ПРИМЕНЕНИЕ АППАРАТА МАТЕМАТИЧЕСКОГО ПРОГРАММИРОВАНИЯ ДЛЯ ПОСТРОЕНИЯ НЕЭЛЕМЕНТАРНЫХ КВАЗИЛИНЕЙНЫХ РЕГРЕССИЙ

Базилевский М. П.[1]
(*ФГБОУ ВО Иркутский государственный университет путей сообщения, Иркутск*)

В неэлементарных квазилинейных регрессиях объясняющие переменные сначала преобразуются с помощью элементарных функций, после чего пары полученных факторов снова преобразуются с помощью неэлементарных функций min *и* max. *Такие модели нелинейны как по факторам, так и по параметрам, поэтому даже их оценивание представляется сложной вычислительной задачей. А если неизвестен состав входящих в модель переменных, а также их элементарные и неэлементарные преобразования, то сложность задачи существенно возрастает. На решение этой проблемы направлено данное исследование. Вместо трудоёмких переборных процедур использован хорошо развитый за последнее время аппарат математического программирования. Метод построения неэлементарных квазилинейных регрессий формализован в виде задачи частично-булевого линейного программирования. Предложенный метод реализован в специальной компьютерной программе. Её достоинство в том, что пользователь может регулировать в процессе построения число преобразованных переменных, поэтому программой можно пользоваться как для решения простых задач управления на обычных персональных компьютерах, так и для обработки массивов больших данных с помощью облачных сервисов. Неэлементарные квазилинейные регрессии могут быть использованы для решения задач управления в технических, социально-экономических, медицинских и других системах.*

Ключевые слова: регрессионный анализ, нелинейная регрессия, неэлементарная квазилинейная регрессия, метод наименьших квадратов, задача частично-булевого линейного программирования.

1. Введение

С помощью машинного обучения [15, 18] в настоящее время решается множество различных задач управления в технике, экономике, медицине, строительстве и других областях человеческой деятельности. Обученные модели используются, в частности, для прогнозирования будущего состояния функциониро-

[1] *Михаил Павлович Базилевский, к.т.н., доцент* (*mik2178@yandex.ru*).

вания исследуемого объекта или процесса. При этом для обучения с целью прогнозирования существует множество типов моделей: деревья решений, алгоритмы кластеризации и пр., но исследователи зачастую выбирают искусственные нейронные сети (ИНС) [22], либо регрессионные модели [16]. Нельзя сказать однозначно, прогнозы по какой из этих двух разновидностей моделей точнее на практике, поскольку все зависит от конкретной ситуации. Данная статья посвящена вопросам автоматизации процесса построения регрессионных моделей.

Давно миновали те времена, когда оценивание линейной регрессии с помощью метода наименьших квадратов (МНК), например, по выборке из ста наблюдений для пяти объясняющих переменных, считалось весьма солидной вычислительной задачей. На современных среднестатистических компьютерах такая задача решается практически мгновенно. Однако линейные регрессии, благодаря своей простой содержательной интерпретации, сегодня всё же находят применение (см., например, [23, 24]). Но большинство реальных объектов и процессов в мире подчиняется нелинейным законам и закономерностям, для описания которых линейных регрессий недостаточно, поэтому приходится прибегать к оцениванию нелинейных зависимостей.

Проведен анализ следующих прикладных научных работ, посвященных нелинейному регрессионному моделированию. В [17] решается проблема поддержки оптимальной температуры внутри помещения, для чего строится нелинейная регрессия зависимости между энергопотреблением здания и температурой наружного воздуха. В [19] разработана трехслойная ИНС и модели нелинейной регрессии для прогнозирования скорости производства биогаза из анаэробного гибридного реактора. При этом оба вида моделей хорошо справились с прогнозированием. В [20] прогнозируются ключевые показатели (осадка гребня, внутренняя осадка и прогиб лицевой плиты) деформаций при строительстве каменно-насыпных плотин с бетонным замком. В [25] представлен анализ и сравнение процесса отслеживания дронов на основе линейных фильтров Калмана по сравнению с нелинейной полиномиальной регрессией. При этом сделан вывод, что оба метода целесообразно использовать в разных усло-

виях шумовых измерений. Работы [8, 11] посвящены прогнозированию прочности бетона на сжатие. В [11] обычный заполнитель в бетонной смеси предлагается заменить переработанными отходам резиновых шин, а в [8] – частично заменить цемент порошком стеклянных отходов. В обоих случаях ИНС оказались несколько лучше по качеству, чем нелинейные регрессии. Хорошие результаты при прогнозировании прочности бетона на сжатие с помощью квазилинейной регрессии продемонстрированы в [5]. Помимо этого учеными ведутся фундаментальные исследования в области нелинейного регрессионного моделирования. Так, в [10] решается проблема усреднения моделей регрессии и исследуется критерий нелинейной информации. А в [9] разработана так называемая оптимальная остаточная регрессия.

Проведенный анализ показывает, что в качестве нелинейных регрессий исследователи зачастую применяют традиционные формы связи между переменными – степенные, полиномиальные, логарифмические и пр. Однако в [2, 3] на основе функции Леонтьева [6] были разработаны новые, показывающие хорошие результаты на практике модели – неэлементарные линейные регрессии (НЛинР), которые линейны по объясняющим переменным, но нелинейны по параметрам. В зависимости от количества объясняющих переменных НЛинР могут иметь большое количество регрессоров, поэтому при моделировании целесообразно решать задачу отбора информативных регрессоров (ОИР) [13, 14]. Для этого в [1] был использован хорошо развитый в последнее время аппарат частично-булевого линейного программирования (ЧБЛП) [21]. Тем самым задача ОИР в НЛинР, оцениваемой с помощью МНК, была сведена к задаче ЧБЛП. Исходя из этого, в работе [4] были предложены неэлементарные квазилинейные регрессии (НКЛинР), в которых объясняющие переменные преобразуются с помощью элементарных математических функций. Там же подробно описана первая версия программы ВИнтер-2 для построения НКЛинР. Однако формализация задачи ОИР в НКЛинР в виде задачи ЧБЛП до сегодняшнего дня нигде не была описана. К тому же ещё никогда, за исключением небольшого примера в [4], не решалась

задача построения НКЛинР по реальным данным. На решение указанных проблем направлено данное исследование.

2. Оптимизационная задача построения НКЛинР

Предположим, что имеется выборочная совокупность объема n, соответствующая измерениям зависимой (объясняемой) переменной y, и l независимых (объясняющих) переменных $x_1, \ldots, x_l$. Тогда предложенная в [1] НЛинР с бинарными операциями min и max и $(1 + l + 4C_l^2)$ неизвестными параметрами α_0, $\alpha_1, \ldots, \alpha_l$, $\alpha_1^{min}, \ldots, \alpha_{C_l^2}^{min}$, $\alpha_1^{max}, \ldots, \alpha_{C_l^2}^{max}$, $k_1^{min}, \ldots, k_{C_l^2}^{min}$, $k_1^{max}, \ldots, k_{C_l^2}^{max}$, имеет вид

$$(1)\quad y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{C_l^2} \alpha_j^{min} \min\left\{x_{i,\mu_{j,1}}, k_j^{min} x_{i,\mu_{j,2}}\right\} + \\ + \sum_{j=1}^{C_l^2} \alpha_j^{max} \max\left\{x_{i,\mu_{j,1}}, k_j^{max} x_{i,\mu_{j,2}}\right\} + \varepsilon_i,\ i = 1, 2, \ldots, n,$$

где C_l^2 – число сочетаний без повторений из l элементов по 2; $\varepsilon_1, \ldots, \varepsilon_n$ – ошибки аппроксимации регрессии; $\mu_{j,1}$, $\mu_{j,2}$, $j = 1, 2, \ldots, C_l^2$, – элементы индексной матрицы **M** размера $C_l^2 \times 2$, содержащей в строках все возможные комбинации (сочетания) пар индексов объясняющих переменных.

Введем множество F, состоящее из числа elem элементарных математических функций, т.е. $F = \{f_1(x), f_2(x), \ldots, f_{elem}(x)\}$. Преобразуя каждую объясняющую переменную с помощью этого множества, получим расширенный набор из $l \times elem$ переменных. Это известный приём, описанный в монографии [7]. В качестве элементарных функций можно использовать x^2, $\sqrt{x}$, $\ln(x)$, e^x и т.д.

Предложенная в [4] для введенного расширенного набора переменных НКЛинР с $(1 + l{\cdot}elem + 4C_{l\cdot elem}{}^2)$ неизвестными параметрами α_0, α_{jk}, $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, elem$, α_j^{min}, α_j^{max}, k_j^{min}, k_j^{max}, $j = 1, 2, \ldots, C_{l\cdot elem}{}^2$, имеет вид

$$(2)\quad y_i = \alpha_0 + \sum_{j=1}^{l}\sum_{k=1}^{elem} \alpha_{jk} f_k(x_{ij}) +$$

$$+ \sum_{j=1}^{C_{l\cdot elem}^2} \alpha_j^{min} \min\left\{ f_{\mu_{j,1,2}}(x_{i,\mu_{j,1,1}}), k_j^{min} f_{\mu_{j,2,2}}(x_{i,\mu_{j,2,1}}) \right\} +$$

$$+ \sum_{j=1}^{C_{l\cdot elem}^2} \alpha_j^{max} \max\left\{ f_{\mu_{j,1,2}}(x_{i,\mu_{j,1,1}}), k_j^{max} f_{\mu_{j,2,2}}(x_{i,\mu_{j,2,1}}) \right\} + \varepsilon_i ,$$

$$i = 1, 2, \ldots, n,$$

где $\mu_{j,1,1}$, $\mu_{j,1,2}$, $\mu_{j,2,1}$, $\mu_{j,2,2}$, $j = 1, 2, \ldots, C_{l\cdot elem}{}^2$ – элементы трехмерного массива (куба) $\mathbf{M}^*$ размера $C_{l\cdot elem}{}^2 \times 2 \times 2$, в котором первое измерение – «Номер пары преобразованных переменных», второе – «Индексы переменных», третье – «Индексы преобразований». Иными словами, горизонтальные срезы куба представляет собой двухмерные массивы (матрицы) двух измерений – «Индексы переменных» и «Индексы преобразований».

НКЛинР (2) также можно называть неэлементарной регрессией с элементарными преобразованиями переменных.

Например, если имеется три переменных x_1, x_2, x_3, а множество $F = \{x^3, 2^x\}$, то без использования операции max НКЛинР (2) примет вид

$$y_i = \alpha_0 + \alpha_{11} x_{i1}^3 + \alpha_{12} 2^{x_{i1}} + \alpha_{21} x_{i2}^3 + \alpha_{22} 2^{x_{i2}} + \alpha_{31} x_{i3}^3 + \alpha_{32} 2^{x_{i3}} +$$
$$+ \alpha_1^{min} \min\left\{ x_{i1}^3, k_1^{min} x_{i2}^3 \right\} + \alpha_2^{min} \min\left\{ x_{i1}^3, k_2^{min} x_{i3}^3 \right\} + \alpha_3^{min} \min\left\{ x_{i1}^3, k_3^{min} 2^{x_{i1}} \right\} +$$
$$+ \alpha_4^{min} \min\left\{ x_{i1}^3, k_4^{min} 2^{x_{i2}} \right\} + \alpha_5^{min} \min\left\{ x_{i1}^3, k_5^{min} 2^{x_{i3}} \right\} + \alpha_6^{min} \min\left\{ x_{i2}^3, k_6^{min} x_{i3}^3 \right\} +$$
$$+ \alpha_7^{min} \min\left\{ x_{i2}^3, k_7^{min} 2^{x_{i1}} \right\} + \alpha_8^{min} \min\left\{ x_{i2}^3, k_8^{min} 2^{x_{i2}} \right\} + \alpha_9^{min} \min\left\{ x_{i2}^3, k_9^{min} 2^{x_{i3}} \right\} +$$
$$+ \alpha_{10}^{min} \min\left\{ x_{i3}^3, k_{10}^{min} 2^{x_{i1}} \right\} + \alpha_{11}^{min} \min\left\{ x_{i3}^3, k_{11}^{min} 2^{x_{i2}} \right\} + \alpha_{12}^{min} \min\left\{ x_{i3}^3, k_{12}^{min} 2^{x_{i3}} \right\} +$$
$$+ \alpha_{13}^{min} \min\left\{ 2^{x_{i1}}, k_{13}^{min} 2^{x_{i2}} \right\} + \alpha_{14}^{min} \min\left\{ 2^{x_{i1}}, k_{14}^{min} 2^{x_{i3}} \right\} +$$
$$+ \alpha_{15}^{min} \min\left\{ 2^{x_{i2}}, k_{15}^{min} 2^{x_{i3}} \right\} + \varepsilon_i , \quad i = 1, 2, \ldots, n,$$

для которой вертикальный срез куба $\mathbf{M}^*$ измерений «Номер пары преобразованных переменных» и «Индексы переменных» –

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 1 & 1 & 2 \\ 2 & 3 & 1 & 2 & 3 & 3 & 1 & 2 & 3 & 1 & 2 & 3 & 2 & 3 & 3 \end{pmatrix}^{\mathrm{T}},$$

а вертикальный срез измерений «Номер пары преобразованных переменных» и «Индексы преобразований» –

$$\begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 2 & 2 & 2 \\ 1 & 1 & 2 & 2 & 2 & 1 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 & 2 \end{pmatrix}^{\mathrm{T}}.$$

Как видно, даже при $l = 3$ и $elem = 2$ в НКЛинР (2) уже содержится слишком много регрессоров (21 штука). А при использовании бинарной операции max их количество ещё существенно возрастет (до 36 штук), что серьезно осложнит процесс оценивания модели. Поэтому возникает необходимость выбора из расширенного набора преобразованных переменных только наиболее информативных из них в некотором смысле. Формализуем далее такую задачу ОИР в виде задачи ЧБЛП так, как это сделано для НЛинР (1) в [1].

Для удобства запишем НКЛинР (2) в виде

$$(3)\quad y_i = \alpha_0 + \sum_{j=1}^{l}\sum_{k=1}^{elem} \alpha_{jk} \cdot w_{ijk} +$$

$$+ \sum_{j=1}^{C_{l\cdot elem}^2} \alpha_j^{min} \min\left\{ w_{i,\mu_{j,1,1},\mu_{j,1,2}}, k_j^{min} \cdot w_{i,\mu_{j,2,1},\mu_{j,2,2}} \right\} +$$

$$+ \sum_{j=1}^{C_{l\cdot elem}^2} \alpha_j^{max} \max\left\{ w_{i,\mu_{j,1,1},\mu_{j,1,2}}, k_j^{max} \cdot w_{i,\mu_{j,2,1},\mu_{j,2,2}} \right\} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

где $w_{ijk} = f_k(x_{ij})$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, elem$.

Из [1] следует, что области возможных значений параметров k_j^{min}, k_j^{max}, $j = 1, 2, \ldots, C_{l\cdot elem}{}^2$, в регрессии (3) при $w_{ijk} \neq 0$ можно записать в виде следующих промежутков:

$$(4)\quad k_j^{\text{нижн}} < k_j^{min} < k_j^{\text{верхн}}, \qquad k_j^{\text{нижн}} < k_j^{max} < k_j^{\text{верхн}}, \quad j = 1, 2, \ldots, C_{l\cdot elem}{}^2,$$

где $k_j^{\text{нижн}} = \min\left\{ \dfrac{w_{1,\mu_{j,1,1},\mu_{j,1,2}}}{w_{1,\mu_{j,2,1},\mu_{j,2,2}}}, \dfrac{w_{2,\mu_{j,1,1},\mu_{j,1,2}}}{w_{2,\mu_{j,2,1},\mu_{j,2,2}}}, \ldots, \dfrac{w_{n,\mu_{j,1,1},\mu_{j,1,2}}}{w_{n,\mu_{j,2,1},\mu_{j,2,2}}} \right\}$,

$$k_j^{\text{верхн}} = \max\left\{ \frac{w_{1,\mu_{j,1,1},\mu_{j,1,2}}}{w_{1,\mu_{j,2,1},\mu_{j,2,2}}}, \frac{w_{2,\mu_{j,1,1},\mu_{j,1,2}}}{w_{2,\mu_{j,2,1},\mu_{j,2,2}}}, \ldots, \frac{w_{n,\mu_{j,1,1},\mu_{j,1,2}}}{w_{n,\mu_{j,2,1},\mu_{j,2,2}}} \right\}.$$

Затем равномерно разобьем каждый из промежутков (4) p точками и перепишем регрессию (3) в виде

$$(5)\quad y_i = \alpha_0 + \sum_{j=1}^{l}\sum_{k=1}^{elem}\alpha_{jk}\cdot w_{ijk} +$$
$$+\sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\alpha_{jk}^{-}\min\left\{w_{i,\mu_{j,1,1},\mu_{j,1,2}},\lambda_{jk}\cdot w_{i,\mu_{j,2,1},\mu_{j,2,2}}\right\}+$$
$$+\sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\alpha_{jk}^{+}\max\left\{w_{i,\mu_{j,1,1},\mu_{j,1,2}},\lambda_{jk}\cdot w_{i,\mu_{j,2,1},\mu_{j,2,2}}\right\}+\varepsilon_i,\ i=1,2,\ldots,n,$$

где λ_{jk}, $j = 1, 2, \ldots, C_{l\cdot elem}^2$, $k = 1, 2, \ldots, p$ – элементы матрицы Λ размера $C_{l\cdot elem}^2 \times p$. Элемент λ_{jk} равен значению k-й точки j-го промежутка (4). Поскольку коэффициенты λ_{jk}, $j = 1, 2, \ldots, C_{l\cdot elem}^2$, $k = 1, 2, \ldots, p$, известны, то регрессия (5) относится к квазилинейным, т.е. становится линейной по неизвестным параметрам α_{jk}, $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, elem$; α_{jk}^{-}, α_{jk}^{+} $j = 1, 2, \ldots, C_{l\cdot elem}^2$, $k = 1, 2, \ldots, p$. Поэтому удобнее записать её в виде

$$(6)\quad y_i = \alpha_0 + \sum_{j=1}^{l}\sum_{k=1}^{elem}\alpha_{jk}\cdot w_{ijk} + \sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\alpha_{jk}^{-}\cdot z_{ijk}^{-} +$$
$$+\sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\alpha_{jk}^{+}\cdot z_{ijk}^{+}+\varepsilon_i,\ i=1,2,\ldots,n,$$

где $z_{ijk}^{-} = \min\left\{w_{i,\mu_{j,1,1},\mu_{j,1,2}},\lambda_{jk}\cdot w_{i,\mu_{j,2,1},\mu_{j,2,2}}\right\}$,

$z_{ijk}^{+} = \max\left\{w_{i,\mu_{j,1,1},\mu_{j,1,2}},\lambda_{jk}\cdot w_{i,\mu_{j,2,1},\mu_{j,2,2}}\right\}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, C_{l\cdot elem}^2$, $k = 1, 2, \ldots, p$.

Сформулируем задачу ОИР для регрессии (6) следующим образом. Требуется выбрать оптимальное число регрессоров так, чтобы коэффициент детерминации R^2 модели был максимален, каждая объясняющая переменная входила в модель не более одного раза, а знаки всех МНК-оценок параметров α_{jk}, α_{jk}^{-}, α_{jk}^{+} удовлетворяли содержательному смыслу факторов. Последнее условие сначала проверяется априори, для чего нужно обратиться к экспертам из данной предметной области, задача которых проанализировать коэффициенты корреляции зависимой переменной со всеми объясняющими переменными w_{jk}, z_{jk}^{-}, $\boldsymbol{z_{jk}^{+}}$.

Не удовлетворяющие смыслу переменные следует исключить, либо, например, дополнить выборку новыми наблюдениями. Затем согласованность содержательному смыслу проверяется уже после оценивания с использованием следующих неравенств:

(7) $\alpha_{jk} \cdot r_{y,w_{jk}} > 0$, $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, elem$;

$\alpha_{jk}^{-} \cdot r_{y,z_{jk}^{-}} > 0$, $\alpha_{jk}^{+} \cdot r_{y,z_{jk}^{+}} > 0$, $j = 1, 2, \ldots, C_{l \cdot elem}^{2}$, $k = 1, 2, \ldots, p$,

где символом r обозначены коэффициенты корреляции между переменными.

Введем принимающие два значения «0» и «1» бинарные переменные

(8) $\delta_{jk} \in \{0,1\}$, $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, elem$,

(9) $\delta_{jk}^{-}, \delta_{jk}^{+} \in \{0,1\}$, $j = 1, 2, \ldots, C_{l \cdot elem}^{2}$, $k = 1, 2, \ldots, p$,

которые отвечают за вхождение регрессоров w_{jk}, z_{jk}^{-}, z_{jk}^{+} в модель. Например, если $\delta_{jk} = 1$, то в модель входит регрессор без бинарной операции (j-я переменная с k-м преобразованием), а если $\delta_{jk} = 0$, то нет. Если $\delta_{jk}^{-} = 1$, то в модель входит регрессор с бинарной операцией min (j-я пара преобразованных переменных для k-й точки разбитого промежутка), а если $\delta_{jk}^{-} = 0$, то нет. Аналогично можно описать переменную δ_{jk}^{+}.

С помощью переменных, преобразованных по правилам

$$y_i^{\bullet} = \frac{y_i - \overline{y}}{\sigma_y},\ i = 1, 2, \ldots, n;\ w_{ijk}^{\bullet} = \frac{w_{ijk} - \overline{w_{jk}}}{\sigma_{w_{jk}}},$$

$i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, elem$;

$$z_{ijk}^{*} = \frac{z_{ijk}^{-} - \overline{z_{jk}^{-}}}{\sigma_{z_{jk}^{-}}},\ z_{ijk}^{**} = \frac{z_{ijk}^{+} - \overline{z_{jk}^{+}}}{\sigma_{z_{jk}^{+}}},\ i = 1, 2, \ldots, n,$$

$j = 1, 2, \ldots, C_{l \cdot elem}^{2}$, $k = 1, 2, \ldots, p$,

перейдем от модели (6) к стандартизованной регрессии с неизвестными коэффициентами β_{jk}, $j = 1, 2, \ldots, l$, $k = 1, 2, \ldots, elem$, β_{jk}^{-}, β_{jk}^{+}, $j = 1, 2, \ldots, C_{l \cdot elem}^{2}$, $k = 1, 2, \ldots, p$, вида

$$(10)\ y_i^{\bullet} = \sum_{j=1}^{l}\sum_{k=1}^{elem}\beta_{jk}\cdot w_{ijk}^{\bullet} + \sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\beta_{jk}^{-}\cdot z_{ijk}^{*} + \\ + \sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\beta_{jk}^{+}\cdot z_{ijk}^{**} + \varepsilon_i^{\bullet},\ i = 1, 2, \ldots, n,$$

где $\varepsilon_1^{\bullet}, \ldots, \varepsilon_n^{\bullet}$ – ошибки аппроксимации модели.

Коэффициент детерминации R^2 регрессий (6) и (10) находится [1] как сумма произведений стандартизованных коэффициентов на соответствующие корреляции регрессоров с переменной y, поэтому введем следующую целевую функцию:

$$(11)\ R^2 = \sum_{j=1}^{l}\sum_{k=1}^{elem}\beta_{jk}\cdot r_{y,w_{jk}^{\bullet}} + \sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\beta_{jk}^{-}\cdot r_{y,z_{jk}^{*}} + \\ + \sum_{j=1}^{C_{l\cdot elem}^2}\sum_{k=1}^{p}\beta_{jk}^{+}\cdot r_{y,z_{jk}^{**}} \to \max.$$

МНК-оценивание модели (10) состоит в решении системы линейных алгебраических уравнений. Однако при комбинировании регрессоров нужно учесть изменение структуры такой системы: обнуление оценок и исключение из системы лишних уравнений. Обнуление оценок, с учётом введенных бинарных переменных и условий (7), осуществляется с помощью следующих ограничений:

$$(12)\ 0 \le \beta_{jk} \le M\cdot\delta_{jk},\quad (j,k) \in \left\{(s_1,s_2)\,|\,r_{y,w_{s_1s_2}} > 0\right\},$$

$$(13)\ -M\cdot\delta_{jk} \le \beta_{jk} \le 0,\ (j,k) \in \left\{(s_1,s_2)\,|\,r_{y,w_{s_1s_2}} < 0\right\},$$

$$(14)\ 0 \le \beta_{jk}^{-} \le M\cdot\delta_{jk}^{-},\quad (j,k) \in \left\{(s_1,s_2)\,|\,r_{y,z_{s_1s_2}^{-}} > 0\right\},$$

$$(15)\ -M\cdot\delta_{jk}^{-} \le \beta_{jk}^{-} \le 0,\ (j,k) \in \left\{(s_1,s_2)\,|\,r_{y,z_{s_1s_2}^{-}} < 0\right\},$$

$$(16)\ 0 \le \beta_{jk}^{+} \le M\cdot\delta_{jk}^{+},\quad (j,k) \in \left\{(s_1,s_2)\,|\,r_{y,z_{s_1s_2}^{+}} > 0\right\},$$

$$(17)\ -M\cdot\delta_{jk}^{+} \le \beta_{jk}^{+} \le 0,\ (j,k) \in \left\{(s_1,s_2)\,|\,r_{y,z_{s_1s_2}^{+}} < 0\right\},$$

где M – большое положительное число, способ выбора которого обсуждается в [1]. Заметим, что использовать ограничения

(12)–(17) на МНК-оценки не обязательно. Однако, как утверждается в [12], их наличие существенно повышает скорость решения задачи.

Исключение из системы лишних уравнений для обнуленных бинарных переменных организуется с помощью ограничений

$$(18)\ -\left(1-\delta_{jk}\right)M \le \sum_{s_1=1}^{l}\sum_{s_2=1}^{elem} r_{w_{jk}w_{s_1s_2}} \cdot \beta_{jk} + \sum_{s_1=1}^{C^2_{l\cdot elem}}\sum_{s_2=1}^{p} r_{w_{jk}z^-_{s_1s_2}} \cdot \beta^-_{s_1s_2} +$$

$$+ \sum_{s_1=1}^{C^2_{l\cdot elem}}\sum_{s_2=1}^{p} r_{w_{jk}z^+_{s_1s_2}} \cdot \beta^+_{s_1s_2} - r_{y,w_{jk}} \le (1-\delta_{jk})M\ ,$$

$j = 1, 2, \ldots, l, k = 1, 2, \ldots, elem,$

$$(19)\ -\left(1-\delta^-_{jk}\right)M \le \sum_{s_1=1}^{l}\sum_{s_2=1}^{elem} r_{w_{s_1s_2}z^-_{jk}} \cdot \beta_{s_1s_2} + \sum_{s_1=1}^{C^2_{l\cdot elem}}\sum_{s_2=1}^{p} r_{z^-_{s_1s_2}z^-_{jk}} \cdot \beta^-_{s_1s_2} +$$

$$+ \sum_{s_1=1}^{C^2_{l\cdot elem}}\sum_{s_2=1}^{p} r_{z^+_{s_1s_2}z^-_{jk}} \cdot \beta^+_{s_1s_2} - r_{yz^-_{jk}} \le (1-\delta^-_{jk})M\ ,$$

$j = 1, 2, \ldots, C_{l\cdot elem}{}^2, k = 1, 2, \ldots, p,$

$$(20)\ -\left(1-\delta^+_{jk}\right)M \le \sum_{s_1=1}^{l}\sum_{s_2=1}^{elem} r_{w_{s_1s_2}z^+_{jk}} \cdot \beta_{s_1s_2} + \sum_{s_1=1}^{C^2_{l\cdot elem}}\sum_{s_2=1}^{p} r_{z^-_{s_1s_2}z^+_{jk}} \cdot \beta^-_{s_1s_2} +$$

$$+ \sum_{s_1=1}^{C^2_{l\cdot elem}}\sum_{s_2=1}^{p} r_{z^+_{s_1s_2}z^+_{jk}} \cdot \beta^+_{s_1s_2} - r_{yz^+_{jk}} \le (1-\delta^+_{jk})M\ ,$$

$j = 1, 2, \ldots, C_{l\cdot elem}{}^2, k = 1, 2, \ldots, p.$

Как видно, если некоторая бинарная переменная «срабатывает», т.е. принимает значение 1, то соответствующее ограничение из множества (18)–(20) трансформируется в строгое равенство. Иначе это ограничение игнорируется.

Ограничения на единственность вхождения каждой объясняющей переменной в модель можно записать в виде

$$(21)\ \sum_{k=1}^{elem}\delta_{jk} + \sum_{i\in\Upsilon_j}\sum_{k=1}^{p}\delta^-_{ik} + \sum_{i\in\Upsilon_j}\sum_{k=1}^{p}\delta^+_{ik} \le 1,\ j = 1, 2, \ldots, l,$$

где Y_j – множество номеров горизонтальных срезов трехмерной матрицы $\mathbf{M}^*$, содержащих среди своих элементов число j.

Таким образом, решение задачи ЧБЛП с целевой функцией (11) и с линейными ограничениями (8), (9), (12)–(21) гарантирует построение НКЛинР (2) с требуемыми свойствами.

3. Пример

Для построения НКЛинР была использована выборочная совокупность объема $n = 21$, по которой в [1] идентифицировалась НЛинР. Из исходного набора были задействованы следующие переменные: y – отправление грузов железнодорожным транспортом общего пользования в Иркутской области (млн тонн); x_2 – процент трудоспособного населения от общей численности; x_3 – численность рабочей силы (тыс. чел.); x_{18} – число предприятий и организаций; x_{22} – производство электроэнергии (млрд кВт*ч).

Выбор именно этих четырех объясняющих переменных продиктован следующими соображениями.

1. На объемы отправленных грузов железнодорожным транспортом, несомненно, влияют объемы производства в регионе. А повышение объемов производства стимулирует, в частности, увеличение мощностей генерации электроэнергии, числа предприятий и численности рабочей силы.

2. Все коэффициенты корреляций объясняющих переменных с y значимы для уровня значимости 0,05.

3. Коэффициенты корреляции всех этих четырех объясняющих переменных с y положительны, что соответствуют экономическому смыслу решаемой задачи.

Заметим, что единственный значимый коэффициент корреляции равен 0,82 между переменными x_2 и x_3. Однако, как будет показано далее, это не повлияет на интерпретацию НКЛинР.

Рассмотрим этапы построения НКЛинР в программе ВИнтер-2.

Поскольку все четыре объясняющих переменных по знаку корреляции с y согласованы, то можно сразу переходить к загрузке выборки в программу. Загрузив статистические данные из текстового файла, на панели «Квазилинейная» сначала нужно выбрать элементарные преобразования объясняющих перемен-

ных. На данный момент в ВИнтер-2 встроено 9 функций: x^{-2}, $x^{-1,5}$, x^{-1}, $x^{-0,5}$, $\sqrt{x}$, x, $x^{1,5}$, x^2, $\ln(x)$. Выбор слишком большого числа преобразований негативно влияет на скорость решения задачи, поэтому было принято решения сформировать множество F из трех функций: x^{-1}, $\sqrt{x}$ и $\ln(x)$. Выбрав преобразования, следует нажать на кнопку «Исключить». В результате нажатия формируются все преобразованные объясняющие переменные и каждая из них автоматически проходит проверку на: 1) согласованность знака корреляции с y смыслу задачи; 2) слишком малое значение коэффициента корреляции с y; 3) слишком высокое значения критерия нелинейности [5]. Для последних двух условий пороговые значения в ВИнтер-2 были назначены 0 и 1 соответственно. Для первого условия ничего задавать не нужно. В итоге ни одна из 12 преобразованных переменных не была исключена.

Затем нужно выбрать неэлементарные преобразования для всех двенадцати элементарно преобразованных переменных. Для этого на панели «Неэлементарные преобразования» нужно указать два параметра: 1) число разбиений; 2) корреляция с y. Первый параметр означает число точек разбиения p промежутков (4). Второй параметр – пороговое значение для абсолютной величины коэффициента корреляции преобразованной переменной с y. Было принято решение задать эти параметры равными 4 и 0,7 соответственно. После чего, нажав кнопку «Исключить», автоматически сформируются все возможные неэлементарные преобразования переменных и каждая из них пройдет проверку на: 1) согласованность знака корреляции с y смыслу задачи; 2) слишком малое значение коэффициента корреляции с y. Всего было сформировано $2C_{l\cdot elem}^2 \cdot p = 2C_{12}^4 \cdot 4 = 528$ неэлементарных преобразований. В результате исключения из них осталось только 80. Тем самым, к построению НКЛинР мы подошли, имея в распоряжении 12 элементарно и 80 не элементарно преобразованных переменных.

Для формирования в ВИнтер-2 задачи ЧБЛП (8), (9), (11)–(21) на панели «Задача ЧБЛП» были выбраны следующие параметры: 1) регрессоры – 0; 2) точность – «0,000000000000»;

3) вхождение – 1. Первый параметр означает, что нет ограничений на число регрессоров в модели, третий – что каждая объясняющая переменная входит в модель не более одного раза. Второй параметр означает, что все величины в задачи должны округляться до 12 знаков после запятой. При нажатии на кнопку «Создать» автоматически сформировалась задача ЧБЛП для пакета LPSolve.

Решение сформированной задачи в LPSolve на обычном персональном компьютере с процессором AMD Ryzen 3 4300 (2,70 ГГц) было получено за 2,724 с. С использованием информации, расположенной в текстовом поле на панели «Неэлементарные преобразования», была произведена расшифровка спецификации модели. Таким образом, построена следующая НКЛинР:

$$(22)\quad \tilde{y} = -618,246 + \underset{(13,97)}{\overset{(0,5363)}{0,43}} \min\{\sqrt{x_{18}}; 32,4774\sqrt{x_{22}}\} + \\ + \underset{(12,11)}{\overset{(0,4244)}{72,594}} \max\{\sqrt{x_2}; 1,1024 \ln x_3\}.$$

Коэффициент детерминации R^2 модели (22) составляет 0,960711, что на 0,014528 выше, чем у построенной в [1] НЛинР. К тому же число степеней свободы НКЛинР (22) $df = n - m - 1 = 21 - 2 - 1 = 18$, что на 2 единицы больше, чем у приведенной в [1] модели. Как видно, каждая объясняющая переменная входит в НКЛинР ровно 1 раз, а знаки всех МНК-оценок в этой регрессии согласуются с содержательным смыслом задачи.

В уравнении (22) в скобках под коэффициентами приведены наблюдаемые значения *t*-критерия Стьюдента, а над оценками – значения абсолютных вкладов переменных в общую детерминацию, которые в сумме дают 0,9607. Оба регрессора в регрессии (22) значимы по *t*-критерию Стьюдента для уровня значимости $\alpha = 0,01$. Причем значение коэффициента корреляции между переменными $\min\{\sqrt{x_{18}}; 32,4774\sqrt{x_{22}}\}$ и $\max\{\sqrt{x_2}; 1,1024 \ln x_3\}$ составляет 0,225, что говорит об отсутствии в модели мультиколлинеарности.

Представим модель (22) в виде кусочно-заданной функции:

$$(23)\quad \tilde{y} = \begin{cases} -618,246 + 13,97\sqrt{x_{22}} + 72,59\sqrt{x_2} & \text{при } \dfrac{x_{18}}{x_{22}} \geq 1054,78,\ \dfrac{\sqrt{x_2}}{\ln x_3} \geq 1,102, \\ -618,246 + 13,97\sqrt{x_{22}} + 80,03\ln x_3 & \text{при } \dfrac{x_{18}}{x_{22}} \geq 1054,78,\ \dfrac{\sqrt{x_2}}{\ln x_3} < 1,102, \\ -618,246 + 0,43\sqrt{x_{18}} + 72,59\sqrt{x_2} & \text{при } \dfrac{x_{18}}{x_{22}} < 1054,78,\ \dfrac{\sqrt{x_2}}{\ln x_3} \geq 1,102, \\ -618,246 + 0,43\sqrt{x_{18}} + 80,03\ln x_3 & \text{при } \dfrac{x_{18}}{x_{22}} < 1054,78,\ \dfrac{\sqrt{x_2}}{\ln x_3} < 1,102. \end{cases}$$

Видно, что функция (23) меняет своё аналитическое выражение в зависимости от значений соотношений x_{18} / x_{22} и $\sqrt{x_2} / \ln x_3$. Если x_{18} / $x_{22} \geq 1054,78$, то на y влияет регрессор $\sqrt{x_{22}}$, а если x_{18} / $x_{22} < 1054,78$, то $\sqrt{x_{18}}$. Если $\sqrt{x_2} / \ln x_3 \geq 1,102$, то на y влияет регрессор $\sqrt{x_2}$, а если $\sqrt{x_2} / \ln x_3 < 1,102$, то $\ln(x_3)$. Условие x_{18} / $x_{22} \geq 1054,78$ сработало для 2003–2009, 2013–2018 гг., а x_{18} / $x_{22} < 1054,78$ – в остальных случаях. Условие $\sqrt{x_2} / \ln x_3 \geq 1,102$ сработало для 2003–2008 гг., а условие $\sqrt{x_2} / \ln x_3 < 1,102$ – в остальных случаях. Для интерпретации коэффициентов при преобразованных переменных $\sqrt{x_2}$, $\sqrt{x_{18}}$, $\sqrt{x_{22}}$ и $\ln(x_3)$ можно использовать приём, предложенный в [5]. Из-за переключения преобразованных переменных x_2 и x_3 высокая степень их корреляции не влияет на интерпретацию оценок.

Далее было принято решение с использованием других известных спецификаций попытаться получить модель с двумя регрессорами, т.е. с числом степеней свободы $df = 18$, превосходящую НКЛинР (22) по значению R^2. В результате были оценены следующие регрессии.

1. Линейная регрессия со всеми четырьмя переменными:

$$\tilde{y} = -76,682 + 1,849x_2 - 0,0162x_3 + 0,0005x_{18} + 0,22x_{22},$$

для которой $R^2 = 0,867334$. Во-первых, в этой регрессии, в отличие от модели (22), знак коэффициента при переменной x_3 не удовлетворяет содержательному смыслу задачи. Во-вторых, если даже линейная регрессия с четырьмя переменными оказалась

хуже модели (22), то двухфакторные линейные регрессии, для которых $df = 18$, тем более окажутся хуже.

2. Квазилинейная модель:

$$\tilde{y} = -126{,}157 - 1{,}77 \cdot 10^6 x_{18}^{-1} + 27{,}434\sqrt{x_2},$$

для которой $R^2 = 0{,}852327$. Эта регрессия получена с использованием той же технологии, что и (22), поэтому знаки коэффициентов в ней корректны, но качество её ниже.

3. Параболический тренд:

$$\tilde{y} = 49{,}2728 + 2{,}9061t - 0{,}1551t^2,$$

для которого $R^2 = 0{,}698581$, $t = 1, 2, \ldots, 21$, – переменная времени.

4. Тренд с фиктивной переменной:

$$\tilde{y} = 44{,}756 + 3{,}379t - 5{,}0275d,$$

для которого $R^2 = 0{,}911306$, а фиктивная переменная задана по правилу $d = \begin{cases} 0, & \text{если } t < 7, \\ t - 7, & \text{если } t \geq 7. \end{cases}$

5. Авторегрессия второго порядка:

$$\tilde{y}_t = 5{,}944 + 0{,}936y_{t-1} - 0{,}0443y_{t-2},$$

для которой $R^2 = 0{,}732612$.

Таким образом, все пять типов построенных моделей проиграли по величине R^2 НКЛинР (22).

4. Заключение

В статье впервые предложен метод построения неэлементарных квазилинейных регрессий, которые могут быть использованы для выработки эффективных управленческих решений в различных системах. Сформулированная задача ЧБЛП содержит линейные ограничения на знаки МНК-оценок и количество вхождений объясняющих переменных в модель. При снятии этих ограничений результаты моделирования могут существенно улучшиться, но при этом увеличится и время решения задачи. Предложенный метод реализован в программе ВИнтер-2, которая с помощью решателя LPSolve по заданным пользователем начальным параметрам автоматически строит оптимальную

по коэффициенту детерминации неэлементарную квазилинейную регрессию, т.е. автоматически определяет состав входящих в неё объясняющих переменных, а также их элементарные и неэлементарные преобразования. В программе имеется возможность управлять количеством преобразованных переменных и тем самым контролировать сложность задачи и время её решения. Построенная с помощью ВИнтер-2 модель по данным о железнодорожных перевозках оказалась лучше, чем построенная ранее неэлементарная линейная регрессия. Однако в построенных нашим методом регрессиях некоторые регрессоры могут быть незначимы по тем или иным критериям, а также может присутствовать мультиколлинеарность. Решению этих проблем будут посвящены дальнейшие исследования.

Литература

1. БАЗИЛЕВСКИЙ М.П. *Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования* // Проблемы управления. – 2022. – №4. – С. 3–14.
2. БАЗИЛЕВСКИЙ М.П. *Отбор информативных операций при построении линейно-неэлементарных регрессионных моделей* // Int. Journal of Open Information Technologies. – 2021. – Т. 9, №5. – С. 30–35.
3. БАЗИЛЕВСКИЙ М.П. *Оценивание линейно-неэлементарных регрессионных моделей с помощью метода наименьших квадратов* // Моделирование, оптимизация и информационные технологии. – 2020. – Т. 8, №4(31).
4. БАЗИЛЕВСКИЙ М.П. *Программа построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессионных моделей* // Труды Института системного программирования РАН. – 2023. – Т. 35, №4. – С. 129–144.
5. БАЗИЛЕВСКИЙ М.П. *Технология построения вполне интерпретируемых квазилинейных регрессионных моделей* // Прикладная математика и вопросы управления. – 2024. – №1. – С. 123–138.

6. КЛЕЙНЕР Г.Б. *Производственные функции: Теория, методы, применение*. – М.: Финансы и статистика, 1986. – 239 с.
7. НОСКОВ С.И. *Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных*. – Иркутск: РИЦ ГП «Облинформпечать», 1996. – 320 с.
8. AHMAD S.A., RAFIQ S.K., AHMED H.U. et al. *Innovative soft computing techniques including artificial neural network and nonlinear regression models to predict the compressive strength of environmentally friendly concrete incorporating waste glass powder* // Innovative Infrastructure Solutions. – 2023. – Vol. 8, No. 4. – P. 119.
9. CHEN D., HU F., NIAN G. et al. *Deep residual learning for nonlinear regression* // Entropy. – 2020. – Vol. 22, No. 2. – P. 193.
10. FENG Y., LIU Q., YAO Q. et al. *Model averaging for nonlinear regression models* // Journal of Business & Economic Statistics. – 2022. – Vol. 40, No. 2. – P. 785–798.
11. JAF D.K.I., ABDALLA A., MOHAMMED A.S. et al. *Hybrid nonlinear regression model versus MARS, MEP, and ANN to evaluate the effect of the size and content of waste tire rubber on the compressive strength of concrete* // Heliyon. – 2024. – Vol. 10, No. 4.
12. KONNO H., YAMAMOTO R. *Choosing the best set of variables in regression analysis using integer programming* // Journal of Global Optimization. – 2009. – Vol. 44. – P. 273–282.
13. MAZUMDER R., RADCHENKO P., DEDIEU A. *Subset selection with shrinkage: Sparse linear modeling when the SNR is low* // Operations Research. – 2023. – Vol. 71, No. 1. – P. 129–147.
14. MILLER A. *Subset selection in regression*. – Chapman and hall/CRC, 2002.
15. MOLNAR C. *Interpretable machine learning*. – Lulu.com, 2020.
16. MONTGOMERY D.C., PECK E.A., VINING G.G. *Introduction to linear regression analysis*. – John Wiley & Sons, 2021.
17. OMOGOROYE O.O., OLANIYI O.O., ADEBIYI O.O. et al. *Electricity consumption (kW) forecast for a building of interest*

based on a time series nonlinear regression model // Asian Journal of Economics, Business and Accounting. – 2023. – Vol. 23, No. 21. – P. 197–207.

18. SHARIFANI K., AMINI M. *Machine learning and deep learning: A review of methods and applications* // World Information Technology and Engineering Journal. – 2023. – Vol. 10, No. 7. – P. 3897–3904.
19. TUFANER F., DEMIRCI Y. *Prediction of biogas production rate from anaerobic hybrid reactor by artificial neural network and nonlinear regressions models* // Clean Technologies and Environmental Policy. – 2020. – Vol. 22. – P. 713–724.
20. WEN L., LI Y., CHAI J. *Multiple nonlinear regression models for predicting deformation behavior of concrete-face rockfill dams* // Int. Journal of Geomechanics. – 2021. – Vol. 21, No. 2. – P. 04020253.
21. WOLSEY L.A. *Integer programming*. – John Wiley & Sons, 2020.
22. YANG G.R., WANG X.J. *Artificial neural networks for neuroscientists: a primer* // Neuron. – 2020. – Vol. 107, No. 7. – P. 1048–1070.
23. YANG Y., DU L., LI Q. et al. *Vibration prediction and analysis of the main beam of the TBM based on a multiple linear regression model* // Scientific Reports. – 2024. – Vol. 14, No. 1. – P. 3498.
24. ZHANG Z., YIN Z., CHEN Y. et al. *Evaluation and prediction of water resources carrying capacity using a multiple linear regression model in Taizhou City, China* // Human and Ecological Risk Assessment: An International Journal. – 2023. – Vol. 29, No. 2.– P. 553–570.
25. ZITAR R.A., MOHSEN A., SEGHROUCHNI A.E. et al. *Intensive review of drones detection and tracking: linear kalman filter versus nonlinear regression, an analysis case* // Archives of Computational Methods in Engineering. – 2023. – Vol. 30, No. 5. – P. 2811–2830.

CONSTRUCTING NON-ELEMENTARY QUASILINEAR REGRESSIONS USING MATHEMATICAL PROGRAMMING APPARATUS

Mikhail Bazilevskiy, Irkutsk State Transport University, Irkutsk, Candidate of Technical Sciences, Assistant Professor (mik2178@yandex.ru).

Abstract: In non-elementary quasilinear regressions, the explanatory variables are first transformed using elementary functions, after which the pairs of resulting factors are again transformed using the non-elementary functions min and max. Such models are nonlinear in both factors and parameters, so even their estimation seems to be a complex computational task. And if the composition of the variables included in the model, as well as their elementary and non-elementary transformations, is unknown, then the complexity of the problem increases significantly. This study aims to solve this problem. Instead of labor-intensive exhaustive search procedures, a well-developed mathematical programming apparatus has been used recently. The method for constructing non-elementary quasilinear regressions is formalized as a mixed 0-1 integer linear programming problem. The proposed method is implemented in a special computer program. Its advantage is that the user can regulate the number of transformed variables during the construction process, so the program can be used both for solving simple control problems on ordinary personal computers and for processing large data arrays using cloud services. Non-elementary quasilinear regressions can be used to solve control problems in technical, socio-economic, medical and other systems.

Keywords: regression analysis, nonlinear regression, non-elementary quasilinear regression, ordinary least squares method, mixed 0-1 linear programming problem.

УДК 519.862.6
ББК 22.18

Статья представлена к публикации
членом редакционной коллегии Р.М. Нижегородцевым.

Поступила в редакцию 24.07.2024.
Опубликована 30.11.2024.