Математические методы исследования
Mathematical methods of investigation

DOI: 10.26896/1028-6861-2018-84-3-68-72

ОШИБКИ ПРИ ИСПОЛЬЗОВАНИИ КОЭФФИЦИЕНТОВ КОРРЕЛЯЦИИ И ДЕТЕРМИНАЦИИ

© **Александр Иванович Орлов**

Институт высоких статистических технологий и эконометрики Московского государственного технического университета им. Н. Э. Баумана, Москва, Россия; e-mail: prof-orlov@mail.ru

Статья поступила 10 октября 2017 г.

Коэффициенты корреляции и детерминации широко используют при статистическом анализе данных. При этом достаточно часто допускают те или иные ошибки. Некоторые из них рассмотрены в данной статье. Ограничимся случаем двух переменных. Наиболее часто используют линейный парный коэффициент корреляции Пирсона и непараметрические ранговые коэффициенты Спирмена и Кендалла. Согласно теории измерений коэффициент корреляции Пирсона можно применять к переменным, измеренным в шкале интервалов (и в шкалах с более узкой группой допустимых преобразований, например, в шкале отношений). Его нельзя применять при анализе порядковых данных. Непараметрические ранговые коэффициенты Спирмена и Кендалла предназначены для оценки связи порядковых переменных. Их можно использовать и в шкалах с более узкой группой допустимых преобразований, например, в шкалах интервалов или отношений. Критическое значение при проверке значимости отличия коэффициента корреляции от нуля зависит от объема выборки и приближается к нулю при его росте. Поэтому использование «шкалы Чеддока» некорректно. При применении пассивного эксперимента коэффициенты корреляции можно обоснованно использовать лишь для прогнозирования, но не для управления. Для получения предназначенных для управления вероятностно-статистических моделей необходим активный эксперимент. Как показал С. Н. Бернштейн, влияние выбросов на коэффициент корреляции Пирсона весьма велико. Эффект «вздувания» коэффициента корреляции состоит в том, что при увеличении числа проанализированных наборов предикторов заметно растет максимальный из соответствующих коэффициентов корреляции — показателей качества приближения. Распространенная ошибка состоит в использовании коэффициента детерминации для оценки качества восстановления зависимости методом наименьших квадратов.

Ключевые слова: математическая статистика; коэффициент корреляции Пирсона; непараметрические ранговые коэффициенты корреляции; выбросы; коэффициент детерминации; распространенные ошибочные выводы.

ERRORS IN THE USE OF CORRELATION AND DETERMINATION COEFFICIENTS

© **Alexander I. Orlov**

Institute of high statistical technologies and econometrics, N. É. Bauman Moscow State Technical University, Moscow, Russia; e-mail: prof-orlov@mail.ru

Submitted October 10, 2017.

Coefficients of correlation and determination are widely used in statistical analysis of data. Some of the errors attributed to their use are considered in this article. We confine ourselves to the case of two variables. The linear Pearson correlation coefficient and nonparametric rank coefficients of Spearman and Kendall are used most commonly. According to the theory of measurements, the Pearson correlation coefficient can be applied to variables measured in the interval scale (and in scales with a narrower group of permissible transformations, for example, in the ratio scale) but it cannot be used in analysis of ordinal data. Spearman and Kendall's nonparametric rank coefficients are designed to evaluate the relationship of ordinal variables. They can also be used in scales with a narrower group of permissible transformations, for example, in the scales of intervals or ratios. The critical value in testing the significance of the difference in the correlation coefficient from zero depends on the sample size and approaches zero as the sample size grows. Therefore, the use of the "Cheddock scale" is incorrect. When

using a passive experiment, the correlation coefficients can be reasonably used only for forecasting, but not for control. To obtain the statistical models valid for control, an active experiment is required. S. N. Bernshtein has shown that the effect of outliers on the Pearson correlation coefficient is very large. The effect of "inflation" of the correlation coefficient is that with increasing number of analyzed sets of predictors, the maximum of the corresponding correlation coefficients, the quality of approximation, increases noticeably. A common mistake is to use the determination coefficient to estimate the quality of the least-squares recovery.

Keywords: mathematical statistics; Pearson correlation coefficient; nonparametric rank correlation coefficients; outliers; determination coefficient; common erroneous conclusions.

Коэффициенты корреляции и детерминации широко используют при статистическом анализе данных. При этом достаточно часто допускают те или иные ошибки. Рассмотрим некоторые из них.

Ограничимся случаем двух переменных. Пусть (X, Y) — двумерный случайный вектор. Наиболее часто используют линейный парный коэффициент корреляции Пирсона и непараметрические ранговые коэффициенты Спирмена и Кендалла.

Согласно теории измерений [1] коэффициент корреляции Пирсона можно применять к переменным, измеренным в шкале интервалов (и в шкалах с более узкой группой допустимых преобразований, например, в шкале отношений). Его нельзя применять при анализе порядковых данных (например, для анализа связи успеваемости по двум учебным предметам). Непараметрические ранговые коэффициенты Спирмена и Кендалла предназначены для оценки связи порядковых переменных. Их можно использовать и в шкалах с более узкой группой допустимых преобразований, например, в шкалах интервалов или отношений.

Исходя из теории устойчивости [2], одни и те же данные целесообразно обработать разными способами и сравнить результаты. В частности, целесообразно рассчитать все упомянутые выше коэффициенты корреляции.

Если X и Y — независимые случайные величины, то коэффициенты корреляции равны нулю. Обратное неверно — из равенства нулю коэффициента корреляции не следует, что случайные величины X и Y независимы.

Значимость отличия от нуля и «шкала Чеддока»

Выборочные коэффициенты корреляции отличаются от теоретических. Их распределения являются асимптотически нормальными.

Часто проверяют нулевую гипотезу о том, что тот или иной теоретический коэффициент корреляции равен нулю. Если эта гипотеза отклоняется, то можно утверждать, что случайные величины X и Y зависимы. Гипотеза отклоняется на уровне значимости α, если выборочный коэффициент корреляции по абсолютной величине больше граничного значения $C(\alpha)f(n)$, где n — объем выборки, C и f — некоторые функции, причем

$$\lim_{n\to\infty} f(n) = 0.$$

Для коэффициента корреляции Пирсона функция f зависит от распределения случайного вектора (X, Y). Распространенные таблицы рассчитаны для случая двумерного нормального распределения (X, Y). Хорошо известно, что распределения подавляющего большинства реальных данных не являются нормальными. Следовательно, применение правил, сформированных для двумерного нормального распределения, как правило, не является обоснованным.

Для непараметрических коэффициентов ранговой корреляции Спирмена и Кендалла свойства правил проверки гипотезы о том, что теоретический коэффициент корреляции равен нулю, не зависят от распределения данных.

Иногда показателям тесноты связи (модулям коэффициентов корреляции) пытаются дать качественную оценку по так называемой шкале Чеддока (см. таблицу). Такая рекомендация не вполне адекватна. При малых объемах выборки значение коэффициента корреляции 0,5 или 0,7 вполне совместимо со справедливостью гипотезы о том, что теоретический коэффициент корреляции равен нулю. А при достаточно большом объеме выборки коэффициент 0,1 может свидетельствовать о необходимости отклонения такой гипотезы.

Шкала Чеддока

Количественная мера тесноты связи	Качественная характеристика силы связи
0,1 – 0,3	Слабая
0,3 – 0,5	Умеренная
0,5 – 0,7	Заметная
0,7 – 0,9	Высокая
0,9 – 0,99	Весьма высокая

Активный и пассивный эксперименты

Вопреки часто встречающимся мнениям и предложениям коэффициенты корреляции можно обоснованно использовать лишь для прогнозирования, но не для управления.

Рассмотрим упрощенный пример. Пусть X — число телевизоров в городе, Y — число преступлений в этом городе, Z — число психических заболеваний в нем. Были собраны данные по нескольким сотням городов (англосаксонских стран). Выборочный коэффициент корреляции между X и Y оказался равным практически единице. Весьма мало отличался от единицы и выборочный коэффициент корреляции между X и Z. С высокой степенью точности справедливы зависимости $Y = aX$ и $Z = bX$. С помощью этих зависимостей можно надежно прогнозировать число преступлений и число психических заболеваний по числу телевизоров в городе.

В подобных ситуациях часто возникает желание использовать зависимости $Y = aX$ и $Z = bX$ для управления. Однако очевидно, что прекращение телевещания (переход к $X = 0$) не приведет к резкому снижению числа преступлений и числа психических заболеваний. В чем причина неудачи, казалось бы, естественного подхода к управлению? Дело в том, что все три рассматриваемые переменные определяются значениями четвертой переменной (латентной, скрытой) — числом жителей города W. А именно, с высокой точностью $X = cW$, $Y = dW$, $Z = eW$, откуда $Y = (d/c)X$, $Z = (e/c)X$.

Проблема в том, что при анализе реальных данных не всегда ясно наличие или отсутствие латентных переменных, определяющих успех управления по регрессионным зависимостям. Полезны понятия «пассивный эксперимент» и «активный эксперимент». При пассивном эксперименте данные накапливаются путем пассивного наблюдения, другими словами, информацию получают в условиях обычного функционирования изучаемых объектов. Активный эксперимент проводят с применением искусственного воздействия на изучаемые объекты по специальной программе.

При пассивном эксперименте существуют только факторы в виде входных контролируемых, но неуправляемых переменных, и экспериментатор находится в положении пассивного наблюдателя. Задача планирования в этом случае сводится к оптимальной организации сбора информации и решению таких вопросов, как выбор количества и частоты измерений, выбор метода обработки результатов измерений.

Наиболее часто целью пассивного эксперимента является построение математической модели объекта. Хорошим примером пассивного эксперимента являются измерения метеорологических параметров (температуры, скорости ветра и т.д.).

Активный эксперимент основан на задании экспериментатором значений факторов. Такой эксперимент позволяет быстрее и эффективнее решать задачи исследования, но более сложен, требует больших материальных затрат и может помешать нормальному ходу технологического процесса. Иногда отсутствует возможность проведения активного эксперимента (например, при исследовании явлений природы). Однако учитывая преимущества активного эксперимента, когда это возможно, предпочтение отдают ему. Теория планирования экспериментов [3, 4] посвящена прежде всего активным экспериментам.

Влияние выбросов на коэффициент корреляции

Еще в 1932 г. С. Н. Бернштейн рассмотрел [5] следующую проблему: «Определить наименьшее возможное значение коэффициента корреляции Пирсона R между величинами X и Y, если известно, что математические ожидания их равны 0 и что существуют две константы L и λ такие, что всегда

$$0 \le \lambda \le Y/X \le L.»$$

Пусть $\sigma^2 = M(X^2)$, $\sigma_1^2 = M(Y^2)$, $\sigma_1/\sigma = u$. В [5] показано, что минимум коэффициента корреляции R достигается при $u = \sqrt{L\lambda}$ и составляет

$$\frac{2\sqrt{L\lambda}}{L+\lambda}.$$

Для достижения минимума необходимо и достаточно, чтобы постоянно выполнялось одно из равенств —

$$Y - Lx = 0 \text{ или } Y - \lambda X = 0.$$

Таким образом, минимум R достигается, когда Y есть функция X, которую можно даже предполагать монотонной, если имеем, например,

$$Y = \begin{cases} \lambda X, |X| < 1, \\ LX, |X| \ge 1. \end{cases}$$

Рассмотрев численный пример, С. Н. Бернштейн заканчивает статью [5] так: «... достаточно, чтобы только один из 701 индивида не подчинился господствующему закону пропорциональности $Y = 0{,}1X$, чтобы коэффициент корреляции понизился до значения 0,198».

Таким образом, влияние выбросов на коэффициент корреляции может быть весьма велико. Следовательно, перед расчетом коэффициента корреляции необходимо исключить выбросы из выборки. Хорошо известно [1], что обоснованное исключение выбросов может быть проведено только на основе соображений предметной области, поскольку математико-статистические алгоритмы являются крайне неустойчивыми по отношению к отклонениям от функции распределения, принятой в вероятностно-статистической модели.

Вздувание коэффициентов корреляции

Это явление обнаружил А. Н. Колмогоров в работе 1933 г. «К вопросу о пригодности найденных статистическим путем формул прогноза» [6]. Предположим, что имеется много наборов предикторов (факторов, признаков). Для каждого из них строится наилучшее приближение отклика с помощью линейной функции от предикторов. Показателем качества приближения служит коэффициент корреляции между откликом и наилучшей линейной функцией от предикторов (в настоящее время чаще используют его квадрат, называемый коэффициентом детерминации). Эффект «вздувания» коэффициента корреляции состоит в том, что при увеличении числа проанализированных наборов предикторов заметно растет максимальный из соответствующих коэффициентов корреляции — показателей качества приближения. Создается впечатление, что тот набор предикторов, на котором достигается рассматриваемый максимум, дает хорошее приближение для отклика. Однако это впечатление развеивается при попытке использовать соответствующую зависимость для прогноза — по новым данным коэффициент корреляции между откликом и ранее найденной линейной функцией от предикторов оказывается значительно меньшим.

В настоящее время весьма популярны методы поиска «наиболее информативного множества признаков» в регрессионном и дискриминантном анализе. Соответствующие алгоритмы, как правило, основаны на переборе большого числа наборов признаков. Поэтому, как показано в [7], актуальность работы А. Н. Колмогорова [6] в настоящее время существенно повысилась. Эффект «вздувания» коэффициента корреляции является одним из проявлений неклассического поведения статистических характеристик в ситуации, когда одна и та же статистическая процедура осуществляется многократно, например, при множественных проверках статистических гипотез [8].

В течение полувека А. Н. Колмогоров интересовался статистическими постановками, в которых число неизвестных параметров растет вместе с объемом данных. К ним относится и кратко рассмотренная выше работа [6]. А в 1970-х годах он стимулировал исследования по так называемой «асимптотике Колмогорова» $p \to \infty$, $n \to \infty$, $p/n \to \lambda$, где p — число параметров, n — объем выборки. Эта асимптотика весьма актуальна как для многомерного статистического анализа [9], так и для статистики объектов нечисловой природы [10], а также для задач статистического приемочного контроля [11].

Коэффициент детерминации

Как уже отмечалось, для модели линейной регрессии с одним признаком (фактором) X коэффициент детерминации равен квадрату линейного парного коэффициента корреляции Пирсона между X и откликом Y. Необходимо подчеркнуть, что такая интерпретация корректна только тогда, когда анализируемые данные являются выборкой из двумерного распределения. Чуть подробнее: исходные данные рассматриваются как независимые одинаково распределенные случайные векторы. Отсюда следует, что если фактор X детерминирован (например, время), то коэффициент детерминации не является квадратом коэффициента корреляции, поскольку понятие коэффициента корреляции для подобной постановки не определено. Следовательно, коэффициент детерминации не является показателем качества зависимости, построенной с помощью метода наименьших квадратов.

Распространенная ошибка состоит в использовании коэффициента детерминации для оценки качества восстановления зависимости методом наименьших квадратов. Часто заявляют, что близость к единице коэффициента детерминации свидетельствует об успешном восстановлении зависимости. При этом взгляд на данные (на корреляционное поле) может дать совершенно иной вывод. Например, все точки, кроме одной, лежат в небольшой по диаметру области и вытянуты вдоль гиперболы. Оставшаяся точка расположена далеко вправо вверху. Формальное применение метода наименьших квадратов приводит к тому, что единственный «выброс» меняет гиперболу на возрастающую линейную зависимость (сопоставьте с примером С. Н. Бернштейна, рассмотренным выше).

Формально рассчитанный коэффициент детерминации в рассматриваемой постановке может быть сколь угодно близким к единице. Однако использование этого факта для обоснования утверждения о высоком качестве восстановления зависимости, скорее всего, является примером неверной интерпретации, во-первых,

из-за неисключенных выбросов, во-вторых, из-за нарушения предпосылок вероятностно-статистической модели выборки (если фактор X детерминирован).

Практическая рекомендация состоит в предварительном проведении отбраковки «выбросов» и проверке выполнения предпосылок вероятностно-статистической модели.

Каждый из рассмотренных выше примеров можно было бы развернуть в отдельную статью, в частности, дав определения используемых понятий и рассмотрев численные примеры. Мы не стали этого делать, чтобы не увеличивать объем статьи. Определения имеются в Интернете, а расчеты читатель может провести самостоятельно.

Как уже отмечалось [12], основная проблема современной науки — недостаточная осведомленность научных работников. Мы постарались показать, что нельзя бездумно применять распространенные программные продукты [13]. Необходимо владеть основами прикладной статистики. Иначе вместо обоснованных результатов статистического анализа данных можно получить ошибочные заключения.

Отметим, что многие важные результаты (в частности, принадлежащие А. Н. Колмогорову и С. Н. Бернштейну) были получены много десятилетий назад. Следовательно, грубо ошибочны встречающаяся иногда ориентация исследователей только на публикации последних пяти лет, а также требования редакций научных журналов не включать в список литературы «старые» источники, которые могут являться основополагающими.

ЛИТЕРАТУРА

1. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.
2. **Орлов А. И.** Устойчивость в социально-экономических моделях. — М.: Наука, 1979. — 296 с.
3. **Налимов В. В.** Теория эксперимента. — М.: Наука, 1971. — 208 с.
4. **Ермаков С. М., Бродский В. З., Жиглявский А. А. и др.** Математическая теория планирования эксперимента. — М.: Физматлит, 1983. — 392 с.
5. **Бернштейн С. Н.** Об одном элементарном свойстве коэффициента корреляции / Зап. Харьк. матем. тов. 1932. № 5. С. 65 – 66; **Бернштейн С. Н.** Собрание сочинений. Т. IV. Теория вероятностей. Математическая статистика. — М.: Наука, 1964. С. 233 – 234.
6. **Колмогоров А. Н.** К вопросу о пригодности найденных статистическим путем формул прогноза / Журн. геофиз. 1933. Т. 3. С. 78 – 82; **Колмогоров А. Н.** Теория вероятностей и математическая статистика. — М.: Наука, 1986. С. 161 – 167.
7. **Орлов А. И.** Методы поиска наиболее информативных множеств признаков в регрессионном анализе / Заводская лаборатория. Диагностика материалов. 1995. Т. 61. № 1. С. 56 – 58.
8. **Орлов А. И.** Проблема множественных проверок статистических гипотез / Заводская лаборатория. Диагностика материалов. 1996. Т. 62. № 5. С. 51 – 54.
9. **Сердобольский В. И., Орлов А. И.** Статистический анализ при большом числе параметров / Программно-алгоритмическое обеспечение прикладного многомерного статистического анализа. Тезисы докладов III Всесоюзной школы-семинара. — М.: ЦЭМИ АН СССР, 1987. С. 151 – 160.
10. **Орлов А. И.** Организационно-экономическое моделирование: учебник. В 3-х ч. Ч. 1. Нечисловая статистика. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2009. — 542 с.
11. **Орлов А. И.** Статистический контроль по двум альтернативным признакам и метод проверки их независимости по совокупности малых выборок / Заводская лаборатория. Диагностика материалов. 2000. Т. 66. № 1. С. 58 – 62.
12. **Лойко В. И., Луценко Е. В., Орлов А. И.** Современные подходы в наукометрии: монография. — Краснодар: КубГАУ, 2017. — 532 с. https://elibrary.ru/item.asp?id=29306423.
13. **Орлов А. И.** Статистические пакеты — инструменты исследователя / Заводская лаборатория. Диагностика материалов. 2008. Т. 74. № 5. С. 76 – 78.

REFERENCES

1. **Orlov A. I.** Applied statistics. — Moscow: Ékzamen, 2006. — 671 p. [in Russian].
2. **Orlov A. I.** Stability in socio-economic models. — Moscow: Nauka, 1979. — 296 p. [in Russian].
3. **Nalimov V. V.** Theory of experiment. — Moscow: Nauka, 1971. — 208 p. [in Russian].
4. **Ermakov S. M., Brodskii V. Z., Zhiglyavskii A. A., et al.** Mathematical theory of design of experiments. — Moscow: Fizmatlit, 1983. — 392 p. [in Russian].
5. **Bernshtein S. N.** On an elementary property of the correlation coefficient / Zap. Khar'k. Matem. Tov. 1932. N 5. P. 65 – 66 [in Russian]; **Bernshtein S. N.** Collected works. Vol. IV. Probability theory. Mathematical statistics. — Moscow: Nauka, 1964. P. 233 – 234 [in Russian].
6. **Kolmogorov A. N.** To the question of the suitability of the predicted formulas found statisticall / Zh. Geofiz. 1933. Vol. 3. P. 78 – 82; **Kolmogorov A. N.** Theory of Probability and Mathematical Statistics. — Moscow: Nauka, 1986. P. 161 – 167 [in Russian].
7. **Orlov A. I.** Methods for finding the most informative sets of characteristics in regression analysis / Zavod. Lab. Diagn. Mater. 1995. Vol. 61. N 1. P. 56 – 58 [in Russian].
8. **Orlov A. I.** The problem of multiple tests of statistical hypotheses / Zavod. Lab. Diagn. Mater. 1996. Vol. 62. N 5. P. 51 – 54.
9. **Serdobol'skii V. I., Orlov A. I.** Statistical analysis with a large number of parameters / Software and algorithmic support of applied multidimensional statistical analysis. Abstracts of the III All-Union School-Seminar. — Moscow: TsÉMI AN SSSR, 1987. P. 151 – 160.
10. **Orlov A. I.** Organizational-economic modeling: textbook. In 3 parts. Part 1. Non-numeric statistics. — Moscow: Izd. MGTU im. N. É. Baumana, 2009. — 542 p. [in Russian].
11. **Orlov A. I.** Statistical control of two alternative variables and a method for verifying their independence from a set of small samples / Zavod. Lab. Diagn. Mater. 2000. Vol. 66. N 1. P. 58 – 62 [in Russian].
12. **Loiko V. I., Lutsenko E. V., Orlov A. I.** Modern approaches in scientometrics: monograph. — Krasnodar: KubGAU, 2017. — 532 p. https://elibrary.ru/item.asp?id=29306423 [in Russian].
13. **Orlov A. I.** Statistical packages — researcher tools / Zavod. Lab. Diagn. Mater. 2008. Vol. 74. N 5. P. 76 – 78 [in Russian].