Модифицированный коэффициент корреляции

© Т.О. Дюкина

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

dtospb@mail.ru t.dukina@spbu.ru

Аннотация. Статья посвящена рассмотрению показателя – коэффициента корреляции Пирсона, его положительным и отрицательным сторонам применения для анализа динамики и связи между явлениями, а также последующей его модификации. Модификация коэффициента корреляции осуществлена на основе замены способа расчета элементов формулы: средних значений. Осуществлена апробация предложенного модифицированного коэффициента корреляции и доказано его преимущество в более точной оценке тесноты связи между вариацией исследуемого фактора и изменением показателя, характеризующего стабильность налоговой системы страны, на эмпирических данных.

Ключевые слова: коэффициент корреляции Пирсона, модифицированный коэффициент корреляции, средняя арифметическая, средняя геометрическая, вариация, динамика, оценка тесноты связи.

The Modified Correlation Coefficient

© Tatiana Dyukina

St. Petersburg State University,
St. Petersburg, Russia

dtospb@mail.ru t.dukina@spbu.ru

Abstract. Article is devoted to consideration of an indicator – coefficient of correlation of Pearson, to his positive and negative sides of application for the analysis of dynamics and communication between the phenomena, and also the subsequent its modification. Modification of the correlation coefficient is carried out on the basis of replacement of a way of calculation of elements of a formula to be performed on the average values. Approbation of the modified correlation coefficient has been carried out and its merits revealed in more exact assessment of the closeness of links between variation of a studied factor and change of the indicator characterizing stability of the country tax system on empirical data have been shown.

Keywords: Pearson correlation coefficient, modified correlation coefficient, arithmetic average, geometric average, variation, dynamics, assessment of closeness of links.

1 Введение

Сегодня вопросам состояния, развития, а также совершенствования статистических методов начинает уделяться повышенное внимание. Это не случайно, так как именно статистические методы предоставляют широкие возможности своевременного и полного анализа разнообразных данных и получения в результате их обработки качественных выводов.

Исследования, которые посвящены решению не только методологических вопросов статистического анализа в различных сферах экономики и общества, но и оценке универсальности и специализации методов, систематизации опыта применения статистических методов при решении различного рода практических задач, а также развитию и созданию новых методов анализа данных, встречались чаще в прошлом столетии по сравнению с сегодняшним днем. В настоящее время такие исследования являются относительно большой редкостью. Таким образом, вопросы совершенствования статистических методов, в том числе отдельных статистических показателей, приобретают еще большую актуальность в свете обозначенных аспектов.

Данная статья посвящена совершенствованию методики расчета одного из наиболее употребляемых в статистической практике анализа данных различных показателей – для оценки тенденции ряда динамики и тесноты связи между показателями. Для статистического анализа тенденции ряда динамики, а также тесноты связи между вариацией исследуемого фактора и изменением изучаемого показателя применение находит широко известный показатель: коэффициент корреляции Пирсона.

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

2 Анализ степени исследования проблемы

2.1 Показатели, применяемые для измерения стабильности (устойчивости) тенденции ряда динамики

Действительно, для измерения стабильности (устойчивости) тенденции ряда динамики среди рекомендованных к применению показателей: коэффициента корреляции рангов Ч. Спирмена (C.E. Spearman) [9, с. 345; (Spearman)] и соотношения между среднегодовым абсолютным изменением и средним квадратическим (либо линейным) отклонением уровней от тренда [9, с. 347] индексу корреляции, показывающему степень сопряженности колебаний фактических уровней с колебаниями теоретических уровней, происходящих под влиянием комплекса основополагающих факторов, и представляющему собой коэффициент корреляции Пирсона (Pearson), или иначе, линейный коэффициент корреляции [7, с. 475] отводится одно из самых важных, можно сказать, эпохальных мест.

Здесь следует акцентировать внимание на том, что, во-первых, коэффициент корреляции Пирсона рекомендуется использовать исключительно в случаях линейной связи. В случаях нелинейной связи, которые встречаются наиболее часто, применение данного показателя нежелательно. Во-вторых, слабым местом данного показателя является его неверное реагирование на выбросы: результаты измерения, выделяющиеся из общей совокупности (слишком большие или малые значения) могут способствовать большим значениям данного показателя. В таком случае, они означают высокую степень сопряженности колебаний фактических уровней с колебаниями теоретических уровней, происходящих под влиянием комплекса основополагающих факторов. В-третьих, в случаях, когда одна из двух переменных не является нормально распределенной (а, как показывает анализ множества эмпирических данных, имеющих экономическую природу, большинство таких данных собственно и не являются нормально распределенными), а также в случаях, когда одна из двух переменных имеет порядковую шкалу измерения, коэффициент корреляции Пирсона неприменим. В этих случаях рекомендуется использовать только ранговые коэффициенты Спирмена и Кендалла.

2.2 Показатели, используемые для количественной оценки влияния отдельных факторов на анализируемый показатель

Для определения количественной оценки влияния отдельных факторов на анализируемый показатель в настоящее время имеется возможность применять различные методы: индексный анализ, дисперсионный анализ, корреляционно-регрессионный, эконометрический анализ и другие. В последнее время наибольшее распространение в научных исследованиях получили методы корреляционно-регрессионного и эконометрического анализа. Обозначенные методы для определения количественной оценки влияния отдельных факторов на уровень стабильности налоговой системы страны требуют осмотрительного, пунктуального и вдумчивого применения, поскольку в процессе их применения могут возникать целый ряд еще неразрешенных в науке в полном объеме проблем:

- использование неполного комплекта влияющих факторов;
- построение моделей, которые содержат ненаблюдаемые факторы;
- ложная причинно-следственная связь, в том числе возникающая из-за употребления в анализе замещающих факторов [1].

Следовательно, широкое применение рассматриваемого в настоящей статье показателя – коэффициента корреляции Пирсона, особенно без учета его особенностей и специфики применения, может привести к неверным расчетам и выводам. Отмеченное становится особенно актуальным в случаях нелинейности развития анализируемых показателей, характеризующих экономическую среду.

3 Методологические вопросы разработки модифицированного коэффициента корреляции

3.1 Особенности экономической среды

Многие экономисты, в числе которых Дж. Кейнс, считают экономическую среду непредсказуемой и изменчивой [5]: «экономическая среда на протяжении некоторого периода времени должна оставаться неизменной и однородной во всех значимых отношениях, за исключением колебаний тех факторов, которые рассматриваются отдельно» [3]. «Но быть уверенными, что такие условия сохранятся в будущем, даже если они обнаруживаются в прошлом, нельзя», – заключает ученый [3].

Действительно, большинство экономических переменных (факторов) взаимодействуют посредством многообразных нелинейных зависимостей. Однако арсенал эконометрической науки сегодня довольно богат, что позволяет успешно решать проблемные вопросы при моделировании социально-экономических процессов и явлений.

3.2 Коэффициент корреляции Пирсона

Как уже отмечено выше, в случае линейных зависимостей широкое применение для определения тесноты связи находит коэффициент корреляции Пирсона (см. формулу 1).

$$K_{xy} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

где K_{xy} – значение коэффициента корреляции Пирсона, $\bar{x}$, $\bar{y}$ – средние значения уровней показателя, рассчитываемые по формуле арифметической средней [8, с. 224].

3.3 Модифицированный коэффициент корреляции

Следует учитывать, что довольно большой объем факторов, вариация которых оказывает влияние на изменение анализируемого показателя, подчиняется законам распределения, характер которых отличен от нормального распределения. Представляется, что среднее значение показателя, рассчитанное по формуле арифметической средней, в этих распределениях не является истинным. В этом случае расчет среднего значения исследуемого показателя по геометрической средней, учитывающим большой разброс значений показателя, представляется более корректным. Вследствие этого, полагаем возможным осуществить модификацию коэффициента корреляции Пирсона посредством введения в нее вместо среднего значения, определяемого по формуле арифметической средней, среднего значения, рассчитанного по формуле геометрической средней Пирсона (см. формулу 2).

$$K_{xy}{}^{M} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x}_{геом})(y_i - \bar{y}_{геом})}{\sum_{i=1}^{n}(x_i - \bar{x}_{геом})^2 \sum_{i=1}^{n}(y_i - \bar{y}_{геом})^2}} \quad (2)$$

где $K_{xy}{}^{M}$ – значение модифицированного коэффициента корреляции, $\bar{x}_{геом}$, $\bar{y}_{геом}$ – средние значения уровней фактора и результативного показателя, определяемые по формуле геометрической средней.

При исследовании совокупностей с качественно разнородными признаками на первый план выступает именно нетипичность средних показателей. Средняя геометрическая величина позволяет осуществить обобщение качественно разнородных значений признаков системных пространственных совокупностей или статистических совокупностей, представленных в динамике (во времени). Она, обнаруживая общие свойства исследуемых совокупностей, которые присущи всем единицам соответствующих совокупностей, позволяет выявить общие закономерности, обусловленные общими причинами, а также избежать случайных влияний.

При модификации коэффициента корреляции Пирсона была выбрана именно средняя геометрическая величина, так как она позволяет наилучшим образом осуществить обобщение значений признака в исследуемой совокупности не только в случаях наличия экстремальных значений отдельных единиц изучаемой статистической совокупности, но и в случаях распределений, принимающих характер, отличающийся от нормального закона распределения.

На наш взгляд, замена средней арифметической величины при модификации коэффициента корреляции Пирсона на другие статистические величины (например, медианное значение, модальное значение, а также иные робастные величины) не рациональна, поскольку не позволит в должной мере обеспечить устойчивость меры среднего. Кроме того, стоит отметить тот факт, что использование Пирсоном модального и медианного значений в известных формулах асимметрии распределений не сделало их более совершенными, наоборот, они общепризнанно считаются весьма приблизительными и довольно часто показывают некорректные значения этого показателя.

Возможности модифицированного коэффициента корреляции (по сравнению с коэффициентом корреляции Пирсона) более обширны: его можно применять для оценки тесноты связи между вариацией исследуемого фактора и изменением показателя, характеризующего анализируемый показатель, в случаях, когда характер распределений исследуемого фактора и (или) показателя, его характеризующего, отличается от закона нормального распределения (поскольку применение среднего значения, рассчитанного по геометрической средней, позволяет корректно учитывать большой разброс значений показателя в распределениях, отличных от нормального закона распределения). В результате модифицированный коэффициент корреляции позволит наиболее точно определять силу влияния вариации фактора на изменение исследуемых показателей.

4 Апробация модифицированного коэффициента корреляции

4.1 Данные и выборка

В настоящем исследовании осуществлена также апробация предложенного модифицированного коэффициента корреляции и эмпирически доказано его преимущество, заключающееся в более точной оценке тесноты связи между вариацией исследуемого фактора и изменением показателя, характеризующего анализируемый показатель.

В качестве исследуемого показателя был выбран показатель, характеризующий стабильность налоговой системы нашей страны – средняя фактическая налоговая нагрузка на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации, а в качестве фактора – уровень заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения.

Исследования осуществлены в динамике за период 2010-2014 гг. на основе официальных статистических данных в разрезе субъектов Российской Федерации, формируемых Федеральной налоговой службой России и Федеральной службой государственной статистики.

Средняя фактическая налоговая нагрузка на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации определена на основе данных ФНС [6]. Федеральная налоговая службой России представляет данные в свободном доступе в целом по Российской Федерации и в разрезе ее субъектов за период с 2007 г. по настоящее время в формах статистической налоговой отчетности.

Уровень заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения, рассчитан на основе данных Федеральной службы государственной статистики [2, 4].

Поскольку данные по исследуемым показателям были взяты в разрезе субъектов Российской Федерации, следовательно, в работе был применен сплошной метод исследования.

4.2 Эмпирические результаты исследования

Предварительно был осуществлен анализ исследуемого показателя, характеризующего стабильность налоговой системы нашей страны – средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации на основе расчета показателей центра, структуры, степени вариации и типа распределения и установлен характер распределения субъектов РФ налоговой системы по показателю средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации за период 2010-2014 гг. (см. Таблицу 1).

Таблица 1 Показатели центра, структуры, степени вариации и типа распределения средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации за период 2010-2014 гг.

Показатели	Годы				
	2010	2011	2012	2013	2014
1	2	3	4	5	6
Средняя арифметическая	62	81	91	91	104
Средняя геометрическая	31	37	42	43	49
Медианное значение	27	31	36	36	41
Размах вариации	901	1 179	1 340	1 224	1 311
Среднее линейное отклонение	61	83	92	91	104
Среднее квадратическое отклонение	141	191	222	217	238
Коэффициент вариации, %	225,7	237,0	244,1	237,9	229,2
Коэффициент асимметрии	4,79	4,72	4,77	4,71	4,57
Коэффициент эксцесса	23,34	22,35	22,37	21,45	20,36

Источник: рассчитано автором

Анализ средних и медианных значений изучаемого показателя за период 2010-2014 гг. (рассчитанных по несгруппированным данным), свидетельствует об их стабильном увеличении на протяжении всего исследуемого периода, что означает положительные изменения исследуемого показателя на макроуровне и, как следствие, направленность изменений в сторону стабильного развития налоговой системы в Российской Федерации (следует отметить, что здесь сказывается влияние инфляционного фактора). Однако все показатели вариации, а также коэффициенты асимметрии и эксцесса (рассчитанные по несгруппированным данным), являются более тонким инструментом, позволяющим учитывать влияние случайных факторов на исследуемый показатель, и указывают на постоянную и довольно существенную вариацию значений рассматриваемого показателя. Анализ коэффициента вариации в исследуемом периоде показал также, что в РФ совокупности субъектов по исследуемому показателю за период 2010-2014 гг., чрезвычайно неоднородные, вариация по субъектам РФ значительная, так как превышает не только 33%, но и 100%, что свидетельствует о крайней нестабильности налоговой системы в пространственном аспекте за анализируемый период. Следует отметить уменьшение значений характеристик распределения (коэффициентов асимметрии и эксцесса) по исследуемому показателю в 2014 г. по сравнению с 2010 г., пусть и незначительное, но, тем не менее, это указывает на позитивные изменения, происходящие в развитии налоговой системы страны.

Среднее значение анализируемого показателя, рассчитанного по формуле геометрической средней в два и более раз меньше, чем аналогичное значение, рассчитанное по формуле арифметической средней,

на протяжении всего исследуемого периода. При этом именно значения показателя, рассчитанные по формуле геометрической средней, наиболее приближены к медианным значениям, что косвенно подтверждает их преимущество в выявлении истинного среднего значения в исследуемой совокупности.

Таким образом, анализ показателей центра, структуры, степени вариации и типа распределения исследуемого показателя за период 2010-2014 гг. позволяет констатировать, что распределение изучаемого показателя на протяжении всего рассматриваемого периода имеет характер гиперэкспоненциального распределения.

На основе данных показателя, характеризующего стабильность налоговой системы нашей страны – средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации, и его фактора – уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения – в динамике за каждый год периода с 2010 по 2014 гг. был рассчитан модифицированный коэффициент корреляции, а также модифицированный коэффициент детерминации (рассчитываемый возведением в квадрат модифицированного коэффициента корреляции). Результаты расчетов эмпирических исследований по расчету коэффициентов корреляции и детерминации (в том числе модифицированных) представлены в таблице 2.

Таблица 2 Коэффициенты корреляции и детерминации взаимосвязи средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения за период 2010-2014 гг.

Показатели	Годы				
	2010	2011	2012	2013	2014
1	2	3	4	5	6
Коэффициент корреляции	0,628	0,600	0,572	0,524	0,457
Коэффициент детерминации	0,395	0,360	0,327	0,274	0,209
Модифицированный коэффициент корреляции	0,410	0,417	0,323	0,203	0,184
Модифицированный коэффициент детерминации	0,168	0,174	0,104	0,041	0,033

Источник: рассчитано автором

Для более наглядного представления информации представим полученные коэффициенты корреляции на графике (см. Рисунок 1) .

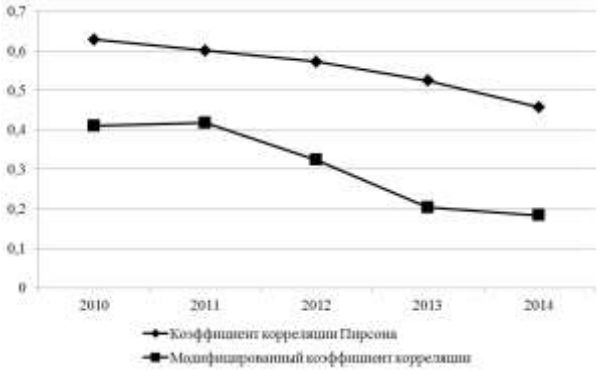

Рисунок 1 Коэффициенты корреляции взаимосвязи средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения за период 2010-2014 гг.

5 Заключение

На основе анализа данных таблицы 2 и рис.1 можно констатировать, что за весь рассматриваемый период уровень значений модифицированного коэффициента корреляции по сравнению с коэффициентом корреляции Пирсона является существенно более низким. Это означает существенно более низкую (в данном случае, в отдельные годы даже более чем в два раза) взаимосвязь вариации средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иным обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения за период 2010-2014 гг.

Аналогичный вывод можно сделать и по рассчитанным коэффициентам детерминации: за весь анализируемый период уровень значений модифицированного коэффициента детерминации оказался меньше, чем у коэффициента корреляции Пирсона.

Кроме того, выявленный характер взаимосвязи вариации исследуемых показателей заметно отличается, что особенно хорошо заметно на графике. Модифицированный коэффициент корреляции имеет более высокие темпы снижения для исследуемых эмпирических показателей по сравнению с коэффициентом корреляции Пирсона, что является следствием более точного учета влияния дисбаланса анализируемой экономической системы, в которой были выявлены нелинейные процессы развития.

Следовательно, использование модифицированных коэффициентов корреляции и детерминации позволяет получить, по нашему мнению, более точную оценку взаимосвязи изменений средней фактической налоговой нагрузки на одного налогоплательщика по налогам, сборам и иных обязательным платежам в бюджетную систему Российской Федерации и уровня заболеваемости всего населения с диагнозом, установленным впервые в жизни на 1000 человек населения.

Литература

[1] Hendri D. Econometrics: alchemy or science? Ekovest, No. 2. pp. 172 – 196 (2003)

[2] Incidence of the population on the main classes of diseases. http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/healthcare/#

[3] KeynesJ. M. Method of professor Tinbergen. Economy questions. No. 4. P. 28 (2007)

[4] Population. http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/statistics/population/demography/#

[5] Rozmainsky I. Methodological bases of the theory of Keynes and his "dispute on a method" with Tinbergen. Economy questions. No. 4. pp. 25-36 (2007)

[6] Summary reports in general on the Russian Federation and in a section of subjects of the Russian Federation. https://www.nalog.ru/rn78/related_activities/statistics_and_analytics/forms/

[7] The tendency of property stratification only accrues. Experts warn about danger of social explosion in Russia because of property stratification [An electronic resource]. http://www.newizv.ru/economics/2014-10-17/209143-tendencija-imushestvennogo-rassloenija-tolko-narastaet.html. – Zagl. from the screen (2014)

[8] Theory of statistics. Under the editorship of the prof. G. L. Gromyko. 2nd prod., reslave. and additional Moscow. 476 p. (2006)

[9] Yeliseyeva I. I., Yuzbashev M. M. General theory of statistics. 4 prod., reslave and additiona. Moscow. 480 p. (1999)