

УДК 519.237.5

DOI 10.34822/1999-7604-2021-4-63-70

ВЫБОР ОПТИМАЛЬНОГО СООТНОШЕНИЯ МЕЖДУ ТОЧНОСТЬЮ И НЕЛИНЕЙНОСТЬЮ ПРИ ПОСТРОЕНИИ КВАЗИЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ

М. П. Базилевский [✉], А. В. Караулова

Иркутский государственный университет путей сообщения, Иркутск, Россия

[✉] E-mail: mik2178@yandex.ru

При построении квазилинейных регрессионных моделей возникает проблема выбора оптимального соотношения между их точностью и нелинейностью. Цель работы заключается во внедрении разработанных ранее критериев нелинейности по площади в технологию организации «конкурса» моделей. Для этого сформулирована двухкритериальная задача, состоящая в максимизации коэффициента детерминации и одновременной минимизации критерия нелинейности, и предложено два способа ее решения. Первый способ основан на формировании множества Парето, с помощью которого можно визуальным образом выбрать наиболее приемлемую альтернативу. Второй способ более эффективен с вычислительной точки зрения, поскольку на первом его этапе исключаются значительно нелинейные переменные. Затем из возможных спецификаций выбирается регрессия с наибольшим значением коэффициента детерминации. Предложенные способы решения двухкритериальной задачи продемонстрированы на примере моделирования работы выпарного аппарата.

Ключевые слова: квазилинейная регрессия, критерий нелинейности по площади, коэффициент детерминации, двухкритериальная задача, множество Парето.

SELECTING THE OPTIMUM RELATIONSHIP BETWEEN ACCURACY AND NON-LINEARITY IN CONSTRUCTING QUASI-LINEAR REGRESSION MODELS

M. P. Bazilevsky [✉], A. V. Karaulova

Irkutsk State Transport University, Irkutsk, Russia

[✉] E-mail: mik2178@yandex.ru

The problem of selecting the optimum relationship between the accuracy and non-linearity is arising when constructing quasi-linear regression models. A two-criterial problem consisting of maximizing the coefficient of determinant and simultaneous minimization of non-linearity criterion was formed to implement the previously developed criteria of non-linearity for area into the technology of arrangement of models "competition". Two methods were proposed to solve the problem. The first method is based on the Pareto set, which makes it possible to visually select the most suitable alternative. The second method is more efficient from the calculating point of view as it eliminates the highly non-linear variables. After that, the regression with the highest value of the determinant coefficient is being selected out of possible specifications. The proposed methods for solution of two-criterial problem are demonstrated on the example of stimulating the operation of an evaporator.

Keywords: quasi-linear regression, non-linearity criterion for area, coefficient of determination, two-criteria problem, Pareto set.

Введение

Как известно, методология математического моделирования [1] состоит в замене исходного объекта его «образом» (математической моделью) и его дальнейшем изучении

с помощью компьютерных программ. Среди многочисленных свойств математических моделей можно выделить адекватность и сложность. Под адекватностью модели понимают ее соответствие изучаемому объекту или процессу. Сложная модель характеризуется громоздкостью, большим количеством входящих в нее факторов, нелинейностью и неудобством в обращении, поэтому помимо адекватности к математической модели предъявляется требование достаточной простоты [2]. Но, как правило, простота и адекватность связаны обратной зависимостью: чем более адекватна модель, тем она менее проста и тем труднее ее интерпретировать. Возникает проблема выбора оптимального в некотором смысле соотношения между простотой и адекватностью математической модели.

Распространенным инструментом построения математических моделей статистического типа является регрессионный анализ [3]. При построении регрессионной модели также следует добиваться компромисса между точностью ее оценивания и сложностью. Известный подход к решению этой проблемы состоит в генерации множества альтернативных вариантов регрессий и выборе оптимального из них либо на основе скорректированного критерия детерминации, либо с помощью информационных критериев Акаике и Шварца [4]. Все перечисленные критерии «штрафуют» модель за дополнительно включенные факторы. В работе [5] рассмотрены алгоритмы поиска компромисса между точностью и сложностью при построении нечетких аппроксиматоров. При этом в качестве критериев выступают среднеквадратичная ошибка и число нечетких правил. В [6] предлагается строить множество Парето, из которого конечный пользователь может выбрать модель в зависимости от своих предпочтений. В [7] для достижения баланса между сложностью и точностью регрессионной модели введен функционал, представляющий собой линейную свертку двух критериев – меры соответствия модели измерениям и меры сложности модели

$\int_{t_{\min}}^{t_{\max}} \left(\frac{d^2 x}{dt^2} \right)^2 dt$ как меры кривизны функции $x(t)$.

Качественным способом выбора математической формы регрессионной зависимости является технология организации «конкурса» моделей [8], в основе которой лежит многокритериальный подход. С помощью этой технологии успешно решено множество прикладных задач анализа данных (см., например, [9–11]). Часто при организации «конкурса» формирование альтернативных вариантов моделей осуществляется в классе квазилинейных регрессий, которые легко оцениваются с помощью метода наименьших квадратов (МНК). К сожалению, выбранную из этого множества альтернатив наилучшую квазилинейную модель весьма затруднительно интерпретировать из-за входящих в нее нелинейных преобразований. Поэтому в работе [12] специально для таких регрессий были разработаны критерии нелинейности по площади и длине дуги. В [13] эти критерии были обобщены на случай многих переменных. В [14] доказано, что критерии нелинейности по площади не зависят от коэффициентов регрессии. Целью данной работы является «внедрение» критериев нелинейности по площади в технологию организации «конкурса» моделей, что позволит строить квазилинейные регрессии с оптимальным соотношением между их точностью и нелинейностью.

Формулировка двухкритериальной задачи

Рассмотрим квазилинейную модель вида

$$y_i = \alpha_0 + \sum_{k=1}^q \sum_{j=1}^m \alpha_{kj} f_k(x_{ij}) + \varepsilon_i, \quad i = \overline{1, n}, \quad (1)$$

где $y_i, i = \overline{1, n}$ – значения объясняемой переменной y ; $x_{ij} > 0, i = \overline{1, n}, j = \overline{1, m}$ – значения m объясняющих переменных; $\varepsilon_i, i = \overline{1, n}$ – ошибки аппроксимации; $f_k(x), k = \overline{1, q}$ – элементар-

ные функции; $\alpha_0, \alpha_{kj}, k = \overline{1, q}, j = \overline{1, m}$ – неизвестные параметры модели; n – количество наблюдений.

Будем считать, что в спецификацию (1) каждая преобразованная объясняющая переменная должна входить только один раз. Тогда модель (1) можно записать в виде

$$y_i = \alpha_0 + \sum_{j=1}^m \alpha_j f_{\omega_j}(x_{ij}) + \varepsilon_i, i = \overline{1, n}, \quad (2)$$

где $1 \leq \omega_j \leq q$ – элемент вектора $\Omega = (\omega_1, \omega_2, \dots, \omega_m)$, указывающий номер преобразования для j -й объясняющей переменной.

Сформулируем следующую однокритериальную задачу. Требуется выбрать такие компоненты вектора Ω в регрессионной модели (2), которые обеспечивали бы наилучшее качество аппроксимации по величине коэффициента детерминации R^2 . Таким образом, целевая функция в этой задаче имеет вид $R^2 \rightarrow \max$.

Точное решение сформулированной задачи может быть найдено с помощью организации «конкурса» моделей по следующему алгоритму:

1. Генерируется множество всех возможных спецификаций модели (2), общее число которых q^m .

2. Каждая регрессия оценивается с помощью МНК.

3. Выбирается наилучшая модель по величине R^2 .

К сожалению, построенная в результате проведенного «конкурса» квазилинейная модель может быть в значительной степени нелинейной, а значит, сложной. Возникает проблема, связанная с оценкой степени ее нелинейности. Понятно, что чем выше степень нелинейной квазилинейной регрессии, тем труднее интерпретировать коэффициенты при нелинейно преобразованных переменных.

Для количественной оценки степени нелинейности квазилинейных регрессий в работах [12–14] были разработаны критерии нелинейности «по площади» и «по длине дуги». В основе создания таких критериев лежит идея, согласно которой в экономике вычисляется коэффициент Джини.

Предположим, что в модели (2) в качестве преобразований $f_{\omega_j}(x)$, $j = \overline{1, m}$ выступают непрерывные и монотонные на отрезках $[x_{\min}^j, x_{\max}^j]$, $j = \overline{1, m}$ элементарные функции, где $x_{\min}^j = \min\{x_{1j}, x_{2j}, \dots, x_{nj}\}$, $x_{\max}^j = \max\{x_{1j}, x_{2j}, \dots, x_{nj}\}$, $j = \overline{1, m}$.

Пусть оцененная, например, с помощью МНК модель (2) имеет вид:

$$\tilde{y} = \tilde{\alpha}_0 + \sum_{j=1}^m \tilde{\alpha}_j f_{\omega_j}(x_j), \quad (3)$$

где $\tilde{\alpha}_0, \tilde{\alpha}_1, \dots, \tilde{\alpha}_m$ – оцененные параметры модели.

Обозначим в (3) $g_j(x) = \tilde{\alpha}_j f_{\omega_j}(x_j)$, $j = \overline{1, m}$. Тогда эти функции также будут непрерывными и монотонными на отрезках $[x_{\min}^j, x_{\max}^j]$, $j = \overline{1, m}$. При этом графики функций $g_j(x_j)$ проходят через точки $M_j(x_{\min}^j, \tilde{\alpha}_j f_{\omega_j}(x_{\min}^j))$ и $N_j(x_{\max}^j, \tilde{\alpha}_j f_{\omega_j}(x_{\max}^j))$. Проведем через пары точек M_j и N_j прямые $\varphi_j(x_j)$, уравнения которых имеют вид:

$$\varphi_j(x_j) = k_j x_j + b_j, j = \overline{1, m}. \quad (4)$$

Критерий нелинейности функции $g_j(x_j)$ «по площади» находится по формуле:

$$NC_s(j) = S_A(j)/S_\Delta(j), \quad (5)$$

где $S_A(j)$ – площадь между прямой $\varphi_j(x_j)$ и кривой $g_j(x_j)$ на отрезке $[x_{\min}^j, x_{\max}^j]$; $S_\Delta(j)$ – площадь прямоугольного треугольника с гипотенузой $M_j N_j$.

Каждый из критериев нелинейности (5) принимает значение от 0 до 1. Если $NC_s(j) = 0$, то преобразование j -й независимой переменной является линейным. Если $NC_s(j) \rightarrow 1$, то преобразование j -й независимой переменной в значительной степени нелинейно.

В случае слабой нелинейности, т. е. при $NC_s(j) \approx 0$, можно считать, что кривая $g_j(x_j)$ практически не отличается от прямой $\varphi_j(x_j)$ на отрезке $[x_{\min}^j, x_{\max}^j]$. Тогда вместо соответствующего неинтерпретируемого коэффициента $\tilde{\alpha}_j$ можно интерпретировать угловой коэффициент k_j прямой (4).

В работе [14] установлено, что критерии нелинейности (5) не зависят от оценок квазилинейной регрессии и находятся по формулам

$$NC_s(j) = \left| \frac{f_{\omega_j}(x_{\max}^j) + f_{\omega_j}(x_{\min}^j)}{f_{\omega_j}(x_{\max}^j) - f_{\omega_j}(x_{\min}^j)} - \frac{2 \int_{x_{\min}^j}^{x_{\max}^j} f_{\omega_j}(x_j) dx_j}{(x_{\max}^j - x_{\min}^j)(f_{\omega_j}(x_{\max}^j) - f_{\omega_j}(x_{\min}^j))} \right|, \quad j = \overline{1, m}. \quad (6)$$

Тем самым можно выявлять степень нелинейности спецификаций квазилинейных регрессий еще до этапа их оценивания. Это обстоятельство можно использовать при организации «конкурса» моделей, исключая на начальной стадии значительно нелинейные регрессии, что сократит время построения наилучшей зависимости.

Отметим, что квазилинейная регрессия (3) характеризуется множеством из m критериев нелинейности (5). Для оценки степени ее нелинейности в целом применяется критерий верхней границы нелинейности:

$$L = \max \{ NC_s(1), NC_s(2), \dots, NC_s(m) \}. \quad (7)$$

Критерий (7) обладает теми же свойствами, что и его компоненты (5).

Сформулируем следующую двухкритериальную задачу. Требуется выбрать такие компоненты вектора Ω в регрессионной модели (2), которые обеспечивали бы как ее высокое качество аппроксимации по величине коэффициента детерминации R^2 , так и низкую степень нелинейности по критерию L . В таком случае целевые функции имеют вид $R^2 \rightarrow \max, L \rightarrow \min$.

Решение сформулированной двухкритериальной задачи возможно двумя способами.

Способ 1. Сначала сгенерировать и оценить с помощью МНК все возможные спецификации модели (2), потом для каждого уравнения вычислить значения критериев R^2 и L , а затем сформировать множество недоминируемых альтернатив (множество Парето). При небольшом количестве образующих множество Парето альтернатив можно путем визуального просмотра значений критериев выбрать наиболее приемлемую альтернативу.

Способ 2. Сначала исключить преобразованные переменные, для которых величина критерия нелинейности $NC_s(j)$ превосходит некоторое наперед заданное значение

$\delta \in [0,1;0,3]$. На основе оставшихся переменных сформировать и оценить все возможные спецификации модели (2), потом для каждого уравнения вычислить значения критерия R^2 и выбрать наилучшую альтернативу.

Первый способ требует оценки всех q^m регрессионных моделей, а второй – $\prod_{j=1}^m q_j$,

где q_j – количество преобразований j -й объясняющей переменной, для которых $NC_s(j) \leq \delta$. Поэтому второй способ с вычислительной точки зрения эффективнее, чем первый.

Моделирование работы выпарного аппарата

Продemonстрируем способы решения сформулированной выше двухкритериальной задачи на примере моделирования работы выпарного аппарата на большом промышленном предприятии. Для этого использовалась известная выборка данных (табл. 1) из монографии [15] по следующим переменным:

y – количество используемого пара в фунтах ежемесячно;

x_1 – средняя температура воздуха по шкале Фаренгейта;

x_2 – число рабочих дней в месяце.

Таблица 1

Статистические данные

№	y	x_1	x_2
1	10,98	35,3	20
2	11,13	29,7	20
3	12,51	30,8	23
4	8,4	58,8	20
5	9,27	61,4	21
6	8,73	71,3	22
7	6,36	74,4	11
8	8,5	76,7	23
9	7,82	70,7	21
10	9,14	57,5	20
11	8,24	46,4	20
12	12,19	28,9	21
13	11,88	28,1	21
14	9,57	39,1	19
15	10,94	46,8	23
16	9,58	48,5	20
17	10,09	59,3	22
18	8,11	70	22
19	6,83	70	11
20	8,8	74,5	23
21	7,68	72,1	20
22	8,47	58,1	21
23	8,86	44,6	20
24	10,36	33,4	20
25	11,08	28,6	22

Примечание: составлено авторами по [15].

В качестве элементарных функций выступали $f_1(x) = x$, $f_2(x) = x^{-1}$, $f_3(x) = x^{-2}$, $f_4(x) = x^{-0,5}$, $f_5(x) = \sqrt{x}$, $f_6(x) = x^2$, $f_7(x) = x^3$, $f_8(x) = \ln x$, $f_9(x) = 2^{0,05x}$.

Сначала задача $R^2 \rightarrow \max$, $L \rightarrow \min$ была решена первым способом. Для этого были сгенерированы и с помощью МНК оценены все возможные спецификации модели (2), общее число которых составило $q^m = 9^2 = 81$. Для каждой такой регрессии были вычислены значения критериев R^2 и L , по которым была построена точечная диаграмма (рис.).

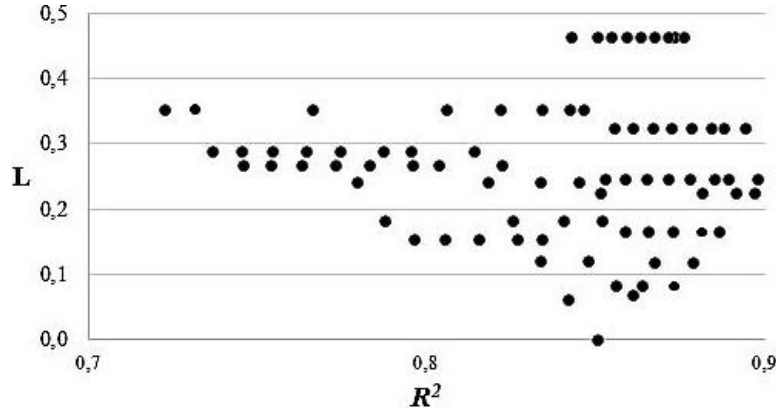


Рис. Точечная диаграмма для критериев R^2 и L
Примечание: составлено авторами.

Как видно по рисунку, критерий R^2 принимает свое наибольшее значение 0,8979 для регрессии с переменными $x_1^{-0,5}$ и x_2^3 , для которой $L = 0,2459$. Для линейной регрессии, у которой $L = 0$, R^2 составляет 0,8505. Понятно, что модель с оптимальным соотношением точности и нелинейности следует искать в нижнем правом углу рисунка.

Далее по значениям критериев R^2 и L было сформировано множество Парето:

$$\tilde{y} = 0,161 + 49,877x_1^{-0,5} + 0,000236x_2^3, R^2 = 0,8979, L = 0,2459, \quad (8)$$

$$\tilde{y} = 21,746 - 3,69 \ln x_1 + 0,000238x_2^3, R^2 = 0,8969, L = 0,2259, \quad (9)$$

$$\tilde{y} = 20,996 - 3,63 \ln x_1 + 0,00619x_2^2, R^2 = 0,8863, L = 0,1646, \quad (10)$$

$$\tilde{y} = 14,312 - 1,047\sqrt{x_1} + 0,00621x_2^2, R^2 = 0,8788, L = 0,1176, \quad (11)$$

$$\tilde{y} = 10,165 - 1,041\sqrt{x_1} + 3,305 \cdot 2^{0,05x_2}, R^2 = 0,8731, L = 0,082, \quad (12)$$

$$\tilde{y} = 6,561 - 0,073x_1 + 3,315 \cdot 2^{0,05x_2}, R^2 = 0,8608, L = 0,069, \quad (13)$$

$$\tilde{y} = 9,167 - 0,0727x_1 + 0,201x_2, R^2 = 0,8505, L = 0. \quad (14)$$

Стоит обратить внимание на последовательность трансформации нелинейной регрессии (8) в линейную регрессию (14). Модель (9) получена из (8) путем замены $x_1^{-0,5}$ на $\ln x_1$, (10) из (9) – x_2^3 на x_2^2 , (11) из (10) – $\ln x_1$ на $\sqrt{x_1}$, (12) из (11) – x_2^2 на $2^{0,05x_2}$, (13) из (12) – $\sqrt{x_1}$ на x_1 , (14) из (13) – $2^{0,05x_2}$ на x_2 .

Выбор наилучшей из альтернатив (8) – (14) в этом случае остается за исследователем.

Затем задача $R^2 \rightarrow \max$, $L \rightarrow \min$ была решена вторым способом. Предварительно были найдены значения критериев нелинейности $NC_s(j)$ для каждой преобразованной переменной. Эти значения приведены в табл. 2.

Таблица 2

Значения критериев нелинейности для преобразованных переменных

NC_S	x	x^{-2}	x^{-1}	$x^{0,5}$	$\sqrt{x}$	x^2	x^3	$\ln x$	$2^{0,05x}$
x_1	0	0,4637	0,3239	0,2459	0,0820	0,1546	0,2885	0,1646	0,2683
x_2	0	0,3529	0,2415	0,1823	0,0608	0,1176	0,2259	0,1218	0,0691

Примечание: составлено авторами на основании данных, полученных в ходе вычислительного эксперимента.

После этого были исключены переменные, для которых величина $NC_S(j)$ превосходит значение $\delta = 0,1$. В результате из 18 преобразованных переменных осталось 5: x_1 , $\sqrt{x_1}$, x_2 , $\sqrt{x_2}$ и $2^{0,05x_2}$. На основе этих пяти преобразованных переменных нужно оценить всего 6 моделей, а не 81, как в предыдущем случае. Лучшей из этих шести регрессий по критерию R^2 оказалась модель (12).

Модель (12) является практически линейной, потому что ее критерий $L = 0,082$. Тогда вместо неинтерпретируемых коэффициентов при переменных $\sqrt{x_1}$ и $2^{0,05x_2}$, равных 1,041 и 3,305 соответственно, можно интерпретировать соответствующие угловые коэффициенты прямых (4): $k_1 = -0,074$ и $k_2 = 0,208$. Заметим, что эти значения незначительно отличаются от коэффициентов линейной регрессии (14). Коэффициент k_1 показывает, что с увеличением средней температуры воздуха x_1 на 1 °F количество используемого пара уменьшается в среднем на 0,074 фунта. А коэффициент k_2 показывает, что с ростом числа рабочих дней в месяце x_2 на 1 единицу количество используемого пара увеличивается в среднем на 0,208 фунта.

Заключение

В работе для выбора оптимального соотношения между точностью и нелинейностью при построении квазилинейных регрессионных моделей сформулирована двухкритериальная задача и предложено 2 способа ее решения. В дальнейшем разработанный математический аппарат планируется реализовать в виде универсального программного комплекса, позволяющего исследователю принимать обоснованные решения об оптимальной структурной спецификации регрессионного уравнения, оцениваемого на основе статистических данных из любой предметной области. Кроме того, вызывает интерес исследование проблемы оценки степени нелинейности квазилинейных регрессий, включающих в себя одну и ту же объясняющую переменную с разными преобразованиями.

Литература

1. Самарский А. А., Михайлов А. П. Математическое моделирование: Идеи. Методы. Примеры. М. : Физматлит, 2001. 320 с.
2. Мышкис А. Д. Элементы теории математических моделей. М. : КомКнига, 2007. 192 с.
3. Westfall P. H., Arias A. L. Understanding Regression Analysis: A Conditional Distribution Approach. Chapman and Hall/CRC, 2020. 514 p.
4. Portet S. A Primer on Model Selection Using the Akaike Information Criterion // Infectious Disease Modelling. 2020. Vol. 5. P. 111–128.
5. Ходашинский И. А., Горбунов И. В. Алгоритмы поиска компромисса между точностью и сложностью при построении нечетких аппроксиматоров // Автометрия. 2013. Т. 49, № 6. С. 51–61.

6. Ходашинский И. А., Горбунов И. В., Синьков Д. С. Алгоритмы генерации структур двухкритериальных Парето-оптимальных нечетких аппроксиматоров // Доклады Томск. гос. ун-та систем управления и радиоэлектроники. 2013. № 1 (27). С. 135–142.
7. Соколов А. В., Волошинов В. В. Выбор математической модели: баланс между сложностью и близостью к измерениям // International Journal of Open Information Technologies. 2018. Т. 6, № 9. С. 33–41.
8. Носков С. И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. Иркутск : Облформпечать, 1996. 321 с.
9. Баенхаева А. В., Базилевский М. П., Носков С. И. Моделирование валового регионального продукта Иркутской области на основе применения методики множественного оценивания регрессионных параметров // Фундамент. исслед. 2016. № 10–1. С. 9–14.
10. Носков С. И., Врублевский И. П. Регрессионная модель динамики эксплуатационных показателей функционирования железнодорожного транспорта // Современные технологии. Системный анализ. Моделирование. 2016. № 2 (50). С. 192–197.
11. Глухов Н. И., Носков С. И., Попов П. Ю. Математическая модель динамики компьютерных преступлений // Информ. технологии и матем. моделирование в управлении сложными системами. 2020. № 1 (6). С. 1–8.
12. Базилевский М. П. Критерии нелинейности квазилинейных регрессионных моделей // Моделирование, оптимизация и информ. технологии. 2018. Т. 6, № 4 (23). С. 185–195.
13. Базилевский М. П. Критерии нелинейности многофакторных квазилинейных регрессий // Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований : материалы II Всерос. национал. науч. конф. студентов, аспирантов и молодых ученых (08–12 апреля 2019 г., г. Комсомольск-на-Амуре). Комсомольск-на-Амуре : Комсомольский-на-Амуре гос. ун-т, 2019. С. 210–213.
14. Базилевский М. П., Караулова А. В. Предварительное оценивание степени нелинейности структурных спецификаций квазилинейных регрессий // Матем. методы в технике и технологиях – ММТТ. 2020. Т. 5. С. 49–52.
15. Draper N. R., Smith H. Applied Regression Analysis. John Wiley & Sons, 1998. 705 p.