2)^2}{z_3 - z_1^2}}\sqrt{z_4 - z_2^2 + \frac{(z_5 - z_1 z_2)^2}{z_4 - z_2^2}}}. \tag{8}$$

Then sample line coefficient $\hat{k}_n$ is $AN(k, n^{-1}\,\boldsymbol{\delta S \delta^T})$, where S is covariance matrix of random vector (X, Y, X^2, Y^2, XY), while vector δ

$$\boldsymbol{\delta} = \left(\frac{\partial g}{\partial z_1} \Big| \boldsymbol{z} = E(\boldsymbol{V}),\ \ldots, \frac{\partial g}{\partial z_5} \Big| \boldsymbol{z} = E(\boldsymbol{V}) \right). \tag{9}$$

One can find proof of this fact in Szego (1975).

Example. Conic dependent coefficient.

Let us assume, that random variables X and Y have positive, finite fourth moment ($0 < E(X^4), E(Y^4) < \infty$). Let real numbers a_1, b_1, c_1, d_1, f_1, a_2, b_2, c_2, e_2, f_2 meet equations given below:

$$\begin{aligned} &\min_{a,b,c,d,f \in R} E\left(aX^2 + bXY + cY^2 + dX - Y + f\right)^2 = \\ &= E\left(a_1X^2 + b_1XY + c_1Y^2 + d_1X - Y + f_1\right)^2, \end{aligned} \tag{10}$$

$$\begin{aligned} &\min_{a,b,c,e,f \in R} E\left(aX^2 + bXY + cY^2 - X + eY + f\right)^2 = \\ &= E(a_2X^2 + b_2XY + c_2Y^2 - X + e_2 + f_2)^2, \end{aligned} \tag{11}$$

Conic regression related with x is algebraic curve, given by:

$$a_1x^2 + b_1xy + c_1y^2 + d_1x - y + f_1 = 0. \tag{12}$$

Conic regression related with y is algebraic curve, given by:

$$a_2x^2 + b_2xy + c_2y^2 - x + e_2y + f_2 = 0. \tag{13}$$

The only difference between (12) and (13) is the linear part. It is analogous to regression lines.

Conic regression parameters meet systems of equations given below:

$$\begin{aligned} &E(X^4)a_1 + E(X^3Y)b_1 + E(X^2Y^2)c_1 + E(X^3)d_1 + E(X^2)f_1 = E(X^2Y) \\ &E(X^3Y)a_1 + E(X^2Y^2)b_1 + E(XY^3)c_1 + E(X^2Y)d_1 + E(XY)f_1 = E(XY^2) \\ &E(X^2Y^2)a_1 + E(XY^3)b_1 + E(Y^4)c_1 + E(XY^2)d_1 + E(Y^2)f_1 = E(Y^3), \\ &E(X^3)a_1 + E(X^2Y)b_1 + E(XY^2)c_1 + E(X^2)d_1 + E(X)f_1 = E(XY) \\ &E(X^2)a_1 + E(XY)b_1 + E(Y^2)c_1 + E(X)d_1 + f_1 = E(Y), \end{aligned} \tag{14}$$

$$\begin{aligned} &E(X^4)a_2 + E(X^3Y)b_2 + E(X^2Y^2)c_2 + E(X^2Y)e_2 + E(X^2)f_2 = E(X^3) \\ &E(X^3Y)a_2 + E(X^2Y^2)b_2 + E(XY^3)c_2 + E(XY^2)e_2 + E(XY)f_2 = E(X^2Y) \\ &E(X^2Y^2)a_2 + E(XY^3)b_2 + E(Y^4)c_2 + E(Y^3)e_2 + E(Y^2)f_2 = (XY^2), \\ &E(X^2Y)a_2 + E(XY^2)b_2 + E(Y^3)c_2 + E(Y^2)e_2 + E(Y)f_2 = E(XY) \\ &E(X^2)a_2 + E(XY)b_2 + E(Y^2)c_2 + E(Y)e_2 + f_2 = E(X). \end{aligned} \tag{15}$$

Conic dependent coefficient k_s of random variables X and Y is defined as:

$$k_s(X,Y) = \cos\alpha, \tag{16}$$

where α is an angle between regression lines at their intersection, nearest to point ($E(X)$, $E(Y)$).
Obviously, we have:

$$k_s = \frac{1 + m_1 m_2}{\sqrt{1 + m_1^2}\sqrt{1 + m_2^2}}, \tag{17}$$

where m_1, m_2 are slopes of straight lines which are tangents to a curves regression.

Example. Parabola dependent coefficient.

Cost function in econometric cost analysis has usually a form of parabola (when we discard random component). Let us assume that parabolas are regression conics, and their axes of symmetry are parallel to OY. Data for this example are given in the following table:

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

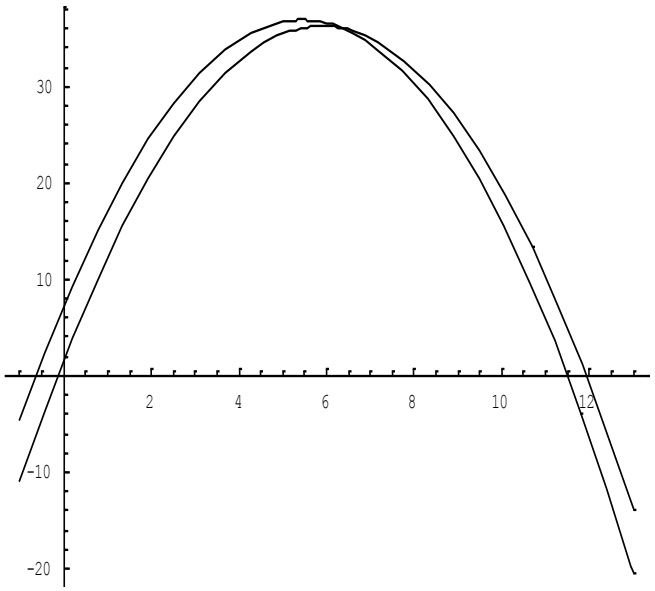

Fig. 1. Regression parabolas

Source: author's own study.

Then regression parabola, in regard to x and y, are described by the following equations:

$$y = -0.9958x^2 + 11.7424x + 1.6392,$$

$$y = -1.0033x^2 + 10.913x + 7.223.$$

Correlation coefficient and conic dependence coefficient are consecutively:

$$r = -0.0332,\ k_s = 0.9478.$$

The following table shows costs of building works in billions of Polish zlotys in the period 1960-1972 (Stanisz, 1993).

X	1	2	3	4	5	6	7	8	9	10	11	12	13
Y	24.34	47.73	79.72	109	144.5	178.9	215.6	279.2	350	414.2	463.8	540	705

Regression parabolas are of the form:

$$y = 3.5394x^2 + 2.7292x + 31.335,$$

$$y = -6.8972x^2 + 157.04x - 391.45.$$

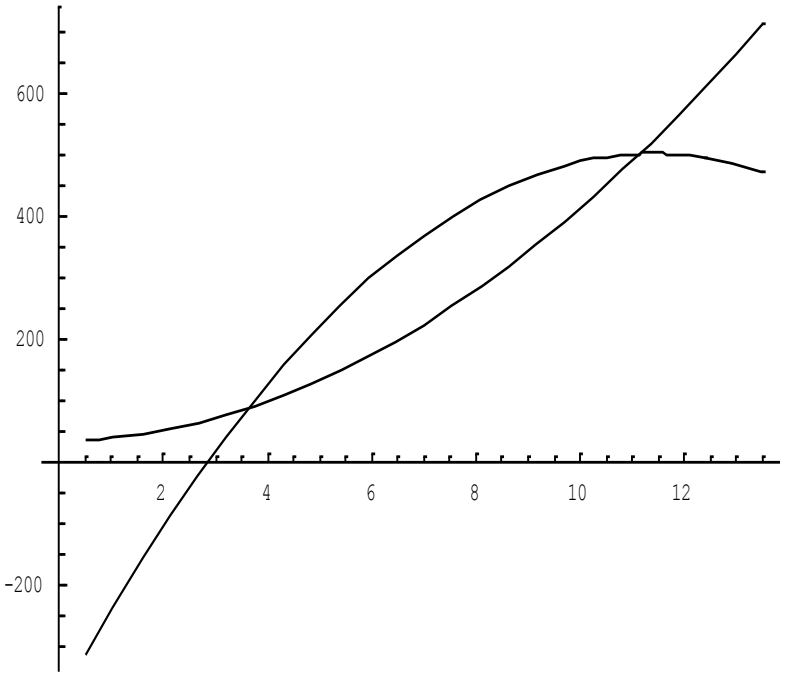

Fig. 2. Regression parabolas

Source: author's own study.

Correlation coefficient and conic dependence coefficient are consecutively:
$r = 0.9619$, $k_s = 0.8832$.

3. Multiaverage

We treat random variables as real measurable functions on probabilistic space (Ω, F, P).One can easily find their moments $m_1, m_2, \ldots$ from:

$$m_k = E(X^k) = \int_{\Omega} X^k(\omega) P(d\omega), \tag{18}$$

where $k = 1,2,\ldots$, while integrals are unconditionally convergent.

Combination of first and second moments, defined as (18), is equal to variance of a random variable:

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

and is called a variance of random variable. Polynomial $X - E(X)$ minimize root-mean-square norm, which is given below:

$$\min_{a \in R} E(X-a)^2 = E(X - E(X))^2 = V_1(X). \tag{19}$$

It can be said that average is the best one-point approximation of random variable.

Let f_X is density of random variable X.
Maximums of density function f_X are quite important. We call them modal values of random variable X. They mark concentration points of probability. In a unimodal density case, average $E(X)$ is a good modal value approximation.

Let random variable X have finite moments:

$$E(X^k) = m_k < \infty, \qquad k = 1,2,\ldots,2n-1. \tag{20}$$

Polynomial p_n minimizing norm:

$$\min_{a,b,\ldots,c \in R} E(X^n + aX^{n-1} + bX^{n-2} + \ldots + c)^2 = E(p_n(X))^2 \tag{21}$$

is given by equation:

$$p_n(x) = K \begin{vmatrix} 1 & m_1 & \ldots & m_n \\ \ldots & \ldots & \ldots & \ldots \\ m_{n-1} & m_n & \ldots & m_{2n-1} \\ 1 & x & \ldots & x^n \end{vmatrix}, \tag{22}$$

where $K \neq 0$, and x is argument of a p_n.

Normed polynomial p_n is an orthogonal polynamial of order n (Szego, 1975). Hence:

$$p_n(x) = (x - s_1) \ldots (x - s_n), \text{ where } s_1 < \ldots < s_n. \tag{23}$$

Ordered $(s_{1, \ldots,} s_n) = M_n(X)$ is called n-average (multiaverage) of random variable X. This vector is an n-point approximation. Variance and standard deviation are, in this case, expressions:

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

$$\sqrt[2n]{V_n(X)} = \sqrt[2n]{E((X - s_1)\ldots(X - s_n))^2} \tag{24}$$

These characteristics measure mean-square deviation of random variable X from n probability concentration points.

In data analysis we usually find average, variance, etc. One can though go one step further. To do this, it is crucial to find moments of random variable X, and then determine 2-average (s_1, s_2), 3-average (s_1, s_2, s_3), etc. Finally, after finding all multiaverages which were needed, one has to find

$$V_1(X) = E(X - E(X))^2, V_2(X) = E((X - s_1)(X - s_2))^2, \text{ etc.}$$

It allows for a more precise data analysis.

Example. Two-modal Weber distribution.

Let random variable X have a density function of this form (we write $X \sim W(\alpha, \beta, \gamma)$)

$$g_{\alpha,\beta,\gamma}(x) = \frac{1}{z(\alpha, \beta)} e^{\alpha(x-\gamma)^2 - \beta(x-\gamma)^4};\ x, \gamma \in R;\ \alpha, \beta > 0. \tag{25}$$

The integration constant $z(\alpha, \beta)$ can be found by special Weber functions (Wilkowski, 2008; Bateman, Erdelyi, 1953) and is equal to:

$$z(\alpha, \beta) = \int_{\Re} e^{\alpha x^2 - \beta x^4} dx = \exp\left(\frac{\alpha^2}{8\beta}\right) \frac{\sqrt{\Pi}}{\sqrt[4]{2\beta}} D_{-\frac{1}{2}}\left(-\frac{\alpha}{\sqrt{2\beta}}\right), \tag{26}$$

where D is a special Weber function.

Obviously, function g has two modal values in:

$$Mo_1 = -\sqrt{\frac{\alpha}{2\beta}} + \gamma \ , \quad Mo_2 = \sqrt{\frac{\alpha}{2\beta}} + \gamma \ , \tag{27}$$

while expectation of random variable X is:

$$E(X) = \gamma \ . \tag{28}$$

Moments of higher rank are given by confluent hypergeometric functions. Let us then assume that $E(X) = 0$. We then have:

$$E(X^{2n-1}) = 0, \tag{29}$$

$$E(X^{2n}) = \frac{1}{2z(\alpha,\beta)} \beta^{\frac{1}{4}(3-2n)} \left[\sqrt{\beta}\Gamma\left(\tfrac{1}{4}+\tfrac{n}{2}\right) H\left(\tfrac{1}{4}+\tfrac{n}{2}, \tfrac{1}{2}, \tfrac{\alpha^2}{4\beta}\right) + \right. \\ \left. + \alpha\Gamma\left(\tfrac{3}{4}+\tfrac{n}{2}\right) H\left(\tfrac{3}{4}+\tfrac{n}{2}, \tfrac{3}{2}, \tfrac{\alpha^2}{4\beta}\right) \right], \tag{30}$$

where $n = 1, 2, \ldots$, and H denote confluent hypergeometric function (Bateman, Erdelyi, 1953).

The following picture shows density function of $W(2,1,1)$ distribution:

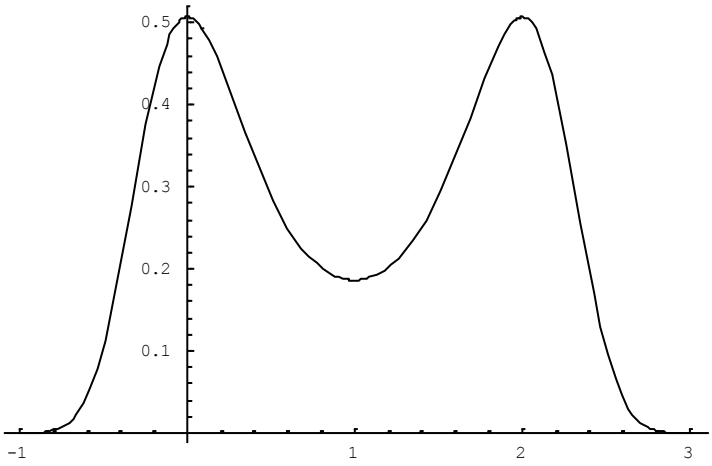

Fig. 3. Density function of $W(2,1,1)$ distribution

Source: author's own study.

Modal values are: $Mo_1 = 0, Mo_2 = 2$, while expectation $E(X) = 1$ is becoming "unexpected value" in this case (probability mass is concentrated around modal values). Polynomial p_2 from (23) is here given by:

$$p_2(x) = x^2 + ax + b = (x - s_1)(x - s_2), \tag{31}$$

where

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

$$s_1 = -0.08746, \quad s_2 = 1.91254. \tag{32}$$

Ordered pair $(s_1, s_2) = (-0.08746; 1.91253)$ is 2-average $M_2(X)$ of random variable X of distribution function $W(2,1,1)$. We also have:

$$V_1(X) = E(X - E(X))^2 = 0.83274, \tag{33}$$

$$V_2(X) = E[(X - s_1)(X - s_2)]^2 = 0.38928. \tag{34}$$

We can see that 2-average is a better approximation of modal values than expectation and $V_2 < V_1$, which could be expected. What is interesting is finding that another n-average will not improve approximation quality (in square-root sense). We have: $M_3(X) = (s_1, s_2, s_3) = (-0.76371; 0.82532; 2.69894)$ and $V_3(X) = E[(X - s_1)(X - s_2)(X - s_3)]^2 = 6.66794$.

Remark 1. If $E(X) = 0$, $E(X^2) = 1$, then

$$M_2(X) = \left(\frac{E(X^3) + \sqrt{E^2(X^3) + 4}}{2}, \frac{E(X^3) - \sqrt{E^2(X^3) + 4}}{2} \right). \tag{35}$$

The equation (35) follows from the fact that the polynomial p_2 from (31) has form given by:

$$p_2(x) = x^2 - E(X^3)x - 1. \tag{36}$$

Remark 2. Let X, Y be independent random variables, and $E(X) = E(Y) = 0$, $E(X^2) = E(Y^2) = 1$. If $M_2(X) = (s_1, s_2)$, $M_2(Y) = (t_1, t_2)$, then $M_2(X + Y) = (k_1, k_2)$, where

$$k_{1,2} = \frac{s_1 + s_2 + t_1 + t_2 \pm \sqrt{s_1 + s_2 + t_1 + t_2 - 16(s_1 s_2 + t_1 t_2)}}{4}. \tag{37}$$

In this case polynomial p_2 from the formula (31) has form:

$$p_2(x) = x^2 - \frac{E(X^3) + E(Y^3)}{2}x - 2. \tag{38}$$

4. Summary

To sum up, the aforementioned heuristic procedure allows for a more precise data analysis than before. Its base is the Pearson's moments method. The main advantage of moments is the fact that they can easily be calculated from a simple sample. Only combinations of moments, for instance variance, have a practical value. Line dependent coefficient (and others, which can be defined thanks to it) and multiaverage are such moments functions. Let us hope that line dependent coefficient and multiaverage will soon become practical tools in statistics and taxonomy.

Literature

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

Antoniewicz R. (2005). *O średnich i przeciętnych*. Wydawnictwo AE we Wrocławiu. Wrocław.

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

Brandt S. (1999). *Data Analysis. Statistical and Computational Methods for Statistics and Engineers*. 3rd editiom. Springer Verlag. New York.

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

McLachlan G., Peel D. (2004). *Finite Mixture Models*. John Wiley & Sons. New York.

Stanisz T. (1993). *Funkcje jednej zmiennej w badaniach ekonomicznych*. PWN. Warszawa.

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

Wilkowski A. (1995). *The coefficient of dependence for consumption curve*. Argumenta Oeconomica. No. 1.

Wilkowski A. (2009). *Uwagi o współczynniku korelacji*. Ekonometria. Vol. 27.

Wilkowski A. (2008). *Notes on normal distribution*. Didactics of Mathematics. No. 5.