УДК 519.213:519.23

Адаптивное восстановление регрессионных зависимостей на основе полупараметрической оценки плотности случайной компоненты*

В.С. ТИМОФЕЕВ

Работа направлена на усиление адаптивных возможностей развиваемых автором алгоритмов оценивания параметров регрессионных моделей, основанных на использовании универсальных семейств распределений. Показано, что использование идей полупараметрического анализа дает возможность достаточно гибко подстраиваться к отклонениям фактического распределения случайной ошибки от постулируемого. Представленная вычислительная схема алгоритма оценивания параметров регрессионных моделей обеспечивают корректную работу в этих условиях.

Ключевые слова: уравнение регрессии, оценивание параметров, универсальные распределения, метод максимального правдоподобия, полупараметрическое оценивание.

ВВЕДЕНИЕ

Классические алгоритмы восстановления регрессионных зависимостей обеспечивают получение корректных и достаточно качественных результатов при справедливости хорошо известных предположений (гипотез) о свойствах случайной компоненты [4]. В этом смысле наиболее жестким является метод максимального правдоподобия, требующий априорной фиксации вида распределения случайной ошибки. Применение нормального распределения обеспечивает существенное упрощение процедуры поиска оценок, позволяя использовать аналитическое решение, приводящее к методу наименьших квадратов. Однако на практике такие ситуации крайне редки. Сегодня можно говорить о том, что нормальное распределение существует лишь как некий теоретический объект, иллюстрирующий способы построения классических статистических выводов.

В связи с этим автором развивается подход, основанный на использовании универсальных семейств распределений, а именно кривых Пирсона, обобщенного лямбда-распределения, устойчивых распределений [3, 9, 11]. Их преимущества весьма очевидны. Во-первых, это очень широкие классы распределений, они описывают огромное число практически реализуемых ситуаций, в том числе сильно засоренные выборки, при этом дисперсия рассматриваемых случайных величин может быть достаточно большой и даже бесконечной. Во-вторых, внутри таких семейств, как правило, содержатся многие хорошо известные классические законы распределения, такие как бета-распределение, гамма-распределение, распределение Стьюдента, нормальное распределение и др.

Очевидно, что чем более широким является взятый за основу базовый класс распределений, тем больше практически реализуемых ситуаций могут быть учтены при проведении анализа. С другой стороны, любое параметрическое семейство распределений (даже универсальное) имеет строго определенные границы варьирования форм, в которых далеко не всегда может быть адекватно представлено то или иное практически реализуемое распределение. Другими словами, ни одно из известных семейств распределений не претендует на описание всех возможных практически реализуемых ситуаций. Следовательно, нужны алгоритмы оценивания параметров регрессионных моделей, обеспечивающие возможность корректной работы при отклонении фактического распределения случайной компоненты от априорно постули-

* *Статья получена 15 февраля 2013 г.*
Работа выполнена при финансовой поддержке РФФИ в рамках научного проекта №13-07-00299а.

руемого, в том числе и универсального распределения. Первой попыткой построения такого алгоритма следует считать использование разложения Грама–Шарлье [2]. В отмеченной работе [2] речь шла об отклонении только от нормального распределения.

1. ПОСТАНОВКА ЗАДАЧИ И ОСНОВНЫЕ ПРЕДПОЛОЖЕНИЯ

Рассмотрим регрессионное уравнение вида

$$y = X\theta + \varepsilon, \tag{1}$$

где $X = \begin{bmatrix} f_1(x_{11}) & \cdots & f_p(x_{1p}) \\ \vdots & \ddots & \vdots \\ f_1(x_{N1}) & \cdots & f_p(x_{Np}) \end{bmatrix}$ – матрица значений регрессионных функций, имеющая полный столбцовый ранг, т. е. $rg(X) = p$, $\theta = (\theta_1,...,\theta_p)^T$ – вектор неизвестных параметров, подлежащих оцениванию, p – количество неизвестных параметров, N – количество проведенных экспериментов; $f_i(x)$ – известные действительные функции вещественного аргумента x, x_{ij} – заданные значения входных факторов в N наблюдениях, $y = (y_1,...,y_N)^T$ – вектор значений отклика, $\varepsilon = (\varepsilon_1,...,\varepsilon_N)^T$ – вектор ошибок наблюдений.

Будем предполагать, что ошибки ε_i наблюдений являются независимыми одинаково распределенными случайными величинами с унимодальной функцией плотности $\psi(x)$, для которых верно, что

$$E(\varepsilon_i) = 0,\ D(\varepsilon_i) = \sigma^2.$$

Задача состоит в том, чтобы по имеющимся исходным данным (значениям отклика и входных факторов) как можно точнее оценить вектор неизвестных параметров уравнения регрессии (1).

2. ПОЛУПАРАМЕТРИЧЕСКАЯ ОЦЕНКА ФУНКЦИИ ПЛОТНОСТИ

Основная идея состоит в соединении двух, казалось бы, абсолютно противоположных подходов: параметрического и непараметрического. В соответствии с первым для неизвестной функции плотности $\psi(x)$ необходимо зафиксировать некое параметрическое семейство распределений. В рамках данной работы предлагается взять некоторое универсальное семейство распределений

$$\{\Psi(x,\beta)\},$$

где β – вектор неизвестных параметров, подлежащих оцениванию.

Пусть для определенности это будет обобщенное лямбда-распределение [13]. Следует напомнить, что функция распределения случайной величины ξ, имеющей лямбда-распределение, зависит от четырех параметров $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)$ и определяется с точки зрения квантилей распределения следующим образом [13]:

$$Q(u, \beta_1, \beta_2, \beta_3, \beta_4) = \beta_1 + \frac{1}{\beta_2}\left[\frac{u^{\beta_3}}{\beta_3} - \frac{(1-u)^{\beta_4}}{\beta_4}\right],\ 0 \le u \le 1,$$

причем $\xi = Q(u, \beta_1, \beta_2, \beta_3, \beta_4)$. Для идентификации обобщенного лямбда-распределения используют различные методы, например, метод моментов [1].

Непараметрический подход предполагает полный отказ от параметрического представления распределения с целью обеспечения максимально возможной «подстройки» под практически реализуемые ситуации. Такая свобода дает большую гибкость в описании данных, но и лишает исследователя информации, получаемой в результате интерпретации значений оценок параметров. Наиболее известная непараметрическая оценка функции плотности предложена Розенблатом–Парзеном [15], которая имеет вид

$$\hat{\psi}(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - x_i}{h}\right), \tag{2}$$

где h – ширина ядра (окна сглаживания), $K(r)$ – функция ядра, в качестве которого могут выступать ядро Епанечникова, Гаусса, квадратичная функция и др.

Существует и другой способ получения непараметрической функции плотности [5, 8]. Он основан на использовании характеристической функции. Хорошо известно [5, 7], что характеристическая функция $\varphi(t)$ некоторой случайной величины ξ с плотностью $\psi(x)$ определяется следующим образом:

$$\varphi(t) = E\left[e^{itx}\right] = \int_{-\infty}^{\infty} e^{itx}\psi(x)\,dx,$$

где $t \in R$, $i = \sqrt{-1}$ – так называемая мнимая единица. Поскольку

$$\left|e^{itx}\right| = 1, \ \forall t \in R,$$

то характеристическая функция существует для любой действительной случайной величины. Данная функция содержит всю информацию о распределении случайной величины и обладает целым рядом важных свойств [5, 7].

На основе имеющейся реализации $x_1, ..., x_N$ случайной величины ξ можно определить выборочную оценку характеристической функции [12]:

$$\hat{\varphi}(t) = \frac{1}{N}\sum_{j=1}^{N} e^{itx_j} = \frac{1}{N}\sum_{j=1}^{N}\left(\cos(tx_j) + i\sin(tx_j)\right). \tag{3}$$

Отметим, что в соответствии с законом больших чисел [1] оценка (3) состоятельна.

Переход от характеристической функции к функции плотности осуществляется посредством преобразования Фурье [6]

$$\psi(x_j) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \varphi(t)e^{-itx_j}\,dt, \ j = 1, ..., N. \tag{4}$$

Искомая непараметрическая оценка $\hat{\psi}(x)$ получается после замены $\varphi(t)$ в (4) на ее эмпирический аналог (3) и замены интеграла (4) конечной суммой.

Имея одновременно эти две оценки для неизвестной функции плотности $\psi(x)$, можно получить некую компромиссную оценку [14]

$$g(x, \alpha, \hat{\beta}) = \alpha\psi(x, \hat{\beta}) + (1 - \alpha)\hat{\psi}(x), \tag{5}$$

где α – неизвестная величина ($0 \le \alpha \le 1$), которую следует оценить по имеющимся данным. В [14] предлагается ее оценивать стандартными методами идентификации распределений, например, методом максимального правдоподобия.

Такое представление позволяет достаточно гибко реагировать на отклонение фактически реализуемого распределения от рассматриваемого параметрического семейства $\{\Psi(x, \beta)\}$.

Очевидно, что случай $\alpha = 1$ соответствует адекватности выбранного параметрического семейства исходным данным, поскольку непараметрическая компонента в (5) не оказывает влияния на итоговую оценку, в то время как при $\alpha = 0$ восстановление распределения происходит только на основе непараметрической компоненты. Фактически по величине α можно делать выводы о степени пригодности рассматриваемого параметрического семейства к изучаемой ситуации.

Для иллюстрации данного факта был проведен ряд вычислительных экспериментов, основанных на технологии статистического моделирования. Будем моделировать значения независимых и одинаково распределенных случайных величин с функцией распределения вида

$$F(x) = (1-\lambda)F_1(x, m_1, \sigma_1) + \lambda F_2(x, m_2, \sigma_2),$$

где $F_i(x, m_i, \sigma_i)$ – функция нормального распределения с математическим ожиданием, равным m_i, и дисперсией σ_i^2, $i = 1,2$, $\lambda \in [0,1]$ – параметр смеси. Во всех проведенных вычислительных экспериментах $m_1 = m_2 = 0$.

Такое представление позволяет моделировать выборки с различной степенью отклонения от нормального распределения, в том числе появление в них отдельных, довольно грубых засоряющих наблюдений – «выбросов». Параметр λ определяет соответствующие доли наблюдений с дисперсиями σ_1^2 и σ_2^2 в выборке. Очевидно, что при $\lambda = 0$ и $\lambda = 1$ имеет место нормальное распределение. В проведенных вычислительных экспериментах полагалось $\sigma_2^2 = 10\sigma_1^2$.

В таблице представлены усредненные по 1000 вычислительным экспериментам значения параметра α из (5) по методу максимального правдоподобия в зависимости от степени загрязненности выборки (уровня выбросов) и ширины окна h (см. (2)). В качестве параметрического семейства рассматривалось стандартное нормальное распределение, в качестве непараметрической компоненты использовалось соотношение (2) с ядром Гаусса. Объем выборки составил 200 элементов.

Таблица

Результаты оценивания параметра α при разной доле выбросов и ширине окна

Ширина окна, *h*	Доля выбросов, λ					
	0.00	0.005	0.01	0.05	0.1	0.2
0.4	0.809	0.890	0.936	0.999	0.999	0.999
0.5	0.407	0.649	0.820	0.998	0.999	0.999
1.0	0.043	0.232	0.417	0.975	0.999	0.999
2.0	0.005	0.038	0.071	0.672	0.974	0.999

Из таблицы видно, что оптимальные значения параметра α восстановленного распределения зависят от доли выбросов и ширины окна сглаживания. Первый факт достаточно очевиден, поскольку с увеличением процента аномальных наблюдений в выборке меняется форма распределения случайной компоненты, которая очень быстро выходит за рамки базового нормального распределения. Алгоритм практически полностью переключается на использование только непараметрической оценки уже при 10 % выбросов. Второй факт весьма интересен, поскольку ширина окна определяет свойства восстановленной функции плотности, а также точность полученных на ее основе оценок параметров регрессионных уравнений. Более подробную информацию об этом можно найти в [10], здесь лишь отметим, что малые значения h при оценивании параметров регрессионных уравнений на основе метода максимального правдоподобия следует использовать очень осторожно.

3. АЛГОРИТМ ОЦЕНИВАНИЯ ПАРАМЕТРОВ РЕГРЕССИОННЫХ МОДЕЛЕЙ

Возвращаясь к задаче восстановления исходного регрессионного уравнения (1), можно предложить новый адаптивный алгоритм, который будет основан на методе максимального правдоподобия. Поскольку речь идет о совместном использовании универсальных семейств распределений и непараметрических оценок, то предположительно данный алгоритм будет обладать очень широкой областью применения. В силу предположений о независимости случайных ошибок и истинности структуры рассматриваемого регрессионного уравнения (1) значения остатков $e_i = y_i - x_i\hat{\theta}$ (x_i – i-я строка матрицы X из (1)) также будут статистически независимыми случайными величинами с плотностью распределения $\psi(z_i, \theta)$, вместо которой будем использовать оценку (5). Тогда для оценивания параметров уравнения (1) можно воспользоваться методом максимального правдоподобия [5]. Учитывая тот факт, что остатки наблюдаемы, т.е. их значения определяются на основе имеющихся исходных данных, запишем логарифмическую функцию правдоподобия

$$l\left(e_1,...,e_N,\hat{\theta}\right) = \ln\left(\prod_{i=1}^{N} g\left(e_i,\alpha,\hat{\beta},\hat{\theta}\right)\right) = \sum_{i=1}^{N} \ln\left(g\left(e_i,\alpha,\hat{\beta},\hat{\theta}\right)\right). \qquad (6)$$

Итерационный алгоритм оценивания неизвестных параметров уравнения регрессии состоит в следующем.

Шаг 1. Определение начального приближения ($k := 0$) вектора неизвестных параметров уравнения (1), в качестве которого можно использовать оценку метода наименьших квадратов, что по сравнению с произвольным начальным приближением позволит сократить число итераций и время вычислений.

Шаг 2. Вычисление остатков регрессионного уравнения.

Шаг 3. По значениям остатков e_i проводится идентификация параметрического распределения $\psi(z_i, \beta)$ любым методом (моментов, максимального правдоподобия и др.).

Шаг 4. По значениям остатков e_i проводится построение непараметрической оценки функции плотности $\hat{\psi}(x)$ на основе соотношения (2) или (4).

Шаг 5. Используя метод максимального правдоподобия, провести выбор оптимального значения α в (5) на основе оценок параметрической и непараметрической частей, полученных на шаге 3 и 4 соответственно.

Шаг 4. Вычисление значения найденной оценки функции плотности $g\left(e_i,\alpha,\hat{\beta},\hat{\theta}\right)$ в точках, соответствующих имеющимся значениям остатков e_i.

Шаг 5. Вычисление значения логарифмической функции правдоподобия (6).

Шаг 6. Поиск очередного значения оценки неизвестных параметров $\hat{\theta}^{k+1}$

$$\hat{\theta}^{k+1} = \arg \max_{\theta} l\left(e_1, e_2, ..., e_N, \hat{\theta}^k\right).$$

Шаг 7. Если $\left\|\hat{\theta}^{k+1} - \hat{\theta}^k\right\| < \varepsilon$, то завершение процесса, в противном случае $k := k+1$ и переход на шаг 2 (ε – заданная погрешность вычисления).

Очевидным преимуществом данного алгоритма является уже отмеченная гибкость, а недостатком достаточно сложная вычислительная схема, предполагающая вложенное использование метода максимального правдоподобия, что естественно сказывается на времени вычислений.

ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача адаптивного оценивания параметров регрессионных зависимостей. Для решения данной задачи предлагается новый универсальный алгоритм, использующий полупараметрическую оценку функции плотности случайной компоненты регресси-

онного уравнения. Использование универсальных семейств распределений обеспечивает возможность построения оценок максимального правдоподобия для большого числа практических ситуаций. Когда фактически реализуемое распределение плохо описывается в рамках выбранного семейства, алгоритм автоматически переключается на непараметрическую оценку, что позволяет его рекомендовать к применению практически в любых ситуациях.

СПИСОК ЛИТЕРАТУРЫ

[1] **Гихман И.И.** Теория вероятностей и математическая статистика / И.И. Гихман, А.В. Скороход, М.И. Ядренко. – Киев, 1979. – 408 с.

[2] **Денисов В.И.** Оценивание параметров регрессионных зависимостей с использованием аппроксимации Грама-Шарлье / В.И. Денисов, В.С. Тимофеев // Автометрия. – Новосибирск: Изд-во СО РАН, 2008.– Т. 44. – № 6. – С. 3–12.

[3] **Денисов В.И.** Устойчивые распределения и оценивание параметров регрессионных зависимостей / В.И. Денисов, В.С. Тимофеев // Известия Томского политехнического университета. – Томск: Изд-во ТПУ. – 2011. – Т. 318, № 2. – С. 10–15.

[4] **Дрейпер Н.** Прикладной регрессионный анализ / Н. Дрейпер, Н. Смит. – М.: Статистика, 1973. – 392 с.

[5] **Кендалл М.** Теория распределений / М. Кендалл, А. Стьарт. – М.: Наука, 1966. – 587 с.

[6] **Оппенгейм А.В.** Цифровая обработка сигналов / А.В. Оппенгейм, Р.В. Шафер. – М.: Связь, 1979. – 416 с.

[7] **Пугачев В.С.** Теория вероятностей и математическая статистика / В.С. Пугачев. – М.: Наука, 1979. – 496 с.

[8] **Тимофеев В.С.** Оценивание параметров регрессионных зависимостей на основе характеристической функции / В.С. Тимофеев // Научный вестник НГТУ. – 2010. – № 2(39). – Новосибирск: НГТУ. – С. 43–52.

[9] **Тимофеев В.С.** Оценивание параметров регрессионных зависимостей с использованием кривых Пирсона. Ч.1 / В.С. Тимофеев // Научный вестник НГТУ. – 2009. – № 4(37). – Новосибирск: Изд-во СО РАН. – С. 57–66.

[10] **Тимофеев В.С.** Ядерные оценки плотности при идентификации уравнений регрессии / В.С. Тимофеев // Научный вестник НГТУ. – 2010. – № 3(40). – Новосибирск: Изд-во СО РАН. – С. 41–50.

[11] **Тимофеев В.С.** Адаптивное оценивание параметров регрессионных моделей с использованием обобщенного лямбда-распределения / В.С. Тимофеев, Е.А. Хайленко // Доклады академии наук высшей школы РФ. – 2010. – № 2(15). – Новосибирск: Изд-во НГТУ. – С. 25–36.

[12] **Feuerverger A.** The emperical characteristic function and its applications / A. Feuerverger, R.A. Mureika // The annals of statistics. – 1977. – Vol. 5. – № 1. – P. 88–97.

[13] **Karian Z.A.** Fitting statistical distributions: the Generalized Lambda Distribution and Generalized Bootstrap methods / Z.A. Karian, E.J. Dudewicz // New York, CRC Press LLC, 2000 – 435 p.

[14] **Olkin I.** A semiparametric approach to density estimation / I. Olkin, C.H. Spiegelman // Journal of the American statistical association. – 1987. – Vol. 82, № 399. – P. 858–865.

[15] **Pagan A.** Nonparametric econometrics / A. Pagan, A. Ullah. – New York, 1999. – 424 p.

REFERENCES

[1] Gihman I.I., Skorohod A.V., Jadrenko M.I. Teorija verojatnostej i matematicheskaja statistika. – Kiev, 1979. – 408 s.

[2] Denisov V.I., Timofeev V.S. Ocenivanie parametrov regressionnyh zavisimostej s ispol'zovaniem approksimacii Grama-Sharl'e // Avtometrija. – Novosibirsk: Izd-vo SO RAN, 2008. – T. 44, № 6, S. 3–12.

[3] Denisov V.I., Timofeev V.S. Ustojchivye raspredelenija i ocenivanie parametrov regressionnyh zavisimostej // Izvestija Tomskogo politehnicheskogo universiteta. – Tomsk: Izd-vo TPU, 2011. – T. 318, №2. – S. 10–15.

[4] Drejper N., Smit N. Prikladnoj regressionnyj analiz. – M.: Statistika, 1973. – 392 s.

[5] Kendall M., St'art A. Teorija raspredelenij. – M.: Nauka, 1966. – 587 s.

[6] Oppengejm A.V., Shafer R.V. Cifrovaja obrabotka signalov. – M.: Svjaz', 1979.

[7] Pugachev V.S. Teorija verojatnostej i matematicheskaja statistika. – M.: Nauka,1979. – 496 s.

[8] Timofeev V.S. Ocenivanie parametrov regressionnyh zavisimostej na osnove harakteristicheskoj funkcii // Nauchn. vestnik NGTU. – Novosibirsk: NGTU. – 2010. – N 2(39). – S. 43–52.

[9] Timofeev V.S. Ocenivanie parametrov regressionnyh zavisimostej s ispol'zovaniem krivyh Pirsona. Ch.1 // Nauchn. vestnik NGTU. – Novosibirsk: Izd-vo SO RAN. 2009. – N 4(37). – S. 57–66.

[10] Timofeev V.S. Jadernye ocenki plotnosti pri identifikacii uravnenij regressii // Nauchn. vestnik NGTU. – Novosibirsk: Izd-vo SO RAN, 2010. – N 3(40). –S. 41–50.

[11] Timofeev V.S., Hajlenko E.A. Adaptivnoe ocenivanie parametrov regressionnyh modelej s ispol'zovaniem obobshhennogo ljambda – raspredelenija // Doklady akademii nauk vysshej shkoly RF. – Novosibirsk: Izd-vo NGTU. 2010. – N2(15). – S. 25–36.

[12] Feuerverger A., Mureika R.A. The emperical characteristic function and its applications // The annals of statistics. – Vol. 5, N.1, 1977. – P. 88–97.

[13] Karian Z.A., Dudewicz E.J. Fitting statistical distributions: the Generalized Lambda Distribution and Generalized Bootstrap methods. – New York: CRC Press LLC, 2000. – 435 p.

[14] Olkin I., Spiegelman C.H. A semiparametric approach to density estimation // Journal of the American statistical association. – Vol. 82, № 399. – P. 858–865.

[15] Pagan A., Ullah A. Nonparametric econometrics. – New York, 1999.

Тимофеев Владимир Семенович, доктор технических наук, профессор кафедры программных систем и баз данных Новосибирского государственного технического университета. Основное направление научных исследований – разработка и исследование устойчивых методов и алгоритмов анализа многофакторных объектов, в том числе с использованием непараметрической статистики. Имеет более 75 публикаций, в том числе один учебник. E-mail: netsc@fpm.ami.nstu.ru

V.S. Timofeev

Adaptive construction of regression models based on semiparametric estimation of disturbance density function

The parameters estimation problem of regression models is considered. The improvement of adaptive features for original author's algorithms for estimation of regression models parameters is carried out. Base of these algorithms consists in using universal distributions, which can to describe a wide area of practice situations. In this paper author uses one class of universal distributions namely generalized lambda distribution. It shows that the semiparametric estimation concept permits very flexible adjustment to small deviations of real distribution of errors from postulate. The parametric part is density function of generalized lambda distribution and nonparametric part is kernel based function (in this study gauss kernel). The calculation sheme of new algorithm supports correct estimation in these conditions. The statistical simulation results should be considered as the basis for algorithm parameters settings.

Key words: regression equation, parameters estimation, universal distributions, maximum likelihood method, semiparametric estimation.