УДК 519.862.6

С.И. Носков[1], А.А. Хоняков[1]

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

ПРОГРАММНЫЙ КОМПЛЕКС ПОСТРОЕНИЯ НЕКОТОРЫХ ТИПОВ КУСОЧНО-ЛИНЕЙНЫХ РЕГРЕССИЙ

Аннотация. В статье рассматриваются способы оценивания параметров трех типов кусочно-линейных регрессий методом наименьших модулей. Впервые рассматривается регрессионная модель, представляющая собой сумму кусочно-линейных регрессий с минимальным и максимальным вкладом независимых переменных. Приводится описание разработанного программного комплекса для автоматизированной оценки параметров этих моделей.

Ключевые слова: кусочно-линейная регрессия, метод наименьших модулей, булевы переменные, задача линейного программирования, задача частично булевого линейного программирования.

S.I. Noskov[1], A.A.Khonyakov[1]

[1]Irkutsk State Transport University, Irkutsk, Russia

SOFTWARE COMPLEX FOR BUILDING SOME TYPES PIECES OF LINEAR REGRESSIONS

Annotation.The article discusses methods for estimating the parameters of three types of piecewise linear regressions based on the least modulus method. For the first time, a regression model is considered, which is the sum of piecewise linear regressions with the minimum and maximum contribution of independent variables. The description of the developed software package for the automated estimation of the parameters of these models is given.

Keywords: piecewise linear regression, least module method, Boolean variables, linear programming problem, partially Boolean linear programming problem.

Введение

Регрессионный анализ [1-5] является важным инструментом построения статистических моделей. Его методы находят широкое применение в различных областях знаний, начиная от сельского хозяйства и заканчивая экономикой.

Основной задачей регрессионного анализа является оценка неизвестных параметров модели при известных значениях зависимой y и независимых x_1, x_2, …, x_m переменных. Разделение переменных на зависимые, также называемые выходными, объясняемыми, эндогенными и независимые (входные, объясняющие, экзогенные) происходит, исходя из соображений содержательного характера, то есть предполагается, что поведение изучаемой переменной y в основном зависит от факторов x_1, x_2, …, x_m.

Регрессионную модель в общем виде можно записать так:

$$y_k = F(\alpha; x_{k1}, x_{k2}, ..., x_{km}) + \varepsilon_k,\ k = \overline{1,n}, \qquad (1)$$

где y - зависимая переменная, x_i, $i = \overline{1,m}$ - независимые переменные, $F(\cdot)$ – некоторая вещественная аппроксимирующая функция; ε_i, $k = \overline{1,n}$ – ошибки аппроксимации; α – вектор неизвестных параметров модели, n - длина выборки.

Наиболее популярной регрессионной моделью можно считать модель линейной регрессии. Основными преимуществами линейной регрессионной модели являются простота оценивания ее параметров, например, с помощью метода наименьших квадратов и возможность добротной содержательной интерпретации. Однако большим недостатком такого вида моделей является их частая неадекватность для реальных физических или социально-экономических процессов, которые зачастую имеют более сложный, нелинейный характер.

Оценка параметров кусочно-линейных регрессий

Одной из таких нелинейных моделей, особенно популярных в экономико-математических моделях, является функция с постоянными пропорциями, которую также называют кусочно-линейной регрессией, или производственной функцией Леонтьева:

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

Предполагается неотрицательность переменных модели. Особенностью аппроксимирующей функции (2) является то, что значение выходного фактора, обычно трактуемого как выпуск продукции, определяется значением лимитирующего входного фактора (ресурса). При этом любое наращивание других факторов не приводит к возрастанию выпуска.

В работе [6] была впервые сформулирована задача точной идентификации параметров $\alpha_i, i = \overline{1,m}$ уравнения (2) с использованием метода наименьших модулей, приводящего к задаче:

$$J(\alpha) = \sum_{k=1}^{n} |\varepsilon_k| \to \min. \quad (3)$$

Было предложено следующее ее решение.

Введем в рассмотрение так называемые расчетные значения выходной переменной z_k:

$$z_k = \min\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, ..., \alpha_m x_{km}\},\ k = \overline{1,n}, \quad (4)$$

после чего регрессия (2) представима в виде

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

Следуя стандартному приему «раскрытия» модулей в (3) (см., например, [7]), введем в рассмотрение переменные u_k и v_k по правилу:

$$u_k = \begin{cases} y_k - z_k, y_k > z_k \\ 0, в \quad пр. \quad случае \end{cases},$$

$$v_k = \begin{cases} z_k - y_k, z_k > y_k \\ 0, в \quad пр. \quad случае \end{cases}$$

Легко видеть, что имеют место тождества

$$z_k + u_k - v_k = y_k,\ k = \overline{1,n}. \quad (6)$$

Из (4) следует справедливость неравенств

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

причем для каждого k, по крайней мере, одно из них должно обращаться в строгое равенство. Для достижения этого требования введем mn булевых переменных $\sigma_{ki}, k = \overline{1,n}$, $i = \overline{1,m}$ и сформируем ограничения:

$$\alpha_i x_{ki} - z_k \le (1 - \sigma_{ki})M,\ k = \overline{1,n},\ i = \overline{1,m} \quad (8)$$

$$\sum_{i=1}^{m} \sigma_{ki} = 1,\ k = \overline{1,n}, \quad (9)$$

где M - заранее выбранное большое положительное число.

Из задания переменных u_k и v_k следуют следующие равенства:

$$|\varepsilon_k| = u_k + v_k,\ u_k v_k = 0,$$

что позволяет представить функционал (3) в виде

$$J(\alpha) = \sum_{k=1}^{n} (u_k + v_k) \to \min. \quad (10)$$

Таким образом, задача (3) поиска значений неизвестных параметров α_i, $i = \overline{1,m}$ кусочно-линейной регрессии (2) по методу наименьших модулей свелась к задаче частично буле-

вого линейного программирования (6) – (10) с $mn+3n+m$ переменными (из которых mn - булевы) и $2(mn+n)$ ограничениями.

В работе [8] была рассмотрена регрессия противоположная по смыслу аппроксимирующей функции (2):

$$y_k = \max\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, ..., \alpha_m x_{km}\} + \varepsilon_k,\ k = \overline{1,n}. \quad (11)$$

В отличие от (2) здесь зависимая переменная имеет негативный характер, например, риск, уязвимость, угроза и т.д., а независимые переменные являются частными показателями этого агрегирующего фактора. Задача оценивания регрессии (11) также может быть сведена к задаче частично булевого линейного программирования, аналогичной задаче (6) – (10).

Действительно, заменим неравенства (7) на противоположные:

$$z_k \geq \alpha_i x_{ki},\ k = \overline{1,n},\ i = \overline{1,m}, \quad (12)$$

а неравенства (8) – на следующие:

$$\alpha_i x_{ki} - z_k \geq (-1+\sigma_{ki})M,\ k = \overline{1,n},\ i = \overline{1,m}. \quad (13)$$

Решение задачи частично булевого линейного программирования (6), (12), (13), (9), (10) как раз и позволит вычислить неизвестные оценки параметров негладкой регрессии (11).

Рассмотрим теперь некий симбиоз регрессий (2) и (11) - регрессионную модель вида:

$$y_k = \min\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, ..., \alpha_m x_{km}\} + \max\{\beta_1 x_{k1}, \beta_2 x_{k2}, ..., \beta_m x_{km}\} + \varepsilon_k,\ k = \overline{1,n}. \quad (14)$$

В данной модели содержится сумма кусочно-линейных регрессий с минимальным и максимальным вкладом независимых переменных.

Оценка параметров для такой модели производится по комбинированному алгоритму.

Расчетные значения выходной переменной z_k можно представить в виде:

$$z_k = z_{1k} + z_{2k} = \min\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, ..., \alpha_m x_{km}\} + \max\{\beta_1 x_{k1}, \beta_2 x_{k2}, ..., \beta_m x_{km}\},\ k = \overline{1,n}. \quad (15)$$

Имеют место тождества:

$$z_{1k} + z_{2k} + u_k - v_k = y_k,\ k = \overline{1,n}. \quad (16)$$

Из (14) следует справедливость неравенств:

$$z_{1k} \leq \alpha_i x_{ki},\ k = \overline{1,n},\ i = \overline{1,m}, \quad (17)$$

$$z_{2k} \geq \beta_i x_{ki},\ k = \overline{1,n},\ i = \overline{1,m}. \quad (18)$$

Причем для каждого k хотя бы одно из них должно обращаться в равенство. Поэтому сформулируем ограничения:

$$\alpha_i x_{ki} - z_{1k} \leq (1-\sigma_{ki})M,\ k = \overline{1,n},\ \ i = \overline{1,m}, \quad (19)$$

$$\sum_{i=1}^{m} \sigma_{ki} = 1,\ k = \overline{1,n}, \quad (20)$$

$$\beta_i x_{ki} - z_{2k} \geq (-1+\upsilon_{ki})M,\ k = \overline{1,n},\ i = \overline{1,m}, \quad (21)$$

$$\sum_{i=1}^{m} \upsilon_{ki} = 1,\ k = \overline{1,n}, \quad (22)$$

где M - заранее выбранное большое положительное число, а σ_{ki} и υ_{ki} - булевы переменные, $k = \overline{1,n}$, $i = \overline{1,m}$.

Таким образом, получаем задачу частично булевого линейного программирования (16)-(22).

Описание программного комплекса построения кусочно-линейных регрессий

Для автоматизированной оценки параметров моделей (2), (11), (14) был разработан программный комплекс на языке программирования Java.

Общий процесс работы программного комплекса можно описать следующей последовательностью шагов.

1. Ввод исходных данных.

2. Выбор типа модели.
3. Формирование задачи частично булевого линейного программирования.
4. Решение задачи частично булевого линейного программирования.
5. Интерпретация полученного решения.
6. Расчет критериев адекватности полученной модели.

Ввод исходных данных в программу осуществляется путем загрузки файла с расширением *.csv. Первая строка файла содержит заголовки столбцов. В первом столбце содержатся значения зависимой переменной y, в остальных – значения независимых переменных. Столбцы разделяются между собой при помощи символа «;», а строки - при помощи знака переноса строки. Пример такого файла представлен на рис. 1.

```
y;x1;x2;x3;
14;1;3;8;
8;2;4;5;
12;3;6;6;
15,7;6;7;8;
21;7;1;9;
20;8;1;8;
10;9;1;3;
11;4;1;6;
```

Рис. 1. Пример файла с исходными данными

Решение задачи частично булевого линейного программирования производится при помощи бесплатного пакета LPSolve. Причиной такого выбора послужило его широкое распространение и эффективность, а также наличие Java-библиотеки.

После подключения Java-библиотеки LPSolve становится доступным её главный класс – LpSolve.

Концептуально формирование задачи линейного программирования при помощи данной библиотеки можно свести к следующим действиям.

1. Присвоение имен переменным.
2. Задание ограничений.
3. Ввод целевой функции.
4. Выбор типа задачи: минимизация или поиск максимума.

Главный класс программы – LpSolver, его диаграмма представлена на рис. 2. Методы этого класса реализуют шаги 3-6 последовательности работы программы. Важный метод этого класса – *solve* позволяет получить параметры модели в виде вектора-строки *solution*, используя класс LPSolve библиотеки, расчетные значения зависимой переменной *calcY*, а также значения ошибок аппроксимации *errorVals* на основе двумерного массива *solveData*, полученного из файла с исходными данными.

Целевая функция и ограничения задаются в виде векторов-строк размерностью по количеству переменных (как было написано выше, количество переменных зависит от количества параметров модели, а также от количества измерений). Для удобства обращения с переменными желательно сначала задать их имена, то есть указать соответствие номера элемента в векторе-строке и его названия. Для этого у класса LpSolve есть специальный метод *setColName(номер_элемента, название)*. Также в ходе задания имен переменным необходимо указать, являются ли они булевыми при помощи метода *setBinary(номер_элемента, флаг_булевого_поля)*.

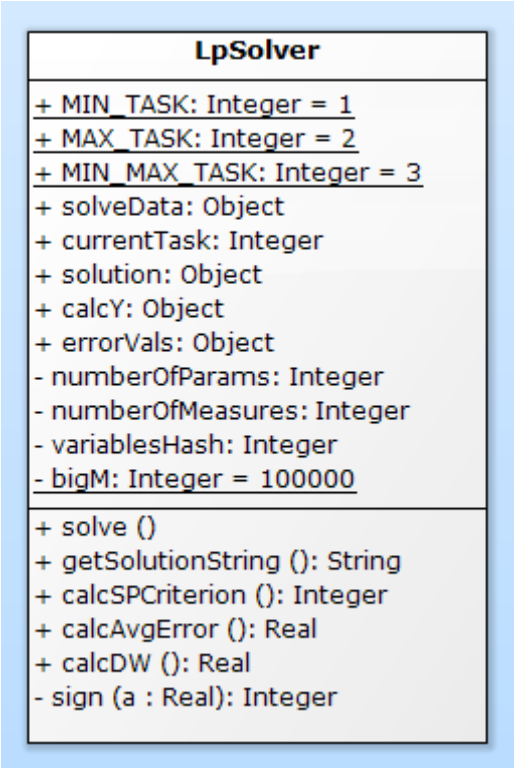

Рис. 2. Главный класс программы

Следующим шагом является формирование векторов-строк ограничений. Для включения режима внесения ограничений предназначен метод *setAddRowmode(флаг_включения)*, для внесения очередного ограничения - *addConstraint(вектор-строка, знак_ограничения, значение)*.

После ввода всех ограничений происходит формирование целевой функции при помощи метода *setObjFn(вектор-строка)*. Процедура решения запускается путем вызова метода *solve()*, а запрос полученного решения производится при помощи метода *getVariables(вектор-строка)*.

Методы *calcSPCriterion*, *calcAvgError*, *calcDW* позволяют выполнить расчет критериев адекватности полученной модели: критерий «согласованности поведения», описанный в [9, 10], среднюю относительную ошибку аппроксимации, критерий Дарбина-Уотсона.

Главное окно программы представлено на рис. 3.

При нажатии кнопки «Выбрать файл» открывается окно проводника, при помощи которого можно выбрать файл с исходными данными.

Далее необходимо выбрать тип регрессии, соответствующий моделям (2), (11) или (14). После чего производится нажатие кнопки «Выполнить оценку параметров модели», результаты вычислений отображаются в области «Результаты вычислений».

Процесс решения задачи в разработанном программном комплексе

По статистическим данным, представленным в таблице 1, покажем процесс решения задачи при помощи разработанной программы.

Сначала по данным из таблицы 1 необходимо сформировать файл, представленный на рис. 1, а затем загрузить его в поле «Исходные данные».

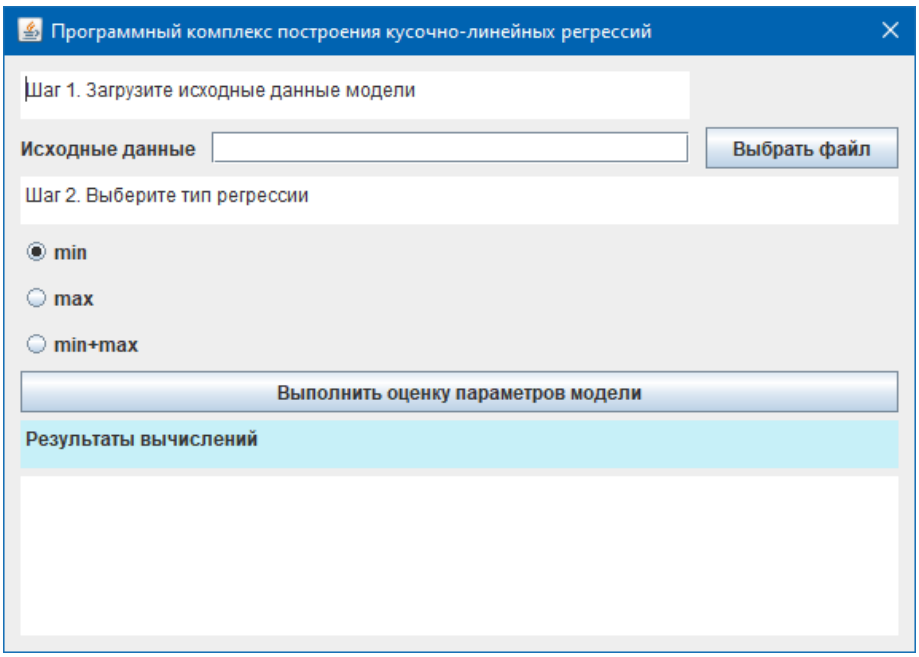

Рис. 3. Главное окно программы

Таблица 1 – Исходные данные

y	x1	x2
7	1	1
26	2	8
19	4	4
30	6	7
45	12	3
107	23	11
150	46	6
190	58	5
199	56	9
80	5	23

В качестве модели выберем сумму кусочно-линейных регрессий с минимальным и максимальным вкладами независимых переменных – опция «min+max». После этого нажмем кнопку «Выполнить оценку параметров модели». Результаты вычислений представлены на рис. 4.

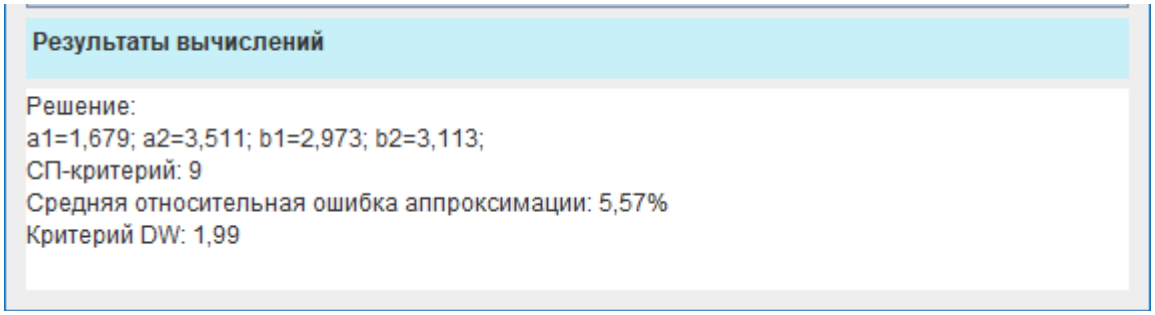

Рис. 4. Результат работы программы

Таким образом, вектор параметров модели получился равным (1,679;3,511;2,973;3,113). Средняя относительная ошибка аппроксимации, равная 5,57%, говорит о хорошем качестве регрессии. Критерий «согласованность поведения», равный 9, служит показателем высокой степень согласованности в характере изменения (поведении) расчетных и фактических значений зависимой переменной на различных наблюдениях выборки.

По таблице распределения Дарбина-Уотсона границы интервала критических значений DW-критерия при уровне значимости 0,01 составляют (0,279;1,873). Следовательно, можно сделать вывод об отсутствии автокорреляции остатков.

В своих последующих работах авторы намерены продолжить исследование регрессий (2), (11), (14) с помощью алгоритмов, представленных в работах[11-19].

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. – М.: Юнити, 1998. – 1022 с.

2. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Книга 1. В 2-х кн. – М.: Финансы и статистика, 1986. – 366 с.

3. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. Книга 2. В 2-х кн. – М.: Финансы и статистика, 1986. – 351 с.

4. Себер Дж. Линейный регрессионный анализ. – М.: Издательство «Мир», 1980. – 456 с.

5. Доугерти К. Введение в эконометрику. – М.: ИНФРА-М, 2009. – 465 с.

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

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

8. 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.

9. Носков С.И. Критерий «согласованность поведения» в регрессионном анализе//Современные технологии. Системный анализ. Моделирование. – 2013. – №1(37). – С.107-110.

10. Носков С.И. Построение эконометрических зависимостей с учетом критерия «согласованность поведения»//Кибернетика и системный анализ. – 1994. – № 1. – С. 177.

11. Kreinovich V., Lakeyev A.V., Noskov S.I. Approximate linear algebra is intractable//Linear Algebra and its Applications. – 1996. – Т. 232. – № 1-3. – С. 45-54.

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

13. Носков С.И. Точечная характеризация множества парето в линейной многокритериальной задаче//Современные технологии. Системный анализ. Моделирование. – 2008. – № 1 (17). – С. 99-101.

14. Lakeyev A.V., Noskov S.I. A description of the set of solutions of a linear equation with interval defined operator and right-hand side//Doklady Mathematics. – 1993. – Т. 47. – № 3. – С. 518.

15. Лакеев А.В., Носков С.И. Метод наименьших модулей для линейной регрессии: число нулевых ошибок аппроксимации//Современные технологии. Системный анализ. Моделирование. – 2012. – № 2 (34). – С. 48-50.

16. Носков С.И. Обобщенный критерий согласованности поведения в регрессионном анализе//Информационные технологии и математическое моделирование в управлении сложными системами. – 2018. – № 1 (1). – С. 14-20.

17. Носков С.И. О методе смешанного оценивания параметров линейной регрессии//Информационные технологии и математическое моделирование в управлении сложными системами. – 2019. – № 1 (2). – С. 41-45.

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

19. Носков С.И., Бутин А.А. Методическое обеспечение оценки уровня уязвимости объектов информатизации//Информационные технологии и проблемы математического моделирования сложных систем. -Иркутск: ИрГУПС, 2015. -Вып. 14. -С. 38-48.

REFERENCES

1. Ayvazyan S.A., Mkhitaryan V.S. Applied statistics and fundamentals of econometrics. – M .: Unity, 1998 . – 1022 p.

2. Draper N., Smith G. Applied regression analysis. Book 1. In 2 book. – M.: Finance and Statistics, 1986. – 366 p.

3. Draper N., Smith G. Applied regression analysis. Book 2. In 2 book. – M.: Finance and Statistics, 1986. – 351 p.

4. Seber J. Linear regression analysis. – M .: Mir Publishing House, 1980. – 456 p.

5. Dougherty K. Introduction to Econometrics. – M .: INFRA-M, 2009 . – 465 p.

6. 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. – Pp. 63-64.

7. Noskov S.I. Technology for modeling objects with unstable functioning and uncertainty in data. – Irkutsk: Oblinformpechat. – 1996. – 320 p.

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

9. Noskov S.I. The criterion of "consistency of behavior" in the regression analysis // Modern technologies. System analysis. Modeling. – 2013. – No. 1 (37). – Pp. 107-110.

10. Noskov S.I. Construction of econometric dependencies taking into account the criterion of "consistency of behavior" // Cybernetics and system analysis. – 1994. – No. 1. – P. 177.

11. Kreinovich V., Lakeyev A.V., Noskov S.I. Approximate linear algebra is intractable // Linear Algebra and its Applications. – 1996. – T. 232. – No. 1-3. – Pp. 45-54.

12. Bazilevsky M.P., Noskov S.I. Algorithm for the formation of a multitude of regression models by transforming a dependent variable // International Journal of Applied and Fundamental Research. – 2011. – No. 3. – Pp. 159-160.

13. Noskov S.I. Point characterization of the Pareto set in a linear multicriteria problem // Modern Technologies. System analysis. Modeling. – 2008. – No. 1 (17). – Pp. 99-101.

14. Lakeyev A.V., Noskov S.I. A description of the set of solutions of a linear equation with interval defined operator and right-hand side // Doklady Mathematics. – 1993. – T. 47. – No. 3. – P. 518.

15. Lakeev A.V., Noskov S.I. The least module method for linear regression: the number of zero approximation errors // Modern Technologies. System analysis. Modeling. – 2012. – No. 2 (34). – Pp. 48-50.

16. Noskov S.I. A generalized criterion for the consistency of behavior in regression analysis // Information technology and mathematical modeling in the management of complex systems. – 2018. – No. 1 (1). – Pp. 14-20.

17. Noskov S.I. On the method of mixed estimation of linear regression parameters // Information technologies and mathematical modeling in the management of complex systems. – 2019 . – No. 1 (2). – Pp. 41-45.

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

19. Noskov S.I., Butin A.A. Methodological support for assessing the level of vulnerability of objects of informatization // Information technologies and problems of mathematical modeling of complex systems. -Irkutsk: IrGUPS, 2015. -№. 14.-C. 38-48.

Информация об авторах

Сергей Иванович Носков – д. т. н., профессор, профессор кафедры «Информационные системы и защита информации», Иркутский государственный университет путей сообщения, г. Иркутск, e-mail: noskov_s@irgups.ru

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

Authors

Sergey Ivanovich Noskov, Doctor of Technical Science, Professor, the Subdepartment Information systems and information security, Irkutsk State Transport University, Irkutsk, e-mail: noskov_s@irgups.ru

Anton Andreyevich Khonyakov, Postgraduate Student, "Information systems and information security", Irkutsk State Transport University, Irkutsk, e-mail: anton_khonyakov@mail.ru

Для цитирования

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)

For citations

Noskov S.I., Khonyakov A.A. Programmnyj kompleks postroeniya nekotoryh tipov kusochno-linejnyh regressij [Software complex for building some types pieces of linear regressions] // Informacionnye tehnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami: ehlektronnyj nauchnyj zhurnal [Information technology and mathematical modeling in the management of complex systems: electronic scientific journal], 2019. No. 3. P. 47-55. [Accessed 20/11/19]