DOI: 10.15514/ISPRAS-2023-35(4)-7

Программа построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессионных моделей

М.П. Базилевский, ORCID: 0000-0002-3253-5697 <*mik2178@yandex.ru*>

Иркутский государственный университет путей сообщения,
664074, Россия, г. Иркутск, ул. Чернышевского, д. 15.

Аннотация. Вполне интерпретируемая линейная регрессия удовлетворяет следующим условиям: знаки её коэффициентов соответствуют содержательному смыслу факторов; мультиколлинеарность незначительна; коэффициенты значимы; качество аппроксимации модели высокое. Ранее для построения таких моделей, оцениваемых с помощью метода наименьших квадратов, была разработана программа ВИнтер-1. В ней по заданным начальным параметрам автоматически формируется задача частично-булевого линейного программирования, в результате решения которой осуществляется отбор наиболее информативных регрессоров. Лежащий в основе этой программы математический аппарат со временем был существенно расширен: были разработаны неэлементарные линейные регрессии, для контроля мультиколлинеарности были предложены линейные ограничения на абсолютные величины интеркорреляций, появились предположения о возможности построения не только линейных, но и квазилинейных регрессий. Данная статья посвящена описанию разработанной второй версии программы построения вполне интерпретируемых регрессий ВИнтер-2. Программа ВИнтер-2 позволяет в зависимости от выбранных пользователем начальных параметров автоматически формулировать для решателя LPSolve задачи частично-булевого линейного программирования для построения как элементарных, так и неэлементарных вполне интерпретируемых квазилинейных регрессий. Предусмотрена возможность задания до девяти элементарных функций и контроля таких параметров, как число регрессоров в модели, число знаков в вещественных числах после запятой, абсолютные вклады переменных в общую детерминацию, число вхождений объясняющих переменных в модель и величины интеркорреляций. В процессе работы с программой также можно контролировать количество элементарно и неэлементарно преобразованных переменных, влияющих на скорость решения задачи частично-булевого линейного программирования. Программа ВИнтер-2 универсальна и может применяться для построения вполне интерпретируемых математических зависимостей в различных предметных областях.

Ключевые слова: линейная регрессия; вполне интерпретируемая регрессия; метод наименьших квадратов; мультиколлинеарность; интеркорреляция; квазилинейная регрессия; неэлементарная регрессия; отбор информативных регрессоров; задача частично-булевого линейного программирования; критерий нелинейности.

Для цитирования: Базилевский М.П. Программа построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессионных моделей. Труды ИСП РАН, том 35, вып. 4, 2023 г., стр. 129–144. DOI: 10.15514/ISPRAS–2023–35(4)–7.

Program for Constructing Quite Interpretable Elementary and Non-elementary Quasi-linear Regression Models

M.P. Bazilevskiy, ORCID: 0000-0002-3253-5697 <*mik2178@yandex.ru*>

Irkutsk State Transport University,
15, Chernyshevskogo st., Irkutsk, 664074, Russia.

Abstract. A quite interpretable linear regression satisfies the following conditions: the signs of its coefficients correspond to the meaningful meaning of the factors; multicollinearity is negligible; coefficients are significant; the quality of the model approximation is high. Previously, to construct such models, estimated using the ordinary least squares, the QInter-1 program was developed. In it, according to the given initial parameters, the mixed integer 0-1 linear programming task is automatically generated, as a result of which the most informative regressors are selected. The mathematical apparatus underlying this program was significantly expanded over time: non-elementary linear regressions were developed, linear restrictions on the absolute values of intercorrelations were proposed to control multicollinearity, assumptions appeared about the possibility of constructing not only linear, but also quasi-linear regressions. This article is devoted to the description of the developed second version of the program for constructing quite interpretable regressions QInter-2. The QInter-2 program allows, depending on the initial parameters selected by the user, to automatically formulate for the LPSolve solver the mixed integer 0-1 linear programming task for constructing both elementary and non-elementary quite interpretable quasi-linear regressions. It is possible to set up to nine elementary functions and control such parameters as the number of regressors in the model, the number of signs in real numbers after the decimal point, the absolute contributions of variables to the overall determination, the number of occurrences of explanatory variables in the model, and the magnitude of intercorrelations. In the process of working with the program, you can also control the number of elementary and non-elementarily transformed variables that affect the speed of solving the mixed integer 0-1 linear programming task. The QInter-2 program is universal and can be used to construct quite interpretable mathematical dependencies in various subject areas.

Keywords: linear regression; quite interpretable regression; ordinary least squares; multicollinearity; intercorrelation; quasi-linear regression; non-elementary regression; subset selection in regression; mixed integer 0-1 linear programming; non-linearity criterion.

For citation: Bazilevskiy M.P. Program for constructing quite interpretable elementary and non-elementary quasi-linear regression models. *Trudy ISP RAN/Proc. ISP RAS*, vol. 35, issue 4, 2023. pp. 129-144 (in Russian). DOI: 10.15514/ISPRAS-2023-35(4)-7.

1. Введение

Построение интерпретируемых моделей машинного обучения [1,2] в настоящее время является актуальной научной задачей. Наиболее просто интерпретируемыми моделями машинного обучения справедливо можно считать линейные регрессионные модели [3]. В результате оценивания линейной регрессии всегда можно объяснить любой её коэффициент, за исключение, быть может, свободного члена. Но, несмотря на это, не каждую оцененную линейную регрессию можно отнести к интерпретируемым. Причины этого состоят в следующем:

1) мультиколлинеарность [4], которая затрудняет оценку степени влияния входных переменных на выходную и может искажать знаки оценок линейной регрессии;
2) неверное направление влияния входных переменных на выходную даже при отсутствии мультиколлинеарности, что говорит об игнорировании исследователем предварительного анализа влияния факторов на отклик;
3) незначимость некоторых коэффициентов линейной регрессии, что делает бессмысленной интерпретацию влияния на зависимую переменную y соответствующих объясняющих переменных;
4) низкое качество аппроксимации построенной линейной регрессии, которое сводит на нет весь процесс интерпретации.

Как отмечено в [1], универсального математического определения интерпретируемости модели машинного обучения не существует. Поэтому для определенности автором был введен термин “вполне интерпретируемая линейная регрессия” [5]. Такая регрессия удовлетворяет следующим условиям:

1) знаки коэффициентов оцененной модели соответствуют содержательному смыслу входящих в уравнение факторов;
2) эффект мультиколлинеарности незначителен.

Это определение можно дополнить ещё двумя условиями: все коэффициенты должны быть значимы, например, по t-критерию Стьюдента; качество аппроксимации линейной регрессии должно быть высоким.

Естественным образом, для построения вполне интерпретируемых линейных регрессий требуется соответствующее математическое и программное обеспечение.

Существующих на сегодняшний день алгоритмов построения интерпретируемых регрессионных моделей мало. Можно выделить лишь алгоритм, предложенный в монографии [6]. Множество работ посвящено лишь отбору информативных регрессоров (ОИР) [7] в линейной регрессии, что не гарантирует возможность её интерпретации. В [6] для решения задачи ОИР применен метод “всех регрессий” [8], который заключается в переборе всех возможных альтернативных вариантов моделей и выборе лучшей из них со всеми значимыми коэффициентами, минимальной остаточной дисперсией отклика и наилучшей интерпретируемостью. При этом линейные регрессии представляются в стандартизованном масштабе, что, как известно [9], упрощает процесс их оценивания с помощью метода наименьших квадратов (МНК). Тем не менее, метод “всех регрессий” трудно назвать эффективным.

За последние десятилетия была существенно развита технология решения задач частично-целочисленного программирования. При этом задача ОИР в линейной регрессии, оцениваемой с помощью МНК, в зарубежной литературе часто сводится к задаче частично-булевого квадратичного программирования [10–19]. Автору такую задачу удалось свести к задаче частично-булевого линейного программирования (ЧБЛП) [5, 20–24]. Постепенно получилось расширить эту задачу линейными ограничениями так, чтобы она гарантировала построение вполне интерпретируемой линейной регрессии. Предложенный метод построения вполне интерпретируемых линейных регрессий был реализован в виде программы ВИнтер-1. После чего задача расширялась, и в [23,24] был предложен метод построения вполне интерпретируемых неэлементарных линейных регрессий (НЛР). Также параллельно возникла идея в возможности построения не только элементарных и неэлементарных линейных, но и квазилинейных регрессий.

Целью данной статьи является описание разработанной второй версии программы ВИнтер-1, предназначенной для построения вполне интерпретируемых элементарных и неэлементарных линейных и квазилинейных регрессионных моделей методом решения автоматически сформулированной задачи ЧБЛП.

2. Элементарные и неэлементарные квазилинейные регрессии

Рассмотрим модель множественной линейной регрессии с объясняющими переменными x_j, $j=\overline{1,l}$:

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \varepsilon_i, \qquad i=\overline{1,n}, \tag{1}$$

где y – объясняемая переменная; n – объем выборки; ε – вектор ошибок аппроксимации; α_0, α_1, ..., α_l – неизвестные параметры.

Для увеличения скорости вычислений МНК-оценок линейной регрессии (1) проведем нормирование всех переменных по формулам

$$y_i^{\bullet} = \frac{y_i - \overline{y}}{\sigma_y},\ x_{i1}^{\bullet} = \frac{x_{i1} - \overline{x_1}}{\sigma_{x_1}},\ ...,\ x_{il}^{\bullet} = \frac{x_{il} - \overline{x_l}}{\sigma_{x_l}},\qquad i = \overline{1,n},$$

где $\overline{y}$, $\overline{x_1}$, ..., $\overline{x_l}$ – средние значения переменных; σ_y, σ_{x_1}, ..., σ_{x_l} – среднеквадратичные отклонения переменных.

С использованием нормированных переменных введем для модели (1) стандартизованную линейную регрессию:

$$y_i^{\bullet} = \beta_1 x_{i1}^{\bullet} + \beta_2 x_{i2}^{\bullet} + ... + \beta_l x_{il}^{\bullet} + \varepsilon_i^{\bullet},\qquad i = \overline{1,n}, \tag{2}$$

где β_1, β_2, ..., β_l – неизвестные параметры; $\varepsilon^{\bullet}$ – вектор ошибок аппроксимации.

В [5] на основе модели (2) предложена следующая задача ЧБЛП построения вполне интерпретируемой линейной регрессии:

$$R^2 = \sum_{j=1}^{l} r_{yx_j} \cdot \beta_j \to \max, \tag{3}$$

$$-(1-\delta_j)\cdot M \le \sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k - r_{yx_j} \le (1-\delta_j)\cdot M,\quad j = \overline{1,l}, \tag{4}$$

$$0 \le \beta_j \le \delta_j \cdot M,\qquad j \in J^+, \tag{5}$$

$$-\delta_j \cdot M \le \beta_j \le 0,\qquad j \in J^-, \tag{6}$$

$$\delta_j \in \{0,1\},\quad j = \overline{1,l}, \tag{7}$$

$$C_{x_j}^{\text{абс}} = r_{yx_j} \cdot \beta_j \ge \theta \cdot \delta_j,\quad j = \overline{1,l}, \tag{8}$$

где R^2 – коэффициент детерминации; $r_{x_i x_j}$ – коэффициент корреляции (интеркорреляция) между i-й и j-й объясняющей переменной; r_{yx_j} – коэффициент корреляции между объясняемой переменной y и j-й объясняющей переменной; M – большое положительное число; δ_j – бинарная переменная, принимающая значение 0, если j-я переменная не входит в модель, и значение 1, если входит; J^- и J^+ – множества объясняющих переменных, удовлетворяющих условиям $r_{yx_j} < 0$ и $r_{yx_j} > 0$ соответственно; $C_{x_j}^{\text{абс}} = r_{yx_j} \cdot \beta_j$ – абсолютный вклад j-й переменной в общую детерминацию R^2; $\theta \ge 0$ – нижняя граница абсолютных вкладов.

Заметим, что ограничения (4) предназначены для включения/исключения уравнений в систему линейных алгебраических уравнений $\sum_{k=1}^{l} r_{x_j x_k} \cdot \beta_k = r_{yx_j}$, $j = \overline{1,l}$, с помощью которой находятся МНК-оценки. Если $\delta_j = 0$, то из этой системы исключается j-е уравнение и коэффициент β_j принимает значение 0, то есть осуществляется МНК-оценивание без участия переменной x_j. Если $\delta_j = 1$, то из системы j-е уравнение не исключается и коэффициент β_j может принимать любое значение, согласованное по знаку с коэффициентом корреляции r_{yx_j}, то есть осуществляется МНК-оценивание с участием переменной x_j.

В [20] задача ЧБЛП (3) – (8) дополнена ограничениями на абсолютные величины интеркорреляций:

$$\left|r_{x_i x_j}\right|\left(\delta_i+\delta_j-1\right)\le r\,,\quad i=\overline{1,l-1}\,,\qquad j=\overline{i+1,l}\,, \tag{9}$$

где $0<r\le 1$ – верхняя граница интеркорреляций.

Решение задачи ЧБЛП (3) – (9) приводит к построению вполне интерпретируемой линейной регрессии с оптимальным по критерию R^2 количеством объясняющих переменных из множества Φ, в которой $\tilde{\beta}_j\cdot r_{yx_j}>0$, $j\in\Phi$, вклады $C_{x_j}^{\text{абс}}\ge\theta$, $j\in\Phi$, и интеркорреляции $\left|r_{x_i x_j}\right|\le r$, $i,j\in\Phi$.

В задаче (3) – (9) для контроля значимости используются абсолютные вклады переменных, а для контроля мультиколлинеарности – величины интеркорреляций. При желании исследователь может дополнить формализацию другими известными критериями. Например, в [21] обсуждается, как в задаче (3) – (9) можно контролировать факторы вздутия дисперсии (VIF), а в [22] – значимость оценок по t-критерию Стьюдента.

Отметим, что выбор в задаче (3) – (9) значения параметра M обсуждается в [5].

Аналогично можно сформулировать задачу ЧБЛП для построения вполне интерпретируемой квазилинейной регрессии. Пусть в распоряжении исследователя имеется elem элементарных функций: $f_1(x)$, $f_2(x)$, ..., $f_{\text{elem}}(x)$. Пусть $x_{ij}>0$, $i=\overline{1,n}$, $j=\overline{1,l}$. С помощью этих функций получим множество элементарно преобразованных переменных z_{jk}, $j=\overline{1,l}$, $k=\overline{1,\text{elem}}$, где z_{jk} – k-е преобразование j-й переменной. Тогда модель квазилинейной регрессии можно записать в виде:

$$y_i=\alpha_0+\sum_{j=1}^{l}\sum_{k=1}^{\text{elem}}\alpha_{jk}\cdot z_{ijk}+\varepsilon_i\,,\qquad i=\overline{1,n}\,. \tag{10}$$

Предварительно для построения модели (10) необходимо сократить список z-переменных по следующим причинам.

1. Иногда знак коэффициента корреляции преобразованной переменной с y может противоречить содержательному смыслу задачи.
2. Преобразование может быть в значительной степени нелинейным, что затрудняет объяснение соответствующего коэффициента.

Допустим, что все элементарные функции монотонны на отрезках $\left[x_{\min}^j,x_{\max}^j\right]$, $j=\overline{1,l}$. Тогда степень нелинейности z-переменных можно оценить с помощью следующих критериев нелинейности [25]:

$$NC_{z_{jk}}=\left|\frac{f_k(x_{\max}^j)+f_k(x_{\min}^j)}{f_k(x_{\max}^j)-f_k(x_{\min}^j)}-\frac{2\int_{x_{\min}^j}^{x_{\max}^j}f_k(x_j)dx_j}{\left(x_{\max}^j-x_{\min}^j\right)\left(f_k(x_{\max}^j)-f_k(x_{\min}^j)\right)}\right|,\ j=\overline{1,l}\,,\ k=\overline{1,\text{elem}}\,. \tag{11}$$

Если $NC_{z_{jk}}\le 0{,}2$, то вместо оценки $\tilde{\alpha}_{jk}$ при k-м преобразовании j-й переменной в модели (10) можно объяснить величину $\tilde{\alpha}_{jk}\dfrac{f_k\left(x_{\max}^j\right)-f_k\left(x_{\min}^j\right)}{x_{\max}^j-x_{\min}^j}$.

Таким образом, сократив список z -переменных, несложно по аналогии с задачей ЧБЛП (3) – (9) сформулировать задачу построения вполне интерпретируемой квазилинейной регрессии. Для возможности её интерпретации необходимо дополнить задачу ЧБЛП ограничением на единственность вхождения каждой объясняющей переменной в модель.

В [23] предложена неэлементарная линейная регрессия (НЛР) вида

$$y_i = \alpha_0 + \sum_{j=1}^{l} \alpha_j x_{ij} + \sum_{j=1}^{p} \alpha_j^{\min} \min\{x_{i,\mu_{j1}}, k_j^{\min} x_{i,\mu_{j2}}\} +$$

$$+ \sum_{j=1}^{p} \alpha_j^{\max} \max\{x_{i,\mu_{j1}}, k_j^{\max} x_{i,\mu_{j2}}\} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{12}$$

где $p = C_l^2$ – количество комбинаций пар объясняющих переменных; min, max – бинарные операции, возвращающие минимум и максимум из двух чисел; μ_{j1}, μ_{j2}, $j = \overline{1,p}$ – элементы матрицы $\mathrm{M}_{p\times 2}$, содержащей в строках все возможные комбинации пар объясняющих переменных; $\alpha_j^{\min}$, $\alpha_j^{\max}$, $k_j^{\min}$, $k_j^{\max}$, $j = \overline{1,p}$ – неизвестные параметры.

Оптимальные оценки параметров $k_j^{\min}$, $k_j^{\max}$, $j = \overline{1,p}$, лежат внутри промежутков [23]

$$\left(k_{\text{нижн}}^j, k_{\text{верхн}}^j\right), \qquad j = \overline{1,p},$$

где $k_{\text{нижн}}^j = \min\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}}, ..., \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\}$, $k_{\text{верхн}}^j = \max\left\{\frac{x_{1,\mu_{j1}}}{x_{1,\mu_{j2}}}, \frac{x_{2,\mu_{j1}}}{x_{2,\mu_{j2}}}, ..., \frac{x_{n,\mu_{j1}}}{x_{n,\mu_{j2}}}\right\}$.

Разбивая эти промежутки на Razb точек, методом полного перебора можно получить приближенные МНК-оценки НЛР (12).

В [23] задача построения НЛР (12) сведена к задаче ЧБЛП. А в [24] эта задача расширена ограничениями, позволяющими строить вполне интерпретируемую НЛР. Для этого предварительно необходимо исключать неэлементарно преобразованные переменные, знак коэффициента корреляции которых с y не согласуется хотя бы с одним из знаков коэффициентов корреляции с y входящих в это преобразование переменных.

На основе (10) введем неэлементарную квазилинейную регрессию (НКР) вида

$$y_i = \alpha_0 + \sum_{j=1}^{l} \sum_{k=1}^{\text{elem}} \alpha_{jk} z_{ijk} + \sum_{j=1}^{p^*} \alpha_j^{\min} \min\{z_{i,\mu^*_{j,1,1},\mu^*_{j,1,2}}, k_j^{\min} z_{i,\mu^*_{j,2,1},\mu^*_{j,2,2}}\} +$$

$$+ \sum_{j=1}^{p^*} \alpha_j^{\max} \max\{z_{i,\mu^*_{j,1,1},\mu^*_{j,1,2}}, k_j^{\max} z_{i,\mu^*_{j,2,1},\mu^*_{j,2,2}}\} + \varepsilon_i, \qquad i = \overline{1,n}, \tag{13}$$

где $p^* = C_{l\cdot\text{elem}}^2$ – количество комбинаций пар элементарно преобразованных переменных; $\mu^*_{j,1,1}$, $\mu^*_{j,1,2}$, $\mu^*_{j,2,1}$, $\mu^*_{j,2,2}$, $j = \overline{1,p^*}$ – элементы трехмерного массива $\mathrm{M}^*_{p^*,2,2}$, содержащего на горизонтальных листах в первом столбце номера переменных, во втором – номера преобразований; $\alpha_j^{\min}$, $\alpha_j^{\max}$, $k_j^{\min}$, $k_j^{\max}$, $j = \overline{1,p^*}$ – неизвестные параметры.

Задача построения вполне интерпретируемой НКР (13) по аналогии с приёмами, рассмотренными в [24], сводится к задаче ЧБЛП. При этом для обеспечения интерпретируемости НКР требуется из трехмерного массива $\mathrm{M}^*_{p^*,2,2}$ исключать горизонтальные листы, содержащие одну и ту же объясняющую переменную.

3. Программа ВИнтер-2

Программа построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессий (ВИнтер-2) была разработана в среде программирования Delphi.

Блок-схема алгоритма работы ВИнтер-2 представлена на рис. 1. После запуска программы сначала нужно загрузить текстовый файл формата ".txt" со статистическими данными. В первом столбце этого файла содержатся значения объясняемой переменной y, во втором – объясняющей переменной x_1 и так далее. Столбцы отделяются друг от друга табуляцией. Имена переменных в первой строке вводить не нужно. Целые и дробные части действительных чисел отделяются друг от друга символом ",". Важно, чтобы знаки коэффициентов корреляции r_{yx_j}, $j=\overline{1,l}$, были согласованы с содержательным смыслом всех расположенных в текстовом файле переменных, т.е. предварительно с факторами должны поработать эксперты из данной предметной области.

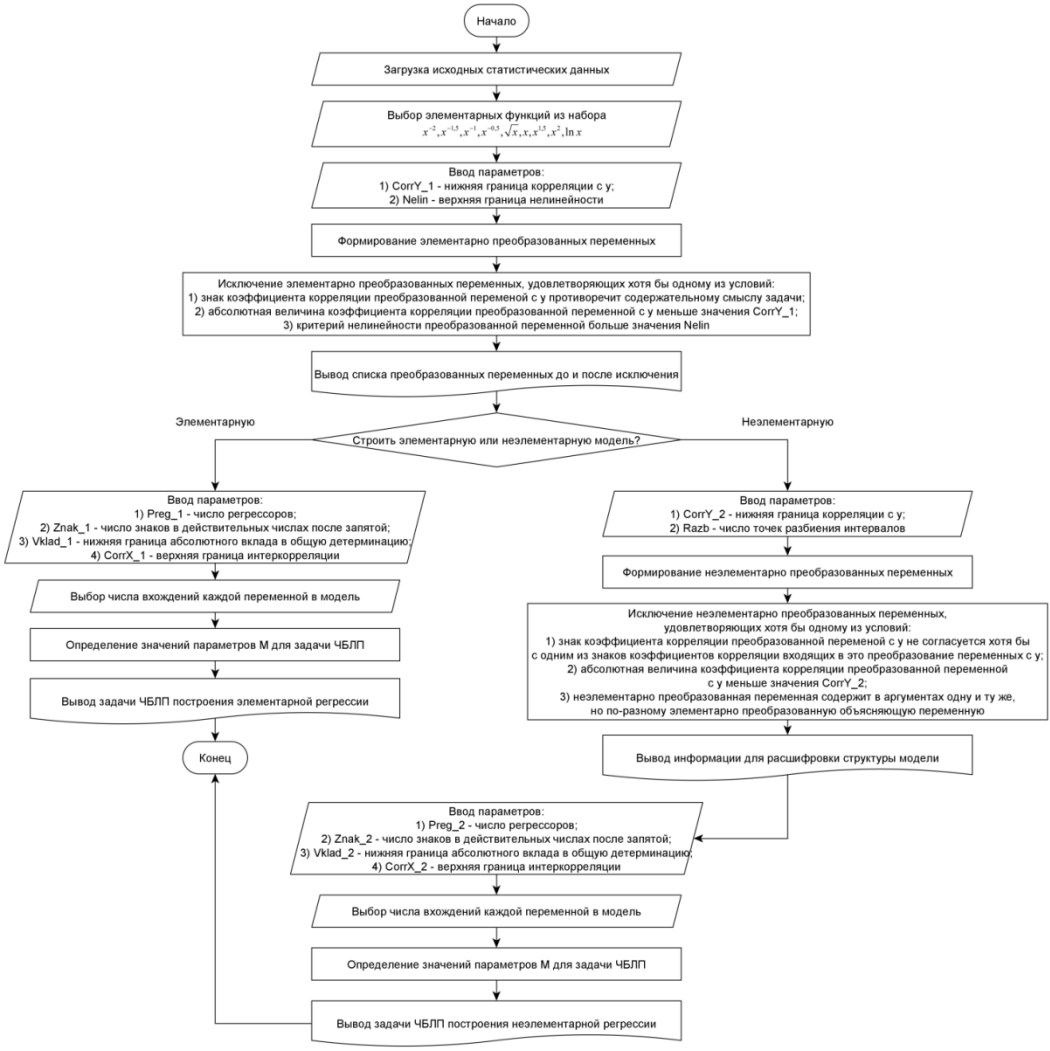

Рис. 1. Блок-схема алгоритма работы программы
Fig. 1. Block diagram of the program operation algorithm

Если всё сделано верно, то данные отобразятся в поле “Данные” главной формы системы (рис. 2). При этом будет сформирована матрица D, содержащая наблюдаемые значения переменных, корреляционная матрица KM для матрицы D, а также матрица MinMax, содержащая в первой и второй строке минимальные и максимальные значения всех переменных, и матрица MNC размера $l \times 9$, содержащая значения всех критериев нелинейности (11) переменных, преобразованных с помощью элементарных функций x^{-2}, $x^{-1,5}$, x^{-1}, $x^{-0,5}$, $x^{0,5}$, x, $x^{1,5}$, x^2 и $\ln x$.

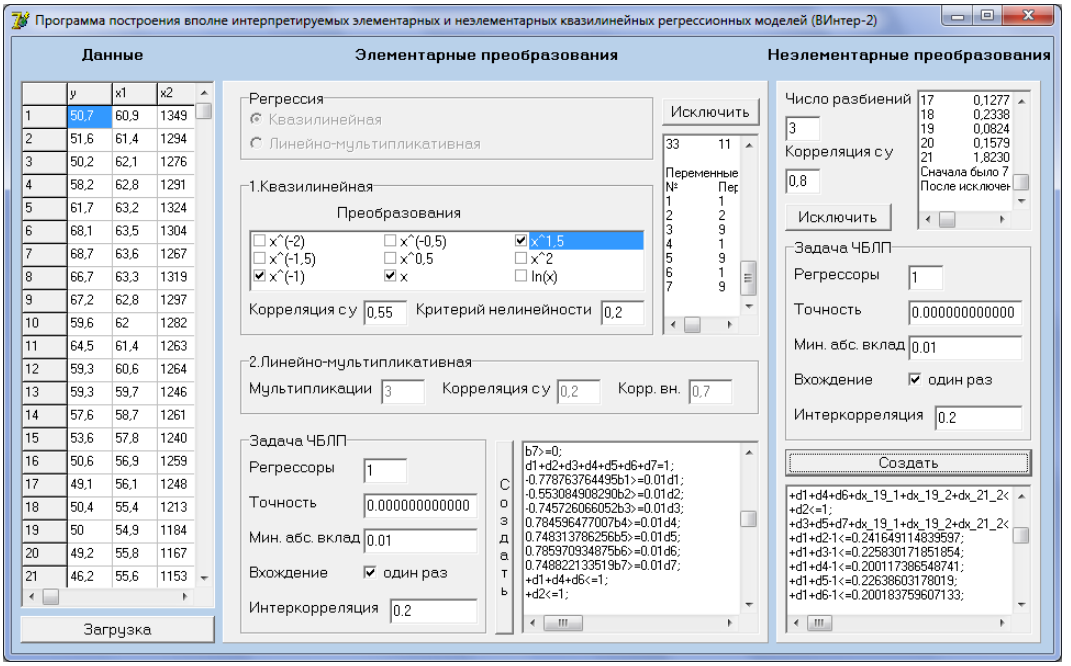

Рис. 2. Главная форма программы
Fig. 2. Main form of the program

Далее на панели “1.Квазилинейная” необходимо выбрать элементарные функции для преобразования переменных, кликнув на соответствующие флажки, а также ввести вручную нижнюю границу корреляции с y – CorrY_1, и верхнюю границу нелинейности – Nelin. Затем нажать кнопку “Исключить”. В результате нажатия программа выполняет следующую последовательность действий.

1. Формируется вектор номеров выбранных элементарных функций CheckPreobr из s элементов.
2. Формируется матрица данных D2 размера $n \times (1 + s \cdot l)$, содержащая значения всех элементарно преобразованных переменных.
3. Формируется корреляционная матрица KM2 по матрице данных D2.
4. Формируется матрица Shifr размера $(l \cdot s) \times 5$, в первом столбце которой содержатся номера исходных переменных, во втором – номера элементарных функций из вектора CheckPreobr. Остальные столбцы заполняются единицами и нулями по следующим правилам. Если знак коэффициента корреляции преобразованной переменной с y не противоречит содержательному смыслу, то в третий столбец вносится “0”, иначе “1”. Если абсолютная величина коэффициента корреляции преобразованной переменной с y не меньше значения CorrY_1, то в четвертый столбец вносится “0”, иначе “1”.

Если значения критерия нелинейности преобразованной переменной не больше значения Nelin, то в пятый столбец вносится “0”, иначе “1”. Например, если матрица Shifr имеет вид

$$\begin{pmatrix} 2 & 3 & 1 & 0 & 1 \\ 4 & 7 & 0 & 1 & 1 \end{pmatrix},$$

то это означает, что преобразование x_2^{-1} противоречит содержательному смыслу задачи, критерий нелинейности выше необходимого значения Nelin, но коэффициент корреляции с y по модулю выше требуемого значения CorrY_1; преобразование $x_4^{1,5}$ слабо коррелирует с y, имеет высокую степень нелинейности, но не противоречит содержательному смыслу задачи.

5. Осуществляется исключение тех преобразований, которые либо противоречат содержательному смыслу задачи, либо слабо коррелируют с y, либо имеют высокую степень нелинейности, т.е. из матрицы Shifr исключаются те строки, в которых сумма элементов последних трех столбцов больше нуля. В результате исключения из матрицы Shifr формируется новая матрица GShifr размера $\text{Lev1Per} \times 5$, где Lev1Per – число преобразований, для которых выполняются все 3 условия. Также формируется новая матрица данных D3.
6. В соответствующем поле главной формы ВИнтер-2 выводятся элементы матриц Shifr и GShifr. Тем самым, регулируя параметры CorrY_1 и Nelin, можно снижать количество участвующих в процессе моделирования элементарно преобразованных переменных, что повышает скорость построения регрессионной модели.

Заметим, во-первых, что пока в систему встроено только 9 элементарных функций. Причём, хотя бы одна из них обязательно должна быть выбрана. Для построения линейной регрессии нужно выбрать только один флажок с именем “x”. Во-вторых, если выбрать CorrY_1 = 0, Nelin = 1, то требования для сильной корреляции и слабой нелинейности выбранных преобразований будут игнорироваться. В-третьих, как видно по главной форме (рис. 2), неактивна панель “2.Линейно-мультипликативная”. Механизм построения такого вида регрессий пока находится на стадии тестирования.

После исключения элементарно преобразованных переменных нужно выбрать, какая модель будет строиться – элементарная (10) или неэлементарная (13). Если пользователь желает построить элементарную регрессию (10), то на панели “Задача ЧБЛП”, расположенной в нижней части главной формы, нужно задать следующие параметры: Preg_1, Znak_1, Vklad_1, CorrX_1. Описание этих параметров представлено на рис. 1. Также для обеспечения интерпретируемости модели нужно выбрать флажок “один раз”, означающий, что каждая объясняющая переменная должна входить в регрессию не более одного раза. Затем нужно нажать кнопку “Создать”. В результате нажатия программа выполняет следующую последовательность действий.

1. Формируется корреляционная матрица KM3 по матрице данных D3.
2. Определяются значения параметров M для линейных ограничений типа (5), (6) [5]. В результате формируется вектор Mb из Lev1Per элементов. Затем автоматически определяются значения параметров M для линейных ограничений типа (4) [5]. Для этого в цикле формируются необходимые задачи линейного программирования, которые последовательно передаются решателю LPSolve IDE, а результаты решения снова возвращаются в ВИнтер-2. В результате формируется матрица Mq размера $\text{Lev1Per} \times 2$, в строках которой содержатся нижние и верхние границы параметров M для ограничений типа (4).
3. Осуществляется формализация задачи ЧБЛП для пакета LPSolve. Формируется целевая функция типа (3) и линейные ограничения типа (4) – (9). Для обеспечения

единственности вхождения в модель каждой объясняющей переменной формируется матрица GVh размера $\text{Lev1Per} \times l$, состоящая из нулей и единиц. Эта матрица связывает преобразованные переменные с исходными факторами и заполняется по следующему правилу: если в k-й строке матрицы GShifr во втором столбце стоит номер h, то в матрице GVh на пересечении k-й строки и h-го столбца ставится “1”. В этой связи линейные ограничения на количество входящих в модель объясняющих переменных имеют вид:

$$\sum_{k=1}^{\text{Lev1Per}} \text{GVh}_{kj} \cdot \delta_k \le 1, \quad j = \overline{1, l}. \tag{14}$$

4. В соответствующем поле главной формы программы ВИнтер-2 выводится сформулированная задача ЧБЛП для построения элементарной регрессии (10).

Если же пользователь желает построить неэлементарную регрессию (13), то сначала необходимо из оставшихся на первом шаге элементарно преобразованных переменных сформировать неэлементарно преобразованные переменные. Для этого на панели, расположенной в правом верхнем углу главной формы, нужно задать следующие параметры: Razb – число точек разбиения интервалов $\left(k^j_{\text{нижн}}, k^j_{\text{верхн}}\right)$, $j = \overline{1, p^*}$, CorrY_2 – нижняя граница корреляции с y. С помощью этих параметров регулируется число неэлементарно преобразованных переменных. Затем нужно нажать кнопку “Исключить”. В результате нажатия система выполняет следующую последовательность действий.

1. Формируется матрица Ind размера $C^2_{\text{Lev1Per}} \times 2$, содержащая по строкам все возможные комбинации пар элементарно преобразованных переменных.
2. С помощью матрицы Ind формируется матрица LAM размера $C^2_{\text{Lev1Per}} \times 2$, содержащая в строках нижнюю и верхнюю границу параметров $k^{\min}_j$, $k^{\max}_j$, $j = \overline{1, p^*}$, для неэлементарных преобразований переменных.
3. С помощью матрицы LAM формируется матрица R размера $C^2_{\text{Lev1Per}} \times \text{Razb}$, содержащая в каждой строке Razb точек, равномерно разбивающих отрезки $\left(k^j_{\text{нижн}}, k^j_{\text{верхн}}\right)$, $j = \overline{1, p^*}$.
4. С помощью матриц Ind и R формируется матрица данных ND2 размера $n \times \left(1 + \text{Lev1Per} + C^2_{\text{Lev1Per}} \cdot \text{Razb} \cdot 2\right)$, содержащая в первом столбце значения переменной y, в последующих Lev1Per столбцах – значения элементарно преобразованных переменных, в последующих $C^2_{\text{Lev1Per}} \times \text{Razb}$ столбцах – значения переменных, преобразованных с помощью неэлементарной функции min, и в последующих $C^2_{\text{Lev1Per}} \times \text{Razb}$ столбцах – значения переменных, преобразованных с помощью неэлементарной функции max.
5. Формируется корреляционная матрица NKM по матрице данных ND2.
6. Формируется матрица NShifr размера $\left(\text{Lev1Per} + 2 \cdot C^2_{\text{Lev1Per}} \cdot \text{Razb}\right) \times 3$, содержащая информацию о всех преобразованных переменных. В первом её столбце содержится информация о типе преобразованной переменной (1 – если элементарная функция, 2 – если неэлементарная функция min, 3 – если неэлементарная функция max), во втором – для элементарной функции содержится номер объясняющей переменной, а для неэлементарной – номер пары объясняющих переменных в матрице Ind, в третьем – для элементарной функции запись отсутствует, а для неэлементарной функции указывается номер точки разбиения в матрице R.

7. Исключаются такие неэлементарно преобразованные переменные, для которых выполняется хотя бы одно из трех условий: знак коэффициента корреляции неэлементарно преобразованной переменной с y не согласуется со знаком хотя бы одного коэффициента корреляции входящей в данное преобразование переменной с y; значение коэффициента корреляции неэлементарно преобразованной переменной с y по модулю меньше величины CorrY_2; неэлементарная функция содержит в аргументах одну и ту же, но преобразованную разными элементарными функциями, объясняющую переменную. В результате исключения вместо матрицы NShifr формируется новая матрица GNShifr размера $(\text{Lev1Per}+\text{numU})\times 3$, где numU – количество неэлементарно преобразованных переменных, для которых не выполнено ни одно из трех перечисленных условий.
8. Из оставшихся после исключения переменных формируется матрица данных ND3 размера $n\times(1+\text{Lev1Per}+\text{numU})$.
9. В правом верхнем углу главной формы программы в соответствующем поле выводится матрица Ind и матрица R, с помощью которых после построения модели (13) можно расшифровать её структурную спецификацию. Также формируется информация о том, сколько неэлементарно преобразованных переменных было изначально, и сколько осталось после исключения.

После формирования неэлементарно преобразованных переменных на панели “Задача ЧБЛП”, расположенной в правой части главной формы ВИнтер-2, необходимо задать следующие параметры: Preg_2, Znak_2, Vklad_2, CorrX_2. Для обеспечения единственности вхождения объясняющих переменных в модель нужно выбрать флажок “один раз”. Затем нужно нажать кнопку “Создать”. В результате нажатия программа выполняет следующую последовательность действий.

1. Формируется корреляционная матрица KM по матрице данных ND3.
2. Определяются значения параметров M [23] для линейных ограничений.
3. Осуществляется формализация задачи ЧБЛП для пакета LPSolve. Имена МНК оценок делятся на 3 типа: “b” – для элементарных преобразований, “bm_” – для неэлементарных преобразований min, “bx_” – для неэлементарных преобразований max. Аналогично создаются имена бинарных переменных – “d”, “dm_” и “dx_”. Для обеспечения единственности вхождения в модель каждой объясняющей переменной с помощью матрицы GNShifr формируется матрица Vh размера $(\text{Lev1Per}+\text{numU})\times 1$, состоящая из нулей и единиц. Единица на пересечении в ней i-й строки и j-го столбца означает, что i-е преобразованием включает в себя j-ю объясняющую переменную. Линейные ограничения на количество вхождений объясняющих переменных в модель составляются аналогично (14). Также с помощью GNShifr формируется матрица Vh2 размера $(\text{Lev1Per}+\text{numU})\times\text{Lev1Per}$, в которой единица на пересечении i-й строки и j-го столбца означает, что i-е преобразование включает в себя j-ю элементарно преобразованную переменную. Помимо этого формируются ограничения на интеркорреляции регрессоров, входящих в НКР (13), и с помощью матрицы Vh2 – на интеркорреляции элементарно преобразованных переменных, входящих в НКР (13).
4. В соответствующем поле главной формы программы ВИнтер-2 выводится сформулированная задача ЧБЛП для построения неэлементарной регрессии (13).

4. Пример

Для демонстрации разработанной программы ВИнтер-2 проводилось моделирование по встроенным в эконометрических пакет Gretl данным о ценах и характеристиках произведенных в Америке автомобилей (data7-12.gdt). Эта выборка содержит одну объясняемую переменную и 10 независимых. Объем выборки составляет 82. Для возможности работы со всеми элементарными преобразованиями, из выборки были исключены фиктивные переменные hatch и trans. Оставшиеся переменные были переобозначены следующим образом: price – y, wbase – x_1, length – x_2, width – x_3, height – x_4, weight – x_5, cyl – x_6, liters – x_7, gasmpg – x_8. Корреляционная матрица для всех этих переменных представлена на рис. 3.

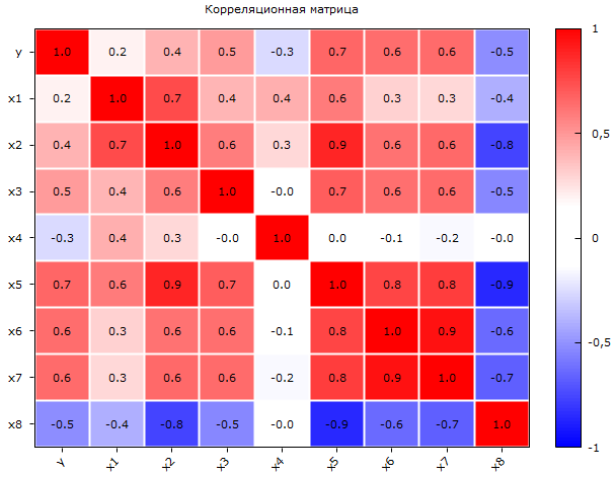

Рис. 3. Корреляционная матрица
Fig. 3. Correlation matrix

Как видно по рис. 3, есть переменные, которые очень слабо коррелируют, например, x_4 и x_5, а есть такие, которые коррелируют сильно, например, x_2 и x_5. Для уровня значимости $\alpha = 0{,}1$ было установлено, что коэффициент корреляции незначим, если его значение по модулю меньше 0,18, и значим в противном случае. Таким образом, все коэффициенты корреляции объясняющих переменных с y значимы на уровне $\alpha = 0{,}1$. Будем считать, что знаки этих коэффициентов соответствуют содержательному смыслу факторов, т.е. чем выше колесная база x_1, тем выше цена y, чем выше длина автомобиля x_2, тем выше цена y и т.д.

Сначала в ВИнтер-2 строилась элементарная линейная регрессия (1). Для исключения переменных были заданы параметры CorrY_1 = 0,18 и Nelin = 0,2. В результате ни одна объясняющая не была исключена. Для формирования задачи ЧБЛП были заданы следующие параметры: Preg_1 = 0 (означает, что нет ограничения на число регрессоров), Znak_1 = 12, Vklad_1 = 0,05, CorrX_1 = 0,18. Флажок “Вхождение” включен. Сформированная в результате задача ЧБЛП была решена в пакете LPSolve IDE. В итоге была построена следующая линейная регрессия:

$$\tilde{y} = 62{,}0819 - \underset{(-3{,}869)}{1{,}53052}\, x_4 + \underset{(8{,}610)}{1{,}26729}\, x_5, \qquad (15)$$

коэффициент детерминации которой $R^2 = 0{,}521219$. Полученное значение гораздо ниже идеального значения 1, но для выборки объема 82 это довольно неплохой результат, тем более, что модель (15) значима по критерию Фишера.

В уравнении (15) в скобках под коэффициентами указаны наблюдаемые значения t-критерия Стьюдента, подтверждающие значимость МНК-оценок для уровня $\alpha = 0,01$.

Знаки коэффициентов уравнения (15) совпадают со знаками соответствующих коэффициентов корреляции (рис. 3). А коэффициент корреляции между x_4 и x_5 незначим (рис. 3), что говорит от отсутствии мультиколлинеарности. Таким образом, выполняются все условия, чтобы можно было отнести линейную регрессию (15) ко вполне интерпретируемым.

Затем строилась элементарная квазилинейная регрессия (10). Для этого в поле "Преобразования" были выбраны все 9 элементарных функций. Для исключения переменных были заданы параметры CorrY_1 = 0,18 и Nelin = 0,2. В результате исключения осталось 50 элементарно преобразованных переменных. Для формирования задачи ЧБЛП были заданы следующие параметры: Preg_1 = 0, Znak_1 = 12, Vklad_1 = 0,05, CorrX_1 = 0,18. Флажок "Вхождение" включен. В результате была построена следующая квазилинейная регрессия:

$$\tilde{y} = -44,0643 + \underset{(4,728)}{118409}\, x_4^{-2} + \underset{(9,409)}{0,022611}\, x_5^2, \qquad (16)$$

для которой $R^2 = 0,577112$, что выше, чем у (15). Коэффициенты модели (16) значимы по t-критерию Стьюдента для уровня $\alpha = 0,01$, а их знаки не противоречат содержательному смыслу факторов. Коэффициент корреляции между регрессорами x_4^{-2} и x_5^2 равен -0,036, поэтому незначим. Из всего этого следует, что квазилинейная регрессия (16) является вполне интерпретируемой.

После чего строилась НЛР (12). Для этого использовались все 8 оставшихся после исключения на первом этапе объясняющих переменных. Для формирования и исключения неэлементарно преобразованных переменных были заданы параметры CorrY_2 = 0,18 и Razb = 3. В результате было сформировано 168 неэлементарно преобразованных переменных, которых после исключения осталось 96. Для формирования задачи ЧБЛП были заданы следующие параметры: Preg_2 = 0, Znak_2 = 12, Vklad_2 = 0,05, CorrX_2 = 0,18. Флажок "Вхождение" включен. В результате была построена следующая НЛР:

$$\tilde{y} = 50,1958 - \underset{(-3,915)}{1,50717}\, x_4 + \underset{(9,088)}{0,608225} \max\{x_3, 2.638063 x_5\}, \qquad (17)$$

для которой $R^2 = 0,546283$. Коэффициенты модели (17) значимы по t-критерию Стьюдента для уровня $\alpha = 0,01$. А если представить её в кусочно-заданной форме [23], то станет ясно, что и знаки коэффициентов не противоречат содержательному смыслу факторов. Коэффициент корреляции между регрессорами x_4 и $\max\{x_3, 2.638063 x_5\}$ равен 0,0417, поэтому незначим. Также незначимы коэффициенты корреляции $r_{x_3 x_4}$ и $r_{x_4 x_5}$ (рис. 3), следовательно, НЛР (17) относится ко вполне интерпретируемым.

Далее строилась НКР (13). Для этого использовались все 50 оставшихся после исключения на первом этапе элементарно преобразованных переменных. Для формирования и исключения неэлементарно преобразованных переменных были заданы параметры CorrY_2 = 0,7 и Razb = 3. В результате было сформировано 7350 неэлементарно преобразованных переменных, которых после исключения осталось 43. Для формирования задачи ЧБЛП были заданы следующие параметры: Preg_2 = 0, Znak_2 = 12, Vklad_2 = 0,05, CorrX_2 = 0,18. Флажок "Вхождение" включен. В результате была построена следующая НКР:

$$\tilde{y} = -247,674 + \underset{(10,78)}{94990,1} \max\{x_4^{-1,5}, 0,000534692 \ln x_2\}, \qquad (18)$$

для которой $R^2 = 0,592322$, что выше, чем у (15) – (17). Коэффициент модели (18) значим по t-критерию Стьюдента для уровня $\alpha = 0,01$, а его знак не противоречат содержательному смыслу факторов. Таким образом, НКР (18) можно считать вполне интерпретируемой.

Представим НКР (18) как кусочно-заданную функцию:

$$\tilde{y} = \begin{cases} -247,674 + 94990,1x_4^{-1,5}, \text{ при } \dfrac{x_4^{-1,5}}{\ln x_2} \geq 0,000534692, \\ -247,674 + 50,79\ln x_2, \text{ при } \dfrac{x_4^{-1,5}}{\ln x_2} < 0,000534692. \end{cases}$$

В таком виде интерпретировать влияние переменных x_2 и x_4 на y затруднительно, поскольку обе они преобразованы с помощью элементарных функций. Но известно, что критерии нелинейности этих преобразованных переменных меньше 0,2, поэтому, как отмечено выше, вместо оценок 50,79 и 94990,1 можно объяснить величины 0,287 и -7,853 соответственно, найденные по формулам $\tilde{\alpha}_{jk} \dfrac{f_k\left(x^j_{\max}\right) - f_k\left(x^j_{\min}\right)}{x^j_{\max} - x^j_{\min}}$. Тогда справедлива следующая интерпретация НКР (18).

1. Длина автомобиля (x_2) влияет на его цену (y) только при условии $\dfrac{x_4^{-1,5}}{\ln x_2} < 0,000534692$, причём, с увеличением x_2 на 10 дюймов цена y возрастает в среднем на 2873 долларов.
2. Высота автомобиля (x_4) влияет на y только при условии $\dfrac{x_4^{-1,5}}{\ln x_2} \geq 0,000534692$, причём, с увеличением x_4 на 1 дюйм цена y убывает в среднем на 7853 долларов.

5. Заключение

В статье представлена разработанная автором программа ВИнтер-2, предназначенная для построения вполне интерпретируемых элементарных и неэлементарных квазилинейных регрессионных моделей. В процессе построения регрессии в ВИнтер-2 можно контролировать количество регрессоров, степень их корреляции с y, степень нелинейности элементарных преобразований, число знаков после запятой в действительных числах, абсолютные вклады переменных в общую детерминацию и величины интеркорреляций. Разработка имеет высокое прикладное значение, поскольку с помощью неё можно решать реальные задачи анализа данных из абсолютно любых предметных областей. Причем, решать их довольно эффективно, поскольку вместо метода “всех регрессий” в ВИнтер-2 формируется задача ЧБЛП, методы решения которой были существенно развиты за последние годы. К тому же, результатом работы ВИнтер-2 является не просто регрессионная модель, по которой можно прогнозировать, но также и интерпретировать влияние каждой входящей в неё переменной на y. Кроме того, ВИнтер-2 позволяет строить уникальные неэлементарные регрессионные модели, позволяющие выявлять новые закономерности функционирования объектов исследования. Например, в [23,24] с помощью ВИнтер-2 уже строились НЛР функционирования железнодорожного транспорта в Иркутской и Тюменской областях. Однако впервые введенные в текущей статье НКР, представляющие собой более сложные зависимости, чем НЛР, а поэтому обладающие большим потенциалом, пока еще ни разу, за исключением небольшого примера в данной работе, не применялись на практике.

В дальнейшем планируется использовать ВИнтер-2 для решения широкого круга реальных прикладных задач, оснастить его возможностью контроля при построении модели факторов вздутия дисперсии [21] и t-критерия Стьюдента [22], а также реализовать в нём возможность построения других вполне интерпретируемых видов регрессий.

Список литературы / References

[1]. Molnar C. Interpretable machine learning. Lulu.com, 2020.

[2]. Doshi-Velez F., Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

[3]. Montgomery D. C., Peck E. A., Vining G. G. Introduction to linear regression analysis. John Wiley & Sons, 2021.

[4]. Shrestha N. Detecting multicollinearity in regression analysis. American Journal of Applied Mathematics and Statistics, vol. 8, no. 2, 2020, pp. 39-42.

[5]. Базилевский М.П. Построение вполне интерпретируемых линейных регрессионных моделей с помощью метода последовательного повышения абсолютных вкладов переменных в общую детерминацию. Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии, ном. 2, 2022, стр. 5-16 / Bazilevskiy M.P. Construction of quite interpretable linear regression models using the method of successive increase the absolute contributions of variables to the general determination. Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies, no. 2, 2022, pp. 5-16. (in Russian).

[6]. Горбач А.Н., Цейтлин Н.А. Покупательское поведение: анализ спонтанных последовательностей и регрессионных моделей в маркетинговых исследованиях. Киев, Освіта УкраЇны, 2011, 220 с. / Gorbach A.N., Tseytlin N.A. Buying Behavior: Analysis of Spontaneous Sequences and Regression Models in Marketing Research. Kyiv, Education of Ukraine, 2011, 220 p. (in Russian).

[7]. Miller A. Subset selection in regression. CRC Press, 2002.

[8]. Себер Дж. Линейный регрессионный анализ. М., Издательство “Мир”, 1980, 456 с. / Seber Dzh. Linear Regression Analysis. Moscow, Mir Publishing House, 1980, 456 p. (in Russian).

[9]. Фёрстер Э., Рёнц Б. Методы корреляционного и регрессионного анализа. М., Финансы и статистика, 1983, 303 с. / Ferster E., Rents B. Methods of Correlation and Regression Analysis. Moscow, Finance and Statistics, 1983, 303 p. (in Russian).

[10]. Konno H., Yamamoto R. Choosing the best set of variables in regression analysis using integer programming. Journal of Global Optimization, 2009, vol. 44, pp. 273-282. DOI: 10.1007/s10898-008-9323-9.

[11]. Miyashiro R., Takano Y. Mixed integer second-order cone programming formulations for variable selection in linear regression. European Journal of Operational Research, 2015, vol. 247, pp. 721-731. DOI: 10.1016/j.ejor.2015.06.081.

[12]. Miyashiro R., Takano Y. Subset selection by Mallows’ Cp: A mixed integer programming approach. Expert Systems with Applications, 2015, vol. 42, pp. 325-331. DOI: 10.1016/j.eswa.2014.07.056.

[13]. Tamura R., Kobayashi K., Takano Y., Miyashiro R., Nakata K., Matsui T. Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. Journal of Global Optimization, 2019, vol. 73, pp. 431-446. DOI: 10.1007/s10898-018-0713-3.

[14]. Park Y.W., Klabjan D. Subset selection for multiple linear regression via optimization. Journal of Global Optimization, 2020, vol. 77, pp. 543-574. DOI: 10.1007/s10898-020-00876-1.

[15]. Takano Y., Miyashiro R. Best subset selection via cross-validation criterion. Top, 2020, vol. 28, no. 2, pp. 475-488. DOI: 10.1007/s11750-020-00538-1.

[16]. Bertsimas D., Li M.L. Scalable holistic linear regression. Operations Research Letters, 2020, vol. 48, no. 3, pp. 203-208. DOI: 10.1016/j.orl.2020.02.008.

[17]. Chung S., Park Y.W., Cheong T. A mathematical programming approach for integrated multiple linear regression subset selection and validation. Pattern Recognition, 2020, vol. 108. DOI: 10.1016/j.patcog.2020.107565.

[18]. Bertsimas D., Gurnee W. Learning sparse nonlinear dynamics via mixed-integer optimization. Nonlinear Dynamics, 2023, vol. 111, no. 7, pp. 6585-6604. DOI: 10.1007/s11071-022-08178-9.

[19]. Watanabe A., Tamura R., Takano Y., Miyashiro R. Branch-and-bound algorithm for optimal sparse canonical correlation analysis. Expert Systems with Applications, 2023, vol. 217, pp. 119530. DOI: 10.1016/j.eswa.2023.119530.

[20]. Базилевский М.П. Формализация процесса отбора информативных регрессоров в линейной регрессии в виде задачи частично-булевого линейного программирования с ограничениями на коэффициенты интеркорреляций. Современные наукоемкие технологии, ном. 8, 2023, стр. 10-14 / Bazilevskiy M.P. Formalization the subset selection process in linear regression as a mixed integer 0-1 linear programming problem with constraints on intercorrelation coefficients. Modern High Technologies, no. 8, 2023, pp. 10-14. (in Russian).

[21]. Базилевский М.П. Отбор информативных регрессоров с учётом мультиколлинеарности между ними в регрессионных моделях как задача частично-булевого линейного программирования. Моделирование, оптимизация и информационные технологии, том 6, ном. 2 (21), 2018, стр. 104-118 / Bazilevskiy M.P. Subset selection in regression models with considering multicollinearity as a task of mixed 0-1 integer linear programming. Modeling, Optimization and Information Technology, vol. 6, no. 2 (21), 2018, pp. 104-118. (in Russian).

[22]. Базилевский М.П. Отбор значимых по критерию Стьюдента информативных регрессоров в оцениваемых с помощью МНК регрессионных моделях как задача частично-булевого линейного программирования. Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии, ном. 3, 2021, стр. 5-16 / Bazilevskiy M.P. Selection of informative regressors significant by Student's t-test in regression models estimated using OLS as a partial Boolean linear programming problem. Proceedings of Voronezh State University. Series: Systems Analysis and Information Technologies, no. 3, 2021, pp. 5-16. (in Russian).

[23]. Базилевский М.П. Метод построения неэлементарных линейных регрессий на основе аппарата математического программирования. Проблемы управления, ном. 4, 2022, стр. 3-14 / Bazilevskiy M.P. A method for constructing non-elementary linear regressions based on mathematical programming. Control Sciences, no. 4, 2022, pp. 3-14. (in Russian).

[24]. Базилевский М.П. Построение вполне интерпретируемых неэлементарных линейных регрессионных моделей. Вестник Югорского государственного университета, ном. 4 (67), 2022, стр. 105-114 / Bazilevskiy M.P. Construction of quite interpretable non-elementary linear regression models. Yugra State University Bulletin, no. 4 (67), 2022, pp. 105-114. (in Russian).

[25]. Базилевский М.П. Критерии нелинейности квазилинейных регрессионных моделей. Моделирование, оптимизация и информационные технологии, том 6, ном. 4 (23), 2018, стр. 185-195 / Bazilevskiy M.P. Nonlinear criteria of quasi-linear regression models. Modeling, Optimization and Information Technology, vol. 6, no. 4 (23), 2018, pp. 185-195. (in Russian).

Информация об авторах / Information about authors

Михаил Павлович БАЗИЛЕВСКИЙ – кандидат технических наук, доцент, доцент кафедры "Математика" Иркутского государственного университета путей сообщения. Сфера научных интересов: математическое моделирование, анализ данных, оптимизация, эконометрика, машинное обучение, искусственный интеллект.

Mikhail Pavlovich BAZILEVSKIY – Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Mathematics of the Irkutsk State Transport University. Research interests: mathematical modeling, data analysis, optimization, econometrics, machine learning, artificial intelligence.