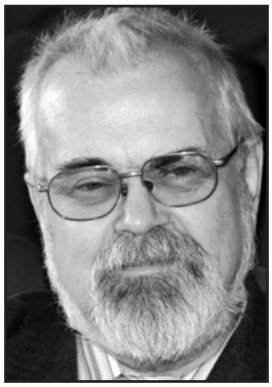

Орлов
Антон Александрович
ассистент кафедры «Экономика и организация производства», МГТУ им. Н.Э. Баумана

Орлов
Александр Иванович
доктор экон. наук, доктор техн. наук, канд. физ.-мат. наук, профессор, зав. НИЛ «Экономико-математические методы в контроллинге», МГТУ им. Н.Э. Баумана

УДК 303.5:519.2

КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ: ШКАЛА ЧЕДДОКА И ЗНАЧИМОСТЬ

Согласно вероятностно-статистической модели исходные данные - выборка из двумерного распределения. Введены коэффициенты корреляции Пирсона, Спирмена и Кендалла. Показана некорректность термина «корреляционно-регрессионный анализ». Корреляционный анализ позволяет оценивать степень связи, прогнозировать значение одной переменной по значению другой, но не позволяет управлять. Рассмотрен ряд вариантов шкалы Чеддока. Выборочные коэффициенты корреляции асимптотически нормальны, когда теоретические равны 0.

Ключевые слова: *корреляция, вероятностно-статистическая модель, коэффициент корреляции Пирсона, коэффициент корреляции Спирмена, коэффициент корреляции Кендалла, шкала Чеддока, проверка гипотез.*

Orlov Anton, *assistant of the Economics and Organization of Production Department, BMSTU*
Orlov Alexander, *Doctor of Economics, Doctor of Technical Sciences, Candidate of Physical and Mathematical Sciences, Professor, head of the Research Laboratory «Economic and Mathematical methods in controlling», BMSTU*

CORRELATION COEFFICIENTS: THE CHEDDOCK SCALE AND SIGNIFICANCE

According to the probabilistic and statistical model, the initial data is a sample from a two-dimensional distribution. Pearson, Spearman and Kendall correlation coefficients are introduced. The incorrectness of the term "correlation and regression analysis" is shown. Correlation analysis allows you to assess the degree of connection, predict the value of one variable from the value of another, but does not allow you to control. Several variants of the Cheddock scale are considered. The sample correlation coefficients are asymptotically normal when the theoretical ones are 0.

Keywords: *correlation, probabilistic statistical model, Pearson correlation coefficient, Spearman correlation coefficient, Kendall correlation coefficient, Cheddock scale, hypothesis testing.*

Введение

Термин «корреляция» означает «связь между переменными». Применительно к анализу данных этот термин обычно используется в сочетании «коэффициент корреляции». Такие коэффициенты применяют для измерения величины и направленности связи между случайными переменными.

В [1] приведены результаты поиска публикаций в научной электронной библиотеке eLIBRARY.RU по ключевым словам: «Корреляция», «Корреляция Пирсона», «Корреляция Спирмена», «Корреляция Кендалла». В табл. 1 дана краткая выдержка.

Данные табл. 1 показывают, что методы изучения корреляции широко применяются при анализе данных в различных областях знаний. Однако, как показано ниже, многие вопросы требуют тщательного рассмотрения. Им и посвящена настоящая статья.

Важно отметить, что большое число авторов не сообщают, какой именно коэффициент корреляции они используют. В таких случаях чаще всего речь идет о коэффициенте корреляции Пирсона.

Коэффициенты корреляции

Как показано в [2], описание методов анализа данных следует начинать с формулировки соответствующей вероятностно-статистической модели.

Пусть $(X, Y) = (X(\omega), Y(\omega))$ – двумерный случайный вектор, т.е. функция, определенная на пространстве элементарных событий $\Omega = \{\omega\}$ со значениями в R^2. Согласно устаревшей парадигме математических методов исследования часто предполагают, что распределение (X, Y) является двумерным нормальным. Однако хорошо известно, что распределения реальных данных, как правило, не являются нормальными (гауссовскими) [3].

Поэтому будем считать, что распределение случайного вектора (X, Y) произвольно, т.е. будем рассматривать непараметрическую модель. При этом примем, что выполнены обычные предположения [3] Центральной предельной теоремы теории вероятностей, позволяющие заключить о справедливости приведенных ниже асимптотических утверждений.

Для измерения связи между координатами случайного вектора (X, Y) используют тот или иной коэффициент корреляции. Среди них наиболее известен линейный парный коэффициент корреляции Пирсона:

$$r = \frac{M\{(X - M(X))(Y - M(Y))\}}{\sigma(X)\sigma(Y)},$$

где $M(X)$ – математическое ожидание случайной величины X, а $\sigma(X)$ – ее среднее квадратическое отклонение (т.е. квадратный корень из дисперсии).

В прикладной статистике исходные данные – это выборка, т.е. набор n пар чисел (т.е. пдвумерных векторов) (x_i, y_i), $i = 1, 2, ..., n$, где n – объем выборки. В рассматриваемой вероятностно-статистической модели элементы выборки – независимые двумерные случайные вектора, одинаково распределенные с $(X, Y) = (X(\omega), Y(\omega))$. Выборочным линейным парным коэффициентом корреляции Пирсона r_n, как известно, называется число:

$$r_n = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Здесь $\overline{x}$ – среднее арифметическое чисел x_i, $i = 1, 2, ..., n$.

Если $r_n = 1$, то $y_i = ax_i + b$, $i = 1, 2, ..., n$, при некоторых числах a и b, причем $a > 0$. Если же $r_n = -1$, то также имеется линейная связь между переменными x_i и y_i, т.е. $y_i = ax_i + b$, $i = 1, 2, ..., n$, но здесь $a < 0$. Из этих утверждений можно сделать вывод, что близость коэффициента корреляции к 1 (по абсолютной величине) говорит о достаточно тесной линейной связи между рассматриваемыми переменными.

В рассматриваемой вероятностно-статистической модели выборочный коэффициент корреляции является состоятельной оценкой теоретического, т.е. сходится (по вероятности) к теоретическому коэффициенту при безграничном возрастании объема выборки:

$$\lim_{n \to +\infty} r_n = r.$$

В теоретических рассмотрениях часто считают, что случайные вектора $(x_i, y_i) = (x_i(\omega), y_i(\omega))$, $i = 1, 2, ..., n$, имеют двумерное нормальное распределение. Как уже говорилось, распределения реальных данных, как правило, отличны от нормальных [3].

Почему же распространено представление о двумерном нормальном распределении? Дело в том, что теория в этом случае проще. В частности, равенство 0 теоретического коэффициента корреляции эквивалентно независимости случайных величин. Если же предположение о двумерной нормальности не выполнено, то из равен-

Количество результатов поиска

Таблица 1

Тематика поиска	Всего	Экономика	Математика
Корреляция	38614	11266	5505
Коэффициент корреляции Пирсона	11922	974	368
Коэффициент корреляции Спирмена	12545	1221	333
Коэффициент корреляции Кендалла	4301	135	87

ства 0 теоретического коэффициента корреляции не вытекает независимость случайных величин. Нетрудно построить пример случайного вектора, для которого коэффициент корреляции равен 0, но координаты зависимы. Поэтому целесообразно использовать непараметрические коэффициенты корреляции, пригодные при любом непрерывном распределении случайного вектора.

Для анализа данных, измеренных в шкалах интервалов и отношений, можно применять коэффициент линейной парной корреляции Пирсона, поскольку его значение не меняется при допустимых преобразованиях в этих шкалах. Но для данных, измеренных в порядковой шкале, его применять нельзя, так как его значение, как правило, меняется при допустимых преобразованиях в порядковой шкале. В таких случаях надо использовать непараметрические ранговые коэффициенты корреляции Спирмена и Кендалла, а также другие коэффициенты, разработанные в теории ранговых корреляций [4]. Можно сказать, что алгоритмы расчетов коэффициента Спирмена и коэффициента Кендалла любые входные данные переводят в порядковые данные (ранги, т.е. места в упорядоченном ряду), а затем дают возможность их исследовать.

В соответствии с теорией устойчивости экономико-математических методов и моделей [5] целесообразно провести расчеты для всех трех коэффициентов корреляции (Пирсона, Спирмена и Кендалла) и сопоставить результаты итоговых расчетов. Если они близки, то нет необходимости выбирать тот или иной из коэффициентов корреляции. Если различны, то следует опираться на коэффициенты ранговой корреляции Спирмена и Кендалла, а при необходимости выбора между ними целесообразно выбрать коэффициент корреляции Кендалла, поскольку он линейно связан с расстоянием Кемени [3].

Для расчета коэффициентов ранговой корреляции Спирмена и Кендалла необходимо предварительно проранжировать значения координат векторов, т.е. построить вариационные ряды для совокупностей $\{x_1, x_2, ..., x_n\}$ и $\{y_1, y_2, ..., y_n\}$ (по отдельности). Пусть p_j – ранг x_j в совокупности $\{x_1, x_2, ..., x_n\}$ и q_j – ранг y_j в совокупности $\{y_1, y_2, ..., y_n\}$, где $j = 1, 2, ..., n$.

Непараметрический выборочный коэффициент корреляции Спирмена для выборки $(x_i, y_i) = (x_i(\omega), y_i(\omega))$, $i = 1, 2, ..., n$, рассчитывается по формуле:

$$\rho_n = 1 - \frac{6\sum_{j=1}^{n}(p_j - q_j)^2}{n^3 - n} \quad (1)$$

Отметим, что ρ_n – это выборочный линейный парный коэффициент Пирсона, построенный по выборке пар рангов (p_i, q_i), $i = 1, 2, ..., n$.

Пример 1. Рассчитаем выборочный коэффициент корреляции Спирмена. Исходные данные – выборка объема $n = 10$, каждый элемент которой – двумерный вектор:

(2,35; 1), (4; 1,5), (3; 1,2), (1; 0), (2,25; 2), (11; 5), (15; 8), (17; 9), (19; 10), (22, 10).

Для расчета выборочного коэффициента корреляции Спирмена:

1. Выпишем значения координат элементов выборки сверху вниз в табл. 2 (в этой таблице i – номер элемента выборки, $i = 1, 2, ..., 10$; x_i – значения его первой координаты, y_i – значения его второй координаты).

2. Проранжируем значения координат. Пусть p_i – ранг x_i среди всех элементов столбца значений первой координаты (т.е. среди всех $x_1, x_2, ..., x_n$), а q_i – ранг y_i среди всех элементов столбца значений первой координаты (т.е. среди всех $y_1, y_2, ..., y_n$), $i = 1, 2, ..., 10$. Для одинаковых элементов в качестве их итоговых рангов (т.н. «связанных рангов») укажем среднее арифметическое положенных им рангов при нумерации – например, если их номера 7 и 8, то каждому присвоим «связанный ранг» 7,5. Отметим, что сумма рангов для каждого набора данных из n элементов должна быть равна сумме всех натуральных чисел от 1 до n, т.е. $n(n + 1)/2$, в рассматриваемом случае – 55. Далее расчет будет

вестись только по значениям рангов, значения исходных данных уже не будут использоваться.

3. Для каждой пары рангов (p_i, q_i), i = 1, 2, ..., n, найдем их разность $p_i - q_i$. (Сумма всех таких разностей рангов элементов с одинаковыми номерами будет всегда равна 0).

4. Каждую разность рангов элементов с одинаковым номером возведем в квадрат.

5. Найдем сумму квадратов разностей рангов $(p_i - q_i)^2$. Для рассматриваемых данных эта сумма равна 18,5 (сумма значений в самом правом столбце табл. 2).

6. Коэффициент ранговой корреляции Спирмена найдем по формуле (1). Для рассматриваемой выборки:

$$\rho_n = 1 - \frac{6 \cdot 18{,}5}{10^3 - 10} = 0{,}88.$$

Иными словами, при расчете коэффициента ранговой корреляции Спирмена берем исходные данные, для каждой координаты элементов выборки выполняем их ранжирование от меньшей величины к большей, тем самым присваивая этим величинам ранги от 1 до объема выборки (в случае одинаковых величин используем связанные ранги, т.е. берем среднее арифметическое для рангов, положенных этим величинам), заменяем все величины их рангами, а затем считаем для получившихся данных выборочный коэффициент корреляции Пирсона.

При расчете коэффициента ранговой корреляции Кендалла необходимо выполнить попарные сравнения – рассмотреть изменение рангов показателей (синонимы – параметров, значений координат) при переходе от одного элемента выборки к другому. Если наблюдается одновременное увеличение или уменьшение рангов по обоим параметрам сравнения, то такую ситуацию называют совпадением (изменения рангов по двум показателям сонаправлено). Если наблюдается увеличение по одному показателю и уменьшение по другому показателю, то такую ситуацию называют инверсией (изменения рангов по двум показателям разнонаправлено). Коэффициент ранговой корреляции Кендалла τ_n для выборки $(x_i, y_i) = (x_i(\omega), y_i(\omega))$, i = 1, 2, ..., n, рассчитывается по формуле:

$$\tau_n = \frac{2(P-Q)}{n(n-1)}, \quad (2)$$

где P – число совпадений; Q – число инверсий [12, 13].

Как и коэффициент Спирмена, коэффициент Кендалла основан на присвоении координатам элементов выборки порядковых номеров по возрастанию или убыванию (рангов) и дальнейших расчетах на их основе.

Ключевой элемент расчета коэффициента Кендалла – поиск пар рангов, имеющих разный порядок в каждом из наборов данных совокупностей $\{x_1, x_2, ..., x_n\}$ и $\{y_1, y_2, ..., y_n\}$. Для того, чтобы коэффициент Кендалла был близок к 1, большие и малые величины в одном наборе данных должны идти максимально в том же порядке, что и в другом (или в прямо противоположном – для обратной корреляции).

Пример 2. Рассмотрим исходные данные примера 1. Алгоритм расчета коэффициента Кендалла может быть следующим:

1. Как и для расчета коэффициента ранговой корреляции Спирмена, пронумеруем в порядке возрастания все элементы обеих наборов данных от 1 до объема выборки – «проставим ранги элементам». Для одинаковых элементов в качестве их итоговых рангов укажем среднее арифметическое положенных им рангов при нумерации (например, если их номера 7 и 8, то каждому элементу присвоим ранг 7,5), т.е. перейдем к связанным рангам. Сумма рангов для каждого набора данных должна быть равна сумме натуральных чисел от 1 до объема выборки. Далее расчет будет вестись только по значениям рангов, значения исходных данных уже не будут использоваться.

2. Упорядочим ранги элементов первого набора (т.е. ранги первой координаты элементов выборки) от меньшего к большему и перестроим ряд рангов набора значений второй координаты в соответствии с этим упорядочением (табл.3). Можно сказать, что строки табл. 2 переставлены в соответствии с возрастанием рангов элементов первой координаты.

3. Рассмотрим получившуюся последовательность рангов второй координаты (табл.3). Для каждого ранга в этой последовательности найдем, сколько величин, больших, чем этот ранг, встречается в последовательности рангов ниже в этом же столбце, и укажем эту величину в табл.4 как «количество совпадений». Также выясним, сколько величин, меньших, чем этот ранг, встречается в последовательности рангов ниже в этом же столбце, и укажем эту величину в табл. 3 как «количество инверсий».

4. Рассчитанные количества совпадений сложим между собой, как и количества инверсий.

Расчет непараметрического коэффициента ранговой корреляции Спирмена

Таблица 2

Номер элемента выборки i	Первая координата x_i	Ранг первой координаты	Вторая координата y_i	Ранг второй координаты	Разность рангов $p_i - q_i$	Квадрат разности рангов $(p_i - q_i)^2$
1	2,35	3	1	2	1	1
2	4	5	1,5	4	1	1
3	3	4	1,2	3	1	1
4	1	1	0	1	0	0
5	2,25	2	2	5	-3	9
6	11	7	5	6	1	1
7	15	8	8	7	1	1
8	10	6	9	8	-2	4
9	19	10	10	9,5	0,5	0,25
10	18	9	10	9,5	-0,5	0,25

Для рассматриваемых данных общее количество совпадений $P = 40$, а общее количество инверсий $Q = 5$.

5. Рассчитаем коэффициент ранговой корреляции Кендалла по формуле (2):

$$\tau_n = \frac{2(P-Q)}{n(n-1)} = \frac{2(40-5)}{10 \cdot 9} = \frac{70}{90} = 0{,}78.$$

Разработаны и другие коэффициенты ранговой корреляции [4].

Поскольку значения коэффициентов ранговой корреляции Спирмена и Кендалла не меняются при любых строго возрастающих преобразованиях шкал измерения исходных выборочных данных $(x_i, y_i) = (x_i(\omega), y_i(\omega))$, $i = 1, 2, ..., n$, т.е. при любых допустимых преобразованиях в порядковых шкалах, то эти коэффициенты используют при анализе данных, измеренных в порядковых шкалах [3].

Коэффициент корреляции Пирсона оценивает отклонение от линейности. Его модуль достигает максимума (равного 1) тогда и только тогда, когда переменные связаны линейной зависимостью. В то же время ранговые коэффициенты корреляции Спирмена и Кендалла оценивают отклонение от монотонности. Они достигают максимума тогда и только тогда, когда значения переменных (синонимы – параметров, значений координат, показателей) одинаково упорядочены. А минимума – когда упорядоченности противоположности, т.е. при перемене знака одной из переменных наблюдается одинаковая упорядоченность.

Для анализа данных, распределения которых не подчиняются нормальному распределению (т.е. для практически всех видов реальных данных [3]), рекомендуем применять коэффициенты ранговой корреляции Спирмена и Кендалла.

В базовой вероятностно-статистической модели совместная функция распределения двумерного случайного вектора $(X, Y) = (X(\omega), Y(\omega))$ предполагается непрерывной, а потому вероятность совпадений рассматриваемых результатов наблюдений равна 0. На практике встречаются случаи, когда некоторые результаты измерений совпадают, т.е. имеются связанные ранги. Модель анализа совпадений при расчете непараметрических ранговых статистик разработана в статье [6]. Применительно к коэффициентам Спирмена и Кендалла проблема учета связанных рангов рассмотрена в [7, разд. 6.10] и [8, с. 207-208].

О термине «корреляционно-регрессионный анализ»

Этот термин широко используется в публикациях. Стоящей за ним тематике посвящены разделы в учебниках, учебные пособия,методические указания к практическим занятиям и для выполнения расчетных заданий, научные публикации, справочные материалы. Поиск в Российском индексе научного цитирования (РИНЦ) по запросу «Корреляционно-регрессионный анализ» в названиях статей и книг за 2019–2024 гг. дал 82 названия. Среди них, например, работы по применению статистического пакета анализа для проведения корреляционно-регрессионного анализа в ходе экономических исследований и по изучению взаимосвязи социально-экономических показателей деятельности организации. Корреляционно-регрессионный анализ применялся при оценке

Расчет совпадений и инверсий при вычислении коэффициента корреляции Кендалла

Таблица 3

Ранг первой координаты	Ранг второй координаты	Совпадения	Инверсии
1	1	9	0
2	5	5	3
3	2	7	0
4	3	6	0
5	4	5	0
6	8	2	2
7	6	3	0
8	7	2	0
9	9.5	1	0
10	9.5	0	0

инвестиционной привлекательности нефтеперерабатывающей отраслии эффективности использования инвестиционных ресурсов нефтяной компании. Он используется как инструментарий оценки инновационной деятельности в регионах, при анализе зависимости выручки предприятия от факторов внешнеэкономической деятельности. С его помощью изучают влияние показателей на выручку предприятия, развития малого бизнеса на уровень жизни населения, инфляции на уровень заработной платы в Российской Федерации. Он используется как инструмент прогнозирования влияния функционирования социально-экономического кластера в сфере ЖКХ на экономику региона. Он оказался полезен для анализа зависимости выручки крупнейших ТНК стран БРИКС в нефтегазовой отрасли от факторов экономической деятельности. На основе данных РИНЦ можно установить, что корреляционно-регрессионный анализ применяется и во многих других областях науки, отраслях народного хозяйства.

Однако необходимо констатировать, что термин «корреляционно-регрессионный анализ» с точки зрения современной прикладной статистики [3] является некорректным. В нем необоснованно механически объединены совершенно разные разделы прикладной статистики – корреляционный анализ и регрессионный анализ.

Коэффициенты корреляции, рассмотренные выше, предназначены для количественной оценки степени связи между двумя случайными переменными, т.е. между координатами двумерного случайного вектора. Цель регрессионного анализа – восстановление зависимости между переменными, по крайней мере одна из которых является случайной. Вторая может быть детерминированной. Например, при изучении динамики показателей финансово-хозяйственной деятельности предприятия. Многообразие моделей регрессионного анализа рассмотрено в статье [9]. Одна из таких моделей предназначена для анализа выборки из распределения двумерного случайного вектора. Эта модель порождения статистических данных является также исходной моделью корреляционного анализа. Она рассмотрена в начале настоящей статьи. Другие модели регрессионного анализа – принципиально иные [9].

Суть в том, что корреляционный анализ позволяет оценивать степень связи, прогнозировать значение одной переменной по значению другой, но не позволяет управлять – изменяя значение одной переменной, целенаправленно менять значение другой.

Этот факт давно известен. Его называют парадоксом корреляции [10]. В качестве примера обсудим анализ данных о росте и весе некоторого количества людей. Можно принять вероятностно-статистическую модель корреляционного анализа, согласно которой указанные данные рассматриваются как независимые одинаково распределенные двумерные случайные вектора. Как показывает практический опыт, линейный парный коэффициент корреляции Пирсона положителен и заметно отличается от 0. С помощью метода наименьших квадратов можно получить линейную зависимость роста от веса, которая позволяет прогнозировать рост по весу человека (с определенной точностью, которая выражается с помощью доверительных границ). Однако очевидно, что эту зависимость

нельзя использовать для управления – изменение (например, уменьшение) веса взрослого человека не приводит к изменению его роста.

Можно обсудить и другие примеры. Так, для города рассмотрим такие показатели, как число телевизоров на определенный год, число убийств в городе, число заболеваний и смертность в том же году. Можно убедиться, что для каждых двух из этих четырех показателей коэффициент корреляции между ними весьма близок к 1. Как следствие, по любому из этих показателей можно достаточно точно спрогнозировать значение любого другого. При этом ясно, что, например, полная ликвидация телевизоров не позволит существенно сократить значения трех других показателей.

Наиболее частой причиной ситуации наличия корреляции между двумя величинами при отсутствии прямой причинно-следственной связи между ними является наличие некоего иного, третьего фактора, от которого и зависят обе эти величины. Например, в рассматриваемом выше случае таким фактором будет численность населения города – при росте этой величины наверняка будет расти и количество телевизоров в городе, и количество убийств, и количество заболеваний просто потому, что покупку телевизоров и убийства совершают люди, они же страдают болезнями, и чем больше людей в городе, тем больше и случаев убийств и заболеваний. Можно сказать, что в ситуации имеется скрытая (латентная) переменная – число жителей в городе, и от этой переменной перечисленные показатели при резком росте численности населения города зависят почти линейно, что и приводит к тому, что коэффициенты корреляции между ними близки к 1.

Реальная связь величин, между которыми наблюдается высокий коэффициент корреляции, далеко не всегда очевидна. Так, по итогам анализа хроник некоторых городов Римской империи, выяснялось, что при почти постоянной численности населения города со временем начинала наблюдаться прямая корреляция между количеством фонтанов (в то время они служили оконечной точкой акведуков, по которым в город подавалась чистая вода) в районах города и смертностью в этих же районах от заболеваний, описания которых соответствовали описаниям инфарктов миокарда и инсультов. С первого взгляда может показаться, что величины количества фонтанов и количества инфарктов связаны тем, что, дескать, вода из фонтанов вызывает инфаркты, но для медицинской науки это нонсенс – никаких механизмов подобных связей выявлено никогда не было. В реальности снабжение районов городов чистой водой из акведуков приводило к резкому снижению смертности населения этих районов от кишечных инфекций и отравлений (альтернативой фонтанам являлись реки в черте города, в которые сливались отходы, а также колодцы, заполненные загрязненными грунтовыми водами), и количество людей, которые в итоге доживали до тех возрастов, в которых смертность от инфарктов миокарда и инсультов головного мозга становилась значимой. Таким образом, как ни парадоксально, рост смертности от таких заболеваний как раз указывал на улучшение здоровья населения городов Римской империи от строительства акведуков и фонтанов (хотя, бесспорно, еще лучше это покажет имевшее место снижение смертности от кишечных инфекций).

Понимание всей цепочки глубинных связей коррелирующих переменных тем не менее иногда дает возможность использовать сведения о корреляции для управления ситуацией. Так, Чарльз Дарвин в своем труде «Происхождение видов путем естественного отбора» отмечал прямую корреляцию между количеством кошек на территориях сельской местности Великобритании и урожайностью красного клевера там же. Можно было бы подумать, что эти два показателя просто одновременно зависят от некоей третьей переменной (например, климатических условий местности), но исследователь выяснил, что цепочка связей на самом деле более сложная: красный клевер подвергался опылению только шмелями (но не пчелами), шмели (в отличие от пчел) жили в гнездах, доступных мышам, которые эти гнезда разоряли, а кошки, в свою очередь, охотились на мышей. Соответственно, чем больше кошек жило на какой-либо территории, тем меньше там оставалось мышей, тем меньше была угроза шмелиным гнездам, тем больше шмелей участвовало в опылении клевера, что и давало повышение его урожайности. Поэтому совет Чарльза Дарвина британским фермерам «Заботиться о кошках и разводить их» был вполне оправданным ответом на просьбу фермеров дать им научную идею повышения урожайности клевера (основной кормовой культуры того времени в Великобритании).

Качество регрессионной модели измеряется коэффициентом детерминации. Если справедлива вероятностно-статистическая модель корреляци-

Степени корреляции

Таблица 4

Степень корреляции	Прямая корреляция	Обратная корреляция
Отсутствует	0	0
Слабая	(0; 0,3)	(0; –0,3)
Умеренная	[0,3; 0,5)	[–0,3; –0,5)
Значительная	[0,5; 0,7)	[–0,5; –0,7)
Сильно выраженная	[0,7; 0,9)	[–0,7; –0,9)
Очень сильная	[0,9; 1)	[–0,9; –1)
Функциональная	1	–1

онного анализа, то для модели парной линейной регрессии коэффициент детерминации равен квадрату обычного коэффициента корреляции между независимой и зависимой переменными. Однако коэффициент детерминации в МНК может быть использован более широко (например, когда одна из переменных детерминирована), чем коэффициент корреляции. Поэтому некорректно говорить, что коэффициент детерминации равен квадрату коэффициента корреляции, хотя расчетные формулы совпадают. Многие ошибки при использовании коэффициентов корреляции и детерминации при анализе конкретных практических данных связаны с неправомерным переносом свойств модели корреляционного анализа на другие модели регрессионного анализа [11].

Для того, чтобы с помощью регрессионной зависимости разрабатывать управленческие решения, необходима серия предварительных экспериментов. В ней исследователь задает (по определенным правилам) значения независимой переменной и измеряет соответствующие значения зависимой. Эта область прикладной статистики называется «планирование эксперимента».

Шкала Чеддока

Она используется для интерпретации полученных результатов расчета коэффициентов корреляции в словесной форме. Другими словами, для удобства представления значений коэффициентов корреляции применяют переход к лингвистической переменной.

При оценке корреляции выделяют ее степени, например, согласно табл. 4 [12].

Впервые подобную лингвистическую шкалу для степеней корреляции предложил американский социолог и статистик Роберт Эммет Чеддок (1879-1940) в 1925 г. [13].В литературных и интернет-источниках встречаются варианты шкалы табл.4, но различия незначительны. Например, для прямой корреляции используют следующие значения и термины: слабая (или очень слабая) связь – от 0,1 до 0,3 (или от 0 до 0,3); умеренная (или слабая) связь – от 0,3 до 0,5; заметная (или средняя) связь – от 0,5 до 0,7; высокая (или сильно выраженная) связь – от 0,7 до 0,9; очень высокая (весьма высокая, сильная) связь – от 0,9 до 1,0 (или от 0,9 до 0,99). Следовательно, при использовании шкалы Чеддока целесообразно указывать, какой именно из вариантов этой шкалы имеется в виду. Разработано много шкал подобного типа с различными названиями (см., например, [14]), обычно в честь тех исследователей, кто их предложил.

Пример 3. Для данных примера 1 коэффициент корреляции Спирмена равен 0,88, а коэффициент корреляции Кендалла равен 0,78. Согласно шкале Чеддока (табл.4) эти коэффициенты корреляции описываются как «сильно выраженные». Таким образом, зависимость между переменными близка к монотонной. Можно сказать, что изменения величин одной переменной следуют за изменениями второй переменной.

Статистическая значимость коэффициентов корреляции

Использование шкалы Чеддока может ввести в заблуждение. Дело в том, что в случае, когда теоретический коэффициент равен 0 (например, когда координаты случайного вектора независимы), его выборочное значение в силу чисто случайных причин может достаточно далеко отстоять от 0 и по шкале Чеддока корреляция окажется, например, значительной.

Рассмотрим статистические гипотезы:

H_0 (нулевая гипотеза): теоретический коэффициент корреляции равен 0;

H_{11} (двусторонняя альтернативная гипотеза): теоретический коэффициент корреляции не равен 0;

H_{12} (односторонняя альтернативная гипотеза): известно, в какую сторону теоретический коэффициент корреляции отклоняется от 0, но неизвестно, на сколько. Пусть для определенности теоретический коэффициент корреляции положителен.

Дальнейшие рассуждения проводятся однотипно для всех трех рассматриваемых коэффициентов корреляции – Пирсона, Спирмена и Кендалла. Обозначим как qn любой из этих коэффициентов. Приведенные ниже рассуждения справедливы для всех трех случаев: $q_n = r_n$ (линейный парный коэффициент корреляции Пирсона), $q_n = \rho_n$ (непараметрический ранговый коэффициент корреляции Спирмена) и, наконец, $q_n = \tau_{1n}$ (нормированный непараметрический ранговый коэффициент корреляции Кендалла), т.е.:

$$\tau_{1n} = \frac{\tau_n}{\sqrt{D(\tau_n)}} = \tau_n\sqrt{\frac{9n(n-1)}{2(2n+5)}} = \frac{P-Q}{\sqrt{\frac{n(n-1)(2n+5)}{18}}}.$$

Теорема. При справедливости нулевой гипотезы (H0) о равенстве 0 теоретического коэффициента корреляции распределение соответствующего выборочного коэффициента корреляции является асимптотически нормальным с математическим ожиданием 0 и дисперсией $1/n$, т.е. при всех x справедливо предельное соотношение:

$$\lim_{n\to+\infty} P(\sqrt{n}\,q_n < x) = \Phi(x),$$

где $\Phi(x)$ – функция стандартного нормального распределения с математическим ожиданием 0 и дисперсией 1.

Доказательство проводится с помощью предельных теорем прикладной математической статистики [3, гл.4].

Эта теорема позволяет построить (асимптотические) решающие правила. Как обычно при проверке статистических гипотез, критические значения в решающих правилах определяются уровнем значимости α, который задает исследователь.

Для двусторонней альтернативной гипотезы решающее правило таково. Если $\sqrt{n}\,|q_n| < C(\alpha)$, то принимается нулевая гипотеза (на уровне значимости α), т.е. нет оснований утверждать, что соответствующий коэффициент корреляции отличается от 0. Если же $\sqrt{n}\,|q_n| > C(\alpha)$, то принимается альтернативная гипотеза H_{11}, т.е. коэффициент корреляции значимо отличается от 0. Здесь критическое значение $C(\alpha) = \Phi^{-1}(1 - \alpha/2)$ Для наиболее часто используемого уровня значимости $\alpha = 0{,}05$ критическое значение равно 1,96.

Для односторонней альтернативной гипотезы решающее правило таково. Если $\sqrt{n}\,q_n < D(\alpha)$, то принимается нулевая гипотеза (на уровне значимости α). Если же $\sqrt{n}\,q_n > D(\alpha)$, то принимается альтернативная гипотеза, т.е. коэффициент корреляции значим и положителен. Здесь критическое значение $D(\alpha) = \Phi^{-1}(1 - \alpha)$. Для наиболее часто используемого уровня значимости $\alpha = 0{,}05$ критическое значение равно 1,64.

Таким образом, результат проверки статистической гипотезы зависит от величины $\sqrt{n}qn$. Пусть рассчитанное по выборке значение коэффициента корреляции равно 0,7. По шкале Чеддока (табл. 4) такая степень корреляции является «сильно выраженной». Однако такой выборочный коэффициент корреляции значимо отличается от 0 лишь в случае $0{,}7\sqrt{n} \geqslant 1{,}96$, т.е. при $n \geqslant 8$.

Аналогично, умеренная корреляция может значимо отличаться от 0 лишь в случае $0{,}3\sqrt{n} \geqslant 1{,}96$, $n \geqslant 43$. Если же $n = 4$, то любое выборочное значение коэффициента корреляции не будет значимо отличаться от 0. Из сказанного ясно, что оценка степени корреляции по шкале Чеддока вплоть до объемов выборки в несколько десятков единиц должна обязательно рассматриваться вместе с результатами проверки на значимость.

Пример 4. Для данных примера 1 объем выборки $n = 10$, коэффициент корреляции Спирмена равен 0,88, а коэффициент корреляции Кендалла равен 0,78. Статистика критерия проверки гипотезы о равенстве 0 коэффициента Спирмена равна $0{,}88\sqrt{10} = 2{,}78$. Поскольку $2{,}78 > 1{,}96$, то принимается альтернативная гипотеза, т.е. теоретический коэффициент корреляции не равен 0. Следовательно, корреляция достоверно имеет место, переменные зависимы. Нормированный непараметрический ранговый коэффициент корреляции Кендалла равен:

$$\tau_{1n} = \tau_n\sqrt{\frac{9n(n-1)}{2(2n+5)}} = 0{,}78\sqrt{\frac{9 \cdot 10 \cdot (10-1)}{2(2 \cdot 10+5)}} =$$

$$= 0{,}78\sqrt{\frac{810}{50}} = 0{,}78 \cdot 4{,}025 = 3{,}14.$$

Поскольку $3{,}14 > 1{,}96$, то нулевая гипотеза отвергается, коэффициент корреляции Кендалла статистически значим, переменные зависимы.

Заключение

В статье проведен анализ связанных с анализом корреляции основных вопросов, возникаю-

щих при статистической обработке реальных данных. Авторами были рассмотрены некоторые направления дальнейших исследований. Полученные рекомендации являются асимптотическими, а их погрешности требуют изучения. Для этого может быть применен метод статистических испытаний.

Важно изучить влияние свойств выборок на значения коэффициентов корреляции. Академик АН СССР С.Н. Бернштейн заканчивает статью [15] так: «... достаточно, чтобы только один из 701 индивида не подчинился господствующему закону пропорциональности Y = 0,1X, чтобы коэффициент корреляции понизился до значения 0,198».

Литература:

1. Шамсувалеева А.М., Орлов А.И. Использование коэффициентов корреляции и конкордации // Тринадцатые Чарновские чтения. Сборник трудов XIII Всероссийской научной конференции по организации производства. – М.: МГТУ им. Н.Э. Баумана, НП «Объединение контроллеров», 2023. С.171-180.

2. Орлов А.И. Контроллинг статистических методов // Журнал «Контроллинг». 2022. №4 (86). С. 2-11.

3. Орлов А.И. Прикладной статистический анализ. – М.: Ай Пи Ар Медиа, 2022. – 812 с.

4. Кендэл М. Ранговые корреляции. – М.: Статистика, 1975.– 216 с.

5. Орлов А.И. Устойчивые экономико-математические методы и модели. – М.: Ай Пи Ар Медиа, 2022. – 337 с.

6. Орлов А.И. Модель анализа совпадений при расчете непараметрических ранговых статистик // Заводская лаборатория. Диагностика материалов. 2017. Т.83. №11. С. 66-72.

7. Большев Л.Н., Смирнов Н.В. Таблицы математической статистики – М.: Наука, 1983. – 416 с.

8.Холлендер М., Вулф Д. Непараметрические методы статистики. – М.: Финансы и статистика, 1983. – 518 с.

9. Орлов А.И. Многообразие моделей регрессионного анализа (обобщающая статья) // Заводская лаборатория. Диагностика материалов. 2018. Т.84. №5. С. 63-73.

10. Секей Г. Парадоксы в теории вероятностей и математической статистике. – М.: Мир, 1990. – 240 с.

11. Орлов А.И. Ошибки при использовании коэффициентов корреляции и детерминации // Заводская лаборатория. Диагностика материалов. 2018. Т.84. № 3. С. 68-72.

12. Шамсувалеева А.М., Прохоров С.Ю., Орлов А.И., Пивкин А.Л., Леус Н.А. Формирование интегрального показателя – индекса готовности стран к космической деятельности // Экономика космоса. 2024. № 1(7). С. 28-42.

13. Chaddock R.E. Principles and methods of statistics. – Boston, New York [etc.]: Houghton Mifflin. 1925. – 471 p.

14. Котеров А.Н. и др. Сила связи. Сообщение 2. Градации величины корреляции // Медицинская радиология и радиационная безопасность. 2019. Т. 64. № 6. С. 12-24.

15. Бернштейн С.Н. Об одном элементарном свойстве коэффициента корреляции // Записки Харьковского математического товарищества. 1932. Т. 5. С. 65-66.

References:

1. Shamsuvaleeva A.M., Orlov A.I. Ispol'zovanie kojefficientov korreljacii i konkordacii // Trinadcatye Charnovskie chtenija. Sbornik trudov XIII Vserossijskoj nauchnoj konferencii po organizacii proizvodstva. – M.: MGTU im. N.Je. Baumana, NP «Ob#edinenie kontrollerov», 2023. S.171-180.

2. Orlov A.I. Kontrolling statisticheskih metodov // Zhurnal «Kontrolling». 2022. № 4 (86). S. 2-11.

3. Orlov A.I. Prikladnoj statisticheskij analiz. – M.: Aj Pi Ar Media, 2022. – 812 s.

4. Kendjel M. Rangovye korreljacii. – M.: Statistika, 1975. – 216 s.

5. Orlov A.I. Ustojchivye jekonomiko-matematicheskie metody i modeli. – M.: Aj Pi Ar Media, 2022. – 337 s.

6. Orlov A.I. Model' analiza sovpadenij pri raschete neparametricheskih rangovyh statistik // Zavodskaja laboratorija. Diagnostika materialov. 2017. T. 83. №11. S. 66-72.

7. Bol'shev L.N., Smirnov N.V. Tablicy matematicheskoj statistiki – M.: Nauka, 1983. – 416 s.

8. Hollender M., Vulf D. Neparametricheskie metody statistiki. – M.: Finansy i statistika, 1983. – 518 s.

9. Orlov A.I. Mnogoobrazie modelej regressionnogo analiza (obobshhajushhaja stat'ja) // Zavodskaja laboratorija. Diagnostika materialov. 2018. T. 84. №5. S. 63-73.

10. Sekej G. Paradoksy v teorii verojatnostej i matematicheskoj statistike. – M.: Mir, 1990. – 240 s.

11. Orlov A.I. Oshibki pri ispol'zovanii kojefficientov korreljacii i determinacii // Zavodskaja laboratorija. Diagnostika materialov. 2018. T.84. № 3. S. 68-72.

12. Shamsuvaleeva A.M., Prohorov S.Ju., Orlov A.I., Pivkin A.L., Leus N.A. Formirovanie integral'nogo pokazatelja – indeksa gotovnosti stran k kosmicheskoj dejatel'nosti // Jekonomika kosmosa. 2024. № 1 (7). S. 28-42.

13. Chaddock R.E. Principles and methods of statistics. – Boston, New York [etc.]: Houghton Mifflin. 1925. – 471 p.

14. Koterov A.N. i dr. Sila svjazi. Soobshhenie 2. Gradacii velichiny korreljacii // Medicinskaja radiologija i radiacionnaja bezopasnost'. 2019. T. 64. № 6. S. 12-24.

15. Bernshtejn S.N. Ob odnom jelementarnom svojstve kojefficienta korreljacii // Zapiski Har'kovskogo matematicheskogo tovarishchestva. 1932. T. 5. S. 65-66.