УДК 303.732.4 : 519.2

08.00.13 Математические и инструментальные методы экономики (экономические науки)

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКИЕ МОДЕЛИ КОРРЕЛЯЦИИ И РЕГРЕССИИ

Орлов Александр Иванович
д.э.н., д.т.н., к.ф.-м.н., профессор
РИНЦ SPIN-код: 4342-4994
Московский государственный технический университет им. Н.Э. Баумана, Россия, 105005, Москва, 2-я Бауманская ул., 5, prof-orlov@mail.ru

Коэффициенты корреляции и детерминации широко используются при статистическом анализе данных. Согласно теории измерений линейный парный коэффициент корреляции Пирсона применим к переменным, измеренным в шкале интервалов. Его нельзя использовать при анализе порядковых данных. Непараметрические ранговые коэффициенты Спирмена и Кендалла оценивают связь порядковых переменных. Критическое значение при проверке значимости отличия коэффициента корреляции от 0 зависит от объема выборки. Поэтому использование "шкалы Чеддока" некорректно. При применении пассивного эксперимента коэффициенты корреляции обоснованно использовать для прогнозирования, но не для управления. Для получения предназначенных для управления вероятностно-статистических моделей необходим активный эксперимент. Влияние выбросов на коэффициент корреляции Пирсона весьма велико. При увеличении числа проанализированных наборов предикторов заметно растет максимальный из соответствующих коэффициентов корреляции - показателей качества приближения (эффект «вздувания» коэффициента корреляции). Рассмотрены четыре основные модели регрессионного анализа. Выделены модели метода наименьших квадратов с детерминированной независимой переменной. Распределение отклонений произвольно, однако для получения предельных распределений оценок параметров и регрессионной зависимости предполагаем выполнение условий центральной предельной теоремы. Второй тип моделей основан на выборке случайных векторов. Зависимость является непараметрической, распределение двумерного вектора - произвольным. Об оценке дисперсии независимой переменной можно говорить только в модели на основе выборки случайных векторов, равно как и о коэффициенте детерминации как критерии качества модели. Обсуждается сглаживание временных рядов. Рассмотрены методы восстановления зависимостей в пространствах общей природы. Показано, что предельное распределение

UDC 303.732.4 : 519.2

08.00.13 - Mathematical and instrumental methods of Economics

PROBABILITY-STATISTICAL MODELS OF CORRELATION AND REGRESSION

Orlov Alexander Ivanovich
Dr.Sci.Econ., Dr.Sci.Tech., Cand.Phys-Math.Sci., professor
Bauman Moscow State Technical University, Moscow, Russia

The correlation and determination coefficients are widely used in statistical data analysis. According to measurement theory, Pearson's linear paired correlation coefficient is applicable to variables measured on an interval scale. It cannot be used in the analysis of ordinal data. The nonparametric Spearman and Kendall rank coefficients estimate the relationship of ordinal variables. The critical value when testing the significance of the difference of the correlation coefficient from 0 depends on the sample size. Therefore, using the Chaddock Scale is incorrect. When using a passive experiment, the correlation coefficients are reasonably used for prediction, but not for control. To obtain probabilistic-statistical models intended for control, an active experiment is required. The effect of outliers on the Pearson correlation coefficient is very large. With an increase in the number of analyzed sets of predictors, the maximum of the corresponding correlation coefficients — indicators of approximation quality noticeably increases (the effect of “inflation” of the correlation coefficient). Four main regression analysis models are considered. Models of the least squares method with a determinate independent variable are distinguished. The distribution of deviations is arbitrary, however, to obtain the limit distributions of parameter estimates and regression dependences, we assume that the conditions of the central limit theorem are satisfied. The second type of model is based on a sample of random vectors. The dependence is nonparametric, the distribution of the two-dimensional vector is arbitrary. The estimation of the variance of an independent variable can be discussed only in the model based on a sample of random vectors, as well as the determination coefficient as a quality criterion for the model. Time series smoothing is discussed. Methods of restoring dependencies in spaces of a general nature are considered. It is shown that the limiting distribution of the natural estimate of the dimensionality of the model is geometric, and the construction of an informative subset of features encounters the effect of "inflation coefficient correlation". Various approaches to the regression analysis of interval data

естественной оценки размерности модели является геометрическим, а построение информативного подмножества признаков наталкивается на эффект "вздувания коэффициентов корреляции". Обсуждаются различные подходы к регрессионному анализ интервальных данных. Анализ многообразия моделей регрессионного анализа приводит к выводу, что не существует единой "стандартной модели"

are discussed. Analysis of the variety of regression analysis models leads to the conclusion that there is no single “standard model”

Ключевые слова: МАТЕМАТИЧЕСКАЯ СТАТИСТИКА, НОВАЯ ПАРАДИГМА ПРИКЛАДНОЙ СТАТИСТИКИ, КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА, НЕПАРАМЕТРИЧЕСКИЕ РАНГОВЫЕ КОЭФФИЦИЕНТЫ КОРРЕЛЯЦИИ, ВЫБРОСЫ, КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ, РЕГРЕССИОННЫЙ АНАЛИЗ, МЕТОД НАИМЕНЬШИХ КВАДРАТОВ, НЕПАРАМЕТРИЧЕСКАЯ СТАТИСТИКА, НЕЧИСЛОВАЯ СТАТИСТИКА, ОЦЕНКА РАЗМЕРНОСТИ МОДЕЛИ, СТАТИСТИКА ИНТЕРВАЛЬНЫХ ДАННЫХ, РАСПРОСТРАНЕННЫЕ ОШИБОЧНЫЕ ВЫВОДЫ

Keywords: MATHEMATICAL STATISTICS, A NEW PARADIGM OF APPLIED STATISTICS, PEARSON CORRELATION COEFFICIENT, NONPARAMETRIC RANK CORRELATION COEFFICIENTS, OUTLIERS, DETERMINATION COEFFICIENT, REGRESSION ANALYSIS, LEAST-SQUARES METHODS, NONPARAMETRIC STATISTICS, NONNUMERIV STATISTICS, ESTIMATION OF THE DIMENSION OF THE MODEL, STATISTICS OF INTERVAL DATA, COMMON ERRONEOUS CONCLUSIONS

DOI: http://dx.doi.org/10.21515/1990-4665-160-011

1. Введение

Коэффициенты корреляции и детерминации широко используются при статистическом анализе данных. При этом достаточно часто допускаются те или иные ошибки. Некоторые из них рассмотрены ниже.

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 0. Обратное неверно - из равенства 0 коэффициента корреляции не следует, что случайные величины X и Y - независимы.

2. Значимость отличия от 0 и "шкала Чеддока"

Выборочные коэффициенты корреляции - случайные величины. Их распределения являются асимптотически нормальными.

Часто проверяют нулевую гипотезу о том, что тот или иной теоретический коэффициент корреляции равен 0. 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 0, не зависят от распределения данных.

Иногда показателям тесноты связи (модулям коэффициентов корреляции) пытаются дать качественную оценку (т.н. шкала Чеддока, см. табл.1):

Таблица 1. 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

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 0. А при достаточно большом объеме выборки коэффициент 0,1 может свидетельствовать о необходимости отклонения такой гипотезы.

3. 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 1. Весьма мало отличался от 1 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 число 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 число преступлений и число психических заболеваний. В чем причина неудачи, казалось бы, естественного подхода к управлению? Дело в том, что значения всех трех рассматриваемых переменных определяются значениями четвертой переменной (латентной, скрытой) - числа 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

проводится 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). Тем не менее, учитывая преимущества активного эксперимента, тогда, когда это возможно, предпочтение отдают ему. Теория планирования экспериментов [3, 4] посвящена прежде всего активным экспериментам.

4. Влияние выбросов на коэффициент корреляции

Акад. АН СССР С.Н. Бернштейн еще в 1932 г. 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

$$0 \le \lambda \le \frac{Y}{X} \le L\,.\text{"}$$

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 равен

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

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

$$Y - Lx = 0,\ Y - \lambda X = 0\,.$$

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* = 0,1*X*, 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.

5. 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 т.н. «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].

6. 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 вектора. 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 1 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 в п.4).

Формально рассчитанный коэффициент детерминации в рассматриваемой постановке может быть сколь угодно близким к 1. 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.

7. Многообразие моделей и методов регрессионного анализа

За столетия разработки математических методов исследования накоплен огромный массив научных результатов. Так, еще 30 лет назад мы оценивали [14] число статей и книг в этой области как 10^6, в том числе актуальных для современных исследователей - как 10^5. Сколькими статьями и книгами может овладеть один человек? Для большинства - 10^3, для отдельных наиболее продвинутых лиц - 10^4, что на порядки меньше, чем объем накопленных научных результатов. Следовательно, необходимы работы по упорядочению накопленных научных результатов. Для успешной работы важно единообразное понимание терминов. Необходимо знание фактов и тенденций развития. Обсудим эти вопросы на примере научной области "модели регрессионного анализа (восстановления зависимостей)" с целью сформировать единую методологическую базу для обсуждения различных частных вопросов этой области. Рассмотрим четыре метода восстановления зависимости.

В простейшем случае есть одна независимая количественная переменная t и одна зависимая количественная переменная x. Требуется указать (как говорят, восстановить) функцию, описывающую зависимость x от t.

В простейшем случае принимают, что эта зависимость - линейная: $x(t) = at + b$. Исходные данные - набор n двумерных векторов ж Предполагается, что имеются отклонения от линейности, т.е. $x_i = at_i + b + e_i$, где e_i, $i = 1, 2, ..., n$, - погрешности (отклонения, невязки). Необходимо оценить неизвестные параметры a и b.

Как известно, оценивание можно провести разными способами. Есть графический метод. Он состоит в том, что точки (t_i, x_i), $i = 1, 2, ..., n$, надо нанести на плоскость и провести с помощью линейки прямую линию, наилучшим образом приближающую эти точки (можно использовать миллиметровую бумагу или опцию "Корреляционное поле" в программном продукте для работы с электронными таблицами EXCEL). Недостатки - субъективизм и невозможность указать точность оценивания зависимости и ее параметров.

Чаще используют расчетные методы. Основная идея состоит в том, чтобы минимизировать одновременно все отклонения $x_i - at_i - b$. Реализовать эту идею можно различными способами. В методе наименьших модулей минимизируют по a и b функцию

$$g(a,b) = \sum_{i=1}^{n} | x_i - at_i - b | .$$

В методе минимакса в качестве показателя суммарного отклонения вместо суммы модулей минимизируют максимальное отклонение

$$h(a,b) = \max_{1 \le i \le n} | x_i - at_i - b | .$$

В 1794 г. К. Гаусс разработал метод наименьших квадратов, основанный на минимизации

$$f(a,b) = \sum_{i=1}^{n} (x_i - at_i - b)^2 .$$

Метод наименьших квадратов выглядит менее естественным, чем метод наименьших квадратов и метод минимакса. Действительно, почему квадрат, а не другая степень? Однако используют и применяют именно метод наименьших квадратов, а остальные два метода - маргинальные, ими занимаются отдельные энтузиасты. Почему в конкурентной борьбе победил именно метод наименьших квадратов? По нашему мнению, дело в том, что оценки параметров a и b метода наименьших квадратов, полученные в результате минимизации $f(a, b)$, задаются элементарными формулами (см., например, [1]), в то время как оценки параметров для двух других методов могут быть найдены лишь с помощью численных алгоритмов [15]. Причина сказанного в том, что для минимизации $f(a, b)$ можно использовать частные производные этой функции по параметрам a и b, в то время как $g(a, b)$ и $h(a, b)$ не дифференцируемы из-за наличия в них модуля. Наличие точных формул не только облегчает вычисление оценок метода наименьших квадратов, но и позволяет глубоко изучить свойства этих оценок.

В проведенных рассуждениях не было никаких вероятностно-статистических моделей. Действительно, метод наименьших квадратов и другие ранее упомянутые методы можно рассматривать в рамках теории приближений. Однако, если целесообразно перенести выводы с набора точек (t_i, x_i), $i = 1, 2, ..., n$, на более широкую совокупность, то необходимо ввести вероятностно-статистические модели, нацеленные на переход от выборки к генеральной совокупности.

Рассмотрим два основных типа вероятностно-статистических моделей.

8. Модели с детерминированной независимой переменной

Широко применяются модели с детерминированной независимой количественной переменной *t*. Для зависимой количественной переменной *x* случайность вводится с помощью равенств $x_i = at_i + b + e_i$, в правой части которых стоят случайные погрешности (отклонения, невязки) $e_1, e_2, \dots, e_n$. Отличительная черта этого типа моделей состоит в том, что независимая переменная является детерминированной, а зависимая - случайной.

В базовой модели случайные величины $e_1, e_2, \dots, e_n$. предполагаются независимыми и одинаково распределенными. Каково их общее распределение? В устаревших литературных источниках часто принимают, что их распределение является нормальным (гауссовским). Однако хорошо известно, что практически все распределения реальных данных не являются нормальными [1, 16]. Поэтому согласно новой парадигме математической статистики [17] следует считать распределение случайные величины $e_1, e_2, \dots, e_n$ произвольным, с одним ограничением - для получения предельных распределений оценок параметров и значений задающей зависимость функции целесообразно предположить выполнение условий центральной предельной теоремы.

Согласно [18] модель восстановления зависимости с независимыми одинаково распределенными случайными погрешностями, имеющими распределения произвольного вида, называется непараметрической. Именно ее следует использовать на практике, поскольку параметрическая модель регрессионного анализа, особенно с нормальными ошибками, не соответствует реальности. Здесь под параметрической моделью понимают модель, в которой распределения погрешностей принадлежат тому или иному параметрическому семейству - подсемейству четырехпараметрического семейства К. Пирсона [19]. Если в описании алгоритма регрессионного анализа используются распределения Стьюдента или Фишера, то необходимо констатировать, что распределения погрешностей предполагаются нормальными,

следовательно, алгоритм не соответствует новой парадигме математической статистики. Отметим, что при непараметрической модели погрешностей сама зависимость может являться параметрической, например, линейной. Как показано в дальнейшем, есть много вариантов постановки задач непараметрической регрессии.

Простейшая модель обобщается в двух направлениях - переход от линейной модели к более общей параметрической зависимости и отказ от независимости и одинаковой распределенности погрешностей. Параметрическая зависимость должна быть линейной по параметрам. Например, типовой является зависимость

$$x_i = a_1 f_1(t_i) + a_2 f_2(t_i) + \ldots + a_m f_m(t_i) + e_i,\ i = 1, 2, \ldots, n, \qquad (1)$$

где функции $f_1(t)$, $f_2(t)$, ... , $f_m(t)$ заданы, а параметры a_1, a_2, ... , a_m подлежат оценке методом наименьших квадратов. В частном случае, когда $f_k(t) = t^{k-1}$, $k = 1, 2, ..., m$, зависимость (1) является многочленом. Если же зависимость не является линейной по параметрам, то минимизацию в методе наименьших квадратов можно провести лишь численно, а теоретическое изучение свойств оценок встречает сложности.

Переход от одной независимой переменной к нескольким не представляет методологических сложностей.

Много постановок порождает отказ от независимости и одинаковой распределенности погрешностей. Например, дисперсии независимых погрешностей могут зависеть от независимой переменной t, например, линейно. Тогда абсолютные отклонения в методе наименьших квадратов заменяют относительными. Отказ от независимости погрешностей приводит к более сложным моделям, поскольку зависимость можно моделировать многими способами. Наиболее простой является модель, в которой все пары погрешностей имеют одинаковые коэффициенты корреляции. В рассматриваемой области необходимы новые исследования.

9. Модели анализа случайных векторов

Второй основной тип вероятностно-статистических моделей основан на выборке случайных векторов. В таких моделях исходные данные в простейшем случае - двумерные случайные вектора $(x_i(\omega), y_i(\omega)), i = 1,2,...,n,$ определенные на одном и том же вероятностном пространстве. В базовой модели все эти случайные вектора независимы и одинаково распределены с вектором $(x(\omega), y(\omega))$. В качестве оцениваемой зависимости рассматривают условное математическое ожидание $y(\omega)$ при условии заданного значения $x(\omega)$.

Пусть случайный вектор $(x(\omega), y(\omega))$ имеет плотность $p(x, y)$. Как известно из теории вероятностей, плотность условного распределения $y(\omega)$ при условии $x(\omega) = x_0$ имеет вид

$$p(y \mid x) = p(y \mid x(\omega) = x_0) = \frac{p(x, y)}{\int_{-\infty}^{+\infty} p(x, y)dy}.$$

Условное математическое ожидание, т.е. регрессионная зависимость y от x, имеет вид

$$f(x) = \int_{-\infty}^{+\infty} yp(y \mid x)dy = \frac{\int_{-\infty}^{+\infty} yp(x, y)dy}{\int_{-\infty}^{+\infty} p(x, y)dy}.$$

Таким образом, для нахождения оценок регрессионной зависимости достаточно найти оценки совместной плотности распределения вероятности $p_n(x, y)$ такие, что

$$p_n(x, y) \to p(x, y)$$

при $n \to \infty$. Тогда непараметрическая оценка регрессионной зависимости

$$f_n(x) = \frac{\int_{-\infty}^{+\infty} yp_n(x, y)dy}{\int_{-\infty}^{+\infty} p_n(x, y)dy}$$

при $n \to \infty$ является состоятельной оценкой регрессии как условного математического ожидания, т.е.

$$f_n(x) \to f(x).$$

Общий подход к построению непараметрических оценок плотности распределения вероятностей в пространствах различной природы развит в ряде публикаций (см., например, [1]), крайняя по времени статья [20].

Таким образом, если выборка $(x_i(\omega), y_i(\omega)), i = 1,2,...,n,$ состоит из случайных векторов, то базовая модель восстановления зависимости является двойной непараметрической, т.е. зависимость является непараметрической и распределение двумерного вектора является произвольным. Как уже отмечалось, принимать гипотезу многомерной нормальности нет оснований. В некоторых случаях полезны параметрические модели зависимости, например,

$$y = b_1\varphi_1(x) + b_2\varphi_2(x) + \ldots + b_m\varphi_m(x) + e_y. \quad (2)$$

где функции $\varphi_1(x)$, $\varphi_2(x)$, ..., $\varphi_m(x)$ заданы, а параметры b_1, b_2, ... , b_m подлежат оценке методом наименьших квадратов. В отличие от (1), в правой части (2) все слагаемые - случайные величины.

Итак, две основные модели основаны на детерминированной независимой переменной и выборке случайных векторов соответственно. Хотя расчетные алгоритмы метода наименьших квадратов во многом совпадают, но интерпретации результатов расчетов могут различаться. Так, об оценке дисперсии независимой переменной можно говорить только в модели на основе выборки случайных векторов, равно как и о коэффициенте детерминации как критерии качества модели. В случае модели на основе детерминированной независимой переменной попытки применять коэффициенте детерминации в качестве критерия качества модели могут привести к грубым ошибкам.

10. Сглаживание временных рядов

С формальной точки зрения временные ряды являются частным случаем моделей с детерминированной независимой переменной, в качестве которой рассматривается время *t*. При этом для зависимой переменной $X(t)$ часто рассматривают аддитивную модель

$$X(t) = T(t) + P(t) + R(t), \quad (3)$$

где $T(t)$ - тренд, задающий центральную тенденцию, $P(t)$ - периодическая составляющая, $R(t)$ - случайная составляющая. Иногда рассматривают мультипликативную модель $X(t) = T(t)\ P(t)\ R(t)$, однако она не имеет самостоятельного значения, поскольку после логарифмирования переходит в модель (3) для логарифмов включенных в модель составляющих.

Для модели (3) рассматривают различные варианты непараметрики. Например, тренд $T(t)$ может задаваться линейной функцией, а периодическая составляющая $P(t)$ быть произвольной. Методы непараметрического оценивания периодической составляющей для такой модели разработаны в [21].

От независимости отклонений приходится отказаться при движении от дискретного времени к непрерывному. В пределе отклонения моделируются случайным процессом с непрерывными траекториями. Так поступают при моделировании динамики курсов акций и валют. Математическая теория оценивания в случае непрерывных случайных процессов существенно отличается от таковой в случае выборок погрешностей.

11. Методы восстановления зависимостей в пространствах общей природы

Обсудим модели регрессионного анализа в общем виде. Сначала рассмотрим параметрические постановки задач регрессионного анализа (восстановления зависимостей) в пространствах произвольной природы,

затем — непараметрические, после чего перейдем к оцениванию нечисловых параметров в классической ситуации, когда отклик и факторы принимают числовые значения.

Задача аппроксимации зависимости (параметрической регрессии). Пусть X и Y — некоторые пространства. Пусть имеются статистические данные — n пар (x_k, y_k), где $x_k \in X$, $y_k \in Y$, $k = 1, 2, \ldots, n$. Задано параметрическое пространство Θ произвольной природы и семейство функций $g(x,\theta)$: $X \times \Theta \to Y$. Требуется подобрать параметр $\theta \in \Theta$ так. чтобы $g(x_k, \theta)$ наилучшим образом приближали y_k, $k = 1, 2, \ldots, n$. Пусть f_k — последовательность показателей различия в Y. При сделанных предположениях параметр θ естественно оценивать путем решения экстремальной задачи:

$$\theta_n = Arg \min_{\theta \in \Theta} \sum_{k=1}^{n} f_k(g(x_k, \theta), y_k). \quad (4)$$

Часто, но не всегда, все f_k совпадают. В классической постановке, когда $X = R^k$, $Y = R^1$, функции f_k различны при неравноточных наблюдениях, например, когда число опытов меняется от одной точки x проведения опытов к другой.

Если $f_k(y_1, y_2) = f(y_1, y_2) = (y_1 - y_2)^2$, то получаем общую постановку метода наименьших квадратов:

$$\theta_n = Arg \min_{\theta \in \Theta} \sum_{k=1}^{n} (g(x_k, \theta) - y_k)^2.$$

В рамках детерминированного анализа данных остается единственный теоретический вопрос — о существовании θ_n. Если все участвующие в формулировке задачи (4) функции непрерывны, а минимум берется по бикомпакту, то θ_n существует. Есть и иные условия существования θ_n [10].

При появлении нового наблюдения x в соответствии с методологией восстановления зависимости рекомендуется выбирать оценку соответствующего y по правилу

$$y^* = g(x, \theta_n).$$

Обосновать такую рекомендацию в рамках детерминированного анализа данных невозможно. Это можно сделать только в вероятностной теории, равно как и изучить асимптотическое поведение θ_n, доказать состоятельность этой оценки.

Как и в классическом случае, вероятностную теорию целесообразно строить для трех различных постановок.

1. Переменная x — детерминированная (например, время), переменная y — случайная, ее распределение зависит от x.

2. Совокупность (x_k, y_k), $k = 1, 2, \ldots, n$, — выборка из распределения случайного элемента со значениями в $X \times Y$.

3. Имеется детерминированный набор пар (x_{k0}, y_{k0}), $k = 1, 2, \ldots, n$, результат наблюдения (x_k, y_k) является случайным элементом, распределение которого зависит от (x_{k0}, y_{k0}). Это — постановка т.н. конфлюэнтного анализа.

Во всех трех случаях

$$f_n(\omega, \theta) = \sum_{k=1}^{n} f_k(g(x_k, \theta), y_k),$$

однако случайность входит в правую часть по-разному в зависимости от постановки, от которой зависит и определение предельной функции $f(\theta)$.

Проще всего выглядит $f(\theta)$ в случае второй постановки при $f_k \equiv f$:

$$f(\theta) = Mf(g(x_1, \theta), y).$$

В случае первой постановки

$$f(\theta) = \lim_{n\to\infty} \sum_{k=1}^{n} Mf_k(g(x_k, \theta), y_k(\omega))$$

в предположении существования указанного предела. Ситуация усложняется для третьей постановки:

$$f(\theta) = \lim_{n\to\infty} \sum_{k=1}^{n} Mf_k(g(x_k(\omega), \theta), y_k(\omega)).$$

Во всех трех случаях на основе общих результатов о поведении решений экстремальных статистических задач можно изучить асимптотику оценок θ_n методами нечисловой статистики [10]. При выполнении соответствующих внутриматематических условий регулярности оценки оказываются состоятельными, т.е. удается восстановить зависимость.

Аппроксимация и регрессия. Соотношение (4) дает решение задачи аппроксимации. Поясним, как эта задача соотносится с нахождением регрессии. Согласно [10] для случайной величины (ξ, η) со значениями в $X \times Y$ регрессией η на ξ относительно меры близости f естественно назвать решение задачи

$$Mf(g(\xi), \eta) \to \min_g, \quad (5)$$

где $f: Y \times Y \to R^1$, $g: X \to Y$, минимум берется по множеству всех измеримых функций.

Можно исходить и из формально другого определения. Для каждого $x \in X$ рассмотрим случайную величину $\eta(x)$, распределение которой является условным распределением η при условии $\xi = x$. В соответствии с определением математического ожидания в пространстве общей природы назовем условным математическим ожиданием решение экстремальной задачи

$$M(\eta \mid \xi = x) = Arg \min\{Mf(y, \eta(x)), y \in Y\}.$$

Оказывается, при обычных предположениях измеримости решение задачи (5) совпадает с $M(\eta \mid \xi = x)$. (Внутриматематические уточнения типа «равенство имеет место почти всюду» здесь опущены.)

Если заранее известно, что условное математическое ожидание $M(\eta \mid \xi = x)$ принадлежит некоторому параметрическому семейству $g(x, \theta)$, то задача нахождения регрессии сводится к оцениванию параметра θ в соответствии с рассмотренной выше второй постановкой вероятностной теории параметрической регрессии.

Если же нет оснований считать, что регрессия принадлежит некоторому параметрическому семейству, можно использовать непараметрические оценки регрессии. Они строятся с помощью непараметрических оценок плотности [1, 20].

Непараметрические методы восстановления зависимости. Пусть ν_1 — мера в X, ν_2 — мера в Y, а их прямое произведение $\nu = \nu_1 \times \nu_2$ — мера в $X \times Y$. Пусть $g(x, y)$ — плотность случайного элемента (ξ, η) по мере ν. Тогда условная плотность $g(y|x)$ распределения η при условии $\xi = x$ имеет вид

$$g(y \mid x) = \frac{g(x, y)}{\int\limits_Y g(x, y)\nu_2(dy)} \qquad (6)$$

(в предположении, что интеграл в знаменателе отличен от 0). Следовательно,

$$Mf(y, \eta(x)) = \int\limits_Y f(y, a)g(a \mid x)\nu_2(da),$$

а потому

$$M(\eta \mid \xi = x) = Arg \min_{y \in Y} Mf(y, \eta(x)) =$$
$$= Arg \min_{y \in Y} \int\limits_Y f(y, a)g(a \mid x)\nu_2(da).$$

Заменяя $g(x,y)$ в (6) непараметрической оценкой плотности $g_n(x,y)$, получаем оценку условной плотности

$$g_n(y \mid x) = \frac{g_n(x, y)}{\int\limits_Y g_n(x, y)\nu_2(dy)}. \qquad (7)$$

Если $g_n(x,y)$ — состоятельная оценка $g(x,y)$, то числитель (7) сходится к числителю (6). Сходимость знаменателя (7) к знаменателю (6) обосновывается с помощью предельной теории статистик интегрального типа [12]. В итоге получаем утверждение о состоятельности непараметрической оценки (7) условной плотности (6).

Непараметрическая оценка регрессии ищется как

$$M_n(\eta \mid \xi = x) = Arg \min_{y \in Y} \int_Y f(y,a) g_n(a \mid x) \nu_2(da).$$

Состоятельность этой оценки следует из приведенных выше общих результатов об асимптотическом поведении решений экстремальных статистических задач.

12. Оценивание объектов нечисловой природы в классических постановках регрессионного анализа

Нечисловая статистика тесно связана с классическими областями прикладной статистики. Ряд трудностей в классических постановках удается понять и разрешить лишь с помощью общих результатов прикладной статистики. В частности, это касается оценивания параметров, когда параметр имеет нечисловую природу.

Рассмотрим типовую прикладную постановку задачи восстановления регрессионной зависимости, линейной по параметрам. Исходные данные имеют вид $(x_i, y_i) \in R^2,\ i = 1, 2, ..., n$. Цель состоит в том, чтобы с достаточной точностью описать *y* как многочлен (полином) от *x*, т.е. модель имеет вид

$$y_i = \sum_{k=0}^{m} a_k x_i^k + \varepsilon_i, \quad i = 1, 2, ..., n, \tag{8}$$

где m — неизвестная степень полинома; a_0, a_1, a_2, …, a_m — неизвестные коэффициенты многочлена; $\varepsilon_i,\ i = 1,\ 2, ..., n$, — погрешности, которые для простоты примем независимыми и имеющими одно и то же нормальное распределение с нулевым математическим ожиданием и дисперсией σ^2.

В прикладной статистике часто используют следующую технологию анализа данных. Сначала пытаются применить модель (8) для линейной функции (m = 1), при неудаче (неадекватности модели) переходят к многочлену второго порядка (m = 2), если снова неудача, то берут модель (8) с m = 3 и т.д. Адекватность модели обычно проверяют по *F*-критерию Фишера, *основанному на предположении нормальности погрешностей*.

Обсудим свойства этой процедуры. Если степень полинома задана ($m = m_0$), то его коэффициенты оценивают методом наименьших квадратов, свойства этих оценок хорошо известны. Однако в рассматриваемой постановке m тоже является неизвестным параметром и подлежит оценке. Таким образом, требуется оценить объект (m, a_0, a_1, a_2, …, a_m), множество значений которого можно описать как $R^1 \cup R^2 \cup R^3 \cup \ldots$. Это — объект нечисловой природы, обычные методы оценивания для него неприменимы. Разработанные к настоящему времени методы оценивания степени полинома носят в основном эвристический характер (см., например, гл. 12 монографии [22]). Рассмотрим некоторые из них.

Замечание. Здесь наглядно проявляется одна из причин живучести вероятностно-статистических моделей на основе нормального распределения. Такие модели, как правило, не адекватны реальной ситуации, о чем сказано выше. Однако с математической точки зрения они позволяют глубже проникнуть в суть изучаемого явления. Поэтому такие модели полезны для первоначального анализа ситуации. В ходе дальнейших исследований необходимо снять нереалистическое предположение нормальности и перейти к непараметрическим моделям.

Оценивание степени полинома. Полезно рассмотреть основной показатель качества регрессионной модели (8). Одни и те же данные можно обрабатывать различными способами. На первый взгляд, показателем отклонений данных от модели может служить остаточная сумма квадратов *SS*. Чем этот показатель меньше, тем приближение лучше, значит, и модель лучше описывает реальные данные. Однако это рассуждение годится только для моделей с одинаковым числом параметров. Ведь если добавляется новый параметр, по которому можно минимизировать, то и минимум, как правило, оказывается меньше.

В качестве основного показателя качества регрессионной модели используют следующую оценку остаточной дисперсии

$$\bar{\sigma}^2(m) = \frac{SS}{n-m-1}.$$

Таким образом, вводят корректировку на число параметров, оцениваемых по наблюдаемым данным. Корректировка состоит в уменьшении знаменателя на указанное число. В модели (8) это число равно (m + 1). В случае задачи восстановления линейной функции одной переменной оценка остаточной дисперсии имеет вид

$$\bar{\sigma}^2 = \frac{SS}{n-2},$$

поскольку число оцениваемых параметров m + 1 = 2.

Еще раз — почему *при подборе вида модели* знаменатель дроби, оценивающей остаточную дисперсию, приходится корректировать на число параметров? Если этого не делать, то придется заключить, что всегда многочлен второй степени лучше соответствует данным, чем линейная функция, многочлен третьей степени лучше приближает исходные данные, чем многочлен второй степени, и т.д. В конце концов доходим до многочлена степени (n – 1) с n коэффициентами, который проходит через все заданные точки. Но его прогностические возможности, скорее всего, существенно меньше, чем даже у линейной функции. *Излишнее усложнение статистических моделей вредно*.

Типовое поведение скорректированной оценки $v(m) = \bar{\sigma}^2(m)$ остаточной дисперсии в случае расширяющейся системы моделей (т.е. при возрастании натурального параметра m) выглядит так. Сначала наблюдаем заметное убывание. Затем оценка остаточной дисперсии колеблется около некоторой константы (дисперсии погрешности). Поясним ситуацию на примере модели восстановления зависимости, выраженной многочленом:

$$x(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \ldots + a_m t^m.$$

Пусть эта модель справедлива при $m = m_0$. При $m < m_0$ в скорректированной оценке остаточной дисперсии учитываются не только погрешности измерений, но и соответствующие (старшие) члены многочлена

(предполагаем, что коэффициенты при них отличны от 0). При $m \geq m_0$ имеем

$$\lim_{n\to\infty} v(m) = \sigma^2.$$

Следовательно, скорректированная оценка остаточной дисперсии будет колебаться около указанного предела. Поэтому представляется естественным, что в качестве оценки неизвестной статистику степени многочлена (полинома) можно использовать первый локальный минимум скорректированной оценки остаточной дисперсии, т.е.

$$m^* = \min\{m : v(m-1) > v(m), \quad v(m) \leq v(m+1)\}.$$

В работе [23] найдено предельное распределение этой оценки параметра, принимающего целые значения - степени многочлена.

Теорема. При справедливости некоторых условий регулярности

$$\lim_{n\to\infty} P(m^* < m_0) = 0, \ \lim_{n\to\infty} P(m^* = m_0 + u) = \lambda(1-\lambda)^u,$$
$$u = 0,1,2,....,$$

где

$$\lambda = \Phi(1) - \Phi(-1) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} \exp\left\{-\frac{x^2}{2}\right\} dx \approx 0,68268.$$

Таким образом, предельное распределение оценки m^* степени многочлена (полинома) является геометрическим. Это означает, в частности, что оценка не является состоятельной. При этом вероятность получить меньшее значение, чем истинное, исчезающе мала. Далее имеем:

$$P(m^* = m_0) \to 0,68268, \ P(m^* = m_0 + 1) \to 0,68268(1-0,68268) = 0,21663,$$

$$P(m^* = m_0 + 2) \to 0,68268(1-0,68268)^2 = 0,068744,$$

$$P(m^* = m_0 + 3) \to 0,68268(1-0,68268)^3 = 0,021814...$$

Разработаны и иные методы оценивания неизвестной степени многочлена, например, путем многократного применения процедуры проверки адекватности регрессионной зависимости с помощью критерия Фишера. Предельное поведение таких оценок — таково же, как в

приведенной выше теореме, только значение параметра λ иное. Для степени многочлена давно предложены состоятельные оценки [24]. Для этого достаточно уровень значимости (при проверке адекватности регрессионной зависимости с помощью критерия Фишера) сделать убывающим при росте объема выборки.

Построение информативного подмножества признаков. В более общем случае многомерной линейной регрессии данные имеют вид (y_i, X_i), $i = 1, 2, \ldots, n$, где $X_i = (x_{i1}, x_{i2}, \ldots, x_{iN}) \in R^N$ — вектор предикторов (факторов, объясняющих переменных), а модель такова:

$$y_i = \sum_{j \in K} a_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \ldots, n \qquad (9)$$

(здесь K — некоторое подмножество множества $\{1, 2, \ldots, n\}$; ε_i — те же, что и в модели (8); a_j — неизвестные коэффициенты при предикторах с номерами из K). Множество K называют *информативным подмножеством признаков*, поскольку согласно формуле (9) остальные признаки можно отбросить без потери информации. Проблема состоит в том, что при анализе реальных данных неизвестно, какие признаки входят в K, а какие нет. Ясна важность оценивания информативного подмножества признаков.

Модель (8) сводится к модели (9), если

$$x_{i1} = 1, x_{i2} = x_i, \; x_{i3} = x_i^2, \; x_{i4} = x_i^3, \; \ldots, x_{ij} = x_i^{j-1}, \; \ldots$$

В модели (8) есть естественный порядок ввода предикторов в рассмотрение — в соответствии с возрастанием степени многочлена, а в модели (9) естественного порядка нет, поэтому здесь приходится рассматривать произвольное подмножество множества предикторов. Есть только частичный порядок — чем мощность подмножества меньше, тем лучше. Модель (9) особенно актуальна в технических исследованиях (см. многочисленные примеры в журнале «Заводская лаборатория. Диагностика материалов»). Она применяется в задачах управления качеством

продукции и других технико-экономических исследованиях, в медицине, экономике, маркетинге и социологии, когда из большого числа факторов, предположительно влияющих на изучаемую переменную, надо отобрать по возможности наименьшее число значимых факторов и с их помощью сконструировать прогнозирующую формулу (9).

Задача оценивания модели (9) разбивается на две последовательные задачи: оценивание множества K — подмножества множества всех предикторов, а затем — неизвестных параметров a_j. Методы решения второй задачи хорошо известны и подробно изучены (обычно используют метод наименьших квадратов). Гораздо хуже обстоит дело с оцениванием объекта нечисловой природы K. Существующие методы — в основном эвристические, они зачастую не являются даже состоятельными. Даже само понятие состоятельности в данном случае требует специального определения.

Определение. Пусть K_0 — истинное подмножество предикторов, т.е. подмножество, для которого справедлива модель (9), а подмножество предикторов K_n — его оценка. Оценка K_n называется состоятельной, если

$$\lim_{n\to\infty} Card(K_n \Delta K_0) = 0,$$

где Δ — символ симметрической разности множеств; $Card(K)$ означает число элементов множества K, а предел понимается в смысле сходимости по вероятности.

Задача оценивания в моделях регрессии, таким образом, разбивается на две — оценивание структуры модели и оценивание параметров при заданной структуре. В модели (8) структура описывается неотрицательным целым числом m, в модели (9) — множеством K. Структура — объект нечисловой природы. Задача ее оценивания сложна, в то время как задача оценивания численных параметров при заданной структуре хорошо изучена, разработаны эффективные (в смысле прикладной математической

статистики) методы. Такова же ситуация и в других методах многомерного статистического анализа — в факторном анализе (включая метод главных компонент) и в многомерном шкалировании, в иных оптимизационных постановках проблем прикладного многомерного статистического анализа.

Множество K и параметры a_j линейной зависимости можно оценивать путем решения задачи оптимизации

$$\sum_{i=1}^{n}\left(y_i - \sum_{j\in K} a_j x_{ij}\right)^2 \to \min, \qquad (10)$$

в которой минимум берется по K, a_j, $j \in K$. Математическая природа множества, по которому проводится минимизация, весьма сложна. Это и объясняет тот факт, что к настоящему времени разработано много эвристических методов оценивания информативного множества параметров K, свойства которых плохо изучены. На основе общих результатов нечисловой статистики об асимптотическом поведении решений экстремальных статистических задач удалось показать, что оценки, полученные путем решения задачи (7), являются состоятельными [7].

К рассматриваемой тематике относится также эффект "вздувания коэффициентов корреляции", рассмотренный выше в п.5.

13. Регрессионный анализ интервальных данных

Иногда рассматривают модели, в которых как входная, так и выходная переменные имеют погрешности, определяемые значениями этих переменных. В простейшем случае вместо "истинных" данных (t_i, x_i), i = 1, 2, ..., n, наблюдают данные с погрешностями (q_i, y_i), i = 1, 2, ..., n. где q_i, = t_i + ε_i, y_i = x_i + δ_i. Здесь ε_i и δ_i - погрешности измерений (наблюдений, регистрации, опытов, анализов). Требуется восстановить зависимость между "истинными" переменными t и x.

Есть несколько подходов к решению. этой задачи. Если заданы ограничения на значения погрешностей, наложенных на случайные величины, то плодотворен подход разработанной нами статистики интервальных данных [25]. Восстановлению линейной зависимости в соответствии с подходом статистики интервальных данных посвящена статья [26]. Подробному изложению статистики интервальных данных посвящены развернутые главы в монографиях [1, 10, 27, 28].

Уходит в прошлое подход т.н. конфлюэнтного анализа, согласно которому погрешности измерений ε_i и δ_i имеют нормальные распределения. Поскольку, как уже отмечалось, распределения практически всех реальных величин не являются нормальными, конфлюэнтный анализ не является адекватным реальным ситуациям и потому не имеет практических перспектив. Точно также распределения Стьюдента и Фишера не адекватны реальности и могут иметь лишь теоретическое значение. Вместе с тем отметим, что, например, неизвестен непараметрический аналог критерия Фишера, предназначенного для проверки адекватности регрессионной модели (скажем, для проверки адекватности линейной модели, когда альтернативой является квадратическая).

14. Заключительные замечания

Как уже отмечалось [12], основная проблема современной науки - всеобщее невежество научных работников. Мы постарались показать, что нельзя бездумно применять распространенные программные продукты (ср. [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 ошибочна встречающаяся иногда ориентация исследователей и редакций научных журналов только на публикации последних 5 лет.

Анализ многообразия моделей регрессионного анализа приводит к выводу, что не существует единой "стандартной модели". В каждом конкретном случае необходимо описывать используемую модель и обосновывать ее.

Исследования в рассматриваемой области прикладной статистики ведутся активно, но много задач всё еще требует решения. Некоторые такие задачи отмечены выше. Например, разработанные в XX в. модели и методы, основанные на предположении нормальности, требуют осмысления и доработки (как теоретической, так и алгоритмической) с позиций непараметрической статистики. Критический разбор устоявшихся взглядов необходим для квалифицированного развития и применения математических методов исследования, в частности, для перехода на современную парадигму математической статистики [17].

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. Т. 5. С. 65-66.
6. Колмогоров А.Н. К вопросу о пригодности найденных статистическим путем формул прогноза / Журн. геофиз. 1933. Т.3. С. 78-82.
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 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 / Под науч. ред. проф. С. Г. Фалько. – Краснодар: КубГАУ, 2017. – 532 с.
13. Орлов А.И. Статистические пакеты – инструменты исследователя / Заводская лаборатория. Диагностика материалов. 2008. Т.74. № 5. С. 76-78.
14. Орлов А.И. Первый Всемирный конгресс Общества математической статистики и теории вероятностей им. Бернулли / Заводская лаборатория. Диагностика материалов. 1987. Т.53. №3. С.90-91.
15. Тырсин А.Н., Максимов К.Е. Оценивание линейных регрессионных уравнений с помощью метода наименьших модулей // Заводская лаборатория. Диагностика материалов. 2012. Том 78. № 7. С. 65-71.
16. Орлов А.И. Распределения реальных статистических данных не являются нормальными // Научный журнал КубГАУ. 2016. № 117. С. 71–90.
17. Орлов А.И. Основные черты новой парадигмы математической статистики / Научный журнал КубГАУ. 2013. № 90. С. 45-71.
18. Орлов А.И. Современное состояние непараметрической статистики / Научный журнал КубГАУ. 2015. № 106. С. 239 – 269.
19. Королюк В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. — М.: Наука, 1985. — 640 с.
20. Орлов А.И. Асимптотика оценок плотности распределения вероятностей / Научный журнал КубГАУ. 2017. № 131. С. 845 – 873.
21. Орлов А.И. Восстановление зависимости методом наименьших квадратов на основе непараметрической модели с периодической составляющей / Научный журнал КубГАУ. 2013. № 91. С. 133-162.
22. Себер Дж. Линейный регрессионный анализ. — М.: Мир, 1980. — 456 с.
23. Орлов А.И. Асимптотика некоторых оценок размерности модели в регрессии / Прикладная статистика. Ученые записки по статистике. Т. 45. — М.: Наука, 1983. — С. 260–265.
24. Орлов А.И. Об оценивании регрессионного полинома / Заводская лаборатория. Диагностика материалов. 1994. Т.60. №5. С. 43-47.
25. Орлов А.И. Статистика интервальных данных (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2015. Т.81. №3. С. 61 - 69.
26. Гуськова Е.А., Орлов А.И. Интервальная линейная парная регрессия (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2005. Т.71. №3. С.57-63.
27. Орлов А.И. Теория принятия решений. — М.: Экзамен, 2006. — 576 с.
28. Орлов А.И., Луценко Е.В. Системная нечеткая интервальная математика. – Краснодар, КубГАУ. 2014. – 600 с.

References

1. Orlov A.I. Prikladnaya statistika. — M.: Ekzamen, 2006. — 671 s.
2. Orlov A.I. Ustojchivost' v social'no-ekonomicheskih modelyah. — M.: Nauka, 1979. — 296 s.
3. Nalimov V.V. Teoriya eksperimenta. — M.: Nauka, 1971. — 208 s.

4. Ermakov S.M., Brodskij V.Z., ZHiglyavskij A.A. i dr. Matematicheskaya teoriya planirovaniya eksperimenta. — M.: Fizmatlit, 1983. — 392 s.
5. Bernshtejn S.N. Ob odnom elementarnom svojstve koefficienta korrelyacii / Zap. Har'k. matem. tov. 1932. T. 5. S. 65-66.
6. Kolmogorov A.N. K voprosu o prigodnosti najdennyh statisticheskim putem formul prognoza / ZHurn. geofiz. 1933. T.3. S. 78-82.
7. Orlov A.I. Metody poiska naibolee informativnyh mnozhestv priznakov v regressionnom analize / Zavodskaya laboratoriya. Diagnostika materialov. 1995. T.61. № 1. S. 56-58.
8. Orlov A.I. Problema mnozhestvennyh proverok statisticheskih gipotez / Zavodskaya laboratoriya. Diagnostika materialov. 1996. T.62. № 5. S. 51-54.
9. Serdobol'skij V.I., Orlov A.I. Statisticheskij analiz pri bol'shom chisle parametrov / Programmno-algoritmicheskoe obespechenie prikladnogo mnogomernogo statisticheskogo analiza. Tezisy dokladov III Vsesoyuznoj shkoly-seminara. — M.: CEMI AN SSSR, 1987. — S. 151-160.
10. Orlov A.I. Organizacionno-ekonomicheskoe modelirovanie: : uchebnik : v 3 ch. CH.1: Nechislovaya statistika. — M.: Izd-vo MGTU im. N. E. Baumana, 2009. — 542 s.
11. Orlov A.I. Statisticheskij kontrol' po dvum al'ternativnym priznakam i metod proverki ih nezavisimosti po sovokupnosti malyh vyborok / Zavodskaya laboratoriya. Diagnostika materialov. 2000. T.66. № 1. S. 58-62.
12. Lojko V. I., Lucenko E. V., Orlov A. I. Sovremennye podhody v naukometrii: monografiya / Pod nauch. red. prof. S. G. Fal'ko. – Krasnodar: KubGAU, 2017. – 532 s.
13. Orlov A.I. Statisticheskie pakety – instrumenty issledovatelya / Zavodskaya laboratoriya. Diagnostika materialov. 2008. T.74. № 5. S. 76-78.
14. Orlov A.I. Pervyj Vsemirnyj kongress Obshchestva matematicheskoj statistiki i teorii veroyatnostej im. Bernulli / Zavodskaya laboratoriya. Diagnostika materialov. 1987. T.53. №3. S.90-91.
15. Tyrsin A.N., Maksimov K.E. Ocenivanie linejnyh regressionnyh uravnenij s pomoshch'yu metoda naimen'shih modulej // Zavodskaya laboratoriya. Diagnostika materialov. 2012. Tom 78. № 7. S. 65-71.
16. Orlov A.I. Raspredeleniya real'nyh statisticheskih dannyh ne yavlyayutsya normal'nymi // Nauchnyj zhurnal KubGAU. 2016. № 117. S. 71–90.
17. Orlov A.I. Osnovnye cherty novoj paradigmy matematicheskoj statistiki / Nauchnyj zhurnal KubGAU. 2013. № 90. S. 45-71.
18. Orlov A.I. Sovremennoe sostoyanie neparametricheskoj statistiki / Nauchnyj zhurnal KubGAU. 2015. № 106. S. 239 – 269.
19. Korolyuk V.S., Portenko N.I., Skorohod A.V., Turbin A.F. Spravochnik po teorii veroyatnostej i matematicheskoj statistike. — M.: Nauka, 1985. — 640 s.
20. Orlov A.I. Asimptotika ocenok plotnosti raspredeleniya veroyatnostej / Nauchnyj zhurnal KubGAU. 2017. № 131. S. 845 – 873.
21. Orlov A.I. Vosstanovlenie zavisimosti metodom naimen'shih kvadratov na osnove neparametricheskoj modeli s periodicheskoj sostavlyayushchej / Nauchnyj zhurnal KubGAU. 2013. № 91. S. 133-162.
22. Seber Dzh. Linejnyj regressionnyj analiz. — M.: Mir, 1980. — 456 s.
23. Orlov A.I. Asimptotika nekotoryh ocenok razmernosti modeli v regressii / Prikladnaya statistika. Uchenye zapiski po statistike. T. 45. — M.: Nauka, 1983. — S. 260–265.
24. Orlov A.I. Ob ocenivanii regressionnogo polinoma / Zavodskaya laboratoriya. Diagnostika materialov. 1994. T.60. №5. S. 43-47.

25. Orlov A.I. Statistika interval'nyh dannyh (obobshchayushchaya stat'ya) / Zavodskaya laboratoriya. Diagnostika materialov. 2015. T.81. №3. S. 61 - 69.
26. Gus'kova E.A., Orlov A.I. Interval'naya linejnaya parnaya regressiya (obobshchayushchaya stat'ya) / Zavodskaya laboratoriya. Diagnostika materialov. 2005. T.71. №3. S.57-63.
27. Orlov A.I. Teoriya prinyatiya reshenij. — M.: Ekzamen, 2006. — 576 s.
28. Orlov A.I., Lucenko E.V. Sistemnaya nechetkaya interval'naya matematika. – Krasnodar, KubGAU. 2014. – 600 s.