Математические методы исследования
Mathematical methods of investigation

DOI: 10.26896/1028-6861-2018-84-5-63-73 УДК (UDC) 519.28

МНОГООБРАЗИЕ МОДЕЛЕЙ РЕГРЕССИОННОГО АНАЛИЗА
(обобщающая статья)

© **Александр Иванович Орлов**

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

Статья поступила 13 декабря 2017 г.

Научные результаты работ необходимо упорядочивать. Важное значение имеют единообразное понимание терминов, знание фактов и тенденций развития. Статья посвящена обсуждению этих вопросов на примере «модели регрессионного анализа (восстановления зависимостей)» в целях формирования единой методологической базы для обсуждения различных частных вопросов в этой области. Рассмотрены четыре метода восстановления зависимости. Выделены модели метода наименьших квадратов с детерминированной независимой переменной. Согласно новой парадигме прикладной статистики следует считать распределение отклонений (погрешностей, невязок) произвольным, с одним ограничением — для получения предельных распределений оценок параметров и зависимости целесообразно предположить выполнение условий центральной предельной теоремы. Другой основной тип вероятностно-статистических моделей метода наименьших квадратов основан на выборке случайных векторов. Зависимость является непараметрической и распределение двумерного вектора — произвольным. Об оценке дисперсии независимой переменной можно говорить только в модели на основе выборки случайных векторов, о коэффициенте детерминации как критерии качества модели — только в случае выборки случайных векторов. Исследованы вопросы сглаживания временных рядов. Рассмотрены методы восстановления зависимостей в пространствах общей природы. Показано, что предельное распределение естественной оценки размерности модели является геометрическим, а построение информативного подмножества признаков наталкивается на эффект «вздувания коэффициентов корреляции». Обсуждаются различные подходы к регрессионному анализу интервальных данных, констатируется уход в прошлое подхода так называемого конфлюэнтного анализа. Многообразие моделей регрессионного анализа приводит к выводу, что не существует единой «стандартной модели». Критический разбор устоявшихся взглядов необходим для квалифицированного развития и применения математических методов исследования, в частности, для перехода на современную парадигму прикладной статистики.

Ключевые слова: математическая статистика; новая парадигма прикладной статистики; регрессионный анализ; метод наименьших квадратов; непараметрическая статистика; нечисловая статистика; оценка размерности модели; статистика интервальных данных.

DIVERSITY OF THE MODELS FOR REGRESSION ANALYSIS
(generalizing article)

© **Alexandr I. Orlov**

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

Submitted December 13, 2017.

Streamlining the results of scientific research entails the necessity of the uniform understanding of terminology, accumulation of facts and insight of the development trend. We consider those issues on the example of “regression analysis model (recovery of the dependencies)” to form a unified methodological base for discussing various particular issues in this field. Four methods are considered. The models of the method of least squares with deterministic independent variable are singled out. Accord-

ing to the new paradigm of applied statistics, the distribution of deviations (errors, discrepancies) should be considered arbitrary, with one restriction, to obtain the limiting distributions of the estimates of parameters and dependencies, it is expedient to assume the fulfillment of conditions of the central limit theorem. The second basic type of probabilistic-statistical models of the method of least squares is based on a sample of random vectors. The dependence is nonparametric and distribution of the two-dimensional vector is arbitrary. Estimate of the variance of the independent variable can be considered only in a model based on a sample of random vectors, as well as the coefficient of determination as a criterion for the quality of the model. The issues of smoothing time series are discussed. Methods of reconstructing dependencies in spaces of general nature are considered. It is shown that the limiting distribution of the natural estimate of the dimensionality of the model is geometric, and construction of the informative subset of features comes across the effect of "inflation of the correlation coefficients." Different approaches to the regression analysis of interval data are discussed: the approach of confluent analysis becomes a thing of the past. An analysis of the variety of models of regression analysis leads to the conclusion that there is no single "standard model." Critical analysis of the hardened beliefs is necessary for competent development and application of mathematical methods of research, in particular, for transition to a modern paradigm of applied statistics.

Keywords: mathematical statistics; new paradigm of applied statistics; regression analysis; least-squares methods; nonparametric statistics; non-numerical statistics; estimation of the dimensionality of the model; statistics of interval data.

За столетия разработки математических методов исследования накоплен огромный массив научных результатов. Так, еще 30 лет назад мы оценивали [1] 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? Большинство — 10^3, отдельные наиболее продвинутые лица — 10^4, что на один-два порядка меньше, чем объем накопленных научных результатов.

Следовательно, необходимы работы по упорядочению накопленных научных результатов. Для успешной работы важно единообразное понимание терминов, знание фактов и тенденций развития. Обсудим эти вопросы на примере «модели регрессионного анализа (восстановления зависимостей)» в целях формирования единой методологической базы для обсуждения различных частных вопросов этой области.

Четыре метода восстановления зависимости

В простейшем случае имеется одна независимая t и одна зависимая x количественные переменные. 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. 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, оценить их можно разными способами. Так, графический метод состоит в следующем. Точки (t_i, x_i), $i = 1, 2, ..., n$, наносят на плоскость и проводят 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 точного 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 модулей и метод минимакса. Действительно, почему квадрат, а не другая степень? Однако используют и применяют именно метод наименьших квадратов. Почему в конкурентной борьбе победил именно этот метод? 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, [2]), в то время как оценки параметров для двух других методов могут быть найдены лишь с помощью численных алгоритмов [3]. 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. Имеется масса литературных источников, посвященных моделям регрессионного анализа (восстановления зависимостей), поэтому даем ссылки лишь на те публикации, которые необходимы по ходу изложения.

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) $e_1, e_2, ..., e_n$. Отличительная черта этого типа моделей состоит в том, что независимая переменная является детерминированной, а зависимая — случайной.

В базовой модели случайные величины $e_1, e_2, ..., e_n$ 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 [2, 4]. Поэтому согласно новой парадигме прикладной статистики [5] следует считать распределение случайных величин $e_1, e_2, ..., e_n$ произвольным, с одним ограничением — для получения предельных распределений оценок параметров и зависимости целесообразно предположить выполнение условий центральной предельной теоремы, точнее — условия Линдеберга [6].

Согласно [7] 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 [8]. Если в описании алгоритма регрессионного анализа используют распределения Стьюдента или Фишера, то необходимо констатировать, что распределения погрешностей предполагаются нормальными, следовательно, алгоритм не соответствует новой парадигме прикладной статистики.

Отметим, что при непараметрической модели погрешностей сама зависимость является параметрической. Как показано в дальнейшем, имеются много вариантов постановки задач непараметрической регрессии.

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

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

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) описывается 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. Возникающие в такой постановке проблемы рассмотрены в [9]. Отказ от независимости приводит к более слож-

ным 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.

Модели анализа случайных векторов

Второй тип вероятностно-статистических моделей основан на выборке случайных векторов. В таких моделях исходные данные в простейшем случае — двумерные случайные векторы $(x_i(\omega), y_i(\omega))$, $i = 1, 2, ..., n$, определенные на одном и том же вероятностном пространстве. В базовой модели все эти случайные векторы 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

$$p(y|x) = p(y|x(\omega) = x_0) = \frac{p(x, y)}{\int\limits_{-\infty}^{+\infty} p(x, y)\mathrm{d}y}.$$

Условное математическое ожидание, т.е. регрессионную зависимость y от x, можно записать как

$$f(x) = \int\limits_{-\infty}^{+\infty} yp(y|x)\mathrm{d}y = \frac{\int\limits_{-\infty}^{+\infty} yp(x, y)\mathrm{d}y}{\int\limits_{-\infty}^{+\infty} p(x, y)\mathrm{d}y}.$$

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

$$f_n(x) = \frac{\int\limits_{-\infty}^{+\infty} yp_n(x, y)\mathrm{d}y}{\int\limits_{-\infty}^{+\infty} p_n(x, y)\mathrm{d}y}$$

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, [2]), последней из которых является статья [10].

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 — непараметрическая и распределение двумерного вектора — произвольное. Как уже отмечалось, принимать гипотезу многомерной нормальности нет оснований. В некоторых случаях полезны параметрические модели зависимости, например,

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

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 коэффициент детерминации в качестве критерия качества модели могут привести к грубым ошибкам.

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 рас-

сматривают 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 [11].

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.

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 $x_k \in X$, $y_k \in Y$, $k = 1, 2, ..., n$. 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, $k = 1, 2, ..., n$. Пусть f_k — последовательность показателей различия в Y. При сделанных предположениях параметр θ естественно оценивать путем решения экстремальной задачи:

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

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 [12].

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

3. Имеется детерминированный набор пар (x_{k0}, y_{k0}), $k = 1, 2, ..., n$, результат наблюдения (x_k, y_k) является случайным элементом, распределение которого зависит от (x_{k0}, y_{k0}). Это постановка так называемого 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, с которой связано 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 [12]. 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 [12] для случайного вектора (ξ, η) со значениями в $X \times Y$ регрессией η на ξ относительно меры близости f естественно назвать решение задачи

$$Mf(g(\xi), \eta) \to \min_{g}, \qquad (5)$$

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

$$M(\eta | \xi = x) = \arg\min Mf(y, \eta(x)), y \in Y.$$

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

Если заранее известно, что условное математическое ожидание $M(\eta | \xi = x)$ 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. Их строят с помощью непараметрических оценок плотности [2, 10].

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

$$g(y | x) = \frac{g(x, y)}{\int_Y g(x, y)\nu_2(dy)} \qquad (6)$$

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

$$Mf(y, \eta(x)) = \int_Y f(y, a)g(a | x)\nu_2(da),$$

а потому

$$M(\eta | \xi = x) = \arg\min_{y\in Y} Mf(y, \eta(x)) =$$
$$= \arg\min_{y\in Y} \int_Y f(y, a)g(a | x)\nu_2(da).$$

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

$$g_n(y | x) = \frac{g_n(x, y)}{\int_Y g_n(x, y)\nu_2(dy)}. \qquad (7)$$

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

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

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

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 част-

ности, 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

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

где m — неизвестная степень полинома; a_0, a_1, a_2, ..., a_m — неизвестные коэффициенты многочлена; ε_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 $R^1 \cup R^2 \cup R^3 \cup ...$ 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 [13]). Рассмотрим некоторые из них.

Замечание. Здесь наглядно проявляется одна из причин живучести вероятностно-статистических моделей на основе нормального распределения. Такие модели, как правило, неадекватны 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:

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

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

$$\hat{\sigma}^2 = \frac{S}{n - 2},$$

поскольку число оцениваемых параметров $m + 1 = 2$.

Еще раз — почему при подборе вида модели знаменатель дроби, оценивающей остаточную дисперсию, приходится корректировать по числу 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 остаточной дисперсии

$$v(m) = \sigma^2(m)$$

в случае расширяющейся системы моделей (т.е. при возрастании натурального параметра m) выглядит так. Сначала наблюдаем заметное убывание. Затем оценка остаточной дисперсии ко-

леблется около некоторой константы (дисперсии погрешности).

Поясним ситуацию на примере модели восстановления зависимости, выраженной многочленом:

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

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

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

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.

$$m^* = \min\{m: \nu(m-1) > \nu(m),$$

$$\nu(m) \leq \nu(m+1)\}.$$

В работе [14] найдено предельное распределение этой оценки 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, ничтожно мала. Далее имеем:

$$P(m^* = m_0) \to 0{,}68268,$$

$$P(m^* = m_0 + 1) \to$$

$$\to 0{,}68268(1 - 0{,}68268) = 0{,}21663,$$

$$P(m^* = m_0 + 2) \to$$

$$\to 0{,}68268(1 - 0{,}68268)^2 = 0{,}068744,$$

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

$$- 0{,}68268)^3 = 0{,}021814...$$

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 — такое же, как в приведенной выше теореме, только значение параметра λ иное. Для степени многочлена давно предложены состоятельные оценки [15]. 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 $(y_i, \mathbf{X}_i)$, $i = 1, 2, ..., n$, где $\mathbf{X}_i = (x_{i1}, x_{i2}, ..., x_{iN}) \in R^N$ — вектор предикторов (факторов, объясняющих переменных), а модель такова:

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

где K — некоторое подмножество множества $\{1, 2, ..., n\}$; 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

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

В модели (8) имеется естественный порядок ввода предикторов в рассмотрение — в соответствии с возрастанием степени многочлена, а в модели (9) такого 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 значи-

мых 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 — 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 [16].

Эффект «вздувания коэффициентов корреляции». 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. Как правило, при этом игнорируют эффект «вздувания коэффициентов корреляции». Это явление обнаружил А. Н. Колмогоров в 1933 г. [17]. 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.

Как отмечено в [16], актуальность работы А. Н. Колмогорова [17] 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 [18].

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 «истин-

ных» данных (t_i, x_i), $i = 1, 2, ..., n$, наблюдают данные с погрешностями (q_i, y_i), $i = 1, 2, ..., n$, где $q_i = t_i + \varepsilon_i$, $y_i = x_i + \delta_i$. Здесь ε_i и δ_i — погрешности измерений (наблюдений, регистрации, опытов, анализов). Требуется восстановить зависимость между «истинными» переменными t и x.

К решению этой задачи есть несколько подходов. Если заданы ограничения на значения погрешностей, наложенных на случайные величины, то плодотворен подход статистики интервальных данных [19]. Восстановлению линейной зависимости в соответствии с подходом статистики интервальных данных посвящена статья [20], подробному изложению статистики интервальных данных — развернутые главы в монографиях [2, 12, 21, 22].

Подход А. П. Вощинина, Н. В Скибицкого и их сподвижников исходит непосредственно из анализа интервальных данных, без использования вероятностно-статистических моделей. Этот подход отражен, например, в работах [23 – 28].

Уходит в прошлое подход так называемого 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 адекватен реальным ситуациям и потому не имеет практических перспектив. Точно так же 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 (например, для проверки адекватности линейной модели, когда альтернативой является квадратическая).

Таким образом, 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 все 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.

Из сказанного следует, что определение понятия «регрессионный анализ» и его содержание целесообразно обсудить подробно на страницах нашего журнала. Критический разбор устоявшихся взглядов необходим для квалифицированного развития и применения математических методов исследования, в частности, для перехода на современную парадигму прикладной статистики [5].

ЛИТЕРАТУРА

1. **Орлов А. И.** Первый Всемирный конгресс Общества математической статистики и теории вероятностей им. Бернулли / Заводская лаборатория. 1987. Т. 53. № 3. С. 90 – 91.
2. **Орлов А. И.** Прикладная статистика. — М.: Экзамен, 2006. — 671 с.
3. **Тырсин А. Н., Максимов К. Е.** Оценивание линейных регрессионных уравнений с помощью метода наименьших модулей / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 7. С. 65 – 71.
4. **Орлов А. И.** Часто ли распределение результатов наблюдений является нормальным? / Заводская лаборатория. 1991. Т. 57. № 7. С. 64 – 66.
5. **Орлов А. И.** Новая парадигма прикладной статистики / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 1. Ч. I. С. 87 – 93.
6. **Гнеденко Б. В.** Курс теории вероятностей. Изд. 8-е, испр. и доп. — М.: Едиториал УРСС, 2005. — 448 с.
7. **Орлов А. И.** Структура непараметрической статистики (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2015. Т. 81. № 7. С. 62 – 72.
8. **Королюк В. С., Портенко Н. И., Скороход А. В., Турбин А. Ф.** Справочник по теории вероятностей и математической статистике. — М.: Наука, 1985. — 640 с.
9. **Копаев Б. В.** В методе наименьших квадратов надо заменить абсолютные отклонения относительными / Заводская лаборатория. Диагностика материалов. 2012. Т. 78. № 7. С. 76 – 76.
10. **Орлов А. И.** Асимптотика оценок плотности распределения вероятностей / Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета. 2017. № 131. С. 845 – 873.
11. **Орлов А. И.** Непараметрический метод наименьших квадратов с периодической составляющей / Заводская лаборатория. Диагностика материалов. 2014. Т. 80. № 1. С. 65 – 75.
12. **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.
13. **Себер Дж.** Линейный регрессионный анализ. — М.: Мир, 1980. — 456 с.
14. **Орлов А. И.** Асимптотика некоторых оценок размерности модели в регрессии / Прикладная статистика. Ученые записки по статистике. Т. 45. — М.: Наука, 1983. С. 260 – 265.
15. **Орлов А. И.** Об оценивании регрессионного полинома / Заводская лаборатория. 1994. Т. 60. № 5. С. 43 – 47.
16. **Орлов А. И.** Методы поиска наиболее информативных множеств признаков в регрессионном анализе / Заводская лаборатория. Диагностика материалов. 1995. Т. 61. № 1. С. 56 – 58.
17. **Колмогоров А. Н.** К вопросу о пригодности найденных статистическим путем формул прогноза / Журнал геофизики. 1933. Т. 3. С. 78 – 82.
18. **Орлов А. И.** Проблема множественных проверок статистических гипотез / Заводская лаборатория. Диагностика материалов. 1996. Т. 62. № 5. С. 51 – 54.
19. **Орлов А. И.** Статистика интервальных данных (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2015. Т. 81. № 3. С. 61 – 69.
20. **Гуськова Е. А., Орлов А. И.** Интервальная линейная парная регрессия (обобщающая статья) / Заводская лаборатория. Диагностика материалов. 2005. Т. 71. № 3. С. 57 – 63.
21. **Орлов А. И.** Теория принятия решений. — М.: Экзамен, 2006. — 576 с.
22. **Орлов А. И., Луценко Е. В.** Системная нечеткая интервальная математика. — Краснодар: КубГАУ, 2014. — 600 с.
23. **Вощинин А. П.** Метод анализа данных с интервальными ошибками в задачах проверки гипотез и оценивания пара-

метров неявных линейно параметризованных функций / Заводская лаборатория. Диагностика материалов. 2000. Т. 66. № 3. С. 51 – 64.
24. **Вощинин А. П.** Интервальный анализ данных: развитие и перспективы / Заводская лаборатория. Диагностика материалов. 2002. Т. 68. № 1. С. 118 – 125.
25. **Вощинин А. П., Скибицкий Н. В.** Интервальный подход к выражению неопределенности измерений и калибровке цифровых измерительных систем / Заводская лаборатория. Диагностика материалов. 2007. Т. 73. № 11. С. 68 – 71.
26. **Скибицкий Н. В., Севальнев Н. В.** Интервальные модели в задачах оптимального управления с дифференциальными связями / Заводская лаборатория. Диагностика материалов. 2015. Т. 81. № 11. С. 73 – 80.
27. **Скибицкий Н. В.** Применение статистического подхода к построению прямых и обратных характеристик объектов / Заводская лаборатория. Диагностика материалов. 2016. Т. 82. № 11. С. 67 – 75.
28. **Скибицкий Н. В.** Построение прямых и обратных статических характеристик объектов по интервальным данным / Заводская лаборатория. Диагностика материалов. 2017. Т. 83. № 1. Ч. 1. С. 87 – 93.

REFERENCES

1. **Orlov A. I.** The First World Congress of the Bernoulli Society for Mathematical Statistics and Probability Theory / Zavod. Lab. 1987. Vol. 53. N 3. P. 90 – 91 [in Russian].
2. **Orlov A. I.** Applied statistics. — Moscow: Ékzamen, 2006. — 671 p. [in Russian].
3. **Tyrsin A. N., Maksimov K. E.** Estimation of the linear regression equations using the least-modules method / Zavod. Lab. Diagn. Mater. 2012. Vol. 78. N 7. P. 65 – 71 [in Russian].
4. **Orlov A. I.** How often the distribution of the results of observations is normal? / Zavod. Lab. 1991. Vol. 57. N 7. P. 64 – 66 [in Russian].
5. **Orlov A. I.** The new paradigm of Applied Statistics / Zavod. Lab. Diagn. Mater. 2012. Vol. 78. N 1. Part I. P. 87 – 93 [in Russian].
6. **Gnedenko B. V.** Course of the of Probability Theory. — Moscow: Editorial URSS, 2005. — 448 p. [in Russian].
7. **Orlov A. I.** Structure of nonparametric statistics (generalizing paper) / Zavod. Lab. Diagn. Mater. 2015. Vol. 81. N 7. P. 62 – 72 [in Russian].
8. **Korolyuk V. S., Portenko N. I., Skorokhod A. V., Turbin A. F.** Handbook on Probability Theory and Mathematical Statistics. — Moscow: Nauka, 1985. — 640 p. [in Russian].
9. **Kopaev B. V.** In the method of least squares, it is necessary to replace the absolute deviations with relative / Zavod. Lab. Diagn. Mater. 2012. Vol. 78. N 7. P. 76 – 76 [in Russian].
10. **Orlov A. I.** Asymptotics of estimates of the probability distribution density / Politem. Set. Élektr. Nauch. Zh. Kuban. Gos. Agrar. Univ. 2017. N 131. P. 845 – 873 [in Russian].
11. **Orlov A. I.** Nonparametric method of least squares with periodic component / Zavod. Lab. Diagn. Mater. 2014. Vol. 80. N 1. P. 65 – 75 [in Russian].
12. **Orlov A. I.** Organizational-economic modeling: Textbook. Part 1. Nonnumeric statistics. — Moscow: Izd. MGTU im. N. É. Baumana, 2009. — 542 p. [in Russian].
13. **Seber G. A. F.** Linear regression analysis. — New York – London – Sydney – Toronto: Jonn Wiley and Sons, 1977. — 456 p.
14. **Orlov A. I.** Asymptotics of some estimates of the dimensionality of the model in regression / Applied statistics. Scientific notes on statistics. Vol. 45. — Moscow: Nauka, 1983. P. 260 – 265 [in Russian].
15. **Orlov A. I.** Regression polynomial estimation / Zavod. Lab. 1994. Vol. 60. N 5. P. 43 – 47 [in Russian].
16. **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].
17. **Kolmogorov A. N.** To the question of the suitability of the predicted formulas found statistically / Zh. Geofiz. 1933. Vol. 3. P. 78 – 82 [in Russian].
18. **Orlov A. I.** The problem of multiple tests of statistical hypotheses / Zavod. Lab. Diagn. Mater. 1996. Vol. 62. N 5. P. 51 – 54 [in Russian].
19. **Orlov A. I.** Statistics of interval data (generalizing article) / Zavod. Lab. Diagn. Mater. 2015. Vol. 81. N 3. P. 61 – 69 [in Russian].
20. **Gus'kova E. A., Orlov A. I.** Interval linear pair regression (generalizing article) / Zavod. Lab. Diagn. Mater. 2005. Vol. 71. N 3. P. 57 – 63 [in Russian].
21. **Orlov A. I.** Decision theory. — Moscow: Ékzamen, 2006. — 576 p. [in Russian].
22. **Orlov A. I., Lutsenko E. V.** System fuzzy interval mathematics. — Krasnodar: KubGAU, 2014. — 600 p. [in Russian].
23. **Voshchinin A. P.** Method of data analysis with interval errors in problems of hypothesis testing and estimation of parameters of implicit linearly parametrized functions / Zavod. Lab. Diagn. Mater. 2000. Vol. 66. N 3. P. 51 – 64 [in Russian].
24. **Voshchinin A. P.** Interval analysis of data: development and prospects / Zavod. Lab. Diagn. Mater. 2002. Vol. 68. N 1. P. 118 – 125 [in Russian].
25. **Voshchinin A. P., Skibitskii N. V.** Interval approach to the expression of measurement uncertainty and calibration of digital measuring systems / Zavod. Lab. Diagn. Mater. 2007. Vol. 73. N 11. P. 68 – 71 [in Russian].
26. **Skibitskii N. V., Seval'nev N. V.** Interval models in optimal control problems with differential constraints / Zavod. Lab. Diagn. Mater. 2015. Vol. 81. N 11. P. 73 – 80 [in Russian].
27. **Skibitskii N. V.** The application of the statistical approach to the construction of forward and backward characteristics of objects / Zavod. Lab. Diagn. Mater. 2016. Vol. 82. N 11. P. 67 – 75 [in Russian].
28. **Skibitskii N. V.** Constructing forward and reverse static characteristics of objects by interval data / Zavod. Lab. Diagn. Mater. 2017. Vol. 83. N 1. Part 1. P. 87 – 93 [in Russian].