Научная статья
УДК 519.852
doi: 10.34822/1999-7604-2022-2-61-65

ПОСТРОЕНИЕ КУСОЧНО-ЛИНЕЙНОЙ РЕГРЕССИИ С ИНТЕРВАЛЬНОЙ НЕОПРЕДЕЛЕННОСТЬЮ В ДАННЫХ ДЛЯ ЗАВИСИМОЙ ПЕРЕМЕННОЙ

Сергей Иванович Носков
Иркутский государственный университет путей сообщения, Иркутск, Россия
sergey.noskov.57@mail.ru, http://orcid.org/0000-0003-4097-2720

Аннотация. В работе рассмотрена задача построения кусочно-линейной регрессионной модели (называемой также производственной функцией Леонтьева, функцией с нулевой эластичностью замены ресурсов, а также функцией с постоянными пропорциями) по данным с интервальной неопределенностью для зависимой переменной. Приведен краткий обзор применения традиционных форм таких моделей, построенных по классическим, точечным данным, для оценки качества воздуха, анализа связи общественного здоровья с сельскохозяйственной деятельностью, оптимизации процессов очистки фрагментов антител, исследования пропускной способности аэропортов и решения некоторых других задач. В качестве функции потерь принята сумма модулей ошибок аппроксимации. Показано, что сформулированная задача сводится к задаче частично-булевого программирования приемлемой размерности. Ее решение не должно вызывать вычислительных трудностей ввиду существующего значительного арсенала соответствующих эффективных программных средств. Результаты работы могут быть полезны при исследовании с помощью методов математического моделирования сложных технических и социально-экономических объектов с интервальной неопределенностью в исходных данных, вызванной сбоями в работе измерительных устройств, ошибками в деятельности статистических служб и другими причинами.

Ключевые слова: кусочно-линейная регрессия, функция Леонтьева, оценивание параметров, линейно-булевое программирование, задача линейного программирования

Для цитирования: *Носков С. И. Построение кусочно-линейной регрессии с интервальной неопределенностью в данных для зависимой переменной // Вестник кибернетики. 2022. № 2 (46). С. 61–65. DOI 10.34822/1999-7604-2022-2-61-65.*

Original article

CONSTRUCTING A DATA-DRIVEN PIECEWISE LINEAR REGRESSION WITH INTERVAL UNCERTAINTY FOR THE DEPENDENT VARIABLE

Sergey I. Noskov
Irkutsk State Transport University, Irkutsk, Russia
sergey.noskov.57@mail.ru, http://orcid.org/0000-0003-4097-2720

Abstract. The article discusses a problem of constructing a data-driven piecewise linear regression model (also known as Leontief production function, zero elasticity of substitution production function, and fixed proportions production function) with interval uncertainty for the dependent variable. A brief review of application of traditional forms of such models constructed according to the classical point data is given for assessing air quality, analyzing public health's relation to the agricultural activity, optimizing processes of antibodies' fragments purification, studying airport capacity, and solving other problems. A sum of approximation errors mode is taken as a loss function. The formulated problem is reduced to the partially Boolean programming problem of acceptable dimension. There should not emerge any calculating difficulties when solving the problem due to the existing large amount of acceptable effective software tools. The results of the study can be applied in research using methods of mathematical simulation of complicated technical and socially economic objects with interval uncertainty in the initial data caused by failures in the operation of measuring devices, errors in the activities of statistical services and other reasons.

Keywords: piecewise linear regression, Leontief function, parameter estimation, linear Boolean programming, linear programming problem

For citation: *Noskov S. I. Constructing a Data-Driven Piecewise Linear Regression with Interval Uncertainty for the Dependent Variable // Proceedings in Cybernetics. 2022. No. 2 (46). P. 61–65. DOI 10.34822/1999-7604-2022-2-61-65.*

ВВЕДЕНИЕ

При построении регрессионных моделей объектов различной природы используются как линейные, так и более сложные конструкции. Одной из них, часто применяемой при анализе экономических систем, является кусочно-линейная модель, называемая также производственной функцией Леонтьева, или функцией с нулевой эластичностью замены ресурсов. Так, в работе [1] применяется кусочно-полиномиальная аппроксимация для формирования точных оценок качества воздуха. В [2] с помощью параметрической модели кусочно-линейной регрессии изучается связь мультиметрического индекса общественного здоровья с сельскохозяйственной деятельностью в прилегающих водосборных бассейнах. Работа [3] посвящена масштабной оптимизации процессов очистки фрагментов антител на основе кусочно-линейного регрессионного моделирования. В статье [4] так называемая кусочно-линейная метарегрессия используется при исследовании предвзятости научных публикаций посредством искажения имеющихся эмпирических данных. В работе [5] описывается применение функций Леонтьева и Кобба – Дугласа при анализе свойств двумерной задачи факторного назначения технологии с учетом технологического меню, понимаемого как выбор фирмой-производителем степени приращения некоторого конкретного фактора или качества товара, востребованного потребителем. В [6] рассматривается кусочно-вогнутая функция полезности Леонтьева, состоящая из набора сегментов леонтьевского типа с убывающей отдачей и верхним пределом полезности на каждом сегменте, изучается сложность вычисления равновесий по Фишеру при задействовании модели биржевого рынка. Наконец, в статье [7] с помощью производственной функции Леонтьева исследуется пропускная способность аэропортов. При этом результаты расчетов показывают, что ее применение позволяет достаточно точно прогнозировать заторы, доступность инфраструктуры, выявлять факторы, блокирующие движение на земле, и определять время занятости взлетно-посадочной полосы.

МАТЕРИАЛЫ И МЕТОДЫ

Наиболее часто при регрессионном моделировании сложных систем применяется линейная модель (уравнение) вида:

$$y_k = \sum_{i=1}^{m} a_i x_{ki} + \varepsilon_k,\ k = \overline{1, n}, \qquad (1)$$

где y – зависимая переменная;

x_i – i-ая независимая переменная;

a_i – i-ый оцениваемый параметр;

ε_k – ошибки аппроксимации;

k – номер наблюдения;

n – число наблюдений (длина выборки).

Представим уравнение (1) в векторной форме:

$$y = Xa + \varepsilon, \qquad (2)$$

где $y = (y_1, ..., y_n)^T$, $a = (a_1, ..., a_m)^T$, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$, $X - (n \times m)$ – матрица с компонентами x_{ki}.

Пару (X, y) будем, в соответствии с принятой терминологией [8, 9], называть выборкой данных.

Несколько менее популярной является кусочно-линейная модель вида:

$$y_k = \min\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, ..., \alpha_m x_{km}\} + \varepsilon_k. \qquad (3)$$

Ее характерной особенностью является то, что производство продукции системой (переменная y) ограничено объемом лимитирующего ресурса, при этом любое наращивание объемов остальных ресурсов не приводит к росту производства.

В работе [10] исследована задача точной идентификации параметров $\alpha_i, i = \overline{1, m}$ уравнения (3) с использованием метода наимень-

ших модулей (далее – МНМ), состоящего в решении задачи:

$$J(\alpha) = \sum_{k=1}^{n} | \varepsilon_k | \to \min. \quad (4)$$

Введем в рассмотрение так называемые расчетные (т. е. вычисленные по модели (3)) значения выходной переменной Z_k:

$$Z_k = \min\{\alpha_1 x_{k1}, \alpha_2 x_{k2}, ..., \alpha_m x_{km}\}, k = \overline{1, n}, \quad (5)$$

после чего регрессия (3) представима в виде:

$$y_k = z_k + \varepsilon_k, k = \overline{1, n}, \quad (6)$$

или в векторной форме:

$$y = z + \varepsilon,$$

где $z = (z_1, ..., z_n)^T$.

Следуя стандартному приему раскрытия модулей в выражении (4) [11], введем в рассмотрение переменные u_k и v_k по правилу:

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

Нетрудно видеть, что имеют место тождества:

$$z_k + u_k - v_k = y_k, k = \overline{1, n}. \quad (7)$$

Из (5) следует справедливость неравенств:

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

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

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

$$\sum_{i=1}^{m} \sigma_{ki} = 1, k = \overline{1, n}, \quad (10)$$

где M – заранее выбранное большое положительное число.

Естественно ввести ограничения неотрицательности переменных:

$$u_k \ge 0, \ v_k \ge 0, k = \overline{1, n}. \quad (11)$$

Из задания переменных u_k и v_k следуют равенства:

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

что позволяет представить функцию (4) в виде:

$$J(\alpha) = \sum_{k=1}^{n} (u_k + v_k) \to \min. \quad (12)$$

Таким образом, задача (4) поиска значений неизвестных параметров $\alpha_i, i = \overline{1, m}$ кусочно-линейной регрессии (3) с помощью МНМ сводится к задаче линейно-булевого программирования (далее – ЛБП) – (7)–(12) с $mn + 3n + m$ переменными (из которых mn – булевы) и $2(mn + n)$ ограничениями.

Пусть теперь часть выборки – вектор y – задана не точно, а с интервальной неопределенностью, а именно: известен интервальный вектор $[y^-, y^+]$, которому принадлежит y. При этом любые соображения, в том числе вероятностные, уточняющие расположение y_{ki} на отрезке $[y_{ki}^-, y_{ki}^+]$, отсутствуют. Вектора y^-, y^+ считаются, таким образом, заданными. Причин проявления интервальной неопределенности в данных может быть несколько (см., например, [12]), основными из них являются погрешность технической измерительной аппаратуры и сбои в работе статистических служб.

Таким образом, по отношению к вычислению неизвестных оценок параметров кусочно-линейной модели (3) задача может быть сформулирована следующим образом: как адаптировать сведение задачи оптимизации (4) к задаче ЛБП (7)–(12) для случая с выборкой данных $(X, [y^-, y^+])$?

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Для решения поставленной задачи воспользуемся приемом, примененным в работе [13] для оценивания неизвестных параметров линейной модели (1) по выборке $(X, [y^-, y^+])$ в которой данные для зависимой переменной y имеют интервальный характер. В соответствии с этим приемом сначала решается задача линейного программирования (далее – ЛП):

$$X(\beta-\gamma)+u\geq y^{-}, \quad (13)$$

$$X(\beta-\gamma)-v\leq y^{+}, \quad (14)$$

$$u\geq 0,\ v\geq 0,\ \beta\geq 0,\ \gamma\geq 0, \quad (15)$$

$$\sum_{k=1}^{n}(u_k+v_k)\to\min, \quad (16)$$

после чего вектор параметров α рассчитывается по формуле:

$$\alpha=\beta-\gamma, \quad (17)$$

где β – положительная часть вектора α, γ – его отрицательная часть.

Если после решения задачи ЛП (13)–(16) окажется, что $u=v=0,$ в [13] предлагается максимизировать разрешающую способность ограничений (13), (14) посредством решения задачи ЛП:

$$X(\beta-\gamma)-u\geq y^{-}, \quad (18)$$

$$X(\beta-\gamma)+v\leq y^{+}, \quad (19)$$

$$u\geq 0,\ v\geq 0,\ \beta\geq 0,\ \gamma\geq 0, \quad (20)$$

$$\sum_{k=1}^{n}(u_k+v_k)\to\min, \quad (21)$$

также с последующим использованием формулы (17).

Займемся теперь анонсируемой выше адаптацией сведения задачи оптимизации (4) к задаче ЛБП (7)–(12) для случая с выборкой данных $(X,[y^{-},y^{+}])$.

Равенства (6) преобразуются в две системы неравенств, аналогичных (13), (14):

$$z+u\geq y^{-}, \quad (22)$$

$$z-v\leq y^{+}. \quad (23)$$

После этого оценки параметров модели (3) рассчитываются посредством решения задачи ЛБП (22), (23), (8)–(12). Если же, как и при решении задачи (13)–(16), окажется, что $u=v=0,$ следует произвести замену ограничений (22), (23) на следующие:

$$z-u\geq y^{-}, \quad (24)$$

$$z+v\leq y^{+} \quad (25)$$

и решать задачу ЛП (8)–(12), (24), (25), (21).

Отметим, что необходимость решения задачи линейно-булевого программирования при оценивании параметров кусочно-линейной модели (3) как для точечной (X, y), так и для интервальной ($X,[y^{-},y^{+}]$) выборки не должно вызывать вычислительных трудностей из-за значительного существующего арсенала соответствующих эффективных программных средств (например, размещенной в Интернете в свободном доступе программы LPsolve, использование которой позволяет решать эту задачу за вполне приемлемое время для размерностей, соответствующих реальным объектам моделирования).

ЗАКЛЮЧЕНИЕ

В работе рассмотрена задача оценивания параметров кусочно-линейной регрессии по данным с интервальной неопределенностью для зависимой переменной. Показано, что эта задача сводится к задаче частично-булевого программирования. Ее решение не должно вызывать затруднений ввиду существующего значительного арсенала соответствующих эффективных программных средств (например, размещенной в Интернете в свободном доступе программы LPsolve и некоторых других разработок [14–16]).

Результаты работы могут быть полезны при исследовании с помощью методов математического моделирования сложных технических и социально-экономических объектов с интервальной неопределенностью в исходных данных, вызванной сбоями в работе измерительных устройств, ошибками в деятельности статистических служб и другими причинами.

Список источников

1. Mo X., Li H., Zhang L., Qu Z. A Novel Air Quality Evaluation Paradigm Based on the Fuzzy Comprehensive Theory // Appl Sci. 2020. Vol. 10, No. 23. P. 8619.

References

1. Mo X., Li H., Zhang L., Qu Z. A Novel Air Quality Evaluation Paradigm Based on the Fuzzy Comprehensive Theory // Appl Sci. 2020. Vol. 10, No. 23. P. 8619.

2. Tomal J. H., Ciborowski J. J. H. Ecological Models for Estimating Breakpoints and Prediction Intervals // Ecol Evol. 2020. Vol. 10, Is. 23. P. 13500–13517.
3. Liu S., Papageorgiou L. G. Optimal Antibody Purification Strategies Using Data-Driven Models // Engineering. 2019. Vol. 5, Is. 6. P. 1077–1092.
4. Bom P. R. D., Rachinger H. A Kinked Meta-Regression Model for Publication Bias Correction // Res Synth Methods. 2019. Vol. 10, Is. 4. P. 497–514.
5. Growiec J. Factor-Specific Technology Choice // Journal of Mathematical Economics. 2018. Vol. 77. P. 1–14.
6. Garg J. Market Equilibrium under Piecewise Leontief Concave Utilities // Theoretical Computer Science. 2017. Vol. 703. P. 55–65.
7. Besma H., Riadh H., Rafaa M. Modeling of the Aerial Capacity through a Leontief Production Function: The Case of Tunisian Airports // Journal of Reviews on Global Economics. 2017. Vol. 6. P. 98–104.
8. Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М : Диалектика, 2017. 912 с.
9. Шуметов В. Г, Крюкова О. А. Методология и практика анализа данных в управлении: методы одномерного и двумерного анализа. Орел : ОФ РАНХиГС, 2013. 178 с.
10. Носков С. И. Оценивание параметров аппроксимирующей функции с постоянными пропорциями // Современ. технологии. Систем. анализ. Моделирование. 2013. № 2. С. 135–136.
11. Носков С. И., Хоняков А. А. Программный комплекс построения некоторых типов кусочно-линейных регрессий // Информ. технологии и математ. моделирование в упр. сложными системами. 2019. № 3 (4). С. 47–55.
12. Вощинин А. П., Сотиров Г. Р. Оптимизация в условиях неопределенности. М. : Изд-во МЭИ, 1989. 224 с.
13. Носков С. И. Построение экспертно-статистических моделей по неполным данным // T-Comm: Телекоммуникации и транспорт. 2021. № 6 (15). С. 33–39.
14. Есиков Д. О., Ивутин А. Н., Ларкин Е. В., Новиков А. С. Программа решения задач целочисленного линейного программирования с булевыми переменными : св-во о гос. регистрации программы для ЭВМ № 2015612409 Российская Федерация. EDN UCCUMV.
15. Фильгус Д. И. Программное обеспечение для решения задач булевого программирования : св-во о гос. регистрации программы для ЭВМ № 2019610724 Российская Федерация. EDN OUCAVZ.
16. Есиков Д. О. Программа распределенного решения задач целочисленного программирования с булевыми переменными островным генетическим алгоритмом на кластере : св-во о государственной регистрации программы для ЭВМ № 2018613135 Российская Федерация. EDN DSFWRF.

Информация об авторе

С. И. Носков – доктор технических наук, профессор, почетный работник сферы образования Российской Федерации.

2. Tomal J. H., Ciborowski J. J. H. Ecological Models for Estimating Breakpoints and Prediction Intervals // Ecol Evol. 2020. Vol. 10, Is. 23. P. 13500–13517.
3. Liu S., Papageorgiou L. G. Optimal Antibody Purification Strategies Using Data-Driven Models // Engineering. 2019. Vol. 5, Is. 6. P. 1077–1092.
4. Bom P. R. D., Rachinger H. A Kinked Meta-Regression Model for Publication Bias Correction // Res Synth Methods. 2019. Vol. 10, Is. 4. P. 497–514.
5. Growiec J. Factor-Specific Technology Choice // Journal of Mathematical Economics. 2018. Vol. 77. P. 1–14.
6. Garg J. Market Equilibrium under Piecewise Leontief Concave Utilities // Theoretical Computer Science. 2017. Vol. 703. P. 55–65.
7. Besma H., Riadh H., Rafaa M. Modeling of the Aerial Capacity through a Leontief Production Function: The Case of Tunisian Airports // Journal of Reviews on Global Economics. 2017. Vol. 6. P. 98–104.
8. Draper N. R., Smith H. Applied Regression Analysis. Moscow : Dialektika, 2017. 912 p. (In Russian).
9. Shumetov V. G., Kryukova O. A. Metodologiia i praktika analiza dannykh v upravlenii: metody odnomernogo i dvumernogo analiza. Orel : Orel Branch of RANEPA, 2013. 178 p. (In Russian).
10. Noskov S. I. Otsenivanie parametrov approksimiruiushchei funktsii s postoiannymi proportsiiami // Modern Technologies. System Analysis. Modeling. 2013. No. 2. P. 135–136. (In Russian).
11. Noskov S. I., Khonyakov A. A. Software Complex for Building Some Types Pieces of Linear Regressions // Information Technology and Mathematical Modeling in the Management of Complex Systems. 2019. No. 3 (4). P. 47–55. (In Russian).
12. Voshchinin A. P., Sotirov G. P. Optimizatsiia v usloviiakh neopredelennosti. Moscow : Publishing House Moscow Power Engineering Institute, 1989. 224 p. (In Russian).
13. Noskov S. I. Construction of Expert-Statistical Models from Incomplete Data // T-Comm. 2021. No. 6 (15). P. 33–39. (In Russian).
14. Esikov D. O., Ivutin A. N., Larkin E. V., Novikov A. S. Program for Problem Solving of Integer Linear Programming with Boolean Variables : Certificate of Registration of a Computer Program No. 2015612409, Russian Federation. EDN UCCUMV. (In Russian).
15. Filgus D. I. Software for Solving Boolean Programming Problems : Certificate of Registration of a Computer Program No. 2019610724, Russian Federation. EDN OUCAVZ. (In Russian).
16. Esikov D. O. Program of Distributed Solution of Problems of Integer Programming with Boolean Variables by Island Genetic Algorithm by Cluster : Certificate of Registration of a Computer Program No. 2018613135, Russian Federation. EDN DSFWRF. (In Russian).

Information about the author

S. I. Noskov – Doctor of Sciences (Engineering), Professor, Honored Worker of Education of the Russian Federation.