ISSN 1814-1196
Научный вестник НГТУ
том 60, № 3, 2015, с. 58–68

http://journals.nstu.ru/vestnik
Science Bulletin of the NSTU
Vol. 60, No. 3, 2015, pp. 58–68

УДК 519.23

Классификация регрессионных моделей на основе объема априорной информации*

В.С. ТИМОФЕЕВ[1], А.В. ФАДДЕЕНКОВ[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, доктор технических наук, доцент. E-mail: v.timofeev@corp.nstu.ru*

[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат технических наук, доцент. E-mail: faddeenkov@corp.nstu.ru*

В данной работе авторами предложена классификация регрессионных моделей. Основанием классификации служит объем априорной информации, доступной исследователю. При этом рассматриваются две основные задачи спецификации модели, с которыми сталкиваются исследователи. С одной стороны, используется информация о структуре регрессионной модели. Выделено три основных уровня: полная определенность, частичная неопределенность и полная неопределенность структуры модели. В первом случае предполагается, что структура модели априорно задана с точностью до неизвестных параметров. Во втором случае структура модели известна не полностью и недостающая часть будет компенсироваться непараметрической составляющей. В третьем случае структура модели не известна, что влечет использование только непараметрических методов.

В качестве второго базиса классификации предлагается использовать полноту априорной информации о распределении случайной составляющей модели. Здесь также предлагается выделять три уровня. Первый уровень соответствует случайным ошибкам с известным (с точностью до параметров) законом распределения. Второй уровень соответствует структурированным ошибкам, образованным линейными комбинациями или смесями случайных величин. Третий уровень является наиболее общим и соответствует отсутствию информации о структуре и распределении случайной составляющей. Различные комбинации уровней информированности о структуре модели и распределении случайной ошибки определяют девять основных групп моделей, каждой из которых соответствуют свои методы идентификации. В простейшем случае рассматриваются классические методы регрессионного анализа, основанные на использовании метода наименьших квадратов. При неполной информации рекомендуется использовать полупараметрические методы, основанные на сплайновой регрессии, моделях структурированной ошибки, универсальных распределениях.

Ключевые слова: регрессионная модель, структура модели, классификация, априорная информация, параметрические методы, непараметрические методы, полупараметрические методы, структурированная ошибка, компоненты дисперсии

DOI: 10.17212/1814-1196-2015-3-58-68

* *Статья получена 1 июля 2015 г.*

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований в рамках научного проекта № 13-07-00299 а.

ВВЕДЕНИЕ И ПОСТАНОВКА ЗАДАЧИ

Регрессионные модели являются одним из наиболее популярных инструментов прикладного статистического анализа. Решение задачи построения регрессионных моделей возможно проводить на основе множества различных подходов и методов. Как правило, эти различия связаны с объемом априорной информации, доступной исследователю на этапе постановки задачи. Естественно предположить, что в каждой конкретной ситуации можно подобрать наиболее подходящие методы решения. Однако на текущий момент нет четкого понимания границ эффективного применения тех или иных методов. В связи с этим авторами была предпринята попытка классификации регрессионных моделей, основанная на объеме априорной информации о структуре модели и свойствах случайной ошибки.

Рассмотрим в общем виде задачу восстановления регрессионной зависимости

$$y = f(x) + \varepsilon, \tag{1}$$

где y – наблюдаемый отклик; $f(x)$ – неизвестная функция; x – вектор значений входных факторов; ε – случайная составляющая. Задача состоит в том, чтобы на основе имеющихся наблюдений за переменными x и y наилучшим образом построить оценку функции $f(x)$.

Очевидно, что в зависимости от объема доступной априорной информации о свойствах функции $f(x)$ и случайной составляющей ε для решения данной задачи могут быть использованы различные методы. В частности, в классическом регрессионном анализе предполагается, что $f(x)$ задана с точностью до неизвестных параметров, т. е. $f(x) = f_0(x, \theta)$. Это приводит к регрессионному уравнению следующего вида:

$$y_i = f_0(x_i, \theta) + \varepsilon_i, \tag{2}$$

где y_i – значение отклика в i-м наблюдении ($i = 1, 2, ..., N$); x_i – значение вектора входных факторов в i-м наблюдении; θ – вектор неизвестных параметров; ε_i – случайная ошибка в i-м наблюдении.

Для модели (2) задача построения функции $f(x)$ сводится к задаче оценивания вектора неизвестных параметров θ. Наиболее известным методом оценивания θ является метод наименьших квадратов (МНК), для корректного применения которого необходимо, чтобы все ошибки ε_i были независимыми и имели одинаковое распределение с нулевым средним и дисперсией σ_ε^2: $\varepsilon_i \sim \left(0, \sigma_\varepsilon^2\right)$, $i = 1, 2, ..., N$ [3].

Однако, на практике исследователь может не обладать полной информацией о структуре модели $f(x)$ и сведение модели к виду (2) с последующим использованием МНК не представляется возможным. Аналогичная ситуация возникает при анализе справедливости предположений о свойствах случайной ошибки. В связи с этим возникает задача анализа возможных уровней неопределенности в модели (1) и привязки к этим уровням соответствующих методов оценивания.

1. ОСНОВАНИЯ КЛАССИФИКАЦИИ

С одной стороны, рассмотрим доступную исследователю информацию о структуре модели $f(x)$. При этом выделим три основных уровня: полную определенность структуры модели, заданной с точностью до неизвестных параметров; частичную неопределенность – структура модели известна не полностью; полную неопределенность структуры модели. С другой стороны, были рассмотрены варианты информированности исследователя о распределении и структуре случайной составляющей модели (1). Здесь также выделим три уровня.

Первый уровень соответствует случайным ошибкам с известным (с точностью до параметров) законом распределения. Второй уровень соответствует структурированным ошибкам, образованным линейными комбинациями или смесями случайных величин. Третий уровень является наиболее общим и соответствует отсутствию информации о структуре и распределении случайной составляющей.

Различные комбинации уровней информированности о структуре модели и распределении случайной ошибки определяют девять основных групп моделей, каждой из которых соответствуют свои методы идентификации (см. таблицу).

Уровни неопределенности структуры модели и ошибок

		Модель		
		Известна (A)	Частично известна (B)	Не известна (C)
Ошибка	Простая ошибка (A)	(AA)	(AB)	(AC)
	Структурированная ошибка (B)	(BA)	(BB)	(BC)
	Неизвестная ошибка (C)	(CA)	(CB)	(CC)

2. ЭЛЕМЕНТЫ КЛАССИФИКАЦИИ

Ситуация (AA). Предполагается, что структура модели (2) известна с точностью до неизвестных параметров и имеется информация о виде распределения случайных ошибок. Классическим методом оценивания в данном случае является широко известный метод максимального правдоподобия (ММП) [4]. При нормальном распределении ММП-оценки совпадают с МНК-оценками. Кроме того, если справедливы условия теоремы Гаусса–Маркова, то оценки метода наименьших квадратов являются наилучшими (в смысле минимума дисперсии оценок) среди всех линейных несмещенных оценок. Это обстоятельство позволяет рекомендовать МНК даже для тех случаев, когда распределение случайных ошибок неизвестно. Недостатком данного метода является слабая устойчивость оценок к нарушениям предположений

теоремы Гаусса–Маркова [3, 4]. Решением проблемы может служить переход к специальным методам оценивания [13–15], в том числе разработанным авторами [2, 5, 6, 9, 11].

Ситуация (АВ). В данном случае относительно ошибки уровень определенности такой же, как в ситуации (АА), но при этом структура регрессионной модели известна только частично. Другими словами, вместо уравнения (2) следует рассматривать уравнение вида

$$y = f_0(x,\ \theta) + \eta(x) + \varepsilon, \tag{3}$$

где $\eta(x)$ – неизвестная компонента модели.

Такие модели получили название полупараметрических и в последнее время завоевывают все большую популярность среди исследователей [20]. При этом $f_0(x,\ \theta)$ называют параметрической частью модели, а $\eta(x)$ – непараметрической частью. Существуют различные подходы к идентификации модели (3), отличающиеся способами восстановления непараметрической части. Одним из вариантов полупараметрической регрессии является сплайновая регрессия

$$y = f_0(x,\ \theta) + \beta_1 f_1 + \beta_2 f_2 + ... + \beta_k f_k + \varepsilon, \tag{4}$$

где $\eta(x) = \beta_1 f_1 + \beta_2 f_2 + ... + \beta_k f_k$ – непараметрическая часть; $f_1,\ f_2,...,\ f_k$ – специальные базисные функции; $\beta_1,...,\ \beta_k$ – неизвестные параметры.

Преимущество данного подхода заключается в том, что сведение модели к виду (4) позволяет использовать известные методы классического регрессионного анализа, включая МНК. Однако, в зависимости от цели исследования в рамках данного подхода возможны различные варианты.

В частности, в случае линейной параметризации функции $f_0(x,\ \theta)$ полупараметрическая регрессионная модель может быть представлена в следующем виде:

$$y_i = \theta_1 x_{i1} + \ldots + \theta_m x_{im} + \beta_1 f_{i1} + \beta_2 f_{i2} + ... + \beta_k f_{ik} + \varepsilon_i, \tag{5}$$

где y_i – значение отклика в i-м наблюдении ($i = 1,\ 2,...,\ N$); x_{ij} – значение j-го регрессора в i-м наблюдении ($j = 1,\ 2,...,\ m$); $\theta_1,...,\ \theta_m$, $\beta_1,...,\ \beta_k$ – неизвестные параметры; $f_{i1},\ f_{i2},...,\ f_{ik}$ – значения базисных функций в i-м наблюдении; ε_i – случайная ошибка в i-м наблюдении ($i = 1,\ 2,...,\ N$).

Естественно, качество воспроизведения исходной зависимости напрямую зависит от количества и вида базисных функций. Однако излишнее усложнение модели может приводить к излишней подгонке линии регрессии под исходные данные.

Данная проблема традиционно решается дополнительным сглаживанием модели с переходом к так называемым «штрафным сплайнам» [16, 20]. Идея этого метода заключается в том, что для снижения излишнего влияния непараметрической части на ее параметры налагается ограничение (штраф) и вектор оценок параметров вычисляется следующим образом:

$$\hat{\Theta} = (X^T X + \lambda^2 D)^{-1} X^T Y, \tag{6}$$

где $\Theta = \begin{bmatrix} \theta_1 & \cdots & \theta_m & \beta_1 & \cdots & \beta_k \end{bmatrix}^T$, $Y = \begin{bmatrix} y_1 & y_2 & \cdots & y_N \end{bmatrix}^T$,

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1m} & f_{11} & \cdots & f_{1k} \\ x_{21} & \cdots & x_{2m} & f_{21} & \cdots & f_{2k} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{Nm} & f_{N1} & \cdots & f_{Nk} \end{bmatrix},$$

λ^2 – параметр сглаживания, D – $(m+k)\times(m+k)$-матрица штрафа:

$$D = \begin{bmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \cdots & 0 & 1 \end{bmatrix} = \begin{bmatrix} 0_{(m\times m)} & 0_{(m\times k)} \\ 0_{(k\times m)} & I_{(k\times k)} \end{bmatrix}.$$

При $\lambda^2 = 0$ сглаживание непараметрической части не проводится и оценка (6) совпадает с обычной МНК-оценкой. Чрезмерное же увеличение параметра сглаживания ($\lambda^2 \to +\infty$) приводит к тому, что регрессионная модель (5) вырождается в модель, состоящую только из параметрической части. В связи с этим выбору величины параметра сглаживания следует уделять особое внимание.

Следует отметить, что спецификация модели (4) напрямую зависит от цели исследования и интерпретации непараметрической компоненты. В отдельных случаях полезно переходить к рассмотрению неизвестных параметров $\beta_1,...,\beta_k$ как случайных величин, что приводит к построению модели со структурированной ошибкой или к ситуации (ВА).

Ситуация (ВА). Предполагается, что случайная составляющая модели обладает внутренней структурой и может быть представлена в виде линейной комбинации случайных величин

$$\varepsilon = \xi_1\nu_1 + \xi_2\nu_2 + ... + \xi_m\nu_m,$$

где ν_1, ν_2,..., ν_m – известные значения переменных, соответствующих случайным эффектам ξ_1, ξ_2,..., ξ_m.

Одним из простейших вариантов подобных моделей является модель стохастического фронта [17]:

$$y = f_0(x, \theta) + \varepsilon, \qquad (7)$$

где $\varepsilon = \xi_1 - \xi_2$, ξ_1 – случайная величина, аналогичная классической ошибке модели (2), часто имеющая нормальное распределение; ξ_2 – неотрицательная случайная величина. В зависимости от предположений о виде распределения величины ξ_2 для идентификации модели (7) используются различные варианты метода максимального правдоподобия.

Другим частным случаем является модель компонент дисперсии [19], в которой предполагается, что случайная составляющая зависит от некоторых случайных факторов

$$Y = X\Theta + e, \quad e = U_1\xi_1 + ... + U_r\xi_r + \varepsilon, \tag{8}$$

где ξ_i, $i = 1, \ldots, r$ – векторы эффектов случайных факторов; U_i, $i = 1, \ldots, r$ – известные матрицы значений переменных, соответствующих r случайным факторам; $\varepsilon = (\varepsilon_1, ..., \varepsilon_N)^T$ – вектор случайных ошибок.

При этом предполагается, что

$$\xi_i \sim \left(0, \sigma_i^2 I_{m_i}\right), \; i = 1, \ldots, r,$$

$$\operatorname{cov}(\xi_i, \xi_j) = 0, \; i \neq j,$$

$$\operatorname{cov}(\xi_i, \varepsilon) = 0, \; \varepsilon \sim \left(0, \sigma_\varepsilon^2 I\right),$$

где m_i – число уровней i-го случайного фактора.

Или в более сжатой форме

$$e \sim (0, \sigma_1^2 V_1 + ... + \sigma_r^2 V_r + \sigma_\varepsilon^2 I), \; V_i = U_i U_i^T, \; i = 1, \ldots, r.$$

Величины $\sigma_1^2, \ldots, \sigma_r^2, \sigma_\varepsilon^2$ получили название «компоненты дисперсии», для их оценивания разработано множество методов [12, 19, 21]. Зная оценки компонент дисперсии, идентификацию модели можно провести, например, с помощью обобщенного МНК:

$$\hat{\Theta} = (X^T \Omega^{-1} X)^- X^T \Omega^{-1} Y, \text{ где } \Omega = \sum_{i=1}^{r} \frac{\hat{\sigma}_i^2}{\hat{\sigma}_\varepsilon^2} U_i U_i^T + I.$$

Примером использования модели со структурированной ошибкой может послужить модель, рассмотренная в разделе 2 настоящего отчета.

Ситуация (АС). При полном отсутствии какой-либо информации о структуре модели (1) восстановление регрессионной зависимости возможно только с использованием непараметрических методов [18]. Наиболее известной и относительно простой оценкой отклика является ядерная оценка Надарая–Уотсона следующего вида:

$$\hat{y} = \hat{f}(x) = \frac{\sum_{i=1}^{N} y_i K\left(\frac{x_i - x}{h_x}\right)}{\sum_{i=1}^{N} K\left(\frac{x_i - x}{h_x}\right)}, \tag{9}$$

где $K(u)$ – функция ядра, h_x – ширина окна.

Точность восстановления отклика с использованием (4.9) напрямую зависит от выбора вида функции ядра и ширины окна. Среди множества из-

вестных видов ядер чаще других используются равномерное, треугольное, Епанечникова и Гаусса [18]. При выборе ширины окна следует принимать во внимание, что при слишком малых значениях h_x возникает эффект «недосглаживания», а при слишком больших значениях – эффект «пересглаживания».

Обобщением оценок Надарая–Уотсона является локальная полиномиальная регрессия [18].

Ситуация (СА). Отсутствие априорной информации о виде распределения случайной компоненты модели (2) не позволяет в явном виде использовать для оценивания неизвестных параметров метод максимального правдоподобия. Кроме того, фактически реализуемые на практике распределения случайных ошибок далеко не всегда удается представить в рамках тех или иных хорошо известных теоретических законов. Исследователь может лишь иметь общие представления о его форме и, возможно, сформулировать отдельные гипотезы о наличии тех или иных особенностей (например, сделать корректное предположение о значении математического ожидания). Следовательно, необходимо использовать алгоритмы идентификации, которые сами извлекают информацию о характере распределения из исходных данных и обладают определенной гибкостью для осуществления подстройки под многообразие фактически реализуемых распределений.

Одним из решений является переход к универсальным распределениям. Их основное преимущество состоит в возможности описания большого круга практических ситуаций. В качестве примера универсальных семейств распределений можно упомянуть кривые Пирсона, которые позволяют проводить анализ ситуаций с такими распределениями, как бета-, гамма-, Стьюдента, экспоненциальное и др. Еще более широким является обобщенное лямбда-распределение, включающее не только хорошо известные в теории вероятности распределения, но и целое множество других. Перспективным также представляется переход в комплексную область посредством построения характеристической функции, что обеспечит привлечение более полной информации и позволит идентифицировать так называемые устойчивые распределения. Они также представляют собой весьма широкий класс распределений, включающий распределения с большой или даже бесконечной дисперсией (например, распределение Коши). Это обстоятельство делает его предпочтительным при исследовании закономерностей на основе сильно засоренных данных.

Процедура оценивания параметров модели $f_0(x, \theta)$ в данной ситуации носит итерационный характер. При этом на каждой итерации происходит идентификация распределения случайных остатков с последующим построением функции правдоподобия и оценивание регрессионных параметров методом максимального правдоподобия. Более подробно с особенностями использования этой процедуры для упомянутых универсальных семейств распределений можно ознакомиться в работах [2, 6, 7, 11].

Другим вариантом оценки распределений остатков является использование непараметрических или полупараметрических методов, позволяющих построить эмпирическую функцию плотности с последующим переходом к функции правдоподобия и итерационной процедуре, описанной выше [8].

Ситуации (BB) и (CB). Специфической особенностью данных ситуаций является одновременная неопределенность структуры модели и структуры ошибки. При этом ошибка в выборе спецификации модели будет сказываться на свойствах остатков, а также на качестве определения структуры ошибки.

Методы идентификации регрессионных зависимостей, разработанные специально для таких условий, авторам неизвестны. Тем не менее в качестве возможного инструмента можно рекомендовать применение алгоритмов, аналогичных ситуациям (AB) и (BA), но с использованием устойчивых методов оценивания. Примеры решения авторами подобных задач можно найти в работах [1, 10].

Ситуации (BC) и (CC). Как и в случае ситуации (AC), для оценивания значений отклика используются непараметрические подходы. Однако, степень информированности о распределении и структуре ошибки позволяет более корректно выбирать методы оценивания и определять их внутренние параметры. Например, при использовании ядерных оценок это влияет на выбор вида функции ядра и ширины окна. Кроме того, по мнению авторов, предпочтение следует отдавать более гибким методам, например, локально полиномиальной регрессии, сплайновой регрессии и др. [16, 18]. Следует отметить, что, как и в предыдущем случае, степень изученности данной постановки задачи очень низка.

ЗАКЛЮЧЕНИЕ

Проведена классификация регрессионных моделей, основанная на объеме априорной информации о структуре модели и свойствах случайной ошибки. Выделено девять основных групп моделей, включая модели с полностью известной структурой, известной частично и полностью неизвестной, а также при известном законе распределения случайной ошибки, известном частично и неизвестном. Для каждой из групп определены базовые методы и алгоритмы идентификации. Данная классификация является более общей по сравнению с традиционными и позволяет принимать более обоснованные решения при выборе инструментов анализа.

СПИСОК ЛИТЕРАТУРЫ

1. Устойчивое оценивание нелинейных структурных зависимостей / В.И. Денисов, А.Ю. Тимофеева, Е.А. Хайленко, О.И. Бузмакова // Сибирский журнал индустриальной математики. – 2013. – № 4. – С. 47–60.

2. *Денисов В.И., Тимофеев В.С.* Устойчивые распределения и оценивание параметров регрессионных зависимостей // Известия Томского политехнического университета. – 2011. – Т. 318, № 2. – С. 10–15.

3. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ: пер. с англ. – М.: Статистика, 1973. – 392 с.

4. *Рао С.Р.* Линейные статистические методы и их применение. – М.: Наука, 1968. – 548 с.

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

6. *Тимофеев В.С.* Оценивание параметров регрессионных зависимостей с использованием кривых Пирсона. Ч. 1 // Научный вестник НГТУ. – 2009. – № 4 (37). – С. 57–67.

7. *Тимофеев В.С.* Оценивание параметров регрессионных зависимостей с использованием кривых Пирсона. Ч. 2 // Научный вестник НГТУ. – 2010. – № 1 (38). – С. 57–63.

8. *Тимофеев В.С.* Ядерные оценки плотности при идентификации уравнений регрессии // Научный вестник НГТУ. – 2010. – № 3 (40). – С. 41–50.

9. *Тимофеев В.С., Вострецова Е.А.* Устойчивое оценивание параметров регрессионных моделей с использованием идей метода наименьших квадратов // Научный вестник НГТУ. – 2007. – № 2 (27). – С. 57–67.

10. *Тимофеев В.С., Фаддеенков А.В., Щеколдин В.Ю.* Исследование алгоритмов оценивания параметров модели со структурированной ошибкой с использованием знакового метода // Научный вестник НГТУ. – 2005. – № 2 (20). – С. 71–84.

11. *Тимофеев В.С., Хайленко Е.А.* Адаптивное оценивание параметров регрессионных моделей с использованием обобщенного лямбда-распределения // Доклады Академии наук высшей школы Российской Федерации. – 2010. – № 2 (15). – С. 25–36.

12. *Фаддеенков А.В.* Алгоритмы анализа линейных регрессионных моделей по панельным данным // Научный вестник НГТУ. – 2007. – № 3 (28). – С. 65–78.

13. Робастность в статистике: подход на основе функций влияния / Ф. Хампель, Э. Рончетти, П. Рауссеу, В. Штаэль. – М.: Мир, 1989. – 512 с.

14. *Хьюбер П.* Робастность в статистике. – М.: Мир, 1984. – 303 с.

15. *Шурыгин А.М.* Прикладная статистика: робастность, оценивание, прогноз. – М.: Финансы и статистика, 2000. – 224 с.

16. *Friedman J.H.* Multivariate adaptive regression splines (with discussion) // Annals of Statistics. – 1991. – N 19. – P. 1–141.

17. *Kumbhakar S.C., Knox Lovell C.A.* Stochastic frontier analysis. – New York: Cambridge University Press, 2003. – 344 p.

18. *Pagan A., Ullah A.* Nonparametric econometrics. – New York: Cambridge University Press, 1999. – 424 p.

19. *Rao C.R., Kleffe J.* Estimation of variance components and applications. – New York: North-Holland, 1988. – 374 p. – (North-Holland series in statistics and probability; vol. 3).

20. *Ruppert D., Wand M.P., Carroll R.J.* Semiparametric regression. – New York: Cambridge University Press, 2003. – 404 p.

21. *Sahai H., Ojeda M.* Analysis of variance for random models: theory, methods, applications, and data analysis. Vol. 2. Unbalanced data. – Boston: Birkhäuser, 2005. – 480 p.

Тимофеев Владимир Семенович, доктор технических наук, профессор кафедры теоретической и прикладной информатики 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 80 публикаций, в том числе один учебник. E-mail: v.timofeev@corp.nstu.ru

Фаддеенков Андрей Владимирович, кандидат технических наук, доцент кафедры теоретической и прикладной информатики Новосибирского государственного технического университета. Основное направление научных исследований – разработка и исследование методов и алгоритмов анализа многофакторных объектов со структурированной ошибкой. Имеет более 40 публикаций, в том числе один учебник. E-mail: faddeenkov@corp.nstu.ru

The classification of regression models based on the amount of a priori information *

V. TIMOFEEV[1], *A. FADDEENKOV*[2]

[1] *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, D. Sc. (Eng.), associate professor. E-mail: v.timofeev@corp.nstu.ru*
[2] *Novosibirsk State Technical University, 20, K. Marx Prospekt, Novosibirsk, 630073, Russian Federation, D. Sc. (Eng.), associate professor. E-mail: faddeenkov@corp.nstu.ru*

In this paper, a new classification of regression models is proposed. The basis of the classification is the capacity of a priori information available to the researcher. Two main tasks of model specification facing researchers are considered. On the one hand, the information about the regression model structure is used. Three main levels of the model structure are considered, namely, a complete certainty, a partial uncertainty and a complete uncertainty. In the first case, it is assumed that the model structure is set up to unknown parameters. In the second, case the model structure is not completely known and the missing part is compensated by a nonparametric component. In the third case, the model structure is unknown, which involves the use of nonparametric methods only. It is proposed to use the completeness of a priori information about the distribution of the random component of the model as a second basis of classification. Here it is also proposed to determine three levels. The first level corresponds to the case of random errors with the known distribution law (accurate within parameters). The second level corresponds to the case of structured errors formed by linear combinations or mixtures of random variables. The third level is general and corresponds to the case with no information about the structure and distribution of the random component. Various combinations of knowledge of the model structure and the distribution of the random error form nine main groups of models with their identification methods. Classical methods of regression analysis based on the least square method are considered in the simplest case. With incomplete information, it is recommended to use semi-parametric methods based on spline regression models, structured error models and universal distributions.

Keywords: regression model, model structure, classification, a priori information, parametric methods, nonparametric methods, semiparametric methods, structured error, variance components

DOI: 10.17212/1814-1196-2015-3-58-68

REFERENCES

1. Denisov V.I., Timofeeva A.Yu., Khailenko E.A., Buzmakova O.I. Ustoichivoe otsenivanie nelineinykh strukturnykh zavisimostei [Robust estimation of nonlinear structural models]. *Sibirskii zhurnal industrial'noi matematiki – Journal of Applied and Industrial Mathematics*, 2013, no. 4, pp. 47–60.

2. Denisov V.I., Timofeev V.S. Ustoichivye raspredeleniya i otsenivanie parametrov regressionnykh zavisimostei [Robust distributions and parameter estimation regression]. *Izvestiya Tomskogo politekhnicheskogo universiteta – Bulletin of the Tomsk Polytechnic University*, 2011, vol. 318, no 2, pp. 10–15.

3 Draper N.R., Smith H. *Applied regression analysis*. New York, John Wiley&Sons, 1966. 407 p. (Russ. ed.: Dreiper N.R., Smit G. *Prikladnoi regressionnyi analiz*. Moscow, Statistika Publ., 1973. 392 p.).

4. Rao C.R. *Linear Statistical Inference and its Applications*. New York, Wiley, 1967. 625 p. (Russ. ed.: Rao S.R. *Lineinye statisticheskie metody i ikh primenenie.* Moscow, Nauka Publ., 1968. 548 p.).

* *Received 1 July 2015.*

The work was supported by the Russian Foundation for Basic Research as part of a research project №13-07-00299 a.

5. Timofeev V.S. Otsenivanie parametrov regressionnykh zavisimostei na osnove kharakteristicheskoi funktsii [The characteristic function in parameter estimation problem for regression model]. *Nauchnyi vestnik NGTU – Science Bulletin of the Novosibirsk State Technical University*, 2010, no. 2 (39), pp. 43–52.

6. Timofeev V.S. Otsenivanie parametrov regressionnykh zavisimostei s ispol'zovaniem krivykh Pirsona. Ch. 1 [The Pirson's curves in parameter estimation problem for regression model. Pt. 1]. *Nauchnyi vestnik NGTU – Science Bulletin of the Novosibirsk State Technical University*, 2009, no. 4 (37), pp. 57–67.

7. Timofeev V.S. Otsenivanie parametrov regressionnykh zavisimostei s ispol'zovaniem krivykh Pirsona. Ch. 2. [The Pirson's curves in parameter estimation problem for regression model. Pt. 2]. *Nauchnyi vestnik NGTU – Science Bulletin of the Novosibirsk State Technical University*, 2010, no. 1 (38), pp. 57–63.

8. Timofeev V.S. Yadernye otsenki plotnosti pri identifikatsii uravnenii regressii [The Kernel estimation of density function in the regression identification problem]. *Nauchnyi vestnik NGTU – Science Bulletin of the Novosibirsk State Technical University*, 2010, no. 3 (40), pp. 41–50.

9. Timofeev V.S., Vostretsova E.A. Ustoichivoe otsenivanie parametrov regressionnykh modelei s ispol'zovaniem idei metoda naimen'shikh kvadratov [Robust estimation of regression model parameters based on ideas of least-square method]. *Nauchnyi vestnik NGTU – Science Bulletin of the Novosibirsk State Technical University*, 2007, no. 2 (27), pp. 57–67.

10. Timofeev V.S., Faddeenkov A.V., Shchekoldin V.Yu. Issledovanie algoritmov otsenivaniya parametrov modeli so strukturirovannoi oshibkoi s ispol'zovaniem znakovogo metoda [Investigation of algorithms for estimating the parameters structured error model using sign method]. *Nauchnyi vestnik NGTU – Science Bulletin of the Novosibirsk State Technical University*, 2005, no. 2 (20), pp. 71–84.

11. Timofeev V.S., Khailenko E.A. Adaptivnoe otsenivanie parametrov regressionnykh modelei s ispol'zovaniem obobshchennogo lyambda-raspredeleniya [Adaptive estimation of regression models parameters using generalized lambda-distribution]. *Doklady Akademii nauk vysshei shkoly Rossiiskoi Federatsii – Proceedings of the Russian higher school Academy of sciences*, 2010, no. 2 (15), pp. 25–36.

12. Faddeenkov A.V. Algoritmy analiza lineinykh regressionnykh modelei po panel'nym dannym [The analysis algorithms of linear regression models on panel data]. *Nauchnyi vestnik NGTU – Science Bulletin of the Novosibirsk State Technical University*, 2007, no. 3 (28), pp. 65–78.

13. Hampel F.R., Ronchetti E.M., Rousseeuw P.J., Stahel W.A. *Robust statistics: the approach based on influence functions*. New York, Wiley, 1986 (Russ. ed.: Khampel' F., Ronchetti E., Rausseu P., Shtael' V. *Robastnost' v statistike: podkhod na osnove funktsii vliyaniya*. Moscow, Mir Publ., 1989. 512 p.).

14. Huber P.J. *Robust Statistics*. New York, Wiley, 1981 (Russ. ed.: Kh'yuber P. *Robastnost' v statistike*. Moscow, Mir Publ., 1984. 303 p.).

15. Shurygin A.M. *Prikladnaya statistika: robastnost', otsenivanie, prognoz* [Applied statistics: robustness, estimation, forecast]. Moscow, Finansy i statistika Publ., 2000. 224 p.

16. Friedman J.H. Multivariate adaptive regression splines (with discussion). *Annals of Statistics*, 1991, no. 19, pp. 1–141.

17. Kumbhakar S.C., Knox Lovell C.A. *Stochastic frontier analysis*. New York, Cambridge University Press, 2003. 344 p.

18. Pagan A., Ullah A. *Nonparametric econometrics*. New York, Cambridge University Press, 1999. 424 p.

19. Rao C.R., Kleffe J. *Estimation of variance components and applications*. *North-Holland Series in Statistics and Probability.* Vol. 3. New York, North-Holland, 1988. 374 p.

20. Ruppert D., Wand M.P., Carroll R.J. *Semiparametric regression*. New York, Cambridge University Press, 2003. 404 p.

21. Sahai H., Ojeda M. *Analysis of variance for random models: theory, methods, applications, and data analysis.* Vol. 2. *Unbalanced data*. Boston, Birkhäuser, 2005. 480 p.