УДК 519.2+004.4 **DOI 10.33042/2522-1809-2019-6-152-84-88**

А.Л. Литвинов

Харківський національний університет міського господарства імені О.М. Бекетова, Україна

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ РЕГРЕСІЙНО-КОРЕЛЯЦІЙНОГО АНАЛІЗУ ДЛЯ КОРЕЛЯЦІЙНОЇ РЕШІТКИ В EXCEL

Розрахунки по регресійно-кореляційному аналізу супроводжуються значною кількістю обчислень. Найбільш доступним додатком для таких розрахунків є Excel, в якому для цього є ряд вбудованих функцій. З їх допомогою можна обробляти тільки вибірки, задані координатами точок (x_i, y_i). У статті пропонується інформаційна технологія, яка розширює можливості Excel по обробці вибірок, заданих кореляційною решіткою.

Ключові слова: змінна, зв'язок, залежність, регресія, кореляція, Excel

Постановка проблеми

Регресійно-кореляційний аналіз знайшов широке використання в комунальному господарстві [1], будівництві [2], статистиці, де на відміну від, наприклад, фізики, залежності між двома змінними носять статистичний характер. Одному значенню змінної відповідає випадкове значення іншої змінної, як показано на рис.1. Це так зване кореляційне поле.

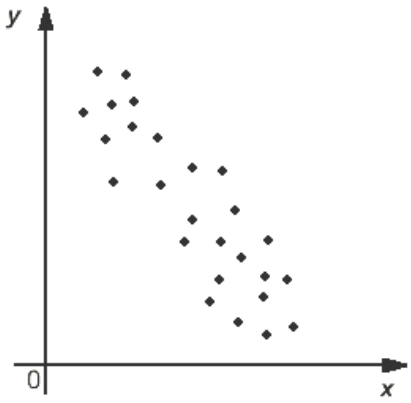

Рис. 1. Кореляційне поле

Розрахунки по регресійно-кореляційному аналізу супроводжуються значною кількістю обчислень за складними формулами. Найбільш доступним додатком для таких розрахунків є Excel, в якому для цього є ряд вбудованих функцій. З їх допомогою можна обробляти тільки вибірки, задані координатами точок (x_i, y_i). Але на практиці досить часто трапляються задачі, в яких вибірка є досить значною і вихідні дані для регресійно-кореляційного аналізу мають вигляд кореляційної решітки (див. табл. 1).

Розробка інформаційної технології для обробки таких вибірок в Excel значно підвищить якість регресійно-кореляційному аналізу і скоротить час на його проведення.

Таблиця1

Двовимірний статистичний розподіл

X	Y						
	y_1	y_2	…	y_i	…	y_m	n_{x_i}
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	n_{x_1}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	n_{x_2}
…	…	…	…		…		…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	n_{x_i}
…	…	…	…	…	…	…	…
x_k	n_{k1}	n_{k2}	…	n_{kj}	…	n_{km}	n_{x_k}
n_{y_j}	n_{y_1}	n_{y_2}	…	n_{y_i}	…	n_{y_m}	N

Аналіз останніх досліджень і публікацій

З моменту своєї появи додаток Excel привернув увагу фахівців з різних галузей своєї функціональної завершеністю, широкими можливостями, легкістю освоєння. В [3] наведені основні відомості по роботі з Microsoft Excel. Значну увагу приділено форматування даних, роботі з елементами таблиці, редагування таблиць. У розділі «Розв'язування задач за допомогою формул» наведені відомості щодо введення формул в клітини таблиці, їх редагування і копіювання, використання вбудованих функцій. Значна увага приділена кореляційному та регресійному аналізу за допомогою вбудованої функції КОРРЕЛ(діапазон1, діапазон2) та команди «Вставка⇒ Диаграммы...». Технології обробки кореляційної решітки немає. У фундаментальній праці [4] розглянуті питання використання Excel в економіці. Книга орієнтована на підготовлених користувачів, що освоїли ази Excel. Викладено технологію моделювання економічних систем і реалізацію моделей в Excel, зокрема фінансове моделювання, моделі прогнозування, моделі масового обслуговування. Особливу увагу приділено питанням оптимізації. Розгля-

 © Литвинов А.Л.

нуто задачі лінійного та нелінійного програмування, пошук екстремуму. Використанню Excel в кореляційному аналізу приділено мало уваги і то в контексті з імовірнісним моделюванням. Стосовно використанню Excel в математичної статистики заслуговує уваги навчальний посібник [5]. У ньому послідовно викладається основний навчальний матеріал з математичної статистики з орієнтацією на студентів інженерного та економічного профілю. Але в ньому немає послідовної технології регресійно-кореляційного аналізу стосовно кореляційної решітки.

Виклад основного матеріалу

Якщо статистична залежність проявляється у тому, що при зміні однієї з випадкових величин змінюється середнє іншої, то в цьому випадку статистична залежність називається ***кореляційною***. Наприклад, нехай X – сума витрат на підготовку лави, а Y – рівень видобутку вугілля. При однакових затратах на підготовку лав видобуток вугілля буде відрізнятися, тобто випадкова величина Y не є функцією від X. Це можна пояснити впливом ряду випадкових факторів (глибиною залягання пласта, його потужністю, сортністю вугілля і т.п.). Проте, середня величина видобутку вугілля є функцією від суми витрат, тобто випадкові величини Y і X пов'язані кореляційною залежністю. Як статистика при описі кореляційної залежності використовують рівняння лінійної регресії [6]:

$$\bar{y}(x) = \rho_{yx} x + b, \qquad (1)$$

де ρ_{xy} - вибірковий коефіцієнт регресії, b - зсув графіка лінії регресії по вісі 0y. Тісноту зв'язку між випадковими величинами Y і X описують вибірковим коефіцієнтом кореляції – $r_в$.

Нехай під час проведення експерименту отримана наступна вибірка, що відображає залежність випадкових величин Y і X (див. табл. 2)

Таблиця 2

Залежність випадкових величин X і Y

X	x_1	x_2	x_3	….	x_{n-1}	x_n
Y	y_1	y_2	y_3		y_{n-1}	y_n

Тоді для розрахунку значень ρ_{xy}, b, r_{xy} при кореляційному аналізі використовують наступні формули [6]:

- вибірковий коефіцієнт регресії

$$\rho_{yx} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})} = \frac{C_{xy}}{\sigma_x^2}, \qquad (2)$$

де $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ вибіркове середнє по x, $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ – вибіркове середнє по y, σ_x – середньо квадратичне відхилення по x, C_{xy} – так звана вибіркова коваріація;

– зміщення лінії регресії по вісі Y

$$b = \bar{y} - \frac{C_{xy}\bar{x}}{\sigma_x^2} \qquad (3)$$

– вибірковий коефіцієнт кореляції

$$r_в = \rho_{yx}\frac{\sigma_x}{\sigma_y} = \frac{C_{xy}}{\sigma_x \sigma_y}. \qquad (4)$$

Регресійно-кореляційний аналіз за вибірками, заданим аналогічно табл.2 можна ефективно проводити в додатку Excel с використанням стандартних функцій.

Для розрахунку середніх і використовуємо функцію = СРЗНАЧ () з категорії «статистичні». Для розрахунку середньоквадратичних відхилень σ_x і σ_y використовуємо комбінацію функцій = КОРІНЬ (ДІСП.В ()). Далі йде розрахунок коваріації, для чого використовується функція =КОВАРІАЦІЯ.В. Після цього за формулами (2) і (3) розраховуються коефіцієнт регресії, зміщення лінії регресії по осі 0y і коефіцієнт кореляції. Слід зауважити, що в Excel є спеціальна функція розрахунку коефіцієнта кореляції: = КОРРЕЛ (B2: L2; B3: L3). Її можна використовувати для контролю правильності обчислень за формулами (2), (3).

На рис.2 зображено робочий лист Excel з розрахунком за вищенаведеними формулами для вибірки невеликого об’єму.

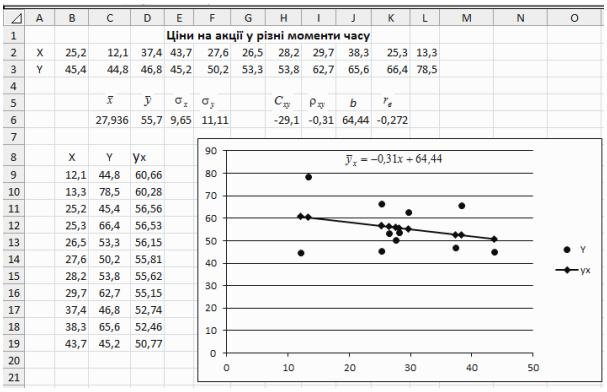

Рис. 2. Розрахунки по кореляції в Excel для малої вибірки

Слід зауважити, що безпосередньо графік рівняння регресії і саме рівняння в Excel можна отримати при побудові точкової діаграми. При побудові діаграми потрібно вибрати тип «Точкова з маркерами», потім на діаграмі клацнути правою кнопкою миші по маркеру і у вікні, що з’явиться вибрати «Додати лінію тренда» [7].

Розрахунки значно ускладнюються, якщо об'єм вибірки значний і спостерігаються пари (x_i, y_j) з однаковими ознаками. Згрупувавши повторювані пари (x_i, y_j) і підрахувавши частоти їхньої появи, одержимо двовимірний статистичний розподіл, який можна оформити в табличному вигляді (табл. 1), яка у кореляційному аналізі називається ***кореляційною решіткою.*** У ній: n_{ij} – частота спільної появи ознак x_i, y_j (пари (x_i, y_j)); $N = \sum_{i=1}^{k}\sum_{j=1}^{m} n_{ij}$ – об'єм вибірки; n_{x_i} – загальна частота появи ознаки x_i, $n_{x_i} = \sum_{j=1}^{m} n_{ij}, (i = 1,2,...,k)$; n_{y_j} – загальна частота появи ознаки y_j, $n_{y_j} = \sum_{i=1}^{k} n_{ij}, (j = 1,2,...,m)$.

За табл.1 можна розрахувати середнє *X* і *Y*, дисперсію та інші характеристики вибірки.

Середнє значення *X* і *Y* обчислюється за формулами [8]:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{k} n_{x_i} x_i ; \ \overline{y} = \frac{1}{N}\sum_{j=1}^{m} n_{y_j} y_j , \qquad (5)$$

Дисперсії *X* і *Y* знаходять за формулами:

$$\begin{aligned} D_x = \sigma_x^2 = \frac{1}{N}\sum_{i=1}^{k} n_{x_i}(x_i - \overline{x})^2 ; \\ D_y = \sigma_y^2 = \frac{1}{N}\sum_{j=1}^{m} n_{y_j}(y_j - \overline{y})^2 \end{aligned} . \qquad (6)$$

Вибірковий коефіцієнт регресії *Y* на *X* – ρ_{yx} у виразу (1) обчислюється за формулою:

$$\rho_{yx} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} n_{x_i y_j} x_i y_j - N \cdot \overline{x} \cdot \overline{y}}{N \cdot \sigma_x^2}, \qquad (7)$$

– величини зміщення лінії регресії по осі $0y$ – *b* обчислюється за формулою:

$$b = \frac{N \cdot \overline{x^2} \cdot \overline{y} - \overline{x} \cdot \sum_{i=1}^{k}\sum_{j=1}^{m} n_{x_i y_j} x_i y_j}{N \cdot \sigma_x^2}, \qquad (8)$$

–вибірковий коефіцієнт кореляції обчислюється за формулою:

$$r_{\text{в}} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} n_{x_i y_j} x_i y_j - N \cdot \overline{x} \cdot \overline{y}}{N \cdot \sigma_x \sigma_y} . \qquad (9)$$

На наступному приладі продемонструємо інформаційну технологію, яка пропонується для виконання регресійно-кореляційного аналізу з використанням програми MS Excel, якщо вихідні дані задані у вигляді кореляційної решітки (див. табл. 3).

Таблиця 3

Кореляційна решітка до прикладу

X	*Y*					n_x
	10	20	30	40	50	
10	5					
20	7	20				
30		23	30	10		
40			47	11	9	
50			2	20	7	
60				3	6	
n_y						

Наведені вище розрахунки можна виконати в системі Excel за такою методикою.

1. У клітинки A1:G9 поміщаємо електронний аналог таблиці 3.

2. У клітинки G3:G8 заносимо суми частот по рядках ($n_{x_i} = \sum_{j=1}^{m} n_{ij}$). Зокрема, в клітинку G3 заносимо = СУММ(B3:F3), в комірку G4 заносимо =СУММ(B4:F4) і т.д.

3. У клітинки B9:F9 заносимо суми частот по стовпцях ($n_{y_j} = \sum_{i=1}^{k} n_{ij}$). Зокрема, у комірку B9 заносимо = СУММ(B3:B8), у комірку C9 заносимо =СУММ(C3:C8) тощо.

4. У комірку G9 заносимо формулу розрахунку всіх частот по параметру *X* (формула =СУММ(G3:G8)), а в клітинку H9 для контролю заносимо формулу розрахунку всіх частот по параметру Y (формула = СУММ(B9:F9)). Частоти повинні співпасти, це буде об'єм вибірки *N*.

5. Розраховуємо середнє значення *X*, для чого в комірку A12 заносимо формулу =СУММПРОИЗВ(G3:G8;A3:A8)/H9.

Розраховуємо середнє значення *Y*, для чого в комірку B12 заносимо формулу =СУММПРОИЗВ(B9:F9;B2:F2)/H9.

Розраховуємо середньоквадратичне відхилення по X, для чого в комірку D12 заносимо формулу

=КОРЕНЬ(СУММПРОИЗВ(G3:G8;(A3:A8–$A$12)^2)/G9).

Розраховуємо середньоквадратичне відхилення по Y, для чого в комірку E12 заносимо формулу

=КОРЕНЬ(СУММПРОИЗВ(G3:G8;(A3:A8–$A$12)^2)/G9).

Розраховуємо другий початковий момент по X, для чого в комірку F12 заносимо формулу =СУММПРОИЗВ(G3:G8;(A3:A8)^2)/G9.

6. Розраховуємо $\sum_{i=1}^{k}\sum_{j=1}^{m} n_{x_i y_j} x_i y_j$. Для цього його перетворимо до наступного вигляду: $\sum_{i=1}^{k} x_i \sum_{j=1}^{m} n_{x_i y_j} y_j$ і спочатку розраховуємо $\sum_{j=1}^{m} n_{x_i y_j} y_j$, $i = 1,2,...,k$. Результат поміщаємо в комірки H3:H8. Розрахункові формули такі: у комірку H3 заносимо =СУММПРОИЗВ(B3:F3;$B$2:$F$2), у комірку H4 – =СУММПРОИЗВ(B4:F4;$B$2:$F$2) тощо. (У комірки H4:H8 формули заносяться шляхом копіювання з комірки H3).

У комірку I3 заносимо остаточне значення $\sum_{i=1}^{k}\sum_{j=1}^{m} n_{x_i y_j} x_i y_j$. Розрахункова формула: =СУММПРОИЗВ(A3:A8;H3:H8).

7. Розраховуємо коефіцієнт регресії Y на X – ρ_{yx}, зміщення лінії регресії по осі 0y – b, формули (7), (8) і вибірковий коефіцієнт кореляції $r_в$ – формула (9).

Для розрахунку ρ_{yx} у комірку H12 заносимо таку формулу:

=(I3-G9*A12*B12)/(G9*D12^2)

Для розрахунку b у комірку I6 заносимо таку формулу:

=(G9*F12*B12-A12*I3)/(G9*D12^2)

Для розрахунку $r_в$ у комірку J9 заносимо наступну формулу:

=H12*D12/E12.

8. Формуємо рівняння регресії Y на X

$$\bar{y}_x = 0,751x + 4,1983.$$

9. У математичній статистиці для оцінки достовірності лінійного регресійного аналізу використовується коефіцієнт детермінації R^2, який ще називають величиною достовірності апроксимації [9]. Для розрахунку коефіцієнта R^2 в Excel використовується функція

КВПИРСОН (масив Y; масив X) [10]. Слід врахувати, що при фіксованому X, Y має кілька значень. Має сенс для кожного значення X_i за формулою =СУММПРОИЗ (масив Y; масив X_i)/nx_i розрахувати середнє значення Y при конкретному X_i. Відповідні розрахунки виконані в клітинках K3: K8. За допомогою формули = КВПИРСОН (K3:K8; A3:A8), яка поміщена в клітинку K12, проведено розрахунок коефіцієнта детермінації R^2. Він дорівнює 0,9895, що свідчить про хорошу величиною достовірності апроксимації.

Результати розрахунків разом з графіком лінії регресії і полем кореляції (без урахування частот окремих точок) наведені на рисунку 3.

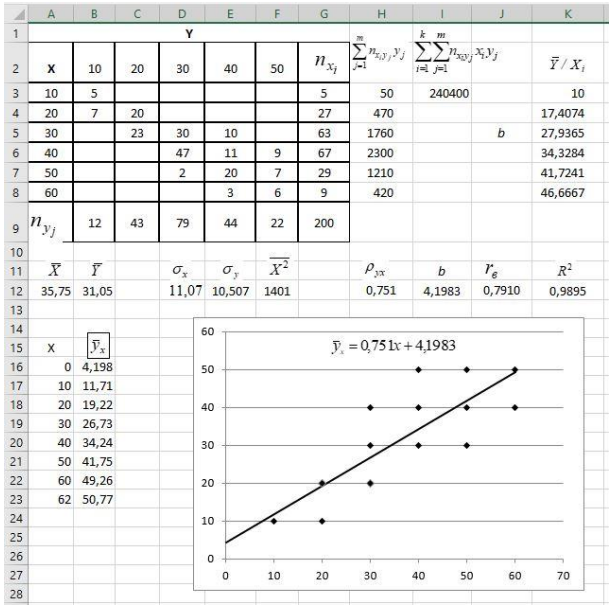

Рис. 3. Розрахунки по регресійно-кореляційному аналізу в Excel

З аналізу випливає: коефіцієнт кореляції досить високий, це обумовлює тісний зв'язок між змінними X і Y. Коефіцієнт детермінації також досить високий, що свідчить про хорошу величиною достовірності апроксимації.

Висновки

Апробація запропонованої інформаційної технології при викладанні курсу «Математична статистика» для інженерних та економічних спеціальностей дозволила значно прискорити проведення розрахунків при виконанні практичних занять та індивідуально-розрахункового завдання, подолати розрив між досить складним теоретичним матеріалом і його практичним застосуванням.

Література

1. Савіна, Г. Г. Практичні аспекти оцінювання ефективності управління підприємством комплексу комунальних послуг [Текст]/ Г. Г. Савіна, Т. І. Скібіна // Інвестиції: практика та досвід. № 24. Київ: Видавництво ТОВ «ДКС-центр», 2016. - С. 37 – 41.

2. Чичулін, В. П. Аналіз кореляційних зв'язків стохастичних характеристик будівельних конструкцій [Текст] / В. П. Чичулін, К. В. Чичуліна // Збірник наукових праць. Серія: галузеве машинобудування, будівництво. Вип. 1(43). Полтава: ПолтНТУ, 2015. С 81 – 86.

3. Завадський, І. О. Microsoft Excel у профільному навчанні : навч. посіб [Текст] / І. О. Завадський, А. П. Забарна. – Київ : Вид. група ВНУ, 2011. – 272 с.

4. Мур, Д. Экономическое моделирование в Microsoft Excel [Текст]/ Д. Мур, Л. Р. Уэдерфорд. – М. : Вильямс, 2004. – 1024 с.

5. Воскобойников, Ю. Е. Математическая статистика (с примерами в Excel) : учеб. пособие [Текст] / Ю. Е. Воскобойников. Е. И. Тимошенко. – Новосибирск : Новосиб. гос. архитектур.-строит. ун-т (Сибстрин), 2006. – 152 с.

6. Волощенко, А. Б. Теорія ймовірностей та математична статистика : навч.-метод. посібник для самост. вивч. Дисц [Текст] / А. Б. Волощенко, І. А. Джалладова. – Київ : КНЕУ, 2003. – 256 с.
7. Карлберг, К. Регрессионный анализ в Microsoft Excel [Текст] / К. Карлсберг. – СПб.: Альфа-книга, 2019. – 396 с.
8. Гмурман, В. Е. Теория вероятностей и математическая статистика [Текст] / В. Е. Гмурман. – М. : Высш. шк., 2002. – 479 с.
9. Кендалл, М. Статистические выводы и связи [Текст] / М. Кендал, А. Стюарт. М.: Наука, 1973. – 899 с.
10. Додж, М. Эффективная работа: Microsoft Office Excel [Текст] / М. Додж, К. Стинсон. СПб.: Питер, 2005. – 1088 с.

References

1. Savina, G. G. (2016). Practical aspects of evaluating the efficiency of enterprise management of a complex of public utilities. *Investment: practice and experience. Kiev, DKS-center, 24,* 37 – 41.
2. Chichulin, V. P. (2015). Correlation analysis of building structures stochastic characteristics. *Collection of scientific papers. Series: industrial engineering, construction. Poltava, PoltSTU, 1(43),* 81 – 86 .
3. Zavadsky, I. O. (2011). Excel in core training: a tutorial. Kiev, VHU, 272 p.
4. Moore, D., & L. R. Weatherford (2004). Decision Modeling in Microsoft Excel . Moscow, Williams.
5. Voskoboinikov, U. E. , & Tymoshenko, E. I.(2006). Mathematical Statistics (with examples in Excel), Novosibirsk, Novosib. state architecture.-builds university.
6. Voloshenko, A. B. (2003). Theory of Probability and Mathematical Statistics: a tutorial, Kiev, KNEU.
7. Carlberg, C.(2019). Regression analysis in Microsoft Excel, SPb., Alpha book 2019 396 c.
8. Gmurman, V. E. (2002). Theory of Probability and Mathematical Statistics, Moscow, Higher school.
9. Kendall, M., & Stuart, A. (1973). Statistical Conclusions and Relationships, Moscow, Science.
10. Dodge, M., & Stinson, C. (2005). Affective work: Microsoft Office Excel, SPb., Piter.

Рецензент: д-р фізмат. наук, проф. А.В. Грицунов, Харківський національний університет радіоелектроніки, Харків, Україна.

Автор: ЛИТВИНОВ Анатолій Леонідович
доктор технічних наук, професор.
Харківський національний університет міського господарства імені О.М. Бекетова
E-mail – litan6996@gmail.com
ID ORCID: http://orcid.org/0000-0001-7063-7814

INFORMATION TECHNOLOGY REGRESSION CORRELATION ANALYSIS FOR THE CORRELATION LATTICE IN EXCEL

A. Litvinov
O.M. Beketov Natioal University of Urban Economy in Kharkiv, Ukraine

Regression-correlation analysis has found wide application in communal economy, buildings and statistics, where, unlike, for example, physics, the dependencies between two variables are functional in nature. One value of a variable corresponds to a random value of another variable, which together forms a so-called correlation field. Calculations by regression-correlation analysis are accompanied by a significant number of calculations using complex formulas, which complicates its application. The most affordable application for such calculations is Microsoft Excel, which has a number of built-in functions for this. However, with their help it is possible to process only the samples defined by the coordinates of the points (x_i, y_i), which are forming a table with two rows. But in practice, quite often there are problems in which the sample is large enough and the initial data for the regression-correlation analysis are in the form of a correlation lattice. The article proposes an information technology that extends Excel's ability to process samples specified by the correlation lattice. It consists of a number of stages, the results of each are used in the following and are based on the electronic analog of the correlation lattice. Firstly, using the sum (array) function, the frequencies of the variables that are correlated are calculated. To calculate the mean and standard deviation, the sumproduct (array 1, array 2) function with the corresponding modification is used. The most difficult procedure in calculating the covariance and correlation coefficient is to calculate the sum of the pairwise products of the studied variables. An algorithm for its implementation has been developed, and standard Excel tools are used to build plots. To assess the reliability of the regression analysis, the determination coefficient R^2 based on the built-in RSQ(array Y; array X) function was used. An algorithm is developed for using this function as applied to the correlation lattice. The operability of the developed information technology for regression-correlation analysis for the samples defined by the correlation lattice is demonstrated by a numerical example.

Keywords: *variable, relationship, dependence, regression, correlation, Excel*