УДК 519.862.6

А. А. Хоняков [1]

[1]*Иркутский государственный университет путей сообщения, г. Иркутск, Российская Федерация*

ОСОБЕННОСТИ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ ДЛЯ ПОСТРОЕНИЯ КУСОЧНО-ЛИНЕЙНЫХ РЕГРЕССИЙ

Аннотация. *В статье рассмотрена разработка программного обеспечения для построения кусочно-линейных регрессий. Выполнен анализ формирования задач частично булевого линейного программирования для различных видов кусочно-линейных регрессий, выделены общие моменты. По результатам анализа представлен вариант реализации программы в виде центрального абстрактного класса, содержащего общую логику для всех моделей, и дочерних классов самих моделей.*

В качестве демонстрации работы программного обеспечения было выполнено моделирование соотношения среднедушевых денежных доходов населения с величиной прожиточного минимума.

Ключевые слова: *кусочно-линейная регрессия, метод наименьших модулей, булевы переменные, задача частично булевого линейного программирования, кусочно-линейная регрессия с максимумом в правой части.*

A. A. Khonyakov [1]

[1]*Irkutsk State Transport University, Irkutsk, Russia*

FEATURES OF SOFTWARE DEVELOPMENT FOR CONSTRUCTING PIECEWISE LINEAR REGRESSIONS

Abstract. *The article deals with the development of software for constructing piecewise linear regressions. The analysis of the formation of partially Boolean linear programming problems for various types of piecewise linear regressions is carried out, common points are highlighted. Based on the analysis results, a variant of the program implementation is presented in the form of a central abstract class containing common logic for all models and child classes of the models themselves.*

As a demonstration of the operation of the software, the modeling of the ratio of the average per capita money income of the population to the value of the subsistence minimum was carried out.

Keywords: *piecewise linear regression, method of least modules, Boolean variables, the task of partial Boolean linear programming, piecewise linear regression with a maximum on the right side.*

Введение

В ходе разработки методического и программного обеспечения построения кусочно-линейных регрессионных моделей были рассмотрены различные варианты реализации таких моделей [1-6].

В первую очередь была рассмотрена кусочно-линейная регрессия с минимумом в правой части или производственная функция с постоянными пропорциями:

$$y_k = \min\left\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, \ldots, \alpha_m x_{km}\right\} + \varepsilon_k,\ k = \overline{1,n}. \quad (1)$$

Функцию (1) можно интерпретировать следующим образом: значение выходного параметра ограничено входными факторами, причём значением наименьшего из них. Пример – для сборки одной единицы ПК необходима одна материнская плата, два модуля оперативной памяти, один процессор, один жесткий диск и один блок питания. Если мы, например, увеличим количество всех имеющихся деталей в два раза, кроме материнской платы, это нам не позволит собрать в два раза больше компьютеров, в данном случае лимитирующим фактором выступит количество имеющихся материнский плат.

Противоположной по смыслу регрессионной моделью является кусочно-линейная регрессия с максимумом в правой части:

$$y_k = \max\left\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, \ldots, \alpha_m x_{km}\right\} + \varepsilon_k,\ k = \overline{1,n}. \quad (2)$$

В этой модели выходной параметр имеет негативный характер – риск, угроза, уязвимость и т.д. Входные же параметры представляют собой некоторые частные показатели этого негативного фактора.

Также были рассмотрена модель вида (3), представляющая собой разность функции минимума и максимума:

$$y_k = \min\left\{\alpha_1 x_{k1}^P, \alpha_2 x_{k2}^P, ..., \alpha_m x_{km}^P\right\} - \max\left\{\beta_1 x_{k1}^N, \beta_2 x_{k2}^N, ..., \beta_h x_{kh}^N\right\} + \varepsilon_k,\ k = \overline{1,n}, \quad (3)$$

где m – количество позитивных факторов, h – количество негативных факторов.

Такая модель позволяет оценить влияние на независимую переменную как позитивных в отношении нее факторов (помещаются в функцию минимума), так и негативных факторов (помещаются в функцию максимума).

Для оценки параметров представленных кусочно-линейных регрессионных моделей необходимо решить задачу частично булевого линейного программирования. Несмотря на то, что для решения подобных задач имеется большой арсенал программных средств, сложности начинаются уже на этапе составления задачи частично булевого линейного программирования. Ручное составление такой задачи представляет собой весьма трудоёмкий процесс, который нуждается в автоматизации. Кроме того, вычисление критериев адекватности полученной модели даже с помощью электронных таблиц тоже нетривиальная задача ввиду необходимости выполнения значительного числа шагов. По всем этим причинам было принято решение о разработке специализированного программного обеспечения для построения кусочно-линейных регрессионных моделей.

В процессе разработки встал вопрос о способе реализации: для каждой разновидности модели писать отдельную подпрограмму (функцию, класс) или постараться обобщить некоторую логику в виде отдельных модулей.

Именно решению этой задачи посвящена данная статья. Целью статьи является описание выбранного подхода для реализации программного обеспечения построения кусочно-линейных регрессионных моделей.

Разработка программного комплекса построения кусочно-линейных регрессий

Сначала необходимо рассмотреть задачи частично булевого линейного программирования [1-5] для моделей (1)-(3), постараться выявить общие их закономерности.

Для решения этих задач применяется метод наименьших модулей, который приводит к задаче частично булевого линейного программирования вида:

$$J(\alpha) = \sum_{k=1}^{n} (u_k + v_k) \to \min, \quad (4)$$

$$y_k = z_k + u_k - v_k,\ k = \overline{1,n}. \quad (5)$$

Для функции минимума вводятся ограничения вида:

$$z_k \le \alpha_i x_{ki},\ k = \overline{1,n},\ i = \overline{1,m}, \quad (6)$$

$$\alpha_i x_{ki} - z_k \le (1 - \sigma_{ki}) M,\ k = \overline{1,n},\ \ i = \overline{1,m}, \quad (7)$$

$$\sum_{i=1}^{m} \sigma_{ki} = 1,\ k = \overline{1,n}. \quad (8)$$

Для функции максимума вводятся ограничения вида:

$$z_k \ge \alpha_i x_{ki},\ k = \overline{1,n},\ i = \overline{1,m}, \quad (9)$$

$$\alpha_i x_{ki} - z_k \ge (-1 + \sigma_{ki}) M,\ k = \overline{1,n}\ ,\ \ i = \overline{1,m}, \quad (10)$$

$$\sum_{i=1}^{m} \sigma_{ki} = 1,\ k = \overline{1,n}, \quad (11)$$

где $\sigma_{ki}, k = \overline{1,n},\ i = \overline{1,m}$ - булевы переменные. В выражениях (5)-(7), (9)-(10) z_k представляет собой замены для функции минимума и максимума. Для разности функций минимума и максимума эта замена принимает вид:

$$z_k = z_{1k} - z_{2k} = \min\left\{\alpha_1 x_{k1}^P, \alpha_2 x_{k2}^P, ..., \alpha_m x_{km}^P\right\} - \max\left\{\beta_1 x_{k1}^N, \beta_2 x_{k2}^N, ..., \beta_h x_{kh}^N\right\}; \quad (12)$$

Очевидно, что целевая функция одинакова для всех рассмотренных выше моделей. Поэтому формирование целевой функции может быть вынесено в некий общий модуль.

Соотношение (5) используется в качестве ограничения. Однако это ограничение для моделей (1) - (2) и для (3) отличаются, поскольку для модели (3) переменная z_k представлена в виде разности (12). Поэтому вынести это в некую общую логику не представляется возможным.

Также в задачу частично булевого линейного программирования вводятся ограничения, связанные с математическим смыслом функций минимума (6)-(8) и максимума (9)-(11). Видно, что функциональность формирования этих ограничений может быть вынесена в общий модуль, причем можно объединить формирование этих ограничений, как для функции минимума, так и для максимума, ввиду их противоположности.

Таким образом, анализ задач частично булевого линейного программирования для нахождения вектора параметров моделей (1)-(3) позволил выделить их общие моменты и различия.

Далее рассмотрим техническую составляющую решения этих задач. За решение задач линейного программирования отвечает библиотека LpSolve [7] для объектно-ориентированного языка Java.

Для корректной работы библиотеки необходимо создать объект класса LpSolve, представляемого библиотекой, заполнить определенные поля этого класса при помощи соответствующих методов (таблица 1).

Таблица 1. Основные методы класса LpSolve

Метод	Описание
LpSolve makeLp(int rows, int columns)	Создаёт новую задачу, количество строк можно указать как 0, количество столбцов – количество переменных в задаче
void addConstraint(double[] row, int constrType, double rh)	Добавляет ограничение к задаче. Первый параметр содержит вектор-строку с коэффициентами левой части, второй параметр – знак ограничения, третий – правое значение.
void setColName(int colNumber, String colName)	Присваивает столбцу (переменной) имя.
void setObjFn(double[] row)	Устанавливает целевую функцию, в качестве входного параметра передаётся вектор-строка с коэффициентами.
void setMinim()	Целевая функция – функция минимума
void setMaxim()	Целевая функция - функция максимума
void setBinary(int colnr, boolean mustBeBin)	Помечает переменную как булеву (двоичную)
int solve()	Запускает решение задачи
void getVariables(double[] var)	Возвращает значения переменных

Общая последовательность действий для решения задачи при помощи этой библиотеки:

1. Создание новой задачи (метод *makeLp*);
2. Именование вектора переменных (метод *setColName*);
3. Добавление ограничений (метод *addConstraint*);
4. Присвоение целевой функции (метод *setObjFn*);
5. Запуск решения (метод *solve*).

С учётом особенностей аналитического вида задач частично булевого линейного программирования и библиотеки функциональность программы была выполнена в виде иерархии классов, представленной на рис. 1.

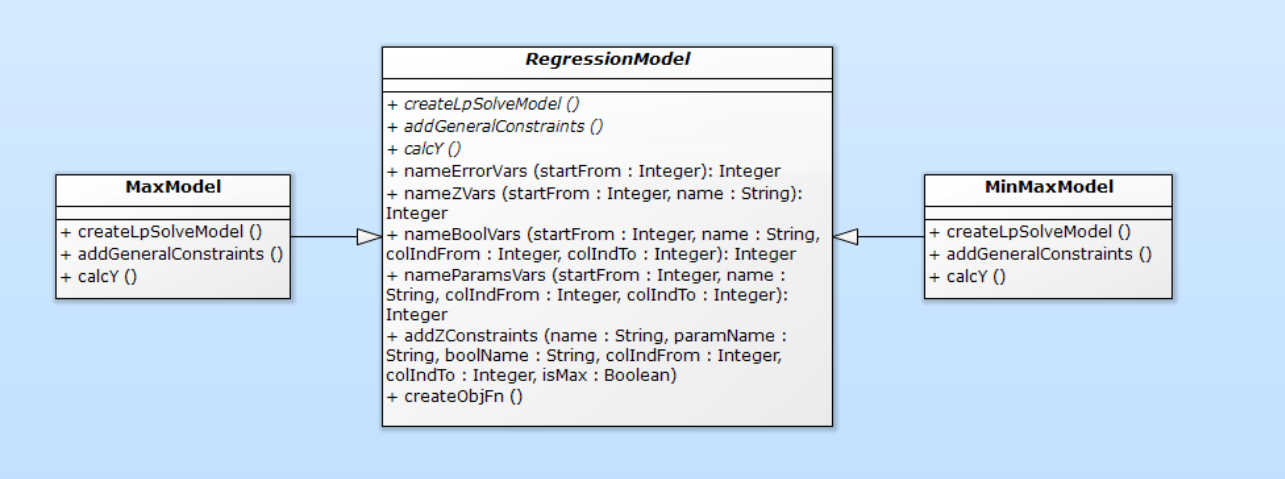

Рис. 1. Диаграмма классов

Центральным классом является абстрактный класс *RegressionModel*, который содержит абстрактные методы *createLpSolveModel*, *addGeneralConstraints*, *calcY*. Каждый из этих методов реализуется в дочерних классах *MinMaxModel*, *MaxModel*.

Метод *createLpSolveModel* отвечает за создание модели, а именно за шаги 1-4 описанной ранее последовательности действий. Из этого метода в дочерних классах вызываются общие методы, реализованные в центральном классе: *nameErrorVars* (шаг 2, задание имен вектору ошибок), *nameZVars* (шаг 2, именование переменных z_k), *nameBoolVars* (шаг 2, именование булевых переменных), *nameParamsVars* (шаг 2, именование вектора параметров). Также вызываются методы, связанные с шагом 3 – *addGeneralConstraints* (отвечает за добавление в модель ограничений вида (5), реализуется в дочерних классах), *addZConstraints* (отвечает за добавление в модель ограничений вида (6)-(11), метод реализован в центральном классе). Наконец, вызывается метод *createObjFn*, который отвечает за формирование целевой функции (4) (шаг 4), который также вынесен в общую логику в класс *RegressionModel*.

Метод *calcY* отвечает за вычисление расчётных значений зависимой переменной в момент калькуляции критериев адекватности.

Демонстрация работы программного обеспечения

Программное обеспечение [8] для построения кусочно-линейных регрессионных моделей выполнено при помощи языка программирования Java, для целей формирования пользовательского интерфейса применена библиотека JavaFX.

Внешний вид приложения представлен на рис. 2.

Загрузка данных в программу осуществляется при помощи меню Импорт через файл или из буфера обмена (например, достаточно выделить нужные ячейки в таблице Excel и скопировать). Далее из выпадающего списка выбирается вид модели. После этого есть возможность либо выгрузить сформированную задачу в файл, чтобы в дальнейшем можно было ее открыть в программе LPSolve IDE, либо запустить решение задачи.

Для демонстрации работы приложения выполним моделирование соотношения среднедушевых денежных доходов населения с величиной прожиточного минимума. В качестве модели выберем кусочно-линейную регрессию с максимумом в правой части.

Для моделирования были использованы данные за 2017-2020 годы (таблица 2), взятые из официальных витрин статистических данных [9-11], по следующим показателям:

- y - соотношение среднедушевых денежных доходов населения с величиной прожиточного минимума, %;
- x_1 - численность выбывших работников в связи с сокращением численности, чел;
- x_2 - численность работников списочного состава, находившихся в простое по вине работодателя и по причинам, не зависящим от работодателя и работника, чел;

- x_3 - численность работников списочного состава, которым были предоставлены отпуска без сохранения заработной платы по письменному заявлению работника, чел;

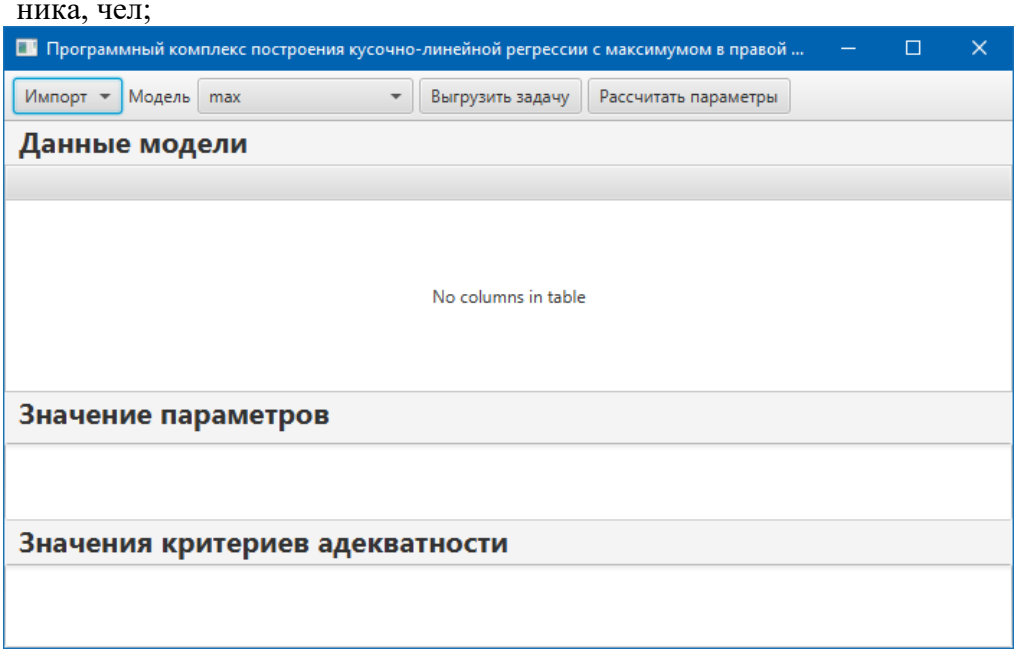

Рис. 2. Внешний вид приложения

Поскольку выходной параметр не является с точки зрения моделирования негативным, применим подход, описанный в [12]. Введем замену вида:

$$y^* = \frac{10^7}{y}. \quad (13)$$

Таблица 2. Исходные данные для моделирования

	y	x_1	x_2	x_3
1 кв. 2017 г.	280,20	71224	176227	2232540
2 кв. 2017 г.	303,10	74624	166399	2490172
3 кв. 2017 г.	303,29	64291	127077,99	2824795
4 кв. 2017 г.	380,40	69407	154985	2513107
1 кв. 2018 г.	288,30	52237	149760	2288825
2 кв. 2018 г.	309,90	56111	142580,99	2668808
3 кв. 2018 г.	311,10	57844	131024,99	2986784
4 кв. 2018 г.	380,40	63132	197582	2685881
1 кв. 2019 г.	280,50	50996	197590,99	2437013
2 кв. 2019 г.	308,30	54646	172680,99	2714613
3 кв. 2019 г.	317,80	52483	130939,99	3077375
4 кв. 2019 г.	389,60	57329	191369,99	2899274
1 кв. 2020 г.	290,30	48054	269296	2555323
2 кв. 2020 г.	285,39	43037	947781,99	2418509
3 кв. 2020 г.	299	41554	331025	2973761

Выполним расчёт при помощи программы, результат представлен на рис. 3.

Таким образом, нами была получена модель:

$$y^* = \max(0{,}4421\,x_1; 0{,}037\,x_2; 0{,}0117\,x_3). \quad (14)$$

Величина погрешности составила допустимые 10,97%. Вектор срабатываний (1, 1, 3, 1, 3, 3, 3, 3, 3, 3, 3, 3, 3, 2, 3) позволяет сделать вывод о том, что все независимые переменные оказали влияние на зависимую переменную.

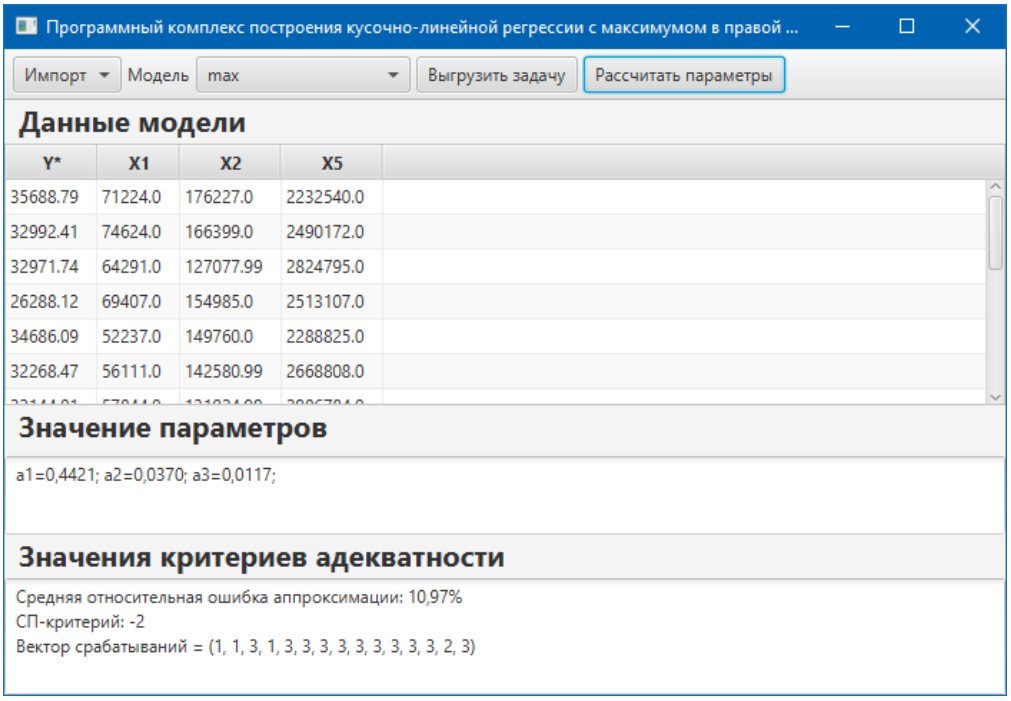

Рис. 3. Результат моделирования

Заключение

В данной работе были рассмотрены особенности разработки программного обеспечения построения кусочно-линейных регрессионных моделей. Проанализирована алгоритмическая и техническая составляющая идентификации параметров таких моделей, описан выбранный подход к разработке. В качестве демонстрации работы программного обеспечения было выполнено моделирование соотношения среднедушевых денежных доходов населения с величиной прожиточного минимума.

Отметим, что предложенный вариант реализации может быть модифицирован в дальнейшем, поскольку автором продолжается изучение и разработка новых кусочно-линейных моделей. Кроме того, планируется развитие описанного программного обеспечения по ряду смежных направлений [13-16].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Ильина Н.К., Лебедева С.А., Носков С.И. Идентификация параметров некоторых негладких регрессий//Информационные технологии и проблемы математического моделирования сложных систем. – 2016. – № 17. – С. 111.

2. Носков С.И., Лоншаков Р.В. Идентификация параметров кусочно-линейной регрессии//Информационные технологии и проблемы математического моделирования сложных систем. – 2008. – № 6. – С. 63-64.

3. Носков С.И. Идентификация параметров кусочно-линейной функции риска// Транспортная инфраструктура Сибирского региона. – 2017. – Т. 1. – С. 417-421.

4. Носков С.И. Технология моделирования объектов с нестабильным функционированием и неопределенностью в данных. – Иркутск: Облинформпечать. – 1996. – 320 с.

5. Носков С.И., Хоняков А.А. Программный комплекс построения некоторых типов кусочно-линейных регрессий // «Информационные технологии и математическое моделирование в управлении сложными системами»: электрон.науч. журн. – 2019. – №3. – С. 47-55 – Режим доступа: http://ismm-irgups.ru/toma/34-2019, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 20.11.2019)

6. Базилевский, М.П. МНК-оценивание параметров специфицированных на основе функций Леонтьева двухфакторных моделей регрессии [Текст] / М.П. Базилевский // Южно-сибирский научный вестник. – 2019. – №2(26). – С.66-70.

7. Официальный сайт программы lpsolve [Электронный ресурс]. – Режим доступа: http://lpsolve.sourceforge.net/

8. Свидетельство о государственной регистрации программ для ЭВМ № 2021613936 Программный комплекс построения кусочно-линейных регрессий с максимумом в правой части/ С.И. Носков, А.А. Хоняков (Россия); Правообладатель: Федеральное государственное бюджетное образовательное учреждение высшего образования «Иркутский государственный университет путей сообщения» (ФГБОУ ВО ИрГУПС); заявка № 2021613052 12.03.2021; дата регистр. 16.03.2021

9. Федеральная служба государственной статистики [Электронный ресурс] – Режим доступа: https://rosstat.gov.ru/statistic

10. Витрина статистических данных статистики [Электронный ресурс] – Режим доступа: https://showdata.gks.ru/finder/

11. Единая межведомственная информационно-статистическая система (ЕМИСС) [Электронный ресурс] – Режим доступа: https://www.fedstat.ru/

12. Носков С.И. Применение функции риска для моделирования экономических систем / С.И. Носков, А.А. Хоняков // Южно-Сибирский научный вестник. – 2020. – № 5. – с. 85-92.

13. Лакеев А. В., Носков С. И. О множестве решений линейного уравнения с интервально заданными оператором и правой частью // Докл. РАН. 1993. Т. 330, №4. С. 430–433.

14. Лакеев А. В., Носков С. И. О множестве решений линейного уравнения с интервально заданными оператором и правой частью // Сиб. мат. журн. 1994. Т. 35, №5. С. 1074–1084.

15. Носков С.И. Точечная характеризация множества Парето в линейной многокритериальной задаче // Современные технологии. Системный анализ. Моделирование. – Иркутск, 2008. – № 17. – С. 99–102.

16. Базилевский М.П., Носков С.И. Алгоритм формирования множества регрессионных моделей с помощью преобразования зависимой переменной // Международный журнал прикладных и фундаментальных исследований. – 2011. – № 3. – С. 159–160.

REFERENCES

1. Ilyina N.K., Lebedeva S.A., Noskov S.I. Identification of parameters of some nonsmooth regressions // Information technologies and problems of mathematical modeling of complex systems. - 2016. - No. 17. - P. 111.

2. Noskov S.I., Lonshakov R.V. Identification of parameters of piecewise linear regression // Information technologies and problems of mathematical modeling of complex systems. - 2008. - No. 6. - S. 63-64.

3. Noskov S.I. Identification of parameters of a piecewise-linear risk function // Transport infrastructure of the Siberian region. - 2017 .- T. 1. - P. 417-421.

4. Noskov S.I. A technology for modeling objects with unstable functioning and uncertainty in the data. - Irkutsk: Oblinform printing. - 1996 .- 320 p.

5. Noskov S.I., Khonyakov A.A. A software package for constructing some types of piecewise linear regressions // "Information technologies and mathematical modeling in the management of complex systems": electronic scientific. zhurn. - 2019. - No. 3. - P. 47-55 - Access mode: http://ismm-irgups.ru/toma/34-2019, free. - Title from the screen. - Yaz. Russian, English (date of access: 20.11.2019)

6. Bazilevsky, M.P. OLS-estimation of parameters of two-factor regression models specified on the basis of Leontiev functions [Text] / M.P. Bazilevsky // South Siberian Scientific Bulletin. - 2019. - No. 2 (26). - Pp.66-70.

7. Official site of the lpsolve program [Electronic resource]. - Access mode: http://lpsolve.sourceforge.net/

8. Certificate of state registration of computer programs No. 2021613936 Software complex for constructing piecewise linear regressions with a maximum on the right side. Noskov, A.A. Khonyakov (Russia); Copyright holder: Federal State Budgetary Educational Institution of Higher

Education "Irkutsk State University of Railways" (FGBOU VO IrGUPS); Application No. 2021613052 03/12/2021; date register. 03/16/2021

9. Federal State Statistics Service [Electronic resource] - Access mode: https://rosstat.gov.ru/statistic

10. Showcase of statistical data of statistics [Electronic resource] - Access mode: https://showdata.gks.ru/finder/

11. Unified interdepartmental information and statistical system (EMISS) [Electronic resource] - Access mode: https://www.fedstat.ru/

12. Noskov S.I. Application of the risk function for modeling economic systems / S.I. Noskov, A.A. Khonyakov // South Siberian Scientific Bulletin. - 2020. - No. 5. - p. 85-92.

13. Lakeev AV and Noskov SI, “On the set of solutions of a linear equation with interval given operator and right-hand side,” Dokl. RAS. 1993. T. 330, No. 4. P. 430-433.

14. Lakeev AV and Noskov SI, “On the set of solutions of a linear equation with interval given operator and right-hand side,” Siberian Math. mat. zhurn. 1994. T. 35, No. 5. S. 1074-1084.

15. Noskov S.I. Point characterization of the Pareto set in a linear multicriteria problem // Sovremennye tekhnologii. System analysis. Modeling. - Irkutsk, 2008. - No. 17. - P. 99-102.

16. Bazilevsky M.P., Noskov S.I. Algorithm for the formation of a set of regression models using the transformation of the dependent variable // International Journal of Applied and Fundamental Research. - 2011. - No. 3. - P. 159–160.

Информация об авторах

Антон Андреевич Хоняков - аспирант, кафедра «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: anton_khonyakov@mail.ru

Authors

Anton Andreevich Khonyakov, Postgraduate Student of Department “Information systems and information security”, Irkutsk State Transport University, Irkutsk, e-mail: anton_khonyakov@mail.ru

Для цитирования

Хоняков А.А. Особенности разработки программного обеспечения для построения кусочно-линейных регрессий [Электронный ресурс] / А.А Хоняков // Молодая наука Сибири: электрон. науч. журн. – 2021. – №2(12) – Режим доступа: http://mnv.irgups.ru/toma/212-2021, свободный. – Загл. с экрана. – Яз. рус., англ. (дата обращения: 03.08.2021)

For citation

Khonyakov A.A. *Osobennosti razrabotki programmnogo obespecheniya dlya postroeniya kusochno-linejnyh regressij* [Features of software development for constructing piecewise linear regressions]. *Molodaya nauka Sibiri: ehlektronnyj nauchnyj zhurnal* [Young science of Siberia: electronic scientific journal], 2021, no. 2. [Accessed 03/08/21]