 DOI 10.25699/SSSB.2021.37.3.015

05.13.18

ПРОГРАММНЫЙ КОМПЛЕКС ПОСТРОЕНИЯ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ МЕТОДОМ СМЕШАННОГО ОЦЕНИВАНИЯ

С.И. Носков, К.С. Перфильева
Иркутский государственный университет путей сообщения, г. Иркутск

В статье описывается программный комплекс, обеспечивающий возможность оценивания параметров линейного регрессионного уравнения методами смешанного оценивания (МСО), наименьших квадратов (МНК) и модулей (МНМ), а также антиробастного оценивания (МАО). Разработанный программный комплекс применен для моделирования объема погрузки основных видов грузов железнодорожным транспортом, в качестве независимых переменных определены объемы погрузки конкурирующими видами транспорта. Это такие факторы, как перевозки автомобильным, морским, трубопроводным и внутренним водным транспортом. Проведён анализ полученных моделей, оценены значения критериев смещения, относительных ошибок аппроксимации, согласованности поведения.

Ключевые слова: линейная регрессия, метод смешанного оценивания, метод наименьших модулей, антиробастное оценивание, программный комплекс.

ВВЕДЕНИЕ

В различных областях человеческой деятельности повседневно возникает необходимость решения задач анализа и выявления явных и скрытых закономерностей функционирования исследуемых процессов. Одним из наиболее востребованных методов анализа данных является регрессионный анализ. Существует множество программных продуктов, где реализованы соответствующие вычислительные процедуры.

Достаточно популярным является программный пакет для эконометрического анализа Gretl, с помощью которого в статье [1] исследуется рынок вторичного жилья г. Биробиджана. В работе [2] численно исследуется эффективность ряда робастных методов оценивания. В [3] установлено, что применение программных продуктов, реализующих методы регрессионного моделирования, эффективно использовать в учебном процессе. В работе [4] подробно рассмотрена технология организации «конкурса» моделей, которая заключается в формировании множества альтернативных вариантов регрессий и последующем выборе лучшей из них с использованием многокритериального подхода. В рамках «конкурса» предложены и рассмотрены четыре формы регрессионных зависимостей: аддитивная, с использованием эффекта запаздывания, с преобразованием зависимой переменной и линейно-мультипликативная. Рассмотренная технология реализована в программном комплексе автоматизации процесса построения регрессионных моделей. В [5] представлен программный комплекс, предназначенный для интеллектуального анализа данных. Он базируется на платформе JAWA и имеет многомодульную структуру.

В данной работе представлено описание программного комплекса построения линейных регрессионных уравнений методом смешанного оценивания параметров с использованием критерия смещения [6-9]. Этот метод предполагает разбиение исходной выборки данных на две подвыборки с номерами наблюдений из множеств N1 и N2, при этом на первой МСО «работает» как МНМ, а на второй – как МАО.

ОСНОВНАЯ ЧАСТЬ

Интерфейс программного комплекса (ПК) реализован с использованием языка программирования Python.

ПК включает в себя две подсистемы:

-подсистема формирования задачи линейного программирования (ЛП);

-подсистема обработки результатов.

Подсистема формирования задачи ЛП на основе заданным пользователем данных формирует ее матрицу ограничений, правую часть и целевой вектор. Подсистема обработки результатов предназначена для поиска решения задачи ЛП и его трансформации в параметры модели.

Общий алгоритм работы программного комплекса представлен на рис. 1.

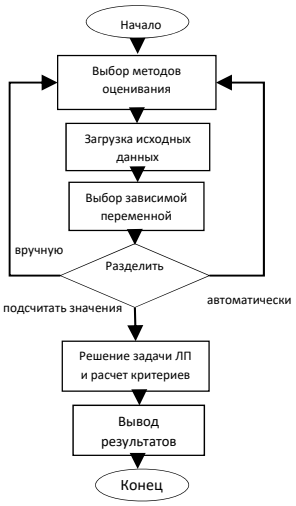

Рис. 1. Общий алгоритм работы ПК

Таким образом, процесс работы ПК начинается с этапа ввода в систему исходных статистических данных и выбора методов оценивания параметров модели. Главное окно ПК представлено на рис. 2 и содержит шесть перекрывающих друг друга страниц: «Основное меню», «Загрузить исходные данные», «Загруженная матрица», «Разделить матрицу», «Результаты решения», «Оценки критерия смещения».

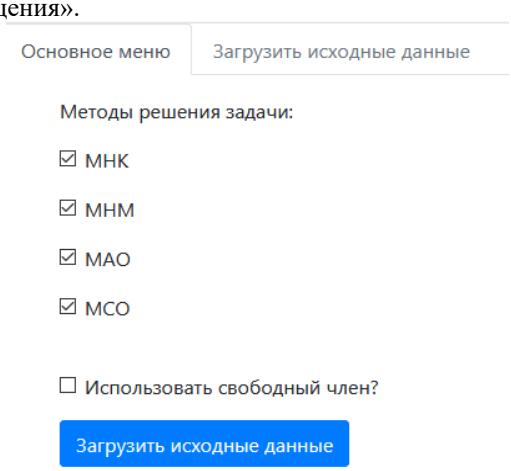

Рис. 2. Главное окно ПК

Первым этапом работы является выбор методов решения задачи и решение вопроса о включении в модель свободного члена. Далее пользователь переходит к вводу в систему исходных статистических данных. Это можно сделать путем импорта из текстового файла с расширение .txt. с помощью кнопок «Обзор» и «Загрузить» (рис. 3). В текстовом файле не должно содержаться никакой информации, кроме выборочных данных. Наблюдения должны отделяться друг от друга клавишей «Tab». Десятичным разделителем вещественных чисел является знак «.». После выполнения всех этих манипуляций на экран выводится таблица исходных статистических данных.

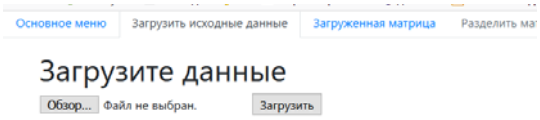

Рис. 3. Загрузка исходных данных

На следующем шаге пользователь должен выбрать зависимую переменную y, остальные загруженные переменные будут считаться независимыми.

Далее пользователю предоставляется выбор:

- автоматически распределить данные на подвыборки с множествами номеров N1 и N2;
- распределить данные на подвыборки вручную.
- использовать при распределении значения критерия смещения (этот вариант предполагает построение всех возможных вариантов модели при разбиении выборки на подвыборки).

Программный код формирования задачи ЛП с последующим применением МСО представлен в листинге 1.

Листинг 1. Описание задачи ЛП для применения МСО

```
class LpSolveMCO(LpSolve):
    def __init__(self, x, y, h1, h2):
        super().__init__(x, y)
        self.H1 = h1
        self.H2 = h2
```

Здесь x- вектор независимых переменных, y- зависимая переменная, h1и h2 – длина первой и второй подвыборок соответсвенно. Класс LpSolveMCO наследуется от LpSolve (Листинг 2).

Листинг 2. Описание класса LpSolve

```
class LpSolve:
    def __init__(self, x, y, ):
        self.x = x
        self.y = y
        self.var = {}
        self.problem = pulp.LpProblem('0', pulp.const.LpMinimize)
        self._create_variable_a()
        self._create_variable_u_v()
        self._create_variable_r()
```

Здесь функции self._create_variable_a(), self._create_variable_u_v(), self._create_variable_r() создают все необходимые данные для последующих вычислений.

Создание задачи ЛП реализовано шаблонным методом и представлено в листинге 3.

Листинг 3. Создание задачи ЛП

```
def build_task_lp(self):
    self.create_c()
    for index in range(len(self.x)):
        self.build_problem_a(index)
    i = 1 + len(self.x)
    for index in range(len(self.x)):
        if index in self.H2:
            self.build_problem_u_v(index, i)
```

```
            i += 1
```

Здесь процедура self.create_c() создаёт целевую функцию, а self.build_problem_a(index), self.build_problem_u_v(index, i) - ограничения задачи.

В листинге 4 описана реализация паттерна для целевой функции и ограничений задачи ЛП.

Листинг 4. Реализация паттерна для целевой функции и ограничений

```
    def create_c(self):
        self.problem += LpAffineExpression(self._build_func_c()), 'Функция цели'
    def _build_func_c(self):
        params = []
        for index in range(len(self.x)):
            if index in self.H1:
                params.append((self._get_u_by_i(index + 1), 1 / len(self.H1),))
                params.append((self._get_v_by_i(index + 1), 1 / len(self.H1),))
        params.append((self._get_r(), 1,))
        return params
    def build_problem_a(self, index_line):
        self.problem += LpAffineExpression(self._init_list(index_line)) == self.y[index_line], str(index_line + 1)
    def _init_list(self, index_line):
        my_list = []
        i = 1
        for item in self.x[index_line]:
  my_list.append(self._create_tuples(self._get_a_by_ij(i, 1), item))
my_list.append(self._create_tuples(self._get_a_by_ij(i, 2), -item))
            i += 1
my_list.append(self._create_tuples(self._get_u_by_i(index_line + 1), 1))
my_list.append(self._create_tuples(self._get_v_by_i(index_line + 1), -1))
        return my_list
    def build_problem_u_v(self, index_line, iteration):
        self.problem += LpAffineExpression(self._init_list_u_v(index_line)) <= 0, str(iteration)
    def _init_list_u_v(self, index_line):
        my_list = [(self._get_u_by_i(index_line + 1), 1,),
                   (self._get_v_by_i(index_line + 1), 1,),
                   (self._get_r(), -1,)]
        return my_list
```

Результаты вычислений открываются в новом окне ПК и представлены в виде таблиц «Найденные значения параметров», «Ошибки аппроксимации», «Значения зависимой переменной», «Критерии».

Помимо построения линейной регрессионной модели с помощью МСО ПК позволяет решать эту задачу с помощью МНК, МНМ, МАО, а также рассчитать значения перечисленных ниже критериев адекватности.

Рассмотрим применение ПК для решения задачи моделирования объема перевозки грузов железнодорожным транспортом (зависимая переменная у). Отметим, что подобная задача решалась в работе [10] путем построения регрессионной функции риска. При этом использовалась официальная статистика за период с января 2019 г. по май 2020 г.

В качестве независимых переменных были назначены следующие:

x_1– объем перевозки грузов автомобильным транспортом;

x_2– объем перевозки грузов внутренним водным транспортом;

x_3– объем перевозки грузов воздушным транспортом;

x_4– объем перевозки грузов морским транспортом;

x_5– объем перевозки грузов трубопроводным транспортом.

В результате бала сформирована выборка из 17 наблюдений по переменным у, x_1, x_2, x_3, x_4, x_5.

Решались следующие задачи.

Задача 1. Построить регрессионные модели с помощью классических методов регрессионного моделирования (МНК, МНМ, МАО).

Задача 2. Построить регрессионную модель с помощью МСО.

Задача 3. Вычислить значения некоторых критериев адекватности, в том числе смещения и согласованности поведения.

В табл.1 представлены вычисленные значения параметров для каждого метода по модели

$$y = \alpha_0 + \sum_{i=1}^{5} \alpha_i x_i$$

Табл. 1. Значения оценок параметров

Метод	a0	a1	a2	a3	a4	a5
МНК	95.080	0.030	-0.039	-728.524	8.378	0.387
МНМ	0.000	0.027	0.215	471.466	4.831	0.373
МАО	0.000	0.033	-0.214	419.224	10.071	0.369
МСО	25.923	0.030	-0.015	0.000	12.745	0.484

В табл. 2 представлены следующие критерии адекватности, значения которых рассчитаны с помощью программного комплекса для каждого метода:

- E - средняя относительная ошибка аппроксимации;
- M - сумма модулей ошибок;
- K - сумма квадратов ошибок;
- O- максимальная по модулю ошибка;
- Ф - обобщенный критерий согласованности поведения.

Табл.2. Значения критериев адекватности

Метод	E	M	K	O	Ф
МНК	18.58	332.78	6630.37	25.15	103.00
МНМ	1.83	33.38	155.94	8.24	102.00
МАО	2.41	42.81	139.84	3.75	98.00
МСО	2.21	39.60	131.95	4.49	101.00

Отметим, что с помощью программного комплекса можно построить все возможные варианты модели с помощью МСО. Это позволяет, в частности, выявить диапазоны значений используемых критериев, что может быть полезно при оценке качества полученных вариантов.

Исследователю предлагается остановить свой выбор на том разбиении выборки на подвыборки, у которого значение критерия смещения будет максимальным. В работах [8,9] подробно описан этот критерий и проведен анализ эмпирических свойств метода смешанного оценивания параметров.

Таким образом, в результате применения МСО была построена следующая модель:

$$y= 78.03 - 0.05\ x_1 - 0.53x_2 +1.794357x_4+0.05x_5$$

ЗАКЛЮЧЕНИЕ

Представленный выше программный комплекс позволяет осуществлять построение линейной регрессионной модели с помощью четырех методов оценивания неизвестных параметров: наименьших квадратов, модулей, антиробастного и смешанного оценивания. Помимо этого, для каждого из указанных методов существует возможность рассчитать значения критериев адекватности: среднюю относительныую ошибку аппроксимации, сумму модулей и квадратов ошибок, максимальную по модулю ошибку, согласованность поведения расчетных и фактических значений зависимой переменной.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1.Лагунова А.А., Баженов Р.И. Разработка в среде Gretl регрессионной модели рынка вторичного жилья г.Биробиджана//NAUKA-RASTUDENT.RU.2015.№ 1(13). с. 40.

2.Валеев С.Г., Куваискова Ю.Е., Юдкова М.В.. Робастные методы оценивания :программное обеспечение, эффективность//Вестник УлГТУ. 2010. с.29-33

3.Базилевский М.П., Носков С.И. Программный комплекс построения линейной регрессионной модели с учетом критерия согласованности поведения фактической и расчетной траекторий изменения значений объясняемой переменной // Вестник Иркутского государственного технического университета. 2017. Т. 21. № 9. С. 37-44

4.Базилевский М.П., Носков С.И. Методические и инструментальные средства построения некоторых типов регрессионных моделей// Системы.Методы.Технологии. 2012. №1 (13). с. 80-87.

5. Шаталова Ю.Г. Программный комплекс анализа данных для дистанционного обучения студентов// Ломоносовские чтения 2018. Сборник материалов ежегодной научной конференции. – 2018 – с.79-80.

6. Носков С.И. О методе смешанного оценивания параметров линейной регрессии// Информационные технологии и математическое моделирование в управлении сложными системами»: электрон. науч. журн. – 2019. – №1. – С. 14-20

7. Носков С.И. Метод смешанного оценивания параметров линейной регрессии: особенности применения// Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2021. №1. С.126-132.

8. Носков С.И., Перфильева К.С. Эмпирический анализ некоторых свойств метода смешанного оценивания параметров линейного регрессионного уравнения// Наука и бизнес. 2020. №6. с.62-66.

9. Носков С.И., Перфильева К.С. Моделирование объема погрузки на железнодорожном транспорте методом смешанного оценивания// Известия Тульского государственного университета. Технические науки. 2021.- №2. С.148-153

10. Носков С.И., Хоняков А.А. Применение функции риска для моделирования экономических систем// Южно-Сибирскй научный вестник,2020. №5(33). с. 85-92.

Сергей Иванович Носков, д-р техн. наук, проф., проф. кафедры «Информационные системы и защита информации», sergey.noskov.57@mail.ru, Россия, Иркутск, Иркутский государственный университет путей сообщения.

Перфильева Карина Сергеевна – аспирант кафедры «Информационные системы и защита информации», 552649-171233@mail.ru, Россия, Иркутск, Иркутский государственный университет путей сообщения.

SOFTWARE PACKAGE FOR CONSTRUCTING LINEAR REGRESSION MODELS USING THE MIXED ESTIMATION METHOD

S. I. Noskov, K. S. Perfileva
Irkutsk, Irkutsk state University of railway transport, Irkutsk

Abstract – The article considers a software package that provides the ability to estimate the parameters of a linear regression equation using mixed estimation (MSO), least squares (OLS) and modules (MCM), as well as anti-blast estimation (MAO). The developed software package was used to model the volume of loading of the main types of cargo by rail, and the competing types of transportation in relation to rail transportation were selected as independent variables. These are such factors as transportation by road, transportation by sea, transportation by pipeline, transportation by inland water transport. The analysis of the obtained models is carried out, the bias criteria, the relative approximation errors, and the generalized behavior consistency criterion (RSPC) are evaluated.

Index terms. Linear regression, mixed estimation method, least modules method, anti-robust estimation, software package.

REFERENCES

1. Lagunova A. A., Bazhenov R. I. Development of a regression model of the secondary housing market in Birobidzhan in the Gretl environment / / NAUKA-RASTUDENT. RU. 2015. no. 1(13). p.40.

2.Valeev S. G., Kuvaiskova Yu. E., Yudkova M. V. Robust evaluation methods: software, efficiency//Bulletin of UlSTU. 2010. p. 29-33

3. Bazilevsky M. P., Noskov S. I. Software package for constructing a linear regression model taking into account the criterion of consistency of the behavior of the actual and calculated trajectories of changes in the values of the explained variable // Bulletin of the Irkutsk State Technical University. 2017. Vol. 21. no. 9. pp. 37-44

4.Bazilevsky M. P., Noskov S. I. Methodological and instrumental means of constructing some types of regression models.Methods.Technologies. 2012. No. 1 (13). pp. 80-87.

5. Shatalova Yu. G. Software package for data analysis for distance learning of students// Lomonosov Readings 2018. Collection of materials of the annual scientific Conference. - 2018-p. 79-80.

6. Noskov S. I. On the method of mixed estimation of linear regression parameters / / Information technologies and mathematical modeling in the management of complex systems": electron. scientific journal-2019. - No. 1. - p. 14-20

7. Noskov S. I. Method of mixed estimation of linear regression parameters: application features// Bulletin of the VSU. Series: System Analysis and Information Technologies. 2021. No. 1. pp. 126-132.

8. Noskov S. I., Perfileva K. S. Empirical analysis of some properties of the method of mixed estimation of parameters of a linear regression equation// Science and Business. 2020. No. 6. pp.

62-66. 9. Noskov S. I., Perfileva K. S. Modeling of the volume of loading on railway transport by the method of mixed estimation// Proceedings of the Tula State University. Technical sciences. 2021. - No. 2. pp. 148-153

10. Noskov S. I., Khonyakov A. A. Application of the risk function for modeling economic systems / / Yuzhno-Sibirsky nauchnyj vestnik, 2020. No. 5(33). pp. 85-92.

Sergey Noskov - doctor of technical Sciences, Professor, Professor of the Department "Information systems and information protection", sergey.noskov.57@mail.ru, Russia, Irkutsk, Irkutsk state University of railway transport.

Perfileva Karina Sergeevna-post-graduate student of the Department "Information systems and information protection", 552649-171233@mail.ru, Russia, Irkutsk, Irkutsk state University of railway transport.