

Раздел 2 МОДЕЛИ, СИСТЕМЫ, СЕТИ В ТЕХНИКЕ

Section 2 MODELS, SYSTEMS, NETWORKS IN THE TECHNIQUE

УДК 519.852
doi: 10.21685/2227-8486-2024-3-8

МЕТОД СМЕШАННОГО ОЦЕНИВАНИЯ ПАРАМЕТРОВ РЕГРЕССИОННОЙ МОДЕЛИ С ПРОИЗВОЛЬНЫМ СОСТАВОМ ГРУПП НАБЛЮДЕНИЙ

С. И. Носков¹, Е. С. Попов²

^{1, 2} Иркутский государственный университет путей сообщения, Иркутск, Россия
¹sergey.noskov.57@mail.ru, ²eglr5732@mail.ru

Аннотация. *Актуальность и цели.* Построение регрессионных моделей различных сложных объектов в соответствии с задачами исследования иногда сопровождается необходимостью разбиения исходной выборки данных на подвыборки (группы наблюдений) или использованием так называемых панельных данных. Цель исследования состоит в модификации разработанного и развитого ранее одним из авторов метода вычисления оценок параметров линейного регрессионного уравнения с помощью метода смешанного оценивания. *Материалы и методы.* Для достижения поставленной цели применялся математический аппарат решения задач линейного и линейно-булева программирования. *Результаты.* Эта цель формализована путем постановки задачи автоматизации формирования состава индексного множества номеров наблюдений для реализации антиробастности в рамках метода смешанного оценивания путем решения задачи линейно-булева программирования. *Выводы.* Описанный в работе подход позволяет совместить при моделировании достоинства методов наименьших модулей и антиробастного оценивания. Построена адекватная анализируемому объекту регрессионная модель грузооборота автомобильного транспорта Российской Федерации.

Ключевые слова: регрессионная модель, методы оценивания параметров, метод смешанного оценивания, задачи линейного и линейно-булева программирования, грузооборот автомобильного транспорта

Для цитирования: Носков С. И., Попов Е. С. Метод смешанного оценивания параметров регрессионной модели с произвольным составом групп наблюдений // Модели, системы, сети в экономике, технике, природе и обществе. 2024. № 3. С. 98–104. doi: 10.21685/2227-8486-2024-3-8

© Носков С. И., Попов Е. С., 2024. Контент доступен по лицензии Creative Commons Attribution 4.0 License / This work is licensed under a Creative Commons Attribution 4.0 License.

THE METHOD OF MIXED ESTIMATION OF REGRESSION MODEL PARAMETERS WITH AN ARBITRARY COMPOSITION OF OBSERVATION GROUPS

S.I. Noskov¹, E.S. Popov²

^{1, 2} Irkutsk State Transport University, Irkutsk, Russia

¹sergey.noskov.57@mail.ru, ²eglr5732@mail.ru

Abstract. *Background.* The construction of regression models of various complex objects in accordance with the objectives of the study is sometimes accompanied by the need to split the original data sample into subsamples (groups of observations) or the use of the so-called panel data. The purpose of the study is to modify the method for calculating estimates of the parameters of a linear regression equation developed earlier by one of the authors, using the mixed estimation method. *Materials and methods.* To achieve this goal, the mathematical apparatus for solving linear and linear Boolean programming problems was used. *Results.* This goal is formalized by setting the task of automating the formation of the composition of the index set of observation numbers to implement the antirobust estimation method within the framework of the mixed estimation method by solving a linear Boolean programming problem. *Conclusions.* The approach described in the work allows us to combine the advantages of the least modulus and anti-robust estimation methods when modeling. A regression model of freight turnover of road transport in the Russian Federation has been constructed that is adequate to the analyzed object.

Keywords: regression model, parameters estimation methods, mixed estimation method, linear and linear Boolean programming problems, road transport freight turnover

For citation: Noskov S.I., Popov E.S. The method of mixed estimation of regression model parameters with an arbitrary composition of observation groups. *Modeli, sistemy, seti v ekonomike, tekhnike, prirode i obshchestve* = *Models, systems, networks in economics, technology, nature and society*. 2024;(3):98–104. (In Russ.). doi: 10.21685/2227-8486-2024-3-8

Введение

Построение регрессионных моделей различных сложных объектов в соответствии с целями исследования иногда сопровождается разбиением исходной выборки данных на подвыборки (группы наблюдений) или использованием так называемых панельных данных. Так, в работе [1] этот прием используется при разработке нового подхода к калибровке условной стоимости риска финансовых учреждений на основе квантильной регрессии нейронной сети. В работе [2] вводится понятие модальной регрессии с фиксированными эффектами для панельных данных, в которой исследуется, как условный режим переменной отклика зависит от ковариат. Предлагаемая модальная регрессия может дополнять регрессию среднего и квантиля и обеспечивать лучшую оценку основной тенденции и эффективность прогнозирования, когда данные искажены. Исследование [3] посвящено выводу условной меры причинности Грейнджера на основе модели параметрической динамической квантильной регрессии. Предложенный метод позволяет более детально охарактеризовать причинные взаимодействия на всем распределении. Предлагаемые тесты имеют разумный размер и мощность по отношению к множеству эмпирически правдоподобных альтернатив в конечных выборках. В статье [4] предлагаются два новых алгоритма для оценки последовательности квантильных регрессий по многим квантильным индексам. В статье [5] предлагаются

быстрые алгоритмы формирования подвыборок для эффективного приближения оценки максимального правдоподобия в логистической регрессии. Работа [6] посвящена методам формирования подвыборок, основанным на моделях логистической регрессии, модели регрессии softmax, квантильной регрессии и оценке квазиправдоподобия. В исследовании [7] описан способ разбиения исходной выборки на подвыборки и восстановления регрессионной модели для каждой из них. В статье [8] изучается влияние такого разбиения на основные характеристики законов распределений. Исследование [9] посвящено проблеме применения теории планирования эксперимента к пассивному эксперименту. Описан алгоритм формирования подвыборок из результатов пассивного эксперимента с необходимыми свойствами. В работе [10] рассмотрены новые подходы к обработке массивов статистических данных, а также к использованию метода складного ножа и бутстреп-процедур в регрессионных моделях прогноза погоды на территории Казахстана.

Материалы и методы

В работах [11–14] описан и развит разработанный одним из авторов метод смешанного оценивания (МСО) параметров линейного регрессионного уравнения (модели):

$$y_k = \sum_{i=1}^m \alpha_i x_{ki} + \varepsilon_k, \quad k \in P = \{1, 2, \dots, n\}, \quad (1)$$

где y – зависимая (объясняемая); x_i – i -я независимая (объясняющая) переменные; α_i – i -й неизвестный параметр; ε_k – ошибки аппроксимации; k – номер наблюдения; n – их количество (длина выборки); P – индексное множество номеров наблюдений.

Пусть, исходя из соображений самого разного характера (общий случай описан в работе [11]), обрабатываемая информация (выборка P) может быть разбита на две части – P_1 и P_2 :

$$P_1 \cup P_2 = P, \quad P_1 \cap P_2 = \emptyset.$$

При этом поиск неизвестных параметров на выборке с номерами из P_1 осуществляется по методу наименьших модулей (МНМ):

$$J^1(\alpha) = \sum_{k \in P_1} |\varepsilon_k| \rightarrow \min,$$

а с номерами из P_2 – по методу антиробастного оценивания (МАО):

$$J^2(\alpha) = \max_{k \in P_2} |\varepsilon_k| \rightarrow \min.$$

Обратим внимание, что МНМ игнорирует возможные выбросы информации, МАО же, напротив, на них сильно реагирует и отклоняет в их направлении гиперплоскость регрессии.

Суть МСО заключается в *одновременном* применении МНМ на подвыборке P_1 и МАО – на P_2 . Реализация МСО, таким образом, приводит к решению следующей задачи линейного программирования (ЛП):

$$\sum_{i=1}^m \alpha_i x_{ki} + u_k - v_k = y_k, \quad k \in P, \quad (2)$$

$$u_k + v_k - r \leq 0, \quad k \in P_2, \quad (3)$$

$$u_k \geq 0, v_k \geq 0, \quad k \in P, \quad (4)$$

$$\sum_{k \in P_1} (u_k + v_k) / d + r + \delta \sum_{k \in P_2} (u_k + v_k) \rightarrow \min, \quad (5)$$

где d – число элементов в множестве P_1 ($d = |P_1|$); δ – заранее выбранное малое положительное число.

Результаты

Сформулируем идею метода смешанного оценивания несколько более общим образом, а именно, пусть исследователь не формирует состав индексного множества P_1 (а значит, и P_2 , поскольку $P_2 = P \setminus P_1$), а указывает лишь число h элементов в P_2 . В этом случае метод смешанного оценивания может быть сведен к последовательному решению двух задач – линейно-булева программирования (ЛБП) и ЛП.

Сформируем системы ограничений:

$$u_k + v_k - r_1 + M s_k \leq M, \quad k \in P, \quad (6)$$

$$\sum_{k \in P} s_k = h, \quad (7)$$

$$s_k \in \{0, 1\}, \quad k \in P. \quad (8)$$

Тогда после решения задачи ЛБП с ограничениями (2), (4), (6)–(8) и целевой функцией

$$r_1 + \delta \sum_{k \in P} (u_k + v_k) \rightarrow \min \quad (9)$$

состав индексного множества P_2 может быть сформирован по правилу:

$$s_k = 1 \Rightarrow k \in P_2,$$

после чего множество P_1 формируется по правилу: $P_1 = P \setminus P_2$.

После решения задачи ЛБП (2), (4), (6)–(9) параметры линейной регрессионной модели (1) рассчитываются посредством решения задачи ЛП (2)–(5), где $d = n - h$.

Обсуждение

Проиллюстрируем предложенный способ оценивания параметров линейной регрессии на данных из работы [15], сформированных на основе статистической информации за 2007–2020 гг. (т.е. $n = 14$), используя обозначения: y – грузооборот автомобильного транспорта Российской Федерации, млрд т.км; x_1 – протяженность автомобильных дорог общего пользования федерального, регионального или межмуниципального и местного значения, тыс.

км; x_2 – объем погрузки основных видов грузов на автомобильном транспорте, тыс. т.

Будем строить линейное регрессионное уравнение

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \varepsilon \quad (10)$$

тремя методами – наименьших квадратов (МНК), модулей и смешанного оценивания с описанным выше обобщением.

1. МНК:

$$y = 9,11 + 0,094x_1 + 0,018x_2 + \varepsilon, \quad (11)$$

$R = 0,97$, $F = 205,4$, $T = (0.43, 18.37, 6.45)$, $E = 3,27$.

Здесь R – критерий множественной детерминации; F – критерий Фишера; T – вектор значений t -критерия Стьюдента; E – средняя абсолютная ошибка аппроксимации:

$$E = \sum_{k \in P} |\varepsilon_k| / n.$$

Очевидно, что модель (11) обладает высокими аппроксимационными характеристиками.

2. МНМ:

$$y = -8.52 + 0.099x_1 + 0.02x_2 + \varepsilon, \quad (12)$$

$$E = 3.02.$$

3. МСО:

$$y = -1.73 + 0.098x_1 + 0.019x_2 + \varepsilon, \quad (13)$$

$$E = 3.12.$$

Анализ моделей (11)–(13) показывает, что они довольно близки и сопоставимы по точности. Это объясняется отсутствием в данных явных выбросов – наблюдений, не согласующихся с выборкой в целом.

Заключение

В работе предложена модификация разработанного и развитого ранее одним из авторов метода вычисления оценок параметров линейного регрессионного уравнения с помощью метода смешанного оценивания. Она состоит в автоматизации формирования состава индексного множества номеров наблюдений для реализации МАО в рамках МСО путем решения задачи ЛБП. Решена задача построения регрессионной модели грузооборота автомобильного транспорта Российской Федерации.

Список литературы

1. Keilbar G., Wang W. Modelling systemic risk using neural network quantile regression // Empirical Economics. 2022. Vol. 62. P. 93–118.
2. Ullah A., Wang T., Yao W. Modal regression for fixed effects panel data // Empirical Economics. 2021. Vol. 60. P. 261–308.
3. Cheng H., Wang Yu., Wang Y., Yang T. Inferring Causal Interactions in Financial Markets Using Conditional Granger Causality Based on Quantile Regression // Computational Economics. 2022. Vol. 59. P. 719–748.

4. Chernozhukov V., Fernández-Val I., Melly B. Fast algorithms for the quantile regression process // *Empirical Economics*. 2022. Vol. 62. P. 7–33.
5. Wang H., Zhu R., Ma P. Optimal Subsampling for Large Sample Logistic Regression // *Journal of the American Statistical Association*. 2018. Vol. 113. P. 829–844.
6. Yao Y., Wang H. A Review on Optimal Subsampling Methods for Massive Datasets // *Journal of Data Science*. 2021. № 1. P. 151–172.
7. Стрижов В. В. Задача выбора многоуровневых моделей с анализом ковариационной матрицы параметров // *Математические методы распознавания образов*. 2011. Т. 15, № 1. С. 154–157.
8. Акимов С. С., Трипкош В. А. Исследование влияния объема массива данных на ключевые выборочные параметры различных законов распределения // *Современные наукоемкие технологии*. 2023. № 11. С. 10–15.
9. Лапач С. Н. Планирование в пассивном эксперименте // *Математические машины и системы*. 2013. № 4. С. 156–160.
10. Боголюбова Е. В. Применение метода складного ножа и бутстреп-процедуры в статистических моделях прогноза // *Гидрометеорология и экология*. 2004. № 1. С. 9–17.
11. Носков С. И. О методе смешанного оценивания параметров линейной регрессии // *Информационные технологии и математическое моделирование в управлении сложными системами*. 2019. № 1. С. 41–45.
12. Носков С. И. Метод смешанного оценивания параметров линейной регрессии: особенности применения // *Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии*. 2021. № 1. С. 126–132.
13. Носков С. И., Перфильева К. С. Эмпирический анализ некоторых свойств метода смешанного оценивания параметров линейного регрессионного уравнения // *Наука и бизнес: пути развития*. 2020. № 6. С. 62–66.
14. Носков С. И. Метод смешанного оценивания параметров линейной регрессии для данных с интервальной неопределенностью // *Известия Тульского государственного университета. Технические науки*. 2022. № 9. С. 274–277.
15. Перфильева К. С. Моделирование грузооборота на автомобильном транспорте с помощью метода смешанного оценивания регрессионных уравнений // *Молодая наука Сибири*. 2022. № 2. С. 248–251.

References

1. Keilbar G., Wang W. Modelling systemic risk using neural network quantile regression. *Empirical Economics*. 2022;62:93–118.
2. Ullah A., Wang T., Yao W. Modal regression for fixed effects panel data. *Empirical Economics*. 2021;60:261–308.
3. Cheng H., Wang Yu., Wang Y., Yang T. Inferring Causal Interactions in Financial Markets Using Conditional Granger Causality Based on Quantile Regression. *Computational Economics*. 2022;59:719–748.
4. Chernozhukov V., Fernández-Val I., Melly B. Fast algorithms for the quantile regression process. *Empirical Economics*. 2022;62:7–33.
5. Wang H., Zhu R., Ma P. Optimal Subsampling for Large Sample Logistic Regression. *Journal of the American Statistical Association*. 2018;113:829–844.
6. Yao Y., Wang H. A Review on Optimal Subsampling Methods for Massive Datasets. *Journal of Data Science*. 2021;(1):151–172.
7. Strizhov V.V. The problem of choosing multilevel models with the analysis of covariance matrix of parameters. *Matematicheskie metody raspoznavaniya obrazov = Mathematical methods of pattern recognition*. 2011;15(1):154–157. (In Russ.)
8. Akimov S.S., Tripkosh V.A. Investigation of the influence of the volume of the data array on the key sample parameters of various distribution laws. *Sovremennye naukoemkie tekhnologii = Modern high-tech technologies*. 2023;(11):10–15. (In Russ.)

9. Lapach S.N. Planning in a passive experiment. *Matematicheskie mashiny i sistemy = Mathematical machines and systems*. 2013;(4):156–160. (In Russ.)
10. Bogolyubova E.V. Application of the folding knife method and bootstrap procedure in statistical forecasting models. *Gidrometeorologiya i ekologiya = Hydrometeorology and Ecology*. 2004;(1):9–17. (In Russ.)
11. Noskov S.I. On the method of mixed estimation of linear regression parameters. *Informatsionnye tekhnologii i matematicheskoe modelirovanie v upravlenii slozhnymi sistemami = Information technologies and mathematical modeling in the management of complex systems*. 2019;(1):41–45. (In Russ.)
12. Noskov S.I. Method of mixed estimation of linear regression parameters: application features. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Sistemnyy analiz i informatsionnye tekhnologii = Bulletin of the Voronezh State University. Series: System analysis and Information Technology*. 2021;(1):126–132. (In Russ.)
13. Noskov S.I., Perfil'eva K.S. Empirical analysis of some properties of the method of mixed estimation of parameters of a linear regression equation. *Nauka i biznes: puti razvitiya = Science and business: ways of development*. 2020;(6):62–66. (In Russ.)
14. Noskov S.I. Method of mixed estimation of linear regression parameters for data with interval uncertainty. *Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskie nauki = Proceedings of the Tula State University. Technical sciences*. 2022;(9): 274–277. (In Russ.)
15. Perfil'eva K.S. Modeling of cargo turnover in road transport using the method of mixed estimation of regression equations. *Molodaya nauka Sibiri = Young Science of Siberia*. 2022;(2):248–251. (In Russ.)

Информация об авторах / Information about the authors

Сергей Иванович Носков

доктор технических наук, профессор,
профессор кафедры информационных
систем и защиты информации,
Иркутский государственный
университет путей сообщения
(Россия, г. Иркутск, ул. Чернышевского, 15)
E-mail: sergey.noskov.57@mail.ru

Sergey I. Noskov

Doctor of technical sciences, professor,
professor of the sub-department
of information systems
and information security,
Irkutsk State Transport University
(15 Chernyshevskogo street,
Irkutsk, Russia)

Егор Сергеевич Попов

магистрант кафедры информационных
систем и защиты информации,
Иркутский государственный
университет путей сообщения
(Россия, г. Иркутск, ул. Чернышевского, 15)
E-mail: eglir5732@mail.ru

Egor S. Popov

Master degree student
of the sub-department of information
systems and information security,
Irkutsk State Transport University
(15 Chernyshevskogo street,
Irkutsk, Russia)

**Авторы заявляют об отсутствии конфликта интересов /
The authors declare no conflicts of interests.**

Поступила в редакцию/Received 06.03.2024

Поступила после рецензирования/Revised 22.07.2024

Принята к публикации/Accepted 17.09.2024